数据治理中心 DataArts Studio

用户操作指南

天翼云科技有限公司

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。

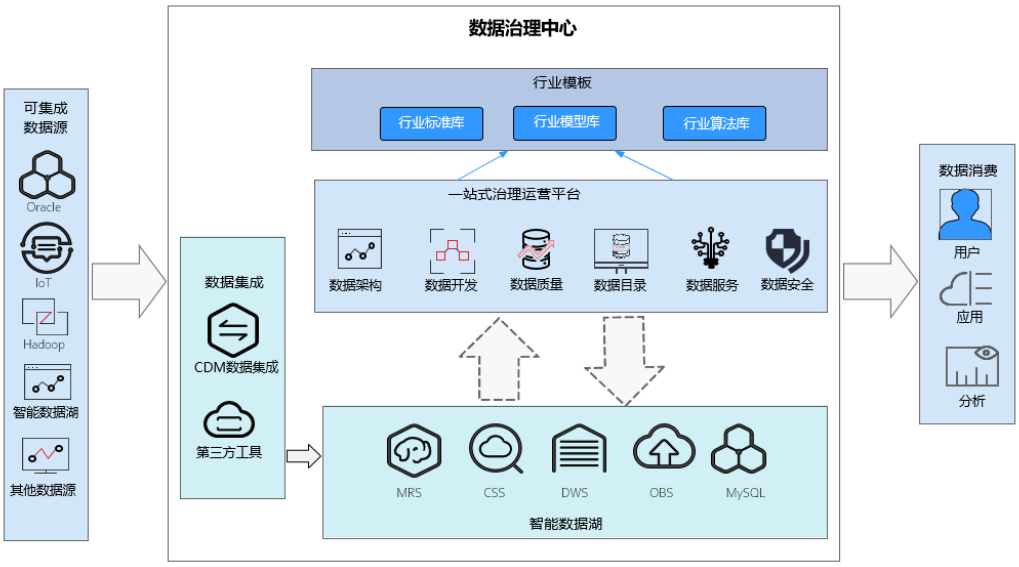
- 数据治理的挑战
  - 缺乏企业数据体系标准和数据规范定义的方法论，数据语言不统一。
  - 缺乏面向普通业务人员的高效、准确的数据搜索工具，数据找不到。
  - 缺乏技术元数据与业务元数据的关联，数据读不懂。
  - 缺乏数据的质量管控和评估手段，数据不可信。
- 数据运营的挑战
  - 数据运营效率低，业务环境的快速变化带来大量多样化的数据分析报表需求，因为缺乏高效的数据运营工具平台，数据开发周期长、效率低，不能满足业务运营决策人员的诉求。
  - 数据运营成本高，数据未服务化，导致数据拷贝多、数据口径不一致，同时数据重复开发，造成资源浪费。
- 数据创新的挑战
  - 企业内部存在大量数据孤岛，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、数据架构、数据质量监控、数据资产管理、数据服务、数据安全等功能，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，用于数据开发与数据治理等活动。

- **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。

- **数据架构**

  作为数据治理的一个核心模块，承担数据治理过程中的数据加工并业务化的功能，提供智能数据规划、自定义主题数据模型、统一数据标准、可视化数据建模、标注数据标签等功能，有利于改善数据质量，有效支撑经营决策。

- **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。

- **数据质量**

  数据全生命周期管控，数据处理全流程质量监控，异常事件实时通知。

- **数据目录**

提供企业级的元数据管理，厘清信息资产。通过数据地图，实现数据血缘和数据全景可视，提供数据智能搜索和运营监控。

- **数据服务**

  数据服务定位于标准化的数据服务平台，提供一站式数据服务开发、测试部署能力，实现数据服务敏捷响应，降低数据获取难度，提升数据消费体验和效率，最终实现数据资产的变现。

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、数据目录、数据质量、数据架构、数据服务、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主机之上，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。

数据治理

数据资源及其应用过程中相关管控活动、绩效和风险管理的集合。

数据调研

基于现有业务数据、行业现状进行数据调查、需求梳理、业务调研，输出企业业务流程以及数据主题划分。

主题设计

通过分层架构表达对数据的分类和定义，帮助厘清数据资产，明确业务领域和业务对象的关联关系。

主题域分组

基于业务场景对主题域分组。

主题域

互不重叠数据的高层面的数据分类，用于管理其下一级的业务对象。

业务对象

指企业运作和管理中不可缺少的重要人、事、物信息。

流程设计

流程设计是针对流程的一个结构化的整体框架，描述了企业流程的分类、层级以及边界、范围、输入/输出关系等，反映了企业的商业模式及业务特点。

数据标准

数据标准用于描述公司层面需共同遵守的数据含义和业务规则。其描述了公司层面对某个数据的共同理解，这些理解一旦确定下来，就应作为企业层面的标准在企业内被共同遵守。

码表

通常只包括一系列允许的值和附加文本描述，与数据标准关联用于生成值域校验质量监控。

SDI

Source Data Integration (SDI)又称贴源数据层。SDI 是源系统数据的简单落地。

DWI

Data Warehouse Integration (DWI)又称数据整合层。DWI 整合多个源系统数据，源系统进来数据会有整合、清洗，基于三范式关系建模。

DWR

Data Warehouse Report (DWR)又称数据报告层。 DWR 基于多维模型，和 DWI 层数据粒度保持一致。

DM

Data Mart (DM) 又称数据集市。DM 面向展现层，数据有多级汇总。

关系建模

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

维度建模

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

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

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

指标（数据架构）

指标是衡量目标总体特征的统计数值，是能表征企业某一业务活动中业务状况的数值指示器。指标一般由指标名称和指标数值两部分组成，指标名称及其涵义体现了指标质的规定性和量的规定性两个方面的特点，指标数值反映了指标在具体时间、地点、条件下的数量表现。

度量

度量是用于衡量业务状况的可量化的数值表现，通常为数字，如：金额、数量、周期等。指标与度量的关系：度量是一个数值型数据，其本身不带有业务含义，只有将度量放在业务语境下，方能体现出业务含义，才能成为指标。

维度

维度是用于观察和分析业务数据的视角，支撑对数据汇聚、钻取、切片分析，用于SQL 中的 GROUP BY 条件。多数维度具有层级结构，如：地理维度(其中包括国家、地区、省以及城市等级别的内容)、时间维度(其中包括年度、季度、月度等级别的内容)。

原子指标

原子指标包含属性和度量，是基于业务活动下某一业务对象所产生的业务状况的度量，以及和所有相关的属性。通过原子指标数据旨在用于支撑衍生指标的敏捷自助消费，其与多维模型中的最细数据粒度保持一致，如：零售门店数量(包含门店名称、门店等级等属性)。多维模型是由数字型度量值组成的一张事实表连接到一组包含描述属性的多张维度表，事实表与维度表通过主外键实现关联。

原子指标中的度量和属性来源于多维模型中的维度表和事实表，与多维模型所属的业务对象保持一致，与多维模型中的最细数据粒度保持一致。原子指标中仅含有唯一度量，所含其它所有与该度量、该业务对象相关的属性，旨在用于支撑指标的敏捷自助消费。

衍生指标

衍生指标是原子指标通过添加口径/修饰词、维度卷积而成，口径/修饰词、维度均来源于原子指标中的属性。例如：促销员门店覆盖率。

复合指标

复合指标由一个或多个衍生指标叠加计算而成，其中的维度、口径/修饰词均继承于衍生指标（不能脱离衍生指标维度和口径/修饰词的范围，去产生新的维度和口径/修饰词）。

数据质量规则

不符合业务实质、不满足数据应用要求的数据判断逻辑。

数据资产

由企业拥有或者控制的，能够为企业带来未来经济利益的，以物理或电子的方式记录的数据资源。在企业中并非所有的数据都构成数据资产，数据资产是能够为企业产生价值的数据资源。

数据地图

以数据搜索为核心，通过可视化方式，综合反映有关数据来源、数量、分布、标准、流向、关联关系、数据质量。让用户找到数据、读懂数据、消费数据，致力于为用户提供高效率的数据消费产品。

元数据

元数据是关于数据的组织、数据域及其关系的信息，简言之，元数据就是关于数据的数据。元数据包括元数据实体和元数据元素。元数据元素是元数据的基本单元，若干个相关的元数据元素构成了元数据实体。

元数据采集

支持创建自定义策略的采集任务，采集数据源中的技术元数据。

数据资产报告

数据资产总览与统计信息展示。

数据服务

数据服务是基于数据分发、发布的框架，将数据作为一种服务产品提供，满足客户的实时数据需求，能复用并符合企业和工业标准，兼顾数据共享和安全。

API 网关

API 网关（API Gateway）提供 API 托管服务，涵盖 API 发布、管理、运维、售卖的全生命周期管理。帮助您简单、快速、低成本、低风险地实现微服务聚合、前后端分离、系统集成，向合作伙伴、开发者开放功能和数据。

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。

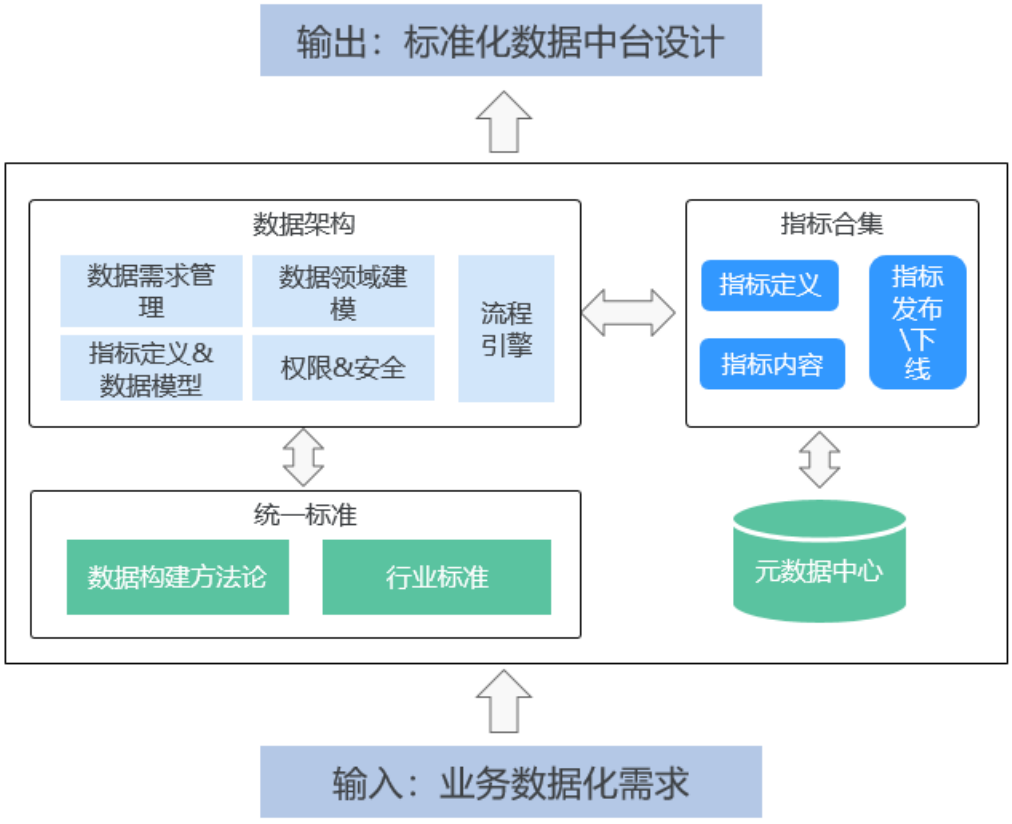
- **脏数据归档**

  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。

数据架构：数据建模可视化、自动化、智能化

DataArts Studio 数据架构践行数据治理方法论，将数据治理行为可视化，打通数据基础层到汇总层、集市层的数据处理链路，落地数据标准和数据资产，通过关系建模、维度建模实现数据标准化，通过统一指标平台建设，实现规范化指标体系，消除歧义、统一口径、统一计算逻辑，对外提供主题式数据查询与挖掘服务。

图1-2 数据架构



DataArts Studio 数据架构主要包括以下三个部分：

- **主题设计**

  构建统一的数据分类体系，用于目录化管理所有业务数据，便于数据的归类，查找，评价，使用。通过分层架构对数据分类和定义，可帮助用户厘清数据资产，明确业务领域和业务对象的关联关系。

- **数据标准**

  构建统一的数据标准体系，数据标准流程化、系统化。用户可基于国家标准或行业标准，对每一行数据、每一个字段的具体取值进行标准化，从而提升数据质量和易用性。

- **数据建模**

  构建统一的数据模型体系，通过规范定义和数据建模，自顶向下构建企业数据分层体系，沉淀企业数据公共层和主题库，便于数据的流通、共享、创造、创新，提升数据使用效率，极大的减少数据冗余，混乱，隔离，不一致以及谬误等。

  DataArts Studio 数据架构支持的数据建模方法有：

  - **关系建模**

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

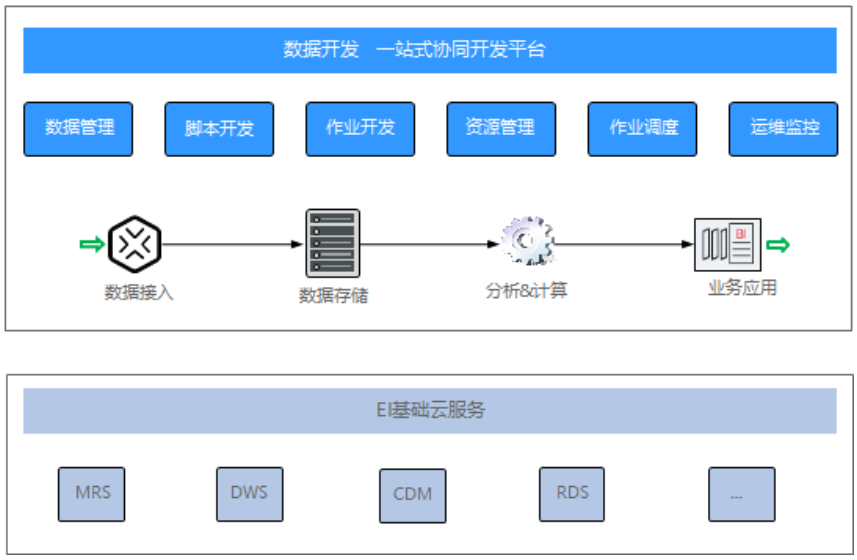
  - **维度建模**

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

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1-3 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。

数据质量：可控可检验

数据质量模块支持对业务指标和数据质量进行监控，数据质量可检验，帮助用户及时发现数据质量问题。

- **业务指标监控**

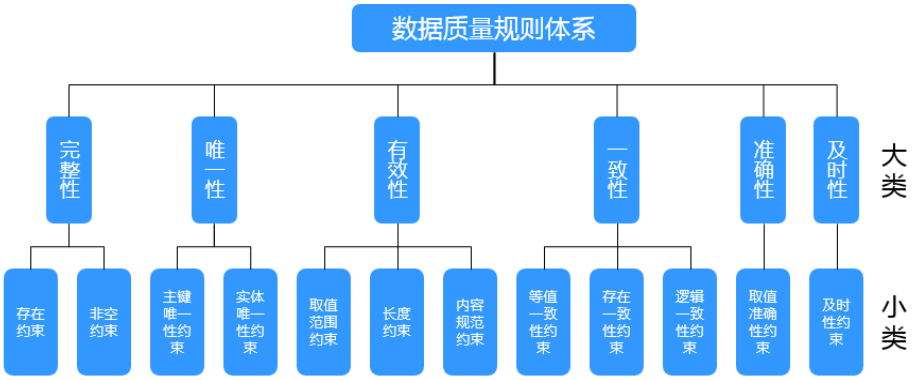
  业务指标监控是对业务指标数据进行质量管理的有效工具，可以灵活的创建业务指标、业务规则和业务场景，实时、周期性进行调度，满足业务的数据质量监控需求。

- **数据质量监控**

  数据质量监控是对数据库里的数据质量进行质量管理的工具，您可以配置数据质量检查规则，在线监控数据准确性。

  数据质量可以从完整性、有效性、及时性、一致性、准确性、唯一性六个维度进行单列、跨列、跨行和跨表的分析，也支持数据的标准化，能够根据数据标准自动生成标准化的质量规则，支持周期性的监控。

图1-4 数据质量规则体系



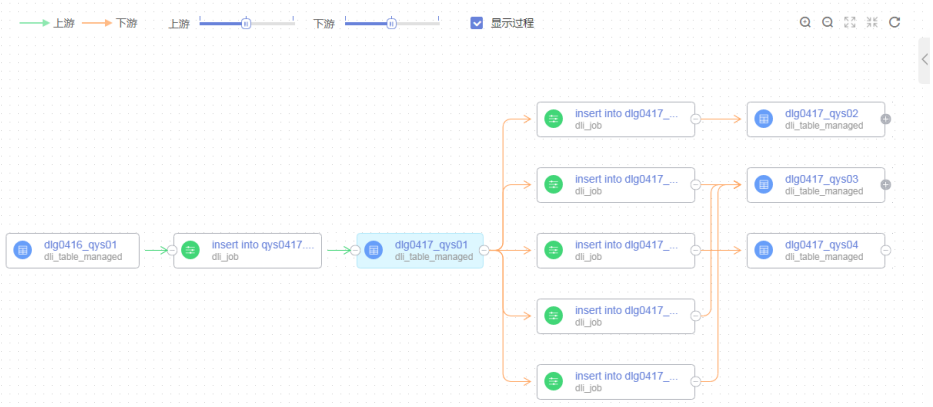
数据资产管理：360 度全链路数据资产可视化

DataArts Studio 提供企业级的元数据管理，厘清信息资产。数据资产管理可视，支持钻取、溯源等。通过数据地图，实现数据资产的数据血缘和数据全景可视，提供数据智能搜索和运营监控。

- **元数据管理**

  元数据管理模块是数据湖治理的基石，支持创建自定义策略的采集任务，可采集数据源中的技术元数据。支持自定义业务元模型，批量导入业务元数据，关联业务和技术元数据、全链路的血缘管理和应用。

图1-5 全链路数据血缘



- **数据地图**

  数据地图围绕数据搜索，服务于数据分析、数据开发、数据挖掘、数据运营等数据表的使用者和拥有者，提供方便快捷的数据搜索服务，拥有功能强大的血缘信息及影响分析。

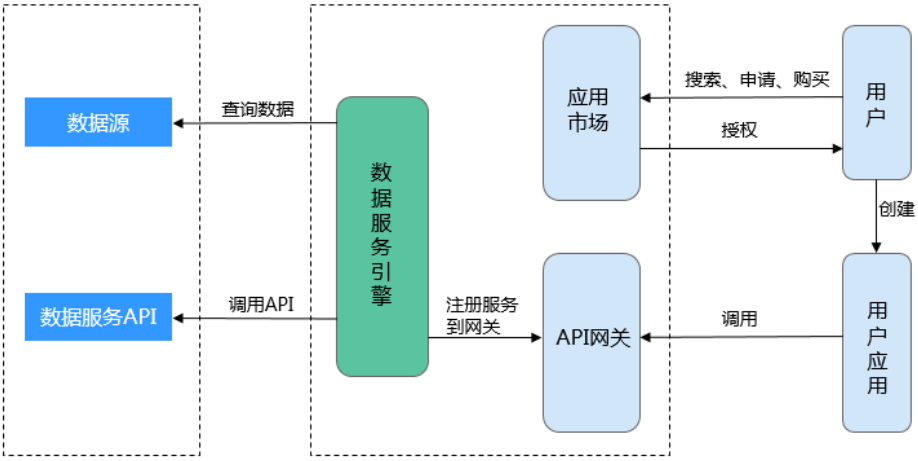
  - 在数据地图中，可通过关键词搜索数据资产，支持模糊搜索，快速检索，定位数据。
  - 使用数据地图根据表名直接查看表详情，快速查阅明细信息，掌握使用规则。获得数据详细信息后，可添加额外描述。
  - 通过数据地图的血缘分析可以查看每个数据表的来源、去向，并查看每个表及字段的加工逻辑。
  - 对数据资产，可以从业务角度定义分类或标签。

数据服务：提升访问查询检索效率

DataArts Studio 数据服务旨在为企业搭建统一的数据服务总线，帮助企业统一管理对内对外的 API 服务，支撑业务主题/画像/指标的访问、查询和检索，提升数据消费体验和效率，最终实现数据资产的变现。数据服务为您提供快速将数据表生成数据 API 的能力，同时支持您将现有的 API 快速注册到数据服务平台以统一管理和发布。

数据服务采用 Serverless 架构，您只需关注 API 本身的查询逻辑，无需关心运行环境等基础设施，数据服务会为您准备好计算资源，并支持弹性扩展，零运维成本。

图1-6 数据服务架构图



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、标准数据架构、连接并萃取数据价值、全流程数据质量监控、统一数据资产管理、数据开发服务等，帮助企业构建完整的数据中台解决方案。

全链路数据治理管控

数据全生命周期管控，提供数据架构定义及可视化的模型设计，智能化的帮助用户生成数据处理代码，数据处理全流程质量监控，异常事件实时通知。

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。

可复用行业知识库

提供垂直行业可复用的领域知识库，涵盖行业数据标准、行业领域模型、行业数据主题库、行业算法库和行业指标库等，支持智慧政务、智慧税务、智慧园区等行业，帮助企业快速定制数据运营端到端解决方案。

统一数据资产管理

全局资产视图、快速查看、智能管理、数据溯源和数据开放共享，从业务视角管理和查看数据，定义业务架构、业务分类和业务术语，统一管理资产访问权限。

数据运营全场景可视

数据治理运营过程可视，托拉拽配置，无需编码；处理结果可视，更直观，便于交互和探索；数据资产管理可视，支持钻取、溯源等。

全方位的安全保障

统一的安全认证，租户隔离，数据的分级分类管理，数据的全生命周期管理，保证数据的隐私合规、可审计、可回溯。

基于角色的访问控制，用户通过角色与权限进行关联，并支持细粒度权限策略，可满足不同的授权需求。

1.5 应用场景

一站式的数据运营治理平台

从数据采集->数据架构->质量监控->数据清洗->数据建模->数据联接->数据整合->数据消费->智能分析，一站式数据智能运营平台，帮助企业快速构建数据运营能力。

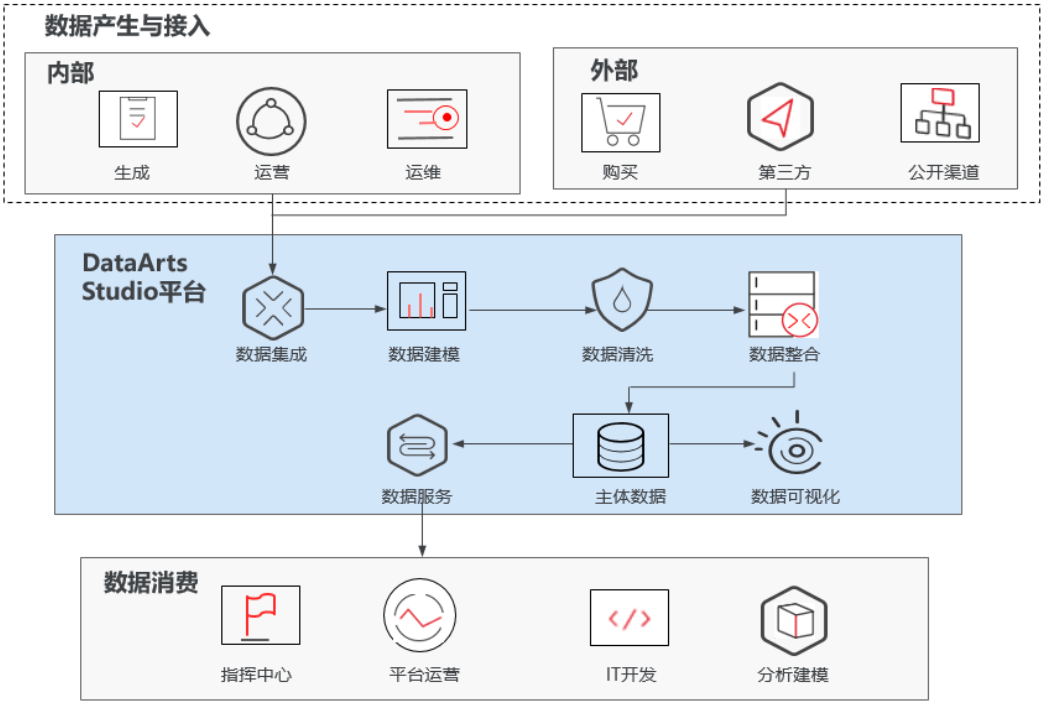
优势

- 多种云服务作业编排
- 全链路数据治理管控
- 丰富数据引擎支持

  支持对接数据湖与数据库云服务，也支持对接企业传统数据仓库，比如 Oracle 等。
- 简单易用

  图形化编排，即开即用，轻松上手。

图1-7 一站式数据运营治理平台



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。

基于行业领域知识库快速构建数据中台

通过应用在企业业务领域积累的丰富的行业领域模型和算法，帮助企业构建数据中台，快速提升数据运营能力。

优势

- 多行业支持

  覆盖政务/税务/城市/交通/园区等各行业。

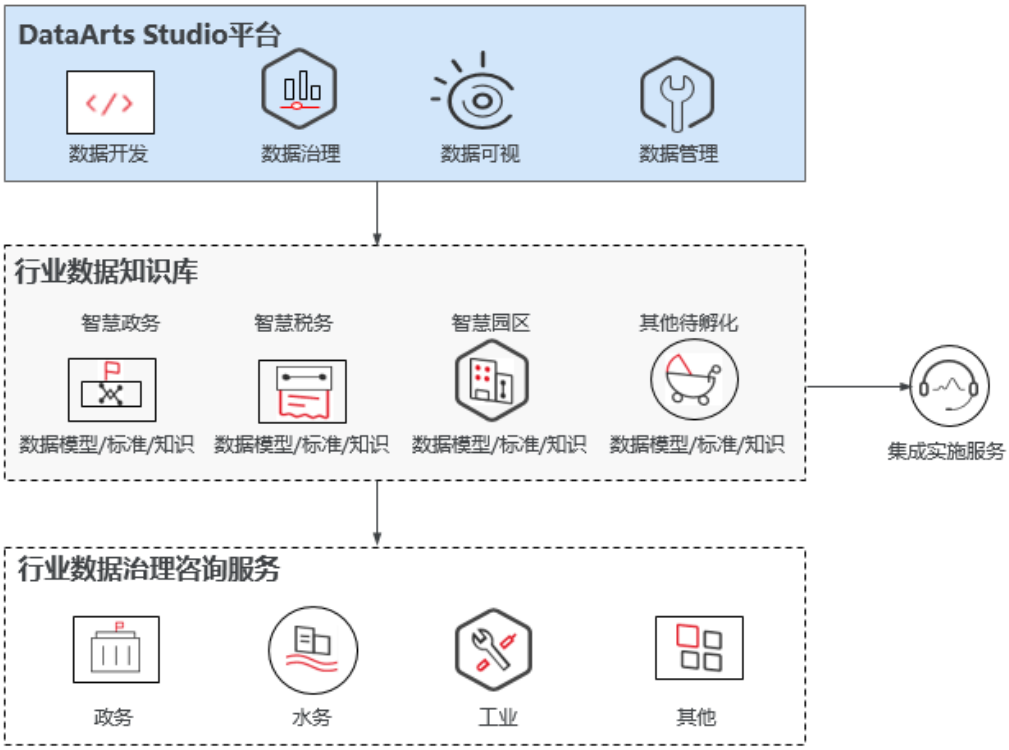
- 标准规范支持

  支持分层结构的行业数据标准。
- 领域模型丰富

  支持包含人员/组织/事件/时空/车辆/资产/设备/资源等八大类数据以及相互之间关系的行业领域模型。
- 快速应用行业库

  支持快速应用的行业主题库、行业算法库、行业指标库。

图1-8 数据中台



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，1.7 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 1.7 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	系统角色

系统角色名称	描述	类别
	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。	

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

云帐号、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			

权限点	管理员	开发者	运维者	访客
	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
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	Y	Y	N	N

权限点	管理员	开发者	运维者	访客
务				
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

数据架构

权限点	管理员	开发者	运维者	访客
查看统计信息	Y	Y	Y	Y
导入主题设计	Y	Y	Y	N
导出主题设计	Y	Y	Y	N
查看主题设计	Y	Y	Y	Y
创建主题设计	Y	Y	N	N
查询所有表	Y	Y	Y	Y
查看业务表	Y	Y	Y	Y
创建、编辑业务表	Y	Y	N	N
删除业务表	Y	Y	N	N
业务表关联主题设计	Y	Y	N	N
关系建模导入	Y	Y	Y	N
关系建模导出	Y	Y	Y	N
逆向数据库	Y	Y	N	N
创建标签	Y	Y	N	N
关联、移除标签	Y	Y	N	N
关联数据标准	Y	Y	N	N

权限点	管理员	开发者	运维者	访客
查看 DDL 模板	Y	Y	Y	Y
更新、恢复默认 DDL 模板	Y	Y	N	N
预览 SQL	Y	Y	Y	Y
查看原子指标	Y	Y	Y	Y
创建、编辑原子指标	Y	Y	N	N
删除原子指标	Y	Y	N	N
查看维度列表	Y	Y	Y	Y
创建、编辑维度	Y	Y	N	N
删除维度	Y	Y	N	N
查看衍生指标列表	Y	Y	Y	Y
创建、编辑衍生指标	Y	Y	N	N
删除衍生指标	Y	Y	N	N
查看复合指标列表	Y	Y	Y	Y
创建、编辑复合指标	Y	Y	N	N
删除复合指标	Y	Y	N	N
查看时间限定	Y	Y	Y	Y
创建、编辑时间限定	Y	Y	N	N
删除时间限定	Y	Y	N	N
编辑、删除系统默认时间限定	Y	N	N	N
查看维度表	Y	Y	Y	Y
删除维度表	Y	Y	N	N
查看事实表	Y	Y	Y	Y
创建、编辑事实表	Y	Y	N	N
删除事实表	Y	Y	N	N
查看汇总表	Y	Y	Y	Y

权限点	管理员	开发者	运维者	访客
创建、编辑汇总表	Y	Y	N	N
删除汇总表	Y	Y	N	N
查看目录	Y	Y	Y	Y
创建目录	Y	Y	N	N
编辑目录	Y	Y	N	N
删除目录	Y	Y	N	N
查看码表	Y	Y	Y	Y
创建、编辑码表	Y	Y	N	N
删除码表	Y	Y	N	N
导入导出码表	Y	Y	Y	N
添加码表数据	Y	Y	N	N
查看数据标准	Y	Y	Y	Y
创建、编辑数据标准	Y	Y	N	N
删除数据标准	Y	Y	N	N
查看数据标准模板	Y	Y	Y	Y
编辑数据标准模板	Y	N	N	N
查看模型	Y	Y	Y	Y
创建、编辑模型	Y	Y	N	N
删除模型	Y	Y	N	N
查看 OBS 文件夹、详情	Y	Y	Y	Y
创建、更新、删除 OBS 目录	Y	Y	N	N
查看审核人	Y	Y	Y	Y
添加审核人	Y	N	N	N
删除审核人	Y	N	N	N
发布、下线各项定义	Y	Y	Y	N
同步表模型	Y	N	N	N

权限点	管理员	开发者	运维者	访客
试运行衍生指标	Y	Y	N	N
运行衍生指标	Y	Y	Y	N
编辑调度	Y	Y	Y	N
启动、停止调度	Y	Y	Y	N
查看调度列表、实例信息、运行日志	Y	Y	Y	N
字段类型的增、删、改、查	Y	N	N	N

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
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

权限点	管理员	开发者	运维者	访客
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
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
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

权限点	管理员	开发者	运维者	访客
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
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

权限点	管理员	开发者	运维者	访客
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
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

权限点	管理员	开发者	运维者	访客
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

数据质量

权限点	管理员	开发者	运维者	访客
获取目录	Y	Y	Y	Y
创建目录	Y	Y	N	N
修改目录	Y	Y	N	N
删除目录	Y	Y	N	N
查看规则列表	Y	Y	Y	Y
查看规则详情	Y	Y	Y	Y
创建规则	Y	Y	N	N
编辑规则	Y	Y	N	N
删除规则	Y	Y	N	N
运行规则	Y	Y	Y	N
停止规则运行实例	Y	Y	Y	N
启动调度	Y	Y	Y	N
停止调度	Y	Y	Y	N
查看运行结果	Y	Y	Y	N
查看总览数据	Y	Y	Y	Y
处理问题	Y	Y	Y	N
查看运行历史	Y	Y	Y	Y

权限点	管理员	开发者	运维者	访客
查看运行详情	Y	Y	Y	Y
查看日志	Y	Y	Y	Y
校验规则模板	Y	N	N	N
新建规则模板	Y	N	N	N
删除规则模板	Y	N	N	N
查看规则模板列表	Y	Y	Y	Y
编辑规则模板	Y	N	N	N
查看规则模板详情	Y	Y	Y	Y
业务指标				
查看目录	Y	Y	Y	Y
创建目录	Y	Y	N	N
修改目录	Y	Y	N	N
删除目录	Y	Y	N	N
查看指标	Y	Y	Y	Y
创建指标	Y	Y	N	N
修改指标	Y	Y	N	N
删除指标	Y	Y	N	N
查看规则	Y	Y	Y	Y
创建规则	Y	Y	N	N
修改规则	Y	Y	N	N
删除规则	Y	Y	N	N
查看业务场景	Y	Y	Y	Y
创建业务场景	Y	Y	N	N
修改业务场景	Y	Y	N	N
删除业务场景	Y	Y	N	N
运行业务场景	Y	Y	Y	N
查看日志	Y	Y	Y	Y
启动调度	Y	Y	Y	N
停止调度	Y	Y	Y	N

权限点	管理员	开发者	运维者	访客
查看运行结果	Y	Y	Y	Y
查看总览数据	Y	Y	Y	Y
处理问题	Y	Y	Y	N

数据目录

权限点	管理员	开发者	运维者	访客
创建采集任务	Y	Y	N	N
更新采集任务	Y	Y	N	N
删除采集任务	Y	Y	N	N
查看采集任务	Y	Y	Y	Y
运行、启动调度、停止调度采集任务	Y	Y	Y	N
查询任务实例运行状态	Y	Y	Y	Y
停止任务实例运行状态	Y	Y	Y	N
重跑任务实例	Y	Y	Y	N
创建任务目录	Y	Y	N	N
修改任务目录	Y	Y	N	N
删除任务目录	Y	Y	N	N
获取任务目录列表	Y	Y	Y	Y
技术资产高级搜索	Y	Y	Y	Y
获取保存的搜索条件	Y	Y	Y	Y
保存搜索条件	Y	Y	N	N
删除搜索条件	Y	Y	N	N
业务资产搜索	Y	Y	Y	Y
获取业务资产目录树	Y	Y	Y	Y
获取技术资产查询	Y	Y	Y	Y

权限点	管理员	开发者	运维者	访客
条件				
获取实体详情	Y	Y	Y	Y
删除实体指定的关联分类	Y	Y	N	N
批量删除实体	Y	Y	N	N
创建类别	Y	Y	N	N
创建标签	Y	Y	N	N
获取类别详情	Y	Y	Y	Y
更新类别	Y	Y	N	N
删除类别	Y	Y	N	N
获取标签详情	Y	Y	Y	Y
更新标签	Y	Y	N	N
删除实体关联的标签	Y	Y	N	N
实体关联标签	Y	Y	N	N
获取数据开发模块算子的血缘信息	Y	Y	Y	Y
资产统计接口	Y	Y	Y	Y
技术资产历史统计	Y	Y	Y	Y
业务资产统计	Y	Y	Y	Y
技术资产统计	Y	Y	Y	Y
技术资产业务资产总量统计	Y	Y	Y	Y
导入采集任务	Y	Y	Y	N
导出采集任务	Y	Y	Y	N
导入分类标签	Y	Y	Y	N
导出分类标签	Y	Y	Y	N
创建标签	Y	Y	N	N
获取标签列表	Y	Y	Y	Y
给实体添加标签	Y	Y	N	N

权限点	管理员	开发者	运维者	访客
删除实体的标签	Y	Y	N	N
创建分类	Y	Y	N	N
删除分类	Y	Y	N	N
更新分类	Y	Y	N	N
获取分类列表	Y	Y	Y	Y
添加分类至资产	Y	Y	N	N
获取列的完整性	Y	Y	Y	Y
获取列的合法性	Y	Y	Y	Y
创建数据权限规则	Y	N	N	N
删除数据权限规则	Y	N	N	N
查看数据权限规则	Y	Y	Y	Y
修改数据权限规则	Y	N	N	N
查看数据权限规则列表	Y	Y	Y	Y
设置数据权限生效状态	Y	N	N	N

数据服务

权限点	管理员	开发者	运维者	访客
创建数据服务专享版集群	具备以下权限之一的用户才能进行该操作： • DAYU Administrator 并且拥有 VPCEndpoint Administrator 权限 • Tenant Administrator 并且拥有 VPCEndpoint Administrator 权限			
删除数据服务专享版集群	具备以下权限之一的用户才能进行该操作： • DAYU Administrator 并且拥有 VPCEndpoint Administrator 权限 • Tenant Administrator 并且拥有 VPCEndpoint Administrator 权限			
全量导出 API	具备以下权限之一的用户才能进行该操作： • DAYU Administrator 权限 • Tenant Administrator 权限			

权限点	管理员	开发者	运维者	访客
修改 API 配额	具备以下权限之一的用户才能进行该操作： • DAYU Administrator 权限 • Tenant Administrator 权限			
查询数据服务专享版集群	Y	Y	Y	Y
查看 API	Y	Y	Y	Y
创建 API	Y	Y	N	N
注册 API	Y	Y	N	N
删除 API	Y	Y	N	N
复制 API	Y	Y	N	N
导入 API	Y	Y	Y	N
导出 API	Y	Y	Y	N
编辑 API	Y	Y	N	N
调试 API	Y	Y	Y	N
发布 API	Y	Y	Y	N
下线 API	Y	Y	Y	N
添加授权	Y	Y	Y	N
查看 API 授权信息	Y	Y	Y	Y
查看 API 调用信息	Y	Y	Y	Y
查看 API DashBoard	Y	Y	Y	Y
查看 API 分组	Y	Y	Y	Y
创建 API 分组	Y	Y	N	N
删除 API 分组	Y	Y	N	N
编辑 API 分组	Y	Y	N	N
查看分组内 API	Y	Y	Y	Y
查看域名列表	Y	Y	Y	Y
绑定域名	Y	Y	Y	N
解绑域名	Y	Y	Y	N
上传证书	Y	Y	Y	N

权限点	管理员	开发者	运维者	访客
删除证书	Y	Y	Y	N
查看应用列表	Y	Y	Y	Y
创建应用	Y	Y	N	N
删除应用	Y	Y	N	N
编辑应用	Y	Y	N	N
查看应用授权	Y	Y	Y	Y
解绑应用	Y	Y	Y	N
续约应用	Y	Y	Y	N
查看服务市场	Y	Y	Y	Y
申请权限	Y	Y	Y	N
监控 API 开发服务	Y	Y	Y	Y
监控 API 调用服务	Y	Y	Y	Y
查看审核信息列表	Y	Y	Y	Y
审核上线申请	Y	Y	Y	N
审核下线申请	Y	Y	Y	N
审核授权申请	Y	Y	Y	N
审核取消授权申请	Y	Y	Y	N
创建审核人	Y	N	N	N
删除审核人	Y	N	N	N
查看流控策略列表	Y	Y	Y	Y
查看流控策略信息	Y	Y	Y	Y
创建流控策略	Y	Y	N	N
绑定流控策略	Y	Y	Y	N
解绑流控策略	Y	Y	Y	N
编辑流控策略	Y	Y	N	N
删除流控策略	Y	Y	N	N
获取数据源类型	Y	Y	N	N
获取数据源连接列表	Y	Y	N	N

权限点	管理员	开发者	运维者	访客
获取数据库列表	Y	Y	N	N
获取数据库中表的列表	Y	Y	N	N
获取表的字段	Y	Y	N	N
获取队列列表	Y	Y	N	N
导入 Excel	Y	Y	Y	N
导出 Excel	Y	Y	Y	N
解析 SQL	Y	Y	N	N
API 取消授权	Y	Y	Y	N
撤销权限申请	Y	Y	Y	N
审核信息发送	Y	Y	Y	N
获取所有用户	Y	Y	Y	N
获取审核人	Y	Y	Y	N
查看 API 分组信息	Y	Y	Y	Y
查看未授权应用	Y	Y	Y	Y

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。

弹性云主机服务

DataArts Studio 使用弹性云主机（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。

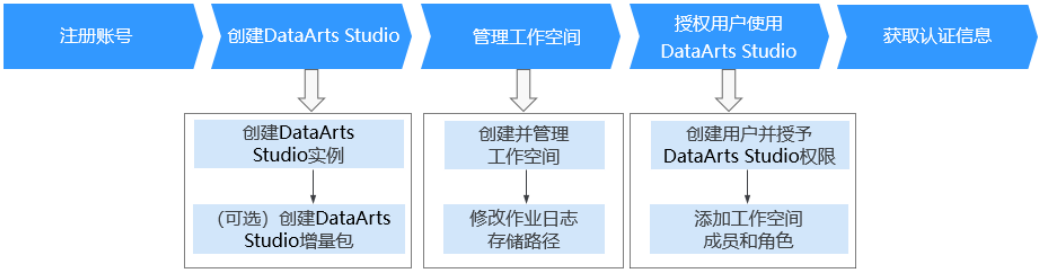
2 准备工作

2.1 准备工作简介

您需要完成创建云帐号、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

只有云帐号、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。
- 数据服务专享集群增量包

  数据服务当前提供共享版与专享版两种服务方式。其中共享版数据服务即开即用；专享版数据服务需要在创建 DataArts Studio 基础包实例后，创建专享版集群增量包。

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

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。
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

参数	说明
	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。
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主机的访问规则的集合，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主机提供访问策略。 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。

数据服务专享集群

步骤 1 单击已开通实例卡片上的“创建增量包”。

步骤 2 进入创建 DataArts Studio 增量包页面，参见表 2-3 进行配置。

表2-3 创建数据服务专享版实例参数说明

参数项	说明
增量包类型	选择数据服务专享集群增量包。
工作空间	选择需要使用数据服务专享集群增量包的工作空间。例如需要在 DataArts Studio 实例的工作空间 A 中使用数据服务专享版，则此处工作空间应选择为 A。集群创建成功后，即可通过在工作空间 A 查看到创建好的数据服务专享集群。
可用区	第一次创建 DataArts Studio 实例或批增量包时，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

参数项	说明
	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。
集群名称	-
集群描述	可以自定义对当前数据服务专享版集群的描述。
版本	当前数据服务专享版的集群版本。
集群规格	不同实例规格，对 API 请求的并发支持能力不同。
公网入口	开启“公网入口”，即允许外部服务通过公网地址，调用专享版实例创建的 API。
带宽大小	可配置公网带宽范围。
虚拟私有云	VPC 即虚拟私有云，是通过逻辑方式进行网络隔离，提供安全、隔离的网络环境。 在相同虚拟私有云中的云服务资源（如 ECS），可以使用数据服务专享版实例的私有地址调用 API。 建议将专享版实例和您的其他关联业务配置一个相同的虚拟私有云，确保网络安全的同时，方便网络配置。 说明 目前 DLM 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	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。 建议将专享版实例和您的其他关联业务配置相同的虚拟私有云下相同的子网，确保网络安全的同时，方便网络配置。 说明 目前 DLM 实例创建完成后不支持切换子网，请谨慎选择所属子网。
安全组	安全组用于设置端口访问规则，定义哪些端口允许被外部访问，以及允许访问外部哪些地址与端口。 例如，后端服务部署在外部网络，则需要设置相应的安全组规则，允许访问后端服务的地址及其监听端口。 说明 1. 如果开启公网入口，安全组入方向需要放开 80（HTTP）和 443（HTTPS）端口的访问权限。 2. 目前 DLM 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。
企业项目	DataArts Studio 专享版集群关联的企业项目。企业项目管理是一种按企业项目管理云资源的方式，具体请参见《企业管理用户指南》。

步骤 3 单击“立即创建”，确认规格后提交。

----结束

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、数据架构、数据开发、数据目录、数据质量、数据服务、数据安全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 2.2.1 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 2-4 所示，配置完成后，单击“确定”完成工作空间的创建。

表2-4 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。
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。
DLM 专享版 API 配额	该参数对应值分别表示已使用配额/已分配配额/总使用配额/总分配配额/总配额。 初始工作空间具有 10 个 API 的试用额度。已分配配额可以修改，但不能小于已使用配额，不能大于总配额-总分配配额+已分配配

参数名	说明
	额。

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，在您的云帐号中，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、高效的其他云帐号或者云服务，这些帐号或者云服务可以根据权限进行代运维。

如果云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。使用云帐号登录 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 2.4.1 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。

1. 登录控制台，在用户名下拉列表中选择“我的凭证”。
2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图 2-3 所示。

图2-3 单击新增访问密钥

3. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id 和 Secret Access Key）。

说明

- 每个用户仅允许新增两个访问密钥。
- 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台界面获取。请在生成后妥善保管。

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2-4 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 2-5 所示，实例 ID 为 **6b88…2688**，工作空间 ID 为 **1dd3bc…d93f0**。

图2-5 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。

您可以从企业管理员处获取。

获取数据目录的 guid

每个业务资产、技术资产或指标资产都具备 guid，guid 是资产的唯一标识符。在调用数据目录接口时，部分 URL 中需要填入 guid。

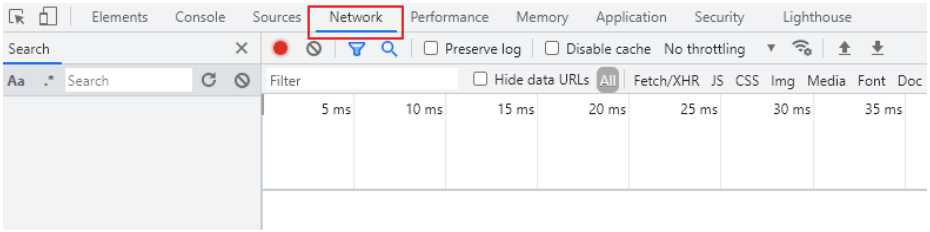
数据目录 guid 获取步骤如下：

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图2-6 选择数据目录

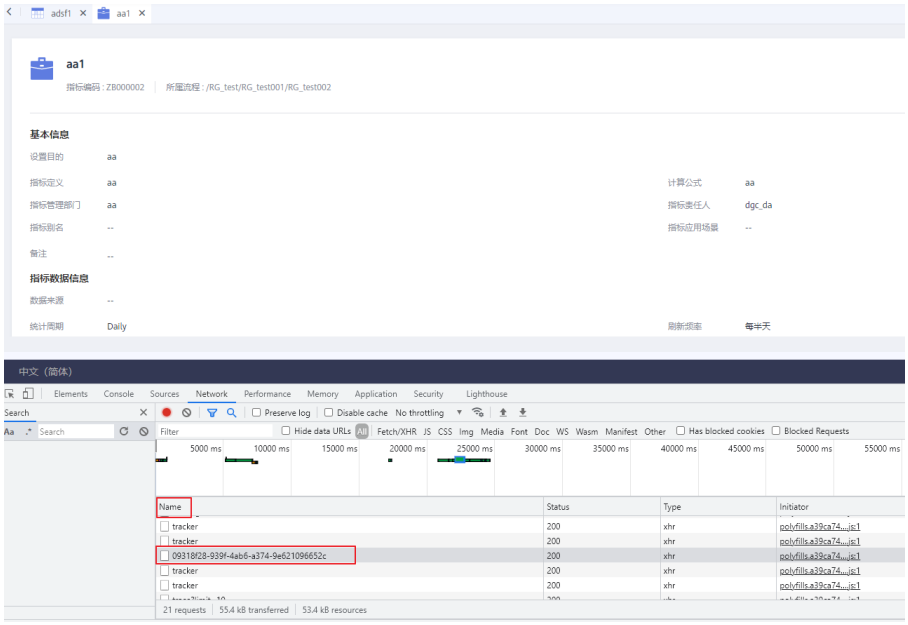
2. 按下 F12，打开开发者调试工具，然后选择 Network 功能。

图2-7 选择 Network



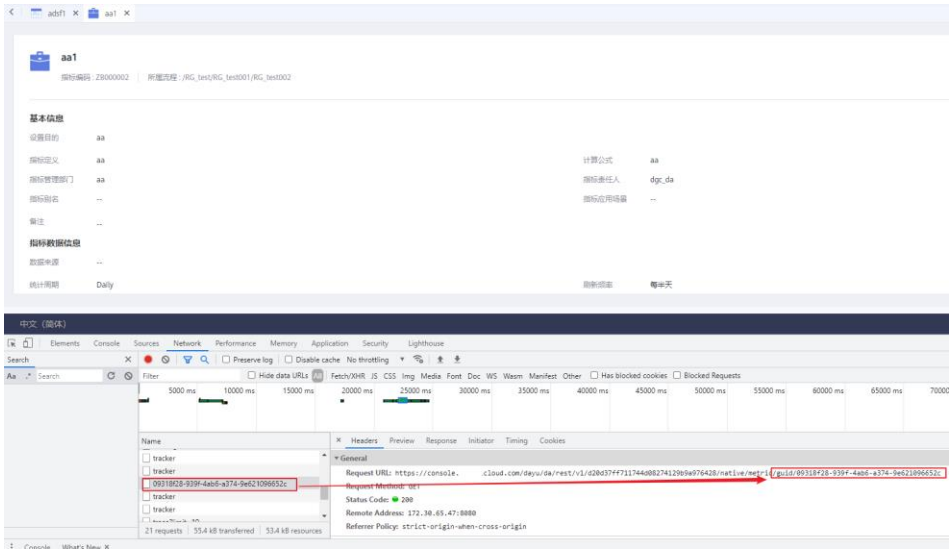
3. 在数据目录的“总览”界面或“数据目录”界面，选择需要查询 guid 的资产，并点击资产名进入资产详情页。
   a. 在 Network 请求中，寻找 Name 形如“09318f28-939f-4ab6-a374-9e621096652c”的长字符串。

图2-8 寻找长字符串



b. 点击该字符串，弹出具体请求的弹窗。在 **Request URL** 中，可以确认该字符串即为该资产的 guid。

图2-9 获取 guid



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 3.2.1 DataArts Studio 支持的数据源。

准备好数据湖之后，您可以通过 3.2.2 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 3.3.3 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	数据架构	数据开发	数据目录[1]	数据质量[2]	数据服务
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）	√	×	√	×	×	×
MapReduce 服务（MRS Ranger）	√	×	×	×	×	×
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）	√	×	√	×	×	×

注释：

[1] 数据目录：数据目录组件除了上表中列出的数据源外，还支持采集以下数据源的元数据：

1. 关系型数据库如 MySQL/PostgreSQL/达梦数据库 DM 等（可使用 RDS 类型连接，采集其元数据）
2. 云搜索服务 CSS
3. 图引擎服务 GES
4. 对象存储服务 OBS

[2] 数据质量：数据质量组件中的质量作业和对账作业功能不支持对接 MRS 集群存算分离的场景。

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。
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	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

数据源类型	简介
Kafka）	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。
MapReduce 服务（MRS Ranger）	Ranger 提供一个集中式安全管理框架，提供统一授权和统一审计能力。它可以对整个 Hadoop 生态中如 HDFS、Hive、HBase、Kafka、Storm 等进行细粒度的数据访问控制。用户可以利用 Ranger 提供的前端 WebUI 控制台通过配置相关策略来控制用户对这些组件的访问权限 。
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。
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 Ranger、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。
DLI	请参见表 3-10。
ORACLE	请参见表 3-11
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 3-14。

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。
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

参数	是否必选	说明
		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。 说明 为保证数据架构、数据质量、数据目录、数据服务等组件能够使用此 MRS 连接，此处连接方式推荐配置为“通过代理连接”。
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

参数	是否必选	说明
		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。
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

参数	是否必选	说明
		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。
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

参数	是否必选	说明
		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。
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

参数	是否必选	说明
		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。
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

参数	是否必选	说明
		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。 **说明** **为保证数据架构、数据质量、数据目录、数据服务等组件能够使用此 MRS 连接，此处连接方式推荐配置为“通过代理连接”。**
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

参数	是否必选	说明
		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.html，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。

参数	是否必选	说明
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。

表3-10 DLI 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。

表3-11 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

参数	是否必选	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。
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3-12 MRS Ranger 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 Ranger 所属的 MRS 安全集群。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。 说明 DataArts Studio 目前仅支持与安全模式集群的 MRS Ranger 创建连接。
用户名	是	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

参数	是否必选	说明
		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3-13 MRS Presto 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 Presto 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。

参数	是否必选	说明
		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。
描述	否	可自定义填写相关连接的描述。

表3-14 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《弹性云主机用户指南》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	主机的登陆用户名。
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”

参数	是否必选	说明
		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数据服务、元数据分类、元数据标签、元数据采集任务和管理中心数据连接。

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 3.2.2 创建数据连接创建数据连接，3.7.1.4 标签管理对元数据进行分类和标签的添加，3.7.4.2 任务管理完成采集任务的创建和 3.8.3.4 发布 API 中发布的 API。

约束条件

- 名称相同的采集任务不支持被重复迁移。
- 名称相同的分类和标签不支持被重复迁移。
- 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。

图3-7 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3-8 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，请联系客服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。

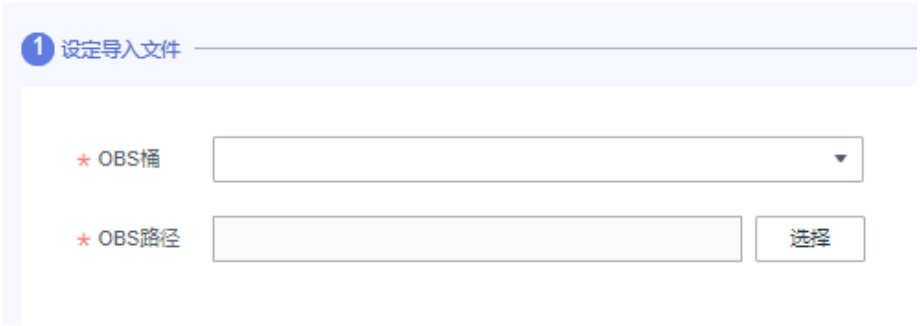
图3-9 选择管理中心

2. 在管理中心页面，单击“资源迁移”，进入资源迁移页面。

图3-10 资源迁移

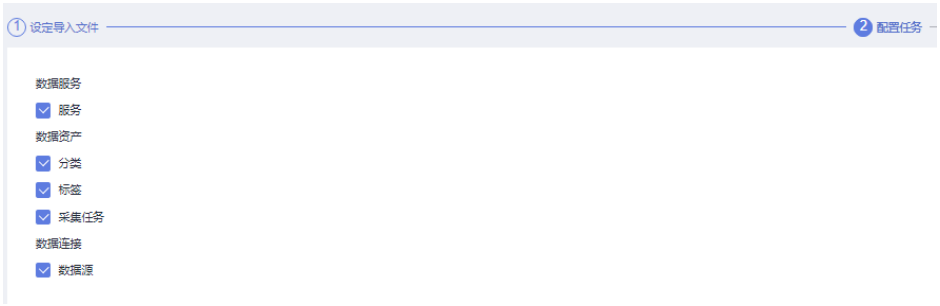
3. 单击“新建导入”，配置待导入资源的 OBS 存储路径。如无 OBS 服务，需要从本地路径选择待上传的资源包。

图3-11 配置待导入的资源存储路径



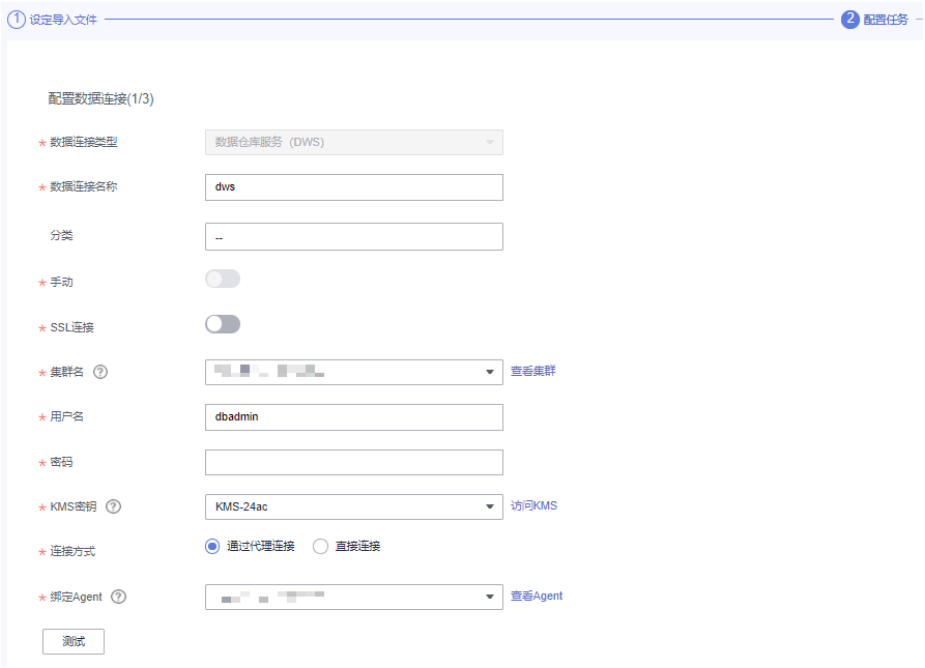
4. 单击“下一步”，勾选导入的资源类型。

图3-12 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3-13 配置数据连接



6. 单击“下一步”，等待导入完成。

图3-14 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，请联系客服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 Ranger、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3-15 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3-16 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 3-15 配置相关参数。

图3-17 创建数据连接

图3-18 MRS Hive 连接配置参数

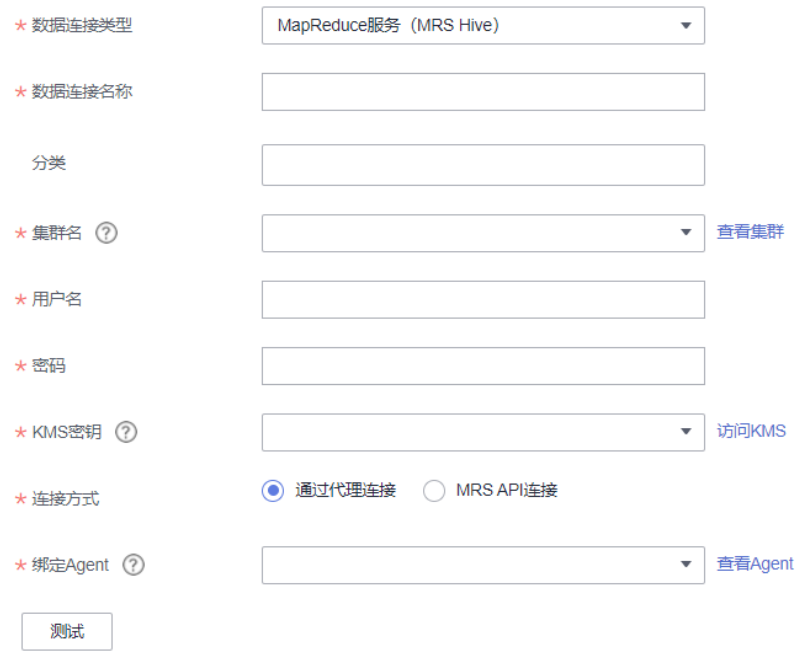


表3-15 MRS Hive 数据连接

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

参数	是否必选	说明
		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。 **说明** **为保证数据架构、数据质量、数据目录、数据服务等组件能够使用此 MRS 连接，此处连接方式推荐配置为“通过代理连接”。**
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**

参数	是否必选	说明
		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 Ranger、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3-19 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3-20 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 3-16 配置相关参数。

图3-21 创建数据连接

图3-22 DWS 连接配置参数

表3-16 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。

参数	是否必选	说明
		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。
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。

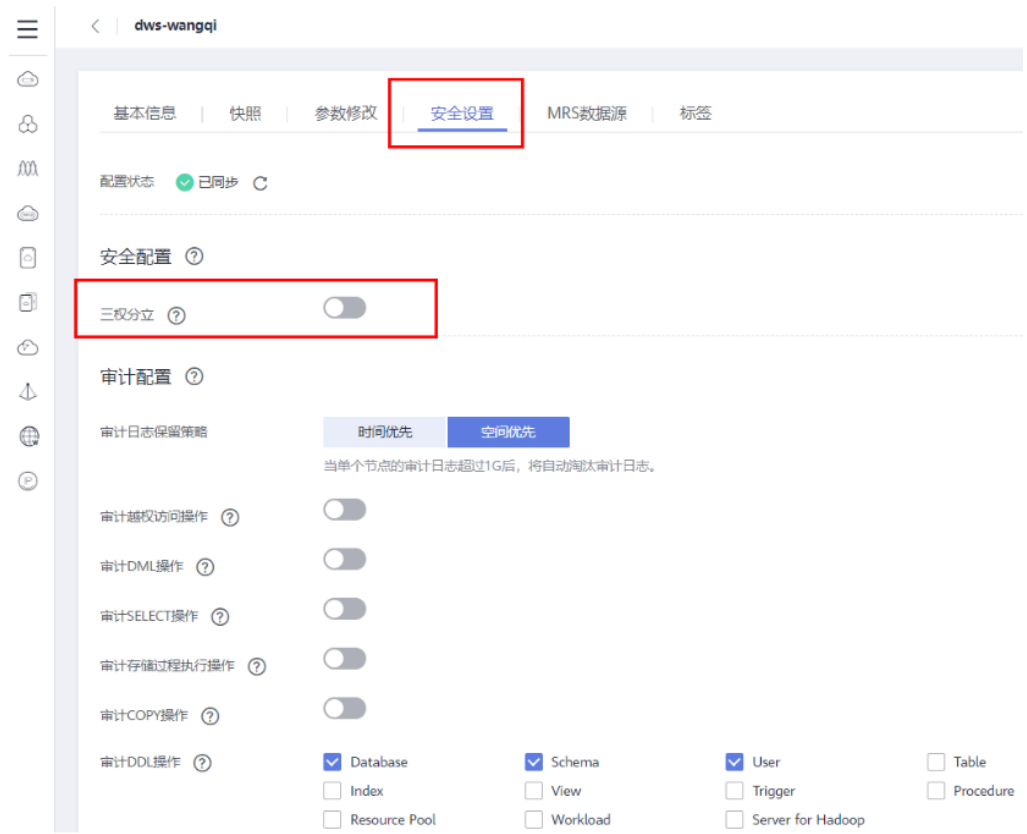
4. 单击“测试”，测试数据连接的连通性。如果无法连通，数据连接将无法创建。
5. 测试通过后，单击“确定”，创建数据连接。

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3-23 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 Ranger、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3-24 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3-25 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 3-17 配置相关参数。

图3-26 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3-27 RDS 连接配置参数

表3-17 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.html，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。

参数	是否必选	说明
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。
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：

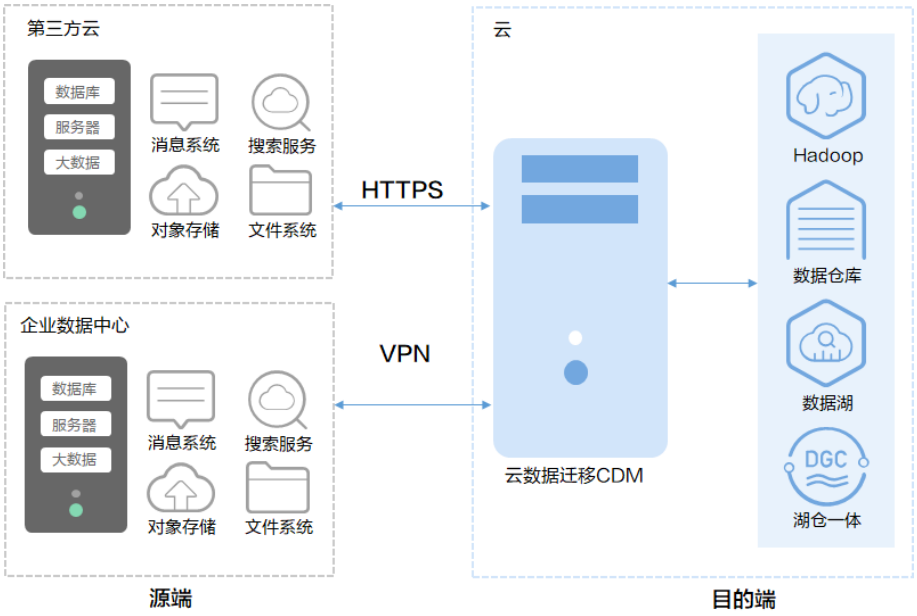
- 登录 CDM 控制台，单击“集群管理”，进入到 CDM 主界面。
- 登录 DataArts Studio 控制台。选择对应工作空间的“数据集成”模块，进入 CDM 主界面。

图3-28 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3-29 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**

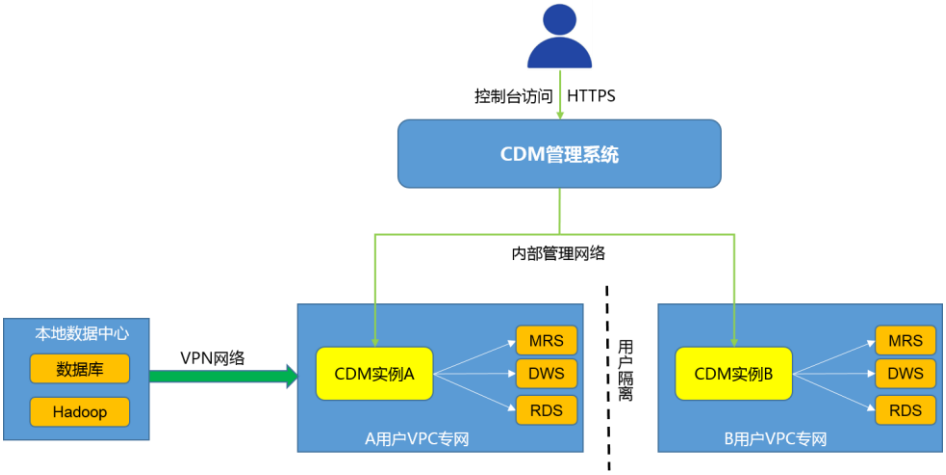
  支持将迁移过程中处理失败的、被清洗过滤掉的、不符合字段转换或者不符合清洗规则的数据单独归档到脏数据日志中，便于用户查看。并支持设置脏数据比例阈值，来决定任务是否成功。

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3-30 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。
3. 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。
4. 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）。
5. 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。
6. 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。
7. 仅支持数据库字符集为 GBK 和 UTF-8。
8. 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 3-18 所示。

表3-18 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，	不支持 DWS 物理机纳管模式。
	数据湖探索（DLI）		-

数据源分类	源端数据源	对应的目的端数据源	说明
		PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
Hadoop	MRS HDFS MRS HBase MRS 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） • 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
	FusionInsight HDFS FusionInsight HBase FusionInsight 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）	• 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。 • 不支持 Ranger 场景。 • 不支持 ZK 开启 SSL 场景。 • FusionInsight HDFS 建议使用

数据源分类	源端数据源	对应的目的端数据源	说明
			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 Apache 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） • 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： – 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）	对象存储服务之间的迁移，推荐使用对象存储迁移服务

数据源分类	源端数据源	对应的目的端数据源	说明
		• Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	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		
	SAP HANA	• 数据仓库：数据湖探索（DLI）	SAP HANA 数据源

数据源分类	源端数据源	对应的目的端数据源	说明
		• Hadoop：MRS Hive	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		
	表格存储服	• 数据仓库：数据仓库服务	

数据源分类	源端数据源	对应的目的端数据源	说明
	务（CloudTable）	（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。
搜索	Elasticsearch	• 数据仓库：数据仓库服务	Elasticsearch 仅支持

数据源分类	源端数据源	对应的目的端数据源	说明
		（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）	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 3-19 所示。

表3-19 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
	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

数据源分类	数据源	读取	写入	说明
				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	
	Oracle	支持	不支持	
	SAP HANA	支持	不支持	• 仅支持 2.00.050.00.1592305219 版

数据源分类	数据源	读取	写入	说明
				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。
	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3-20 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	CHAR	固定长度的字符串是以	‘a’或	CHAR	CHAR

类别	类型	简要释义	存储格式示例	Hive	DWS
串	(M)	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。	‘aaaaa’		
	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
	INT UNSIGNED	与 INTEGER UNSIGNED 相同	-	BIGINT	INTEGER

类别	类型	简要释义	存储格式示例	Hive	DWS
	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
	BOOL	MySQL 的 bool 实际上就是 tinyint(1)	-128、127	SMALLINT	BYTEA

类别	类型	简要释义	存储格式示例	Hive	DWS
	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
日期时间	DATE	以 YYYY-MM-DD 格式的日期，在 1000-01-01 和 9999-12-31 之	1999-10-01	DATE	TIMESTAMP

类别	类型	简要释义	存储格式示例	Hive	DWS
		间。 例如，1973 年 12 月 30 日将被存储为 1973-12-30。			
	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
	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

类别	类型	简要释义	存储格式示例	Hive	DWS
	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
	TINY TEXT	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
	LONG TEXT	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
	TINY BLOB	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
	LONG BLOB	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
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）

类别	类型	简要释义	存储格式示例	Hive	DWS
	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3-21 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
	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	DATE	TIMESTAMP

类别	类型	简要释义	Hive	DWS
		哪一天、小时、分钟、秒。		
	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
	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	不支持	不支持

类别	类型	简要释义	Hive	DWS
		的 rowid 的表。		

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3-22 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
	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	DOUB	FLOA	NUMB

类别	类型	简要释义	Hive	DWS	Oracle
		值。	LE	T4	ER
	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
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	不支持	不支持	不支持

类别	类型	简要释义	Hive	DWS	Oracle
		数据，其中 n 是从 1~8000 的值。	（String）	（String）	
	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
	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 集群与其他弹性云主机共享弹性 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3-31 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 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3-32 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3-33 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 3.3.6.8 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3-34 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3-35 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。

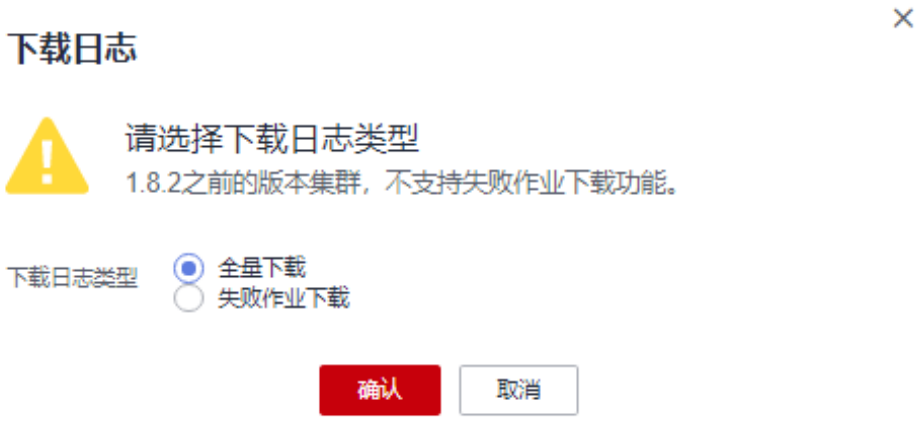
图3-36 集群列表

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3-37 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，云帐号下的其他 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，云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。

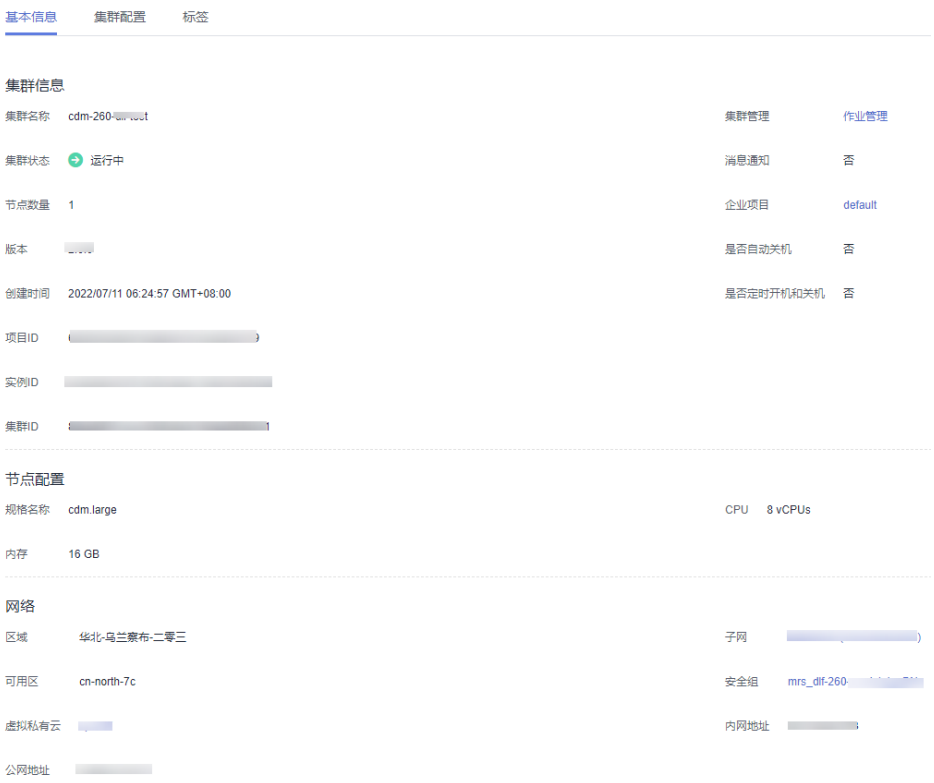
图3-38 集群列表

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，可查看集群的基本信息。

图3-39 CDM 集群的配置信息



----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3-40 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。

图3-41 修改集群配置

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3-42 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

管理 CDM 集群标签

步骤 1　单击左侧导航上的“集群管理”，进入集群管理界面。

图3-43 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“标签”页签。

图3-44 修改集群配置

步骤 3　单击“添加标签”，通过添加标签为 CDM 集群设置资源标识。

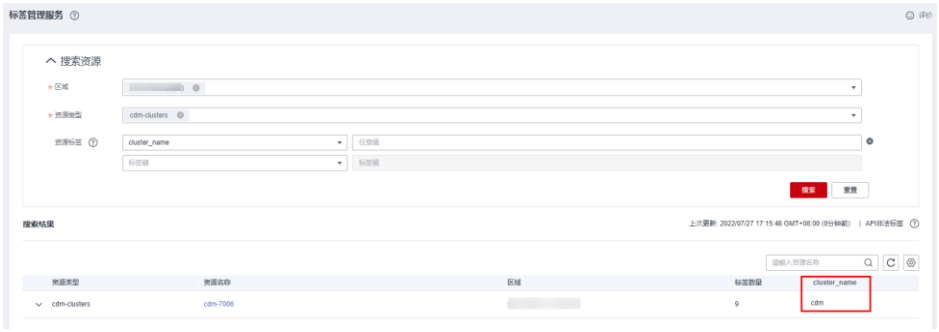
图3-45 添加标签

说明

- 一个集群最多可添加 10 个标签。
- 标签键（key）的最大长度为 36 个字符，标签值（value）的最大长度为 43 个字符。

步骤 4 （可选）在标签列表中，单击标签操作列“编辑”或者“删除”，修改/删除 CDM 集群标签。

步骤 5 在标签管理服务中，选择资源搜索条件，单击“搜索”即可筛选出所设置标签的资源。



----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 3-23 所示。

表3-23 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

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_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主机内部采集的数据。 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主机内部采集的数据。 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主机内部采集的数据。 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
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	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）
		连接数。 单位：Count/个。			
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主机内部采集的数据。 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。

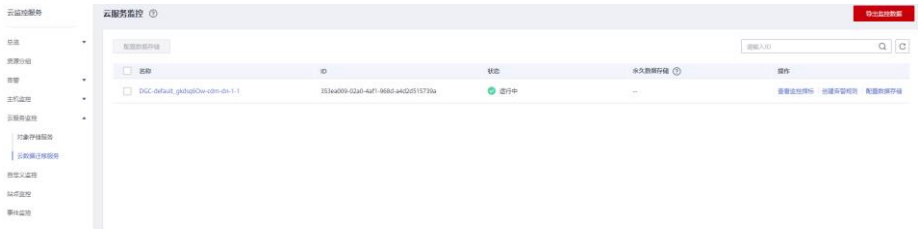
设置 CDM 集群的告警规则包括设置告警规则名称、监控对象、监控指标、告警阈值、监控周期和是否发送通知等参数。本节介绍了设置 CDM 集群告警规则的具体方法。

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3-46 “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）。

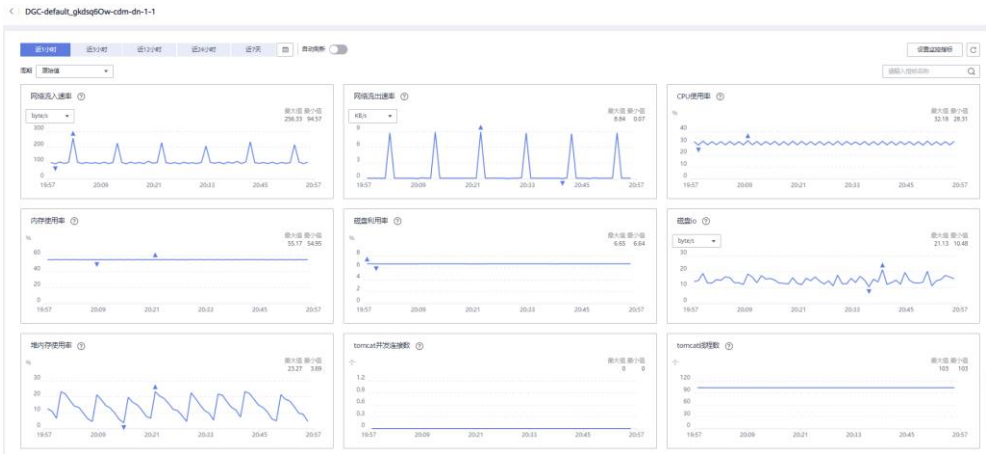
  对于新创建的集群，需要等待一段时间，才能查看上报的监控数据和监控视图。

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3-47 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 3.3.3 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3-48 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 3-24 中查找对应参数。

表3-24 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 3.3.5.5 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 3.3.5.7 配置 MySQL 数据库连接。
Oracle	连接 Oracle 数据库时，具体参数请参见 3.3.5.8 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 3.3.5.6 配置分库连接。
对象存储服务（OBS）	连接 OBS 时，具体参数请参见 3.3.5.13 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 3.3.5.12 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 3.3.5.11 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 3.3.5.10 配置 Hive 连接。
表格存储服务（CloudTable）	连接 CloudTable 时，具体参数请参见 3.3.5.17 配置 CloudTable 连接。
• FTP • SFTP	连接 FTP 或 SFTP 服务器时，具体参数请参见 3.3.5.14 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 3.3.5.19 配置 MongoDB 连接。
文档数据库服务（DDS）	连接 DDS 时，具体参数请参见 3.3.5.16 配置

连接器类型	参数说明
	DDS 连接。
• Redis • 分布式缓存服务（DCS）	连接 Redis 或 DCS 时，具体参数请参见 3.3.5.15 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 3.3.5.21 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 3.3.5.23 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 3.3.5.9 配置 DLI 连接。
DMS Kafka	连接 DMS 的 Kafka 队列时，具体参数请参见 3.3.5.22 配置 DMS Kafka 连接。
Cassandra	连接 Cassandra 时，具体参数请参见 3.3.5.20 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 3-25 下载对应的驱动。
- 已参见 3.3.5.14 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 3-25 获取建议版本的 JDK8 .jar 格式驱动即可。

表3-25 获取驱动

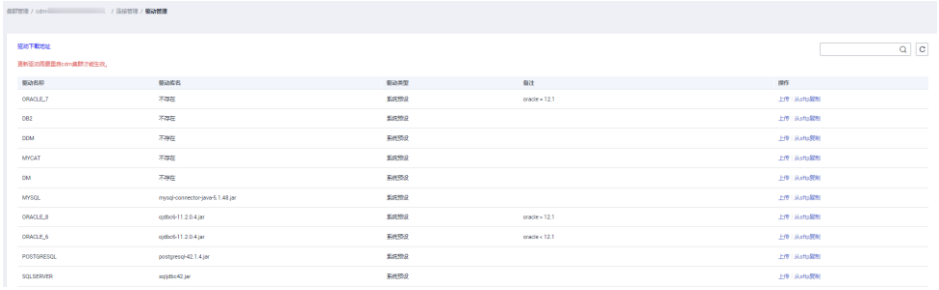
关系数据库类型	驱动名称	获取地址	建议版本
• 云数据库 MySQL • MySQL	MYSQL	https://downloads.mysql.com/archives/c-j/	5.1.48，获取 mysql-connector-java-5.1.48.jar

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3-49 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。

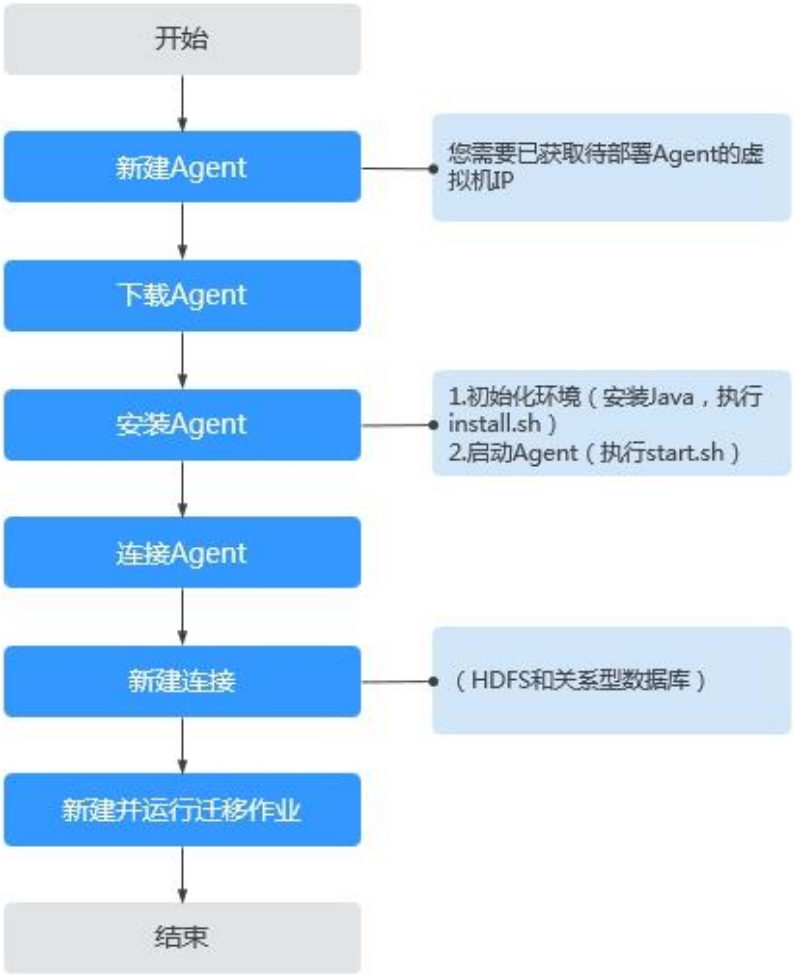
----结束

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 3-50 所示。

图3-50 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3-51 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 3-52 所示，下载 Agent。

图3-52 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 3.3.5.2 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 3-53 所示，连接 Agent。

图3-53 连接 Agent

步骤 2 Agent 连接成功后，即可在创建连接中选择 Agent。

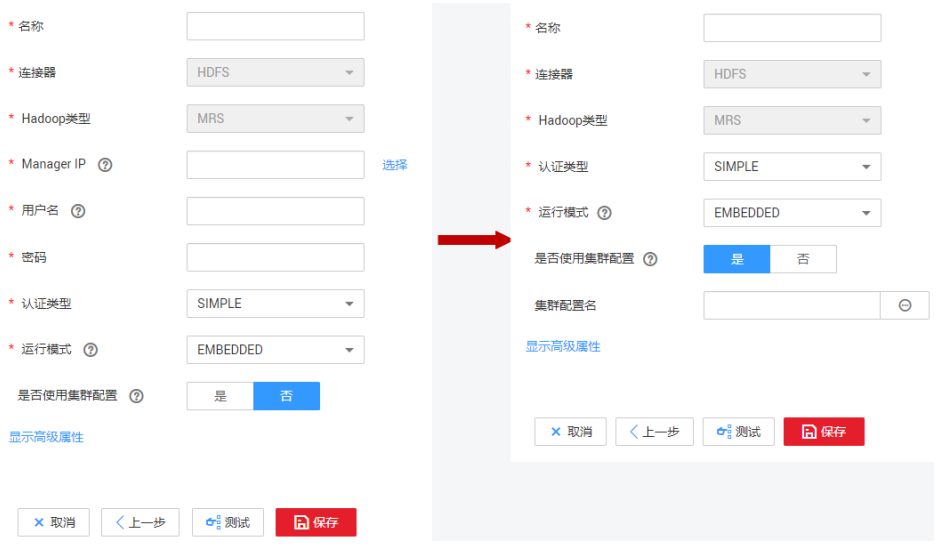
----结束

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 3-54 所示。

图3-54 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 3-26 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 3-26 获取相应 Hadoop 集群配置文件和 Keytab 文件。

表3-26 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	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
	档。	
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。
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	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。

Hadoop 类型连接	集群配置文件获取方式	Keytab 文件获取方式
	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）	

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

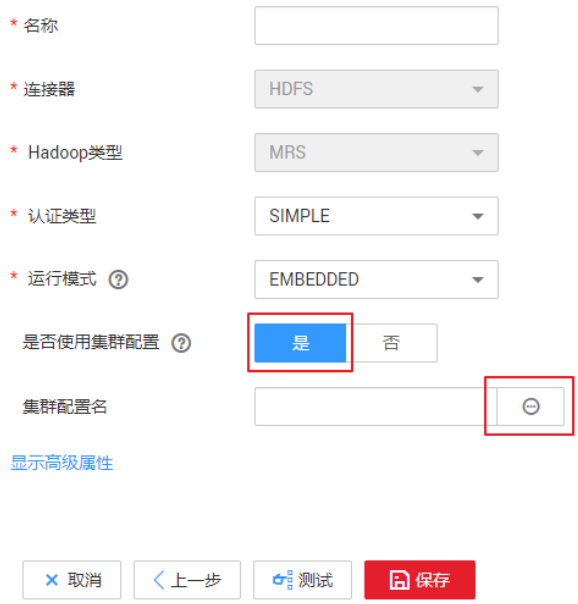
1. 进入 CDM 主界面，进入集群管理界面。选择 CDM 集群后的“作业管理 > 连接管理 > 集群配置管理”。
2. 在集群配置管理界面，选择“新建集群配置”，配置参数填写如下：

图3-55 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3-56 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 3.3.5.2 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 3-27 所示。

表3-27 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
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：

参数名	说明	取值样例
		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 3.3.5.3 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 3-28 所示。

表3-28 分库连接参数

参数名	说明	取值样例

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 3.3.5.3 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 MySQL 数据库连接

连接 MySQL 数据库连接时，相关参数如表 3-29 所示。

表3-29 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 3.3.5.3 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。	-

参数名	说明	取值样例
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
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

3.3.5.8 配置 Oracle 数据库连接

连接 Oracle 数据库时，连接参数如表 3-30 所示。

表3-30 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

参数名	说明	取值样例
	• SID：通过 SID 连接 Oracle 数据库。	
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
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 3.3.5.3 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。 • 阿里云 AnalyticDB(ADB)数据库对 useCursorFetch 参数有兼容问题，在配置与 ADB 连接时需要关闭 useCursorFetch 开关，即添加连接属性“useCursorFetch”，对应属性值设置为“false”。如不进行此设置，会出现日期转换出错的情况。	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 DLI 连接

连接数据湖探索（DLI）服务时，相关参数如表 3-31 所示。

表3-31 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。 您需要先创建当前账号的访问密钥，并获得对应的 AK 和 SK。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图 3-57 所示。 图3-57 单击新增访问密钥  3. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id 和 Secret Access Key）。 说明 • 每个用户仅允许新增两个访问密钥。 • 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台界面获取。请在生成后妥善保管。	-
密钥(SK)		-
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。	-

3.3.5.10 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 3-32 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3-32 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
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

参数名	说明	取值样例
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
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	EMBEDDED

参数名	说明	取值样例
	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。	
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。	-

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 3-33 所示。

表3-33 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
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。	-

参数名	说明	取值样例
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。 **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。	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 3-34 所示。

表3-34 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	-

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
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。 **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。	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。

3.3.5.11 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 3-35 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3-35 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
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	cdm

参数名	说明	取值样例
	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 MRS 组件的库、表、列进行操作，还需要参考 MRS 文档添加对应组件的库、表、列操作权限。 • 如果 CDM 集群为 2.9.0 之前的版本，或 MRS 集群为 3.1.0 之前的版本，则所创建的用户需要具备 Manager_administrator 或 System_administrator 权限，才能在 CDM 创建连接。 • 仅具备 Manager_tenant 或 Manager_auditor 权限，无法创建连接。	
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
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。 **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	STANDALONE

参数名	说明	取值样例
	导致迁移失败。	
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。	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 3-36 所示。

表3-36 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。	-
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	STANDALONE

参数名	说明	取值样例
	式，只能使用 STANDALONE 模式。 **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。	
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。	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 3-37 所示。

表3-37 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
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	-

参数名	说明	取值样例
	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。 **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。	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。

3.3.5.12 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 3-38 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3-38 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
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	cdm

参数名	说明	取值样例
	建连接；如果需要对 MRS 组件的库、表、列进行操作，还需要参考 MRS 文档添加对应组件的库、表、列操作权限。 • 如果 CDM 集群为 2.9.0 之前的版本，或 MRS 集群为 3.1.0 之前的版本，则所创建的用户需要具备 Manager_administrator 或 System_administrator 权限，才能在 CDM 创建连接。 • 仅具备 Manager_tenant 或 Manager_auditor 权限，无法创建连接。	
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
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	hdfs_01

参数名	说明	取值样例
	数有效。此参数用于选择用户已经创建好的集群配置。	

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 3-39 所示。

表3-39 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
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	STANDALONE

参数名	说明	取值样例
	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。	
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。	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 3-40 所示。

表3-40 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	-

参数名	说明	取值样例
	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
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。	hdfs_01

3.3.5.13 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 3-41 所示。

表3-41 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

参数名	说明	取值样例
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 可从企业管理员处获取。 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。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图 3-58 所示。 图3-58 单击新增访问密钥  3. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id 和 Secret Access Key）。 说明 • 每个用户仅允许新增两个访问密钥。 • 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台界面获取。请在生成后妥善保管。	-
密钥(SK)		-

3.3.5.14 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 3-42 所示。

表3-42 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。	-

3.3.5.15 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 3-43 所示。

表3-43 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

3.3.5.16 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 3-44 所示。

表3-44 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。	-

3.3.5.17 配置 CloudTable 连接

连接 CloudTable 时，相关参数如表 3-45 所示。

表3-45 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
IAM 统一身	如果所需连接的 CloudTable 集群在创建时开启了	否

参数名	说明	取值样例
份认证	“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。	
用户名	登录 CloudTable 集群的用户名。	admin
AK	登录 CloudTable 集群的访问标识。 您需要先创建当前账号的访问密钥，并获得对应的 AK 和 SK。	-
SK	登录 CloudTable 集群的密钥。 您需要先创建当前账号的访问密钥，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。	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。

3.3.5.18 配置 CloudTable OpenTSDB 连接

连接 CloudTable OpenTSDB 时，相关参数如表 3-46 所示。

表3-46 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
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	-

参数名	说明	取值样例
	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。 您需要先创建当前账号的访问密钥，并获得对应的 AK 和 SK。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图 3-59 所示。 图3-59 单击新增访问密钥  3. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id 和 Secret Access Key）。 说明 • 每个用户仅允许新增两个访问密钥。 • 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台界面获取。请在生成后妥善保管。	-
密钥(SK)		-

3.3.5.19 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 3-47 所示。

表3-47 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。	-

3.3.5.20 配置 Cassandra 连接

表3-48 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
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

3.3.5.21 配置 Kafka 连接

MRS Kafka

连接 MRS 上的 Kafka 数据源时，相关参数如表 3-49 所示。

表3-49 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。	-
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 MRS 组件的库、表、列进行操作，还需要参考 MRS 文档添加对应组件的库、表、列操作权限。 • 如果 CDM 集群为 2.9.0 之前的版本，或 MRS 集群为 3.1.0 之前的版本，则所创建的用户需要具备 Manager_administrator 或 System_administrator 权限，才能在 CDM 创建连接。 • 仅具备 Manager_tenant 或 Manager_auditor 权限，无法创建连接。	-
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 3-50 所示。

表3-50 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。

3.3.5.22 配置 DMS Kafka 连接

连接 DMS 的 Kafka 队列时，相关参数如表 3-51 所示。

表3-51 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。	-

3.3.5.23 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 3-52 所示。

表3-52 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 3-53 所示。

表3-53 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

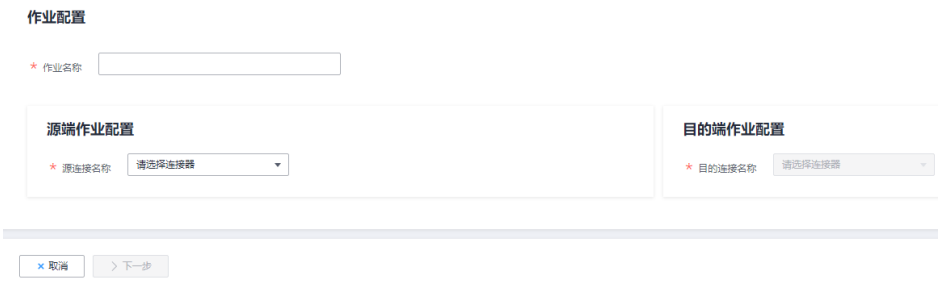
- 已 3.3.5.1 新建连接。
- CDM 集群与待迁移数据源可以正常通信。

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3-60 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 3-54 和表 3-55 选择。

表3-54 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 3.3.6.3.1 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 3.3.6.3.2 配置 HDFS 源端参数。
• 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 3.3.6.3.3 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 3.3.6.3.4 配置 Hive 源端参数。
DLI	支持从 DLI 导出数据。	参见 3.3.6.3.5 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 3.3.6.3.6 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 3.3.6.3.7 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 3.3.6.3.8 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	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	

源端类型	说明	参数配置
• Microsoft SQL Server • SAP HANA • MYCAT • 分库	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 3.3.6.3.12 配置 MongoDB/DDS 源端参数。
Redis	支持从开源 Redis 导出数据。	参见 3.3.6.3.13 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 3.3.6.3.14 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 3.3.6.3.15 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 3-55 所示。

表3-55 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 3.3.6.4.1 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 3.3.6.4.2 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 3.3.6.4.3 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 3.3.6.4.4 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 3.3.6.4.10 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 3.3.6.4.5 配置常见关系数据库目的端参数。

目的端类型	说明	参数配置
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 3.3.6.4.7 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 3.3.6.4.8 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 3.3.6.4.9 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3-61 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：
- 有主键可以使用主键作为分布列。

1. 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。
2. 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”。

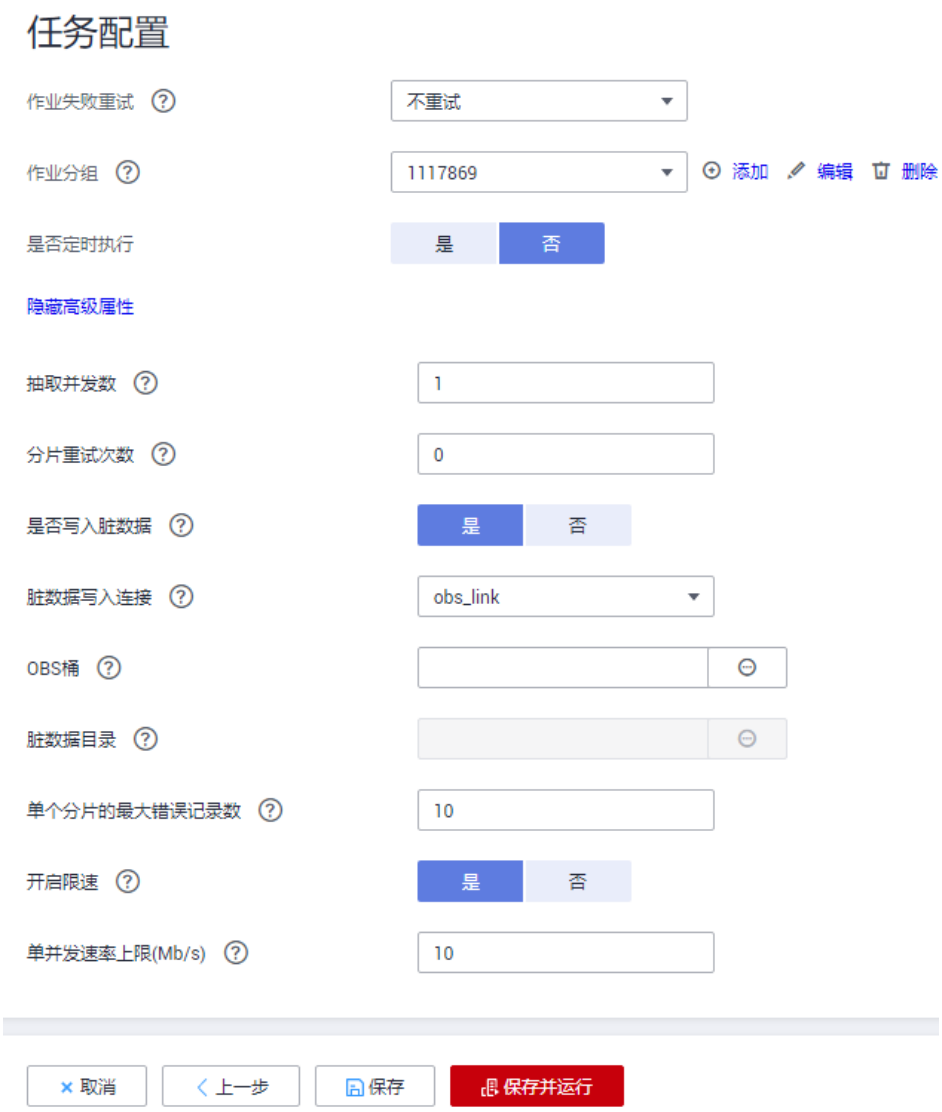
图3-62 新建转换器

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3-63 任务参数



各参数说明如表 3-56 所示。

表3-56 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	不重试

参数	说明	取值样例
	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。	
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 3.3.6.5 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
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

参数	说明	取值样例
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。	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。	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 3.3.6.1 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 3-64 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3-64 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 3.3.5.1 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 3-57 所示。

表3-57 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

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 "${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}"	-
Redis	键过滤字符	填写键过滤字符后，将迁移符合条件的键。例如：a*，迁移所有:*	-
DDS MongoDB	数据库名称	待迁移的数据库名称，源连接中配置的用户需要拥有读取该数据库的权限。	mongodb
	查询筛选	创建用于匹配文档的筛选器。例如：{HTTPStatusCode:{$gt:"400",$lt:"500"	-

源端数据库类型	源端参数	参数说明	取值样例
		},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 3-58 所示。

表3-58 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 3.3.6.4.5 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 3.3.6.4.4 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 3.3.6.4.3 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 3.3.6.4.2 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 3.3.6.4.1 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 3.3.6.4.8 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	mongodb

源端数据库类型	源端参数	参数说明	取值样例
		的用户需要拥有读取该数据库的权限。	
	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3-65 任务参数

各参数说明如表 3-59 所示。

表3-59 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 3.3.5.13 配置 OBS 连接时，源端作业参数如表 3-60 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3-60 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
	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	/0521/Lists.txt

参数类型	参数名	说明	取值样例
		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。	
	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
	编码类型	文件编码类型，例如：“UTF-8”或	GBK

参数类型	参数名	说明	取值样例
		“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。	
	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。	10

参数类型	参数名	说明	取值样例
		单位：秒。	
	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	AES-256-GCM

参数类型	参数名	说明	取值样例
		密，在源端为解密。	
	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 3.3.5.12 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 3-61 所示。

表3-61 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
	源目录或文件	“列表文件”选择为“否”时，才有该参数。	/user/cdm/

参数类型	参数名	说明	取值样例
		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。	
	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	\n

参数类型	参数名	说明	取值样例
		参数。	
	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
	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	2019-07-01 00:00:00

参数类型	参数名	说明	取值样例
		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**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近 90 天内的文件。	
	终止时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${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 3.3.5.11 配置 HBase 连接或 3.3.5.17 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 3-62 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3-62 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	可选参数，用于拆分 Rowkey 的分隔符，若不设置	|

参数名	说明	取值样例
隔符	则不切分。	
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 3.3.5.10 配置 Hive 连接时，源端作业参数如表 3-63 所示。

表3-63 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
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	• 单/多值过滤： "${dateformat(yyyyMMdd, -1, DAY)}

参数名	说明	取值样例
	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。	${dateformat(yyyyMMdd)}" • 范围过滤： "${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 3.3.5.9 配置 DLI 连接时，源端作业参数如表 3-64 所示。

表3-64 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 3.3.5.14 配置 FTP/SFTP 连接时，源端作业参数如表 3-65 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3-65 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
	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	是

参数类型	参数名	说明	取值样例
		数。	
	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
	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	ok.txt

参数类型	参数名	说明	取值样例
		任务。该文件本身不会被迁移。	
	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 3-66 所示。当前只支持从 HTTP URL 导出数据，不支持导入。

表3-66 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、FusionInsight LibrA、PostgreSQL、Microsoft SQL Server、SAP HANA。

从以上数据库导出数据时，源端作业参数如表 3-67 所示。

表3-67 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	select id,name from sqoop.user;

参数类型	参数名	说明	取值样例
		• into outfile	
	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
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	id

参数类型	参数名	说明	取值样例
		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，不支持中文字符。	
	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。	-
	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	否

参数类型	参数名	说明	取值样例
		dba_tab_partitions 和 **dba_tab_subpartitions** 的 **SELECT** 权限。	

3.3.6.3.9 配置 MySQL 源端参数

作业中源连接为 3.3.5.7 配置 MySQL 数据库连接，源端作业参数如表 3-68 所示。

表3-68 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;
	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

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 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，不支持中文字符。	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	使用该字段将作业拆分为多个子作业并发执	-

参数类型	参数名	说明	取值样例
	字段	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：
- 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。

1. 当前版本不支持一次性删除、更新万条记录。
2. 不支持整库迁移。
3. 不支持 DDL 操作。
4. 不支持事件（event）迁移。
5. 当选择增量迁移时，源 MySQL 数据库的“binlog_format”需要设置为“ROW”。
6. 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。
7. 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。
8. 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 3.3.5.8 配置 Oracle 数据库连接，源端作业参数如表 3-69 所示。

表3-69 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_E

参数类型	参数名	说明	取值样例
		• ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	
	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，不支持中文字符。	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	DS='${dateformat(yyyy-

参数类型	参数名	说明	取值样例
		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。	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
	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 3.3.5.6 配置分库连接，源端作业参数如表 3-70 所示。

表3-70 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 3.3.5.19 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 3-71 所示。

表3-71 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 3-72 所示。

表3-72 Redis 作为源端时的作业参数

参数名	说明	取值样例
Redis 键前	键的前缀，类似关系型数据库的表名。	TABLE

参数名	说明	取值样例
缀		
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。	;
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 3.3.5.21 配置 Kafka 连接或 3.3.5.22 配置 DMS Kafka 连接时，源端作业参数如表 3-73 所示。

表3-73 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

参数	说明	取值样例
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 3.3.5.23 配置 Elasticsearch/云搜索服务（CSS）连接时，源端作业参数如表 3-74 所示。

表3-74 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
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	last_name:Smith

参数名	说明	取值样例
	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。	
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3-66 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 3.3.5.18 配置 CloudTable OpenTSDB 连接时，源端作业参数如表 3-75 所示。

表3-75 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 3.3.5.13 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 3-76 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3-76 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	KMS

参数类型	参数名	说明	取值样例
		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 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	否

参数类型	参数名	说明	取值样例
		入的桶时，必须选择非归档存储的桶。	
	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。	,
	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	从关系型数据库导出数据到 OBS，“文件格	否

参数类型	参数名	说明	取值样例
	行	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。	
	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
	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 3.3.5.12 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 3-77 所示。

表3-77 HDFS 作为目的端时的作业参数

参数名	说明	取值样例

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
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	否

参数名	说明	取值样例
	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"。	
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
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 3.3.5.11 配置 HBase 连接或 3.3.5.17 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 3-78 所示。

表3-78 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	否

参数名	说明	取值样例
	字符串写入 HBase。	

3.3.6.4.4 配置 Hive 目的端参数

作业中目的连接为 3.3.5.10 配置 Hive 连接时，目的端作业参数如表 3-79 所示。

表3-79 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 3-80 所示。

表3-80 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	不自动创建

参数类型	参数名	说明	取值样例
		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。	
	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	否

参数类型	参数名	说明	取值样例
		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。	
	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 3.3.5.5 配置常见关系数据库连接，目的端作业参数如表 3-81 所示。

表3-81 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

参数类型	参数名	说明	取值样例
		• 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。	
	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	清除部分数据

参数类型	参数名	说明	取值样例
		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
	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。 应用场景主要是将有中文内容的字符字段导入到 DWS 时，需要自动将字符长度放大 3 倍。 在导入中文内容的字符到 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。	否

参数类型	参数名	说明	取值样例
		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。	
	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 3-67 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3-67 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 3.3.5.16 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 3-82 所示。

表3-82 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 3-83 所示。

表3-83 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 3.3.5.23 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 3-84 所示。

表3-84 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，	index

参数名	说明	取值样例
	索引和类型名称只能全部小写，不能有大写。	
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
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 3.3.5.9 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 3-85 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3-85 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

参数名	说明	取值样例
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
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 3.3.5.18 配置 CloudTable OpenTSDB 连接时，目的端作业参数如表 3-86 所示。

表3-86 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 3-87 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3-87 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 3.3.5.13 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3-88 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。

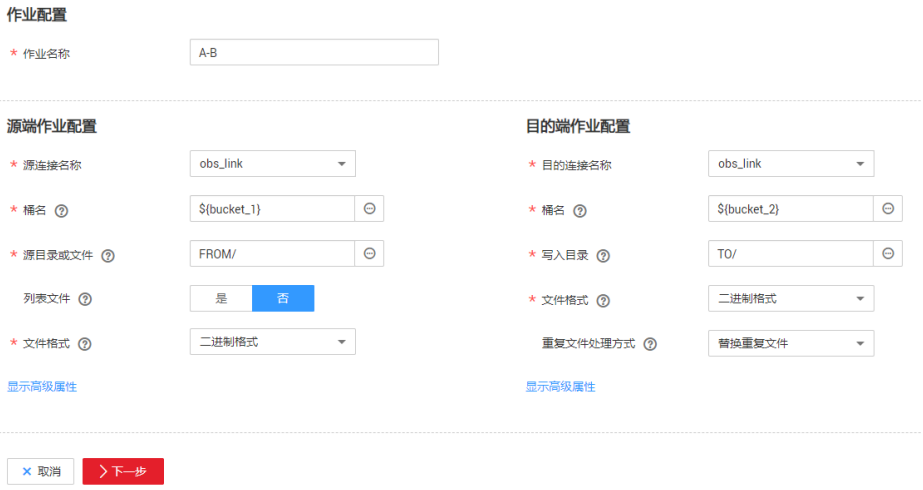
```
bucket_1=A
bucket_2=B
```

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3-68 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 3.3.6.5 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 3-69。

图3-69 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3-89 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
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、管理员账号和密码，且该账号拥有数据导入、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。

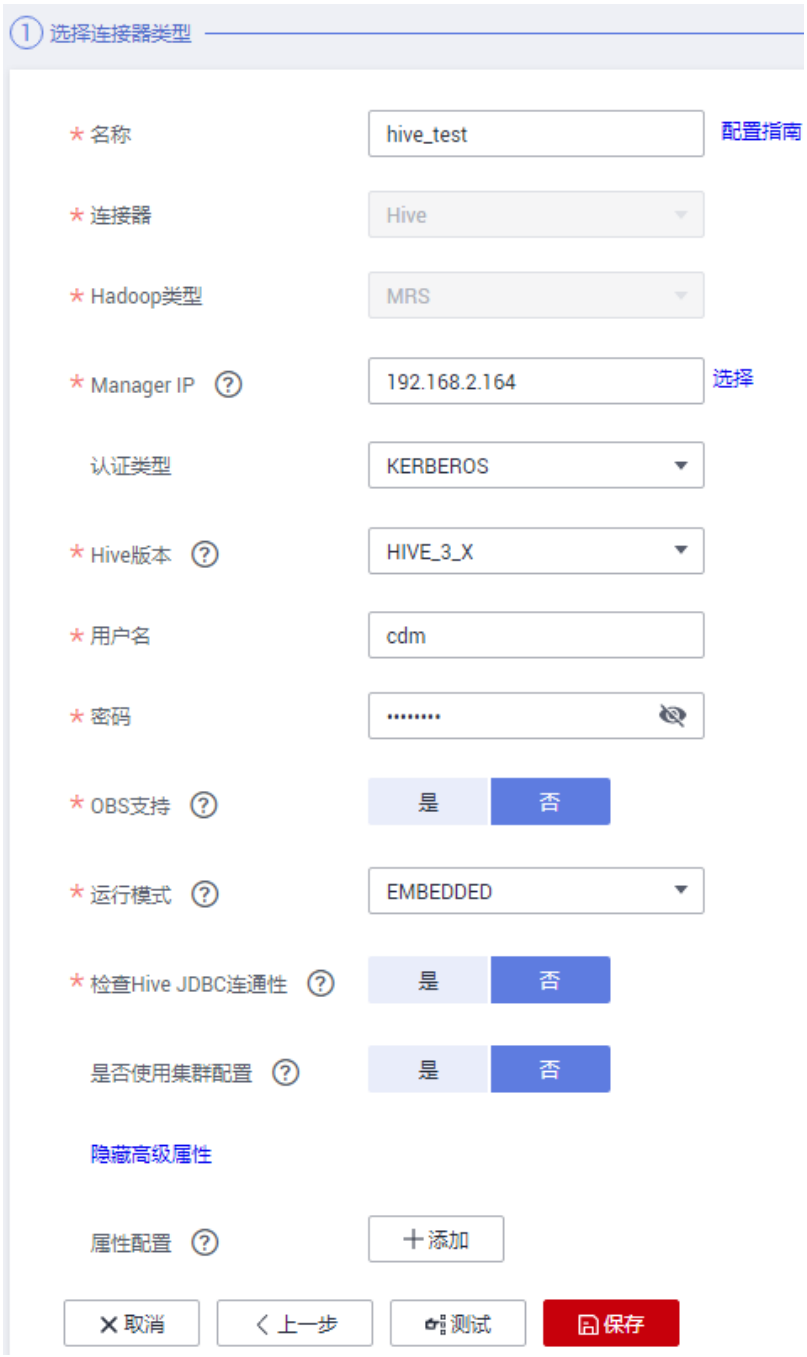
新建 MRS hive 连接

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 3-70 所示。

图3-70 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 3-71 所示。

图3-71 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。这里保持默认，必填参数如表 3-90 所示。

表3-90 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	mrs-link

参数名	说明	取值样例
	义便于记忆、区分的连接名。	
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 MRS 组件的库、表、列进行操作，还需要参考 MRS 文档添加对应组件的库、表、列操作权限。 • 如果 CDM 集群为 2.9.0 之前的版本，或 MRS 集群为 3.1.0 之前的版本，则所创建的用户需要具备 Manager_administrator 或 System_administrator 权限，才能在 CDM 创建连接。 • 仅具备 Manager_tenant 或 Manager_auditor 权限，无法创建连接。	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
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	EMBEDDED

参数名	说明	取值样例
	认证模式又有 SIMPLE 认证模式，只能使用 STANDALONE 模式。 **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。	
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。

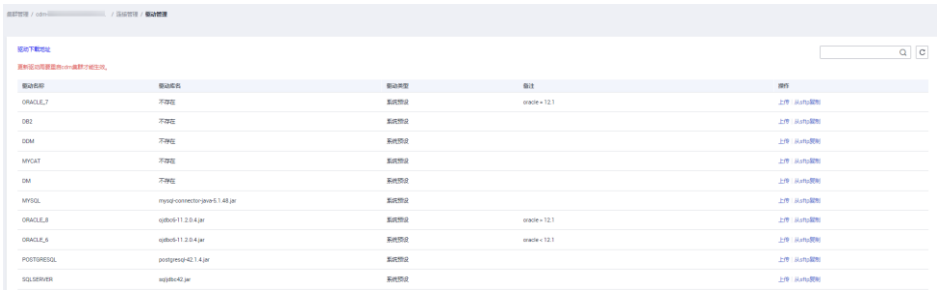
前提条件

- 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3-72 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 3-73 所示。

图3-73 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 3-91 所示。

图3-74 创建 MySQL 连接

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

参数名	说明	取值样例
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 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 3-75 所示。

图3-75 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 3.3.5.2 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 3.3.4.1 创建 CDM 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3-76 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 3.3.5.5 配置常见关系数据库连接。这里保持默认，必填参数如表 3-92 所示。

表3-92 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。

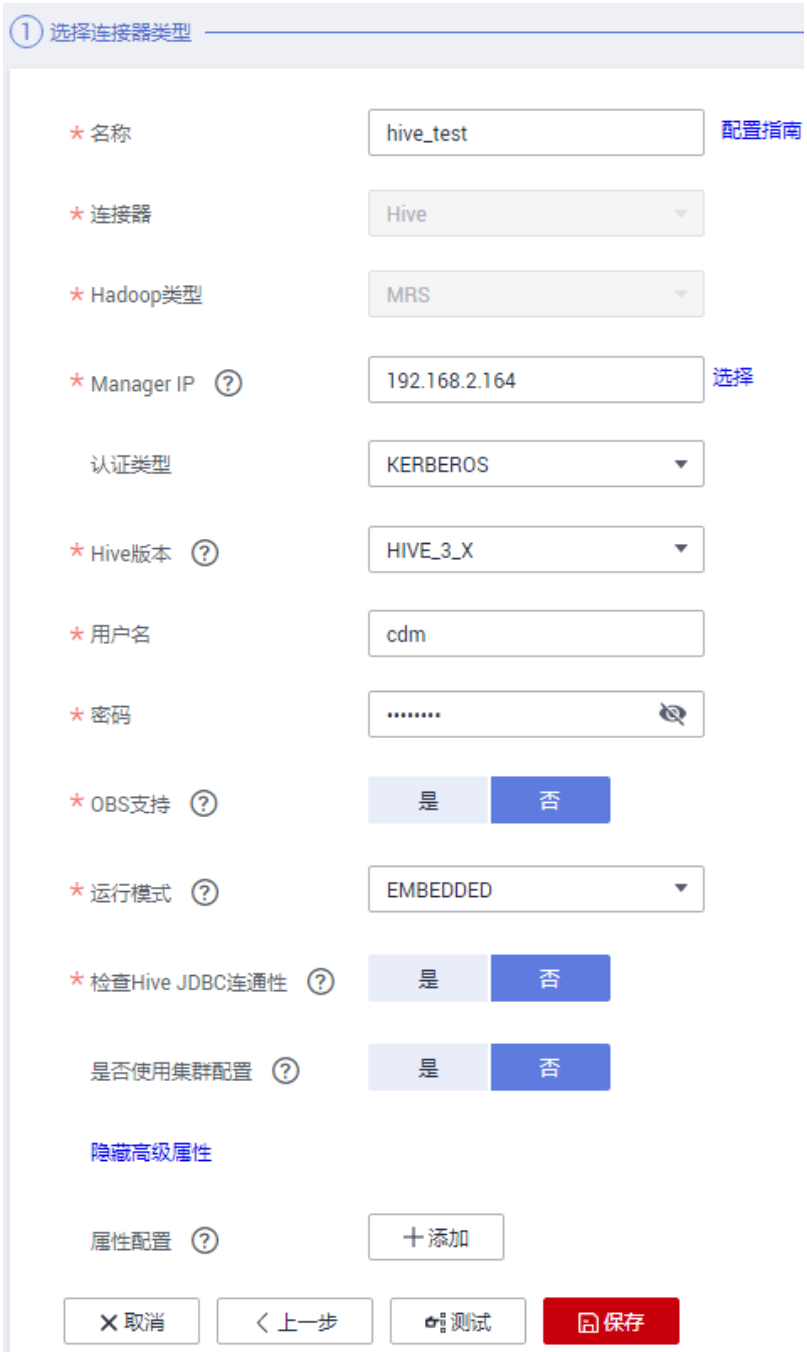
----结束

创建 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 3-77 所示。

图3-77 创建 MRS Hive 连接



各参数说明如表 3-93 所示，需要您根据实际情况配置。

表3-93 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
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
运行模式	“HIVE_3_X”版本支持该参数。支持以下模式：	EMBEDDED

参数名	说明	取值样例
	• 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。	
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。	-

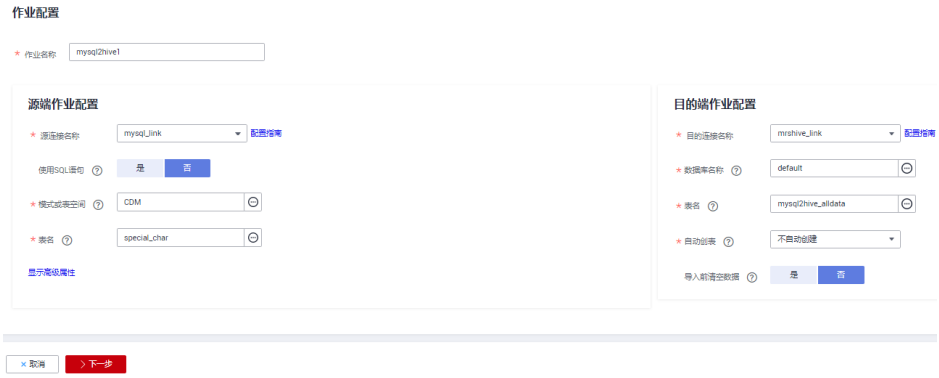
步骤 3 单击“保存”回到连接管理界面。

----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 3-78 所示。

图3-78 创建 MySQL 到 Hive 的迁移任务



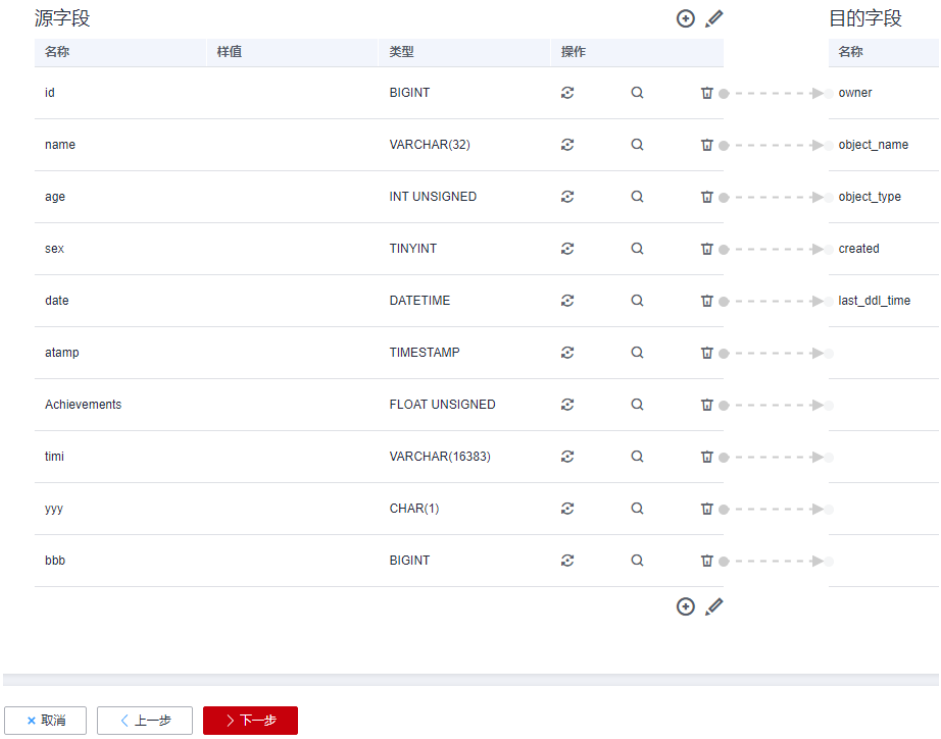
说明

“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 3-79 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3-79 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 3-80 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3-80 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 3.3.6.5 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 3.3.5.2 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 3.3.4.1 创建 CDM 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 3.3.5.5 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

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 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3-81 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。

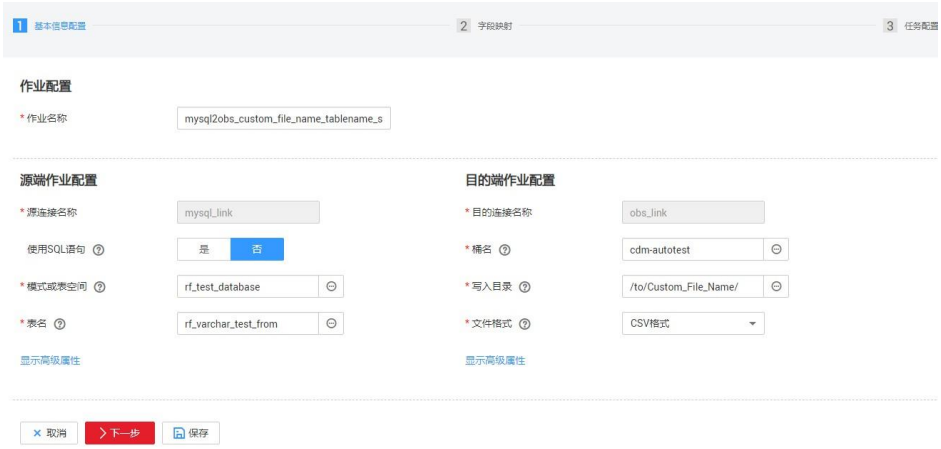
步骤 3　单击“保存”回到连接管理界面。

----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3-82 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 3.3.6.3.8 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 3.3.6.4.1 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 3-83 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3-83 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 3.3.6.5 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 3.3.5.2 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 3.3.4.1 创建 CDM 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 3.3.5.5 配置常见关系数据库连接。这里保持默认，必填参数如表 3-95 所示。

表3-95 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 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 3-96 所示，可选参数保持默认即可。

表3-96 DWS 连接参数

参数名	说明	取值样例

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

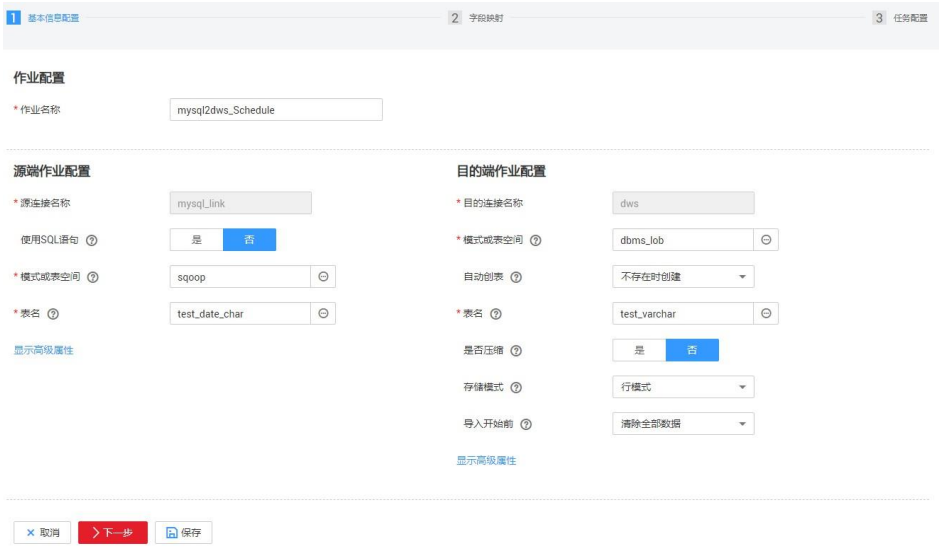
步骤 3 单击“保存”完成创建连接。

----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3-84 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 3.3.6.3.8 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 3-85 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3-85 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 3.3.6.5 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 3.3.5.2 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 3.3.4.1 创建 CDM 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3-86 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 3.3.5.5 配置常见关系数据库连接。这里保持默认，必填参数如表 3-97 所示。

表3-97 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 3-87 所示。

图3-87 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 3.3.5.2 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 3.3.4.1 创建 CDM 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。

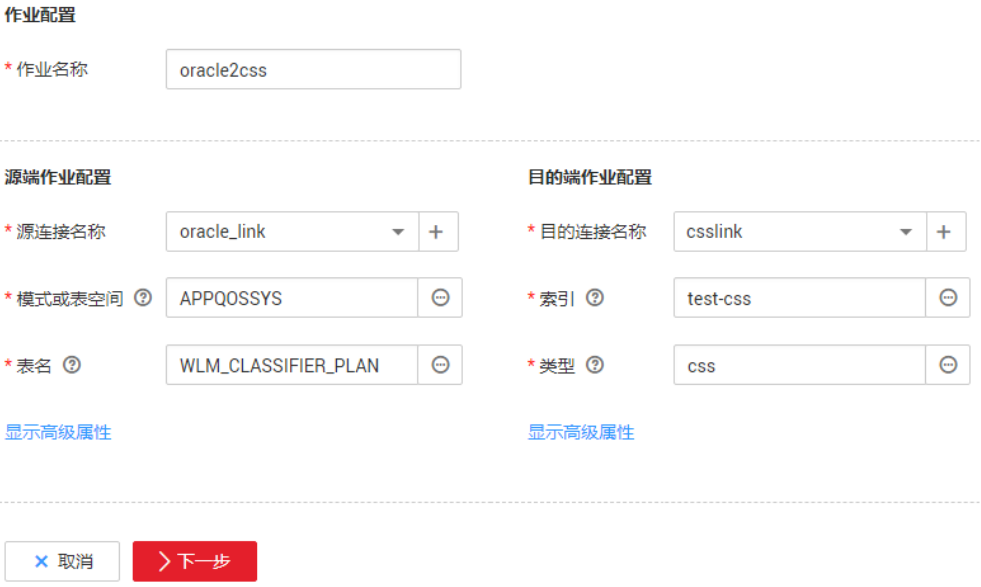
步骤 3 单击“保存”回到连接管理界面。

----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3-88 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 3.3.6.3.8 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 3.3.6.4.9 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 3-89 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3-89 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 3.3.6.5 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

- 已购买 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 3.3.5.2 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 3.3.4.1 创建 CDM 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3-90 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 3-98 所示。

图3-91 创建 Oracle 连接

表3-98 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 3-99 所示，可选参数保持默认即可。

表3-99 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

参数名	说明	取值样例
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

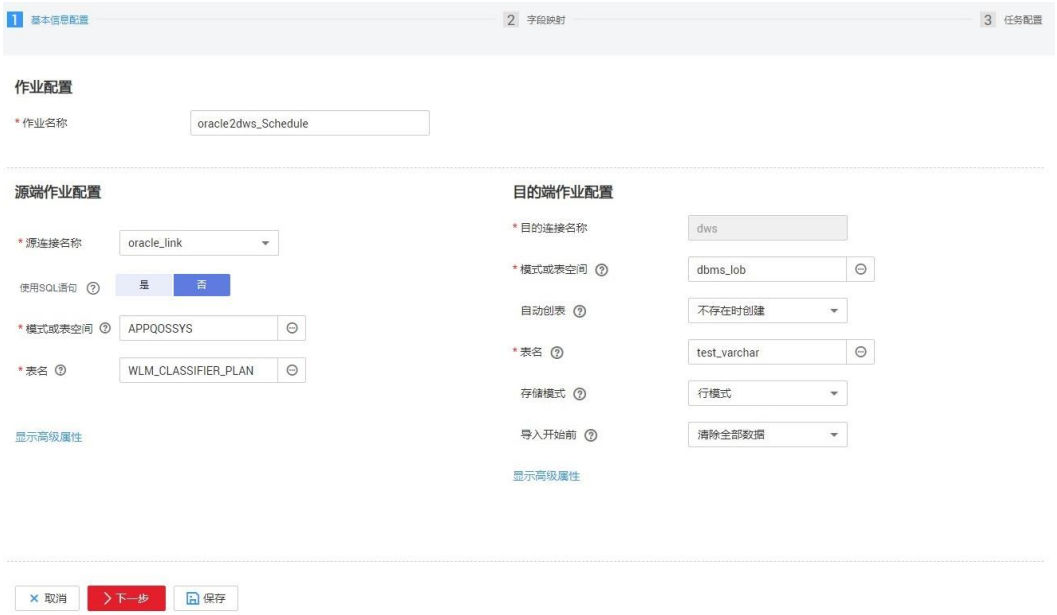
步骤 3 单击“保存”完成创建连接。

----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 Oracle 导出数据到 DWS 的任务。

图3-92 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 3.3.6.3.8 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 3-93 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3-93 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 3.3.6.5 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 3.3.4.1 创建 CDM 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3-94 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3-95 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 3.3.6.3.1 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 3.3.6.4.9 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 3-96 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3-96 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 3.3.6.5 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 3.3.4.1 创建 CDM 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 3-97 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3-97 创建 DLI 连接

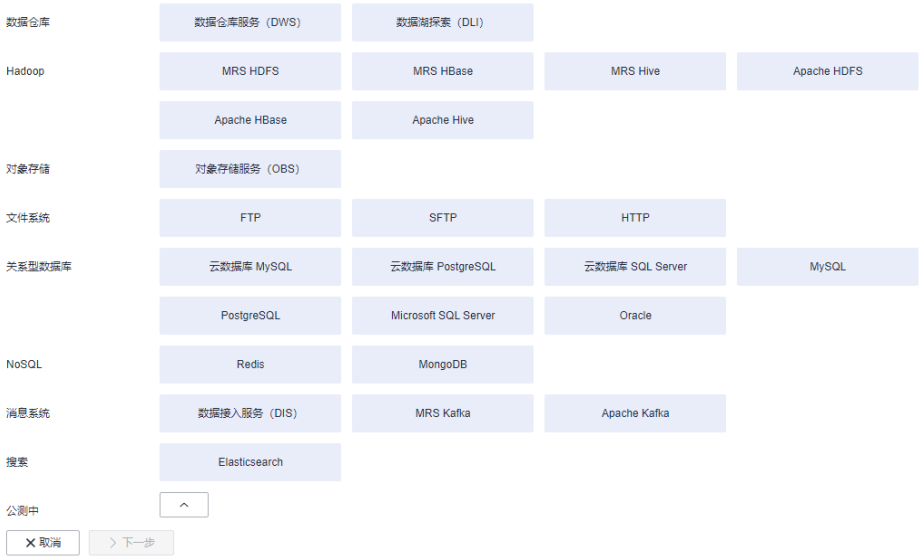
步骤 3 单击“保存”回到连接管理界面。

----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3-98 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 3-99 所示。

图3-99 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 3.3.6.3.1 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 3.3.6.5 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 3.3.4.1 创建 CDM 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3-100 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。

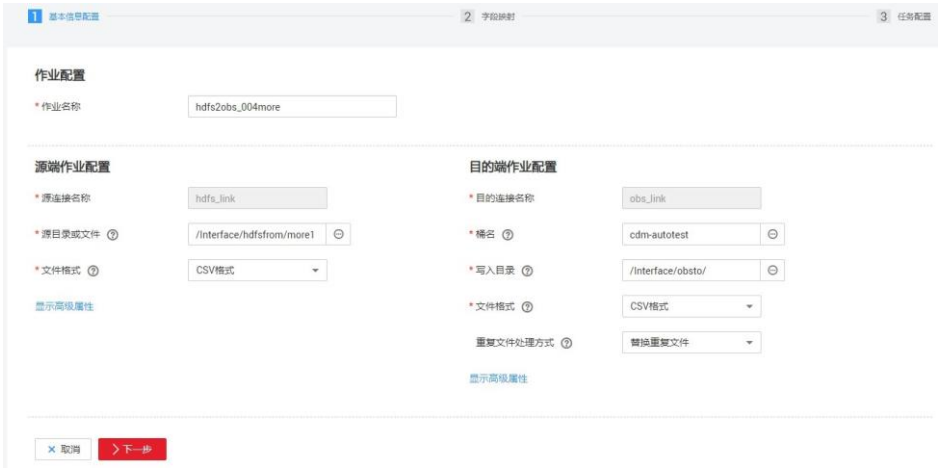
步骤 3 单击“保存”回到连接管理界面。

----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3-101 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 3.3.6.3.2 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 3.3.6.4.1 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 3.3.6.5 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 3.3.4.1 创建 CDM 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。

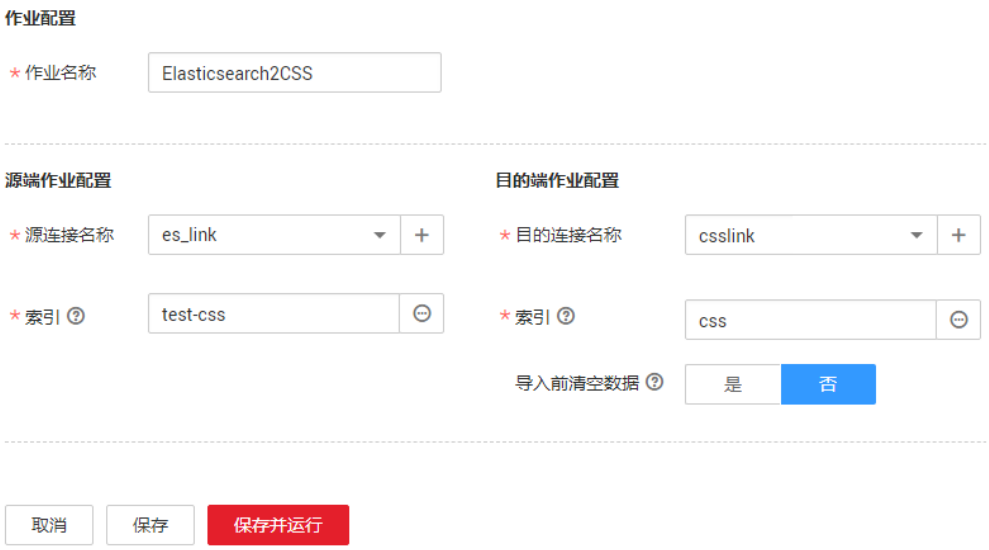
步骤 3 单击“保存”回到连接管理界面。

----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3-102 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3-103 作业执行记录

----结束

3.3.8.13 DDS 数据迁移到 DWS

操作场景

CDM 支持迁移文档数据库服务（Document Database Service，简称 DDS）的数据到其他数据源，这里以数据仓库服务（Data Warehouse Service，简称 DWS）为例，介绍如何使用 CDM 将 DDS 数据迁移到 DWS，流程如下：

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

前提条件

- 已 DWS/DDS。
- 已获取 DWS/DDS 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/DDS 数据库的读、写和删除权限。

创建 CDM 集群并绑定 EIP

步骤 1 参考 3.3.4.1 创建 CDM 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。
- 如果 DDS 和 DWS 属于相同的 VPC，则创建 CDM 集群时选择同一个 VPC，不用绑定 EIP。子网、安全组可以选择与其中一个（DDS 或 DWS）集群的保持一致，再配置安全组规则允许 CDM 集群访问另一个服务（DWS 或 DDS）的集群。
- 如果 DDS 和 DWS 不在同一个 VPC，则创建 CDM 集群时选择与 DDS 相同的 VPC，再将 CDM 集群 3.3.4.2 解绑/绑定集群的 EIP，CDM 通过 EIP 访问 DWS 集群。

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 DWS。如果 DDS 与 DWS 在同一个 VPC，则不用为 CDM 集群绑定 EIP。

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 DDS 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“文档数据库服务（DDS）”后单击“下一步”配置连接参数，参数说明如表 3-100 所示。

表3-100 DDS 连接参数

参数名	说明	取值样例
名称	根据连接的数据源，用户自定义便于记忆、区分的连接名称。	mongo_link
服务器列表	DDS 集群的地址列表，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_mongodb
用户名	登录 DDS 数据库的用户名。	cdm
密码	登录 DDS 数据库的密码。	-

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 3-101 所示，可选参数保持默认即可。

表3-101 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。	-

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”，开始创建数据迁移任务。

步骤 2 配置作业基本信息：

- 作业名称：输入便于记忆、区分的作业名称。
- 源端作业配置
  - 源连接名称：选择创建 DDS 连接中的“mongo_link”。
  - 数据库名称：选择待迁移数据的数据库。
  - 集合名称：DDS 中 MongoDB 的集合，类似于关系型数据库中的表名。
- 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。
  - 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。
  - 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 3 单击“下一步”进入字段映射界面，CDM 会自动匹配源端和目的端的数据表字段，需用户检查字段映射关系是否正确。

- 如果字段映射关系不正确，用户单击字段所在行选中后，按住鼠标左键可拖拽字段来调整映射关系。
- 导入到 DWS 时需要手动选择 DWS 的分布列，建议按如下顺序选取：

  a. 有主键可以使用主键作为分布列。

b. 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。

c. 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。

- 如果需要转换源端字段内容，可在该步骤配置，这里选择不进行字段转换。

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 3.3.6.5 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

3.3.9 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 3.3.9.1.3 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3-104 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 3.3.9.1.3 时间宏变量使用解析，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 3-105 所示，参数配置如下：

  图3-105 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 3.3.9.1.3 时间宏变量使用解析，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 3-102 所示。

表3-102 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 3-106 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3-106 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 3-107 所示。

图3-107 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3-108 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 3.3.9.1.3 时间宏变量使用解析，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3-109 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。

KMS 密钥需要先在数据加密服务创建，具体操作请参见《数据加密服务用户指南》。

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。

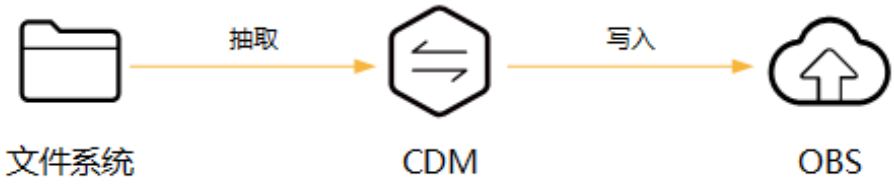
说明

- 如果选择使用 KMS 加密，则无法 3.3.9.4 MD5 校验文件一致性。
- 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 3-110 所示。

图3-110 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。

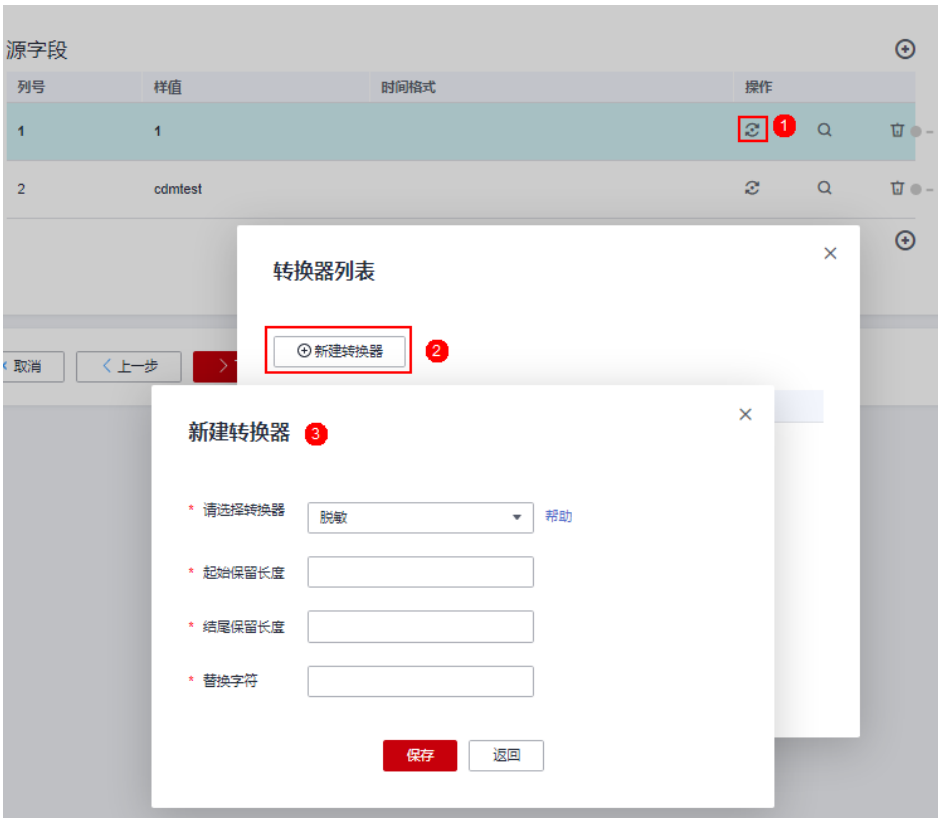
说明

- 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 3.3.9.3 迁移文件时加解密。

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 3-111 所示。

图3-111 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3-112 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3-113 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(SqoopConfigurati
on.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3-103 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3-104 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)

列号	样值
4	sqoop.anonymous.user
5	189.xxx.xxx.75
6	show
7	version
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3-105 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3-106 Django 日志解析结果

列号	样值
1	08/Jan/2018 20:59:07

列号	样值
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3-107 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。

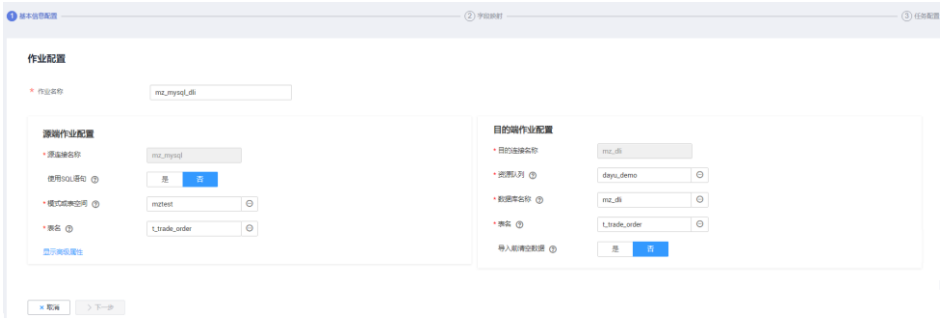
前提条件

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

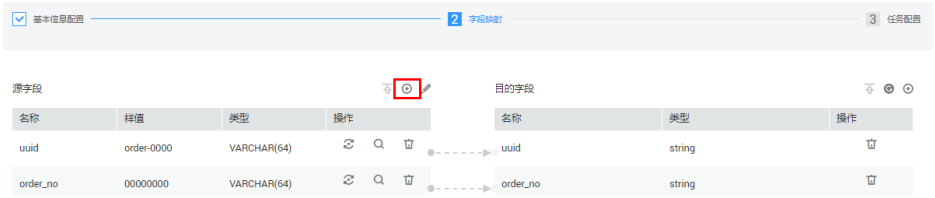
步骤 1 在创建表/文件迁移作业时，选择已创建的源端连接器、目的端连接器。

图3-114 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3-115 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 3-108。

图3-116 添加字段

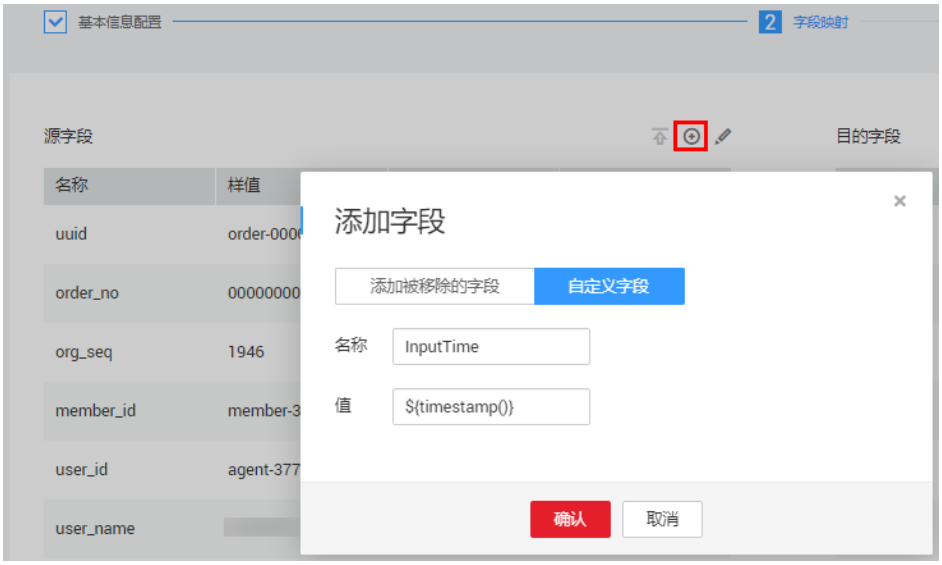


表3-108 时间变量宏定义具体展示

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

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3-109 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 3-109。

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 3-117 所示，数据库的 name 字段的值中包含了字段分隔符逗号：

     图3-117 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：3.3.9.7 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

took	timedOut	total	maxScore
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3-118 添加字段

ii. 在添加的新字段后面，单击 添加字段转换器。

图3-119 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3-120 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 3.3.9.1.1 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 3-109。

   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 数据架构

3.4.1 数据架构概述

数据架构简介

DataArts Studio 数据架构以关系建模、维度建模理论支撑，实现规范化、可视化、标准化数据模型开发，定位于数据治理流程设计落地阶段，输出成果用于指导开发人员实践落地数据治理方法论。

数据架构作为数据治理的一个核心模块，承担数据治理过程中的数据加工并业务化的功能。数据架构主要包括数据调研、标准设计、模型设计和指标设计四个部分。数据架构支持 DLI、POSTGRESQL、DWS、MRS_Hive 数据连接类型。

DataArts Studio 数据架构致力于：

- 构建统一的数据分类体系，用于目录化管理所有业务数据，便于数据的归类、查找、评价和使用。
- 构建统一的数据标准体系，基于国家或行业标准，用于标准化每一行数据，每一个字段的具体取值，提升数据质量和易用性。
- 构建统一的数据模型体系，通过规范定义和数据建模，自顶向下构建企业数据分层体系，沉淀企业数据公共层和主题库，便于数据的流通、共享、创造、创新，提升数据使用效率，极大的减少数据冗余、混乱、隔离、不一致以及谬误等。

模型设计方法概述

根据业务需求抽取信息的主要特征，模拟和抽象出一个能够反映业务信息（对象）之间关联关系的模型，即数据模型。数据模型也是可视化的展现企业内部信息如何组织的蓝图。数据模型应满足三方面要求：能比较真实地模拟业务（场景）；容易为人所理解；便于在 IT 系统中实现。

在 DataArts Studio 数据架构的数据建模过程中，用到的建模方法主要有以下两种：

- **关系建模**

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

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

  - **概念模型**：是从用户的视角，主要从业务流程、活动中涉及的主要业务数据出发，抽象出关键的业务实体，并描述这些实体间的关系。

- **逻辑模型**：是概念模型的进一步细化，通过实体、属性和关系勾勒出企业的业务信息蓝图，是 IT 和业务人员沟通的桥梁。逻辑数据模型是一组规范化的逻辑表结构，逻辑数据模型是根据业务规则确定的，关于业务对象、业务对象的数据项及业务对象之间关系的基本蓝图。
- **物理模型**：是在逻辑数据模型的基础上，考虑各种具体的技术实现因素，进行数据库体系结构设计，真正实现数据在数据库中的存放，例如：所选的数据仓库是 DWS 或 MRS_Hive。

- **维度建模**

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

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

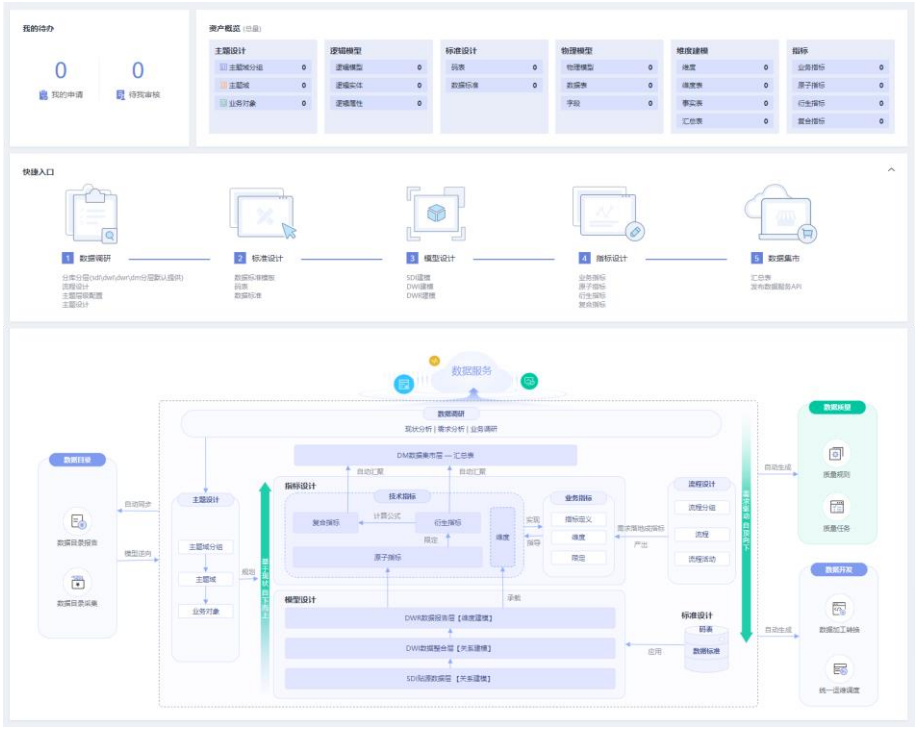
  典型的维度模型有星形模型，以及在一些特殊场景下使用的雪花模型。

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

数据架构总览

在 DataArts Studio 控制台首页，选择对应工作空间的“数据架构”模块，进入数据架构页面，查看“总览”，如图 3-121 所示。

图3-121 数据架构总览



- **我的待办**
  - 显示“我的申请”和“待我审核”的数量。
  - 单击每一项上面统计数量将分别跳转到“我的申请”和“待我审核”页面。
- **资产概览**
  - 显示数据架构中所有对象的总量。
  - 单击每个对象名称后的统计数量将跳转的该对象的管理页面。
- **快捷入口**

  显示数据架构数据治理方法的整体流程。单击流程下的具体操作，可以跳转到对应的界面。
- **数据架构流程**
  - 显示数据架构流程以及与 DataArts Studio 其他模块间的交互关系。关于数据架构流程的详细描述，请参见 3.4.2 数据架构使用流程。
  - 将鼠标移至流程图上的对象名称之上，页面上将显示对象的描述信息。
  - 对于 DataArts Studio 已支持的对象，单击对象名称，可跳转至该对象的管理页面。

数据架构信息架构

信息架构是以结构化的方式描述在业务运作和管理决策中所需要的各类信息及其关系的一套整体组件规范。在数据架构的“信息架构”页面，可以查看和管理所有的表，包括业务表、维度表、事实表、汇总表等资源。

在 DataArts Studio 控制台首页，选择对应工作空间的“数据架构”模块，进入数据架构页面，查看“信息架构”。

在信息架构页面，可以执行以下操作：

- **搜索**

  在“信息架构”列表右上方，单击“高级搜索”，设置表名、类型、数据源等筛选条件，然后单击“搜索”可以查找指定的表，单击“表名称”，可以进入表的详情页面，查看表的详细信息。
- **新建**

  单击“新建”，可以新建业务表-逻辑模型、业务表-物理模型、维度表、事实表和汇总表。创建的过程可以参见 3.4.6.1.1 逻辑模型设计、3.4.6.1.2 物理模型设计、3.4.6.2.1 新建维度、3.4.6.2.3 新建事实表、3.4.8.1 新建汇总表。
- **导入**

  单击“更多 > 导入”，当前仅支持导入业务表。下载表导入模板，填写模板后，先添加再上传，上传成功后，然后单击“关闭”。有关导入业务表的更多信息，请参见 3.4.9.3 导入导出表。
- **导出**

  单击“更多 > 导出”，可以导出业务表-物理模型或 DDL。有关导出的更多信息，请参见导出表或 DDL。
- **同步**

单击“更多 > 同步”，可以同步表到数据目录，作为技术资产；同步逻辑模型到数据目录，作为业务资产。

- **修改主题**

  单击“更多 > 修改主题”，可以将选中的表更改到其它主题。

- **删除**

  单击“更多 > 删除”，可以删除数据表，其中待发布，已发布和待下线状态的数据表不可被删除。且数据被引用的数据表不可被删除。

- **下线**

  单击“更多 > 下线”，可以下线已发布且不带下展的数据表。数据被引用的数据表不支持下线。

说明

“带下展”，指待发布后又重新编辑的数据。

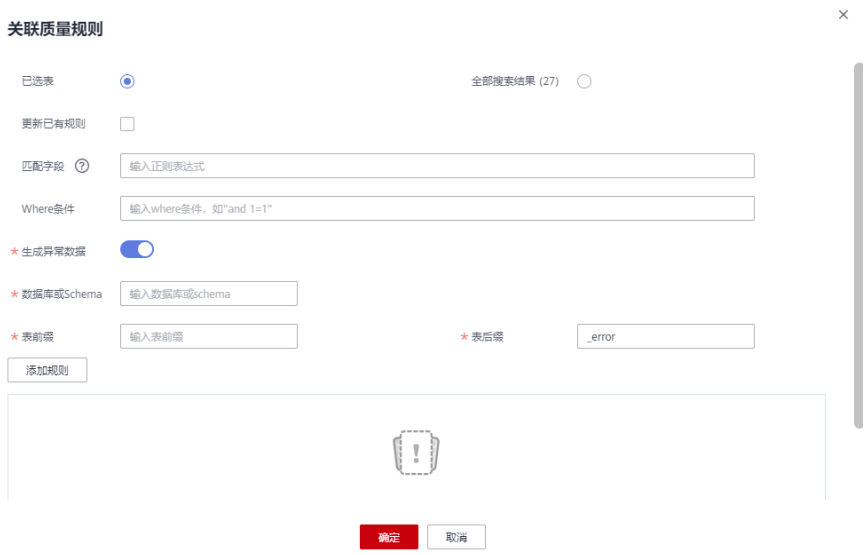
- **发布**

  单击“发布”，可发布数据表。待发布、待下线、已发布（不带下展）状态的数据表不支持发布。

- **关联质量规则**

  单击“关联质量规则”，配置下图所示的相关参数，完成质量规则的关联。有关关联质量规则的更多信息，您也可以参考 3.4.9.4 关联质量规则。

图3-122 关联质量规则

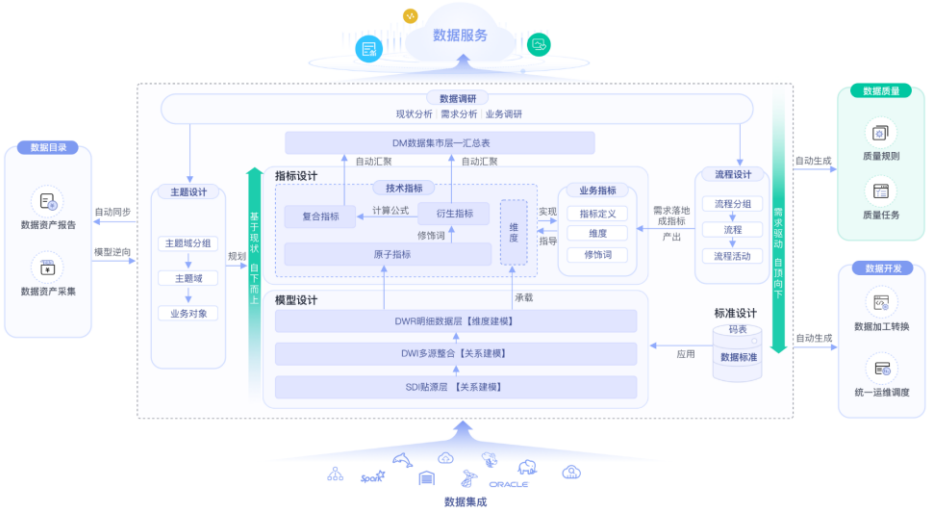


生成异常数据：勾选此项，表示异常数据将按照配置的参数存储到规定的库中。

3.4.2 数据架构使用流程

DataArts Studio 数据架构的流程如下：

图3-123 数据架构流程



1. **准备工作：**
   - **添加审核人：**在数据架构中，业务流程中的步骤都需要经过审批，因此，需要先添加审核人。只有工作空间管理员角色的用户才具有添加审核人的权限。
   - **管理配置中心：**数据架构中提供了丰富的自定义选项，统一通过配置中心提供，您需要根据自己的业务需要进行自定义配置。
2. **数据调研：**基于现有业务数据、行业现状进行数据调查、需求梳理、业务调研，输出企业业务流程以及数据主题划分。
   - **主题设计：**通过分层架构表达对数据的分类和定义，帮助厘清数据资产，明确业务领域和业务对象的关联关系。
     - **主题域分组：**基于业务场景对主题域进行分组。
     - **主题域：**互不重叠数据的高层面的数据分类，用于管理其下一级的业务对象。
     - **业务对象：**指企业运作和管理中不可缺少的重要人、事、物信息。
   - **流程设计：**针对流程的一个结构化的整体框架，描述了企业流程的分类、层级以及边界、范围、输入/输出关系等，反映了企业的商业模式及业务特点。
3. **标准设计：**新建码表&数据标准。
   - **新建码表：**通常只包括一系列允许的值和附加文本描述，与数据标准关联用于生成值域校验质量监控。
   - **新建数据标准：**用于描述公司层面需共同遵守的属性层数据含义和业务规则。其描述了公司层面对某个数据的共同理解，这些理解一旦确定下来，就应作为企业层面的标准在企业内被共同遵守。
4. **模型设计：**应用关系建模和维度建模的方法，进行分层建模。
   - **关系建模：**基于关系建模，新建 SDI 层和 DWI 层两个模型。
     - **SDI：**Source Data Integration，又称贴源数据层。SDI 是源系统数据的简单落地。

    - **DWI**：Data Warehouse Integration，又称数据整合层。DWI 整合多个源系统数据，对源系统进来的数据进行整合、清洗，并基于三范式进行关系建模。
  - **维度建模**：基于维度建模，新建 DWR 层模型并发布维度和事实表。
    - **DWR**：Data Warehouse Report，又称数据报告层。DWR 基于多维模型，和 DWI 层数据粒度保持一致。
    - **维度**：维度是用于观察和分析业务数据的视角，支撑对数据进行汇聚、钻取、切片分析，用于 SQL 中的 GROUP BY 条件。
    - **事实表**：归属于某个业务过程的事实逻辑表，可以丰富具体业务过程所对应事务的详细信息。
5. **指标设计**：新建业务指标和技术指标，技术指标又分为原子指标、衍生指标和复合指标。
  - **指标**：指标一般由指标名称和指标数值两部分组成，指标名称及其涵义体现了指标质的规定性和量的规定性两个方面的特点，指标数值反映了指标在具体时间、地点、条件下的数量表现。

    业务指标用于指导技术指标，而技术指标是对业务指标的具体实现。
  - **原子指标**：原子指标中的度量和属性来源于多维模型中的维度表和事实表，与多维模型所属的业务对象保持一致，与多维模型中的最细数据粒度保持一致。

    原子指标中仅含有唯一度量，所含其它所有与该度量、该业务对象相关的属性，旨在用于支撑指标的敏捷自助消费。
  - **衍生指标**：是原子指标通过添加限定、维度卷积而成，限定、维度均来源于原子指标关联表的属性。
  - **复合指标**：由一个或多个衍生指标叠加计算而成，其中的维度、限定均继承于衍生指标。

    注意，不能脱离衍生指标、维度和限定的范围，去产生新的维度和限定。
6. **数据集市建设**：新建 DM 层并发布汇总表。
  - **DM (Data Mart)**：又称数据集市。DM 面向展现层，数据有多级汇总。
  - **汇总表**：汇总表是由一个特定的分析对象（如会员）及其相关的统计指标组成的。组成一个汇总逻辑表的统计指标都具有相同的统计粒度（如会员），汇总逻辑表面向用户提供了以统计粒度（如会员）为主题的所有统计数据（如会员主题集市）。

3.4.3 准备工作

发布衍生指标后，您就可以在运维中心对衍生指标进行运行或调度。

操作步骤

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-124 选择数据架构

2. 在左侧导航栏中，单击“运维中心”，进入运维中心页面，如下图所示。

图3-125 运维中心页面 1

3. 您可以根据实际需要选择如下操作。

当需要...	则...
编辑调度	执行 4。
运行	执行 5。
停止调度	执行 6。
查看运行日志	执行 7。

4. 编辑调度
   a. 在需要编辑的对象右侧，单击“编辑调度”，弹出“编辑调度”对话框。
   b. 根据实际情况设置相关信息。
   c. 单击“确定”。
5. 运行
   在需要启动的对象右侧，单击“运行”，系统提示运行成功。
6. 停止调度

在需要停止的对象右侧，单击“停止调度”，系统提示停止调度成功。

7. 查看运行日志

   a. 在对象列表中，单击所需查看的对象名称前的 按钮展开对象，展开后可以查看该对象的运行实例。

   b. 找到所需查看的实例，单击“查看运行日志”，系统将显示“运行日志”页面，可以查看运行日志和运行结果。

3.4.3.1 添加审核人

在数据架构中，业务流程中的步骤都需要经过审批，因此，需要先添加审核人。只有工作空间管理员角色的用户才具有添加审核人的权限。

添加审核人

审核人必须是当前工作空间下具有审核权限的成员，需要先在“DataArts Studio 首页-空间管理”的工作空间内编辑并添加空间成员。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

   图3-126 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“配置中心”，进入相应页面后，选择“审核人管理”页签。
3. 在“审核人管理”页面，单击“添加”按钮。
4. 在弹出的添加对话框中，选择审核人，输入正确的手机号码和电子邮箱，单击“确定”完成审核人添加。

   审核人必须是当前工作空间下具有审核权限的成员，只有管理员和开发者才具有审核权限。

说明

- 审核人不支持手工添加，需要先在“DataArts Studio 首页-空间管理”的工作空间内编辑并添加空间成员，以便添加审核人时进行选择。
- 勾选短信通知或邮件通知，并添加审核人后，DataArts Studio 将自动在消息通知服务（SMN）中创建对应的主题。
- 主题的显示名格式为：DataArts_主题_审核人_***项目名称_项目ID***-dlg_ds_***审核人名称***。

图3-127 添加审核人

5. 根据需要，可以添加多个审核人。

相关操作

进入数据架构的“配置中心 > 审核人管理”页面，可以对审核人进行管理。

图3-128 审核人管理

- **查找审核人**

  在审核人列表的右上方，输入所要查找的审核人名称，然后单击 按钮，即可查找指定的审核人。

- **删除审核人**

  在审核人列表中，查找所要删除的审核人，然后选中该审核人，再单击“删除”按钮，即可删除指定的审核人。

3.4.3.2 管理配置中心

主题配置

主题配置用于自定义主题设计中的主题层级和自定义属性。系统默认有三个层级，从上到下分别命名为主题域分组（L1）、主题域（L2）、业务对象（L3）。您可以自定义的主题层级限制在最大 7 层，最少 2 层。自定义属性最多可以配置 10 个。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

   图3-129 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“配置中心”，进入相应页面后再单击“主题配置”页签。
3. 在主题层级区域，可对主题层级进行增加、删除和编辑操作。
   - 在“操作”栏中单击 + 按钮可以添加自定义主题层级项，完成后单击“确定”。
   - 在“操作”栏中单击 按钮可以删除主题层级项，完成后单击“确定”。
   - 除最后一层业务对象外，其它层级均可以通过单击对应的层级名称实现“编辑”操作。
4. 在属性自定义项区域，可对属性进行增加、删除和编辑操作。
   - 在“属性自定义项”右侧，单击“新建”可新增一条自定义属性。
   - 在“操作”栏中单击 按钮可以删除一条自定义属性。
   - 单击对应的属性名称、属性名称（英文），是否必填，实现“编辑”操作。

标准模板管理

标准模板管理用于自定义数据标准的默认选项。首次进入数据架构的数据标准页面，也会显示制定数据标准模板的页面。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

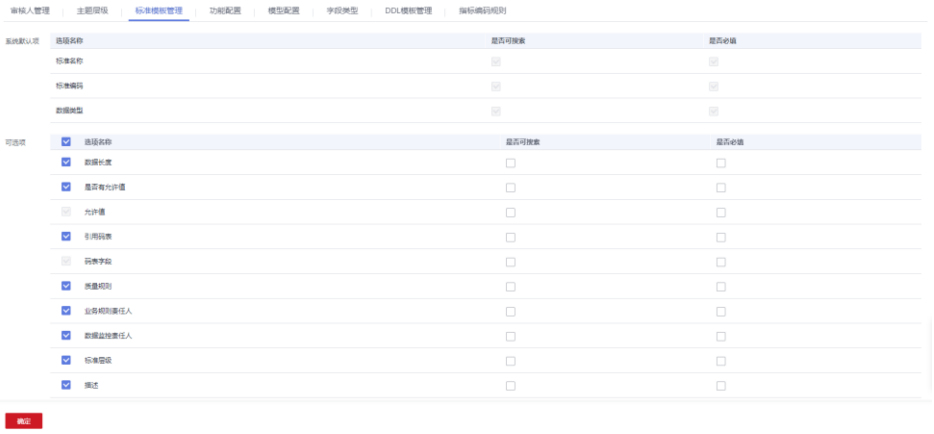
   图3-130 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“配置中心”，进入相应页面后再单击“标准模板管理”。
3. 如下图所示，在“可选项”中勾选所需要的选项，单击“新建”按钮可以添加自定义项，完成后单击“确定”。

说明

- 标准模板支持“是否可搜索”、“是否必填”。
- 保存模板后，在新建数据标准时需要设置此处模板中选中选项的参数值。

图3-131 标准模板管理



功能配置

功能配置用于自定义数据架构中的各项功能。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-132 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“配置中心”，进入相应页面后再单击“功能配置”。
3. 在功能配置页面，可根据用户具体的功能需求配置参数，然后单击“确定”。如果单击“重置”可恢复默认设置。

图3-133 功能配置

- **模型设计业务流程步骤**：此处勾选的流程，在关系建模或维度建模的对象发布上线时，系统会依次自动执行。一般建议全部勾选。
  - **创建表**：当数据架构中的表发布并通过审核后，系统将自动在对应的数据源中创建相应的物理表。在表删除时，系统也会自动删除物理表。
  - **同步技术资产**：关系建模或维度建模中的表发布后，同步表到数据目录模块作为技术资产，同时同步标签到对应技术资产。

说明

若开启“同步技术资产”功能，您必须预先在 DataArts Studio 数据目录模块中对表所属的数据库创建数据目录采集任务并采集成功，否则同步技术资产将会执行失败。

  - **同步业务资产**：同步逻辑模型到数据目录，作为业务资产，同时同步标签到对应业务资产。
  - **资产关联**：实现业务资产与技术资产的关联。业务资产与技术资产同步完成后，在数据目录模块中查看对应的业务资产或技术资产详情时，可以看到相关联的技术资产或业务资产。该功能要求表信息中含有数据源信息。
  - **创建质量作业**：当关系建模或维度建模中的表发布并通过审核后，对于关联数据标准（包含数据长度或允许值）或关联质量规则的表，系统将自动在 DataArts Studio 数据质量模块中创建一个质量作业。
  - **创建数据开发作业**：汇总表发布后，自动生成端到端的全流程数据开发作业。
  - **发布数据服务 API**：汇总表发布后，自动生成数据服务 API，此功能仅当数据服务支持汇总表的数据连接时生效。
  - **数据落库**：码表维度发布后，会自动将码表的数值填入维度表中。
- **模型下线流程**：选择当模型下线时，是否同步删除技术资产、业务资产、质量作业、数据开发作业。
- **数据表更新方式**：当数据架构中的表在发布后进行了修改，是否同时更新数据库中的表。默认为“不更新”，但在配置中心可以依据自己的需求设置更新动作。依据 DDL 模板，在模板里面配置对应的更新语句即可。
  - **不更新**：不更新数据库中的表。

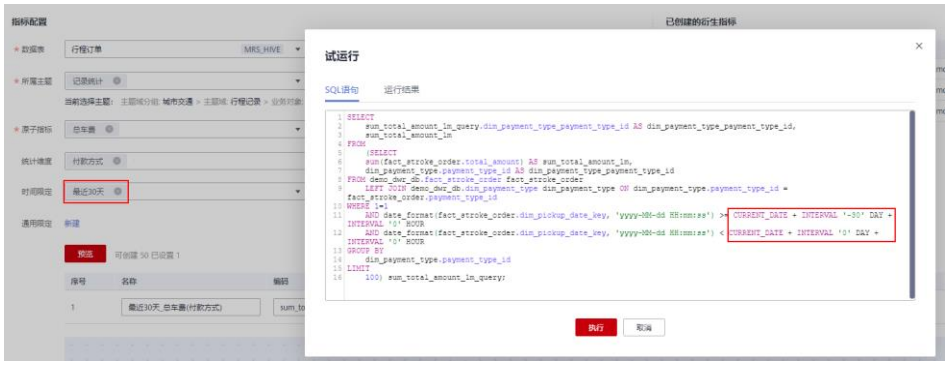
- **依据 DDL 更新模板**：依据 DDL 模板管理中配置的 DDL 更新模板，更新数据库中的表，但能否更新成功是由底层数仓引擎的支持情况决定的。由于不同类型的数仓支持的更新表的能力不同，在数据架构中所做的表更新操作，如果数仓不支持，则无法确保数据库中的表和数据架构中的表是一致的。例如，DLI 类型的表更新操作不支持删除表字段，如果在数据架构的表中删除了表字段，则无法在数据库中相应的删除表字段。

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

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

- **数据表不区分大小写**：对于选中的连接类型，在发布相应类型的表时，同步技术资产时名称将不区分大小写，找到相同的即认为已存在。
- **业务表映射使用新版本**：系统默认为新版本映射。新版本映射功能支持 join 等操作，推荐使用新版本映射。
- **汇总表自动汇聚**：发布衍生指标或复合指标时，系统支持自动生成汇总表，一个统计维度对应一个汇总表。自动生成的汇总表可在汇总表页面下选择“自动汇聚”页签查看。
- **数据标准是否重名**：默认关闭，打开后数据标准可以重名。
- **时间限定生成使用动态表达式**：开关打开后，则使用动态时间表达式；如开关关闭，则默认使用原有的静态时间表达式。例如时间限定设置为最近 30 天：如果使用静态表达式，如果当前为 9 月，生成的最近 30 天的数据就是 8 月，即使当前到了 10 月，生成的数据还是 8 月，不能自动更新；如果使用动态表达式，当前到了 10 月，最近 30 天自动更新为 9 月。动态表达式时间函数举例如下所示：

图3-134 动态表达式



说明

如果第一次打开开关，需重置 DLL 模板中的衍生指标。如之前有修改过 DLL 模板，请先做好模板备份。重置模板会将原来修改过的模板覆盖，重置后需要将原来修改的内容重新编辑一次。

- **信息架构页面表查询时，主题支持并列查询个数**：默认为 1 个，暂不支持设置。
- **码表数据落库并行行数**：码表维度发布后，设置将码表的数值填入维度表中时的并行操作行数。当码表数值较多时，会导致落库失败，可以适当调小该参数。
- **码表生成质量规则**：下拉选择即可。当码表的数据量较小时，选择“枚举值校验”即可；否则选择字段一致性校验。

模型配置

当您在主题设计、模型设计等过程中，如果需要进行如下操作，您可以通过本页面进行配置：

- 增加主题别名、表模型别名、字段别名。
- 设置维度（维度表）、事实表、汇总表的默认表编码前缀。
- 增加表的自定义字段。
- 增加属性的自定义字段。

图3-135 模型配置

在数据架构控制台，单击左侧导航树中的“配置中心”，进入相应页面后再单击“模型配置”页签。

- 启用别名。在“模型配置”页面，您可以增加别名。
  - 选项说明如下：
    - 主题设计：选择之后，在新建、编辑主题时，必须输入别名。
    - 表模型：选择之后，在新建、编辑表时，必须输入别名。会影响业务表、维度（维度表）、事实表和汇总表等。
    - 字段：选择之后，在新建、编辑表字段时，必须输入别名。
- 表名管理。设置维度（维度表）、事实表和汇总表的默认表编码前缀。
- 表自定义项。在新建、编辑表时，可以在表的基本设置中设置自定义的字段。会影响业务表、维度（维度表）、事实表和汇总表等。

字段类型管理

当您执行新建表、逆向数据库或模型转换等操作时，如果系统默认的数据类型或不同数据源之间的数据类型映射关系无法满足需求，您可以增加、删除或修改数据类型。系统默认的数据类型不支持删除。

步骤 1 在数据架构控制台，单击左侧导航树中的“配置中心”，进入相应页面后再单击“字段类型”页签。

步骤 2 在“字段类型”页面，您可以查看数据类型及不同数据源之间的数据类型映射关系，其中“创建人”为 SYSTEM 的类型为系统默认的字段类型。

类型分组说明如下：

- DEFAULT：通用数据类型，未指定数据源类型时建表所用的字段类型。例如，新建逻辑模型的表时，就是使用 DEFAULT 分组中的数据类型。
- DLI：DLI 连接类型的表的数据类型。
- DWS：DWS 连接类型的表的数据类型。
- MRS_HIVE：MRS_HIVE 连接类型的表的数据类型。
- POSTGRESQL：POSTGRESQL 连接类型的表的数据类型。

图3-136 字段类型

步骤 3 管理字段类型。

- **新建类型**

  如果要增加数据类型，单击“新建”按钮。在弹出对话框中，配置如下参数，然后单击“确定”。

  图3-137 新建类型

表3-110 基本配置

参数名称	说明
类型分组	选择新建类型所属的类型分组。
名称	数据类型的名称。只能包含中文、英文字母、数字、左右括号、空格和下划线，且以中文或英文字母开头。
编码	数据类型的编码，必须为数仓支持的类型。只能包含大写英文字母和下划线。
所属域	选择新建类型所属的域。
数仓对应类型	选择新建类型所映射连接的数据类型。
DLI	选择新建类型所映射的 DLI 连接的数据类型。
DWS	选择新建类型所映射的 DWS 连接的数据类型。
MRS_HIVE	选择新建类型所映射的 MRS_HIVE 连接的数据类型。
POSTGRESQL	选择新建类型所映射的 POSTGRESQL 连接的数据类型。
是否有拓展	对于某些数据类型，需要设定数据的长度范围时，可以打开“是否有拓展”开关，并配置对应的拓展。 例如高精度数据类型 DECIMAL(p,s)，需要分别指定小数的最大位数(p)和小数位的数量(s)，则数据类型 DECIMAL 的默认拓展可填写为“(10,2)”，指的是小数点左侧的位数为 2，小数点右侧的最大位数为 10-2=8；又如数据类型 VACHAR 也需要指定位数，当默认拓展填写为“10”，指的是最大长度为 10 字符。

- **编辑类型**

  在字段类型列表中，找到需要编辑的字段类型，然后单击  按钮进行编辑，参数说明请参见表 3-110。

- **删除类型**

  仅支持对于用户新建的数据类型进行删除操作。“创建人”为 SYSTEM 的类型为系统默认的字段类型，不支持删除操作。

  在字段类型列表中，找到需要删除的字段类型，单击  按钮，然后在弹出对话框中单击“确定”完成删除。

- **重置**

  单击“字段类型”页面底部的“重置”按钮，可恢复系统默认配置。

----结束

DDL 模板管理

在 DataArts Studio 数据架构中，支持修改各种类型（例如 DLI、POSTGRESQL、DWS、Hive）的表或 DLI 视图的 DDL 模板。如果您需要将已创建的某一类型的表生成其他数据源的 DDL 语句，您就可以根据目标数据源的 DDL 语法，修改该类型的表的 DDL 模板。

1. 在数据架构控制台，单击左侧导航树中的“配置中心”，进入相应页面后再单击“DDL 模板管理”。
2. 在“DDL 模板管理”页面，您可以配置各种类型的表或 DLI 视图的 DDL 模板，您可以参考该页面中的“填写说明”修改 DDL 模板，修改完成后单击“确定”。如果单击“重置”可恢复默认设置。

   如图 3-138 所示，说明如下：

   - 新建：可查看或编辑新建表或 DLI 视图的 DDL 模板。
   - 更新：可查看或编辑更新表或 DLI 视图的 DDL 模板。
   - 删除：可查看或编辑删除表或 DLI 视图的 DDL 模板。
   - 衍生指标：可以查看或编辑衍生指标的 SQL 模板。
   - 复合指标：可以查看或编辑复合指标的 SQL 模板。
   - 汇总表：可以查看或编辑汇总表的 SQL 模板。
   - “参考数据”区域：显示了一个表详情的示例，示例中的变量定义了表的详细信息。
   - “模板代码编辑”区域：可以编辑 DDL 模板。如果您需要将所选类型的表，生成其他类型的数据库的 DDL 语句，您可以根据目标数据源的 DDL 语法，修改 DDL 模板。
   - “预览结果”区域：编辑 DDL 模板后，可以预览按模板生成的 DDL 语句。

   图3-138 DDL 模板管理

编码规则

1. 在数据架构控制台，单击左侧导航树中的“配置中心”，然后再选择“编码规则”页签。
2. 管理编码规则。

- 添加编码规则

  如果需要自定义编码规则，在“编码规则”列表上方，单击“添加”，在弹出对话框中，配置如下参数，然后单击“确定”。

图3-139 添加编码规则

表3-111 添加编码规则说明

参数名称	说明
类型	选择编码规则的类型，当前支持如下四种： 业务指标，逻辑实体，逻辑属性， 数据标准。
生效范围	生效范围默认是全局。可以选择 主题、流程、码表、数据标准下一级路径。
系统规则	是否为系统规则。自定义的编码规则系统预置为否，不能修改。
编码规则	采用前缀+数字码的方式，不能修改。
前缀	可以是“英文字符”+“数字”的方式，但不能以数字结尾。支持修改。

参数名称	说明
数字码	支持顺序码和随机码两种方式。
起始码	数字码范围的起始值。
结束码	数字码范围的终止值。
编码示例	根据前缀动态修改后，可以更新展示。

- 删除编码规则

  如果需要删除自定义编码规则，在“编码规则”列表勾选待删除的编码规则，单击列表上方的“删除”，在弹出对话框中，单击“是”即可删除。

说明

系统预置的四个编码规则（逻辑实体、 数据标准、逻辑实体属性、业务指标），不可以删除。

- 编辑编码规则

  如果需要修改自定义编码规则，单击“编码规则”列表中待修改编码规则的“编辑”，弹出“修改编码规则”对话框，修改完成后，单击“确定”。

3.4.4 数据调研

3.4.4.1 流程设计

流程架构基于价值流产生，属于业务架构的流程处理模块，指导并规范 BT&IT 需求的管理，确保业务需求受理、分析、交付等过程的高效运作；并聚焦高价值需求，实现业务价值最大化，支撑业务运作及目标的达成。

新建流程

根据业务需求设计流程，流程支持 L1~L3 三层。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-140 选择数据架构

2. 单击左侧导航栏中“流程设计”，进入流程设计页面，在流程树中选中一个流程，单击 ⊕ 按钮在所选流程下新建流程。首次新建流程时，可选择在流程的根节点下新建流程。

图3-141 流程设计

3. 在弹出对话框中配置如下参数，然后单击“确定”完成流程的创建。

图3-142 新建流程

表3-112 新建流程参数说明

参数名	说明
流程名称	流程名称，只能包含中文、英文字母、数字和下划线。
责任人	流程的责任人，可以手动输入名字或直接选择已有的责任人。
上级流程	选择所属的上级流程。
描述	流程的描述信息。

4. 依次新建更多的流程或子流程。一般需要设计 L1~L3 三层流程。第一层标识为 L1 层，第二层标识为 L2 层，第三层标识为 L3。

   示例如下：

图3-143 流程设计示例

导出流程

您可以将数据架构中已创建的流程导出到文件中。

步骤 1 在数据架构控制台，单击左侧导航树中的“流程设计”，进入流程设计页面。

步骤 2 单击流程列表上方的“导出”按钮，等待几秒钟后，页面右上角提示“流程导出成功”，可以查看导出的流程。

说明

“主题”或“流程”作为层级联动性质，导出均默认为全量导出，不支持筛选。

----结束

导入流程

步骤 1 在数据架构控制台，单击左侧导航树中的“流程设计”，进入流程设计页面。

步骤 2 单击流程列表上方的“导入”按钮导入流程。

步骤 3 在“导入流程”对话框中，根据页面提示配置如下参数，然后先单击“添加文件”后，再单击“上传文件”。

图3-144 导入流程

表3-113 导入配置参数说明

参数名	说明
更新已有数据	如果所要导入的流程，在 DataArts Studio 数据架构中已经存在，是否更新已有的流程。支持以下选项： • **不更新**：当流程已存在时，将直接跳过，不处理。 • **更新**：当流程已存在时，更新已有的流程信息。 在导入流程时，只有创建或更新操作，不会删除已有的流程。
上传模板	选择所需导入的流程设计文件。 所需导入的流程设计文件，可以通过以下两种方式获得。 • **下载流程模板并填写模板** 在“导入配置”页签内，单击“下载流程模板”下载模板，然后根据业务需求填写好模板中的相关参数并保存后，先添加再上传，完成模板上传。模板参数的详细描述请参见表 3-114。 • **导出的流程** 您可以将某个 DataArts Studio 实例的数据架构中已建立的流程设计信息导出到 Excel 文件中。导出后的文件可用于导入。导出流程的操作请参见导出流程。

下载的流程模板参数如表 3-114 所示，其中名称前带“*”的参数为必填参数，名称前未带“*”的参数为可选参数。一个流程需要填写一条记录。

表3-114 流程导入参数说明

参数名	说明
上级流程	第一层的流程，其上级流程为空，不用填。 非第一层的流程，其上级流程不能为空。上级流程为多级流程时，流程之间以“/”分隔。例如“集成产品开发/开发生命周期”。
*名称	流程名称。
*责任人	流程的责任人，可以手动输入名字或直接选择已有的责任人。
描述	流程的描述信息。

步骤 4 导入结果会在“导入流程”对话框的“上次导入”中显示。如果导入结果为“成功”，单击“关闭”完成导入。如果导入失败，您可以在“备注”列查看失败原因，将模板文件修改正确后，再重新上传。

----结束

删除流程

您可以将无用的流程删除，注意，删除后无法恢复，请谨慎操作。当流程下面存在子流程时，需先删除子流程。

步骤 1 在数据架构控制台，单击左侧导航树中的“流程设计”，进入流程设计页面。

步骤 2 在流程列表中，选中要删除的流程，然后单击上方的“删除”按钮。

步骤 3 在弹出的“删除流程”对话框中，确认删除流程信息正确后，单击“是”删除流程。

----结束

3.4.4.2 主题设计

主题设计是通过分层架构表达对数据的分类和定义，帮助厘清数据资产，明确业务领域和业务对象的关联关系。

您可以通过以下两种方式进行主题设计：

- 新建主题

  手动新建主题。

- 导入主题设计信息

  如果主题信息比较复杂，建议采用导入方式批量导入主题信息。

  - 您可以下载系统提供的主题设计模板，在模板文件中填写主题的相关参数后，使用模板批量导入主题信息。
  - 您可以预先将某个 DataArts Studio 实例的数据架构中已建立的主题设计信息导出到 Excel 文件中。导出后的文件可用于导入。导出主题设计信息的操作，请参见导出主题设计信息。

建立好主题设计信息后，可以对主题信息进行查找、编辑或删除操作，详情请参见管理主题设计。

主题设计概述

默认情况下，系统预设了“L1-主题域分组”、“L2-主题域”和“L3-业务对象”三层主题层级。

- **主题域分组**：主题域分组是基于业务场景对主题域进行分组。
- **主题域**：主题域是根据数据的性质对数据进行划分，性质相同的数据划分为一类，其划分后得出的各数据集合叫做主题域，主题域是信息需求范围的上层级数据集合。
- **业务对象**：业务对象是指企业运作和管理中不可缺少的重要人、事、物等信息。

您也可以根据您的实际情况，参考主题配置对主题层级进行自定义配置。

新建主题

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

   图3-145 选择数据架构

2. 单击左侧导航栏中“主题设计”，进入主题设计页面，单击左上角的“新建”。

图3-146 主题设计

3. 在“新建主题域分组”对话框中，配置如下参数，然后单击“确定”完成主题域分组的创建。

表3-115 主题域分组参数说明

参数名	说明
*名称	只允许除/、\、<、>以外的字符。
*编码	英文名称。只允许字母、数字、空格、下划线、中划线、左右括号以及&符号。
别名	只允许除\、<、>以外的字符。 说明 您需提前在配置中心的“模型配置”页签中启用主题设计别名，这里才可配置别名。
*上级主题	选择所属的上级主题。
数据 owner 部门	数据的拥有者所在部门。
*数据 owner 人员	在下拉框中选择需要的数据 owner 人员，支持多选和自定义输入。
描述	主题域分组的描述信息。

图3-147 新建主题域分组

4. 在一个主题下，还可以新建多个主题。

   主题层级数目由用户在配置中心的主题层级中自定义，系统默认有三个层级，从上到下分别命名为主题域分组（L1）、主题域（L2）、业务对象（L3）。

导入主题设计信息

步骤 1 在数据架构控制台，单击左侧导航树中的“主题设计”，进入主题设计页面。

步骤 2 单击左上方的“导入”按钮，弹出导入主题对话框。

图3-148 导入主题设计

步骤 3 在“导入主题”对话框中，根据页面提示配置如下参数，然后先单击“添加文件”后，再单击“上传文件”。

图3-149 导入配置

表3-116 导入配置参数说明

参数名	说明
更新已有数据	在导入时是否更新已有的主题信息（主题域分组、主题域或业务对象）。在导入时，系统将按编码判断将要导入的主题信息在系统中是否已存在。 • **不更新**：当主题信息已存在时，将直接跳过，不更新。 • **更新**：当主题信息已存在时，更新已有的主题信息。 在导入主题信息时，只有创建或更新操作，不会删除已有的主题信息。
上传模板	选择所需导入的主题设计文件。 所需导入的主题设计文件，可以通过以下两种方式获得。 • **下载主题导入模板并填写模板** 在“导入配置”页签内，单击“下载主题导入模板”下载模板，然后根据业务需求填写好模板中的相关参数并保存。模板参数的详细描述请参见表 3-117。 • **导出的主题设计信息** 您可以将某个 DataArts Studio 实例的数据架构中已建立的主题设计信息导出到 Excel 文件中。导出后的文件可用于导入。关于导出主题设计的更多信息，请参见导出主题设计信息。

下载的主题导入模板参数如表 3-117 所示，其中名称前带“*”的参数为必填参数，名称前未带“*”的参数为可选参数。一个主题对象需要填写一行信息。

表3-117 模板参数说明

参数名	说明

参数名	说明
上级主题	上层主题的编码路径，以/分隔。
*名称	中文名称。只允许除/、\、<、>以外的字符。
*编码	英文名称。只允许英文字母、数字、空格、下划线、中划线、左右括号以及&符号。
别名	主题对象的别名。
描述	主题对象的描述信息。 对于最低层级主题，此项参数为必选。您在导入文件中应补充最低层级主题的描述信息。
数据 owner 部门	数据的拥有者所在部门。 对于最低层级主题，此项参数为必选。您在导入文件中应补充最低层级主题的数据 owner 部门信息。
*数据 owner 人员	数据的拥有者，支持填写多个，中间以逗号分隔。

步骤 4 导入结果会在“上次导入”页面中显示。如果导入成功，单击“关闭”完成导入。如果导入失败，您可以查看失败原因，将模板文件修改正确后，再重新上传。

图3-150 上次导入页面

----结束

导出主题设计信息

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航树中的“主题设计”，进入主题设计页面。

步骤 2 单击左上方的“导出”将当前已有的主题设计导出到 Excel 文件中。导出后的文件可用于导入。

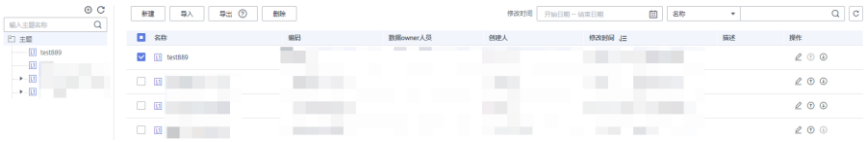
说明

“主题”或“流程”作为层级联动性质，导出均默认为全量导出，不支持筛选。

----结束

管理主题设计

图3-151 主题设计区域



- 查找

  您可以在主题的搜索框中，输入所需查找的关键字进行查找。
- 编辑

  您可以在主题列表中，选择一个对象，然后单击其名称右侧的 按钮进行编辑。
- 删除

  您可以在主题列表中，选择一个对象，单击上方“删除”。
- 上移/下移

  您可以在主题列表中，选择一个对象，然后单击其名称右侧的 按钮进行下移，或单击其名称右侧的 按钮进行上移。

3.4.5 标准设计

3.4.5.1 新建码表

码表，也称 lookup 表、数据字典表，一般由中英文名称编码组成，由可枚举数据构成，存储枚举数据名称与编码的映射关系。码表的作用主要有：

- 在数据清洗中用于标准化业务数据以及补充映射字段。
- 在质量监控中用于监控业务数据的值域范围。
- 在维度建模中可以引申为枚举维度。

新建码表并发布

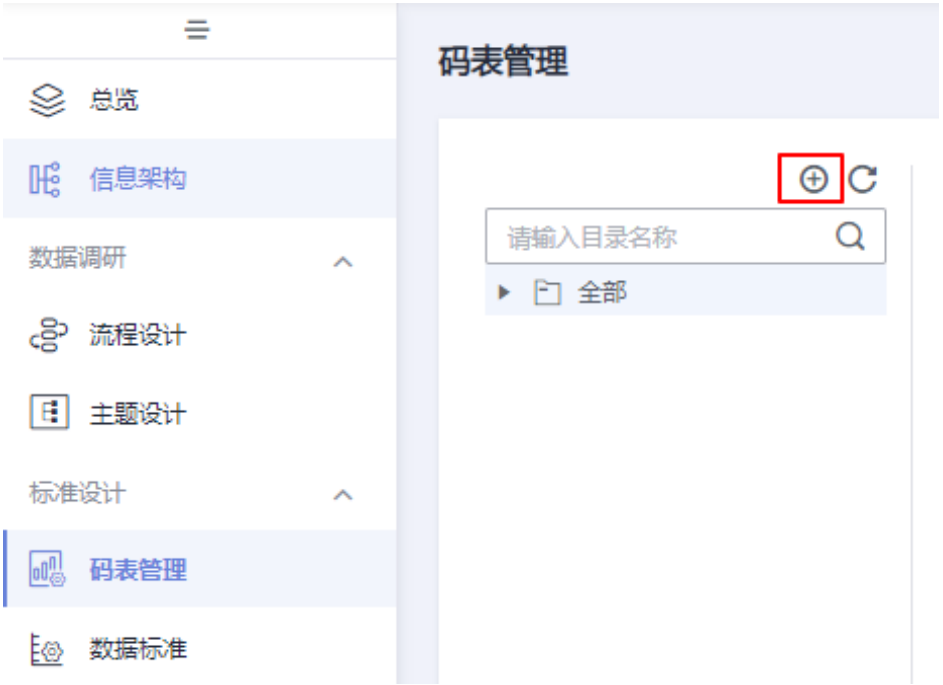
手动新建码表，完成新建后可以参考填写数值到码表中添加码表记录。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-152 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“码表管理”。
3. 在“码表管理”页面的码表目录树中，选择一个目录，然后单击 ⊕ 按钮在所选目录下新建目录。首次新建目录时，可选择在根目录下新建目录。

图3-153 码表管理页面



4. 在弹出窗口中进行参数配置，单击“确定”。

图3-154 新建码表目录

表3-118 参数描述

参数名称	说明
目录名称	只能包含中文、英文字母、数字和下划线。
选择目录	在已有的目录中选择一个目录，新建的目录将创建在所选择的目录中。

5. 在目录树中单击刚建好的目录，然后单击“新建”按钮新建一个码表。
6. 在“新建码表”页面中，做如下配置：
   在“基础配置”区域，配置如下参数：

图3-155 基础配置

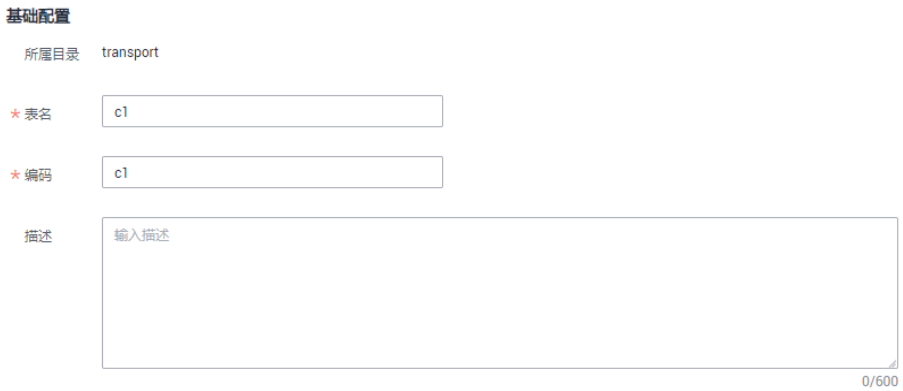


表3-119 基础配置

参数名称	说明
表名	码表名称。只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文或英文字母开头。
编码	码表的英文名称。只能包含英文字母、数字和下划线，且以英文字母开头。
描述	描述信息。支持的长度为 0~600 个字符。

在“建表配置”中添加所需要的表字段，单击“新建”或 可以添加新的字段，单击某个字段后的 按钮可删除该字段。

图3-156 建表配置

7. 单击“发布”，在提交发布对话框中，选择审核人，再单击“确认提交”提交审核。审核通过后，返回“码表管理”页面，在列表中可以查看已建好的码表且状态显示为“已发布”，已发布的码表才可被使用。

说明

如果当前用户已被添加为审核人，则可以勾选“自助审批”，单击“确认提交”后，码表状态显示为“已发布”。

填写数值到码表中

对于已创建的码表，您可以通过填写数值，增加码表记录。

步骤 1 在数据架构控制台，单击左侧的“码表管理”，进入码表管理页面。

步骤 2 在码表列表，找到所需要的码表，单击其所在行的“更多 > 填写数值”。

步骤 3 进入相应页面后，单击“新建”，并在弹出窗口中设置各字段的值。

图3-157 填写数值

步骤 4 完成后单击“确定”。或者您也可以单击“确定并继续”继续添加更多码表记录。

----结束

导入码表

Excel 导入码表时，码表名称需要限制在 32 个字符以内。

通过导入码表，可以导入新的码表，也可以往已有的码表中批量导入码表记录。如果码表记录比较多，建议采用导入方式。

步骤 1 在数据架构控制台，单击左侧的“码表管理”，进入码表管理页面。

步骤 2 在左侧的目录树中，选择一个目录，再单击“更多 > 导入”。您也可以在所选择的码表目录上单击鼠标右键，然后选择菜单“导入”。

图3-158 码表页面

步骤 3　在“导入码表”对话框中，根据页面提示配置参数，然后单击“上传文件”。

图3-159 导入码表

表3-120 导入配置参数说明

参数名	说明
更新已有表	在导入时是否更新已有的码表信息。在导入时，系统将按编码进行判断将要导入的码表在系统中是否已存在。支持以下选项： • **不更新**：当码表已存在时，将直接跳过，不更新。 • **更新**：当码表已存在时，更新已有的码表信息。如果码表处于“已发布”状态，码表更新后，您需要重新发布码表，才能使更新后的表生效。 在导入码表时，只有创建或更新操作，不会删除已有的码表。
上传模板	选择所需导入的码表文件。所需导入的码表文件，可以通过以下两种方式获得。 • **下载码表模板并填写模板**

参数名	说明
	在“导入配置”页签内，单击“下载码表导入模板”下载模板，然后根据业务需求填写好模板中的相关参数并保存。模板参数的详细描述，请参见表 3-121。 码表模板填写说明： – 模板中参数名称前带“*”的参数为必填参数，名称前未带“*”的参数为可选参数。 – 一个码表可以添加多个字段。 – 如果要导入多个码表，可以在模板文件中添加多个 Sheet 页，Sheet 页的名称即为码表名称。 – 如果码表名称已存在，当“更新已有数据”设置为“更新”时，导入时会更新已有的码表。 – 如果码表名称不存在，导入时会新建该码表。 • **导出的码表文件** 您可以将某个 DataArts Studio 实例的数据架构中已创建的码表导出到 Excel 文件中。导出后的文件可用于码表导入。码表导出操作请参见管理码表。

表3-121 码表导入模板参数

参数名称	说明
目录	码表所属的目录。多级目录以“/”分隔，例如“dir01/dir02”。
*表名称	码表名称。只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文或英文字母开头。
*表编码	码表的英文名称。只能包含英文字母、数字、下划线，且以英文字母开头。
表描述	码表的描述信息。支持的长度 0~600 个字符。
*字段名称	字段名称。只能包含中文、英文字母、数字、左右括号、空格、中划线和下划线，且以中文或英文字母开头。
*字段编码	字段编码。只能包含英文字母、数字、下划线，且以英文字母开头。
*字段数据类型	支持的数据类型有：STRING、BIGINT、DOUBLE、TIMESTAMP、DATE、BOOLEAN、DECIMAL。
字段描述	字段的描述信息。支持的长度 0~600 个字符。
是否生成标准	• true：生成数据标准。 • false：不生成数据标准。默认为 false。 注意：如果要自动生成数据标准，还需在“配置中心 > 标准模板

参数名称	说明
	管理”中勾选上“引用码表”选项。

如果导入时，需要同时导入码表记录，请在码表导入模板中新建一个命名为码表名称的 Sheet 页，并在该 Sheet 页中增加码表字段，每个字段为一列，列名由字段名称、换行、字段编码组成，然后再填写所需导入的码表数值。如果码表导入模板中已有码表名称的 Sheet 页，则无需再新建该 Sheet 页，您可以直接在该 Sheet 中填写所需导入的码表数值。

步骤 4 导入结果会在“上次导入”页面中显示。如果导入成功，单击“关闭”完成导入。如果导入失败，您可以查看失败原因，将模板文件修改正确后，再重新上传。

----结束

通过逆向数据库导入码表

通过逆向数据库，您可以从其他数据源中将一个或多个已创建的数据库表导入到码表目录中，使其变成码表。

步骤 1 在数据架构控制台，单击左侧导航树中的“码表管理”，进入码表管理页面。

步骤 2 在左侧的码表目录树中，选中一个目录，然后在码表列表上方，单击“逆向数据库”。

步骤 3 在“逆向数据库”对话框中，配置如下参数，然后单击“确定”。

表3-122 逆向数据库配置

参数名称	说明
数据连接类型	在下拉列表中将显示逆向数据库支持的数据连接类型，请选择所需要的数据连接类型。
数据连接	选择数据连接。 如需从其他数据源逆向数据库到码表目录中，需要先在 DataArts Studio 管理中心创建一个数据连接，以便连接数据源。创建数据连接的操作，请参见 3.2.2 创建数据连接。
数据库	选择数据库。
Schema	下拉选择 Schema。该参数仅 DWS 模型的表有效。
队列	DLI 队列。仅当“数据连接类型”选择“DLI”时，该参数有效。
更新已有表	如果从其他数据源逆向过来的表，在码表中已存在同名的表，选择是否更新已有的码表。
逆向表数据	• **不逆向**：逆向数据库时，将表导入到码表目录中，但是不导入表数据。您可以在完成逆向数据库后，参考填写数值到码表中添加记录到码表中。 • **覆盖**：逆向数据库时，将表导入到码表目录中，同时将表数据

参数名称	说明
	导入到该码表中。
数据表	选择一个或多个需导入的数据表。

步骤 4 逆向数据库的结果会在“上次逆向”页面中显示。如果逆向成功，单击“关闭”。如果逆向失败，您可以查看失败原因，问题解决后，选中失败的表，然后单击“重新逆向”进行重试。

图3-160 逆向结果

----结束

导出码表

Excel 导出码表时，码表名称需要限制在 32 个字符以内。

步骤 1 在数据架构控制台，单击左侧导航树中的“码表管理”，进入码表管理页面。

步骤 2 导出码表。

- **导出码表**

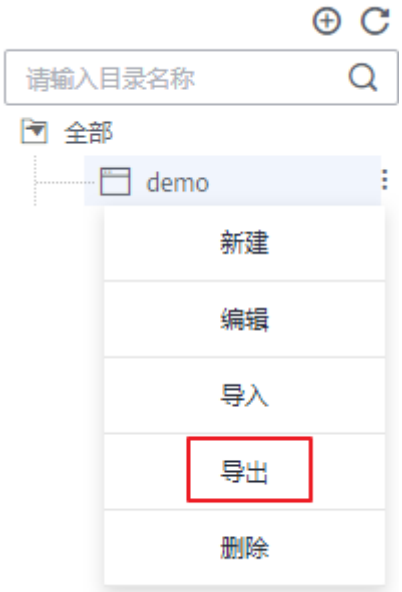
  在码表列表中，选中所需导出的码表，然后单击“更多 > 导出”。

图3-161 码表列表

- **导出码表目录中的所有表**

  在码表目录树中，选中一个目录，单击鼠标右键，选择“导出”菜单。

  图3-162 导出码表目录



----结束

删除码表

码表被删除后，将无法恢复，请谨慎操作。删除码表时，如果码表为待发布、已发布或待下线状态，则无法删除。您需要对码表进行操作，使其变为其他状态时，才能删除该码表。

步骤 1 在数据架构控制台，单击左侧导航树中的“码表管理”，进入码表管理页面。

步骤 2 在码表列表中，选择要删除的码表，然后在列表上方单击“更多 > 删除”。

步骤 3 在弹出的确认对话框中，单击“是”进行删除。

----结束

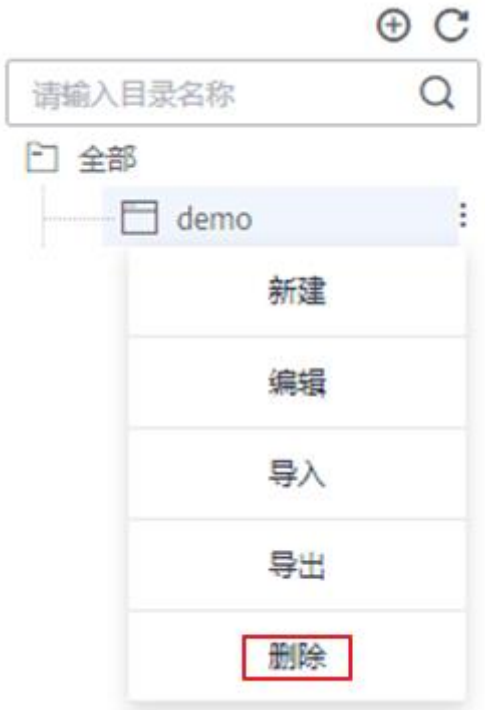
删除码表目录

删除码表目录时，如果该目录或其子目录包含码表，则无法删除。您需要先删除其中的码表后，才能删除该目录。

步骤 1 在数据架构控制台，单击左侧导航树中的“码表管理”，进入码表管理页面。

步骤 2 在左侧码表目录树中，选择要删除的目录，单击鼠标右键，选择“删除”菜单。

图3-163 管理码表目录



步骤 3 在弹出的确认对话框中，单击“是”进行删除。

----结束

管理码表

建立好码表后，可以对码表进行查找、编辑、下线或删除等操作。

在数据架构控制台，单击左侧导航树中的“码表管理”，进入码表管理页面。您可以对码表进行管理。

图3-164 码表管理

- **编辑**

  在码表列表中，找到所需要的码表，单击其所在行的“编辑”，即可编辑指定的码表。

- **发布**

  在码表列表中，对于状态为“草稿”或“已驳回”的码表，单击其所在行的“发布”，并在弹出框中选择审核人并单击“确认提交”，即可发布该码表提交审核。等待审核人员审核通过后，码表就发布成功了。

- **下线**

  在码表列表中，对于状态为“已发布”的码表，单击其所在行的“更多-下线”，并在弹出框中选择审核人并单击“确认提交”，即可提交下线申请。等待审核人员审核通过后，码表就下线成功了。

- **填写数值**

  在码表列表中，找到所需要的码表，单击其所在行的“更多-填写数值”，可以快速设置各字段的值。

- **发布历史**

  在码表列表中，找到所需要的码表，单击其所在行的“更多-发布历史”，可以查看码表的发布历史和变更详情，并支持进行版本对比。

3.4.5.2 新建数据标准

数据标准是用于描述公司层面需共同遵守的数据含义和业务规则，它描述了公司层面对某个数据的共同理解，这些理解一旦确定下来，就应作为企业层面的标准在企业内被共同遵守。

数据标准，也称数据元，由一组属性规定其定义、标识、表示和允许值的数据单元，是不可再分的最小数据单元。您可以将数据标准关联到各个业务上的数据库中。其中，标识符、数据类型、表示格式、值域是数据交换的基础，它们用于描述表的字段元信息，规范字段所存储的数据信息。

本章节介绍如何创建数据标准，创建好的数据标准，可用于在关系建模中新建业务表时与业务表中的字段相关联，从而约束业务表中的字段遵从指定的数据标准。

新建数据标准目录

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-165 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“数据标准”。
3. 首次进入数据治理中心的数据标准页面，会显示制定数据标准模板的页面，在“可选项”中勾选所需要的选项，添加自定义项，完成后单击“确定”。

   保存模板后，如需修改，您也可以进入“配置中心 > 标准模板管理”页面修改模板，详情请参见标准模板管理。在新建数据标准时，将需要设置此处模板中选中的选项。
4. 在“数据标准”页面，在目录树上，单击一个目录，然后单击 ⊕ 按钮在该目录下新建一个目录。首次新建目录时选择在根目录下新建目录。

图3-166 数据标准页面

5. 在弹出窗口中配置如下参数，然后单击“确定”。

图3-167 新建数据标准目录

表3-123 参数描述

参数名称	说明

参数名称	说明
目录名称	只能包含中文、英文字母、数字和下划线。
选择目录	在已有的目录中选择一个目录，新建的目录将创建在所选择的目录中。

新建数据标准

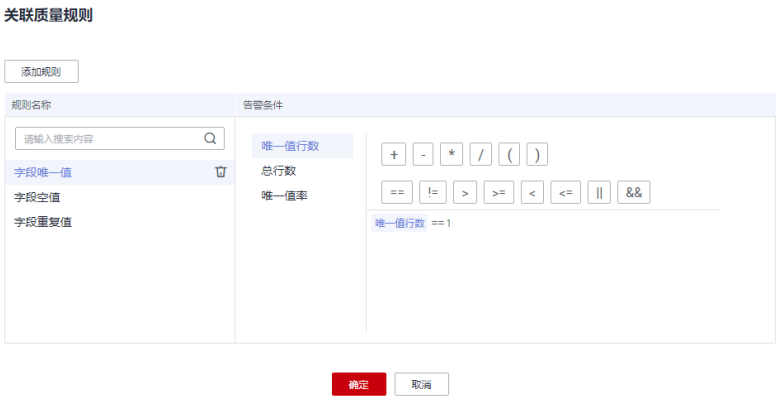
步骤 1 在“数据标准”页面的目录树中，选择一个目录，然后单击“新建”按钮新建一个数据标准。

步骤 2 在新建数据标准页面中，请参考表 3-124 配置参数，然后单击“发布”。

在新建数据标准页面中，仅显示在“配置中心 > 标准模板管理”中已勾选的参数和已添加的自定义参数。表 3-124 中所示为选中数据标准模板中的所有参数并添加了一个自定义参数的场景。有关配置数据标准模板的详细信息，请参见标准模板管理。

表3-124 数据标准参数说明

参数名称	说明
标准名称	只能包含中文、英文字母、数字、左右括号、空格、中划线和下划线，且以中文或英文字母开头。如果未开启“数据标准是否重名”，需要确保标准名称在本工作空间内唯一。请在“数据架构”模块，“配置中心”的“功能配置”页签下查看“数据标准是否重名”是否开启。
标准编码	支持自动生成和自定义两种方式。 自定义的标准编码要求本工作空间内唯一，用于唯一标识一条数据标准记录。详情参考表 3-111。
数据类型	数据类型有：STRING、BIGINT、DOUBLE、TIMESTAMP、DATE、BOOLEAN、DECIMAL。 不同的系统数据类型可能存在差异，系统内部会做类型转换。如果未找到所需要的数据类型，您可以参考字段类型管理添加数据类型。
数据长度	设置数据长度： • 可以为空。数据长度为空时，对数据长度不做限制。 • 可以设置为具体的数值。输入 1~10000 之间的数值。 • 可以设置为一个范围。输入数据范围的临界值。 如果设置了数据长度标准，当数据类型为 STRING 时，会为关联该标准的属性创建数据质量作业，其他类型暂不支持创建质量作业。
是否有允许值	当开启时，请输入允许值。
引用码表	选择已创建的码表并选择相应的“码表字段”，这样就可以将码表字段和数据标准相关联。如果未创建码表，请参见 3.4.5.1 新建码

参数名称	说明
	表进行创建。在“配置中心 > 功能配置”页面中的“模型设计业务流程步骤 > 创建质量作业”勾选的情况下，当引用码表的数据标准被关系建模的业务表关联后，如果表发布成功，系统将会在DataArts Studio 数据质量中自动创建一个质量作业，并根据数据标准以及码表分别生成相应的质量规则。如果当前表已经发布已有质量作业，则系统会自动更新质量作业，新增根据数据标准以及码表生成的质量规则。
质量规则	在“配置中心 > 功能配置”页面中的“模型设计业务流程步骤 > 创建质量作业”勾选的情况下，数据标准设置质量规则后，如果将表关联该数据标准，表在发布成功后，系统将会在 DataArts Studio 数据质量中自动创建一个质量作业，质量作业中会包含此处添加的质量规则。如果当前表已经发布，则系统会自动更新质量作业。 单击  弹出“关联质量规则”对话框，单击“添加规则”进行设置。 例如，添加名称为“字段唯一值”规则，选中该规则后单击“确定”，在“告警条件”中输入告警条件表达式，然后按照此方法添加其他规则后，单击“确定”。 告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。 在此处的“关联质量规则”对话框中，每个质量规则的“告警参数”以按钮形式列出。 图3-168 关联质量规则界面 
业务规则责任人	在下拉框中选择业务规则责任人。该责任人为质量规则制定责任人，可以手动输入名字或直接选择已有的责任人。
数据监控责任人	在下拉框中选择数据监控责任人。该责任人为质量规则实施责任人，可以手动输入名字或直接选择已有的责任人。

参数名称	说明
标准层级	• global：全局级别。 • domain：非全局级别。
用户自定义字段	该配置项是在 DataArts Studio 数据架构的“配置中心 > 标准模板管理”中添加的自定义项。您可以根据实际情况添加一个或多个自定义项，名称可以自己定义。有关添加自定义项的更多信息，请参见标准模板管理。
描述	描述信息。支持的长度为 0~600 个字符。

步骤 3 单击“保存”，完成新建数据标准操作。

----结束

导入数据标准

步骤 1 在数据架构控制台，单击左侧的“数据标准”，进入数据标准页面。

步骤 2 在数据标准的目录结构中，选择一个指定的目录名称，然后单击上方“更多 > 导入”，弹出对话框如下图所示。

图3-169 导入数据标准

步骤 3 在导入配置页签内，选择是否“更新已有数据”。已有数据是通过标准编码唯一标识的，即如果导入模板中的某个标准编码在当前工作空间下已经存在，则系统会认为导入模板中标准编码所在的这组数据为已有数据。

步骤 4 在导入配置页签内，单击“下载数据标准导入模板”下载模板。打开模板，请根据业务需求填写好模板中的相关参数并保存。

模板中的参数说明如表 3-125、表 3-126 所示，其中名称前带“*”的参数为必填参数，名称前未带“*”的参数为可选参数。

表3-125 标准 Sheet 页参数说明

参数名称	说明
目录	导入的数据标准所属的目录。
*标准名称	数据标准的中文名称。只能包含中文、英文字母、数字、左右括号、空格、中划线和下划线，且以中文或英文字母开头。
*标准编码	支持自动生成和自定义两种方式。 自定义的标准编码要求本工作空间内唯一，用于唯一标识一条数据标准记录。详情参考表 3-111。
*数据类型	数据类型有：STRING、BIGINT、DOUBLE、TIMESTAMP、DATE、BOOLEAN、DECIMAL。 不同的系统数据类型可能存在差异，系统内部会做类型转换。如果未找到所需要的数据类型，您可以参考字段类型管理添加数据类型。
数据长度	可以为空，或者输入 1~10000 之间的数值。数据长度为空时，对数据长度不做限制。 如果输入了数据长度标准，当数据类型为 STRING 时，会为关联该标准的属性创建数据质量作业，其他类型暂不支持创建质量作业。
是否有允许值	true 表示有允许值，false 表示没有允许值。
允许值	当参数“是否有允许值”为 true 时，必须设置“允许值”。 支持添加多个允许值，最多支持 20 个。多个允许值之间以逗号分隔，例如“1,2,3”。
引用码表	填写已创建的码表名称。
码表字段	当“引用码表”不为空时，请设置该引用码表中的“码表字段”，这样就可以将码表字段和数据标准相关联。
业务规则责任人	填写业务规则责任人，可以手动输入名字或直接选择已有的责任人。
数据监控责任人	填写数据监控责任人，可以手动输入名字或直接选择已有的责任人。
标准层级	• global：全局级别。 • domain：非全局级别。
描述	描述信息。支持的长度 0~600 字符。
用户自定义字段（可选）	如果在定制数据标准模板时，您添加了一个或多个自定义字段，则在导入模板中也需要填写相应的字段，如果未添加自定义字段，则无需填写。关于定制数据标准模板的更多信息，请参见标准模板管理。

在“质量规则”Sheet 页中，可以配置数据标准所需添加的质量规则。在“配置中心 > 功能配置”页面中的“模型设计业务流程步骤 > 创建质量作业”勾选的情况下，数据标准设置质量规则后，如果将表关联该数据标准，表在发布成功后，系统将会在 DataArts Studio 数据质量模块中自动创建一个质量作业，质量作业中会包含此处添加的质量规则。如果当前表已经发布，则系统会自动更新质量作业。

表3-126 质量规则 Sheet 页参数说明

参数名称	说明
*标准编码	需要添加质量规则的数据标准编码
规则名称	填写已有的规则名称。在 DataArts Studio 控制台左上角的模块下拉列表中选择“数据质量”进入 DataArts Studio 数据质量控制台，然后您可以进入“规则模板”页面查看已有的规则名称。
告警配置	告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。 在告警条件表达式中，告警参数以${1}、${2}、${3}等变量名称表示，变量名即代表所指定的质量规则的告警参数，变量$1 代表第一个告警参数，$2 代表第二个告警参数，以此类推。在 DataArts Studio 控制台左上角的模块下拉列表中选择“数据质量”进入 DataArts Studio 数据质量控制台，然后您可以进入“规则模板”页面在“结果说明”一列中查看质量规则支持的告警参数。 例如：${1}>100
正则表达式	只有当“规则名称”配置为“正则表达式校验”或者“合法性校验”时，需要配置正则表达式。

步骤 5 返回“导入数据标准”对话框，选择上一步配置好的数据标准模板文件，然后单击“上传文件”。

如果上传的模板文件校验不通过，请修改正确后，再重新上传。

步骤 6 在导入对话框中，导入结果会在“上次导入”页面中显示。如果导入成功，单击“关闭”完成导入。如果导入失败，您可以查看失败原因，将模板文件修改正确后，再重新上传。

图3-170 上次导入结果

----结束

管理数据标准

在 DataArts Studio 数据架构控制台，单击左侧导航树中的“数据标准”，进入数据标准页面。您可以对数据标准进行管理。

图3-171 数据标准列表

在数据标准页面，可以执行以下操作：

- **搜索**

  在数据标准上方，设置标准、数据类型、创建人、审核人等筛选条件，然后单击“搜索”可以查找指定的数据标准。

  找到指定的数据标准后，可以执行以下操作：

  - 编辑
  - 发布
  - 下线

- **导入**

  单击“更多 > 导入”，可以导入数据标准，下载导入模板，填写模板并上传，然后单击“确定”。

- **导出**

- 导出指定目录中的数据标准

  在数据标准目录结构中，选中一个目录，单击数据标准列表上方的“更多 > 导出”，可以导出该目录下的所有的数据标准。

- 导出指定的数据标准

  在数据标准列表中，选中需要导出的数据标准，然后单击列表上方的“更多 > 导出”，可以导出所选中的数据标准。

- **删除**

  勾选标准后单击“更多 > 删除”，可以删除数据标准，其中待发布，已发布和待下线状态的数据标准不可被删除。且被引用的数据标准不可被删除。

- **发布**

  选中需要发布的数据标准，单击“发布”，弹出“提交发布”对话框，下列两种方式任选其一。

  - 选择审核人。如果下拉列表里无审核人，可单击旁边的+进行添加。
  - 勾选“自主审批”。

  说明

  如果当前账号在审批人列表中，才会有“自主审批”功能。

  单击“确认提交”，如果选择了审核人，需要审核通过后才能发布上线。如果勾选了“自主审批”，会立即发布上线。

导出数据标准

步骤 1 在数据架构控制台，单击左侧的“数据标准”，进入数据标准页面。

步骤 2 在数据标准的目录结构中，选择一个指定的目录名称并单击右键，然后单击“导出”即可。

----结束

3.4.6 模型设计

3.4.6.1 关系建模

3.4.6.1.1 逻辑模型设计

逻辑模型是利用实体及相互之间的关系，准确描述业务规则的实体关系图。逻辑模型要保证业务所需数据结构的正确性及一致性，使用一系列标准的规则将各种对象的特征体现出来，并对各实体之间的关系进行准确定义。

同时，逻辑模型也为构建物理模型提供了有力的参考依据，并支持转换为物理模型，是最终成功设计数据库过程中必不可少的一个阶段。

本章节主要介绍以下内容：

- 逻辑模型设计注意事项
- 新建逻辑模型
- 新建逻辑实体并发布

- 逻辑模型转换为物理模型

逻辑模型设计注意事项

- 不只针对当前业务现状，还要考虑业务将来的发展计划。
- 必须有熟知业务的人员参与建模，将实际业务所需内容充分反映在模型中。
- 必须要考虑设计的逻辑模型在向物理模型转换时具有较高的效率。
- 物理特性放在物理建模阶段考虑。
- 各个实体、属性、关系等必须要与实际业务中的信息能够对应。

新建逻辑模型

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

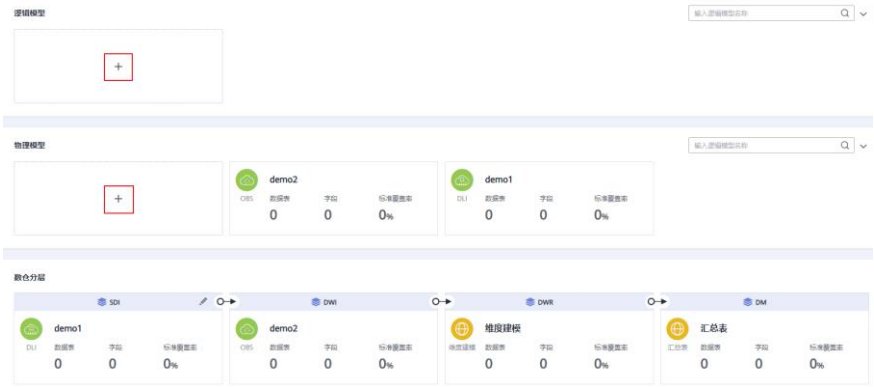
   图3-172 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“关系建模”。
3. 在“关系建模”页面，如果当前未创建过关系模型，系统会弹出“新建分层治理模型”提示框。如果不是首次创建，单击+按钮新建模型。

图3-173 新建分层治理模型

图3-174 关系建模页面



4. 在弹出窗口中配置如下参数，然后单击“确定”。

图3-175 新建逻辑模型

表3-127 参数描述

参数名称	说明

参数名称	说明
模型名称	只能包含中文、英文字母、数字和下划线。
描述	逻辑模型的描述信息。

新建逻辑实体并发布

逻辑实体即逻辑表。当您完成逻辑模型的创建之后，您就可以在逻辑模型中新建逻辑实体。

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“关系建模”进入关系建模页面。

步骤 2 在关系模型中选择所需要的逻辑模型，单击该模型进入管理页面，然后单击“新建”按钮新建一个逻辑实体。

步骤 3 在“新建逻辑实体”页面，根据页面提示完成相关配置。

1. 填写基本配置参数：

图3-176 基本配置 - 逻辑实体

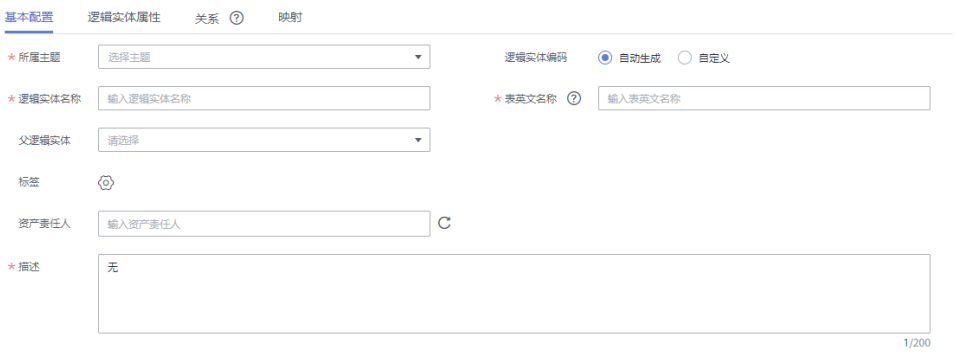


表3-128 基本配置

参数名称	说明
*所属主题	单击“选择主题”选择所属的主题信息。
逻辑实体编码	支持自动生成和自定义两种方式。
*逻辑实体名称	逻辑实体的名称。只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文或英文字母开头。
*表英文名称	逻辑实体转换为物理表的名称。只能包含英文字母、数字、下划线、$、{、}，且不能以数字开头。

参数名称	说明
父逻辑实体	设置一个父逻辑实体。本模块的父逻辑实体、子逻辑实体表示一个继承的概念，公共使用的逻辑实体及属性在逻辑上可以提炼为一个逻辑实体的就是父逻辑实体， 子逻辑实体是在父逻辑实体的基础上增加了特有属性，父逻辑实体属性的修改会影响所有继承它的子逻辑实体。
标签	标签是用户自定义的标识，它可以帮助用户对数据资产进行分类和搜索。添加标签后，您就可以在 DataArts Studio 数据目录模块中通过标签搜索相关的数据资产。 单击 按钮可以为表添加标签，在弹出框中可以选择一个或多个已有的标签，或者输入一个新的标签名称后按回车键。您也可以前往 DataArts Studio 数据目录模块的“标签管理”页面添加新的标签，详情请参见 3.7.1.4 标签管理，然后再返回此页面，就可以在标签的下拉列表中选择新添加的标签。
资产责任人	在下拉列表中选择用户，可以手动输入名字或直接选择已有的责任人。
*描述	描述信息。支持的长度 1~200 字符。

2. 在“逻辑实体属性”页面添加所需要的逻辑实体属性，逻辑实体属性参数说明参考表 3-129。

图3-177 添加逻辑实体属性

表3-129 逻辑实体属性参数

参数名称	说明
*名称	只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文或英文字母开头。
*英文名称	只能包含英文字母、数字、下划线，且以英文字母开头。
*编码	逻辑属性的编码，当逻辑实体为自定义编码时，逻辑属性可以自定义编码，也可以自动编码。
数据类型	设置属性的数据类型。如果在下拉列表中未找到所需要的数据类型，您可以参考字段类型管理添加数据类型。
数据标准	如果您已创建数据标准，单击 按钮可以选择一个数据标准与逻辑实体属性相关联。在“配置中心 > 功能配置”页面中的“模型设计业

参数名称	说明
	务流程步骤 > 创建质量作业”勾选的情况下，将逻辑实体属性关联数据标准后，逻辑实体发布上线后，就会自动生成一个质量作业，每个关联了数据标准的逻辑实体会生成一个质量规则，基于数据标准对属性进行质量监控，您可以前往 DataArts Studio 数据质量模块的“质量作业”页面进行查看。 如果您还未创建数据标准，请参见 3.4.5.2 新建数据标准进行创建。
主键	选中时为主键。
分区	选中时为分区字段。
不为空	是否限制该字段不为空。
标签	单击 按钮可以为逻辑实体属性添加标签。 • 在弹出框中可以选择一个或多个已有的标签。如果尚未添加标签，您也可以前往 DataArts Studio 数据目录模块的“标签管理”页面添加新的标签，详情请参见 3.7.1.4 标签管理。 • 在弹出框中，您也可以输入一个新的标签名称然后按回车键。标签名称只能包含中文、英文字母、数字和下划线，且不能以下划线开头。
描述	描述信息。

3. 在“关系”页面，单击“新建”新建关系。

   关系用于两个父、子实体（有时也称为主、从实体）之间的主外键关联关系，即描述实体与实体是以何种形态关联在一起，或者描述一个实体本身的行为会对另外一个实体产生何种影响。数据模型内实体之间的关系尤为重要，必须要对其准确定义。否则，无法在数据模型中准确描述实际的业务规则，而且很大程度上破坏数据的一致性。

   例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则其关系为：

   - 子逻辑实体：成绩表
   - 子逻辑实体属性 FK：学号
   - 子对父： 1
   - 父逻辑实体：学生表
   - 父逻辑实体属性 PK：学号
   - 父对子： 1

图3-178 新建关系

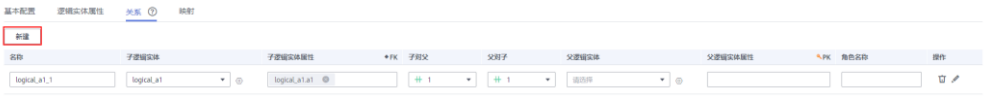


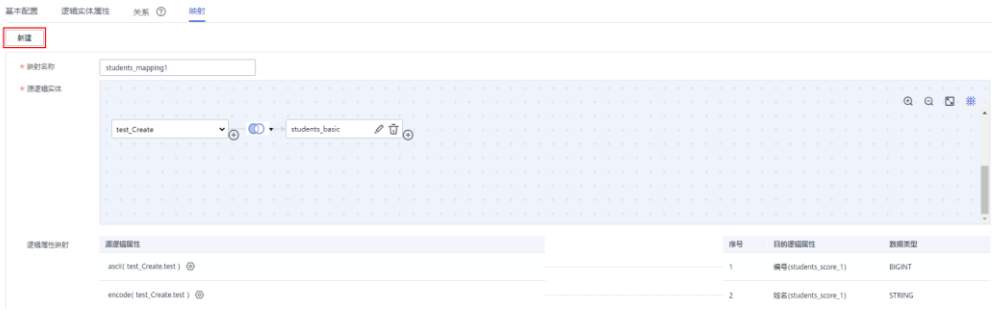
表3-130 新建关系参数说明

参数名称	说明
名称	通过名称来描述该关系。
子逻辑实体	单击该属性在下拉列表中选择子逻辑实体。单击可设置当前逻辑实体为子逻辑实体。 例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则子逻辑实体应为“成绩表”，对应父逻辑实体应为“学生表”。
子逻辑实体属性 FK	选择子逻辑实体属性，FK 表示外键 Foreign Key。该子逻辑实体的属性应为父逻辑实体的外键。 例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则此子逻辑实体属性 FK 应为“成绩表”的“学号”。
子对父	1 ：表示每条子逻辑实体数据在父逻辑实体中有且只有一条数据与之对应。 0,1 ：表示每条子逻辑实体数据在父逻辑实体中最多有一条数据与之对应。 0..n ：表示每条子逻辑实体数据在父逻辑实体中可能有多条数据与之对应。 1..n ：表示每条子逻辑实体数据在父逻辑实体中至少有一条数据与之对应。
父对子	1 ：表示每条父逻辑实体数据在子逻辑实体中有且只有一条数据与之对应。 0,1 ：表示每条父逻辑实体数据在子逻辑实体中最多有一条数据与之对应。 0..n ：表示每条父逻辑实体数据在子逻辑实体中可能有多条数据与之对应。 1..n ：表示每条父逻辑实体数据在子逻辑实体中至少有一条

参数名称	说明
	数据与之对应。
父逻辑实体	选择与所选子逻辑实体有逻辑关系的逻辑实体。 例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则父逻辑实体应为“学生表”，对应子逻辑实体应为“成绩表”。
父逻辑实体属性 PK	选择父逻辑实体的属性，PK 表示主键 Primary Key。该父逻辑实体的属性应为父逻辑实体的主键。 例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则此父逻辑实体属性 PK 应为“学生表”的“学号”。
角色名称	可以自定义一个角色名称，用于标识该关系。
操作	单击 可删除一条关系。单击 可编辑关系。

4. 在“映射”页面，单击“新建”新建映射，创建完成后单击“保存”。映射指的是给两个逻辑实体（源逻辑实体和目的逻辑实体）建立起属性的对应关系。

图3-179 新建映射



- **映射名称**：新建映射时会自动生成，用户可以手动修改。
- **源逻辑实体**：如果数据来源于一个模型中的多个逻辑实体，可以单击逻辑实体后的按钮为该逻辑实体和其他逻辑实体之间设置 JOIN。

图3-180 设置源表 JOIN 条件

表3-131 JOIN 条件参数说明

参数名	参数说明
JOIN 逻辑实体	下拉选择需要和源逻辑实体建立 JOIN 关系的逻辑实体。
JOIN 方式	从左到右依次表示 left JOIN、right JOIN、inner JOIN、outer JOIN。
JOIN 属性	JOIN 属性一般选择源逻辑实体和 JOIN 逻辑实体中含义相同的属性，单击 + 或 — 按钮增加或删除 JOIN 属性。JOIN 属性之间是 and 的关系。

- **逻辑属性映射**：为来源于当前映射的属性，依次选择一个含义相同的源属性。

步骤 4 单击“发布”，选择审核人，再单击“确认提交”提交审核。

等待审核人员审核，审核通过后，返回模型页面，在列表中可以查看建好的逻辑实体。

说明

系统默认在“配置中心>功能配置>模型设计业务流程步骤”中勾选了“同步业务资产”：

- 对于新建的逻辑模型，单击“发布”可直接将逻辑模型同步到数据目录模块中的业务资产中。
- 对于历史发布的逻辑模型，单击列表上方的“更多>同步”可将逻辑模型同步到数据目录模块的业务资产中。

----结束

逻辑模型转换为物理模型

完成逻辑模型的创建后，您可以将逻辑模型转换为物理模型，支持转换为新的物理模型或已有的物理模型。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-181 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“关系建模”。
3. 在总览图中找到所需要的逻辑模型，将光标移动到该卡片上，单击该模型的转换按钮。

图3-182 逻辑模型转换

4. 在“转换为物理模型”对话框中，配置如下参数，然后单击“确定”。

图3-183 转换为物理模型

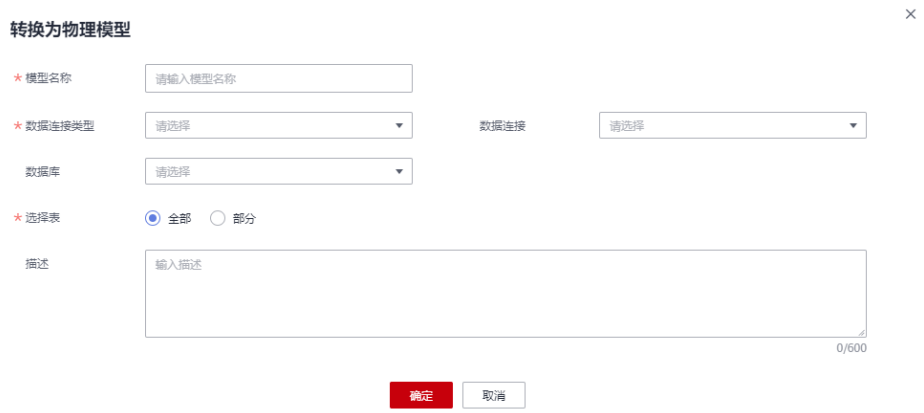
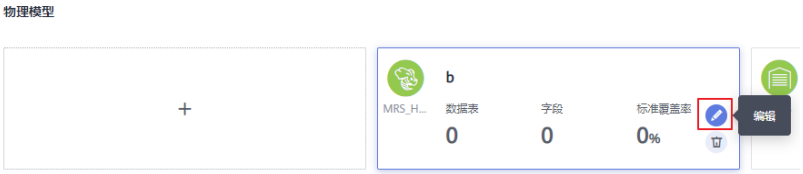


表3-132 参数描述

参数名称	说明
*模型名称	逻辑模型所需转换的物理模型的名称。您可以输入一个新的模型名称，系统会创建该新模型，也可以在下拉列表中选择一个已有的模型。 模型名称只能包含中文、英文字母、数字和下划线。
*数据连接类型	在下拉列表中选择数据连接类型。如果没有所需要的数据类型，您可以参考字段类型管理添加新的数据类型。
数据连接	选择所需要的数据连接。同一个关系模型一般建议使用统一的数据连接。 如果您还未创建与数据源之间的数据连接，请前往 DataArts Studio 管理中心控制台进行创建，详情请参见 3.2.2 创建数据连接。
数据库	选择数据库。如果您还未创建数据库，可以前往 DataArts Studio 数据开发控制台进行创建，详情请参见 3.5.2.3 新建数据库。
选择表	• 全部：将所有的逻辑实体转换为物理表。 • 部分：将选择的部分逻辑实体转换为物理表。
队列	DLI 队列。该参数仅 DLI 连接类型有效。
Schema	DWS 和 POSTGRESQL 的模式。该参数仅支持 DWS 和 POSTGRESQL 连接类型。
描述	描述信息。支持的长度为 0~600 个字符。

5. 转换为物理模型后，您可以为该物理模型设置分层，您可以选择 SDI 层或 DWI 层。如图 3-184，在物理模型中找到转换后的物理模型，将光标移动到该卡片上，单击该模型的编辑按钮，进入“编辑物理模型”弹窗。

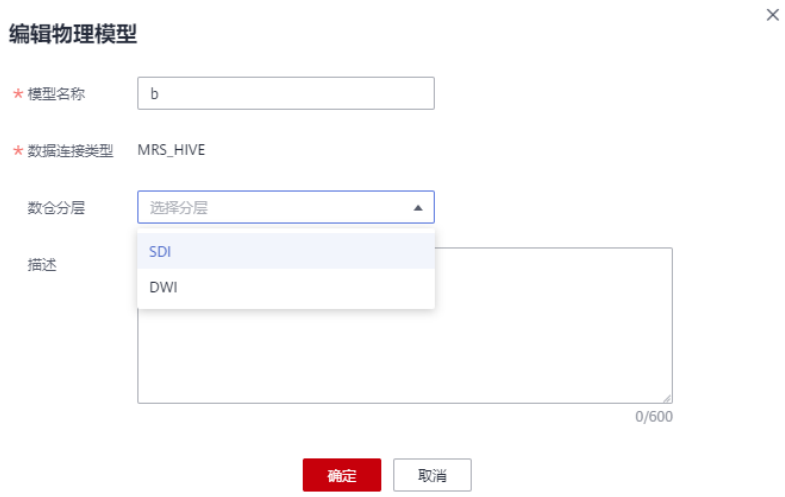
图3-184 设置物理模型分层



进入“编辑物理模型”弹窗后，选择“数仓分层”，下拉选择 SDI 或 DWI 分层。

- **SDI**：Source Data Integration，又称贴源数据层。SDI 是源系统数据的简单落地。
- **DWI**：Data Warehouse Integration，又称数据整合层。DWI 整合多个源系统数据，对源系统进来的数据进行整合、清洗，并基于三范式进行关系建模。

图3-185 编辑物理模型



3.4.6.1.2 物理模型设计

物理模型是指按照一定规则和方法，将逻辑模型中所定义的实体、属性、属性约束、关系等要素转换为数据库软件所能够识别的表关系图(Table Relationship Diagram)的一种物理描述。

在关系建模中，您可以新建 SDI 层和 DWI 层两个模型，模型最终是通过物理建模进行落地的。除了将逻辑模型转换外为物理模型外，您也可以参考本章节直接新建一个物理模型。

本章节主要介绍以下内容：

- 物理模型设计时的考虑事项
- 新建物理模型
- 新建表并发布

物理模型设计时的考虑事项

- 物理模型要确保业务需求及业务规则所要求的功能得到满足，性能得到保障。
- 物理模型要确保数据的一致性及数据的质量。
- 新业务或新功能增加时能够以较少的改动或不改动就能够满足需求的扩展。

新建物理模型

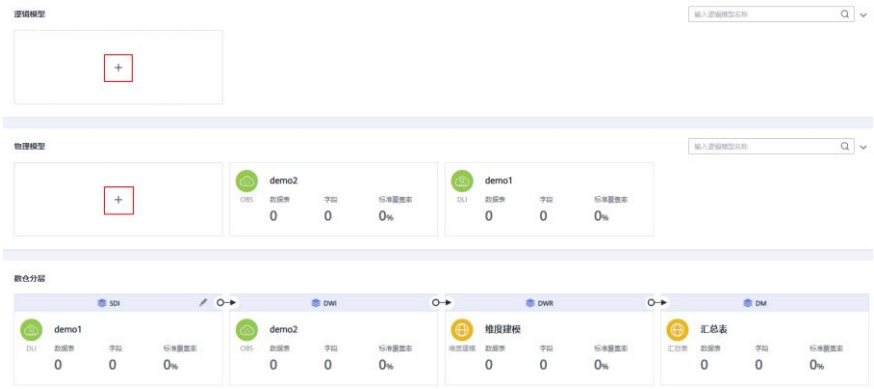
1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-186 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“关系建模”。
3. 在“关系建模”页面，如果当前未创建过关系模型，系统会弹出“新建分层治理模型”提示框。如果不是首次创建，单击+按钮新建模型。

图3-187 新建分层治理模型

图3-188 关系建模页面



4. 在弹出窗口中配置如下参数，然后单击“确定”。

图3-189 新建模型

表3-133 参数描述

参数名称	说明
模型名称	只能包含中文、英文字母、数字和下划线。
数据连接类型	下拉选择数据连接类型。
数仓分层	下拉选择 SDI 或 DWI 分层。 • **SDI**：Source Data Integration，又称贴源数据层。SDI 是源系统数据的简单落地。

参数名称	说明
	• **DWI**：Data Warehouse Integration，又称数据整合层。DWI 整合多个源系统数据，对源系统进来的数据进行整合、清洗，并基于三范式进行关系建模。
描述	描述信息。支持的长度 0~600 字符。

新建表并发布

当您完成 DLI/POSTGRESQL/DWS/MRS_HIVE 类型的关系模型的创建之后，您就可以在关系模型中新建业务表。

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“关系建模”进入关系建模页面。

步骤 2 选择所需要建表的物理模型，单击进入，然后单击上方“新建”按钮新建一个表。

图3-190 入口

步骤 3 在“新建表”页面，根据页面提示完成建表的配置。

1. 填写基本配置参数。

图3-191 表基本配置

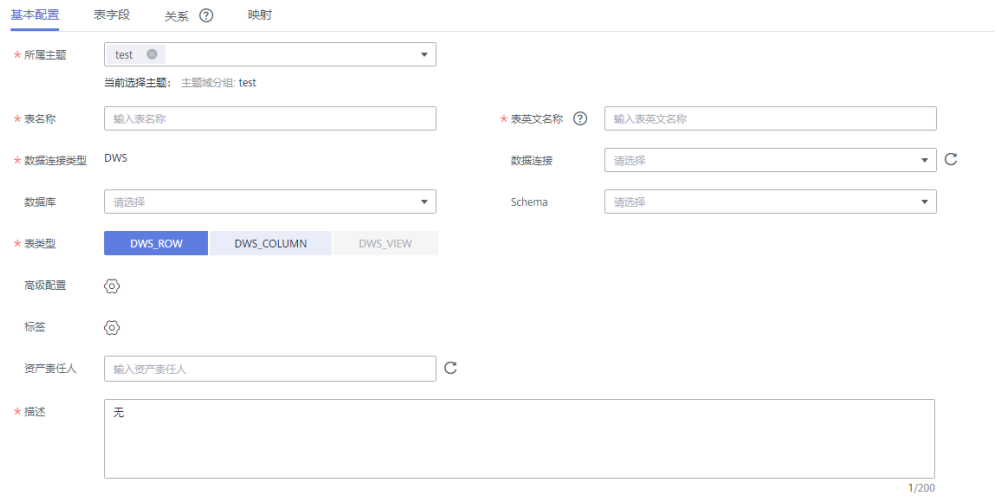


表3-134 基本配置

参数名称	说明
所属主题	单击“选择主题”选择所属的主题信息。
表名称	表的名称。只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文或英文字母开头。
表英文名称	表的英文名称。只能包含英文字母、数字、下划线、$、{、}，且不能以数字开头。
数据连接类型	-
数据连接	选择所需要的数据连接。同一个关系模型一般建议使用统一的数据连接。 如果您还未创建与数据源之间的数据连接，请前往 DataArts Studio 管理中心进行创建，详情请参见 3.2.2 创建数据连接。
数据库	选择数据库。
队列	DLI 队列。该参数仅 DLI 模型的表有效。
Schema	DWS 和 POSTGRESQL 的模式。该参数仅 DWS 和 POSTGRESQL 模型的表有效。
表类型	DLI 模型的表支持以下表类型： • Managed：数据存储位置为 DLI 的表。 • External：数据存储位置为 OBS 的表。当“表类型”设置为 External 时，需设置“OBS 路径””参数。OBS 路径格式如：/bucket_name/filepath。 DWS 模型的表支持以下表类型： • DWS_ROW：行存表。行存储是指将表按行存储到硬盘分区上。 • DWS_COLUMN：列存表。列存储是指将表按列存储到硬盘分区上。 • DWS_VIEW：视图存表。视图存储是指将表按视图存储到硬盘分区上。 MRS_HIVE 模型仅支持 HIVE_TABLE。
数据格式	该参数仅 DLI 模型的表有效。DLI 模型的表支持以下数据格式： • Parquet：DLI 支持读取不压缩、snappy 压缩、gzip 压缩的 parquet 数据。 • CSV：DLI 支持读取不压缩、gzip 压缩的 csv 数据。 • ORC：DLI 支持读取不压缩、snappy 压缩的 orc 数据。 • JSON：DLI 支持读取不压缩、gzip 压缩的 json 数据。 • Carbon：DLI 支持读取不压缩的 carbon 数据。 • Avro：DLI 支持读取不压缩的 avro 数据。

参数名称	说明
高级配置	设置自定义项，以对表进行描述。自定义项设置完成后仅可用于在表详情中进行查看，无特殊需求时无需设置。 例如您需要标识该表的来源时，可以设置自定义项配置名为“来源”，值为对应的表来源信息。配置完成后可以在表详情中查看该信息。
标签	标签是用户自定义的标识，它可以帮助用户对数据资产进行分类和搜索。添加标签后，您就可以在 DataArts Studio 数据目录模块中通过标签搜索相关的数据资产。 单击 按钮可以为表添加标签，在弹出框中可以选择一个或多个已有的标签，或者输入一个新的标签名称后按回车键。您也可以前往 DataArts Studio 数据目录模块的“标签管理”页面添加新的标签，详情请参见 3.7.1.4 标签管理，然后再返回此页面，就可以在标签的下拉列表中选择新添加的标签。
资产责任人	在下拉列表中选择用户，可以手动输入名字或直接选择已有的责任人。
描述	描述信息。支持的长度 1~600 字符。

2. 在“表字段”页面添加所需要的字段。

图3-192 添加所需表字段

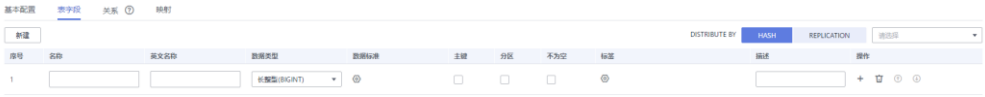


表3-135 表字段参数

参数名称	说明
名称	只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文或英文字母开头。
英文名称	只能包含英文字母、数字、下划线，且以英文字母开头。
数据类型	设置字段的数据类型。如果在下拉列表中未找到所需要的数据类型，您可以参考字段类型管理添加数据类型。
数据标准	如果您已创建数据标准，单击 按钮可以选择一个数据标准与字段相关联。在“配置中心 > 功能配置”页面中的“模型设计业务流程步骤 > 创建质量作业”勾选的情况下，将字段关联数据标准后，表发布上线后，就会自动生成一个质量作业，每个关联了数据标准的字段会生成一个质量规则，基于数据标准对字段进行质量监控，您可以前往

参数名称	说明
	DataArts Studio 数据质量模块的“质量作业”页面进行查看。 如果您还未创建数据标准，请参见 3.4.5.2 新建数据标准进行创建。
主键	选中时为主键。
分区	选中时为分区字段。
不为空	是否限制该字段不为空。
标签	单击 按钮可以为表字段添加标签。 • 在弹出框中可以选择一个或多个已有的标签。如果尚未添加标签，您也可以前往 DataArts Studio 数据目录模块的“标签管理”页面添加新的标签，详情请参见 3.7.1.4 标签管理。 • 在弹出框中，您也可以输入一个新的标签名称然后按回车键。标签名称只能包含中文、英文字母、数字和下划线，且不能以下划线开头。
描述	描述信息。

3. （可选）在“关系”页面，单击“新建”新建关系。

   关系用于两个父、子表（有时也称为主、从表）之间的主外键关联关系，即描述表与表是以何种形态关联在一起，或者描述一个表本身的行为会对另外一个表产生何种影响。数据模型内表之间的关系尤为重要，必须要对其准确定义。否则，无法在数据模型中准确描述实际的业务规则，而且很大程度上破坏数据的一致性。

   例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则其关系为：

   - 子表：成绩表
   - 子表字段 FK：学号
   - 子对父： 1
   - 父表：学生表
   - 父表字段 PK：学号
   - 父对子： 1

   图3-193 新建关系（可选）

表3-136 新建关系参数说明

参数名称	说明
名称	通过名称来描述该关系。
子表	单击该字段可在下拉列表中选择表。单击可设置当前表为子表。 例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则子表应为“成绩表”，对应父表应为“学生表”。
子表字段 FK	选择子表的字段，FK 表示外键 Foreign Key。该子表的字段应为父表的外键。 例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则此子表字段 FK 应为“成绩表”的“学号”。
子对父	1 ：表示每条子表数据在父表中有且只有一条数据与之对应。 0,1 ：表示每条子表数据在父表中最多有一条数据与之对应。 0..n ：表示每条子表数据在父表中可能有多条数据与之对应。 1..n ：表示每条子表数据在父表中至少有一条数据与之对应。
父对子	1 ：表示每条父表数据在子表中有且只有一条数据与之对应。 0,1 ：表示每条父表数据在子表中最多有一条数据与之对应。 0..n ：表示每条父表数据在子表中可能有多条数据与之对应。 1..n ：表示每条父表数据在子表中至少有一条数据与之对应。
父表	选择与所选子表对应的父表。 例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则父表应为“学生表”，对应子表应为“成绩表”。
父表字段 PK	选择父表的字段，PK 表示主键 Primary Key。该父表的字段应为父表的主键。 例如，对于根据 3NF 范式设计的“成绩表”和“学生表”，成绩表中的“学号”属性为学生表的主键。则此父表字段 PK 应为“学生表”

参数名称	说明
	的“学号”。
角色名称	可以自定义一个角色名称，用于标识该关系。
操作	单击 可删除一条关系。单击 可编辑关系。

4. （可选）在“映射”页面，单击“新建”可以新建一个映射，通过新建映射设计当前表的数据来源。
   - 如果表中的字段数据来源于不同的关系模型，您需要创建多个映射。

     当前支持表数据来源于不同连接类型的关系模型。在每个映射中，您只需要为来源于当前映射的字段设置源字段，其他字段可以不设置。

     例如，假设当前表的前面 5 个字段和后 5 个字段数据来源于 2 个不同的模型，您可以新建如下两个映射：

     - **map1**：设置“来源”为关系模型 A 的表 table01，在“字段映射”中依次设置第 1~5 个字段的源字段为 table01 中含义相同的相应字段，后 5 个字段不用设置。
     - **map2**：设置“来源”为关系模型 B 的表 table02，在“字段映射”中依次设置第 6~10 个字段的源字段为 table02 中含义相同的相应字段，前 5 个字段不用设置。
   - 如果表中的字段数据来源于同一个关系模型中的多个表，您可以新建一个映射。

     在该映射的“源表”中，您可以将多个表设置 Join，然后在“字段映射”区域依次为表中的字段设置源字段，所选择的源字段应与表中的字段代表相同含义，一一对应。

     例如，假设当前表的字段都来源于关系模型 d1，第 1 个字段来源于表 vendor，第 2 个字段来源于表 payment_type，第 3 个字段来源于表 rate，其余字段来源于 dwd_taxi_trip_data。

     您可以新建一个映射，如图 3-194 所示，设置表 dwd_taxi_trip_data 和 vendor、payment_type、rate 做 Join，然后在字段映射中，依次设置源字段。

   新建映射的参数说明，可以参考表 3-137。

图3-194 配置映射关系

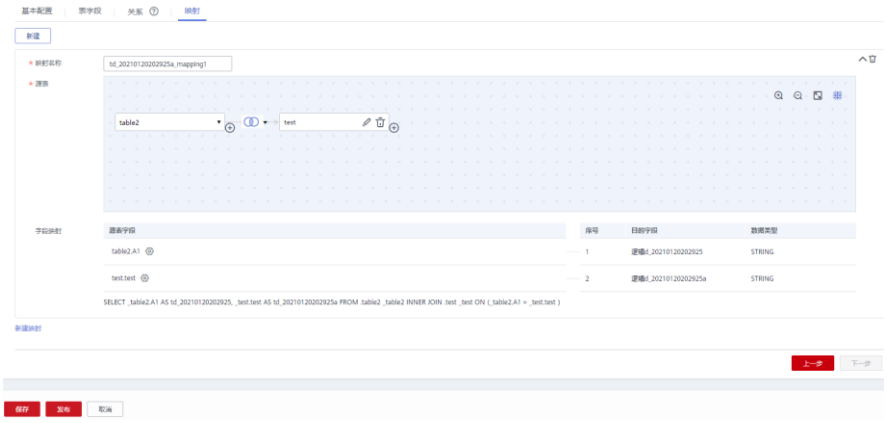


表3-137 映射参数

参数名称	说明
映射名称	只能包含中文、英文字母、数字和下划线。
来源模型	在下拉列表中选择一个已创建的关系模型。如果未创建关系模型，请参见 3.4.6.1.2 物理模型设计进行创建。
源表	选择数据来源的表，如果数据来源于一个模型中的多个表，可以单击表名后的按钮 为该表和其他表之间设置 JOIN。 1. 选择一种“JOIN 方式”，“JOIN 方式”从左到右依次表示 left JOIN、right JOIN、inner JOIN、outer JOIN。 2. 在“JOIN 字段”中设置 JOIN 条件，JOIN 条件一般选择源表和 JOIN 表中含义相同的字段，单击 + 或 — 按钮增加或删除 JOIN 条件。JOIN 条件之间是 and 的关系。 3. 单击“确定”完成设置。 4. 设置 JOIN 后，如果想删除 JOIN 表，单击所需删除的表名后的 按钮就可以删除该 JOIN 表。 图3-195 JOIN 条件 

参数名称	说明
字段映射	为来源于当前映射的字段，依次选择一个含义相同的源字段。如果表字段来源于多个模型，您需要新建多个映射，每个映射中，您只需要为来源于当前映射的字段设置源字段，其他字段可以不设置。

在映射区域的右上角，单击 按钮，可以删除指定的映射，单击 可以收起映射区域。

5. （可选）新建表的“表类型”为“DWS_VIEW”时，在“视图定义”页面，单击“新建”可以新建一个视图。

图3-196 新建视图

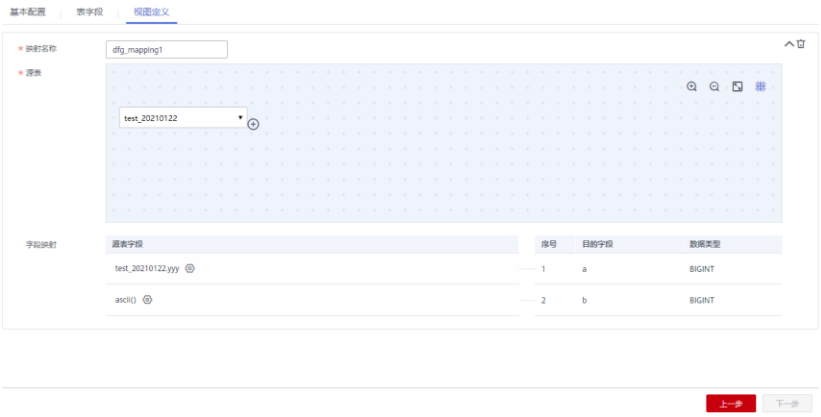


表3-138 视图定义参数

参数名称	说明
映射名称	只能包含中文、英文字母、数字和下划线。
源表	选择数据来源的表，如果数据来源于一个模型中的多个表，可以单击表名后的按钮 为该表和其他表之间设置 JOIN。 1. 选择一种“JOIN 方式”，“JOIN 方式”从左到右依次表示 left JOIN、right JOIN、inner JOIN、outer JOIN。 2. 在“JOIN 字段”中设置 JOIN 条件，JOIN 条件一般选择源表和 JOIN 表中含义相同的字段，单击 或 按钮增加或删除 JOIN 条件。JOIN 条件之间是 and 的关系。 3. 单击“确定”完成设置。 4. 设置 JOIN 后，如果想删除 JOIN 表，单击所需删除的表名后的 按钮就可以删除该 JOIN 表。 图3-197 JOIN 条件界面

参数名称	说明
	
字段映射	为来源于当前映射的字段，依次选择一个含义相同的源字段。如果表字段来源于多个模型，您需要新建多个映射，每个映射中，您只需要为来源于当前映射的字段设置源字段，其他字段可以不设置。

在映射区域的右上角，单击 按钮，可以删除指定的映射，单击 可以收起映射区域。

步骤 4 完成表的配置后，单击“发布”，选择审核人，再单击“确认提交”提交审核。

步骤 5 等待审核人员审核。当审核人审批通过后，返回“关系建模”页面可以查看表的“状态”和“同步状态”。

发布是一个异步操作，您可以单击 按钮刷新状态。表发布并通过审核后，系统会依据“配置中心 > 功能配置”页面中的“模型设计业务流程步骤”进行创建表、同步技术资产、同步业务资产等操作，在表的“同步状态”一列中将显示同步状态。

- “同步状态”若均显示成功，则说明表发布成功。鼠标移至“同步状态”中的 图标之上，若显示“创建表: 创建成功”说明该表在对应的数据源下已经创建成功。
- “同步状态”若显示某一项或某几项失败，可以先刷新状态。如果仍失败，可以单击“更多 > 发布历史”，然后进入“发布日志”页签查看日志。

  请根据错误日志定位失败原因，问题解决后，再单击“发布日志”页面中的“重新同步”再次下发同步命令。如果仍同步失败，请联系技术支持人员协助解决。

----结束

3.4.6.2 维度建模

3.4.6.2.1 新建维度

维度是用于观察和分析业务数据的视角，支撑对数据汇聚、钻取、切片分析，用于 SQL 中的 GROUP BY 条件。维度多数具有层级结构，如：地理维度(其中包括国家、

地区、省以及城市等级别的内容)、时间维度(其中包括年度、季度、月度等级别的内容)。创建维度，即从顶层规范业务中实体（或称主数据）的存在性及唯一性。

对系统的影响

维度发布并通过审核后，系统会自动创建与维度相对应的维度表，维度表的名称和编码均与维度相同。

新建维度并发布

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

   图3-198 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“维度建模”，选择“维度”页签进入维度页面。
3. 在左侧主题目录中选中一个对象，然后单击“新建”开始新建维度。

   在新建维度之前，如果您尚未添加主题信息，请先参考 3.4.4.2 主题设计添加主题信息。
4. 在“新建维度”页面，根据页面提示配置参数。

   “基本配置”和“物化配置”，设置如下：

图3-199 配置参数

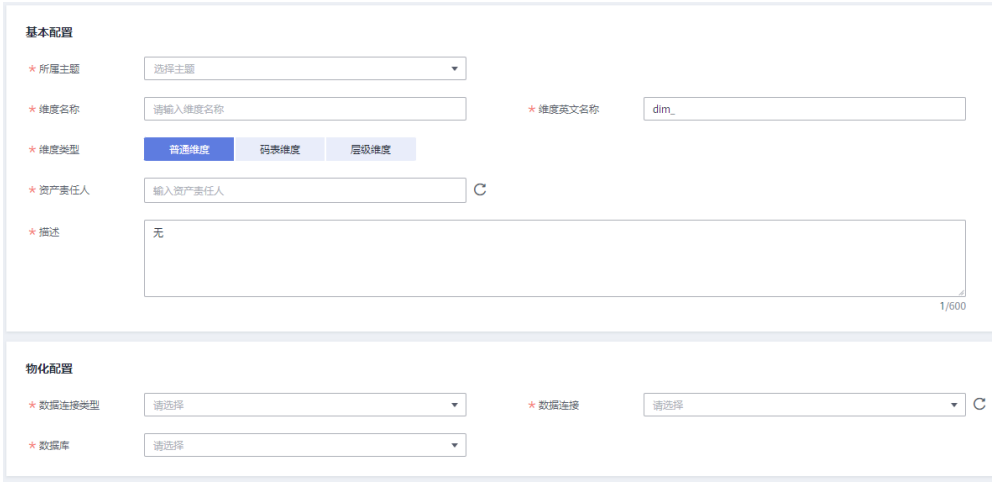


表3-139 基本配置

参数名称	说明
所属主题	下拉框中选择相应的主题。
维度名称	只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文和英文字母开头。
维度英文名称	只能包含英文字母、数字和下划线，且英文字母开头。
维度类型	• **普通维度**：不具有层级结构的维度。 • **码表维度**：基于码表创建的维度，其字段信息、数据与码表保持一致，表示内容是可枚举的维度。 • **层级维度**：属性之间具有层级结构的维度。
资产责任人	在下拉列表中选择维度所属的资产责任人，可以手动输入名字或直接选择已有的责任人。
描述	描述信息。支持的长度为 0~600 个字符。

表3-140 物化配置

参数名称	说明
数据连接类型	在下拉列表中选择数据连接类型。
数据连接	选择所需要的数据连接。 如果您还未创建与数据源之间的数据连接，请前往 DataArts Studio 管理中心控制台进行创建，详情请参见 3.2.2 创建数据连接。
数据库	选择数据库。如果您还未创建数据库，可以前往 DataArts Studio 数

参数名称	说明
	据开发控制台进行创建，详情请参见 3.5.2.3 新建数据库。
队列	DLI 队列。该参数仅 DLI 连接类型有效
Schema	DWS 或 POSTGRESQL 的模式。该参数在 DWS 或 POSTGRESQL 连接类型有效。
表类型	DWS 表类型有： • DWS_ROW：行存表。行存储是指将表按行存储到硬盘分区上。 • DWS_COLUMN：列存表。列存储是指将表按列存储到硬盘分区上。 MRS_HIVE 类型仅支持 HIVE_TABLE。
DISTRIBUTE BY	该参数仅 DWS 连接类型有效。可选取多个字段。 • **REPLICATION**：在每一个 DN 节点上存储一份全量表数据。这种存储方式的优点是每个 DN 上都有此表的全量数据，在 join 操作中可以避免数据重分布操作，从而减小网络开销；缺点是每个 DN 都保留了表的完整数据，造成数据的冗余。一般情况下只有较小的维度表才会定义为 Replication 表。 • **HASH**：采用这种分布方式，需要为用户表指定一个分布列（distribute key）。当插入一条记录时，系统会根据分布列的值进行 hash 运算后，将数据存储在对应的 DN 中。对于 Hash 分布表，在读/写数据时可以利用各个节点的 IO 资源，大大提升表的读/写速度。一般情况下大表（1000000 条记录以上）定义为 Hash 表。

在“属性配置”中添加维度属性，单击“新建”按钮，可以添加多个维度属性。

图3-200 属性配置

表3-141 属性配置

参数名称	说明
属性名称	只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文和英文字母开头。
属性编码	只能包含英文字母、数字和下划线，且英文字母开头。
数据标准	单击 按钮可以选择一个数据标准与字段相关联。在“配置中

参数名称	说明
	心 > 功能配置”页面中的“模型设计业务流程步骤 > 创建质量作业”勾选的情况下，将字段关联数据标准后，维度发布上线后，就会自动生成一个质量作业，每个关联了数据标准的字段会生成一个质量规则，基于数据标准对字段进行质量监控，您可以前往DataArts Studio 数据质量模块的“质量作业”页面进行查看。 如果您还未创建数据标准，请参见 3.4.5.2 新建数据标准进行创建。
数据类型	根据原始数据定义数据类型。
代理键	请根据业务需求选择合适的字段作为代理键。系统默认第一个维度属性为代理键。
主键	请根据业务需求选择合适的字段作为主键。
分区	是否设置为分区字段。
不为空	是否限制该字段不为空。
描述	输入维度属性的描述信息。

在“映射配置”页签，单击“新建映射”，创建维度与事实表的映射。需配置如下参数：

图3-201 映射配置

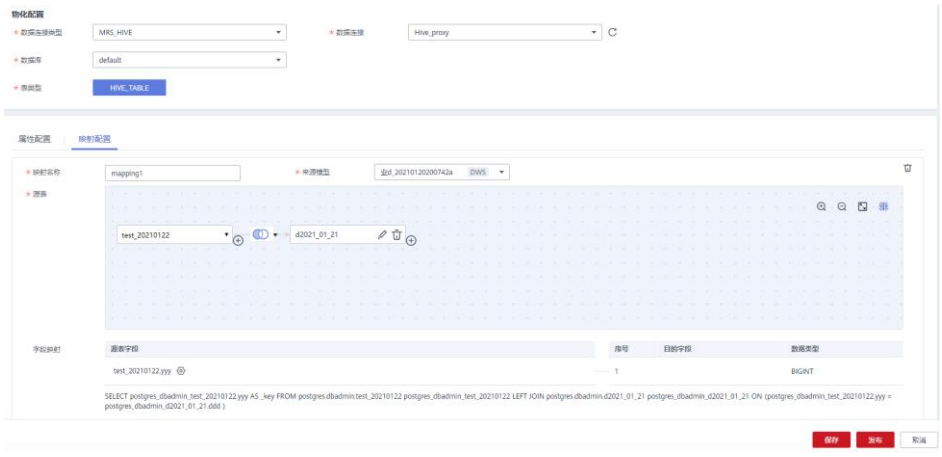


表3-142 映射参数

参数名称	说明
映射名称	只能包含中文、英文字母、数字和下划线。
来源模型	在下拉列表中选择一个已创建的关系模型。如果未创建关系模型，请参见 3.4.6.1.2 物理模型设计进行创建。

参数名称	说明
源表	选择数据来源的表，如果数据来源于一个模型中的多个表，可以单击表名后的按钮为该表和其他表之间设置 JOIN。 1. 选择一种“JOIN 方式”，“JOIN 方式”从左到右依次表示 left JOIN、right JOIN、inner JOIN、outer JOIN。 2. 在“JOIN 字段”中设置 JOIN 条件，JOIN 条件一般选择源表和 JOIN 表中含义相同的字段，单击 + 或 — 按钮增加或删除 JOIN 条件。JOIN 条件之间是 and 的关系。 3. 单击“确定”完成设置。 4. 设置 JOIN 后，如果想删除 JOIN 表，单击所需删除的表名后的按钮就可以删除该 JOIN 表。 图3-202 JOIN 条件 
字段映射	为来源于当前映射的字段，依次选择一个含义相同的源字段。如果表字段来源于多个模型，您需要新建多个映射，每个映射中，您只需要为来源于当前映射的字段设置源字段，其他字段可以不设置。

在映射区域的右上角，单击按钮，可以删除指定的映射，单击可以收起映射区域。

5. 配置完成后，单击“发布”。
6. 在弹出对话框中，选择审核人，单击“确认提交”，完成发布。
7. 可以参照步骤 3~步骤 6，完成其他维度的创建和发布。
8. 完成所有维度的新建和发布之后，需要等待审核人员审核。

   审核通过后，系统会自动创建与维度相对应的维度表，维度表的名称和编码均与维度相同。在“维度建模”页面，选择“维度表”页签，可以查看建好的维度表。

   在维度表列表中，在“同步状态”一列中可以查看维度表的同步状态。

图3-203 维度表的同步状态

- 如果同步状态均显示成功，则说明维度发布成功，维度表在数据库中创建成功。
- 如果同步状态中存在失败，可单击该维度表所在行的“发布历史”，然后在展开的页面中进入“发布日志”页签查看日志。请根据错误日志定位失败原因，问题解决后，您可以返回维度表页面勾选该维度表，再单击列表上方的“同步”按钮尝试重新同步。如果问题仍未能解决，请联系技术支持人员协助处理。

编辑维度

步骤 1 在数据架构控制台，单击左侧导航树中的“维度建模”，进入相应页面后，选择“维度”页签。

步骤 2 在维度列表中找到需要编辑的维度，单击“编辑”，进入编辑维度页面。

图3-204 编辑维度

步骤 3 根据实际需要编辑维度的相关信息，参数配置请参考配置参数。

步骤 4 单击“保存”，保存所做的修改。或者，单击“发布”，发布修改后维度。

----结束

发布维度

如果新建了维度但并未发布，可以执行以下步骤发布维度：

步骤 1 在数据架构控制台，单击左侧导航树中的“维度建模”，进入相应页面后，选择“维度”页签。

步骤 2 在维度列表中找到需要发布的维度，单击“发布”。

图3-205 发布维度

步骤 3 在弹出对话框中，选择审核人，单击“确认提交”，完成发布。

----结束

您也可以执行以下步骤批量发布维度：

步骤 1 在数据架构控制台，单击左侧导航树中的“维度建模”，进入相应页面后，选择“维度”页签。

步骤 2 在维度列表中勾选需要发布的维度，单击列表上方的“发布”。

图3-206 批量发布维度

步骤 3 在弹出对话框中，选择审核人和作业调度时间，单击“确认提交”，完成发布。

注意，此处“作业调度时间”指的是维度发布后，自动创建质量作业的调度时间。

图3-207 批量发布维度

----结束

下线维度

对于已发布的维度，可以执行以下步骤下线维度：

步骤 1 在数据架构控制台，单击左侧导航树中的“维度建模”，进入相应页面后，选择“维度”页签。

步骤 2 在维度列表中找到需要下线的维度，单击“更多-下线”。

图3-208 下线维度

步骤 3 在弹出对话框中，选择审核人，然后单击“确认提交”，完成维度的下线。

----结束

删除维度

如果您已不再需要某个维度，可以删除该维度。如果待删除的维度已发布，则无法执行删除操作，您必须先将该维度下线后，才能执行删除操作，具体操作请参见下线维度。

步骤 1 在数据架构控制台，单击左侧导航树中的“维度建模”，进入相应页面后，选择“维度”页签。

步骤 2 在维度列表中找到需要删除的维度，勾选该维度，然后单击维度列表上方“更多”中的“删除”按钮。

图3-209 删除

步骤 3 在系统弹出的“删除”对话框中，确认无误后，单击“确定”将维度删除。

删除弹框中的“删除物理表”勾选后，删除时将同步删除数据库里的物理表

----结束

3.4.6.2.2 管理维度表

维度表与维度一一对应，通过丰富维度中的属性信息构建形成。维度表的生命周期（包括新建、发布、编辑、下线操作）通过维度进行管理，在维度发布成功后，系统会自动创建并发布对应的维度表。

查看维度表发布历史

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“维度表”页签，进入维度表页面。
3. 在列表中，找到所需要的维度表，在右侧单击“发布历史”，将显示“发布历史”页面。

图3-210 维度表页面

4. 在“发布历史”中，您可以查看维度表的发布历史、版本对比信息以及发布日志。

   如果“发布日志”中有错误日志，说明发布失败。您可以单击“重新同步”进行重试，将表同步到 DataArts Studio 的其他模块中。

查看预览 SQL

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“维度表”页签，进入维度表页面。
3. 在维度表列表中，找到所需要的维度表，在右侧单击“预览 SQL”，弹出“预览 SQL”对话框。

图3-211 预览 SQL

4. 在“预览 SQL”中，您可以查看 SQL 语句，也可以复制 SQL。

同步维度表

当您新建或编辑维度后，对维度进行发布，如果同步状态中存在失败，可以对维度表手动进行同步。

说明

- 同步时，系统将根据“配置中心 > 功能配置”页面中的“数据表更新方式”执行相应的同步操作，详情请参见功能配置。

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“维度表”页签，进入维度表页面。
3. 在维度表列表中，勾选需要同步的维度表，单击列表左上方的“同步”按钮，系统弹出“批量同步”对话框。

图3-212 同步维度表

4. 确认无误后，单击“确认提交”，完成后界面将显示同步结果。

同步后，您可以在维度表列表中，查看维度表的同步状态。单击列表右上方的刷新按钮，可以刷新状态。

维度表关联质量规则

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“维度表”页签，进入维度表页面。
3. 在维度表列表中，勾选需要关联质量规则的维度表。单击“关联质量规则”。

图3-213 关联维度表质量规则

4. 在弹出的页面中配置关联质量规则参数。配置完成单击确定。
   - **更新已有规则**：若勾选此项，新添加的规则会覆盖旧规则。
   - **匹配字段**：此参数默认应用于所有字段，依据用户输入的正则表达式对字段进行过滤。
   - **Where 条件**：可依据用户输入的 where 条件对字段进行过滤。
   - **生成异常数据**：开启此项，表示异常数据讲按照配置的参数存储到规定的库中。
   - **数据库或 Schema**：开启“生成异常数据”时显示此项，表示存储异常数据的数据库或 Schema
   - **表前缀**：开启“生成异常数据”时显示此项，表示存储异常数据的表的前缀。
   - **表后缀**：开启“生成异常数据”时显示此项，表示存储异常数据的表的后缀。
   - **添加规则**：单击“添加规则”进行设置。例如，添加名称为“字段唯一值”规则，选中该规则后单击“确定”，在“告警条件”中输入告警条件表达式，然后按照此方法添加其他规则后，单击“确定”。
   - 告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。在此处的“关联质量规则”对话框中，每个质量规则的“告警参数”以按钮形式列出。

单个字段关联质量规则

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“维度表”页签，进入维度表页面。
3. 在维度表列表中，单击需要关联质量规则的维度表名称。
4. 在维度表的详情页的表字段列表中，查找字段并单击 ，配置单个表字段关联质量规则。

图3-214 维度表单个字段关联质量规则

5. 配置完成后，单击“确定”，完成维度表字段关联质量规则。
   - **更新已有规则**：若勾选此项，新添加的规则会覆盖旧规则。
   - **添加规则**：单击“添加规则”进行设置。例如，添加名称为“字段唯一值”规则，选中该规则后单击“确定”，在“告警条件”中输入告警条件表达式，然后按照此方法添加其他规则后，单击“确定”。
   - 告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。在此处的“关联质量规则”对话框中，每个质量规则的“告警参数”以按钮形式列出。

图3-215 添加规则界面

表字段批量关联质量规则

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。

2. 单击“维度表”页签，进入维度表页面。
3. 在维度表列表中，单击需要关联质量规则的维度表名称。
4. 在维度表的详情页的表字段列表中，勾选需要关联质量规则的表字段，单击关联质量规则。

图3-216 维度表批量字段关联质量规则

5. 在弹出的界面中添加规则，完成规则参数配置。
   - **更新已有规则**：若勾选此项，新添加的规则会覆盖旧规则。
   - **添加规则**：单击“添加规则”进行设置。例如，添加名称为“字段唯一值”规则，选中该规则后单击“确定”，在“告警条件”中输入告警条件表达式，然后按照此方法添加其他规则后，单击“确定”。
   - 告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。在此处的“关联质量规则”对话框中，每个质量规则的“告警参数”以按钮形式列出。

图3-217 规则设置界面

6. （可选）如需要将质量作业中不符合设定规则的异常数据存储在异常表中，可以打开“异常数据输出配置”开关。

图3-218 异常数据输出开关

点击开关，并打开“生成异常数据”按钮，表示异常数据将按照配置的参数存储到规定的库中。

图3-219 异常数据输出配置

各参数具体含义如下：

- 数据库或 Schema：表示存储异常数据的数据库或 Schema。

- 表前缀：表示存储异常数据的表的前缀。
- 表后缀：表示存储异常数据的表的后缀。

配置完成后点击 保存配置。

7. （可选）质量规则的检查范围默认是全表，如需要精确定位分区查询数据，请填写 where 条件。

图3-220 where 条件开关

8. 配置完成后，单击“确定”，完成维度表字段批量关联质量规则。

删除维度表

如果待删除的维度表处于待发布、已发布或待下线状态，则无法删除。用户可以通过维度管理来删除维度表。

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“维度表”页签，进入维度表页面。
3. 在维度表列表中，勾选需要删除的维度表，单击列表左上方的“删除”按钮，系统弹出“删除”对话框。

图3-221 删除维度表

4. 如果确认要删除，单击“是”。

3.4.6.2.3 新建事实表

归属于某个业务过程的事实逻辑表，可以丰富具体业务过程所对应事务的详细信息。创建事实逻辑表即完成公共的事务明细数据沉淀，从而便于提取业务中事务相关的明细数据。

新建事实表并发布

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-222 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“维度建模”，选择“事实表”页签。
3. 在左侧主题树中选中一个主题，然后单击“新建”按钮。

4. 在“新建事实表”页面，完成如下配置：

   a. 设置“基本配置”参数：

图3-223 事实表基本配置

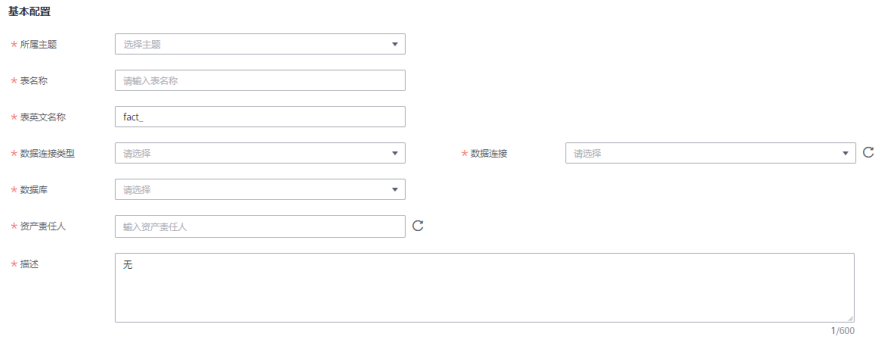


表3-143 基本配置参数说明

参数名称	说明
所属主题	单击“选择主题”，选择表所属的主题域分组、主题域和业务对象。
表名称	只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文和英文字母开头。
表英文名称	只能以英文字母开头，支持英文字母、数字、下划线。
数据连接类型	在下拉框中选择对应的数据连接类型。
数据连接	在下拉框中选择对应的数据连接。维度建模建议使用统一的数据连接。
数据库	在下拉框中选择对应的数据库。
队列	DLI 队列。该参数仅 DLI 连接类型有效。
Schema	DWS 或 POSTGRESQL 的模式。该参数在 DWS 或 POSTGRESQL 连接类型有效。
表类型	DWS 连接类型的表支持以下表类型： • DWS_ROW：行存表。行存储是指将表按行存储到硬盘分区上。 • DWS_COLUMN：列存表。列存储是指将表按列存储到硬盘分区上。 MRS_HIVE 仅支持 HIVE_TABLE 类型。
DISTRIBUTE BY	该参数仅 DWS 连接类型有效，为非必选项。您需要先添加表字段，才能在此下拉列表中选择某一个表字段作为 DISTRIBUTE BY 字段，可选取多个字段。 DWS 表当前支持复制（Replication）和散列（Hash）两种分布策

参数名称	说明
	略。 • **REPLICATION**：在每一个 DN 节点上存储一份全量表数据。这种存储方式的优点是每个 DN 上都有此表的全量数据，在 join 操作中可以避免数据重分布操作，从而减小网络开销；缺点是每个 DN 都保留了表的完整数据，造成数据的冗余。一般情况下只有较小的维度表才会定义为 Replication 表。 • **HASH**：采用这种分布方式，需要为用户表指定一个分布列（distribute key）。当插入一条记录时，系统会根据分布列的值进行 hash 运算后，将数据存储在对应的 DN 中。对于 Hash 分布表，在读/写数据时可以利用各个节点的 IO 资源，大大提升表的读/写速度。一般情况下大表（1000000 条记录以上）定义为 Hash 表。
资产责任人	根据下拉框选择对应的资产责任人，可以手动输入名字或直接选择已有的责任人。
描述	描述信息。支持的长度 0~600 字符。

b. 在“字段配置”区域，单击“新建”添加维度或度量字段。
   - 选择新建“维度”字段，可勾选一个或多个已创建的维度，单击“确定”后，会将维度的代理键字段添加到列表中；
   - 选择新建“度量”字段，需要新建度量字段。

   字段配置参数请参见表 3-144。字段配置完成后，单击字段后的或可以调整字段的顺序。

图3-224 配置维度或度量字段

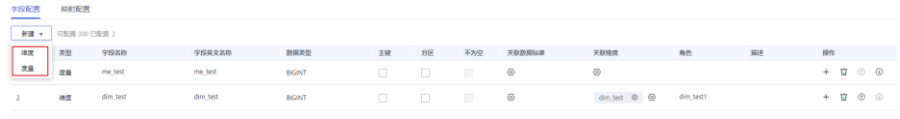


表3-144 字段配置参数

参数名称	说明
字段名称	只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文和英文字母开头。 维度属性的字段会自动显示所添加的维度的代理键名称，一般不需要修改。
字段英文名称	只能以英文字母开头，支持英文字母、数字、下划线。
数据类型	显示该维度的数据类型。

参数名称	说明
主键	选中时表示该字段为主键。
分区	选中时表示该字段为分区字段。
不为空	是否限制该字段不为空。
关联数据标准	如果您已创建数据标准，在“数据标准”列，单击 按钮可以选择一个数据标准与字段相关联。在“配置中心 > 功能配置”页面中的“模型设计业务流程步骤 > 创建质量作业”勾选的情况下，将字段关联数据标准后，表发布上线后，就会自动生成一个质量作业，每个关联了数据标准的字段会生成一个质量规则，基于数据标准对字段进行质量监控，您可以前往 DataArts Studio 数据质量模块的“质量作业”页面进行查看。 如果您还未创建数据标准，请参见 3.4.5.2 新建数据标准进行创建。
关联维度	只有维度属性的字段需要绑定维度，度量属性的字段不需要进行此操作。 显示当前关联的维度名称。单击 可以更换关联的维度。
角色	只有维度属性的字段被添加多次时需要设置角色区分，度量属性的字段不需要进行此操作。 当同一个维度被添加多次时，需要设置不同的角色来加以区分。
描述	描述信息。

c. 在“映射配置”页签，单击“新建映射”，配置映射参数。

图3-225 配置映射

表3-145 映射参数

参数名称	说明
映射名称	只能包含中文、英文字母、数字和下划线。

参数名称	说明
来源模型	在下拉列表中选择一个已创建的关系模型。如果未创建关系模型，请参见 3.4.6.1.2 物理模型设计进行创建。
源表	选择数据来源的表，如果数据来源于一个模型中的多个表，可以单击表名后的按钮为该表和其他表之间设置 JOIN。 1. 选择一种“JOIN 方式”，“JOIN 方式”从左到右依次表示 left JOIN、right JOIN、inner JOIN、outer JOIN。 2. 在“JOIN 字段”中设置 JOIN 条件，JOIN 条件一般选择源表和 JOIN 表中含义相同的字段，单击 + 或 − 按钮增加或删除 JOIN 条件。JOIN 条件之间是 and 的关系。 3. 单击“确定”完成设置。 4. 设置 JOIN 后，如果想删除 JOIN 表，单击所需删除的表名后的按钮就可以删除该 JOIN 表。 图3-226 JOIN 条件 
字段映射	为来源于当前映射的字段，依次选择一个含义相同的源字段。如果表字段来源于多个模型，您需要新建多个映射，每个映射中，您只需要为来源于当前映射的字段设置源字段，其他字段可以不设置。

5. 单击“发布”，提交事实表的发布审核。
6. 等待审核人员审核事实表。

   审核通过后，事实表就会在数据库中自动创建。
7. 返回“维度建模 > 事实表”页面，在列表中找到刚发布的事实表，在“同步状态”一列中可以查看事实表的同步状态。
   - 如果同步状态均显示成功，则说明事实表发布成功，事实表在数据库中已创建成功。
   - 如果同步状态中存在失败，可单击该事实表所在行的“更多 > 发布历史”，然后在展开的页面中进入“发布日志”页签查看日志。请根据错误日志定位失败原因，问题解决后，您可以在事实表页面勾选该事实表，再单击列表上方的“更多 > 同步”尝试重新同步。如果问题仍未能解决，请联系技术支持人员协助处理。

管理事实表

事实表创建好之后，进入数据架构的“维度建模 > 事实表”页面，您可以对事实表进行编辑、发布、下线、查看发布历史或删除操作。

图3-227 事实表管理

- **编辑事实表**
  a. 在事实表列表中，找到需要编辑的事实表，单击“编辑”，进入编辑事实表页面。
  b. 根据实际需要编辑相关内容。
  c. 单击“保存”，保存设置的信息；单击“发布”，发布设置的信息。
- **发布事实表**
  a. 在事实表列表中，勾选需要发布的事实表，单击“发布”按钮，弹出“批量发布”对话框。
  b. 在下拉菜单中选择审核人。
  c. 单击“确认提交”。
- **查看发布历史**
  a. 在事实表列表中，找到所需要的事实表，在右侧单击“更多 > 发布历史”，将显示“发布历史”页面。
  b. 在“发布历史”中，您可以查看事实表的发布历史、版本对比信息以及发布日志。

     如果“发布日志”中有错误日志，说明发布失败。您可以单击“重新同步”将表同步到 DataArts Studio 的其他模块中。
- **关联质量规则**
  a. 在事实表列表中，勾选所需要的关联质量规则事实表，在上方单击“关联质量规则”，弹出“关联质量规则”对话框。
  b. 在“关联质量规则”对话框中，您可以批量给事实表的字段添加规则并关联到字段。
  c. 单击“确定”。
- **预览 SQL**
  a. 在事实表列表中，找到所需要的事实表，在右侧单击“更多 > 预览 SQL”，弹出“预览 SQL”对话框。
  b. 在“预览 SQL”中，您可以查看 SQL 语句，也可以复制 SQL。
- **下线事实表**

a. 在事实表列表中，勾选需要下线的事实表，单击“下线”，系统弹出“批量下线”对话框。

b. 在下拉菜单中选择审核人。

c. 单击“确认提交”。

说明

- “下线”及“删除”事实逻辑表的前提是无依赖引用，例如事实表未被原子指标等使用时，才能进行删除操作。

- **删除事实表**

  如果您不再需要某一个事实表，您可以将它删除。当事实表处于待发布、已发布或待下线状态时，无法删除。

  a. 在事实表列表中，勾选需要删除的事实表，在列表上方选择“更多 > 删除”，系统弹出“删除”对话框。

  b. 单击“是”。

- **导入事实表**

  可通过导入的方式将事实表批量快速的导入到系统中。

  a. 在事实表上方，单击“更多>导入”，进入“导入配置”页签。

  图3-228 导入表

  b. 下载事实表导入模板，编辑完成后保存至本地。

  c. 选择是否更新已有数据。

  说明

  如果系统中已有的编码和模板中的编码相同，系统则认为是数据重复。

  - 不更新：当数据重复时，不会替换系统中原有的数据。
  - 更新：当数据重复时
    - 系统中的原有数据为草稿状态，则会覆盖生成新的草稿数据。
    - 系统中的原有数据为发布状态，则会生成下展数据。

  d. 单击“添加文件”，选择编辑完成的导入模板。

  e. 单击“上传文件”，上传完成后，自动跳转到“上次导入”页签，查看已导入的数据。

f. 单击“关闭”。

- **导出事实表**

  可通过导出的方式将事实表导出到本地。在事实表上方，单击“更多>导出”，即可将系统中的事实表导出到本地。

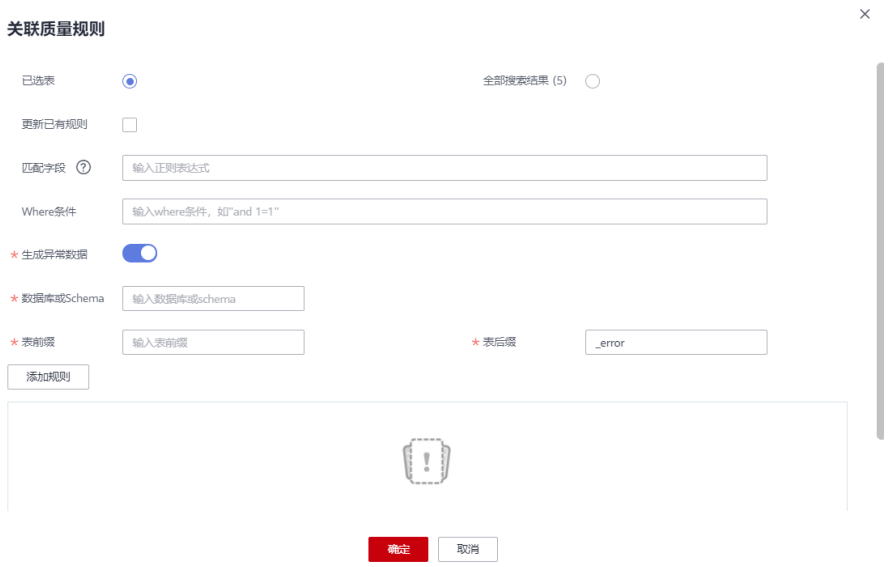
事实表关联质量规则

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“事实表”页签，进入事实表页面。
3. 在事实表列表中，勾选需要关联质量规则的事实表。单击“关联质量规则”。

图3-229 关联事实表质量规则

4. 在弹出的页面中配置关联质量规则参数。配置完成单击确定。
   - **更新已有规则**：若勾选此项，新添加的规则会覆盖旧规则。
   - **匹配字段**：此参数默认应用于所有字段，依据用户输入的正则表达式对字段进行过滤。
   - **Where 条件**：可依据用户输入的 where 条件对字段进行过滤。
   - **生成异常数据**：开启此项，表示异常数据讲按照配置的参数存储到规定的库中。
   - **数据库或 Schema**：开启“生成异常数据”时显示此项，表示存储异常数据的数据库或 Schema
   - **表前缀**：开启“生成异常数据”时显示此项，表示存储异常数据的表的前缀。
   - **表后缀**：开启“生成异常数据”时显示此项，表示存储异常数据的表的前缀。
   - **添加规则**：单击“添加规则”进行设置。例如，添加名称为“字段唯一值”规则，选中该规则后单击“确定”，在“告警条件”中输入告警条件表达式，然后按照此方法添加其他规则后，单击“确定”。
   - 告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。在此处的“关联质量规则”对话框中，每个质量规则的“告警参数”以按钮形式列出。

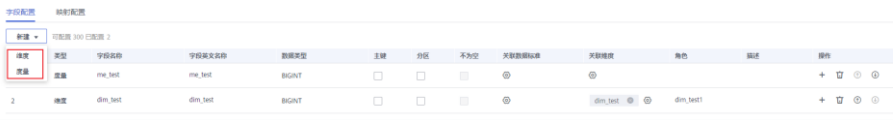
图3-230 事实表关联质量规则



事实表新建字段

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“事实表”页签，进入事实表页面。
3. 在事实表列表中，查找需要新建字段的表名称，单击其“编辑”，进入编辑页。
4. 单击字段配置处的新建，在展开的下拉框选择新建字段类型，并配置相关参数。

图3-231 新建字段

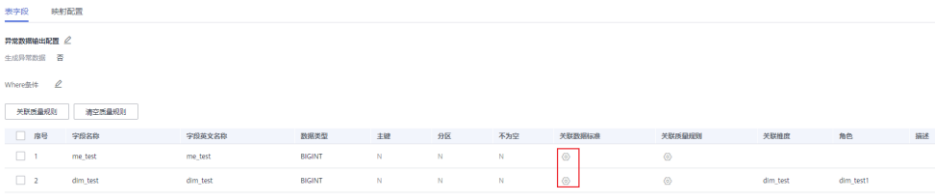


5. 配置完成后，单击“确定”，完成事实表新建字段。

事实表字段关联数据标准

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“事实表”页签，进入事实表页面。
3. 在事实表列表中，单击需要关联数据标准的事实表名称。
4. 在事实表的详情页的表字段列表中，查找需要关联数据标准的字段，单击其所属的 ，配置单个表字段关联数据标准。数据标准的来源请参考新建数据标准

图3-232 事实表字段关联数据标准



5. 配置完成后，单击“确定”，完成事实表字段关联数据标准。

图3-233 设置数据标准

事实表字段单个关联质量规则

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“事实表”页签，进入事实表页面。
3. 在事实表列表中，单击需要关联质量规则的事实表名称。
4. 在事实表的详情页的表字段列表中，单击 ，配置单个表字段关联质量规则。

图3-234 事实表单个字段关联质量规则

5. 配置完成后，单击“确定”，完成事实表字段关联质量规则。

图3-235 添加事实表质量规则

事实表字段批量关联质量规则

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“事实表”页签，进入事实表页面。
3. 在事实表列表中，单击需要关联质量规则的事实表名称。
4. 在事实表的详情页的表字段列表中，勾选需要关联质量规则的表字段，单击关联质量规则。

图3-236 事实表批量字段关联质量规则

5. 在弹出的界面中添加规则，完成规则参数配置。

图3-237 规则配置页

6. 配置完成后，单击“确定”，完成事实表字段批量关联质量规则。

3.4.7 指标设计

3.4.7.1 业务指标

经过数据调研和需求分析之后，您需要根据需求落地指标。指标是衡量目标总体特征的统计数值，是能表征企业某一业务活动中业务状况的数值指示器。指标一般由指标名称和指标数值两部分组成，指标名称及其涵义体现了指标质的规定性和量的规定性两个方面的特点，指标数值反映了指标在具体时间、地点、条件下的数量表现。业务指标用于指导技术指标，而技术指标是对业务指标的具体实现。

前提条件

在新建业务指标之前，您需要先完成流程设计，具体操作请参见 3.4.4.1 流程设计。

新建业务指标并发布

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-238 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“业务指标”，进入业务指标页面。
3. 在左侧的流程树中，选中一个流程，单击“新建”开始新建业务指标。
4. 在“新建业务指标”页面，请根据以下步骤配置参数，配置完成后，单击“发布”。
   a. 填写“基本信息”参数。

图3-239 新建业务指标

表3-146 指标基本信息参数

参数说明	说明
指标名称	业务指标的名称。
指标编码	• 指标编码是自动生成的，生成规则可以在 DataArts Studio 数据架构的“配置中心”页面进行配置，详情请参见编码规则。
指标别名	可选参数。
所属流程	选择指标所属的业务流程。如果您还未创建业务流程，请参见 3.4.4.1 流程设计进行创建。
设置目的	描述设置该指标的目的。
指标定义	需准确描述指标的定义。
备注	备注信息。

b. 配置指标数据信息。

图3-240 指标数据信息

表3-147 指标数据信息参数

参数说明	说明
计算公式	定义业务指标的计算逻辑，以便指导开发者根据计算公式设计原子指标、衍生指标。业务指标是为了指导技术指标的落地，实际并不做运算。
统计周期	指定指标的统计周期，以便指导开发者根据统计周期设计时间限定。
统计维度	可以在下拉列表中选择已经创建的维度。维度的创建请参见 3.4.6.2.1 新建维度。

参数说明	说明
统计口径和修饰词	限定是对业务场景限定抽象，用于度量范围的圈定。
刷新频率	指标数据的刷新频率。开发者或运维者可以依据指标的刷新频率，合理设置衍生指标的调度频率。
指标应用场景	描述指标的应用场景。
关联技术指标类型	下拉选择与业务指标关联的技术指标类型，包含衍生指标和复合指标。
关联技术指标	下拉选择与业务指标关联的技术指标。
度量对象	衡量该指标的度量字段。
计量单位	指标的计量单位。

c. 配置管理信息。

图3-241 管理信息

表3-148 管理信息参数说明

参数说明	说明
数据来源	描述数据来源，也就是数据的产生者。
指标管理部门	指标的管理部门。
指标责任人	指标的责任人，可以手动输入名字或直接选择已有的责任人。

5. 在弹出对话框中，选择审核人，单击“确认提交”，完成发布。
6. 可以参照步骤 3~步骤 5，完成其他业务指标的创建和发布。
7. 完成所有业务指标的新建和发布之后，需要等待审核人员审核。
   审核通过后，业务指标创建完成。

编辑业务指标

1. 在数据架构控制台，单击左侧导航树中的“业务指标”，进入业务指标页面。

图3-242 管理业务指标

2. 在业务指标列表中找到需要编辑的指标，单击“编辑”，进入编辑业务指标页面。
3. 根据实际需要编辑业务指标的相关信息。
4. 单击“保存”，保存所做的修改。或者，单击“发布”，发布修改后的业务指标。

发布业务指标

如果新建了业务指标但并未发布，可以执行以下步骤发布业务指标：

步骤 1 在数据架构控制台，单击左侧导航树中的“业务指标”，进入业务指标页面。

步骤 2 在业务指标列表中找到需要发布的指标，单击“发布”。

步骤 3 在弹出对话框中，选择审核人，单击“确认提交”，完成发布。

图3-243 提交发布

----结束

下线业务指标

对于已发布的业务指标，可以执行以下步骤下线业务指标：

步骤 1 在数据架构控制台，单击左侧导航树中的“业务指标”，进入业务指标页面。

步骤 2 在业务指标列表中找到需要下线的业务指标，单击“下线”。

步骤 3 在弹出对话框中，选择审核人，然后单击“确认提交”，审核通过后，完成业务指标的下线。

图3-244 提交下线

----结束

删除业务指标

如果您已不再需要某个业务指标，可以删除该业务指标。如果待删除的业务指标已发布，则无法执行删除操作，您必须先将该业务指标下线后，才能执行删除操作。

1. 在数据架构控制台，单击左侧导航树中的“业务指标”，进入业务指标页面。
2. 在维度列表中找到需要删除的业务指标度，勾选该业务指标，然后单击业务指标列表上方“更多”中的“删除”。

图3-245 删除业务指标

3. 在系统弹出的“删除”对话框中，确认无误后，单击“是”将业务指标删除。

3.4.7.2 技术指标

3.4.7.2.1 新建原子指标

原子指标是对指标统计逻辑、具体算法的一个抽象。为了从根源上解决定义、研发不一致的问题，指标定义明确设计统计逻辑（即计算逻辑），不需要 ETL 二次或者重复研发，从而提升了研发效率，也保证了统计结果的一致性。

背景信息

原子指标来源于事实表：

- 原子指标是为了构建应用统计分析所需的衍生指标，而定义的数据组件，因此只可以基于事实逻辑表明细数据表来创建。
- 衍生指标无来源表，它归属于每个组合成它的原始的原子指标的来源表。

原子指标与衍生指标的关系：

- 原子指标的计算逻辑修改生效后，会直接更新应用于相关的衍生指标。
- 原子指标删除英文名，需要校验下游是否有衍生指标使用，如果有，则无法删除。
- 目前原子指标在被下游使用的情况下，支持变更英文名。
- 原子指标的更改会影响下游衍生指标。

前提条件

您已创建并发布事实表，且事实表已通过审核，具体操作请参见 3.4.6.2.3 新建事实表。

新建原子指标并发布

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-246 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“技术指标”，选择“原子指标”页签进入原子指标页面。
3. 在左侧主题目录中选中一个主题，然后单击“新建”按钮，开始新建原子指标。
4. 在新建原子指标页面，参考表 3-149 配置参数，然后单击“发布”。

图3-247 新建原子指标

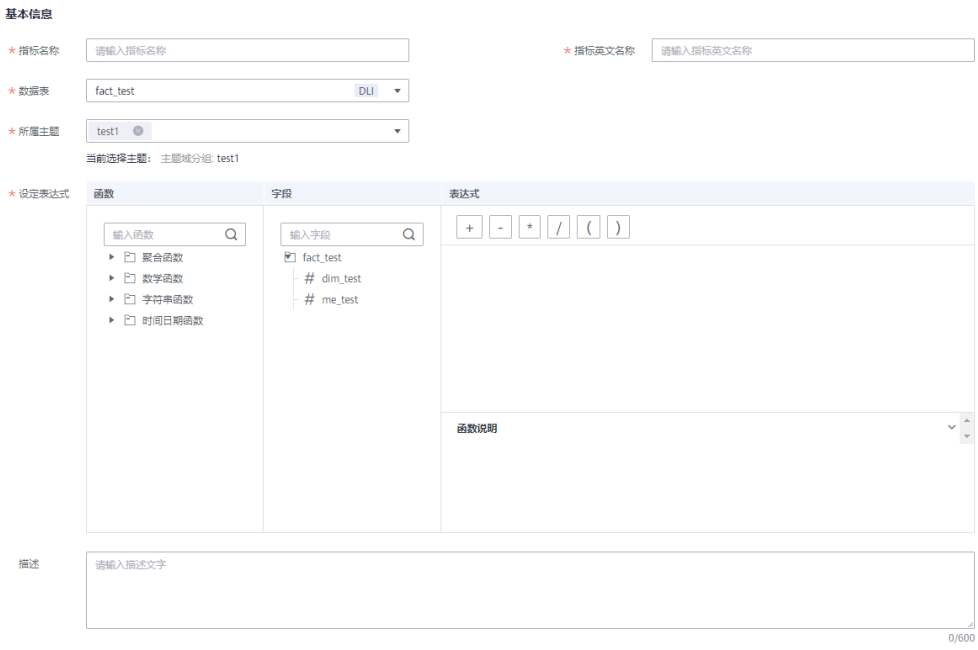


表3-149 新建原子指标参数说明

参数名称	说明
*指标名称	只能包含中文、英文字母、数字和下划线，且以中文或英文字母开头。
*指标英文名称	只能包含英文字母、数字和下划线，且以英文字母开头。
*数据表	在下拉列表中选择一个已发布的事实表，如果表很多，您也可以在下拉列表的输入框中输入表名称搜索事实表。如果您尚未创建事实表，请参见新建事实表并发布进行创建并发布。
*所属主题	原子指标所属的主题信息。当“数据表”选择事实表后，将自动显示事实表所属的主题信息，您也可以单击“选择主题”进行选择。
*设定表达式	根据实际情况选择所需要的函数和字段，并设定表达式。
描述	描述信息。支持的长度为0~600个字符。

5. 在弹出框中单击“确认提交”，提交审核。
6. （可选）参考步骤3~步骤5，完成其他原子指标的发布。
7. 等待审核人员审核。

   审核通过后，原子指标创建完成。

管理原子指标

1. 在数据架构控制台，单击左侧导航树中的“技术指标”，选择“原子指标”页签，进入原子指标页面。

   图3-248 管理原子指标

2. 您可以根据实际需要选择如下操作。

当需要...	则...
新建	执行新建原子指标并发布。
编辑	执行3。
发布	执行4。
查看发布历史	执行5。

当需要...	则...
下线	执行 6。
删除	执行 7。
导入	执行 8。
导出	执行 9。

3. 编辑
   a. 在需要编辑的原子指标右侧，单击“编辑”，进入编辑原子指标页面。
   b. 根据实际需要编辑相关内容。
   c. 单击“发布”。如果您暂时不想发布，可以先单击“保存”，稍后再发布。
4. 发布
   a. 在需要发布的原子指标右侧，单击“发布”，弹出“提交发布”对话框。
   b. 在下拉菜单中选择审核人。
   c. 单击“确认提交”。
5. 查看发布历史
   a. 在列表中，找到所需查看的原子指标，单击“更多 > 发布历史”，将显示“发布历史”页面。
   b. 在“发布历史”中，您可以查看原子指标的发布历史和版本对比信息。
6. 下线
   a. 在需要下线的原子指标右侧，单击“更多 > 下线”，系统弹出“提交下线”对话框。
   b. 在下拉菜单中选择审核人。
   c. 单击“确认提交”。

说明

下线及删除原子指标的前提是无依赖引用，即无衍生指标引用。

7. 删除
   a. 勾选需要删除的原子指标，单击上方“更多 > 下线”，系统弹出“删除”对话框。
   b. 单击“是”。
8. 导入

   可通过导入的方式将原子指标批量快速的导入到系统中。
   a. 在原子指标列表上方，单击“更多>导入”，进入“导入配置”页签。

图3-249 导入原子指标

b. 下载原子指标导入模板，编辑完成后保存至本地。

c. 选择是否更新已有数据。

说明

如果系统中已有的编码和模板中的编码相同，系统则认为是数据重复。

- 不更新：当数据重复时，不会替换系统中原有的数据。
- 更新：当数据重复时
  - 系统中的原有数据为草稿状态，则会覆盖生成新的草稿数据。
  - 系统中的原有数据为发布状态，则会生成下展数据。

d. 单击“添加文件”，选择编辑完成的导入模板。

e. 单击“上传文件”，上传完成后，自动跳转到“上次导入”页签，查看已导入的数据。

f. 单击“关闭”。

9. 导出

可通过导出的方式将原子指标导出到本地。

a. 在原子指标列表选中待导出的指标。

b. 在列表上方，单击“更多>导出”，即可将系统中的原子指标导出到本地。

说明

- 在左侧主题树中选中某个主题，可以导出该主题下的所有原子指标；
- 当该空间下不超过 500 条原子指标数据时可以全部导出。

3.4.7.2.2 新建衍生指标

衍生指标是原子指标通过添加限定、维度卷积而成，限定、维度均来源于原子指标中的属性。发布衍生指标时，会自动生成一张汇总表，可在“汇总表-自动汇聚”下查看。

衍生指标=原子指标+统计维度+时间限定+通用限定。

- **原子指标**：明确统计口径，即计算逻辑。

- **统计维度**：用于观察和分析业务数据的视角，支撑对数据进行汇聚、钻取、切片分析，用于 SQL 中的 GROUP BY 条件。
- **时间限定**：时间限定是时间条件限制的标准化定义。
- **通用限定**：统计的业务范围，筛选出符合业务规则的记录（类似于 SQL 中 where 后面的条件，不包括时间区间）。

前提条件

- 在新建衍生指标之前，请先确认原子指标已经新建并通过审核。
- 如果衍生指标将使用统计维度或时间限定，请先确认维度或时间限定已经新建并通过审核。

新建衍生指标并发布

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

   图3-250 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“技术指标”，选择“衍生指标”页签进入衍生指标页面。
3. 在左侧的主题目录中选中一个主题，然后单击“新建”按钮，开始新建衍生指标。
4. 在新建衍生指标页面，根据页面提示配置参数。

图3-251 新建衍生指标

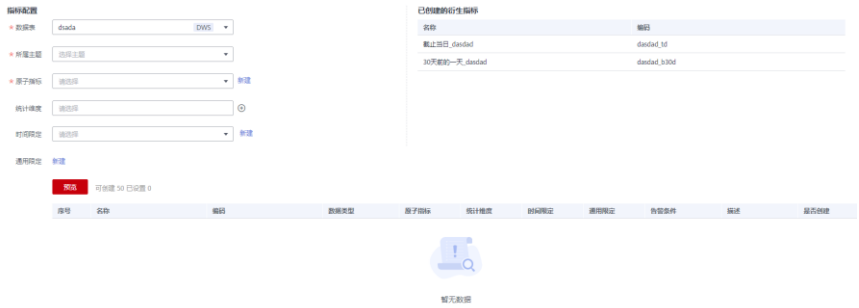
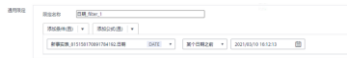


表3-150 新建衍生指标参数说明

参数名称	说明
*数据表	在下拉列表中选择即可。
*所属主题	显示所属的主题信息。
*原子指标	选择原子指标。
统计维度	在下拉列表中，选择一个或多个维度。此处只能选择原子指标所关联的事实表中的属性。
时间限定	在下拉框中选择所需要的时间限定，并选择关联的字段。系统预置了一些时间限定，如果不能满足需求，请参考 3.4.7.2.4 新建时间限定进行创建。
通用限定	如需设置通用限定，可以单击“新建”按钮新建一个或多个通用限定。 如图 3-252 所示，在新建通用限定区域，通过以下配置新建一个通用限定。 • **限定名称**：指定通用限定的名称。 • **添加条件(且)**：单击该下拉框，选择“且条件”或者“或条件”可以添加相应的条件，然后在字段下拉框中选择一个字段，并根据页面提示设置条件。您可以添加多个条件。 在某个条件后面单击删除按钮 ，可以将该条件删除。 • **添加公式(且)**：单击该下拉框，选择“且公式”或者“或公式”可以添加相应的公式，然后再单击“编辑公式”按钮，在弹出对话框中选择所需要的“函数”和“字段”，并设置“表达式”。 在某个公式后面单击删除按钮 ，可以将该公式删除。 图3-252 通用限定

参数名称	说明
	
告警配置	由衍生指标和表达式组成，表达式由告警参数和逻辑运算符组成。在指标运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。
描述	描述信息。支持的长度为 0~600 个字符。

5. 参数配置完成后，单击“预览”，可以查看该衍生指标的相关信息，并定义名称、编码、数据类型、告警条件和描述等信息。

表3-151 预览衍生指标参数说明

参数名称	说明
名称	系统已根据原子指标、统计维度、时间限定等参数自动生成，您也可以自定义。
编码	系统已根据原子指标、统计维度、时间限定等参数编码自动生成，您也可以自定义。
数据类型	系统已根据原子指标的数据类型自动生成，您也可以自定义。
告警条件	告警条件表达式由告警参数和逻辑运算符组成。在指标运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。
描述	描述信息。支持的长度为 0~600 个字符。

6. 在页面下方，单击“试运行”按钮，然后在弹出框中单击“试运行”按钮，测试所设置的衍生指标是否可以正常运行。

   如果试运行失败，请根据错误提示定位错误原因，将配置修改正确后，再单击“试运行”按钮进行重试。
7. 如果试运行成功，单击“发布”，提交发布审核。
8. 在弹出框中单击“确认提交”，提交审核。
9. （可选）参考步骤 2~步骤 8，完成其他衍生指标的发布。
10. 等待审核人员审核。

    审核通过后，衍生指标创建完成。

管理衍生指标

进入数据架构的“技术指标 > 衍生指标”页面，您可以对衍生指标进行编辑、发布、下线、查看发布历史或删除操作。

图3-253 管理衍生指标

1. 在数据架构控制台，单击左侧导航树中的“技术指标”，选择“衍生指标”页签，进入衍生指标页面。
2. 您可以根据实际需要选择如下操作。

当需要...	则...
新建	执行新建衍生指标并发布。
编辑	执行 3。
发布	执行 4。
查看发布历史	执行 5
预览 SQL	执行 6
下线	执行 7。
查看汇总表	执行 8。
删除	执行 9。
导入	执行 10。
导出	执行 11。

3. 编辑
   a. 在需要编辑的衍生指标右侧，单击“编辑”，进入编辑衍生指标页面。
   b. 根据实际需要编辑相关内容。
   c. 在页面下方，单击“试运行”按钮，然后在弹出框中单击“试运行”按钮，测试所设置的衍生指标是否可以正常运行。

      如果试运行失败，请根据错误提示定位错误原因，将配置修改正确后，再单击“试运行”按钮进行重试。
   d. 如果试运行成功，单击“发布”，提交发布审核。
4. 发布
   a. 在需要发布的衍生指标右侧，单击“发布”，弹出“提交发布”对话框。
   b. 在下拉菜单中选择审核人。

c. 单击“确认提交”。

5. 查看发布历史

a. 在列表中，找到需要查看的衍生指标，在右侧单击“更多 > 发布历史”，将显示“发布历史”页面。

b. 在“发布历史”中，您可以查看衍生指标的发布历史和版本对比信息。

6. 预览 SQL

a. 在列表中，找到所需要的衍生指标，在右侧单击“更多 > 预览 SQL”，弹出“预览 SQL”对话框。

b. 在“预览 SQL”中，您可以查看 SQL 语句，也可以复制 SQL。

7. 下线

说明

下线衍生指标的前提是无依赖引用，即无复合指标引用。

a. 在需要下线的衍生指标右侧，单击“更多 > 下线”，系统弹出“提交下线”对话框。

b. 在下拉菜单中选择审核人。

c. 单击“确认提交”。

8. 查看汇总表

当前仅支持查看自动汇聚的汇总表详情。在需要查看汇总表的指标右侧，选择“更多 > 查看汇总表”，跳转到汇总表详情页面。

9. 删除

说明

删除衍生指标的前提是无依赖引用，即无复合指标引用。

a. 在衍生指标列表中，勾选需要删除的衍生指标，单击页面上方“更多 > 删除”，系统弹出“删除”对话框。

b. 单击“是”。

10. 导入

可通过导入的方式将衍生指标批量快速的导入到系统中。

a. 在汇总表上方，单击“更多>导入”，进入“导入配置”页签。

图3-254 导入衍生指标

b. 下载衍生指标导入模板，编辑完成后保存至本地。

c. 选择是否更新已有数据。

说明

如果系统中已有的编码和模板中的编码相同，系统则认为是数据重复。

- 不更新：当数据重复时，不会替换系统中原有的数据。
- 更新：当数据重复时
  - 系统中的原有数据为草稿状态，则会覆盖生成新的草稿数据。
  - 系统中的原有数据为发布状态，则会生成下展数据。

d. 单击“添加文件”，选择编辑完成的导入模板。

e. 单击“上传文件”，上传完成后，自动跳转到“上次导入”页签，查看已导入的数据。

f. 单击“关闭”。

11. 导出

可通过导出的方式将衍生指标导出到本地。

a. 在衍生指标列表选中待导出的指标。

b. 在列表上方，单击“更多>导出”，即可将系统中的衍生指标导出到本地。

说明

- 在左侧主题树中选中某个主题，可以导出该主题下的所有衍生指标；
- 当该空间下不超过 500 条衍生指标数据时可以全部导出。

3.4.7.2.3 新建复合指标

复合指标是由一个或多个衍生指标叠加计算而成，其中的维度、限定均继承于衍生指标。注意，不能脱离衍生指标、维度和限定的范围，去产生新的维度和限定。

前提条件

您已新建衍生指标，并且衍生指标已通过审核，具体操作请参见 3.4.7.2.2 新建衍生指标。

新建复合指标

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-255 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“技术指标”，选择“复合指标”页签。
3. 在“复合指标”页面，在左侧的主题目录中选中一个主题，然后单击“新建”按钮。
4. 在新建复合指标页面，根据页面提示配置以下参数。

图3-256 新建复合指标

表3-152 新建复合指标参数说明

参数名称	说明
*复合指标名称	只能包含中文、英文字母、数字和下划线，且必须以中文或英文字母开头。
*复合指标英文名称	只能包含英文字母、数字和下划线，且必须以英文字母开头。

参数名称	说明
*所属主题	显示所属的主题信息。您也可以单击“选择主题”进行选择。
*统计维度	选择来源于衍生指标的统计维度。
*数据类型	选择复合指标的数据类型。
*设定表达式	选择所需要的衍生指标，并根据实际需求在“表达式”中设置表达式。
告警配置	由衍生指标和表达式组成，表达式由告警参数和逻辑运算符组成。在指标运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。 说明 当前暂不支持触发质量告警。
描述	描述信息。支持的长度为 0~600 个字符。

5. 在页面下方，单击“试运行”按钮，然后在弹出框中单击“试运行”按钮，测试所设置的复合指标是否可以正常运行。

   如果试运行失败，请根据错误提示定位错误原因，将配置修改正确后，再单击“试运行”按钮进行重试。
6. 如果试运行成功，单击“发布”，提交发布审核。
7. 在弹出框中单击“确认提交”，提交审核。
8. 等待审核人员审核。

   审核通过后，复合指标创建完成。

编辑复合指标

1. 在数据架构控制台，单击左侧导航树的“技术指标”，然后选择“复合指标”页签，进入复合指标页面。

   图3-257 复合指标

2. 在复合指标列表中，找到需要编辑的复合指标，单击“编辑”，进入“编辑复合指标”页面。
3. 根据实际需要修改配置参数。参数说明请参见表 3-152。
4. 在页面下方，单击“试运行”按钮，然后在弹出框中单击“试运行”按钮，测试所设置的复合指标是否可以正常运行。

如果试运行失败，请根据错误提示定位错误原因，将配置修改正确后，再单击“试运行”按钮进行重试。

5. 如果试运行成功，单击“发布”，提交发布审核。
6. 在弹出框中单击“确认提交”，提交审核。

发布复合指标

当您新建或编辑复合指标后，需要发布复合指标，才能使其生效。如果复合指标处于待发布、已发布或待下线状态，则无法发布。

1. 在数据架构控制台，单击左侧导航树的“技术指标”，然后选择“复合指标”页签，进入复合指标页面。
2. 在复合指标列表中，勾选需要发布的复合指标，单击“发布”按钮，弹出“批量发布”对话框。
3. 确认无误后，单击“确认提交”，提交审核。

查看发布历史

1. 在数据架构控制台，单击左侧导航树的“技术指标”，然后选择“复合指标”页签，进入复合指标页面。
2. 在复合指标列表中，找到需要查看的复合指标，单击“更多 > 发布历史”，将显示“发布历史”页面。
3. 在“发布历史”中，您可以查看复合指标的发布历史和版本对比信息。

预览 SQL

1. 在数据架构控制台，单击左侧导航树的“技术指标”，然后选择“复合指标”页签，进入复合指标页面。
2. 在复合指标列表中，找到需要查看的复合指标，单击“更多 > 预览 SQL”，弹出“预览 SQL”对话框。
3. 在“预览 SQL”中，您可以查看 SQL 语句，也可以复制 SQL。

下线复合指标

对于已发布的复合指标，如果不在需要使用，可以将其下线。

说明

下线复合指标的前提是无依赖引用，即无汇总表引用。

1. 在数据架构控制台，单击左侧导航树的“技术指标”，然后选择“复合指标”页签，进入复合指标页面。
2. 在复合指标列表中，勾选需要下线的复合指标，单击“下线”按钮，弹出“批量下线”对话框。
3. 确认无误后，单击“确认提交”。

删除复合指标

说明

删除复合指标的前提是无依赖引用，即无汇总表引用。

1. 在数据架构控制台，单击左侧导航树的“技术指标”，然后选择“复合指标”页签，进入复合指标页面。
2. 在复合指标列表中，勾选需要删除的复合指标，单击列表上方的“更多 >删除”按钮，系统弹出“删除”对话框。
3. 单击“确定”。

导入复合指标

可通过导入的方式将复合指标批量快速的导入到系统中。

1. 在复合指标列表上方，单击“更多>导入”，进入“导入配置”页签。

图3-258 导入复合指标

2. 下载复合指标导入模板，编辑完成后保存至本地。
3. 选择是否更新已有数据。

说明

如果系统中已有的编码和模板中的编码相同，系统则认为是数据重复。

- 不更新：当数据重复时，不会替换系统中原有的数据。
- 更新：当数据重复时
  - 系统中的原有数据为草稿状态，则会覆盖生成新的草稿数据。
  - 系统中的原有数据为发布状态，则会生成下展数据。

4. 单击“添加文件”，选择编辑完成的导入模板。
5. 单击“上传文件”，上传完成后，自动跳转到“上次导入”页签，查看已导入的数据。
6. 单击“关闭”。

导出复合指标

可通过导出的方式将复合指标导出到本地。

1. 在复合指标列表选中待导出的指标。
2. 在列表上方，单击“更多>导出”，即可将系统中的复合指标导出到本地。

说明

- 在左侧主题树中选中某个主题，可以导出该主题下的所有复合指标；
- 当该空间下不超过 500 条复合指标数据时可以全部导出。

3.4.7.2.4 新建时间限定

原子指标是计算逻辑的标准化定义，时间限定则是条件限制的标准化定义。为保障所有统计指标统一、标准、规范地构建，时间限定在业务板块内唯一，并唯一归属于一个来源逻辑表，计算逻辑也以该来源逻辑表模型的字段为基础进行定义。由于一个时间限定的定义可能来自于归属不同数据域的多个逻辑表，因此一个时间限定可能归属于多个数据域。

新建时间限定并发布

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-259 选择数据架构

2. （可选）在数据架构控制台，单击左侧导航树中的“配置中心”，在功能配置下选择是否开启“时间限定生成使用动态表达式”功能，默认关闭。

图3-260 功能配置

3. 在数据架构控制台，单击左侧导航树中的“技术指标”，选择“时间限定”页签。
4. 进入时间限定页面后，单击“新建”按钮。
5. 在新建时间限定页面，参考表 3-153 配置参数，然后单击“发布”。

图3-261 时间限定

表3-153 新建时间限定参数说明

参数名称	说明
*限定名称	只能包含中文、英文字母、数字和下划线，且必须以中文或英文字母开头。
*限定英文名称	只能包含英文字母、数字和下划线。
*时间配置	可选择“按年”、“按月”、“按日”、“按小时”或“按分钟”，然后根据需要选择“快速选择”或“自定义”进行时间条件的设置。 自定义时，“-”表示从当前时间向前的时间段，“+”表示从当前时间向后的时间段。例如，过去一年到未来三年，可以按年自定义为“-1 到+3”或“+3 到-1”。
描述	描述信息。支持的长度 0~490 字符。

6. 在弹出框中单击“确认提交”，提交发布审核。
7. 等待审核人员审核。

   审核通过后，时间限定创建完成。

管理时间限定

1. 在数据架构控制台，单击左侧导航树中的“技术指标”，选择“时间限定”页签，进入时间限定页面。

   图3-262 时间限定页面

2. 您可以根据实际需要选择如下操作。

当需要...	则...
新建	执行新建时间限定并发布。
编辑	执行 3。
发布	执行 4。
发布历史	执行 5。
下线	执行 6。
删除	执行 7。

3. 编辑
   a. 在需要编辑的时间限定右侧，单击“编辑”，进入编辑时间限定页面。
   b. 根据实际需要编辑相关内容。
   c. 单击“保存”，保存该时间限定信息；或者单击“发布”，发布该时间限定信息。
4. 发布
   a. 在需要发布的时间限定右侧，单击“发布”，弹出“提交发布”对话框。
   b. 在下拉菜单中选择审核人。
   c. 单击“确认提交”。
5. 发布历史
   a. 在列表中，找到所需查看的时间限定，单击“更多 > 发布历史”，将显示“发布历史”页面。
   b. 在“发布历史”中，您可以查看时间限定的发布历史和版本对比信息。

6. 下线
   a. 在需要下线的时间限定右侧，单击“更多 > 下线”，系统弹出“提交下线”对话框。
   b. 在下拉菜单中选择审核人。
   c. 单击“确认提交”。

说明

下线及删除时间限定的前提是无依赖引用，即衍生指标引用。

7. 删除
   a. 勾选需要删除的时间限定，单击页面上方“删除”，系统弹出“删除”对话框。
   b. 单击“是”。

3.4.8 数据集市建设

3.4.8.1 新建汇总表

汇总逻辑表是由一个特定的分析对象（如会员）及其相关的统计指标组成的。组成一个汇总逻辑表的统计指标都具有相同的统计粒度（如会员），汇总逻辑表面向用户提供了以统计粒度（如会员）为主题的所有统计数据（如会员主题集市）。

汇总表分为“手工创建”和“自动汇聚”，此处仅描述手工创建场景。

说明

如果在“数据架构 > 配置中心 > 功能配置”页面中开启了“模型设计业务流程步骤 > 创建数据开发作业”（默认为关闭），发布汇总表时，系统将在数据开发中自动创建一个数据开发作业，作业名称以“*数据库名称_表编码*”开头。您可以进入“数据开发 > 作业开发”页面查看作业。该作业默认没有调度配置，需要您自行在数据开发模块中设置。

前提条件

在创建汇总表之前，请先确认您已完成维度、维度表、事实表和衍生指标/复合指标的新建、发布与审核。

新建汇总表并发布

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-263 选择数据架构

2. 在数据架构控制台，单击左侧导航树中的“维度建模”，然后选择“汇总表”页签。
3. 在左侧主题目录中选中一个主题，然后单击“新建”按钮，开始创建汇总表。
4. 在“新建汇总表”页面，完成如下配置：
   a. 设置“基本配置”参数：

图3-264 汇总表基本配置

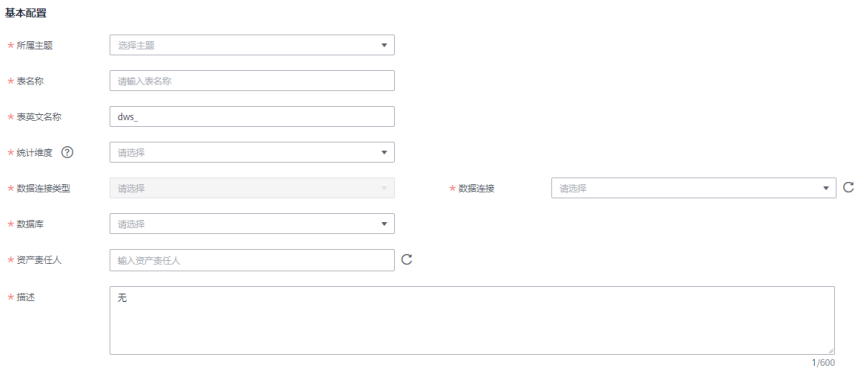


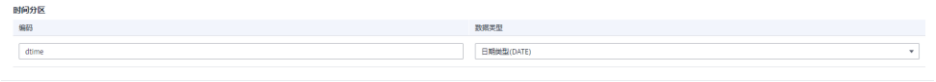
表3-154 基本配置参数说明

参数说明	说明
*所属主题	单击“选择主题”，选择表所属的主题域分组、主题域和业务对象。
*表名称	设置表名称。只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文和英文字母开头。

参数说明	说明
*表英文名称	设置表英文名称。只能包含英文字母、数字和下划线，且以英文字母开头。
*统计维度	选择统计维度。在下拉列表中只会显示来源于衍生指标的统计维度。如果下拉列表中没有统计维度，请参考 3.4.6.2.1 新建维度进行创建。 当汇总表完成新建后，“统计维度”指定的维度的所有维度属性，都将被自动添加到汇总表中，成为汇总表中的字段。您可以在完成汇总表的新建后，进入汇总表页面，单击汇总表名称查看详情中的表字段信息。
*数据连接类型	请选择和维度表、事实表相同的数据连接类型。
*数据连接	维度建模建议使用统一的数据连接。
*数据库	选择数据库。
队列	DLI 队列。该参数仅 DLI 连接类型有效。
Schema	DWS 或 POSTGRESQL 的模式。该参数在 DWS 或 POSTGRESQL 连接类型有效。
表类型	DWS 连接类型的表支持以下表类型： • DWS_ROW：行存表。行存储是指将表按行存储到硬盘分区上。 • DWS_COLUMN：列存表。列存储是指将表按列存储到硬盘分区上。 MRS_HIVE 仅支持 HIVE_TABLE 类型。
DISTRIBUTE BY	该参数仅 DWS 连接类型有效。DWS 表当前支持复制（Replication）和散列（Hash）两种分布策略。用户可选取多个字段。 • **REPLICATION 方式**：在每一个 DN 节点上存储一份全量表数据。这种存储方式的优点是每个 DN 上都有此表的全量数据，在 join 操作中可以避免数据重分布操作，从而减小网络开销；缺点是每个 DN 都保留了表的完整数据，造成数据的冗余。一般情况下只有较小的维度表才会定义为 Replication 表。 • **HASH 方式**：采用这种分布方式，需要为用户表指定一个分布列（distribute key）。当插入一条记录时，系统会根据分布列的值进行 hash 运算后，将数据存储在对应的 DN 中。对于 Hash 分布表，在读/写数据时可以利用各个节点的 IO 资源，大大提升表的读/写速度。一般情况下大表（1000000 条记录以上）定义为 Hash 表。
*资产责任人	在下拉框中选择资产责任人，可以手动输入名字或直接选择已有的责任人。
*描述	描述信息。支持的长度为 1~600 个字符。

b. 在“时间分区”区域，输入字段编码以及选择数据类型。当表发布成功后，在往表里写数据时，将根据该时间分区字段进行分区。

图3-265 时间周期配置



c. 在“指标配置”区域，单击“添加”，可以添加一个或多个与所指定的“统计维度”相关联的衍生指标或复合指标。

图3-266 指标配置

5. 单击“发布”，并在弹出框中单击“确认提交”，提交审核。
6. 请联系审核人员审核汇总表，等待审核通过。

   审核通过后，汇总表就会在数据库中自动创建。
7. 返回“维度建模 > 汇总表”页面，在列表中找到刚发布的汇总表，在“同步状态”一列中可以查看汇总表的同步状态。
   - 如果同步状态均显示成功，则说明汇总表发布成功，汇总表在数据库中已创建成功。
   - 如果同步状态中存在失败，可单击该汇总表所在行的“更多 > 发布历史”，然后在展开的页面中进入“发布日志”页签查看日志。请根据错误日志定位失败原因，问题解决后，您可以在汇总表页面勾选该汇总表，再单击列表上方的“更多 > 同步”尝试重新同步。如果问题仍未能解决，请联系技术支持人员协助处理。

管理汇总表

1. 在数据架构控制台，单击左侧导航树中的“维度建模”，选择“汇总表”页签，进入汇总表页面。

   图3-267 汇总表页面

2. 您可以根据实际需要选择如下操作。

当需要...	则...

当需要...	则...
新建	执行新建汇总表并发布。
编辑	执行 3。
发布	执行 4。
发布历史	执行 5。
预览 SQL	执行 6。
下线	执行 7。
关联质量规则	执行 8。
删除	执行 9。
导入	执行 10。
导出	执行 11。

3. 编辑
   a. 在需要编辑的汇总表右侧，单击“编辑”，进入编辑汇总表页面。
   b. 根据实际需要编辑相关内容。
   c. 单击“发布”。
4. 发布
   a. 在需要发布的汇总表右侧，单击“发布”，弹出“提交发布”对话框。
   b. 在下拉菜单中选择审核人。
   c. 单击“确认提交”。
5. 查看发布历史
   a. 在汇总列表中，找到所需要的汇总表，在右侧单击“更多 > 发布历史”，将显示“发布历史”页面。
   b. 如果“发布历史”中，您可以查看汇总表的发布历史记录、版本对比信息以及发布日志。

      如果“发布日志”中有错误日志，说明发布失败。您可以单击“重新同步”进行重试。
6. 预览 SQL
   a. 在汇总表列表中，找到所需要的汇总表，在右侧单击“更多 > 预览 SQL”，弹出“预览 SQL”对话框。
   b. 在“预览 SQL”中，您可以查看 SQL 语句，也可以复制 SQL。
7. 下线
   a. 在需要下线的汇总表右侧，单击“下线”，系统弹出“提交下线”对话框。
   b. 在下拉菜单中选择审核人。
   c. 单击“确认提交”。

说明

汇总表下线后，API 的如何处理由客户在数据服务中根据实际情况决定，数据架构侧不会对 API 做任何处理。

8. 关联质量规则
   a. 在汇总表列表中，勾选所需要的关联质量规则汇总表，在上方单击“关联质量规则”，弹出“关联质量规则”对话框。
   b. 在“关联质量规则”对话框中，您可以批量给汇总表的字段添加规则并关联到字段。
   c. 单击“确定”。
9. 删除
   a. 勾选需要删除的汇总表，单击上方“更多 > 删除”，系统弹出“删除”对话框。
   b. 单击“是”。
10. 导入

可通过导入的方式将汇总表批量快速的导入到系统中。

a. 在汇总表上方，单击“更多>导入”，进入“导入配置”页签。

图3-268 导入汇总表

b. 下载汇总表导入模板，编辑完成后保存至本地。
c. 选择是否更新已有数据。

说明

如果系统中已有的编码和模板中的编码相同，系统则认为是数据重复。

- 不更新：当数据重复时，不会替换系统中原有的数据。
- 更新：当数据重复时
  - 系统中的原有数据为草稿状态，则会覆盖生成新的草稿数据。
  - 系统中的原有数据为发布状态，则会生成下展数据。

d. 单击“添加文件”，选择编辑完成的导入模板。
e. 单击“上传文件”，上传完成后，自动跳转到“上次导入”页签，查看已导入的数据。

f. 单击“关闭”。

11. 导出

可通过导出的方式将汇总表导出到本地。

a. 在手工创建或自动汇聚列表选中待导出的汇总表。

b. 在列表上方，单击“更多>导出”，即可将系统中的汇总表导出到本地。

说明

- 在左侧主题树中选中某个主题，可以导出该主题下的所有汇总表；
- 当该空间下不超过 500 条汇总表数据时可以全部导出。

汇总表关联质量规则

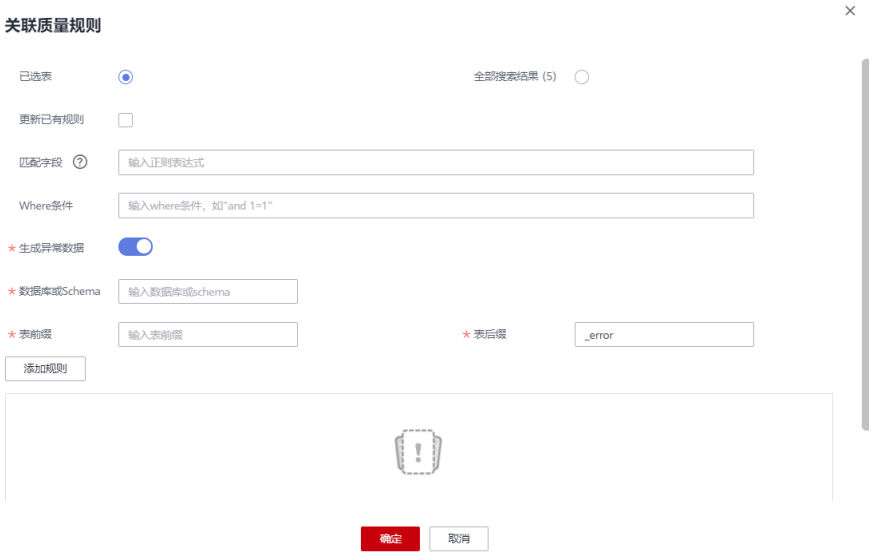
1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“汇总表”页签，进入汇总表页面。
3. 在汇总表列表中，勾选需要关联质量规则的汇总表。单击“关联质量规则”。

图3-269 关联汇总表质量规则

4. 在弹出的页面中配置关联质量规则参数。配置完成单击确定。
   - **更新已有规则**：若勾选此项，新添加的规则会覆盖旧规则。
   - **匹配字段**：此参数默认应用于所有字段，依据用户输入的正则表达式对字段进行过滤。
   - **Where 条件**：可依据用户输入的 where 条件对字段进行过滤。
   - **生成异常数据**：勾选此项，表示异常数据讲按照配置的参数存储到规定的库中。
   - **数据库或 Schema**：勾选“生成异常数据”时显示此项，表示存储异常数据的数据库或 Schema
   - **表前缀**：勾选“生成异常数据”时显示此项，表示存储异常数据的表的前缀。
   - **表后缀**：勾选“生成异常数据”时显示此项，表示存储异常数据的表的前缀。
   - **添加规则**：单击“添加规则”进行设置。例如，添加名称为“字段唯一值”规则，选中该规则后单击“确定”，在“告警条件”中输入告警条件表达式，然后按照此方法添加其他规则后，单击“确定”。告警表达式举例如下：

- 告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。在此处的“关联质量规则”对话框中，每个质量规则的“告警参数”以按钮形式列出。

图3-270 汇总表关联质量规则



汇总表字段关联数据标准

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“汇总表”页签，进入汇总表页面。
3. 在汇总表列表中，单击需要关联数据标准的汇总表名称。
4. 在汇总表的详情页的表字段列表中，查找需要关联数据标准的字段，单击其所属的 ，配置单个表字段关联数据标准。

图3-271 汇总表字段关联数据标准

5. 配置完成后，单击“确定”，完成汇总表字段关联数据标准。数据标准的来源请参考新建数据标准。

图3-272 配置数据标准

单个表字段关联质量规则

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“汇总表”页签，进入汇总表页面。
3. 在汇总表列表中，单击需要关联质量规则的汇总表名称。
4. 在汇总表的详情页的表字段列表中，单击 ，配置单个表字段关联质量规则。

图3-273 汇总表单个字段关联质量规则

5. 配置完成后，单击“确定”，完成汇总表字段关联质量规则。
   - **更新已有规则**：若勾选此项，新添加的规则会覆盖旧规则。
   - **添加规则**：单击“添加规则”进行设置。例如，添加名称为“字段唯一值”规则，选中该规则后单击“确定”，在“告警条件”中输入告警条件表达式，然后按照此方法添加其他规则后，单击“确定”。
   - 告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。在此处的“关联质量规则”对话框中，每个质量规则的“告警参数”以按钮形式列出。

图3-274 配置质量规则

表字段批量关联质量规则

1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“汇总表”页签，进入汇总表页面。
3. 在汇总表列表中，单击需要关联质量规则的汇总表名称。
4. 在汇总表的详情页的表字段列表中，勾选需要关联质量规则的表字段，单击关联质量规则。

图3-275 汇总表批量字段关联质量规则

5. 在弹出的界面中添加规则，完成规则参数配置。
   - **更新已有规则**：若勾选此项，新添加的规则会覆盖旧规则。
   - **添加规则**：单击“添加规则”进行设置。例如，添加名称为“字段唯一值”规则，选中该规则后单击“确定”，在“告警条件”中输入告警条件表达式，然后按照此方法添加其他规则后，单击“确定”。
   - 告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。在此处的“关联质量规则”对话框中，每个质量规则的“告警参数”以按钮形式列出。

图3-276 添加汇总表质量规则

6. 配置完成后，单击“确定”，完成汇总表字段批量关联质量规则。

3.4.9 通用操作

3.4.9.1 逆向数据库（关系建模）

通过逆向数据库，您可以将其他数据源的数据库中的表导入到指定的关系模型中。

前提条件

在逆向数据库之前，请先在 DataArts Studio 数据目录模块中对数据库进行元数据采集，以便同步数据目录时可以同步成功，否则同步数据目录将执行失败。有关数据目录元数据采集的具体操作，请参见 3.7.4.2 任务管理。

逆向数据库导入表到模型中

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“关系建模”进入关系建模页面。

步骤 2 在模型总览中找到所需要执行逆向数据库导入的模型，单击卡片进入模型，单击上方的“逆向数据库”。

图3-277 逆向数据库

步骤 3　在“逆向数据库”对话框中配置如下参数。

图3-278 配置逆向数据库参数

表3-155 逆向数据库

参数名称	说明
所属主题	单击“选择主题”按钮选择所属的主题信息。
数据连接类型	如果逆向到逻辑模型，请在下拉列表中选择所需要的连接类型。 如果逆向到物理模型，将显示当前模型的连接类型。

参数名称	说明
数据连接	选择所需要的数据连接。 如需从其他数据源逆向数据库到关系模型中，需要先在 DataArts Studio 管理中心创建一个数据连接，以便连接数据源。创建数据连接的操作，请参见 3.2.2 创建数据连接。
数据库	选择数据库。
队列	仅限 DLI 连接类型，需选择 DLI 队列。
Schema	仅限 DWS 连接类型，需设置 DWS 模式。
更新已有表	在导入时，如果所要导入的表在关系模型中已存在，是否更新已有的表。在导入时，系统将按表编码进行判断将要导入的表在当前的关系模型中是否已存在。在导入时，只有创建或更新操作，不会删除已有的表。 • **不更新**：如果表已存在，将直接跳过，不更新。 • **更新**：如果表已存在，更新已有的表信息。如果表处于“已发布”状态，表更新后，您需要重新发布表，才能使更新后的表生效。
数据表	选择“全部”时，将数据库中的所有的表都导入关系模型中。 选择“部分”时，请选择需要导入关系模型的表。

步骤 4 单击“确定”开始执行逆向数据库操作。

----结束

3.4.9.2 逆向数据库（维度建模）

通过逆向数据库，您可以将其他数据源的数据库中的表导入到指定的关系模型中。

前提条件

在逆向数据库之前，请先在 DataArts Studio 数据目录模块中对数据库进行元数据采集，以便同步数据目录时可以同步成功，否则同步数据目录将执行失败。有关数据目录元数据采集的具体操作，请参见 3.7.4.2 任务管理。

逆向数据库导入表到维度模型中

步骤 1 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-279 选择数据架构

步骤 2 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“维度建模”进入维度建模页面。

步骤 3 打开需要逆向数据库导入的维度或表的页签，然后单击列表上方的“逆向数据库”。

图3-280 选中对象

步骤 4 在“逆向数据库”对话框中配置参数。

表3-156 逆向数据库

参数名称	说明
所属主题	单击“选择主题”按钮选择所属的主题信息。
数据连接类型	选择维度建模的逆向数据库。
数据连接	选择所需要的数据连接。 如需从其他数据源逆向数据库到关系模型中，需要先在 DataArts Studio 管理中心创建一个数据连接，以便连接数据源。创建数据连接的操作，请参见 3.2.2 创建数据连接。

参数名称	说明
数据库	选择数据库。
队列	仅限 DLI 连接类型，需选择 DLI 队列。
Schema	DWS 或 POSTGRESQL 的模式。该参数在 DWS 或 POSTGRESQL 连接类型有效。
更新已有表	在导入时，只有创建或更新操作，不会删除已有的表。 • **不更新**：如果表已存在，将直接跳过，不更新。 • **更新**：如果表已存在，更新已有的表信息。如果表处于“已发布”状态，表更新后，您需要重新发布表，才能使更新后的表生效。
数据表	选择“全部”时，将数据库中的所有的表都导入。 选择“部分”时，请选择需要导入的表。

步骤 5 单击“确定”开始执行逆向数据库操作。等待操作执行完成，即可在“上次逆向”中查看结果或者执行重新逆向操作。

----结束

3.4.9.3 导入导出表

您可以通过导入表，可以将表批量导入到模型中。您也可以将已有的表导出，导出后的表可用于导入到其他模型中。

导入表到逻辑模型

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“关系建模”进入关系建模页面。

步骤 2 在模型总览中，找到所需要的逻辑模型，单击模型卡片进入，在主题目录中选中一个对象，然后单击“更多 -> 导入”。

步骤 3 在“导入表”对话框中，单击“下载关系建模导入模板”。

图3-281 导入表

表3-157 导入配置参数说明

参数名	说明
更新已有表	如果所要导入的表，在模型中已经存在，是否更新已有的表。系统将根据表编码判断将要导入的表在关系模型中是否已存在。在导入时，只有创建或更新操作，不会删除已有的表。支持以下选项： • **不更新**：如果表已存在，将直接跳过，不处理。 • **更新**：如果表已存在，更新已有的表信息。如果表处于“已发布”状态，表更新后，您需要重新发布表，才能使更新后的表生效。
上传模板	选择所需导入的文件。所需导入的文件，可以通过以下两种方式获得。 • **下载关系建模导入模板并填写模板** 在“导入配置”页签内，单击“下载关系建模导入模板”下载模板，然后根据业务需求填写好模板中的相关参数并保存。 • **导出的表文件** 您可以将某个 DataArts Studio 实例的数据架构中已创建的表导出到 Excel 文件中。导出后的文件可用于导入到关系模型中。导出模型的操作请参见导出表或 DDL。

步骤 4 打开下载的模板，请根据业务需求填写好模板中的相关参数并保存，模板中的“填写说明”Sheet 页供参考。

模板中的参数，其中名称前带“*”的参数为必填参数，名称前未带“*”的参数为可选参数。

在模板的“表模型”Sheet 页中，所需填写的参数，说明如下：

表3-158 表模型 Sheet 页参数说明

参数名	参数说明
所属主题	需填写已有的主题的编码路径，以/分隔。如果您未新建主题信息，请参见 3.4.4.2 主题设计进行新建。
*逻辑实体名称	表名称，只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文和英文字母开头。
*表名称	表英文名称，只能包含英文字母、数字、下划线、$、{、}，且不能以数字开头。
表别名	用户在配置中心打开了“表别名”时显示此项，名称别名。
表级标签	给表添加的标签，请输入已有的标签或新的标签名称。您也可以先前往 DataArts Studio 数据目录模块的“标签管理”页面添加标签，然后再回到此处设置相应的标签。添加标签的具体操作，请参见 3.7.1.4 标签管理。
*描述	表的描述信息。
资产责任人	需输入 DataArts Studio 实例当前工作空间中的用户名，可以手动输入名字或直接选择已有的责任人。
父表	只能填写为本模型中的其他表的表名称。
DWS 表 DISTRIBUTE BY	仅 DWS 连接支持，支持 HASH(属性名称)、REPLICATION2 种方式分布。
*属性名称（CHN）	表中的属性字段的中文名称。只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文或英文字母开头。
*属性名称（ENG）	表中的属性字段的英文名称。只能包含英文字母、数字和下划线，且以英文字母开头。
属性别名	用户在配置中心打开了“属性别名”时显示此项，属性别名。
顺序	属性字段在表中的顺序，从 1 开始。可以不填，不填时属性字段默认按模板中的顺序在表中排列。
属性描述	属性字段的描述信息。
*数据类型	逻辑模型的数据类型，请参见字段类型管理中的 DEFAULT 类型分组。
数据长度	数据的长度。对于不定长的数据类型，如果所指定的数据连接类型支持对其指定数据长度，请指定数据长度。 例如，DWS 连接类型，如果字段类型为 CHAR(10)，需要在“数据类型”中填写“CHAR”，在“数据长度”中填写“10”。
是否分区	填写“Y”表示该字段为分区字段，填写“N”表示不是分区字段。

参数名	参数说明
是否主键	填写“Y”表示该字段为主键，填写“N”表示不是主键。
不为空	填写“Y”表示该字段不为空，填写“N”表示字段允许为空。
引用的数据标准编码	填写需要引用的数据标准的编码，也可以不填。如果未创建数据标准，请参见 3.4.5.2 新建数据标准进行创建。
属性标签	为属性字段添加的标签，请输入已有的标签或新的标签名称。您也可以先前往 DataArts Studio 数据目录模块的“标签管理”页面添加标签，然后再回到此处设置相应的标签。添加标签的具体操作，请参见 3.7.1.4 标签管理。

在“关系”Sheet 页中，所需填写的参数，请参考表 3-130 中的说明。

暂不支持导入映射，“映射”Sheet 页无需填写。

步骤 5 导入结果会在导入对话框的“上次导入”中显示。如果导入成功，单击“关闭”完成导入。如果导入失败，您可以查看失败原因，将模板文件修改正确后，再重新上传。

图3-282 上次导入

----结束

导入表到物理模型

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“关系建模”进入关系建模页面。

步骤 2 在关系模型树中，找到所需要的物理模型，单击进入，展开模型，选中一个对象，然后单击“导入”。

步骤 3 在“导入表”对话框中，单击“下载关系建模导入模板”。

图3-283 导入表

表3-159 导入配置参数说明

参数名	说明
更新已有表	如果所要导入的表，在模型中已经存在，是否更新已有的表。系统将根据表编码判断将要导入的表在关系模型中是否已存在。在导入时，只有创建或更新操作，不会删除已有的表。支持以下选项： • **不更新**：如果表已存在，将直接跳过，不处理。 • **更新**：如果表已存在，更新已有的表信息。如果表处于“已发布”状态，表更新后，您需要重新发布表，才能使更新后的表生效。
上传模板	选择所需导入的文件。所需导入的文件，可以通过以下两种方式获得。 • **下载关系建模导入模板并填写模板** 在“导入配置”页签内，单击“下载关系建模导入模板”下载模板，然后根据业务需求填写好模板中的相关参数并保存。 • **导出的表文件** 您可以将某个 DataArts Studio 实例的数据架构中已创建的表导出到 Excel 文件中。导出后的文件可用于导入到关系模型中。导出模型的操作请参见导出表或 DDL。

步骤 4 打开下载的模板，请根据业务需求填写好模板中的相关参数并保存，模板中的“填写说明”Sheet 页供参考。

模板中的参数，其中名称前带“*”的参数为必填参数，名称前未带“*”的参数为可选参数。

在模板的“表模型”Sheet 页中，所需填写的参数，说明如下：

表3-160 表模型 Sheet 页参数说明

参数名	参数说明（导入 DLI/POSTGRESQL/DWS/MRS_HIVE 类型的表）
所属主题	需填写已有的主题的编码路径，以/分隔。如果您未新建主题信息，请参见 3.4.4.2 主题设计进行新建。
*逻辑实体名称	表名称，只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文和英文字母开头。
*表名称	表英文名称，只能包含英文字母、数字、下划线、$、{、}，且不能以数字开头。
表别名	用户在配置中心打开了“表别名”时显示此项，名称别名。
表级标签	给表添加的标签，请输入已有的标签或新的标签名称。您也可以先前往 DataArts Studio 数据目录模块的“标签管理”页面添加标签，然后再回到此处设置相应的标签。添加标签的具体操作，请参见 3.7.1.4 标签管理。
*描述	表的描述信息。
资产责任人	需输入 DataArts Studio 实例当前工作空间中的用户名。只有工作空间管理员或开发者、运维者角色的用户才可以设置为责任人。
数据连接类型	支持以下连接类型：DLI、POSTGRESQL、DWS、MRS_HIVE。
*表类型	DLI 模型的表支持以下表类型： • Managed：数据存储位置为 DLI 的表。 • External：数据存储位置为 OBS 的表。当“表类型”设置为 External 时，需设置“OBS 路径””参数。 • DLI_VIEW：该类型只支持导入，不支持在控制台页面创建。 DWS 模型的表支持以下表类型： • DWS_ROW：行类型。 • DWS_COLUMN：列类型。 • DWS_VIEW：视图类型。 MRS_HIVE 模型的表不支持该参数。
OBS 路径	DLI 模型的表类型为 DLI_EXTERNAL 时，需填写与表相关联的存放源数据的 OBS 路径。OBS 路径格式如：bucket_name/filepath。
数据格式	该参数仅 DLI 模型的表有效。 表类型为 DLI_MANAGED 的表支持的数据格式有：Parquet、Carbon。

参数名	参数说明（导入 DLI/POSTGRESQL/DWS/MRS_HIVE 类型的表）
	表类型为 DLI_EXTERNAL 的表支持的数据格式有：Parquet、Carbon、CSV、ORC、JSON、Avro。
表所属的数据连接	输入已创建的数据连接名称。
表所属的数据库	输入已创建的数据库名称。
数据连接扩展信息	连接类型为 DLI 时，输入 DLI 队列名称。连接类型为 DWS 或 POSTGRESQL 时，输入 Schema 名称。
*属性名称（CHN）	表中的属性字段的中文名称。只能包含中文、英文字母、数字、左右括号、中划线和下划线，且以中文或英文字母开头。
*属性名称（ENG）	表中的属性字段的英文名称。只能包含英文字母、数字和下划线，且以英文字母开头。
顺序	属性字段在表中的顺序，从 1 开始。可以不填，不填时属性字段默认按模板中的顺序在表中排列。
属性描述	属性字段的描述信息。
*数据类型	不同的数据连接类型支持的数据类型不一样，请参见字段类型管理。
数据长度	对于不定长的数据类型，如果所指定的数据连接类型支持对其指定数据长度，请指定数据长度。 例如，DWS 连接类型，如果字段类型为 CHAR(10)，需要在“数据类型”中填写“CHAR”，在“数据长度”中填写“10”。
是否分区	填写“Y”表示该字段为分区字段，填写“N”表示不是分区字段。
是否主键	填写“Y”表示该字段为主键，填写“N”表示不是主键。
不为空	填写“Y”表示该字段不为空，填写“N”表示字段允许为空。
引用的数据标准编码	填写需要引用的数据标准的编码，也可以不填。如果未创建数据标准，请参见 3.4.5.2 新建数据标准进行创建。
属性标签	为属性字段添加的标签，请输入已有的标签或新的标签名称。您也可以先前往 DataArts Studio 数据目录模块的“标签管理”页面添加标签，然后再回到此处设置相应的标签。添加标签的具体操作，请参见 3.7.1.4 标签管理。
其他配置	为 JSON 格式，用于存放表额外配置信息。格式如下： { "option_name1": "value",

参数名	参数说明（导入 DLI/POSTGRESQL/DWS/MRS_HIVE 类型的表）

	"option_name2": "value"
	……
	}
	例如：
	{
	"a1": "100",
	"a2": "30"
	}

版本号	可选参数。

在“关系”Sheet 页中，所需填写的参数，请参考表 3-136 中的说明。

暂不支持导入映射，“映射”Sheet 页无需填写。

步骤 5 导入结果会在导入对话框的“上次导入”页面中显示。如果导入成功，单击“关闭”完成导入。如果导入失败，您可以查看失败原因，将模板文件修改正确后，再重新上传。

----结束

导出表或 DDL

步骤 1 在 DataArts Studio 数据架构主界面，单击左侧导航栏的“关系建模”进入关系建模页面。

步骤 2 在模型总览中，找到所需要的逻辑模型，单击模型卡片进入，在主题目录中选择对象，然后单击“更多 -> 导出”。

图3-284 导出表或 DDL

步骤 3 在弹出对话框中，选择需要导出的对象。

导出的 Excel 表可以用于导入操作。

图3-285 导出表

导出 DDL 时，会将所选表的 DDL 语句导出成 txt 文件。

图3-286 导出 DDL

步骤 4 单击“确定”。

----结束

3.4.9.4 关联质量规则

当您完成表的新建和发布后，您可以在表中关联质量规则。在“配置中心 > 功能配置”页面中的“模型设计业务流程步骤 > 创建质量作业”勾选的情况下，完成质量规则的关联后，表发布后就会在 DataArts Studio 数据质量中自动创建质量作业，如果当前表已经发布，则系统会自动更新质量作业。

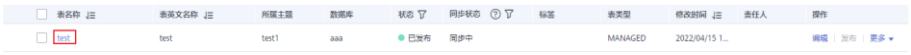
关联质量规则并查看质量作业

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“关系建模”进入关系建模页面。

步骤 2 在页面选择所需要的模型单击进入，在右侧的列表中将显示该模型下面所有的表。您也可以展开主题结构，选中一个对象，右侧的列表中将显示该对象下所有的表。

步骤 3 在列表中，找到所需要的表，单击表名称进入表详情页面。

图3-287 关系模型列表



步骤 4 在详情页的表字段区域，选中需要关联质量规则的的字段，然后单击“关联质量规则”按钮。

图3-288 关联质量规则

异常数据输出配置：勾选此项，并勾选生成异常数据，表示异常数据将按照配置的参数存储到规定的库中。

步骤 5 在弹出的“关联质量规则”对话框中，单击“添加规则”。

图3-289 添加质量规则页

此时，系统将弹出“添加规则”对话框，在规则列表中将显示 **DataArts Studio 数据质量**中默认的质量规则，选中所需要的规则，然后单击“确定”。如果列表中的规则不满足业务需求，您也可以创建自定义规则，单击“新建规则”可以跳转到 **DataArts Studio 数据质量**页面，请参考 3.6.2.2 新建规则模板新建规则。

图3-290 添加规则

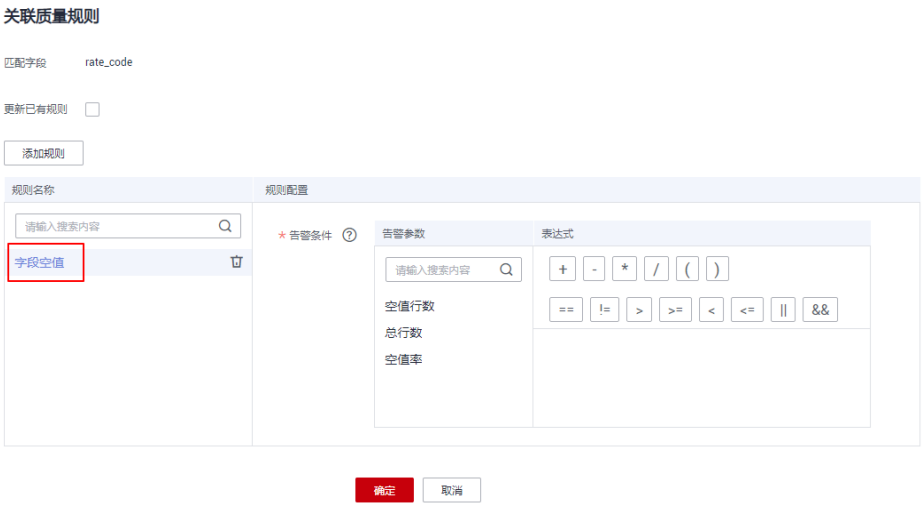
添加规则完成后，将返回“关联质量规则”对话框，在“规则名称”列表中，选中一条规则，然后设置告警条件，设置完所有规则的告警条件后单击“确定”。

- 在“告警条件”输入框中，请输入告警条件表达式，在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。

- 告警条件表达式由告警参数和逻辑运算符组成。

  每个规则的告警参数会在“告警参数”中以按钮形式列出。单击这些按钮，在“告警条件”中将按告警参数的排列顺序显示为${1}、${2}、${3}等变量名称，以此类推，变量名即代表告警参数。也就是说，在设置“告警条件”时，使用变量${1}代表第一个告警参数，${2}代表第二个告警参数，以此类推。

图3-291 设置告警条件

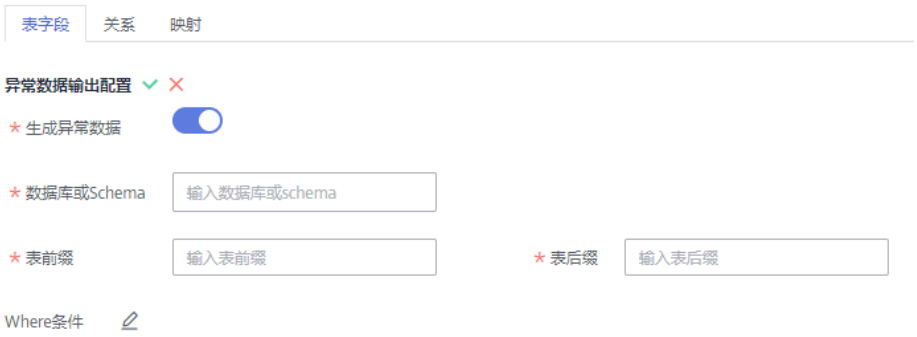


步骤 6 （可选）如需要将质量作业中不符合设定规则的异常数据存储在异常表中，可以打开“异常数据输出配置”开关。

图3-292 异常数据输出开关

点击开关，并打开“生成异常数据”按钮，表示异常数据将按照配置的参数存储到规定的库中。

图3-293 异常数据输出配置



各参数具体含义如下：

- 数据库或 Schema：表示存储异常数据的数据库或 Schema。
- 表前缀：表示存储异常数据的表的前缀。
- 表后缀：表示存储异常数据的表的后缀。

配置完成后点击 保存配置。

步骤 7 （可选）质量规则的检查范围默认是全表，如需要精确定位分区查询数据，请填写 where 条件。

图3-294 where 条件开关

步骤 8 查看关联质量规则的结果，如果显示成功，单击“确定”。如果显示失败，请查看失败原因，等问题处理后，再重新关联质量规则。

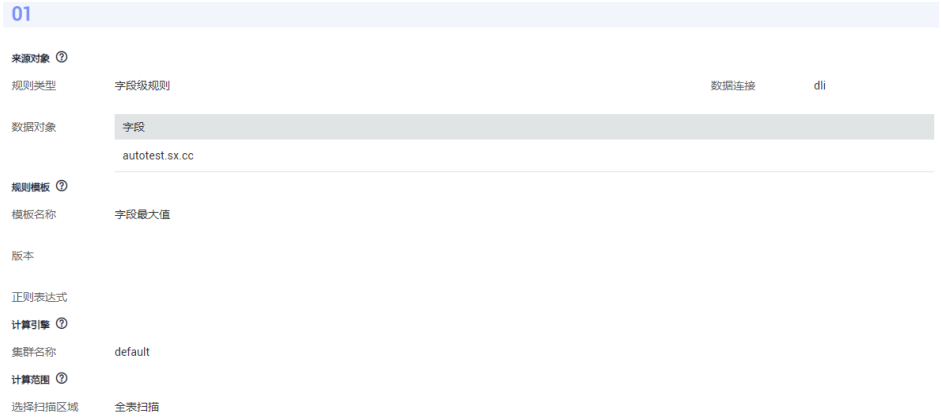
图3-295 关联结果

步骤 9 返回关系模型列表页面，找到已关联质量规则的表，在“同步状态”列中，鼠标移至创建质量作业的图标 上，单击“查看”进入质量作业页面查看已添加的质量规则。

图3-296 质量作业同步状态

步骤 10 进入质量作业的“规则配置”页面，可以查看刚才添加的质量规则。

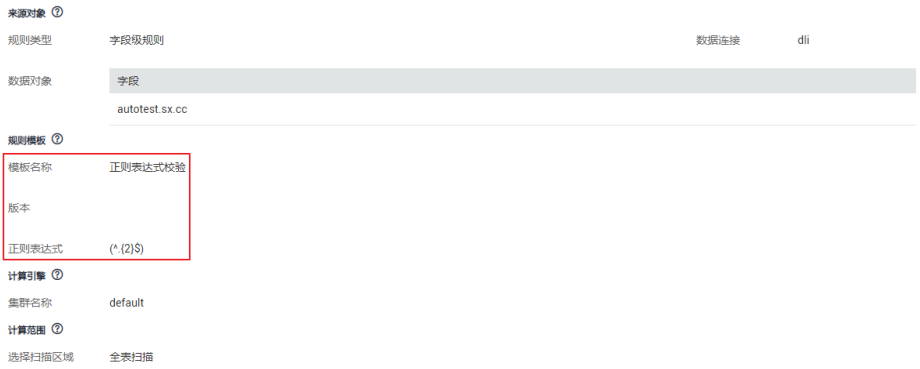
图3-297 质量规则



此外，在建表时已关联的数据标准，在表发布后也会在上图中生成相应的质量规则，您可以在质量作业中进行查看。

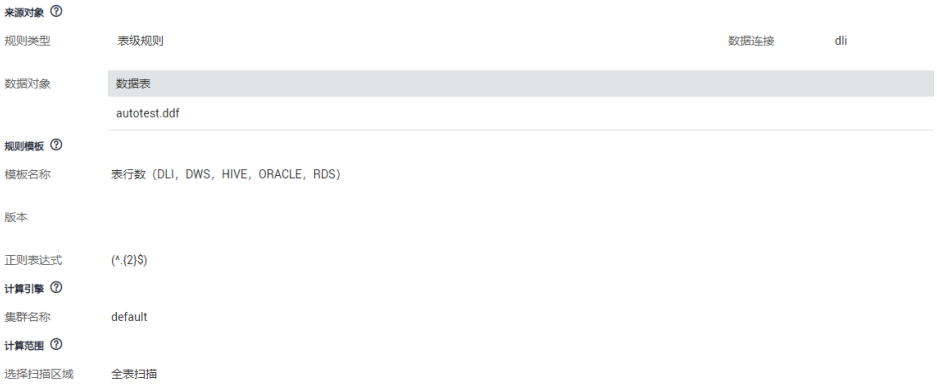
字段关联的数据标准生成的质量规则，示例如下：

图3-298 字段关联的质量规则



字段关联了数据标准，数据标准关联的码表生成的质量规则，示例如下：

图3-299 码表的质量规则



----结束

3.4.9.5 查看表

对于关系模型中的表，您可以查看模型视图、表详情、关系图、预览 SQL 以及发布历史。

查看模型视图

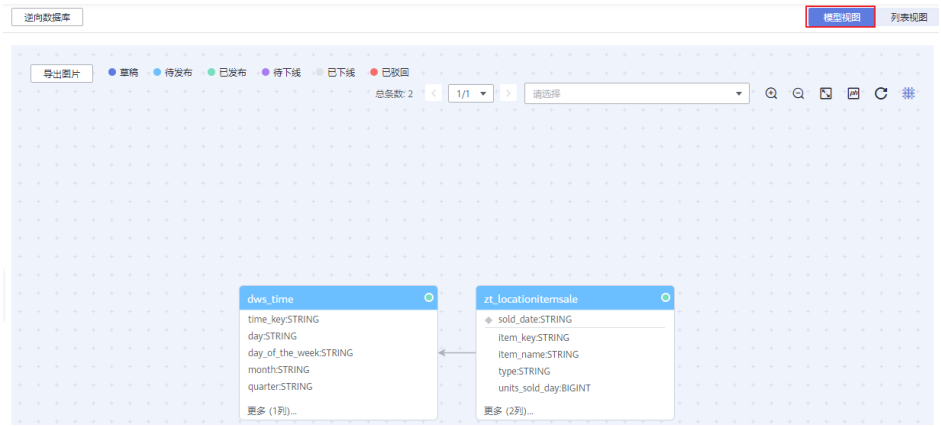
当您在关系模型中完成表的新建后，就可以通过列表视图和模型视图两种形式查看表模型。关系模型页面默认显示为列表视图，您可以切换为模型视图进行查看。

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“关系建模”进入关系建模页面。

步骤 2 在关系模型树中选择所需要模型，展开模型树，选择一个对象。

步骤 3 关系模型页面默认显示为列表视图，单击列表右上方的“模型视图”按钮，切换为模型视图，如下图所示。单击“列表视图”则可以重新切换回列表视图。

图3-300 模型视图



在模型视图中支持以下功能：

- 双击表名，可显示表的详情信息。
- 单击左上角的“导出图片”按钮，可以将模型视图导出成图片。
- 在右上角的搜索框中输入表名，可以快速找到的所要查看的表。
-  功能依次为放大、缩小、全屏、物理模型/逻辑模型切换、刷新、显示画布。

----结束

查看表详情以及预览 SQL

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“关系建模”进入关系建模页面。

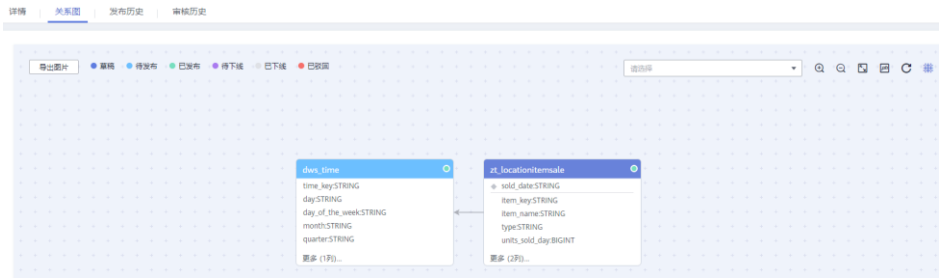
步骤 2 在模型总览中，找到所需要的逻辑模型，单击模型卡片进入，在主题目录中选中一个主题，右侧的列表中将显示该主题下所有的表。

步骤 3 在表的列表中，找到需要查看详情以及预览 SQL 的表，在表所在行，单击“更多 > 预览 SQL”可以预览 SQL 或复制 SQL。完成预览后单击“确定”返回关系模型的列表页面。

图3-301 关系模型列表 2

步骤 4 在表的列表中，单击表名称进入表详情页面，可以查看表的详情、关系图、发布历史和审核历史。

图3-302 关系图



----结束

查看发布历史

表发布后，您可以查看表的发布历史、版本对比和发布日志。如果表发布失败，或者数据目录、数据质量同步失败，您可以通过查看发布日志定位问题、重新同步。

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏的“关系建模”进入关系建模页面。

步骤 2 在模型总览中，找到所需要的逻辑模型，单击模型卡片进入，在主题目录中选中一个主题，右侧的列表中将显示该主题下所有的表。

步骤 3 在列表中，找到所需要的表，然后在表所在行，单击“更多 > 发布历史”，查看表的发布历史、版本对比和发布日志。

图3-303 发布历史

----结束

3.4.9.6 批量修改主题/目录/流程

批量修改主题

当前仅支持信息架构、关系建模、维度、事实表、汇总表、技术指标模块进行批量修改主题操作，操作流程相同。

此处以批量修改信息架构为例，展示如下：

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏中的“信息架构”。

步骤 2 进入后，在页面选择所需要批量修改主题的项，单击“更多 > 修改主题”，可以将选中的项更改到其它主题。配置完成点击“确定”。

图3-304 批量修改主题

----结束

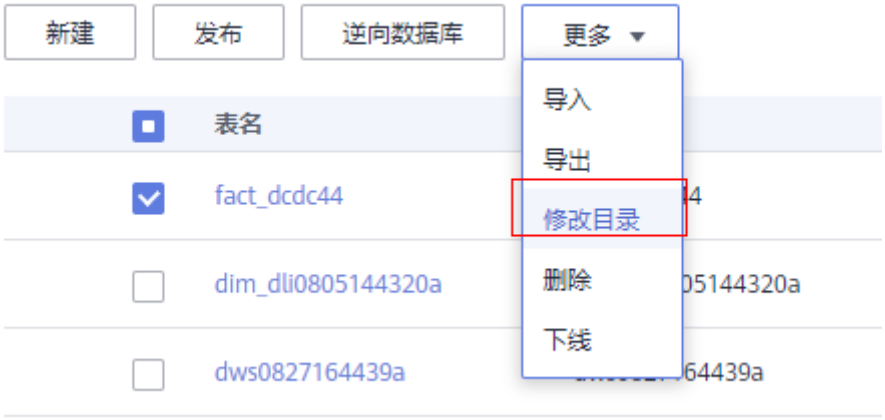
批量修改目录

当前仅支持码表管理、数据标准进行批量修改目录操作。

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏中的码表管理或数据标准。

步骤 2 进入后，在页面选择所需要批量修改目录的项，单击“更多 > 修改目录”，可以将选中的项更改到其它目录。

图3-305 批量修改目录（此处以码表管理模块为例）



----结束

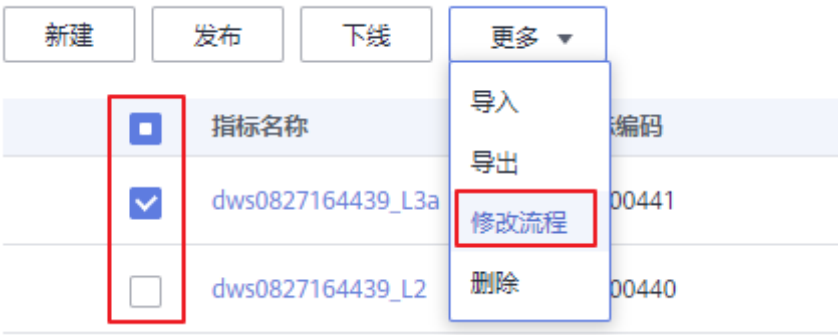
批量修改流程

当前仅支持业务指标进行批量修改流程操作。

步骤 1 在 DataArts Studio 数据架构控制台，单击左侧导航栏中的业务指标。

步骤 2 进入业务指标页面后，在页面选择所需要批量修改流程的指标，单击“更多 > 修改流程”，可以将选中的项更改到其它流程。

图3-306 批量修改流程



----结束

3.4.9.7 审核中心

开发环境生成的规范建模、数据处理类任务提交后，都会存储在审核中心页面，然后在审核中心页面进行任务发布，这些任务才会在生产环境上线。

审核人员审核对象

如果您是审核人员，请使用审核人员的帐号参考以下步骤审核对象。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-307 选择数据架构

2. 在左侧导航树中，单击“审核中心”，选择“待我审核”页签，在列表中找到需要审核的对象，然后在该对象所在行单击“审核”。

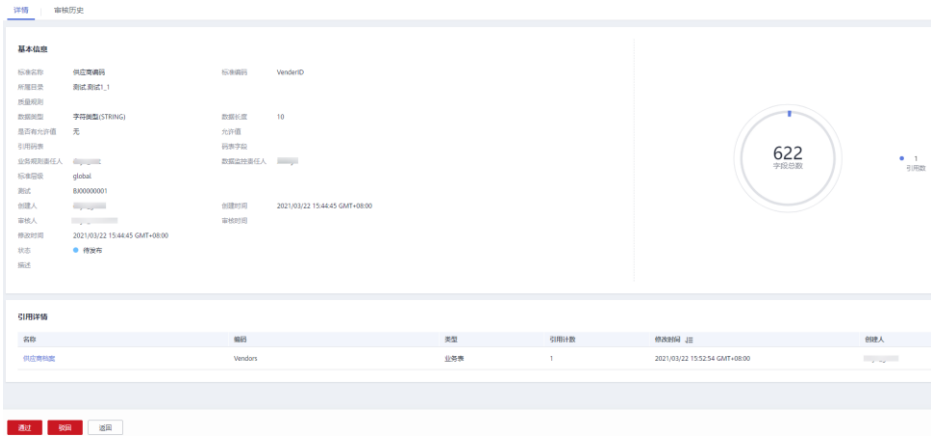
   您也可以勾选多个待审核的对象，然后单击“批量审核”按钮进行批量审核。

图3-308 审核

3. 在审核的详情页面，确认信息无误后，单击“通过”，然后在弹出对话框中输入审核意见并单击“确定”完成审核。

如果信息有误，请单击“驳回”，然后在弹出对话框中输入审核意见并单击“确定”完成审核。

图3-309 审核信息



查看已审核、待审核、我的申请

- **待我审核**

  在 DataArts Studio 数据架构的左侧导航树中，单击“审核中心”，选择“待我审核”页签，可以查看待审批的对象。

- **已审核**

  在 DataArts Studio 数据架构的左侧导航树中，单击“审核中心”，选择“已审核”页签，可以查看已通过审批的对象。

- **我的申请**

  在 DataArts Studio 数据架构的左侧导航树中，单击“审核中心”，选择“我的申请”页签，可以查看自己提交审批的对象。

待我审核

步骤 1 在 DataArts Studio 数据架构控制台的左侧导航栏中，单击“审核中心”，进入审核中心页面，系统默认显示待审核页面，如下图所示。

图3-310 待审核页面

功能区域	说明
1	批量审核： 1. 勾选多个待审核信息。

功能区域	说明
	2. 单击 批量审核 ，弹出“批量审核”对话框。 3. 输入有效的审核意见。 4. 单击“批量通过”，所选审核信息通过审核；单击“批量驳回”，所选审核信息被驳回。
2	单个审核： 1. 单击操作列“审核”，进入指定待审核信息的审核页面。 2. 据实际情况勾选审核结果、输入有效的审核的意见。 3. 单击“确定”，完成审核。
3	• 提交时间 开始日期 – 结束日期 ：通过该按钮过滤出修改时间段内的待审核信息。 • 对象 ：通过该按钮查询对象和创建人的待审核信息。 • ：通过该按钮设置待审核表的表列项。 • ：刷新按钮。

----结束

我的申请

步骤 1 在数据架构控制台，单击“审核中心”，进入审核中心页面。

步骤 2 单击“我的申请”，进入我的申请页面，如下图所示。

图3-311 我的申请页面

您可以进行如下操作：

- 通过操作列“查看”，查看指定行信息。
- 通过操作列“撤回”，撤回申请。

----结束

3.4.10 使用教程

3.4.10.1 数据架构示例

DataArts Studio 数据架构以关系建模、维度建模理论支撑，实现规范化、可视化、标准化数据模型开发，定位于数据治理流程设计落地阶段，输出成果用于指导开发人员实践落地数据治理方法论。

本章节操作场景如下：

- 对 MRS Hive 数据湖中的出租车出行数据进行数据模型设计。
- 数据库 demo_sdi_db 中已具备出租车出行原始数据表 sdi_taxi_trip_data。
- 原始数据表 sdi_taxi_trip_data 的数据字段介绍如下：

  数据说明如下：

表3-161 出租车行程数据

序号	字段名称	字段描述
1	VendorID	供应商编号 取值如下： 1=A Company 2=B Company
2	tpep_pickup_datetime	上车时间
3	tpep_dropoff_datetime	下车时间
4	passenger_count	乘客人数
5	trip_distance	行驶距离
6	ratecodeid	费率代码 取值如下： 1=Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
7	store_fwd_flag	存储转发标识
8	PULocationID	上车地点
9	DOLocationID	下车地点
10	payment_type	付款方式代码 取值如下： 1=Credit card

序号	字段名称	字段描述
		2=Cash 3=No charge 4=Dispute 5=Unknown 6=Voided trip
11	fare_amount	车费
12	extra	加收
13	mta_tax	MTA 税
14	tip_amount	手续费
15	tolls_amount	通行费
16	improvement_surcharge	改善附加费
17	total_amount	总车费

数据架构的流程如下：

1. **准备工作**：
   - **添加审核人**：在数据架构中，业务流程中的步骤都需要经过审批，因此，需要先添加审核人。只有工作空间管理员角色的用户才具有添加审核人的权限。
   - **管理配置中心**：数据架构中提供了丰富的自定义选项，统一通过配置中心提供，您需要根据自己的业务需要进行自定义配置。
2. **数据调研**：基于现有业务数据、行业现状进行数据调查、需求梳理、业务调研，输出企业业务流程以及数据主题划分。
   - **主题设计**：通过分层架构表达对数据的分类和定义，帮助厘清数据资产，明确业务领域和业务对象的关联关系。
   - **流程设计**：本例暂不涉及。流程设计是针对流程的一个结构化的整体框架，描述了企业流程的分类、层级以及边界、范围、输入/输出关系等，反映了企业的商业模式及业务特点。
3. **标准设计**：新建码表&数据标准。
   - **新建码表并发布**：通常只包括一系列允许的值和附加文本描述，与数据标准关联用于生成值域校验质量监控。
   - **新建数据标准并发布**：用于描述公司层面需共同遵守的属性层数据含义和业务规则。其描述了公司层面对某个数据的共同理解，这些理解一旦确定下来，就应作为企业层面的标准在企业内被共同遵守。
4. **模型设计**：应用关系建模和维度建模的方法，进行分层建模。
   - **关系建模：新建 SDI 层和 DWI 层两个模型**。
     - **SDI**：Source Data Integration，又称贴源数据层。SDI 是源系统数据的简单落地。

    - **DWI**：Data Warehouse Integration，又称数据整合层。DWI 整合多个源系统数据，对源系统进来的数据进行整合、清洗，并基于三范式进行关系建模。
  - **维度建模：在 DWR 层新建并发布维度&维度建模：在 DWR 层新建并发布事实表。**
    - **DWR**：Data Warehouse Report，又称数据报告层。DWR 基于多维模型，和 DWI 层数据粒度保持一致。
    - **维度**：维度是用于观察和分析业务数据的视角，支撑对数据进行汇聚、钻取、切片分析，用于 SQL 中的 GROUP BY 条件。
    - **事实表**：归属于某个业务过程的事实逻辑表，可以丰富具体业务过程所对应事务的详细信息。
5. **指标设计：新建并发布技术指标**：新建业务指标（本例不涉及）和技术指标，技术指标又分为原子指标、衍生指标和复合指标。
   - **指标**：指标一般由指标名称和指标数值两部分组成，指标名称及其涵义体现了指标质的规定性和量的规定性两个方面的特点，指标数值反映了指标在具体时间、地点、条件下的数量表现。

     业务指标用于指导技术指标，而技术指标是对业务指标的具体实现。
   - **原子指标**：原子指标中的度量和属性来源于多维模型中的维度表和事实表，与多维模型所属的业务对象保持一致，与多维模型中的最细数据粒度保持一致。

     原子指标中仅含有唯一度量，所含其它所有与该度量、该业务对象相关的属性，旨在用于支撑指标的敏捷自助消费。
   - **衍生指标**：是原子指标通过添加限定、维度卷积而成，限定、维度均来源于原子指标关联表的属性。
   - **复合指标**：由一个或多个衍生指标叠加计算而成，其中的维度、限定均继承于衍生指标。

     注意，不能脱离衍生指标、维度和限定的范围，去产生新的维度和限定。
6. **维度建模：在 DM 层新建并发布汇总表**。
   - **DM (Data Mart)**：又称数据集市。DM 面向展现层，数据有多级汇总。
   - **汇总表**：汇总表是由一个特定的分析对象（如会员）及其相关的统计指标组成的。组成一个汇总逻辑表的统计指标都具有相同的统计粒度（如会员），汇总逻辑表面向用户提供了以统计粒度（如会员）为主题的所有统计数据（如会员主题集市）。

添加审核人

在数据架构中，数据建模流程中的步骤都需要经过审批，因此，需要先添加审核人。DAYU Administrator 角色或该工作空间管理员，具备对应的添加审核人的权限。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据架构”模块，进入数据架构页面。

图3-312 选择数据架构

2. 单击左侧导航树中的“配置中心”，进入相应页面后，在“审核人管理”页签，单击“添加”按钮。
3. 选择审核人（工作空间管理员或开发者），输入正确的电子邮箱和手机号，单击“确定”完成审核人添加。

   您也可以添加自己当前帐号为审核人，在后续提交审批的相关操作中，支持进行“自助审批”。根据需要，可以添加多个审核人。

图3-313 添加审核人

管理配置中心

数据架构中提供了丰富的自定义选项，统一通过配置中心提供，您可有根据自己的业务需要进行自定义配置。

1. 在数据架构控制台，单击左侧菜单栏的“配置中心”，进入配置中心页面。
2. 进入“功能配置”页签，按照您的需求，进行自定义设置。

图3-314 功能配置

3. 单击“确定”完成配置。

主题设计

在本示例中，主题设计如表 3-162 所示，说明如下：

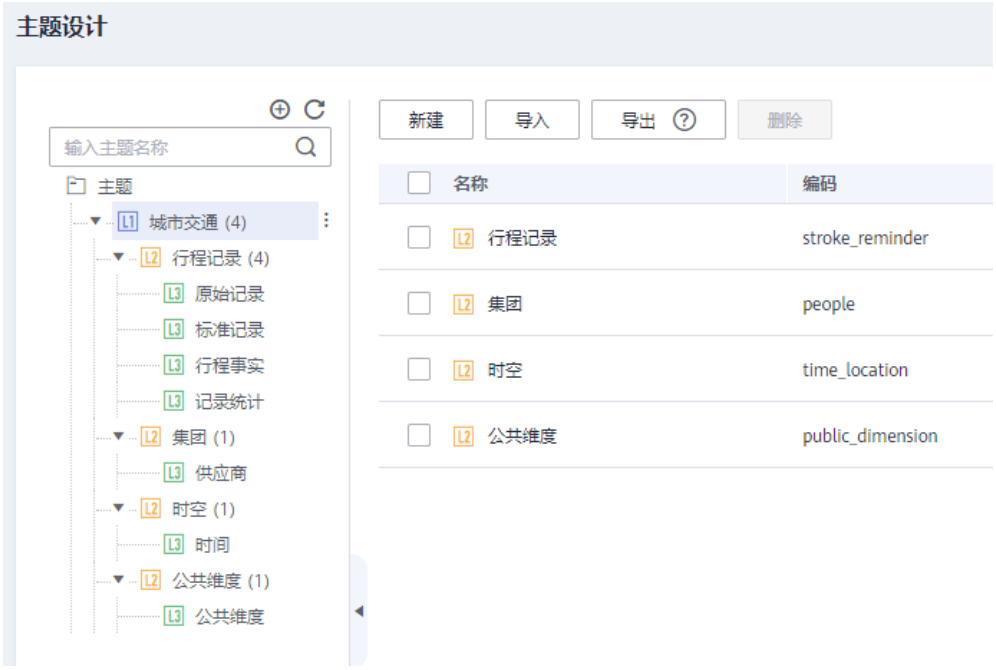
- 新建 1 个主题域分组：城市交通。
- 在主题域分组“城市交通”下，新建 4 个主题域：行程记录、集团、时空、公共维度。
- 在主题域“行程记录”下，新建 4 个业务对象：原始记录、标准记录、行程事实、记录统计。
- 在主题域“集团”下，新建 1 个业务对象：供应商。
- 在主题域“时空”下，新建 1 个业务对象：时间。
- 在主题域“公共维度”下，新建 1 个业务对象：公共维度。

表3-162 主题设计信息

主题域分组名称（L1）	主题域分组编码（L1）	主题域名称（L2）	主题域编码（L2）	业务对象名称（L3）	业务对象编码（L3）
城市交通	city_traffic	行程记录	stroke_reminder	原始记录	origin_stroke
				标准记录	stand_stroke
				行程事实	stroke_fact

主题域分组名称（L1）	主题域分组编码（L1）	主题域名称（L2）	主题域编码（L2）	业务对象名称（L3）	业务对象编码（L3）
				记录统计	stroke_statistic
		集团	people	供应商	vendor
		时空	time_location	时间	date
		公共维度	public_dimension	公共维度	public_dimension

图3-315 主题设计



操作步骤如下：

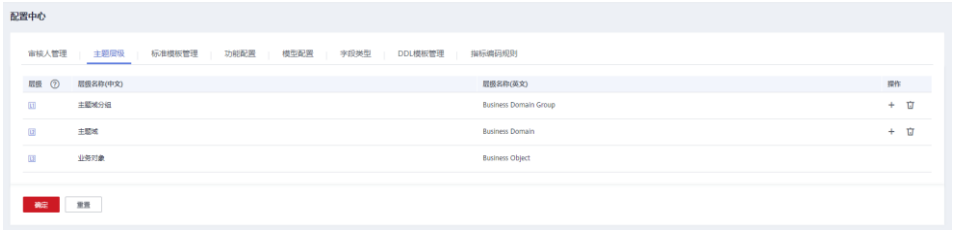
步骤 1 登录 DataArts Studio 控制台。找到已创建的 DataArts Studio 实例，单击实例卡片上的“进入控制台”。

步骤 2 在工作空间概览列表中，找到所需要的工作空间，单击“数据架构”，进入数据架构控制台。

步骤 3 在数据架构控制台，单击左侧菜单栏的“配置中心”。选择“ 主题层级”，使用默认的 3 层层级。

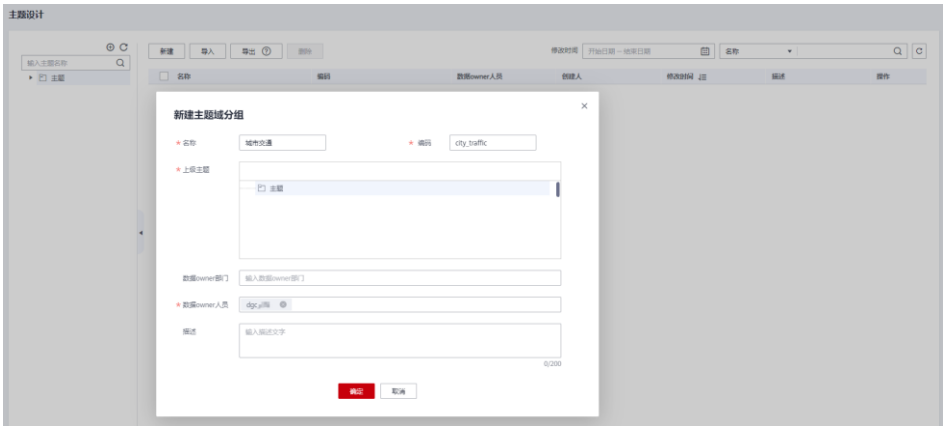
L1-L7 表示主题层级，默认 3 层，最大 7 层，最少 2 层，最后一层是业务对象，其他层级名称可编辑修改。配置中心配置的层级数，将在“主题设计”模块生效。

图3-316 配置主题层级



步骤 4 在数据架构控制台，单击左侧菜单栏的“主题设计”，进入相应页面后，单击“新建”创建 L1 层主题，即主题域分组。

图3-317 新建 L1 层主题



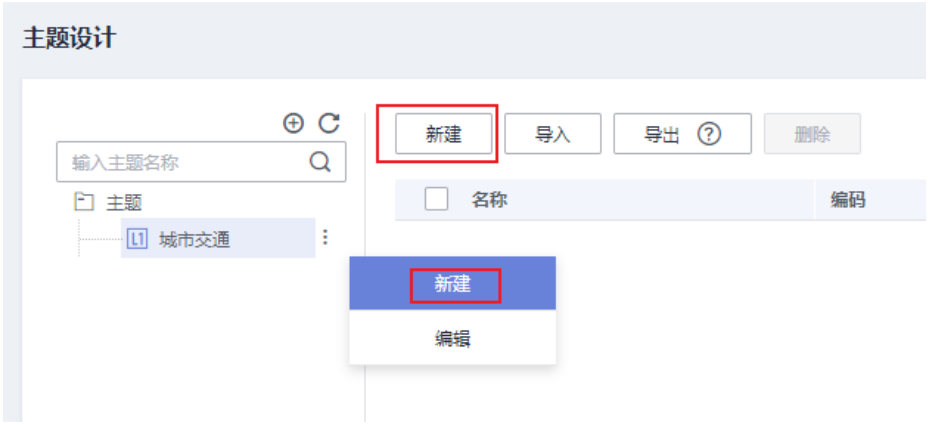
在弹出窗口中，按图 3-317 所示填写参数，然后单击“确定”完成主题域分组的创建。

步骤 5 在 L1 层主题“城市交通”下，依次新建 4 个 L2 层主题，即主题域：行程记录、集团、时空、公共维度。

以主题域“行程记录”为例，新建主题域的步骤如下，其他主题域也请参照以下步骤进行添加：

1. 选中已创建的 L1 层主题“城市交通”。单击右键，选择“新建”。或者单击右侧的“新建”按钮。

图3-318 创建 L2 层主题



2. 在弹出窗口中，“名称”和“编码”请参照表 3-162 中的“主题域名称”和“主题域编码”进行填写，其他参数可根据实际情况进行填写，配置完成后单击“确定”完成主题域的新建。

步骤 6 新建业务对象。

- 在主题域“行程记录”下，新建 4 个业务对象：原始记录、标准记录、行程事实、记录统计。
- 在主题域“集团”下，新建 1 个业务对象：供应商。
- 在主题域“时空”下，新建 1 个业务对象：时间。
- 在主题域“公共维度”下，新建 1 个业务对象：公共维度。

以在主题域“行程记录”下新建业务对象“原始记录”为例，新建业务对象的步骤如下，其他业务对象也请参照以下步骤进行添加：

1. 选中已创建的 L2 层主题“行程记录”。单击右键，选择“新建”。或者单击右侧的“新建”按钮。
2. 在弹出窗口中，“名称”和“编码”请参照表 3-162 中的“业务对象名称”和“业务对象编码”进行填写，其他参数可根据实际情况进行填写，配置完成后单击“确定”完成业务对象新建。

----结束

新建码表并发布

在本示例中，您需要新建如表 3-163 所示的 3 个码表：

表3-163 码表

目录	*表名称	*表编码	表描述	*字段名称	*字段编码	*字段数据类型	字段描述
付款	付款	payment_type	无	付款方式	payment_type_id	BIGINT	无

目录	*表名称	*表编码	表描述	*字段名称	*字段编码	*字段数据类型	字段描述
方式	方式			编码			
				付款方式值	payment_type_value	STRING	无
供应商	供应商	vendor	无	供应商id	vendor_id	BIGINT	无
				供应商	vendor_value	STRING	无
费率	费率代码	rate_code	无	费率 id	rate_code_id	BIGINT	无
				费率说明	rate_code_value	STRING	无

操作步骤如下：

步骤 1 在数据架构控制台，单击左侧导航树中的“码表管理”，进入码表管理页面。

步骤 2 新建 3 个码表目录：付款方式、供应商、费率。

以新建“付款方式”目录为例，新建目录步骤如下，其他目录也请参照以下步骤进行新建。

1. 在码表管理页面，单击码表目录树中上方的 新建目录。

图3-319 码表目录树

2. 在弹出框中，输入目录名称，选择目录，然后单击“确定”。

图3-320 新建码表目录

步骤 3 新建 3 个码表：付款方式、供应商、费率代码。

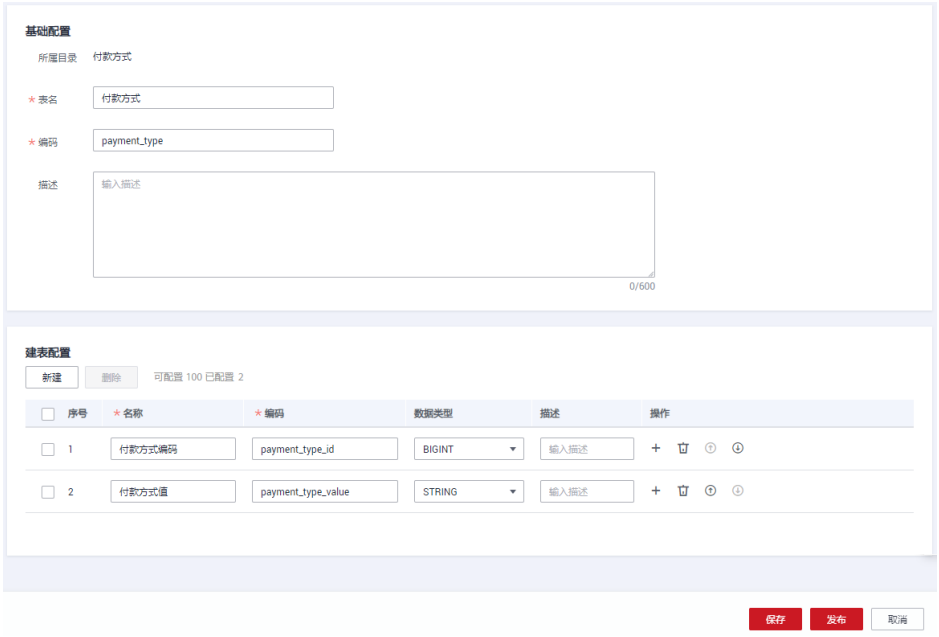
以新建“付款方式”码表为例，新建码表步骤如下，其他码表也请参照以下步骤完成新建：

1. 在码表管理页面，在码表目录树中选择一个目录，然后在右侧单击“新建”按钮。

图3-321 码表管理

2. 在新建码表页面中，请参考表 3-163 配置参数，然后单击“保存”。

图3-322 新建码表



3. 参考步骤步骤 3.1~步骤 3.2，在供应商目录下创建供应商码表，在费率目录下创建费率码表。

图3-323 供应商码表

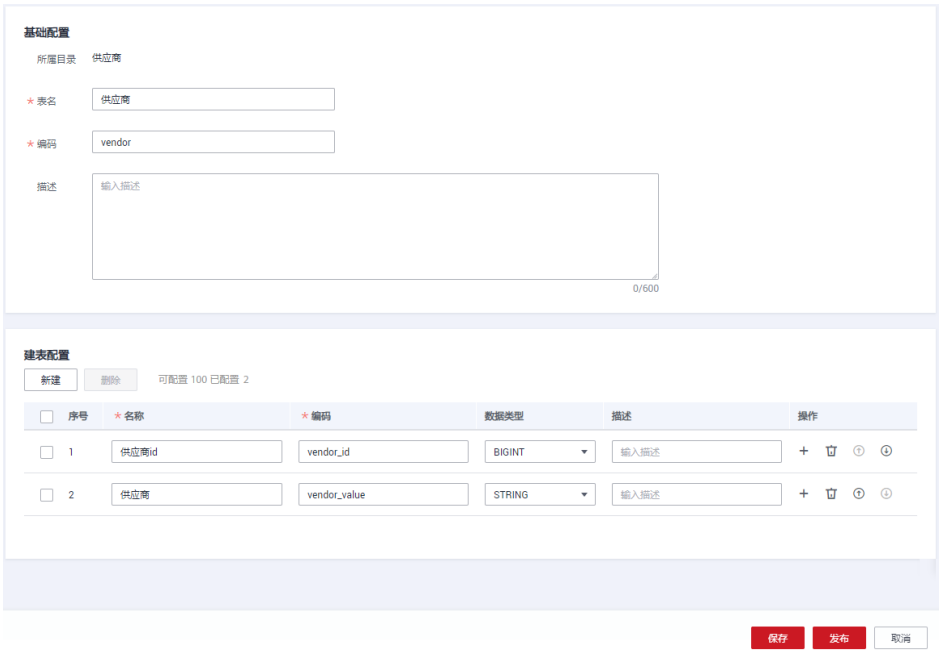
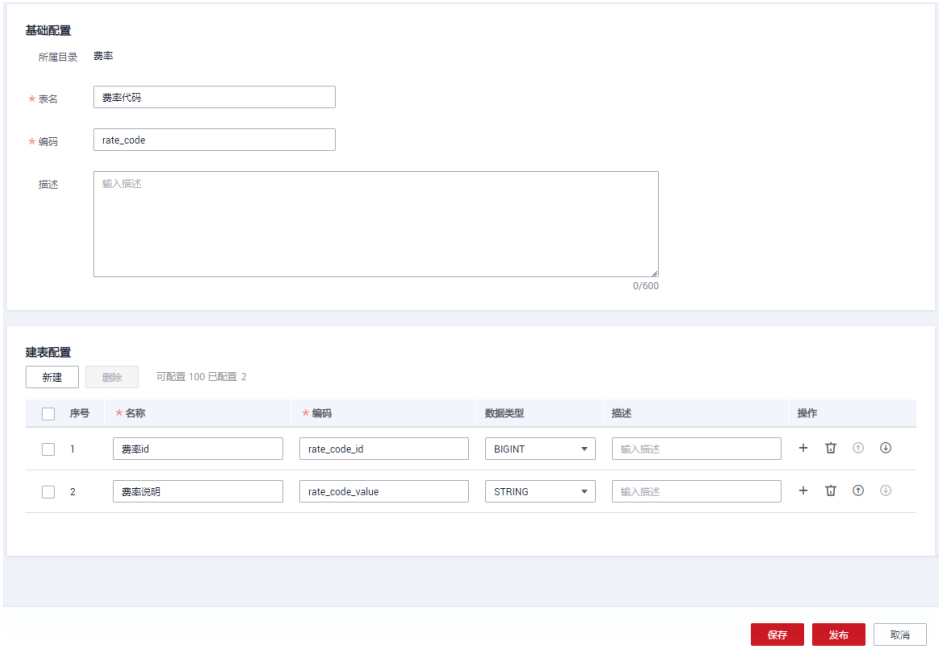


图3-324 费率码表



步骤 4 分别为付款方式、供应商、费率 3 个码表填写数值。

在“码表管理”页面，找到码表“付款方式”，然后在该码表所在行选择“更多 > 填写数值”。在填写数值页面，依次单击“新建”添加如表 3-164 所示的数值。

表3-164 付款方式码表的数值

付款方式编码 payment_type_id	付款方式值 payment_type_value
1	Credit card
2	Cash
3	No charge
4	Dispute
5	Unknown
6	Voided trip

返回“码表管理”页面，找到码表“供应商”，然后在该码表所在行选择“更多 > 填写数值”。在填写数值页面，依次单击“新建”添加如表 3-165 所示的数值。

表3-165 供应商码表的数值

供应商 id vendor_id	供应商 vendor_value

供应商 id vendor_id	供应商 vendor_value
1	A Company
2	B Company

返回“码表管理”页面，找到码表“费率代码”，然后在该码表所在行选择“更多 > 填写数值”。在填写数值页面，依次单击“新建”添加如表 3-166 所示的数值。

表3-166 费率码表的数值

费率 id rate_code_id	费率说明 rate_code_value
1	Standard rate
2	JFK
3	Newark
4	Nassau or Westchester
5	Negotiated fare
6	Group ride

步骤 5 返回码表管理页面后，在码表列表中，选中刚才新建的 3 个码表，然后单击“发布”发布码表。

步骤 6 在“批量发布”对话框中选择审核人，再单击“确认提交”，等待审核人员审核通过后，码表发布成功。如果当前帐号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

----结束

新建数据标准并发布

在本示例中，您需要新建如表 3-167 所示的 3 个数据标准：

表3-167 数据标准

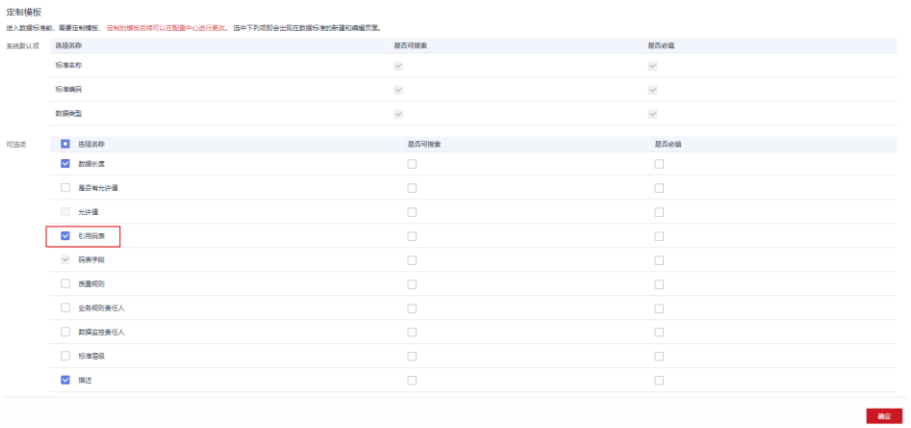
目录	*标准名称	*标准编码（自定义）	*数据类型	数据长度	引用码表	*码表字段	描述
付款方式	付款方式	payment_type	长整型(BIGINT)	无	付款方式	付款方式编码	无

目录	*标准名称	*标准编码（自定义）	*数据类型	数据长度	引用码表	*码表字段	描述
供应商	供应商	vendor	长整型(BIGINT)	无	供应商	供应商 id	无
费率	费率代码	rate_code	长整型(BIGINT)	无	费率代码	费率 id	无

步骤 1 在数据架构控制台，单击左侧导航树中的“数据标准”，进入数据标准页面。

步骤 2 首次进入“数据标准”页面，需要定制模板，定制的模板后续可以在配置中心进行更改。本示例需要额外勾选“引用码表”，如图所示。

图3-325 新建数据标准目录



步骤 3 请参考以下步骤，分别新建 3 个数据标准的目录：付款方式、供应商、费率。

在数据标准页面的目录树上方，单击 新建目录，然后在弹出框中输入目录名称“付款方式”并选择目录，单击“确定”完成目录的新建。

图3-326 新建数据标准目录

步骤 4 请参考以下步骤，分别新建 3 个数据标准：付款方式、供应商、费率。

1. 在数据标准页面的目录树中，选中所需要的目录，然后在右侧页面中单击“新建”。
2. 在新建数据标准页面中，3 个数据标准可分别参考如下配置，配置完成后单击“保存”。在本示例中，数据标准模板只选取了几个参数，您可以参考用户指南中的“数据架构> 配置中心”的“标准模板管理”定制数据标准模板。

图3-327 数据标准-付款方式

图3-328 数据标准-供应商

图3-329 数据标准-费率代码

步骤 5 返回数据标准页面后，在列表中勾选刚才新建的 3 个数据标准，然后单击“发布”发布数据标准。

步骤 6 在“批量发布”对话框中选择审核人，再单击“确认提交”，等待审核人员审核通过后，数据标准发布成功。如果当前帐号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

----结束

关系建模：新建 SDI 层和 DWI 层两个模型

在关系建模中，分别新建 SDI 层和 DWI 层两个关系模型，并通过逆向数据库导入原始数据表到 SDI 层的关系模型中，在 DWI 层模型中新建一个“标准出行数据”的标准化的业务表。

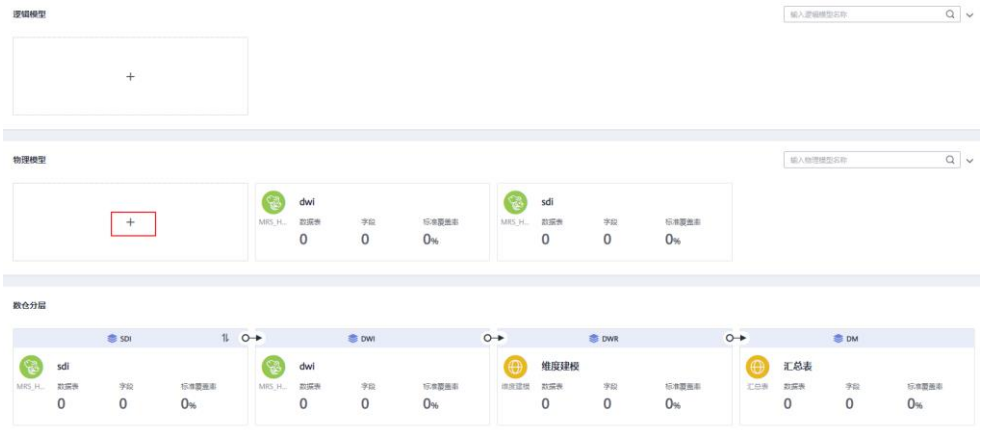
步骤 1 在数据架构控制台，单击左侧导航树中的“关系建模”。

- 如果当前未创建过关系模型，系统会弹出“新建分层治理模型”提示框。您可以新建一个 SDI 层关系模型，命名为“sdi”，再新建一个 DWI 层关系模型，命名为“dwi”。单击“确定”即可。

图3-330 “新建分层治理模型”提示框

- 如果不是首次创建，单击+新建物理模型，如下图所示。

图3-331 关系建模页面



a. 先新建一个 SDI 层关系模型，命名为“sdi”。在物理模型页签中，单击+，新建模型，配置如下参数，单击“确定”。

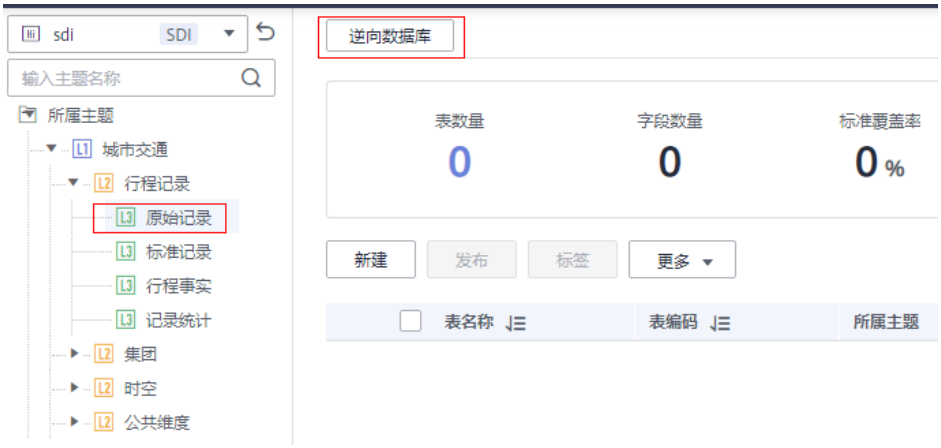
图3-332 新建 SDI 物理模型

b. 再新建一个 DWI 层关系模型，命名为“dwi”。在物理模型页签中，单击 +，新建模型，配置如下参数，单击“确定”。

图3-333 新建 DWI 模型

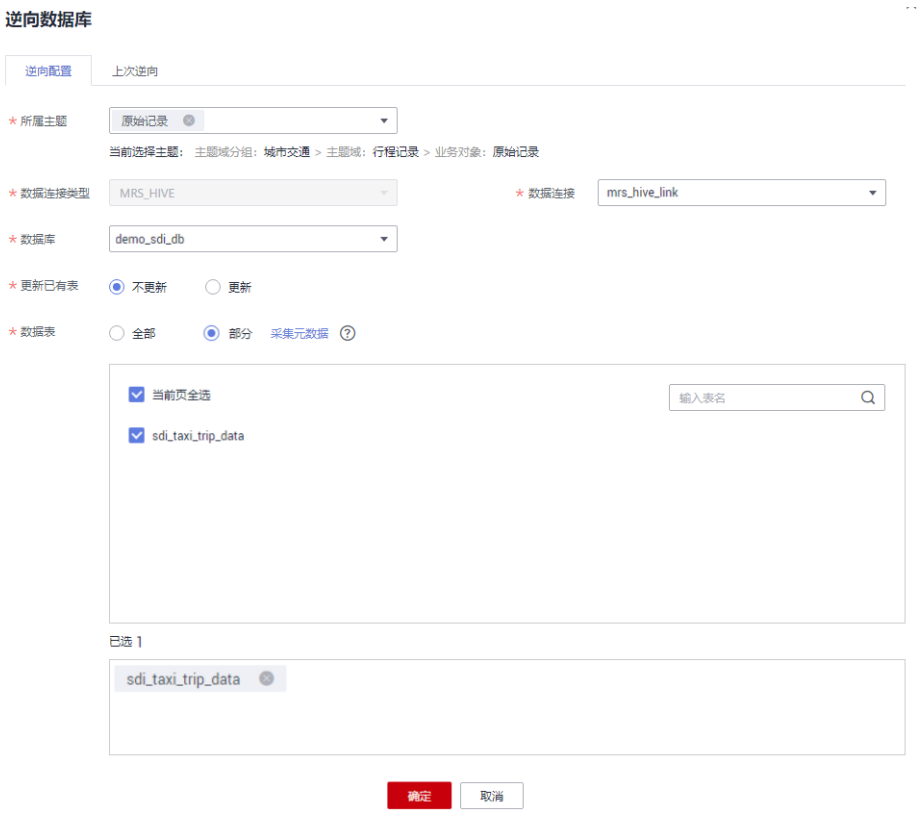
步骤 2 在“数仓分层”页签中，单击新建的 SDI 关系模型，展开，选中业务对象“城市交通 > 行程记录 > 原始记录”，单击“逆向数据库”，通过逆向数据库，导入原始表。

图3-334 模型目录



在“逆向数据库”窗口中，配置如下所示参数，然后单击“确定”。在本示例中选择贴源层数据库 demo_sdi_db 中的原始数据表。

图3-335 逆向数据库



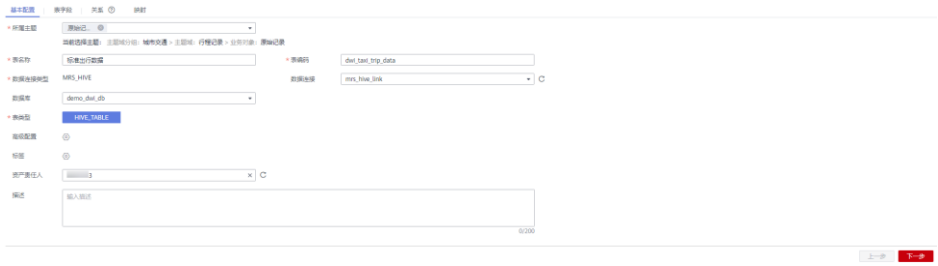
逆向数据库成功后，单击“关闭”。您可以在列表中查看导入的表：

图3-336 查看表

步骤 3 请参照以下步骤，新建一个“标准出行数据”的标准化的业务表。

1. 在“数仓分层”页签中，单击新建的 DWI 关系模型，展开，选中 DWI 模型中的业务对象“城市交通 > 行程记录 > 原始记录”，然后在右侧列表上方单击“新建”按钮，进入新建表页面。
2. 在新建表的“基本配置”标签页中，配置如下：

图3-337 行程数据表基本配置



3. 进入“表字段”标签页，单击“新建”，在标准出行数据表中，依次添加如表 3-168 所示的字段，并单击字段供应商编号、费率代码、付款方式的“数据标准”列中的 按钮，分别关联数据标准“供应商”、“费率代码”和“付款方式”。添加完成后如图 3-338 所示。

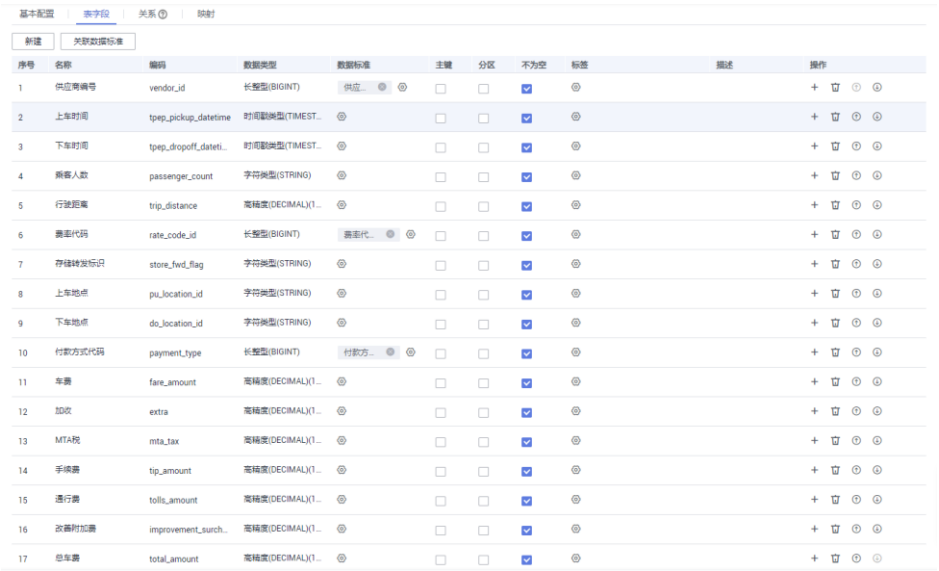
表3-168 标准出行数据表字段

序号	名称	编码	数据类型	数据标准	主键	分区	不为空	标签
1	供应商编号	vendor_id	长整型(BIGINT)	供应商	不勾选	不勾选	勾选	-
2	上车时间	tpep_pickup_datetime	时间戳类型(TIMESTAMP)	-	不勾选	不勾选	勾选	-
3	下车时间	tpep_dropoff_datetime	时间戳类型(TIMESTAMP)	-	不勾选	不勾选	勾选	-

序号	名称	编码	数据类型	数据标准	主键	分区	不为空	标签
4	乘客人数	passenger_count	字符类型(STRING)	-	不勾选	不勾选	勾选	-
5	行驶距离	trip_distance	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	-
6	费率代码	rate_code_id	长整型(BIGINT)	费率代码	不勾选	不勾选	勾选	-
7	存储转发标识	store_fwd_flag	字符类型(STRING)	-	不勾选	不勾选	勾选	-
8	上车地点	pu_location_id	字符类型(STRING)	-	不勾选	不勾选	勾选	-
9	下车地点	do_location_id	字符类型(STRING)	-	不勾选	不勾选	勾选	-
10	付款方式代码	payment_type	长整型(BIGINT)	付款方式	不勾选	不勾选	勾选	-
11	车费	fare_amount	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	-
12	加收	extra	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	-
13	MTA税	mta_tax	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	-
14	手续费	tip_amount	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	-
15	通行费	tolls_amount	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	-
16	改善附加	improvement_surcharge	高精度(DECIMAL)(	-	不勾选	不勾选	勾选	-

序号	名称	编码	数据类型	数据标准	主键	分区	不为空	标签
	费		10,2)					
17	总车费	total_amount	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	-

图3-338 行程数据表字段



对于标准出行数据表中的字段，您可以执行以下操作。

- **关联数据标准**

  在新建表或编辑表时，进入“表字段”标签页，在字段所在行的“数据标准”列，单击 按钮可以选择一个数据标准与字段相关联。将字段关联数据标准后，表发布上线后，就会自动生成一个质量作业，每个关联了数据标准的字段会生成一个质量规则，基于数据标准对字段进行质量监控，您可以前往 DataArts Studio 数据质量模块的“质量作业”页面进行查看。有关关联数据标准的更多信息，请参见用户指南中的“数据架构> 关系建模 > 物理模型设计”中的“新建表并发布”。

- **添加标签**

  标签是用户自定义的标识。添加标签后，您就可以在 DataArts Studio 数据目录模块中通过标签搜索相关的数据资产。

  在新建表或编辑表时，进入“表字段”标签页，在字段所在行的“标签”列，单击 按钮可以添加标签，在弹出框中，您可以输入新的标签名称后按回车，也可以在下拉列表中选择已有标签。

- **关联质量规则**

  完成表的新建后，您可以在表中为字段关联质量规则，完成关联后，当表发布成功后，就会在 DataArts Studio 数据质量中自动创建质量作业，如果当前表已经发布，则系统会自动更新质量作业。有关关联质量规则的更多信息，请参见用户指南中的“数据架构> 关系建模 > 关联质量规则”。

4. 接下来，进入“映射”标签页，通过新建映射设计表的数据来源。
   - 如果表中的字段数据来源于不同的关系模型，您需要创建多个映射。在每个映射中，您只需要为来源于当前映射的字段设置源字段，其他字段可以不设置。
   - 如果表中的字段数据来源于同一个关系模型中的多个表，您可以新建一个映射。在该映射的“源表”中，您可以将多个表设置 Join，然后再为表中的字段设置源字段。

   本示例只需要新建一个映射。单击“新建”，新建一个映射，如图 3-339。
   - **映射名称**：新建映射时会自动生成，您也可以修改。
   - **来源模型**：本示例选择“sdi”。
   - **源表**：本示例选择原始数据表“sdi_taxi_trip_data”，标准出行数据表的数据均来源于该原始数据表。

     图3-339 新建映射

   - **字段映射**：

     在“字段映射”区域，依次为表中的字段设置源字段，所选择的源字段应与表中的字段代表相同含义，一一对应。如图 3-340 所示，在字段映射的底部，会显示生成的 SQL 语句，可供参考。

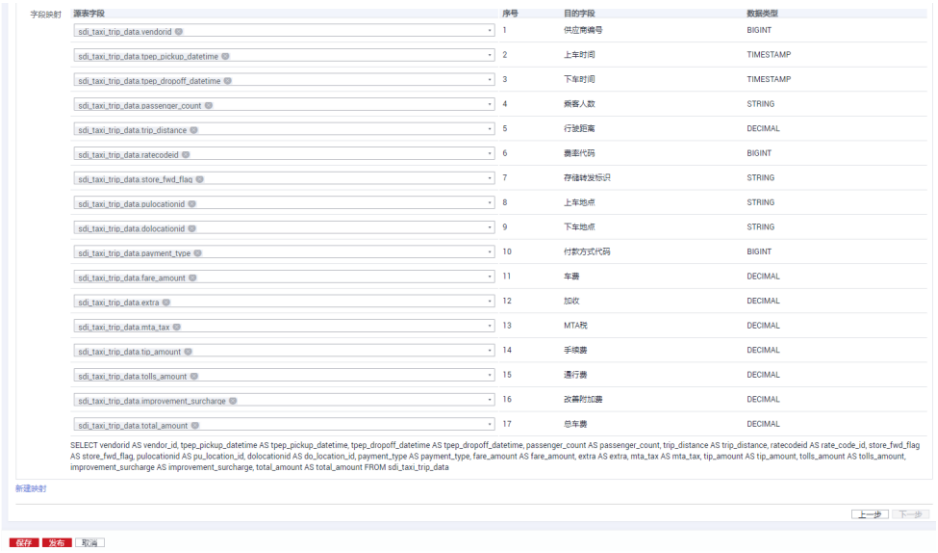
说明

- 如果在“数据架构 > 配置中心 > 功能配置”页面中开启了“模型设计业务流程步骤 > 创建数据开发作业”（默认为关闭），发布表时，系统支持根据表的映射信息，在数据开发中自动创建一个 ETL 作业，每一个映射会生成一个 ETL 节点，作业名称以“*数据库名称_表编码*”开头。当前该功能处于内测阶段，仅支持 DLI->DLI 和 DLI->DWS 两种映射的作业创建。

  已创建的 ETL 作业可以进入“数据开发 > 作业开发”页面查看。ETL 作业默认每天 0 点启动调度。

- 在本示例中，不支持自动创建 ETL 作业，映射信息仅为数据开发提供数据的 ETL 流向。在做数据开发的过程中，可以参考此处的映射关系编写 SQL 脚本。

图3-340 字段映射



5. 完成映射的配置后，出租车行程数据表配置完成，单击“保存”。

步骤 4 模型创建好之后，勾选已创建的模型，选择“更多 > 导出”，然后在弹出框中选中“表”并单击“确定”，可以将整个模型导出。参考同样的方法导出模型“sdi”。导出后的模型，可以作为备份，今后可用于模型导入。

图3-341 导出模型

步骤 5 发布表模型。

1. 发布步骤 2 中通过逆向数据库导入 SDI 模型的原始表，发布后，就可以通过 DataArts Studio 对原始表进行管理和监控。

   返回关系建模页面，在模型目录选择“sdi”模型，然后在右侧的列表中，勾选表 sdi_taxi_trip_data，再单击“发布”，然后在弹出框中选择审核人，单击“确认提交”，等待审核人员审核通过后，“sdi”模型发布成功。如果当前帐号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

2. 发布 DWI 模型中的表。

   返回关系建模页面，在模型目录中选择“dwi”模型，然后在右侧的列表中，勾选表“标准出行数据”，再单击“发布”，然后在弹出框中选择审核人，单击“确认提交”，等待审核人员审核通过后，“dwi”模型发布成功。如果当前帐号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

步骤 6 当表模型发布成功后，进入数据架构的“关系建模”页面可以查看表的“状态”和“同步状态”。

发布是一个异步操作，您可以单击 按钮刷新状态。表发布并通过审核后，系统会依据“配置中心 > 功能配置”页面中的“模型设计业务流程步骤”进行创建表、同步技术资产、同步业务资产等操作，在表的“同步状态”一列中将显示同步状态。

- “同步状态”若均显示成功，则说明表发布成功。鼠标移至“同步状态”中的 图标之上，若显示“创建表：创建成功”说明该表在对应的数据源下已经创建成功。
- “同步状态”若显示某一项或某几项失败，可以先刷新状态。如果仍失败，可以选择操作列的“更多 > 发布历史”，然后进入“发布日志”标签页查看日志。

  请根据错误日志定位失败原因，问题解决后，再返回“关系建模”页面，在列表中勾选需同步的表，然后选择“更多 > 同步”尝试重新同步。如果仍同步失败，请联系技术支持人员协助解决。

图3-342 查看表状态

在列表中单击表名，可以查看表的详情，其中“数据源”显示了表的位置。

图3-343 表详情

----结束

维度建模：在 DWR 层新建并发布维度

在维度建模中，在 DWR 数据报告层中新建 3 个码表维度（供应商、费率代码和付款方式）和 1 个层级维度（日期维度）。

步骤 1 在数据架构控制台，单击左侧导航树中的“维度建模”，进入维度建模页面。

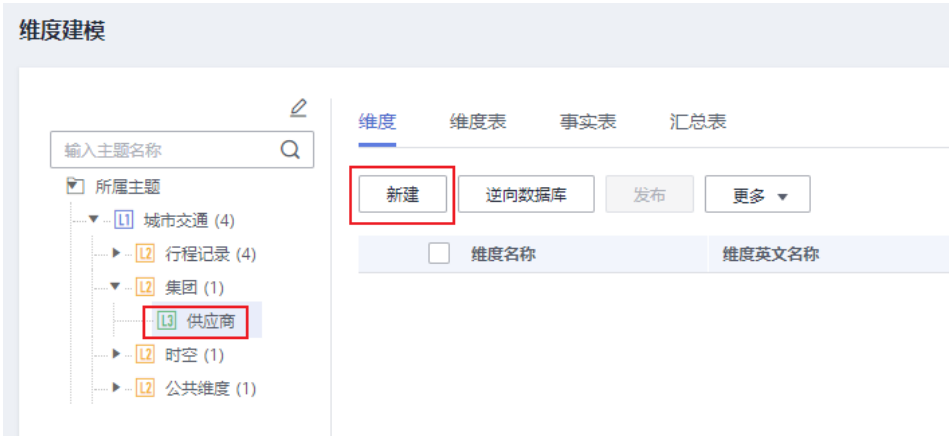
步骤 2 新建如表 3-169 所示的 3 个码表维度。

表3-169 码表维度

*所属主题	*维度名称	*维度英文名称	*维度类型	*资产责任人	描述	*数据连接类型	*数据连接	*数据库	选择码表
供应商	供应商	dim_vendor	码表维度	-	无	MRS_HIVE	mrs_hive_link	demo_dwr_db	供应商
公共维度	费率代码	dim_rate_code	码表维度	-	无	MRS_HIVE	mrs_hive_link	demo_dwr_db	费率
公共维度	付款方式	dim_payment_type	码表维度	-	无	MRS_HIVE	mrs_hive_link	demo_dwr_db	付款方式

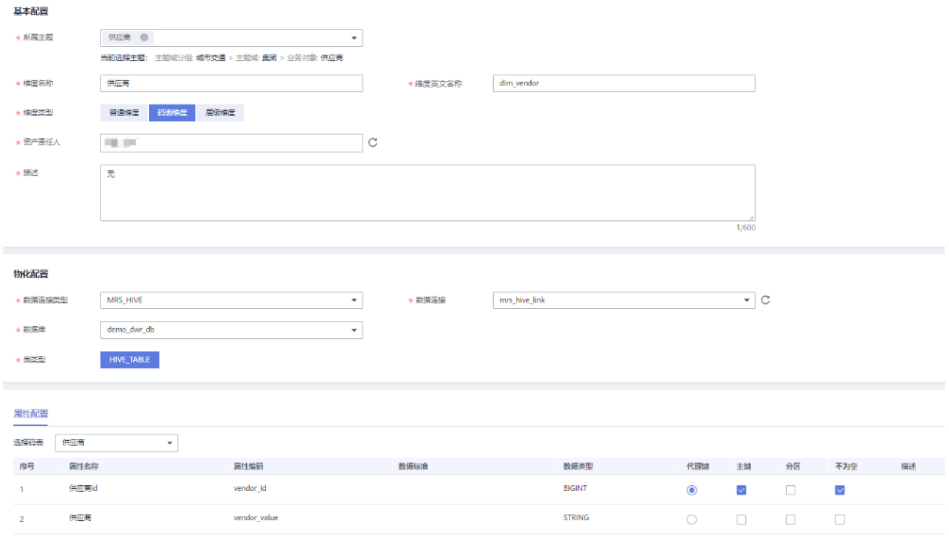
1. 在“维度建模”页面进入“维度”标签页，在主题树中选中“城市交通 > 集团 > 供应商”，然后单击“新建”新建供应商维度。

图3-344 维度建模



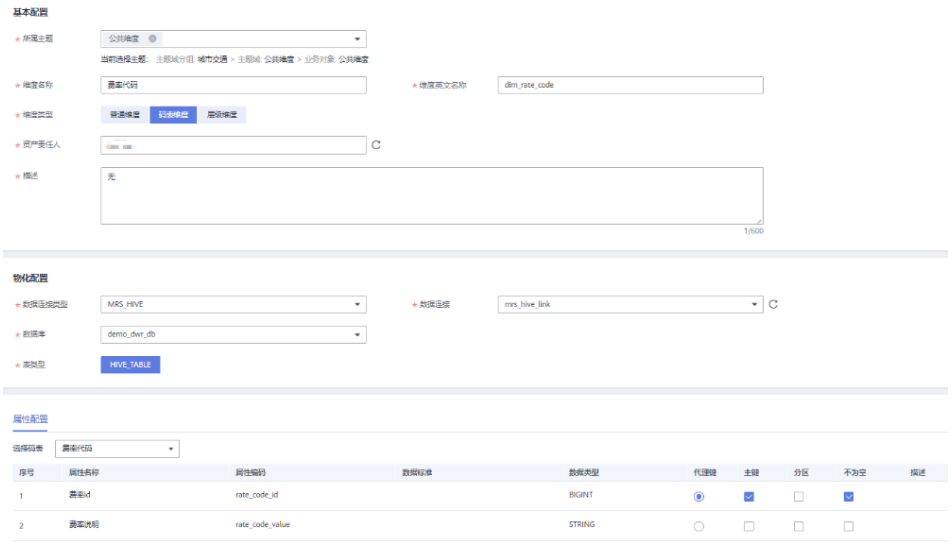
2. 在新建维度页面，如下图所示配置参数，然后单击“保存”完成维度的新建。

图3-345 新建维度



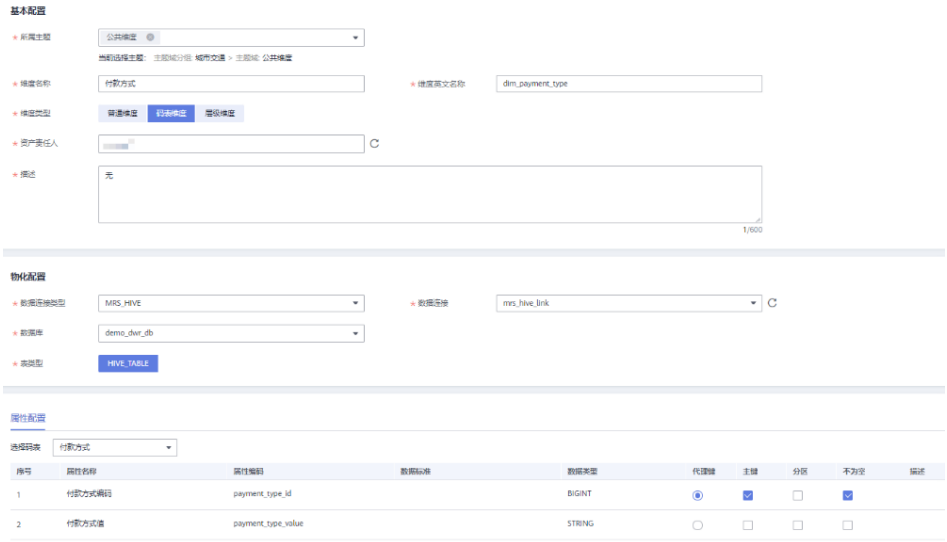
3. 在“维度建模”页面进入“维度”标签页，在主题树中选中“城市交通 > 公共维度 > 公共维度”，然后单击“新建”新建费率代码维度。在新建维度页面，配置如下，配置完成后单击“保存”。

图3-346 费率代码维度



4. 在“维度建模”页面进入“维度”标签页，在主题树中选中“城市交通 > 公共维度 > 公共维度”，然后单击“新建”新建付款方式维度。在新建维度页面，维度配置如下，配置完成后单击“保存”。

图3-347 付款方式维度



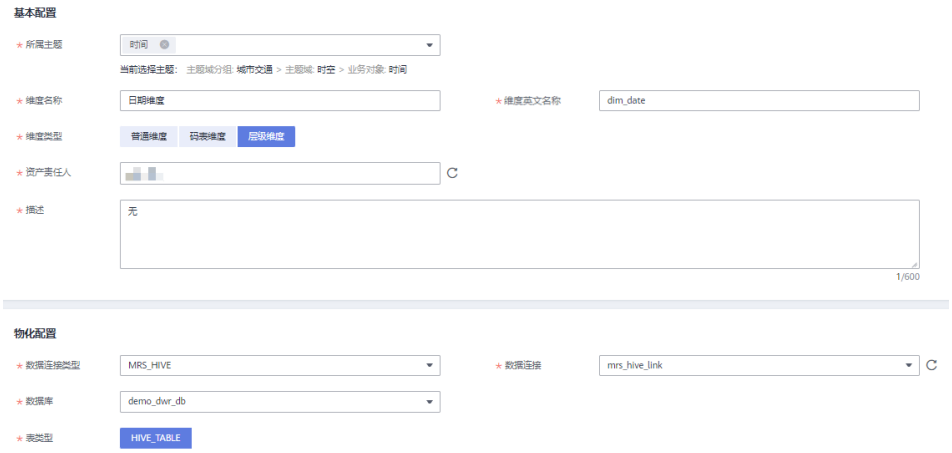
步骤 3 新建一个层级维度“日期维度”。

1. 在“维度建模”页面进入“维度”标签页，在主题树中选中“城市交通 > 时空 > 时间”，然后单击“新建”新建日期维度。
2. 基本配置和物化配置如下：

表3-170 日期维度

*所属主题	*维度名称	*维度英文名称	*维度类型	*资产责任人	描述	*数据连接类型	*数据连接	*数据库
时间	日期维度	dim_date	层级维度	-	无	MRS_HIVE	mrs_hive_link	demo_dwr_db

图3-348 日期维度



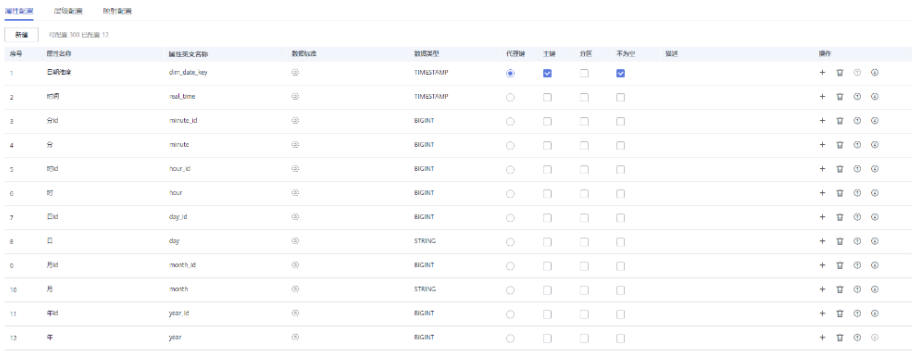
3. 属性配置如下：

表3-171 属性配置

序号	属性名称	属性英文名称	数据标准	数据类型	代理键	主键	分区	不为空
1	日期维度	dim_date_key	-	TIMESTAMP	选中	选中	不勾选	勾选
2	时间	real_time	-	TIMESTAMP	不选	不选	不勾选	不勾选
3	分 id	minute_id	-	BIGINT	不选	不选	不勾选	不勾选
4	分	minute	-	BIGINT	不选	不选	不勾选	不勾选
5	时 id	hour_id	-	BIGINT	不选	不选	不勾选	不勾选
6	时	hour	-	BIGINT	不选	不选	不勾选	不勾选
7	日 id	day_id	-	BIGINT	不选	不选	不勾选	不勾选
8	日	day	-	STRING	不选	不选	不勾选	不勾选
9	月 id	month_id	-	BIGINT	不选	不选	不勾选	不勾选
10	月	month	-	STRING	不	不选	不勾选	不勾选

序号	属性名称	属性英文名称	数据标准	数据类型	代理键	主键	分区	不为空
					选			
11	年 id	year_id	-	BIGINT	不选	不选	不勾选	不勾选
12	年	year	-	BIGINT	不选	不选	不勾选	不勾选

图3-349 属性配置



4. 在层级配置区域，单击“新建”，新建如下 2 个层级：

图3-350 层级 1

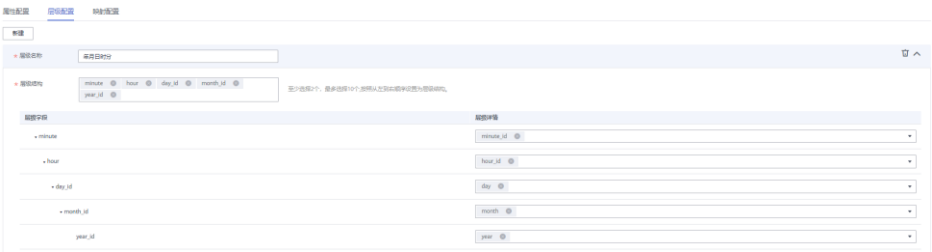


图3-351 层级 2

5. 新建维度页面配置完成后，单击“保存”。

步骤 4 返回维度页面后，在维度列表中，勾选刚才新建的 4 个维度，再单击“发布”。

步骤 5 在“批量发布”对话框中，选择审核人，单击“确认提交”，等待审核人员审核通过后，维度发布成功。如果当前帐号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

步骤 6 完成所有维度的新建和发布，待审核通过后，系统会自动创建与维度相对应的维度表，维度表的名称和编码均与维度相同。在“维度建模”页面，选择“维度表”页签，可以查看建好的维度表。

在维度表列表中，在“同步状态”一列中可以查看维度表的同步状态。

- 如果同步状态均显示成功，则说明维度发布成功，维度表在数据库中创建成功。
- 如果同步状态中存在失败，可单击该维度表所在行的“发布历史”，然后在展开的页面中进入“发布日志”页签查看日志。请根据错误日志定位失败原因，问题解决后，您可以勾选该维度表，再单击列表上方的“同步”按钮尝试重新同步。如果问题仍未能解决，请联系技术支持人员协助处理。

图3-352 维度表同步状态

----结束

维度建模：在 DWR 层新建并发布事实表

在维度建模中，在 DWR 数据报告层中新建一个事实表“行程订单”。

步骤 1 在数据架构控制台，单击左侧导航树中的“维度建模”，进入维度建模页面。

步骤 2 单击“事实表”页签，进入事实表页面。在左侧的主题树中选择业务对象“城市交通 > 行程记录 > 行程事实”，然后单击“新建”按钮开始新建行程订单表。

在新建事实表页面的“基本配置”区域，配置如下：

- 所属主题：主题域分组：城市交通>主题域：行程记录>业务对象：行程事实
- 表名称：行程订单
- 表英文名称：fact_stroke_order
- 数据连接类型：MRS_HIVE
- 数据连接：mrs_hive_link
- 数据库：demo_dwr_db
- 表类型：HIVE_TABLE
- 资产责任人：在下拉列表中选择一个责任人。

- 描述：无

在“字段配置”区域，选择 " 新建 > 维度"，在弹出框中选择维度“费率代码”、“供应商”、“付款方式”、“日期维度”，单击“确定”。再次选择 " 新建 > 维度"，在弹出框中选择“日期维度”并单击“确定”。然后，在维度字段列表中，调整维度字段的顺序，并修改 2 个日期维度的信息，如表 3-172 所示。

表3-172 维度字段

序号	字段名称	字段英文名称	数据类型	主键	分区	不为空	关联数据标准	关联维度	角色	描述
1	费率 id	rate_code_id	BIGINT	不勾选	不勾选	不勾选	-	费率代码	dim_	-
2	供应商 id	vendor_id	BIGINT	不勾选	不勾选	不勾选	-	供应商	dim_	-
3	付款方式编码	payment_type_id	BIGINT	不勾选	不勾选	不勾选	-	付款方式	dim_	-
4	上车时间	dim_pickup_date_key	TIMESTAMP	不勾选	不勾选	不勾选	-	日期维度	dim_pickup	日期层维表
5	下车时间	dim_dropoff_date_key	TIMESTAMP	不勾选	不勾选	不勾选	-	日期维度	dim_dropoff	日期层维表

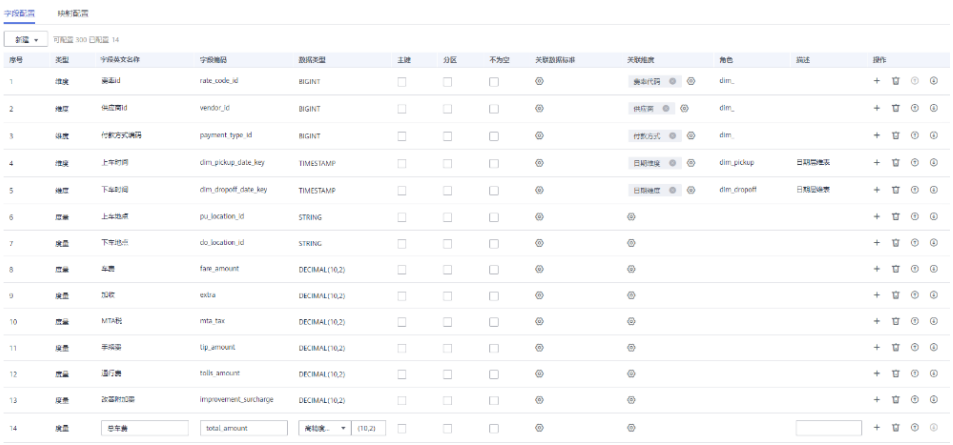
在“字段配置”区域，选择“新建 > 度量”，依次新建如表 3-173 所示的字段。

表3-173 度量属性

序号	字段名称	字段英文名称	数据类型	主键	分区	不为空	关联数据标准

序号	字段名称	字段英文名称	数据类型	主键	分区	不为空	关联数据标准
6	上车地点	pu_location_id	字符类型(STRING)	不勾选	不勾选	不勾选	-
7	下车地点	do_location_id	字符类型(STRING)	不勾选	不勾选	不勾选	-
8	车费	fare_amount	高精度(DECIMAL)(10,2)	不勾选	不勾选	不勾选	-
9	加收	extra	高精度(DECIMAL)(10,2)	不勾选	不勾选	不勾选	-
10	MTA 税	mta_tax	高精度(DECIMAL)(10,2)	不勾选	不勾选	不勾选	-
11	手续费	tip_amount	高精度(DECIMAL)(10,2)	不勾选	不勾选	不勾选	-
12	通行费	tolls_amount	高精度(DECIMAL)(10,2)	不勾选	不勾选	不勾选	-
13	改善附加费	improvement_surcharge	高精度(DECIMAL)(10,2)	不勾选	不勾选	不勾选	-
14	总车费	total_amount	高精度(DECIMAL)(10,2)	不勾选	不勾选	不勾选	-

图3-353 事实表字段配置



步骤 3 新建事实表页面配置完成后，单击“发布”提交审核。

步骤 4 在弹出框中选择审核人，单击“确认提交”，等待审核人员审核通过后，事实表发布成功。如果当前帐号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

步骤 5 返回“维度建模 > 事实表”页面，在列表中找到刚发布的事实表，在“同步状态”一列中可以查看事实表的同步状态。

- 如果同步状态均显示成功，则说明事实表发布成功，事实表在数据库中已创建成功。
- 如果同步状态中存在失败，可单击该事实表所在行的“更多 > 发布历史”，然后在展开的页面中进入“发布日志”页签查看日志。请根据错误日志定位失败原因，问题解决后，您可以在事实表页面勾选该事实表，再单击列表上方的“更多 > 同步”尝试重新同步。如果问题仍未能解决，请联系技术支持人员协助处理。

----结束

指标设计：新建并发布技术指标

在本示例中，您需要新建如表 3-174 和表 3-175 所示的技术指标：

表3-174 原子指标

*指标名称	*指标英文名称	数据表	*所属主题	*设定表达式	描述
总车费	sum_total_amount	行程订单	行程事实	sum (总车费)	无

表3-175 衍生指标

指标	*数据表	*所属主题	*原子指标	统计维度	时间限定	通用限定
基于付款方式维度统计总车费	行程订单	记录统计	总车费	付款方式	无	无
基于费率代码维度统计总车费	行程订单	记录统计	总车费	费率代码	无	无
基于供应商和下车时间维度统计总车费	行程订单	记录统计	总车费	供应商，行程订单.下车时间	无	无

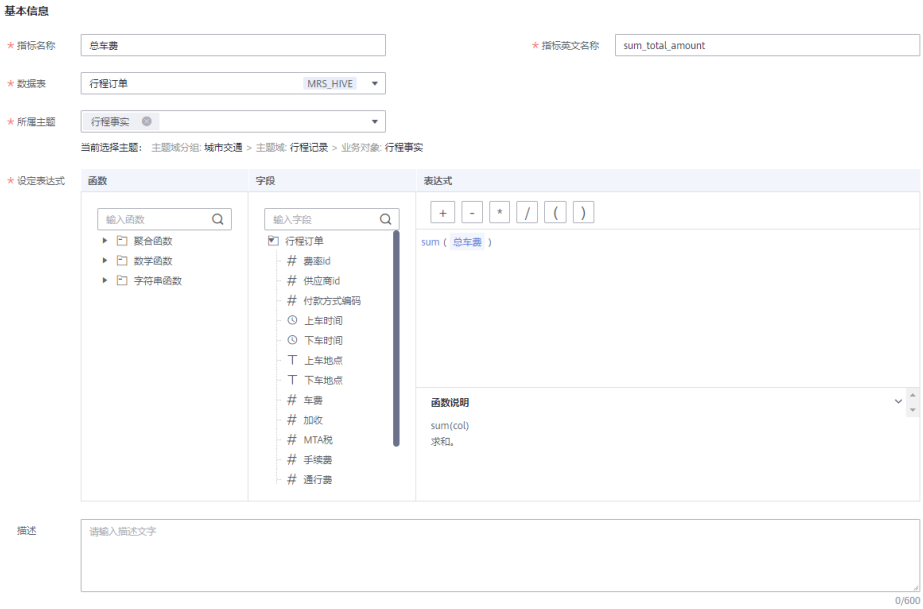
步骤 1 在数据架构控制台，单击左侧导航树中的“技术指标”，进入技术指标页面。

步骤 2 新建一个原子指标“总车费”，用于统计总车费。

1. 在技术指标页面，进入“原子指标”标签页，然后单击“新建”按钮。

2. 在新建原子指标页面配置如下，配置完成后单击“发布”。

图3-354 原子指标



3. 等待审核人审核通过。审核通过后，原子指标就创建好了。

步骤 3 当原子指标通过审核后，新建以下 3 个衍生指标。

- 总车费(付款方式)：基于付款方式维度统计总车费

  在技术指标页面，进入“衍生指标”标签页，然后单击“新建”按钮，在新建衍生指标页面，配置如下。配置完成后，单击“试运行”，并在弹出窗口中单击“执行”，如果运行通过单击“保存”。

图3-355 总车费（付款方式）

- 总车费(费率代码)：基于费率代码维度统计总车费

在技术指标页面，进入“衍生指标”标签页，然后单击“新建”按钮，在新建衍生指标页面，配置如下。配置完成后，单击“试运行”，并在弹出窗口中单击“执行”，如果运行通过单击“保存”。

图3-356 总车费(费率代码)

- 截止当日_总车费(供应商,行程订单.下车时间)：基于供应商维度统计总车费

  在技术指标页面，进入“衍生指标”标签页，然后单击“新建”按钮，在新建衍生指标页面，配置如下。配置完成后，单击“试运行”，并在弹出窗口中单击“执行”，如果运行通过单击“保存”。

图3-357 总车费(供应商)

步骤 4 返回技术指标页面的“衍生指标”标签页后，勾选建好的 3 个衍生指标，单击“发布”，在弹出框中选择审核人，单击“确认提交”，等待审核人员审核通过后，事实表发布成功。如果当前帐号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

----结束

维度建模：在 DM 层新建并发布汇总表

在 DM 数据集市层，您需要新建如表 3-176 所示的汇总表。

表3-176 汇总表

*所属主题	*表名称	*表英文名称	统计维度	数据连接类型	*数据连接	*数据库	资产责任人	描述
记录统计	付款方式统计汇总	dws_payment_type	付款方式	MRS_HIVE	mrs_hive_link	demo_dm_db	-	无
记录统计	费率统计汇总	dws_rate_code	费率代码	MRS_HIVE	mrs_hive_link	demo_dm_db	-	无
记录统计	供应商统计汇总	dws_vendor	供应商,行程订单.下车时间	MRS_HIVE	mrs_hive_link	demo_dm_db	-	无

步骤 1 在数据架构控制台，单击左侧导航树中的“维度建模”，进入维度建模页面。

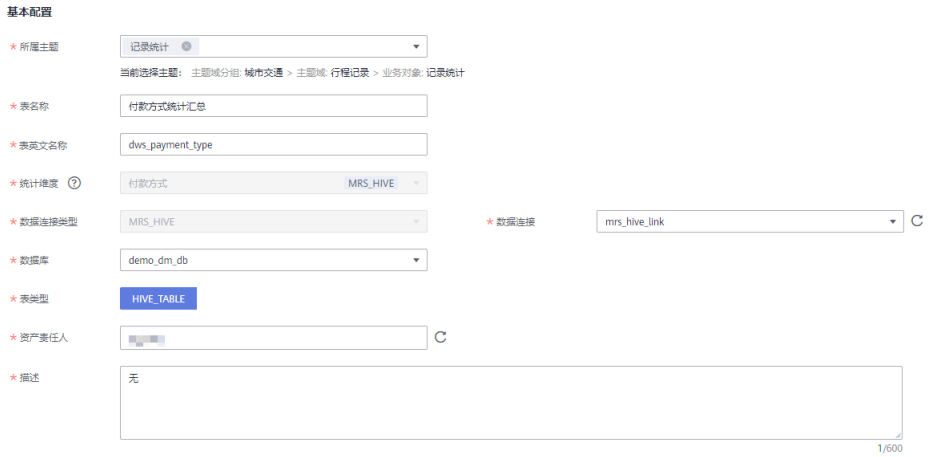
步骤 2 单击“汇总表”页签，进入汇总表页面。

步骤 3 新建 3 个汇总表：付款方式统计汇总表、费率统计汇总表、供应商统计汇总表。

1. 在“汇总表”页面，在主题树中选中“城市交通 > 行程记录 > 记录统计”，然后单击“新建”新建付款方式统计汇总表。在新建汇总表页面，配置如下，配置完成后单击“保存”。

   在新建汇总表页面，基本配置如下：

   图3-358 付款方式统计汇总

   在“时间分区”区域，输入字段编码以及选择数据类型。当表发布成功后，在往表里写数据时，将根据该时间分区字段进行分区。

图3-359 时间分区配置

在“指标配置”区域，单击“添加”，添加衍生指标“总车费(付款方式)”。此处只能添加与所指定的“统计维度”相关联的并且已发布的衍生指标或复合指标。

图3-360 指标配置

完成上述配置后，单击“保存”。

2. 在“汇总表”页面，在主题树中选中“城市交通 > 行程记录 > 记录统计”，然后单击“新建”新建费率统计汇总表。在新建汇总表页面，配置如下，配置完成后单击“保存”。

图3-361 费率统计汇总-基本配置

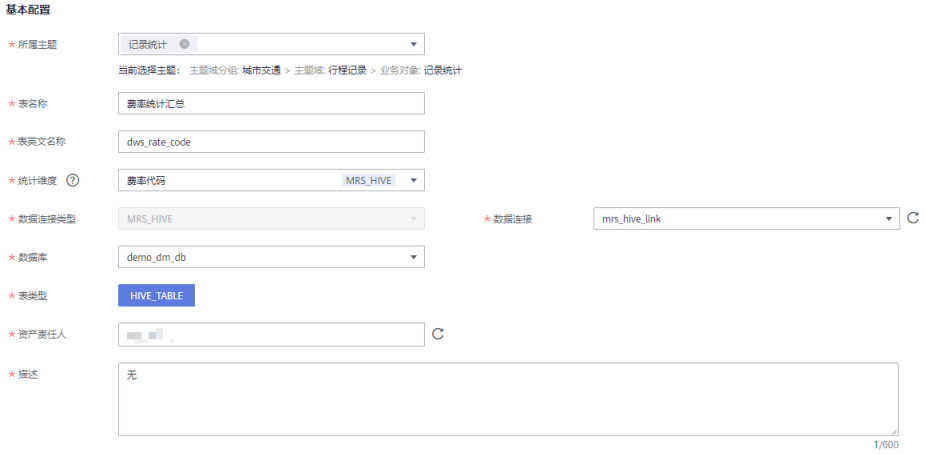


图3-362 费率统计汇总-指标配置

3. 在“汇总表”页面，在主题树中选中“城市交通 > 行程记录 > 记录统计”，然后单击“新建”新建供应商统计汇总表。在新建汇总表页面，配置如下，配置完成后单击“保存”。

图3-363 供应商统计汇总-基本配置

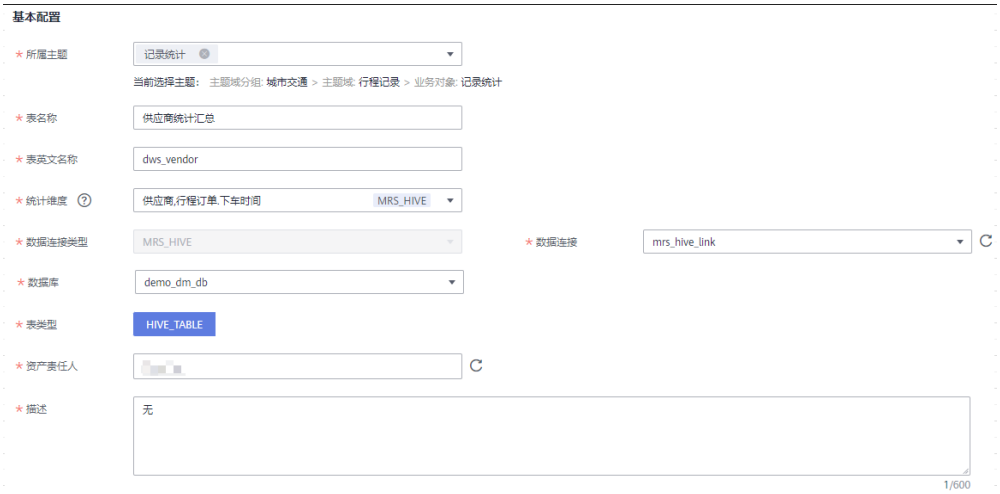


图3-364 供应商统计汇总-指标配置

步骤 4 返回维度建模页面的“汇总表”标签页后，勾选建好的 3 个汇总表，单击“发布”。

步骤 5 在弹出框中选择审核人，单击“确认提交”，等待审核人员审核通过后，汇总表会自动创建。如果当前帐号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

步骤 6 返回“维度建模 > 汇总表”页面，在列表中找到刚发布的汇总表，在“同步状态”一列中可以查看汇总表的同步状态。

- 如果同步状态均显示成功，则说明汇总表发布成功，汇总表在数据库中已创建成功。
- 如果同步状态中存在失败，可单击该汇总表所在行的“更多 > 发布历史”，然后在展开的页面中进入“发布日志”页签查看日志。请根据错误日志定位失败原因，问题解决后，您可以在汇总表页面勾选该汇总表，再单击列表上方的“更多 > 同步”尝试重新同步。如果问题仍未能解决，请联系技术支持人员协助处理。

----结束

3.5 数据开发

3.5.1 数据开发概述

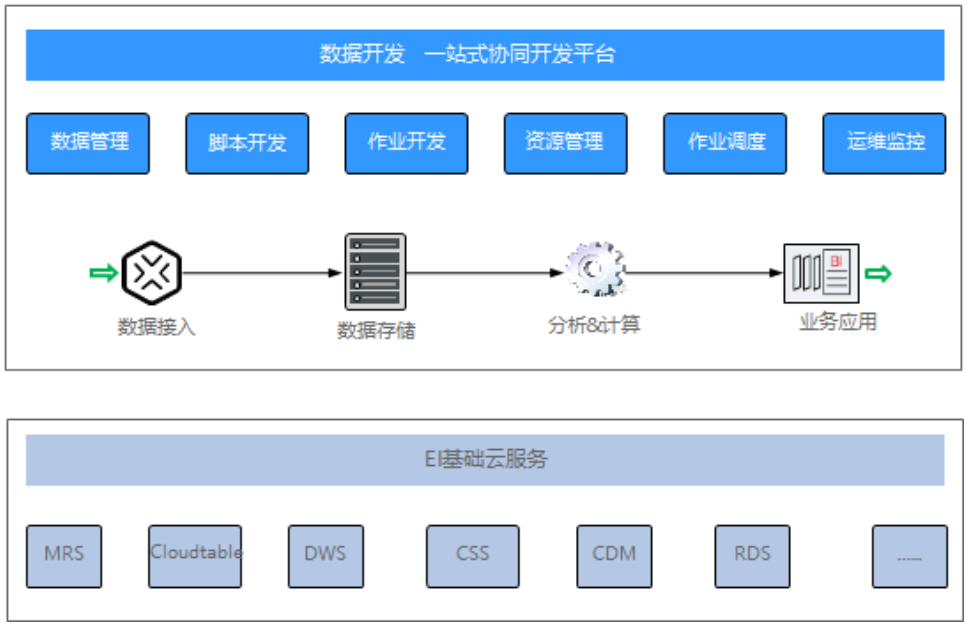
数据开发是一个一站式的大数据协同开发平台，提供全托管的大数据调度能力。它可管理多种大数据服务，极大降低用户使用大数据的门槛，帮助您快速构建大数据处理中心。

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3-365 数据开发模块架构



数据开发的主要功能

表3-177 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。

支持的功能	说明
	• 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。

3.5.2 数据管理

3.5.2.1 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：

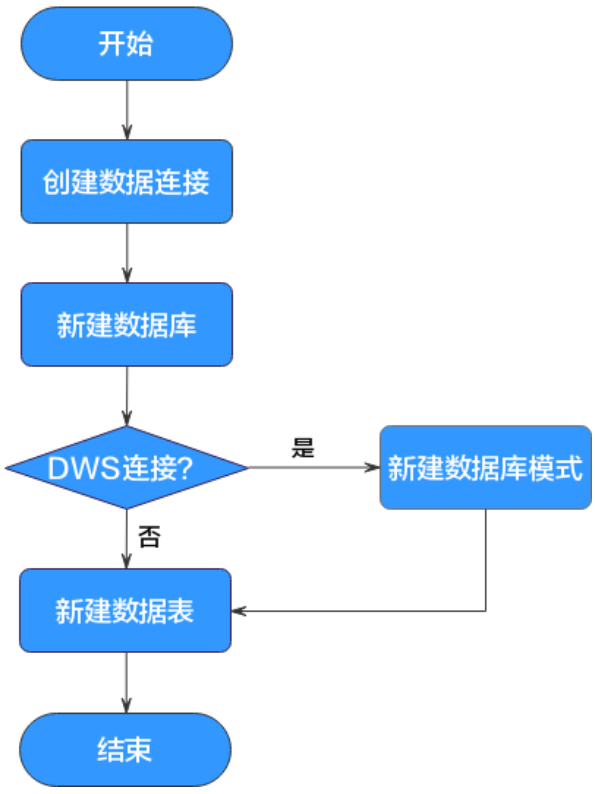
- 支持管理 DWS、MRS Hive 等多种数据湖。
- 支持可视化和 DDL 方式管理数据库表。

说明

如果您在使用数据开发前，已参考 3.1 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 3.5.3.1 脚本开发流程或 3.5.4.1 作业开发流程。

数据管理的使用流程如下：

图3-366 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 3.5.2.2 新建数据连接。
2. 基于相应服务，新建数据库。具体请参见 3.5.2.3 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 3.5.2.4 （可选）新建数据库模式。
4. 新建数据表。具体请参见 3.5.2.5 新建数据表。

3.5.2.2 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 3.2.2 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3-367 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。

3.5.2.3 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 3.5.2.2 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3-368 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 3-178 所示的参数。

表3-178 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。

3.5.2.4（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 3.5.2.2 新建数据连接。
- 已新建 DWS 数据库，请参见 3.5.2.3 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3-369 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 3-179 所示的参数。

表3-179 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。

3.5.2.5 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 3.5.2.2 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3-370 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 3-180 配置相关参数。

表3-180 基本属性

数据连接类型	参数说明
DLI	请见表 3-184 的“基本属性”部分
DWS	请见表 3-185 的“基本属性”部分
MRS Hive	请见表 3-186 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 3-181 所示的参数。

表3-181 表结构

数据连接类型	参数说明
DLI	请见表 3-184 的“表结构”部分
DWS	请见表 3-185 的“表结构”部分
MRS Hive	请见表 3-186 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3-371 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 3-182 所示的参数，并在下方的编辑器中输入 SQL 语句。

表3-182 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 3-183 所示的数据表信息。

表3-183 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3-184 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

参数	是否必选	说明
		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3-185 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

参数	是否必选	说明
		“_”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。
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。

参数	是否必选	说明
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3-186 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

参数	是否必选	说明
		• 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。

3.5.3 脚本开发

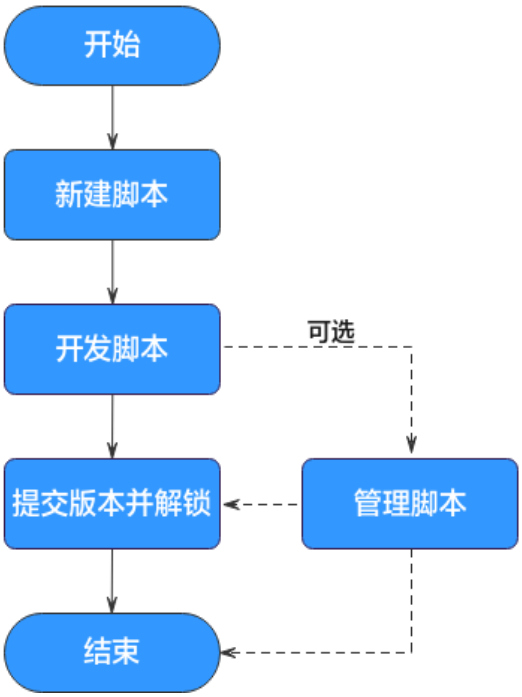
3.5.3.1 脚本开发流程

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3-372 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 3.5.3.2 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 3.5.3.3 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 3.5.3.4 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 3.5.3.5 （可选）管理脚本。

3.5.3.2 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 3.5.2.2 新建数据连接和 3.5.2.3 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3-373 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 3-187 所示的参数。

表3-187 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 3.5.3.3.1 开发 SQL 脚本、3.5.3.3.2 开发 Shell 脚本、3.5.3.3.3 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。

3.5.3.3 开发脚本

3.5.3.3.1 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 3.5.4.3 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 3.5.2.2 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3-374 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 3-188 所示的属性。创建 Flink SQL 脚本时请跳过此步骤。

表3-188 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）

属性	说明
	• 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 3.5.10.1 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3-375 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 3-189 所示的参数。

表3-189 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 3-190。

说明

Flink SQL 脚本、RDS SQL 脚本、Shell 脚本的执行结果，不支持转储。

表3-190 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。

参数	是否必选	说明
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。

3.5.3.3.2 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 3.5.4.3 开发作业）。

前提条件

- 已新增 Shell 脚本，请参见 3.5.3.2 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 3-14。
- 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3-376 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 3-191 所示的属性。

表3-191 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	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

参数	说明	示例

		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 3-192 所示的参数。

表3-192 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。

3.5.3.3.3 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 3.5.4.3 开发作业）。

前提条件

- 已新增 Python 脚本，请参见 3.5.3.2 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 3-14。
- 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3-377 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 3-193 所示的参数。

表3-193 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。

参数	是否必选	说明
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。

3.5.3.4 提交版本并解锁

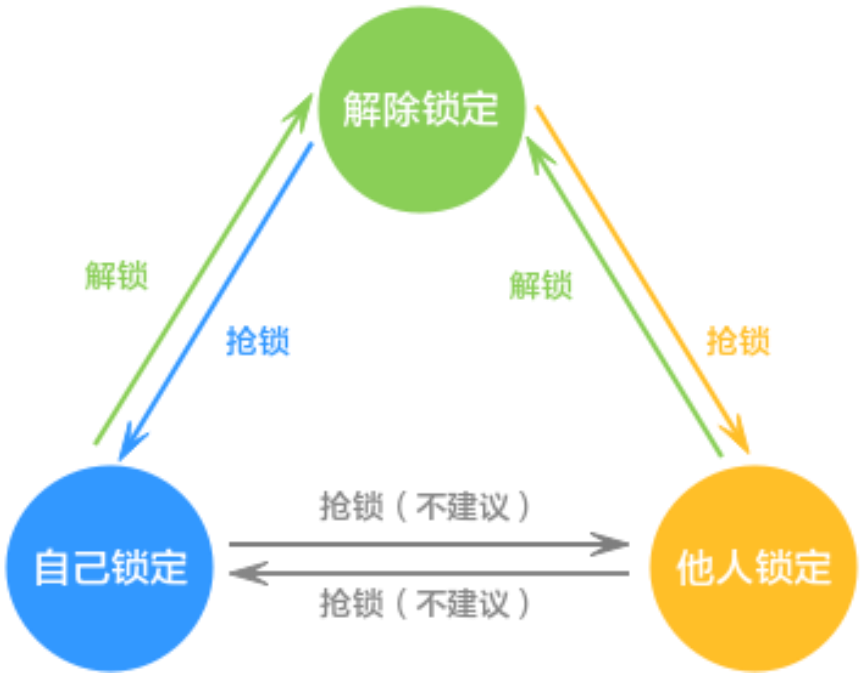
提交版本并解锁，涉及到数据开发的版本管理和编辑锁定功能。

- 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 3.5.8.1.5 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3-378 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。

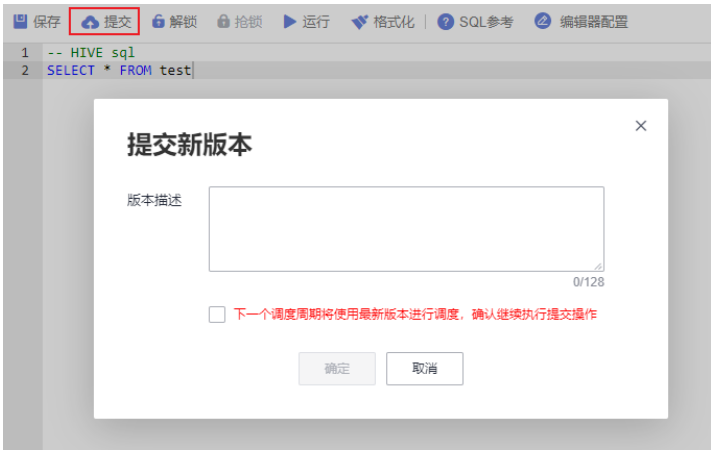
图3-379 选择数据开发

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3-380 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3-381 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3-382 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3-383 版本回滚

版本对比

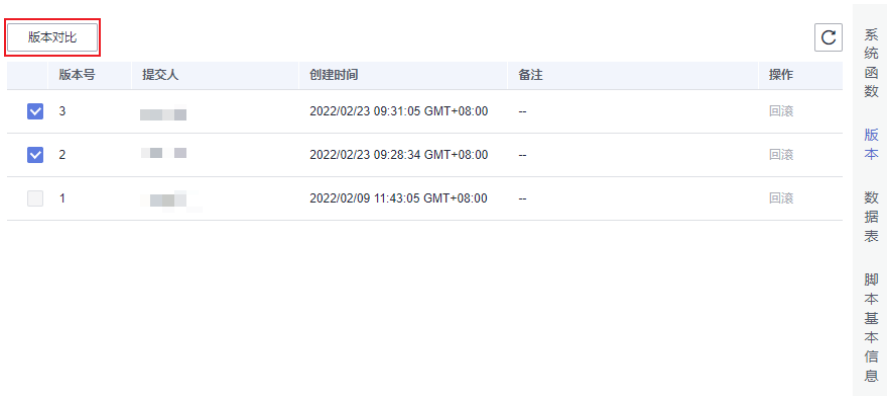
支持对比两个不同版本的脚本内容。如果只勾选一个版本，则对比该版本和开发态的脚本内容；如果勾选两个版本，则对比选中的两个版本的脚本内容。

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3-384 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3-385 版本对比详情

3.5.3.5（可选）管理脚本

3.5.3.5.1 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 3.5.3.3 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3-386 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 3-194 所示的参数。

表3-194 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。

3.5.3.5.2 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 3.5.3.3 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3-387 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3-388 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3-389 重命名脚本名称

表3-195 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。

3.5.3.5.3 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 3.5.3.3 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3-390 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，配置如表 3-196 所示的参数。

表3-196 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。

3.5.3.5.4 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3-391 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3-392 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3-393 导入脚本

2. 单击“下一步”，根据提示导入脚本。

3.5.3.5.5 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 3.5.3.3 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3-394 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。

3.5.3.5.6 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3-395 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3-396 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。

3.5.3.5.7 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3-397 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”。

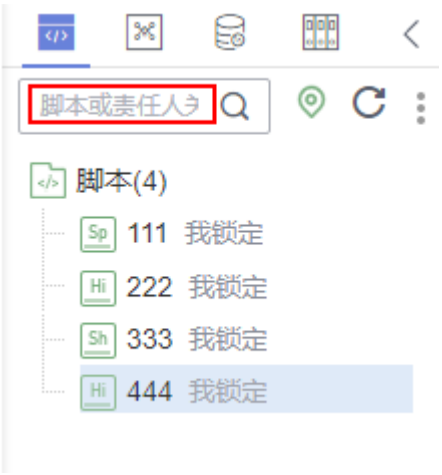
图3-398 责任人配置

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3-399 根据脚本责任人筛选脚本



3.5.3.5.8 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3-400 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3-401 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3-402 批量解锁

3.5.4 作业开发

3.5.4.1 作业开发流程

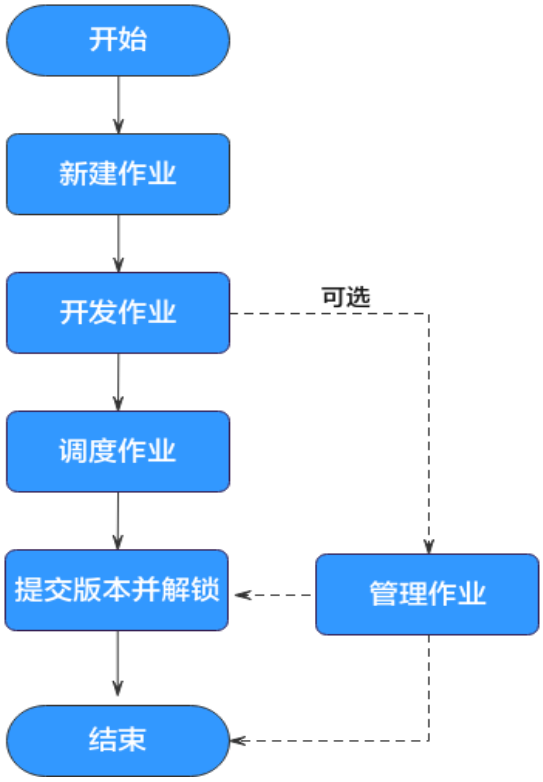
作业开发功能提供如下能力：

- 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 3-403 了解数据开发模块作业开发的基本流程。

图3-403 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 3.5.4.2 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 3.5.4.3 开发作业。
3. 调度作业：配置作业调度任务。具体请参见 3.5.4.4 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 3.5.4.5 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 3.5.4.6 （可选）管理作业。

3.5.4.2 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3-404 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 3-197 所示的参数。

表3-197 作业目录参数

参数	说明

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。

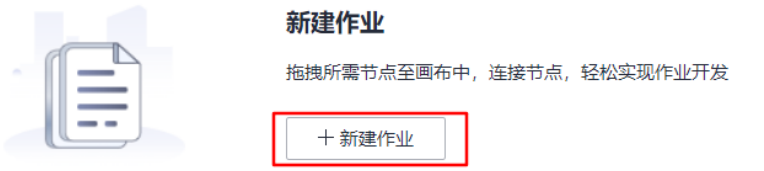
1. 登录 DataArts Studio 控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图3-405 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：

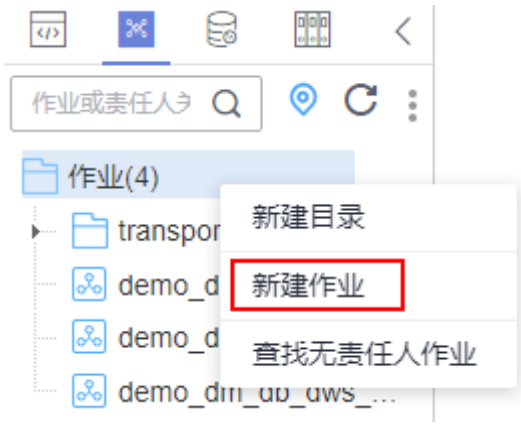
   方式一：在“作业开发”界面中，单击“新建作业”。

图3-406 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3-407 新建作业（方式二）



4. 在弹出的“新建作业”页面，配置如表 3-198 所示的参数。

表3-198 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。

参数	说明
	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。

3.5.4.3 开发作业

对已新建的作业进行开发和配置。

前提条件

- 已 3.5.4.2 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3-408 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3-409 编排作业

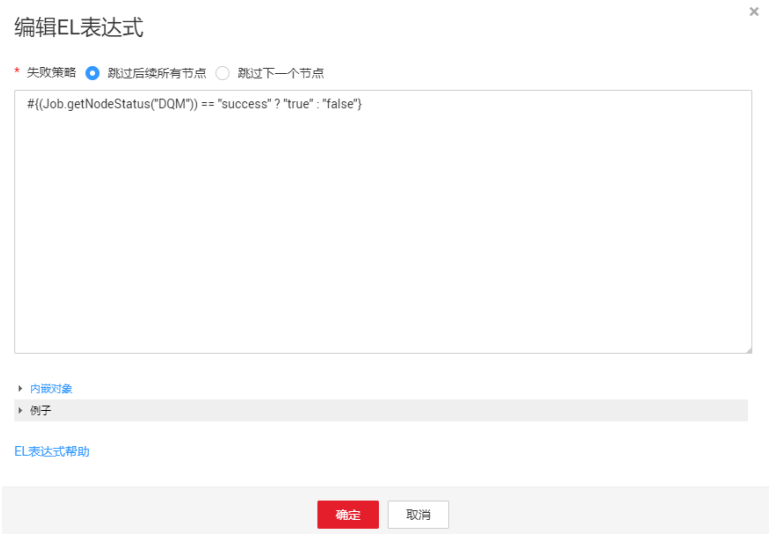
5. 配置节点功能。右键单击画布中的节点图标，根据实际需要选择如表 3-199 所示的功能。

表3-199 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。
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 3.5.10.1 表达式概述。

7. 请参见 3.5.9.1 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 3.5.9.1 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 3-200 所示的参数。

表3-200 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 3.5.8.1.3 管理

参数	说明
	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 3-201 所示的参数。

表3-201 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：${参数名称}
修改	在参数名和参数值的文本框中直接修改。
掩码显示	在参数值为密钥等情况下，从安全角度，请单击  将参数值掩码显示。
删除	在参数值文本框后方，单击，删除作业参数。
常量	
新增	单击“新增”，在文本框中填写作业常量的名称和参数值。 • 参数名称 名称只能包含字符：英文字母、数字、中划线和下划线。 • 参数值 – 字符串类的参数直接填写字符串，例如：str1 – 数值类的参数直接填写数值或运算表达式。 参数配置完成后，在作业中的引用格式为：${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。

功能	说明
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

3.5.4.4 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 3.5.4.3 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 3-202 所示的参数。

表3-202 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 3-203。 • 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 3-204。
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3-203“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。
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 3.5.11.1 作业依赖详解。

参数	说明
	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。
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 3.5.8.1.5 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

参数	说明
	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3-204“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。
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 3-205 所示的参数。

表3-205 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。

参数	说明
	• 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）
“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	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

参数	说明
隔	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

3.5.4.5 提交版本并解锁

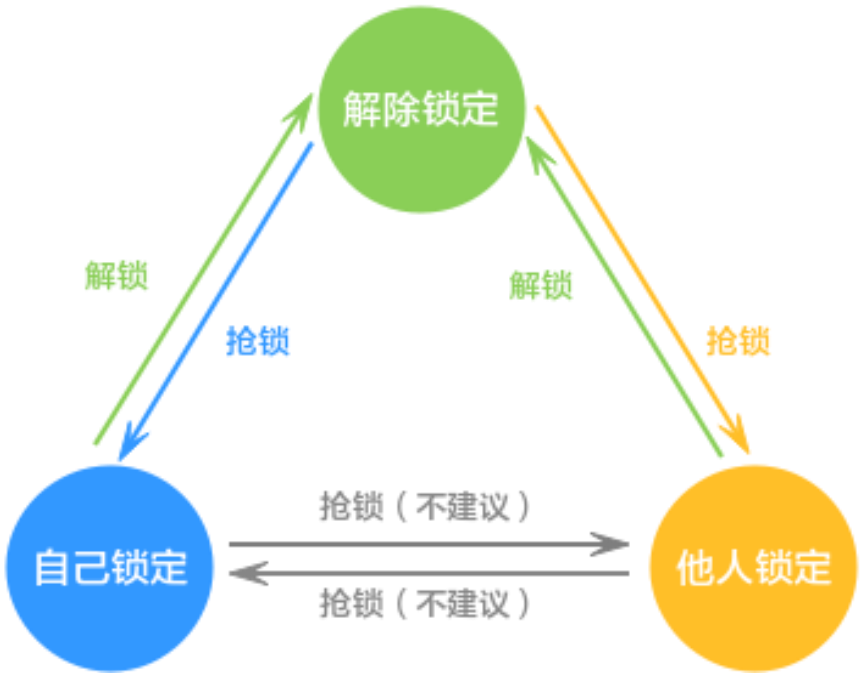
提交版本并解锁，涉及到数据开发的版本管理和编辑锁定功能。

- 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 3.5.8.1.5 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3-410 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3-411 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3-412 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3-413 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3-414 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3-415 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3-416 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3-417 版本详情查看

版本对比

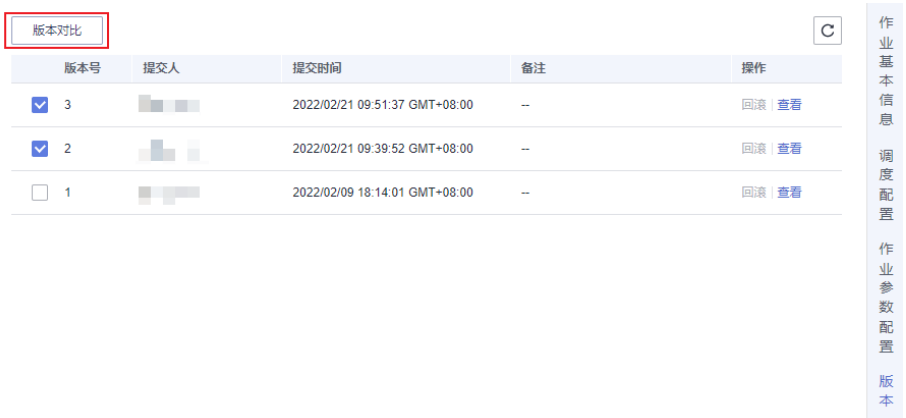
1. 登录 DataArts Studio 控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图3-418 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3-419 对比版本操作界面



3.5.4.6（可选）管理作业

3.5.4.6.1 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 3.5.4.3 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3-420 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 3-206 所示的参数。

表3-206 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。

3.5.4.6.2 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 3.5.4.3 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3-421 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3-422 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3-207 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。

3.5.4.6.3 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 3.5.4.3 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3-423 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3-208 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。

3.5.4.6.4 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3-424 导出作业（方式一）

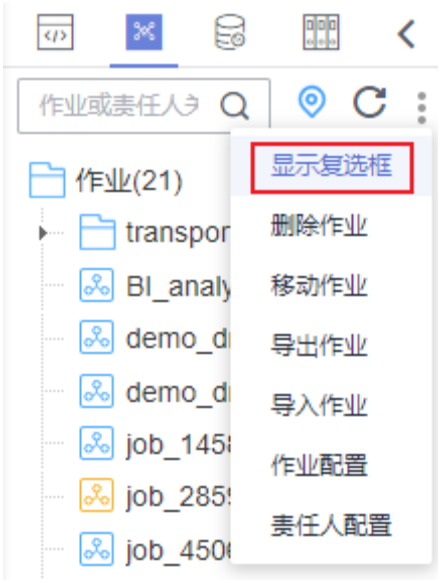
步骤 2 单击“确定”，导出所需的作业文件。

----结束

方式二：在作业目录中导出一个或多个作业

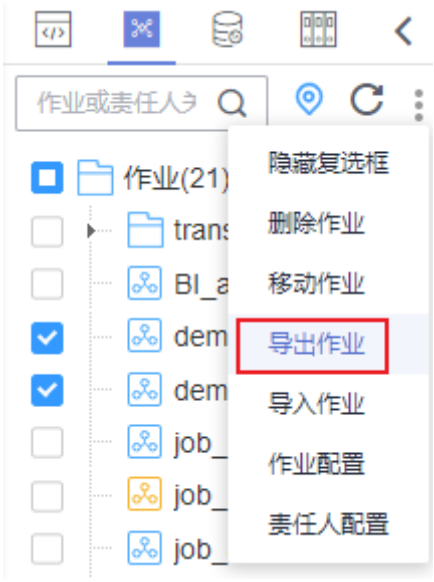
步骤 1 单击作业目录中的，选择“显示复选框”。

图3-425 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3-426 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3-427 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3-428 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”。

3.5.4.6.5 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3-429 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3-430 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。

3.5.4.6.6 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3-431 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”。

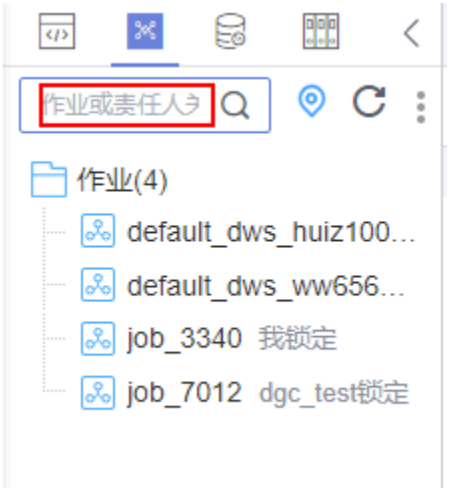
图3-432 责任人配置

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3-433 根据作业责任人筛选作业



3.5.4.6.7 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。

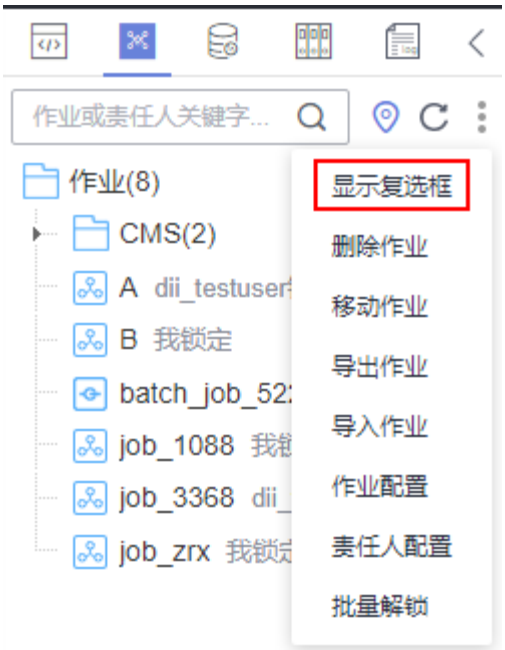
操作步骤

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3-434 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3-435 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3-436 批量解锁

3.5.5 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3-437 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 3-209 所示的参数。

表3-209 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。

3.5.6 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 3.5.8.1.2 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3-438 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3-439 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。

3.5.7 运维调度

3.5.7.1 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。

3.5.7.2 作业监控

3.5.7.2.1 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 3-210 所示的操作。

图3-440 批作业监控

表3-210 批作业监控支持的操作项

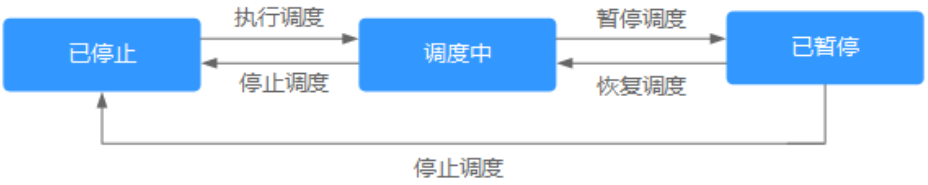
序号	支持的操作项	说明
1	根据“作业名”或“责任人名”搜索作业	-
2	根据“作业是否配置通知”、“调度状态”、“作业标签”或“下次计划时间”范围，筛选作业	-
3	批量配置作业	通过勾选作业名称前的复选框，支持批量执行操作。
4	查看作业实例状态	单击作业名称前方的，显示“最近的实例”页面，查看该作业最近的实例信息。
5	查看作业的节点信息	单击作业名称，在打开的页面中点击作业节点，查看该节点的相关关联作业/脚本与监控信息。
6	调度作业相关	在作业的“操作”列，支持执行调度、暂停调度、恢复调度、停止调度、调度配置等，详情请参见批作业监控：调度作业。
7	通知设置	在作业的“操作”列，选择“更多 > 通知设置”，弹出“新建通知”页面，参考表 3-220 配置通知参数。
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。

序号	支持的操作项	说明
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3-441 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。

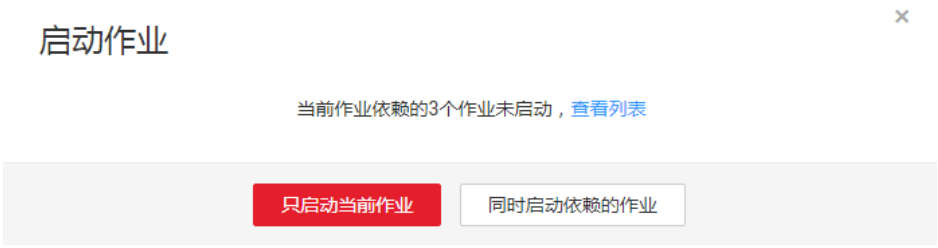
图3-442 选择数据开发

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3-443 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 3.5.7.4 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3-444 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 3-211 所示的参数。

图3-445 补数据参数

表3-211 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

参数	说明
	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3-446 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3-447 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。

3.5.7.2.2 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 3-212 所示的操作。

图3-448 实时作业监控

表3-212 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

序号	支持的操作项	说明
		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3-449 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 3-213 所示的参数。

表3-213 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3-450 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 3-214 所示的参数。

图3-451 调度配置

表3-214 调度配策略参数说明

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。
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

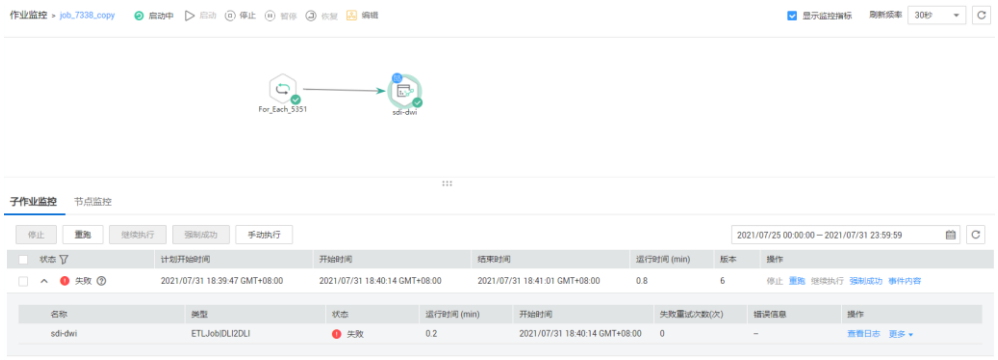
当用户配置的作业中某个节点配置有事件调度时，单击此节点可以查询子作业监控。在“子作业监控”页面可以对子作业设置停止、重跑、继续执行、强制成功、查看事件内容等操作。

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3-452 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。

图3-453 子作业监控



在“子作业监控”页面的“操作”列，提供如表 3-215 所示的操作。

表3-215 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。
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 3-216 所示的操作。

表3-216 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。

操作项	说明
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。

3.5.7.3 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3-454 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 3-217。

   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 3-217 所示的操作。

表3-217 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。
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。

操作项	说明
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 3-218 所示的操作。

表3-218 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3-455 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3-456 设置作业重跑

表3-219 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。

参数	说明
重跑作业实例列表	选择需要重跑的上下游作业，支持多选。
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。

3.5.7.4 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。

3.5.7.5 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。

3.5.7.5.1 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3-457 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 3-220 所示的参数。

表3-220 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。
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

参数	是否必选	说明
		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 3-220 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。

3.5.7.5.2 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3-458 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 3-221 所示的参数。

表3-221 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。

参数	是否必选	说明
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。

3.5.7.6 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3-459 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3-460 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3-461 还原资产

步骤 4 单击“确定”。

----结束

3.5.8 配置管理

3.5.8.1 配置

3.5.8.1.1 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3-462 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3-463 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3-464 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 3-222 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3-222 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 3.5.10.1 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

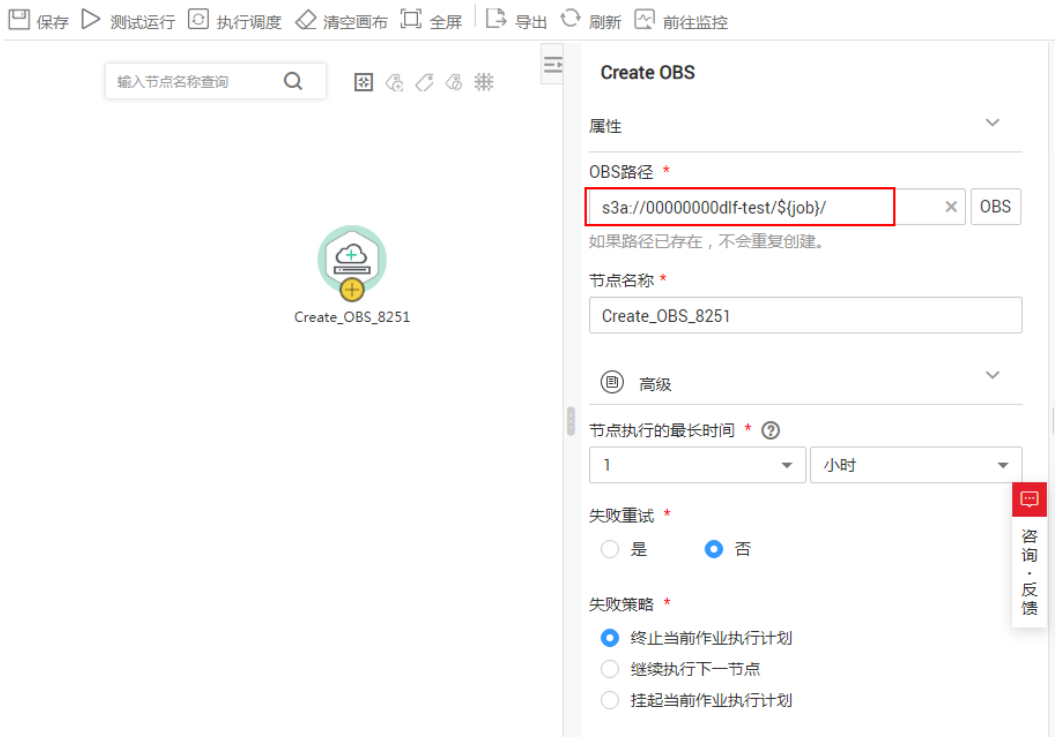
背景信息：

- 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3-465 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

3.5.8.1.2 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3-466 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3-467 配置 OBS 桶

步骤 5 单击“保存”，完成配置。

----结束

3.5.8.1.3 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3-468 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

3.5.8.1.4 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”。
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 3-469 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3-469 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 3-223 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 3-224 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3-223 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

节点名称	系统权限	权限描述
		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3-224 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)
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)

节点名称	授权项
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。

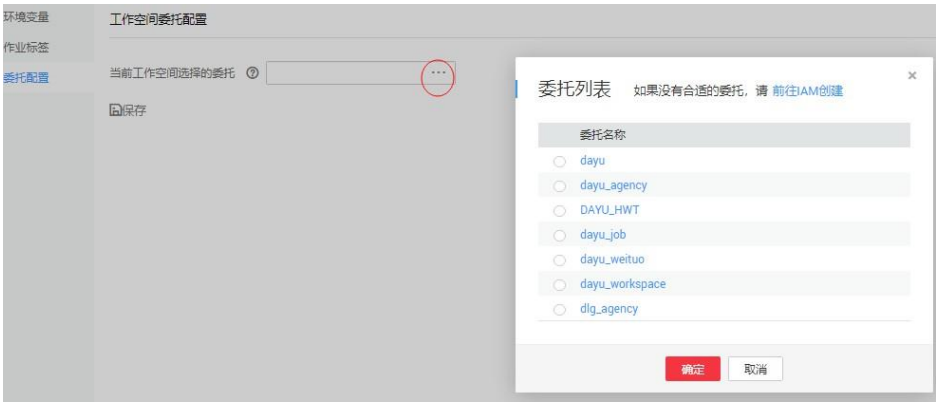
1. 登录 DataArts Studio 控制台。选择实例，点击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。

图3-470 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3-471 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3-472 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3-473 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。

3.5.8.1.5 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

3.5.8.2 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 3.5.11.8 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3-474 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 3-225 所示的参数。

表3-225 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3-475 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 3-226 所示的参数。单击“确定”，新建资源。

表3-226 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。
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

参数	是否必选	说明
		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3-476 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 3-226 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3-477 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3-478 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3-479 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3-480 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。

3.5.9 节点参考

3.5.9.1 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 3.5.10.1 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3-481 连接示意图

3.5.9.2 节点数据血缘

3.5.9.2.1 方案概述

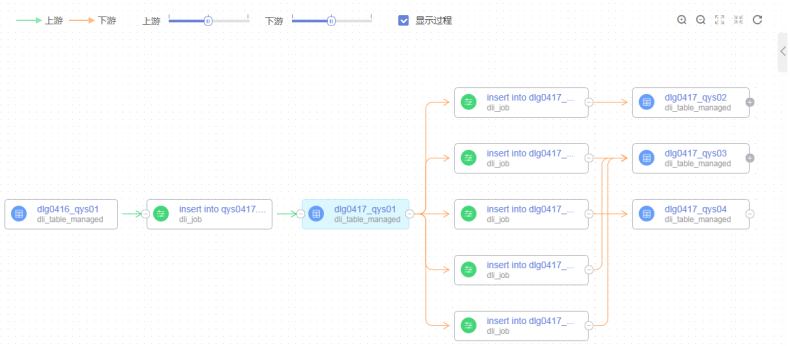
什么是数据血缘

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

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

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

图3-482 数据血缘关系示例



DataArts Studio 数据血缘实现方案

- 数据血缘的产生：

  在 DataArts Studio 平台，自动分析血缘是通过在数据开发模块中配置数据处理迁移类型的节点产生的，当前支持采集节点静态配置产生的血缘和部分节点实例上的血缘。详情请参见自动分析血缘。

  另外，DataArts Studio 平台还支持手动配置血缘方式，当用户手动配置血缘时，自动分析血缘将不生效。详情请参见手动配置血缘。

- 数据血缘的展示：

  当数据开发模块中的作业已完成血缘关系配置后，启动作业调度，并在数据目录模块进行元数据采集任务，则可以在数据目录模块可视化查看数据血缘关系。

3.5.9.2.2 配置数据血缘

在 DataArts Studio 平台，自动分析血缘是通过在数据开发模块中配置数据处理迁移类型的节点产生的，当前支持采集节点静态配置产生的血缘和部分节点实例上的血缘。详情请参见自动分析血缘。

另外，DataArts Studio 平台还支持手动配置血缘方式，当用户手动配置血缘时，自动分析血缘将不生效。详情请参见手动配置血缘。

自动分析血缘

自动分析血缘是通过在数据开发模块中配置数据处理迁移类型的节点产生的，当作业中包含如下节点时，系统支持自动解析血缘。

- **SQL 类型节点**

DataArts Studio 目前支持对 DLI SQL、DWS SQL 和 MRS Hive SQL 节点的血缘解析，可以支持多 SQL 解析及列级血缘解析，当语句中有临时表时，会自动在数据目录中创建相关的临时表实体。

- 3.5.9.10 DLI SQL
  - 支持解析 DLI 中表与表之间数据插入产生的血缘。
  - 支持通过建表语句产生的 OBS 文件到 DLI 表之间的血缘。
- 3.5.9.12 DWS SQL
  - 支持 Create table like/as 等 DDL 操作产生的 DWS 表之间的血缘。
  - 支持 Insert into 等 DML 操作产生的 DWS 表之间的血缘。
- 3.5.9.14 MRS Hive SQL
  - 支持 Create table like/as 等 DDL 操作产生的 MRS 表之间的血缘。
  - 支持 Insert into/overwrite 等 DML 操作产生的 MRS 表之间的血缘。

- **数据集成类型节点**

  目前支持对 CDM Job 节点、ETL Job 节点和 OBS Manager 节点的血缘解析。

  - 3.5.9.3 CDM Job

    支持 MRS Hive、DLI、DWS、RDS、OBS 以及 CSS 之间表文件迁移所产生的血缘。
  - 3.5.9.23 ETL Job

    支持 DLI、OBS、MySQL 以及 DWS 之间的 ETL 任务产生的血缘。
  - 3.5.9.27 OBS Manager

    支持 OBS 之间目录和文件复制迁移产生的血缘。

说明

当前血缘解析能力，单条 sql 语句不支持 sql 中含有分号的场景。

手动配置血缘

在 DataArts Studio 数据开发中，用户也可以自己定义节点的输入、输出血缘关系。当用户手动配置血缘时，自动分析血缘将不生效。手动配置血缘不会影响作业的运行。

目前手动配置血缘时输入、输出数据源支持 DLI、DWS、Hive、CSS、OBS 和 CUSTOM。CUSTOM 即自定义类型，在手动配置血缘时，对于不支持的数据源，您可以添加为自定义类型。

支持手动配置血缘的节点类型如下所示，关于手动配置血缘的更多内容，请参见相关节点的详细介绍。

- 3.5.9.3 CDM Job
- 3.5.9.4 Rest Client
- 3.5.9.10 DLI SQL
- 3.5.9.11 DLI Spark
- 3.5.9.12 DWS SQL
- 3.5.9.13 MRS Spark SQL
- 3.5.9.14 MRS Hive SQL
- 3.5.9.15 MRS Presto SQL

- 3.5.9.16 MRS Spark
- 3.5.9.17 MRS Spark Python
- 3.5.9.23 ETL Job
- 3.5.9.27 OBS Manager

3.5.9.2.3 查看数据血缘

当数据开发模块中的作业已完成血缘关系配置后，启动作业调度，并在数据目录模块进行元数据采集任务，则可以在数据目录模块可视化查看数据血缘关系。

前提条件

已完成血缘关系的自动配置或手动配置，请参见 3.5.9.2.2 配置数据血缘。

启动作业调度

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3-483 选择数据开发

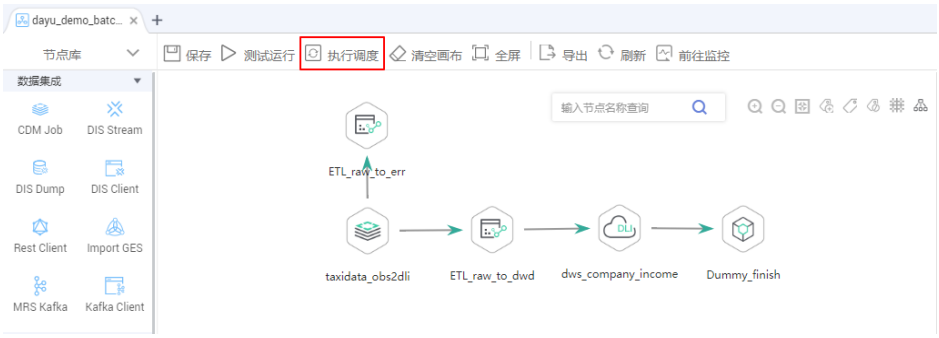
步骤 2 在数据开发控制台，单击左侧导航栏中的作业开发按钮，进入作业开发页面后，打开已完成血缘配置的作业。

步骤 3 在数据开发中，当作业进行“执行调度”时，系统开始解析血缘关系。

说明

测试运行不会解析血缘。

图3-484 作业调度



----结束

新建元数据采集任务

如果已创建元数据采集任务，此操作可跳过。

步骤 1 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-485 选择数据目录

步骤 2 请参见 3.7.4.2 任务管理，新建元数据采集任务。

----结束

查看数据血缘关系

步骤 1 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-486 选择数据目录

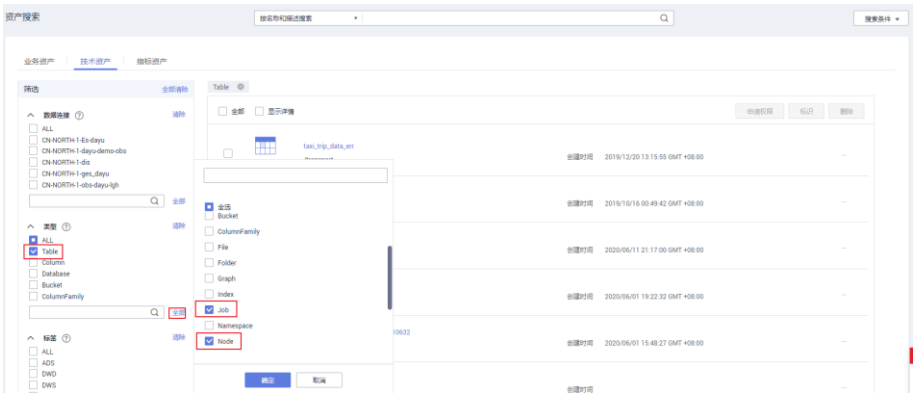
步骤 2 在“数据目录 > 技术资产”页面，可以对数据开发的作业、节点、表进行查询。

在“类型”筛选区域，单击“全部”按钮并勾选“Job”、“Node”和“Table”类型，然后单击“确定”。数据开发中的作业对应于 Job 类型，节点对应于 Node 类型，表对应于 Table 类型。

说明

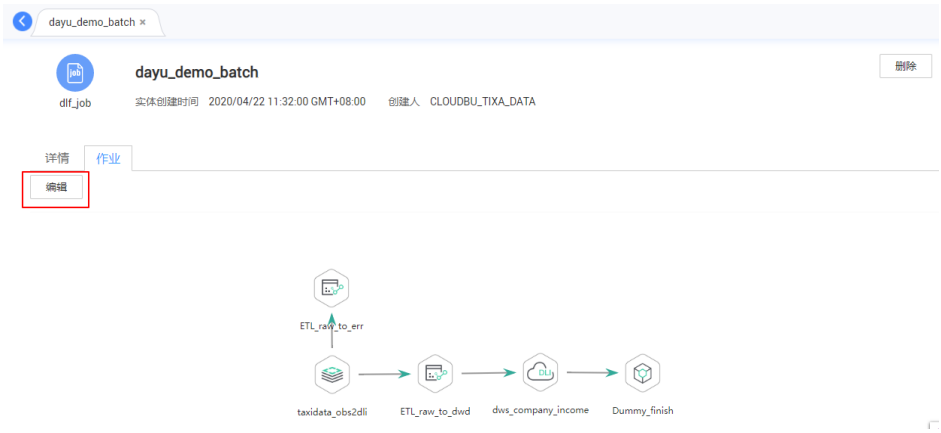
数据开发中的作业信息不属于任何一个数据连接，故如果在搜索条件中勾选数据连接，则查询不到结果。

图3-487 选择类型



步骤 3 在数据资产搜索结果中，类型名称末尾带“_job”的数据资产为作业，单击某一作业名称，可以查看该作业的详情。在作业的详情页面进入“作业”页签，单击“编辑”可跳转到数据开发的作业编辑页面。

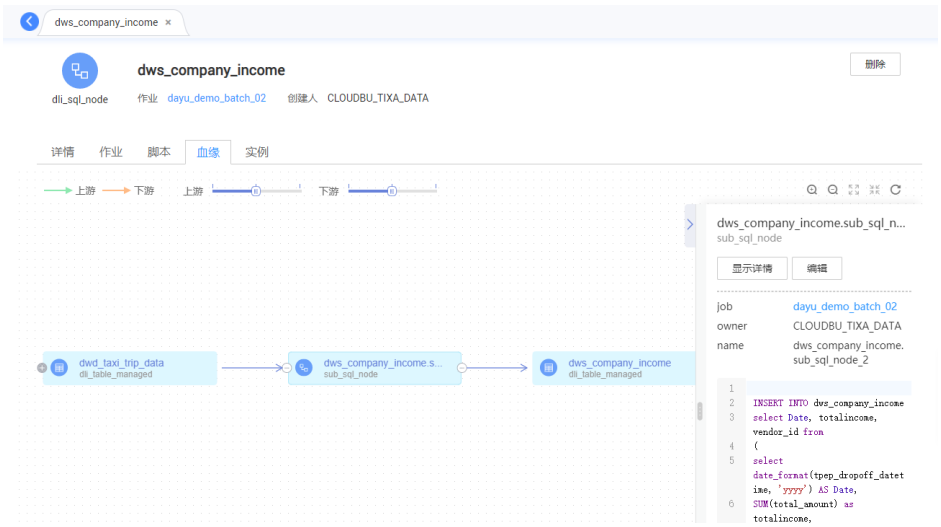
图3-488 查看作业



步骤 4 在数据资产搜索结果中，类型名称末尾带“_node”的数据资产为节点，单击某一节点名称，可以查看节点的详情。在节点（需是支持血缘的节点类型）详情页面，可以查看节点的血缘信息。

- 单击血缘图中节点左右两端“+”、“-”图标，可以进一步展开查看血缘的上下链路。
- 单击血缘图中的某一个节点，可以查看该节点的详情。
- 进入“作业”页签，单击“编辑”可跳转到数据开发的作业编辑页面。

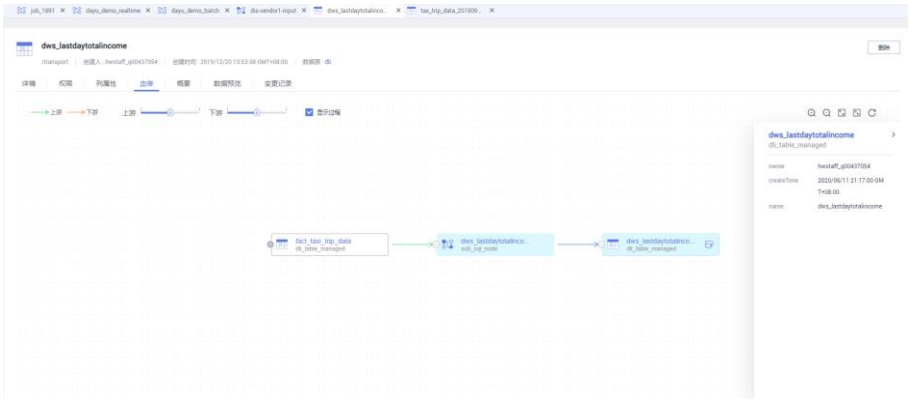
图3-489 查看节点血缘



步骤 5 在数据资产搜索结果中，图标为表格的数据资产为表，单击某一表名称，可以查看表的详情。在详情页面，可以查看表的血缘信息。

- 单击血缘图中表左右两端“+”、“-”图标，可以进一步展开查看血缘的上下链路。
- 单击血缘图中的某一个表，可以查看该表的详情。

图3-490 查看表血缘



----结束

3.5.9.3 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 3-227，表 3-228 和表 3-229 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3-227 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

参数	是否必选	说明
		到备集群，实现作业同步，具体操作方法请参见3.3.6.8 批量管理作业。
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3-228 高级参数

参数	是否必选	说明

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。 • 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3-229 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

参数	说明
	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。
删除	单击  ，删除节点输入功能的参数配置。

参数	说明
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。
确定	单击“确认”，保存节点输出功能的参数配置。

参数	说明
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。

3.5.9.4 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 弹性云主机，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 3-230，表 3-231 和表 3-232 配置 Rest Client 节点的参数。

表3-230 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
HTTP 方法	是	选择请求的类型： • GET • POST • PUT

参数	是否必选	说明
		• 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）。
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"`

参数	是否必选	说明
		} 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3-231 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）

参数	是否必选	说明
		• 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3-232 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。

参数	说明
	– 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。
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。

参数	说明
	• 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。
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。

3.5.9.5 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 3-233 和表 3-234 配置 Import GES 节点的参数。

表3-233 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
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3-234 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。

3.5.9.6 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 3-235 和表 3-236 配置 MRS Kafka 的参数。

表3-235 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3-236 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。

3.5.9.7 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 3-237 配置 Kafka Client 节点的参数。

表3-237 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。
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3-238 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

参数	是否必选	说明
		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。

3.5.9.8 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3-239 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3-240 高级参数

参数	是否必选	说明

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。

3.5.9.9 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 3-241。
- **Flink SQL 作业**：请参见表 3-242。
- **Flink 自定义作业**：请参见表 3-243。

- 表 3-244

表3-241 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3-242 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 3.5.3.2 新建脚本和 3.5.3.3.1 开发 SQL 脚本创建和开发 Flink SQL 脚本。
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。

参数	是否必选	说明
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3-243 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。
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

参数	是否必选	说明
		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3-244 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。

参数	是否必选	说明
		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。

3.5.9.10 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 3-245，表 3-246 和表 3-247 配置 DLI SQL 节点的参数。

表3-245 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

参数	是否必选	说明
		本未创建，请参考 3.5.3.2 新建脚本和 3.5.3.3.1 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 3.5.10.1 表达式概述。 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3-246 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。

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3-247 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

参数	说明
	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。
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

参数	说明
	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。
删除	单击，删除节点输出功能的参数配置。

参数	说明
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。

3.5.9.11 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 3-248，表 3-249 和表 3-250 配置 DLI Spark 节点的参数。

表3-248 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。
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。

参数	是否必选	说明
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。
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

参数	是否必选	说明
		• 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3-249 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

参数	是否必选	说明
		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3-250 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

参数	说明
	– 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。
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

参数	说明
	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。
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。

3.5.9.12 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 3.5.11.6 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 3-251，表 3-252 和表 3-253 配置 DWS SQL 节点的参数。

表3-251 属性参数

参数	是否必选	说明

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 3.5.3.2 新建脚本和 3.5.3.3.1 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。
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 3.5.10.1 表达式概述。 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3-252 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

参数	是否必选	说明
		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3-253 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。

参数	说明
	• 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。
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。

参数	说明
	– 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。
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。

3.5.9.13 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 3-254，表 3-255 和表 3-256 配置 MRS Spark SQL 节点的参数。

表3-254 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 3.5.3.2 新建脚本和 3.5.3.3.1 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 3.5.10.1 表达式概述。 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 ”。
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 节点名称不得包含中文字符、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3-255 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3-256 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

参数	说明
	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。
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

参数	说明
	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。

3.5.9.14 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 3-257，表 3-258 和表 3-259 配置 MRS Hive SQL 节点的参数。

表3-257 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 3.5.3.2 新建脚本和 3.5.3.3.1 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 3.5.10.1 表达式概述。 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

参数	是否必选	说明
		且长度为 1～64 个字符。 说明 节点名称不得包含中文字符、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3-258 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3-259 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	
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

参数	说明
	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。
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。

3.5.9.15 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 3-260，表 3-261 和表 3-262 配置 MRS Presto SQL 节点的参数。

表3-260 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 3.5.3.2 新建脚本和 3.5.3.3.1 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。

参数	是否必选	说明
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 3.5.10.1 表达式概述。 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。 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 节点名称不得包含中文字符、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3-261 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”。

参数	是否必选	说明
		• 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3-262 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。

参数	说明
	– 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。
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。

参数	说明
	– 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。
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。

3.5.9.16 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 3-263，表 3-264 和表 3-265 配置 MRS Spark 节点的参数。

表3-263 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

参数	是否必选	说明
		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 作业名称不得包含中文字符、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。
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3-264 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。

表3-265 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

参数	说明
	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。
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。

参数	说明
	– 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。
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。

3.5.9.17 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 3.5.11.10 开发一个 MRS Spark Python 作业。

参数

用户可参考表 3-266，表 3-267 和表 3-268 配置 MRS Spark Python 节点的参数。

表3-266 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 作业名称不得包含中文字符、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3-267 高级参数

参数	是否必选	说明

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3-268 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

参数	说明
	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。
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

参数	说明
	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。
删除	单击，删除节点输出功能的参数配置。

参数	说明
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。

3.5.9.18 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 3-269 和表 3-270 配置 MRS Flink 节点的参数。

表3-269 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 作业名称不得包含中文字符、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。
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。

参数	是否必选	说明
		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3-270 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。

3.5.9.19 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 3-271 和表 3-272 配置 MRS MapReduce 节点的参数。

表3-271 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 作业名称不得包含中文字符、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3-272 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	是	设置节点执行的超时时间，如果节点配置了重试，在超

参数	是否必选	说明
时间		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。

3.5.9.20 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 3-273 和表 3-274 配置 CSS 节点的参数。

表3-273 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。

参数	是否必选	说明
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。
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3-274 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

参数	是否必选	说明
		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。

3.5.9.21 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 3-275 和表 3-276 配置 Shell 节点的参数。

表3-275 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

参数	是否必选	说明
		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 3.5.3.2 新建脚本和 3.5.3.3.2 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3-276 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。

3.5.9.22 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 3-277 和表 3-278 配置 RDS SQL 节点的参数。

表3-277 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。
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 3.5.3.2 新建脚本和 3.5.3.3.1 开发

参数	是否必选	说明
		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3-278 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。

3.5.9.23 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 3-279，表 3-280 和表 3-281 配置 ETL Job 节点的参数。

表3-279 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3-280 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3-281 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

参数	说明
	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。
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。

参数	说明
	• 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。

3.5.9.24 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 3-282 和表 3-283 配置 Python 节点的参数。

表3-282 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 3.5.3.2 新建脚本和 3.5.3.3.3 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。
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3-283 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。

3.5.9.25 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 3-284 和表 3-285 配置 Create OBS 节点的参数。

表3-284 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3-285 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。

3.5.9.26 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 3-286 和表 3-287 配置 Delete OBS 节点的参数。

表3-286 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3-287 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。

3.5.9.27 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 3-288，表 3-289 和表 3-290 配置 OBS Manager 节点的参数。

表3-288 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

参数	是否必选	说明
		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3-289 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3-290 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。

参数	说明
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。
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

参数	说明
	– 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。
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。

3.5.9.28 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 3-291 和表 3-292 配置 Open/Close Resource 节点的参数。

表3-291 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

参数	是否必选	说明
		群。

表3-292 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。

3.5.9.29 Data Quality Monitor

功能

通过 Data Quality Monitor 节点可以对运行的数据进行质量监控。

参数

用户可参考表 3-293 和表 3-294 配置 Data Quality Monitor 节点的参数。

表3-293 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。
DQC 作业类型	是	数据质量作业的类型： • 质量作业 • 对账作业
质量作业名称	是	DQC 作业类型为质量作业时需要配置。选择在数据质量模块中创建的质量作业名称。
是否忽略质量作业告警	是	DQC 作业类型为质量作业时需要配置。 • 是：如果该质量作业处于告警状态时，当前节点的状态将被设置为成功，继续执行后续节点。 • 否：如果该质量作业处于告警状态时，则当前节点的状态将被设置为失败。
对账作业名称	是	DQC 作业类型为对账作业时需要配置。选择在数据质量模块中创建的对账作业名称。
是否忽略对账作业告警	是	DQC 作业类型为对账作业时需要配置。 • 是：如果该对账作业处于告警状态时，当前节点的状态将被设置为成功，继续执行后续节点。 • 否：如果该对账作业处于告警状态时，则当前节点的状态将被设置为失败。

表3-294 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

参数	是否必选	说明
		– 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。

3.5.9.30 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 3-295 和表 3-296 配置 Sub Job 节点的参数。

表3-295 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

参数	是否必选	说明
		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3-296 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。

3.5.9.31 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 3-297 配置 For Each 节点的参数。

表3-297 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']] 。
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3-298 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。

3.5.9.32 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 3-299 和表 3-300 配置 SMN 节点的参数。

表3-299 属性参数

参数	是否必选	说明

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。

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":{`

参数	是否必选	说明
		"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3-300 高级参数

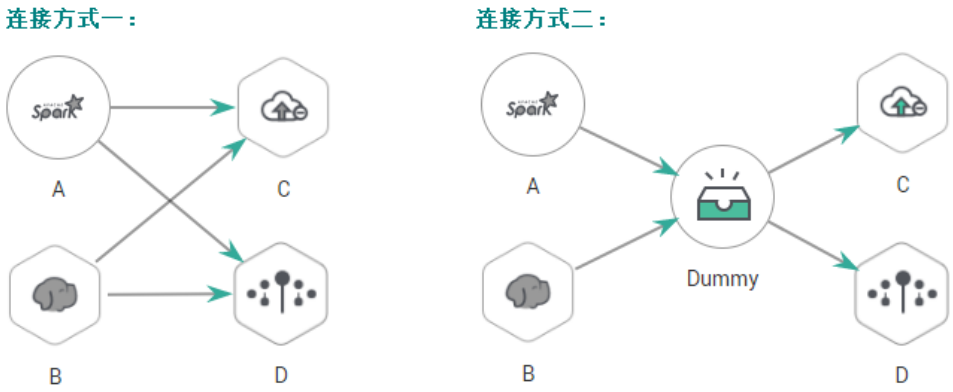
参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将不会再重试，直接置为失败状态。
失败重试	是	节点执行失败后，是否重新执行节点。 • 是：重新执行节点，请配置以下参数。 – 最大重试次数 – 重试间隔时间（秒） • 否：默认值，不重新执行节点。 说明 如果作业节点配置了重试，并且配置了超时时间，该节点执行超时后将不会再重试，直接置为失败状态。
失败策略	是	节点执行失败后的操作： • 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 • 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“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。

3.5.9.33 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 3-491 所示。

图3-491 连接方式对比



参数

用户可参考表 3-301 配置 Dummy 节点的参数。

表3-301 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。

3.5.10 EL 表达式参考

3.5.10.1 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：

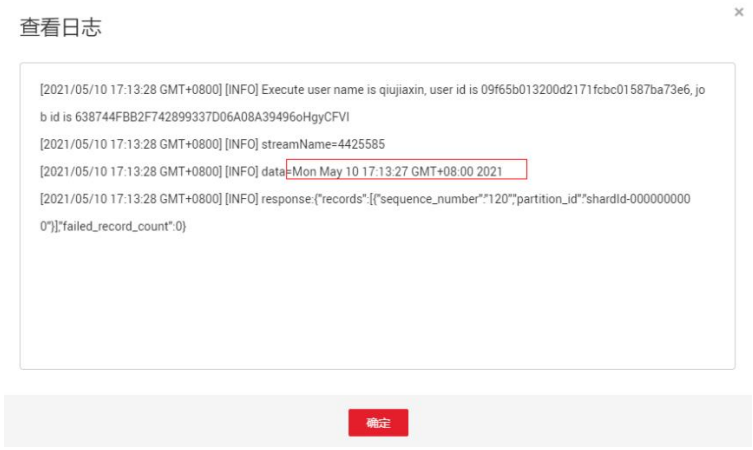
1. 获取作业中“get_cluster”节点的执行结果（“Job.getNodeOutput("get_cluster")”），执行结果是一个 JSON 字符串。
2. 通过 JSON 路径（“tables[0].table_name”），获取 JSON 字符串中字段的值。

调试方法介绍

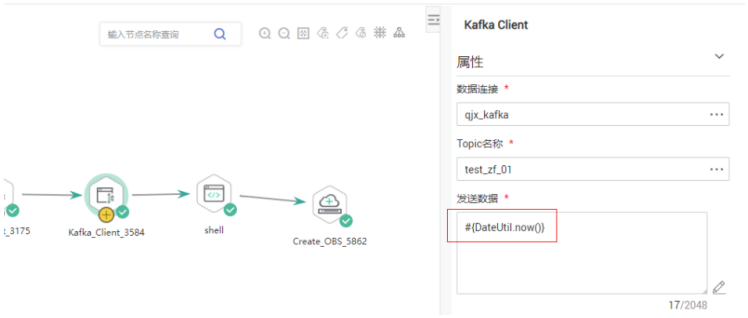
下面为您介绍几种 EL 表达式的调试方法，能够在调试过程中方便地看到替换结果。

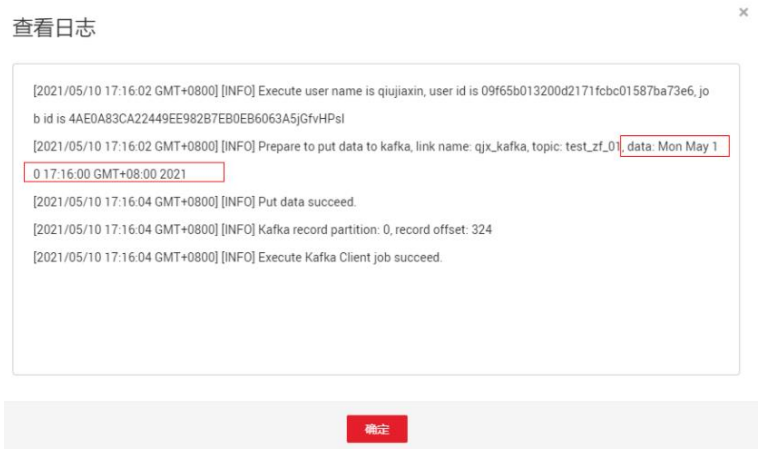
后文以#{DateUtil.now()}表达式为例进行介绍。

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。

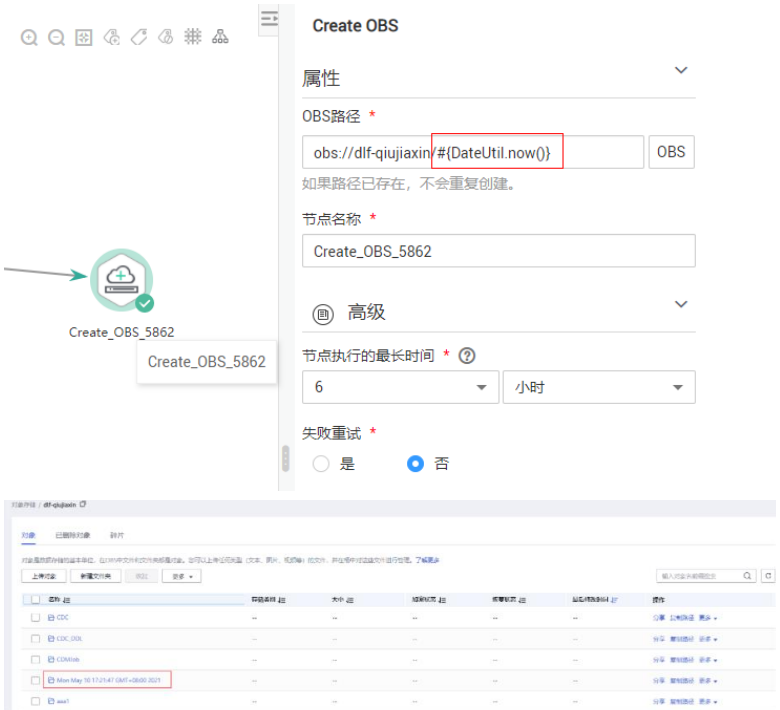




3. 使用 Shell 节点。
   - 前提：您需要具备 ECS 弹性云主机。
   - 方法：创建一个主机连接，将 EL 表达式直接通过 echo 打印出来，点击“测试运行”之后查看日志，日志中会打印出 EL 表达式的值。

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。



3.5.10.2 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3-302 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
<= 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}
```

3.5.10.3 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 3-303 所示。

表3-303 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
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")}
```

3.5.10.4 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3-304 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")}
```

3.5.10.5 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3-305 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3-306 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。

方法	描述
	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 3.5.10.3 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 3.5.10.3 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 3.5.10.3 日期和时间模式。
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")}
```

3.5.10.6 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,".")}
```

3.5.10.7 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3-307 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。
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 3.5.10.3 日期和

方法	描述
	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")}
```

3.5.10.8 JSONUtil 内嵌对象

JSONUtil 内嵌对象提供了 JSON 对象方法。

方法

表3-308 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")}
```

3.5.10.9 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3-309 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]}
```

3.5.10.10 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3-310 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")}

3.5.10.11 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 3.2.2 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 3.5.2.3 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 3.5.2.5 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

3.5.11 使用教程

3.5.11.1 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 3-492 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3-492 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。

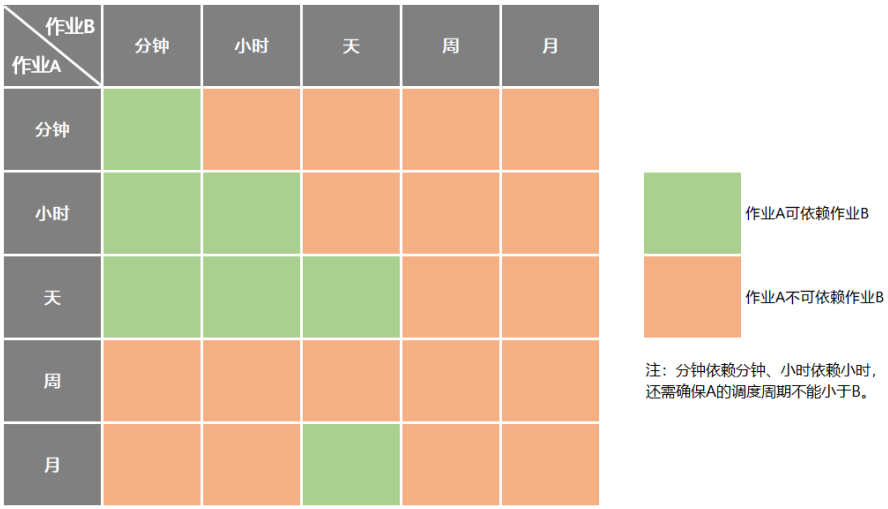
设置依赖作业的条件

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 3-493 所示。

图3-493 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3-311 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	-
...	...	...

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3-312 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3-313 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3-314 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	-

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 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
...	-	...
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
...	...	...

3.5.11.2 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。

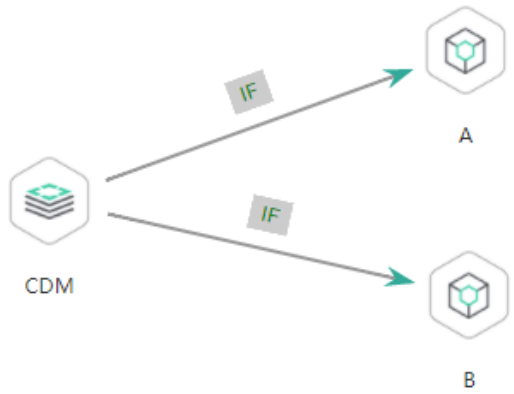
EL 表达式用法可参考 3.5.10.1 表达式概述。

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 3-494 的样例，说明如何设置 IF 条件。

图3-494 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 3-494 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。

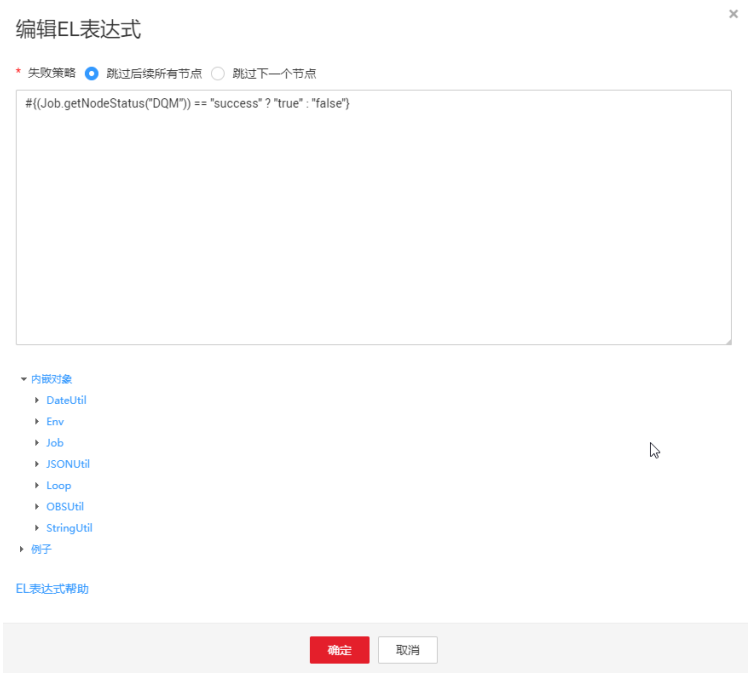
每一个条件分支都需要填写 IF 条件，IF 条件为通过 EL 表达式语法填写三元表达式。当三元表达式结果为 true 的时候，才会执行连线后面的节点，否则后续节点将被跳过。

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3-495 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 3-496 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3-496 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 3-497 所示：

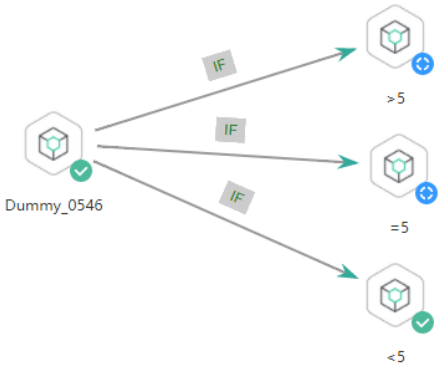
图3-497 主作业样例

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 3-498 所示。

图3-498 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 3-498 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。

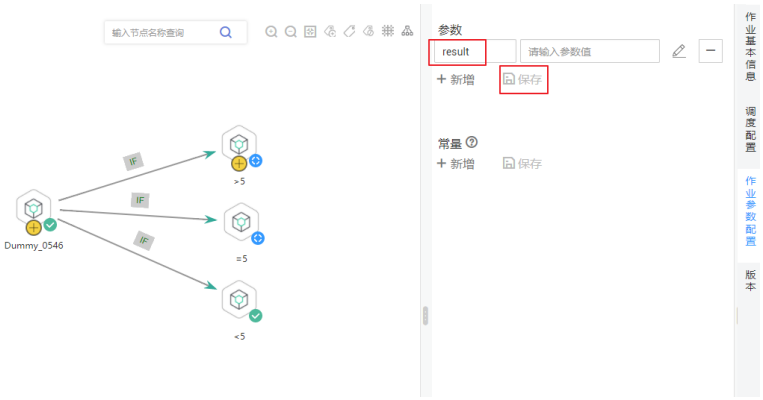
每一个条件分支都需要填写 IF 条件，IF 条件为通过 EL 表达式语法填写三元表达式。当三元表达式结果为 true 的时候，才会执行连线后面的节点，否则后续节点将被跳过。

- 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3-499 配置作业参数



步骤 6 配置完成后保存作业。

----结束

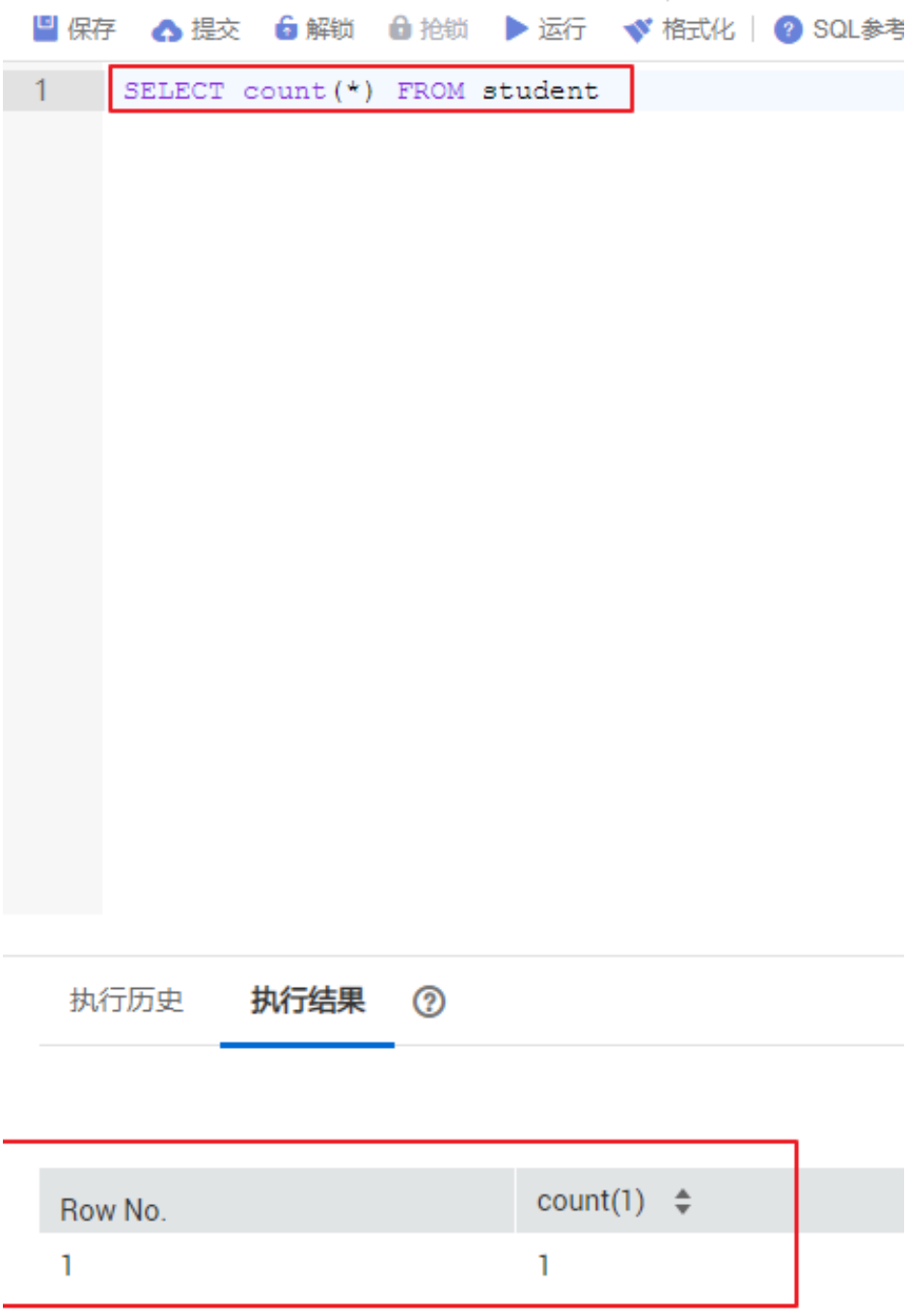
开发主作业

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 3-497 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3-500 HIVE SQL 脚本执行结果



步骤 3 配置 For Each 节点属性，如图 3-501 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3-501 For Each 节点属性

步骤 4 配置完成后保存作业。

----结束

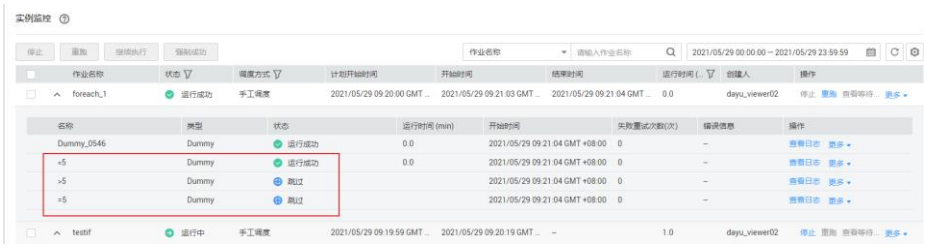
测试运行主作业

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 3-502 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3-502 子作业运行结果



----结束

多 IF 条件下当前节点的执行策略

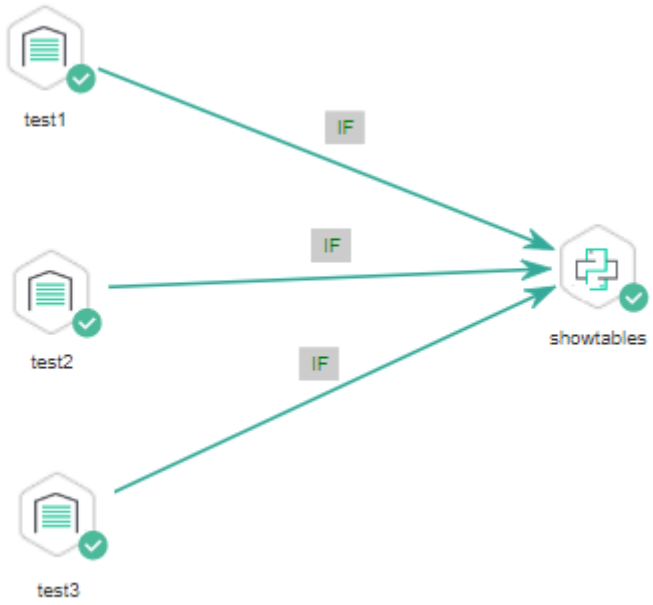
如果当前节点的执行依赖多个 IF 条件的节点，执行的策略包含逻辑或和逻辑与两种。

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3-503 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 3-503 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。

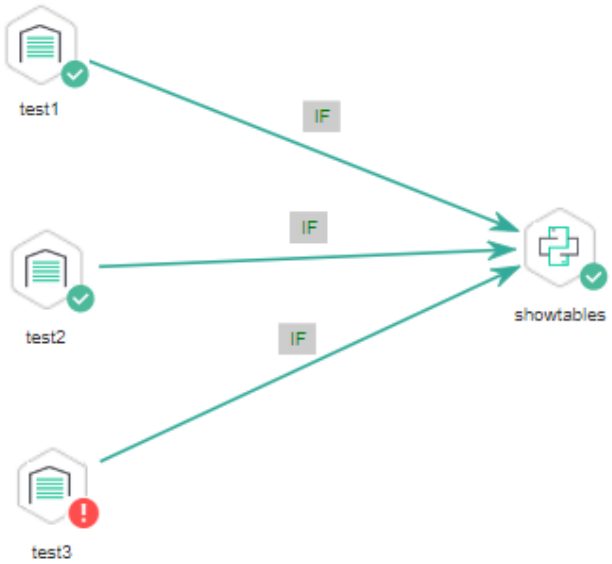
test1 运行成功，则对应的 IF 条件为 true；

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3-504 配置为“逻辑或”的作业运行情况

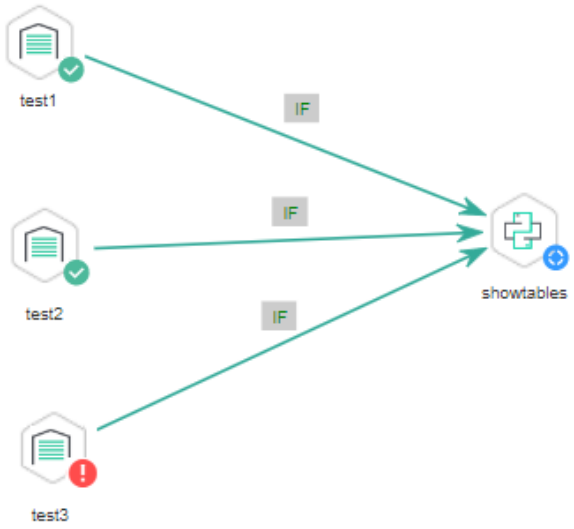


测试运行日志

[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3-505 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

3.5.11.3 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 3-506 所示，第一个 Rest Client 调用了 MRS 服务查询集群列表的 API，图 3-507 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3-506 Rest Clien 作业样例 1

图3-507 JSON 消息体


```
{
    "clusterTotal":31,
    "clusters":[
        {
            "clusterId":"6ea1b5c2-6526-4ef8-9c8f-4105b63fa893",
            "clusterName":"mrs_hbase22",
            "totalNodeNum":"2",
            "clusterState":"running",
            "stageDesc":null,
            "createAt":"1620378935",
            "updateAt":"1620611307",
            "chargingStartTime":"1620380067",
            "billingType":"Metered",
            "dataCenter":"cn-north-7",
            "vpc":"vpc-dlf",
            "vpcId":"f35aee01-c4a3-47c1-8d92-9df430537de4",
            "duration":"0",
            "fee":"0.0",
            "hadoopVersion":"",
            "componentList":[
                {
                    "id":"218051",
                    "componentId":"MRS 2.1.0_001",
                    "componentName":"Hadoop",
                    "componentVersion":"3.1.1",
                    "external_datasources":null,
                    "componentDesc":"A distributed data storage and processing framework for large da
ta sets, including core components such as HDFS, YARN, and MapReduce.",
                    "componentDescEn":null,
                    "multi_service_name":null
                },
```

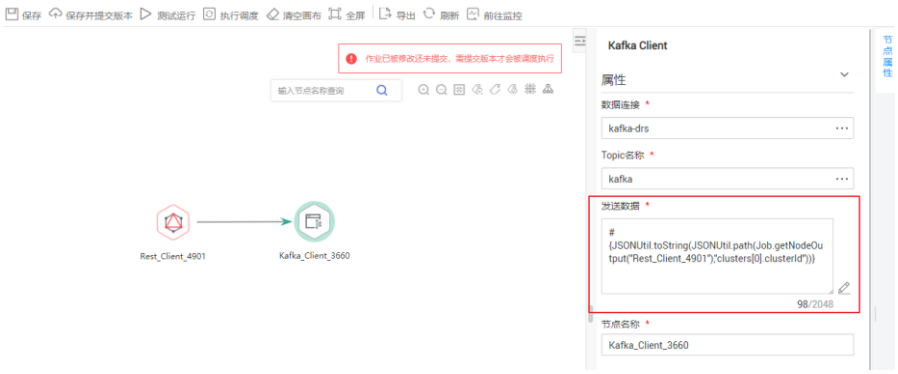
通过 EL 表达式获取返回值

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 3.5.10.1 表达式概述。

如图 3-508 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3-508 Rest Client 作业样例 2



3.5.11.4 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 3-509 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3-509 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 3.5.10.1 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 3-315 所示。

表3-315 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

源数据表名	目的表名
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）服务控制台中的 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）服务控制台中的 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 3-315 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 3-315 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。DLI 创建外表请参见，DWS 创建外表请参见。

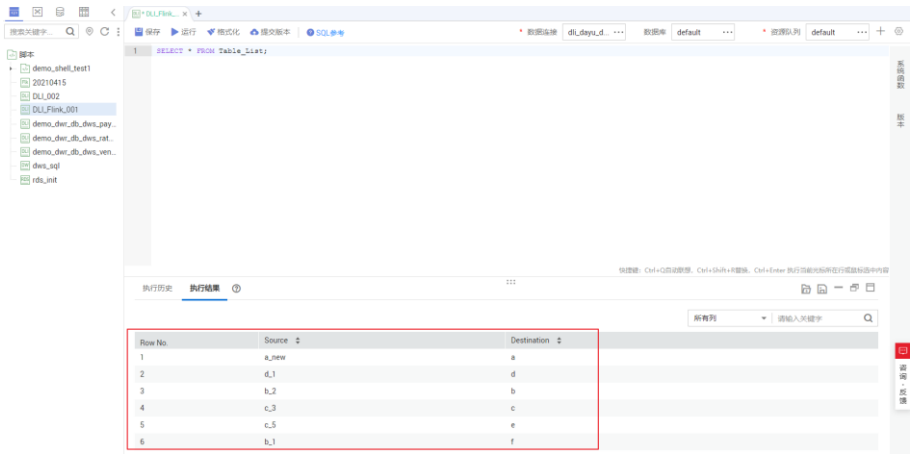
3. 您可以将表 3-315 数据保存在 HDFS 的 CSV 文件中，然后通过 HIVE SQL 创建 Hive 外表关联这个 CSV 文件，然后将 HIVE 外表查询的结果作为数据集。DLI 创建外表请参见。

本例以方式 1 进行说明，将表 3-315 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）服务控制台中的 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3-510 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 3-511 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：

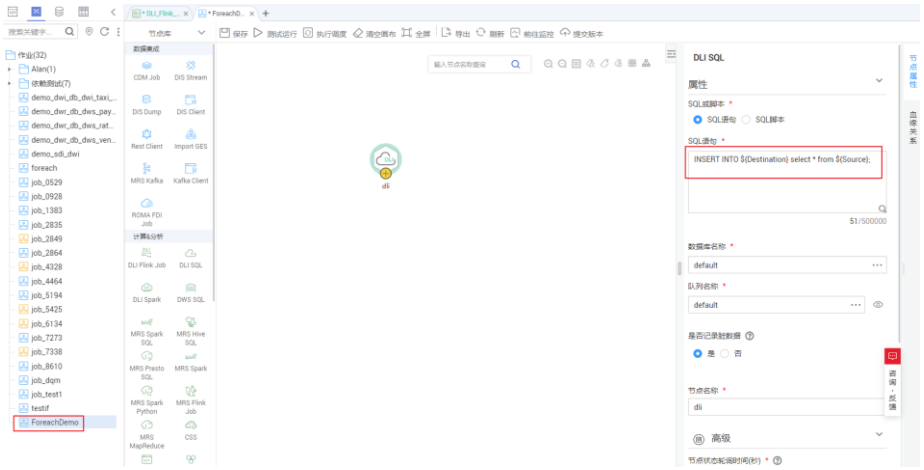
   ```
   INSERT INTO ${Destination} select * from ${Source};
   ```

   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。

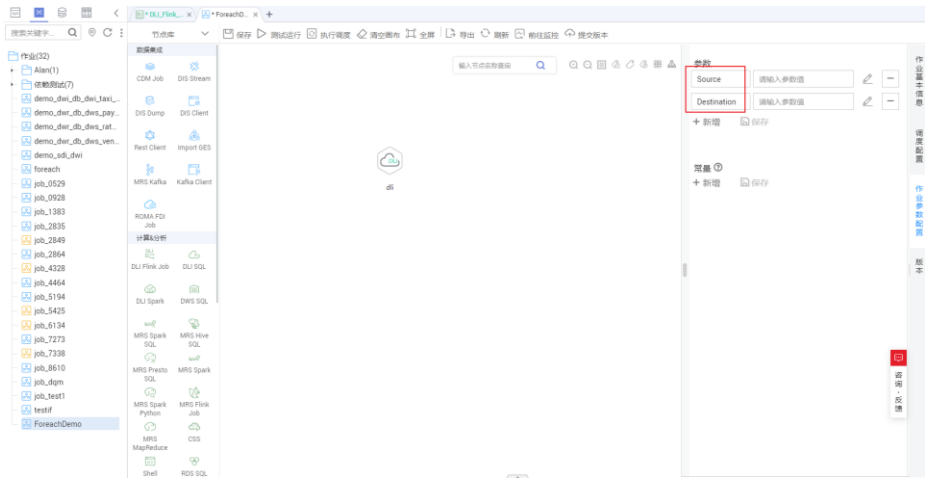
   而表达式${job_param_name}，既可以获取到父作业传递过来的参数值，也可以获取到全局配置的变量。

图3-511 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3-512 配置子作业参数

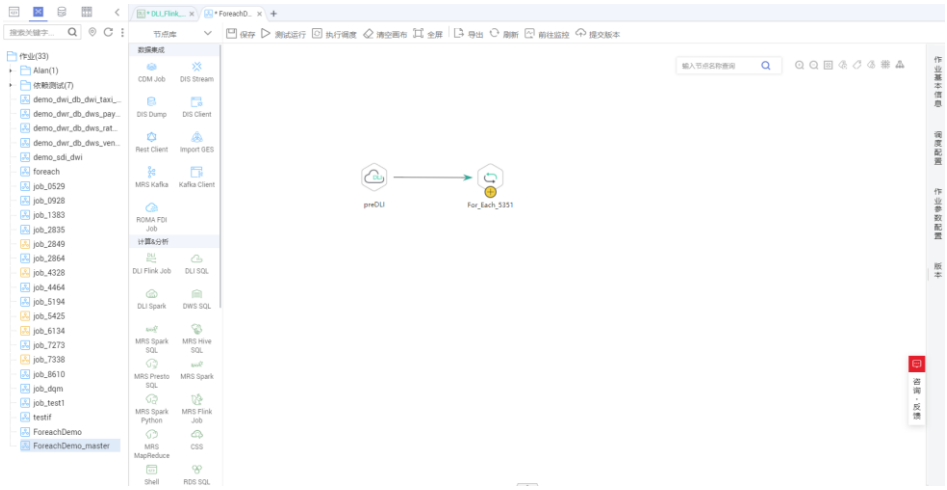


3. 配置完成后保存作业。

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 3-513 所示的作业。

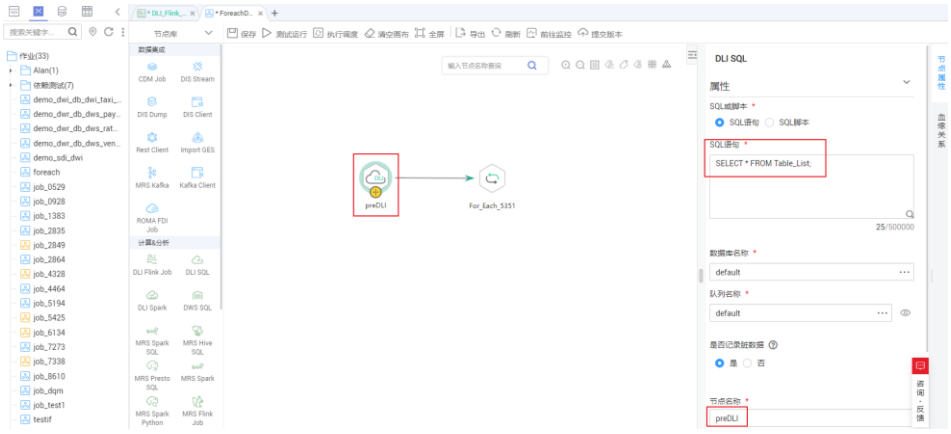
图3-513 编排作业



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。

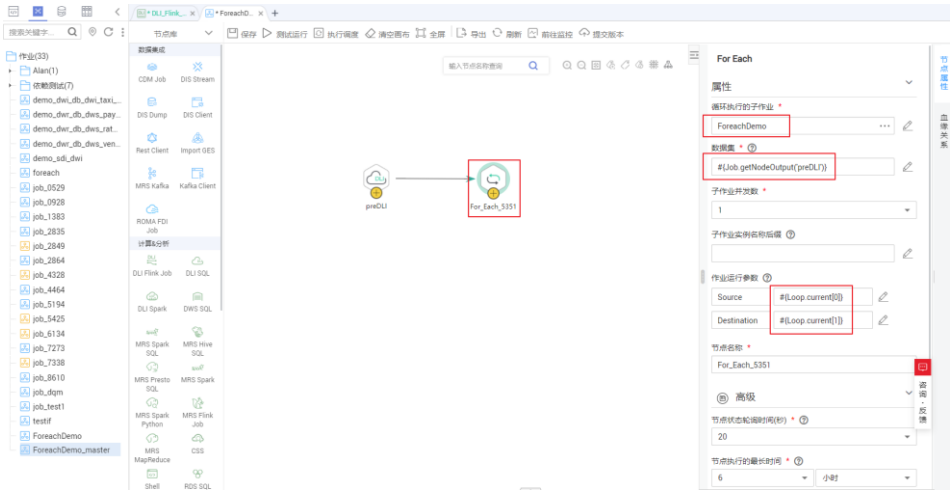
```
SELECT * FROM Table_List;
```

图3-514 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3-515 配置 For Each 算子

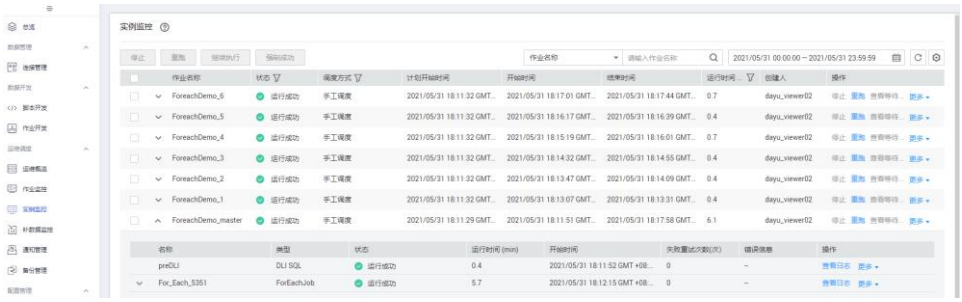


4. 配置完成后保存作业。

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3-516 查看作业实例

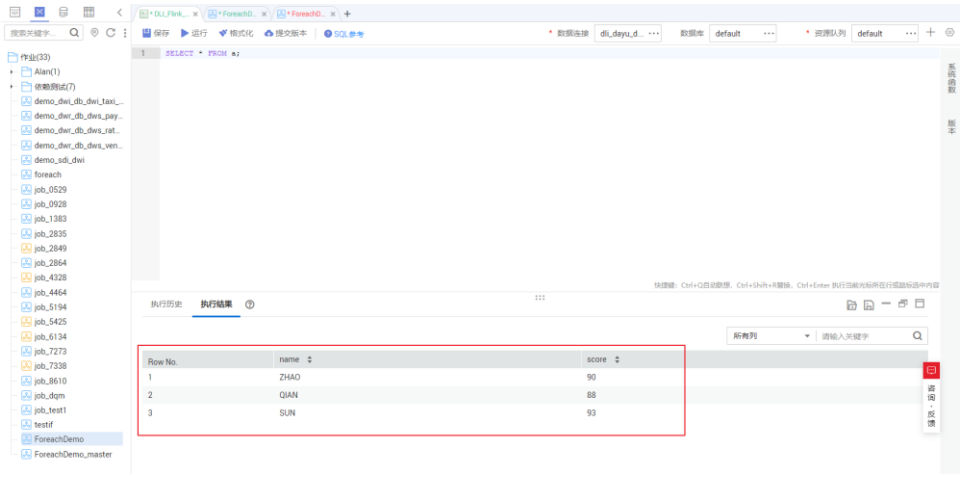


3. 查看对应的 6 个 DLI 目的表中是否已被插入预期的数据。您可以在 DataArts Studio 数据开发模块中，新建 DLI SQL 脚本执行以下 SQL 命令导入数据，也可以在数据湖探索（DLI）服务控制台中的 SQL 编辑器中执行以下 SQL 命令：

```
/* 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3-517 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

3.5.11.5 开发一个 Python 脚本

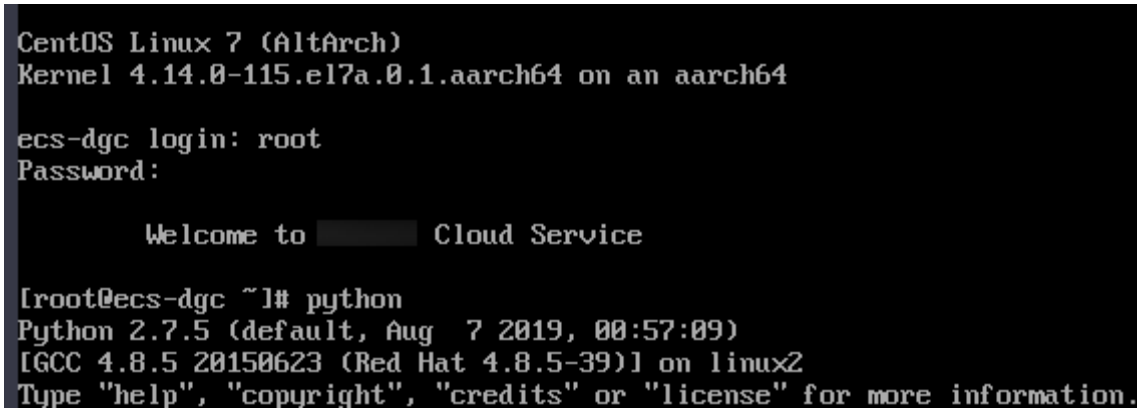
本章节介绍如何在数据开发模块上开发并执行 Python 脚本示例。

环境准备

- 已开通弹性云主机，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，您可登陆主机后使用 **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主机 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3-518 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3-519 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3-520 创建数据连接

步骤 4 参见表 3-316 配置相关参数，创建主机连接名称为“python_test”的数据连接，如所示。

表3-316 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《弹性云主机用户指南》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	主机的登陆用户名。
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”

参数	是否必选	说明
		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”。

说明

- 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。
- 主机连接：建立主机数据连接中创建的连接。

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

3.5.11.6 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 3.5.2.3 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3-521 开发脚本

关键说明：

- 图 3-521 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

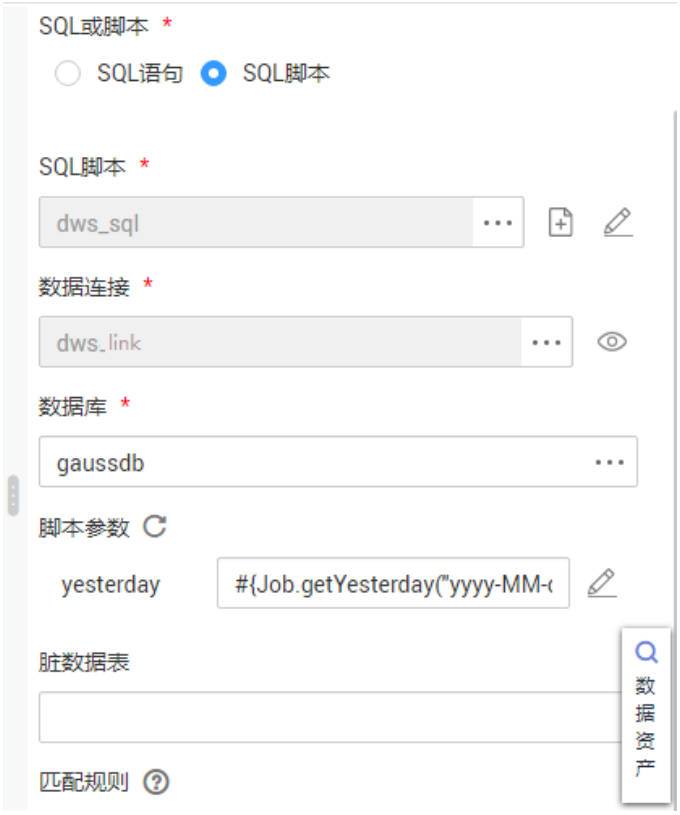
DWS SQL 脚本开发完成后，我们为 DWS SQL 脚本构建一个周期执行的作业，使得该脚本能定期执行。

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3-522 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3-523 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3-524 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

3.5.11.7 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”。

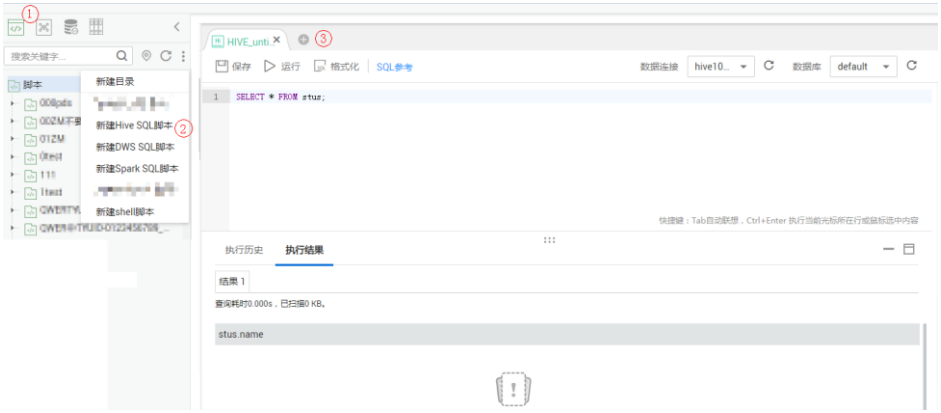
关键参数说明：

- 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3-525 开发脚本



关键说明：

- 图 3-525 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

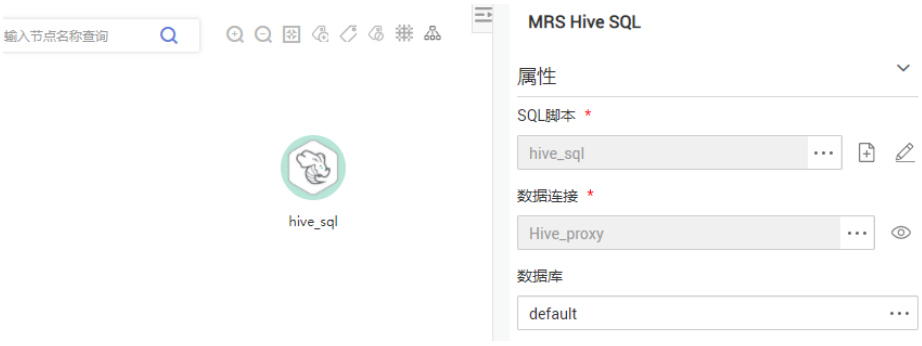
Hive SQL 脚本开发完成后，我们为 Hive SQL 脚本构建一个周期执行的作业，使得该脚本能定期执行。

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3-526 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3-527 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3-528 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

3.5.11.8 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。

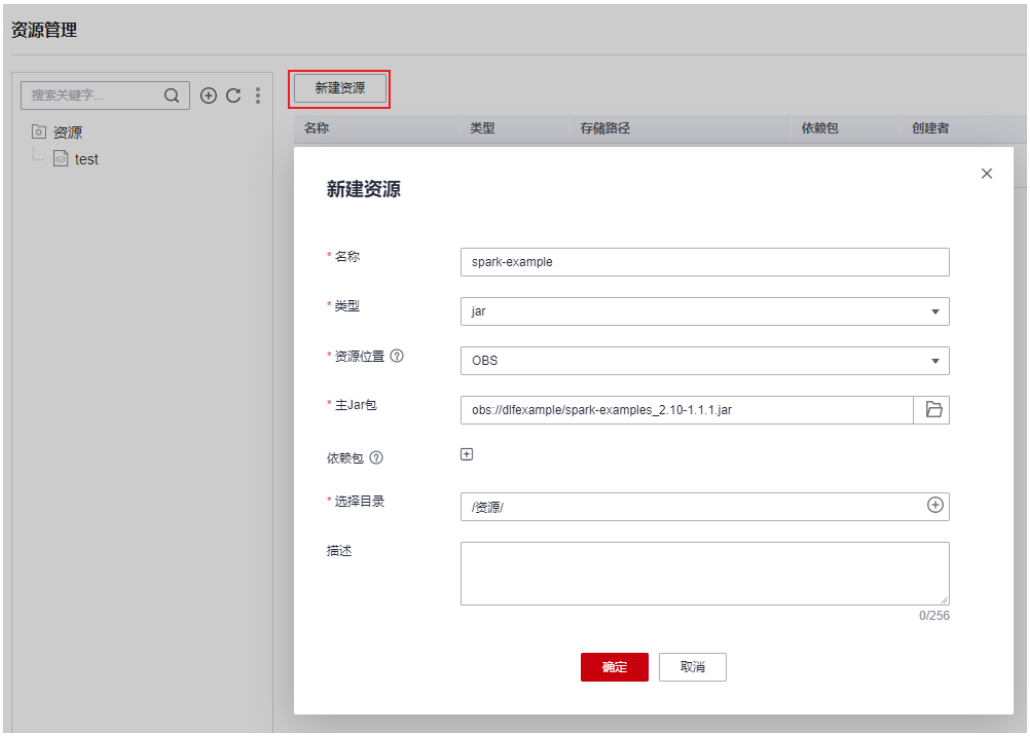
步骤 1 获取 Spark 作业代码 JAR 包后，将 JAR 包上传到 OBS 桶中，存储路径为“obs://dlfexample/spark-examples_2.10-1.1.1.jar”。

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3-529 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3-530 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3-531 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3-532 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3-533 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

3.5.11.9 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

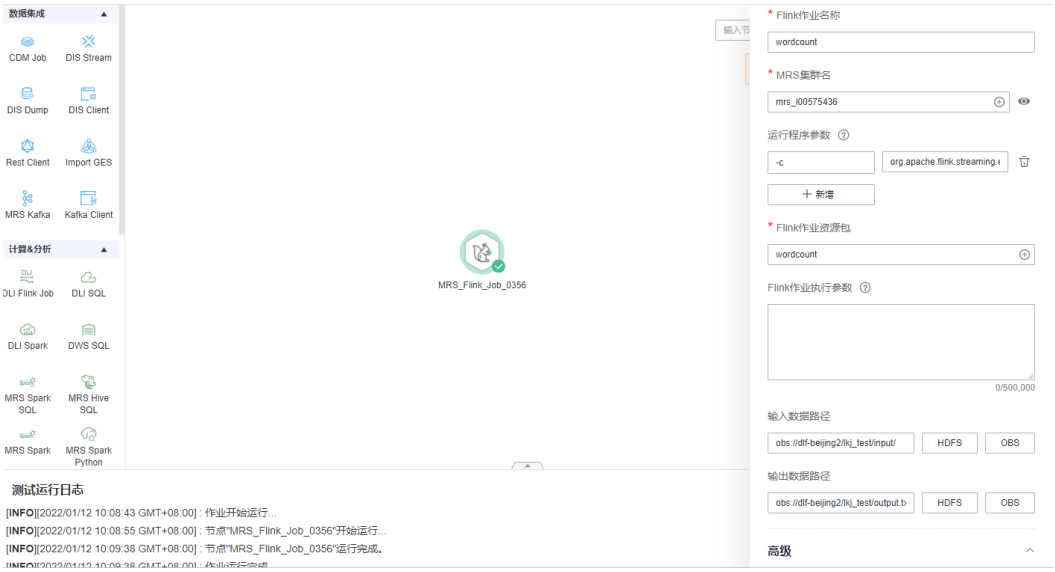
本例中，**WordCount.jar** 文件上传路径为：lkj_test/WordCount.jar；**word.txt** 文件上传路径为：lkj_test/input/word.txt。

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3-534 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3-535 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-region1/lkj_test/input/word.txt
--输出数据路径
obs://dlf-region1/lkj_test/output.txt
```

其中：

obs://dlf-region1/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-region1/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

3.5.11.10 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3-536 上传文件至 OBS 桶

说明

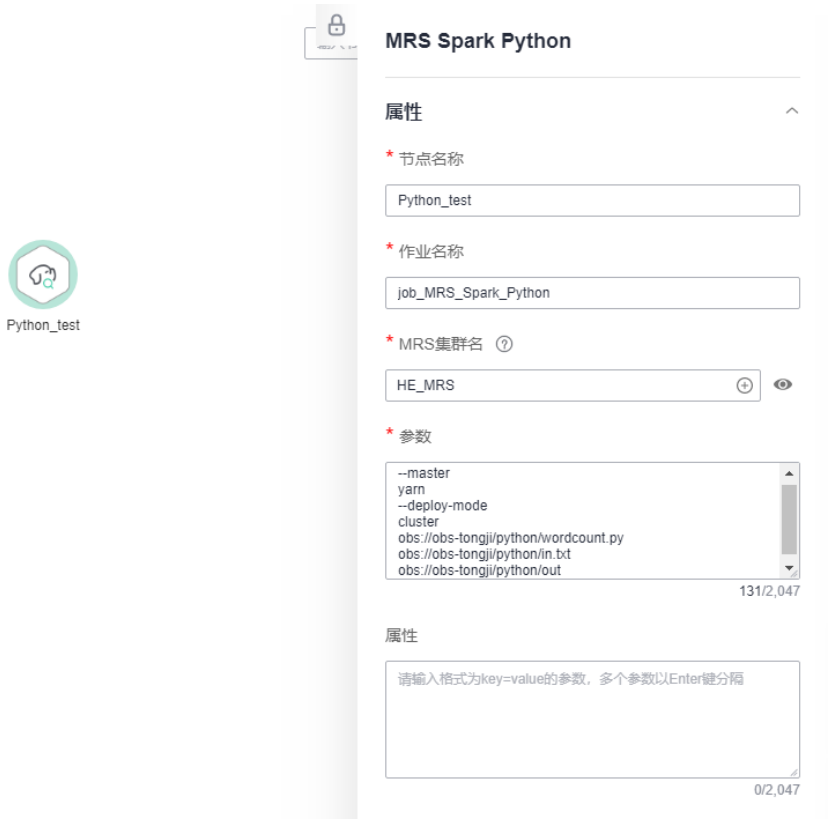
本例中，**wordcount.py** 和 **in.txt** 文件上传路径为：obs://obs-tongji/python/

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3-537 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3-538 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3-539 查看作业执行结果

作业日志中显示已运行成功

图3-540 作业运行日志

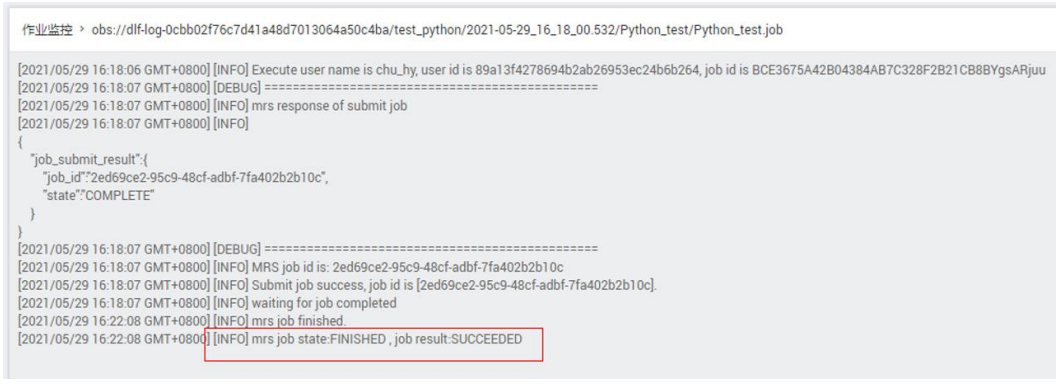


图3-541 作业运行状态

步骤 7 查看 OBS 桶中返回的记录。（没设置返回可跳过）

图3-542 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
```

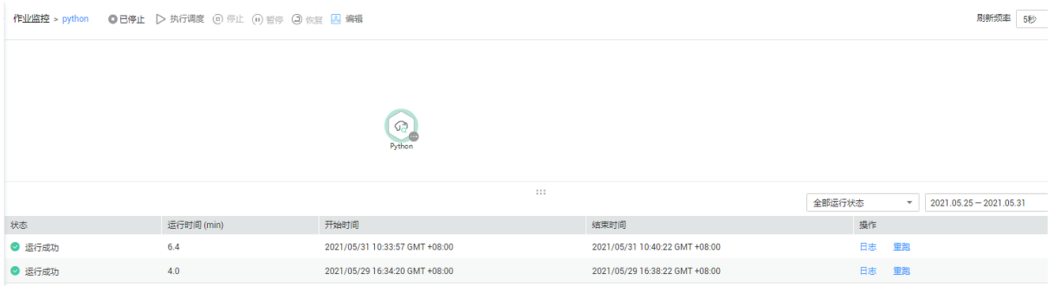
其中：zt_test_sparkPython1.py 为脚本所在路径

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。

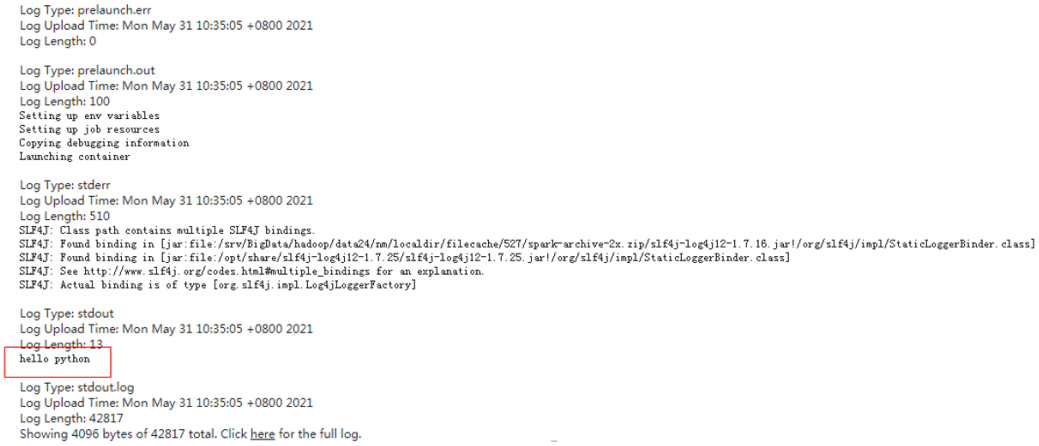
图3-543 查看作业执行结果



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3-544 查看 YARN 上日志



----结束

3.6 数据质量

3.6.1 业务指标监控

3.6.1.1 业务指标监控概述

业务指标监控模块是对业务指标进行质量管理的工具。

为了进行业务指标监控，您可以先自定义 SQL 指标，然后通过指标的逻辑表达式定义规则，最后新建并调度运行业务场景。通过业务场景的运行结果，您可以判断业务指标是否满足质量规则。业务场景的运行结果说明如下：

- 正常：表示实例正常结束，且执行结果符合预期。
- 告警：表示实例正常结束，但执行结果不符合预期。

- 异常：表示实例未正常结束。
- --：表示实例正在运行中，无执行结果。

业务指标监控主界面包括以下功能模块。

功能	说明
总览	默认首页是总览页面，显示了业务场景实例的运行状态和告警状态。 主要包括以下几部分内容： • 快速入门，介绍业务指标监控的业务流。 • 最近 7 天内的业务场景实例运行分布情况、实例告警运行分布情况。 • 可选周期内的告警趋势图、业务场景看板图、指标看板图。
指标管理	指标管理是业务指标监控的核心功能模块，是配置指标的主要入口。
规则管理	规则管理是配置规则的主要入口，支持通过指标的逻辑表达式定义规则。
业务场景管理	业务场景可以认为是业务指标质量作业，将创建的规则组进行调度运行。
运维管理	运维管理用于查看业务场景运行状态，处理运维问题。其中我的订阅中显示了所有订阅的任务运行情况。

3.6.1.2 新建指标

管理所有业务指标，包括指标的来源、定义等，使用目录维护业务指标。

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

前提条件

在 DataArts Studio 控制台的“*实例* > 进入控制台 > 空间管理 > 数据质量 > 业务指标监控 > 指标管理”页面创建归属目录。基于某个数据连接创建指标，需要选择指标目录，请参见图 3-545 创建归属目录。

图3-545 新建指标的归属目录

表3-317 导航栏按键说明

序号	说明
1	新建目录。
2	刷新目录。
3	选择全部，单击右键，可新建目录、重命名目录和删除目录。

新建指标

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

图3-546 选择数据质量

2. 选择“业务指标监控 > 指标管理”。
3. 单击“新建”，在弹出的对话框中，参见表 3-318 配置相关参数。

表3-318 配置业务指标参数

参数名	说明
指标名称	业务指标的名称，只能包含中文、英文字母、数字、“_”，且长度为 1~64 个字符。
数据连接	从下拉列表中选择已创建的数据连接。 说明 • 支持的数据连接类型：DWS、PostgreSQL、MRS Hive、DLI 和 MySql。 • 指标都是基于数据连接的，所以在建立指标之前需要先到元数据管理模块中建立数据连接。
数据库/队列	选择指标运行的数据库。 说明 当数据源为 DLI 时，需要选择运行的队列。
描述	为更好的识别业务指标 ，此处加以描述信息。描述信息长度不能超过 4096 个字符。
所属目录	业务指标的存储目录，可选择已创建的目录。目录创建请参见图 3-545。
来源类型	支持“自定义”。 用户自定义 SQL 语句，定义指标的来源。

3.6.1.3 新建规则

管理所有业务规则，规则定义了指标间或者指标和数值间的关系，使用目录维护业务规则。

前提条件

在 DataArts Studio 控制台的“*实例* > 进入控制台 > 空间管理 > 数据质量 > 业务指标监控 > 规则管理”页面创建归属目录。基于指标创建业务规则，需要选择规则归属目录，请参见图 3-547 创建归属目录。

图3-547 新建规则的归属目录

表3-319 导航栏按键说明

序号	说明
1	新建目录。
2	刷新目录。
3	选择全部，单击右键，可新建目录、重命名目录和删除目录。

新建规则

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

图3-548 选择数据质量

2. 选择“业务指标监控 > 规则管理”。
3. 单击“新建”，在弹出的对话框中，参见表 3-320 配置相关参数，新建规则。

表3-320 配置业务规则参数

参数名	说明
规则名称	业务规则的名称，只能包含中文、英文字母、数字、“_”，且长度为 1~64 个字符。
描述	为更好的识别业务规则，此处加以描述信息。描述信息长度不能超过 4096 个字符。
所属目录	业务规则的存储目录，可选择已创建的目录。目录创建请参见图 3-547。
定义关系	关系是定义指标和数值间或者指标和指标间的逻辑表达式，可以包含算术运算。指标使用小写字母 a-z 代替它的缩写，按添加指标的顺序依次为 a,b,c,...。 说明 只支持一个合法逻辑表达式，支持简单的四则算术运算。

3.6.1.4 新建业务场景

管理所有业务场景，场景定义了规则间的逻辑关系，使用目录维护业务场景。

前提条件

在 DataArts Studio 控制台的“*实例* > 进入控制台 > 空间管理 > 数据质量 > 业务指标监控 > 业务场景管理”页面创建归属目录。基于规则创建业务场景，需要选择业务场景归属目录，请参见图 3-549 创建归属目录。

图3-549 新建业务场景的归属目录

表3-321 导航栏按键说明

序号	说明
1	新建目录。
2	刷新目录。
3	选择全部，单击右键，可新建目录、重命名目录和删除目录。

新建业务场景

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

图3-550 选择数据质量

2. 选择“业务指标监控 > 业务场景管理”。
3. 单击“新建”，在弹出的对话框中，参见表 3-322 配置相关参数，新建场景。

表3-322 配置业务场景参数

参数名	说明
基本配置	
业务场景名称	业务场景的名称，只能包含中文、英文字母、数字、“_”，且长度为 1~64 个字符。
描述	为更好的识别业务场景 ，此处加以描述信息。描述信息长度不能超过 256 个字符。
所属目录	业务场景的存储目录，可选择已创建的目录。目录创建请参见图 3-549。
业务级别	支持提示、一般、严重和致命四种业务级别，业务级别决定发出通知消息的模板样式。
规则组配置	
定义规则组	规则组包含一个或者多个规则，规则间是逻辑表达式。
定义规则 A	支持从下拉框中选择已定义的规则。 单击⊕，可插入多条规则。
订阅配置	
通知状态	通过单击 或 来关闭或开启通知开关。

参数名	说明
通知类型	包含如下类型： • 触发告警 • 运行成功
选择主题	选择消息通知的主题。

4. 单击“下一步”，选择调度方式，支持单次调度和周期调度两种方式，周期调度的相关参数配置请参见表 3-323。

表3-323 配置周期调度参数

参数名	说明
生效日期	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
间隔时间	调度任务的间隔时间。
调度时间	设置调度任务的起始时间和结束时间。

3.6.1.5 查看业务场景实例

管理所有运行的业务场景，查看运行状态、运行日志、问题处理等。

界面说明

介绍“业务指标运维管理”页面中的区域和按键功能。

图3-551 运维管理页面

表3-324 运维管理页面说明

序号	区域	描述
1	菜单栏	运维管理的菜单栏，包括业务场景实例和我的订阅。

序号	区域	描述
		• 业务场景实例：展示当前用户的所有业务场景实例内容。 • 我的订阅：展示被当前用户设置订阅的业务场景信息列表。“我的订阅”较“业务场景实例”增加了“通知状态”信息。该信息展示了业务场景实例的运行结果是否被成功订阅，例如，发送告警邮件。
2	导航栏	左侧导航栏，包括数据业务场景的存储目录。 用户可以根据实际需要对业务场景进行分目录存放，每级目录旁边的数字代表属于该级目录的业务场景的个数。
3	业务场景实例列表	展示实例名称、运行状态、运行结果等信息。
4	搜索区域	• 可以选择性的展示业务场景实例，例如运行的开始时间和结束时间处于某一时间区间业务场景。 • 根据处理人、创建人、实例名称进行筛选展示业务场景实例的列表信息，输入内容支持模糊搜索。

表3-325 业务场景实例列表说明

菜单/按键	说明
运行状态	展示实例运行状态。 • 成功：表示实例运行成功。 • 失败：表示实例运行失败。 • 运行中：表示实例正在运行中。
运行结果	展示实例运行是否正常结束。 • 正常：表示实例正常结束，且执行结果符合预期。 • 告警：表示实例正常结束，但执行结果不符合预期。 • 异常：表示实例未正常结束。 • --：表示实例正在运行中，无执行结果。
重跑	再次运行业务场景实例。
运行日志	查看规则实例的详细运行日志信息。
更多 > 处理问题	对当前业务场景实例进行进一步处理。支持填写处理意见，关闭问题和移交他人。 如果实例的处理人是当前登录用户则可以对业务场景实例进行处理操作，包括填写意见和转交给他人处理。
更多 > 处理	可查看历史处理记录。

菜单/按键	说明
日志	

3.6.2 数据质量监控

3.6.2.1 数据质量监控概述

数据质量监控 DQC（Data Quality Control）模块是对数据库里的数据质量进行质量管理的工具。您可从完整性、有效性、及时性、一致性、准确性、唯一性六个维度进行单列、跨列、跨行和跨表的分析。数据质量支持对离线数据的监控，当离线数据发生变化时，数据质量会对数据进行校验，并阻塞生产链路，以避免问题数据污染扩散。同时，数据质量提供了历史校验结果的管理，以便您对数据质量分析和定级。

另外，数据质量监控 DQC 支持根据数据架构中的数据标准，自动生成标准化的质量规则，并进行周期性的监控。

数据质量监控主界面包括以下功能模块。

功能	说明
总览	默认首页是总览页面，显示了数据表的报警和阻塞情况。 主要包括以下几部分内容： • 所选周期内的作业数、实例数、异常表数，以及各种实例运行状态的分布和变化趋势情况。 • 当天告警分类统计、当天数据表告警统计、最近 7 天规则告警分类趋势的统计和最近 7 天规则数量的趋势。
规则模板	质量规则模板是数据质量的核心功能，是配置规则的主要入口。它主要管理规则配置（内置模板和自定义模板）的相关功能。
质量作业	质量作业可将规则模板或自定义规则应用到表中，进行数据质量监控。
对账作业	对账作业可将创建的规则应用到两张表中进行质量监控，并输出对账结果。
运维管理	运维管理用于查看规则运行状态，处理运维问题。
质量报告	系统根据作业的结果，会自动生成质量报告。

3.6.2.2 新建规则模板

数据质量支持对离线数据的监控，质量规则是数据质量的核心。DataArts Studio 系统内置的模板规则共计 25 种，分为库级规则、表级规则、字段级规则和跨字段级规则等规则类型，如表 3-326 所示。

表3-326 系统内置的规则模板一览表

规则类型	维度	模板名称	说明
库级	完整性	数据库空值扫描	计算数据库中所有表字段的空值行数
表级	准确性	表行数	计算数据表的总行数
	完整性	数据表空值扫描	计算数据表中所有表字段的空值行数
字段级	唯一性	字段唯一值	计算数据表中指定字段的唯一值行数
		字段重复值	计算数据表中指定字段的重复值行数
		多字段唯一性校验	校验 DWS 表中多个字段的组合是否唯一，最多支持 10 个字段的组合。
	完整性	字段空值	计算数据表中指定字段的空值行数
	准确性	字段平均值	计算数据表中指定字段的平均值
		字段汇总值	计算数据表中指定字段的汇总值
		字段最大值	计算数据表中指定字段的最大值
		字段最小值	计算数据表中指定字段的最小值
		字段长度校验	通过输入字段长度范围，校验 DWS 表中字段是否在允许范围内。
		字段值范围校验	通过输入字段值范围，校验 DWS 表中字段值是否在允许范围内。
		字段时间校验	通过输入字段时间范围，校验 DWS 表中字段时间是否在允许范围内。 注意，当前仅支持 DATE 和 TIMESTAMP 类型的字段，不支持 TIME 格式。
	有效性	身份证校验	通过内置的正则表达式规则，校验数据表中指定字段的合法情况
		邮箱校验	通过内置的正则表达式规则，校验数据表中指定字段的合法情况
		正则表达式校验	通过输入自定义的正则表达式，校验数据表中指定字段的合法情况
		IP 地址校验	通过内置的正则表达式规则，校验数据表中指定字段的合法情况
		电话格式校验	通过内置的正则表达式规则，校验数据表中指定字段的合法情况

规则类型	维度	模板名称	说明
		邮编格式校验	通过内置的正则表达式规则，校验数据表中指定字段的合法情况
		日期格式校验	通过内置的正则表达式规则，校验数据表中指定字段的合法情况
		合法性校验	通过输入自定义的正则表达式，校验数据表中指定字段的合法情况
		枚举值校验	通过输入自定义的枚举值，校验数据表中指定字段的合法情况
跨字段级	一致性	字段一致性校验	针对相同数据源的不同字段，校验数据表中指定字段是否与参考字段一致
	准确性	跨字段时间校验	针对相同 DWS 数据源的不同字段，通过输入大小关系符号，校验数据表中指定字段是否与参考字段的时间大小关系是否符合预期。 注意，当前仅支持 DATE 和 TIMESTAMP 类型的字段，不支持 TIME 格式。

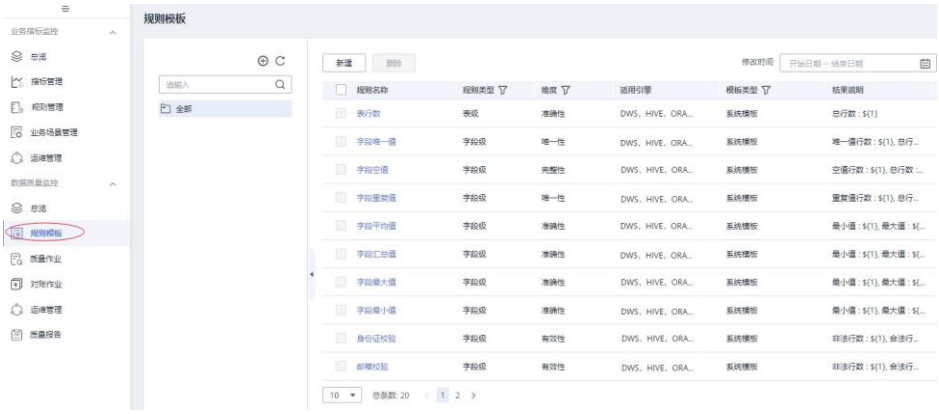
当系统内置规则模板不足以满足您的需求，您可根据实际需要创建规则。目前创建规则的方式包括规则模板和自定义规则：

- 自定义模板：在“数据质量监控 > 规则模板”处，新建规则模板。新建的规则模板系统会自动划分为对应的规则类型，为区分系统内置模板，显示为自定义模板。当前质量作业应用自定义模板时，不支持进行异常数据输出和质量评分。
- 自定义规则：在创建质量作业时，“规则类型”选择为“自定义规则”，然后您可以通过输入完整的 SQL 语句，定义如何对数据对象进行数据质量监控。

本文以新建自定义模板为例，说明如何创建规则。如果您需要新建自定义规则，请直接参考 3.6.2.3 新建质量作业进行自定义规则质量作业的创建。

步骤 1 选择“数据质量监控 > 规则模板”，单击“新建”，在弹出的新建规则模板页面中进行配置。

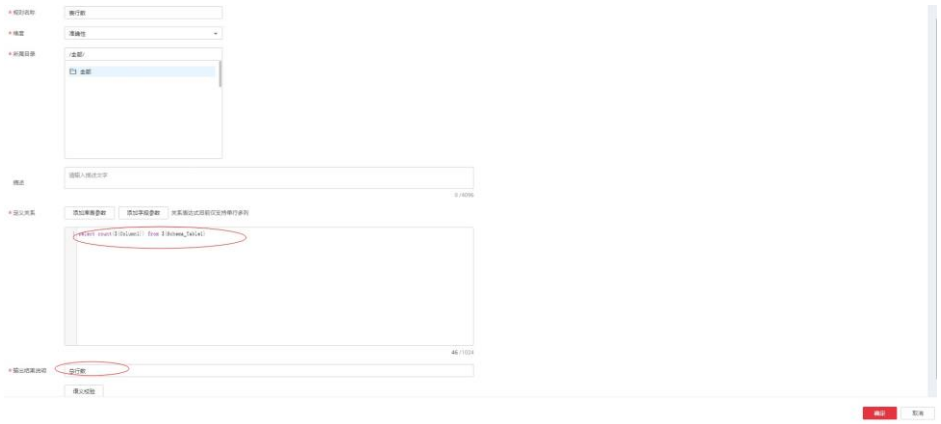
图3-552 新建规则模板



步骤 2 在弹出的新建规则模板页面中输入规则模板名称，选择规则匹配的维度，定义 SQL 模板并对输出结果进行说明。

- 维度：数据质量支持从完整性、有效性、及时性、一致性、准确性、唯一性六个维度进行单列、跨列、跨行和跨表的分析。自定义质量规则时，请对此规则进行维度匹配。
- 定义关系：输入 SQL 语句，实现对数据的查找。
  - 样例：统计表行数，输入 select count(${Column1}) from ${Schema_Table1}。其中${Column1}通过单击“添加字段参数”生成，${Schema_Table1} 通过单击“添加库表参数”生成。
- 输出结果说明：对 SQL 获得结果的每一列进行说明，列说明之间用逗号进行分隔。
  - 样例：当定义关系设置为，select max（${Column1}）,min(${Column2}) from ${Schema_Table1}，则输出结果说明为“最大值，最小值”。结果说明应该与关系定义的输出结果顺序一一对应。

图3-553 配置规则模板



步骤 3 单击“确定”后，系统默认发布此规则模板，版本名称默认为 V1.0。

----结束

管理规则模板

自定义规则模板不支持直接修改已发布的历史版本。当您有修改需求，可以通过发布新版本以修改规则模板，并可以选择下线历史版本且将待下线历史版本关联的作业迁移到新版本上。具体请参见如下操作。

步骤 1 选择“数据质量监控 > 规则模板”，在规则模板列表中找到待修改的规则模板，单击操作列的“发布”。

图3-554 发布规则模板

步骤 2 支持修改维度，修改输出结果说明和重新定义关系。

步骤 3 单击“发布新版本”，在提交发布对话框中，重新设置版本名称，并确认发布。

图3-555 发布新版本

步骤 4 提交发布后，单击操作列的“发布历史”，可以查看该规则模板的发布记录，支持查看变化信息、修改版本名称、下线对应版本等。

图3-556 发布历史界面

步骤 5 如需下线历史版本，点击历史版本最右侧的“下线”按钮。

- 如果该版本没有关联作业，点击确认即可下线。
- 如果该版本存在关联作业，需要选择迁移版本，将新版本与作业关联后，点击确认才能完成下线。

图3-557 迁移版本并下线

步骤 6 发布历史处支持进行版本比对，直观展示修改点。

图3-558 比对版本

----结束

导出规则模板

系统支持将自定义的规则模板批量导出，一次最多可导出 200 个规则模板。

步骤 1 选择“数据质量监控 > 规则模板”，选择要导出的自定义规则模板。

步骤 2 单击“导出”，弹出“导出规则模板”对话框。

步骤 3 单击“导出”，切换到“导出记录”页签。

步骤 4 在导出文件列表中，单击最新导出文件对应的“下载”，可将规则模板的 Excel 表格下载到本地。

----结束

导入规则模板

系统支持将自定义的规则模板批量导入，一次最大可导入 1M 数据的文件，并且最多 200 个规则模板。

步骤 1 选择“数据质量监控 > 规则模板”，单击“导入”，弹出“导入规则模板”对话框。

步骤 2 在“导入配置”页签，选择模板名称重名策略。

- 终止：如果模板名称有重复，则全部导入失败。
- 跳过：如果模板名称有重复，会忽略后继续导入。

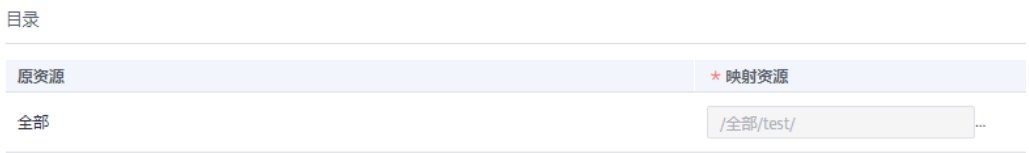
步骤 3 单击“上传文件”，选择准备好的数据文件。

说明

可通过如下两种方式填写数据文件：

- (推荐使用) 通过“导出”功能，可将数据直接/或修改后批量导入系统。
- 通过“下载 Excel 模板”，将数据填写好再导入至系统中。

步骤 4 配置目录的映射资源信息，选择导入后的规则模板存储目录。



步骤 5 单击“导入”，将填好的 Excel 表格模板导入到系统。

步骤 6 单击“导入记录”页签，可查看对应的导入记录。

----结束

3.6.2.3 新建质量作业

质量作业可将创建的规则应用到建好的表中进行质量监控。

前提条件

在 DataArts Studio 控制台数据质量模块，“数据质量监控 > 质量作业”页面创建归属目录。基于某个数据连接创建质量作业，需要选择作业归属目录，请参见图 3-559 创建归属目录。

图3-559 新建质量作业的归属目录

表3-327 导航栏按键说明

序号	说明
1	新建目录。
2	刷新目录。
3	选择目录，单击右键，可新建目录、删除目录和对目录重命名。

配置流程

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

图3-560 选择数据质量

2. 选择“数据质量监控 > 质量作业”。
3. 单击“新建”，在弹出的对话框中，参见表 3-328 配置相关参数。

表3-328 配置作业参数

参数名	说明
作业名称	质量作业的名称，只能包含中文、英文字母、数字、“_”，且长度为 1~64 个字符。
描述	为更好的识别数据质量作业，此处加以描述信息。描述信息长度不能超过 256 个字符。
所属目录	数据质量作业的存储目录，可选择已创建的目录。目录创建请参见图 3-559。
作业级别	支持提示、一般、严重和致命四种级别，作业级别决定发出通知消息的模板样式。

4. 单击“下一步”，进入规则配置页面。您需要点击规则卡片中的 ，然后参见表 3-329 配置数据质量规则。默认规则配置完成后，您也可选择继续添加更多的质量规则，创建完成后单击下一步，即可将创建的所有规则应用到已建好的库或表中。

图3-561 打开质量作业规则配置

表3-329 配置模板规则

添加方式	配置	说明
基本信息	子作业名称	在作业的执行结果中，每条规则对应一个子作业。为便于结果查看和日志定位，建议您补充子作业信息。
	描述	为更好的识别子作业，此处加以描述信息。
来源对象	规则类型	包括库级规则、表级规则、字段级规则、跨字段级规则和自定义规则，自定义规则可针对表中的具体字段配置监控规则。
	数据连接	来源对象/目的对象支持的数据源类型：DWS，MRS Hive，DLI，ORACLE、RDS（MySQL、PostgreSQL）。 从下拉列表中选择已创建的数据连接。 说明 • 规则都是基于数据连接的，所以在建立数据质量规则之前需要先到管理中心模块中建立数据连接。 • 针对通过代理连接的 MRS HIVE，需要选择 MRS API 方式或者代理方式提交： • MRS API 方式：通过 MRS API 的方式提交。历史作业默认是 MRS API 提交，编辑作业时建议不修改。 • 代理方式：通过用户名、密码访问的方式提交。新建作业建议选择代理提交，可以避免权限问题导致的作业提交失败。
	数据库	选择配置的数据质量规则所应用到的数据库。 说明 • 数据库基于已建立的数据连接。 • 当“规则类型”选择“库级规则”，数据对象选择对应的数据库即可。
	数据表	选择配置的数据质量规则所应用到的表。

添加方式	配置	说明
		说明 • 数据表与数据库强相关，基于已选择的数据库。 • 当“规则类型”选择“表级规则”，数据对象选择对应的数据表。
	SQL	当“规则类型”选择“自定义规则”时，需要配置该参数。此处需输入完整的 SQL 语句，定义如何对数据对象进行数据质量监控。
	失败策略	选择是否勾选“忽略规则错误”。
	选择字段	当“规则类型”选择“字段级规则”，需要配置该参数。此处选择对应数据表中的字段。 说明 数据质量字段级别校验不支持对字段名为单个字母（例如：a,b,c,d...等）的字段进行校验。
	参考数据对象	当“规则类型”选择“跨字段级规则”，需要配置该参数。此处选择参考的数据字段。
	维度	当“规则类型”选择“自定义规则”时，需要配置该参数。将该自定义规则与质量六性（完整性、有效性、及时性、一致性、准确性、唯一性）进行关联。
计算引擎	集群名称	选择运行质量作业的引擎。仅数据连接为 DLI 类型时，此参数有效。
规则模板	模板名称	选择系统内置的或者用户自定义的规则模板。 说明 模板类型与规则类型强相关，详情请参见表 3-326。除去系统内置规则模板外，您也可关联在 3.6.2.2 新建规则模板中新建的自定义模板。
	版本	仅“模板名称”选择为自定义的规则模板时，需要配置该参数。自定义的规则模板发布后，会产生对应的版本号，此处选择所需的版本。
	权重	设置规则的权重，支持按照字段级别设置权重。权重范围：【1-9】，整数。默认值为 5。
计算范围	选择扫描区域	支持选择“全表扫描”或“条件扫描”，默认为全表扫描。 当仅需计算一部分数据，或需周期性按时间戳运行质量作业时，建议通过设置 where 条件进行条件扫描。
	where 条件	输入 where 子句，系统会选择符合条件的数据进行扫描。 例如需要筛选数据表中“age”字段在 (18, 60] 区间范围内

添加方式	配置	说明

		的数据时，where 条件可设置为如下内容：
		`age > 18 and age <= 60`
		where 条件还支持输入为 SQL 动态表达式，例如当需要根据“time”字段筛选数据表中 24 小时前的数据时，where 条件可设置为如下内容：
		`time >= (date_trunc('hour', now()) - interval '24 h')`
		`and time <= (date_trunc('hour', now()))`

告警条件	告警表达式	此参数可选，如果您需要针对当前规则设定告警条件，则可以在此配置告警条件的表达式。如果您需要通过多条规则的逻辑运算统一设置告警条件的表达式，此处无需设置，可在下一步的告警配置中统一设置。 配置规则的告警条件后，系统通过“告警参数”的值，结合告警条件进行真假判断，如果结果为真则进行告警。另外，除了单一告警表达式的结果，您还可以通过逻辑运算符组成组成更复杂的告警条件进行告警。当前表达式中支持如下逻辑运算符，且可以通过“(”和“)”进行包围： • +：相加 • -：相减 • *：相乘 • /：相除 • ==：等于 • !=：不等于 • >：大于 • <：小于 • >=：大于等于 • <=：小于等于 • !：非 • ||：或 • &&：与 例如，“规则模板”为“字段空值”时，您可以参考如下样例进行配置： • 需要配置字段空值大于 10 时告警，则此处可设置为“${1}>10”，其中“${1}”为通过告警参数配置的“空值行数”。 • 需要配置有字段空值率大于 80%时告警，则此处可设置为“${3}>0.8”，其中“${3}”为通过告警参数配置的“空值率”。 • 需要配置字段空值大于 10 或字段空值率大于 80%时告警，则此处可设置为“(${1}>10)\|\|(${3}>0.8)”，其中“${1}”和“${3}”分别为通过告警参数配置的“空值行数”和“空值率”，“||”表示满足两个条件之一即

添加方式	配置	说明
		会告警。
	告警参数	此参数来源于规则模板的输出结果。您可以单击界面显示的参数从而输入告警表达式中的告警参数，单击后系统会在“告警表达式”输入框给出参数的表达式。 例如“规则模板”为“字段空值”时，点击告警参数“空值行数”，在“告警表达式”输入框会显示为“${1}”。
	逻辑运算符	可选，本参数支持将单一告警表达式的结果进行逻辑运算，组成更复杂的告警条件。 您可以将鼠标光标放在“告警表达式”输入框处需要进行逻辑运算的两个告警表达式之间，然后单击输入如下之一运算符。另外，您也可以手动输入，当前表达式中支持如下逻辑运算符，且可以通过“(”和“)”进行包围： • +：相加 • -：相减 • *：相乘 • /：相除 • ==：等于 • !=：不等于 • >：大于 • <：小于 • >=：大于等于 • <=：小于等于 • !：非 • ||：或 • &&：与 例如，“规则模板”为“字段空值”，需要配置字段空值大于 10 或字段空值率大于 80%时告警，则“告警表达式”可设置为“(${1}>10)\|\|(${3}>0.8)”，其中“${1}”和“${3}”分别为通过告警参数配置的“空值行数”和“空值率”，“||”表示满足两个条件之一即会告警。
	质量评分	当“规则类型”选择“自定义规则”时，需要配置该参数。
	生成异常数据	开启“生成异常数据”开关，单击“选择库表”可将质量作业中不符合设定规则的异常数据存储在异常表中。 说明 • 自定义模板不支持生成异常数据，自定义规则可通过自定义异常表 SQL 生成异常数据。 • 系统内置模板，“表级规则”中的“表行数”模板。“字段级规则”中的“字段平均值”、“字段汇总值”、“字段最大值”、

添加方式	配置	说明
		“字段最小值”模板不支持生成异常数据。 • 当质量作业设置周期调度或重跑时，每次实例运行的扫描的异常数据会持续插入该异常表。建议您定期到该数据湖中清理异常表数据，避免异常数据表超大带来的成本与性能问题。
	异常表	单击选择库表，可以配置输出表名的前后缀。
	输出配置	• 输出规则配置：勾选，则可在异常表中显示质量作业的配置信息，方便查看异常数据产生的源头。 • 输出空值：勾选，则当空值不满足设定规则时，可在异常表中输出空值。
	异常数据数量	可选择输出全部的异常数据，或者设定数量的异常数据。
	异常表SQL	当“规则类型”选择“自定义规则”时，需要配置该参数。此处需输入完整的 SQL 语句，指定输出哪些数据是异常数据。
	查看相同规则	单击，创建质量作业时， • 能够根据表和字段判断规则的重复性。 • 提示已存在相关子规则和质量作业，您可看到已有规则。
计算范围	选择扫描区域	用来确定所配置的某条规则应检查的范围。 • 勾选全表扫描，则遍历所有表。 • 勾选条件扫描，输入 where 条件后，精确定位分区查询数据，不需要全表扫描查询。

5. 单击“下一步”，设置告警配置信息。如果您在上一步的规则配置中已配置告警表达式，此处会自动带出已配置的表达式；如果未配置，则您可在此进行配置。多条（2 条及以上）子规则时，则可以选择如下两种告警配置方式之一进行配置：
   a. 支持通过子规则的告警条件，分别上报告警。
   b. 将子规则之间的告警参数值通过数学运算和逻辑运算，设置一个统一的告警条件表达式来表示作业是否告警。

   当前表达式中支持如下逻辑运算符，且可以通过“(”和“)”进行包围：
   - +：相加
   - -：相减
   - *：相乘
   - /：相除
   - ==：等于
   - !=：不等于
   - >：大于

- <：小于
- >=：大于等于
- <=：小于等于
- !：非
- ||：或
- &&：与

6. 单击“下一步”，设置订阅配置信息，如果需要接收 SMN 通知，打开通知状态，选择通知类型和 SMN 服务主体。
7. 单击“下一步”，选择调度方式，支持单次调度和周期调度两种方式，周期调度的相关参数配置请参见表 3-330。配置完成后单击“提交”。

说明

1. 单次调度会产生手动任务的实例，手动任务的特点是没有调度依赖，只需要手动触发即可。
2. 周期调度会产生周期实例，周期实例是周期任务达到启用调度所配置的周期性运行时间时，被自动调度起来的实例快照。
3. 周期任务每调度一次，便生成一个实例工作流。您可以对已调度起的实例任务进行日常的运维管理，如查看运行状态，对任务进行终止、重跑等操作。
4. 只有支持委托提交作业的 MRS 集群，才支持质量作业周期调度。 支持委托方式提交作业的 MRS 集群有：

- MRS 的非安全集群。
- MRS 的安全集群，集群版本大于 2.1.0，并且安装了 MRS 2.1.0.1 以上的补丁。

表3-330 配置周期调度参数

参数名	说明
生效日期	调度任务的生效日期。
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 说明 • 调度周期选择分钟/小时，需配置调度的开始时间、间隔时间和结束时间。开始时间目前支持设置到分钟级别，进行错峰调度。 • 调度周期选择天，需要配置调度时间，即确定了调度任务于每天的几时几分启用。 • 调度周期选择周，需要配置生效时间和调度时间，即确定了调度任务于周几的几时几分启用。

导出质量作业

系统支持批量导出质量作业，一次最多可导出 200 个质量作业。

步骤 1 选择“数据质量监控 > 质量作业”，选择要导出的质量作业。

步骤 2 单击“导出”，弹出“导出质量作业”对话框。

步骤 3 单击“导出”，切换到“导出记录”页签。

步骤 4 在导出文件列表中，单击最新导出文件对应的“下载”，可将质量作业的 Excel 表格下载到本地。

----结束

导入质量作业

系统支持批量导入质量作业，一次最大可导入 1M 数据的文件，并且最多 200 个质量作业。

步骤 1 选择“数据质量监控 > 质量作业”，单击“导入”，弹出“导入质量作业”对话框。

步骤 2 在“导入配置”页签，选择模板名称重名策略。

- 终止：如果质量作业名称有重复，则全部导入失败。
- 跳过：如果质量作业名称有重复，会忽略后继续导入。
- 覆盖：如果质量作业名称有重复，会覆盖现有同名作业。

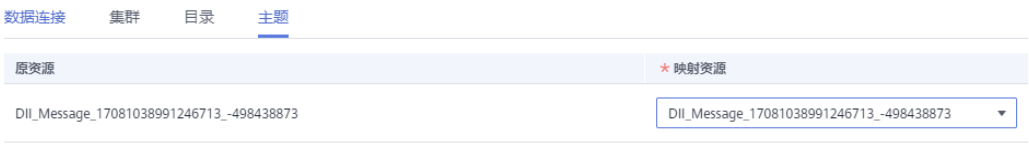
步骤 3 单击“上传文件”，选择准备好的数据文件。

说明

可通过如下两种方式填写数据文件：

- (推荐使用) 通过“导出”功能，可将数据直接/或修改后批量导入系统。
- 通过“下载 Excel 模板”，将数据填写好，再导入至系统中。

步骤 4 分别配置数据连接、集群、目录、主题的映射资源信息。



- 数据连接：选择导入后的数据连接类型。
- 集群：如果数据连接类型是 DLI，需要选择对应的队列。
- 目录：选择导入后的质量作业存储目录。
- 主题：如果配置了消息通知，需要选择主题。

步骤 5 单击“导入”，将填好的 Excel 表格模板导入到系统。

步骤 6 单击“导入记录”页签，可查看对应的导入记录。

----结束

3.6.2.4 新建对账作业

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

数据质量监控中的对账作业支持跨源数据对账能力，可将创建的规则应用到两张表中进行质量监控，并输出对账结果。

前提条件

在 DataArts Studio 控制台的数据质量模块，“数据质量监控 > 对账作业”页面创建归属目录。基于某个数据连接创建对账作业，需要选择作业归属目录，请参见图 3-562 创建归属目录。

目录相关操作参考表 3-331。

图3-562 新建对账作业的归属目录

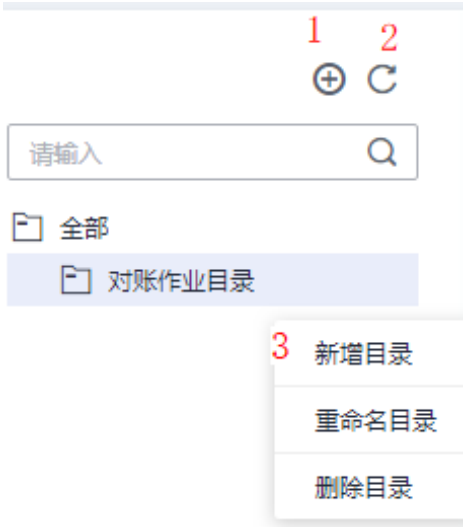


表3-331 目录导航栏按键说明

序号	说明
1	新建目录
2	刷新目录
3	选择目录，单击右键，可新建目录、删除目录和对目录重命名。

创建作业

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

   图3-563 选择数据质量

2. 选择“数据质量监控 > 对账作业”。
3. 单击“新建”，在弹出的对话框中，参见表 3-332 配置相关参数。

表3-332 配置作业参数

参数名	说明
作业名称	对账作业的名称，只能包含中文、英文字母、数字、“_”，且长度为 1~64 个字符。
描述	为更好的识别数据对账作业 ，此处加以描述信息。描述信息长度不能超过 256 个字符。
所属目录	数据对账作业的存储目录，可选择已创建的目录。目录创建请参见图 3-562。

参数名	说明
作业级别	支持提示，一般，严重和致命四种级别，作业级别决定发出通知消息的模板样式。

4. 单击“下一步”，进入规则配置页面。您需要点击规则卡片中的 ，然后参见表3-333 配置数据对账规则。您也可选择添加对账规则。

图3-564 打开对账作业规则配置

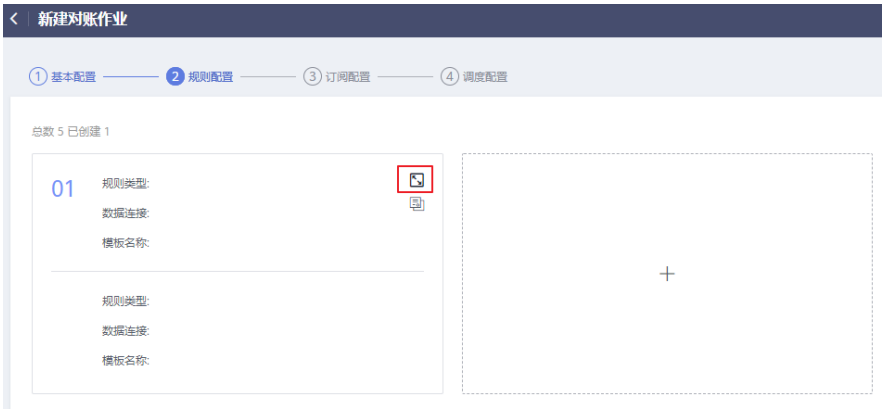


表3-333 配置模板规则

模块	参数名	说明
基本信息	子作业名称	在作业的执行结果中，每条规则对应一个子作业。为便于结果查看和日志定位，建议您补充子作业信息。
	描述	为更好的识别子作业，此处加以描述信息。
来源对象/目的对象	规则类型	来源对象的“规则类型”包括“表级规则”，“字段级规则”和“自定义规则”。字段级规则可针对表中的具体字段配置监控规则。此处选择为表级规则，页面中其他设置项对应为表级规则配置项。 目的对象的“规则类型”由来源对象的规则类型自动生成。
	数据连接	来源对象/目的对象支持的数据源类型：DWS，MRS Hive，DLI，ORACLE、RDS（MySQL、PostgreSQL）。 从下拉列表中选择已创建的数据连接。 说明 • 规则都是基于数据连接的，所以在建立数据质量规则之前需要先到管理中心模块中建立数据连接。 • 针对通过代理连接的 MRS HIVE，需要选择 MRS API 方式或

模块	参数名	说明
		者代理方式提交： • MRS API 方式：通过 MRS API 的方式提交。历史作业默认是 MRS API 提交，编辑作业时建议不修改。 • 代理方式：通过用户名、密码访问的方式提交。新建作业建议选择代理提交，可以避免权限问题导致的作业提交失败。
	数据对象	在来源对象选择的数据表将和右侧目的对象的数据表做结果比较。选择配置的数据对账规则所应用到的表。 说明 数据表与数据库强相关，基于已选择的数据库。数据库基于已建立的数据连接。
	SQL	当“规则类型”选择“自定义规则”时，需要配置该参数。此处需输入完整的 SQL 语句，定义如何对数据对象进行数据质量监控。
计算引擎	集群名称	选择运行对账作业的引擎。仅数据连接为 DLI 类型时，此参数有效。
规则模板	模板名称	该参数定义如何对数据对象做数据质量监控。 来源对象的模板名称包含内置的规则模板和用户自定义的规则模板。 目的对象的“模板名称”由来源对象的规则类型自动生成。 说明 模板类型与规则类型强相关，详情请参见表 3-326。除去系统内置规则模板外，您也可关联在 3.6.2.2 新建规则模板中新建的自定义模板。
	版本	仅“模板名称”选择为自定义的规则模板时，需要配置该参数。自定义的规则模板发布后，会产生对应的版本号，此处选择所需的版本。
计算范围	选择扫描区域	支持选择“全表扫描”或“条件扫描”，默认为全表扫描。 当仅需计算一部分数据，或需周期性按时间戳运行质量作业时，建议通过设置 where 条件进行条件扫描。
	where 条件	输入 where 子句，系统会选择符合条件的数据进行扫描。 例如需要筛选数据表中“age”字段在 (18, 60] 区间范围内的数据时，where 条件可设置为如下内容： `age > 18 and age <= 60` where 条件还支持输入为 SQL 动态表达式，例如当需要根据“time”字段筛选数据表中 24 小时前的数据时，where 条件可设置为如下内容：

模块	参数名	说明
		`time >= (date_trunc('hour', now()) - interval '24 h') and time <= (date_trunc('hour', now()))`
告警条件	告警表达式	此参数可选，如果您需要针对当前规则设定告警条件，则可以在此配置告警条件的表达式。 配置规则的告警条件后，系统通过“告警参数”的值，结合告警条件进行真假判断，如果结果为真则进行告警。另外，除了单一告警表达式的结果，您还可以通过逻辑运算符组成组成更复杂的告警条件进行告警。当前表达式中支持如下逻辑运算符，且可以通过“(”和“)”进行包围： • +：相加 • -：相减 • *：相乘 • /：相除 • ==：等于 • !=：不等于 • >：大于 • <：小于 • >=：大于等于 • <=：小于等于 • !：非 • ||：或 • &&：与 例如，对账作业的来源侧和目的侧的“规则模板”为“表行数”时，您可以参考如下样例进行配置： • 需要配置来源侧表行数小于 100 时告警，则此处可设置为“${1_1}<100”，其中“${1_1}”为通过告警参数配置的来源侧表“总行数”。 • 需要配置来源侧表行数不等于目的侧表行数时告警，则此处可设置为“${1_1}!=${2_1}”，其中“${1_1}”为通过告警参数配置的来源侧表“总行数”，“${2_1}”为通过告警参数配置的目的侧表“总行数”。 • 需要配置来源侧表行数小于 100 或来源侧表行数不等于目的侧表行数时告警，则此处可设置为“(${1_1}<100)\|\|(${1_1}!=${2_1})”，其中“${1_1}”和“${2_1}”分别为通过告警参数配置的来源侧表和目的侧表的“总行数”，“||”表示满足两个条件之一即会告警。
	告警参数	此参数来源于规则模板的输出结果。您可以单击界面显示的参数从而输入告警表达式中的告警参数，单击后系统会

模块	参数名	说明
		在“告警表达式”输入框给出参数的表达式。 例如“规则模板”为“表行数”时，点击告警参数“总行数”，在“告警表达式”输入框会显示为“${1_1}”。
	逻辑运算符	可选，本参数支持将单一告警表达式的结果进行逻辑运算，组成更复杂的告警条件。 您可以将鼠标光标放在“告警表达式”输入框处需要进行逻辑运算的两个告警表达式之间，然后单击输入如下之一运算符。另外，您也可以手动输入，当前表达式中支持如下逻辑运算符，且可以通过“(”和“)”进行包围： • +：相加 • -：相减 • *：相乘 • /：相除 • ==：等于 • !=：不等于 • >：大于 • <：小于 • >=：大于等于 • <=：小于等于 • !：非 • ||：或 • &&：与 例如，“规则模板”为“表行数”，需要配置来源侧表行数小于 100 或来源侧表行数不等于目的侧表行数时告警，则此处可设置为“(${1_1}<100)\|\|(${1_1}!=${2_1})”，其中“${1_1}”和“${2_1}”分别为通过告警参数配置的来源侧表和目的侧表的“总行数”，“||”表示满足两个条件之一即会告警。

5. 单击“下一步”，设置订阅配置信息，如果需要接收 SMN 通知，打开通知状态，选择通知类型和 SMN 服务主体，如图 3-565。

图3-565 订阅配置

6. 单击“下一步”，选择调度方式，支持单次调度和周期调度两种方式，周期调度的相关参数配置请参见表 3-334。配置完成后单击“提交”。

说明

1. 单次调度会产生手动任务的实例，手动任务的特点是没有调度依赖，只需要手动触发即可。
2. 周期调度会产生周期实例，周期实例是周期任务达到启用调度所配置的周期性运行时间时，被自动调度起来的实例快照。
3. 周期任务每调度一次，便生成一个实例工作流。您可以对已调度起的实例任务进行日常的运维管理，如查看运行状态，对任务进行终止、重跑等操作。
4. 只有支持委托提交作业的 MRS 集群，才支持对账作业周期调度。 支持委托方式提交作业的 MRS 集群有：

- MRS 的非安全集群。
- MRS 的安全集群，集群版本大于 2.1.0，并且安装了 MRS 2.1.0.1 以上的补丁。

表3-334 配置周期调度参数

参数名	说明
生效日期	调度任务的生效日期。
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 说明 • 调度周期选择分钟/小时，需配置调度的开始时间、间隔时间和结束时间。 • 调度周期选择天，需要配置调度时间，即确定了调度任务于每天的几时几分启用。 • 调度周期选择周，需要配置生效时间和调度时间，即确定了调度任务于周几的几时几分启用。

导出对账作业

系统支持批量导出对账作业，一次最多可导出 200 个对账作业。

步骤 1 选择“数据质量监控 > 对账作业”，选择要导出的对账作业。

步骤 2 单击“导出”，弹出“导出对账作业”对话框。

步骤 3 单击“导出”，切换到“导出记录”页签。

步骤 4 在导出文件列表中，单击最新导出文件对应的“下载”，可将质量作业的 Excel 表格下载到本地。

----结束

导入对账作业

系统支持批量导入对账作业，一次最大可导入 1M 数据的文件，并且最多 200 个对账作业。

步骤 1 选择“数据质量“监控” > 对账作业”，单击“导入”，弹出“导入对账作业”对话框。

步骤 2 在“导入配置”页签，选择模板名称重名策略。

- 终止：如果对账作业名称有重复，则全部导入失败。
- 跳过：如果对账作业名称有重复，会忽略后继续导入。
- 覆盖：如果对账作业名称有重复，会覆盖现有同名作业。

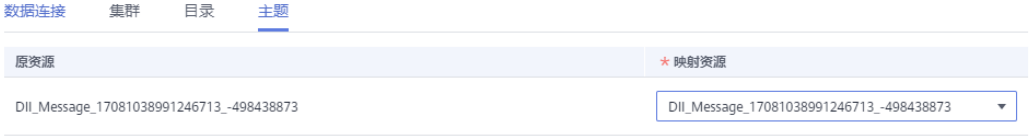
步骤 3 单击“上传文件”，选择准备好的数据文件。

说明

可通过如下两种方式填写数据文件：

- (推荐使用) 通过“导出”功能，可将数据直接/或修改后批量导入系统。
- 通过“下载 Excel 模板”，将数据填写好，再导入至系统中。

步骤 4 分别配置数据连接、集群、目录、主题、的映射资源信息。单击“导入”，将填好的 Excel 表格模板导入到系统。



- 数据连接：选择导入后的数据连接类型。
- 集群：如果数据连接类型是 DLI，需要选择对应的队列。

- 目录：选择导入后的对账作业存储目录。
- 主题：如果配置了消息通知，需要选择主题。

步骤 5 单击“导入记录页签”，可查看对应的导入记录。

----结束

3.6.2.5 查看规则实例

界面说明

介绍“数据质量规则运维管理”页面中的区域和按键功能。

图3-566 页面区域说明

表3-335 数据质量规则运维管理页面

序号	区域	描述
1	导航栏	左侧导航栏，包括数据质量规则的存储目录。 用户可以根据实际需要对规则进行分目录存放，每级目录旁边的数字代表属于该级目录的规则实例的个数。
2	规则实例列表	展示实例名称、运行状态、运行结果等信息。
3	搜索区域	• 可以选择性的展示规则实例，例如运行的开始时间和结束时间处于某一时间区间实例。 • 根据处理人、创建人、实例名称进行搜索展示规则实例的列表信息，输入内容支持模糊搜索。

表3-336 规则实例列表说明

菜单/按键	说明
实例名称	由“规则名称-数字”组成，数字越大，表示该实例创建的时间越近。
类型	显示作业类型，当前包含质量作业和对账作业。
运行状态	展示实例运行状态，包含成功、失败和运行中、告警。右侧弹窗分选项卡可查看规则实例的详细运行日志信息。 • 成功：表示实例正常结束，且执行结果符合预期。

菜单/按键	说明
	• 失败：表示实例未正常结束。 • 告警：表示实例正常结束，但执行结果不符合预期。 • 运行中：表示实例正在运行中，无执行结果。
通知状态	展示实例通知状态，包含成功、失败和未触发。
开始时间	展示实例开始运行的时间。
运行时间	展示实例的运行时长。
重跑	再次运行规则实例。
结果&日志	详细展示作业实例的运行结果和日志。 • 对账作业结果 对账作业运行结果中，左侧表示源端表行数规则运行结果，右侧表示目的端表行数规则运行结果，误差率表示两端数据行数的差异比率，误差率为 0 表示两端一致。
处理&记录	对当前规则实例进行进一步处理。支持填写处理意见，关闭问题和移交他人。 如果实例的处理人是当前登录用户则可以对规则实例进行处理操作，包括填写意见和转交给他人处理。

3.6.2.6 查看质量报告

您可以查询业务指标、数据质量中数据对象的质量评分，来判断各个对象是否质量达标。

查询业务质量评分

质量评分的满分可设置为 5 分，10 分，100 分。默认为 5 分制，是以表关联的规则为基础进行评分的。而表、业务对象、主题域等不同维度的评分，本质上是基于规则评分在不同维度下的加权平均值进行计算的。

您可以查询主题域分组、主题域、业务对象、表以及表关联的规则评分，具体评分对象的计算公式，请参见表 3-337。

表3-337 对象评分计算公式

对象	评分计算公式
规则	创建质量作业时，包含“比率”、“值率”的系统内置规则及用户自定义规则可以生成质量评分报告。 • 包含“比率”、“值率”的规则可以分为正向规则及反向规则，正向规则即比值越高，代表数据质量越好；反向规则即比值越高，则数据质量越差。 正向规则包含唯一值率、重复值率、合法比率规则，反向规则包

对象	评分计算公式
	含空值率规则。 • 正向规则评分=满足规则的数据行数/数据总行数*满分（5，10，100）。 • 反向规则评分=（1-满足规则的数据行数/数据总行数）*满分（5，10，100）。 • 当表为空，即总行数为 0 时，正向规则评分固定为满分，反向评分固定为 0 分。
表	表评分计算公式：Σ(表关联的所有规则评分*规则权重)/Σ规则权重
业务对象	业务对象下所有表评分的加权求平均值，即：Σ业务对象下所有表评分/表的数量。
主题域	主题域下所有业务对象评分的加权求平均值，即：Σ主题域下所有业务对象评分/业务对象的数量。
主题域分组	分组下所有主题域评分的加权求平均值，即：Σ分组下所有主题域评分/主题域的数量。

步骤 1 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

图3-567 选择数据质量

步骤 2 选择“数据质量监控 > 质量报告”。

步骤 3 单击“业务报告”页签，选择主题及截至日期，查询截至日期前 7 天的数据质量评分，如图 3-568 所示。

图3-568 业务对象

说明

- 以评分满分为 5 分为例。其中 4-5 分评价为优秀，3-4 分为良好，2-3 分为中等，1-2 分为及格，0-1 分为不及格。
- 当天质量评分数据在次日凌晨生成。
- 质量评分历史趋势中的实线为截至日期前 7 天质量评分组成的连线，虚线为这 7 天质量评分的平均分。
- 若一天多次运行该作业，当天的质量评分为最后一次的得分。

步骤 4 单击“表评分”列的评分值链接，展开该表关联的规则评分。

步骤 5 单击“规则评分”列的评分值链接，展开该规则关联的字段评分，如图 3-569 所示。

图3-569 表关联规则评分

子规则字段评分

字段名称	规则描述	分数	字段权重	空值行数	总行数	空值率	告警状态
autotest....	字段空值...	85.7142	5	1	7	0.1428	false
autotest....	字段空值...	71.4285	5	2	7	0.2857	true

----结束

查看数据质量评分

质量评分的满分可设置为 5 分，10 分，100 分。默认为 5 分制，是以表关联的规则为基础进行评分的。而表、数据库等不同维度的评分均基于规则评分，本质上是基于规则评分在不同维度下的加权平均值进行计算的。

您可以查询所创建数据连接下数据库、数据库下的数据表以及数据表所关联规则的评分，具体评分对象的计算公式，请参见表 3-338。

表3-338 对象评分计算公式

对象	评分计算公式
规则	创建质量作业时，作业关联的规则中结果说明列包含“比率”、“值率”的系统内置规则及用户自定义规则可以生成质量评分报告。 • 包含“比率”、“值率”的规则可以分为正向规则及反向规则，正向规则即比值越高，代表数据质量越好；反向规则即比值越高，则数据质量越差。 正向规则包含唯一值率、重复值率、合法比率规则，反向规则包含空值率规则。 • 正向规则评分=满足规则的数据行数/数据总行数*满分（5，10，100）。 • 反向规则评分=（1-满足规则的数据行数/数据总行数）*满分（5，10，100）。
数据表	表评分计算公式：Σ(表关联的所有规则评分*规则权重)/Σ规则权重
数据库	数据库下所有数据表评分的加权求平均值，即：Σ数据库下所有数据表评分/表的数量。

步骤 1 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

图3-570 选择数据质量

步骤 2 选择“数据质量监控 > 质量报告”。

步骤 3 在“技术报告”页签，选择数据连接及截至日期，查询截至日期前 7 天的数据质量评分，如图 3-571 所示。

图3-571 选择数据连接

说明

- 以评分满分为 5 分为例。其中 4-5 分评价为优秀，3-4 分为良好，2-3 分为不及格，1-2 分为较差，0-1 分为极差。
- 当天质量评分数据在次日凌晨生成。
- 质量评分历史趋势中的实线为截至日期前 7 天质量评分组成的连线，虚线为这 7 天质量评分的平均分。
- 若一天多次运行该作业，当天的质量评分为最后一次的得分。

步骤 4 单击“表评分”列的评分值链接，展开该表关联的规则评分。

步骤 5 单击“规则评分”列的评分值链接，展开该规则关联的字段评分，如图 3-572 所示。

图3-572 表关联规则评分界面

----结束

3.6.3 使用教程

3.6.3.1 新建一个业务场景

场景说明

业务场景用于监控业务指标。本例以新建一个业务场景为例，介绍如何使用业务指标监控功能。

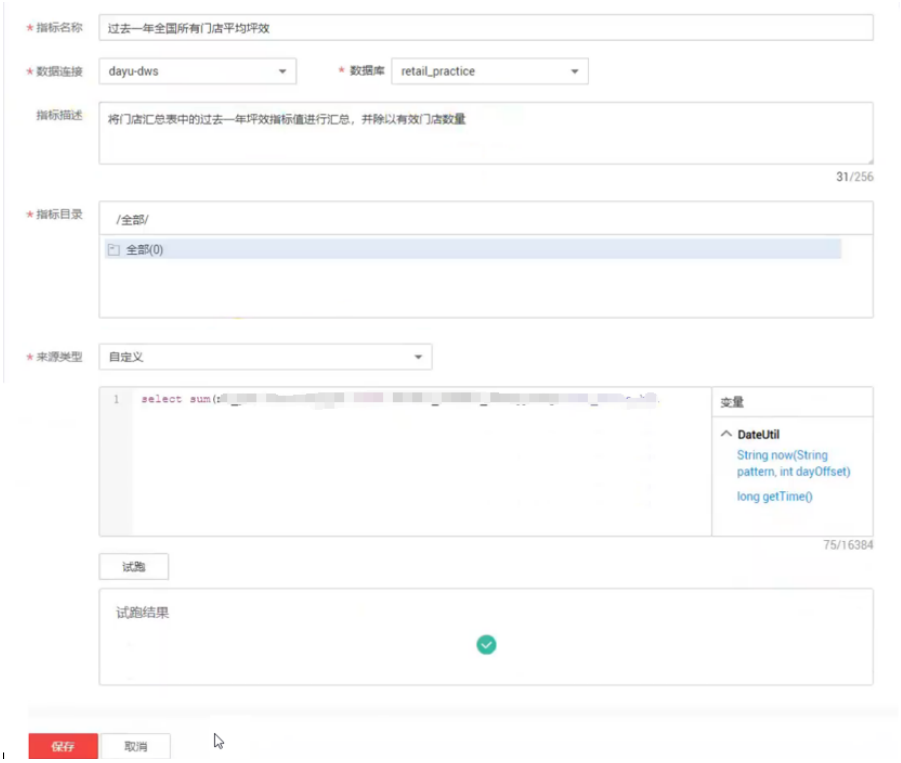
操作步骤

步骤 1 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

图3-573 选择数据质量

步骤 2 新建业务指标。

1. 单击左侧导航“指标管理”。
2. 单击页面上方的“新建”，如下图所示。

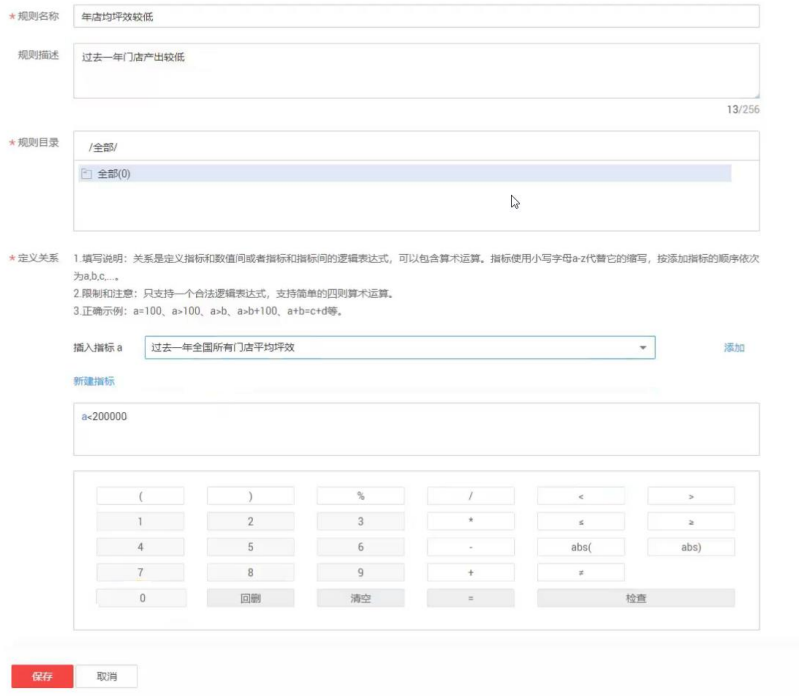


3. 单击“试跑”，查看试跑运行成功的结果。
4. 单击“保存”，完成指标的创建。

步骤 3 新建规则。

1. 单击左侧导航“规则管理”。
2. 单击页面上方的“新建”，创建第一条规则。
3. 输入参数值，如下图所示。

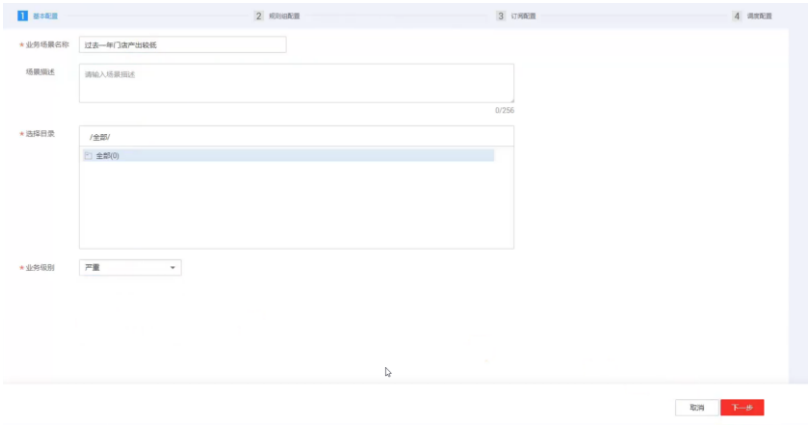
4. 单击“保存”。
5. 单击页面上方的“新建”，创建第二条规则。
6. 输入参数值，如下图所示。



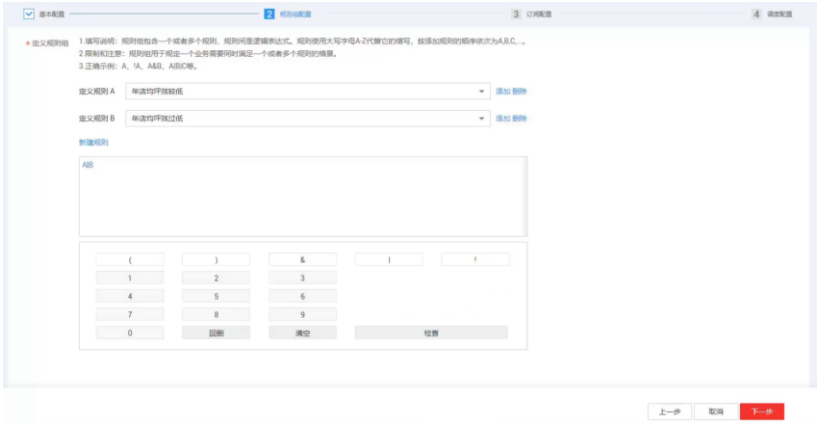
7. 单击“保存”。

步骤 4 新建业务场景。

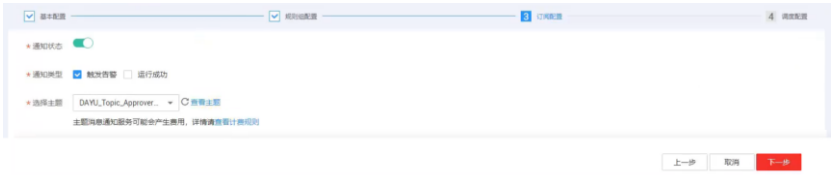
1. 单击左侧导航“业务场景管理”。
2. 单击页面上方的“新建”，输入场景的基本配置参数，如下图所示。



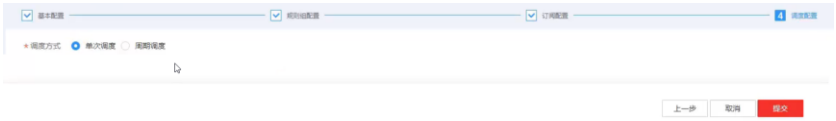
3. 单击“下一步”，输入规则组的配置参数，如下图所示。



4. 单击“下一步”，配置订阅信息，如下图所示。



5. 单击“下一步”，配置调度信息，如下图所示。



6. 单击“提交”，完成作业场景的创建。

步骤 5 在业务场景管理列表中，单击操作列的“运行”，跳转到运维管理模块。

1. 单击右上角的刷新按钮，可以查看业务场景的运行状态为成功。
2. 单击运行结果，可查看具体的坪效结果。

----结束

3.6.3.2 新建一个质量作业

场景说明

开发质量作业是为了监控数据质量。本章以新建一个质量作业为例，介绍如何开发质量作业。

操作步骤

步骤 1 在 DataArts Studio 控制台首页，选择实例，单击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

步骤 2 创建规则模板。

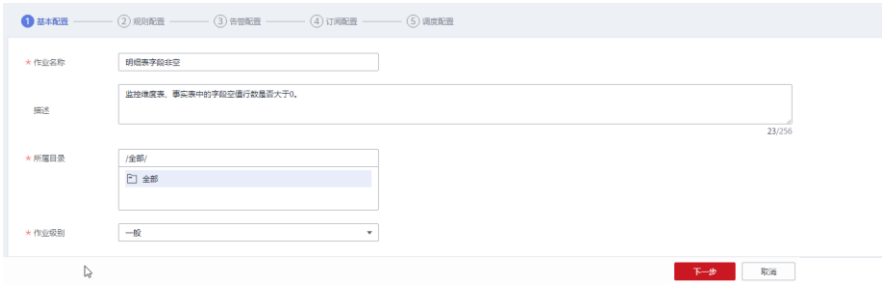
1. 单击左侧导航“规则模板”，默认展示系统自定义的规则。数据质量的规则包含 6 个维度，分别是：完整性、唯一性、及时性、有效性、准确性、一致性。
2. **可选：**单击“新建”，可自定义创建规则。

说明

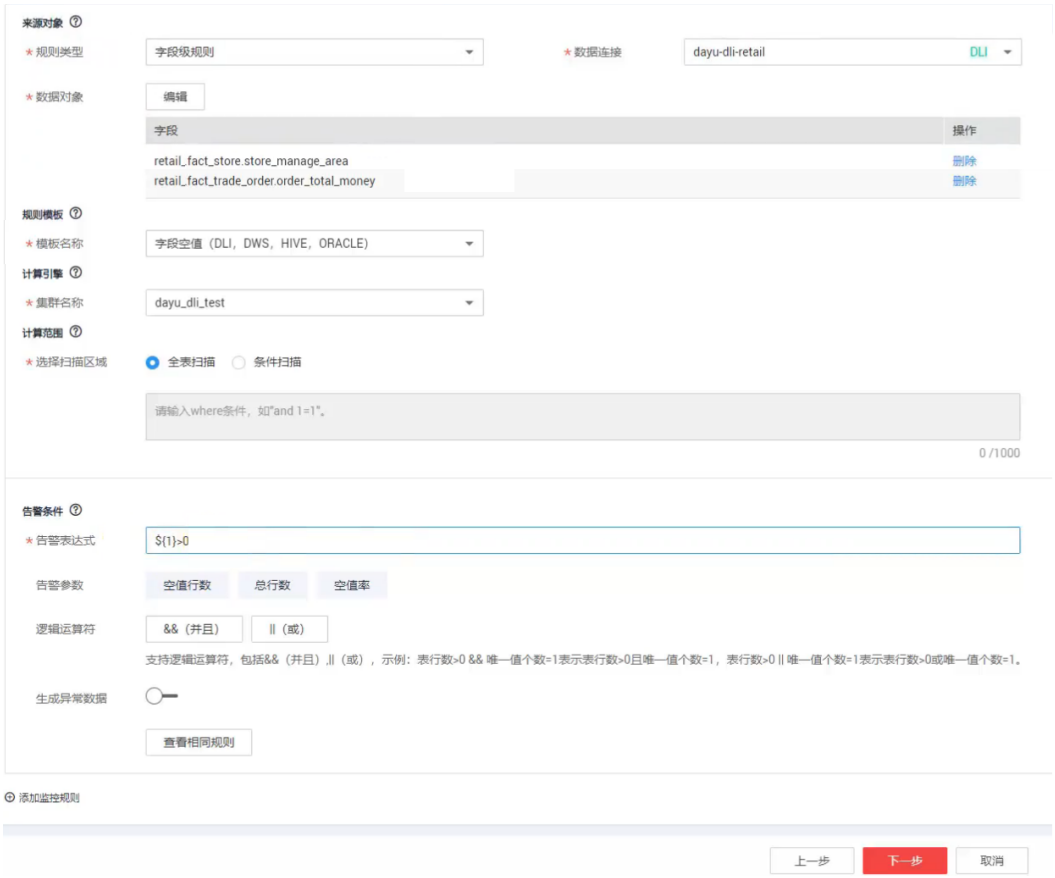
本例使用系统自定义的规则即可。

步骤 3 创建质量作业。

1. 单击左侧导航“质量作业”。
2. 单击“新建”，配置质量作业的基本信息，如下图所示。



3. 单击“下一步”，进入规则配置页面。您需要单击规则卡片中的  ，然后配置规则信息，如下图所示。



4. 单击“下一步”，配置告警信息，如下图所示。

5. 单击“下一步”，配置订阅信息，如下图所示。

6. 单击“下一步”，配置调度信息，如下图所示。

7. 单击“提交”，完成质量作业的创建。

步骤 4 在质量作业表中，单击操作列的“运行”，跳转到运维管理模块。

1. 待质量作业运行成功后，单击左侧导航菜单的“质量报告”
2. 默认展示技术报告，如下图所示。

3. 单击“业务报告”页签，查看业务报告，如下图所示。

----结束

3.6.3.3 新建一个对账作业实例

场景说明

数据对账对于数据开发和数据迁移流程中的数据一致性至关重要，而跨源数据对账的能力是检验数据迁移或数据加工前后是否一致的关键指标。本章分别以 DLI 和 DWS 作为数据源，介绍如何通过 DataArts Studio 中的数据质量模块实现跨源数据对账的基本一致性校验。

环境准备

需要准备好对账的数据源，即通过管理中心分别创建数据连接，用于跨源数据对账。

操作步骤

步骤 1 建立跨源数据连接。

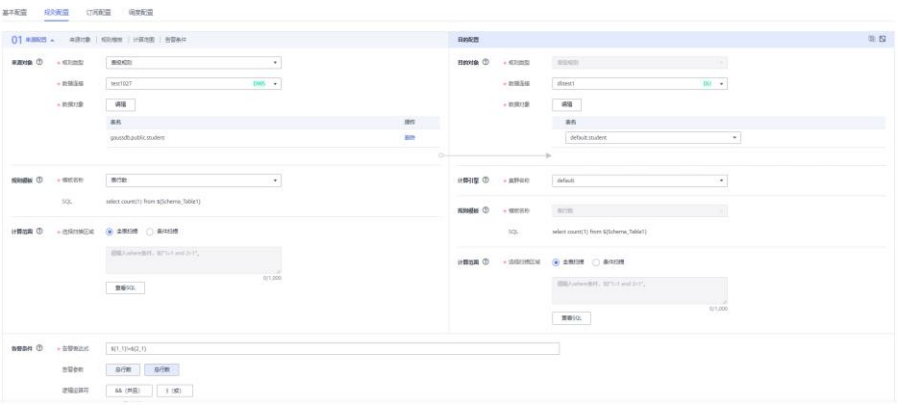
1. 创建 DLI 数据连接。在 DataArts Studio 管理中心模块，单击创建数据连接，数据连接类型选择“数据湖探索（DLI）”，输入数据连接名称，单击“测试”，提示连接成功，单击“确定”。

2. 创建 DWS 数据连接。在 DataArts Studio 管理中心模块，单击创建数据连接，数据连接类型选择“数据仓库服务（DWS）”，输入数据连接名称，设置其他参数，如下图所示，单击“测试”，提示连接成功，单击“确定”。

步骤 2 创建对账作业。

1. 在 DataArts Studio 数据质量模块，单击左侧导航菜单“对账作业”。
2. 单击“新建”，配置对账作业的基本信息，如下图所示。

3. 单击“下一步”，进入规则配置页面。您需要点击规则卡片中的  ，然后配置对账规则，如下图所示。



说明

- 需要分别配置源端和目的端的信息。
- 配置告警条件，其中点击左侧的表行数（${1_1}）表示左侧源端选中表的行数，点击右侧表行数（${2_1}）表示目的端表行数。此处配置告警条件为${1_1}!=${2_1}，表示当左侧表行数与右侧表行数不一致时，触发报警并显示报警状态。

4. 单击“下一步”，配置订阅信息，如下图所示。

说明

勾选触发告警表示作业报警时发送通知到对应的 smn 主题，勾选运行成功表示不报警时发送通知到 SMN 主题。

5. 单击“下一步”，配置调度方式，如下图所示。

说明

单次调度表示需要手动触发运行，周期性调度表示会按照配置定期触发作业运行。此处以当天配置为例，设置每 15 分钟触发运行一次对账作业为例的配置。

6. 单击“提交”，对账作业创建完成。

步骤 3 查看对账作业。

1. 单击对应的对账作业操作列中的运行链接，运行对账作业后，自动跳转到运维管理页面。
2. 单击结果&日志查看运行结果和运行日志，等待作业运行结束后，如下图所示。

----结束

结果分析

至此，完成了通过 DataArts Studio 数据质量模块中的对账作业功能实现了 DLI 和 DWS 两种不同数据源中的表行数一致性对账功能。

运行结果中，左侧表示源端表行数规则运行结果，右侧表示目的端表行数规则运行结果。

误差率表示两端数据行数的差异比率，此处误差率为 0 表示两端一致。

3.7 数据目录

该模块提供企业级的元数据管理，厘清信息资产。通过数据地图，实现数据资产的数据血缘和数据全景可视，提供数据智能搜索和运营监控。

3.7.1 数据地图

3.7.1.1 简介

数据地图围绕数据搜索，服务于数据分析、数据开发、数据挖掘、数据运营等数据表的使用者和拥有者，提供方便快捷的数据搜索服务，拥有功能强大的血缘信息及影响分析。

- 搜索：在进行数据分析前，使用数据地图进行关键词搜索，帮助快速缩小范围，找到对应的数据。
- 详情：使用数据地图根据表名直接查看表详情，快速查阅明细信息，掌握使用规则。
- 血缘：通过数据地图的血缘分析可以查看每个数据表的来源、去向，并查看每个表及字段的加工逻辑。

3.7.1.2 资产总览

通过总览，可以查看数据资产全局视图及资产报告，包括资产总数量和总大小，以及按照数据连接分类的数据库数量和数据表数量。

前提条件

- 已创建数据连接，如何创建数据连接请参见 3.2.2 创建数据连接。
- 已配置采集任务，如何创建采集任务请参见 3.7.4.2 任务管理。
- 采集任务已成功运行，查看采集任务运行状况请参见 3.7.4.3 任务监控。

操作步骤

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-574 选择数据目录

2. 选择“数据地图 > 总览”。
3. 单击“业务资产”，业务资产显示业务对象、逻辑实体、业务属性的数量及其详情。
4. 单击“技术资产”，技术资产显示数据库、数据表、数据量的数量及其详情。
5. 单击“指标资产”，指标资产显示业务指标及其详情。

标签和分类

“标签”是相关性很强的关键字，帮助用户对资产进行分类和描述，以便于检索。

“分类”是指按照种类、等级或性质分别归类。分类是自上而下的，通过对事物进行分析，按照一定的标准，划分出不同的类别。

二者主要区别如下：

表3-339 标签和分类区别

属性	分类	标签
排他性	有	无
关系	从属	相关（关联）
创建	事前规划	任意时间
代价	高	低

3.7.1.3 数据目录

通过数据目录可以对资产进行搜索、过滤、查看详情、查看血缘、查看关系、添加分类与标签等操作。

资产搜索

通过资产名称和描述的关键字或按所有属性搜索资产，支持模糊搜索。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

   图3-575 选择数据目录

2. 选择“数据地图 > 数据目录”。
3. 在资产搜索输入框输入需要查找的数据关键字进行搜索，搜索结果以列表方式显示。

   按名称和描述搜索：表示按照资产的名称和描述进行搜索。

   按所有属性搜索：表示按照资产的全部属性（即详情页中展示的属性）进行搜索。

说明

- 支持保存当前设置的搜索条件。
- 支持导入搜索条件。

资产筛选

对资产搜索结果，可以基于条件进行筛选，支持的筛选条件类别如下：

- 数据连接：数据资产所属数据连接名称。

- 类型：数据资产所属类型。
- 分类：数据资产所属分类。
- 标签：数据资产所包含的标签。
- 密级：数据资产所属密级。

如下通过资产类型过滤搜索结果，其他类同。

步骤 1 在类型过滤区域，选择“Table”，搜索结果显示属于 Table 类型的资产。

步骤 2 类型过滤条件按照名称排序，默认只显示前五种类型，单击“全部”，显示系统目前支持的所有资产类型。

----结束

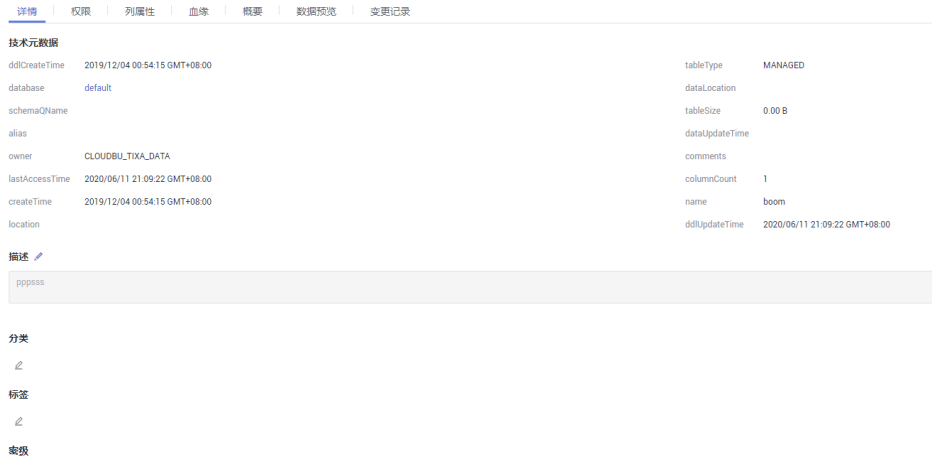
资产详情

以查看数据表详情为例。

步骤 1 在资产搜索结果列表，单击任意数据表，进入数据表详情页面。

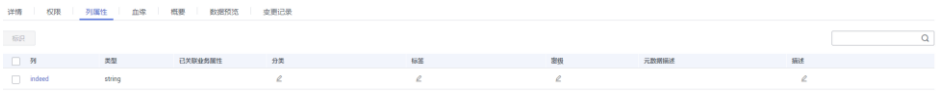
步骤 2 在“详情”页签，可查看技术元数据基本属性、编辑描述；可给数据表添加标签和密级；可给数据表的列和 OBS 对象添加或删除分类、标签和密级。

图3-576 查看详情



步骤 3 在“列属性”页签，可查看数据表的列属性，给数据列添加或删除分类、标签和密级，并编辑描述。

图3-577 管理列属性



步骤 4 在“血缘”页签，可查看数据表的血缘关系，包括血缘和影响。如何配置数据血缘请参见 3.5.9.2 节点数据血缘。节点配置血缘关系后，作业执行时可以自动解析，然后数据目录采集元数据时会采集上来，在数据目录中展示。

步骤 5 在“概要”页签，查看数据表的概要信息（当前仅支持 DWS、DLI 类型数据表查看概要）。

单击“更新”，可更新概要信息。

步骤 6 在“数据预览”页签，查看数据表脱敏后的效果。

步骤 7 在“变更记录”页签，查看数据表变更详情。

----结束

3.7.1.4 标签管理

标签是用来标识数据的业务含义，是相关性很强的关键字，可以帮助您对资产进行分类和描述，以便于检索。

为方便管理技术资产，可以从业务角度定义标签，并与技术资产关联，比如标识某个表是 SDI 贴源数据层、DWI 数据整合层等。

管理标签

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-578 选择数据目录

2. 选择“数据地图 > 标签管理”。
3. 单击“新建”，新建标签。

- **标签名称**：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。
- **描述**：标签的描述信息，长度不能超过 255 个字符。

4. 勾选标签，单击“删除”，可删除标签。
5. 单击标签后的“编辑”，可修改标签描述。

标识数据：添加标签

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-579 选择数据目录

2. 选择“数据地图 > 数据目录”。
3. 在资产搜索输入框输入需要添加标签的数据的关键字，然后单击“搜索”，搜索结果以列表方式显示。
4. 勾选需要添加标签的资产，单击右上角“标识”。在添加标识对话框中配置标签。

图3-580 添加标识

5. 选择标识种类为标签，并配置标签，单击“确定”提交。

说明

此处支持全新添加标签，也支持选择已有标签。已有标签来源于管理标签。

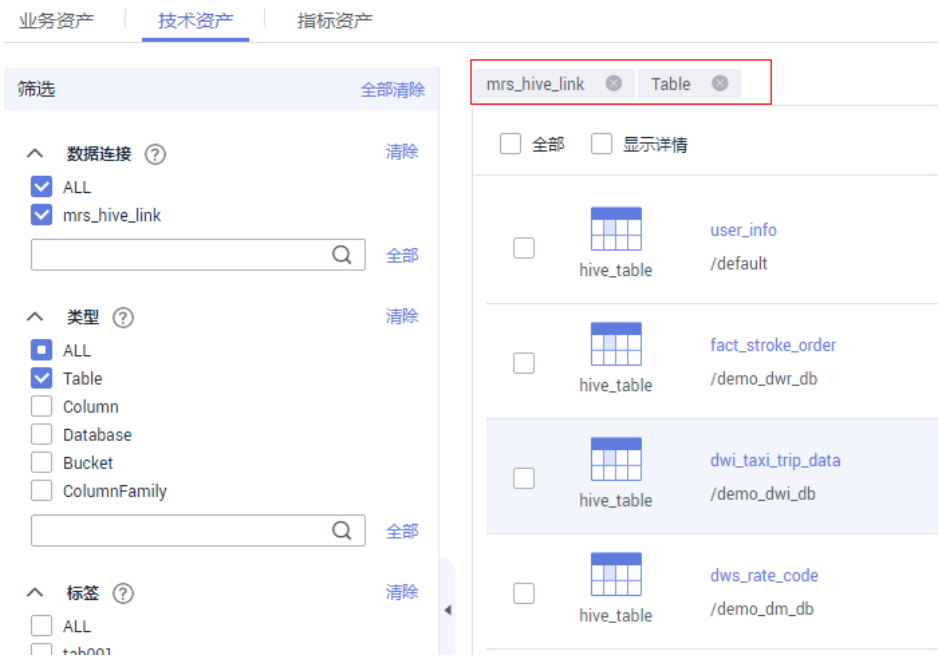
查看详情

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-581 选择数据目录

2. 单击相应工作空间的“数据目录”，系统跳转至“数据目录”页面。
3. 选择“数据地图 > 数据目录”，在“技术资产”页签，选择“数据连接”，类型选择“Table”。筛选出该数据连接下的所有表，参考图 3-582。

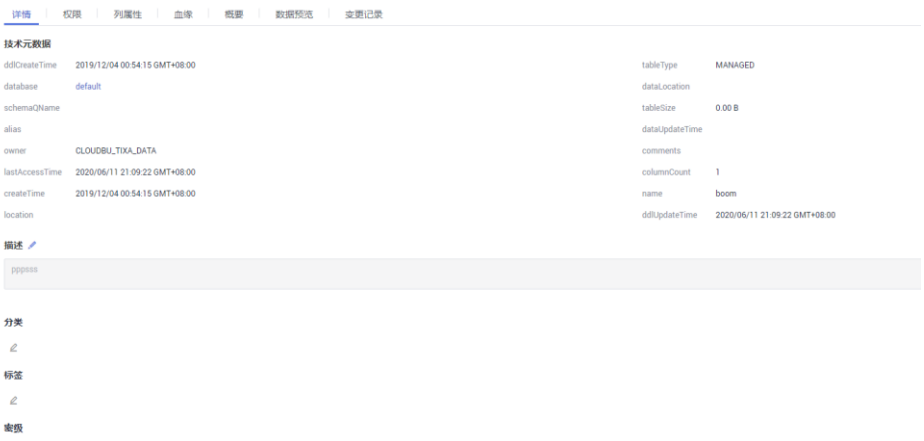
图3-582 筛选表



4. 单击相应的表名称，即可跳转至表详情页面，参考图 3-583。

   表详情页为用户展示表的详情、权限信息、列属性、血缘信息、概要、数据预览和变更记录。

图3-583 表详情页



数据预览

用户可以通过表的数据预览模块，预览当前表的业务数据。根据列的分类信息，支持对预览的数据进行实时脱敏。

- 数据预览支持的数据源类型：DWS、DLI、Hive、MySQL。

- 列的分类信息支持在新建采集任务时自动设置和在数据分类菜单中手动添加两种方式。其中仅 DWS、DLI 支持新建采集任务时自动设置分类。

3.7.2 数据权限

3.7.2.1 数据权限简介

为确保数据使用安全可控，使用数据表需要先申请权限。数据权限模块为用户提供便捷的权限管控能力，提供可视化申请审批流程，并可以进行权限的审计和管理。提高数据安全的同时，还可以方便用户进行数据权限管控。

3.7.2.2 数据目录权限

本章节主要介绍数据目录权限管理。

约束与限制

- 仅管理员角色的用户支持创建、删除、修改数据目录权限规则和设置数据目录权限生效状态。
- 开发者、运维者和访客角色的用户仅支持查看数据目录权限规则和规则列表。

管理数据目录权限规则

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

   图3-584 选择数据目录

2. 选择“数据权限 > 数据目录权限”，单击“新建”，配置数据目录权限规则。
   a. 规则名称：设置数据权限规则的名称。
   b. 类型：当前支持从标签、密级和分类的维度进行过滤筛选。

c. 范围：选择实际的标签、密级和分类。

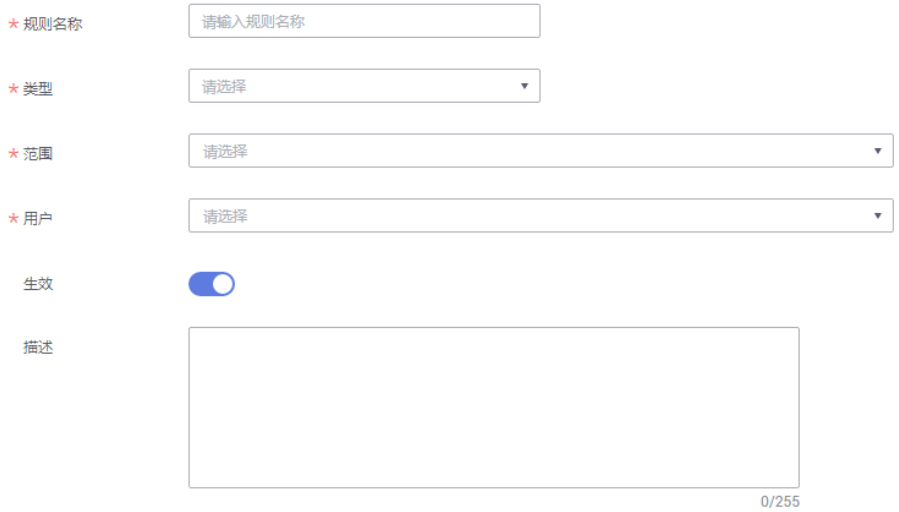
d. 用户：配置的数据目录权限规则所适配的用户。

e. 生效：打开，表示该数据目录权限规则生效。反之，不生效。

说明

数据目录权限规则生效后，仅该数据目录权限规则所适配的用户，可管理限定标签或者分类的数据资产。例如设置类型为标签，范围选择 test，用户设置为 A，当开启权限规则后，A 用户只可管理 test 标签的资产。

图3-585 新建规则



3. 在数据权限规则列表中，选择对应规则后的编辑和删除，可修改和删除数据权限规则。

3.7.2.3 数据表权限

用户可以在“我的权限”页面，查看工作空间内自己拥有的表和列权限，并对表和列的权限进行申请或交还。

管理员角色的用户具备管理“用户权限”的功能，即管理员可查看已在该工作空间内申请过权限的所有用户的资源权限。

申请表/列权限

说明

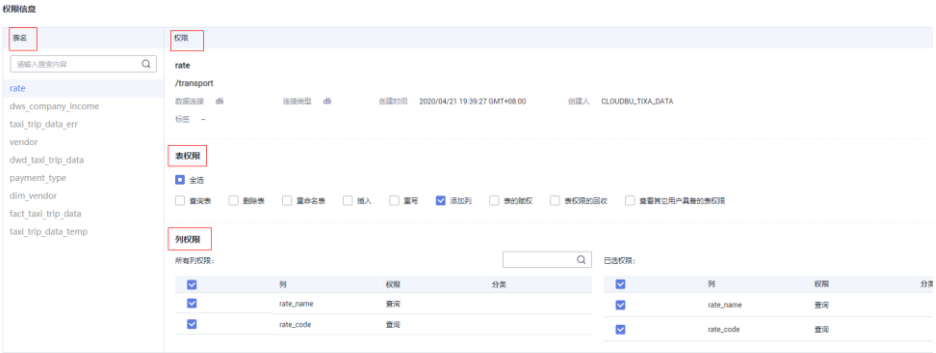
- 当前版本仅支持 DLI 数据表权限控制。
- 因申请表/列权限，需要审批人审批后方生效。所以申请表/列权限前，请先参见•管理审批人 新建审批人。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-586 选择数据目录

2. 选择“数据权限 > 数据表权限”，在“我的权限”页签中单击“申请”。
3. 输入使用场景说明，选择对应数据连接、数据库和数据表。
4. 选择需要申请的表/列权限。
   - 申请单张表/列权限。
     勾选自己当前无权限但需要使用的表权限/列权限。
   - 申请多张表/列权限。
     批量选择多张表后，在权限信息页面依次勾选需要使用的表/列权限。

图3-587 申请表/列权限信息



5. 单击“确定”，系统弹出提交对话框。配置审批人后，单击“确定”。
6. 等待审批人审批。待审批人审批后，权限即生效。

管理自有表权限

当用户需要对已申请的表/字段权限进行管理，包含查看、编辑和交还权限，请参见本节进行操作。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

   图3-588 选择数据目录

2. 选择“数据权限 > 数据表权限”，在“我的权限”页签中，支持如下操作：
   - 操作 > 查看，查看用户已申请的权限详情。
   - 操作 > 编辑，可修改用户已申请的数据表权限。
   - 操作 > 交还，可交还用户已申请的数据表权限。

   图3-589 管理表权限

审计用户权限

管理员可在“用户权限”页面查看同一工作空间内，分别有哪些帐号拥有表和字段的权限，并可回收不必要的表和字段的权限，也可对用户进行批量授权。

说明

仅空间管理员可审计用户权限，包含查看用户列表、回收用户权限、对用户进行授权。

- 查看拥有表权限的帐号和对应的资产列表

  选择“数据表权限 > 用户权限”，查看同一工作空间内，已申请表权限的帐号。

  图3-590 查看拥有表权限的帐号

- 回收用户的资产权限
  - 选择“数据表权限 > 用户权限”，单击帐号后操作列的“回收”，可回收该帐号所有的资产权限。
  - 选择“数据表权限 > 用户权限”，勾选用户名前的复选框，单击左上角“回收”，支持批量回收用户资产权限。

  图3-591 回收用户的资产权限

- 对用户授权

  图3-592 授权

- 在资产上管理用户的权限

  选择“数据表权限 > 用户权限”，单击帐号前的下拉列表，展开该用户所拥有的资产。单击对应特定资产操作列的“查看”、“编辑”和“回收”，完成在资产上管理用户的权限。

  图3-593 基于资产管理用户权限

3.7.2.4 审批中心

约束与限制

仅管理员角色的用户支持管理审批人，可新建和删除审批人。

审批管理

用户可在审批中心页面，查看自己提交的申请及进度，查看待自己审批的申请，查看已审批的历史记录并对审批人进行管理。

- 审批人管理

  选择“数据权限 > 审批中心”，在“审批人管理”页签“新建”和“删除”审批人，如图 3-594。审批人数据来源于工作空间中添加的人。

  图3-594 管理审批人

- 待我审批
  a. 选择“数据权限 > 审批中心”，单击“待我审批”页签。

     在此页面查看当前需要用户审批的申请单。
  b. 单击操作栏的“审批”，查看申请单的详细信息并进行审批。
  c. 填写审批意见后，根据实际情况同意或拒绝该申请。
- 我已审批
  a. 选择“数据权限 > 审批中心”，单击“我已审批”页签。
  b. 单击操作栏中的“查看”，即可查看申请单的审批记录和申请内容等详细信息。
- 我的申请
  a. 选择“数据权限 > 审批中心”，单击“我的申请”页签。
  b. 单击操作栏中的“查看”，即可查看申请单的详细信息。
  c. 单击操作栏中的“重新申请”，即可重新授权。

3.7.3 数据安全（待下线）

3.7.3.1 数据安全简介

应用背景

数据安全为数据湖提供数据生命周期内统一的数据使用保护能力。通过敏感数据识别、分级分类、隐私保护、资源权限控制、数据加密传输、加密存储、数据风险识别

以及合规审计等措施，帮助用户建立安全预警机制，增强整体安全防护能力，让数据可用不可得和安全合规。

说明

在已上线数据安全组件的区域，数据安全功能已由数据安全组件提供，不再作为数据目录组件能力。

功能模块

数据安全包括：

- 数据密级

  对数据进行等级划分，方便数据的管理。
- 数据分类

  基于数据密级，可以进行数据分类，来有效识别数据库内的敏感数据。
- 脱敏策略

  基于数据分类，可以通过创建脱敏策略，实现数据资产的脱敏和隐私保护。

3.7.3.2 数据密级

本章主要介绍数据密级管理，包括密级的创建、删除和调整优先级。

只有在创建密级之后，您才可以创建数据分类，进而创建脱敏策略进行数据脱敏。

前提条件

无。

进入数据密级管理页面

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-595 选择数据目录

2. 选择“数据安全 > 数据密级”，用户可以在该页面新建、管理和删除分级，也可以调整分级的优先级。
   - 创建分级：单击“数据密级”页签左上角的“新建”，输入名称和描述。
   - 删除：在“数据密级”页签，勾选不需要的分级，单击左上角的“删除”。
   - 调整优先级：在“数据密级”页签，单击相应分级后的上移（提高优先级）和下移（降低优先级）。

3.7.3.3 数据分类

本章主要介绍如何创建数据分类规则。

只有在创建数据分类规则之后，您才可以创建数据脱敏策略进行数据脱敏。

前提条件

数据密级定义已完成，请参见 3.7.3.2 数据密级。

新建分类规则

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-596 选择数据目录

2. 选择“数据安全 > 数据分类”，在“分类规则”页签中，单击“新建”。

   系统弹出“新建分类”对话框，填写相关配置，完成创建分类规则。支持按模板创建（内置）规则和自定义规则两种方式。

图3-597 配置分类规则

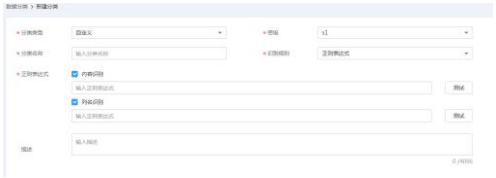


表3-340 配置分类规则参数说明

配置	说明
分类类型	即规则所属分类，支持内置（按模板添加）和自定义添加。
密级	对配置的数据进行等级划分。如果现有的分级不满足需求，请进入数据密级管理页面进行设置，详情请参见 3.7.3.2 数据密级。
分类模板	分类类型选择“内置”，呈现此参数。如果选择“内置”，用户可以根据实际需要选择系统内置的敏感数据识别定义模板，例如：时间、手机号、车牌号。
分类名称	• 分类类型选择“内置”，分类名称自动关联分类模板生成。 • 分类类型选择“自定义”，用户可以自行填写分类名称。 说明

配置	说明
	定义数据分类规则，名称必须唯一。
识别规则	分类类型选择“自定义”，呈现此参数，支持正则表达式。
正则表达式	• 内容识别：提供的数据识别方式之一，自定义正则表达式。 • 列名识别：提供字段名精确匹配和模糊匹配方式，支持多个字段匹配。
描述	对当前规则进行简单描述。

新建分组

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

   图3-598 选择数据目录

2. 选择“数据安全 > 数据分类”，在“分组”页签中，单击“新建”。
   系统弹出“新建分组”对话框，填写相关配置，单击“确定”，完成创建分组。
   参数设置参考表 3-341，并勾选左侧列表中的分类规则。
   用户所勾选的规则将显示在右侧列表中。

表3-341 参数配置表

配置	说明
名称	规则组名称只能包含中文、英文字母、数字和下划线。

配置	说明
描述	为更好的识别规则组，此处加以描述信息。描述信息长度不能超过4096 个字符。

3.7.3.4 脱敏策略

本节介绍如何创建数据脱敏策略，然后在数据目录中进行脱敏查询。

前提条件

- 数据分类规则已创建，数据分类规则的创建请参见 3.7.3.3 数据分类。
- 数据连接，数据表已创建成功，敏感数据已被数据目录采集。

创建脱敏策略

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

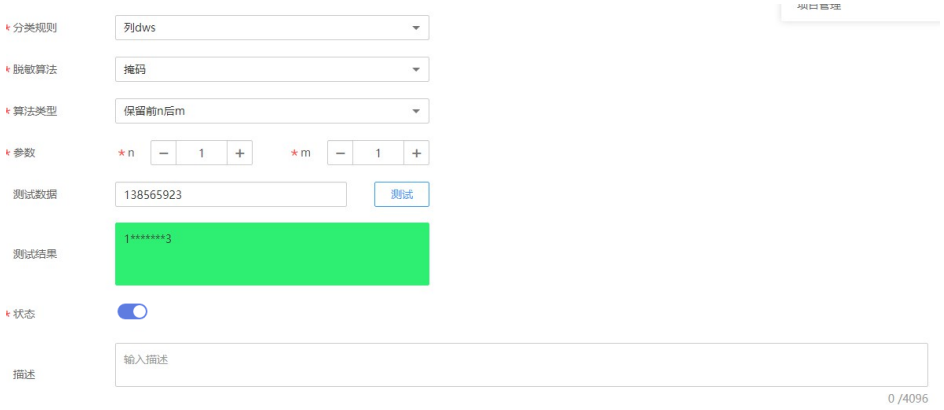
图3-599 选择数据目录

2. 选择“数据安全 > 脱敏策略”，在“脱敏策略”页面中，单击“新建”。
3. 绑定分类规则，配置脱敏算法并适配对应的算法类型。脱敏算法包含掩码，截断和哈希。每种脱敏算法对应多种算法类型，请根据产品界面进行选择，这里不再赘述。配置完成后单击“确定”。

说明

已被绑定脱敏算法的分类规则不支持被重复绑定。

图3-600 新建脱敏



4. 适配脱敏算法后，支持用户在线进行测试。输入测试数据，单击“测试”，在测试结果文本框中进行验证。
5. 开启或关闭状态，只有启用状态下的脱敏策略才可生效。

查看数据脱敏效果

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-601 选择数据目录

2. 选择“数据地图 > 数据目录”。
3. 在资产搜索结果列表，搜索脱敏后的数据表，进入数据表详情页面。
4. 单击“数据预览”，查看数据脱敏后的效果。

3.7.4 元数据采集

3.7.4.1 元数据简介

按照传统的定义，元数据（Metadata）是关于数据的数据。元数据打通了源数据、数据仓库、数据应用，记录了数据从产生到消费的全过程。元数据主要记录数据仓库中模型的定义、各层级间的映射关系、监控数据仓库的数据状态及 ETL 的任务运行状态。在数据仓库系统中，元数据可以帮助数据仓库管理员和开发人员非常方便地找到他们所关心的数据，用于指导其进行数据管理和开发工作，提高工作效率。

将元数据按用途的不同分为两类：技术元数据（Technical Metadata)和业务元数据（Business Metadata）。

- 技术元数据是存储关于数据仓库系统技术细节的数据，是用于开发和管理数据仓库使用的数据。
- 业务元数据从业务角度描述了数据仓库中的数据，它提供了介于使用者和实际系统之间的语义层，使得不懂计算机技术的业务人员也能够“读懂”数据仓库中的数据。

元数据管理模块是数据湖治理的基石，支持创建自定义策略的采集任务，采集数据源中的技术元数据。支持自定义业务元模型、批量导入业务元数据、关联业务和技术元数据、全链路的血缘管理和应用。

3.7.4.2 任务管理

本章主要介绍如何通过配置元数据采集策略新建采集任务，不同类型的数据源对应的采集策略不尽相同。元数据管理依据采集任务的配置策略，采集对应的技术元数据信息。

前提条件

元数据采集支持丰富的数据源类型，对于 DWS、DLI、MRS HBase、MRS Hive、RDS（MySQL）、RDS（PostgreSQL）和 ORACLE 类型的数据源，首先需要在管理中心创建数据连接。

新增采集任务

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-602 选择数据目录

2. 选择“元数据采集 > 任务管理”。
3. 选择采集任务所归属的目录。如果未新建目录请参见图 3-603 创建进行。

图3-603 新建采集任务的归属目录

4. 单击页面上方“新建”或者右键单击任务菜单，单击“新增任务”，在弹出的对话框中，配置相关参数，新建采集任务。

   新建任务有如图 3-604 所示的两个入口。

图3-604 新建采集任务入口

a. 配置基本参数，参考表 3-342。

表3-342 基本配置说明

参数名	说明
任务名称	采集任务的名称，只能包含中文、英文字母、数字和下划线，且长度不能超过 62 个字符。
描述	为更好的识别采集任务，此处加以描述信息。描述信息长度不能超过 255 个字符。
选择目录	采集任务的存储目录，可选择已创建的目录。目录创建请参见图 3-603。

b. 配置数据源信息，参考表 3-343。

表3-343 数据源信息参数说明

参数名	说明
数据源类型	从下拉列表中选择数据源类型。 说明 元数据采集支持丰富的数据源类型，对于 DWS、DLI、MRS HBase、MRS Hive、RDS（MySQL）、RDS（PostgreSQL）和 ORACLE 类型的数据源，首先需要在管理中心创建数据连接。
OBS 桶	选择待采集数据归属的 OBS 桶，仅数据源类型为 OBS 时，呈现此参数。
数据连接	• 所选数据连接类型中已创建数据连接，支持从下拉列表中选择。 • 所选数据连接类型中未创建数据连接，请单击“新建”，创建

参数名	说明
	新的数据连接。
OBS 路径	选择待采集数据在 OBS 桶中的存储路径，仅数据源类型为 OBS 时，呈现此参数。
采集范围	选择待采集数据的采集范围，仅数据源类型为 OBS 时，呈现此参数，原因是 obs 桶中是分目录层级的。 • 选择“当前文件夹”，采集任务仅采集 obs 路径中设置的文件夹下的对象。 • 选择“当前文件夹和所有子文件夹”，采集任务会采集 obs 路径中设置的文件夹下所有的对象，包括其子文件夹下的对象
采集内容	选择待采集数据的采集内容，仅数据源类型为 OBS 时，呈现此参数，原因是 obs 桶中是分目录层级的。 • 选择“文件夹和对象”，采集任务采集文件夹和对象。 • 选择“ 文件夹”，采集任务仅采集文件夹。
数据库和 schema	仅数据源类型为 DWS 时，呈现此参数。 单击数据库和 schema 后的“设置”，设置采集任务扫描的数据库和 schema 范围。当不进行设置时，默认选择该数据连接下的所有数据库和 schema 。 单击“清除”，可对已选择的数据库和 schema 进行修改。
命名空间	仅数据源类型为 MRS HBase 时，呈现此参数。 单击命名空间后的“设置”，设置采集任务扫描的命名空间范围。当不进行设置时，默认选择该数据连接下的所有命名空间 。 单击“清除”，可对已选择的命名空间进行修改。
数据库 数据表	呈现待采集的数据库和数据表。 • 单击数据库后的“设置”，设置采集任务扫描的数据库范围。当不进行设置时，默认选择该数据连接下的所有数据库 。 • 单击数据表后的“设置”，设置采集任务扫描的数据表范围。当不进行设置时，默认选择数据库下的所有数据表。针对数据连接类型为 Mysql、Oracle 和 DLI 的数据表，支持按照正则表达式过滤需要采集的表。 • 当数据库和数据表均不设置时，则采集任务扫描的数据范围为该数据连接下的所有数据表。 • 单击“清除”，可对已选择的数据库和数据表进行修改。
选择图	仅数据源类型为 GES 时，呈现此参数。 选择存储了以“关系”为基础的结构数据的图。
选择集群	仅数据源类型为 CSS 时，呈现此参数。 选择待采集数据存储的 CSS 集群。 您也可以单击“新建”，创建 CSS 集群，创建完成后单击“刷

参数名	说明
	新”，选择新建的 CSS 集群即可。
绑定 Agent	管理 CloudTable/GES/CSS 类型的数据连接，请选择 CDM 集群提供的 Agent。 用户也可以单击“新建”，创建新的 Agent，创建完成后单击“刷新”，选择新的 Agent 即可。
索引	仅数据源类型为 CSS 时，呈现此参数。 用于存储 Elasticsearch 的数据，类似关系型数据库的 Database。是一个或多个分片分组在一起的逻辑空间。

c. 元数据采集参数配置，参考表 3-344。

表3-344 元数据采集参数说明

参数名	说明
数据源元数据已更新	当数据连接中元数据发生变化时，通过配置更新策略，设置数据目录中元数据的更新方式。 需要注意的是配置的更新、删除策略是作用在用户配置的数据库、数据表的范围内的。 • 勾选“仅更新数据目录中的元数据”：采集任务仅更新数据目录已经采集到的元数据 • 勾选“仅添加新元数据”：采集任务仅采集数据源中存在，但是数据目录中不存在的元数据 • 勾选“更新数据目录中的元数据、添加新元数据”：采集任务全量同步数据源中的元数据 • 勾选“忽略更新、添加操作”：不采集数据源中的元数据
数据源元数据已删除	当数据连接中元数据发生变化时，通过配置删除策略，设置数据目录中元数据的更新方式。 • 勾选“从数据目录中删除元数据”：当数据源中的某些元数据已经被删除，数据目录中也将同步删除对应的元数据 • 勾选“忽略删除”：当数据源中的某些元数据已经被删除，数据目录中不同步删除对应元数据。

d. 勾选数据概要时的参数配置，参考表 3-345。

表3-345 数据概要参数说明

参数名	说明
基于全量数	基于已采集的全量数据在数据目录中生成数据概要。

参数名	说明
据	适用于数据量较少（100W 以下）的情况。
基于采样数据，采样数量为 *x* 条	基于已采集的全量数据在数据目录中生成数据概要。 适用于数据量较多的情况。
基于全量数据，随机取*x*%的数据	基于已采集的全量数据在数据目录中生成数据概要。 适用于数据量较多的情况。
DLI 队列	选择获取 profile 数据，执行 DLI　SQL 用的队列。 勾选“采集唯一值”表示只统计已采集的表中的唯一值的个数，并在数据目录中的概要页签呈现。
数据格式	当存储在 OBS 桶中的数据为 CSV 格式，请依据数据的实际属性进行勾选是否有表头，是否自定义分隔符，是否自定义引用字符，是否自定义转义字符。
日期格式	当存储在 OBS 桶中的数据为 CSV 格式，请依据实际属性配置日期格式，以免影响数据被错误解析。
时间戳格式	当存储在 OBS 桶中的数据为 CSV 格式，请依据实际属性配置时间戳格式，以免影响数据被错误解析。

e. 数据分类配置说明（仅当数据目录组件中具备数据安全功能时，支持配置该选项；当前暂不支持关联独立数据安全组件中的敏感数据识别规则）

- 数据分类：勾选此项参见 3.7.3.3 数据分类新建分类规则组或者选中已有分类规则组，实现自动识别数据并添加分类。
- 数据分级：勾选“根据数据分类结果更新数据表密级”，表示可根据匹配的分类规则中，将密级最高的设置为表的密级。
- 数据同步：勾选“手动同步分类结果”，表示“数据目录 > 数据目录 > 列属性”中呈现的数据列，在采集任务执行完毕后，不会自动添加分类和密级属性。需要用户前往“元数据采集 > 任务监控 ”页面，找到任务实例，选择“操作 > 更多 > 扫描结果”，查看采集任务的执行结果，确认分类结果是否匹配。勾选分类匹配字段前的复选框，单击“同步”，即可将分类和密级属性手动同步到资产。

说明

仅 DWS、DLI 数据源支持创建采集任务时添加数据分类，实现自动识别。另外，只可给数据表的列和 OBS 对象添加分类。

5. 单击“下一步”，选择调度方式，支持单次调度和周期调度两种方式。

   单次调度：超时时间表示如果任务运行的时长超过了设置的超时时间，任务会被认定运行失败。

   周期调度的相关参数配置请参见表 3-346。

说明

1. 单次调度会产生手动任务的实例，手动任务的特点是没有调度依赖，只需要手动触发即可。
2. 周期调度会产生周期实例，周期实例是周期任务达到启用调度所配置的周期性运行时间时，被自动调度起来的实例快照。
3. 周期任务每调度一次，便生成一个实例工作流。用户可以对已调度起的实例任务进行日常的运维管理，如查看运行状态，对任务进行终止、重跑等操作。

表3-346 配置周期调度参数

参数名	说明
生效日期	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
开始时间	周期调度开始的具体时间，与生效日期中的开始时期配合使用。
间隔时间	两次周期调度之间的间隔时间。 即使上一次调度任务实例未结束，从上次调度开始时间达到间隔时间后，新的调度任务实例也会开始。当前采集任务支持多实例并发运行。
结束时间	周期调度结束的具体时间，与生效日期中的结束时期配合使用。
超时时间	单次任务实例的运行超时时间，如果运行时长超过了此处设置，任务会被认定运行失败。
启动调度	勾选复选框，则表示立即启动此调度任务。

6. 单击“提交”，采集任务创建成功。

管理采集任务

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-605 选择数据目录

2. 选择“元数据采集 > 任务管理”。

在采集任务页面，可查看所有已创建的采集任务。

表3-347 管理采集任务

参数名	说明
任务名称	采集任务的名称。 单击采集任务名称，可查看该采集任务的采集策略和调度属性。
数据源类型	数据连接的名称。
调度状态	显示采集任务的调度方式，单击，可进行筛选。
调度周期	显示采集任务的调度频率，单击，可进行筛选。
描述	展示采集任务的描述信息。
创建人	展示采集任务的创建人。
最近运行时间	展示采集任务的最近运行时间。
操作	对已创建的采集任务可进行如下操作： • 编辑：支持对采集任务（状态为已启动、未启动、运行失败）的采集策略强相关参数进行修改，不支持修改数据源类型。 • 运行：单击“运行”，可运行此采集任务，并可在“任务监控”页面查看其状态和相关日志信息。 • 启动调度：当其状态为“已停止”，则可重新启动调度。

参数名	说明
	• 停止调度：当调度状态为“调度中”，则可停止调度。

3.7.4.3 任务监控

监控元数据采集任务运行情况，查看采集日志，支持重跑采集任务。

在数据目录页面，选择“元数据采集 > 任务监控”。在任务监控页面，对采集任务进行监控，参考表 3-348。

表3-348 监控采集任务

参数名	说明
任务名称	采集任务的名称。
实例状态	实例（即采集任务）的状态。 • 成功 • 部分成功 • 执行中 • 失败 • 运行异常 • 暂停：因管理面升级，监控任务暂停，升级完成后监控继续执行。
调度方式	展示采集任务的调度状态，分为单次调度和周期调度。
调度周期	展示采集任务的调度周期。
开始时间	重跑采集任务的启动时间。
结束时间	重跑采集任务的结束时间。
运行时间	采集任务的运行时间。
操作	对被纳入监控的采集任务可进行如下操作： • 重跑：实例状态为失败和成功状态的实例，支持重跑。 • 日志：查看实例日志。 说明 单击“日志”，可实时查看元数据采集、数据概要、数据分类三类任务的运行日志。 • 更多 > 取消：创建采集任务的时候，配置“数据分类”为“手动同步分类结果”时，才可进行此操作。状态为执行中的实例，单击取消，可终止重跑此实例。 • 更多 > 扫描结果：创建采集任务的时候，配置“数据分类”为“手动同步分类结果”时，才可进行此操作。可用于查看采集任务实例执行结果，确认分类结果是否匹配。勾选分类匹配字段前的

参数名	说明
	复选框，单击“同步”，即可将分类和密级属性手动同步到资产。

3.7.5 使用教程

3.7.5.1 开发一个增量元数据采集任务

配置、运行采集任务是构建数据资产的前提，下面举例说明如何通过配置采集任务达到灵活采集元数据的目的。

场景一：仅添加新元数据

用户的数据库中新增的数据表，采集任务仅采集新增的表。

例如新增 table4 的情况下：

- 采集前的数据表元数据：table1，table2，table3
- 采集后的数据表元数据：table1，table2，table3，**table4**

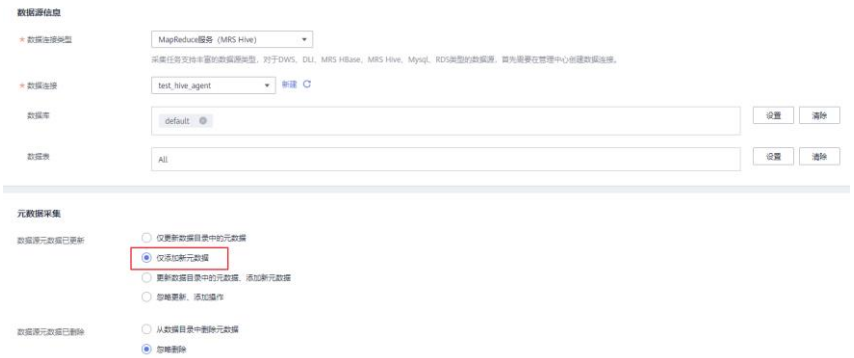
按照下面的配置，采集任务仅会采集 table4。（前提：table1-table3 已经在数据目录中）

步骤 1 进入 DataArts Studio 控制台首页的数据目录模块。

步骤 2 单击左侧导航的“任务管理”。

步骤 3 单击“新建”。

步骤 4 配置任务信息，如下图所示。



步骤 5 单击“下一步”，配置调度属性如下图所示。

步骤 6 单击“提交”，完成采集任务的创建。

步骤 7 单击任务管理列表中的“运行”或“启动调度”，跳转到任务监控页面并查看任务状态。

----结束

场景二：更新数据目录中的元数据，添加新元数据

用户的数据库中新增了数据表，采集数据源中指定的所有表。

例如新增 table4 的情况下：

- 采集前的数据表元数据：table1，table2，table3
- 采集后的数据表元数据：**table1**，**table2**，**table3**，**table4**

按照如下配置，采集任务会采集 default 下所有的表（table1-table4）。

步骤 1 进入 DataArts Studio 控制台首页的数据目录模块。

步骤 2 单击左侧导航的“任务管理”。

步骤 3 单击“新建”。

步骤 4 配置任务信息，如下图所示。

步骤 5 单击“下一步”，配置调度属性如下图所示。

步骤 6 单击“提交”，完成采集任务的创建。

步骤 7 单击任务管理列表中的“运行”或“启动调度”，跳转到任务监控页面并查看任务状态。

----结束

场景三：仅更新数据目录中的元数据

用户的数据库中数据表有新增的情况，采集任务仅采集数据目录中已经存在的表。

例如新增 table4 的情况下：

- 采集前的数据表元数据：table1，table2，table3
- 采集后的数据表元数据：**table1**，**table2**，**table3**

按照如下配置，采集任务仅采集 table1，table2 和 table3。

步骤 1 进入 DataArts Studio 控制台首页的数据目录模块。

步骤 2 单击左侧导航的“任务管理”。

步骤 3 单击“新建”。

步骤 4 配置任务信息，如下图所示。

步骤 5 单击“下一步”，配置调度属性如下图所示。

步骤 6 单击“提交”，完成采集任务的创建。

步骤 7 单击任务管理列表中的“运行”或“启动调度”，跳转到任务监控页面并查看任务状态。

----结束

场景四：更新数据目录中的元数据，添加新元数据 ，并从数据目录中删除元数据

用户的数据库中数据表有删除的情况，采集任务能够删除数据目录中对应的数据表。

例如数据库删除 table1 的情况下：

- 采集前的数据表元数据：table1，table2，table3

- 采集后的数据表元数据：**table2，table3**

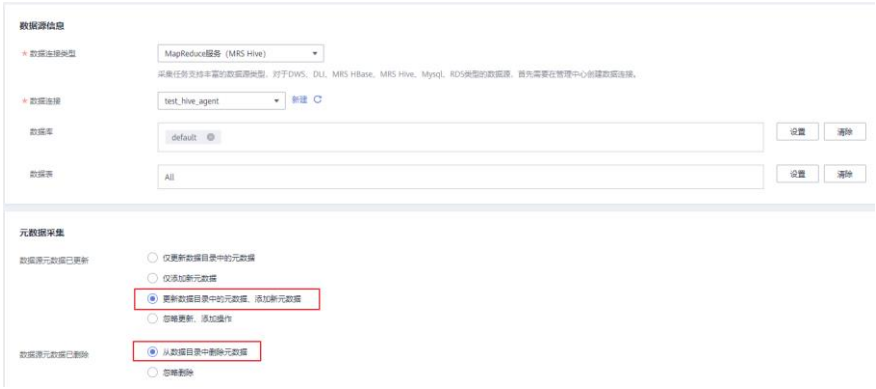
按照如下配置，采集任务会删除数据目录中的 table1。

步骤 1 进入 DataArts Studio 控制台首页的数据目录模块。

步骤 2 单击左侧导航的“任务管理”。

步骤 3 单击“新建”。

步骤 4 配置任务信息，如下图所示。



步骤 5 单击“下一步”，配置调度属性如下图所示。

步骤 6 单击“提交”，完成采集任务的创建。

步骤 7 单击任务管理列表中的“运行”或“启动调度”，跳转到任务监控页面并查看任务状态。

----结束

3.7.5.2 通过数据地图查看数据血缘关系

3.7.5.2.1 方案概述

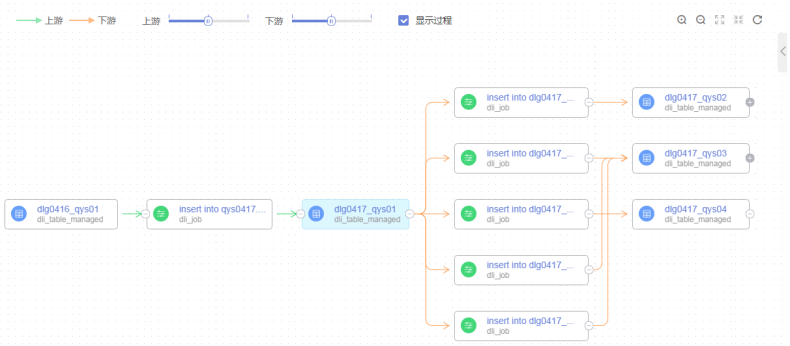
什么是数据血缘

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

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

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

图3-606 数据血缘关系示例



DataArts Studio 数据血缘实现方案

- 数据血缘的产生：

  在 DataArts Studio 平台，自动分析血缘是通过在数据开发模块中配置数据处理迁移类型的节点产生的，当前支持采集节点静态配置产生的血缘和部分节点实例上的血缘。详情请参见自动分析血缘。

  另外，DataArts Studio 平台还支持手动配置血缘方式，当用户手动配置血缘时，自动分析血缘将不生效。详情请参见手动配置血缘。

- 数据血缘的展示：

  当数据开发模块中的作业已完成血缘关系配置后，启动作业调度，并在数据目录模块进行元数据采集任务，则可以在数据目录模块可视化查看数据血缘关系。

3.7.5.2.2 配置数据血缘

在 DataArts Studio 平台，自动分析血缘是通过在数据开发模块中配置数据处理迁移类型的节点产生的，当前支持采集节点静态配置产生的血缘和部分节点实例上的血缘。详情请参见自动分析血缘。

另外，DataArts Studio 平台还支持手动配置血缘方式，当用户手动配置血缘时，自动分析血缘将不生效。详情请参见手动配置血缘。

自动分析血缘

自动分析血缘是通过在数据开发模块中配置数据处理迁移类型的节点产生的，当作业中包含如下节点时，系统支持自动解析血缘。

- **SQL 类型节点**

DataArts Studio 目前支持对 DLI SQL、DWS SQL 和 MRS Hive SQL 节点的血缘解析，可以支持多 SQL 解析及列级血缘解析，当语句中有临时表时，会自动在数据目录中创建相关的临时表实体。

- 3.5.9.10 DLI SQL
  - 支持解析 DLI 中表与表之间数据插入产生的血缘。
  - 支持通过建表语句产生的 OBS 文件到 DLI 表之间的血缘。
- 3.5.9.12 DWS SQL
  - 支持 Create table like/as 等 DDL 操作产生的 DWS 表之间的血缘。
  - 支持 Insert into 等 DML 操作产生的 DWS 表之间的血缘。
- 3.5.9.14 MRS Hive SQL
  - 支持 Create table like/as 等 DDL 操作产生的 MRS 表之间的血缘。
  - 支持 Insert into/overwrite 等 DML 操作产生的 MRS 表之间的血缘。

- **数据集成类型节点**

  目前支持对 CDM Job 节点、ETL Job 节点和 OBS Manager 节点的血缘解析。

  - 3.5.9.3 CDM Job

    支持 MRS Hive、DLI、DWS、RDS、OBS 以及 CSS 之间表文件迁移所产生的血缘。
  - 3.5.9.23 ETL Job

    支持 DLI、OBS、MySQL 以及 DWS 之间的 ETL 任务产生的血缘。
  - 3.5.9.27 OBS Manager

    支持 OBS 之间目录和文件复制迁移产生的血缘。

说明

当前血缘解析能力，单条 sql 语句不支持 sql 中含有分号的场景。

手动配置血缘

在 DataArts Studio 数据开发中，用户也可以自己定义节点的输入、输出血缘关系。当用户手动配置血缘时，自动分析血缘将不生效。手动配置血缘不会影响作业的运行。

目前手动配置血缘时输入、输出数据源支持 DLI、DWS、Hive、CSS、OBS 和 CUSTOM。CUSTOM 即自定义类型，在手动配置血缘时，对于不支持的数据源，您可以添加为自定义类型。

支持手动配置血缘的节点类型如下所示，关于手动配置血缘的更多内容，请参见相关节点的详细介绍。

- 3.5.9.3 CDM Job
- 3.5.9.4 Rest Client
- 3.5.9.10 DLI SQL
- 3.5.9.11 DLI Spark
- 3.5.9.12 DWS SQL
- 3.5.9.13 MRS Spark SQL
- 3.5.9.14 MRS Hive SQL
- 3.5.9.15 MRS Presto SQL

- 3.5.9.16 MRS Spark
- 3.5.9.17 MRS Spark Python
- 3.5.9.23 ETL Job
- 3.5.9.27 OBS Manager

3.7.5.2.3 查看数据血缘

当数据开发模块中的作业已完成血缘关系配置后，启动作业调度，并在数据目录模块进行元数据采集任务，则可以在数据目录模块可视化查看数据血缘关系。

前提条件

已完成血缘关系的自动配置或手动配置，请参见 3.7.5.2.2 配置数据血缘。

启动作业调度

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3-607 选择数据开发

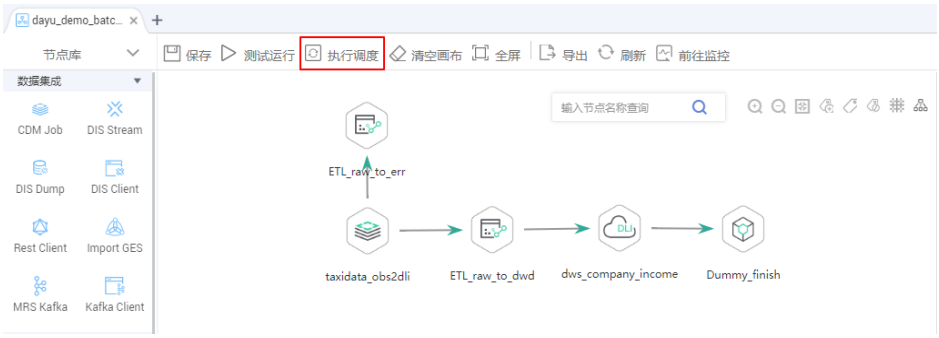
步骤 2 在数据开发控制台，单击左侧导航栏中的作业开发按钮，进入作业开发页面后，打开已完成血缘配置的作业。

步骤 3 在数据开发中，当作业进行“执行调度”时，系统开始解析血缘关系。

说明

测试运行不会解析血缘。

图3-608 作业调度



----结束

新建元数据采集任务

如果已创建元数据采集任务，此操作可跳过。

步骤 1 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-609 选择数据目录

步骤 2 请参见 3.7.4.2 任务管理，新建元数据采集任务。

----结束

查看数据血缘关系

步骤 1 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图3-610 选择数据目录

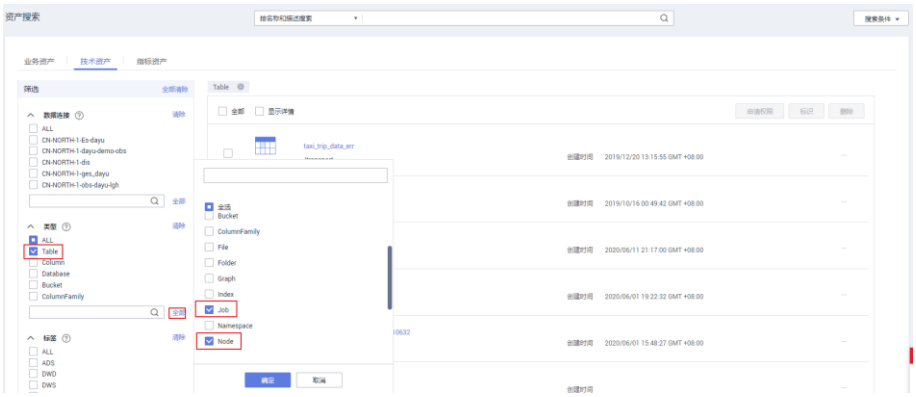
步骤 2 在“数据目录 > 技术资产”页面，可以对数据开发的作业、节点、表进行查询。

在“类型”筛选区域，单击“全部”按钮并勾选“Job”、“Node”和“Table”类型，然后单击“确定”。数据开发中的作业对应于 Job 类型，节点对应于 Node 类型，表对应于 Table 类型。

说明

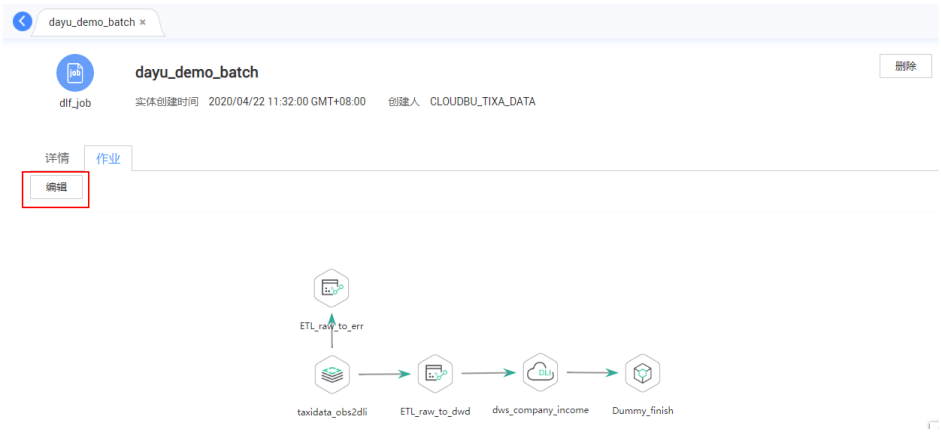
数据开发中的作业信息不属于任何一个数据连接，故如果在搜索条件中勾选数据连接，则查询不到结果。

图3-611 选择类型



步骤 3 在数据资产搜索结果中，类型名称末尾带“_job”的数据资产为作业，单击某一作业名称，可以查看该作业的详情。在作业的详情页面进入“作业”页签，单击“编辑”可跳转到数据开发的作业编辑页面。

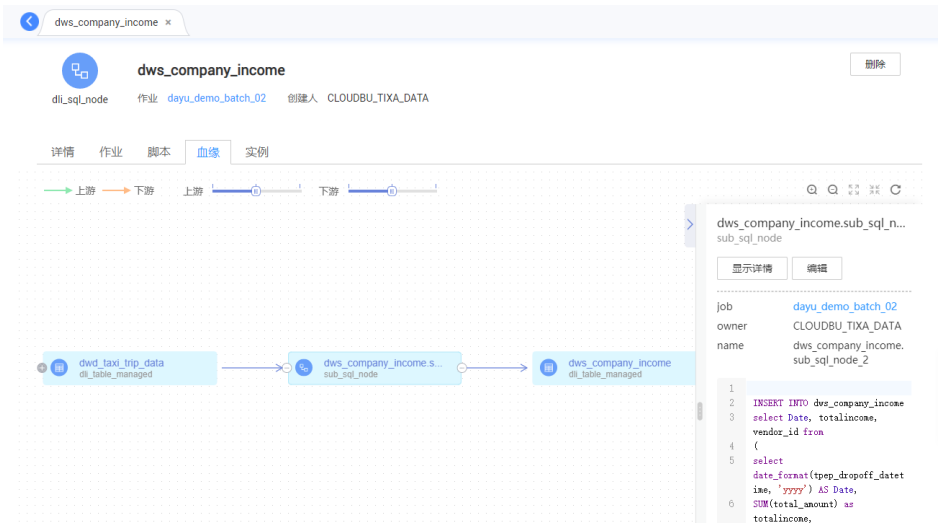
图3-612 查看作业



步骤 4 在数据资产搜索结果中，类型名称末尾带“_node”的数据资产为节点，单击某一节点名称，可以查看节点的详情。在节点（需是支持血缘的节点类型）详情页面，可以查看节点的血缘信息。

- 单击血缘图中节点左右两端“+”、“-”图标，可以进一步展开查看血缘的上下链路。
- 单击血缘图中的某一个节点，可以查看该节点的详情。
- 进入“作业”页签，单击“编辑”可跳转到数据开发的作业编辑页面。

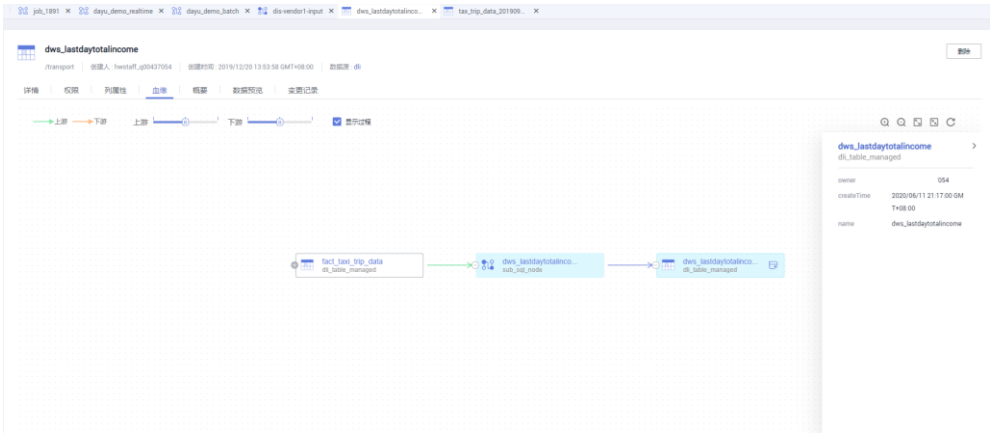
图3-613 查看节点血缘



步骤 5 在数据资产搜索结果中，图标为表格的数据资产为表，单击某一表名称，可以查看表的详情。在详情页面，可以查看表的血缘信息。

- 单击血缘图中表左右两端“+”、“-”图标，可以进一步展开查看血缘的上下链路。
- 单击血缘图中的某一个表，可以查看该表的详情。

图3-614 查看表血缘



----结束

3.8 数据服务

3.8.1 数据服务概览

DataArts Studio 数据服务旨在为企业搭建统一的数据服务总线，帮助企业统一管理对内对外的 API 服务。数据服务为您提供快速将数据表生成数据 API 的能力，涵盖 API 发布、管理、运维的全生命周期管理，帮助您简单、快速、低成本、低风险地实现微服务聚合、前后端分离、系统集成，向合作伙伴、开发者开放功能和数据。

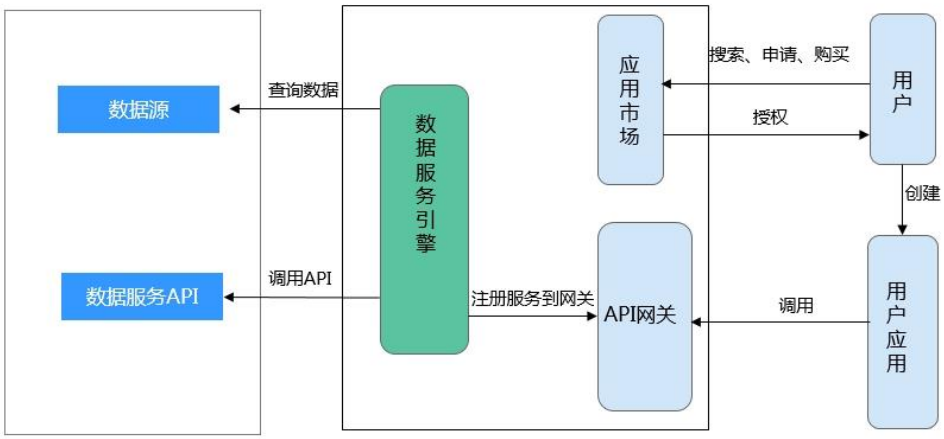
相对于数据共享交换或其他数据开放形式，使用数据服务进行数据开放具备如下优势：

- 统一接口标准，减少上层应用对接工作量。
- 将数据逻辑沉淀至数据平台，实现应用逻辑与数据逻辑解耦，在减少数据模型的重复开发的同时，避免数据逻辑调整带来的“散弹式修改”。
- 将数据逻辑相关的存储与计算资源下沉到数据平台，降低应用侧的资源消耗。
- 减少大量明细、敏感数据在应用侧的暴露，同时通过 API 审核发布、鉴权流控、动态脱敏等手段，提升数据安全能力。

值得注意的是，数据服务是通过将数据逻辑封装成统一标准的 Restful 风格 API 从而实现数据开放，适用于小批量数据的快速响应交互场景。如果为大量数据开放的场景，更适于通过数据共享交换或其他方案实现。

数据服务采用 Serverless 架构，您只需关注 API 本身的查询逻辑，无需关心运行环境等基础设施，数据服务会为您准备好计算资源，并支持弹性扩展，零运维成本。

图3-615 数据服务架构图



API 开放方使用流程

您作为 API 提供者，需要实现一个或一组 API 的开放，那么您需要先后完成以下工作：

1. 3.8.3.1 准备工作

   如果您需要使用数据服务，需要先 3.8.3.1.1 创建专享版集群。

   另外，在创建 API 前，您还需要 3.8.3.1.2 新建审核人。

2. 3.8.3.2 创建 API

   创建 API 即 3.8.3.2 创建 API 和 3.8.3.2.3 注册 API。其中，生成 API 支持两种方式（3.8.3.2.1 配置模式生成 API 和 3.8.3.2.2 脚本模式生成 API）。

3. 3.8.3.3 调试 API

   API 创建后需要验证服务是否正常，管理控制台提供了调试功能。

4. 3.8.3.4 发布 API

   只有将 API 发布后，API 才支持被调用。

5. 3.8.3.5 管理 API

   您可以根据您的需要，对已创建发布的 API 进行管理。

6. 3.8.3.6 流量控制

   为了保护后端服务的稳定的考虑，您可以对 API 进行流量控制。

API 调用方使用流程

您作为 API 调用者，需要实现一个 API 的调用，那么您需要完成以下工作：

1. 获取 API

   从服务目录获取需要调用 API。仅在 API 发布后，才支持被调用。

2. （可选）创建应用并获取授权

对于使用 APP 和 IAM 认证的 API，需要完成创建应用和将 API 授权给应用。在 API 调用过程中，使用所创建应用的密钥对（AppKey、AppSecret），数据服务根据密钥对进行身份核对，完成鉴权。

3. 调用 API

   API 调用者完成以上步骤后，可以进行 API 调用。

总览页面说明

在总览页用户可以看到丰富的监控数据视图。数据服务总览页面分别从 API 开放方和 API 调用方的视角，统计了 API 的相关度量数据。

图3-616 API 总览

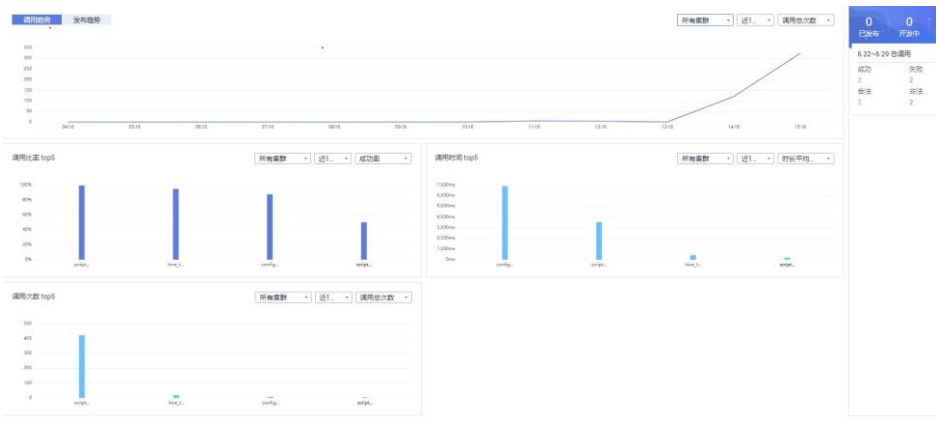


表3-349 API 开放方数据统计

界面	说明
发布趋势	从每天、每周、每月、每年的维度统计了 API 开放方发布的 API 数量。
调用趋势	从半天、每天、每周、每月的维度统计了 API 开放方发布的 API，被调用的次数。
调用比率 TOP5	统计 API 开放方所开放的 API，被调用的比率，包含成功率、失败率、合法率和非法率。
调用时间 TOP5	统计 API 开放方所开放的 API，被调用时长，支持统计的维度包含时长平均总值、成功时长平均总值，失败时长平均总值。
调用次数 TOP5	统计 API 开放方所开放的 API，被调用次数排名 TOP5 的，支持统计的维度包含调用总次数、成功次数、失败次数、合法次数和非法次数。
已发布	统计 API 开放方已成功发布至服务市场的 API 数量。
开发中	统计 API 开放方开发中的 API 数量。

界面	说明
申请者	统计 API 开放方已成功发布的 API，被申请调用的应用数量。
调用成功	统计 API 开放方已成功发布的 API，被应用成功调用的次数。
调用失败	统计 API 开放方已成功发布的 API，被应用调用失败的次数。
总调用次数	统计 API 开放方已成功发布的 API，被应用调用的总次数。

图3-617 调用总览

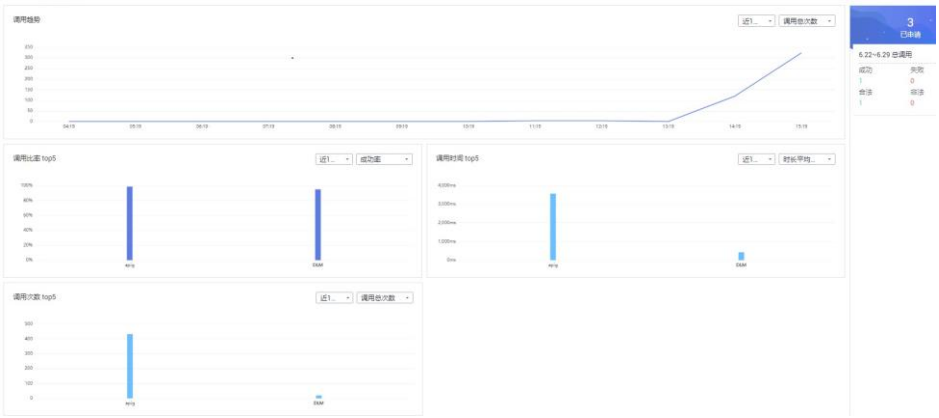


表3-350 API 调用方数据统计

界面	说明
调用趋势	从每天、每周、每月、每年的维度统计了 API 调用方调用的 API 数量。
调用比率	统计调用方最近 7 天内调用 API 的成功和失败比例。
已申请	统计 API 调用方，在数据服务平台申请调用的 API 数量。
调用成功	统计 API 调用方在数据服务平台成功调用 API 的次数。
总调用次数	统计 API 调用方在数据服务平台调用 API 的总次数。

3.8.2 规格说明

专享版规格

数据服务专享版的实例规格，如表 3-351 所示。

表3-351 专享版实例规格说明

实例规格	最大支持发布的 API 数量
small (X86 或 ARM)	500
medium (X86 或 ARM)	1000
large (X86 或 ARM)	2000

API 返回数据规格

数据服务适用于小批量数据的快速响应交互场景，不适用于将大量数据通过 API 的方式返回。当前通过数据服务 API 返回数据的规格如下表所示。

表3-352 API 的返回数据条数限制

API 分类	使用场景	数据源	默认规格
配置类 API	调试 API	DLI/MySQL/RDS/DWS	10
	调用 API	DLI/MySQL/RDS/DWS	100
脚本类 API	测试 SQL	-	10
	调试 API	DLI	• 默认分页：100 • 自定义分页：1000
		MySQL/RDS/DWS	• 默认分页：10 • 自定义分页：2000
	调用 API	DLI	• 默认分页：100 • 自定义分页：1000
		MySQL/RDS/DWS	• 默认分页：10 • 自定义分页：2000

3.8.3 开发 API

3.8.3.1 准备工作

3.8.3.1.1 创建专享版集群

本小节指导您顺利创建专享版实例，实例创建完成后，才能在数据服务专享版创建 API 并对外提供服务。

须知

如果需要创建、删除专享版集群或修改 API 配额，则需具备以下权限之一的账号才能进行操作：

- DAYU Administrator 并且拥有 VPCEndpoint Administrator 权限。
- Tenant Administrator 并且拥有 VPCEndpoint Administrator 权限。

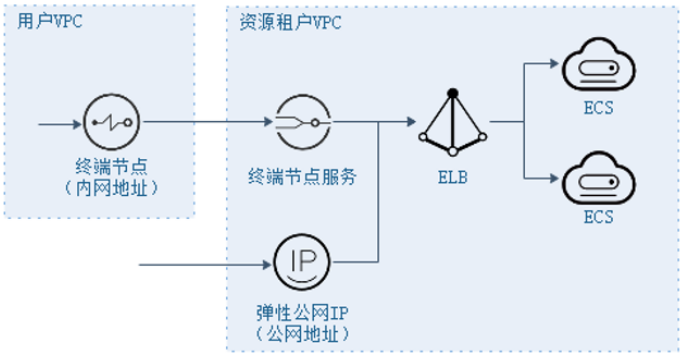
网络环境准备

如图 3-618 所示，专享版集群创建后，资源位于资源租户区，由 ELB 统一对集群节点进行负载均衡。

用户可以通过两种途径访问集群：

- 内网地址：内网地址为用户 VPC 内的终端节点 IP 地址。
- 外网地址（可选）：外网地址为绑定在 ELB 上的 EIP 地址。EIP 仅在创建数据服务集群时，勾选开启公网入口，才会具备。

图3-618 专享版集群网络架构说明



因此，为了保证创建的专享版集群能够被用户访问，创建中需要注意如下网络配置：

- VPC

  虚拟私有云。专享版实例需要配置虚拟私有云（VPC），在同一 VPC 中的资源（如 ECS），可以使用专享版实例的私有地址调用 API。

  在创建时专享版实例时，建议配置和您其他关联业务相同 VPC，确保网络安全的同时，方便网络配置。

- 弹性公网 IP

  专享版实例的 API 如果要允许外部调用，则需要购买一个弹性公网 IP，并在购买时绑定给实例，作为实例的公网入口。

- 安全组

  安全组类似防火墙，控制谁能访问实例的指定端口，以及控制实例的通信数据流向指定的目的地址。安全组入方向规则建议按需开放地址与端口，这样可以最大程度保护实例的网络安全。

专享版实例绑定的安全组有如下要求：

- 入方向：如果需要从公网调用 API，或从其他安全组内资源调用 API，则需要为专享版实例绑定的安全组的入方向放开 80（HTTP）、443（HTTPS）两个端口。
- 出方向：如果后端服务部署在公网，或者其他安全组内，则需要为专享版实例绑定的安全组的出方向放开后端服务地址与 API 调用监听端口。
- 如果 API 的前后端服务与专享版实例绑定了相同的安全组、相同的虚拟私有云，则无需专门为专享版实例开放上述端口。

- 路由配置

  在物理机纳管场景下，如果物理机纳管网段与集群网段不一致，需要配置路由。

  进入集群“基本信息”页面，可以增加或删除路由，如下图所示。

说明

如果数据服务集群无路由配置功能，可联系相关支撑人员修改配置项 dlm.instance.route.action.support 启用此功能。

操作步骤

创建数据服务专享集群增量包，系统会按照您所选规格自动创建一个数据服务专享集群。

步骤 1 单击已开通实例卡片上的“创建增量包”。

步骤 2 进入创建 DataArts Studio 增量包页面，参见表 3-353 进行配置。

表3-353 创建数据服务专享版实例参数说明

参数项	说明
增量包类型	选择数据服务专享集群增量包。
工作空间	选择需要使用数据服务专享集群增量包的工作空间。例如需要在 DataArts Studio 实例的工作空间 A 中使用数据服务专享版，则此处工作空间应选择为 A。集群创建成功后，即可通过在工作空间 A 查看到创建好的数据服务专享集群。

参数项	说明
可用区	第一次创建 DataArts Studio 实例或批增量包时，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。
集群名称	-
集群描述	可以自定义对当前数据服务专享版集群的描述。
版本	当前数据服务专享版的集群版本。
集群规格	不同实例规格，对 API 请求的并发支持能力不同。
公网入口	开启“公网入口”，即允许外部服务通过公网地址，调用专享版实例创建的 API。
带宽大小	可配置公网带宽范围。
虚拟私有云	VPC 即虚拟私有云，是通过逻辑方式进行网络隔离，提供安全、隔离的网络环境。 在相同虚拟私有云中的云服务资源（如 ECS），可以使用数据服务专享版实例的私有地址调用 API。 建议将专享版实例和您的其他关联业务配置一个相同的虚拟私有云，确保网络安全的同时，方便网络配置。 说明 目前 DLM 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	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。 建议将专享版实例和您的其他关联业务配置相同的虚拟私有云下相同的子网，确保网络安全的同时，方便网络配置。 说明 目前 DLM 实例创建完成后不支持切换子网，请谨慎选择所属子网。
安全组	安全组用于设置端口访问规则，定义哪些端口允许被外部访问，以及允许访问外部哪些地址与端口。 例如，后端服务部署在外部网络，则需要设置相应的安全组规则，允许访问后端服务的地址及其监听端口。 说明 1. 如果开启公网入口，安全组入方向需要放开 80（HTTP）和 443（HTTPS）端口的访问权限。 2. 目前 DLM 实例创建完成后不支持切换安全组，请谨慎选择所属安全

参数项	说明
	组。
企业项目	DataArts Studio 专享版集群关联的企业项目。企业项目管理是一种按企业项目管理云资源的方式，具体请参见《企业管理用户指南》。

步骤 3 单击“立即创建”，确认规格后提交。

----结束

设置 API 分配配额

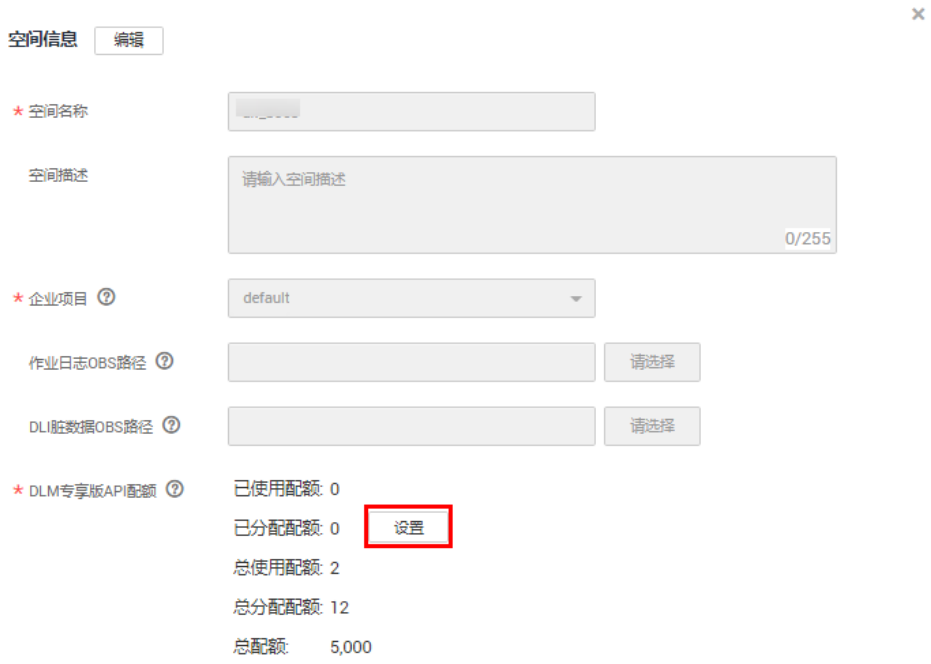
专享版集群创建成功后，需要设置 API 分配配额，当分配配额之后，才能创建相应的 API，配额设置参考如下步骤。

步骤 1 在“空间管理”页签中，单击工作空间操作列“编辑”链接。

图3-619 编辑空间管理

步骤 2 在“空间信息”中，单击“设置”按钮对已分配配额进行配置。

图3-620 设置已分配配额

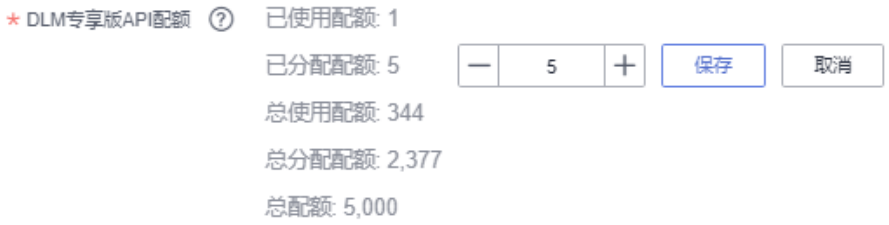


说明

数据服务已创建的 API 属于计费项，当前操作正在增加 API 配额，这会使工作空间下可以创建更多的 API，同时可能使收费增加，请确认。

步骤 3 设置专享版 API 已分配配额。

图3-621 设置配额



说明

已分配配额不能小于已使用配额，不能大于总配额-总分配配额+已分配配额。

----结束

3.8.3.1.2 新建审核人

在发布 API 时，会触发审核，审核机制如下：

- 当发布人不具备审核人权限时，发布 API 时需要提交给审核人审核。

- 当发布人具备审核人权限时，可无需审批直接发布 API。工作空间管理员角色的用户默认具备审核人权限。

因此，如果不具备审核人权限的用户需要发布 API 时，请先添加审核人。只有工作空间管理员角色的用户才具有添加审核人的权限。

操作步骤

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-622 选择数据服务

2. 在左侧导航选择服务版本（例如：专享版），进入总览页。
3. 单击左侧导航栏中的“审核中心”，进入相应页面后，选择“审核人管理”页签，然后单击“新建”按钮。

图3-623 新建审核人界面

4. 选择审核人（此处的帐户列表来自于工作空间成员），输入正确的手机号码和电子邮箱，单击“确认”完成审核人的添加。
5. 根据需要，可以添加多个审核人。

3.8.3.2 创建 API

3.8.3.2.1 配置模式生成 API

本节介绍如何通过配置模式生成 API。

使用配置模式生成数据 API 简单且容易上手，您不需编写任何代码，通过产品界面进行勾选配置即可快速生成 API。推荐对 API 功能的要求不高或者无代码开发经验的用户使用。

前提条件

已在“管理中心 > 数据连接”页面，完成数据源的配置。

新建 API 目录

API 目录是按一定次序编排记录的 API 索引，是反映类别、指导使用、检索 API 的工具，帮助 API 开发者对 API 服务进行有效的分类和管理。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

   图3-624 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 进入“开发 API > > API 目录”页面，单击 。

   输入新建 API 目录名称，可新建 API 目录。
4. 对应已成功创建的 API 目录的操作列，可重新编辑 API 目录或者管理 API。

   单击“编辑”，可修改 API 目录名称信息。仅当 API 处于已创建、已驳回、已下线、已停用的情况下才能进行 API 修改。

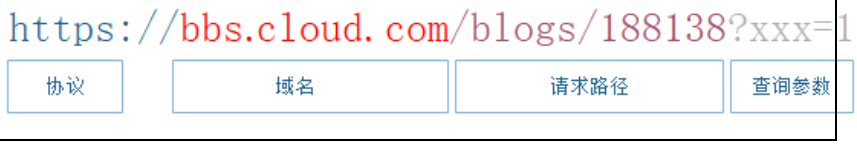
配置 API 基本信息

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-625 选择数据服务

2. 在左侧导航选择服务版本（例如：专享版），进入总览页。
3. 进入“API 管理”页面，单击“新建”，填写 API 基本信息。

表3-354 API 基本信息

配置	说明
API 名称	支持中文、英文、数字、下划线，且只能以英文或中文开头，3-64 个字符。
API 目录	一个特定功能或场景的 API 集合，类似文件夹，是数据服务中 API 的最小组织单元，也是 API 网关中的最小管理单元。 您可单击“新建”进行新建，也可选择新建 API 目录已创建的 API 分组。
请求 Path	API 访问路径，例如：/v2/{project_id}/streams。 请求 Path 即完整的 url 中，域名之后、查询参数之前的部分，如图 3-626 中的“/blogs/188138”。 图3-626 统一资源定位符 url 说明 

配置	说明
	在请求路径中，可以使用大括号{}标识路径中的参数作为通配符。如“/blogs/{blog_id}”表示/blogs 后可以携带任何参数，例如“/blogs/188138”和“/blogs/0”均会匹配至/blogs/{blog_id}，由此 API 统一处理。 此外，相同域名下，不允许重复的请求路径出现。路径参数作为通配符时，名称不具备唯一性，例如“/blogs/{blog_id}”和“/blogs/{xxxx}”，会被视作相同路径。
参数协议	用于传输请求的协议，支持 HTTP 和 HTTPS 协议。 • HTTP 属于基础的网络传输协议，无状态、无连接、简单、快速、灵活、使用明文传输，在使用上较为便捷，但是安全性欠佳。 • HTTPS 是在 HTTP 协议上进行了 SSL 或 TLS 加密校验的协议，能够有效验证身份以及保护数据完整性。相对的，访问 HTTPS 的 API，需要配置相关的 SSL 证书或跳过 SSL 校验，否则将无法访问。
请求方式	HTTP 请求方式，表示请求什么类型的操作，包含 GET、POST 等，遵循 resultful 风格。 • GET：请求服务器返回指定资源，推荐使用 GET 请求。 • POST：请求服务器新增资源或执行特殊操作，仅在注册 API 时使用。POST 请求当前不支持 body 体，而是直接透传。
描述	对 API 进行简要描述。
标签	对 API 设置标签。用于标记当前 API 的属性，创建后可以通过标签快速检索定位 API。单个 API 最多可设置 20 个标签。
审核人	拥有 API 的审核权限。 单击“新建”，进入“审核中心 > 审核人管理”页面，新建审核人。
安全认证	API 认证方式： • APP 认证：表示由 API 网关服务负责接口请求的安全认证，安全级别最高。 • IAM 认证：表示借助 IAM 服务进行安全认证，安全级别中等。 • 无认证：属于无防护的模式，无需认证即可访问，安全级别低，不推荐使用。
服务目录可见性	发布后，所选范围内的用户均可以在服务目录中看到此 API。 • 当前工作空间可见 • 当前项目可见 • 当前租户可见
访问日志	勾选，则此 API 的查询结果将会产生记录并被保留 7 天，可以在

配置	说明
	“运营管理 > 访问日志”处通过选择“请求日期”的方式查看对应日期的日志。
最低保留期限	API 解绑前预留的最低期限。0 表示不设限制。 API 进行停用/下线/取消授权时，会通知已授权用户，并为用户预留至少 X 小时，直到所有授权用户均完成解除或处理，或者到达截止时间，API 才会停用/下线/取消授权。
入参定义	配置 API 请求中的参数，包含资源路径中的动态参数，请求 URI 中的查询参数和 Header 参数。 添加入参定义时，如果参数设定为必填，则 API 在访问时，必须传入指定参数；如果非必填，则在 API 访问时，未传入的参数，会使用默认值进行代替。 参数大小限制如下： • query+path，url 最大 32KB • header，最大 128KB • body， 最大 128KB 以配置资源路径中的动态参数为例进行说明，例如资源路径（请求 Path）设置为： /v2/{project_id}/streams，资源路径中的{project_id}为动态参数，需要在此进行配置。 1. 单击“添加”，参数名配置为 project_id。 2. 参数位置选择 PATH。 3. 类型设置为 STRING。 4. 选择性配置示例值和描述。

4. 配置好 API 基本信息后，单击“下一步”，即可进入 API 取数逻辑页面。

配置取数逻辑

“取数方式”选择“配置方式”：

1. 选择数据源、数据连接、数据库和数据表，获取到需要配置的表。

说明

数据服务仅支持部分数据源，详情请参见 3.2.1 DataArts Studio 支持的数据源。您需提前在 DataArts Studio 管理中心中配置好数据源，数据表支持表名搜索。

2. 配置参数字段

选择好数据表之后，单击“参数设置”后的“添加”，添加参数页面自动列出这个表的所有字段，分别勾选需要设置为请求参数、返回参数和排序参数的字段，分别添加到请求参数、返回参数和排序参数列表当中。

图3-627 添加参数

3. 编辑请求参数信息

请求参数主要分为三部分，绑定参数、绑定字段、操作符。在请求参数列表中，需要设置绑定参数和操作符。

- 绑定参数对外开放，是用户访问 API 时，直接使用的参数。
- 绑定字段对外不可见，是 API 调用时，实际访问的内容。
- 操作符则是用户访问时，对参数的的处理方式。当前支持的操作符及含义如下：

表3-355 支持的操作符

操作符	描述
=	检查两个操作数的值是否相等，如果相等则条件为真。
<>	检查两个操作数的值是否相等，如果不相等则条件为真。
>	检查左操作数的值是否大于右操作数的值，如果是则条件为真。
>=	检查左操作数的值是否大于等于右操作数的值，如果是则条件为真。
<	检查左操作数的值是否小于右操作数的值，如果是则条件为真。
<=	检查左操作数的值是否小于等于右操作数的值，如果是则条件为真。
%like%	%like%表示忽略前后缀，进行字符匹配。
%like	%like 表示忽略前缀，进行字符匹配。
like%	like%表示忽略后缀，进行字符匹配。
in	in 运算符用于把某个值与一系列指定列表的值进行比较。
not in	in 运算符的对立面，用于把某个值与不在一系列指定列表的值进行比较。

4. 编辑返回参数信息

在返回参数列表中，设置参数的名称、参数类型、示例值和描述。

返回参数主要分为三部分，参数名、绑定字段、参数类型。参数名对外开放，是 API 返回时，最终展示给用户的参数名称；绑定字段对外不可见，是 API 调用时，实际返回的内容；参数类型则是 API 调用时，数据的呈现格式，分为数值型和字符型两类。

5. 编辑排序参数信息

   在排序参数列表中，设置排序字段是否可选，排序方式和描述。

   排序参数主要分为四部分，参数名、字段名称、是否可选以及排序方式。参数名对外开放，是 API 返回时，最终展示给用户的参数名称；字段名称对外不可见，是 API 调用时，实际访问的内容；是否可选决定了当前排序条件是否允许移除，勾选则表示可以不使用此参数；排序方式分为升序、降序以及自定义，表示了当前参数允许使用的排序形式。

   通过排序参数列表中的操作列的向上、向下和删除按钮，可调整排序参数的顺序或者删除某排序参数。

6. 单击“下一步”，设置 pre_order_by 的值为所有排序参数的描述，以“英文分号”进行分隔。

   以如下样例数据为例进行说明：

表3-356 排序字段对应的参数描述

排序字段	对应的排序参数描述
id	a:asc 其中，a 是参数名，asc 代表升序。
name	• b:asc • b • b:desc 其中，b 是参数名，因为排序方式是自定义，所以有如上三种参数描述。
age	c:desc 其中，c 是参数名，desc 代表降序。

依据表 3-356，分析得出各个字段对应的排序参数描述，则 pre_order_by 的设置方式有如下几种情况，包含所有排序参数的描述。

表3-357 配置 pre_order_by

pre_order_by	对应的后端 order by 语句	备注
a:asc;b;c:desc	order by id ASC, name, age DESC	-
b;c:desc	order by name, age DESC	因 a 是可选排序字段，所以可以不填。
b:asc;c:desc	order by name ASC; age DESC	b 排序方式是自定义，排

pre_order_by	对应的后端 order by 语句	备注
		序时可选择升序。
b:desc;c:desc	order by name DESC; age DESC	b 排序方式是自定义，排序时可选择降序。

图3-628 配置排序参数值

说明

- pre_order_by 是非必填参数，当未配置 pre_order_by 参数值时，则选取非可选排序字段作为排序的依据。
- 当配置 pre_order_by 参数值时，配置类 API 需严格按照排序参数顺序进行设置。例如 a:asc;b;c:desc ，可行。当设置为 b;a:asc;c:desc，则报错。

测试 API

完成 API 参数的配置并保存后，单击左下角的“开始测试”，可进入 API 测试环节。

填写参数值，单击“开始测试”，即可在线发送 API 请求，在右侧可以看到 API 请求详情及返回内容。如果测试失败，请仔细查看错误提示并做相应的修改重新测试。 配置过程中需要注意正常返回示例的设置。

完成 API 测试之后，单击“确定”，即成功生成了一个数据 API。

修改 API

生成 API 后，如果您需要修改 API 内容，可在“开发 API > API 目录”或“开发 API > API 管理”处选择对应 API，点击“编辑”按钮进行修改 API 的相关操作。

说明

仅当 API 处于已创建、已驳回、已下线、已停用的情况下才能进行 API 修改。

3.8.3.2.2 脚本模式生成 API

本文将为您介绍如何通过脚本模式生成 API。

为了满足高阶用户的个性化查询需求，数据服务也提供了自定义 SQL 的脚本模式，允许您自行编写 API 的查询 SQL，并支持多表关联、复杂查询条件以及聚合函数等能力。

配置 API 基本信息

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-629 选择数据服务

2. 在左侧导航选择服务版本（例如：专享版），进入总览页。
3. 进入“API 管理”页面，单击“新建”，填写 API 基本信息。

表3-358 API 基本信息

配置	说明
API 名称	支持中文、英文、数字、下划线，且只能以英文或中文开头，3-64 个字符。
API 目录	一个特定功能或场景的 API 集合，类似文件夹，是数据服务中 API 的最小组织单元，也是 API 网关中的最小管理单元。 您可单击“新建”进行新建，也可选择已创建的 API 分组。
请求 Path	API 访问路径，例如：/v2/{project_id}/streams。 请求 Path 即完整的 url 中，域名之后、查询参数之前的部分，如图 3-630 中的“/blogs/188138”。

配置	说明
	图3-630 统一资源定位符 url 说明  在请求路径中，可以使用大括号{}标识路径中的参数作为通配符。如“/blogs/{blog_id}”表示/blogs 后可以携带任何参数，例如“/blogs/188138”和“/blogs/0”均会匹配至/blogs/{blog_id}，由此 API 统一处理。 此外，相同域名下，不允许重复的请求路径出现。路径参数作为通配符时，名称不具备唯一性，例如“/blogs/{blog_id}”和“/blogs/{xxxx}”，会被视作相同路径。
参数协议	用于传输请求的协议，支持 HTTP 和 HTTPS 协议。 • HTTP 属于基础的网络传输协议，无状态、无连接、简单、快速、灵活、使用明文传输，在使用上较为便捷，但是安全性欠佳。 • HTTPS 是在 HTTP 协议上进行了 SSL 或 TLS 加密校验的协议，能够有效验证身份以及保护数据完整性。相对的，访问 HTTPS 的 API，需要配置相关的 SSL 证书或跳过 SSL 校验，否则将无法访问。
请求方式	HTTP 请求方式，表示请求什么类型的操作，包含 GET、POST 等，遵循 resultful 风格。 • GET：请求服务器返回指定资源，推荐使用 GET 请求。 • POST：请求服务器新增资源或执行特殊操作，仅在注册 API 时使用。POST 请求当前不支持 body 体，而是直接透传。
描述	对 API 进行简要描述。
标签	对 API 设置标签。用于标记当前 API 的属性，创建后可以通过标签快速检索定位 API。单个 API 最多可设置 20 个标签。
审核人	拥有 API 的审核权限。 单击“新建”，进入“审核中心 > 审核人管理”页面，新建审核人。
安全认证	API 认证方式： • APP 认证：表示由 API 网关服务负责接口请求的安全认证，安全级别最高。 • IAM 认证：表示借助 IAM 服务进行安全认证，安全级别中等。 • 无认证：属于无防护的模式，无需认证即可访问，安全级别低，不推荐使用。

配置	说明
服务目录可见性	发布后，所选范围内的用户均可以在服务目录中看到此 API。 • 当前工作空间可见 • 当前项目可见 • 当前租户可见
访问日志	勾选，则此 API 的查询结果将会产生记录并被保留 7 天，可以在“运营管理 > 访问日志”处通过选择“请求日期”的方式查看对应日期的日志。
最低保留期限	API 解绑前预留的最低期限。0 表示不设限制。 API 进行停用/下线/取消授权时，会通知已授权用户，并为用户预留至少 X 小时，直到所有授权用户均完成解除或处理，或者到达截止时间，API 才会停用/下线/取消授权。
入参定义	配置 API 请求中的参数，包含资源路径中的动态参数，请求 URI 中的查询参数和 Header 参数。 添加入参定义时，如果参数设定为必填，则 API 在访问时，必须传入指定参数；如果非必填，则在 API 访问时，未传入的参数，会使用默认值进行代替。 参数大小限制如下： • query+path，url 最大 32KB • header，最大 128KB • body， 最大 128KB 以配置资源路径中的动态参数为例进行说明，例如资源路径（请求 Path）设置为： /v2/{project_id}/streams，资源路径中的{project_id}为动态参数，需要在此进行配置。 1. 单击“添加”，参数名配置为 project_id。 2. 参数位置选择 PATH。 3. 类型设置为 STRING。 4. 选择性配置示例值和描述。

4. 配置好 API 基本信息后，单击“下一步”，即可进入 API 取数逻辑页面。

配置取数逻辑

“取数方式”选择“脚本方式”：

1. 选择数据源、数据连接、数据库和队列，获取到需要配置的表。

说明

数据服务仅支持部分数据源，详情请参见 3.2.1 DataArts Studio 支持的数据源。您需提前在 DataArts Studio 管理中心中配置好数据源，按照脚本编辑提示要求输入 SQL 语句。

2. 编写 API 查询 SQL。

   在脚本编辑页面，按照脚本编辑提示要求输入 SQL 语句。

   说明

   - SELECT 查询的字段即为 API 返回参数，支持返回别名。
   - WHERE 条件中的参数为 API 请求参数，参数格式为${参数名}。

3. 选择分页方式。

   - 默认分页是指在创建 API 时输入了 SQL，数据服务会自动基于 SQL 外层包装分页逻辑。例如输入的 SQL 脚本为：

```
SELECT name as Student_Name FROM tableofresults
```

     数据服务在处理调试或者调用时，将自动在用户 SQL 外层包装分页逻辑，从而变成以下脚本：

```
SELECT * FROM (SELECT name as Student_Name FROM tableofresults) LIMIT {pageSize} OFFSET {offsetValue}
```

     其中 pageNum、offsetValue 为调试或者调用时传入参数的值。如果未定义 pageNum 参数的话，数据服务将默认给 API 设置一个 pageNum 参数；offsetValue 是调试或者调用时传入参数 pageSize 的值计算得到，如果用户未定义 pageSize 参数的话，数据服务将默认给 API 设置一个 pageSize 参数。

   - 自定义分页是指用户在创建 API 时，数据服务将不对用户 SQL 进行处理，分页逻辑完全由用户定义。如果用户需要创建分页的 API 的话，则可以在写 SQL 时加入分页逻辑。例如：

```
SELECT name as Student_Name FROM tableofresults LIMIT {pageSize} OFFSET {offsetValue}
```

4. 添加排序参数。

   在排序参数列表中，设置排序字段是否可选，排序方式和描述。

   单击 ⇒，将入参和排序参数添加为 SQL 语句的 API 请求参数。

   说明

   添加排序参数前，请确保 SQL 语句正确。

5. 编辑请求参数信息

   编写好 API 查询 SQL 后，单击“测试 SQL”，在数据库字段页签内绑定 HTTP 入参。参见配置取数逻辑中的 6 配置 pre_order_by 参数。

   说明

   pre_order_by 是非必填参数，当未配置 pre_order_by 参数值时，则选取非可选排序字段作为排序的依据。

测试 API

完成 API 参数的配置并保存后，单击左下角的“开始测试”，可进入 API 测试环节。

填写参数值，单击“开始测试”，即可在线发送 API 请求，在右侧可以看到 API 请求详情及返回内容。如果测试失败，请仔细查看错误提示并做相应的修改重新测试。 配置过程中需要注意正常返回示例的设置。

完成 API 测试之后，单击“确定”，即成功生成了一个数据 API。

修改 API

生成 API 后，如果您需要修改 API 内容，可在“开发 API > API 目录”或“开发 API > API 管理”处选择对应 API，点击“编辑”按钮进行修改 API 的相关操作。

说明

仅当 API 处于已创建、已驳回、已下线、已停用的情况下才能进行 API 修改。

3.8.3.2.3 注册 API

本文将为您介绍如何注册 API，与通过数据表生成的 API 统一管理和发布到 API 网关。

目前数据服务共享版支持 Restful 风格的 API 注册，包含 GET/POST 常见请求方式。

配置 API 基本信息

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-631 选择数据服务

2. 在左侧导航栏选择服务共享版，进入总览页。
3. 进入“数据服务 > 开发 API > API 管理”页面，单击“注册现有 API”，填写 API 基本信息。

表3-359 API 基本信息

配置	说明

配置	说明
API 名称	支持中文、英文、数字、下划线，且只能以英文或中文开头，3-64 个字符。
API 目录	一个特定功能或场景的 API 集合，是数据服务中 API 的最小组织单元，也是 API 网关中的最小管理单元。 您可单击“新建”进行新建，也可选择新建 API 目录已创建的 API 分组。
请求 Path	资源路径，也即 API 访问路径。 示例：/v2/{project_id}/streams
协议	用于传输请求的协议，支持 HTTP 和 HTTPS 协议。
请求方式	HTTP 请求方法（也称为操作或动词），它告诉服务你正在请求什么类型的操作。 GET：请求服务器返回指定资源。 POST：请求服务器新增资源或执行特殊操作，注册 API 时推荐使用。POST 请求当前不支持 body 体，而是直接透传。
描述	对 API 进行简要描述。
标签	用户自定义输入，只能包含中文、英文字母、数字和下划线，且不能以下划线开头。
审核人	拥有 API 的审核权限。 单击“新建”，进入“审核中心 > 审核人管理”页面，新建审核人。
安全认证	API 认证方式： • APP 认证：表示由 API 网关服务负责接口请求的安全认证。 • IAM 认证：表示借助 IAM 服务进行安全认证。 • 无认证：表示不需要认证。
服务目录可见性	发布后，所选范围内的用户均可以在服务目录中看到此 API。 • 当前工作空间可见 • 当前项目可见 • 当前租户可见
访问日志	勾选，则此 API 的查询结果将会产生记录并被保留 7 天，可以在“运营管理 > 访问日志”处通过选择“请求日期”的方式查看对应日期的日志。
最低保留期限	API 解绑前预留的最低期限。 API 进行停用/下线/取消授权时，会通知已授权用户，并为用户预留至少 X 小时，直到所有授权用户均完成解除或处理，或者到达截止时间，API 才会停用/下线/取消授权。0 表示不设限制。
入参定义	配置 API 请求中的参数，包含资源路径中的动态参数，请求 URI 中

配置	说明
	的查询参数和 Header 参数。 以配置资源路径中的动态参数为例进行说明，例如资源路径（请求 Path）设置为： /v2/{project_id}/streams，资源路径中的{project_id}为动态参数，需要在此进行配置。 1. 单击“添加”，参数名配置为 project_id。 2. 参数位置选择 PATH。 3. 类型设置为 STRING。 4. 选择性配置示例值和描述。

4. 配置好 API 基本信息后，单击“下一步”，即可进入 API 取数逻辑页面。

配置 API 参数

配置 API 基本信息后，即可配置 API 参数。这里将配置 API 的后端服务和请求参数。

表3-360 API 参数配置说明

配置	说明
协议	用于传输请求的协议，支持 HTTP 和 HTTPS 协议。 用于数据服务模块向待注册 API 服务传输请求。
请求方式	HTTP 请求方法（也称为操作或动词），它告诉服务你正在请求什么类型的操作。用于数据服务模块向待注册 API 服务传输请求。 GET：请求服务器返回指定资源。 POST：请求服务器新增资源或执行特殊操作。
后端服务 HOST	待注册 API 服务的 Host，不能以 http://或 https://开头，并且不包含 Path。
后端服务 PATH	待注册 API 服务的 Path，Path 中支持参数，参数要放在{}中，如 /user/{userid}。
后端超时（ms）	设置后端超时时间。
后端服务参数	请求参数位置支持 Path、Header、Query，不同的请求方式所支持的可选参数位置不一样，请根据产品上提供的可选项按需选择。
常量参数	常量参数即参数值是固定的参数，对调用者不可见，API 调用时不需传入常量参数，但后台服务始终接收这里定义好的常量参数及参数值。适用于当您希望把 API 的某个参数的取值固定为某个值以及要对调用者隐藏参数的场景。

API 测试

完成 API 参数的配置并保存后，单击左下角的“开始测试”，即可进入 API 测试环节。填写参数值，单击“开始测试”，即可在线发送 API 请求，在右侧可以看到 API 请求详情及返回内容。如果测试失败，请仔细查看错误提示并做相应的修改重新测试。配置过程中需要注意正常返回示例的设置。

完成 API 测试之后，单击“确定”，即完成注册数据 API。

3.8.3.3 调试 API

操作场景

API 创建后需要验证服务是否正常，管理控制台提供调试功能，您可以添加 HTTP 头部参数与 body 体参数，调试 API 接口。

说明

- 后端路径中含有环境变量的 API，不支持调试。
- API 绑定签名密钥时，不支持调试。
- 如果 API 已绑定流控策略，在调试 API 时，流控策略无效。

前提条件

- 已创建 API。
- 已搭建完成后端服务。

操作步骤

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-632 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“开发 API > API 管理”，进入到 API 管理信息页面。
4. 通过以下任意一种方法，进入 API 调试页面。
   - 在待调试的 API 所在行，单击“更多 > 调试”。
   - 单击“*API 名称*”，进入 API 详情页面，单击“调试”。

   左侧为 API 请求参数配置区域，参数说明如表 3-361 所示。右侧为 API 发送的请求信息和 API 请求调用后的返回结果回显。

表3-361 调试 API

参数名称	说明
参数配置	Query 的参数与参数值。
集群配置	仅专享版支持，选择调试 API 所依托的实例。

说明

不同类型的请求，调试界面展现的信息项有差异。

5. 添加请求参数后，单击“开始测试”。

   右侧返回结果回显区域打印 API 调用的 Response 信息。
   - 调用成功时，返回 HTTP 状态码为“200”和 Response 信息。
   - 调试失败时，返回 HTTP 状态码为 4*xx* 或 5*xx*。
6. 您可以通过调整请求参数与参数值，发送不同的请求，验证 API 服务。

说明

如果需要修改 API 参数，请在右上角单击“编辑”，进入 API 编辑页面。

后续操作

API 调试成功后，您可以将 API3.8.3.4 发布 API，以便 API 调用者调用。

3.8.3.4 发布 API

本文将为您介绍如何将数据服务中的 API 发布到服务目录。

操作场景

数据服务是数据对外开放的最后一道防线，为了安全起见，在数据服务中生成的 API 以及注册的 API，都需要发布到服务目录中才能对外提供服务。

操作步骤

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-633 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 进入“数据服务 > 总览 > 开发 API > API 管理”页面，在 API 服务列表操作列中，选择“更多 > 发布”。
4. 在确认发布界面，您可以点击“更多”，选择发布详情。

图3-634 发布详情

- 专享版默认发布到数据服务专享版集群上，发布成功后 API 调用者可以通过内网调用该 API。您也可以选择“更多”，将 API 发布到 APIG 专享版或 ROMA Connect 实例上。
  - APIG 专享版：如果您需要将 API 发布到 APIG 专享版上，则您需要提前在 API 网关服务上创建一个 APIG 实例。实例创建后，有一个默认 API 分组，系统为分组自动分配一个内部测试用的调试域名，此调试域名唯一且不可更改，每天最多可以访问 1000 次。如果您不希望与其他 API 共享此规格，可以在 APIG 控制台新建一个 API 分组（详情请参考“API 网关 APIG 用户指南> API 分组管理> 创建 API 分组”章节），然后在数据服务发布时选择对应 API 分组，独享每天最多访问 1000 次的规格。另外，您还可以为 API 分组绑定一个或多个独立域名（详情请参考“API 网关 APIG 用户指南> API 分组管理> 绑定域名”章节），API 调用者通过访问独立域名来调用您开放的 API，这样即可不受每天最多访问 1000 次的规格限制。
  - ROMA Connect 实例：如果您需要将 API 发布到 ROMA Connect 实例上，则您需要提前在 ROMA Connect 服务上创建一个 ROMA 实例，并创建 API 分组（详情请参考“应用与数据集成平台 ROMA Connect 用户指南> 服务集成指导> 开放 API> 创建 API 分组”章节）。API 分组创建后，系统为分组自动分配一个内部测试用的子域名，此子域名每天最多可以访问 1000 次。为了不受此规格限制，您可以为 API 分组绑定独立域名（详情请参考“应用与数据集成平台 ROMA Connect 用户指南> 服务集成指导> 开放 API> 绑定域名”章节），API 调用者通过访问独立域名来调用您开放的 API。

5. 在发布 API 时，会触发审核，审核机制如下：
   - 当发布人不具备审核人权限时，发布 API 时需要提交给审核人审核。
   - 当发布人具备审核人权限时，可无需审批直接发布 API。工作空间管理员角色的用户默认具备审核人权限。

   如果非审核人权限的用户发布 API 时，待审核人审核通过后，即可发布完成。

后续操作

发布完成后，您可以进入到“服务目录”，查看 API 信息。

也可以对 API 进行管理，请参见 3.8.3.5 管理 API；或进一步在“运营管理 > 流控策略”页面设置流量控制等功能，请参见 3.8.3.6 流量控制。

3.8.3.5 管理 API

3.8.3.5.1 设置 API 可见

操作场景

当需要修改 API 的在服务目录中的可见范围时，可以通过“设置可见”功能或编辑 API 中的“服务目录可见性”参数进行设置。

前提条件

已创建 API。

通过“设置可见”功能修改 API 可见范围

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-635 选择数据服务

1. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
2. 进入“开发 API > API 目录”或“开发 API > API 管理”页面，在待修改的 API 所在行，选择“更多 > 设置可见”。
3. 在弹出的窗口中点击添加，填写项目 ID 并确认，即可设置此 API 在服务目录中额外对以该项目下的用户可见。

图3-636 设置可见

通过“服务目录可见性”参数修改 API 可见范围

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-637 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 进入“开发 API > API 目录”或“开发 API > API 管理”页面，在 API 列表操作列中，选择“编辑”。注意，仅当 API 处于已创建、已驳回、已下线、已停用的情况下才能进行 API 修改。

4. 在基本配置处，修改“服务目录可见性”参数的取值，可以选择为“当前工作空间可见”、“当前项目可见”或“当前租户可见”。然后保存修改。
5. 修改完成后，重新恢复或发布 API，即可修改此 API 在服务目录中的可见范围。

3.8.3.5.2 停用/恢复 API

操作场景

已发布的 API 因为其他原因（如需要编辑 API 等）需要暂停对外提供服务，可以暂时将 API 从相关环境中停用。停用后您可以通过恢复 API，使该 API 继续对外提供服务。

说明

- 停用 API 会保留原有的授权信息，在停用期间您可以对 API 进行编辑、调试等操作。
- 停用 API 将导致此 API 在指定的时间无法被访问，请确保已经告知使用此 API 的用户。

前提条件

- 已创建 API。
- API 已发布到该环境。

停用 API

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-638 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“开发 API > API 管理”，进入到 API 管理信息页面。

4. 在待停用的 API 所在行，单击“更多 > 停用”，弹出“停用”对话框。
5. 选择 API 需要停用的时间，单击“确定”，完成 API 定时停用。

恢复 API

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-639 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 在待恢复的 API 所在行，单击“更多 > 恢复”，完成 API 恢复。

3.8.3.5.3 下线/删除 API

操作场景

已发布的 API 因为其他原因需要停止对外提供服务，可以将 API 从相关环境中下线，相关操作请参见下线 API。

- 下线后的 API 如果要继续使用，需要重新进行发布操作，但需注意下线 API 不会保留原有的授权信息。
- 下线后的 API 如果确认不再提供服务，可以将 API 删除，相关操作请参见删除 API。

说明

- 下线 API 不会保留原有的授权信息。
- 下线将导致此 API 在指定的时间无法被访问，请确保已经告知使用此 API 的用户。
- 删除 API 导致此 API 无法恢复，请确认后谨慎操作。

前提条件

- 已创建 API。
- API 已发布到该环境。

下线 API

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

   图3-640 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“开发 API > API 管理”，进入到 API 管理信息页面。
4. 在待下线的 API 所在行，单击“更多 > 下线”，弹出“下线 API”对话框。
5. 选择 API 需要下线的时间，单击“确定”，完成 API 定时下线。

删除 API

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-641 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 在左侧选择“API 目录”，进入 API 列表页，勾选需要删除的 API，单击“删除”。
4. 单击“确定”，完成 API 删除。

说明

如果需要批量删除 API，则勾选待删除的 API，单击“删除”。最多同时删除 1000 个 API。

3.8.3.5.4 复制 API

操作场景

您可以通过复制 API 功能，得到与原 API 配置相同的 API。

前提条件

已创建 API。

操作步骤

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-642 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“开发 API > API 管理”页面，进入 API 管理页面。
4. 勾选待复制的 API 所在行，在 API 列表上方，选择“更多 > 复制”，弹出复制窗口。
5. 在弹出的窗口中输入新 API 的名称和请求 path，点击确认即可完成 API 复制。

图3-643 复制 API

3.8.3.5.5 全量导出/导出/导入 API

操作场景

数据服务支持全量导出/批量导出/导入 API，可以快速复制或迁移现有的 API。

前提条件

- 已创建 API。
- 执行全量导出 API 必须具备 DAYU Administrator 或 Tenant Administrator 权限。
- 同时只能有一个全量导出任务执行。
- 每个工作空间每分钟仅能全量导出一次。

全量导出 API

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

   图3-644 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“开发 API > API 管理”页面，进入 API 管理页面。
4. 在 API 列表上方，选择“更多 > 全量导出”，弹出导出确认窗口。

   说明

   - 执行全量导出 API 必须具备 DAYU Administrator 或 Tenant Administrator 权限。
   - 同时只能有一个全量导出任务执行。
   - 每个工作空间每分钟仅能全量导出一次。

   在导出窗口中单击“确认”导出全量 API，点击确认即可以 Excel 文件的形式导出 API。

图3-645 全量导出 API

5. 打开下载到本地的 Excel 文件，可以查看导出的 API。不同类型的 API 会分别导出到文件页签中，点击下方页签可以切换查看并编辑。

图3-646 Excel 文件样式

导出 API

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-647 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。

3. 单击“开发 API > API 管理”页面，进入 API 管理页面。
4. 勾选待导出的 API 所在行，在 API 列表上方，选择“更多 > 导出”，弹出导出窗口。
5. 在导出窗口中确认待导出的 API，点击确认即可以 Excel 文件的形式导出 API。

图3-648 导出 API

6. 打开下载到本地的 Excel 文件，可以查看导出的 API。不同类型的 API 会分别导出到文件页签中，点击下方页签可以切换查看并编辑。

图3-649 Excel 文件样式

导入 API

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-650 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“开发 API > API 管理”页面，进入 API 管理页面。
4. 在 API 列表上方，选择“更多 > 导入”，进入导入 API 窗口。
5. 在导入窗口中点击“选择 Excel 文件”，选择后点击导入，导入结果中可以展示导入状态。

说明

待导入的 API 文件可以是从其他项目直接导出的 API 文件，也可以是通过模板填写的 Excel 文件，需要确保符合模板规范要求。

图3-651 导入 API

6. 导入成功后，即可在 API 列表中查看导入的 API。

3.8.3.6 流量控制

操作场景

DataArts Studio 数据服务的 API 流量控制基于指定规则对 API 的访问流量进行调节控制的限流策略，能够提供多种维度的后端服务保护功能。当前 API 流控支持通过用户、应用和时间段等不同维度限制 API 的调用次数。

为了提供持续稳定的服务，您需要通过创建并选择流控策略，针对部分 API 进行流量控制。流控策略和 API 本身是相互独立的，只有将流控策略绑定 API 后，流控策略才对绑定的 API 生效。

说明

同一个环境中一个 API 只能被一个流控策略绑定，一个流控策略可以绑定多个 API。

前提条件

需要绑定的 API 已发布。

创建流控策略

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

   图3-652 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“运营管理 > 流控策略”，进入到流量控制信息页面。
4. 单击“创建流控策略”，弹出“创建流控策略”对话框。输入如表 3-362 所示信息。

表3-362 流控策略信息

信息项	描述
策略名称	API 流控策略名称。
时长	流量限制的时长。 • 与“API 流量限制”配合使用，表示单位时间内的单个 API 请求次数上限。 • 与“用户流量限制”配合使用，表示单位时间内的单个用户请求次数上限。 • 与“应用流量限制”配合使用，表示单位时间内的单个 APP 请求次数上限。
API 流量限制	单个 API 被调用次数上限。 与“时长”配合使用，表示单位时间内的单个 API 请求次数上限。
用户流量限制	单个用户调用 API 次数上限。 • 不超过“API 流量限制”。 • 与“时长”配合使用，表示单位时间内的单个用户请求次数上限。
应用流量限制	单个应用调用 API 次数上限。 • 不超过“用户流量限制”。 • 与“时长”配合使用，表示单位时间内的单个应用请求次数上限。
描述	关于控制策略的描述。

5. 单击“确定”，完成流量控制策略的创建。

   创建成功后，策略信息页面增加显示新创建的策略，您可以将相关 API 绑定到该策略，以实现流量控制。

绑定 API

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-653 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“运营管理 > 流控策略”，进入到流量控制信息页面。
4. 通过以下任意一种方法，进入“绑定 API”页面。
   - 在待绑定的流量控制策略所在行，单击“绑定 API”。
   - 单击策略名称，进入策略详情页面。在“绑定的 API 列表”页签中单击“绑定 API”。
5. 选择“API 分组”和“API 名称”，筛选所需的 API。
6. 勾选 API，单击“绑定”，完成 API 绑定策略。

说明

在流控策略绑定 API 后，如果 API 不需要调用此策略，单击“解除”，解除绑定。如果需要批量解绑 API，则勾选待解绑的 API，单击“解除”。最多同时解绑 1000 个 API。

删除流控策略

当已创建的流控策略不再提供服务时，可以将此流控策略删除。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-654 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“运营管理 > 流控策略”，进入到流量控制信息页面。
4. 在待删除的流控策略所在行，单击“删除”。

说明

- 仅在流控策略未绑定任何 API 时，支持删除，否则请先解绑 API。
- 如果需要批量删除流控策略，则勾选待删除的流控策略，单击“删除”。最多同时删除 1000 个流控策略。

5. 单击“确定”，完成流控策略的删除。

3.8.4 调用 API

概述

您作为 API 调用者，需要实现一个 API 的调用，那么您需要完成以下工作：

1. 获取 API

   从服务目录获取需要调用 API。仅在 API 发布后，才支持被调用。

2. （可选）创建应用并获取授权

   对于使用 APP 和 IAM 认证的 API，需要完成创建应用和将 API 授权给应用。在 API 调用过程中，使用所创建应用的密钥对（AppKey、AppSecret），数据服务根据密钥对进行身份核对，完成鉴权。

3. 调用 API

   API 调用者完成以上步骤后，可以进行 API 调用。

创建应用

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-655 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“调用 API > 应用管理”，进入到应用管理页面。单击“新建”，弹出“新建应用”对话框。填写如表 3-363 所示信息。

表3-363 应用信息

信息项	描述
应用名称	应用名称。
应用类型	IAM：使用 IAM 认证，即 token 访问。 APP：对接 APP，通过 APP 认证方式访问。
描述	对应用的介绍。

4. 单击“确定”，创建应用。

   创建应用成功后，在“应用管理”页面的列表中显示新创建的应用和应用 ID。
5. 单击“应用名称”，进入应用详情页面，查看 AppKey 和 AppSecret。

将 API 授权给应用

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-656 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 通过以下任意一种方式，将 API 授权给应用。

   主动授权：

   a. 单击“开发 API > API 管理”，进入到 API 管理页面。
   b. 在待绑定应用的 API 所在行，单击“查看授权”，进入 API 完整信息界面。在“授权信息”页签中，单击“添加授权”。
   c. 设置授权的截止时间，勾选应用名称，单击“确认授权”，完成 API 的授权。

   申请授权：

   a. 单击“调用 API > 服务目录”，进入服务市场主页面。
   b. 单击待绑定应用的 API 名称，进入 API 完整信息页面。
   c. 在“调用信息”页面，单击“申请权限”。
   d. 设置使用截止时间并选择应用名称，单击“确认”。
   e. 申请后，需要等待审核中心审核，方可授权成功。
4. 授权成功后，可以在应用管理详情页面查看已绑定的 API。

说明

- 如果已绑定 API 列表中包含无需绑定的 API，在此 API 所在行的操作列，单击“解绑”，将无需绑定的 API 删除。
- 如果需要调试已绑定的 API，单击“测试”，进入调试页面。
- 如果需要对已绑定的 API 延长授权时间，单击“续约”。

调用 API

以下三种认证方式区别仅在于认证的内容不同，调用 API 的方式是相同的。

- “IAM 认证”：需要借助 IAM 服务进行安全认证。
- “无认证”：不需要认证，直接调用 API 即可。
- “APP 认证方式”：API 调用者通过 APP 认证方式调用 API。
  - 使用 APP 认证时，需要通过 SDK 访问。
  - 其中 SDK 访问提供了基于 Java、Go、Python、JavaScript、C#、PHP、C++、C、Android 等多种语言的 SDK 包。
  - 各个语言调用 API 示例请参考《数据治理中心 2.9.2 SDK 参考》“使用 APP 认证调用 API”的 Java、Go、Python、C#、JavaScript、PHP、C++、C、Android、curl 章节。

3.8.5 审核中心操作说明

数据服务平台的审核中心，提供给 API 开放方和 API 调用方用以审核 API 的上线、下线、申请授权、续约等操作。

- 当 API 开放方需要将 API 发布至服务市场，从服务市场下线，取消对某个应用的授权，基于数据服务平台提交操作后，均需要等待审核中心责任人审核后方生效。
- 当 API 调用方申请 API 授权和授权续约时，基于数据服务平台提交操作后，也需要等待审核中心责任人审核后方生效。
- 待审核的 API 可在审核中心由发起者执行撤销操作。
- 审核人在审核中心看到的审核申请，只有当非审核人发布 API 的时候才能看到审核申请，并且发布到 APIG 的话只能发布到 release 运行环境。

审核方式

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-657 选择数据服务

2. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
3. 单击“运营管理 > 审核中心”，选择“待审核”页签。
4. 根据审核类型、提交时间等筛选条件，筛选出待处理任务，选择操作列的审核，即生效。

说明

勾选多个 API 名称前的复选框，支持批量审核。

管理审核人

数据服务平台提供管理审核人的功能，您可在审核中心新建和删除审核人。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-658 选择数据服务

2. 在左侧导航选择服务版本（例如：专享版），进入总览页。
3. 单击“运营管理 > 审核中心”，选择“审核人管理”页签。
4. 单击“新建”，在“新建审核人”对话框中设置基本信息。

撤销 API 申请

数据服务平台提供撤销待审核申请的功能，您可在“审核中心 >申请列表”撤销待审核申请。

1. 在 DataArts Studio 控制台首页，选择实例，点击“进入控制台”，选择对应工作空间的“数据服务”模块，进入数据服务页面。

图3-659 选择数据服务

2. 在左侧导航选择服务版本（例如：专享版），进入总览页。
3. 单击“运营管理 >审核中心”，选择“申请列表 > 调用”页签。
4. 查找需要撤销的 API 名称，单击“撤销”。

4 常见问题

4.1 产品咨询类

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 可从企业管理员处获取。

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/购买成功后，可以转移到其他账号下吗？

不可以，实例试用/购买后不能转移到另一个账户。

4.1.5 DataArts Studio 是否支持版本降级？

已购买 DataArts Studio 实例后，不支持降级版本。

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：

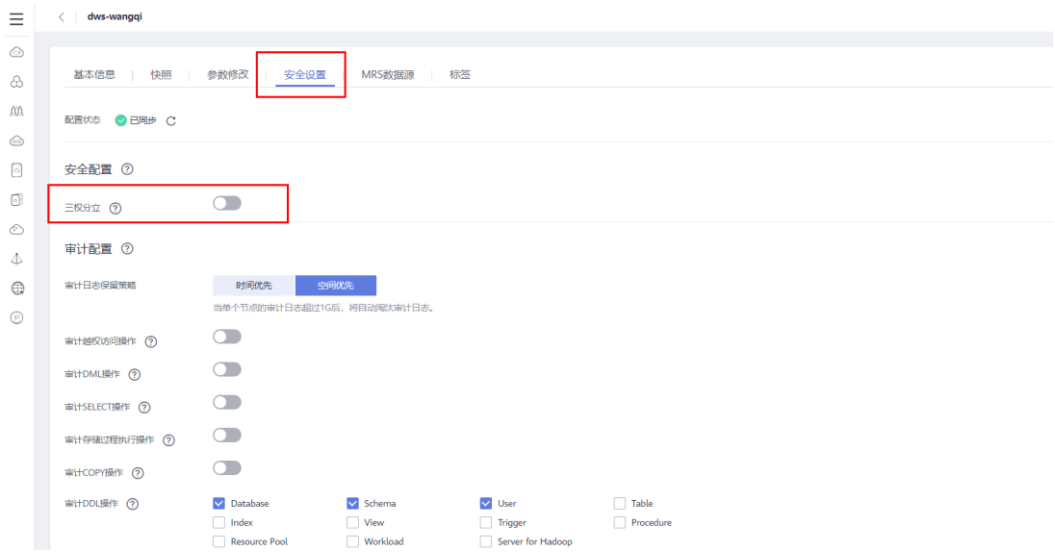
- 创建 MRS 集群时未选择 Hive/HBase 组件。
- 创建 MRS 数据连接时所选择的 CDM 集群和 MRS 集群网络不互通。

  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。

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

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主机共享弹性 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？

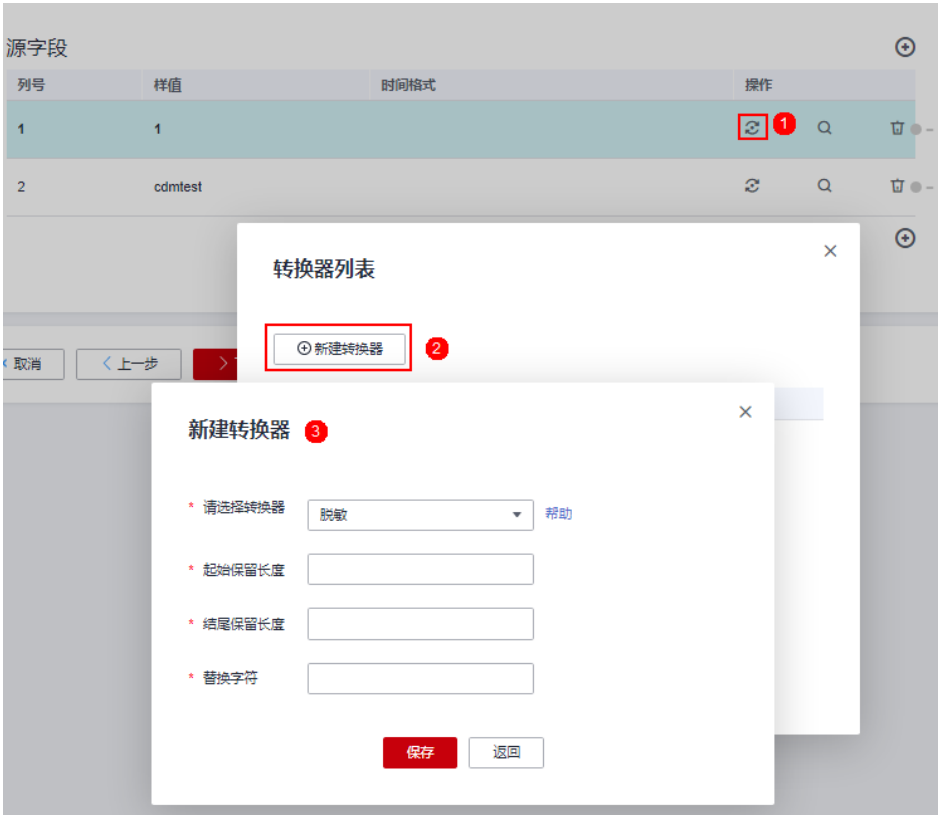
支持，CDM 支持以下字段转换器：

- 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
	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/"+CLU
STER_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/"+CLUS
TER_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/"+CLUS
TER_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/"+C
LUSTER_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

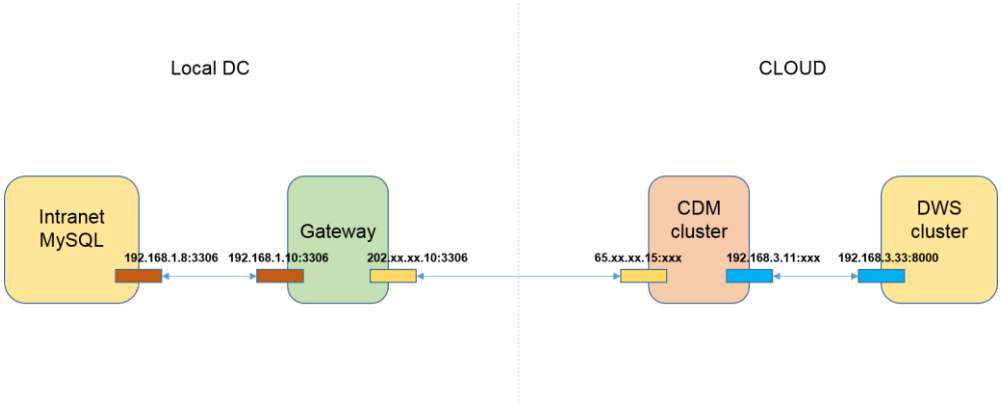
长时间将内网数据库暴露在公网会有安全风险，迁移数据完成后，请及时停止端口映射。

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。

图4-4 网络拓扑样例



操作步骤

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

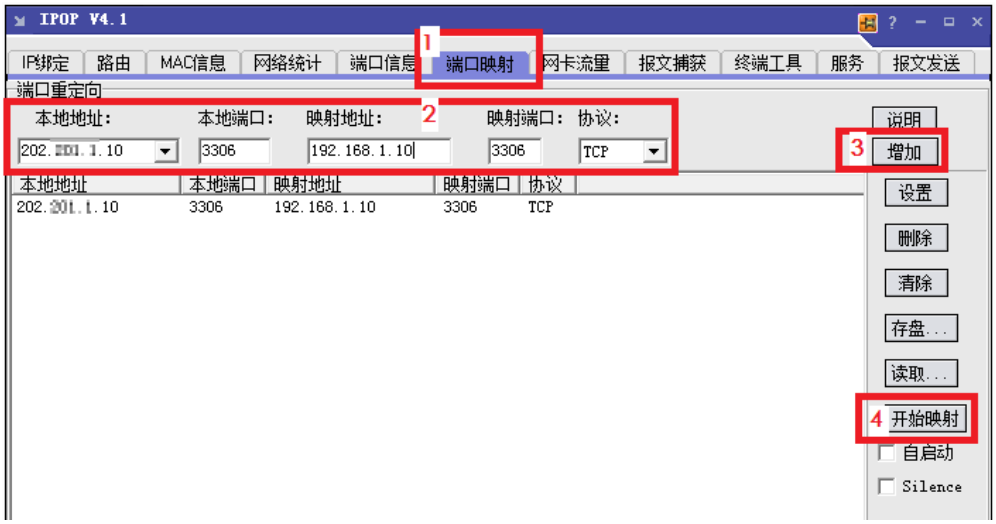
1. 在该机器上 **ping** 内网 MySQL 地址可以 **ping** 通，例如：**ping 192.168.1.8**。
2. 在另外一台可上网的机器上 **ping** 网关机的公网地址可以 **ping** 通，例如 **ping 202.*xx*.*xx*.10**。

步骤 2 下载端口映射工具 IPOP，在网关机上安装 IPOP。

步骤 3 运行端口映射工具，选择“端口映射”，如图 4-5 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4-5 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 4-6 所示。

图4-6 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_2097677531$" 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$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:748)
Caused by: oracle.jdbc.OracleDatabaseException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYSSMU3_2097677531$" 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。

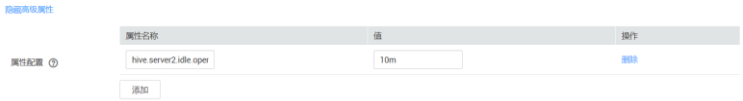
当用户选择 MongoDB 连接迁移失败时，用户需查看 MongoDB 连接中用户的权限信息，确保对指定库具备 ReadWrite 权限。

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 4-7 所示 c_date 字段为非索引列。

图4-7 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。

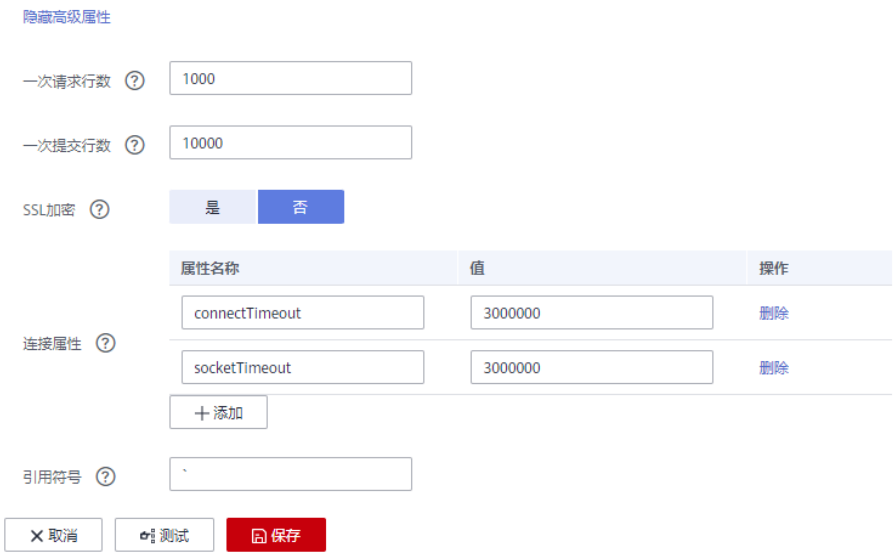
图4-8 连接信息

3. 单击“连接管理”，在“操作”列中，单击“连接”进行编辑。

图4-9 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4-10 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 MRS 组件的库、表、列进行操作，还需要参考 MRS 文档添加对应组件的库、表、列操作权限。
- 如果 CDM 集群为 2.9.0 之前的版本，或 MRS 集群为 3.1.0 之前的版本，则所创建的用户需要具备 Manager_administrator 或 System_administrator 权限，才能在 CDM 创建连接。
- 仅具备 Manager_tenant 或 Manager_auditor 权限，无法创建连接。

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 数据架构

4.4.1 码表和数据标准有什么关系？

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

4.4.2 关系建模和维度建模的区别？

- 关系建模为事务性模型，对应三范式建模。

- 维度建模为分析性模型，主要包括事实表、维度表的设计，多用于实现多角度、多层次的数据查询和分析。

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

4.4.3 数据架构支持哪些数据建模方法？

DataArts Studio 数据架构支持的建模方法有以下两种：

- **关系建模**

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

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

  - **概念模型**：是从用户的视角，主要从业务流程、活动中涉及的主要业务数据出发，抽象出关键的业务实体，并描述这些实体间的关系。
  - **逻辑模型**：是概念模型的进一步细化，通过实体、属性和关系勾勒出企业的业务信息蓝图，是 IT 和业务人员沟通的桥梁。逻辑数据模型是一组规范化的逻辑表结构，逻辑数据模型是根据业务规则确定的，关于业务对象、业务对象的数据项及业务对象之间关系的基本蓝图。
  - **物理模型**：是在逻辑数据模型的基础上，考虑各种具体的技术实现因素，进行数据库体系结构设计，真正实现数据在数据库中的存放，例如：所选的数据仓库是 DWS。

- **维度建模**

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

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

  典型的维度模型有星形模型，以及在一些特殊场景下使用的雪花模型。

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

4.4.4 规范化的数据如何使用？

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

4.4.5 数据架构支持逆向数据库吗？

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

4.4.6 数据架构中的指标与数据质量的指标的区别？

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

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

4.4.7 为什么数据架构更新表后无变化？

用户在数据架构中更新了表，但实际上表数据并无变化，这是因为在更新前未对数据表更新方式做配置。配置数据表更新方式操作如下：

1. 单击“数据架构 > 配置中心”。
2. 单击“功能配置”页签。
3. “数据表更新方式”选择“重建数据表”。
4. 单击“确定”，完成配置。

4.4.8 表是否可配置生命周期管理？

目前暂不支持表生命周期管理的配置。

4.5 数据开发

4.5.1 数据开发可以创建多少个作业，作业中的节点数是否有限制？

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

另外，系统支持用户根据实际需求调整最大配额。如有需求，请进行申请。

4.5.2 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。

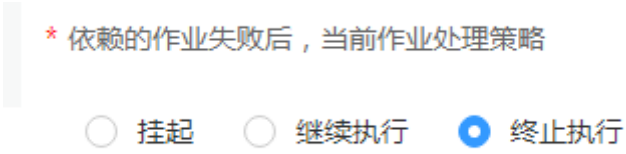
这是因为在数据开发中，单个作业最多允许 5 个实例并行执行，如果作业实际执行时间大于作业配置的调度周期，会导致后面批次的作业实例堆积，从而出现上述问题。

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。

4.5.3 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4-11 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。

4.5.4 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。

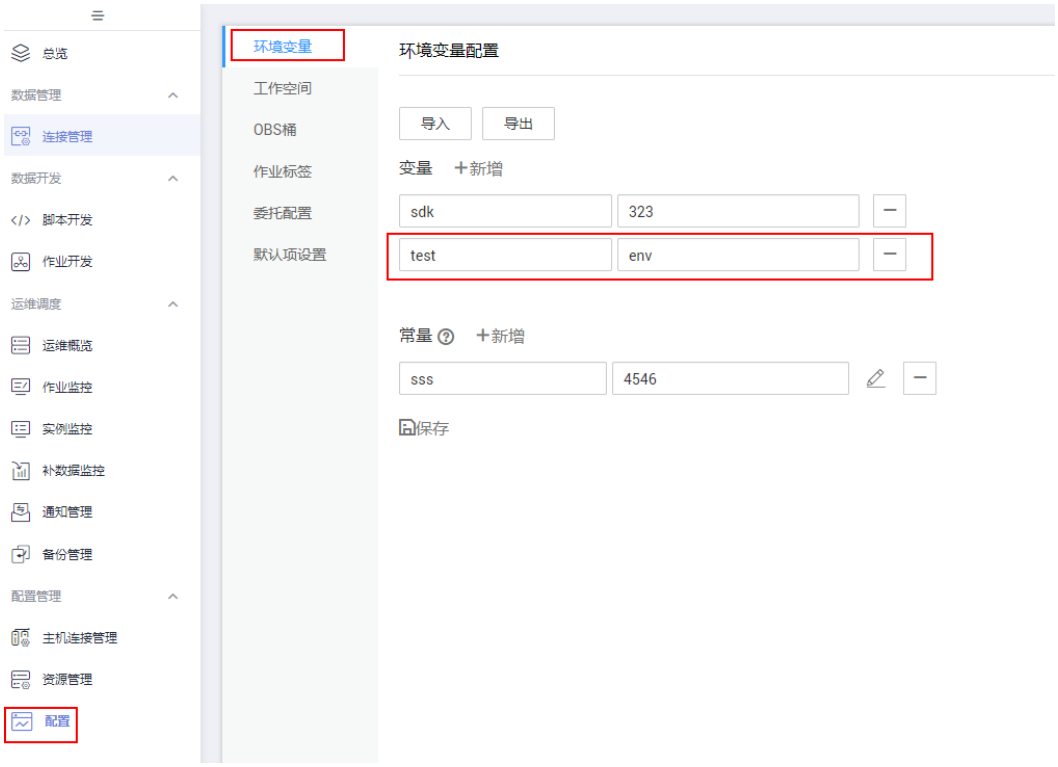
4.5.5 环境变量、作业参数、脚本参数有什么区别和联系？

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：

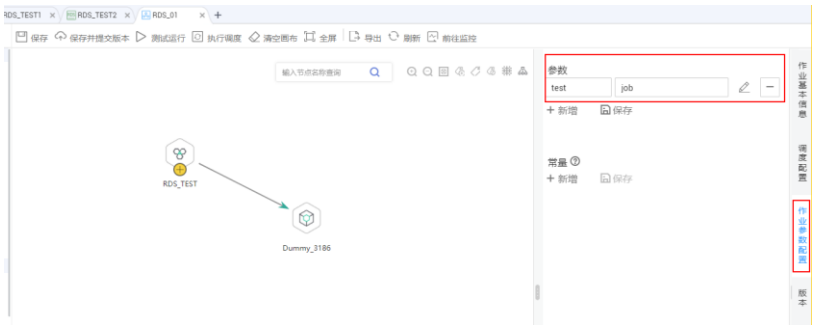
- 环境变量中支持定义变量和常量，环境变量的作用范围为当前工作空间。
  - 变量是指不同的空间下取值不同，需要重新配置值，比如“工作空间名称”变量，这个值在不同的空间下配置不一样，导出导入后需要重新进行配置。
  - 常量是指在不同的空间下都是一样的，导入的时候，不需要重新配置值。

图4-12 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4-13 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 4-14 所示；通过作业调度时可通过节点属性赋值，如图 4-15 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4-14 独立执行时的脚本参数

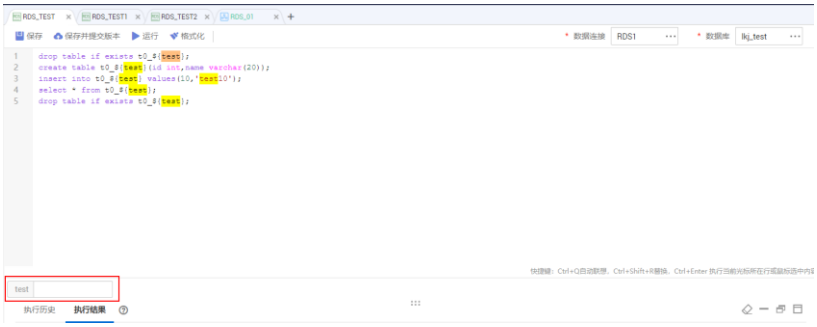


图4-15 作业调度时的脚本参数

4.5.6 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。

4.5.7 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。

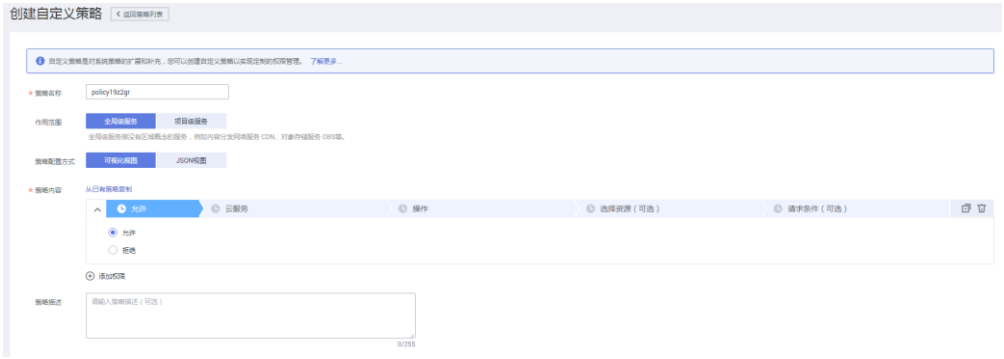
先创建自定义策略（查询指定条件下的委托列表），再通过给用户组授予自定义策略来进行精细的访问控制。

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

4.5.8 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4-16 查看调度周期

4.5.9 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。

4.5.10 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4-17 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。

4.5.11 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。

4.5.12 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。

4.5.13 作业开发时，测试运行后如何查看运行日志？

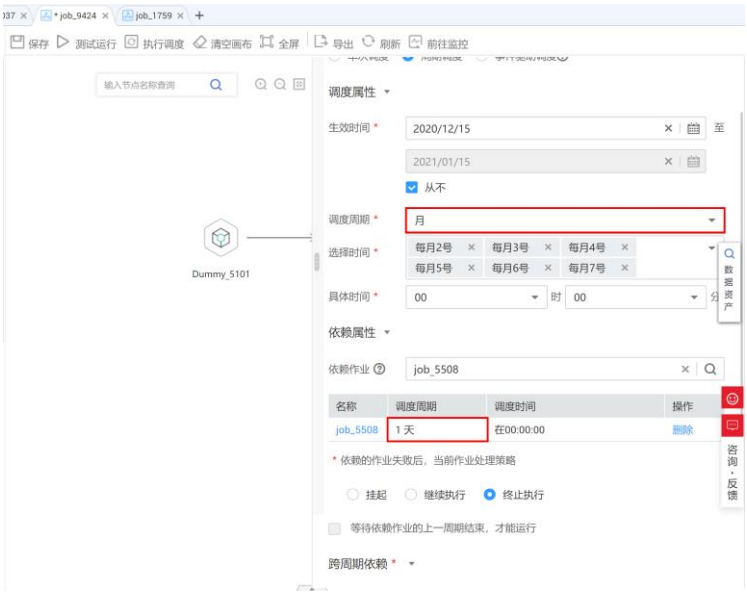
方式 1：待节点测试运行完成后，在当前节点鼠标右键选择查看日志。

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。

4.5.14 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4-18 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。

4.5.15 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。

4.5.16 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。

4.5.17 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。

4.5.18 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。

4.5.19 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。

4.5.20 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4-19 提交版本

4.5.21 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。

4.5.22 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。

4.5.23 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。

4.5.24 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。

4.5.25 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。

4.5.26 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4-20 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。

4.5.27 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

4.5.28 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。

4.6 数据质量

4.6.1 质量作业和对账作业有什么区别？

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

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

4.6.2 如何确认质量作业或对账作业已经阻塞？

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

4.6.3 如何手工重启阻塞的质量作业或对账作业？

阻塞的作业需要进行手工重启，如不重启 1 天内也会因作业超时自动结束该作业。

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

4.6.4 怎样查看质量规则模板关联的作业？

步骤 1 单击待操作规则模板操作列的“发布历史”。

图4-21 发布历史

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

图4-22 查看关联作业

----结束

4.6.5 用户在执行质量作业时提示无 MRS 权限怎么办？

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

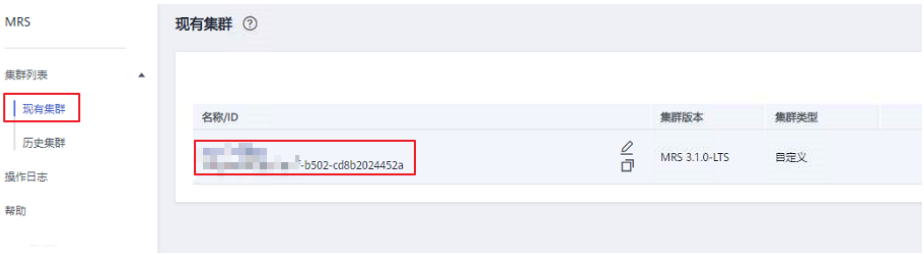
此类问题一般是由于用户不具备 MRS 集群操作权限导致的。

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

操作如下：

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

图4-23 MRS 集群实例



步骤 2 在“IAM 用户同步”处，单击同步。

图4-24 单击同步

步骤 3 在操作日志处查看操作结果。

图4-25 操作日志

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

注意

在步骤 4 中，需要创建一个与当前主账号同名的账号。

----结束

4.7 数据目录

4.7.1 数据目录组件有什么用？

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

4.7.2 数据目录支持采集哪些对象的资产？

数据目录目前支持采集的资产有：数据仓库服务（DWS）、MapReduce 服务（MRS HBase）、MapReduce 服务（MRS Hive）、MySQL、云数据库 RDS（DataArts Studio 仅支持 MySQL 和 PostgreSQL 数据库）。

4.7.3 什么是数据血缘关系？

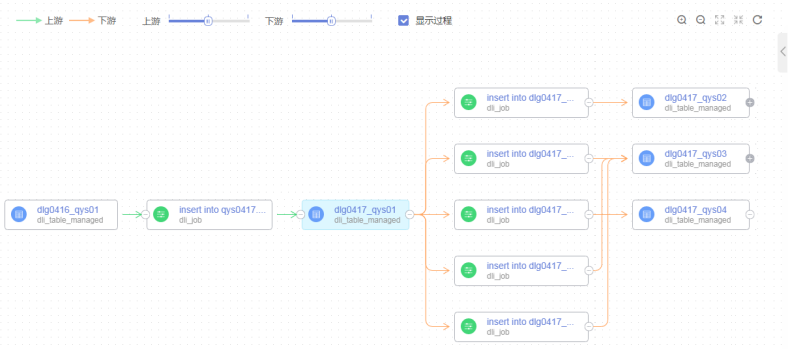
大数据时代，数据爆发性增长，海量的、各种类型的数据在快速产生。这些庞大复杂的数据信息，通过联姻融合、转换变换、流转流通，又生成新的数据，汇聚成数据的海洋。

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

- **归属性**：一般来说，特定的数据归属特定的组织或者个人，数据具有归属性。
- **多源性**：同一个数据可以有多个来源（多个父亲）。一个数据可以是多个数据经过加工而生成的，而且这种加工过程可以是多个。
- **可追溯性**：数据的血缘关系，体现了数据的生命周期，体现了数据从产生到消亡的整个过程，具备可追溯性。

- **层次性**：数据的血缘关系是有层次的。对数据的分类、归纳、总结等对数据进行的描述信息又形成了新的数据，不同程度的描述信息形成了数据的层次。

图4-26 数据血缘关系示例



4.7.4 数据目录如何可视化展示数据血缘？

数据血缘展示，首先要需要有相关的作业调度，其次要进行元数据采集。

4.8 数据服务

4.8.1 创建 API 时提示代理调用失败，怎么办？

需要在空余时间对 CDM 集群进行重启释放内存。

4.8.2 数据服务 API 接口，访问“测试 APP”，填写了相关参数，但是后台报错要怎么处理？

在调用 API 时配置参数 header parameter。

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

4.8.3 使用 API 时报错，请问有什么办法可以解决？

使用 API 时需注意，每个子域名每天最多可以访问 1000 次。

4.8.4 API 传参是否支持传递操作符？

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

4.8.5 数据服务专享版提供的 API 配额已满怎么解决？

如果数据服务专享版提供的 API 配额已满，无法创建新的 API 时可修改 API 配额。

A 修订记录

发布日期	修订说明
2023-02-28	产品版本升级。
2023-02-28	第一次正式发布。