数据治理中心

用户指南

文档版本 01

发布日期 2025-02-27

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

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

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

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：
https://www.huawei.com/cn/psirt/vul-response-process
如企业客户须获取漏洞信息，请参见如下网址：
https://securitybulletin.huawei.com/enterprise/cn/security-advisory

目 录

1 DataArts Studio 使用流程

数据治理中心DataArts Studio是具有数据全生命周期管理、智能数据管理能力的一站式治理运营平台，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，加快数据变现，实现数字化转型。

DataArts Studio 使用流程简介

使用DataArts Studio平台，通常包括以下步骤：

表 1-1 DataArts Studio 全流程开发

主流程	说明	子任务	操作指导
流程设计	在使用DataArts Studio前，建议您通过流程设计提前分析业务情况，明确业务诉求，并结合DataArts Studio服务的能力进行业务流程设计。 1. 需求分析。分析业务情况，明确业务诉求，并提炼出数据治理流程的实现框架，支撑具体数据治理实施流程的设计。 2. 业务调研。明确DataArts Studio服务的能力边界，并分析后续的业务负载情况。 3. 流程设计。以实际业务情况结合DataArts Studio服务的业务能力，完成数据治理业务流程设计，后续的数据治理操作均基于所设计的业务流程完成。	1. 需求分析 2. 业务调研 3. 流程设计	流程设计与实际业务强相关，您可以参考**基于出租车出行数据的数据治理流程设计**进行流程设计，或通过**咨询**了解。
购买并配置DataArts Studio	如果您是第一次使用DataArts Studio，需要先完成注册华为账号、购买DataArts Studio实例、创建工作空间等一系列操作。	购买并配置DataArts Studio	**购买并配置DataArts Studio**

主流程	说明	子任务	操作指导
授权用户使用DataArts Studio	如果您需要授权其他IAM用户使用DataArts Studio，则需要完成创建用户并授权的操作。	授权用户使用DataArts Studio	授权用户使用DataArts Studio
管理中心	根据自身的业务特点和源数据类型，进行数据存储与分析系统的选型，选取合适的云服务用于存储源数据并进行数据查询和分析。然后，创建该云服务相应的数据连接。	新建数据连接	创建DataArts Studio数据连接
数据集成	通过DataArts Studio平台将源数据上传或者接入到云上。 数据集成提供同构/异构数据源之间批量数据迁移的服务，支持自建和云上的文件系统，以及关系数据库，数据仓库，NoSQL，大数据云服务，对象存储等数据源。	数据集成	支持的数据源 创建CDM集群 创建CDM与数据源之间的连接 新建表/文件迁移作业
数据目录（元数据采集）	为了在DataArts Studio对迁移到云上的原始数据层进行管理和监控，先对其元数据进行采集并监控。	元数据采集	采集数据源的元数据
数据架构	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。 根据业务需求设计关系模型、维度模型，在数据架构模块中，逐步建立模型中的对象，例如维度、事实表、指标、汇总表等。	添加审核人	添加审核人
		管理配置中心	管理配置中心
		流程设计	流程设计
		主题设计	主题设计
		码表管理	新建码表
		制定数据标准	新建数据标准
		关系建模	关系建模
		维度建模	维度建模
		业务指标	业务指标
		技术指标	技术指标
		数据集市建设	数据集市

主流程	说明	子任务	操作指导
数据开发	可管理多种大数据服务，提供一站式的大数据开发环境。 使用DataArts Studio数据开发，用户可进行数据管理、数据集成、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。	数据管理	数据管理流程
		脚本开发	脚本开发流程
		作业开发	作业开发流程
		运维调度	运维概览
数据质量	对业务指标和数据指标进行监控。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。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。	数据质量监控	新建数据质量规则 新建数据质量作业 新建数据对账作业
数据目录	在DataArts Studio数据目录模块中，您可以查看数据地图。	数据地图	查看工作空间内的数据资产
数据安全	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。在数据安全模块，您可以进行访问权限管理、敏感数据识别、隐私保护管理等操作。	统一权限治理	权限治理使用流程
		敏感数据治理	敏感数据治理流程
		隐私保护管理	隐私数据保护简介
数据服务	统一管理对内对外的API服务，提供快速将数据表生成数据API的能力。	开发API	购买并管理专享版集群 新建数据服务审核人 创建API 调试API 发布API 管理API 编排API 配置API调用流控策略 授权API调用
		调用API	申请API授权 通过不同方式调用API

2 购买并配置 DataArts Studio

2.1 购买 DataArts Studio 实例

DataArts Studio采用基础包+增量包的计费模式，其中基础包即DataArts Studio实例，购买方法请参见**购买DataArts Studio基础包**。

背景信息

- 只有拥有**DAYU Administrator**或**Tenant Administrator**权限的用户才可以购买DataArts Studio实例或DataArts Studio增量包。如需购买，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。

- 只有拥有**Security Administrator**权限的用户才创建云服务委托。云服务委托可将相关云服务的操作权限委托给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。

前提条件

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

VPC、子网、安全组的详细操作，请参见《**虚拟私有云用户指南**》。

登录 DataArts Studio 控制台

1. 登录华为云控制台。
2. 在控制台左上方，单击“服务列表”按钮 ☰，选择“ ”，进入DataArts Studio控制台。

购买 DataArts Studio 基础包

步骤1 进入**购买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。 选择区域时，您需要考虑以下几个因素： • 地理位置 一般情况下，建议就近选择靠近您或者您的目标用户的区域，这样可以减少网络时延，提高访问速度。 • 云服务之间的关系 如果多个云服务一起搭配使用，需要注意不同区域的云服务内网不互通。 例如DataArts Studio（包括管理中心、CDM等组件）需要与MRS、OBS等服务互通时，如果DataArts Studio与其他云服务处于不同区域的情况下，需要通过公网或者专线打通网络；而在同区域情况下，同子网、同安全组的不同实例默认网络互通。 • 资源的价格 不同区域的资源价格可能有差异，请参见**华为云服务价格详情**。 详情请参见**什么是可用区**。
企业项目	default	DataArts Studio实例默认工作空间关联的企业项目。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**》。 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工作空间的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购买一个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。 **说明** 未开通企业项目时，则每个IAM项目只允许创建1个DataArts Studio实例。
版本	初级版	选择需要购买的DataArts Studio版本，版本差异请参见**版本规格说明**。 **说明** 购买DataArts Studio实例时，会默认包含一个数据集成CDM集群，此集群规格建议用于作为连接代理。如需用于数据迁移作业，请购买更高规格的批量数据迁移增量包，详情请参考**购买批量数据迁移增量包**。

参数名称	样例	说明
计费方式	包年包月	当前DataArts Studio基础包仅支持包年包月计费方式。
实例名称	DataArts Studio-test	自定义DataArts Studio实例名称。实例名称不支持修改，请提前合理规划。
可用区	可用区1	选择DataArts Studio实例可用区，即数据集成CDM集群所在可用区。DataArts Studio实例通过数据集成CDM集群与其他服务实现网络互通。 第一次购买DataArts Studio实例或增量包时，可用区无要求。再次购买DataArts Studio实例或增量包时，是否将资源放在同一可用区内，主要取决于您对容灾能力和网络时延的要求。 ● 如果您的应用需要较高的容灾能力，建议您将资源部署在同一区域的不同可用区内。 ● 如果您的应用要求实例之间的网络延时较低，则建议您将资源创建在同一可用区内。 详情请参见**什么是可用区**。
虚拟私有云	vpc1	DataArts Studio实例中的数据集成CDM集群所属的VPC、子网、安全组。DataArts Studio实例通过数据集成CDM集群与其他服务实现网络互通。 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、子网、安全组的详细操作，请参见《**虚拟私有云用户指南**》。 **说明** ● 目前DataArts Studio实例创建完成后不支持切换默认CDM集群的VPC、子网、安全组，请谨慎选择。 ● 此处支持选择共享VPC子网，即由VPC的所有者将VPC内的子网共享给当前账号，由当前账号在购买DataArts Studio时选择共享VPC子网。通过共享VPC子网功能，可以简化网络配置，帮助您统一配置和运维多个账号下的资源，有助于提升资源的管控效率，降低运维成本。如何共享VPC子网，请参考《**共享VPC**》。
子网	subnet-1	
安全组	sg-1	
购买时长	1年	按您的需求选择购买的时长。
自动续费	-	勾选自动续费前的复选框，可实现自动按月或者按年续费。 购买时长为按月购买时，自动续费周期为1个月；购买时长为按年购买时，自动续费周期为1年。

参数名称	样例	说明
标签	标签键：key1 标签值：asd	通过为资源添加标签，可以对资源进行自定义标记，实现资源的分类。 **说明** 如您的账号归属某个**组织**，且该组织已经设定DataArts Studio服务的相关**标签策略**，则需按照标签策略规则添加标签。标签如果不符合标签策略的规则，则可能会导致实例创建失败，请联系组织管理员了解标签策略详情。 当前DataArts Studio实例标签支持的使用场景如下： • 当拥有大量云资源时，您可以按使用者、维护者或用途等各类维度为云资源（包括DataArts Studio实例）添加标签，最后您可以在标签管理服务（简称TMS）通过标签识别、管理多种云资源，使资源管理变得更加轻松。 • 当拥有多个DataArts Studio实例时，您可以按使用者、维护者或用途等各类维度为各实例添加标签，然后在DataArts Studio实例列表页面，可以通过标签搜索、识别DataArts Studio实例。 标签由标签键和标签值组成。在添加标签时，标签键和标签值可以选择在标签管理服务（简称TMS）中创建的预定义标签，也可以直接输入自定义的标签。然后单击输入框右侧的“添加”，即可成功添加一条标签。 **说明** 预定义标签需要预先在标签管理服务中创建好，然后才能进行选择。您可以通过单击“查看预定义标签”进入标签管理服务的“预定义标签”页面，然后单击“创建标签”来创建新的预定义标签，具体请参见《标签管理服务用户指南》中的“**创建预定义标签**”章节。 另外，DataArts Studio实例最多支持添加20个标签，标签的键名不能重复，一个“标签键”只能添加一个对应“标签值”。

步骤3 查看当前配置，确认无误后单击“立即购买”。

步骤4 单击“提交订单”，付款成功后等待实例创建成功，即可在首页看到已开通的实例。

图 2-1 查看 DataArts Studio 实例

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。

须知

只有拥有**Security Administrator**权限的用户才创建云服务委托。云服务委托可将相关云服务的操作权限委托给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”，系统会在IAM服务自动创建dlg_agency默认委托。

- 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。

图 2-2 云资源访问授权

步骤6 在已购买的实例中单击“进入控制台”，进入DataArts Studio控制台。

----结束

2.2 购买 DataArts Studio 增量包

2.2.1 如何选择增量包

DataArts Studio采用基础包+增量包的计费模式。如果购买的基础包无法满足您的使用需求，则您需要额外购买增量包。

DataArts Studio 增量包

当前DataArts Studio支持的增量包如表2-2所示。

表 2-2 增量包介绍

增量包类型	增量包说明	购买场景说明	购买方式
批量数据迁移增量包	批量数据迁移增量包对应数据集成CDM集群。 • 通过购买一个按需计费方式的批量数据迁移增量包，系统会按照您所选规格自动创建一个数据集成CDM集群。 • 通过购买一个套餐包方式的批量数据迁移增量包，系统不自动创建CDM集群，而是在生效期内的每个计费月内按月提供745小时/月的使用时长，在绑定区域为在DataArts Studio控制台购买的对应实例规格的CDM集群使用。 数据集成CDM集群可用于如下场景： • 用于创建并运行数据迁移作业，提供数据上云和数据入湖的集成能力。 • 作为在管理中心创建连接时的Agent代理，为DataArts Studio实例和数据源直接提供网络通道。	DataArts Studio实例中已经包含一个仅用于测试、试用等非正式业务场景的CDM集群。 • 如果该集群已经满足您的使用需求，则无需再购买批量数据迁移增量包。 • 如果您需要CDM集群用于满足业务需求，请通过按需计费方式购买批量数据迁移增量包。 • 如果您需要为购买的CDM集群匹配套餐包用于降低使用成本，请通过套餐包方式购买批量数据迁移增量包。 **说明** DataArts Studio实例赠送的CDM集群，由于规格限制，仅用于测试、试用等非正式业务场景。用于业务场景的CDM集群可以通过“批量数据迁移增量包”进行购买，且不建议同时作为数据连接Agent代理和运行数据迁移作业使用。	• 按需计费 • 套餐包

增量包类型	增量包说明	购买场景说明	购买方式
数据集成资源组增量包	数据集成资源组增量包对应数据集成实时作业所需的资源组。数据集成资源组提供数据上云和数据入湖出湖的集成能力，全向导式配置和管理，支持单表、整库、分库分表、全量及增量、实时数据集成。 • 通过购买一个按需计费方式的数据集成资源组增量包，系统会按照您所选规格自动创建一个数据集成实时作业所需的资源组。 • 通过购买一个套餐包方式的数据集成资源组增量包，系统不自动创建新的资源组，而是在生效期内的每个计费月内按月提供745小时/月的使用时长，在绑定区域为在DataArts Studio控制台购买的对应资源组使用。 数据集成资源组可用于如下场景： 用于创建并运行数据迁移作业，提供数据上云和数据入湖的集成能力。	DataArts Studio实例中默认不包含数据集成资源组，如果您需要使用数据离线、实时迁移功能，请创建数据集成资源组增量包。	• 按需计费 • 套餐包
数据服务专享集群增量包	数据服务专享集群增量包对应数据服务专享版集群。创建一个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。 数据服务定位于标准化的数据服务平台，提供了快速将数据表生成数据API的能力，帮助您简单、快速、低成本、低风险地实现数据开放。数据服务需要在创建数据服务专享集群后才能使用。	DataArts Studio实例中默认不包含数据服务专享集群，如果您需要使用数据服务，请创建数据服务专享集群增量包。	包年包月

增量包类型	增量包说明	购买场景说明	购买方式
作业节点调度次数/天增量包	作业节点调度次数/天增量包用于扩充作业节点调度次数/天配额。 不同版本的DataArts Studio实例，默认提供了不同的作业节点调度次数/天规格限制。该规格是以每天执行的数据开发作业、质量作业、对账作业、业务场景和元数据采集作业的调度次数之和计算的。其中数据开发作业的每天调度次数，是以节点（包含Dummy节点）为粒度进行度量的，另外补数据任务也会计入度量次数，但测试运行、失败重试不会计入。您可以在DataArts Studio实例卡片上通过“更多 > 配额使用量”查看该配额情况。 **说明** DataArts Studio实例中数据开发作业节点运行的并行数上限，与当前实例的作业节点调度次数/天配额有关。 • 当“作业节点调度次数/天配额<=500”时，节点运行的并行数上限为10。 • 当“500<作业节点调度次数/天配额<=5000”时，节点运行的并行数上限为50。 • 当“5000<作业节点调度次数/天配额<=20000”时，节点运行的并行数上限为100。 • 当“20000<作业节点调度次数/天配额<=40000”时，节点运行的并行数上限为200。 • 当“40000<作业节点调度次数/天配额<=80000”时，节点运行的并行数上限为300。 • 当“作业节点调度次数/天配额>80000”时，节点运行的并行数上限为400。	当您的每日作业节点调度次数接近、达到该规格，或需要扩充数据开发作业节点运行的并行数上限时，建议购买作业节点调度次数/天增量包，以避免作业调度和运行并发数受限。 **说明** 当作业节点调度的已使用次数+运行中次数+本日将运行次数之和大于此版本规格，执行调度批处理作业或者启动实时作业时就会提示作业节点调度次数/天超过配额。	包年包月
技术资产数量增量包	技术资产数量增量包用于扩充技术资产数量配额。 不同版本的DataArts Studio实例，默认提供了不同的技术资产数量规格限制。该规格是以数据目录中表和OBS文件的数量之和计算的。您可以在DataArts Studio实例卡片上通过“更多 > 配额使用量”查看该配额情况。	当您的技术资产数量接近或达到该规格时，建议购买技术资产数量增量包，以避免资产采集受限。	包年包月

增量包类型	增量包说明	购买场景说明	购买方式
数据模型数量增量包	数据模型数量增量包用于扩充数据模型数量配额。 不同版本的DataArts Studio实例，默认提供了不同的数据模型数量规格限制。该规格是以数据架构中逻辑模型、物理模型、维度表、事实表和汇总表的数量之和计算的。您可以在DataArts Studio实例卡片上通过“更多 > 配额使用量”查看该配额情况。	当您的数据模型数量接近或达到该规格时，建议购买数据模型数量增量包，以避免数据架构设计受限。	包年包月

2.2.2 购买批量数据迁移增量包

批量数据迁移增量包对应数据集成CDM集群。

- 通过购买一个按需计费方式的批量数据迁移增量包，系统会按照您所选规格自动创建一个数据集成CDM集群。
- 通过购买一个套餐包方式的批量数据迁移增量包，系统不自动创建CDM集群，而是在生效期内的每个计费月内按月提供745小时/月的使用时长，在绑定区域为在DataArts Studio控制台购买的对应实例规格的CDM集群使用。

数据集成CDM集群可用于如下场景：

- 用于创建并运行数据迁移作业，提供数据上云和数据入湖的集成能力。
- 作为在管理中心创建连接时的Agent代理，为DataArts Studio实例和数据源直接提供网络通道。

DataArts Studio实例中已经包含一个仅用于测试、试用等非正式业务场景的CDM集群。

- 如果该集群已经满足您的使用需求，则无需再购买批量数据迁移增量包。
- 如果您需要CDM集群用于满足业务需求，请通过按需计费方式购买批量数据迁移增量包。
- 如果您需要为购买的CDM集群匹配套餐包用于降低使用成本，请通过套餐包方式购买批量数据迁移增量包。

说明

DataArts Studio实例赠送的CDM集群，由于规格限制，仅用于测试、试用等非正式业务场景。用于业务场景的CDM集群可以通过“批量数据迁移增量包”进行购买，且不建议同时作为数据连接Agent代理和运行数据迁移作业使用。

背景信息

- 套餐包（按需资源包）方式购买批量数据迁移增量包时，需注意如下几点：
  - 套餐包（按需资源包）方式购买批量数据迁移增量包后，系统不自动创建CDM集群，而是在生效期内的每个计费月内按月提供745小时/月的使用时长，在绑定区域为在DataArts Studio控制台购买的对应实例规格的CDM集群使用。
  - 套餐包（按需资源包）仅支持给DataArts Studio控制台购买的CDM集群使用；在CDM控制台购买的CDM集群，不支持使用DataArts Studio增量包形式

购买的套餐包（按需资源包），仅支持使用在云数据迁移CDM服务控制台购买的折扣套餐（按需资源包）。

- 如果当前绑定区域有1个或多个对应实例规格的CDM集群，则扣费方式是先扣除已购买资源包内的时长额度，超出部分以按需计费的方式进行结算（资源包对应多个集群时，会出现每月订购周期内可使用时长不足的情况）。

  例如购买了1个月的套餐包（745小时/月），按区域和实例规格匹配到两个CDM集群后，从当前开始的1个月订购有效期内，两个集群同时使用只能使用745/2=372.5小时，约15.5天，剩余时间内两个集群按照按需计费的方式结算费用。
- 如果当前绑定区域没有对应实例规格的CDM集群，购买套餐包后不会消耗所购买的时长；但在生效期内，若未使用CDM集群，套餐包也不会延期。建议您先安排好服务使用计划，再购买套餐包。
- 如果您希望享受套餐包的优惠价格，需要先购买一个“套餐包”增量包，再购买一个和套餐包具有相同区域和规格的“按需计费”增量包。
- 如果您先购买一个“按需计费”增量包，再购买一个相同区域和规格的“套餐包”增量包，则在购买套餐包之前已经产生的费用按“按需计费”计费，购买套餐包之后的费用按“套餐包”计费。

- 您可以在DataArts Studio实例卡片上，通过“更多 > 查看增量包”，查看已购买的增量包。
- 数据集成CDM集群可以在DataArts Studio控制台以增量包的形式购买，也可以在云数据迁移CDM服务控制台直接购买。二者差异体现在如下方面：

  a. 套餐计费：在DataArts Studio控制台购买的CDM集群，套餐计费时仅支持在DataArts Studio控制台购买的套餐包；在CDM控制台购买的CDM集群，套餐计费仅支持在云数据迁移CDM服务控制台购买的折扣套餐。

  b. 权限控制：在DataArts Studio控制台购买的CDM集群，按照DataArts Studio的权限体系进行权限管理；在CDM控制台购买的CDM集群，按照云数据迁移CDM服务的权限体系进行权限管理。

  c. 使用场景：在DataArts Studio控制台购买的CDM集群按工作空间隔离，需要在关联的工作空间使用；在CDM控制台购买的CDM集群，不支持DataArts Studio工作空间级别的资源隔离，所有DataArts Studio工作空间均可使用。

  推荐您在DataArts Studio控制台以增量包的形式购买，本章节以此为例进行说明。

按需计费方式购买数据集成集群

购买“按需计费”增量包，系统会按照您所选规格自动创建一个数据集成CDM集群。

1. 单击已开通实例卡片上的“购买增量包”。
2. 进入购买DataArts Studio增量包页面，参见**表2-3**进行配置。

表 2-3 配置数据集成的增量包

参数	说明
增量包类型	选择批量数据迁移增量包。
计费方式	选择按需计费。

参数	说明
可用区	第一次购买DataArts Studio实例或增量包时，可用区无要求。 再次购买DataArts Studio实例或增量包时，是否将资源放在同一可用区内，主要取决于您对容灾能力和网络时延的要求。 ● 如果您的应用需要较高的容灾能力，建议您将资源部署在同一区域的不同可用区内。 ● 如果您的应用要求实例之间的网络延时较低，则建议您将资源创建在同一可用区内。 详情请参见**什么是可用区**。
工作空间	选择需要使用批量数据迁移增量包的工作空间。只有在关联了工作空间后，才能在此工作空间中使用创建的CDM集群。
企业项目	当关联了多个工作空间后，需要为CDM集群指定一个企业项目。
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：大规格，8核CPU、16G内存的虚拟机，最大带宽/基准带宽为3/0.8 Gbps，集群作业并发数上限为16。 ● cdm.xlarge：超大规格，16核CPU、32G内存的虚拟机，最大带宽/基准带宽为10/4 Gbps，集群作业并发数上限为32，适合使用10GE高速带宽进行TB级以上的数据量迁移。 ● cdm.4xlarge：4倍超大规格，64核CPU、128G内存的虚拟机，最大带宽/基准带宽为40/36 Gbps，集群作业并发数上限为128。
虚拟私有云 子网 安全组	DataArts Studio实例中的数据集成CDM集群所属的VPC、子网、安全组。 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、子网、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VPC、子网、安全组，请谨慎选择。 ● 此处支持选择共享VPC子网，即由VPC的所有者将VPC内的子网共享给当前账号，由当前账号在购买CDM集群时选择共享VPC子网。通过共享VPC子网功能，可以简化网络配置，帮助您统一配置和运维多个账号下的资源，有助于提升资源的管控效率，降低运维成本。如何共享VPC子网，请参考《**共享VPC**》。

参数	说明
IPv6双栈支持	当配置的子网支持IPv6后，可选择是否开启IPv6双栈支持。 开启IPv6双栈后，集群内网IP支持IPv4和IPv6，可通过IPv4或IPv6内网地址访问集群。 **说明** • 开启IPv6双栈后，仅支持选择已开启IPv6网段的子网。如果待选择的子网未开启IPv6网段，则需要在虚拟私有云VPC服务中开启。 • 仅支持为内网IP启用IPv6双栈，暂不支持为公网IP启用IPv6双栈。

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“立即购买”，确认规格后单击“创建”。
4. 购买成功后，即可返回对应的工作空间查看已购买的数据集成集群。

套餐包方式购买数据集成集群

如果您希望享受“套餐包”的优惠价格，您需要先购买一个“套餐包”增量包，再购买一个和“套餐包”增量包具有相同区域和规格的“按需计费”增量包。

1. 单击已开通实例卡片上的“购买增量包”。
2. 进入购买DataArts Studio增量包页面，按照如下配置：
   a. 增量包类型：选择批量数据迁移增量包。
   b. 计费方式：选择套餐包。
   c. 购买时长：表示此套餐包的有效时长。
   d. 购买数量：表示购买套餐包的数量。例如当购买时长选择1个月，购买数量选择2，那么您将拥有1490小时的额度，有效期是1个月。
3. 单击“立即购买”，确认规格后提交订单。
4. 购买套餐包成功后，系统不会自动创建数据集成集群。此时您还需要参考**按需计费方式购买数据集成集群**再购买一个和“套餐包”具有相同区域和规格的“按需计费”增量包，创建成功后您即可享受套餐包的优惠价格。

2.2.3 购买数据集成资源组增量包

数据集成资源组增量包对应数据集成实时作业所需的资源组。数据集成资源组提供数据上云和数据入湖出湖的集成能力，全向导式配置和管理，支持单表、整库、分库分表、全量+增量及增量同步等不同场景的数据迁移。

- 通过购买一个按需计费方式的数据集成资源组增量包，系统会按照您所选规格自动创建一个数据集成实时作业所需的资源组。
- 通过购买一个套餐包方式的数据集成资源组增量包，系统不自动创建新的资源组，而是在生效期内的每个计费月内按月提供745小时/月的使用时长，在绑定区域为在DataArts Studio控制台购买的对应资源组使用。

数据集成资源组可用于如下场景：

用于创建并运行实时处理集成作业，提供数据上云和数据入湖的集成能力。

DataArts Studio实例中默认不包含数据集成资源组，如果您需要使用数据实时迁移功能，请创建数据集成资源组增量包。

背景信息

- 创建数据集成资源组增量包，系统会按照您所选规格自动创建一个数据集成实时作业所需的资源组。
- 套餐包（按需资源包）方式购买数据集成资源组增量包时，需注意如下几点：
  - 套餐包（按需资源包）方式购买数据集成资源组增量包后，系统不自动创建新的数据集成实时作业所需的资源组，而是在生效期内的每个计费月内按月提供745小时/月的使用时长，在绑定区域为在DataArts Studio控制台购买的对应资源组使用。
  - 如果当前绑定区域有一个或多个对应资源组，则扣费方式是先扣除已购买资源包内的时长额度，超出部分以按需计费的方式进行结算（资源包对应多个集群时，会出现每月订购周期内可使用时长不足的情况）。

    例如购买了1个月的套餐包（745小时/月），按区域和实例规格匹配到两个资源后，从当前开始的1个月订购有效期内，两个资源组同时使用只能使用745/2=372.5小时，约15.5天，剩余时间内两个资源组按照按需计费的方式结算费用。
  - 如果当前绑定区域没有对应资源组，购买套餐包后不会消耗所购买的时长；但在生效期内，若未使用资源组，套餐包也不会延期。建议您先安排好服务使用计划，再购买套餐包。
  - 如果您希望享受套餐包的优惠价格，需要先购买一个“套餐包”增量包，再购买一个和套餐包具有相同区域和规格的“按需计费”增量包。
  - 如果您先购买一个“按需计费”增量包，再购买一个相同区域和规格的“套餐包”增量包，则在购买套餐包之前已经产生的费用按“按需计费”计费，购买套餐包之后的费用按“套餐包”计费。
- 您可以在DataArts Studio实例卡片上，通过“更多 > 查看增量包”，查看已购买的增量包。
- 不同规格类型的资源组，计费不同，详情请查看**计费说明**，您也可以通过DataArts Studio提供的**价格计算器**，选择您需要的区域、规格，快速计算出购买DataArts Studio资源组的参考价格。

按需计费方式购买数据集成资源组

购买“按需计费”增量包，系统会按照您所选规格自动创建一个数据集成实时作业所需的资源组。

1. 通过以下方式购买资源组。
   - 方式一：

     单击已开通实例卡片上的“购买增量包”。

图 2-3 购买增量包

- 方式二：
  i. 选择实例，单击“进入控制台”。
  ii. 单击右上角“购买增量包”，进入购买DataArts Studio增量包页面。
- 方式三：
  i. 选择实例，单击“更多 > 资源管理”，进入资源管理页面。

图 2-4 进入资源管理

ii. 在“实时资源管理”页签，单击“购买资源组”，进入购买DataArts Studio增量包页面。

图 2-5 购买资源组

2. 进入购买DataArts Studio增量包页面，参见**表2-4**进行配置。

图 2-6 创建数据集成资源组增量包

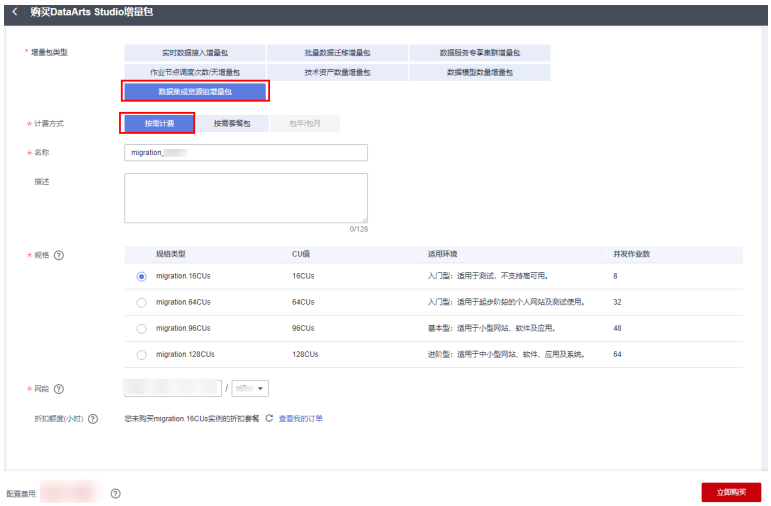


表 2-4 配置数据集成的增量包

参数	说明
增量包类型	选择数据集成资源组增量包。
计费方式	选择按需计费。
名称	自定义数据集成资源组名称。
描述	可以自定义对当前数据集成资源组的描述。
规格	选择资源组的规格类型，即CU值、适用环境、创建作业的最大数量。 不同规格的资源组支持迁移的任务数或创建作业的最大数存在上限，您需要根据业务需要选择合适规格的资源组。单个作业（最少需要2 CU）最多支持创建50张表。 小规格：16CUs，创建作业的最大数量为7。适用于测试，不支持高可用，不建议选择。 中规格：64CUs，创建作业的最大数量为32。 大规格：96CUs，创建作业的最大数量为48。 超大规格：128CUs，创建作业的最大数量为64。
网段	建议使用网段范围： • 10.0.0.0~10.255.0.0/8~19 • 172.16.0.0~172.31.0.0/12~19 • 192.168.0.0~192.168.0.0/16 ~19 **说明** • 为了后续使用对等连接打通网络，这里需设置与源端、目的端集群或实例不重叠的网段。如果重叠会导致网络不通。 • 受CCE底层逻辑限制，网段掩码最高为19位，20位之后不可选。
折扣额度(小时)	折扣套餐是按月或按年预先支付费用，相比按需计费节省15%到29%的费用。

须知

- 资源组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。
- 使用中的按需资源包不支持退订，详情可查看**不可退订**。

3. 单击“立即购买”，确认规格后提交。创建资源组失败时，如果显示配额问题，请联系工作人员申请配额。
4. 购买成功后，即可返回对应的工作空间查看已购买的数据集成资源组。

套餐包方式购买数据集成资源组

如果您希望享受“套餐包”的优惠价格，您需要先购买一个“套餐包”增量包，再购买一个和“套餐包”增量包具有相同区域和规格的“按需计费”增量包。

1. 进入购买DataArts Studio增量包页面，按照如下配置：
   a. 增量包类型：选择数据集成资源组增量包。
   b. 计费方式：选择套餐包。
   c. 购买时长：表示此套餐包的有效时长。
   d. 购买数量：表示购买套餐包的数量。例如当购买时长选择1个月，购买数量选择2，那么您将拥有1490小时的额度，有效期是1个月。
2. 单击“立即购买”，确认规格后提交订单。
3. 购买套餐包成功后，系统不会自动创建数据集成资源组。此时您还需要参考**按需计费方式购买数据集成资源组**再购买一个和“套餐包”具有相同区域和规格的“按需计费”增量包，创建成功后您即可享受套餐包的优惠价格。

2.2.4 购买数据服务专享集群增量包

数据服务专享集群增量包对应数据服务专享版集群。创建一个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。

数据服务定位于标准化的数据服务平台，提供了快速将数据表生成数据API的能力，帮助您简单、快速、低成本、低风险地实现数据开放。数据服务需要在创建数据服务专享集群后才能使用。

DataArts Studio实例中默认不包含数据服务专享集群，如果您需要使用数据服务，请创建数据服务专享集群增量包。

背景信息

- 您可以在DataArts Studio实例卡片上，通过“更多 > 查看增量包”，查看已购买的增量包。

购买数据服务专享集群

步骤1 单击已开通实例卡片上的“购买增量包”。

步骤2 进入购买DataArts Studio增量包页面，参见**表2-5**进行配置。

表 2-5 购买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。
计费方式	实例收费方式，当前支持“包年包月”。
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。集群购买成功后，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。 如果需要在其他工作空间内使用该集群，您可以在集群创建成功后，参考**管理集群共享**将该集群共享给其他工作空间。
可用区	选择数据服务专享集群所在的可用区。 支持单AZ和多AZ两种部署方式。推荐使用多AZ方式。 ● 单AZ：仅可以选择1个AZ，集群节点部署在同一AZ上。 ● 多AZ：可选择2-10个AZ，集群节点部署在不同AZ上，以提升集群的容灾能力。 详情请参见**什么是可用区**。
集群名称	集群名称必须以字母开头,可以包含字母、数字、中划线或者下划线,不能包含其他的特殊字符。输入长度不能小于5个字符。
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数量的支持能力不同。
公网入口	开启“公网入口”，创建集群时会为集群自动绑定一个新建的弹性公网IP，后续可以通过此公网IP地址调用专享版API。该功能新建的弹性公网IP不会计入收费项。 如果您存在需要本地调用或跨网调用API的使用场景，建议开启。如果在创建集群时未开启公网入口，后续则不再支持绑定EIP。
带宽大小	可配置公网带宽范围。

参数项	说明
虚拟私有云 子网 安全组	DataArts Studio实例中的数据服务专享版集群所属的VPC、子网、安全组。 在相同VPC、子网、安全组中的云服务资源（如ECS），可以使用数据服务专享版实例的私有地址调用API。建议将专享版集群和您的其他关联业务配置一个相同的VPC、子网、安全组，确保网络安全的同时，方便网络配置。 VPC、子网、安全组的详细操作，请参见《**虚拟私有云用户指南**》。 **说明** • 目前专享版集群创建完成后不支持切换VPC、子网、安全组，请谨慎选择。 • 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。 • 此处支持选择共享VPC子网，即由VPC的所有者将VPC内的子网共享给当前账号，由当前账号在购买数据服务专享版集群时选择共享VPC子网。通过共享VPC子网功能，可以简化网络配置，帮助您统一配置和运维多个账号下的资源，有助于提升资源的管控效率，降低运维成本。如何共享VPC子网，请参考《**共享VPC**》。
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**。
节点数量	-
购买时长	-

步骤3 单击“立即购买”，确认规格后提交。

----结束

2.2.5 购买作业节点调度次数/天增量包

作业节点调度次数/天增量包用于扩充作业节点调度次数/天配额。

不同版本的DataArts Studio实例，默认提供了不同的作业节点调度次数/天规格限制。该规格是以每天执行的数据开发作业、质量作业、对账作业、业务场景和元数据采集作业的调度次数之和计算的。其中数据开发作业的每天调度次数，是以节点（包含Dummy节点）为粒度进行度量的，另外补数据任务也会计入度量次数，但测试运行、失败重试不会计入。您可以在DataArts Studio实例卡片上通过“更多 > 配额使用量”查看该配额情况。

说明

DataArts Studio实例中数据开发作业节点运行的并行数上限，与当前实例的作业节点调度次数/天配额有关。

- 当“作业节点调度次数/天配额<=500”时，节点运行的并行数上限为10。
- 当“500<作业节点调度次数/天配额<=5000”时，节点运行的并行数上限为50。
- 当“5000<作业节点调度次数/天配额<=20000”时，节点运行的并行数上限为100。
- 当“20000<作业节点调度次数/天配额<=40000”时，节点运行的并行数上限为200。
- 当“40000<作业节点调度次数/天配额<=80000”时，节点运行的并行数上限为300。
- 当“作业节点调度次数/天配额>80000”时，节点运行的并行数上限为400。

当您的每日作业节点调度次数接近、达到该规格，或需要扩充数据开发作业节点运行的并行数上限时，建议购买作业节点调度次数/天增量包，以避免作业调度和运行并发数受限。

说明

当作业节点调度的已使用次数+运行中次数+本日将运行次数之和大于此版本规格，执行调度批处理作业或者启动实时作业时就会提示作业节点调度次数/天超过配额。

背景信息

- 不同版本的DataArts Studio实例的规格请参见**版本规格说明**。
- 您可以在DataArts Studio实例卡片上，通过“更多 > 查看增量包”，查看已购买的增量包。
- 您可以在DataArts Studio实例卡片上，通过“更多 > 配额使用量”，查看当前实例的配额使用量。也支持可以在空间管理处，通过对应空间的“操作 > 配额使用量”，查看每个工作空间的配额使用量。

设置配额使用量阈值告警

购买配额扩充增量包前，您可以设置配额使用量阈值告警。当触发告警时，表明您应当购买配额扩充增量包，否则随着业务量增长，您的业务可能会受到影响。

设置配额使用量阈值告警的操作方法如下所示：

步骤1 在DataArts Studio实例卡片上，单击选择“更多 > 告警阈值”。

图 2-7 告警阈值

步骤2 配置告警阈值，取值范围在0-100之间，设置为0表示不告警。当配额使用量超出设置的告警阈值时，会触发SMN短信或邮件告警。

步骤3 进入消息通知服务SMN控制台，单击进入“主题管理 > 主题”，找到主题名称“DGC_Topic_Manager_Schedule_Alarm_*项目名称_实例ID*”。

- 项目名称可以参考如下步骤进行获取：
  a. 注册并登录管理控制台。
  b. 在用户名的下拉列表中单击“我的凭证”。
  c. 在“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。
- 实例ID可参考如下步骤进行获取：

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

b. 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2-8所示，实例ID为**6b88…2688**，工作空间ID为**1dd3bc…d93f0**。

图 2-8 获取实例 ID 和工作空间 ID

步骤4 在对应主题的操作栏，选择“添加订阅”。然后协议选择“短信”或“邮件”，输入接收告警通知的手机号或邮箱即可。

图 2-9 添加订阅

----结束

购买作业节点调度次数/天增量包

1. 单击已开通实例卡片上的“购买增量包”。
2. 进入购买DataArts Studio增量包页面，按照如下配置：
   - 增量包类型：选择作业节点调度次数/天增量包。
   - 计费方式：当前仅支持套餐包。
   - 增量包规格：请根据您的业务情况选择合适的增量包规格。
   - 购买时长：表示此套餐包的有效时长。
   - 自动续费：勾选自动续费前的复选框，可实现自动按月或者按年续费。购买时长为按月购买时，自动续费周期为1个月；购买时长为按年购买时，自动续费周期为1年。
3. 单击“立即购买”，确认规格后提交订单。
4. 购买套餐包成功后，系统配额会在默认规格基础上，增加增量包的规格。

2.2.6 购买技术资产数量增量包

技术资产数量增量包用于扩充技术资产数量配额。

不同版本的DataArts Studio实例，默认提供了不同的技术资产数量规格限制。该规格是以数据目录中表和OBS文件的数量之和计算的。您可以在DataArts Studio实例卡片上通过“更多 > 配额使用量”查看该配额情况。

当您的技术资产数量接近或达到该规格时，建议购买技术资产数量增量包，以避免资产采集受限。

背景信息

- 不同版本的DataArts Studio实例的规格请参见**版本规格说明**。
- 您可以在DataArts Studio实例卡片上，通过“更多 > 查看增量包”，查看已购买的增量包。
- 您可以在DataArts Studio实例卡片上，通过“更多 > 配额使用量”，查看当前实例的配额使用量。也支持可以在空间管理处，通过对应空间的“操作 > 配额使用量”，查看每个工作空间的配额使用量。

设置配额使用量阈值告警

购买配额扩充增量包前，您可以设置配额使用量阈值告警。当触发告警时，表明您应当购买配额扩充增量包，否则随着业务量增长，您的业务可能会受到影响。

设置配额使用量阈值告警的操作方法如下所示：

步骤1 在DataArts Studio实例卡片上，单击选择“更多 > 告警阈值”。

图 2-10 告警阈值

步骤2 配置告警阈值，取值范围在0-100之间，设置为0表示不告警。当配额使用量超出设置的告警阈值时，会触发SMN短信或邮件告警。

步骤3 进入消息通知服务SMN控制台，单击进入“主题管理 > 主题”，找到主题名称“DGC_Topic_Manager_Schedule_Alarm_***项目名称_实例ID***”。

- 项目名称可以参考如下步骤进行获取：
  a. 注册并登录管理控制台。
  b. 在用户名的下拉列表中单击“我的凭证”。
  c. 在“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。
- 实例ID可参考如下步骤进行获取：
  a. 在DataArts Studio控制台首页，选择对应工作空间，并单击任一模块，如“管理中心”。

b. 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2-11所示，实例ID为**6b88…2688**，工作空间ID为**1dd3bc…d93f0**。

图 2-11 获取实例 ID 和工作空间 ID

步骤4 在对应主题的操作栏，选择“添加订阅”。然后协议选择“短信”或“邮件”，输入接收告警通知的手机号或邮箱即可。

图 2-12 添加订阅

----结束

购买技术资产数量增量包

1. 单击已开通实例卡片上的“购买增量包”。
2. 进入购买DataArts Studio增量包页面，按照如下配置：
   - 增量包类型：选择技术资产数量增量包。
   - 计费方式：当前仅支持套餐包。
   - 增量包规格：请根据您的业务情况选择合适的增量包规格。
   - 购买时长：表示此套餐包的有效时长。
   - 自动续费：勾选自动续费前的复选框，可实现自动按月或者按年续费。购买时长为按月购买时，自动续费周期为1个月；购买时长为按年购买时，自动续费周期为1年。
3. 单击“立即购买”，确认规格后提交订单。
4. 购买套餐包成功后，系统配额会在默认规格基础上，增加增量包的规格。

2.2.7 购买数据模型数量增量包

数据模型数量增量包用于扩充数据模型数量配额。

不同版本的DataArts Studio实例，默认提供了不同的数据模型数量规格限制。该规格是以数据架构中逻辑模型、物理模型、维度表、事实表和汇总表的数量之和计算的。您可以在DataArts Studio实例卡片上通过“更多 > 配额使用量”查看该配额情况。

当您的数据模型数量接近或达到该规格时，建议购买数据模型数量增量包，以避免数据架构设计受限。

背景信息

- 不同版本的DataArts Studio实例的规格请参见**版本规格说明**。
- 您可以在DataArts Studio实例卡片上，通过“更多 > 查看增量包”，查看已购买的增量包。
- 您可以在DataArts Studio实例卡片上，通过“更多 > 配额使用量”，查看当前实例的配额使用量。也支持可以在空间管理处，通过对应空间的“操作 > 配额使用量”，查看每个工作空间的配额使用量。

设置配额使用量阈值告警

购买配额扩充增量包前，您可以设置配额使用量阈值告警。当触发告警时，表明您应当购买配额扩充增量包，否则随着业务量增长，您的业务可能会受到影响。

设置配额使用量阈值告警的操作方法如下所示：

步骤1 在DataArts Studio实例卡片上，单击选择“更多 > 告警阈值”。

图 2-13 告警阈值

步骤2 配置告警阈值，取值范围在0-100之间，设置为0表示不告警。当配额使用量超出设置的告警阈值时，会触发SMN短信或邮件告警。

步骤3 进入消息通知服务SMN控制台，单击进入“主题管理 > 主题”，找到主题名称“DGC_Topic_Manager_Schedule_Alarm_***项目名称_实例ID***”。

- 项目名称可以参考如下步骤进行获取：
  a. 注册并登录管理控制台。
  b. 在用户名的下拉列表中单击“我的凭证”。
  c. 在“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。
- 实例ID可参考如下步骤进行获取：
  a. 在DataArts Studio控制台首页，选择对应工作空间，并单击任一模块，如“管理中心”。

b. 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2-14所示，实例ID为**6b88…2688**，工作空间ID为**1dd3bc…d93f0**。

图 2-14 获取实例 ID 和工作空间 ID

步骤4 在对应主题的操作栏，选择“添加订阅”。然后协议选择“短信”或“邮件”，输入接收告警通知的手机号或邮箱即可。

图 2-15 添加订阅

----结束

购买数据模型数量增量包

1. 单击已开通实例卡片上的“购买增量包”。
2. 进入购买DataArts Studio增量包页面，按照如下配置：
   - 增量包类型：选择数据模型数量增量包。
   - 计费方式：当前仅支持套餐包。
   - 增量包规格：请根据您的业务情况选择合适的增量包规格。
   - 购买时长：表示此套餐包的有效时长。
   - 自动续费：勾选自动续费前的复选框，可实现自动按月或者按年续费。购买时长为按月购买时，自动续费周期为1个月；购买时长为按年购买时，自动续费周期为1年。
3. 单击“立即购买”，确认规格后提交订单。
4. 购买套餐包成功后，系统配额会在默认规格基础上，增加增量包的规格。

2.3 访问 DataArts Studio 实例控制台

前提条件

请参见**购买DataArts Studio实例**，确认已购买DataArts Studio实例。

操作步骤

步骤1 登录华为云控制台，在左上角的服务列表中选择“数据治理中心DataArts Studio”，进入DataArts Studio实例控制台。

- 如果当前区域下有多个DataArts Studio实例，则默认进入实例列表。请单击所需实例卡片上的“进入控制台”，进入DataArts Studio控制台首页。

图 2-16 实例列表

- 如果当前区域下仅有一个DataArts Studio实例，则默认进入DataArts Studio控制台首页。

图 2-17 控制台首页

----结束

2.4 创建并配置简单模式工作空间

2.4.1 创建简单模式工作空间

购买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”，并赋予该用户为管理员角色。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，其他账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、数据安全组件的操作权限。

约束限制

- DataArts Studio实例下允许创建的工作空间数量配额暂无限制，请您根据业务需求自行规划。
- 存储作业日志和脏数据依赖于OBS服务。

前提条件

请参见**购买DataArts Studio实例**，确认已购买DataArts Studio实例。

背景说明

- 购买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”，并赋予该用户为管理员角色。
- 在主账号创建的DataArts Studio实例中，该账号下的IAM用户如需创建工作空间，需要由主账号给IAM用户赋予**DAYU Administrator**或**Tenant Administrator**权限。在子用户创建的DataArts Studio实例中，主账号默认具有该DataArts Studio实例的所有执行权限。
- 具备DAYU User账号权限的用户，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

步骤1 参考**访问DataArts Studio实例控制台**，以**DAYU Administrator**或**Tenant Administrator**账号登录DataArts Studio管理控制台。

步骤2 在“空间管理”页签，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2-6**所示。

图 2-18 新建空间

表 2-6 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。
描述	空间的描述信息。
空间模式	选择新建工作空间的模式。 ● 简单模式：即传统的DataArts Studio工作空间模式，使用方便，但无法对数据开发流程和表权限进行强管控。 ● 企业模式：企业模式下DataArts Studio数据开发组件以及对应管理中心组件数据连接支持设置开发环境和生产环境，有效隔离开发者对生产环境业务的影响。企业模式的相关介绍请参见**企业模式概述**。
企业项目	DataArts Studio实例默认工作空间关联的企业项目。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**》。 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工作空间的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购买一个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。 **说明** 未开通企业项目时，则每个IAM项目只允许创建1个DataArts Studio实例。

参数名	说明
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，您可以参考如下步骤进行获取。 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“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。 **说明** 数据开发作业运行完以后，运行结果日志会存储在OBS桶路径下面，便于查看运行历史记录，文件格式为**xxxxx.log**的文件就是作业运行日志，xxxxx表示作业id。已经运行完的SQL结果，删除历史记录后，不会影响业务。
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。
标签	通过为资源添加标签，可以对资源进行自定义标记，实现资源的分类。 **说明** 如您的账号归属某个**组织**，且该组织已经设定DataArts Studio服务的相关**标签策略**，则需按照标签策略规则添加标签。标签如果不符合标签策略的规则，则可能会导致实例创建失败，请联系组织管理员了解标签策略详情。 当拥有多个工作空间时，您可以按使用者、维护者或用途等各类维度为各工作空间添加标签，然后在工作空间列表页面，可以通过标签搜索、识别不同类型的工作空间。 标签由标签键和标签值组成。在添加标签时，标签键和标签值可以选择在标签管理服务（简称TMS）中创建的预定义标签，也可以直接输入自定义的标签。然后单击输入框右侧的“添加”，即可成功添加一条标签。 **说明** 预定义标签需要预先在标签管理服务中创建好，然后才能进行选择。您可以通过单击“查看预定义标签”进入标签管理服务的“预定义标签”页面，然后单击“创建标签”来创建新的预定义标签，具体请参见《标签管理服务用户指南》中的“**创建预定义标签**”章节。 另外，工作空间最多支持添加20个标签，标签的键名不能重复，一个“标签键”只能添加一个对应“标签值”。

步骤3 配置完成后，单击“确定”完成工作空间的创建。

----结束

相关操作

- 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，可以将工作空间禁用，以后仍可以将其重新启用。

  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“更多 > 禁用”。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”。

  说明

  工作空间被禁用后，您将无法再访问工作空间，无法编辑工作空间或查看配额，工作空间内调度作业将停止运行。

  工作空间内购买的数据集成集群仍会继续计费。

- 启用工作空间：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“更多 > 启用”。在“启用”对话框中，如果确认启用，请单击“确定”。
- 编辑工作空间：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。在“空间信息”页面，您可以参考**表2-6**修改工作空间的相关参数，最后单击“确定”保存修改的配置。

  值得注意的是，与新建工作空间相比，编辑工作空间不支持修改标签（详见**添加/编辑标签**），支持添加工作空间成员（详见**添加工作空间成员和角色**）和配置工作空间配额（详见**设置工作空间配额**）。
- 添加/编辑标签：在“空间管理”页面，找到所需添加/编辑标签的工作空间，单击其所在行的“更多 > 标签”。在“标签”对话框中，单击“添加/编辑标签”为工作空间关联标签。
- 查看配额使用量：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“配额使用量”，此时显示“配额使用量”页面。在“配额使用量”页面，您可以查看当前空间内，各配额规格的使用量。
- 置顶工作空间：在“空间管理”页面，找到所需置顶的工作空间，单击其所在行的“更多 > 置顶”，完成置顶。
- 删除工作空间：在“空间管理”页面，找到所需删除的工作空间，单击其所在行的“更多 > 删除”。在“删除工作空间”对话框中，如果确认删除，请单击“确定”。

说明

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

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

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

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

图 2-19 删除失败提示

2.4.2 设置工作空间配额

使用DataArts Studio前，您需要为当前工作空间设置工作空间配额（当前仅支持数据服务专享版API配额）。如果当前工作空间的“已使用配额”超出“已分配配额”，或者“总使用配额”超出“总分配配额”，则会导致相应业务使用受限，例如无法再新建数据服务专享版API。

- 已使用配额：表示当前工作空间下已使用的配额，由系统自动统计。
- 已分配配额：表示分配给当前工作空间可使用的配额，需要由管理员为每个工作空间分配。
- 总使用配额：表示当前实例下已使用的总配额，由系统自动统计。
- 总分配配额：表示当前实例下分配给所有工作空间可使用的总配额，由系统自动统计。
- 总配额：表示当前实例所拥有的最大总配额，固定值不可修改。

前提条件

修改工作空间的用户账号，需要满足如下任一条件：

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

操作步骤

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

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

图 2-20 空间信息

步骤3 在“空间信息”中，单击“数据服务专享版API配额”中对应配额的“设置”按钮，对已分配配额进行配置。配置完成后单击“保存”，保存当前配置。

已分配配额表示分配给当前工作空间下可使用的配额。注意，已分配配额不能小于已使用配额，不能大于未分配配额（即总配额-总分配配额）。

说明

数据服务专享版在每个DataArts Studio实例下具有创建10个专享版API免费试用额度，超出试用配额后会产生数据服务专享版API的费用，所创建的超出试用配额API按每天每个进行收费。

图 2-21 设置已分配配额

步骤4 已分配配额设置完成后，单击“空间信息”中的“确定”，完成配置。

----结束

2.4.3 （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和DLI脏数据存储路径，支持基于工作区全局配置OBS桶。

约束限制

- 该功能依赖于OBS服务。
- OBS路径仅支持OBS桶，不支持并行文件系统。

前提条件

修改工作空间的用户账号，需要满足如下任一条件：

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

修改方法

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

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

图 2-22 空间信息

步骤3 在“空间信息”中，单击“作业日志OBS路径”和“DLI脏数据OBS路径”后的“请选择”按钮，选择日志和DLI脏数据存储路径，可选择某个具体的目录。

图 2-23 修改日志和 DLI 脏数据存储路径

步骤4 修改完成后，单击“确定”，即完成作业日志和DLI脏数据存储路径的修改。

----结束

2.5 （可选）升级企业模式工作空间

2.5.1 企业模式简介

为方便不同安全管控要求的用户生产数据，DataArts Studio为您提供简单模式和企业模式两种工作空间模式。本文从简单模式工作空间与企业模式工作空间物理形态、对开发行为的影响等多个维度为您介绍两种模式工作空间的区别。

须知

目前，仅管理中心和数据开发组件支持企业模式。

简单模式下为实现开发和生产环境隔离，需要创建两个工作空间，一个是开发环境工作空间，一个是生产环境工作空间，然后将开发工作空间导出的脚本或作业，导入到生产工作空间。在这种方式下，无法简单便捷地完成生产和开发环境同步，缺少审批管控环节。针对以上问题，可以通过企业空间模式，在一个工作空间实现开发与生产环境隔离，通过一键发布和审批流程，快速且高效的发布任务，极大提高了工作效率。

建议您将简单模式工作空间升级为企业模式工作空间，以便获得更好的开发流程管控。详情请参见**创建企业模式工作空间**。

背景信息

本文内容由以下几部分构成，从不同角度分别为您解决企业模式不同的问题。

表 2-7 了解企业模式

分类	说明
简单模式与企业模式介绍	不同工作空间模式的介绍。
不同模式工作空间对生产任务开发与运维的影响	DataArts Studio建立于对应工作空间物理属性之上的任务开发与运维机制介绍。
不同模式工作空间的优劣势对比	不同工作空间模式的优劣势对比。
企业模式对使用流程的影响	介绍企业模式工作空间下的流程管控。

分类	说明
不同工作空间模式下，DataArts Studio模块对应操作	简单模式仅有生产环境，企业模式有开发环境和生产环境，此部分为您介绍各个环境与DataArts Studio模块的对应关系。

注意事项

- 不同工作空间模式对于数据湖引擎存在一定的要求，企业模式工作空间需要分别为开发环境和生产环境进行数据湖引擎配置，才可以实现开发生产环境隔离。配置开发生产环境隔离包含以下三种方式：

  图 2-24 配置开发生产环境隔离

  - 配置两套数据湖服务，进行开发与生产环境隔离。

    对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS等），DataArts Studio通过管理中心的创建数据连接区分开发环境和生产环境的数据湖服务，在开发和生产流程中自动切换对应的数据湖。因此您需要准备两套数据湖服务，且两套数据湖服务的版本、规格、组件、区域、VPC、子网以及相关配置等信息，均应保持一致，详细操作请参见**创建DataArts Studio数据连接**。

    创建数据连接时，通过不同的集群来进行开发与生产环境的隔离，如**图2-25**所示。

    图 2-25 创建数据连接时选择不同集群

- 配置DLI环境隔离。

  配置企业模式环境隔离，包含DLI队列配置和DB配置。

  对于Serverless服务（例如DLI），DataArts Studio通过管理中心的环境隔离来配置生产环境和开发环境数据湖服务的对应关系，在开发和生产流程中自动切换对应的数据湖。因此您需要在Serverless数据湖服务中准备两套队列、两套数据库资源，建议通过名称后缀进行区分，详细操作请参见**配置DataArts Studio企业模式环境隔离**。

- 配置DB，在同一个数据湖服务下配置两套数据库，进行开发与生产环境隔离。

  对于DWS、MRS Hive和MRS Spark这三种数据源，如果在创建数据连接时选择同一个集群，如**图2-26**所示，则需要配置数据源资源映射的DB数据库映射关系进行开发生产环境隔离，如**图2-27**所示。详细操作请参见**DB配置**。

图 2-26 创建数据连接时选择同一个集群

图 2-27 DB 配置

- 企业模式工作空间下，开发环境的数据开发作业默认不进行调度，仅发布至生产环境后可进行调度。

简单模式与企业模式介绍

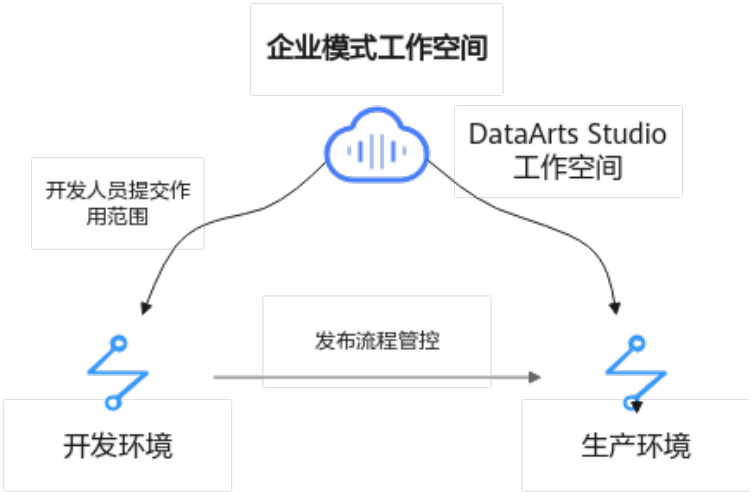
简单模式：传统的DataArts Studio工作空间模式。简单模式工作空间下，DataArts Studio数据开发组件以及对应管理中心组件无法设置开发环境和生产环境，只能进行简单的数据开发，无法对数据开发流程和表权限进行强管控。一个数据湖作为DataArts Studio的生产环境。

图 2-28 简单模式工作空间

企业模式：为解决简单模式存在的风险，DataArts Studio工作空间新增支持企业模式。企业模式下，DataArts Studio数据开发组件以及对应管理中心组件的数据连接支持设置开发环境和生产环境，有效隔离开发者对生产环境业务的影响。需要两个数据湖中，其中一个数据湖作为DataArts Studio开发环境，另一个作为DataArts Studio生产环境。其中：

- 开发环境只针对开发人员开放，只用于脚本或作业开发，开发完后发布到生产环境中。
- 生产环境内不能做任何修改，只对最终用户开放，任何修改必须回退到开发环境中重新修改发布。

图 2-29 企业模式工作空间



说明

- 您可选择创建任意模式工作空间体验DataArts Studio，使用企业模式工作空间实现DataArts Studio开发环境与生产环境代码隔离、不同环境计算资源隔离、权限隔离、任务发布流程管控等需求。
- 若您已在使用简单模式工作空间，并且希望保留当前简单模式工作空间的代码时，可选择工作空间模式升级，详情请参见**创建企业模式工作空间**。

不同模式工作空间对生产任务开发与运维的影响

表 2-8 不同模式工作空间对生产任务开发与运维的影响

对比	简单模式	企业模式（推荐）
生产任务开发流程管控差异	任务提交后，无需发布，即可进入调度系统周期性执行，产出结果数据。 （提交-->生产）	• 任务需要先提交至开发环境，再执行发布操作，将任务发布至生产环境，才可以自动调度运行任务。 （提交-->发布-->生产） • 生产环境内不能做任何修改，只对最终用户开放，任何修改必须回退到开发环境中重新修改发布。
生产任务运维权限管控差异	开发人员可直接编辑生产任务的脚本和作业。	开发人员只能在数据开发界面编辑代码并且提交，但是不能将代码直接发布到生产环境，发布生产的操作需要有运维权限（部署者、管理员、运维者这几类角色拥有此权限）。 • 所有脚本和作业仅支持在开发环境编辑，无法修改生产环境的代码。 • 您可基于企业模式工作空间特性，以及DataArts Studio角色权限体系来规划与管控DataArts Studio上任务开发与运维流程。详情请参见**企业模式业务流程**。
生产数据权限管控差异	开发人员可直接使用生产数据进行测试，无法保障生产数据安全。	开发人员在开发环境可使用测试数据进行测试，生产环境数据可读。

不同模式工作空间的优劣势对比

表 2-9 不同模式工作空间的优劣势对比

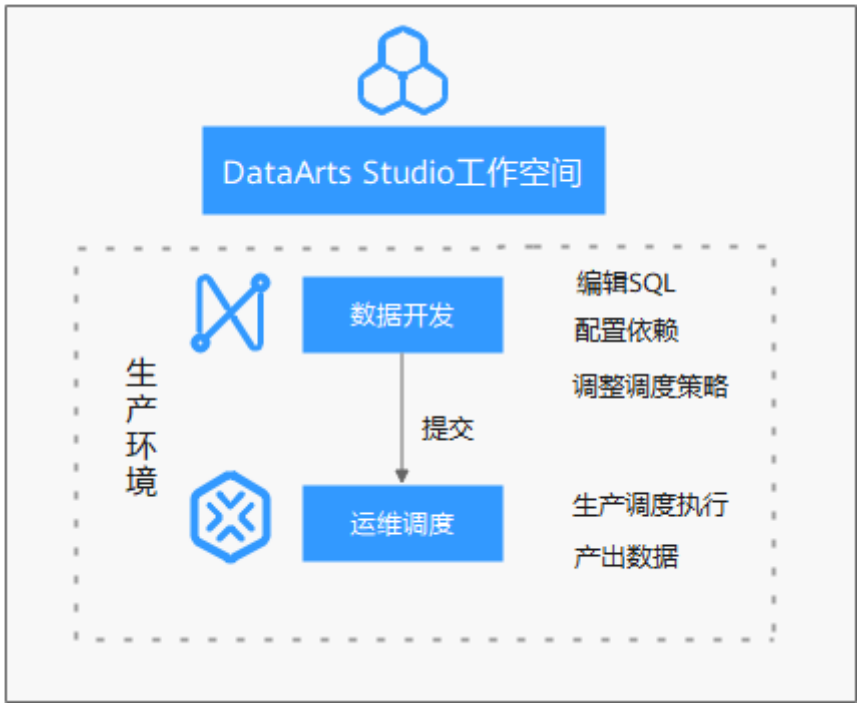
对比	简单模式	企业模式
优势	简单、方便、易用。 ● 仅需要授权数据开发人员“开发者”角色即可完成所有数据开发工作。 ● 提交脚本或作业后，您无需发布，脚本或作业即可进入调度系统周期性执行，产出结果数据。	安全、规范。 ● 具备安全、规范的代码发布管控流程（包含代码评审、代码DIFF查看等功能），保障生产环境稳定性，避免不必要的因代码逻辑引起的脏数据蔓延或任务报错等非预期情况。 ● 数据访问得到有效管控，数据安全得以保障。 ● 所有脚本或作业仅支持在开发环境编辑，开发者无法修改生产环境的脚本或作业。 ● 开发环境和生产环境的数据隔离，开发者无法影响生产环境的数据。 ● 开发环境下，脚本、作业以当前开发者的身份执行；生产环境下，脚本、作业则使用空间级的公共IAM账号或公共委托执行。 ● 如果需要对生产环境进行变更，必须在开发环境通过开发者的发布操作才能将变更提交到生产环境，需要管理者或部署者审批通过，才能发布成功。

对比	简单模式	企业模式
劣势	存在不稳定、不安全的风险。 • 无法设置开发环境和生产环境隔离，只能进行简单的数据开发。 • 无法对生产表权限进行控制。 **说明** 开发调测阶段，开发者可直接访问生产数据湖的数据，随意对表进行增加、删除和修改等操作，存在数据安全风险。 • 无法对数据开发流程进行管控。 **说明** 开发者可以不经过任何人审批，随时新增、修改脚本或作业并提交至调度系统，给业务带来不稳定因素。	流程相对复杂，一般情况下无法一人完成所有数据开发、生产流程。

企业模式对使用流程的影响

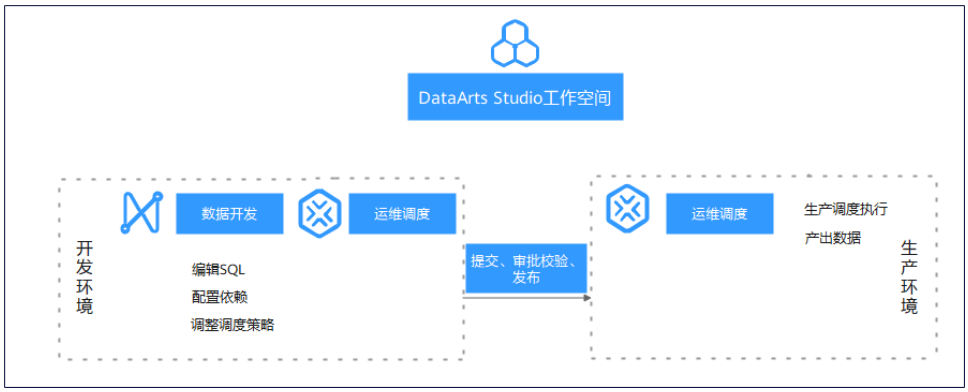
- 简单模式工作空间下，DataArts Studio数据开发组件以及对应管理中心组件无法设置开发环境和生产环境，只能进行简单的数据开发，无法对数据开发流程和表权限进行强管控。提交脚本或作业后，您无需发布，脚本或作业即可进入调度系统周期性执行，产出结果数据。

图 2-30 简单模式流程



- 企业模式下，DataArts Studio数据开发组件以及对应管理中心组件的数据连接支持设置开发环境和生产环境，有效隔离开发者对生产环境业务的影响。其中：开发环境只针对开发人员开放，只用于脚本或作业开发，开发完后发布到生产环境中。生产环境内不能做任何修改，只对最终用户开放，任何修改必须回退到开发环境中重新修改发布。

图 2-31 企业模式流程



不同工作空间模式下，DataArts Studio 模块对应操作

表 2-10 不同工作空间模式下对应模块的操作

DataArts Studio模块	简单模式	企业模式
管理中心	操作生产环境（数据连接、数据导入导出）	操作开发环境+生产环境（数据源资源映射配置、数据连接、数据导入导出）
数据开发	操作生产环境（实例、数据库）	操作开发环境+生产环境（实例、数据库）

2.5.2 创建企业模式工作空间

若您当前使用简单模式工作空间，但希望使用开发与生产环境隔离机制，您可以将简单模式工作空间升级为企业模式工作空间，如果您之前未使用过简单模式、无需继承业务数据，则可以直接新建新企业模式工作空间，本文为您介绍如何创建工作空间模式。

使用限制

只有DAYU Administrator、Tenant Administrator可升级企业模式或创建企业模式。

前提条件

创建工作空间模式前，您需要先了解以下内容：

- 已了解简单模式与企业模式工作空间的区别，包括不同工作空间的开发流程等差异，详情请参见**简单模式与企业模式介绍**。

- 已配置空间级的身份调度，包含公共委托和公共IAM账号，详情请参见**配置公共委托**和**配置公共IAM账号**。
- 已准备好两套相互隔离的数据湖引擎，用于隔离开发和生产环境。
  - 配置两套数据湖服务，进行开发与生产环境隔离。

    对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS等），DataArts Studio通过管理中心的创建数据连接区分开发环境和生产环境的数据湖服务，在开发和生产流程中自动切换对应的数据湖。因此您需要准备两套数据湖服务，且两套数据湖服务的版本、规格、组件、区域、VPC、子网以及相关配置等信息，均应保持一致，详细操作请参见**创建DataArts Studio数据连接**。

    创建数据连接时，通过不同的集群来进行开发与生产环境的隔离，如**图2-32**所示。

    图 2-32 创建数据连接时选择不同集群

  - 配置DLI环境隔离。

    配置企业模式环境隔离，包含DLI队列配置和DB配置。

    对于Serverless服务（例如DLI），DataArts Studio通过管理中心的环境隔离来配置生产环境和开发环境数据湖服务的对应关系，在开发和生产流程中自动切换对应的数据湖。因此您需要在Serverless数据湖服务中准备两套队列、两套数据库资源，建议通过名称后缀进行区分，详细操作请参见**配置DataArts Studio企业模式环境隔离**。
  - 配置DB，在同一个数据湖服务下配置两套数据库，进行开发与生产环境隔离。

    对于DWS、MRS Hive和MRS Spark这三种数据源，如果在创建数据连接时选择同一个集群，如**图2-33**所示，则需要配置数据源资源映射的DB数据库映射关系进行开发生产环境隔离，如**图2-34**所示。详细操作请参见**DB配置**。

    图 2-33 创建数据连接时选择同一个集群

图 2-34 DB 配置

- 数据准备与同步
  - 数据湖服务创建完成后，您需要按照项目规划（例如数据开发需要操作的库表等），分别在开发和生产环境的数据湖服务中，新建数据库、数据库模式（仅DWS需要）、数据表等。
    - 对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS），使用两套集群资源，两套环境中的数据库、数据库模式（仅DWS需要）和数据表必须保持同名。
    - 对于Serverless服务（例如DLI），两套队列和两套数据库建议通过名称和后缀（开发环境添加后缀“_dev”，生产环境无后缀）进行关联与区分，数据表必须保持同名。
    - 对于DWS、MRS Hive和MRS Spark数据源，如果使用一套相同的集群资源，通过两个数据库（开发环境添加后缀“_dev”，生产环境无后缀）进行开发生产环境隔离，两套环境中数据库模式（仅DWS需要）和数据表必须保持同名。
  - 数据库、数据库模式（仅DWS需要）、数据表等新建完成后，如果涉及原始数据表等，您还需要将两套数据湖服务之间的数据进行同步：
    - 数据湖中已有数据：通过CDM或DRS等数据迁移服务，在数据湖间批量同步数据。
    - 数据源待迁移数据：通过对等的CDM或DRS等数据迁移服务作业进行同步，保证生产环境和开发环境的数据湖服务数据一致。

变更内容

工作空间模式升级后会在原简单模式工作空间对应的生产环境基础上，增加了与生产环境隔离的开发环境。

简单模式升级企业模式

对于简单模式的工作空间，DAYU Administrator、Tenant Administrator可以直接将其升级为企业模式。

- 升级前操作

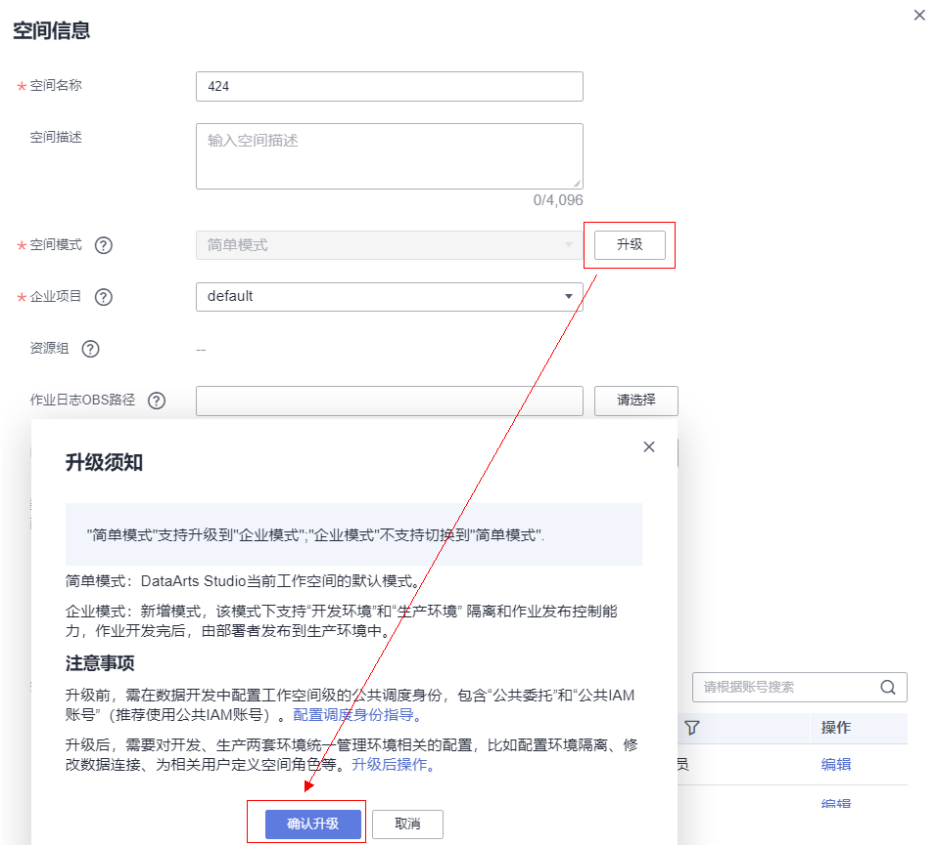
  如果您需要升级工作空间模式，需要在数据开发中配置空间级别的公共委托或公共IAM账号，避免升级失败。

  配置委托的操作详情可参见**配置调度身份**。

图 2-35 配置工作空间委托

- 升级操作
  a. 登录DataArts Studio控制台。
  b. 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。
  c. 在“空间管理”页面，找到需要升级模式的工作空间，单击其所在行的“编辑”，此时显示“空间信息”页面。
  d. 在“空间信息”页面，单击“空间模式”后的“升级”按钮，弹出确认界面后，单击“确认升级”您就可以将该工作空间升级为企业模式。

图 2-36 升级企业模式



- 升级后操作

  升级后需要管理员手工修改数据连接、配置环境隔离，并按照组织分工在工作空间处定义管理员、开发者、部署者、运维者等角色。

  a. 修改数据连接：请参考**创建DataArts Studio数据连接**。
  b. 配置环境隔离：请参考**配置DataArts Studio企业模式环境隔离**。
  c. 为其他用户定义工作空间角色：请参见**添加工作空间成员和角色**章节添加工作空间成员和角色。

新建企业模式工作空间

如果您之前未使用过简单模式、无需继承业务数据，则可以直接新建新企业模式工作空间。

- 创建工作空间

  a. 使用具有DAYU Administrator、Tenant Administrator权限的账号进入DataArts Studio控制台。
  b. 单击控制台的“空间管理”页签，进入工作空间页面。
  c. 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2-11**所示，配置完成后，单击“确定”完成工作空间的创建。

图 2-37 空间信息

表 2-11 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。
空间模式	选择工作空间为简单模式还是企业模式。新建企业模式工作空间时，此处需配置为企业模式。

参数名	说明
企业项目	DataArts Studio实例默认工作空间关联的企业项目。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》。 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工作空间的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购买一个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。 **说明** 未开通企业项目时，则每个IAM项目只允许创建1个DataArts Studio实例。
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。 **说明** 数据开发作业运行完以后，运行结果日志会存储在OBS桶路径下面，便于查看运行历史记录，文件格式为**xxxxx.log**的文件就是作业运行日志，xxxxx表示作业id。已经运行完的SQL结果，删除历史记录后，不会影响业务。
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。

参数名	说明
标签	通过为资源添加标签，可以对资源进行自定义标记，实现资源的分类。 **说明** 如您的账号归属某个**组织**，且该组织已经设定DataArts Studio服务的相关**标签策略**，则需按照标签策略规则添加标签。标签如果不符合标签策略的规则，则可能会导致实例创建失败，请联系组织管理员了解标签策略详情。 当拥有多个工作空间时，您可以按使用者、维护者或用途等各类维度为各工作空间添加标签，然后在工作空间列表页面，可以通过标签搜索、识别不同类型的工作空间。 标签由标签键和标签值组成。在添加标签时，标签键和标签值可以选择在标签管理服务（简称TMS）中创建的预定义标签，也可以直接输入自定义的标签。然后单击输入框右侧的“添加”，即可成功添加一条标签。 **说明** 预定义标签需要预先在标签管理服务中创建好，然后才能进行选择。您可以通过单击“查看预定义标签”进入标签管理服务的“预定义标签”页面，然后单击“创建标签”来创建新的预定义标签，具体请参见《标签管理服务用户指南》中的“**创建预定义标签**”章节。 另外，工作空间最多支持添加20个标签，标签的键名不能重复，一个“标签键”只能添加一个对应“标签值”。

- 创建后操作

  创建后需要管理员手工新建数据连接、配置环境隔离，并按照组织分工在工作空间处定义管理员、开发者、部署者、运维者等角色。

  a. 新建数据连接：请参考**创建DataArts Studio数据连接**。
  b. 配置环境隔离：请参考**配置DataArts Studio企业模式环境隔离**。
  c. 为其他用户定义工作空间角色：请参见**添加工作空间成员和角色**章节添加工作空间成员和角色。
  d. 另外，新建企业模式工作空间，还需要您在数据开发中配置空间级别的公共委托或公共IAM账号。配置委托的操作详情可参见**配置调度身份**。

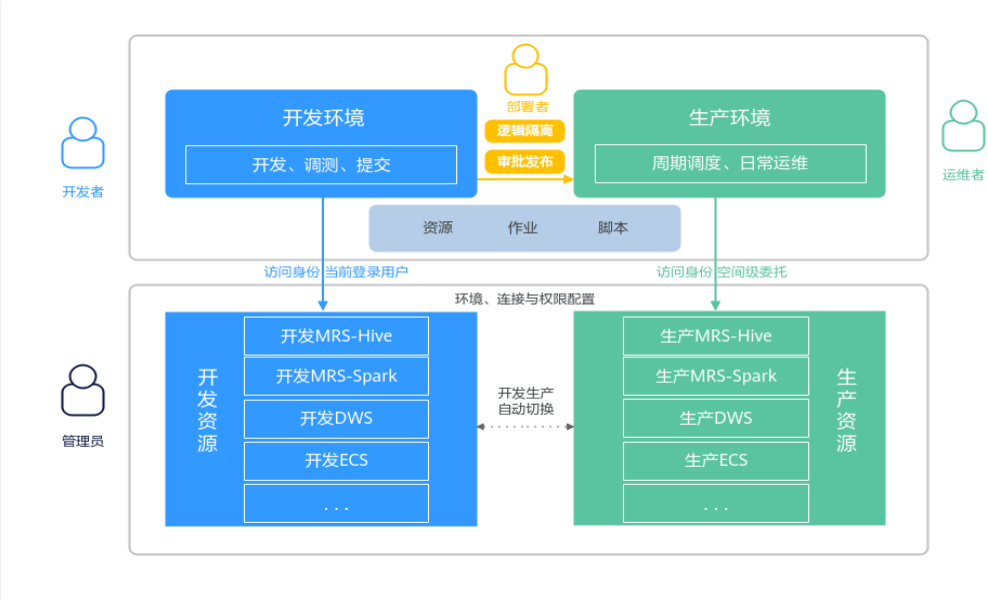
图 2-38 配置工作空间委托

2.5.3 企业模式角色操作

2.5.3.1 企业模式业务流程

当前DataArts Studio企业模式，主要涉及管理中心和数据开发组件，业务流程由管理员、开发者、部署者、运维者等角色共同完成。

图 2-39 企业模式架构



- 管理员：管理员需要进行准备数据湖、配置数据连接和环境隔离、数据的导入导出、配置项目用户权限等操作。
- 开发者：开发者需要在数据开发的开发环境，进行脚本、作业开发等操作，开发完成后进行测试运行、提交版本，最终提交发布任务。
- 部署者：部署者需要在数据开发的开发环境，查看待审批任务，并进行发布审批操作。
- 运维者：部署者需要在数据开发的生产环境，基于开发者发布的资源，进行作业监控、通知管理、备份等操作。
- 自定义角色：用户可以对需要的操作权限进行自定义设置，来满足实际业务的需要。
- 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。

表 2-12 企业模式内的权限

-	简单空间	企业空间
管理者	拥有生产环境管理中心的所有权限，包含连接配置、数据导入导出等。	● 增加了部署相关的新操作 ● 进行管理中心的连接配置、环境隔离配置，数据导入导出等 ● 进行数据开发配置，比如环境配置、调度身份配置、配置默认项等
开发者	拥有生产环境的作业及脚本开发的所有权限。	● 开发环境：所有操作 ● 生产环境：只读操作 ● 部署：增加了打包、查看发布项、查看发布项列表、查看发布包内容 ● 环境信息配置：只读操作
部署者	无	● 查看发布包 ● 查看发布项列表 ● 发布包：只有部署者和管理者可以操作 ● 撤销发布：只有部署者和管理者可以操作
运维者	拥有生产环境的作业及脚本实例的运行监控、运维调度等权限。	● 开发环境：只读操作 ● 生产环境：所有操作 ● 部署：查看发布包内容 ● 环境信息配置：只读操作
访客	仅只读	仅只读

2.5.3.2 管理员操作

管理员作为项目负责人或开发责任人，需要为企业模式的环境配置、人员角色等进行统一管控，相关操作如下表所示。

表 2-13 管理员操作

操作	说明
准备工作	包含数据湖准备以及数据准备与同步。 **数据湖准备：** 由于企业模式下需要区分开发环境和生产环境，因此您需要分别准备对应生产环境和开发环境的两套数据湖服务，用于隔离开发和生产环境： ● 对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS），DataArts Studio通过管理中心的创建数据连接区分开发环境和生产环境的数据湖服务，在开发和生产流程中自动切换对应的数据湖。因此您需要准备**两套数据湖服务（即两个集群）**，且两套数据湖服务的版本、规格、组件、区域、VPC、子网以及相关配置等信息，均应保持一致。 例如，当您的数据湖服务为MRS集群时，需要准备两套MRS集群，且版本、规格、组件、区域、VPC、子网等保持一致。如果某个MRS集群修改了某些配置，也需要同步到另一套MRS集群上。 ● 对于Serverless服务（例如DLI），DataArts Studio通过管理中心的环境隔离来配置生产环境和开发环境数据湖服务的对应关系，在开发和生产流程中自动切换对应的数据湖。因此您需要在Serverless数据湖服务中准备两套队列、数据库资源，建议通过名称后缀进行区分。 ● 特别的，对于DWS、MRS Hive和MRS Spark数据源，如果使用**一套相同的集群**，则需要配置数据源资源映射的DB数据库映射关系进行开发生产环境隔离。 **数据准备与同步：** ● 数据湖服务创建完成后，您需要按照项目规划（例如数据开发需要操作的库表等），分别在开发和生产环境的数据湖服务中，新建数据库、数据库模式（仅DWS需要）、数据表等。 – 对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS），使用两套集群资源，两套环境中的数据库、数据库模式（仅DWS需要）和数据表必须保持同名。 – 对于Serverless服务（例如DLI），两套队列和数据库建议通过名称和后缀（开发环境添加后缀“_dev”，生产环境无后缀）进行关联与区分，数据表必须保持同名。 – 对于DWS、MRS Hive和MRS Spark数据源，使用一套集群资源，通过两个数据库（开发环境添加后缀“_dev”，生产环境无后缀）进行开发生产环境隔离，两套环境中的数据库模式（仅DWS需要）和数据表必须保持同名。 ● 数据库、数据库模式（仅DWS需要）、数据表等新建完成后，如果涉及原始数据表等，您还需要将两套数据湖服务之间的数据进行同步： – 数据湖中已有数据：通过CDM或DRS等数据迁移服务，在数据湖间批量同步数据。 – 数据源待迁移数据：通过对等的CDM或DRS等数据迁移服务作业进行同步，保证生产环境和开发环境的数据湖服务数据一致。

操作	说明
创建企业模式数据连接	对于所有的数据湖引擎，都需要创建数据连接。 对于集群化的数据源，如果使用不同的集群，支持同时创建DataArts Studio与开发环境数据湖、DataArts Studio与生产环境数据湖之间的数据连接。 具体请参见**创建DataArts Studio数据连接**。
配置企业模式环境隔离	配置开发、生产环境的DLI队列和DB映射配置的环境隔离。 对于DWS、MRS Hive和MRS Spark这三种数据源，如果在创建数据连接时选择同一个集群资源，则需要在同一个数据湖服务下配置两套数据库，进行开发与生产环境隔离，具体请参见**DB配置**。 对于数据源为DLI时，可以通过企业模式环境隔离配置两套DLI队列和DB数据库进行生产与开发环境的隔离。具体请参见**配置DataArts Studio企业模式环境隔离**。
授权用户使用DataArts Studio	为协同使用DataArts Studio的项目成员创建具备“DAYU User”权限的IAM账号，并匹配对应的工作空间角色。 具体请参见**授权用户使用DataArts Studio**章节创建用户并授予权限。

2.5.3.3 开发者操作

开发者作为任务开发与处理的人员，需要开发脚本、开发作业等，相关操作如下表所示。

表 2-14 开发者操作

操作	说明
脚本开发	选择开发环境的数据湖引擎，在开发环境下的调测并发布数据开发脚本，发布到生产环境后系统会自动替换为对应生产环境引擎。 具体请参见**脚本开发**。
作业开发	选择开发环境的数据湖引擎，在开发环境下的调测并发布数据开发作业，发布到生产环境后系统会自动替换为对应生产环境引擎。 具体请参见**作业开发**。

2.5.3.4 部署者操作

- 部署者作为管理开发任务上线的人员，需要审批待发布任务，相关操作如下文所示。
- 部署者审批开发者提交的发布任务，审批通过后才能将修改后的作业同步到生产环境。

在企业模式中，开发者提交脚本或作业版本后，系统会对应产生发布任务。开发者确认发包后，需要部署者审批通过，才能将修改后的作业同步到生产环境。

前提条件

开发者已完成**脚本任务发布**或**作业任务发布**。

操作步骤

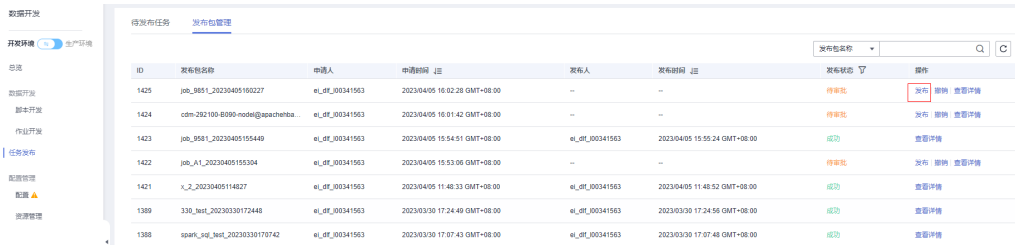
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤2 在数据开发主界面的左侧导航栏，选择开发环境，选择“数据开发 > 任务发布”。

步骤3 在任务发布界面，切换到“发布包管理”，您可以看到开发者提交的待审批任务。您可以通过“查看详情”操作，查看当前任务相比上一版本的修改点。

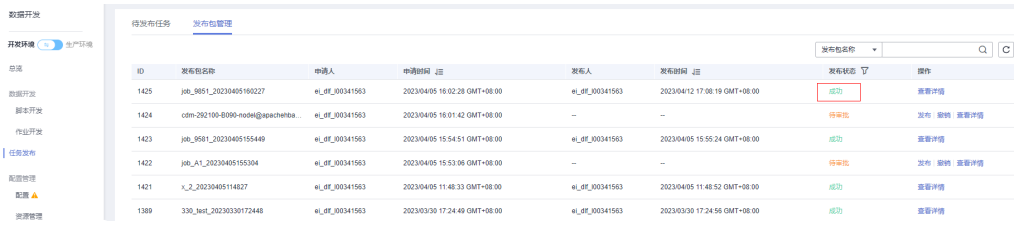
- 如有问题，可以通过“撤销”驳回发布任务。由开发者修改后重新提交发布任务，再度进行审批。
- 确认发布任务没有问题后，请通过“发布”操作，将任务审批通过。

图 2-40 审批发布



步骤4 成功发布之后，您可以查看任务的发布状态。任务发布成功后，开发者的修改将同步到生产环境。

图 2-41 查看任务状态



----结束

2.5.3.5 运维者操作

运维者作为运维管理的负责人，需要对生产环境的作业、实例、通知、备份等进行统一管控，相关操作如下表所示。

表 2-15 运维者操作

操作	说明
作业监控	包含对批作业、实时作业的监控。 具体请参见**作业监控**。
实例监控	对作业实例进行监控，作业每次运行，都会对应产生一次作业实例记录。 具体请参见**实例监控**。
补数据监控	对补数据作业运行情况进行监控。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。 具体请参见**补数据监控**。
通知管理	配置在作业运行异常或成功时能接收到通知。 具体请参见**通知管理**。
备份管理	支持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。 具体请参见**备份管理**。

2.6 管理 DataArts Studio 资源

资源管理提供对DataArts Studio资源的统一管理。

离线资源管理

离线资源管理为您提供查看当前DataArts Studio实例下所有CDM集群的功能，并支持为CDM集群关联不同的工作空间。

说明

只有当CDM集群在关联了工作空间后，才能在所关联的工作空间中使用该CDM集群。

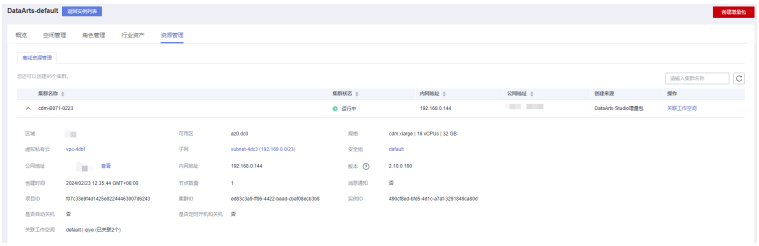
步骤1 参考**访问DataArts Studio实例控制台**，以**DAYU Administrator**或**Tenant Administrator**账号登录DataArts Studio管理控制台。

步骤2 单击控制台的“资源管理”页签，进入资源管理页面。

步骤3 在默认的离线资源管理页签，您可以查看当前实例下的所有CDM集群及其状态、内网地址、公网地址等信息。

步骤4 单击CDM集群列表中集群名称列的按钮开 ，可查看该CDM集群的详情信息，例如可用区、虚拟私有云、子网和安全组等网络相关信息，以及规格、集群ID、关联的工作空间等信息。

图 2-42 查看集群详情信息



步骤5 您可以单击CDM集群列表中操作列的“关联工作空间”，在弹窗中勾选或去勾选该CDM集群关联的工作空间，单击确认即可完成CDM集群与工作空间的关联。

注意，只有当CDM集群在关联了工作空间后，才能在所关联的工作空间中使用该CDM集群。

图 2-43 关联工作空间

----结束

2.6.1 实时集成资源组关联工作空间

进行实时数据集成任务配置前，您需要将数据集成资源组与将要使用的DataArts Studio工作空间进行关联，以确保在配置实时集成作业时可以选到指定的计算资源组。

前提条件

已购买资源组，详情请参见**购买数据集成资源组**。

操作步骤

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择实例，单击“进入控制台”。

步骤3 单击“资源管理”，进入资源管理页面。

步骤4 在“实时资源管理”页签中，找到指定的数据集成资源组，单击右侧操作栏中的“关联工作空间”。

图 2-44 关联工作空间入口

步骤5 在弹出框中，搜索需要使用的DataArts Studio工作空间，单击“关联”按钮，即可在对应工作空间中选到该数据集成资源组。

说明

一个数据集成资源组可以关联到多个DataArts Studio工作空间。

图 2-45 关联工作空间

----结束

3 授权用户使用 DataArts Studio

3.1 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，在您的华为账号中，给企业中不同职能部门的员工创建IAM用户，让员工拥有唯一安全凭证，并使用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、高效的其他华为账号或者云服务，这些账号或者云服务可以根据权限进行代运维。

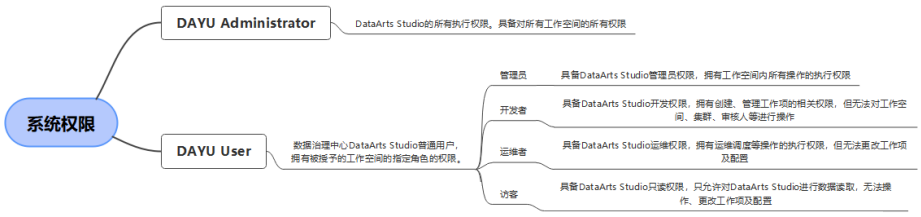
如果华为账号已经能满足您的要求，不需要创建独立的IAM用户，您可以跳过本章节，不影响您使用DataArts Studio服务的其它功能。

本章节为您介绍用户授权的方法，操作流程如**操作步骤**所示。

背景信息

- 给用户组授权之前，请您了解DataArts Studio的权限体系，并结合实际需求选择对应的权限。关于DataArts Studio权限的详细描述，请参见**DataArts Studio权限管理**。

  图 3-1 权限体系

- 若您需要对除DataArts Studio之外的其它服务授权，IAM支持服务的所有权限请参见**系统权限**。

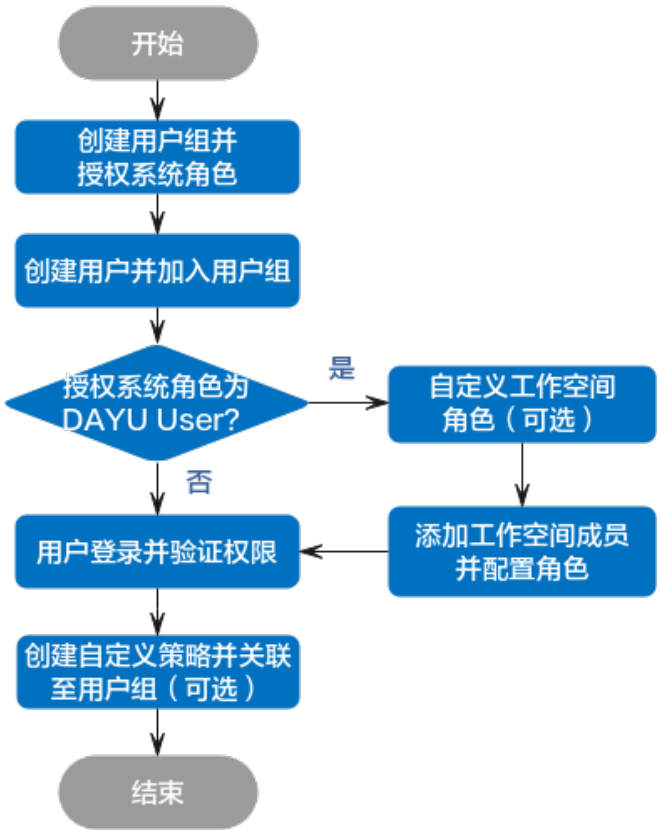
约束与限制

- DAYU User系统角色为用户提供了实例及工作空间和依赖服务的相关权限，具体工作空间内的业务操作权限由工作空间角色提供。

- IAM提供了以下两种授权机制。注意，DataArts Studio仅支持其中的IAM角色方式，不支持IAM策略。
  - **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。传统的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。
  - **IAM策略**：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。

操作步骤

图 3-2 授权流程



步骤1 创建用户组并授权系统角色。

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

创建用户组并授权的具体操作，请参见**创建用户组并授权**。

说明

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

步骤2 创建用户并加入用户组。

在IAM控制台创建用户，并将其加入**步骤1**中创建的用户组。

创建用户并加入用户组的具体操作，请参见**创建用户并加入用户组**。

说明

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

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

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

角色的权限说明请参见**权限列表**章节。

步骤4 用户登录并验证权限

新创建的用户登录控制台，切换至授权区域，验证权限，例如：

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

----结束

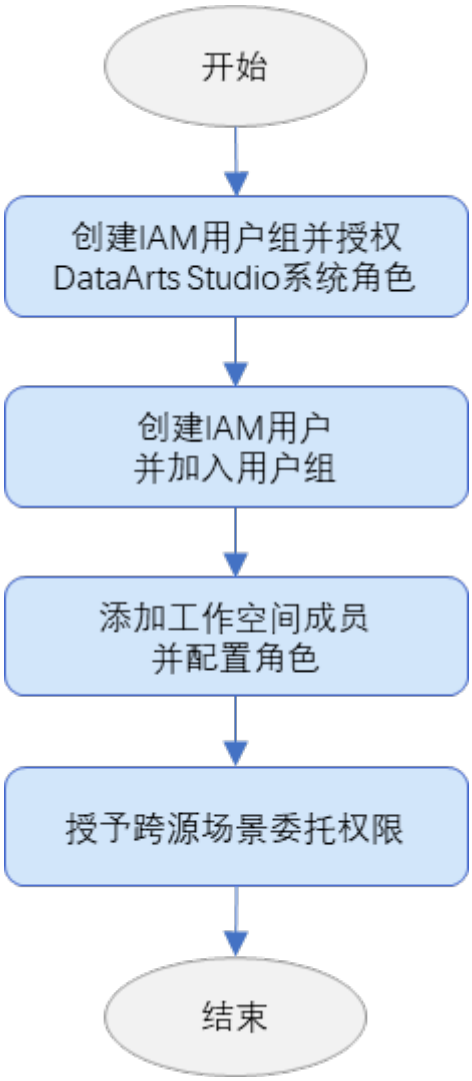
后续操作

依赖服务权限最小化：DAYU User系统角色预置的依赖服务相关权限过大，可能导致相关安全风险。您可以参考**如何最小化授权IAM用户使用DataArts Studio**，手动调整过大的预置依赖服务权限，使依赖服务权限最小化。

3.2 授权使用实时数据集成

DataArts Studio提供实时数据同步能力，如果您期望使用该功能，本章节为您介绍相关用户授权的方法，操作流程如下。

图 3-3 实时数据集成授权流程



约束与限制

- 已购买并配置DataArts Studio实例，并创建了可供使用的工作空间。
- 已创建IAM用户并授权使用DataArts Studio权限，详情请参见**创建IAM用户并授予DataArts Studio权限**。

操作步骤

步骤1 当前IAM用户需要加入拥有DataArts Studio的系统角色（如“DAYU Administrator”或“DAYU User”）的用户组，详情请参见**创建IAM用户并授予DataArts Studio权限**。

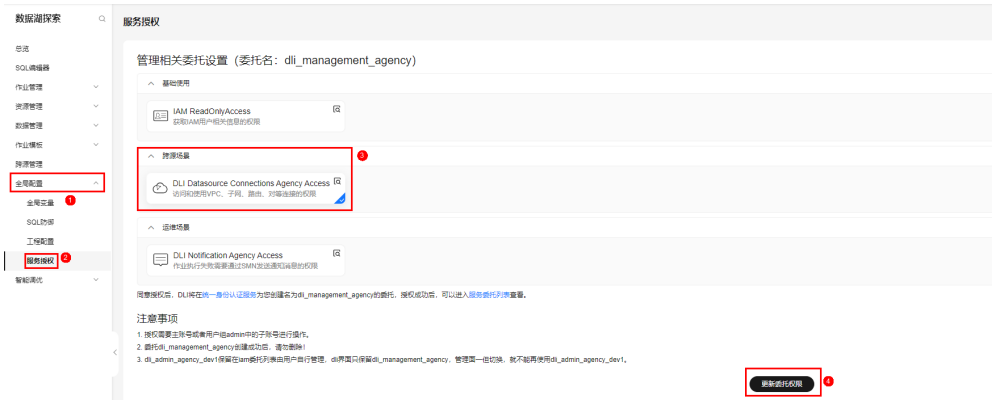
步骤2 当前IAM用户需要配置DataArts Studio工作空间的角色成员，且该空间角色成员需要拥有数据开发、管理中心服务的类管理员或开发者权限，用于查看、创建与操作数据连接、数据集成任务，角色的权限说明请参见**权限列表**。

步骤3 配置DLI云服务跨源场景委托权限。

实时数据集成与数据湖探索（DLI）云服务底层使用统一纳管集群资源，首次使用时需要通过DLI云服务创建跨源场景委托，用于底层计算资源访问和使用本租户VPC、子网、路由、对等连接等权限，详细请参见**配置DLI云服务委托权限**。

1. 搜索并进入DLI云服务控制台。
2. 在DLI控制台左侧导航栏中单击“全局配置 > 服务授权”。
3. 在委托设置页面，“管理相关委托设置”中勾选“跨源场景”权限，并单击“更新委托权限”。
4. 查看并了解更新委托的提示信息，单击“确定”。完成DLI委托权限的更新。

图 3-4 配置 DLI 云服务跨源场景委托权限

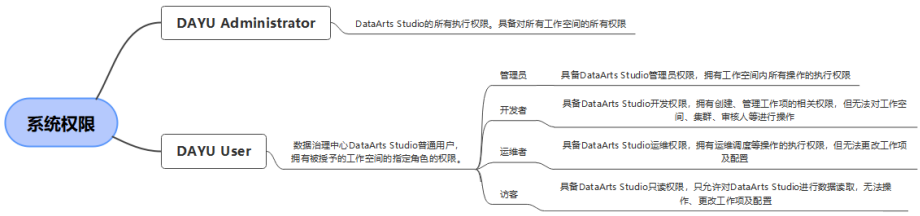


----结束

3.3 添加工作空间成员和角色

对于**DAYU User**账号权限的IAM用户而言，DataArts Studio工作空间角色决定了其在工作空间内的权限。如果您需要与**DAYU User**账号权限的IAM用户协同使用DataArts Studio实例，请参考**创建IAM用户并授予DataArts Studio权限**的操作准备必要的IAM用户，然后参考本章节将该用户添加为工作空间成员并配置工作空间角色。

图 3-5 权限体系



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、部署者、运维者和访客这几种预置角色可被分配。各角色权限的详细说明请参见**权限列表**章节。

- 管理员：工作空间管理员，拥有工作空间内所有的业务操作权限。建议将项目负责人、开发责任人、运维管理员设置为管理员角色。
- 开发者：开发者拥有工作空间内创建、管理工作项的业务操作权限。建议将任务开发、任务处理的用户设置为开发者。

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

背景信息

如果创建的IAM用户被授权**DAYU User**权限，则还需要添加工作空间成员和角色，否则会导致IAM用户无法查看已有的DataArts Studio工作空间。

约束与限制

由于鉴权缓存机制的限制，工作空间成员的角色发生变更后，不会直接生效。需要在工作空间成员暂停访问DataArts Studio控制台并等待6分钟后，才能使角色变更生效。

前提条件

修改工作空间的用户账号，需要满足如下任一条件：

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

添加成员和角色

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

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

图 3-6 空间信息

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

图 3-7 添加成员

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

----结束

相关操作

- 移除空间成员：通过空间编辑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”。在“移除”对话框中，如果确认要移除成员，请单击“确定”。

  说明

  工作空间的所有者不能被删除。

图 3-8 移除成员

4 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。

4.1 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库作为数据底座，由数据底座提供存储和计算的能力，DataArts Studio基于数据底座进行一站式数据开发、治理和服务。

DataArts Studio 支持的数据源

在本章节中，DataArts Studio支持的数据源是指除数据集成之外其他各组件支持的数据源情况，各组件支持程度各有不同，详情请参见**表4-1**。

另外，除数据集成之外其他各组件所使用的数据连接，均来自于管理中心已勾选对应组件的数据连接（只有勾选适用组件后，在相应组件内才能使用对应的连接）。因此如需对接这些数据源，请前往“DataArts Studio控制台 > 管理中心”创建数据连接。

说明

数据集成组件中集成作业支持的数据源与其他组件数据源情况维度不同，因此在数据集成章节内呈现，不在本章节内进行说明。当前集成作业包含CDM作业、离线作业和实时作业三种场景，支持的数据源情况如下：

- 数据集成（CDM作业）的数据连接在CDM集群中创建，CDM集成作业支持的数据源与CDM集群版本相关，详情请参见**数据集成（CDM作业）支持的数据源**。
- 数据集成（离线作业）的数据连接来自于管理中心中适用组件已勾选“数据集成”的数据连接，离线集成作业支持的数据源详情请参见**离线集成作业支持的数据源**。
- 数据集成（实时作业）的数据连接来自于管理中心中适用组件已勾选“数据集成”的数据连接，实时集成作业支持的数据源详情请参见**实时集成作业支持的数据源**。

表 4-1 DataArts Studio 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[2]	数据质量[3]	数据服务	数据安全
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 Spark）[1]	√	√	√	×	√	×	×
MapReduce服务（MRS ClickHouse）	√	√	√	√	×	√	×
MapReduce服务（MRS Hetu）	√	×	√	×	√	√	√
MapReduce服务（MRS Impala）	√	×	√	×	×	×	×
MapReduce服务（MRS Ranger）	√	×	×	×	×	×	√
MapReduce服务（MRS Presto）	√	×	√	×	×	×	×
MapReduce服务（MRS Doris）	√	√	√	√	×	√	×
云数据库 RDS（云数据库MySQL）	√	√	√	√	√	√	×
云数据库 RDS（云数据库PostgreSQL）	√	√	√	√	√	×	×
云数据库 RDS（云数据库SQL Server）	√	×	×	√	×	×	×
MySQL	√	√	×	×	√	√	×
Oracle	√	√	×	√	√	×	×
实时数据接入 DIS	√	×	√	√	×	×	×

数据源类型	管理中心	数据架构	数据开发	数据目录[2]	数据质量[3]	数据服务	数据安全
主机连接	√	×	√	×	×	×	×

说明

当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。

注释：

[1] MapReduce服务（MRS Spark）：数据架构和数据质量组件通过MRS Spark连接支持MRS Hudi数据源。MRS Hudi作为一种数据格式，元数据存放在Hive中，操作通过Spark进行。因此数据目录通过MRS Hive采集Hudi元数据，数据架构和数据质量通过MRS Spark对Hudi数据源进行治理（数据质量业务指标监控暂不支持Hudi数据源）。

[2] 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
5. MRS Hudi组件（MRS Hudi作为一种数据格式，元数据存放在Hive中，操作通过Spark进行。在Hudi表开启“**同步hive表配置**”后，可通过采集MRS Hive元数据的方式采集Hudi表的元数据）

[3] 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 4-2 数据源简介

数据源类型	简介
数据仓库服务（DWS）	华为云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）	华为云DLI是完全兼容Apache Spark和Apache Flink生态，实现批流一体的Serverless大数据计算分析服务。DLI支持多模引擎，企业仅需使用SQL或程序就可轻松完成异构数据源的批处理、流处理、内存计算、机器学习等，挖掘和探索数据价值。

数据源类型	简介
MapReduce服务（MRS HBase）	HBase是一个开源的、面向列（Column-Oriented）、适合存储海量非结构化数据或半结构化数据的、具备高可靠性、高性能、可灵活扩展伸缩的、支持实时数据读写的分布式存储系统。 使用MRS HBase可实现海量数据存储，并实现毫秒级数据查询。选择MRS HBase可以实现物流数据毫秒级实时入库更新，并支持百万级时序数据查询分析。
MapReduce服务（MRS Hive）	Hive是一种可以存储、查询和分析存储在Hadoop中的大规模数据的机制。Hive定义了简单的类SQL查询语言，称为HiveQL，它允许熟悉SQL的用户查询数据。 使用MRS Hive可实现TB/PB级的数据分析，快速将线下Hadoop大数据平台（CDH、HDP等）迁移上云，业务迁移 “0”中断，业务代码 “0”改动。
MapReduce服务（MRS Kafka）	华为云MapReduce服务可提供专属MRS Kafka集群。Kafka是一个分布式的、分区的、多副本的消息发布-订阅系统，它提供了类似于JMS的特性，但在设计上完全不同，它具有消息持久化、高吞吐、分布式、多客户端支持、实时等特性，适用于离线和在线的消息消费，如常规的消息收集、网站活性跟踪、聚合统计系统运营数据（监控数据）、日志收集等大量数据的互联网服务的数据收集场景。
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。
MapReduce服务（MRS Clickhouse）	ClickHouse是一款开源的面向联机分析处理的列式数据库，其独立于Hadoop大数据体系，最核心的特点是极致压缩率和极速查询性能。同时，ClickHouse支持SQL查询，且查询性能好，特别是基于大宽表的聚合分析查询性能非常优异，比其他分析型数据库速度快一个数量级。 当前ClickHouse被广泛的应用于互联网广告、App和Web流量、电信、金融、物联网等众多领域，非常适用于商业智能化应用场景。
MapReduce服务（MRS Impala）	Impala直接对存储在HDFS、HBase或对象存储服务（OBS）中的Hadoop数据提供快速、交互式SQL查询。除了使用相同的统一存储平台之外，Impala还使用与Apache Hive相同的元数据，SQL语法（Hive SQL），ODBC驱动程序和用户界面（Hue中的Impala查询UI）。这为实时或面向批处理的查询提供了一个熟悉且统一的平台。作为查询大数据的工具的补充，Impala不会替代基于MapReduce构建的批处理框架，例如Hive。基于MapReduce构建的Hive和其他框架最适合长时间运行的批处理作业。

数据源类型	简介
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 。
MapReduce服务（ MRS Hudi ）	Hudi是一种数据湖的存储格式，在Hadoop文件系统之上提供了更新数据和删除数据的能力以及消费变化数据的能力。支持多种计算引擎，提供IUD接口，在HDFS的数据集上提供了插入更新和增量拉取的流原语。 Hudi的元数据存放在Hive中，操作通过Spark进行。
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。
MapReduce服务（ MRS Doris ）	Doris是一个高性能、实时的分析型数据库，仅需亚秒级响应时间即可返回海量数据下的查询结果，不仅可以支持高并发的点查询场景，也能支持高吞吐的复杂分析场景。因此，Apache Doris能够较好的满足报表分析、即时查询、统一数仓构建、数据湖联邦查询加速等使用场景。
云数据库 RDS	华为云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。
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。
ORACLE	ORACLE数据库系统是以分布式数据库为核心的一组软件产品，是目前最流行的客户/服务器(CLIENT/SERVER)或B/S体系结构的数据库之一。 ORACLE数据库是目前世界上使用最为广泛的数据库管理系统，作为一个通用的数据库系统，它具有完整的数据管理功能；作为一个关系数据库，它是一个完备关系的产品；作为分布式数据库它实现了分布式处理功能。
实时数据接入 DIS	使用实时数据接入通道，可实现跨空间作业调度。若使用数据通道连接，可以向其他账号的DIS通道发送消息；若不使用，仅能给本账号下所有region的通道发送消息。
Rest Client	通过Rest Client执行一个RESTful请求。目前支持IAM Token、用户名密码两种认证鉴权方式的RESTful请求。

数据源类型	简介
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。

4.2 创建 DataArts Studio 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。

配置开发和生产环境的数据连接后，数据开发时脚本/作业中的开发环境数据连接通过发布流程后，将自动切换对应生产环境的数据连接。

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。
- 主机连接当前仅支持Linux系统主机。
- 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。
- 数据连接中的数据湖认证信息如果发生变化（如密码过期）时，此连接会失效。建议您将数据湖认证信息设定为永久有效，避免由于连接失败导致业务受损。
- 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
- CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。

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等MRS类型的数据连接前，需确保您已购买MRS集群，集群的“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”，并且集群中包含所需要的组件。
- 在创建数据连接前，请确保您已具备连接所需的Agent代理（即CDM集群，如果无可用CDM集群请参考**创建CDM集群**进行创建），且待连接的数据湖与CDM集群之间网络互通。
  - 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。请确保数据源所在的主机和CDM集群均能访问公网，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：
    - 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。
    - 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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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。
- 如果使用企业模式，您还需要注意以下事项：

  由于企业模式下需要区分开发环境和生产环境，因此您需要分别准备对应生产环境和开发环境的两套数据湖服务，用于隔离开发和生产环境：
  - 对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS），**如果使用两套集群**，DataArts Studio通过管理中心的创建数据连接区分开发环境和生产环境的数据湖服务，在开发和生产流程中自动切换对应的数据湖。因此您需要准备两套数据湖服务，且两套数据湖服务的版本、规格、组件、区域、VPC、子网以及相关配置等信息，均应保持一致。创建数据连接的详细操作请参见**创建DataArts Studio数据连接**。
  - 对于Serverless服务（例如DLI），DataArts Studio通过管理中心的环境隔离来配置生产环境和开发环境数据湖服务的对应关系，在开发和生产流程中自动切换对应的数据湖。因此您需要在Serverless数据湖服务中准备两套队列、数据库资源，建议通过名称后缀进行区分，详细操作请参见**配置DataArts Studio企业模式环境隔离**。
  - 对于DWS、MRS Hive和MRS Spark这三种数据源，如果在创建数据连接时**选择同一个集群**，则需要配置数据源资源映射的DB数据库映射关系进行开发生产环境隔离，详细操作请参见**DB配置**。
  - 离线处理集成作业不支持在企业模式下运行。

  例如，当您的数据湖服务为MRS集群时，需要准备两套MRS集群，且版本、规格、组件、区域、VPC、子网等保持一致。如果某个MRS集群修改了某些配置，也需要同步到另一套MRS集群上。

创建数据连接

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

步骤3 在管理中心页面，单击“数据连接”，进入数据连接页面并单击“创建数据连接”。

图 4-1 创建数据连接

步骤4 在创建连接页面中，选择“数据连接类型”，并参见**表4-3**配置相关参数。

说明

- 对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS），**如果使用两套集群**，DataArts Studio通过管理中心的创建数据连接区分开发环境和生产环境的数据湖服务，在开发和生产流程中自动切换对应的数据湖。因此您需要准备两套数据湖服务，且两套数据湖服务的版本、规格、组件、区域、VPC、子网以及相关配置等信息，均应保持一致。创建数据连接的详细操作请参见**创建DataArts Studio数据连接**。
- 对于Serverless服务（例如DLI），DataArts Studio通过管理中心的环境隔离来配置生产环境和开发环境数据湖服务的对应关系，在开发和生产流程中自动切换对应的数据湖。因此您需要在Serverless数据湖服务中准备两套队列、数据库资源，建议通过名称后缀进行区分，详细操作请参见**配置DataArts Studio企业模式环境隔离**。
- 对于DWS、MRS Hive和MRS Spark这三种数据源，如果在创建数据连接时**选择同一个集群**，则需要配置数据源资源映射的DB数据库映射关系进行开发生产环境隔离，详细操作请参见**DB配置**。
- 离线处理集成作业不支持在企业模式下运行。

表 4-3 数据连接

数据连接类型	参数说明
DWS	请参见**DWS数据连接参数说明**。
DLI	请参见**DLI数据连接参数说明**。
MRS Hive	请参见**MRS Hive数据连接参数说明**。
MRS HBase	请参见**MRS HBase数据连接参数说明**。
MRS Kafka	请参见**MRS Kafka数据连接参数说明**。
MRS Spark	请参见**MRS Spark数据连接参数说明**。
MRS Clickhouse	请参见**MRS Clickhouse数据连接参数说明**。
MRS Hetu	请参见**MRS Hetu数据连接参数说明**。
MRS Impala	请参见**MRS Impala数据连接参数说明**。
MRS Presto	请参见**MRS Presto数据连接参数说明**。

数据连接类型	参数说明
MRS Doris	请参见**Doris数据连接参数说明**。
OpenSource Clickhouse	请参见**OpenSource ClickHouse数据连接参数说明**。
RDS	请参见**RDS数据连接参数说明**。 RDS连接类型支持连接RDS中的MySQL/PostgreSQL/达梦数据库 DM/SQL Server/SAP HANA等关系型数据库。
MySQL(待下线)	不建议使用MySQL(待下线)连接器，推荐使用RDS连接MySQL数据源，请参见**RDS数据连接参数说明**。
ORACLE	请参见**ORACLE数据连接参数说明**。
DIS	请参见**DIS数据连接参数说明**。
主机连接	请参见**主机连接参数说明**。
Rest Client	请参见**Rest Client数据连接参数说明**。
Redis	请参见**Redis数据连接参数说明**。
SAP HANA	请参见**SAP HANA数据连接参数说明**。

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“保存”，完成数据连接的创建。

----结束

相关操作

- 编辑数据连接：在数据连接页面的连接列表中，找到所需编辑的连接，然后单击“编辑”。根据需要修改连接参数，参数描述可参考**表4-3**。

  说明

  编辑时如果不涉及修改密码，可不填写此项，系统会自动带入上次连接创建时的密码。

  完成修改后，单击“测试”去测试数据连接是否可以正常连接，如果可以正常连接，单击“保存”。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。
- 删除数据连接：在数据连接页面的连接列表中，找到所需删除的连接，然后单击“删除”。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”。

  如果待删除的连接已被引用，则不可直接删除。删除前需要根据删除提示窗口中的数据连接引用列表，到各组件中解除对该连接的引用，然后再尝试重新删除。

  说明

  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。

4.3 配置 DataArts Studio 数据连接参数

4.3.1 DWS 数据连接参数说明

表 4-4 DWS 数据连接

参数	是否必选	说明
数据连接类型	是	DWS连接固定选择为数据仓库服务（DWS）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** ● 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 ● 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
SSL加密	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方式进行通信。 ● 开关关闭，SSL通道加密和证书认证两种方式均可进行通信。
手动	是	选择连接模式。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。 ● 使用连接串模式时，手动填写对应集群的IP或域名、端口进行连接配置，且需打通本连接Agent（即CDM集群）和DWS集群之间的网络。 **说明** 数据安全组件不支持连接串模式的DWS连接。
DWS集群名	是	“手动”选择为“集群名模式”时需要配置本参数。 选择DWS集群，系统会显示所有项目ID和企业项目相同的DWS集群。

参数	是否必选	说明
IP或域名	是	“手动”选择为“连接串模式”时需要配置本参数。 “IP或域名”如果手动填写，必须写内网IP，端口必须为对资源组网段放开的端口，否则可能导致网络连接不通。 表示通过内部网络访问集群数据库的访问地址，可填写为IP或域名。内网访问IP或域名地址在创建集群时自动生成，您可以通过管理控制台获取访问地址： 1. 根据注册的账号登录DWS云服务管理控制台。 2. 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IP、域名和端口等信息。
端口	是	“手动”选择为“连接串模式”时需要配置本参数。 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。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
绑定Agent	是	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，请参考**创建CDM集群**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DWS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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
数据源认证及其他功能配置		
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。

参数	是否必选	说明
元数据采集范围	否	配置元数据实时同步的数据库和数据表范围，不填写默认不筛选。 可填写为如下两种形式之一： ● database_name：筛选数据库名包含“database_name”的数据库 ● database_name.table_name：筛选数据库名包含“database_name”的数据库，在匹配到的数据库中再匹配表名包含“table_name”的数据表 例如： ● 填写为“datatest”，则元数据实时同步将同步数据库名包含“datatest”的数据库中的数据表。 ● 填写为“datatest.table1”，则元数据实时同步将同步如下数据表：数据库名包含“datatest”的数据库，其中表名包含“table_name”的数据表。

4.3.2 DLI 数据连接参数说明

表 4-5 DLI 数据连接

参数	是否必选	说明
数据连接类型	是	DLI连接固定选择为数据湖探索（DLI）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** ● 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 ● 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。

参数	是否必选	说明
元数据采集范围	否	配置元数据实时同步的数据库和数据表范围，不填写默认不筛选。 可填写为如下两种形式之一： ● database_name：筛选数据库名包含“database_name”的数据库 ● database_name.table_name：筛选数据库名包含“database_name”的数据库，在匹配到的数据库中再匹配表名包含“table_name”的数据表 例如： ● 填写为“datatest”，则元数据实时同步将同步数据库名包含“datatest”的数据库中的数据表。 ● 填写为“datatest.table1”，则元数据实时同步将同步如下数据表：数据库名包含“datatest”的数据库，其中表名包含“table_name”的数据表。
基础与网络连通配置		
项目ID	否	适用组件勾选数据集成后，呈现此参数。 DLI服务所在区域的项目ID。 项目ID表示租户的资源，账号ID对应当前账号，IAM用户ID对应当前用户。用户可在对应页面下查看不同Region对应的项目ID、账号ID和用户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“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。

4.3.3 MRS Hive 数据连接参数说明

表 4-6 MRS Hive 数据连接

参数	是否必选	说明
数据连接类型	是	MRS Hive连接固定选择为MapReduce服务（ MRS Hive ）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** ● 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 ● 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
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. MRS API连接仅支持在数据开发组件使用，其他组件例如数据架构、数据质量、数据目录等无法使用此连接。 2. 在数据开发组件不支持通过可视化方式查看与管理该连接下的数据库、数据表和字段。特别的，仅当连接MRS 3.2.1以及之后版本的MRS集群时，支持通过可视化方式查看数据库、数据表和字段，但仍不支持可视化方式管理。 3. 在数据开发组件的SQL编辑器运行SQL时，只能以日志形式显示执行结果。 **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”。
手动	是	通过代理连接时，是必选项。 选择连接模式。如无访问其他项目或企业项目下MRS集群的需求，使用集群名模式即可。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。仅可选择本项目内且企业项目相同的MRS集群进行连接。 ● 使用连接串模式时，通过手动输入Manager IP，并打通本连接Agent（即CDM集群）和MRS集群之间的网络，则可以访问其他项目或企业项目的MRS集群。

参数	是否必选	说明
Manager IP	是	使用连接串模式时，是必选项。 此参数填写为MRS Manager的浮动IP地址。仅支持连接MRS云服务，自建Hadoop集群必须先纳管到MRS云服务才能连接。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 注意，通过输入框后的“选择”按钮仅能获取本项目内且企业项目相同的MRS集群，如果需要访问其他项目或企业项目的MRS集群，则需要获取MRS Manager的浮动IP地址并手动输入，并确保已打通本连接Agent（即CDM集群）和MRS租户面集群之间的网络。Manager的浮动IP地址可通过登录MRS集群主Master节点获取，执行ifconfig命令，回显中eth0:wsom的IP就是MRS Manager的浮动IP。登录MRS集群Master节点请参见**登录集群节点**章节，如果登录的是非主Master节点无法查询，请切换到另一个Master节点查询。 手动填写IP时请根据场景和顺序填写，多个IP之间使用","分隔。例如: 127.0.0.1或127.0.0.1,127.0.0.2,127.0.0.3。 ● 填写单个IP，IP应为MRS集群管理面的浮动IP。 ● 填写3个IP时，应填写MRS集群业务面的主节点IP、备节点IP和MRS集群管理面的浮动IP。

参数	是否必选	说明
MRS集群名	是	通过MRS API连接或使用集群名模式时，是必选项。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。
KMS密钥	否	通过代理连接时，是必选项。 通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	通过代理连接时，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，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
数据源认证及其他功能配置		
认证类型	是	使用连接串模式时，是必选项。 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。
用户名	是	MRS集群的人机用户，通过代理连接时是必选项。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**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。

参数	是否必选	说明
密码	是	MRS集群的访问密码，通过代理连接的时候，是必选项。
开启ldap	否	当“连接方式”参数选择为“通过代理连接”时，显示该配置项。 当MRS Hive对接外部LDAP开启了LDAP认证时，连接Hive时需要使用LDAP账号与密码进行认证，此时必须开启此参数，否则会连接失败。
ldap用户名	是	当“开启ldap”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的用户名。
ldap密码	是	当“开启ldap”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的密码。
元数据采集范围	否	配置元数据实时同步的数据库和数据表范围，不填写默认不筛选。 可填写为如下两种形式之一： ● database_name：筛选数据库名包含“database_name”的数据库 ● database_name.table_name：筛选数据库名包含“database_name”的数据库，在匹配到的数据库中再匹配表名包含“table_name”的数据表 例如： ● 填写为“datatest”，则元数据实时同步将同步数据库名包含“datatest”的数据库中的数据表。 ● 填写为“datatest.table1”，则元数据实时同步将同步如下数据表：数据库名包含“datatest”的数据库，其中表名包含“table_name”的数据表。
OBS支持	否	适用组件勾选数据集成后，呈现此参数。 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。
使用委托	否	适用组件勾选数据集成后，呈现此参数。 开启委托功能，即可以在无需持有永久AKSK的情况下创建数据连接，根据DLF配置的调度身份执行CDM作业。
公共委托	否	适用组件勾选数据集成且“使用委托”选择“是”时，呈现此参数。 仅涉及用于测试该连接委托功能是否正常，作业运行将根据DLF配置的调度身份执行CDM作业。

参数	是否必选	说明
访问标识(AK)	-	适用组件勾选数据集成且“OBS支持”选择“是”时，呈现此参数。 AK和SK分别为登录OBS服务器的访问标识与密钥。 您需要先创建当前账号的访问密钥，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4-2所示。 **图 4-2** 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)	-	

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账户登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 安全策略 > 密码策略”，单击“新增密码策略”，添加一个永不过期的密码策略。
   - “密码策略名”可配置为“neverexp”。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为kerberos认证用户，密码策略选择为永不过期策略“neverexp”，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。

4. 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。
5. 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账户登录MRS Manager页面。
2. 在Manager页面的“系统设置”中，单击“密码策略配置”，修改密码策略。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 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，添加一个专有的人机用户作为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。

4. 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。
5. 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。

4.3.4 MRS HBase 数据连接参数说明

表 4-7 MRS HBase 数据连接

参数	是否必选	说明
数据连接类型	是	MRS HBase连接固定选择为MapReduce服务（MRS HBase）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。
基础与网络连通配置		
手动	是	选择连接模式。如无访问其他项目或企业项目下MRS集群的需求，使用集群名模式即可。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。仅可选择本项目内且企业项目相同的MRS集群进行连接。 ● 使用连接串模式时，通过手动输入Manager IP，并打通本连接Agent（即CDM集群）和MRS集群之间的网络，则可以访问其他项目或企业项目的MRS集群。

参数	是否必选	说明
Manager IP	是	使用连接串模式时，是必选项。 此参数填写为MRS Manager的浮动IP地址。仅支持连接MRS云服务，自建Hadoop集群必须先纳管到MRS云服务才能连接。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 注意，通过输入框后的“选择”按钮仅能获取本项目内且企业项目相同的MRS集群，如果需要访问其他项目或企业项目的MRS集群，则需要获取MRS Manager的浮动IP地址并手动输入，并确保已打通本连接Agent（即CDM集群）和MRS租户面集群之间的网络。Manager的浮动IP地址可通过登录MRS集群主Master节点获取，执行ifconfig命令，回显中eth0:wsom的IP就是MRS Manager的浮动IP。登录MRS集群Master节点请参见**登录集群节点**章节，如果登录的是非主Master节点无法查询，请切换到另一个Master节点查询。 手动填写IP时请根据场景和顺序填写，多个IP之间使用","分隔。例如: 127.0.0.1或127.0.0.1,127.0.0.2,127.0.0.3。 ● 填写单个IP，IP应为MRS集群管理面的浮动IP。 ● 填写3个IP时，应填写MRS集群业务面的主节点IP、备节点IP和MRS集群管理面的浮动IP。

参数	是否必选	说明
MRS集群名	是	使用集群名模式时，是必选项。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
数据源认证及其他功能配置		
认证类型	是	使用连接串模式时，是必选项。 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。
用户名	是	MRS集群的用户名。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**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。
密码	是	MRS集群的访问密码。

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账户登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 安全策略 > 密码策略”，单击“新增密码策略”，添加一个永不过期的密码策略。
   - “密码策略名”可配置为“neverexp”。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为kerberos认证用户，密码策略选择为永不过期策略“neverexp”，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。
4. 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。
5. 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账户登录MRS Manager页面。
2. 在Manager页面的“系统设置”中，单击“密码策略配置”，修改密码策略。

- “密码有效期（天）”配置为“0”，表示永不过期。
- “密码失效提前提醒天数”配置为“0”。
- 其他参数保持默认即可。

3. 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，添加一个专有的人机用户作为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。

4. 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。
5. 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。

4.3.5 MRS Kafka 数据连接参数说明

表 4-8 MRS Kafka 数据连接

参数	是否必选	说明
数据连接类型	是	MRS Kafka连接固定选择为MapReduce服务（MRS Kafka）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。

参数	是否必选	说明
基础与网络连通配置		
手动	是	选择连接模式。如无访问其他项目或企业项目下MRS集群的需求，使用集群名模式即可。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。仅可选择本项目内且企业项目相同的MRS集群进行连接。 ● 使用连接串模式时，通过手动输入Manager IP，并打通本连接Agent（即CDM集群）和MRS集群之间的网络，则可以访问其他项目或企业项目的MRS集群。
Manager IP	是	使用连接串模式时，是必选项。 此参数填写为MRS Manager的浮动IP地址。仅支持连接MRS云服务，自建Hadoop集群必须先纳管到MRS云服务才能连接。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 注意，通过输入框后的“选择”按钮仅能获取本项目内且企业项目相同的MRS集群，如果需要访问其他项目或企业项目的MRS集群，则需要获取MRS Manager的浮动IP地址并手动输入，并确保已打通本连接Agent（即CDM集群）和MRS租户面集群之间的网络。Manager的浮动IP地址可通过登录MRS集群主Master节点获取，执行ifconfig命令，回显中eth0:wsom的IP就是MRS Manager的浮动IP。登录MRS集群Master节点请参见**登录集群节点**章节，如果登录的是非主Master节点无法查询，请切换到另一个Master节点查询。 手动填写IP时请根据场景和顺序填写，多个IP之间使用","分隔。例如: 127.0.0.1或127.0.0.1,127.0.0.2,127.0.0.3。 ● 填写单个IP，IP应为MRS集群管理面的浮动IP。 ● 填写3个IP时，应填写MRS集群业务面的主节点IP、备节点IP和MRS集群管理面的浮动IP。

参数	是否必选	说明
MRS集群名	是	使用集群名模式时，是必选项。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
数据源认证及其他功能配置		
认证类型	是	使用连接串模式时，是必选项。 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。
用户名	是	MRS集群的用户名。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**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。
密码	是	MRS集群的访问密码。

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账户登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 安全策略 > 密码策略”，单击“新增密码策略”，添加一个永不过期的密码策略。
   - “密码策略名”可配置为“neverexp”。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为kerberos认证用户，密码策略选择为永不过期策略“neverexp”，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。
4. 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。
5. 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账户登录MRS Manager页面。
2. 在Manager页面的“系统设置”中，单击“密码策略配置”，修改密码策略。

- “密码有效期（天）”配置为“0”，表示永不过期。
- “密码失效提前提醒天数”配置为“0”。
- 其他参数保持默认即可。

3. 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，添加一个专有的人机用户作为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。

4. 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。
5. 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。

4.3.6 MRS Spark 数据连接参数说明

表 4-9 MRS Spark 数据连接

参数	是否必选	说明
数据连接类型	是	MRS Spark连接固定选择为MapReduce服务（MRS Spark）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。

参数	是否必选	说明
基础与网络连通配置		
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. MRS API连接仅支持在数据开发组件使用，其他组件例如数据架构、数据质量、数据目录等无法使用此连接。 2. 在数据开发组件不支持通过可视化方式查看与管理该连接下的数据库、数据表和字段。特别的，仅当连接MRS 3.2.1以及之后版本的MRS集群时，支持通过可视化方式查看数据库、数据表和字段，但仍不支持可视化方式管理。 3. 在数据开发组件的SQL编辑器运行SQL时，只能以日志形式显示执行结果。 **说明** MRS API连接方式的MRS Spark数据连接适用于数据开发场景，代理连接方式的MRS Spark数据连接适用于数据治理场景。 ● 为保证数据开发场景下，支持为每个Spark SQL作业独立配置需要的资源（例如线程、内存、CPU核数并指定MRS资源队列等），连接方式需要配置为“MRS API连接”。注意，代理连接不支持为每个Spark SQL作业独立配置资源。 ● 为保证数据架构等其他组件能够使用此连接，连接方式需要配置为“通过代理连接”。
手动	是	通过代理连接时，是必选项。 选择连接模式。如无访问其他项目或企业项目下MRS集群的需求，使用集群名模式即可。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。仅可选择本项目内且企业项目相同的MRS集群进行连接。 ● 使用连接串模式时，通过手动输入Manager IP，并打通本连接Agent（即CDM集群）和MRS集群之间的网络，则可以访问其他项目或企业项目的MRS集群。

参数	是否必选	说明
Manager IP	是	使用连接串模式时，是必选项。 此参数填写为MRS Manager的浮动IP地址。仅支持连接MRS云服务，自建Hadoop集群必须先纳管到MRS云服务才能连接。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 注意，通过输入框后的“选择”按钮仅能获取本项目内且企业项目相同的MRS集群，如果需要访问其他项目或企业项目的MRS集群，则需要获取MRS Manager的浮动IP地址并手动输入，并确保已打通本连接Agent（即CDM集群）和MRS租户面集群之间的网络。Manager的浮动IP地址可通过登录MRS集群主Master节点获取，执行ifconfig命令，回显中eth0:wsom的IP就是MRS Manager的浮动IP。登录MRS集群Master节点请参见**登录集群节点**章节，如果登录的是非主Master节点无法查询，请切换到另一个Master节点查询。 手动填写IP时请根据场景和顺序填写，多个IP之间使用","分隔。例如: 127.0.0.1或127.0.0.1,127.0.0.2,127.0.0.3。 ● 填写单个IP，IP应为MRS集群管理面的浮动IP。 ● 填写3个IP时，应填写MRS集群业务面的主节点IP、备节点IP和MRS集群管理面的浮动IP。

参数	是否必选	说明
MRS集群名	是	通过MRS API连接或使用集群名模式时，是必选项。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。
KMS密钥	是	通过代理连接时，是必选项。 通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	通过代理连接时，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，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
数据源认证及其他功能配置		
认证类型	是	使用连接串模式时，是必选项。 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。
MRS版本	否	使用连接串模式时，是必选项。 选择MRS集群的版本。
组件名	否	使用连接串模式时，是必选项。 选择Spark组件的版本。

参数	是否必选	说明
用户名	是	MRS集群的人机用户，通过代理连接时是必选项。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**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。
密码	是	MRS集群的访问密码，通过代理连接的时候，是必选项。

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账户登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 安全策略 > 密码策略”，单击“新增密码策略”，添加一个永不过期的密码策略。
   - “密码策略名”可配置为“neverexp”。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为kerberos认证用户，密码策略选择为永不过期策略“neverexp”，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。

4. 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。
5. 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账户登录MRS Manager页面。
2. 在Manager页面的“系统设置”中，单击“密码策略配置”，修改密码策略。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 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，添加一个专有的人机用户作为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。

4. 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。
5. 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。

4.3.7 MRS Clickhouse 数据连接参数说明

表 4-10 MRS Clickhouse 数据连接

参数	是否必选	说明
数据连接类型	是	MRS Clickhouse连接固定选择为MapReduce服务（MRS Clickhouse）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。
基础与网络连通配置		
手动	是	选择连接模式。如无访问其他项目或企业项目下MRS集群的需求，使用集群名模式即可。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。仅可选择本项目内且企业项目相同的MRS集群进行连接。 ● 使用连接串模式时，通过手动输入Manager IP，并打通本连接Agent（即CDM集群）和MRS集群之间的网络，则可以访问其他项目或企业项目的MRS集群。

参数	是否必选	说明
Manager IP	是	使用连接串模式时，是必选项。 此参数填写为MRS Manager的浮动IP地址。仅支持连接MRS云服务，自建Hadoop集群必须先纳管到MRS云服务才能连接。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 注意，通过输入框后的“选择”按钮仅能获取本项目内且企业项目相同的MRS集群，如果需要访问其他项目或企业项目的MRS集群，则需要获取MRS Manager的浮动IP地址并手动输入，并确保已打通本连接Agent（即CDM集群）和MRS租户面集群之间的网络。Manager的浮动IP地址可通过登录MRS集群主Master节点获取，执行ifconfig命令，回显中eth0:wsom的IP就是MRS Manager的浮动IP。登录MRS集群Master节点请参见**登录集群节点**章节，如果登录的是非主Master节点无法查询，请切换到另一个Master节点查询。 手动填写IP时请根据场景和顺序填写，多个IP之间使用","分隔。例如: 127.0.0.1或127.0.0.1,127.0.0.2,127.0.0.3。 ● 填写单个IP，IP应为MRS集群管理面的浮动IP。 ● 填写3个IP时，应填写MRS集群业务面的主节点IP、备节点IP和MRS集群管理面的浮动IP。

参数	是否必选	说明
MRS集群名	是	使用集群名模式时，是必选项。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 2.9.2及以后的CDM版本才支持MRS Clickhouse连接。 • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
数据源认证及其他功能配置		
认证类型	是	使用连接串模式时，是必选项。 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。
用户名	是	MRS集群的用户名。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**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。
密码	是	MRS集群的访问密码。

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账户登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 安全策略 > 密码策略”，单击“新增密码策略”，添加一个永不过期的密码策略。
   - “密码策略名”可配置为“neverexp”。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为kerberos认证用户，密码策略选择为永不过期策略“neverexp”，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。
4. 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。
5. 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账户登录MRS Manager页面。
2. 在Manager页面的“系统设置”中，单击“密码策略配置”，修改密码策略。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 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，添加一个专有的人机用户作为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。
4. 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。
5. 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。

4.3.8 MRS Hetu 数据连接参数说明

表 4-11 MRS Hetu 数据连接

参数	是否必选	说明
数据连接类型	是	MRS Hetu连接固定选择为MapReduce服务（MRS Hetu）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。
基础与网络连通配置		
手动	是	选择连接模式。如无访问其他项目或企业项目下MRS集群的需求，使用集群名模式即可。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。仅可选择本项目内且企业项目相同的MRS集群进行连接。 ● 使用连接串模式时，通过手动输入Manager IP，并打通本连接Agent（即CDM集群）和MRS集群之间的网络，则可以访问其他项目或企业项目的MRS集群。
Manager IP	是	使用连接串模式时，是必选项。 此参数填写为MRS Manager的浮动IP地址。仅支持连接MRS云服务，自建Hadoop集群必须先纳管到MRS云服务才能连接。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 **说明** ● 仅支持连接MRS 3.1.1及以上版本的MRS集群。 ● 若需要连接MRS 3.2.1版本集群，则需通过HetuEngine WebUI界面，给计算实例添加如下自定义参数：参数名为“protocol.v1.alternate-header-name”，值为“Presto”，参数文件为“coordinator.config.properties”和“worker.config.properties”。 注意，通过输入框后的“选择”按钮仅能获取本项目内且企业项目相同的MRS集群，如果需要访问其他项目或企业项目的MRS集群，则需要获取MRS Manager的浮动IP地址并手动输入，并确保已打通本连接Agent（即CDM集群）和MRS租户面集群之间的网络。Manager的浮动IP地址可通过登录MRS集群主Master节点获取，执行ifconfig命令，回显中eth0:wsom的IP就是MRS Manager的浮动IP。登录MRS集群Master节点请参见**登录集群节点**章节，如果登录的是非主Master节点无法查询，请切换到另一个Master节点查询。 手动填写IP时请根据场景和顺序填写，多个IP之间使用","分隔。例如: 127.0.0.1或127.0.0.1,127.0.0.2,127.0.0.3。 ● 填写单个IP，IP应为MRS集群管理面的浮动IP。 ● 填写3个IP时，应填写MRS集群业务面的主节点IP、备节点IP和MRS集群管理面的浮动IP。

参数	是否必选	说明
MRS集群名	是	使用集群名模式时，是必选项。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 **说明** • 仅支持连接MRS 3.1.1及以上版本的MRS集群。 • 若需要连接MRS 3.2.1版本集群，则需通过HetuEngine WebUI界面，给计算实例添加如下自定义参数：参数名为“protocol.v1.alternate-header-name”，值为“Presto”，参数文件为“coordinator.config.properties”和“worker.config.properties”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 2.9.2及以后的CDM版本才支持MRS Hetu连接。 • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
hsbroker IP列表	是	MRS Hetu组件的hsbroker节点ip列表，多个ip用“,”分隔。 获取方法： 1. 登录MRS FusionInsight Manager。 2. 选择“集群 > 服务 > HetuEngine > 角色 > HSBroker”，获取HSBroker所有实例的业务IP。
hsbroker端口	是	MRS Hetu组件的hsbroker节点端口号。 获取方法： 1. 登录MRS FusionInsight Manager。 2. 选择“集群 > 服务 > HetuEngine > 配置 > 全部配置”，在右侧搜索“server.port”，获取HSBroker的端口号。
数据源认证及其他功能配置		
认证类型	是	使用连接串模式时，是必选项。 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。

参数	是否必选	说明
用户名	是	MRS集群的用户名，该用户需要具有HetuEngine组件的权限。 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。您可以参考**创建MRS安全集群的kerberos认证用户**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。 **须知** 创建完Hetu用户后，您还需参考**从零开始使用HetuEngine**，完成该章节中的所有配置才能创建Hetu连接。
密码	是	MRS集群的访问密码。

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账户登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 安全策略 > 密码策略”，单击“新增密码策略”，添加一个永不过期的密码策略。
   - “密码策略名”可配置为“neverexp”。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为kerberos认证用户，密码策略选择为永不过期策略“neverexp”，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。

4. 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。
5. 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账户登录MRS Manager页面。
2. 在Manager页面的“系统设置”中，单击“密码策略配置”，修改密码策略。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 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，添加一个专有的人机用户作为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。

4. 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。
5. 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。

4.3.9 MRS Impala 数据连接参数说明

表 4-12 MRS Impala 数据连接

参数	是否必选	说明
数据连接类型	是	MRS Impala连接固定选择为MapReduce服务（MRS Impala）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。
基础与网络连通配置		
手动	是	选择连接模式。如无访问其他项目或企业项目下MRS集群的需求，使用集群名模式即可。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。仅可选择本项目内且企业项目相同的MRS集群进行连接。 ● 使用连接串模式时，通过手动输入Manager IP，并打通本连接Agent（即CDM集群）和MRS集群之间的网络，则可以访问其他项目或企业项目的MRS集群。

参数	是否必选	说明
Manager IP	是	使用连接串模式时，是必选项。 此参数填写为MRS Manager的浮动IP地址。仅支持连接MRS云服务，自建Hadoop集群必须先纳管到MRS云服务才能连接。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 注意，通过输入框后的“选择”按钮仅能获取本项目内且企业项目相同的MRS集群，如果需要访问其他项目或企业项目的MRS集群，则需要获取MRS Manager的浮动IP地址并手动输入，并确保已打通本连接Agent（即CDM集群）和MRS租户面集群之间的网络。Manager的浮动IP地址可通过登录MRS集群主Master节点获取，执行ifconfig命令，回显中eth0:wsom的IP就是MRS Manager的浮动IP。登录MRS集群Master节点请参见**登录集群节点**章节，如果登录的是非主Master节点无法查询，请切换到另一个Master节点查询。 手动填写IP时请根据场景和顺序填写，多个IP之间使用","分隔。例如: 127.0.0.1或127.0.0.1,127.0.0.2,127.0.0.3。 ● 填写单个IP，IP应为MRS集群管理面的浮动IP。 ● 填写3个IP时，应填写MRS集群业务面的主节点IP、备节点IP和MRS集群管理面的浮动IP。

参数	是否必选	说明
MRS集群名	是	使用集群名模式时，是必选项。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	MRS为非全托管服务，DataArts Studio无法直接与非全托管服务进行连接。CDM集群提供了DataArts Studio与非全托管服务通信的代理，所以创建MRS的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 2.9.2及以后的CDM版本才支持MRS Impala连接。 • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
Impalad IP列表	是	MRS Impala组件Impalad角色的管理IP。 获取方法： 1. 登录MRS FusionInsight Manager。 2. 选择“集群 > 服务 > Impala > 实例”，可查看Impalad管理IP地址。
数据源认证及其他功能配置		
认证类型	是	使用连接串模式时，是必选项。 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。

参数	是否必选	说明
用户名	是	MRS集群的人机用户，通过代理连接时是必选项。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**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。
密码	是	MRS集群的访问密码，通过代理连接的时候，是必选项。
开启ldap	否	通过代理连接的时候，此项可配置。 当MRS Impala对接外部LDAP开启了LDAP认证时，连接MRS Impala时需要使用LDAP账号与密码进行认证，此时必须开启此参数，否则会连接失败。
ldap用户名	是	当“开启ldap”参数选择为“是”时，此参数是必选项。 填写为MRS Impala开启LDAP认证时配置的用户名。
ldap密码	是	当“开启ldap”参数选择为“是”时，此参数是必选项。 填写为MRS Impala开启LDAP认证时配置的密码。

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账户登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 安全策略 > 密码策略”，单击“新增密码策略”，添加一个永不过期的密码策略。

- “密码策略名”可配置为“neverexp”。
- “密码有效期（天）”配置为“0”，表示永不过期。
- “密码失效提前提醒天数”配置为“0”。
- 其他参数保持默认即可。

3. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为kerberos认证用户，密码策略选择为永不过期策略“neverexp”，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。

4. 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。
5. 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账户登录MRS Manager页面。
2. 在Manager页面的“系统设置”中，单击“密码策略配置”，修改密码策略。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 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，添加一个专有的人机用户作为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。

4. 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。
5. 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。

4.3.10 MRS Ranger 数据连接参数说明

表 4-13 MRS Ranger 数据连接

参数	是否必选	说明
数据连接类型	是	MRS Ranger连接固定选择为MapReduce服务（MRS Ranger）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。
基础与网络连通配置		

参数	是否必选	说明
手动	是	选择连接模式。如无访问其他项目或企业项目下MRS集群的需求，使用集群名模式即可。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。仅可选择本项目内且企业项目相同的MRS集群进行连接。 ● 使用连接串模式时，通过手动输入Manager IP，并打通本连接Agent（即CDM集群）和MRS集群之间的网络，则可以访问其他项目或企业项目的MRS集群。
Manager IP	是	使用连接串模式时，是必选项。 此参数填写为MRS Manager的浮动IP地址。仅支持连接MRS云服务，自建Hadoop集群必须先纳管到MRS云服务才能连接。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 注意，通过输入框后的“选择”按钮仅能获取本项目内且企业项目相同的MRS集群，如果需要访问其他项目或企业项目的MRS集群，则需要获取MRS Manager的浮动IP地址并手动输入，并确保已打通本连接Agent（即CDM集群）和MRS租户面集群之间的网络。Manager的浮动IP地址可通过登录MRS集群主Master节点获取，执行ifconfig命令，回显中eth0:wsom的IP就是MRS Manager的浮动IP。登录MRS集群Master节点请参见**登录集群节点**章节，如果登录的是非主Master节点无法查询，请切换到另一个Master节点查询。 手动填写IP时请根据场景和顺序填写，多个IP之间使用","分隔。例如: 127.0.0.1或127.0.0.1,127.0.0.2,127.0.0.3。 ● 填写单个IP，IP应为MRS集群管理面的浮动IP。 ● 填写3个IP时，应填写MRS集群业务面的主节点IP、备节点IP和MRS集群管理面的浮动IP。 **说明** 当绑定Agent选择的CDM集群为2.9.3.300及以下版本时，仅支持与安全模式集群的MRS Ranger创建连接。 如果需要与非安全模式集群的MRS Ranger创建连接，则需要确保CDM集群为2.10.0.300及以上版本，或联系客服或技术支持人员升级CDM集群中的dlg-agent组件版本。

参数	是否必选	说明
MRS集群名	是	使用集群名模式时，是必选项。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。 **说明** 当绑定Agent选择的CDM集群为2.9.3.300及以下版本时，仅支持与安全模式集群的MRS Ranger创建连接。 如果需要与非安全模式集群的MRS Ranger创建连接，则需要确保CDM集群为2.10.0.300及以上版本，或联系客服或技术支持人员升级CDM集群中的dlg-agent组件版本。
IP	是	MRS Ranger组件中RangerAdmin角色的管理IP。如果有多个ip用“,”分隔。 获取方法： 1. 登录MRS FusionInsight Manager。 2. 选择“集群 > 服务 > Ranger > 实例”，可查看RangerAdmin角色对应的管理IP地址。

参数	是否必选	说明
端口	是	MRS Ranger组件的实例端口号。 获取方法： 1. 登录MRS FusionInsight Manager。 2. 选择“集群 > 服务 > Ranger > 配置 > 基础配置”，非安全模式MRS集群查看“ranger.service.http.port”参数对应的端口，安全模式MRS集群查看“ranger.service.https.port”参数对应的端口。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
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，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
数据源认证及其他功能配置		
认证类型	是	使用连接串模式时，是必选项。 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。

参数	是否必选	说明
用户名	是	MRS集群的用户名。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**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。
密码	是	MRS集群的访问密码。

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账户登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 安全策略 > 密码策略”，单击“新增密码策略”，添加一个永不过期的密码策略。
   - “密码策略名”可配置为“neverexp”。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为kerberos认证用户，密码策略选择为永不过期策略“neverexp”，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。

4. 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。
5. 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账户登录MRS Manager页面。
2. 在Manager页面的“系统设置”中，单击“密码策略配置”，修改密码策略。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 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，添加一个专有的人机用户作为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。

4. 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。
5. 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。

4.3.11 MRS Presto 数据连接参数说明

表 4-14 MRS Presto 数据连接

参数	是否必选	说明
数据连接类型	是	MRS Presto连接固定选择为MapReduce服务（MRS Presto）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。
基础与网络连通配置		

参数	是否必选	说明
MRS集群名	是	选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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。
数据源认证及其他功能配置		
描述	否	可自定义填写相关连接的描述。

4.3.12 Doris 数据连接参数说明

表 4-15 Doris 数据连接

参数	是否必选	说明
数据连接类型	是	Doris连接固定选择为Doris。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** • 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 • 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
Doris类型	是	可选择MRS Doris和CloudTable Doris。
MRS集群名	是	当选择MRS Doris时有效。 **说明** 目前仅支持MRS 3.2.0及以上MRS集群版本。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。

参数	是否必选	说明
FE IP	是	MRS集群Doris或者Cloud组件frontend节点的IP，可以填写一个或多个IP。如果有多个ip用“,”分隔。 获取方法： 1. 登录MRS FusionInsight Manager。 2. 选择“集群 > 服务 > Doris > 实例”，获取FE角色的管理IP。
端口	是	Doris FE通过mysql协议查询连接端口。 MRS Doris获取方法： 1. 登录MRS FusionInsight Manager。 2. 选择“集群 > 服务 > Doris > 配置 > 基础配置”，搜索“query_port”查看端口值。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
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，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
SSL加密	否	支持对RDS服务启用SSL加密传输。默认开启SSL，如源端SSL未开启，则需手动关闭SSL加密。
数据源驱动配置		
驱动程序名称	是	驱动程序名称，目前支持MySQL jdbc驱动，驱动名为：com.mysql.jdbc.Driver。
驱动文件来源	是	选择驱动文件的来源方式。

参数	是否必选	说明
驱动文件路径	是	“驱动文件来源”选择“OBS路径”时配置。 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版本及以上版本，如果低于5.1.48版本则连接会报错“The db user or password invalid”。 **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。
驱动文件	是	“驱动文件来源”选择“本地文件”时配置。不同类型的关系数据库，需要适配不同类型的驱动。
数据源认证及其他功能配置		
用户名	是	MRS集群或CloudTable集群的用户名。 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**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。
密码	是	MRS集群或CloudTable集群的访问密码。

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账户登录MRS服务的Manager页面。
2. 在Manager页面选择“系统 > 权限 > 安全策略 > 密码策略”，单击“新增密码策略”，添加一个永不过期的密码策略。
   - “密码策略名”可配置为“neverexp”。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为kerberos认证用户，密码策略选择为永不过期策略“neverexp”，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。
4. 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。
5. 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账户登录MRS Manager页面。
2. 在Manager页面的“系统设置”中，单击“密码策略配置”，修改密码策略。
   - “密码有效期（天）”配置为“0”，表示永不过期。
   - “密码失效提前提醒天数”配置为“0”。
   - 其他参数保持默认即可。
3. 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，添加一个专有的人机用户作为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。

4. 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。
5. 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。

4.3.13 OpenSource ClickHouse 数据连接参数说明

表 4-16 OpenSource ClickHouse 数据连接

参数	是否必选	说明
数据连接类型	是	OpenSource ClickHouse连接固定选择为MapReduce服务（OpenSource ClickHouse）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** • 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 • 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		

参数	是否必选	说明
IP	是	填写ClickHouseServer所在节点IP。
端口	是	默认使用ClickHouseServer的配置参数http_port，用于接收JDBC请求的端口。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中已创建的密钥。
绑定Agent	是	选择CDM集群作为网络代理，必须和ClickHouseServer网络互通才可以成功创建连接。如果没有可用的CDM集群，请参考**创建CDM集群**进行创建。 **说明** CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。

4.3.14 RDS 数据连接参数说明

RDS数据连接支持连接MySQL、PostgreSQL、SQL Server等数据库。

表 4-17 RDS 数据连接

参数	是否必选	说明
数据连接类型	是	RDS连接固定选择为RDS。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** • 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 • 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		

参数	是否必选	说明
IP或域名	是	关系型数据库数据源的访问地址，可填写为IP或域名。 “IP或域名”如果手动填写，必须写内网IP，端口必须为对资源组网段放开的端口，否则可能导致网络连接不通。 ● 如果为RDS或GaussDB等云上数据源，可以通过管理控制台获取访问地址： 1. 根据注册的账号登录对应云服务的管理控制台。 2. 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IP、域名和端口等信息。 **说明** 仅GaussDB数据源支持多域名的方式，多个域名之间用“,”分隔。 ● 如果为MySQL、PostgreSQL或达梦数据库 DM等线下数据源，可以通过数据库管理员获取相应的访问地址。
端口	是	关系型数据库数据源的访问端口。 ● 如果为RDS或GaussDB等云上数据源，可以通过管理控制台获取访问地址： 1. 根据注册的账号登录对应云服务的管理控制台。 2. 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IP、域名和端口等信息。 **说明** 仅GaussDB数据源支持多域名的方式，多个域名之间用“,”分隔。 ● 如果为MySQL、PostgreSQL或达梦数据库 DM等线下数据源，可以通过数据库管理员获取相应的访问地址。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	RDS类型数据源为非全托管服务，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类型的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**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RDS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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
SSL加密	否	支持对RDS服务启用SSL加密传输。
数据源驱动配置		
驱动程序名称	是	驱动程序名称： • com.mysql.jdbc.Driver：连接RDS for MySQL或MySQL数据源时，选择此驱动程序名称。 • org.postgresql.Driver：连接RDS for PostgreSQL或PostgreSQL数据源时，选择此驱动程序名称。 • com.microsoft.sqlserver.jdbc.SQLServerDriver: 连接RDS for SQL Server数据源时，选择此驱动名称。 • dm.jdbc.driver.DmDriver：连接达梦数据库 DM数据源时，选择此驱动程序名称。 • com.huawei.opengauss.jdbc.Driver：连接GaussDB数据源时，选择此驱动程序名称。
驱动文件来源	是	选择驱动文件的来源方式。

参数	是否必选	说明
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://mvnrepository.com/artifact/org.postgresql/postgresql**，建议42.3.4版本。 • SQL Server驱动：获取地址**https://learn.microsoft.com/zh-cn/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver16**，建议8.4.1版本。 • 达梦数据库驱动：DM JDBC驱动jar包请从DM安装目录/dmdbms/drivers/jdbc中获取DmJdbcDriver18.jar。 • GaussDB驱动：请在**GaussDB官方文档**中搜索“JDBC包、驱动类和环境类”，然后选择实例对应版本的资料，参考文档获取驱动包。 **说明** • 驱动文件所在的OBS路径中不能包含中文。 • 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。
数据源认证及其他功能配置		
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。

4.3.15 ORACLE 数据连接参数说明

表 4-18 Oracle 数据连接

参数	是否必选	说明
数据连接类型	是	ORACLE连接固定选择为ORACLE。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** • 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 • 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
IP或域名	是	待连接的数据库的访问地址，可填写为IP或域名，其中公网IP和内网IP地址均支持。
端口	是	待连接的数据库端口。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
绑定Agent	是	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，请参考**创建CDM集群**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Oracle连接。
数据源认证及其他功能配置		
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。

参数	是否必选	说明
数据库连接类型	是	选择所需的连接方式。 ● SID： SID即Oracle数据库实例ID。一个实例只能对应一个数据库，但是一个数据库可以由多个实例对应。 ● Service Name： Service Name参数是由oracle8i开始引进的，即Oracle数据库对外服务名，标识整个数据库。
SID	是	“Connection type”配置为“SID”时，为必选项。 SID即Oracle数据库实例ID。一个实例只能对应一个数据库，但是一个数据库可以由多个实例对应。
Service Name	是	“Connection type”配置为“Service Name”时，为必选项。 Service Name参数是由oracle8i开始引进的，即Oracle数据库对外服务名，标识整个数据库。

4.3.16 DIS 数据连接参数说明

表 4-19 DIS 连接

参数	是否必选	说明
数据连接类型	是	DIS连接固定选择为DIS。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。
基础与网络连通配置		
目标项目ID	是	使用DIS Client节点发送消息至目标DIS通道时，目标通道所在的项目ID。
目标Region	是	使用DIS Client节点发送消息至目标DIS通道时，目标通道所在的Region。

参数	是否必选	说明
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
数据源认证及其他功能配置		
访问标识(AK)	是	使用DIS Client节点发送消息至目标DIS通道时，创建目标通道的租户AK。
密钥(SK)	是	使用DIS Client节点发送消息至目标DIS通道时，创建目标通道的租户SK。
描述	否	支持添加该连接的相关描述。

4.3.17 主机连接参数说明

表 4-20 主机连接

参数	是否必选	说明
数据连接类型	是	主机连接固定选择为主机连接。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。
基础与网络连通配置		
主机地址	是	Linux操作系统主机的IP地址。 请参考**查看云服务器详细信息**获取。

参数	是否必选	说明
绑定Agent	是	选择CDM集群，CDM集群提供Agent。如果没有可用的CDM集群，请参考**创建CDM集群**进行创建。 **说明** • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。 • 在调度Shell、Python脚本时，Agent会访问ECS主机，如果Shell、Python脚本的调度频率很高，ECS主机会将Agent的内网IP加入黑名单。为了保障作业的正常调度，强烈建议您使用ECS主机的**root**用户将绑定Agent（即CDM集群）的内网IP加到**/etc/hosts.allow**文件里面。 CDM集群的内网IP获取方式请参见**查看并修改CDM集群配置**。
端口	是	主机的SSH端口号。 Linux操作系统主机的默认登录端口为22，如有修改可通过主机路径“/etc/ssh/sshd_config”文件中的port字段确认端口号。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
数据源认证及其他功能配置		
用户名	是	主机的登录用户名。
登录方式	是	选择主机的登录方式： • 密钥对 • 密码
密钥对	是	“登录方式”为“密钥对”时，显示该配置项。 主机的登录方式为密钥对时，您需要获取并上传其私钥文件至OBS，在此处选择对应的OBS路径（OBS路径中不能存在中文字符）。 **说明** 此处上传的私钥文件应和主机上配置的公钥是一个密钥对，详情请参见**密钥对使用场景介绍**。
密钥对密码	是	如果密钥对未设置密码，则不需要填写该配置项。
密码	是	“登录方式”为“密码”时，显示该配置项。 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。

须知

- Shell或Python脚本可以在该ECS主机上运行的最大并发数由ECS主机的**/etc/ssh/sshd_config**文件中MaxSessions的配置值确定。请根据Shell或Python脚本的调度频率合理配置MaxSessions的值。
- 连接主机的用户需要具有主机**/tmp**目录下文件的创建与执行权限。
- Shell和Python脚本都是发往ECS主机的**/tmp**目录下去运行的，需要确保**/tmp**目录磁盘不被占满。

4.3.18 Rest Client 数据连接参数说明

表 4-21 Rest Client 连接

参数	是否必选	说明
数据连接类型	是	Rest Client连接固定选择为Rest Client。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** • 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 • 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
连接地址前缀	是	适用组件勾选数据集成时显示该参数。 连接地址前缀，测试以及作业时调用接口，会自动拼接此前缀。https仅支持TLSv1.2协议。 例如: https://xxx.com/prefix。
默认Header参数	是	适用组件勾选数据集成时显示该参数。 默认header参数，作业时调用接口都会携带此header。例如：{"Content-Type":"application/json"}
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中已创建的密钥。

参数	是否必选	说明
绑定Agent	是	适用组件勾选数据集成时显示该参数。 DataArts Studio无法直接与非全托管服务进行连接，需要提供DataArts Studio与非全托管服务通信的代理。CDM集群可以提供通信代理服务，请选择一个CDM集群，如果没有可用的CDM集群，请参考**创建CDM集群**进行创建。
数据源认证及其他功能配置		
认证方式	是	认证方法。包括： ● NONE：无认证。 ● BASIC_AUTH：基础验证。 如果数据源API支持用户名和密码的方式进行验证，您可选择此种验证方式，并在选择完成后配置用于验证的用户名和密码，后续数据集成过程中对接数据源时，通过Basic Auth协议传递给RESTful地址，完成验证。格式：{"Authorization":"Basic base64(username:password)"} ● TOKEN_AUTH：Token验证（token为静态token，永不过期，否则token过期会导致作业失败）。 如果数据源API支持Token的方式进行验证，您可选择此种验证方式，并在选择完成后配置用于验证的固定Token值，后续数据集成过程中对接数据源时，通过传入header中进行验证，格式：{"Authorization":"Bearer <token>"}。 ● OAUTH_CODE_GRANT Oauth 2.0（Authorization Code）：Oauth2.0认证。 Oauth2.0授权码模式，使用账号密码换取accessToken，再使用获取的accessToken访问接口。
用户名	否	认证方式为BASIC_AUTH模式时显示该参数。 可以通过#username获取该值，放到body、header中传递。
密码	否	认证方式为BASIC_AUTH模式时显示该参数。 可以通过#password获取该值，放到body、header中传递。
Token	否	认证方式为TOKEN_AUTH模式时显示该参数。 可以通过#token获取该值，放到body、header中传递。
认证地址	否	认证方式为OAUTH_CODE_GRANT模式时显示该参数。 Oauth 2.0模式认证地址，该接口支持Oauth2.0，使用认证凭据换取令牌，在进行测试连接以及作业前会调用此接口获取令牌，并且在【认证令牌】中定义该令牌在后续接口中携带的位置、名称、和取值方式。 例如: https://xxx.com/auth/token

参数	是否必选	说明
认证请求方法	否	认证方式为OAUTH_CODE_GRANT模式时显示该参数。 Oauth 2.0模式认证请求方法，GET/POST。在填写了认证地址的情况下，必填。 例如：GET
认证账号	否	认证方式为OAUTH_CODE_GRANT模式时显示该参数。 Oauth 2.0模式需要填写账号，可以用#authUsername获取此参数，填写到authHeader参数或者authbody参数中。
认证密码	否	认证方式为OAUTH_CODE_GRANT模式时显示该参数。 Oauth 2.0模式需要填写密码，可以用#authPassword获取此参数，填写到authHeader参数或者authbody参数中
认证请求header	否	认证方式为OAUTH_CODE_GRANT模式时显示该参数。 Oauth 2.0模式请求header，支持通过#authUsername、#authPassword获取认证账号和认证密码。 例如：{"username": "#authUsername","password": "#authPassword","Content-Type":"application/json"}
认证请求body	否	认证方式为OAUTH_CODE_GRANT模式时显示该参数。 Oauth 2.0模式请求body，get请求不支持此参数，可以支持通过#authUsername、#authPassword获取认证账号和认证密码。 例如：{"username": "#authUsername","password": "#authPassword"}

参数	是否必选	说明

认证令牌	否	认证方式为OAUTH_CODE_GRANT模式时显示该参数。
		认证令牌，可以从认证接口响应体中获取token，并在测试连接以及作业时携带，仅支持放到header中。此参数定义了参数名称（name）、参数值（value），参数值支持spel表达式。
		例如：
		认证响应体为：
		{
		"code" : 200,
		"data" : {
		"access_token" : "DSFSDFWE87WE9089W9EW9ER898WER9W89ER8",
		"expired":1000
		}
		}
		如果我们要获取access_token的值，并且满足Bearer <token>的格式，则填写格式为：
		NAME： Authenrization
		VALUE：'Bearer ' + #response.data.access_token

认证令牌有效时间	否	认证方式为OAUTH_CODE_GRANT模式时显示该参数。 认证令牌有效时间，单位s，支持el表达式，0代表永久有效，默认为0。 例1：300，有效时间为300秒。 例2：#response.data.expired，从认证接口返回的json中获取expired属性的值，支持int类型，默认单位为秒，如果不是此格式，请手动输入有效时间。

4.3.19 Redis 数据连接参数说明

表 4-22 Redis 数据连接

参数	是否必选	说明
数据连接类型	是	Redis连接固定选择为Redis。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100个字符。

参数	是否必选	说明
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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** ● 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 ● 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
手动	否	通过代理连接的时候，此项可配置，通过勾选按钮来选择集群名模式或连接串模式。 ● 使用集群名模式时通过选择填写集群名称进行连接配置。 ● 使用连接串模式填写对应集群的IP和端口进行连接配置。

参数	是否必选	说明
MRS集群名	是	选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。
服务器列表	是	手动参数为连接串模式时显示该参数。 一个或多个通过逗号分割的服务器列表（服务器域名或IP地址:服务器端口）。 例如: 192.168.0.1:27017,192.168.0.2:27017
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
绑定Agent	是	CDM集群提供了DataArts Studio与Redis通信的代理。创建Redis类型的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**进行创建。
数据源认证及其他功能配置		

参数	是否必选	说明
认证类型	是	手动参数选择连接串模式时的必选项。 选择数据库的认证类型。 包括SIMPLE类型、KERBEROS类型。
密码	是	数据库的访问密码，创建集群的时候，输入的密码。

4.3.20 SAP HANA 数据连接参数说明

表 4-23 SAP HANA 数据连接

参数	是否必选	说明
数据连接类型	是	RDS(SAP HANA)连接固定选择为RDS(SAP HANA)。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** • 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 • 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		

参数	是否必选	说明
IP或域名	是	关系型数据库数据源的访问地址，可填写为IP或域名。 ● 如果为RDS或GaussDB等云上数据源，可以通过管理控制台获取访问地址： 1. 根据注册的账号登录对应云服务的管理控制台。 2. 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IP、域名和端口等信息。 **说明** 仅GaussDB数据源支持多域名的方式，多个域名之间用“,”分隔。 ● 如果为MySQL、PostgreSQL或达梦数据库 DM等线下数据源，可以通过数据库管理员获取相应的访问地址。
端口	是	关系型数据库数据源的访问端口。 ● 如果为RDS或GaussDB等云上数据源，可以通过管理控制台获取访问地址： 1. 根据注册的账号登录对应云服务的管理控制台。 2. 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IP、域名和端口等信息。 **说明** 仅GaussDB数据源支持多域名的方式，多个域名之间用“,”分隔。 ● 如果为MySQL、PostgreSQL或达梦数据库 DM等线下数据源，可以通过数据库管理员获取相应的访问地址。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
绑定Agent	是	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SAP HANA类型的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**进行创建。
数据源驱动配置		

参数	是否必选	说明
驱动程序名称	是	驱动程序名称： • com.mysql.jdbc.Driver：连接RDS for MySQL或MySQL数据源时，选择此驱动程序名称。 • org.postgresql.Driver：连接RDS for PostgreSQL或PostgreSQL数据源时，选择此驱动程序名称。 • com.microsoft.sqlserver.jdbc.SQLServerDriver: 连接RDS for SQL Server数据源时，选择此驱动名称。 • dm.jdbc.driver.DmDriver：连接达梦数据库 DM数据源时，选择此驱动程序名称。 • com.huawei.opengauss.jdbc.Driver：连接GaussDB数据源时，选择此驱动程序名称。
驱动文件来源	是	选择驱动文件的来源方式。
驱动文件路径	是	“驱动文件来源”选择“OBS路径”时配置。 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://mvnrepository.com/artifact/org.postgresql/postgresql**，建议42.3.4版本。 • SQL Server驱动：获取地址**https://learn.microsoft.com/zh-cn/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver16**，建议8.4.1版本。 • 达梦数据库驱动： DM JDBC驱动jar包请从DM安装目录/dmdbms/drivers/jdbc中获取DmJdbcDriver18.jar。 • GaussDB驱动：请在**GaussDB官方文档**《GaussDB 用户指南》中搜索“JDBC包、驱动类和环境类”，然后选择实例对应版本的资料，参考文档获取驱动包。 **说明** • 驱动文件所在的OBS路径中不能包含中文。 • 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。
驱动文件	是	“驱动文件来源”选择“本地文件”时配置。 驱动文件请根据驱动类型去相关官网上下载，并在选择弹窗中上传驱动，或在该弹窗中指定已上传的驱动文件。
数据源认证及其他功能配置		
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。

4.3.21 LTS 数据连接参数说明

表 4-24 LTS 数据连接

参数	是否必选	说明
数据连接类型	是	LTS连接固定选择为LTS。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** • 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 • 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
项目ID	是	适用组件勾选数据集成后，呈现此参数。 DLI服务所在区域的项目ID。 项目ID表示租户的资源，账号ID对应当前账号，IAM用户ID对应当前用户。用户可在对应页面下查看不同Region对应的项目ID、账号ID和用户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“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	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LTS数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**进行创建。
数据源认证及其他功能配置		
访问标识（AK）	是	OBS服务访问标识（AK）。 例如：HCXUET8G37MWF。
密钥（SK）	否	OBS服务访问标识对应的密钥（SK）。

4.4 配置 DataArts Studio 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。

资源导入可以基于OBS服务，也支持从本地导入。支持迁移的资源包含如下业务数据：

- 管理中心组件中创建的数据连接。
- 数据集成组件中创建的CDM作业，包含作业中的CDM连接。
- 数据开发组件中已提交版本的脚本和作业。导出作业时默认只导出作业，不包含其依赖的脚本和资源。
- 数据架构组件中创建的主题、流程、码表、数据标准、关系建模模型、维度、业务指标、原子指标、衍生指标、复合指标和汇总表，不包含事实表。
- 数据目录组件中创建的元数据采集任务，以及定义的元数据分类和标签。
- 数据服务组件中发布的API。

约束条件

- 对于数据目录组件中名称相同的元数据采集任务、元数据分类和标签，不支持被重复迁移。
- 待导入的资源应为通过导出获取的zip文件，导入时系统会进行资源校验。
- 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。
- 仅企业版支持数据目录（分类、标签、采集任务）导出，专家版暂不支持。
- 导入文件时，OBS和本地方式均限制文件大小不超过10MB。

导出资源

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

步骤3 在管理中心页面，单击“资源迁移”，进入资源迁移页面。

图 4-3 资源迁移

步骤4 单击“导出文件”，配置文件的OBS存储位置和文件名称。

图 4-4 选择导出文件

步骤5 单击“下一步”，勾选导出的模块。

图 4-5 勾选导出的模块

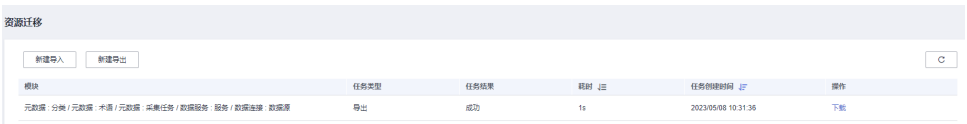
步骤6 单击“下一步”，等待导出完成，资源包导出到所设置的OBS存储位置。

图 4-6 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。

步骤7 导出完成后可在资源迁移任务列表中，单击对应任务的“下载”按钮，本地获取导出的资源包。

图 4-7 下载导出结果



----结束

导入资源

步骤1 在管理中心页面，单击“资源迁移”，进入资源迁移页面。

图 4-8 资源迁移

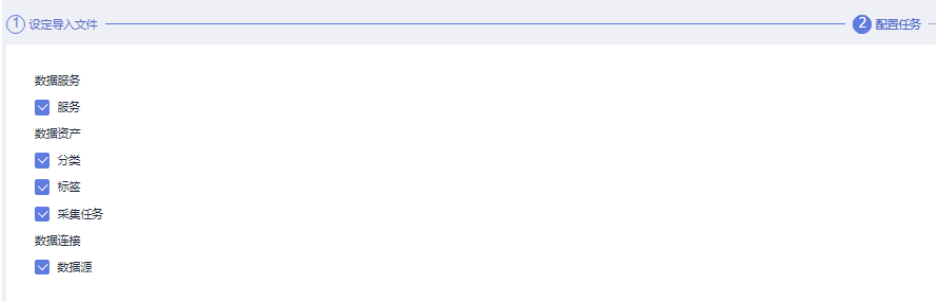
步骤2 单击“导入文件”，选择导入方式后，配置待导入资源的OBS或本地路径。待导入的资源应为通过导出获取的zip文件。

图 4-9 配置待导入的资源存储路径

步骤3 单击“导入文件”，上传待导入资源。待导入的资源应为通过导出获取的zip文件

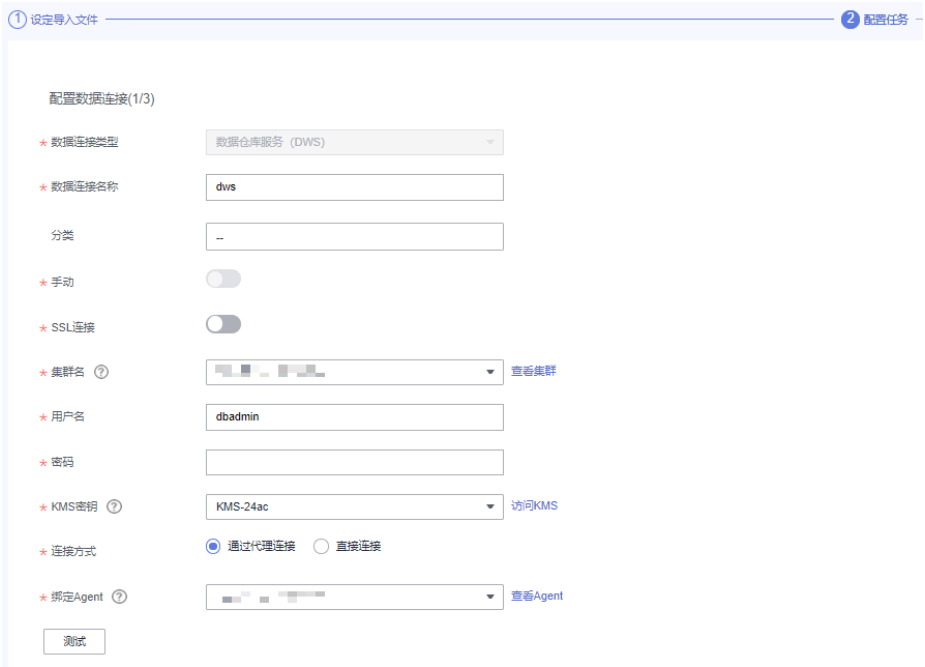
步骤4 单击“下一步”，勾选导入的资源类型。

图 4-10 勾选导入的资源类型



步骤5 如果选择导入数据源，则单击“下一步”需要配置数据连接。

图 4-11 配置数据连接

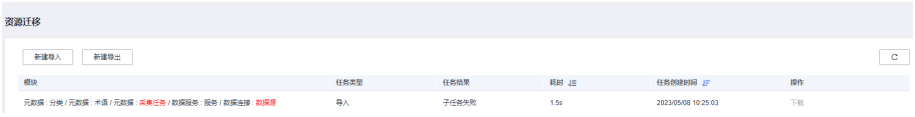


步骤6 单击“下一步”，等待导入任务下发，导入任务成功下发后系统提示“导入开始”。

图 4-12 导入开始

步骤7 系统提示“导入开始”后，单击“确定”，可在资源迁移任务列表中查看导入结果。其中存在子任务失败时，可单击红色子任务名，查看失败原因。

图 4-13 查看导入结果



----结束

4.5 配置 DataArts Studio 企业模式环境隔离

- 管理中心的环境隔离，当前支持配置DLI和DB配置的开发、生产环境隔离。
- 配置环境隔离后，数据开发时脚本/作业中的开发环境数据连接通过发布流程后，将自动切换对应生产环境的数据连接。

前提条件

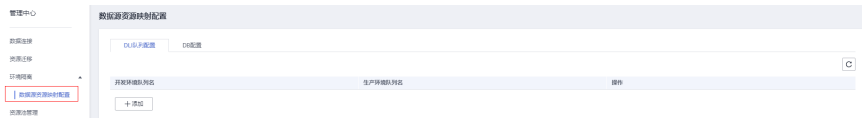
- 创建DLI环境隔离前，应已创建DLI的**数据连接**。

（可选）创建 DLI 环境隔离

仅Serverless服务（当前即DLI）需要配置环境隔离。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 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。
3. 在管理中心页面，单击“数据源资源映射配置”，进入数据源资源映射配置页面。

图 4-14 数据源资源映射配置



4. 单击“DB配置”下的“添加”，然后分别配置开发环境数据库名和生产环境数据库名，完成后单击“保存”。通过  和  可以进行编辑和删除操作。

数据库名需配置为已创建完成的数据库名。建议在创建数据库时，开发环境数据库名和生产环境数据库名保持一致，开发环境数据库名带上“_dev”后缀，以与生产环境数据库名进行区分。

图 4-15 DB 配置

5. 单击“DLI队列配置”下的“添加”，然后分别配置开发环境队列名和生产环境队列名，完成后单击“保存”。通过  和  可以进行编辑和删除操作。

队列名需配置为已在DLI创建完成的队列名。建议开发环境队列名和生产环境队列名保持一致，开发环境队列名带上“_dev”后缀，以与生产环境队列名进行区分。

图 4-16 DLI 队列配置

6. “DB配置”和“DLI队列配置”完成后，DLI环境隔离创建成功。

DB 配置

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。
2. 在管理中心页面，单击“数据源资源映射配置”，进入数据源资源映射配置页面。
3. 单击“DB配置”下的“添加”，然后分别配置开发环境数据库名和生产环境数据库名，完成后单击“保存”。通过  和  可以进行编辑和删除操作。

   数据库名需配置为已创建完成的数据库名。建议在创建数据库时，开发环境数据库名和生产环境数据库名保持一致，开发环境数据库名带上“_dev”后缀，以与生产环境数据库名进行区分。

须知

对于DWS、MRS Hive和MRS Spark这三种数据源，如果在创建数据连接时**选择同一个集群**，则需要配置数据源资源映射的DB数据库映射关系进行开发生产环境隔离。

图 4-17 DB 配置

4.6 管理中心典型场景教程

4.6.1 新建 DataArts Studio 与 MRS Hive 数据湖的连接

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等MRS类型的数据连接前，需确保您已购买MRS集群，集群的“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”，并且集群中包含所需要的组件。

- 在创建数据连接前，请确保您已具备连接所需的Agent代理（即CDM集群，如果无可用CDM集群请参考**创建CDM集群**进行创建），且待连接的数据湖与CDM集群之间网络互通。
  - 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。请确保数据源所在的主机和CDM集群均能访问公网，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：
    - 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。
    - 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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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。
- 如果使用企业模式，您还需要注意以下事项：

  由于企业模式下需要区分开发环境和生产环境，因此您需要分别准备对应生产环境和开发环境的两套数据湖服务，用于隔离开发和生产环境：
  - 对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS），**如果使用两套集群**，DataArts Studio通过管理中心的创建数据连接区分开发环境和生产环境的数据湖服务，在开发和生产流程中自动切换对应的数据湖。因此您需要准备两套数据湖服务，且两套数据湖服务的版本、规格、组件、区域、VPC、子网以及相关配置等信息，均应保持一致。创建数据连接的详细操作请参见**创建DataArts Studio数据连接**。
  - 对于Serverless服务（例如DLI），DataArts Studio通过管理中心的环境隔离来配置生产环境和开发环境数据湖服务的对应关系，在开发和生产流程中自动切换对应的数据湖。因此您需要在Serverless数据湖服务中准备两套队列、数据库资源，建议通过名称后缀进行区分，详细操作请参见**配置DataArts Studio企业模式环境隔离**。
  - 对于DWS、MRS Hive和MRS Spark这三种数据源，如果在创建数据连接时**选择同一个集群**，则需要配置数据源资源映射的DB数据库映射关系进行开发生产环境隔离，详细操作请参见**DB配置**。
  - 离线处理集成作业不支持在企业模式下运行。

  例如，当您的数据湖服务为MRS集群时，需要准备两套MRS集群，且版本、规格、组件、区域、VPC、子网等保持一致。如果某个MRS集群修改了某些配置，也需要同步到另一套MRS集群上。

创建数据连接

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

步骤3 在管理中心页面，单击“数据连接”，进入数据连接页面并单击“创建数据连接”。

图 4-18 创建数据连接

步骤4 单击“创建数据连接”，在弹出的页面中，选择“数据连接类型”为“MapReduce服务（MRS Hive）”，并参见**表4-25**配置相关参数。

图 4-19 MRS Hive 连接配置参数

表 4-25 MRS Hive 数据连接

参数	是否必选	说明
数据连接类型	是	MRS Hive连接固定选择为MapReduce服务（MRS Hive）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** ● 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 ● 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
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. MRS API连接仅支持在数据开发组件使用，其他组件例如数据架构、数据质量、数据目录等无法使用此连接。 2. 在数据开发组件不支持通过可视化方式查看与管理该连接下的数据库、数据表和字段。特别的，仅当连接MRS 3.2.1以及之后版本的MRS集群时，支持通过可视化方式查看数据库、数据表和字段，但仍不支持可视化方式管理。 3. 在数据开发组件的SQL编辑器运行SQL时，只能以日志形式显示执行结果。 **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”。

参数	是否必选	说明
手动	是	通过代理连接时，是必选项。 选择连接模式。如无访问其他项目或企业项目下MRS集群的需求，使用集群名模式即可。 ● 使用集群名模式时，通过选择已有集群名称进行连接配置。仅可选择本项目内且企业项目相同的MRS集群进行连接。 ● 使用连接串模式时，通过手动输入Manager IP，并打通本连接Agent（即CDM集群）和MRS集群之间的网络，则可以访问其他项目或企业项目的MRS集群。
Manager IP	是	使用连接串模式时，是必选项。 此参数填写为MRS Manager的浮动IP地址。仅支持连接MRS云服务，自建Hadoop集群必须先纳管到MRS云服务才能连接。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 注意，通过输入框后的“选择”按钮仅能获取本项目内且企业项目相同的MRS集群，如果需要访问其他项目或企业项目的MRS集群，则需要获取MRS Manager的浮动IP地址并手动输入，并确保已打通本连接Agent（即CDM集群）和MRS租户面集群之间的网络。Manager的浮动IP地址可通过登录MRS集群主Master节点获取，执行ifconfig命令，回显中eth0:wsom的IP就是MRS Manager的浮动IP。登录MRS集群Master节点请参见**登录集群节点**章节，如果登录的是非主Master节点无法查询，请切换到另一个Master节点查询。 手动填写IP时请根据场景和顺序填写，多个IP之间使用","分隔。例如: 127.0.0.1或127.0.0.1,127.0.0.2,127.0.0.3。 ● 填写单个IP，IP应为MRS集群管理面的浮动IP。 ● 填写3个IP时，应填写MRS集群业务面的主节点IP、备节点IP和MRS集群管理面的浮动IP。

参数	是否必选	说明
MRS集群名	是	通过MRS API连接或使用集群名模式时，是必选项。 选择所属的MRS集群。仅支持连接MRS云服务，自建Hadoop集群必须在纳管到MRS云服务后才可以选择。系统会显示所有项目ID和企业项目相同的MRS集群。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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** 当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。
KMS密钥	否	通过代理连接时，是必选项。 通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。

参数	是否必选	说明
绑定Agent	是	通过代理连接时，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，请参考**创建CDM集群**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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** • 对于多个开启Kerberos认证的MRS集群，如果在创建数据连接时使用同一个CDM集群作为Agent，则会导致作业运行失败。建议您按照业务情况规划多个CDM集群。 • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
数据源认证及其他功能配置		
认证类型	是	使用连接串模式时，是必选项。 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。
用户名	是	MRS集群的人机用户，通过代理连接时是必选项。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**创建一个新的密码永不过期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。 • 建议用户名的密码策略设置为永不过期，避免由于密码过期导致连接失败，引起业务受损。

参数	是否必选	说明
密码	是	MRS集群的访问密码，通过代理连接的时候，是必选项。
开启ldap	否	当“连接方式”参数选择为“通过代理连接”时，显示该配置项。 当MRS Hive对接外部LDAP开启了LDAP认证时，连接Hive时需要使用LDAP账号与密码进行认证，此时必须开启此参数，否则会连接失败。
ldap用户名	是	当“开启ldap”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的用户名。
ldap密码	是	当“开启ldap”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的密码。
元数据采集范围	否	配置元数据实时同步的数据库和数据表范围，不填写默认不筛选。 可填写为如下两种形式之一： ● database_name：筛选数据库名包含“database_name”的数据库 ● database_name.table_name：筛选数据库名包含“database_name”的数据库，在匹配到的数据库中再匹配表名包含“table_name”的数据表 例如： ● 填写为“datatest”，则元数据实时同步将同步数据库名包含“datatest”的数据库中的数据表。 ● 填写为“datatest.table1”，则元数据实时同步将同步如下数据表：数据库名包含“datatest”的数据库，其中表名包含“table_name”的数据表。
OBS支持	否	适用组件勾选数据集成后，呈现此参数。 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。
使用委托	否	适用组件勾选数据集成后，呈现此参数。 开启委托功能，即可以在无需持有永久AKSK的情况下创建数据连接，根据DLF配置的调度身份执行CDM作业。
公共委托	否	适用组件勾选数据集成且“使用委托”选择“是”时，呈现此参数。 仅涉及用于测试该连接委托功能是否正常，作业运行将根据DLF配置的调度身份执行CDM作业。

参数	是否必选	说明
访问标识(AK)	-	适用组件勾选数据集成且“OBS支持”选择“是”时，呈现此参数。 AK和SK分别为登录OBS服务器的访问标识与密钥。 您需要先创建当前账号的访问密钥，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4-20所示。 **图 4-20** 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)	-	

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”，创建数据连接。

----结束

参考

1. 为什么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集群时所选择的企业项目与工作空间的企业项目不同。
   – 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。

4.6.2 新建 DataArts Studio 与 DWS 数据湖的连接

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等MRS类型的数据连接前，需确保您已购买MRS集群，集群的“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”，并且集群中包含所需要的组件。
- 在创建数据连接前，请确保您已具备连接所需的Agent代理（即CDM集群，如果无可用CDM集群请参考**创建CDM集群**进行创建），且待连接的数据湖与CDM集群之间网络互通。
  - 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。请确保数据源所在的主机和CDM集群均能访问公网，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：
    - 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。
    - 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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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。
- 如果使用企业模式，您还需要注意以下事项：

  由于企业模式下需要区分开发环境和生产环境，因此您需要分别准备对应生产环境和开发环境的两套数据湖服务，用于隔离开发和生产环境：
  - 对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS），**如果使用两套集群**，DataArts Studio通过管理中心的创建数据连接区分开发环境和生产环境的数据湖服务，在开发和生产流程中自动切换对应的数据湖。因此您需要准备两套数据湖服务，且两套数据湖服务的版本、规格、组件、区域、VPC、子网以及相关配置等信息，均应保持一致。创建数据连接的详细操作请参见**创建DataArts Studio数据连接**。
  - 对于Serverless服务（例如DLI），DataArts Studio通过管理中心的环境隔离来配置生产环境和开发环境数据湖服务的对应关系，在开发和生产流程中自动切换对应的数据湖。因此您需要在Serverless数据湖服务中准备两套队列、数据库资源，建议通过名称后缀进行区分，详细操作请参见**配置DataArts Studio企业模式环境隔离**。
  - 对于DWS、MRS Hive和MRS Spark这三种数据源，如果在创建数据连接时**选择同一个集群**，则需要配置数据源资源映射的DB数据库映射关系进行开发生产环境隔离，详细操作请参见**DB配置**。
  - 离线处理集成作业不支持在企业模式下运行。

例如，当您的数据湖服务为MRS集群时，需要准备两套MRS集群，且版本、规格、组件、区域、VPC、子网等保持一致。如果某个MRS集群修改了某些配置，也需要同步到另一套MRS集群上。

创建数据连接

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

步骤3 在管理中心页面，单击“数据连接”，进入数据连接页面并单击“创建数据连接”。

图 4-21 创建数据连接

步骤4 单击“创建数据连接”，在弹出的页面中，选择“数据连接类型”为“数据仓库服务（DWS）”，并参见**表4-26**配置相关参数。

图 4-22 DWS 连接配置参数

表 4-26 DWS 数据连接

参数	是否必选	说明
数据连接类型	是	DWS连接固定选择为数据仓库服务（DWS）。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** • 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 • 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
SSL加密	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方式进行通信。 • 开关关闭，SSL通道加密和证书认证两种方式均可进行通信。
手动	是	选择连接模式。 • 使用集群名模式时，通过选择已有集群名称进行连接配置。 • 使用连接串模式时，手动填写对应集群的IP或域名、端口进行连接配置，且需打通本连接Agent（即CDM集群）和DWS集群之间的网络。 **说明** 数据安全组件不支持连接串模式的DWS连接。
DWS集群名	是	“手动”选择为“集群名模式”时需要配置本参数。 选择DWS集群，系统会显示所有项目ID和企业项目相同的DWS集群。
IP或域名	是	“手动”选择为“连接串模式”时需要配置本参数。 “IP或域名”如果手动填写，必须写内网IP，端口必须为对资源组网段放开的端口，否则可能导致网络连接不通。 表示通过内部网络访问集群数据库的访问地址，可填写为IP或域名。内网访问IP或域名地址在创建集群时自动生成，您可以通过管理控制台获取访问地址： 1. 根据注册的账号登录DWS云服务管理控制台。 2. 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IP、域名和端口等信息。
端口	是	“手动”选择为“连接串模式”时需要配置本参数。 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。

参数	是否必选	说明
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
绑定Agent	是	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，请参考**创建CDM集群**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DWS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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
数据源认证及其他功能配置		
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。
元数据采集范围	否	配置元数据实时同步的数据库和数据表范围，不填写默认不筛选。 可填写为如下两种形式之一： • database_name：筛选数据库名包含“database_name”的数据库 • database_name.table_name：筛选数据库名包含“database_name”的数据库，在匹配到的数据库中再匹配表名包含“table_name”的数据表 例如： • 填写为“datatest”，则元数据实时同步将同步数据库名包含“datatest”的数据库中的数据表。 • 填写为“datatest.table1”，则元数据实时同步将同步如下数据表：数据库名包含“datatest”的数据库，其中表名包含“table_name”的数据表。

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”，创建数据连接。

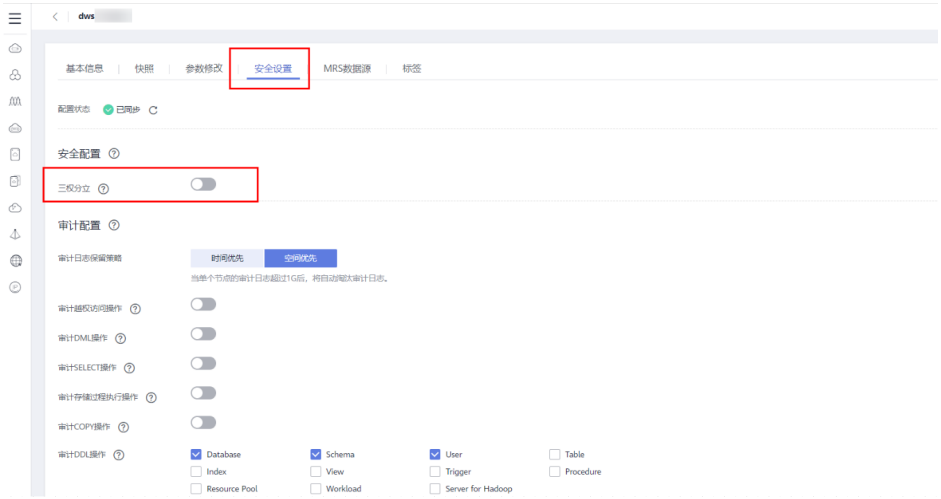
----结束

参考

1. 创建DWS数据连接，开启SSL连接时测试连接失败?

   请在DWS控制台，单击进入对应的DWS集群后，选择“安全设置”，然后关闭三权分立功能。

   图 4-23 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。

4.6.3 新建 DataArts Studio 与 MySQL 数据库的连接

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等MRS类型的数据连接前，需确保您已购买MRS集群，集群的“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”，并且集群中包含所需要的组件。
- 在创建数据连接前，请确保您已具备连接所需的Agent代理（即CDM集群，如果无可用CDM集群请参考**创建CDM集群**进行创建），且待连接的数据湖与CDM集群之间网络互通。
  - 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。请确保数据源所在的主机和CDM集群均能访问公网，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：
    - 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。

    - 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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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。

- 如果使用企业模式，您还需要注意以下事项：

  由于企业模式下需要区分开发环境和生产环境，因此您需要分别准备对应生产环境和开发环境的两套数据湖服务，用于隔离开发和生产环境：

  - 对于集群化的数据源（例如MRS、DWS、RDS、MySQL、Oracle、DIS、ECS），**如果使用两套集群**，DataArts Studio通过管理中心的创建数据连接区分开发环境和生产环境的数据湖服务，在开发和生产流程中自动切换对应的数据湖。因此您需要准备两套数据湖服务，且两套数据湖服务的版本、规格、组件、区域、VPC、子网以及相关配置等信息，均应保持一致。创建数据连接的详细操作请参见**创建DataArts Studio数据连接**。
  - 对于Serverless服务（例如DLI），DataArts Studio通过管理中心的环境隔离来配置生产环境和开发环境数据湖服务的对应关系，在开发和生产流程中自动切换对应的数据湖。因此您需要在Serverless数据湖服务中准备两套队列、数据库资源，建议通过名称后缀进行区分，详细操作请参见**配置DataArts Studio企业模式环境隔离**。
  - 对于DWS、MRS Hive和MRS Spark这三种数据源，如果在创建数据连接时**选择同一个集群**，则需要配置数据源资源映射的DB数据库映射关系进行开发生产环境隔离，详细操作请参见**DB配置**。
  - 离线处理集成作业不支持在企业模式下运行。

  例如，当您的数据湖服务为MRS集群时，需要准备两套MRS集群，且版本、规格、组件、区域、VPC、子网等保持一致。如果某个MRS集群修改了某些配置，也需要同步到另一套MRS集群上。

创建数据连接

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

步骤3 在管理中心页面，单击“数据连接”，进入数据连接页面并单击“创建数据连接”。

图 4-24 创建数据连接

步骤4 单击“创建数据连接”，在弹出的页面中，选择“数据连接类型”为“RDS”，并参见表4-27配置相关参数。

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 4-25 RDS 连接配置参数

表 4-27 RDS 数据连接

参数	是否必选	说明
数据连接类型	是	RDS连接固定选择为RDS。

参数	是否必选	说明
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。 **说明** • 当开启数据集成作业特性后，可勾选数据集成组件，勾选后在数据开发组件创建集成作业时支持选择本数据连接。 • 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
基础与网络连通配置		
IP或域名	是	关系型数据库数据源的访问地址，可填写为IP或域名。 “IP或域名”如果手动填写，必须写内网IP，端口必须为对资源组网段放开的端口，否则可能导致网络连接不通。 • 如果为RDS或GaussDB等云上数据源，可以通过管理控制台获取访问地址： 1. 根据注册的账号登录对应云服务的管理控制台。 2. 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IP、域名和端口等信息。 **说明** 仅GaussDB数据源支持多域名的方式，多个域名之间用“,”分隔。 • 如果为MySQL、PostgreSQL或达梦数据库 DM等线下数据源，可以通过数据库管理员获取相应的访问地址。

参数	是否必选	说明
端口	是	关系型数据库数据源的访问端口。 ● 如果为RDS或GaussDB等云上数据源，可以通过管理控制台获取访问地址： 1. 根据注册的账号登录对应云服务的管理控制台。 2. 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IP、域名和端口等信息。 **说明** 仅GaussDB数据源支持多域名的方式，多个域名之间用“,”分隔。 ● 如果为MySQL、PostgreSQL或达梦数据库 DM等线下数据源，可以通过数据库管理员获取相应的访问地址。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
绑定Agent	是	RDS类型数据源为非全托管服务，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类型的数据连接时，请选择一个CDM集群。如果没有可用的CDM集群，请参考**创建CDM集群**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RDS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，且使用同一个VPC和子网，安全组规则需允许两者网络互通。 **说明** CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。
SSL加密	否	支持对RDS服务启用SSL加密传输。
数据源驱动配置		

参数	是否必选	说明
驱动程序名称	是	驱动程序名称： • com.mysql.jdbc.Driver：连接RDS for MySQL或MySQL数据源时，选择此驱动程序名称。 • org.postgresql.Driver：连接RDS for PostgreSQL或PostgreSQL数据源时，选择此驱动程序名称。 • com.microsoft.sqlserver.jdbc.SQLServerDriver: 连接RDS for SQL Server数据源时，选择此驱动名称。 • dm.jdbc.driver.DmDriver：连接达梦数据库 DM数据源时，选择此驱动程序名称。 • com.huawei.opengauss.jdbc.Driver：连接GaussDB数据源时，选择此驱动程序名称。
驱动文件来源	是	选择驱动文件的来源方式。
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://mvnrepository.com/artifact/org.postgresql/postgresql**，建议42.3.4版本。 • SQL Server驱动：获取地址**https://learn.microsoft.com/zh-cn/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver16**，建议8.4.1版本。 • 达梦数据库驱动： DM JDBC驱动jar包请从DM安装目录/dmdbms/drivers/jdbc中获取DmJdbcDriver18.jar。 • GaussDB驱动：请在**GaussDB官方文档**中搜索“JDBC包、驱动类和环境类”，然后选择实例对应版本的资料，参考文档获取驱动包。 **说明** • 驱动文件所在的OBS路径中不能包含中文。 • 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。
数据源认证及其他功能配置		
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。

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”，创建数据连接。

----结束

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。

5 数据集成（CDM 作业）

5.1 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”。

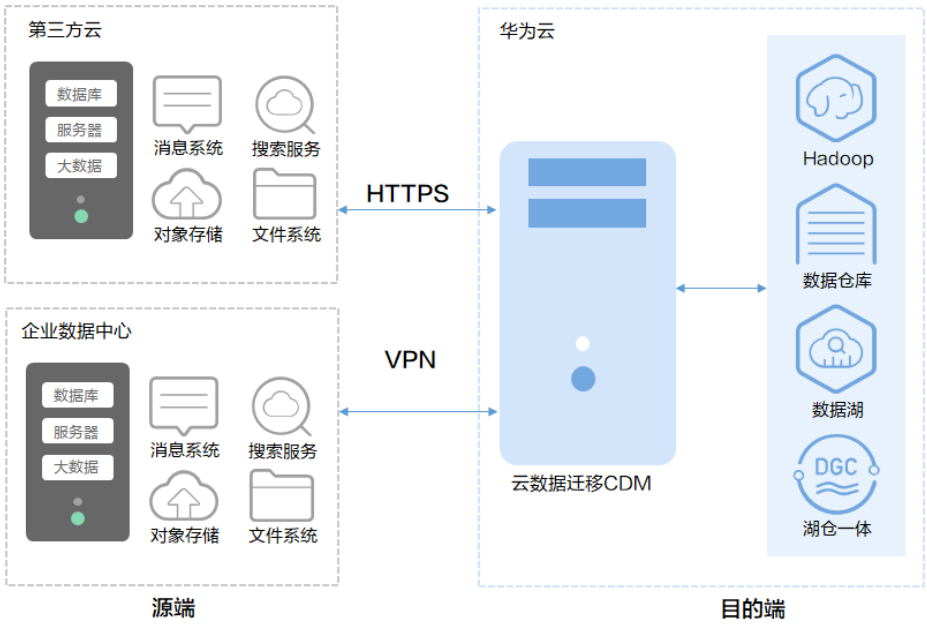
您可以通过以下方式之一进入CDM主界面：

- 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。

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 5-1 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。

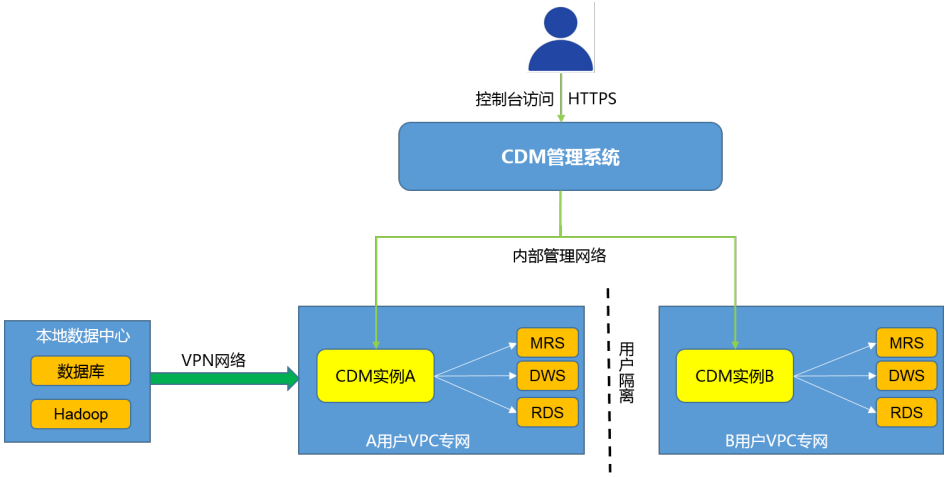
CDM 迁移原理

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 5-2 CDM 迁移原理



5.2 约束与限制

CDM 系统级限制和约束

1. DataArts Studio实例赠送的数据集成集群，由于规格限制，仅用于测试业务、数据连接代理场景。
2. 用于运行数据迁移作业的其他规格CDM集群可以在DataArts Studio控制台以增量包的形式购买，也可以在云数据迁移CDM服务控制台直接购买。二者差异体现在如下方面：
   a. 套餐计费：在DataArts Studio控制台购买的CDM集群，套餐计费时仅支持在DataArts Studio控制台购买的套餐包；在CDM控制台购买的CDM集群，套餐计费仅支持在云数据迁移CDM服务控制台购买的折扣套餐。
   b. 权限控制：在DataArts Studio控制台购买的CDM集群，按照DataArts Studio的权限体系进行权限管理；在CDM控制台购买的CDM集群，按照云数据迁移CDM服务的权限体系进行权限管理。
   c. 使用场景：在DataArts Studio控制台购买的CDM集群按工作空间隔离，需要在关联的工作空间使用；在CDM控制台购买的CDM集群，不支持DataArts Studio工作空间级别的资源隔离，所有DataArts Studio工作空间均可使用。
3. 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。
4. CDM集群为ARM或X86版本，依赖于底层资源的架构。
5. 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。
6. 在迁移数据时CDM会对数据源端产生压力。建议创建新的数据库账号用于数据迁移，并配置账号策略用于以限制迁移对数据源端的资源消耗。例如可配置当CPU使用率超30%就清理该账号下的连接，从而避免影响业务。

7. 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。
8. 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。
9. 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。
10. 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。
11. 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。
12. 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。
13. 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。
14. 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。
15. 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。
16. 作业运行过程中，任务异常中断，目标端已写入的部分数据不会清理，需手动清理。
17. 单文件传输大小不超过1TB。

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，推荐使用数据复制服务（DRS）来实现数据库增量迁移到RDS。
2. CDM支持的数据库整库迁移，仅支持数据表迁移，不支持存储过程、触发器、函数、视图等数据库对象迁移。

   CDM仅适用于一次性将数据库迁移到云上的场景，包括同构数据库迁移和异构数据库迁移，不适合数据同步场景，比如容灾、实时同步。
3. CDM迁移数据库整库或数据表失败时，已经导入到目标表中的数据不会自动回滚，对于需要事务模式迁移的用户，可以配置“先导入到阶段表”参数，实现迁移失败时数据回滚。

   极端情况下，可能存在创建的阶段表或临时表无法自动删除，也需要用户手工清理（阶段表的表名以“_cdm_stage”结尾，例如：cdmtet_cdm_stage）。
4. CDM访问用户本地数据中心数据源时（例如本地自建的MySQL数据库），需要用户的数据源可支持Internet公网访问，并为CDM集群实例绑定弹性IP。这种方式下安全实践是：本地数据源通过防火墙或安全策略仅允许CDM弹性IP访问。

5. 仅支持常用的数据类型，字符串、数字、日期，对象类型有限支持，如果对象过大会出现无法迁移的问题。
6. 仅支持数据库字符集为GBK和UTF-8。
7. 字段名不可包含&和%。
8. jdbc2hive，hive2jdbc整库迁移的实现机制就是按字段名称映射的，不支持字段名称不一致的迁移场景。

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。

推荐使用**云专线服务**，解决网络访问的同时，还可以提升迁移速度。

数据仓库服务(DWS)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

   说明

   2.9.1.200及之前版本的集群，DWS源端暂不支持NVARCHAR2数据类型。
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。
- 目的端为DLI数据源时，抽取并发数建议配置为1，否则可能会导致写入失败。

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. 第三方云的Redis服务无法支持作为源端。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**字段转换**功能转换为上述格式。

数据接入服务（DIS）和 Kafka 数据源约束

- 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。
- 设置为长久运行的任务，如果DIS系统发生中断，任务也会失败结束。
- 迁移作业源端为MRS Kafka时，字段映射不支持自定义字段。
- 迁移作业源端为DMS kafka时，字段映射支持自定义字段。

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中使用Parquet格式存储时间戳数据时，时间戳的精度为纳秒级别（即精确到毫微秒），即2023-03-27 00:00:00.000。当源端数据精度大于纳秒级别时，字段映射时会对数据进行截取。例如源端数据为2023-03-27 00:00:00.12345，目的端数据会被截取为2023-03-27 00:00:00.123。
- 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;
```

5.3 支持的数据源

5.3.1 支持的数据源（2.10.0.300）

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**。

说明

本文介绍2.10.0.300版本CDM集群所支持的数据源。因各版本集群支持的数据源有所差异，其他版本支持的数据源仅做参考。

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5-1所示。

表 5-1 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），MRS ClickHouse，Doris • 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 • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	不支持DWS物理机纳管模式。
	数据湖探索（DLI）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS ClickHouse，Doris • 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：MongoDB • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	MongoDB建议使用的版本：4.2。 用户需要具备DLI数据源所有字段的“查询表”权限，即SELECT权限。

数据源分类	源端数据源	对应的目的端数据源	说明
	MRS ClickHouse	数据仓库：MRS ClickHouse，数据湖探索（DLI）	• MRS ClickHouse建议使用的版本：21.3.4.X。 • 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
	Doris	数据仓库：Doris	当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。

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	• 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 • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	• 支持本地存储，仅MRS Hive、MRS Hudi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、MRS Hudi暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X • 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
	MRS Hive	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS Clickhouse、Doris • 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 • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable），SAP HANA	
	MRS Hudi	• 数据仓库：数据仓库服务（DWS） • Hadoop：MRS HBase	

数据源分类	源端数据源	对应的目的端数据源	说明
	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） • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	• 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 HBase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） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	• 对象存储服务之间的迁移，推荐使用对象存储迁移服务OMS。 • 不支持二进制文件导入到数据库或NoSQL。

数据源分类	源端数据源	对应的目的端数据源	说明
文件系统	FTP SFTP	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 搜索：Elasticsearch • 对象存储：对象存储服务（OBS） • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	• 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。 • FTP/SFTP到OBS的迁移仅支持二进制文件。 • HTTP到OBS的迁移推荐使用obsutil工具，请参见**obsutil简介**。
	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），Doris • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 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 • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable），SAP HANA	• OLTP数据库之间的迁移推荐通过数据复制服务DRS进行迁移。 • Microsoft SQL Server建议使用的版本：2005以上。 • 金仓和GaussDB数据源可通过PostgreSQL连接器进行连接，支持的迁移作业的源端、目的端情况与PostgreSQL数据源一致。
	云数据库SQL Server 云数据库PostgreSQL	• 数据仓库：数据仓库服务（DWS） • 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 • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	

数据源分类	源端数据源	对应的目的端数据源	说明
	MySQL PostgreSQL Oracle	• 数据仓库：数据仓库服务（DWS） • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 对象存储：对象存储服务（OBS） • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	
	Microsoft SQL Server	• 数据仓库：数据仓库服务（DWS） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	
NoSQL	分布式缓存服务（DCS） Redis MongoDB	• Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS）	NoSQL数据源不支持作为目的端。 Redis到DCS的迁移，可以通过其他方式进行，请参见**自建Redis迁移至DCS**。
消息系统	数据接入服务（DIS） Apache Kafka DMS Kafka	公测中：云搜索服务（CSS）	消息系统不支持作为目的端。

数据源分类	源端数据源	对应的目的端数据源	说明
	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 • 搜索：Elasticsearch， • 公测中：表格存储服务（CloudTable），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。 • 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
搜索	Elasticsearch	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • 搜索：Elasticsearch • 公测中：表格存储服务（CloudTable），云搜索服务（CSS）	Elasticsearch仅支持非安全模式。

数据源分类	源端数据源	对应的目的端数据源	说明
公测中	表格存储服务（CloudTable HBase）	• 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 • 搜索：Elasticsearch • 公测中：表格存储服务（CloudTable），云搜索服务（CSS）	-
	云搜索服务（CSS）	• 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 • 搜索：Elasticsearch • 公测中：表格存储服务（CloudTable），云搜索服务（CSS）	导入数据到CSS推荐使用Logstash，请参见**使用Logstash导入数据到Elasticsearch**。

数据源分类	源端数据源	对应的目的端数据源	说明
	SAP HANA	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS Hive	SAP HANA数据源存在如下约束： • SAP HANA不支持作为目的端。 • 仅支持2.00.050.00.1592305219版本。 • 仅支持Generic Edition。 • 不支持BW/4 FOR HANA。 • 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 • 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 • 迁移时不支持目的端自动建表。

数据源分类	源端数据源	对应的目的端数据源	说明
	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） • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	• 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
	分库	• 分库数据仓库：数据湖探索（DLI） • Hadoop：MRS HBase，MRS Hive • 搜索：Elasticsearch • 对象存储：对象存储服务（OBS） • 公测中：云搜索服务（CSS）	分库数据源不支持作为目的端。
	达梦数据库DM	• 数据仓库：数据仓库服务（DWS） • Hadoop：MRS Hive，MRS Hudi	-
	神通（ST）	Hadoop：MRS Hive，MRS Hudi	-
	文档数据库服务（DDS）	Hadoop：MRS HDFS，MRS HBase，MRS Hive	-

数据源分类	源端数据源	对应的目的端数据源	说明
	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） • 搜索：Elasticsearch • 公测中：云搜索服务（CSS），表格存储服务（CloudTable）	-
	GBASE8S	• Hadoop：MRS HDFS，MRS HBase • 消息系统：MRS Kafka	-
	GBASE8A	• Hadoop：MRS HDFS，MRS Hive，MRS HBase • 消息系统：MRS Kafka	-

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5-2所示。

表 5-2 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	-

数据源分类	数据源	读取	写入	说明
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 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
	MRS Hive	支持	支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
	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
	MRS Hudi	支持	支持	支持本地存储、存算分离场景。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
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	

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

数据源分类	数据源	读取	写入	说明
	文档数据库服务（DDS）	支持	支持	仅支持DDS和MRS之间迁移。

5.3.2 支持的数据源（2.9.3.300）

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**。

说明

本文介绍2.9.3.300版本CDM集群所支持的数据源。因各版本集群支持的数据源有所差异，其他版本支持的数据源仅做参考。

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5-3**所示。

表 5-3 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），MRS ClickHouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	不支持DWS物理机纳管模式。

数据源分类	源端数据源	对应的目的端数据源	说明
	数据湖探索（DLI）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS ClickHouse • 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），MongoDB • 搜索：Elasticsearch，云搜索服务（CSS）	MongoDB建议使用的版本：4.2。 用户需要具备DLI数据源所有字段的“查询表”权限，即SELECT权限。
	MRS ClickHouse	数据仓库：MRS ClickHouse，数据湖探索（DLI）	• MRS ClickHouse建议使用的版本：21.3.4.X。 • 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。

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	• 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、MRS Hudi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、MRS Hudi暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X • 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
	MRS Hive	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS Clickhouse • 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）	
	MRS Hudi	• 数据仓库：数据仓库服务（DWS） • Hadoop：MRS HBase	

数据源分类	源端数据源	对应的目的端数据源	说明
	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

数据源分类	源端数据源	对应的目的端数据源	说明
	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 HBase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） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 对象存储服务之间的迁移，推荐使用对象存储迁移服务OMS。 • 不支持二进制文件导入到数据库或NoSQL。

数据源分类	源端数据源	对应的目的端数据源	说明
文件系统	FTP SFTP	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS） • 对象存储：对象存储服务（OBS）	• 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。 • FTP/SFTP到OBS的迁移仅支持二进制文件。 • HTTP到OBS的迁移推荐使用obsutil工具，请参见**obsutil简介**。
	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，MRS Hudi • 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）	• Microsoft SQL Server建议使用的版本：2005以上。 • 金仓和GaussDB数据源可通过PostgreSQL连接器进行连接，支持的迁移作业的源端、目的端情况与PostgreSQL数据源一致。
	云数据库SQL Server 云数据库PostgreSQL	• 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）	

数据源分类	源端数据源	对应的目的端数据源	说明
	MySQL PostgreSQL Oracle	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 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）	
	Microsoft SQL Server	• 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）	

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

数据源分类	源端数据源	对应的目的端数据源	说明
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到DCS的迁移，可以通过其他方式进行，请参见**自建Redis迁移至DCS**。
	Redis		
	文档数据库服务（DDS）		
	MongoDB		
	表格存储服务（CloudTable HBase）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	Cassandra	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
消息系统	数据接入服务（DIS）	搜索：云搜索服务（CSS）	消息系统不支持作为目的端。
	Apache Kafka		
	DMS Kafka		

数据源分类	源端数据源	对应的目的端数据源	说明
	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。 • 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
搜索	Elasticsearch	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	Elasticsearch仅支持非安全模式。
	云搜索服务（CSS）		导入数据到CSS推荐使用Logstash，请参见**使用Logstash导入数据到Elasticsearch**。

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5-4所示。

表 5-4 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	-
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 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。

数据源分类	数据源	读取	写入	说明
	MRS Hive	支持	支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
	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

数据源分类	数据源	读取	写入	说明
	MRS Hudi	支持	支持	支持本地存储、存算分离场景。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
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	

数据源分类	数据源	读取	写入	说明
	SAP HANA	支持	不支持	• 仅支持2.00.050.00.1592305219版本。 • 仅支持Generic Edition。 • 不支持BW/4 FOR HANA。 • 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 • 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 • 迁移时不支持目的端自动建表。
	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	-

5.3.3 支持的数据源（2.9.2.200）

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**。

说明

本文介绍2.9.2.200版本CDM集群所支持的数据源。因各版本集群支持的数据源有所差异，其他版本支持的数据源仅做参考。

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5-5所示。

表 5-5 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），MRS ClickHouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	不支持DWS物理机纳管模式。
	数据湖探索（DLI）		用户需要具备DLI数据源所有字段的“查询表”权限，即SELECT权限。
	MRS ClickHouse	数据仓库：MRS ClickHouse，数据湖探索（DLI）	• MRS ClickHouse建议使用的版本：21.3.4.X。 • 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。

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	• 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、MRS Hudi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、MRS Hudi暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X • 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
	MRS Hive	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS Clickhouse • 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）	
	MRS Hudi	数据仓库：数据仓库服务（DWS）	

数据源分类	源端数据源	对应的目的端数据源	说明
	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

数据源分类	源端数据源	对应的目的端数据源	说明
	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 HBase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） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 对象存储服务之间的迁移，推荐使用对象存储迁移服务OMS。 • 不支持二进制文件导入到数据库或NoSQL。

数据源分类	源端数据源	对应的目的端数据源	说明
文件系统	FTP SFTP	• 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。 • HTTP到OBS的迁移推荐使用obsutil工具，请参见**obsutil简介**。
	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，MRS Hudi • 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）	• Microsoft SQL Server建议使用的版本：2005以上。 • 金仓和GaussDB数据源可通过PostgreSQL连接器进行连接，支持的迁移作业的源端、目的端情况与PostgreSQL数据源一致。
	云数据库SQL Server 云数据库PostgreSQL	• 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）	

数据源分类	源端数据源	对应的目的端数据源	说明
	MySQL PostgreSQL Oracle	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 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）	
	Microsoft SQL Server	• 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）	

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

数据源分类	源端数据源	对应的目的端数据源	说明
	表格存储服务（CloudTable HBase）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	Cassandra	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
消息系统	数据接入服务（DIS）	搜索：云搜索服务（CSS）	消息系统不支持作为目的端。
	Apache Kafka		
	DMS Kafka		

数据源分类	源端数据源	对应的目的端数据源	说明
	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。 • 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
搜索	Elasticsearch	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	Elasticsearch仅支持非安全模式。
	云搜索服务（CSS）		导入数据到CSS推荐使用Logstash，请参见**使用Logstash导入数据到Elasticsearch**。

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5-6**所示。

表 5-6 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	-
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 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。

数据源分类	数据源	读取	写入	说明
	MRS Hive	支持	支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
	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

数据源分类	数据源	读取	写入	说明
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	Redis	支持	支持	-
	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	-

5.3.4 支持的数据类型

配置字段映射时，数据源支持的数据类型请参见**表5-7**，以确保数据完整导入到目的端。

表 5-7 支持的数据类型

数据连接类型	数据类型说明
MySQL	请参见**MySQL数据库迁移时支持的数据类型**。
SQL Server	请参见**SQL Server数据库迁移时支持的数据类型**。
Oracle	请参见**Oracle数据库迁移时支持的数据类型**。
PostgreSQL	请参见**PostgreSQL数据库迁移时支持的数据类型**。
神通（ST）	请参见**神通（ST）数据库迁移时支持的数据类型**。
SAP HANA	请参见**SAP HANA数据库迁移时支持的数据类型**。
DWS	请参见**DWS数据库迁移时支持的数据类型**。
达梦	请参见**达梦数据库迁移时支持的数据类型**。
DLI	请参见**DLI数据库迁移时支持的数据类型**。
Elasticsearch/云搜索服务（CSS）	请参见**Elasticsearch/云搜索服务（CSS）数据库迁移时支持的数据类型**。

MySQL 数据库迁移时支持的数据类型

源端为MySQL数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 5-8 开源 MySQL 数据库作为源端时支持的数据类型

类别	类型	简要释义	存储格式示例	Hive	DWS
字符串	CHAR（M）	固定长度的字符串是以长度为1到255之间的字符长度（例如：CHAR（5）），存储右空格填充到指定的长度。 限定长度不是必需的，它会默认为1。	‘a’或‘aaaaa’	CHAR	CHAR

类别	类型	简要释义	存储格式示例	Hive	DWS
	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92233720368547758 07。如果无符号，允许的范围是从0到184467440737095516 15. 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

类别	类型	简要释义	存储格式示例	Hive	DWS
日期时间	DATE	以YYYY-MM-DD格式的日期，在1000-01-01和9999-12-31之间。 例如，1973年12月30日将被存储为1973-12-30。	1999-10-01	DATE	TIMESTAMP
	TIME	用于存储时、分、秒信息。	'09:10:21'或'9:10:21'	不支持（String）	TIME
	DATETIME	日期和时间组合以YYYY-MM-DD HH:MM:SS格式，在1000-01-01 00:00:00到9999-12-31 23:59:59之间。例如，1973年12月30日下午3:30，会被存储为1973-12-30 15:30:00。	'1973-12-30 15:30:00'	TIMESTAMP	TIMESTAMP
	TIMESTAMP	1970年1月1日午夜之间的时间戳，到2037的某个时候。这看起来像前面的DATETIME格式，无需只是数字之间的连字符; 1973年12月30日下午3点30分将被存储为19731230153000（YYYYMMDDHHMMSS）。	19731230153000	TIMESTAMP	TIMESTAMP
	YEAR（M）	以2位或4位数字格式来存储年份。如果长度指定为2（例如YEAR（2）），年份就可以为1970至2069（70~69）。如果长度指定为4，年份范围是1901-2155，默认长度为4。	2000	不支持（String）	不支持
多媒体（二进制）	BINARY（M）	字节数为M，允许长度为0-M的变长二进制字符串，字节数为值的长度加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
	TINYTEXT	0-255字节短文本二进制字符串。	-	-	不支持
	MEDIUMTEXT	0-167772154字节中等长度文本二进制字符串。	-	-	不支持
	LONGTEXT	0-4294967295字节极大长度文本二进制字符串。	-	-	不支持
	BLOB	字段的最大长度是65535个字符。 BLOB是“二进制大对象”，并用来存储大的二进制数据，如图像或其他类型的文件。BLOB大小写敏感。	0x5236（二进制数据）	不支持	不支持
	TINYBLOB	0-255字节短文本二进制字符串。	-	不支持	不支持
	MEDIUMBLOB	0-167772154字节中等长度文本二进制字符串。	-	不支持	不支持
	LONGBLOB	0-4294967295字节极大长度文本二进制字符串。	0x5236（二进制数据）	不支持	不支持
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 5-9 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、月中的哪一天、小时、分钟、秒。	DATE	TIMESTAMP

类别	类型	简要释义	Hive	DWS
	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。	String	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，用户可以通过它来读取这个文件。	不支持	不支持
其他类型	rowid	实际上是数据库表中行的地址，它有10字节长。	不支持	不支持
	urowid	是一个通用的rowid，没有固定的rowid的表。	不支持	不支持

SQL Server 数据库迁移时支持的数据类型

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

表 5-10 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值。	DOUBLE	FLOAT4	NUMBER
	float	其中为用于存储float数值尾数的位数（以科学计数法表示），因此可以确定精度和存储大小。	FLOAT	FLOAT8	binary_float

类别	类型	简要释义	Hive	DWS	Oracle
	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
	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数据，其中n是从1~8000的值。	不支持（String）	不支持（String）	不支持
	varbinary	可变长度二进制数据。	不支持（String）	不支持（String）	不支持
货币数据类型	money	用于存储货币值。	不支持（String）	不支持（String）	不支持

类别	类型	简要释义	Hive	DWS	Oracle
	smallmoney	与money类型相似，输入数据时在前面加上一个货币符号，如美元为$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

PostgreSQL 数据库迁移时支持的数据类型

源端为PostgreSQL数据库，目的端为Hive、DWS、DLI时，支持的数据类型如下：

表 5-11 PostgreSQL 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	DLI
字符	char	定长字符串，存储右空格填充到指定的长度。	CHAR	CHAR	不支持（String）
	varchar	变长字符串，不会用空格将字段或变量填充至最大长度。	CARCHAR	CARCHAR	不支持（String）

类别	类型	简要释义	Hive	DWS	DLI
数值	smallint	拓展名 int2，存储在2个字节中，它允许的范围是从-32768到32767。	SMALLINT	SMALLINT	SMALLINT
	int	拓展名 int4，存储在4个字节中，它允许的范围是从-2147483648到2147483647。	INTEGER	INT	INT
	bigint	拓展名 int8，存储在8个字节中，允许范围为-9223372036854775808到9223372036854775807。	BIGINT	BIGINT	BIGINT
	decimal（p，s）	精度p表示为值存储的有效位数，刻度s表示可以在小数点后存储的位数。p最大位数是1000。	DECIMAL（P，S）	DECIMAL（P，S）	DECIMAL（P，S）
	float	4字节或8字节存储。float（n）：n取值在1-24内，精度有效位数为6 位数，长度4 个字节，是单精度，n取值在25-53内，精度有效位数为15 位数，长度8 字节，是双精度。	FLOAT/DOUBLE	FLOAT/DOUBLE	FLOAT/DOUBLE
	smallserial	序列数据类型，以smallint格式存储。	SMALLINT	SMALLINT	SMALLINT
	serial	序列数据类型，以int格式存储。	INTEGER	INT	INT
	bigserial	序列数据类型，以bigint格式存储。	BIGINT	BIGINT	BIGINT
日期时间	date	存储日期数据。	DATE	DATE	DATE
	timestamp	存储日期和时间数据，无时区。	TIMESTAMP	TIMESTAMP	不支持（String）
	timestamptz	存储日期和时间数据，有时区。	TIMESTAMP	TIMESTAMPZ	不支持（String）

类别	类型	简要释义	Hive	DWS	DLI
	time	只用于一日内时间，无时区。	不支持（String）	TIME	不支持（String）
	timez	只用于一日内时间，有时区。	不支持（String）	TIMEZ	不支持（String）
	interval	时间间隔。	不支持（String）	不支持（String）	不支持（String）
位串类型	bit	定长位串，例如：b'000101'。	不支持（String）	不支持（String）	不支持（String）
	varbit	可变长位串，例如：b'101'。	不支持（String）	不支持（String）	不支持（String）
货币类型	money	存储在8个字节中，它允许的范围是从-92233720368547 7.5808到922337203685477.5807。	DOUBLE	MONEY	DECIMAL（P，S）
布尔类型	boolean	存储在1个字节中，可以取值为 1、0 或 NULL。	BOOLEAN	BOOLEAN	BOOLEAN
文本类型	text	变长文本，无长度限制。	不支持（String）	不支持（String）	不支持（String）

DWS 数据库迁移时支持的数据类型

源端为DWS数据库时，支持的数据类型如下：

表 5-12 DWS 数据库作为源端时支持的数据类型

类别	类型	简要释义
字符	char	定长字符串，存储右空格填充到指定的长度。
	varchar	变长字符串，不会用空格将字段或变量填充至最大长度。
数值	double	用于存储指明双精度的浮点数。
	decimal（p，s）	精度p表示为值存储的有效位数，刻度s表示可以在小数点后存储的位数。p最大位数是1000。

类别	类型	简要释义
	numeric	用于存储零、正负定点数。
	real	与double相同。
	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。
	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。
	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。
	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。
日期时间	date	存储日期数据。
	timestamp	存储日期和时间数据，无时区。
	time	只用于一日内时间，无时区。
位串类型	bit	定长位串，例如： b'000101'。
布尔类型	boolean	存储在1个字节中，可以取值为 1、0 或 NULL。
文本类型	text	变长文本，无长度限制。

神通（ST）数据库迁移时支持的数据类型

源端为神通（ST）数据库，目的端为MRS Hive、MRS Hudi时，支持的数据类型如下：

表 5-13 神通（ST）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	MRS Hive	MRS Hudi
字符	VARCHAR	用于存储指定定长字符串。	‘a’或‘aaaaa’	VARCHAR（765）	STRING
	BPCHAR	用于存储指定变长字符串。	‘a’或‘aaaaa’	VARCHAR（765）	STRING

类别	类型	简要释义	存储格式示例	MRS Hive	MRS Hudi
数值	NUMERIC	用于存储零、正负定点数。	52.36	DECIMAL（10，0）	DECIMAL（18，0）
	INT	用于存储零、正负定点数。	5236	INT	INT
	BIGINT	用于存储有符号整数，精度为19，标度为0。	5236	BIGINT	BIGINT
	TINYINT	用于存储有符号整数，精度为3，标度为0。	100	SMALLINT	INT
	BINARY	用于存储定长二进制数据。	0x2A3B4058	不支持	FLOAT
	VARBINARY	用于存储可变长二进制数据。	0x2A3B4058	不支持	BINARY
	FLOAT	用于存储带二进制精度的浮点数。	52.36	FLOAT	FLOAT
	DOUBLE	用于存储指明双精度的浮点数。	52.3	DOUBLE	DOUBLE
日期时间	DATE	用于存储年、月、日信息。	'1999-10-01'、'1999/10/01'或'1999.10.01'	DATE	DATE
	TIME	用于存储时、分、秒信息。	'09:10:21'或'9:10:21'	STRING	STRING
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。	'2002-12-12 09:10:21'、'2002-12-12 9:10:21'、'2002/12/12 09:10:21' 或'2002.12.12 09:10:21'	TIMESTAMP	TIMESTAMP
多媒体	CLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。	0x5236（二进制数据）	STRING	STRING
	BLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。	0x5236（二进制数据）	不支持	BINARY

类别	类型	简要释义	存储格式示例	MRS Hive	MRS Hudi
布尔类型	BOOLEAN	存储在1个字节中，可以取值为 1、0 或 NULL。	1	BOOLEAN	BOOLEAN

SAP HANA 数据库迁移时支持的数据类型

源端为SAP HANA数据库时，支持的数据类型如下：

表 5-14 SAP HANA 数据库作为源端时支持的数据类型

类别	类型	简要释义
字符	VARCHAR	用于存储指定定长字符串。
	NVARCHAR	包含unicode格式数据的变长字符串。
	TEXT	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。
数值	BIGINT	用于存储有符号整数，精度为19，标度为0。
	TINYINT	用于存储有符号整数，精度为3，标度为0。
	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。
	REAL	可以存储正的或者负的十进制数值。
	DECIMAL	带固定精度和小数位数的数值数据类型。
	FLOAT	用于存储带二进制精度的浮点数。
	DOUBLE	用于存储指明双精度的浮点数。
日期时间	DATE	用于存储年、月、日信息。
	TIME	用于存储时、分、秒信息。
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。
多媒体	CLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。
	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。
布尔类型	BOOLEAN	存储在1个字节中，可以取值为 1、0 或 NULL。

DLI 数据库迁移时支持的数据类型

源端为DLI数据库时，支持的数据类型如下：

表 5-15 DLI 数据库作为源端时支持的数据类型

类别	类型	简要释义
字符	CHAR	用于存储指定定长字符串。
	VARCHAR	与CHAR相同。
	STRING	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。
数值	BIGINT	用于存储有符号整数，精度为19，标度为0。
	TINYINT	用于存储有符号整数，精度为3，标度为0。
	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。
	INT	用于存储有符号整数，精度为10，标度为0。
	DECIMAL	带固定精度和小数位数的数值数据类型。
	FLOAT	用于存储带二进制精度的浮点数。
	DOUBLE	用于存储指明双精度的浮点数。
日期时间	DATE	用于存储年、月、日信息。
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。
布尔类型	BOOLEAN	存储在1个字节中，可以取值为 1、0 或 NULL。

Elasticsearch/云搜索服务（CSS）数据库迁移时支持的数据类型

源端为Elasticsearch/云搜索服务（CSS）数据库时，支持的数据类型如下：

表 5-16 Elasticsearch/云搜索服务（CSS）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	MySQL
字符	keyword	用于存储字符串。	“keyword”	String
	text	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。	“long string”	TEXT

类别	类型	简要释义	存储格式示例	MySQL
	string	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。	“a string”	String
整数	short	用于存储16位有符号整数，取值范围为-32768至32767。	32765	smallInt
	integer	用于存储32位有符号整数，取值范围为-2^{31}至2^{31}-1。	3276566	int
	long	用于存储64位有符号整数，取值范围为-2^{63}至2^{63}-1。	3276566666	BIGINT
数值	double	64位双精度IEEE 754浮点类型。	21.333	double
	float	32位单精度IEEE 754浮点类型 。	21.333	double
布尔类型	boolean	存储在1个字节中，可以取值为 1、0 或 NULL。	1	Boolean
对象	object	扁平化存储对象的字符串。	{"users.name": ["John", "Smith"], users.age": [26, 28], "users.gender":[1, 2]}	TEXT
嵌套	nested	嵌套存储对象的字符串。	{"users.name" : "John" , "users.age" : 26, "users.gender" : 1} { "users.name" : "Smith", "users.age" : 28, "users.gender" : 2}	TEXT

类别	类型	简要释义	存储格式示例	MySQL
日期	date	日期格式的字符串。	“2018-01-13”或“2018-01-13 12:10:30”	DATE或timeStamp
特殊	ip	Ip地址格式的字符串。	“192.168.127.100”	String
数组	string_array	全部是字符串的数组。	[“str”，“str”]	TEXT
	short_array	全部是16位整数的数组。	[1，1，1]	TEXT
	integer_array	全部是32位整数的数组。	[1，1，1]	TEXT
	long_array	全部是64位整数的数组。	[1，1，1]	TEXT
	float_array	全部是32位浮点数的数组。	[1.0，1.0，1.0]	TEXT
	double_array	全部是64位浮点数的数组。	[1.0，1.0，1.0]	TEXT
范围	completion	自动补全的字符串。	“string”	TEXT

Doris 数据库迁移时支持的数据类型

源端为Doris数据库时，支持的数据类型如下：

表 5-17 Doris 作为源端时支持的数据类型

类别	类型	简要释义
字符串	CHAR（M）	范围：char[（length）]，定长字符串，长度length范围是1~255，默认为1。
	VARCHAR（M）	范围：char（length），变长字符串，长度length范围是1~65535。
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。

类别	类型	简要释义
数值类型	TINYINT	长度：长度为1个字节的有符号整型。 范围：[-128， 127]。
	SMALLINT	长度：长度为2个字节的有符号整型。 范围：[-32768， 32767]。
	INT	长度：长度为4个字节的有符号整型。 范围：[-2147483648， 2147483647]。
	BIGINT	长度：长度为8个字节的有符号整型。 范围：[-9223372036854775808，9223372036854775807]。
	LARGEINT	长度：长度为16个字节的有符号整型。 范围：[-2^127， 2^127-1]。
	FLOAT	长度：长度为4字节的浮点类型。 范围：-3.40E+38 ~ +3.40E+38。
	DOUBLE	长度：长度为8字节的浮点类型。 范围：-1.79E+308 ~ +1.79E+308。
	DECIMAL[M， D]	保证精度的小数类型。M代表一共有多少个有效数字，D代表小数点后最多有多少数字。M的范围是[1，27]，D的范围是[1，9]，另外，M必须要大于等于D的取值。默认取值为decimal[10，0]。 precision: 1 ~ 27。 scale: 0 ~ 9。
日期类型	DATE	范围：['1000-01-01'， '9999-12-31']。默认的打印形式是'YYYY-MM-DD'。
	DATETIME	范围：['1000-01-01 00:00:00'， '9999-12-31 00:00:00']。默认的打印形式是’YYYY-MM-DD HH:MM:SS’。
特殊类型	HLL	HLL（HyperLogLog） 类型是一个二进制类型。HLL 类型只能用于聚合类型的表（Aggregation Table），并且必须指定聚合类型为 HLL_UNION。 HLL 类型主要用于非精确快速去重场景下，对数据进行预聚合。 HLL列只能通过配套的 hll_union_agg、hll_cardinality、hll_hash 进行查询或使用。
	BITMAP	BITMAP 类型是一个二进制类型。BITMAP 类型只能用于聚合类型的表（Aggregation Table），并且必须指定聚合类型为 BITMAP_UNION。 BITMAP 类型主要用于精确去重场景下，对数据进行预聚合。同时也可以用于如用户画像场景存放用户ID等。 BITMAP 列只能通过配套的 BITMAP 函数进行查询和使用。

达梦数据库迁移时支持的数据类型

源端为达梦数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 5-18 达梦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	用于存储指定定长字符串。	‘a’或‘aaaaa’	CHAR	CHAR
	CHARACTER	与 CHAR 相同。	‘a’或‘aaaaa’	CHAR	CHAR
	VARCHAR	用于存储指定变长字符串.	‘a’或‘aaaaa’	VARCHAR	VARCHAR
	VARCHAR2	与 VARCHAR 相同。	‘a’或‘aaaaa’	VARCHAR	VARCHAR
数值	NUMERIC	用于存储零、正负定点数。	52.36	DECIMAL	NUMERIC
	DECIMAL	与 NUMERIC 相似。	52.36	DECIMAL	NUMERIC
	DEC	与 DECIMAL 相同。	52.36	DECIMAL	NUMERIC
	INTEGER	用于存储有符号整数，精度为10，标度为0。	5236	INT	INTEGER
	INT	与 INTEGER 相同。	5236	INT	INTEGER
	BIGINT	用于存储有符号整数，精度为19，标度为0。	5236	BIGINT	BIGINT
	TINYINT	用于存储有符号整数，精度为3，标度为0。	100	TINYINT	SMALLINT
	SMALLINT	用于存储有符号整数，精度为5，标度为0。	9999	SMALLINT	SMALLINT
	BYTE	与 TINYINT 相似，精度为3，标度为0。	100	TINYINT	SMALLINT
	BINARY	用于存储定长二进制数据。	0x2A3B4058	BINARY（NULL）	BYTEA（NULL）

类别	类型	简要释义	存储格式示例	Hive	DWS
	VARBINARY	用于存储可变长二进制数据。	0x2A3B4058	BINARY（NULL）	BYTEA（NULL）
	FLOAT	用于存储带二进制精度的浮点数。	52.36	FLOAT	FLOAT8
	DOUBLE	与FLOAT类似。	52.36	DOUBLE	FLOAT8
	REAL	用于存储带二进制精度的浮点数，但它不能由用户指定使用的精度。	52.3	FLOAT	FLOAT4
	DOUBLE PRECISION	用于存储指明双精度的浮点数。	52.3	DOUBLE	FLOAT8
位串	BIT	用于存储整数数据1、0 或 NULL。	1、0 或 NULL	TINYINT（1 0 NULL）	BOOLEAN（true false NULL）
日期时间	DATE	用于存储年、月、日信息。	1999-10-01'、'1999/10/01'或'1999.10.01'	DATE	TIMESTAMP
	TIME	用于存储时、分、秒信息。	'09:10:21'或'9:10:21'	不支持（String）	TIME
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。	2002-12-12 09:10:21'，'2002-12-12 9:10:21' '2002/12/12 09:10:21' 或'2002.12.12 09:10:21'	TIMESTAMP	TIMESTAMP
	TIME WITH TIME ZONE	用于存储一个带时区的 TIME 值，其定义是在 TIME 类型的后面加上时区信息。	'09:10:21+8:00'，'09:10:21+8:00'或'9:10:21+8:00'	不支持（String）	TIME WITH TIME ZONE

类别	类型	简要释义	存储格式示例	Hive	DWS
	TIMESTAMP WITH TIME ZONE	用于存储一个带时区的 TIMESTAMP 值，其定义是TIMESTAMP类型的后面加上时区信息。	2002-12-12 09:10:21 +8:00'，'2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00'或'2002.12.12 09:10:21 +8:00'	TIMESTAMP	TIMESTAMP WITH TIME ZONE
	TIMESTAMP WITH LOCAL TIME ZONE	用于存储一个本地时区的 TIMESTAMP 值，能够将标准时区类型 TIMESTAMP WITH TIME ZONE 类型转化为本地时区类型。	2002-12-12 09:10:21 +8:00'，'2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00'或'2002.12.12 09:10:21 +8:00'	不支持（String）	不支持（TEXT）
	DATETIME WITH TIME ZONE	同TIMESTAMP WITH TIME ZONE。	2002-12-12 09:10:21 +8:00'，'2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00'或'2002.12.12 09:10:21 +8:00'	TIMESTAMP	TIMESTAMP WITH TIME ZONE
	INTERVAL YEAR	描述一个若干年的间隔，引导精度规定了年的取值范围。	INTERVAL '0015' YEAR	不支持（String）	不支持（VARCHAR）
	INTERVAL YEAR TO MONTH	描述一个若干年若干月的间隔，引导精度规定了年的取值范围。	INTERVAL '0015-08' YEAR TO MONTH	不支持（String）	不支持（VARCHAR）

类别	类型	简要释义	存储格式示例	Hive	DWS
	INTERVAL MONTH	描述一个若干月的间隔，引导精度规定了月的取值范围。	INTERVAL '0015' MONTH	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY	描述一个若干日的间隔，引导精度规定了日的取值范围。	INTERVAL '150' DAY	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY TO HOUR	描述一个若干日若干小时的间隔，引导精度规定了日的取值范围。	INTERVAL '923' DAY TO HOUR	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY TO MINUTE	描述一个若干日若干小时若干分钟的间隔，引导精度规定了日的取值范围。	INTERVAL '09 23:12' DAY TO MINUTE	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY TO SECOND	描述一个若干日若干小时若干分钟若干秒的间隔，引导精度规定了日的取值范围。	INTERVAL '09 23:12:01.1' DAY TO SECOND	不支持（String）	不支持（VARCHAR）
	INTERVAL HOUR	描述一个若干小时的间隔，引导精度规定了小时的取值范围。	INTERVAL '150' HOUR	不支持（String）	不支持（VARCHAR）
	INTERVAL HOUR TO MINUTE	描述一个若干小时若干分钟的间隔，引导精度规定了小时的取值范围。	INTERVAL '23:12' HOUR TO MINUTE	不支持（String）	不支持（VARCHAR）
	INTERVAL HOUR TO SECOND	描述一个若干小时若干分钟若干秒的间隔，引导精度规定了小时的取值范围。	INTERVAL '23:12:01.1' HOUR TO SECOND	不支持（String）	不支持（VARCHAR）
	INTERVAL MINUTE	描述一个若干分钟的间隔，引导精度规定了分钟的取值范围。	INTERVAL '150' MINUTE	不支持（String）	不支持（VARCHAR）
	INTERVAL MINUTE TO SECOND	描述一个若干分钟若干秒的间隔，引导精度规定了分钟的取值范围。	INTERVAL '12:01.1' MINUTE TO SECOND	不支持（String）	不支持（VARCHAR）

类别	类型	简要释义	存储格式示例	Hive	DWS
	INTERVAL SECOND	描述一个若干秒的间隔，引导精度规定了秒整数部分的取值范围。	INTERVAL '51.1' SECOND	不支持（String）	不支持（VARCHAR）
多媒体	IMAGE	IMAGE 用于指明多媒体信息中的图像类型。 图像由不定长的像素点阵组成，长度最大为 2G-1 字节。该类型除了存储图像数据之外，还可用于存储任何其它二进制数据。	0x2A3B4058（二进制数据）	不支持	不支持
	LONGVARBINARY	与IMAGE相同。	0x2A3B4059（二进制数据）	不支持	不支持
	TEXT	用于存储长字符串类型，其字符串的长度 最大为 2G-1，存储长的文本串。	0x5236（二进制数据）	不支持	不支持
	LONGVARCHAR	与 TEXT 相似。	0x5236（二进制数据）	不支持	不支持
	BLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。	0x5236（二进制数据）	不支持	不支持
	CLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。	0x5236（二进制数据）	不支持	不支持
	BFILE	用于指明存储在操作系统中的二进制文件， 文件存储在操作系统而非数 据库中，仅能进行只读访问。	-	不支持	不支持

5.4 创建并管理 CDM 集群

5.4.1 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。

前提条件

已申请VPC、子网和安全组。CDM集群连接云上其它服务时，需确保CDM集群与待连接的云服务在同一个VPC。如果CDM集群与其它云服务所属不同VPC，则CDM集群需要通过EIP连接云服务。

说明

- 当CDM集群与其他云服务所在的区域、VPC、子网、安全组一致时，可保证CDM集群与其他云服务内网互通，无需专门打通网络。
- 当CDM集群与其他云服务所在的区域和VPC一致、但子网或安全组不一致时，需配置路由规则及安全组规则以打通网络。配置路由规则请参见**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**章节。
- 当CDM集群与其他云服务所在的区域一致、但VPC不一致时，可以通过对等连接打通网络。配置对等连接请参见**如何配置对等连接**章节。

  注：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。
- 当CDM集群与其他云服务所在的区域不一致时，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。
- 另外，如果创建了企业项目，则企业项目也会影响CDM集群与其他云服务的网络互通，只有企业项目一致的云服务才能打通网络。

操作场景

DataArts Studio实例中已经包含一个仅用于测试、试用等非正式业务场景的CDM集群。

- 如果该集群已经满足您的使用需求，则无需再购买批量数据迁移增量包。
- 如果您需要CDM集群用于满足业务需求，请通过按需计费方式购买批量数据迁移增量包，详情请参考**按需计费方式购买数据集成集群**。
- 如果您需要为购买的CDM集群匹配套餐包用于降低使用成本，请通过套餐包方式购买批量数据迁移增量包，详情请参考**套餐包方式购买数据集成集群**。

说明

DataArts Studio实例赠送的CDM集群，由于规格限制，仅用于测试、试用等非正式业务场景。用于业务场景的CDM集群可以通过“批量数据迁移增量包”进行购买，且不建议同时作为数据连接Agent代理和运行数据迁移作业使用。

5.4.2 解绑/绑定 CDM 集群的 EIP

操作场景

CDM集群创建完成后，支持解绑或绑定EIP。EIP即弹性公网IP，由虚拟私有云（Virtual Private Cloud，简称VPC）负责其计费。

如果CDM需要访问本地数据源、Internet的数据源，或者跨VPC的云服务，则必须要为CDM集群绑定一个弹性IP，或者使用NAT网关让CDM集群与其他弹性云服务器共享弹性IP访问Internet，具体操作请见**添加SNAT规则**。

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。

或参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。在DataArts Studio控制台首页，选择对应工作空间的“数据集成”模块，进入CDM首页。

图 5-3 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

5.4.3 重启 CDM 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。

须知

重启CDM集群进程或集群VM都会导致正在运行的作业失败，重启期间也无法调度新的作业，请谨慎操作！

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。

或参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。在DataArts Studio控制台首页，选择对应工作空间的“数据集成”模块，进入CDM首页。

图 5-4 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 5-5 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，选择完成并单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

5.4.4 删除 CDM 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。
- 待删除集群中所需的连接和作业数据已通过**批量管理CDM作业**中的导出作业功能进行备份。
- 对于购买DataArts Studio服务时系统赠送的CDM集群，非常不建议您进行删除操作。该集群删除后无法再次赠送，只能另外购买。
- 删除集群后，CDM集群不再按需计费或扣除套餐时长。如果您为删除的CDM集群购买了CDM折扣套餐或包年包月形式的DataArts Studio数据集成增量包，则请参考**云服务退订**章节进行套餐包退订。

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。

或参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。在DataArts Studio控制台首页，选择对应工作空间的“数据集成”模块，进入CDM首页。

图 5-6 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 通过以下两种方式进入删除集群确认界面。

- 选择集群操作列中的“更多 > 删除”。
- 选中需要删除的集群，单击删除按钮。

步骤3 输入“DELETE”后单击“确定”，即开始删除CDM集群。

图 5-7 删除集群 1

----结束

5.4.5 下载 CDM 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。

或参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。在DataArts Studio控制台首页，选择对应工作空间的“数据集成”模块，进入CDM首页。

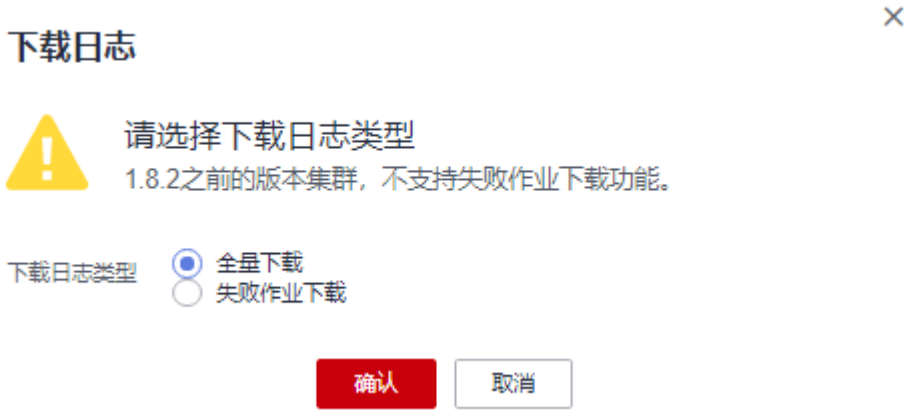
图 5-8 集群列表

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 5-9 下载日志类型



步骤3 确认后，即可下载日志到本地。

----结束

5.4.6 查看并修改 CDM 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。该功能产生的消息通知不会计入收费项。
  - 用户隔离：控制其他用户是否能够查看、操作该集群中的迁移作业和连接。
    - 开启该功能时，该集群中的迁移作业、连接会被隔离，华为账号下的其他IAM用户无法查看、操作该集群中的迁移作业和连接。

说明

按组批量启动作业会运行组内所有作业。如果开启了用户隔离功能，即使华为账号下的其他IAM用户无法查看到组内作业，按组批量启动作业依然会将组内作业运行，因此在用户隔离场景不建议使用按组批量启动作业功能。

- 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，华为账号下的所有拥有相应权限的IAM用户可以查看、操作迁移作业和连接。

  注意，用户隔离关闭后需要重启集群VM才能生效。

- 最大抽取并发数：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出限制，超出部分将排队等待。

  注意，最大抽取并发数取值范围为1-1000，建议根据集群规格进行配置，建议值详见**最大抽取并发数**。过高的并发数可能导致内存溢出，请谨慎修改。

说明

此处的“最大抽取并发数”参数与作业配置管理处的“最大抽取并发数”参数同步，在任意一处修改即可生效。

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。

或参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。在DataArts Studio控制台首页，选择对应工作空间的“数据集成”模块，进入CDM首页。

图 5-10 集群列表

说明

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

步骤2 单击集群名称，可查看集群的基本信息。

----结束

修改集群配置

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

或参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。在DataArts Studio控制台首页，选择对应工作空间的“数据集成”模块，进入CDM首页。

图 5-11 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、用户是否隔离以及最大抽取并发数的配置。

步骤3 修改完成后单击“保存”，返回集群管理界面。

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

图 5-12 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

5.4.7 管理集群标签

操作场景

CDM集群已经创建成功后，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标签系统或者CDM集群管理列表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。

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 5-13 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 5-14 修改集群配置

步骤3 单击“添加/编辑标签”，通过添加、修改标签为CDM集群设置资源标识。

图 5-15 添加标签

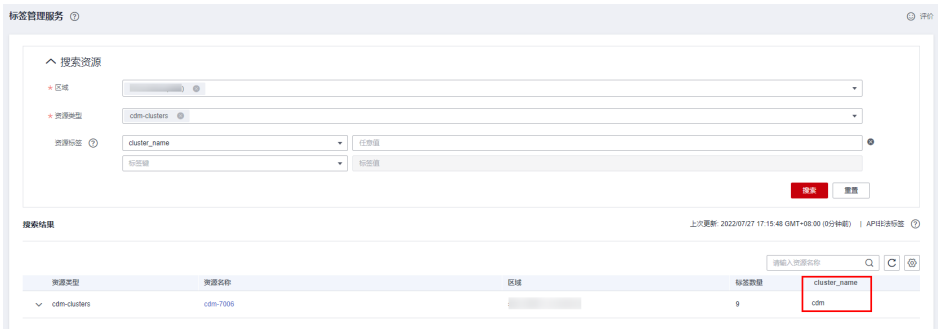
说明

- 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“删除”，删除CDM集群标签。

步骤5 通过以下两种方式筛选出所配置标签的资源。

- 在标签管理服务中，选择资源搜索条件，单击“搜索”即可筛选出所配置标签的资源。



- 在集群列表中，单击标签搜索，筛选出所配置标签的资源。

----结束

5.4.8 管理并查看 CDM 监控指标

5.4.8.1 CDM 支持的监控指标

功能说明

云监控服务（Cloud Eye）可以监控和查看云服务的运行状态、各个指标的使用情况，并对监控项创建告警规则。

当您创建了CDM集群后，云监控服务会自动关联CDM的监控指标，帮助您实时掌握CDM集群的各项性能指标，精确掌握CDM集群的运行情况。

- 本章节描述了CDM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。
- 如果您需要查看CDM相关的监控指标，请参见**查看CDM监控指标**。
- 如果您需要在监控数据满足指定条件时发送报警通知，可参见**设置CDM告警规则**。

前提条件

使用CDM监控功能，需获取CES相关权限。

命名空间

SYS.CDM

监控指标

CDM集群支持的监控指标如表5-19所示。

表 5-19 CDM 支持的监控指标

指标ID	指标名称	指标含义	取值范围	测量对象	监控周期（原始指标）
bytes_in	网络流入速率	该指标用于统计每秒流入测量对象的网络流量。 单位：字节/秒。	≥ 0 bytes/s	CDM集群实例	1分钟
bytes_out	网络流出速率	该指标用于统计每秒流出测量对象的网络流量。 单位：字节/秒。	≥ 0 bytes/s	CDM集群实例	1分钟
cpu_usage	CPU使用率	该指标用于统计测量对象的CPU使用率。 单位：%。	0%～100%	CDM集群实例	1分钟
mem_usage	内存使用率	该指标用于统计测量对象的内存使用率。 单位：%。	0%～100%	CDM集群实例	1分钟
pg_pending_job	排队作业数	该指标用于统计该CDM实例中处于PENDING状态的作业数。 单位：Count/个。 **说明** 2.10.0.300版本及以上版本支持该指标。	>=0	CDM集群实例	1分钟
pending_threads	排队抽取并发数	该指标用于统计该CDM实例中处于Waiting状态的抽取并发线程数。 单位：Count/个。 **说明** 2.10.0.300版本及以上版本支持该指标。	>=0	CDM集群实例	1分钟
disk_usage	磁盘利用率	该指标为从物理机层面采集的磁盘使用率，数据准确性低于从弹性云服务器内部采集的数据。 单位：%。	0.001%~90%	CDM集群实例	1分钟

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）
disk_io	磁盘io	该指标为从物理机层面采集的磁盘每秒读取和写入的字节数，数据准确性低于从弹性云服务器内部采集的数据。 单位：Byte/sec	0~10GB	CDM集群实例	1分钟
tomcat_heap_usage	堆内存使用率	该指标为从物理机层面采集的堆内存使用率，数据准确性低于从弹性云服务器内部采集的数据。 单位：%。	0.001%~90%	CDM集群实例	1分钟
tomcat_connect	tomcat并发连接数	该指标为从物理机层面采集的tomcat并发连接数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟
tomcat_thread_count	tomcat线程数	该指标为从物理机层面采集的tomcat所占线程数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟
pg_connect	数据库连接数	该指标为从物理机层面采集的postgres数据库连接数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟
pg_submission_row	历史记录表行数	该指标为从物理机层面采集的postgres数据库submission表行数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟
pg_failed_job_rate	失败作业率	该指标为从物理机层面sqoop进程采集的失败作业率。 单位：%。	0.001%~100%	CDM集群实例	1分钟
inodes_usage	Inodes利用率	该指标为从物理机层面采集的磁盘inodes使用率，数据准确性低于从弹性云服务器内部采集的数据。 单位：%。	0.001%~100%	CDM集群实例	1分钟

维度

Key	Value
instance_id	云数据迁移服务实例

5.4.8.2 设置 CDM 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。

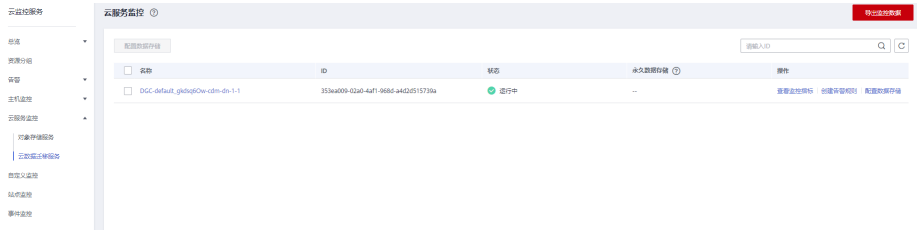
设置CDM集群的告警规则包括设置告警规则名称、监控对象、监控指标、告警阈值、监控周期和是否发送通知等参数。本节介绍了设置CDM集群告警规则的具体方法。

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 单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 5-16 “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

5.4.8.3 查看 CDM 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）。

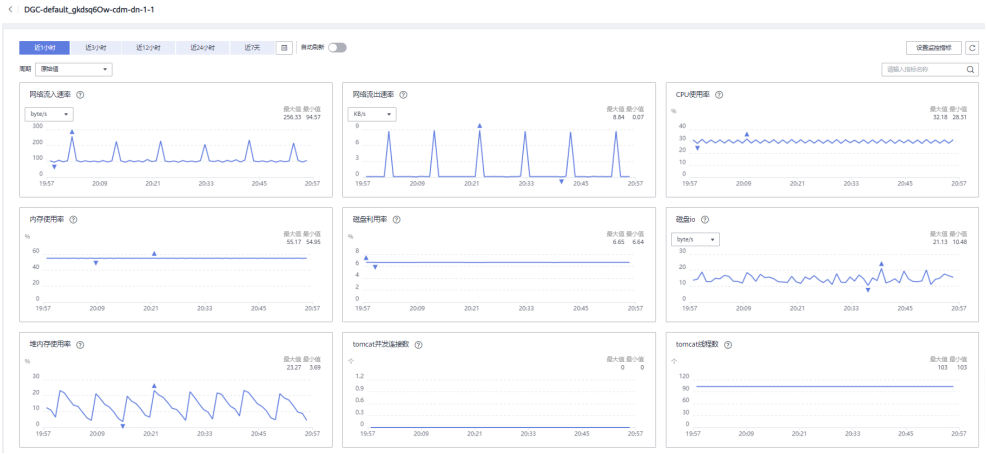
  对于新创建的集群，需要等待一段时间，才能查看上报的监控数据和监控视图。

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 5-17 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

5.5 在 CDM 集群中创建连接

5.5.1 创建 CDM 与数据源之间的连接

操作场景

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

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

约束限制

- 当所连接的数据源发生变化（如MRS集群扩容等情况）时，您需要重新编辑并保存该连接。
- 作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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。

新建连接

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。

或参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。在DataArts Studio控制台首页，选择对应工作空间的“数据集成”模块，进入CDM首页。

图 5-18 集群列表

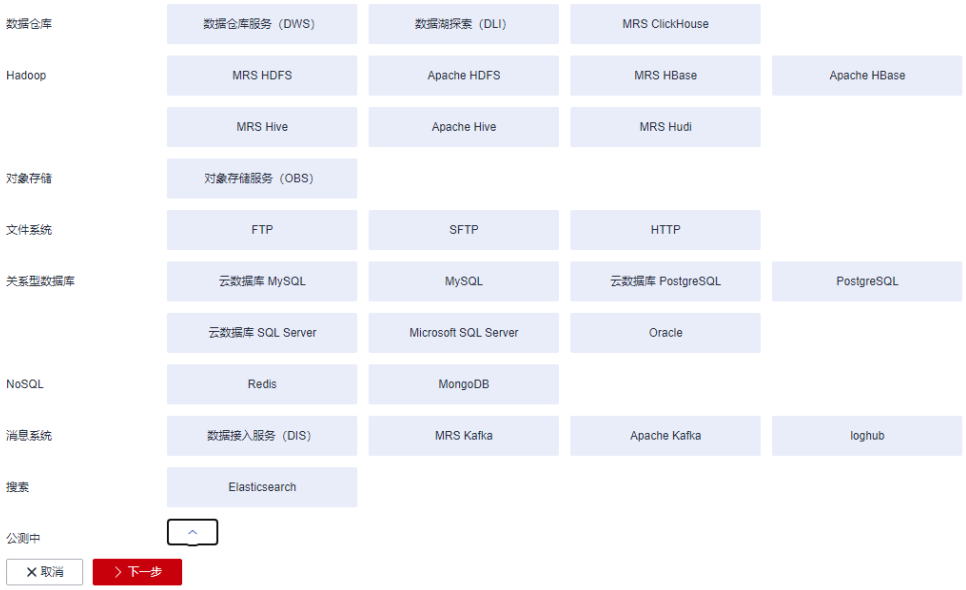
说明

“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 在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，如**图5-19**所示。

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 5-19 选择连接器类型



步骤3 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5-20**中查找对应参数。

表 5-20 连接参数分类

连接器类型	参数说明
• 云数据库 PostgreSQL • 云数据库 SQL Server • PostgreSQL • Microsoft SQL Server	由于关系型数据库所采用的JDBC驱动相同，所以连接参数也一样，具体参数请参见**PostgreSQL/SQLServer连接参数说明**。
数据仓库服务（DWS）	连接数据仓库服务（DWS）时，具体参数请参见**数据仓库服务（DWS）连接参数说明**。
SAP HANA	连接SAP HANA时，具体参数请参见**SAP HANA连接参数说明**。
达梦数据库 DM	连接达梦数据库时，具体参数请参见**达梦数据库DM连接参数说明**。
MySQL	连接MySQL数据库时，具体参数请参见**云数据库MySQL/MySQL数据库连接参数说明**。
Oracle	连接Oracle数据库时，具体参数请参见**Oracle数据库连接参数说明**。

连接器类型	参数说明
分库	连接达梦数据库时，具体参数请参见**分库连接参数说明**。
对象存储服务（OBS）	连接OBS时，具体参数请参见**OBS连接参数说明**。
• 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**HDFS连接参数说明**。
• 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**HBase连接参数说明**。
• 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**Hive连接参数说明**。
表格存储服务（CloudTable）	连接CloudTable时，具体参数请参见**CloudTable连接参数说明**。
• FTP • SFTP	连接FTP或SFTP服务器时，具体参数请参见**FTP/SFTP连接参数说明**。
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**MongoDB连接参数说明**。
文档数据库服务（DDS）	连接DDS时，具体参数请参见**DDS连接参数说明**。
• Redis • 分布式缓存服务（DCS）	连接Redis或DCS时，具体参数请参见**Redis连接参数说明**。
• MRS Kafka • Apache Kafka	连接MRS Kafka或Apache Kafka数据源时，具体参数请参见**Kafka连接参数说明**。
数据接入服务（DIS）	连接DIS时，具体参数请参见**DIS连接参数说明**。
云搜索服务 Elasticsearch	连接云搜索服务或Elasticsearch时，具体参数请参见**云搜索服务（CSS）连接参数说明**。
数据湖探索（DLI）	连接数据湖探索服务时，具体参数请参见**DLI连接参数说明**。
DMS Kafka	连接DMS的Kafka队列时，具体参数请参见**DMS Kafka连接参数说明**。

连接器类型	参数说明
Cassandra	连接Cassandra时，具体参数请参见**Cassandra连接参数说明**。 **说明** 2.9.3.300以上版本不支持Cassandra。
MRS Hudi	连接MRS Hudi时，具体参数请参见**MRS Hudi连接参数说明**。
MRS ClickHouse	连接MRS ClickHouse时，具体参数请参见**MRS ClickHouse连接参数说明**。
神通数据库（ST）	连接神通数据库（ST）时，具体参数请参见**神通（ST）连接参数说明**。

说明

目前以下数据源处于公测阶段：FusionInsight HDFS、FusionInsight HBase、FusionInsight 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、神通数据库（ST）。

步骤4 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，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，避免影响现有作业运行。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

5.5.2 配置连接参数

5.5.2.1 OBS 连接参数说明

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。

说明

- CDM集群和OBS桶不在同一个Region时，不支持跨Region访问OBS桶。
- 作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

连接OBS时，相关连接参数如**表5-21**所示。

表 5-21 OBS 连接的参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	obs_link
OBS终端节点	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。您可以通过以下方式获取OBS桶的Endpoint信息： OBS桶的Endpoint，可以进入OBS控制台概览页，单击桶名称后查看桶的基本信息获取。 **说明** • CDM集群和OBS桶不在同一个Region时，不支持跨Region访问OBS桶。 • 作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。	obs.myregion.mycloud.com
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

参数名	说明	取值样例
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。 您需要先创建当前账号的访问密钥，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5-20所示。 **图 5-20** 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)		-
连接属性	可选参数，单击“显示高级属性”后显示。 自定义连接属性，单击“添加”可增加多个属性。 只支持配置connectionTimeout，socketTimeout和idleConnectionTime。 常见配置举例如下： • socketTimeout: Socket层传输数据的超时时间，单位为毫秒。 • connectionTimeout: 建立HTTP/HTTPS连接的超时时间，单位为毫秒。	-

5.5.2.2 PostgreSQL/SQLServer 连接参数说明

连接PostgreSQL/SQLServer时，相关参数如表5-22所示，金仓和GaussDB数据源可通过PostgreSQL连接器进行连接，支持的迁移作业的源端、目的端情况与PostgreSQL数据源一致。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-22 PostgreSQL/SQLServer 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。	sql_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： 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	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
驱动类名	根据上传驱动选择对应驱动类名。 当前支持postgresql和kingbase8两种驱动类名。	-
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。	"
驱动版本	不同类型的关系数据库，需要适配不同的驱动，更多详情请参见**如何获取驱动**。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
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。	10000
SSL加密	可选参数，控制是否通过SSL加密方式连接数据库。	是

参数名	说明	取值样例
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。 常见配置举例如下： • **connectTimeout=60**与**socketTimeout=300**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位s），避免超时导致失败。 • **useCursorFetch=false**：CDM作业默认打开了JDBC连接器与关系型数据库通信使用二进制协议开关，即useCursorFetch=true。部分第三方可能存在兼容问题导致迁移时间转换出错，可以关闭此开关。 • **trustServerCertificate=true**：在创建安全连接的时候可能会报PKIX错误，建议设置为true。	sslmode=require
连接私密属性	可选参数，单击“显示高级属性”后显示。 自定义私密连接属性。	sk=09fUgD5WOF1L6f

5.5.2.3 数据仓库服务（DWS）连接参数说明

连接数据仓库服务（DWS）时，相关参数如表5-23所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-23 数据仓库服务（DWS）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。	dws_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”，可获取用户的实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	不同的数据库端口不同，请根据具体情况配置。
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm

参数名	说明	取值样例
密码	用户名密码。	-
使用Agent	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
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。	10000
SSL加密	可选参数，控制是否通过SSL加密方式连接数据仓库。	是 **说明** 启用SSL加密需确保DWS本身已启用SSL加密。
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。 常见配置举例如下： • **connectTimeout=60与socketTimeout=300**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位s），避免超时导致失败。 • **useCursorFetch=false**：CDM作业默认打开了JDBC连接器与关系型数据库通信使用二进制协议开关，即useCursorFetch=true。部分第三方可能存在兼容问题导致迁移时间转换出错，可以关闭此开关；开源MySQL数据库支持useCursorFetch参数，无需对此参数进行设置。	sslmode=require **说明** 启用SSL加密后sslmode值不设置可能会导致连接失败。
连接私密属性	可选参数，单击“显示高级属性”后显示。 自定义私密连接属性。	sk=09fUgD5WOF1L6f

5.5.2.4 云数据库 MySQL/MySQL 数据库连接参数说明

连接MySQL数据库连接时，相关参数如表5-24所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-24 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。注意，开启本参数后，日期类型将不符合格式的会存储为0000-00-00，更多详细信息可在MySQL官网文档查看。 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	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
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。	10000
SSL加密	可选参数，控制是否通过SSL加密方式连接数据库，创建云数据MySQL连接时显示该参数。	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。 常见配置举例如下： ● **connectTimeout=600000**与**socketTimeout=300000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。 ● **tinyInt1isBit=false**或**mysql.bool.type.transform=false**：MySQL默认开启配置tinyInt1isBit=true，将TINYINT(1)当作BIT也就是Types.BOOLEAN来处理，会将1或0读取为true或false从而导致迁移失败，此时可关闭配置避免迁移报错。 ● **useCursorFetch=false**：CDM作业默认打开了JDBC连接器与关系型数据库通信使用二进制协议开关，即useCursorFetch=true。部分第三方可能存在兼容问题导致迁移时间转换出错，可以关闭此开关；开源MySQL数据库支持useCursorFetch参数，无需对此参数进行设置。 ● **allowPublicKeyRetrieval=true**：MySQL默认关闭允许公钥检索机制，因此连接MySQL数据源时，如果TLS不可用、使用RSA公钥加密时，可能导致连接报错。此时可打开公钥检索机制，避免连接报错。 ● **useSSL=false**：CDM集群版本为2.10.0.300且MySQL版本为mysql5.7.43以上时，可以通过添加连接属性**useSSL=false**打开SSL加密开关。	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。	`

参数名	说明	取值样例
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

5.5.2.5 Oracle 数据库连接参数说明

连接Oracle数据库时，连接参数如表5-25所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-25 Oracle 数据库连接参数

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

参数名	说明	取值样例
驱动版本	不同类型的关系数据库，需要适配不同的驱动，更多详情请参见**如何获取驱动**。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。 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“单次请求行数”参数值的设置。	1000
单次提交行数	可选参数，单击“显示高级属性”后显示。 指定单次批量提交的行数。	10000
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。 常见配置举例如下： • socketTimeout：配置JDBC连接超时时间，单位为毫秒。 • mysql.bool.type.transform：配置mysql读取时，是否将tinyint(1)解析成boolean类型，默认为true。	-
连接私密属性	可选参数，单击“显示高级属性”后显示。 自定义私密连接属性。	sk=09fUgD5WOF1L6f

5.5.2.6 DLI 连接参数说明

连接数据湖探索（DLI）服务时，相关参数如**表5-26**所示。

说明

- 作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。
- 迁移数据到DLI时，DLI要在OBS的*dli-trans*内部临时桶生成数据文件，因此在需要赋予DLI连接中使用AK/SK所在用户对*dli-trans*桶的读、写、创建目录对象等权限，否则会导致迁移失败。*dli-trans*内部临时桶的权限策略添加请参见**新增dli-trans*内部临时桶授权策略**。

表 5-26 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

参数名	说明	取值样例
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5-21所示。 **图 5-21** 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，账号ID对应当前账号，IAM用户ID对应当前用户。用户可在对应页面下查看不同Region对应的项目ID、账号ID和用户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“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。	-
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。	50000

新增 *dli-trans-*内部临时桶授权策略

步骤1 登录统一身份认证服务IAM控制台。

步骤2 在左侧导航窗格中，选择“权限管理>权限”页签，单击右上方的“创建自定义策略”。

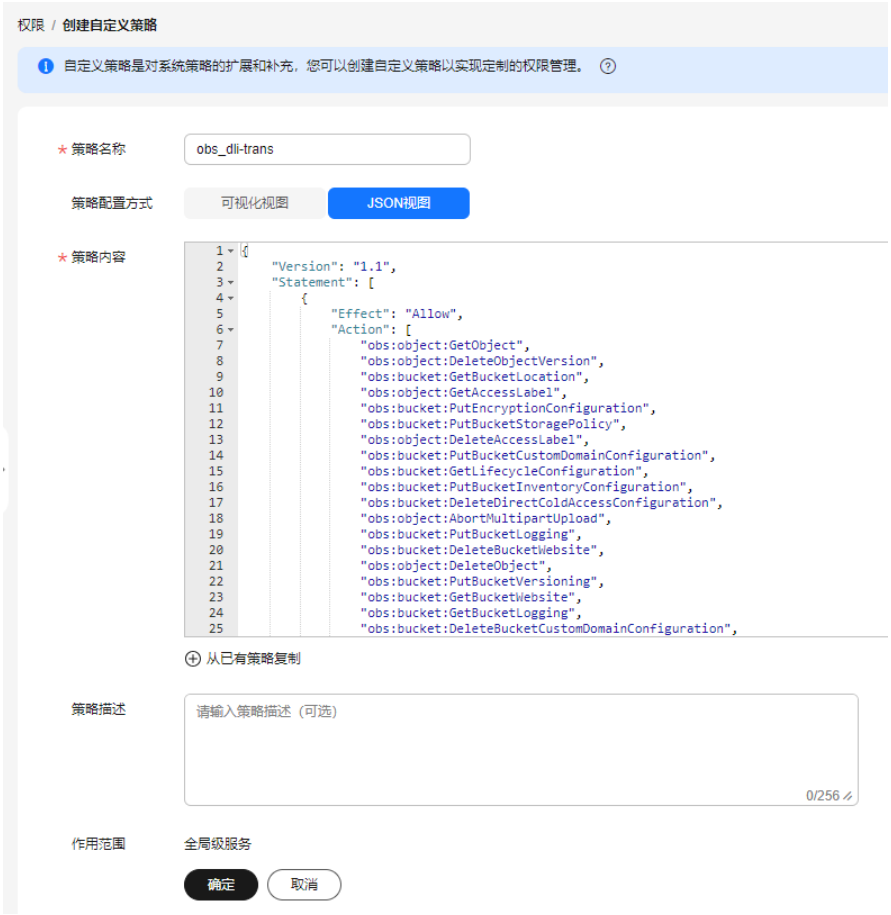
图 5-22 创建自定义策略

步骤3 在自定义策略配置页面，策略配置方式切换至JSON视图，然后按照如下策略内容，创建**obs_dli-trans**自定义策略。

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "obs:object:GetObject",
                "obs:object:DeleteObjectVersion",
                "obs:bucket:GetBucketLocation",
                "obs:object:GetAccessLabel",
                "obs:bucket:PutEncryptionConfiguration",
                "obs:bucket:PutBucketStoragePolicy",
                "obs:object:DeleteAccessLabel",
                "obs:bucket:PutBucketCustomDomainConfiguration",
                "obs:bucket:GetLifecycleConfiguration",
                "obs:bucket:PutBucketInventoryConfiguration",
                "obs:bucket:DeleteDirectColdAccessConfiguration",
                "obs:object:AbortMultipartUpload",
                "obs:bucket:PutBucketLogging",
                "obs:bucket:DeleteBucketWebsite",
                "obs:object:DeleteObject",
                "obs:bucket:PutBucketVersioning",
                "obs:bucket:GetBucketWebsite",
                "obs:bucket:GetBucketLogging",
                "obs:bucket:DeleteBucketCustomDomainConfiguration",
                "obs:object:PutObject",
                "obs:object:RestoreObject",
                "obs:bucket:PutReplicationConfiguration",
                "obs:bucket:GetBucketQuota",
                "obs:object:GetObjectVersionAcl",
                "obs:bucket:DeleteBucket",
                "obs:bucket:CreateBucket",
                "obs:bucket:GetDirectColdAccessConfiguration",
                "obs:bucket:PutDirectColdAccessConfiguration",
                "obs:bucket:GetBucketAcl",
                "obs:bucket:GetBucketVersioning",
                "obs:bucket:GetBucketInventoryConfiguration",
                "obs:bucket:GetBucketStoragePolicy",
                "obs:bucket:GetEncryptionConfiguration",
                "obs:bucket:PutBucketCORS",
                "obs:bucket:PutBucketTagging",
                "obs:bucket:GetBucketTagging",
                "obs:bucket:PutLifecycleConfiguration",
                "obs:bucket:GetBucketCustomDomainConfiguration",
                "obs:object:ListMultipartUploadParts",
                "obs:object:ModifyObjectMetaData",
                "obs:bucket:ListBucketVersions",
                "obs:bucket:PutBucketQuota",
                "obs:object:PutAccessLabel",
                "obs:bucket:ListBucket",
                "obs:bucket:GetBucketCORS",
                "obs:bucket:DeleteBucketInventoryConfiguration",
                "obs:object:GetObjectVersion",
                "obs:bucket:PutBucketWebsite",
                "obs:bucket:DeleteReplicationConfiguration",
                "obs:object:GetObjectAcl",
                "obs:bucket:GetBucketNotification",
                "obs:bucket:PutBucketNotification",
                "obs:bucket:GetReplicationConfiguration",
```

```
            "obs:bucket:GetBucketPolicy",
            "obs:bucket:DeleteBucketTagging",
            "obs:bucket:GetBucketStorage"
        ],
        "Resource": [
            "OBS:*:*:object:*",
            "OBS:*:*:bucket:dli-trans*"
        ]
    }
  ]
}
```

图 5-23 配置 **obs_dli-trans** 自定义策略



步骤4 单击“确定”，完成**obs_dli-trans**自定义策略创建。

步骤5 在IAM左侧导航窗格中，选择“用户组”，找到DLI连接中使用AK/SK所在用户的归属用户组，单击授权，将**obs_dli-trans**自定义策略授权给该用户。

图 5-24 为用户组授权 **obs_dli-trans** 自定义策略

----结束

5.5.2.7 Hive 连接参数说明

目前CDM支持连接的Hive数据源有以下几种：

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

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

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连接适用于华为云上的MapReduce服务。MRS Hive的连接参数如**表5-27**所示。

说明

- 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
- 新建MRS Hive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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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 5-27 MRS Hive 连接参数

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

参数名	说明	取值样例
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。	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。	-
开启LDAP认证	通过代理连接的时候，此项可配置。 当MRS Hive对接外部LDAP开启了LDAP认证时，连接Hive时需要使用LDAP账号与密码进行认证，此时必须开启此参数，否则会连接失败。	否
LDAP用户名	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的用户名。	-

参数名	说明	取值样例
LDAP密码	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 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5-25所示。 **图 5-25** 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)		-

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

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

常见配置举例如下：

- **connectTimeout=360000与socketTimeout=360000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。
- **hive.server2.idle.operation.timeout=360000**：为避免Hive迁移作业长时间卡住，可自定义operation超时时间（单位ms）。
- **hive.storeFormat=textfile**：关系型数据库迁移到Hive时，自动建表默认为orc格式。如果需要指定为textfile格式，可增加此配置。parquet格式同理，hive.storeFormat属性值指定为parquet格式即可。
- **fs.defaultFS=obs://hivedb**：对接的MRS Hive为存算分离模式时，可通过此配置获取更佳兼容性。

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5-28**所示。

表 5-28 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。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否

参数名	说明	取值样例
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 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5-26所示。 **图 5-26** 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。	- -
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。 集群配置的创建方法请参见**管理集群配置**。	hive_01

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

常见配置举例如下：

- **connectTimeout=360000与socketTimeout=360000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。
- **hive.server2.idle.operation.timeout=360000**：为避免Hive迁移作业长时间卡住，可自定义operation超时时间（单位ms）。

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5-29所示。

表 5-29 Apache Hive 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hivelink
URI	NameNode URI地址。	hdfs://hacluster
Hive元数据地址	设置Hive元数据地址，参考 hive.metastore.uris配置项。例如：thrift://host-192-168-1-212:9083	-
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。	HIVE_3_X
IP与主机名映射	如果Hadoop配置文件使用主机名，需要配置IP与主机的映射。格式：IP与主机名之间使用空格分隔，多对映射使用分号或回车换行分隔。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否

参数名	说明	取值样例
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 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5-27所示。 **图 5-27** 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)		-
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

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

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

常见配置举例如下：

- **connectTimeout=360000与socketTimeout=360000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。
- **hive.server2.idle.operation.timeout=360000**：为避免Hive迁移作业长时间卡住，可自定义operation超时时间（单位ms）。

5.5.2.8 HBase 连接参数说明

目前CDM支持连接的HBase数据源有以下几种：

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

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

MRS HBase

连接MRS上的HBase数据源时，相关参数如**表5-30**所示。

说明

- 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
- 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。
- 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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

  当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。

表 5-30 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。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。	127.0.0.1

参数名	说明	取值样例
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。	-
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
HBase版本	HBase版本。	HBASE_2_X
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	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

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

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

FusionInsight HBase

连接FusionInsight HD上的HBase数据源时，相关参数如**表5-31**所示。

表 5-31 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

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

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

Apache HBase

连接Apache Hadoop上的HBase数据源时，相关参数如**表5-32**所示。

表 5-32 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

参数名	说明	取值样例
IP与主机名映射	输入IP和主机名。 如果配置文件使用主机名，需要配置所有IP与主机的映射，多个主机之间使用空格进行分隔。	IP：10.3.6.9 主机名：hostname01
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。 集群配置的创建方法请参见**管理集群配置**。	hbase_01

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

5.5.2.9 HDFS 连接参数说明

目前CDM支持连接的HDFS数据源有以下几种：

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

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

MRS HDFS

连接MRS上的HDFS数据源时，相关参数如表5-33所示。

说明

- 当前暂不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。
- 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。
- 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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

  当同一Agent连接多个MRS集群时，如果其中一个MRS集群被删除或状态异常，会影响另外一个正常的MRS集群数据连接。因此建议一个Agent对应一个MRS集群数据连接。

表 5-33 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。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。	127.0.0.1

参数名	说明	取值样例
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。	-
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。 **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。 若在一个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	Agent功能待下线，无需配置。	-

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

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

FusionInsight HDFS

连接FusionInsight HD上的HDFS数据源时，相关参数如**表5-34**所示。

表 5-34 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

参数名	说明	取值样例
运行模式	选择HDFS连接的运行模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式。 **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
Agent	Agent功能待下线，无需配置。	-
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hdfs_01

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

Apache HDFS

连接Apache Hadoop上的HDFS数据源时，相关参数如**表5-35**所示。

表 5-35 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

参数名	说明	取值样例
运行模式	选择HDFS连接的运行模式： • EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 • STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式。 **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
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	Agent功能待下线，无需配置。	-
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	当“是否使用集群配置”为“是”或“认证类型”为“KERBEROS”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hdfs_01

5.5.2.10 FTP/SFTP 连接参数说明

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。
- 作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

连接FTP或SFTP服务器时，连接参数相同，如**表5-36**所示。

表 5-36 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

参数名	说明	取值样例
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，FTP默认值为21；SFTP默认值为22。	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。	-
FTP文件名编码	FTP时显示该参数。 ftp-client的controlEncoding文件名编码配置，默认为ISO-8859-1，目前支持ISO-8859-1/UFT8。	ISO-8859-1

5.5.2.11 Redis 连接参数说明

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。

Redis连接不支持SSL加密的Redis数据源。

连接本地Redis数据库时，相关参数如表5-37所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-37 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	Redis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。	-
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

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
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
集群配置名称	仅当认证类型为KERBEROS时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hdfs_01

5.5.2.12 DDS 连接参数说明

DDS连接适用于华为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5-38**所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-38 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

参数名	说明	取值样例
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。	-
直连模式	适用于主节点网络通，副本节点网络不通场景。 **说明** • 直连模式服务器列表只能配一个ip。 • 直连适用于主节点网络通，副本节点网络不通场景。	否

5.5.2.13 CloudTable 连接参数说明

连接CloudTable时，相关参数如表5-39所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-39 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统一身份认证	如果所需连接的CloudTable集群在创建时开启了“IAM统一身份认证”，该参数需设置为“是”，否则设置为“否”。 当选择IAM统一身份认证时，需要输入用户名、AK和SK。	否
用户名	登录CloudTable集群的用户名。	admin
AK	登录CloudTable集群的访问标识。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。	-

参数名	说明	取值样例
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。 集群配置的创建方法请参见**管理集群配置**。	hadoop_01

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

5.5.2.14 MongoDB 连接参数说明

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5-40**所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-40 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。	-
直连模式	适用于主节点网络通，副本节点网络不通场景。 **说明** • 直连模式服务器列表只能配一个ip。 • 直连适用于主节点网络通，副本节点网络不通场景。	否

参数名	说明	取值样例
连接属性	自定义连接属性，支持MongoDB属性，单位为ms。连接属性如下： ● socketTimeout，默认socketTimeout=60000 ● maxWaitTime，默认maxWaitTime=10000 ● connectTimeout，默认connectTimeout=10000 ● serverSelectionTimeout，默认serverSelectionTimeout=5000	socketTimeout=60000

5.5.2.15 Cassandra 连接参数说明

说明

- 2.9.3.300以上版本不支持Cassandra。
- 作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-41 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

5.5.2.16 DIS 连接参数说明

连接DIS时，相关参数如表5-42所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-42 DIS 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。	dis_link
区域	DIS所在的区域。	-
终端节点	待连接DIS的URL，URL一般格式为：https://Endpoint。 终端节点（Endpoint）即调用API的**请求地址**，不同服务不同区域的终端节点不同。本服务的Endpoint可从**终端节点Endpoint**获取。	-
访问标识(AK)	登录DIS服务器的访问标识。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。	-
密钥(SK)	登录DIS服务器的密钥。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。	-
项目ID	DIS的项目ID。	-

5.5.2.17 Kafka 连接参数说明

MRS Kafka

连接MRS上的Kafka数据源时，相关参数如**表5-43**所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-43 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。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。	127.0.0.1

参数名	说明	取值样例
用户名	需要配置MRS Manager的用户名和密码。 如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。 **说明** • 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对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5-44所示。

表 5-44 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。

5.5.2.18 DMS Kafka 连接参数说明

连接DMS的Kafka队列时，相关参数如表5-45所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-45 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。当DMS Kafka实例的连接信息中启用的安全协议为“SASL_SSL”时需要开启。 开启Kafka SASL_SSL，则数据加密传输，安全性更高，但性能会下降。 **说明** 启用SSL认证后，Kafka会将Kafka Broker连接地址视做域名不断进行解析，导致性能消耗。建议修改CDM集群对应的ECS主机（通过集群IP查找对应的ECS主机）中的“/etc/hosts”文件，为其添加Broker连接地址的自映射，以便客户端能够快速解析实例的Broker。例如Kafka Broker地址配置为10.154.48.120时，hosts文件中的自映射配置为： `10.154.48.120 10.154.48.120`	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。	-

参数名	说明	取值样例
属性配置	• 当DMS Kafka实例的连接信息中启用的安全协议后，需要添加数据加密方式属性：属性名称填写为**security.protocol**，值根据Kafka实例中的安全协议填写为**SASL_SSL**或**SASL_PLAINTEXT**。 • 当DMS Kafka实例的连接信息中配置SASL认证机制后，需要添加认证方式的属性：属性名称填写为**sasl.mechanism**，值根据Kafka实例中配置的SASL认证机制填写为**PLAIN**或**SCRAM-SHA-512**（同时支持时选择其中任意一种填写即可）。	-

5.5.2.19 云搜索服务（CSS）连接参数说明

华为云的云搜索服务（CSS）是一个基于Elasticsearch且完全托管的在线分布式搜索服务，CSS连接适用于将各类日志文件、数据库记录迁移到CSS，Elasticsearch引擎进行搜索和分析的场景。

说明

- 导入数据到CSS推荐使用Logstash，请参见**使用Logstash导入数据到Elasticsearch**。
- 作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

连接云搜索服务(CSS)时，相关参数如**表5-46**所示。

表 5-46 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。	-

参数名	说明	取值样例
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

5.5.2.20 Elasticsearch 连接参数说明

Elasticsearch连接适用于第三方云的Elasticsearch服务，以及用户在本地数据中心或ECS上自建的Elasticsearch。

说明

- Elasticsearch连接器仅支持非安全模式的Elasticsearch集群。
- 作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

连接Elasticsearch时，相关参数如表5-47所示。

表 5-47 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。	es_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

5.5.2.21 达梦数据库 DM 连接参数说明

连接达梦数据库 DM时，相关参数如表5-48所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-48 达梦数据库 DM 连接参数

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

参数名	说明	取值样例
端口	配置为要连接的数据库的端口。	不同的数据库端口不同，请根据具体情况配置。
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。	-
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
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。	'

5.5.2.22 SAP HANA 连接参数说明

连接SAP HANA时，相关参数如表5-49所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-49 SAP HANA 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。	sap_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”，可获取用户的实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	不同的数据库端口不同，请根据具体情况配置。
数据库名称	配置为要连接的数据库名称。	dbname

参数名	说明	取值样例
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
使用Agent	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
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
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。 常见配置举例如下： • **connectTimeout=360000**与**socketTimeout=360000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。 • **useCursorFetch=false**：CDM作业默认打开了JDBC连接器与关系型数据库通信使用二进制协议开关，即useCursorFetch=true。部分第三方可能存在兼容问题导致迁移时间转换出错，可以关闭此开关；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。	'

5.5.2.23 分库连接参数说明

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5-50所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-50 分库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	my_link
用户名	待连接数据库的用户。 仅当“数据源列表”中某个后端数据库A未配置用户名密码时，该配置对A生效。如果后端数据库B已配置用户名密码，此处配置不对B生效。	cdm
密码	待连接数据库的用户密码。 仅当“数据源列表”中某个后端数据库A未配置用户名密码时，该配置对A生效。如果后端数据库B已配置用户名密码，此处配置不对B生效。	-
使用Agent	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
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“192.168.3.0:3306:cdm|192.168.2.2:3306:cdm:user:password”表示，第一个后端数据库IP为192.168.3.0，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”。	192.168.3.0:3306:cdm|192.168.2.2:3306:cdm:user:password
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。	'

5.5.2.24 MRS Hudi 连接参数说明

连接MRS Hudi时，相关参数如**表5-51**所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-51 Hudi 连接参数

参数名	说明	取值样例
名称	连接名称。	Hudilink
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。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。	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。	KERBEROS
账号	登录MRS Manager的账号。	cdm
密码	登录MRS Manager的密码。	-
OBS支持	是否支持OBS存储，如果hudi表数据存储在OBS，需要打开此开关。	是

参数名	说明	取值样例
访问标识（AK）	“OBS支持”设置为“是”时，呈现此参数。 AK和SK分别为登录OBS服务器的访问标识与密钥。 您需要先创建当前账号的访问密钥，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5-28所示。 **图 5-28** 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）		-
OBS测试路径	“OBS支持”设置为“是”时，呈现此参数。 请填写完整的文件路径，将调用元数据查询接口来校验路径的访问权限。 **说明** • 如果是对象存储，路径需要填写到对象级别，否则会报错404，例如：“obs://bucket/dir/test.txt”。 • 如果是并行文件系统，则可以只填写到目录级别。例如：“obs://bucket/dir”。	obs://bucket/dir/test.txt
属性配置	需要集成的表名，多个表名使用英文逗号“,”分开，请务必配置，不要有空格，默认无需配置。	-

5.5.2.25 MRS ClickHouse 连接参数说明

连接MRS ClickHouse时，相关参数如表5-52所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-52 ClickHouse 连接参数

参数名	说明	取值样例
名称	连接名称。	cklink
数据库服务器	配置为要连接的数据库的IP地址或域名。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。 登录MRS ClickHouse数据源所在集群的Manager页面，选择“集群 > 服务 > ClickHouse > 实例”，查看ClickHouseServer所在的“业务IP”。	192.168.0.1
端口	配置为要连接的数据库的端口。 **说明** ● 如果使用Server节点，开启“SSL加密”，配置默认端口。登录MRS ClickHouse数据源所在集群的Manager页面，选择“集群 > 服务 > ClickHouse > 实例”，配置ClickHouseServer的默认端口，非安全模式MRS集群配置“http_port”参数对应的端口，安全模式MRS集群配置“https_port”参数对应的端口。 ● 如果使用Balancer节点，开启“SSL加密”，配置默认端口。登录MRS ClickHouse数据源所在集群的Manager页面，选择“集群 > 服务 > ClickHouse > 实例”，配置ClickHouseBalancer的默认端口，非安全模式MRS集群配置“lb_http_port”参数对应的端口，安全模式MRS集群配置“lb_https_port”参数对应的端口。 ● 如果MRS ClickHouse是安全集群，则需配置为https默认端口。	8123
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。	-
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。	否
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。	'

5.5.2.26 神通（ST）连接参数说明

连接神通（ST）数据库连接时，相关参数如表5-53所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-53 神通（ST）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。	st_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Agent	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
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。	'
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
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。 常见配置举例如下： • **connectTimeout=360000**与**socketTimeout=360000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。	sslmode=require

5.5.2.27 CloudTable OpenTSDB 连接参数说明

连接CloudTable OpenTSDB时，相关参数如表5-54所示。

表 5-54 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，账号ID对应当前账号，IAM用户ID对应当前用户。用户可在对应页面下查看不同Region对应的项目ID、账号ID和用户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“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。	-
用户名	访问CloudTable服务的用户名。	admin

参数名	说明	取值样例
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5-29所示。 **图 5-29** 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)		-

5.5.2.28 GBASE 连接参数说明

连接GBASE连接时，相关参数如表5-55所示。

表 5-55 GBASE 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。	adb_link
连接器	默认为关系数据库，不可更改。	-
数据库服务器	配置为要连接的数据库的IP地址或域名，多个值以;分隔。	192.168.0.1;192.168.0.2
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

参数名	说明	取值样例
用户名	待连接数据库的用户。 数据库用户名。新建分库连接时，此配置对数据源列表中所有未配置用户名密码的后端连接生效；编辑分库连接时，如需修改已存在的后端连接，请在数据源列表中单独指定用户名密码。	cdm
密码	数据库密码。	-
使用Agent	Agent功能待下线，无需配置。 GBASE为GBASE8A时显示该参数。	-
Agent	Agent功能待下线，无需配置。 GBASE为GBASE8A时显示该参数。	-
引用符号	可选参数，数据库包围标识符。对某些数据库意味着大小写敏感，如不需用请置空。	"
驱动版本	不同类型的关系数据库，需要适配不同的驱动，更多详情请参见**如何获取驱动**。 GBASE为GBASE8A时显示该参数。	-
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。	10000
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。	1000
连接属性	自定义连接属性。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。 常见配置举例如下： • socketTimeout：配置JDBC连接超时时间，单位为毫秒。 • mysql.bool.type.transform：配置mysql读取时，是否将tinyint(1)解析成boolean类型，默认为true。	-

5.5.2.29 YASHAN 连接参数说明

连接YASHAN时，相关参数如**表1 YASHAN连接参数**所示。

说明

作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

表 5-56 YASHAN 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。	yashan_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”，可获取用户的实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	1688
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	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
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。	"
驱动版本	不同类型的关系数据库，需要适配不同的驱动，更多详情请参见**如何获取驱动**。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
SSL加密	可选参数，单击“显示高级属性”后显示。 支持启用SSL加密传输。	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。 常见配置举例如下： ● socketTimeout：配置JDBC连接超时时间，单位为毫秒。 ● mysql.bool.type.transform：配置mysql读取时，是否将tinyint(1)解析成boolean类型，默认为true。	socketTimeout=300
连接私密属性	自定义私密连接属性。	xxx=xxx

5.5.3 上传 CDM 连接驱动

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5-57**下载对应的驱动。
- 已参见**FTP/SFTP连接参数说明**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5-57**获取建议版本的JDK8 .jar格式驱动即可。

表 5-57 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版本，获取mysql-connector-java-5.1.48.jar
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://mvnrepository.com/artifact/org.postgresql/postgresql**	PostgreSQL推荐使用42.3.4版本，获取postgresql-42.3.4.jar

关系数据库类型	驱动名称	获取地址	建议版本
YASHAN	Yashan DB 23.2.4	**https://download.yashandb.com/download**	YASHAN推荐使用23.2.4版本，获取： • Linux X86：yashandb-23.2.4.100-linux-x86_64.tar • Linux ARM：yashandb-23.2.4.100-linux-aarch64.tar
金仓数据库	POSTGRESQL	**https://mvnrepository.com/artifact/org.postgresql/postgresql**	金仓数据库推荐使用42.2.9版本PostgreSQL驱动，获取postgresql-42.2.9.jar
GaussDB数据库	POSTGRESQL	GaussDB JDBC驱动请在**GaussDB官方文档**中搜索“JDBC包、驱动类和环境类”，然后选择实例对应版本的文档，参考文档获取gsjdbc4.jar。	请从对应版本的发布包中获取gsjdbc4.jar
• 云数据库 SQL Server • Microsoft SQL Server	SQLServer	**https://docs.microsoft.com/en-us/sql/connect/jdbc/release-notes-for-the-jdbc-driver?view=sql-server-ver15#previous-releases**	4.2版本，获取sqljdbc42.jar
达梦数据库 DM	DM	DM JDBC驱动jar包请从DM安装目录/dmdbms/drivers/jdbc中获取DmJdbcDriver18.jar。	请从对应版本的安装目录中获取DmJdbcDriver18.jar
POSTGRESQL_KINGBASE	POSTGRESQL_KINGBASE	**https://www.kingbase.com.cn/rjcxxz/index.htm**	与KINGBASE数据库版本配套的驱动版本

关系数据库类型	驱动名称	获取地址	建议版本
GBASE	• GBASE8A • GBASE8S	• GBASE8A：https://www.gbase.cn/download/gbase-8a?category=DRIVER_PACKAGE • GBASE8S：https://www.gbase.cn/download/gbase-8s-1?category=DRIVER_PACKAGE	• GBASE8A：GBase 8a MPP Cluster V9版本，获取gbase-connector-java-9.5.0.7-build1-bin.jar • GBASE8S：GBase 8s V8.8版本，获取gbasedbtjdbc_3.5.1_3X1_3.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”，进入驱动管理页面上传驱动。

图 5-30 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。

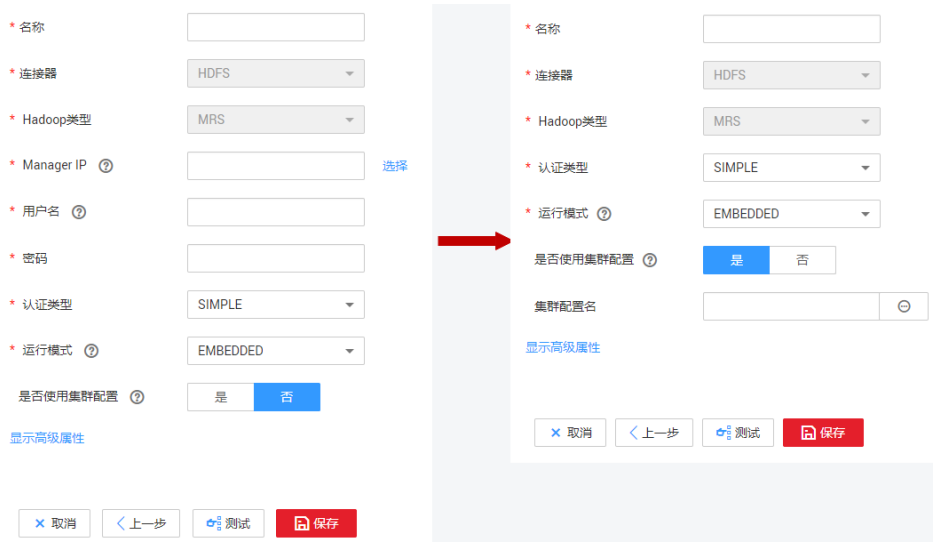
----结束

5.5.4 新建 Hadoop 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5-31所示。

图 5-31 使用集群配置前后对比



CDM支持的Hadoop类型连接主要包括以下几类：

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

操作场景

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

前提条件

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

获取集群配置文件和 Keytab 文件

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

表 5-58 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 ● MRS Hudi ● MRS ClickHouse	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，其他保持默认即可，单击确认后进行本地下载。 4. 获取下载的tar包，此即为FusionInsight集群配置文件。 针对MRS 2.x及之前版本集群： 1. 登录MRS管理控制台。 2. 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面，单击“组件管理”。 3. 单击“下载客户端”。“客户端类型”选择“仅配置文件”，“下载路径”选择“服务器端”或“远端主机”，自定义文件保存路径后，单击“确定”开始生成客户端配置文件。 4. 将生成的配置文件，保存到本地路径。 具体可参见MapReduce服务文档。	针对MRS 3.x版本集群： 1. 登录FusionInsight Manager。 2. 通过“系统 > 权限 > 用户”，选择所需用户所在行，单击“更多 > 下载认证凭据”下载认证凭据文件。 3. 获取下载的tar包，此即为FusionInsight集群Keytab文件。 针对MRS 2.x及之前版本集群： 1. 登录MRS服务的Manager，单击“系统设置”。在“权限配置”区域，单击“用户管理”。 2. 在需导出keytab文件用户所在的行，选择“更多 > 下载认证凭据”下载认证文件，待文件自动生成后指定保存位置，并妥善保管该文件。 具体可参见MapReduce服务文档。

Hadoop类型连接	集群配置文件获取方式	Keytab文件获取方式
FusionInsight集群 ● FusionInsight HDFS ● FusionInsight HBase ● FusionInsight Hive	1. 登录FusionInsight Manager。 2. 选择“集群 > 待操作的集群名称 > 概览 > 更多 > 下载客户端”，界面显示“下载集群客户端”对话框。 3. 对话框中选择“仅配置文件”，平台类型和服务端保持一致，其他保持默认即可，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，单击“更多 > 下载认证凭据”下载认证凭据文件。 3. 获取下载的tar包，此即为FusionInsight集群Keytab文件。 具体可参见FusionInsight文档。

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

说明

- 集群配置文件包含集群的配置参数。如果修改了集群的配置参数，需重新获取配置文件。
- 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

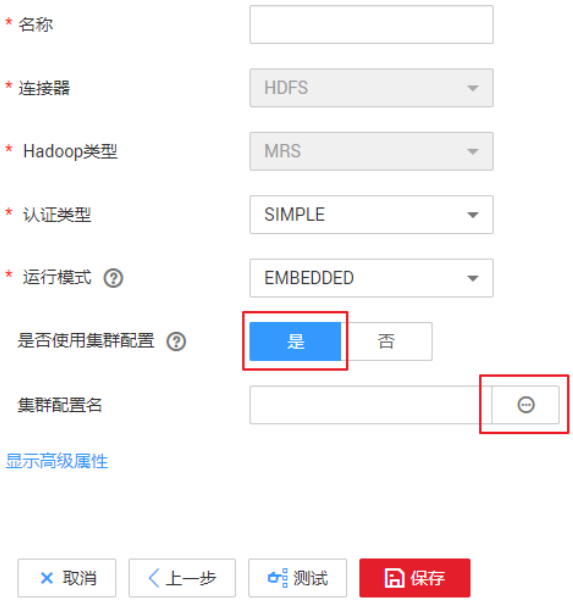
1. 进入CDM主界面，进入集群管理界面。选择CDM集群后的“作业管理 > 连接管理 > 集群配置管理”。
2. 在集群配置管理界面，选择“新建集群配置”，配置参数填写如下：

   图 5-32 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 5-33 使用集群配置



5.6 在 CDM 集群中创建作业

5.6.1 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**支持的数据源**。

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。
- 单文件传输大小不超过1TB。
- 配置源端和目的端参数时，字段名不可包含&和%。

前提条件

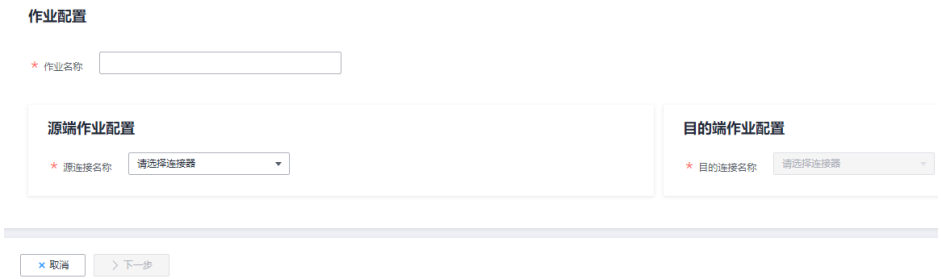
- 已新建连接，详情请参见**创建CDM与数据源之间的连接**。
- CDM集群与待迁移数据源可以正常通信。

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 5-34 新建表/文件迁移的作业

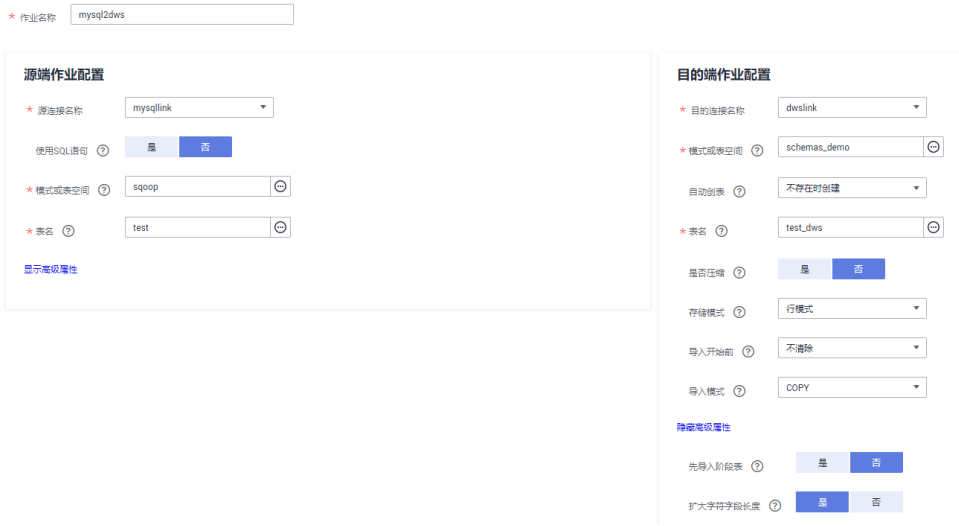


步骤3 选择源连接、目的连接：

- 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“oracle2rds_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，例如迁移MySQL到DWS时，如图5-35所示。

图 5-35 新建作业



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5-59和表5-60选择。

表 5-59 源端作业参数说明

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

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

源端类型	说明	参数配置
• MongoDB • 文档数据库服务（DDS）	支持从MongoDB或DDS导出数据。	参见**配置MongoDB/DDS源端参数**。
Redis	支持从开源Redis导出数据。	参见**配置Redis源端参数**。
数据接入服务（DIS）	仅支持导出数据到云搜索服务。	参见**配置DIS源端参数**。
• Apache Kafka • DMS Kafka • MRS Kafka	仅支持导出数据到云搜索服务。	参见**配置Kafka/DMS Kafka源端参数**。
• 云搜索服务 • Elasticsearch	支持从云搜索服务或Elasticsearch导出数据。	参见**配置Elasticsearch/云搜索服务源端参数**。
MRS Hudi	支持从MRS Hudi导出数据。	参见**配置MRS Hudi源端参数**。
MRS ClickHouse	支持从MRS ClickHouse导出数据。	参见**配置MRS ClickHouse源端参数**。
神通（ST）	支持从神通（ST）导出数据。	参见**配置神通（ST）源端参数**。
达梦数据库 DM	支持从达梦数据库 DM导出数据。	参见**配置达梦数据库DM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5-60**所示。

表 5-60 目的端作业参数说明

目的端类型	说明	参数配置
OBS	支持使用CSV或二进制格式批量传输大量文件到OBS。	参见**配置OBS目的端参数**。
MRS HDFS	导入数据到HDFS时，支持设置压缩格式。	参见**配置HDFS目的端参数**。
MRS HBase CloudTable	支持导入数据到HBase，创建新HBase表时支持设置压缩算法。	参见**配置HBase/CloudTable目的端参数**。
MRS Hive	支持快速导入数据到MRS的Hive。	参见**配置Hive目的端参数**。

目的端类型	说明	参数配置
• MySQL • SQL Server • PostgreSQL	支持导入数据到云端的数据库服务。	使用JDBC接口导入数据，参见**配置MySQL/SQL Server/PostgreSQL目的端参数**。
DWS	支持导入数据到数据仓库DWS。	参见**配置DWS目的端参数**。
Oracle	支持导入数据到Oracle。	参见**配置Oracle目的端参数**。
数据湖探索（DLI）	支持导入数据到DLI服务。	参见**配置DLI目的端参数**。
Elasticsearch或云搜索服务	支持导入数据到云搜索服务。	参见**配置Elasticsearch/云搜索服务（CSS）目的端参数**。
MRS Hudi	支持快速导入数据到MRS的Hudi。	参见**配置MRS Hudi目的端参数**。
MRS Clickhouse	支持快速导入数据到MRS的Clickhouse。	参见**配置MRS ClickHouse目的端参数**。
MongoDB	支持快速导入数据到MongoDB。	参见**配置MongoDB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 5-36 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，以及SFTP/FTP迁移数据到DLI的链路场景），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。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- 支持通过字段映射界面的，可自定义添加常量、变量及表达式。
- 当作业源端为OBS、迁移CSV文件时，并且配置“解析首行为列名”参数的场景下显示列名。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datatime）。
- Hive作为源端数据源时，支持array、map类型的数据读取。
- 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。
- 如果是导入到数据仓库服务（DWS），则还需在目的字段中选择分布列，建议按如下顺序选取分布列：
  1. 有主键可以使用主键作为分布列。
  2. 多个数据段联合做主键的场景，建议设置所有主键作为分布列。
  3. 在没有主键的场景下，如果没有选择分布列，DWS会默认第一列作为分布列，可能会有数据倾斜风险。

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

图 5-37 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。

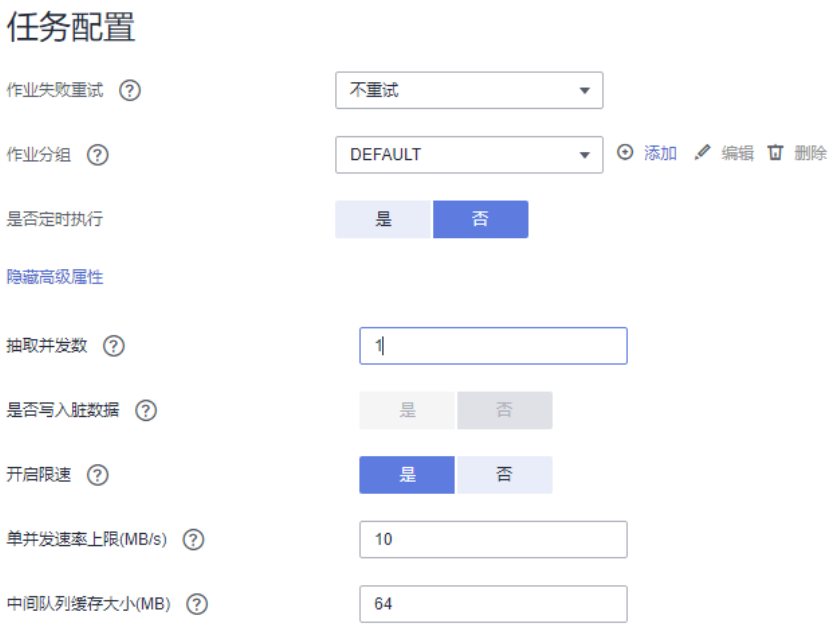
- 字符串反转：自动反转字符串，例如将“ABC”转换为“CBA”。
- 字符串替换：将选定的字符串替换。
- 表达式转换：使用JSP表达式语言（Expression Language）对当前字段或整行数据进行转换，详细请参见**字段转换**。
- 去换行：将字段中的换行符（\n、\r、\r\n）删除。

说明

作业源端开启“使用SQL语句”参数时不支持配置转换器。

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 5-38 任务参数



各参数说明如**表5-61**所示。

表 5-61 任务配置参数

参数	说明	取值样例
作业失败重试	如果作业执行失败，可选择自动重试三次或者不重试。 建议仅对文件类作业或启用了导入阶段表的数据库作业配置自动重试，避免自动重试重复写入数据导致数据不一致。 **说明** 如果通过DataArts Studio数据开发使用参数传递并调度CDM迁移作业时，不能在CDM迁移作业中配置“作业失败重试”参数，如有需要请在数据开发中的CDM节点配置“失败重试”参数。	不重试

参数	说明	取值样例
作业分组	选择作业的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。	DEFAULT
是否定时执行	如果选择“是”，可以配置作业自动启动的时间、重复周期和有效期，具体请参见**配置CDM作业定时任务**。 **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	当前任务从源端进行读取最大线程数。 **说明** 由于数据源限制，实际执行时并发的线程数可能小于等于此处配置的并发数，如CSS，ClickHouse数据源不支持多并发抽取。 CDM通过数据迁移作业，将源端数据迁移到目的端数据源中。其中，主要运行逻辑如下： 1. 数据迁移作业提交运行后，CDM会根据作业配置中的“抽取并发数”参数，将每个作业拆分为多个Task，即作业分片。 **说明** 不同源端数据源的作业分片维度有所不同，因此某些作业可能出现未严格按作业“抽取并发数”参数分片的情况。 2. CDM依次将Task提交给运行池运行。根据集群配置管理中的“最大抽取并发数”参数，超出规格的Task排队等待运行。 因此作业抽取并发数和集群最大抽取并发数参数设置为适当的值可以有效提升迁移速度。 作业抽取并发数的配置原则如下： 1. 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。 2. 表中每行数据大小为1MB以下的可以设置多并发抽取，超过1MB的建议单线程抽取数据。 3. 作业抽取并发数可参考集群最大抽取并发数配置，但不建议超过集群最大抽取并发数上限。 4. 目的端为DLI数据源时，抽取并发数建议配置为1，否则可能会导致写入失败。 其中，集群最大抽取并发数的设置与CDM集群规格有关，并发数上限建议配置为vCPU核数*2。例如8核16GB规格集群的最大抽取并发数上限为16。	1

参数	说明	取值样例
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。 **说明** 脏数据当前仅支持写入到OBS桶路径中。因此仅当已具备OBS连接时，此参数才可以配置。	是
脏数据写入连接	当“是否写入脏数据”为“是”才显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	obs_link
OBS桶	当“脏数据写入连接”为OBS类型的连接时，才显示该参数。 写入脏数据的OBS桶的名称。	dirtydata
脏数据目录	“是否写入脏数据”选择为“是”时，该参数才显示。 OBS上存储脏数据的目录，只有在配置了脏数据目录的情况下才会记录脏数据。 用户可以进入脏数据目录，查看作业执行过程中处理失败的数据或者被清洗过滤掉的数据，针对该数据可以查看源数据中哪些数据不符合转换、清洗规则。	/user/dirtydir
单个分片的最大错误记录数	当“是否写入脏数据”为“是”才显示该参数。 单个map的错误记录超过设置的最大错误记录数则任务自动结束，已经导入的数据不支持回退。推荐使用临时表作为导入的目标表，待导入成功后再改名或合并到最终数据表。	0
开启限速	设置限速可以保护源端读取压力，速率代表CDM传输速率，而非网卡流量。 **说明** • 支持对非二进制文件迁移的作业进行单并发限速。 • 如果作业配置多并发则实际限制速率需要乘以并发数。 • 文件到文件的二进制传输不支持限速功能。	是

参数	说明	取值样例
单并发速率上限(MB/s)	CDM限速并查看作业读写速率。 支持对到HIVE\DLI\JDBC\OBS\HDFS的作业进行单并发限速，如果配置多并发则实际速率限制需要乘以并发数。 **说明** 限制速率为大于1的整数。	20
中间队列缓存大小(MB)	数据写入时中间队列缓存大小，取值范围为1-500，默认值为64。 如果单行数据超过该值，可能会导致迁移失败。如果该值设置过大时，可能会影响集群正常运行。请酌情设置，无特殊场景请使用默认值。例如：64	64

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，stopped。

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

5.6.2 新建整库迁移作业

操作场景

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

说明

整库迁移作业每次运行，会根据整库作业的配置重建子任务，不支持修改子任务后再重新运行主作业。

支持整库迁移的数据源请参见**支持的数据源**。

约束限制

配置源端和目的端参数时，字段名不可包含&和%。

前提条件

- 已新建连接，详情请参见**创建CDM与数据源之间的连接**。
- 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。

图 5-39 创建整库迁移作业

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5-62**所示。

表 5-62 源端作业参数

源端数据库类型	源端参数	参数说明	取值样例
• DWS • MySQL • PostgreSQL • SQL Server • Oracle • SAP HANA	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账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，详细说明请参见**关系数据库增量迁移**。	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

源端数据库类型	源端参数	参数说明	取值样例
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 ${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	"2017-12-31 20:00:00"
	终止时间	终止时间（不包含该值）。格式为'yyyy-MM-dd hh:mm:ss', 支持dateformat时间宏变量函数。 例如："2018-01-01 20:00:00" 或 "${dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00" 或 "${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}"	"2018-01-01 20:00:00"
Redis	键过滤字符	填写键过滤字符后，将迁移符合条件的键。 例如：a*, 迁移所有: *	a*
DDS	数据库名称	待迁移的数据库名称，源连接中配置的用户需要拥有读取该数据库的权限。	ddsdb
	查询筛选	创建用于匹配文档的筛选器。 例如：{HTTPStatusCode:{$gt:"400",$lt:"500"},HTTPMethod:"GET"}。	-

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5-63**所示。

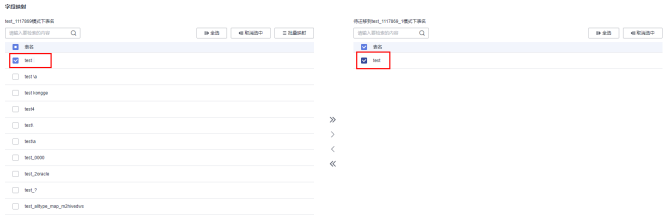
表 5-63 目的端作业参数

目的端数据库类型	目的端参数	参数说明	取值样例
• 云数据库MySQL • 云数据库PostgreSQL • 云数据库SQL Server	-	整库迁移到RDS关系数据库时，目的端作业参数请参见**配置MySQL/SQL Server/PostgreSQL目的端参数**。	schema
DWS	-	整库迁移到DWS时，目的端作业参数请参见**配置DWS目的端参数**。	-
MRS Hive	-	整库迁移到MRS Hive时，目的端作业参数请参见**配置Hive目的端参数**。	hivedb

目的端数据库类型	目的端参数	参数说明	取值样例
MRS HBase CloudTable	-	整库迁移到MRS HBase或CloudTable时，目的端作业参数请参见**配置HBase/CloudTable目的端参数**。	是
Redis	清除数据库	在导入数据前清除数据库数据。	是
DDS	数据库名称	待迁移的数据库名称，源连接中配置的用户需要拥有读取该数据库的权限。	mongodb
	迁移行为	选择新增或替换。	-

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，选择的目的库表名须与源端库表名保持一致。如源端表名为test，则目的端表名只能选择test。

图 5-40 字段映射关系



步骤6 单击“下一步”配置任务参数。

图 5-41 任务参数

各参数说明如表5-64所示。

表 5-64 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	当前任务从源端进行读取最大线程数。 **说明** 由于数据源限制，实际执行时并发的线程数可能小于等于此处配置的并发数，如CSS，ClickHouse数据源不支持多并发抽取。	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

说明

Oracle整库迁移作业场景下，如果源端选择视图或无主键表，且目标端为hudi时，不支持自动建表。

5.6.3 配置 CDM 作业源端参数

5.6.3.1 配置 OBS 源端参数

作业中源连接为**OBS连接**时，源端作业参数如**表5-65**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 5-65 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，具体请参见**文件列表迁移**。 待迁移数据的目录，将迁移目录下的所有文件（包括所有嵌套子目录及其子文件）。 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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

参数类型	参数名	说明	取值样例
	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OBS桶中的绝对路径或目录。 此处建议选择为文件的绝对路径。当选择为目录时，也支持迁移子目录中的文件，但如果目录下文件量过大，可能会导致集群内存不足。	/0521/Lists.txt
	JSON类型	当“文件格式”选择为“JSON格式”时，才有该参数。JSON文件中存储的JSON对象的类型，可以选择“JSON对象”或“JSON数组”。	JSON对象
	记录节点	当“文件格式”选择为“JSON格式”并且“JSON类型”为“JSON对象”时，才有该参数。对该JSON节点下的数据进行解析，如果该节点对应的数据为JSON数组，那么系统会以同一模式从该数组中提取数据。多层嵌套的JSON节点以字符“.”分隔。	data.list
高级属性	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n
	字段分隔符	文件中的字段分隔符，使用Tab键作为分隔符请输入“\t”。当“文件格式”选择为“CSV格式”时，才有该参数。	,
	使用包围符	选择“是”时，包围符内的字段分隔符会被视为字符串值的一部分，目前CDM默认的包围符为："。	否

参数类型	参数名	说明	取值样例
	使用转义符	选择“是”时，CSV数据行中的\作为转义符使用。 选择“否”时，CSV中的\作为数据不会进行转义。CSV只支持\作为转义符。	是
	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，正则表达式写法请参考**正则表达式分隔半结构化文本**。	^(\d.*\d) (\w*) \[(.*) \] ([\w\.]*) (\w.*).*
	前N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文件的前N行数据作为标题行，不写入目的端的表。	否
	标题行数	“前N行为标题行”选择“是”时才有该参数。抽取数据时将被跳过的标题行数。 **说明** 标题行数不为空，取值为1-99之间的整数。	1
	解析首行为列名	“前N行为标题行”选择“是”时才有该参数。选择是否将标题的首行解析为列名，在配置字段映射时会在原字段中显示该列名。 **说明** • 标题行数大于1时，当前仅支持解析标题的首行作为列名。 • 列名不支持“&”字符，否则会导致作业迁移失败，需修改CSV文件“&”字符即可正常迁移。	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。	GBK
	压缩格式	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

参数类型	参数名	说明	取值样例
	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，否则保持原样传输。当输入*或为空时，所有文件都会被解压。	*
	启动作业标识文件	选择是否开启作业标识文件的功能。当源端路径下存在启动作业的标识文件时才启动作业，否则会挂起等待一段时间，等待时长在下方“等待时间”中配置。	否
	标识文件名	选择开启作业标识文件的功能时，需要指定启动作业的标识文件名。指定文件后，只有在源端路径下存在该文件的情况下才会运行任务。该文件本身不会被迁移。	ok.txt
	等待时间	选择开启作业标识文件的功能时，如果源路径下不存在启动作业的标识文件，作业挂机等待的时长，当超时后任务会失败。 等待时间设置为0时，当源端路径下不存在标识文件，任务会立即失败。 单位：秒。	10
	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。具体使用方法可参见**文件增量迁移**。	通配符
	目录过滤器	“过滤类型”选择“通配符”、“正则表达式”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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*input

参数类型	参数名	说明	取值样例
	文件过滤器	“过滤类型”选择“通配符”、“正则表达式”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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*.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等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**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-06-01 00:00:00
	终止时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-01 00:00:00
	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

参数类型	参数名	说明	取值样例
	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，详细请参见**MD5校验文件一致性**。	.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。

5.6.3.2 配置 HDFS 源端参数

作业中源连接为**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5-66**所示。

表 5-66 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。 待迁移数据的目录或单个文件路径。 待迁移数据的目录，将迁移目录下的所有文件（包括所有嵌套子目录及其子文件）。 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	/user/cdm/

参数类型	参数名	说明	取值样例
	文件格式	传输数据时所用的文件格式，可选择以下文件格式： ● CSV格式：以CSV格式解析源文件，用于迁移文件到数据表的场景。 ● 二进制格式：选择“二进制格式”时不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。 ● Parquet格式：以Parquet格式解析源文件，用于HDFS数据导到表的场景。	CSV格式
	列表文件	当“文件格式”选择为“二进制格式”时，才有该参数。 打开列表文件功能时，支持读取OBS桶中文件（如txt文件）的内容作为待迁移文件的列表。该文件中的内容应为待迁移文件的绝对路径（不支持目录），文件内容示例如下： /mrs/job-properties/application_1634891604621_0014/job.properties /mrs/job-properties/application_1634891604621_0029/job.properties	是
	列表文件源连接	当“列表文件”选择为“是”时，才有该参数。可选择列表文件所在的OBS连接。	OBS_test_link
	列表文件OBS桶	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶名。	01
	列表文件或目录	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶中的绝对路径或目录。	/0521/Lists.txt
高级属性	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n
	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。	,

参数类型	参数名	说明	取值样例
	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文件的前N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”。只有文本文件可以设置编码类型，当“文件格式”选择为“二进制格式”时，该参数值无效。	GBK
	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。具体使用方法可参见**文件增量迁移**。	-
	目录过滤器	“过滤类型”选择“通配符”、“正则表达式”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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*input
	文件过滤器	“过滤类型”选择“通配符”、“正则表达式”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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*.csv
	时间过滤	选择“是”时，可以根据文件的修改时间，选择性的传输文件。	是

参数类型	参数名	说明	取值样例
	起始时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间大于等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**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-01 00:00:00
	终止时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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

参数类型	参数名	说明	取值样例
	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。 详细使用方法请参见**迁移文件时加解密**。	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位的十六进制数组成，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，详细请参见**MD5校验文件一致性**。	.md5

5.6.3.3 配置 HBase/CloudTable 源端参数

作业中源连接为**HBase连接**或**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5-67**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 5-67 HBase/CloudTable 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	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_2
	列族	可选参数，导出数据所属的列族。	CF1&CF2
高级属性	切分Rowkey	可选参数，选择是否拆分Rowkey，默认为“否”。	是
	Rowkey分隔符	可选参数，用于拆分Rowkey的分隔符，若不设置则不切分。	|
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-01-01 20:00:00

参数类型	参数名	说明	取值样例
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-02-01 20:00:00

5.6.3.4 配置 Hive 源端参数

作业中源连接为**Hive连接**时，源端作业参数如**表5-68**所示。

表 5-68 Hive 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	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_E

参数类型	参数名	说明	取值样例
	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条件进行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条件进行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条件进行数据过滤及在字段映射页面添加新字段。 **说明** 源端为Hive数据源且使用JDBC方式读取数据时，CDM不支持多并发，即后续操作中抽取并发数只能设置为1。	HDFS
	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
		● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。
			select id,name from sqoop.user;

参数类型	参数名	说明	取值样例
高级属性	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，属性名称为分区名称，属性值可以配置多个值（空格分隔），也可以配置为字段取值范围，接受时间宏函数。详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	• 单/多值过滤场景属性值：${dateformat(yyyyMMdd, -1, DAY)} ${dateformat(yyyyMMdd)} • 范围过滤场景属性值：${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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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。

5.6.3.5 配置 DLI 源端参数

作业中源连接为**DLI连接**时，源端作业参数如**表5-69**所示。

表 5-69 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	用于抽取分区的信息。	• ['year=2020'] • ['year=2020,location=sun'] • ['year=2020,location=sun', 'year=2021,location=earth'] • 读取前一天数据：当前日期为2024-07-16，则['DS=${dateformat(yyyy-MM-dd, -1, DAY)}']表示抽取DS分区值为2024-07-15的数据。 其他场景请参见**时间宏变量使用解析**。

5.6.3.6 配置 FTP/SFTP 源端参数

作业中源连接为**FTP/SFTP连接**时，源端作业参数如**表5-70**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 5-70 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，具体请参见**文件列表迁移**。 待迁移数据的目录，将迁移目录下的所有文件（包括所有嵌套子目录及其子文件）。 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	/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。 **说明** 当目的端为OBS数据源时，仅支持配置二进制格式。	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

参数类型	参数名	说明	取值样例
	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“.”分隔。	data.list
高级属性	使用rfc4180解析器	当“文件格式”选择为“CSV格式”时，才有该参数。是否使用rfc4180解析器解析CSV文件。	否
	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
	使用转义符	选择“是”时，CSV数据行中的\作为转义符使用。 选择“否”时，CSV中的\作为数据不会进行转义。CSV只支持\作为转义符。	是
	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	当“使用正则表达式分隔字段”选择为“是”时，才有该参数。 分隔字段的正则表达式，正则表达式写法请参考**正则表达式分隔半结构化文本**。	^(\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文件的前N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

参数类型	参数名	说明	取值样例
	压缩格式	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，否则保持原样传输。当输入*或为空时，所有文件都会被解压。	*
	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
	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|。	|
	标识文件名	选择开启作业标识文件的功能时，需要指定启动作业的标识文件名。指定文件后，只有在源端路径下存在该文件的情况下才会运行任务。该文件本身不会被迁移。	ok.txt
	等待时间	选择开启作业标识文件的功能时，如果源路径下不存在启动作业的标识文件，作业挂机等待的时长，当超时后任务会失败。 等待时间设置为0时，当源端路径下不存在标识文件，任务会立即失败。 单位：秒。	10
	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。具体使用方法可参见**文件增量迁移**。	无
	目录过滤器	“过滤类型”选择“通配符”和“正则表达式”时，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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*input,*out

参数类型	参数名	说明	取值样例
	文件过滤器	“过滤类型”选择“通配符”和“正则表达式”时，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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*.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	“时间过滤”选择“是”时，可以指定一个时间值，当文件的修改时间大于等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**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-01 00:00:00
	终止时间	“时间过滤”选择“是”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-30 00:00:00
	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

参数类型	参数名	说明	取值样例
	标识文件类型	选择开启作业标识文件的功能时，该参数才显示。 • MARK_DONE：只有在源端路径下存在标识文件的情况下才会执行迁移任务。 • MARK_DOING：只有在源端路径下不存在标识文件的情况下才会执行迁移任务。	MARK_DOING
	是否跳过空行	“文件格式”选择“CSV格式”时，该参数才显示。 如果某行数据为空，则跳过此行。	否
	null值	“文件格式”选择“二进制格式”时，该参数才显示。 由于文本文件中无法用字符串定义null值，此配置项定义将何种字符串标识为null。	否
	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，详细请参见**MD5校验文件一致性**。	.md5

5.6.3.7 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5-71**所示。当前只支持从HTTP URL导出数据，不支持导入。

表 5-71 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。	https://bucket.obs.myhuaweicloud.com/object-key
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中存储文本文件的文件自定义目录，多级目录可用“/”进行分隔。	test1

参数名	说明	取值样例
文件格式	传输数据时使用的格式。其中CSV和JSON仅支持迁移到数据表场景，二进制格式适用于文件迁移场景。	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
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
MD5文件名后缀	校验CDM抽取的文件，是否与源文件一致，详细请参见**MD5校验文件一致性**。	.md5
QUERY参数	此字段为true时，则上传对象时使用的对象名为去掉query参数的字符。	否

5.6.3.8 配置 PostgreSQL/SQL Server 源端参数

作业中源连接为从云数据库 PostgreSQL、云数据库 SQL Server、PostgreSQL、Microsoft SQL Server导出的数据时，源端作业参数如**表5-72**所示。

表 5-72 PostgreSQL/SQL Server 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

参数类型	参数名	说明	取值样例
	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 ● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。	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账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： ● **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 ● ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 ● ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E

参数类型	参数名	说明	取值样例
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **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。	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'

参数类型	参数名	说明	取值样例
	分区字段是否允许空值	是否允许分区字段包含空值。	是
	按表分区抽取	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”参数指定具体的表分区。 • 该功能不支持非分区表。 • 仅支持源端数据源为PostgreSQL时配置该参数。 • 数据库用户需要具有系统视图**dba_tab_partitions**和**dba_tab_subpartitions**的**SELECT**权限。	否
	拆分作业	选择“是”，会根据“作业拆分字段”值，将作业拆分为多个子作业并发执行。 **说明** 仅支持目的端为DLI和Hive时配置该参数及*作业拆分字段*、*拆分字段最小值*、*拆分字段最大值*、*子作业个数*参数。	是
	作业拆分字段	“拆分作业”选择“是”时，显示该参数，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	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最小值。	-
	拆分字段最大值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最大值。	-
	子作业个数	“拆分作业”选择“是”时，显示该参数，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。	-

5.6.3.9 配置 DWS 源端参数

作业中源连接为**DWS连接**时，源端作业参数如**表5-73**所示。

表 5-73 DWS 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

参数类型	参数名	说明	取值样例
	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 ● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。	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账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： ● **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 ● ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 ● ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E

参数类型	参数名	说明	取值样例
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **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	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	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。	id

参数类型	参数名	说明	取值样例
	分区字段含有空值	是否允许分区字段包含空值。	是
	拆分作业	选择“是”，会根据“作业拆分字段”值，将作业拆分为多个子作业并发执行。 **说明** 仅支持目的端为DLI和Hive时配置该参数及*作业拆分字段*、*拆分字段最小值*、*拆分字段最大值*、*子作业个数*参数。	是
	作业拆分字段	“拆分作业”选择“是”时，显示该参数，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	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最小值。	-
	拆分字段最大值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最大值。	-
	子作业个数	“拆分作业”选择“是”时，显示该参数，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。	-

5.6.3.10 配置 SAP HANA 源端参数

SAP HANA作为源端作业参数如**表5-74**所示。

表 5-74 SAP HANA 作源端时的作业参数

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

参数类型	参数名	说明	取值样例
	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 ● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。	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账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： ● **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 ● ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 ● ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E

参数类型	参数名	说明	取值样例
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **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	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	抽取区分字段	“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。	id

5.6.3.11 配置 MySQL 源端参数

作业中源连接为**云数据库MySQL/MySQL数据库连接**时，源端作业参数如**表5-75**所示。

表 5-75 MySQL 作为源端时的作业参数

参数名	说明	取值样例
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
	• 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。
		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账号是否有元数据查询的权限。 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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 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。例如：表名配置为*user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
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。	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
date类型值是否保留一位精度	date类型值是否保留一位精度。	是
分区字段含有空值	是否允许分区字段包含空值。	是

参数名	说明	取值样例
拆分作业	选择“是”，会根据“作业拆分字段”值，将作业拆分为多个子作业并发执行。 **说明** 仅支持目的端为DLI和Hive时配置该参数及*作业拆分字段*、*拆分字段最小值*、*拆分字段最大值*、*子作业个数*参数。	是
作业拆分字段	“拆分作业”选择“是”时，显示该参数，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	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最小值。	-
拆分字段最大值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最大值。	-
子作业个数	“拆分作业”选择“是”时，显示该参数，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

5.6.3.12 配置 Oracle 源端参数

作业中源连接为**Oracle数据库连接**，源端作业参数如**表5-76**所示。

表 5-76 Oracle 作为源端时的作业参数

参数名	说明	取值样例
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

参数名	说明	取值样例
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 ● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。	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账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： ● **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 ● ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 ● ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E

参数名	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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **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
抽取分区字段	“按表分区抽取”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。	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
分区字段含有空值	“按表分区抽取”选择“否”时，显示该参数，表示是否允许分区字段包含空值。	是

参数名	说明	取值样例
按表分区抽取	从Oracle导出数据时，支持从分区表的各个分区并行抽取数据。启用该功能时，可以通过下面的“表分区”参数指定具体的Oracle表分区。 • 该功能不支持非分区表。 • 数据库用户需要具有系统视图**dba_tab_partitions**和**dba_tab_subpartitions**的**SELECT**权限。	否
表分区	输入需要迁移数据的Oracle表分区，多个分区以&分隔，不填则迁移所有分区。 如果有子分区，以“分区.子分区”的格式填写，例如“P2.SUBP1”。	P0&P1&P2.SUBP1&P2.SUBP3
拆分作业	选择“是”，会根据“作业拆分字段”值，将作业拆分为多个子作业并发执行。 **说明** 仅支持目的端为DLI和Hive时配置该参数及*作业拆分字段*、*拆分字段最小值*、*拆分字段最大值*、*子作业个数*参数。	是
作业拆分字段	“拆分作业”选择“是”时，显示该参数，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	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最小值。	-
拆分字段最大值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最大值。	-
子作业个数	“拆分作业”选择“是”时，显示该参数，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。

5.6.3.13 配置分库源端参数

作业中源连接为**分库连接**，源端作业参数如**表5-77**所示。

表 5-77 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账号是否有元数据查询的权限。 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。例如：表名配置为 *user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。	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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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)}

5.6.3.14 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**MongoDB连接**时，即从本地MongoDB或DDS导出数据时，源端作业参数如**表5-78**所示。

表 5-78 MongoDB/DDS 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	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，或者对应连接里的账号是否有元数据查询的权限。	COLLECTION
高级属性	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'}

5.6.3.15 配置 Redis 源端参数

第三方云的Redis服务无法支持作为源端。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。

作业中源连接为从本地Redis导出的数据时，源端作业参数如**表5-79**所示。

表 5-79 Redis 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	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
高级属性	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

5.6.3.16 配置 DIS 源端参数

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。

作业中源连接为**DIS连接**时，源端作业参数如所**表5-80**示。

表 5-80 DIS 作为源端时的作业参数

参数类型	参数	说明	取值样例
基本参数	DIS通道	DIS的通道名。	dis
	是否持久运行	用户自定义是否永久运行。设置为长久运行的任务，如果DIS系统发生中断，任务也会失败结束。	是
	DIS分区ID	DIS分区ID，该参数支持输入多个分区ID，使用英文逗号（,）分隔。	0,1,2

参数类型	参数	说明	取值样例
	偏移量参数	设置从DIS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。 ● 上次停止处：从上次停止处继续读取。 ● 最早：最小偏移量，即拉取最早的数据。	最新
	APP名字	配置用户数据消费程序的唯一标识符，不存在时会自动创建。	cdm
	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格式：以JSON格式解析源数据。	二进制格式
	字段分隔符	数据格式为“CSV格式”时呈现此参数。默认为逗号，使用Tab键作为分隔符请输入“\t”。	,
	记录分隔符	数据格式为“CSV格式”或“JSON格式”时呈现此参数。用于配置每条记录之间的分隔符。	,
高级属性	最大消息数/poll	可选参数，每次向DIS请求数据限制最大请求记录数。	100

5.6.3.17 配置 Kafka/DMS Kafka 源端参数

作业中源连接为**Kafka连接**或**DMS Kafka连接**时，源端作业参数如**表5-81**所示。

表 5-81 Kafka 作为源端时的作业参数

参数类型	参数	说明	取值样例
基本参数	Topics	支持单个或多个topic。	est1,est2

参数类型	参数	说明	取值样例
	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）：以DRS格式解析源数据。 ● CDC（JSON）：以JSON格式解析源数据。 ● CDC（DRS_AVRO）：以DRS_AVRO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
	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
	抽取数据最大运行时间	持续拉取数据时间。如天调度作业，根据每天topic产生的数据量，配置足够的拉取时间。单位：分钟。	60
	等待时间	当配置为60时，如果消费者60s内从Kafka拉取数据返回一直为空（一般是已经读完主题中的全部数据，也可能是网络或者Kafka集群可用性原因），则立即停止任务，否则持续重试读取数据。单位：秒。	60
	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
	开始时间(>=)	“偏移量参数”选择为“时间范围”时配置。拉取数据的开始时间，包含设置时间点的数据。	2020-12-20 12:00:00
	结束时间(<)	“偏移量参数”选择为“时间范围”时配置。拉取数据的结束时间，不包含设置时间点的数据。	2020-12-20 20:00:00
	字段分隔符	“数据格式”选择为“CSV格式”时配置。默认为空格，使用Tab键作为分隔符请输入“\t”。	,
	记录分隔符	“数据格式”选择为“CSV格式”、“JSON”时配置。默认为空格，使用Tab键作为分隔符请输入“\t”。	,

参数类型	参数	说明	取值样例
高级参数	使用配置文件	“数据格式”选择为“CDC场景”时配置，用于配置OBS文件。	否
	OBS链接	选择OBS连接器信息。	obs_link
	OBS桶	选择OBS桶。	obs_test
	配置文件	选择OBS的配置文件。	/obs/config.csv
	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
	通知Topic	发送通知数据到通知Topic中。在CDC场景中，通知的内容是记录生成文件列表的文件名。	notice

5.6.3.18 配置 Elasticsearch/云搜索服务源端参数

作业中源连接为**Elasticsearch连接参数说明**或**云搜索服务（CSS）连接参数说明**时，源端作业参数如**表5-82**所示。

表 5-82 Elasticsearch/云搜索服务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	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。 **说明** Elasticsearch搜索引擎7.x及以上版本不支持自定义类型，只能使用_doc类型。此处即使自定义也不会生效。	_doc
高级属性	拆分nested类型字段	可选参数，选择是否将nested字段的json内容拆分，例如：将“a:{ b:{ c:1, d:{ e:2, f:3 } } }”拆成三个字段“a.b.c”、“a.b.d.e”、“a.b.d.f”。	否

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

参数类型	参数名	说明	取值样例
		• 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
	分页大小	Elasticsearch分页查询，用来设置分页size的大小。	1000
	ScrollId超时时间配置	Elasticsearch scroll查询时会记录一个scroll_id，超时或者scroll查询结束后会清除请求的srcoll_id，通过设置这个超时时间配置，来指定scroll_id超时时间。	5

5.6.3.19 配置 OpenTSDB 源端参数

作业中源连接为**CloudTable OpenTSDB连接**时，源端作业参数如**表5-83**所示。

表 5-83 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

5.6.3.20 配置 MRS Hudi 源端参数

作业中源连接为**MRS Hudi连接**时，源端作业参数如**表5-84**所示。

表 5-84 MRS Hudi 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	选择已配置的MRS Hudi连接。	hudi_from_cdm
	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Hudi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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_E
高级属性	Where子句	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	age > 18 and age <= 60

5.6.3.21 配置 MRS ClickHouse 源端参数

作业中源连接为**MRS ClickHouse连接**时，源端作业参数如**表5-85**所示。

表 5-85 MRS ClickHouse 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	选择已配置的MRS ClickHouse连接。	ck_from_cdm

参数类型	参数名	说明	取值样例
	模式或表空间	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账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。	default
	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，或者对应连接里的账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。	TBL_E
高级属性	Where子句	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	age > 18 and age <= 60

5.6.3.22 配置神通（ST）源端参数

从神通（ST）导出数据时，源端作业参数如**表5-86**所示。

表 5-86 神通（ST）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

参数类型	参数名	说明	取值样例
	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 ● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。	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账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： ● **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 ● ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 ● ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E

参数类型	参数名	说明	取值样例
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **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。	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'

参数类型	参数名	说明	取值样例
	分区字段含有空值	是否允许分区字段包含空值。	是

5.6.3.23 配置达梦数据库 DM 源端参数

从达梦数据库 DM导出数据时，源端作业参数如表5-87所示。

表 5-87 达梦数据库 DM 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
		• 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。
			select id,name from sqoop.user;

参数类型	参数名	说明	取值样例
	模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的账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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **说明** 表名支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有表（要求表中的字段个数和类型都一样）。例如： • **table***表示导出所有以“table”开头的表。 • ***table**表示导出所有以“table”结尾的表。 • ***table***表示表名中只要有“table”字符串，就全部导出。	table

参数类型	参数名	说明	取值样例
高级属性	抽取分区字段	“使用SQL语句”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** • 抽取分区字段支持CHAR、VARCHAR、LONGVARCHAR、TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。 • 当选择CHAR、VARCHAR、LONGVARCHAR抽取分区字段类型时，字段值不支持ASCII字符代码表之外的字符，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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	分区字段含有空值	是否允许分区字段包含空值。	是

5.6.3.24 配置 YASHAN 源端参数

作业中源连接从YASHAN导出的数据时，源端作业参数如**表5-88**所示。

表 5-88 YASHAN 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

参数类型	参数名	说明	取值样例
	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 ● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。	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账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： ● **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 ● ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 ● ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E

参数类型	参数名	说明	取值样例
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **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	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	date类型值是否保留一位精度	date类型值是否保留一位精度。	否

参数类型	参数名	说明	取值样例
	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。	id
	分区字段含有空值	是否允许分区字段包含空值。 多并发抽取时，若确定分区字段不含Null，将该值设为“否”可提升性能，若不确定，请设为“是”，否则可能会丢数据。	否
	拆分作业	选择“是”，会根据“作业拆分字段”值，将作业拆分为多个子作业并发执行。 **说明** 仅支持目的端为DLI和Hive时配置该参数及*作业拆分字段*、*拆分字段最小值*、*拆分字段最大值*、*子作业个数*参数。	否
	作业拆分字段	“拆分作业”选择“是”时，显示该参数，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	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最小值。	-
	拆分字段最大值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最大值。	-
	子作业个数	“拆分作业”选择“是”时，显示该参数，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。	-

5.6.4 配置 CDM 作业目的端参数

5.6.4.1 配置 OBS 目的端参数

作业中目的连接为**OBS连接**时，即导入数据到云服务OBS时，目的端作业参数如**表5-89**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 5-89 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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。 **说明** • 当源端为MRS Hive数据源时，仅支持配置CSV格式。 • 当源端为FTP/SFTP数据源时，仅支持配置二进制格式。	CSV格式
	重复文件处理方式	当源端为HDFS数据源时配置。 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 具体使用方法可参见**文件增量迁移**。	跳过重复文件

参数类型	参数名	说明	取值样例
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。 详细使用方法请参见**迁移文件时加解密**。	KMS
	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
	复制Content-Type属性	“文件格式”为“二进制”，且源端、目的端都为对象存储时，才有该参数。 选择“是”后，迁移对象文件时会复制源文件的Content-Type属性，主要用于静态网站的迁移场景。 归档存储的桶不支持设置Content-Type属性，所以如果开启了该参数，目的端选择写入的桶时，必须选择非归档存储的桶。	否
	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。“文件格式”为“二进制格式”时该参数值无效。	\n
	字段分隔符	文件中的字段分隔符。“文件格式”为“二进制格式”时该参数值无效。	,
	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

参数类型	参数名	说明	取值样例
	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，具体请参见**MD5校验文件一致性**。	是
	记录校验结果	当选择校验MD5值时，可以选择是否记录校验结果。	是
	校验结果写入连接	可以指定任意一个OBS连接，将MD5校验结果写入该连接的桶下。	obslink
	OBS桶	写入MD5校验结果的OBS桶。	cdm05
	写入目录	写入MD5校验结果的目录。	/md5/
	编码类型	文件编码类型，例如：“UTF-8”或“GBK”。“文件格式”为“二进制格式”时该参数值无效。	GBK
	使用包围符	“文件格式”为“CSV格式”，才有该参数，用于将数据库的表迁移到文件系统的场景。 选择“是”时，如果源端数据表中的某一个字段内容包含字段分隔符或换行符，写入目的端时CDM会使用双引号（"）作为包围符将该字段内容括起来，作为一个整体存储，避免其中的字段分隔符误将一个字段分隔成两个，或者换行符误将字段换行。例如：数据库中某字段为hello,world，使用包围符后，导出到CSV文件的时候数据为"hello,world"。	否
	首行为标题行	从关系型数据库导出数据到OBS，“文件格式”为“CSV格式”时，才有该参数。 在迁移表到CSV文件时，CDM默认是不迁移表的标题行，如果该参数选择“是”，CDM在才会将表的标题行数据写入文件。	否
	作业成功标识文件	当作业执行成功时，会在写入目录下生成一个标识文件，文件名由用户指定。不指定时默认关闭该功能。	finish.txt
	文件夹模式	从关系型数据库导出数据到OBS，才有该参数。 启用后将会以根目录-表名-数据类型-数据的文件夹模型生成文件。例如：raw_schema/tbl_student/datas/tbl_student_1.csv	是

参数类型	参数名	说明	取值样例
	Blog/Clog文件扩展名	“文件夹模式”为“是”时，才有该参数。文件夹模式下自定义Blob/Clog数据的文件扩展名。	.dat/.jpg/.png
	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。 **说明** 源端为关系型数据库数据源时，目录层次为源端表名+自定义目录，其他场景下为自定义目录。	${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}
	自定义文件名	从关系型数据库导出数据到OBS，且“文件格式”为“CSV格式”时，才有该参数。 用户可以通过该参数自定义OBS端生成的文件名，支持以下自定义方式： ● 字符串，支持特殊字符。例如“cdm#”，则生成的文件名为“cdm#.csv”。 ● 时间宏，例如“${timestamp()}”，则生成的文件名为“1554108737.csv”。 ● 表名宏，例如“${tableName}”，则生成的文件名为源表名“sqltabname.csv”。 ● 版本宏，例如“${version}”，则生成的文件名为集群版本号“2.9.2.200.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_2.9.2.200.csv”。	cdm

5.6.4.2 配置 HDFS 目的端参数

作业中目的连接为**HDFS连接**时，目的端作业参数如**表5-90**所示。

表 5-90 HDFS 作为目的端时的作业参数

参数名	说明	取值样例
写入目录	写入数据到HDFS服务器的目录。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	/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	当源端为文件类数据源（HTTP/FTP/SFTP/HDFS/OBS）时配置。 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。	,

参数名	说明	取值样例
使用包围符	“文件格式”为“CSV格式”，才有该参数，用于将数据库的表迁移到文件系统的场景。 选择“是”时，如果源端数据表中的某一个字段内容包含字段分隔符或换行符，写入目的端时CDM会使用双引号（"）作为包围符将该字段内容括起来，作为一个整体存储，避免其中的字段分隔符误将一个字段分隔成两个，或者换行符误将字段换行。例如：数据库中某字段为hello,world，使用包围符后，导出到CSV文件的时候数据为"hello,world"。	否
首行为标题行	在迁移表到CSV文件时，CDM默认是不迁移表的标题行，如果该参数选择“是”，CDM在才会将表的标题行数据写入文件。	否
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。 **说明** 源端为关系型数据库数据源时，目录层次为源端表名+自定义目录，其他场景下为自定义目录。	${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。 详细使用方法请参见**迁移文件时加解密**。	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

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。 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。

5.6.4.3 配置 HBase/CloudTable 目的端参数

作业中目的连接为**HBase连接**或**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5-91**所示。

表 5-91 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，支持从源端复制字段名。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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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
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。 **说明** Hbase自动建表包含列族与协处理器Coprocessor信息。其他属性按默认值设置，不跟随源端。	不自动创建

参数名	说明	取值样例
Row key拼接分隔符	可选参数，用于多列合并作为rowkey，默认为空格。	,
Rowkey冗余	可选参数，是否将选做Rowkey的数据同时写入HBase的列，默认值“否”。	否
压缩算法	可选参数，创建新HBase表时采用的压缩算法，默认为值“NONE”。 ● NONE：不压缩。 ● SNAPPY：压缩为Snappy格式。 ● GZ：压缩为GZ格式。	NONE
WAL开关	选择是否开启HBase的预写日志机制（WAL，Write Ahead Log）。 ● 是：开启后如果出现HBase服务器宕机，则可以从WAL中回放执行之前没有完成的操作。 ● 否：关闭时能提升写入性能，但如果HBase服务器宕机可能会造成数据丢失。	否
匹配数据类型	● 是：源端数据库中的Short、Int、Long、Float、Double、Decimal类型列的数据，会转换为Byte[]数组（二进制）写入HBase，其他类型的按字符串写入。如果这几种类型中，有合并做rowkey的，则依然当字符串写入。 该功能作用是：降低存储占用空间，存储更高效；特定场景下rowkey分布更均匀。 ● 否：源端数据库中所有类型的数据，都会按照字符串写入HBase。	否

5.6.4.4 配置 Hive 目的端参数

作业中目的连接为**Hive连接**时，目的端作业参数如**表5-92**所示。

表 5-92 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

参数名	说明	取值样例
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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_X
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。 **说明** ● 自动建表只同步列注释，表注释不会被同步。 ● 自动建表不支持同步主键。	不自动创建
源端null值转换值	将源端null值转换为其他值。 ● TO_NULL ● TO_EMPTY_STRRING ● TO_NULL_STRING	TO_NULL
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
换行符处理方式	对于写入Hive textfile格式表的数据中存在换行符的场景，指定对换行符的处理策略。 ● 删除 ● 替换为其他字符串 ● 不处理	删除
Hive表分区字段	“自动创建”设置为“不自动创建”时，无该此参数。 对Hive建表设置分区字段，多个值以逗号隔开。	A,B

参数名	说明	取值样例
表路径	“自动创建”设置为“不自动创建”时，无该此参数。 表路径。	-
存储格式	“自动创建”设置为“不自动创建”时，无该此参数。 选择存储格式。 ● 行式存储格式：TEXTFILE。 ● 列式存储格式：ORC、RCFILE、PARQUET。 TEXTFILE使用明文存储，当数据存在特殊字符的场景下可能会导致数据写入错乱，请谨慎使用。建议优先使用ORC存储格式。	ORC
hive表清理数据模式	“导入前清空数据”设置为“是”时，呈现此参数。 选择Hive表清理数据模式。 ● LOAD_OVERWRITE模式：将生成一个临时数据文件目录，使用Hive的load overwrite语法将临时目录加载到Hive表中。 ● TRUCATE模式：只清理分区下的数据文件，不删除分区。 **说明** 目的端为分区表时，Hive表清理数据模式建议设置为LOAD_OVERWRITE模式，否则可能会有集群内存过载/磁盘过载的风险。	TRUCATE
分区信息	“导入前清空数据”设置为“是”时，呈现此参数。目的端为分区表时，必须指定分区。 ● 当使用TRUCATE模式：只清理分区下的数据文件。 ● 当使用LOAD_OVERWRITE模式：覆盖写入到指定分区，仅支持指定单分区。	单分区： year=2020,location=sun; 多分区： ['year=2020,location=sun', 'year=2021,location=earth']. 前一日分区： day='${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}',

参数名	说明	取值样例
执行Analyze语句	数据全部写入完成后会异步执行ANALYZE TABLE语句，用于优化Hive表查询速度。 执行的SQL如下： ● 非分区表：ANALYZE TABLE tablename COMPUTE STATISTICS ● 分区表：ANALYZE TABLE tablename PARTITION(partcol1[=val1], partcol2[=val2], ...) COMPUTE STATISTICS **说明** “执行Analyze语句”参数配置仅用于单表迁移场景。 执行Analyze语句可能会对Hive造成压力。	是
内部写队列内存最大值	当出现内存不足场景时，请酌情修改该参数，当参数过小时，会影响迁移速率。 取值范围是1-128，默认为空，不做限制，单位为MB，超出范围会设置为不限制。	16
内部转换队列内存最大值	当出现内存不足场景时，请酌情修改该参数，当参数过小时，会影响迁移速率。 取值范围是1-128，默认为空，不做限制，单位为MB，超出范围会设置为不限制。	16

说明

- 源端Hive包含array和map类型时，目的端表格式只支持ORC和parquet复杂类型。若目的端表格式为RC和TEXT时，会对源数据进行处理，支持成功写入。
- 因map类型为无序的数据结构，迁移到目的端的数据类型可能跟源端顺序不一致。
- 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;
```

5.6.4.5 配置 MySQL/SQL Server/PostgreSQL 目的端参数

当作业将数据导入到MySQL/SQL Server/PostgreSQL时，目的端作业参数如表5-93所示。

表 5-93 MySQL、SQL Server、PostgreSQL 作为目的端时的作业参数

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

参数类型	参数名	说明	取值样例
	表名	写入数据的目标表名，单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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	导入数据到云数据库 MySQL且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
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，具体请参见**事务模式迁移**。 默认为“否”，CDM直接将数据导入到目的表。这种情况下如果作业执行失败，已经导入到目标表中的数据不会自动回滚。 **说明** 如果“导入开始前”选择“清除部分数据”或“清除全部数据”，CDM的事务模式不会回滚已经删除的数据。	否

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

5.6.4.6 配置 Oracle 目的端参数

作业中目的连接为**Oracle数据库连接**时，目的端作业参数如**表5-94**所示。

表 5-94 Oracle 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

参数类型	参数名	说明	取值样例
	表名	写入数据的目标表名，单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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
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，具体请参见**事务模式迁移**。 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
	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
	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

5.6.4.7 配置 DWS 目的端参数

作业中目的连接为**DWS连接**时，目的端作业参数如**表5-95**所示。

表 5-95 目的端为 DWS 时的作业参数

参数名	说明	取值样例
模式或表空间	待写入数据的数据库名称，支持自动创建Schema。单击输入框后面的按钮可选择模式或表空间。	schema
自动创表	只有当源端为关系数据库时，才有该参数。表示写入表数据时，用户选择的操作： ● 不自动创建：不自动建表。 ● 不存在时创建：当目的端的数据库没有“表名”参数中指定的表时，CDM会自动创建该表。如果“表名”参数配置的表已存在，则不创建，数据写入到已存在的表中。 ● 先删除后创建：CDM先删除“表名”参数中指定的表，然后再重新创建该表。 当选择在DWS端自动创表时，DWS的表与源表的字段类型映射关系见**在DWS端自动建表时的字段类型映射**。 **说明** 自动建表只同步列注释，表注释不会被同步。	不自动创建
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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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

参数名	说明	取值样例
导入模式	导入数据到DWS时，用户可以指定导入模式： ● COPY模式，源数据经过管理节点后，复制到DWS的DataNode节点。 ● UPSERT模式，数据发生主键或唯一约束冲突时，更新除了主键和唯一约束列的其他列数据。	COPY
导入开始前	导入数据前，选择是否清除目的表的数据： ● 不清除：写入数据前不清除目标表中数据，数据追加写入。 ● 清除全部数据：写入数据前会清除目标表中数据。 ● 清除部分数据：需要配置“where条件”参数，CDM根据条件选择性删除目标表的数据。	清除部分数据
where条件	**“导入开始前”**参数选择为“清除部分数据”时配置，配置后导入前根据where条件删除目的表的数据。	age > 18 and age <= 60
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	当选择自动创表时，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。 **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

参数名	说明	取值样例
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
loader线程数	每个loader内部启动的线程数，可以提升写入并发数。	1

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5-42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 5-42 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	SAP HANA	DWS
NUMBER(p,0) (p=3 or p=5)	SMALLINT,TINYINT	SMALLINT,TINYINT	SMALLINT	SMALLINT,TINYINT	SMALLINT
NUMBER(10,0)	INT	INT	INTEGER	INTEGER	INTEGER
NUMBER(19,0)	BIGINT	BIGINT	BIGINT	BIGINT	BIGINT
无	无	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(p,s) (0 < p <= 38)	NUMERIC(p,s) (p <= 1000)
RAW	BINARY	BINARY	BYTEA	BINARY	BYTEA
CHAR	CHAR	CHAR	CHAR	CHAR(p) (p <= 2000)	CHAR
NCHAR	NCHAR	NCHAR	NCHAR	NCHAR(p) (p <= 5000)	NCHAR
DATE	DATE	DATE	DATE	DATE	DATE
DATE	DATETIME	DATETIME2	TIMESTAMP	TIMESTAMP	TIMESTAMP
VARCHAR2(p) (p <= 4000)	VARCHAR	VARCHAR(p) (if p >= 8000 p=max)	VARCHAR(p) (p <= 10485760)	VARCHAR(p) (p <= 5000)	VARCHAR(p) (p <= 10485760)
FLOAT	DOUBLE	FLOAT	DOUBLE PRECISION	DOUBLE	DOUBLE PRECISION
FLOAT	REAL	FLOAT	REAL	REAL	REAL
CLOB	TEXT	TEXT	TEXT	CLOB	TEXT
DATE	无	TIME	TIME	TIME	TIME
BOOLEAN	无	无	BOOLEAN	BOOLEAN	BOOLEAN

说明

自动建表场景不支持创建索引。

5.6.4.8 配置 DDS 目的端参数

作业中目的连接为**DDS连接**时，即导入数据到文档数据库服务（DDS）时，目的端作业参数如**表5-96**所示。

表 5-96 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。	ddsdb
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，或者对应连接里的账号是否有元数据查询的权限。	COLLECTION

5.6.4.9 配置 Redis 目的端参数

当作业将数据导入到Redis时，目的端作业参数如**表5-97**所示。

表 5-97 Redis 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
是否以列值作为field	当值存储类型为HASH时显示此参数。仅支持Hash，如果打开开关，除主键列外，按字段顺序交替取值作为field和value。	是
写入前将相同的键删除	写入前将相同的键删除。 • 否：如果原来Redis已存在类型不同的同名key，则迁移作业会跳过该key。 • 是：Redis会先删除原有的同名key，再执行迁移。	否
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。	;
key值有效期	用于设置统一的生存时间，单位：秒。	300

5.6.4.10 配置 Elasticsearch/云搜索服务（CSS）目的端参数

作业中目的连接为**Elasticsearch连接参数说明**或**云搜索服务（CSS）连接参数说明**时，即将数据导入到Elasticsearch/云搜索服务（CSS）时，目的端作业参数如**表5-98**所示。

须知

表/文件迁移和整库迁移时需配置的参数不同，下表参数为表/文件迁移时的全量参数，实际参数以界面显示为准。

表 5-98 Elasticsearch/云搜索服务（CSS）作为目的端时的作业参数

参数名	说明	取值样例
索引	待写入数据的Elasticsearch的索引，类似关系数据库中的数据库名称。CDM支持自动创建索引和类型，索引和类型名称只能全部小写，不能有大写。	index
类型	待写入数据的Elasticsearch的类型，类似关系数据库中的表名称。类型名称只能全部小写，不能有大写。 **说明** Elasticsearch搜索引擎7.x及以上版本不支持自定义类型，只能使用_doc类型。此处即使自定义也不会生效。	type
操作	操作类型。 • INDEX：不指定主键，es内部生成id，使得每次写入都是不同id的新增数据文件。 • CREATE：需要指定主键。如果主键已经存在，写入失败。 • UPDATE：需要指定主键。如果主键已经存在，覆盖原有数据。 • UPSERT：需要指定主键。如果主键已经存在，覆盖原有数据；如果主键不存在，则新建文档写入。	INDEX
管道ID	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。 目的端为Elasticsearch时需要先在kibana中创建管道ID。 目的端为CSS时不需要创建管道ID，此参数填写配置文件名称，默认为name。	目的端为Elasticsearch时：pipeline_id 目的端为CSS时：name（name为配置文件名称）
开启路由	开启路由后，支持指定某一列的值作为路由写入Elasticsearch。 **说明** 开启路由前建议先建好目的端索引，可提高查询效率。	否

参数名	说明	取值样例
路由字段	**“开启路由”**参数选择为“是”时配置，用于配置目的端路由字段。目的端索引存在但是获取不到字段信息时，支持手动填写字段。路由字段允许为空，为空时写入Elasticsearch不指定routing值。	value1
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

5.6.4.11 配置 DLI 目的端参数

作业中目的连接为**DLI连接**时，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5-99**所示。

⚠ 注意

使用CDM服务迁移数据到DLI时，DLI要在OBS的*dli-trans**内部临时桶生成数据文件，因此在需要赋予DLI连接中使用AK/SK所在用户对*dli-trans**桶的读、写、创建目录对象等权限，否则会导致迁移失败。*dli-trans**内部临时桶的权限策略添加请参见**新增dli-trans*内部临时桶授权策略**。

表 5-99 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。 新建队列操作请参考**创建队列**。	cdm
数据库名称	写入数据的数据库名称。	dli
表名	写入数据的表名。	car_detail
导入前清空数据	选择导入前是否清空目的表的数据。 如果设置为是，任务启动前会清除目标表中数据。	否
空字符串作为null	如果设置为true，空字符串将作为null。	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。 **说明** 当源端为Kafka时，如果DLI导入前清空数据，则不支持INSERT_OVERWRITE。	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

新增 *dli-trans*内部临时桶授权策略

步骤1 登录统一身份认证服务IAM控制台。

步骤2 在左侧导航窗格中，选择“权限管理>权限”页签，单击右上方的“创建自定义策略”。

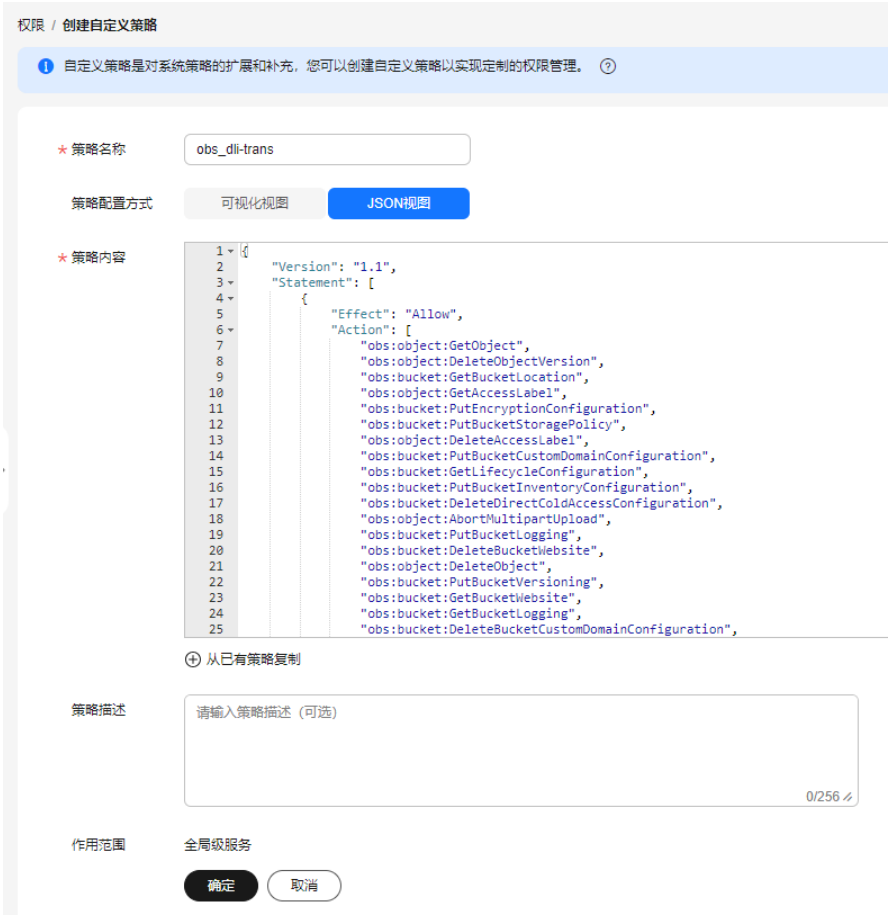
图 5-43 创建自定义策略

步骤3 在自定义策略配置页面，策略配置方式切换至JSON视图，然后按照如下策略内容，创建**obs_dli-trans**自定义策略。

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "obs:object:GetObject",
                "obs:object:DeleteObjectVersion",
                "obs:bucket:GetBucketLocation",
```

```
            "obs:object:GetAccessLabel",
            "obs:bucket:PutEncryptionConfiguration",
            "obs:bucket:PutBucketStoragePolicy",
            "obs:object:DeleteAccessLabel",
            "obs:bucket:PutBucketCustomDomainConfiguration",
            "obs:bucket:GetLifecycleConfiguration",
            "obs:bucket:PutBucketInventoryConfiguration",
            "obs:bucket:DeleteDirectColdAccessConfiguration",
            "obs:object:AbortMultipartUpload",
            "obs:bucket:PutBucketLogging",
            "obs:bucket:DeleteBucketWebsite",
            "obs:object:DeleteObject",
            "obs:bucket:PutBucketVersioning",
            "obs:bucket:GetBucketWebsite",
            "obs:bucket:GetBucketLogging",
            "obs:bucket:DeleteBucketCustomDomainConfiguration",
            "obs:object:PutObject",
            "obs:object:RestoreObject",
            "obs:bucket:PutReplicationConfiguration",
            "obs:bucket:GetBucketQuota",
            "obs:object:GetObjectVersionAcl",
            "obs:bucket:DeleteBucket",
            "obs:bucket:CreateBucket",
            "obs:bucket:GetDirectColdAccessConfiguration",
            "obs:bucket:PutDirectColdAccessConfiguration",
            "obs:bucket:GetBucketAcl",
            "obs:bucket:GetBucketVersioning",
            "obs:bucket:GetBucketInventoryConfiguration",
            "obs:bucket:GetBucketStoragePolicy",
            "obs:bucket:GetEncryptionConfiguration",
            "obs:bucket:PutBucketCORS",
            "obs:bucket:PutBucketTagging",
            "obs:bucket:GetBucketTagging",
            "obs:bucket:PutLifecycleConfiguration",
            "obs:bucket:GetBucketCustomDomainConfiguration",
            "obs:object:ListMultipartUploadParts",
            "obs:object:ModifyObjectMetaData",
            "obs:bucket:ListBucketVersions",
            "obs:bucket:PutBucketQuota",
            "obs:object:PutAccessLabel",
            "obs:bucket:ListBucket",
            "obs:bucket:GetBucketCORS",
            "obs:bucket:DeleteBucketInventoryConfiguration",
            "obs:object:GetObjectVersion",
            "obs:bucket:PutBucketWebsite",
            "obs:bucket:DeleteReplicationConfiguration",
            "obs:object:GetObjectAcl",
            "obs:bucket:GetBucketNotification",
            "obs:bucket:PutBucketNotification",
            "obs:bucket:GetReplicationConfiguration",
            "obs:bucket:GetBucketPolicy",
            "obs:bucket:DeleteBucketTagging",
            "obs:bucket:GetBucketStorage"
        ],
        "Resource": [
            "OBS:*:*:object:*",
            "OBS:*:*:bucket:dli-trans*"
        ]
    }
  ]
}
```


图 5-44 配置 **obs_dli-trans** 自定义策略



步骤4 单击“确定”，完成**obs_dli-trans**自定义策略创建。

步骤5 在IAM左侧导航窗格中，选择“用户组”，找到DLI连接中使用AK/SK所在用户的归属用户组，单击授权，将**obs_dli-trans**自定义策略授权给该用户。

图 5-45 为用户组授权 **obs_dli-trans** 自定义策略

----结束

5.6.4.12 配置 OpenTSDB 目的端参数

作业中目的连接为**CloudTable OpenTSDB连接**时，目的端作业参数如**表5-100**所示。

表 5-100 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

5.6.4.13 配置 MRS Hudi 目的端参数

作业中目的连接为**MRS Hudi连接**时，目的端作业参数如**表5-101**所示。

表 5-101 MRS Hudi 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	目的连接名称	选择已配置的MRS Hudi连接。	hudi_to_cdm
	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。	dbadmin
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	cdm
	自动创表	是否自动创建Hudi表。 • 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。	不自动创表

参数类型	参数名	说明	取值样例
	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。	否
	全量模式写Hoodie	选择写Hoodie模式，默认选“是”表示全量模式，“否”表示微批模式。 ● 全量模式为异步分片写入Hoodie，适用于一次全量写入场景。 ● 微批模式为异步分批写入Hoodie，适用于对入库时间SLA要求较为严格的场景，以及对资源消耗较小，对MOR表存储类型在线进行压缩的场景。 **说明** 运行-失败重试期间不允许修改此模式。	是
	批次数据大小	“全量模式写Hoodie”设置为“否”时，使用微批模式呈现此参数。 用于设置单个批次写Hoodie的数据行数，默认100000行。	100000
	使用入库时间字段	将一个字段标记为入库时间字段，自动建表时将此字段自动加到建表语句中，写入Hudi时将把此字段的值替换为当前时间，不自动建表时选择已经存在的入库时间字段。	是
	入库时间字段名称	“使用入库时间字段”设置为“是”时，呈现此参数。 用于记录写入Hudi的时间。 **说明** ● 对于已存在目的端表中带有入库时间字段的，可以直接使用已有的timestamp类型字段。 ● 对于自动建表的场景，该字段会被拼接到建表语句中，类型为timestamp，该字段名称不能与源端的字段有重复（包括自定义字段）。	cdc_last_update_date
Hudi建表参数	Location	存储在OBS或HDFS上数据库表的文件路径。	-

参数类型	参数名	说明	取值样例
	Hudi表类型	Hudi表存储类型。 • MOR表：数据先写入avro格式的日志文件，读取时合并到parquet文件。 • COW表：数据直接写入parquet文件。	MOR
	Hudi表主键	对Hudi建表设置主键，多个值以逗号隔开。	-
	Hudi表生成器类	主键生成类型，实现org.apache.hudi.keygen.KeyGenerator从传入记录中提取键值。	-
	Hudi表预聚合键	对Hudi建表设置预聚合键，当两个记录拥有相同的主键时，保留precombine字段值较大的记录。 **说明** 如果没有时间字段，可以设置和主键一样的字段，当遇到主键冲突时，保留最新的记录。	ts
	Hudi表分区字段	对Hudi建表设置分区字段，多个值以逗号隔开。	-
	Hudi表压缩策略（是否开启写入压缩）	在线进行压缩，仅对MOR表生效。	是
	Hudi表清除策略（保留提交数）	清除时保留的提交数。	1
	Hudi表归档策略（最小保留提交数）	归档时保留的最小提交数。	1
	Hudi表归档策略（最大保留提交数）	归档时保留的最大提交数。	100
	Hudi表配置	对Hudi建表设置自定义参数属性，此处填入的参数将会在options中生效。例如：主键、combineKey、索引。	-

5.6.4.14 配置 MRS ClickHouse 目的端参数

作业中目的连接为**MRS ClickHouse连接**时，目的端作业参数如**表5-102**所示。

说明

当作业源端为MRS ClickHouse、DWS及Hive时：

- 若int及float类型字段为null时，创建MRS ClickHouse表格时字段类型需设置为nullable()，否则写入到MRS ClickHouse的值会为0。
- 请确认目的端表引擎是否为ReplicatedMergeTree引擎，该引擎自带去重机制，且去重数据不能准确预测，选用该引擎应保证数据唯一性，否则会造成不唯一数据被忽略写入，或尝试替换其他表引擎，例如MergeTree。

表 5-102 MRS ClickHouse 作为目的端时的作业参数

参数名	说明	取值样例
模式或表空间	单击输入框后面的按钮可选择模式或表空间。	schema
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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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
是否在集群操作	**“导入开始前”**参数选择为“清除部分数据”或“清除全部数据”时，显示该参数。如果设置为是，将对集群中的所有节点进行全部/部分数据清除操作。	是
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

5.6.4.15 配置 MongoDB 目的端参数

作业中目的连接为**MongoDB连接**时，目的端作业参数如**表5-103**所示。

表 5-103 MongoDB 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。	mddb
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，或者对应连接里的账号是否有元数据查询的权限。	COLLECTION
迁移行为	将记录迁移到MongoDB目的端时，选择需要进行的插入行为操作。 ● 新增：将文件记录直接插入指定的集合。 ● 有则新增，无则替换：以指定的过滤键作为查询条件。如果在集合中找到匹配的记录，则替换该记录（找到多条匹配记录时，只会替换找到的第一条记录）。如果不存在，则添加新记录。 ● 替换：使用指定的过滤键作为查询条件。如果在集合中找到匹配的记录，则替换该记录（找到多条匹配记录时，只会替换找到的第一条记录）。如果没有，则不会添加新记录。	新增
导入前准备语句	执行任务前需要先执行的MongoDB查询语句。 **说明** ● “导入前准备语句”格式是json，只有两个键值对，第一个键值对是配置操作类别，key是"type"，value只支持"remove"和"drop"。第二个键值对是针对不同操作类别，需要配置的数据条件或者集合名称。 ● 导入前准备语句的执行不会影响即将写入的数据内容。	{"type":"remove","json":"{$or:[{Pid:{$gt:'0',$lt:'2'}},{X:{$gt:'50',$lt:'80'}}]}"}

5.6.5 配置 CDM 作业字段映射

操作场景

- 作业参数配置完成后，将进行字段映射的配置，您可以通过字段映射界面的可自定义新增字段，也可单击操作列下 创建字段转换器。
- 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。
- 自动创表场景下，需在目的端表中提前手动新增字段，再在字段映射里新增字段。

约束限制

- 作业源端开启“使用SQL语句”参数时不支持配置转换器。
- 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。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datetime）。
- 当作业源端为OBS、迁移CSV文件时，并且配置“解析首行为列名”参数的场景下显示列名。
- 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。
- 自动创表场景下，需在目的端表中提前手动新增字段，再在字段映射里新增字段。
- 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。
- 如果字段映射关系不正确，您可以通过拖拽字段、单击对字段批量映射两种方式来调整字段映射关系。
- 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：
  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CDM不支持源端迁移字段类型，请参见**不支持数据类型转换规避指导**将字段类型转换为CDM支持的类型。

新增字段

您可以单击字段映射界面的选择“添加新字段”自定义新增字段，通常用于标记数据库来源，以确保导入到目的端数据的完整性。

图 5-46 字段映射

目前支持以下类型自定义字段：

- **常量**

  常量参数即参数值是固定的参数，不需要重新配置值。例如“lable”=“friends”用来标识常量值。

- **变量**

您可以使用时间宏、表名宏、版本宏等变量来标记数据库来源信息。变量的语法：${variable}，其中“variable”指的是变量。例如“input_time”=“${timestamp()}”用来标识当前时间的时间戳。

- **表达式**

  您可以使用表达式语言根据运行环境动态生成参数值。表达式的语法：#{expr}，其中“expr”指的是表达式。例如“time”=“#{DateUtil.now()}”用来标识当前日期字符串。

新建转换器

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 5-47 新建转换器

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**

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

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

- **去前后空格**

  自动去字符串前后的空值，不需要配置参数。

- **字符串反转**

  自动反转字符串，例如将“ABC”转换为“CBA”，不需要配置参数。

- **字符串替换**

  替换字符串，需要用户配置被替换的对象，以及替换后的值。

- **去换行**

  将字段中的换行符（\n、\r、\r\n）删除。

- **表达式转换**

  数据进行转换过程中，替换内容包含特殊字符时，需要先使用\将该字符转义成普通字符。

- 表达式支持以下两个环境变量：
  - value：当前字段值。
  - row：当前行，数组类型。
- 表达式支持的工具类用法罗列如下，未列出即表示不支持：
  i. 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)
  ii. 将当前字段的字符串全部转为大写。

     表达式：StringUtils.upperCase(value)
  iii. 如果想将第1个日期字段格式从“2018-01-05 15:15:05”转换为“20180105”。

     表达式：DateUtils.format(DateUtils.parseDate(row[0],"yyyy-MM-dd HH:mm:ss"),"yyyyMMdd")
  iv. 如果想将时间戳转换成“yyyy-MM-dd hh:mm:ss”格式的日期字符串的类型，例如字段值为“1701312046588”，转换后为“2023-11-30 10:40:46”。

     表达式：DateUtils.format(NumberUtils.toLong(value),"yyyy-MM-dd HH:mm:ss")
  v. 如果想将“yyyy-MM-dd hh:mm:ss”格式的日期字符串转换成时间戳的类型。

     表达式：DateUtils.getTime(DateUtils.parseDate(value,"yyyy-MM-dd hh:mm:ss"))
  vi. 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,"-")
  vii. 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
  viii. 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"
  ix. 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
  x. 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")
  xi. 获取一个36位的UUID（Universally Unique Identifier，通用唯一识别码）。

     表达式：CommonUtils.randomUUID()
  xii. 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)

xiii. 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)

xiv. 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*)

xv. 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)

xvi. 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*")

xvii. 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*")

xviii. 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*")

xix. 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*")

xx. 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*)

xxi. 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*)

xxii. 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*")

xxiii. 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*")

xxiv. 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*")

xxv. 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*)

xxvi. 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*")

xxvii. 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)

xxviii. 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)

xxix. 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)

xxx. 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)

xxxi. 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)

xxxii. 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)

xxxiii. 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)

xxxiv. 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*)

xxxv. 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*)

xxxvi. 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*")

xxxvii. 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*")

xxxviii. 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)

xxxix. 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*")

xl. 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*")

xli. 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.*")

xlii. 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*")

替换内容包含特殊字符时，需要先把该字符转义成普通字符，例如，客户想通过该表达式把字符串中 \t 去掉时，需要配置为：StringUtils.replace(value,"\\t","")（即把 \ 再次转义）。

xliii. 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*")

xliv. 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*")

xlv. 如果当前字段为字符串类型，去除字段中首、尾处所有指定的字符，例如去除“abcyx”中首尾所有的“x”、“y”、“z”和“b”，转换后为“abc”。

表达式：StringUtils.strip(value,"*xyz*b")

xlvi. 如果当前字段为字符串类型，去除字段末尾所有指定的字符，例如去除当前字段末尾的“abc”字符串。

表达式：StringUtils.stripEnd(value,*"abc"*)

xlvii. 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*)

xlviii 如果当前字段为字符串类型，获取字符串指定位置后（索引从0开始，包括指定位置的字符）的子字符串，指定位置如果为负数，则从末尾往前计算位置，末尾第一位为-1。例如获取“abcde”索引为2的字符（即c）及之后的字符串，则转换后为“cde”。

表达式：StringUtils.substring(value,*2*)

xlix. 如果当前字段为字符串类型，获取字符串指定区间（索引从0开始，区间起点包括指定位置的字符，区间终点不包含指定位置的字符）的子字符串，区间位置如果为负数，则从末尾往前计算位置，末尾第一位为-1。例如获取“abcde”第2个字符（即c）及之后、第4个字符（即e）之前的字符串，则转换后为“cd”。

表达式：StringUtils.substring(value,*2*,4)

l. 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*")

li. 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*")

lii. 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*")

liii. 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*")

liv. 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*")

lv. 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)

lvi. 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)

lvii. 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*)

lviii. 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)

lix. 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*)

lx. 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)

lxi. 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*)

lxii. 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)

lxiii. 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*)

lxiv. 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)

lxv. 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*)

lxvi. 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)

lxvii.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*)

lxviii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)

lxix. 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*")

lxx. 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*"))

lxxi. 取出缓存的IP与地址映射对象。

表达式：CommonUtils.getCache("*ipList*")

lxxii.判断是否有IP与地址映射缓存。

表达式：CommonUtils.cacheExists("*ipList*")

lxxiii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*)

lxxiv如果value值为空或者null时，则返回字符串“aaa”，否则返回value。

表达式：StringUtils.defaultIfEmpty(value,"aaa")

特殊链路说明

- 当源端为DLI，目的端为DWS时，DLI的tinyint类型字段映射为DWS的smallint类型字段。
- 当源端为Hudi，目的端为DWS时，Hudi的Double类型字段映射为DWS的Float类型字段。

5.6.6 配置 CDM 作业定时任务

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。
- 定时任务功能原理：采用Java Quartz定时器，类似Cron表达式配置。对起始时间解析出分，小时，天，月。构造出cronb表达式。

  以配置天调度为例：重复周期选择1天：若当前时间2022/10/14 12:00，配置起始时间为2022/10/14 00:00。任务在2022/10/15 00:00执行；若当前时间2022/10/14 12:00，配置起始时间为2022/10/15 00:00。任务在2022/10/15 00:00执行。

  重复周期选择2天：若当前时间2022/10/14 12:00，配置起始时间为2022/10/14 00:00。任务在2022/10/16 00:00执行；若当前时间2022/10/14 12:00，配置起始时间为2022/10/15 00:00。任务在2022/10/16 00:00执行。

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。

图 5-48 重复周期为分钟

例如上图表示：从2023年1月1日0时0分开始第一次自动执行作业，每30分钟自动执行一次，2023年12月31日23时59分之后不再自动执行。

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。

图 5-49 重复周期为小时

例如上图表示：定时配置从2023年1月1日0时0分生效，0:10时开始第一次自动执行作业，0:30第二次，0:50第三次，以后每2小时重复三次，2023年12月31日23时59分之后不再自动执行。

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。

图 5-50 重复周期为天

例如上图表示：从2023年1月1日0时0分开始第一次自动执行，每3天自动执行一次，配置一直有效。

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。

图 5-51 重复周期为周

例如上图表示：在2023年1月1日0时0分以后，每2周的周二、周六、周日的0时0分，便自动执行作业，直到2023年12月31日23时59分不再自动执行。

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。

图 5-52 重复周期为月

例如上图表示：从2023年1月1日0点开始，每月5日、25日的0点自动执行作业，直到2023年12月31日23时59分不再自动执行。

5.6.7 CDM 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：

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

最大抽取并发数

最大抽取并发数即集群最大抽取并发数。

说明

此处的“最大抽取并发数”参数与集群配置处的“最大抽取并发数”参数同步，在任意一处修改即可生效。

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

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

   说明

   不同源端数据源的作业分片维度有所不同，因此某些作业可能出现未严格按作业“抽取并发数”参数分片的情况。

2. CDM依次将Task提交给运行池运行。根据集群配置管理中的“最大抽取并发数”参数，超出规格的Task排队等待运行。

因此作业抽取并发数和集群最大抽取并发数参数设置为适当的值可以有效提升迁移速度，您可参考下文有效配置抽取并发数。

1. 集群最大抽取并发数的上限建议为vCPU核数*2，如**表5-104**所示。

   表 5-104 集群最大抽取并发数配置建议

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

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

定时备份/恢复

该功能依赖于OBS服务。当前定时备份内容不会自动老化删除，您需要定期手动清理备份文件。

- 前提条件

  已创建OBS连接，详情请参见**OBS连接参数说明**。

- 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 5-105 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。

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。

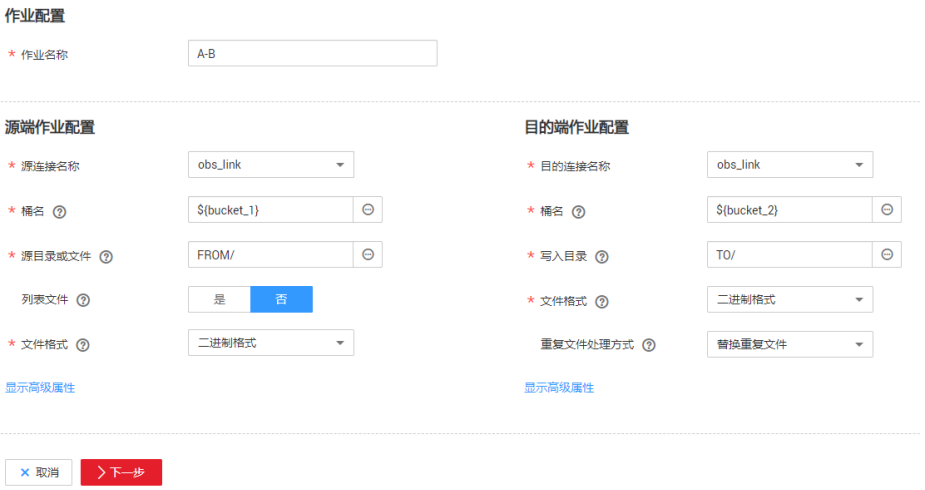
   ```
   bucket_1=A
   bucket_2=B
   ```

   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 5-53 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
```

5.6.8 管理单个 CDM 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，stopped。

  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**

  查看作业执行结果及最近30天内的历史信息，包括历史执行记录、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。可对比源端和目的端的数据，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。

修改

- **修改作业参数**

可重新配置作业参数，支持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“表/文件迁移”显示作业列表，可对单个作业执行如下操作：

- 修改作业参数：单击作业操作列的“编辑”可修改作业参数。
- 运行作业：单击作业操作列的“运行”可手动启动作业。
- 查看历史记录：单击作业操作列的“历史记录”进入历史记录界面，可查看该作业的历史执行记录、读取和写入的统计数据。在历史记录界面单击“日志”，可查看作业执行的日志信息。
- 删除作业：选择作业操作列的“更多 > 删除”可删除作业。
- 停止作业：选择作业操作列的“更多 > 停止”可停止作业。
- 查看作业JSON：选择作业操作列的“更多 > 查看作业JSON”，可查看该作业的JSON定义。
- 编辑作业JSON：选择作业操作列的“更多 > 编辑作业JSON”，可直接编辑该作业的JSON文件，作用等同于修改作业的参数配置。
- 配置定时任务：选择作业操作列的“更多 > 配置定时任务”，可选择在有效期内周期性启动作业，具体请参考**配置CDM作业定时任务**。
- 日志：选择作业操作列的“更多 > 日志”，可查看该作业最近的一次日志。

  也可以在历史记录界面可查看作业所有的日志。
- 失败重试：选择作业操作列的“更多 > 失败重试”，可以对执行失败的作业，选择自动重试三次或者不重试。

步骤3 修改完成后单击“保存”或“保存并运行”。

----结束

5.6.9 批量管理 CDM 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、按分组批量启动作业、按分组导出作业等操作。

  说明

  按组批量启动作业会运行组内所有作业。如果开启了用户隔离功能，即使华为账号下的其他IAM用户无法查看到组内作业，按组批量启动作业依然会将组内作业运行，因此在用户隔离场景不建议使用按组批量启动作业功能。

- **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5-54。

  图 5-54 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。

  说明

  当前导入时不支持覆盖已有作业。

----结束

5.7 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：

- 源端的源目录或文件
- 源端的表名
- “通配符”过滤类型中的目录过滤器和文件过滤器
- “时间过滤”中的起始时间和终止时间
- 分区过滤条件和Where子句
- 目的端的写入目录
- 目的端的表名

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。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。

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），MONTH（月），YEAR（年）。

    说明

    其中MONTH（月），YEAR（年）的偏移量类型存在特殊场景：

    - 对于年、月来说，若进行偏移后实际没有该日期，则按照日历取该月最大的日期。
    - 不支持在源端和目的端的“时间过滤”参数中的起始时间、终止时间使用年、月的偏移。

  例如当前日期为"2023-03-01 09:00:00"，则：

  - **"dateformat(yyyy-MM-dd HH:mm:ss, -1, YEAR)"**表示当前时间的前一年，也就是"2022-03-01 09:00:00"。
  - **"dateformat(yyyy-MM-dd HH:mm:ss, -3, MONTH)"**表示当前时间的前三月，也就是"2022-12-01 09:00:00"。
  - **"dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)"**表示当前时间的前一天，也就是"2023-02-28 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"2023-03-01 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"2023-03-01 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"2023-03-01 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5-106**所示。

说明

表中示例实际使用时必须嵌在''中使用，比如需要以yyyy-MM-dd格式返回当前时间时，参数为'${dateformat(yyyy-MM-dd)}'。

表 5-106 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
${dateformat(yyyy-MM-dd, -1, DAY)} 00:00:00	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前一天0点。	2017-10-15 00:00:00
${dateformat(yyyy-MM-dd, -1, DAY)} 12:00:00	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前一天12点。	2017-10-15 12:00:00
${dateformat(yyyy-MM-dd, -N, DAY)} 00:00:00	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前N天的0点。	N为3时： 2017-10-13 00:00:00
${dateformat(yyyy-MM-dd, -N, DAY)} 12:00:00	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前N天的12点。	N为3时： 2017-10-13 12:00:00
${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5-55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 5-55 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5-56所示。

图 5-56 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。

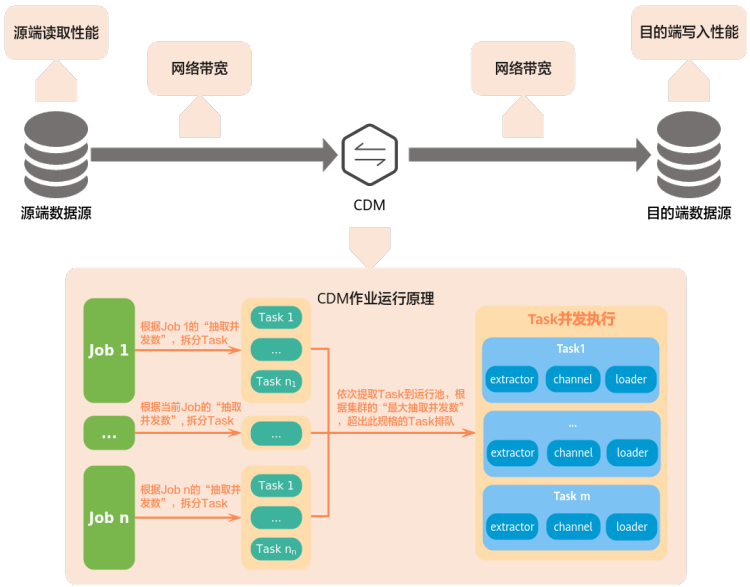
5.8 优化迁移性能

5.8.1 迁移作业原理

数据迁移模型

CDM数据迁移时，简化的迁移模型如图5-57所示。

图 5-57 CDM 数据迁移模型



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

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

   说明

   不同源端数据源的作业分片维度有所不同，因此某些作业可能出现未严格按作业“抽取并发数”参数分片的情况。

2. CDM依次将Task提交给运行池运行。根据集群配置管理中的“最大抽取并发数”参数，超出规格的Task排队等待运行。

性能影响因素

根据迁移模型，可以看出CDM数据迁移的速率受源端读取速度、网络带宽、目的端写入性能、CDM集群和作业配置等因素影响。

表 5-107 性能影响因素

影响因素		说明
业务相关因素	作业抽取并发数配置	创建CDM迁移作业时，支持设置该作业的抽取并发数。 该参数设置为适当的值可以有效提升迁移速度，过小则会限制迁移速度，过大则会导致任务过载、迁移失败。 ● 迁移的目的端为文件时，CDM不支持多并发，此时应配置为单进程抽取数据。 ● 表中每行数据大小为1MB以下的可以设置多并发抽取，超过1MB的建议单线程抽取数据。
	集群最大抽取并发数规格	该参数设置为适当的值可以有效提升迁移速度，过小则会限制迁移速度，过大则会导致源端负载过高、影响系统稳定性。 不同规格的CDM集群支持的最大抽取并发数规格不同，并发数上限建议设置为vCPU核数*2。 ● cdm.large：16 ● cdm.xlarge：32 ● cdm.4xlarge：128
	业务模型	如果大量CDM作业同时执行，当超过当前CDM集群的并发执行作业数时，会导致作业排队，耗时提升。 建议您将迁移作业的运行时间错开，平摊在业务周期内，避免资源紧张导致迁移时间过长。
	数据模型	数据迁移时，对于不同的数据结构，迁移速度也会受到一定影响。例如： ● 对于表迁移，宽表的迁移速度较慢，字符串类型越多（字段大小）迁移速度越慢。 ● 对于文件而言，总大小相同时，大文件迁移较快，多个小文件迁移较慢。 ● 对于消息而言，消息内容越多，所占带宽越高，每秒事务（TPS）越低。

影响因素	说明
源端读取速度	取决于源端数据源的性能。 如需优化，请参见源端数据源的相关说明文档。
网络带宽	CDM集群与数据源之间可以通过内网、公网VPN、NAT或专线等方式互通。 ● 通过内网互通时，网络带宽是根据不同的CDM实例规格的带宽限制的。 – cdm.large实例规格CDM集群网卡的基准/最大带宽为0.8/3 Gbps。 – cdm.xlarge实例规格CDM集群网卡的基准/最大带宽为4/10 Gbps。 – cdm.4xlarge实例规格CDM集群网卡的基准/最大带宽为36/40 Gbps。 ● 通过公网互通时，网络带宽受到公网带宽的限制。CDM侧公网带宽规格受限于CDM集群所绑定的弹性公网IP，数据源侧受限于其所访问的公网带宽规格。 ● 通过VPN、NAT或专线互通时，网络带宽受到VPN、NAT或专线带宽的限制。
目的端写入性能	取决于目的端数据源的性能。 如需优化，请参见目的端数据源的相关说明文档。

5.8.2 性能调优

概述

根据数据迁移模型分析，除了源端读取速度、目的端写入性能、带宽优化外，您也可以通过如下方式优化作业迁移速度：

- **使用大规格CDM集群**

  不同规格的CDM集群网卡带宽、集群最大抽取并发数等有所差异。如果您有较高的迁移速度需求，或当前CDM集群的CPU使用率、磁盘使用率、内存使用率等指标经常在较高区间运行，建议您选用大规格的CDM集群规格进行数据迁移。

- **使用多个CDM集群**

  包含但不限于以下情况时，建议您使用多个CDM集群进行业务分流，提升迁移效率与业务稳定性。

  – 需要作为不同的用途或给多个业务部门使用。例如既需要用于数据迁移作业，又需要作为DataArts Studio管理中心连接代理时，建议各配置至少一个CDM集群。

  – 待迁移任务库表较多，迁移量较大。此时可以使用多个CDM集群同时作业，提升迁移效率。

  – 当前CDM集群的CPU使用率、磁盘使用率、内存使用率等指标经常在较高区间运行。此时建议使用多个CDM集群进行业务分流。

- **错峰执行CDM作业**

如果大量CDM作业同时执行，当超过当前CDM集群的并发执行作业数时，会导致作业排队，耗时提升。

建议您将迁移作业的运行时间错开，平摊在业务周期内，避免资源紧张导致迁移时间过长。

- **调整抽取并发数**

  对于低任务量场景，调整抽取并发数是性能调优的最佳方式。CDM迁移作业支持设置作业抽取并发数，同时也可以设置集群最大抽取并发数。

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

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

     说明

     不同源端数据源的作业分片维度有所不同，因此某些作业可能出现未严格按作业“抽取并发数”参数分片的情况。

  b. CDM依次将Task提交给运行池运行。根据集群配置管理中的“最大抽取并发数”参数，超出规格的Task排队等待运行。

  因此作业抽取并发数和集群最大抽取并发数参数设置为适当的值可以有效提升迁移速度。关于如何调整抽取并发数，详情请参考**如何调整抽取并发数**。

如何调整抽取并发数

1. 集群最大抽取并发数的设置与CDM集群规格有关，并发数上限建议配置为vCPU核数*2，如**表5-108**所示。

   表 5-108 集群最大抽取并发数配置建议

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

图 5-58 集群最大抽取并发数配置

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

图 5-59 作业抽取并发数配置

5.8.3 参考：作业分片维度

CDM在进行作业分片时，根据源端数据源的差异，分片维度有所不同。详情如**表 5-109**所示。

表 5-109 不同源端数据源的作业分片维度

数据源分类	源端数据源	作业分片原理
数据仓库	数据仓库服务（DWS）	• 支持按表字段分片。 • 不支持按表分区分片。
	数据湖探索（DLI）	• 支持分区表的分区信息分片。 • 不支持非分区表分片。
Hadoop	MRS HDFS	支持按文件分片。
	MRS HBase	支持按HBase的Region分片。
	MRS Hive	• HDFS读取方式时，支持按Hive文件分片。 • JDBC读取方式时，不支持分片。
	FusionInsight HDFS	支持按文件分片。
	FusionInsight HBase	支持按HBase的Region分片。
	FusionInsight Hive	• HDFS读取方式时，支持按Hive文件分片。 • JDBC读取方式时，不支持分片。

数据源分类	源端数据源	作业分片原理
	Apache HDFS	支持按文件分片。
	Apache HBase	支持按HBase的Region分片。
	Apache Hive	• HDFS读取方式时，支持按Hive文件分片。 • JDBC读取方式时，不支持分片。
对象存储	对象存储服务（OBS）	支持按文件分片。
文件系统	FTP	支持按文件分片。
	SFTP	支持按文件分片。
	HTTP	支持按文件分片。
关系型数据库	云数据库 MySQL	• 支持按表字段分片。 • 仅当配置“按表分区抽取”时，按表分区分片。
	云数据库 PostgreSQL	• 支持按表字段分片。 • 仅当配置“按表分区抽取”时，按表分区分片。
	云数据库 SQL Server	• 支持按表字段分片。 • 仅当配置“按表分区抽取”时，按表分区分片。
	MySQL	• 支持按表字段分片。 • 仅当配置“按表分区抽取”时，按表分区分片。
	PostgreSQL	• 支持按表字段分片。 • 仅当配置“按表分区抽取”时，按表分区分片。
	Microsoft SQL Server	• 支持按表字段分片。 • 不支持按表分区分片。
	Oracle	• 支持按表字段分片。 • 仅当配置“按表分区抽取”时，按表分区分片。
	SAP HANA	• 支持按表字段分片。 • 不支持按表分区分片。
	分库	每个后端连接一个子作业，子作业支持按主键分片。

数据源分类	源端数据源	作业分片原理
NoSQL	分布式缓存服务（DCS）	不支持分片。
	Redis	不支持分片。
	文档数据库服务（DDS）	不支持分片。
	MongoDB	不支持分片。
	Cassandra	支持按Cassandra的token range分片。
消息系统	数据接入服务（DIS）	支持按topic分片。
	Apache Kafka	支持按topic分片。
	DMS Kafka	支持按topic分片。
	MRS Kafka	支持按topic分片。
搜索	Elasticsearch	不支持分片。
	云搜索服务（CSS）	不支持分片。

5.8.4 参考：CDM 性能实测数据

背景说明

文中提供的性能指标仅用于参考，实际环境会受源或目标数据源性能、网络带宽及时延、数据及业务模型等因素影响。推荐您在正式迁移前，可先用小数据量实测进行速度摸底。

环境信息

- CDM集群为xlarge规格，2.9.1 200版本。
- 性能测试中，表数据规格为5000W行100列，HDFS二进制文件数据规格分别为3597W行100列、6667W行100列和10000W行100列。
- 多并发抽取/写入速率，定义为分别取作业抽取并发数为1、10、20、30、50时，最大的抽取/写入速率。

数据源抽取写入性能实测数据

常见数据源的性能实测结果分别如表5-110和表5-111所示。

表 5-110 读取性能实测数据

数据源	数据源规格	版本	单并发抽取速率（行/s）	多并发抽取速率（行/s）
云数据库 MySQL	8U 32G	MySQL 5.7	42052	195313（并发度：40）
Oracle	8U 16G	19C	18539	18706（并发度：10）
MRS Hbase	master 16U64G *3 node 8U32G *3	MRS 3.1.0	6296	69156（并发度：30）
MRS Hive	master 16U64G *3 node 8U32G *3	MRS 3.1.0	22321	170068（并发度：30）
MRS HDFS（二进制文件）	master 16U64G *3 node 8U32G *3	MRS 3.1.0	138727	141468（并发度：20）
			125556	126990（并发度：10）
			120919	120919（并发度：10）
DWS	8U 16G	8.1.1.300	13434	/
DLI	16U	SQL队列	71023	19290（并发度：20）
MRS Hudi（MOR）	master 16U64G *3 node 8U64G *3	MRS 3.2.0	75187	467289（并发度：30）
MRS Hudi（COW）	master 16U64G *3 node 8U64G *3	MRS 3.2.0	84033	485436（并发度：30）
Clickhouse	node 8U32G *2	clickhouse 22.3.2.2	187265	/
Elasticsearch	4U8G *6	elasticsearch7.10.2	28752	/
RDS（Postgresql）	4U32G（主备模式）	Postgresql 13.12	128865	1351351（并发度：30）

表 5-111 写入性能实测数据

数据源	数据源规格	版本	单并发写入速率（行/s）	多并发写入速率（行/s）
云数据库 MySQL	8U 32G	MySQL 5.7	2658	/
Oracle	8U 16G	19C	/	/
MRS Hbase	master 16U64G *3 node 8U32G *3	MRS 3.1.0	3959	4120（并发度：10）
MRS Hive	master 16U64G *3 node 8U32G *3	MRS 3.1.0	25813	26882（并发度：10）
MRS HDFS（二进制文件）	master 16U64G *3 node 8U32G *3	MRS 3.1.0	65075	90155（并发度：10）
			86248	86248（并发度：1）
			76687	76687（并发度：1）
DWS	8U 16G	8.1.1.300	26624	27902（并发度：10）
DLI	16U	SQL队列	15211	18430（并发度：10）
MRS Hudi（MOR）	master 16U64G *3 node 8U64G *3	MRS 3.2.0	16345	183150（并发度：10）
MRS Hudi（COW）	master 16U64G *3 node 8U64G *3	MRS 3.2.0	21088	88183（并发度：20）
Clickhouse	node 8U32G * 2	clickhouse 22.3.2.2	93984	/
Elasticsearch	4U8G *6	elasticsearch 7.10.2	22271	/
RDS（Postgresql）	4U32G（主备模式）	Postgresql 13.12	34746	153374（并发度：10）

5.9 关键操作指导

5.9.1 增量迁移原理介绍

5.9.1.1 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，当文件的修改时间大于等于指定的起始时间，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。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。

文件/路径过滤器

- 参数位置：在创建表/文件迁移作业时，如果源端数据源为文件类型，那么源端作业参数的高级属性中可以看到“过滤类型”参数，该参数可选择：通配符或正则表达式。
- 参数原理：“过滤类型”选择“通配符”时，CDM就可以通过用户配置的通配符过滤文件或路径，CDM只迁移满足指定条件的文件或路径。
- 配置样例：

  例如源端文件名带有时间字段“2017-10-15 20:25:26”，这个时刻生成的文件为“/opt/data/file_20171015202526.data”，则在创建作业时，参数配置如下：

  a. 过滤类型：选择“通配符”。

  b. 文件过滤器：配置为**“*${dateformat(yyyyMMdd,-1,DAY)}*”**（这是CDM支持的日期宏变量格式，详见**时间宏变量使用解析**）。

图 5-60 文件过滤

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，只有修改时间介于起始时间和终止时间之间（时间区间为左闭右开，即等于起始时间也在区间之内）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 5-61 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。

5.9.1.2 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。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。

Where 子句

- 参数位置：在创建表/文件迁移作业时，如果源端为关系型数据库，那么在源端作业参数的高级属性下面可以看到“Where子句”参数。
- 参数原理：通过“Where子句”参数可以配置一个SQL语句（例如：age > 18 and age <= 60），CDM只导出该SQL语句指定的数据；不配置时导出整表。

  Where子句支持配置为**时间宏变量**，当数据表中有时间日期字段或时间戳字段时，配合定时执行作业，能够实现抽取指定日期的数据。
- 配置样例：

  假设数据库表中存在表示时间的列DS，类型为“varchar(30)”，插入的时间格式类似于“2017-xx-xx”，如**图5-62**所示，参数配置如下：

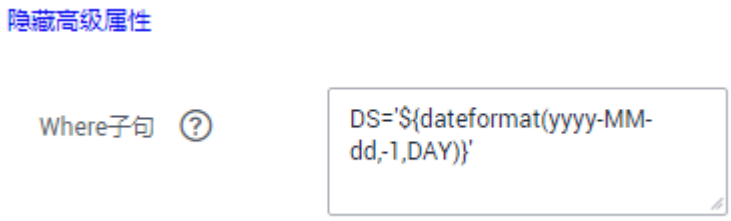
  图 5-62 表数据

  SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	s	2017-05-01
2	5	s	2017-05-01
3	1	g	2017-05-02
4	4	o	2017-05-02
5	6	a	2017-05-02
6	7	n	2017-05-02
7	1	g	2017-05-02
8	4	o	2017-05-02
9	6	a	2017-05-02
10	7	n	2017-05-02
11	2	fa	2017-10-15
12	3	te	2017-10-15
13	2	fa	2017-10-15
14	3	te	2017-10-15

  a. Where子句：配置为**DS='${dateformat(yyyy-MM-dd,-1,DAY)}'**。

图 5-63 Where 子句



b. 配置定时任务：重复周期为1天，每天的凌晨0点自动执行作业。

这样就可以每天0点导出前一天产生的所有数据。Where子句支持配置多种**时间宏变量**，结合CDM定时任务的重复周期：分钟、小时、天、周、月，可以实现自动导出任意指定日期内的数据。

5.9.1.3 HBase/CloudTable 增量迁移

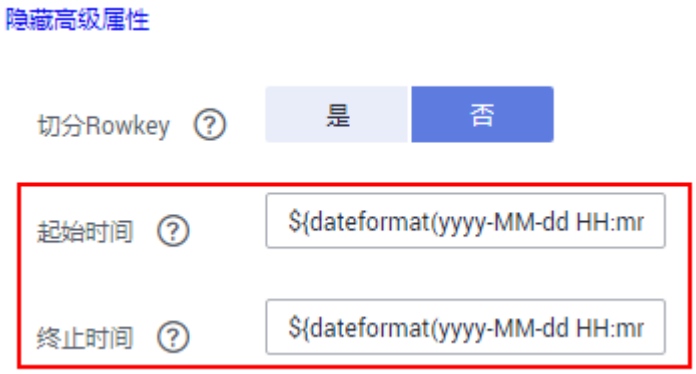
使用CDM导出HBase（包括MRS HBase、FusionInsight HBase、Apache HBase）或者表格存储服务（CloudTable）的数据时，支持导出指定时间段内的数据，配合CDM的定时任务，可以实现HBase/CloudTable的增量迁移。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。

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 5-64 HBase 时间区间



- 起始时间（包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间及以后的数据。
- 终止时间（不包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间以前的数据。

这2个参数支持配置为**时间宏变量**，例如：

- 起始时间配置为**${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}**时，表示只导出昨天以后的数据。

- 终止时间配置为**${dateformat(yyyy-MM-dd HH:mm:ss)}**时，表示只导出当前时间以前的数据。

这2个参数同时配置后，CDM就只导出前一天内的数据，再将该作业配置为每天0点执行一次，就可以增量同步每天新生成的数据。

5.9.1.4 MongoDB/DDS 增量迁移

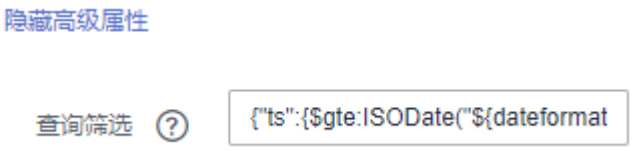
使用CDM导出MongoDB或者DDS的数据时，支持导出指定时间段内的数据，配合CDM的定时任务，可以实现MongoDB/DDS的增量迁移。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。

在创建CDM表/文件迁移的作业，源连接选择为MongoDB连接或者DDS连接时，高级属性的可选参数中可以配置查询筛选。

图 5-65 MongoDB 查询筛选



此参数支持配置为**时间宏变量**，例如起始时间配置为**{"ts":{$gte:ISODate("${dateformat(yyyy-MM-dd'T'HH:mm:ss.SSS'Z',-1,DAY)}")}}**，表示查找ts字段中大于时间宏转换后的值，即只导出昨天以后的数据。

参数配置后，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。

5.9.2 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 5-66 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。

5.9.3 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：HDFS（使用二进制格式传输时支持）。
- 目的端支持的数据源：HDFS（使用二进制格式传输时支持）。

下面分别以HDFS导出加密文件时解密、导入文件到HDFS时加密为例，介绍AES-256-GCM加解密的使用方法。

- **源端配置解密**

  创建从HDFS导出文件的CDM作业时，源端数据源选择HDFS、文件格式选择二进制格式后，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HDFS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HDFS的作业时，目的端数据源选择HDFS、文件格式选择二进制格式后，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HDFS时，目的端HDFS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。

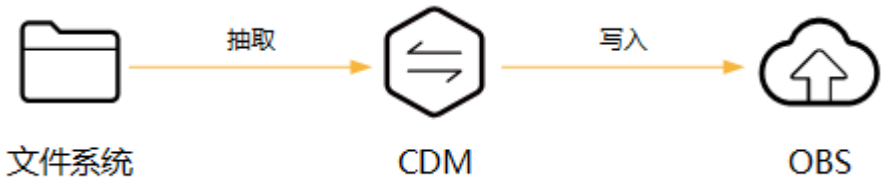
说明

- 如果选择使用KMS加密，则无法**使用MD5校验一致性**。
- 如果这里使用其它项目的KMS ID，则需要修改“项目ID”参数为KMS ID所属的项目ID；如果KMS ID与CDM在同一个项目下，“项目ID”参数保持默认即可。
- 使用KMS加密后，OBS上对象的加密状态不可以修改。
- 使用中的KMS密钥不可以删除，如果删除将导致加密对象不能下载。

5.9.4 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5-67**所示。

图 5-67 迁移文件到 OBS



在这个过程中，CDM支持使用MD5检验文件一致性。

- **抽取时**
  - 该功能支持源端为OBS、HDFS、FTP、SFTP、HTTP。可校验CDM抽取的文件，是否与源文件一致。
  - 该功能由源端作业参数“MD5文件名后缀”控制（“文件格式”为“二进制格式”时生效），配置为源端文件系统中的MD5文件名后缀。
  - 当源端数据文件同一目录下有对应后缀的保存md5值的文件，例如build.sh和build.sh.md5在同一目录下。若配置了“MD5文件名后缀”，则只迁移有MD5值的文件至目的端，没有MD5值或者MD5不匹配的数据文件将迁移失败，MD5文件自身不被迁移。
  - 若未配置“MD5文件名后缀”，则迁移所有文件。
- **写入时**
  - 该功能目前只支持目的端为OBS。可校验写入OBS的文件，是否与CDM抽取的文件一致。
  - 该功能由目的端作业参数“校验MD5值”控制，读取文件后写入OBS时，通过HTTP Header将MD5值提供给OBS做写入校验，并将校验结果写入OBS桶（该桶可以不是存储迁移文件的桶）。如果源端没有MD5文件则不校验。

说明

- 迁移文件到文件系统时，目前只支持校验CDM抽取的文件是否与源文件一致（即只校验抽取的数据）。
- 迁移文件到OBS时，支持抽取和写入文件时都校验。
- 如果选择使用MD5校验，则无法**使用KMS加密**。

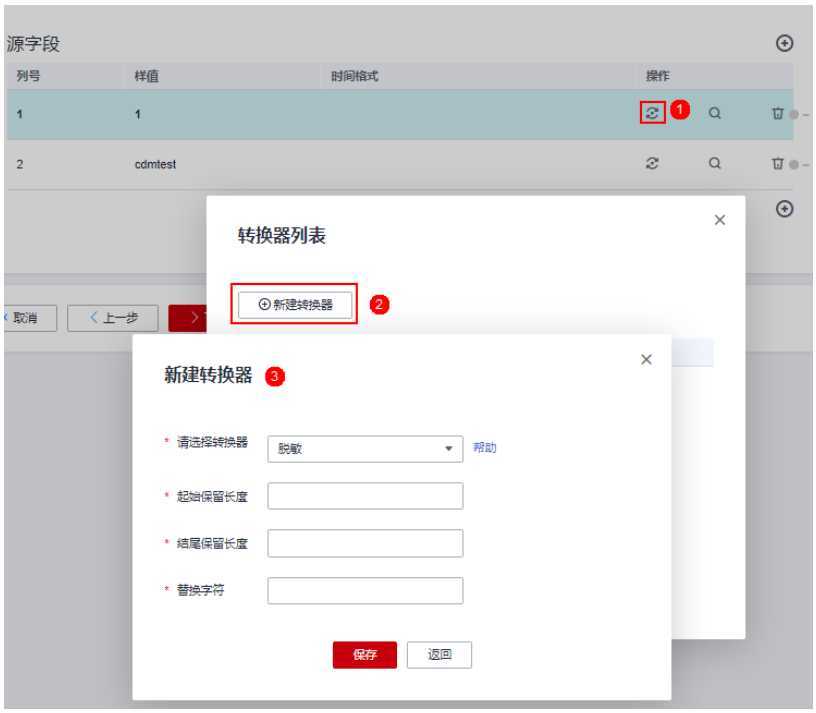
5.9.5 字段转换器配置指导

操作场景

- 作业参数配置完成后，将进行字段映射的配置，您可以单击操作列下 创建字段转换器。
- 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。

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，如下图所示。

图 5-68 新建字段转换器



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**

约束限制

- 作业源端开启“使用SQL语句”参数时不支持配置转换器。
- 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。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datetime）。
- 当作业源端为OBS、迁移CSV文件时，并且配置“解析首行为列名”参数的场景下显示列名。
- 当使用二进制格式进行文件到文件的迁移时，没有配置字段转换器这一步。
- 自动创表场景下，需在目的端表中提前手动新增字段，再在字段映射里新增字段。
- 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。

- 如果字段映射关系不正确，您可以通过拖拽字段、单击  对字段批量映射两种方式来调整字段映射关系。
- 创建表达式转换器时，表达式的功能是对该字段的数据进行处理，故不建议使用时间宏，如需使用，请根据以下场景处理（源端是文件类的配置时仅支持**方式一**）：
  - 方式一：新建表达式转换器时，表达式需要用''包围。

    ${dateformat(yyyy-MM-dd)}不加引号使用时，解析成2017-10-16之后还会进行运算，将'-'识别为减号，导致结果为1991，**须使用'${dateformat(yyyy-MM-dd)}'**，即'2017-10-16'。

    图 5-69 使用''包围表达式

  - 方式二：源字段中新增自定义字段，在样值中填写时间宏变量，重新进行字段映射处理。

    图 5-70 源字段新增自定义字段

    
- 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：
  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。

脱敏

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

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

去前后空格

自动去字符串前后的空值，不需要配置参数。

字符串反转

自动反转字符串，例如将“ABC”转换为“CBA”，不需要配置参数。

字符串替换

替换字符串，需要用户配置被替换的对象，以及替换后的值。

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。

数据进行转换过程中，替换内容包含特殊字符时，需要先使用\将该字符转义成普通字符。

- 表达式支持以下两个环境变量：
  - value：当前字段值。
  - row：当前行，数组类型。
- 表达式支持的工具类用法罗列如下，未列出即表示不支持：
  a. 如果当前字段为字符串类型，将字符串全部转换为小写，例如将“aBC”转换为“abc”。

     表达式：StringUtils.lowerCase(value)
  b. 将当前字段的字符串全部转为大写。

     表达式：StringUtils.upperCase(value)
  c. 如果想将第1个日期字段格式从“2018-01-05 15:15:05”转换为“20180105”。

     表达式：DateUtils.format(DateUtils.parseDate(row[0],"yyyy-MM-dd HH:mm:ss"),"yyyyMMdd")
  d. 如果想将时间戳转换成“yyyy-MM-dd hh:mm:ss”格式的日期字符串的类型，例如字段值为“1701312046588”，转换后为“2023-11-30 10:40:46”。

     表达式：DateUtils.format(NumberUtils.toLong(value),"yyyy-MM-dd HH:mm:ss")
  e. 如果想将“yyyy-MM-dd hh:mm:ss”格式的日期字符串转换成时间戳的类型。

     表达式：DateUtils.getTime(DateUtils.parseDate(value,"yyyy-MM-dd hh:mm:ss"))
  f. 如果当前字段值为“yyyy-MM-dd”格式的日期字符串，需要截取年，例如字段值为“2017-12-01”，转换后为“2017”。

表达式：StringUtils.substringBefore(value,"-")

g. 如果当前字段值为数值类型，转换后值为当前值的两倍。

表达式：value*2

h. 如果当前字段值为“true”，转换后为“Y”，其它值则转换后为“N”。

表达式：value=="true"?"Y":"N"

i. 如果当前字段值为字符串类型，当为空时，转换为“Default”，否则不转换。

表达式：empty value? "Default":value

j. 如果想将日期字段格式从“2018/01/05 15:15:05”转换为“2018-01-05 15:15:05”。

表达式：DateUtils.format(DateUtils.parseDate(value,"yyyy/MM/dd HH:mm:ss"),"yyyy-MM-dd HH:mm:ss")

k. 获取一个36位的UUID（Universally Unique Identifier，通用唯一识别码）。

表达式：CommonUtils.randomUUID()

l. 如果当前字段值为字符串类型，将首字母转换为大写，例如将“cat”转换为“Cat”。

表达式：StringUtils.capitalize(value)

m. 如果当前字段值为字符串类型，将首字母转换为小写，例如将“Cat”转换为“cat”。

表达式：StringUtils.uncapitalize(value)

n. 如果当前字段值为字符串类型，使用空格填充为指定长度，并且将字符串居中，当字符串长度不小于指定长度时不转换，例如将“ab”转换为长度为4的“ab”。

表达式：StringUtils.center(value,*4*)

o. 删除字符串末尾的一个换行符（包括“\n”、“\r”或者“\r\n”），例如将“abc\r\n\r\n”转换为“abc\r\n”。

表达式：StringUtils.chomp(value)

p. 如果字符串中包含指定的字符串，则返回布尔值true，否则返回false。例如“abc”中包含“a”，则返回true。

表达式：StringUtils.contains(value,"*a*")

q. 如果字符串中包含指定字符串的任一字符，则返回布尔值true，否则返回false。例如“zzabyycdxx”中包含“z”或“a”任意一个，则返回true。

表达式：StringUtils.containsAny(value,"*za*")

r. 如果字符串中不包含指定的所有字符，则返回布尔值true，包含任意一个字符则返回false。例如“abz”中包含“xyz”里的任意一个字符，则返回false。

表达式：StringUtils.containsNone(value,"*xyz*")

s. 如果当前字符串只包含指定字符串中的字符，则返回布尔值true，包含任意一个其它字符则返回false。例如“abab”只包含“abc”中的字符，则返回true。

表达式：StringUtils.containsOnly(value,"*abc*")

t. 如果字符串为空或null，则转换为指定的字符串，否则不转换。例如将空字符转换为null。

表达式：StringUtils.defaultIfEmpty(value,*null*)

u. 如果字符串以指定的后缀结尾（包括大小写），则返回布尔值true，否则返回false。例如“abcdef”后缀不为null，则返回false。

表达式：StringUtils.endsWith(value,*null*)

v. 如果字符串和指定的字符串完全一样（包括大小写），则返回布尔值true，否则返回false。例如比较字符串“abc”和“ABC”，则返回false。

表达式：StringUtils.equals(value,"*ABC*")

w. 从字符串中获取指定字符串的第一个索引，没有则返回整数-1。例如从“aabaabaa”中获取“ab”的第一个索引1。

表达式：StringUtils.indexOf(value,"*ab*")

x. 从字符串中获取指定字符串的最后一个索引，没有则返回整数-1。例如从“aFkyk”中获取“k”的最后一个索引4。

表达式：StringUtils.lastIndexOf(value,"*k*")

y. 从字符串中指定的位置往后查找，获取指定字符串的第一个索引，没有则转换为“-1”。例如“aabaabaa”中索引3的后面，第一个“b”的索引是5。

表达式：StringUtils.indexOf(value,"*b*",*3*)

z. 从字符串获取指定字符串中任一字符的第一个索引，没有则返回整数-1。例如从“zzabyycdxx”中获取“z”或“a”的第一个索引0。

表达式：StringUtils.indexOfAny(value,"*za*")

aa. 如果字符串仅包含Unicode字符，返回布尔值true，否则返回false。例如“ab2c”中包含非Unicode字符，返回false。

表达式：StringUtils.isAlpha(value)

ab. 如果字符串仅包含Unicode字符或数字，返回布尔值true，否则返回false。例如“ab2c”中仅包含Unicode字符和数字，返回true。

表达式：StringUtils.isAlphanumeric(value)

ac. 如果字符串仅包含Unicode字符、数字或空格，返回布尔值true，否则返回false。例如“ab2c”中仅包含Unicode字符和数字，返回true。

表达式：StringUtils.isAlphanumericSpace(value)

ad. 如果字符串仅包含Unicode字符或空格，返回布尔值true，否则返回false。例如“ab2c”中包含Unicode字符和数字，返回false。

表达式：StringUtils.isAlphaSpace(value)

ae. 如果字符串仅包含ASCII可打印字符，返回布尔值true，否则返回false。例如“!ab-c~”返回true。

表达式：StringUtils.isAsciiPrintable(value)

af. 如果字符串为空或null，返回布尔值true，否则返回false。

表达式：StringUtils.isEmpty(value)

ag. 如果字符串中仅包含Unicode数字，返回布尔值true，否则返回false。

表达式：StringUtils.isNumeric(value)

ah. 获取字符串最左端的指定长度的字符，例如获取“abc”最左端的2位字符“ab”。

表达式：StringUtils.left(value,*2*)

ai. 获取字符串最右端的指定长度的字符，例如获取“abc”最右端的2位字符“bc”。

表达式：StringUtils.right(value,*2*)

aj. 将指定字符串拼接至当前字符串的左侧，需同时指定拼接后的字符串长度，如果当前字符串长度不小于指定长度，则不转换。例如将“yz”拼接到“bat”左侧，拼接后长度为8，则转换后为“yzyzybat”。

表达式：StringUtils.leftPad(value,*8*,"*yz*")

ak. 将指定字符串拼接至当前字符串的右侧，需同时指定拼接后的字符串长度，如果当前字符串长度不小于指定长度，则不转换。例如将“yz”拼接到“bat”右侧，拼接后长度为8，则转换后为“batyzyzy”。

表达式：StringUtils.rightPad(value,*8*,"*yz*")

al. 如果当前字段为字符串类型，获取当前字符串的长度，如果该字符串为null，则返回0。

表达式：StringUtils.length(value)

am. 如果当前字段为字符串类型，删除其中所有的指定字符串，例如从“queued”中删除“ue”，转换后为“qd”。

表达式：StringUtils.remove(value,"*ue*")

an. 如果当前字段为字符串类型，移除当前字段末尾指定的子字符串。指定的子字符串若不在当前字段的末尾，则不转换，例如移除当前字段“www.domain.com”后的“.com”。

表达式：StringUtils.removeEnd(value,"*.com*")

ao. 如果当前字段为字符串类型，移除当前字段开头指定的子字符串。指定的子字符串若不在当前字段的开头，则不转换，例如移除当前字段“www.domain.com”前的“www.”。

表达式：StringUtils.removeStart(value,"*www.*")

ap. 如果当前字段为字符串类型，替换当前字段中所有的指定字符串，例如将“aba”中的“a”用“z”替换，转换后为“zbz”。

表达式：StringUtils.replace(value,"*a*","*z*")

替换内容包含特殊字符时，需要先把该字符转义成普通字符，例如，客户想通过该表达式把字符串中 \t 去掉时，需要配置为：StringUtils.replace(value,"\\t","")（即把 \ 再次转义）。

aq. 如果当前字段为字符串类型，一次替换字符串中的多个字符，例如将字符串“hello”中的“h”用“j”替换，“o”用“y”替换，转换后为“jelly”。

表达式：StringUtils.replaceChars(value,"*ho*","*jy*")

ar. 如果字符串以指定的前缀开头（区分大小写），则返回布尔值true，否则返回false，例如当前字符串“abcdef”以“abc”开头，则返回true。

表达式：StringUtils.startsWith(value,"*abc*")

as. 如果当前字段为字符串类型，去除字段中首、尾处所有指定的字符，例如去除“abcyx”中首尾所有的“x”、“y”、“z”和“b”，转换后为“abc”。

表达式：StringUtils.strip(value,"*xyz*b")

at. 如果当前字段为字符串类型，去除字段末尾所有指定的字符，例如去除当前字段末尾的“abc”字符串。

表达式：StringUtils.stripEnd(value,*"abc"*)

au. 如果当前字段为字符串类型，去除字段开头所有指定的字符，例如去除当前字段开头的所有空格。

表达式：StringUtils.stripStart(value,*null*)

av. 如果当前字段为字符串类型，获取字符串指定位置后（索引从0开始，包括指定位置的字符）的子字符串，指定位置如果为负数，则从末尾往前计算位

置，末尾第一位为-1。例如获取“abcde”索引为2的字符（即c）及之后的字符串，则转换后为“cde”。

表达式：StringUtils.substring(value,*2*)

aw. 如果当前字段为字符串类型，获取字符串指定区间（索引从0开始，区间起点包括指定位置的字符，区间终点不包含指定位置的字符）的子字符串，区间位置如果为负数，则从末尾往前计算位置，末尾第一位为-1。例如获取“abcde”第2个字符（即c）及之后、第4个字符（即e）之前的字符串，则转换后为“cd”。

表达式：StringUtils.substring(value,*2*,4)

ax. 如果当前字段为字符串类型，获取当前字段里第一个指定字符后的子字符串。例如获取“abcba”中第一个“b”之后的子字符串，转换后为“cba”。

表达式：StringUtils.substringAfter(value,"*b*")

ay. 如果当前字段为字符串类型，获取当前字段里最后一个指定字符后的子字符串。例如获取“abcba”中最后一个“b”之后的子字符串，转换后为“a”。

表达式：StringUtils.substringAfterLast(value,"*b*")

az. 如果当前字段为字符串类型，获取当前字段里第一个指定字符前的子字符串。例如获取“abcba”中第一个“b”之前的子字符串，转换后为“a”。

表达式：StringUtils.substringBefore(value,"*b*")

ba. 如果当前字段为字符串类型，获取当前字段里最后一个指定字符前的子字符串。例如获取“abcba”中最后一个“b”之前的子字符串，转换后为“abc”。

表达式：StringUtils.substringBeforeLast(value,"*b*")

bb. 如果当前字段为字符串类型，获取嵌套在指定字符串之间的子字符串，没有匹配的则返回null。例如获取“tagabctag”中“tag”之间的子字符串，转换后为“abc”。

表达式：StringUtils.substringBetween(value,"*tag*")

bc. 如果当前字段为字符串类型，删除当前字符串两端的控制字符（char≤32），例如删除字符串前后的空格。

表达式：StringUtils.trim(value)

bd. 将当前字符串转换为字节，如果转换失败，则返回0。

表达式：NumberUtils.toByte(value)

be. 将当前字符串转换为字节，如果转换失败，则返回指定值，例如指定值配置为1。

表达式：NumberUtils.toByte(value,*1*)

bf. 将当前字符串转换为Double数值，如果转换失败，则返回0.0d。

表达式：NumberUtils.toDouble(value)

bg. 将当前字符串转换为Double数值，如果转换失败，则返回指定值，例如指定值配置为1.1d。

表达式：NumberUtils.toDouble(value,*1.1d*)

bh. 将当前字符串转换为Float数值，如果转换失败，则返回0.0f。

表达式：NumberUtils.toFloat(value)

bi. 将当前字符串转换为Float数值，如果转换失败，则返回指定值，例如配置指定值为1.1f。

表达式：NumberUtils.toFloat(value,*1.1f*)

bj. 将当前字符串转换为Int数值，如果转换失败，则返回0。

表达式：NumberUtils.toInt(value)

bk. 将当前字符串转换为Int数值，如果转换失败，则返回指定值，例如配置指定值为1。

表达式：NumberUtils.toInt(value,*1*)

bl. 将字符串转换为Long数值，如果转换失败，则返回0。

表达式：NumberUtils.toLong(value)

bm. 将当前字符串转换为Long数值，如果转换失败，则返回指定值，例如配置指定值为1L。

表达式：NumberUtils.toLong(value,*1L*)

bn. 将字符串转换为Short数值，如果转换失败，则返回0。

表达式：NumberUtils.toShort(value)

bo. 将当前字符串转换为Short数值，如果转换失败，则返回指定值，例如配置指定值为1。

表达式：NumberUtils.toShort(value,*1*)

bp. 将当前IP字符串转换为Long数值，例如将“10.78.124.0”转换为Long数值是“172915712”。

表达式：CommonUtils.ipToLong(value)

bq. 从网络读取一个IP与物理地址映射文件，并存放到Map集合，这里的URL是IP与地址映射文件存放地址，例如“http://10.114.205.45:21203/sqoop/IpList.csv”。

表达式：HttpsUtils.downloadMap("*url*")

br. 将IP与地址映射对象缓存起来并指定一个key值用于检索，例如“ipList”。

表达式：CommonUtils.setCache("*ipList*",HttpsUtils.downloadMap("*url*"))

bs. 取出缓存的IP与地址映射对象。

表达式：CommonUtils.getCache("*ipList*")

bt. 判断是否有IP与地址映射缓存。

表达式：CommonUtils.cacheExists("*ipList*")

bu. 根据指定的偏移类型（month/day/hour/minute/second）及偏移量（正数表示增加，负数表示减少），将指定格式的时间转换为一个新时间，例如将“2019-05-21 12:00:00”增加8个小时。

表达式：DateUtils.getCurrentTimeByZone("*yyyy-MM-dd HH:mm:ss*",value, "*hour*", *8*)

bv. 如果value值为空或者null时，则返回字符串“aaa”，否则返回value。

表达式：StringUtils.defaultIfEmpty(value,"aaa")

5.9.6 新增字段操作指导

操作场景

- 作业参数配置完成后，将进行字段映射的配置，您可以通过字段映射界面的可自定义新增字段。

- 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。

您可以单击字段映射界面的选择“添加新字段”自定义新增字段，通常用于标记数据库来源，以确保导入到目的端数据的完整性。

图 5-71 字段映射

目前支持以下类型自定义字段：

- **常量**

  常量参数即参数值是固定的参数，不需要重新配置值。例如“lable”=“friends”用来标识常量值。

- **变量**

  您可以使用时间宏、表名宏、版本宏等变量来标记数据库来源信息。变量的语法：${variable}，其中“variable”指的是变量。例如“input_time”=“${timestamp()}”用来标识当前时间的时间戳。

- **表达式**

  您可以使用表达式语言根据运行环境动态生成参数值。表达式的语法：#{expr}，其中“expr”指的是表达式。例如“time”=“#{DateUtil.now()}”用来标识当前日期字符串。

约束限制

- 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。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datetime）。
- 当作业源端为OBS、迁移CSV文件时，并且配置“解析首行为列名”参数的场景下显示列名。
- 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。
- 自动创表场景下，需在目的端表中提前手动新增字段，再在字段映射里新增字段。
- 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。
- 如果字段映射关系不正确，您可以通过拖拽字段、单击对字段批量映射两种方式来调整字段映射关系。

- 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：
  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CDM不支持源端迁移字段类型，请参见**不支持数据类型转换规避指导**将字段类型转换为CDM支持的类型。

5.9.7 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。

5.9.8 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 5-72 正则表达式参数

在迁移CSV格式的文件时，CDM支持使用正则表达式分隔字段，并按照解析后的结果写入目的端。正则表达式语法请参考对应的相关资料，这里举例下面几种日志文件的正则表达式的写法：

- **Log4J日志**
- **Log4J审计日志**
- **Tomcat日志**
- **Django日志**
- **Apache server日志**

Log4J 日志

- 日志样例：
```
2018-01-11 08:50:59,001 INFO
[org.apache.sqoop.core.SqoopConfiguration.configureClassLoader(SqoopConfiguration.java:251)]
Adding jars to current classloader from property: org.apache.sqoop.classpath.extra
```
- 正则表达式为：
```
^(\d.*\d) (\w*) \[(.*)\] (\w.*).*
```
- 解析出的结果如下：

表 5-112 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 5-113 Log4J 审计日志解析结果

列号	样值
1	2018-01-11 08:51:06,156
2	INFO

列号	样值
3	org.apache.sqoop.audit.FileAuditLogger.logAuditEvent(FileAuditLogger.java:61)
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 5-114 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 5-115 Django 日志解析结果

列号	样值
1	08/Jan/2018 20:59:07
2	settings

列号	样值
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 5-116 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

5.9.9 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。
- 目的端数据表中已有时间日期字段或时间戳字段。如自动创表场景下，需提前在目的端表中手动创建时间日期字段或时间戳字段。

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 5-73 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 5-74 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5-117。

图 5-75 添加字段

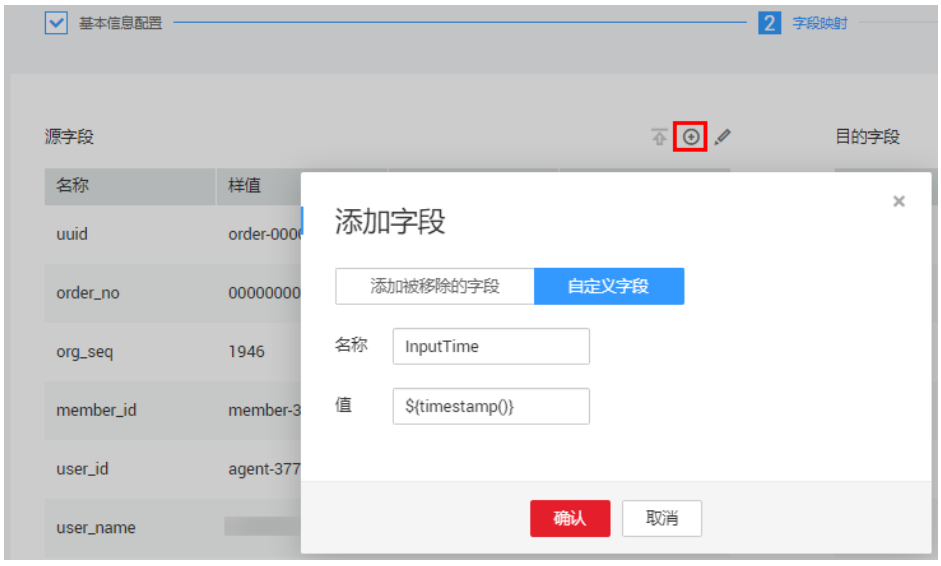


表 5-117 时间变量宏定义具体展示

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

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 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
${dateformat(yyyy-MM-dd, -1, DAY)} 00:00:00	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前一天0点。	2017-10-15 00:00:00
${dateformat(yyyy-MM-dd, -1, DAY)} 12:00:00	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前一天12点。	2017-10-15 12:00:00
${dateformat(yyyy-MM-dd, -N, DAY)} 00:00:00	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前N天的0点。	N为3时： 2017-10-13 00:00:00
${dateformat(yyyy-MM-dd, -N, DAY)} 12:00:00	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前N天的12点。	N为3时： 2017-10-13 12:00:00
${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。
- 添加完字段后，请确保自定义入库时间字段与目的端表字段类型相匹配。

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。

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。

步骤7 前往目的端数据源查看数据迁移的入库时间。

----结束

5.9.10 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 5-118 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5-118。

3. **编码类型**

   文件的编码类型，默认是UTF-8，中文的编码有时会采用GBK。

   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5-76所示，数据库的name字段的值中包含了字段分隔符逗号：

     图 5-76 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文件为源端，并且其中数据是被包围符括起来的时候，如果想把数据正确的导入到数据库，就需要在源端开启“使用包围符”，这样包围符内的值的，会写入一个字段内。

5. **使用正则表达式分隔字段**

   这个功能是针对一些复杂的半结构化文本，例如日志文件的解析，详见**使用正则表达式分隔半结构化文本**。

6. **首行为标题行**

   这个参数是针对CSV文件导出到其它地方的场景，如果源端指定了该参数，CDM在抽取数据时将第一行作为标题行。在传输CSV文件的时候会跳过标题行，这时源端抽取的行数，会比目的端写入的行数多一行，并在日志文件中进行说明跳过了标题行。

7. **写入文件大小**

   这个参数是针对数据库导出到CSV文件的场景，如果一张表的数据量比较大，那么导出到CSV文件的时候，会生成一个很大的文件，有时会不方便下载或查看。这时可以在目的端指定该参数，这样会生成多个指定大小的CSV文件，避免导出的文件过大。该参数的数据类型为整型，单位为MB。

JSON 格式

这里主要介绍JSON文件格式的以下内容：

- **CDM支持解析的JSON类型**
- **记录节点**
- **从JSON文件复制数据**

1. **CDM支持解析的JSON类型：JSON对象、JSON数组。**
   - JSON对象：JSON文件包含单个对象，或者以行分隔/串连的多个对象。
     i. 单一对象JSON

     ```
     {
       "took" : 190,
       "timed_out" : false,
       "total" : 1000001,
       "max_score" : 1.0
     }
     ```

     ii. 行分隔的JSON对象

     ```
     {"took" : 188, "timed_out" : false, "total" : 1000003, "max_score" : 1.0 }
     {"took" : 189, "timed_out" : false, "total" : 1000004, "max_score" : 1.0 }
     ```

     iii. 串连的JSON对象

     ```
     {
       "took": 190,
       "timed_out": false,
       "total": 1000001,
       "max_score": 1.0
     }
     {
       "took": 191,
       "timed_out": false,
       "total": 1000002,
       "max_score": 1.0
     }
     ```

   - JSON数组：JSON文件是包含多个JSON对象的数组。

     ```
     [{
       "took" : 190,
     ```



```
    "timed_out" : false,
    "total" : 1000001,
    "max_score" : 1.0
 },
 {
    "took" : 191,
    "timed_out" : false,
    "total" : 1000001,
    "max_score" : 1.0
 }]
```

2. **记录节点**

   记录数据的根节点。该节点对应的数据为JSON数组，CDM会以同一模式从该数组中提取数据。多层嵌套的JSON节点以字符“.”分隔。

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。

   表 5-119 示例

took	timedOut	total	maxScore
190	false	1000001	1.0
191	false	1000002	1.0
192	false	1000003	1.0

   b. 示例二

   从记录节点中提取数据。JSON文件包含了单个的JSON对象，但是其中有效的数据在一个数据节点下，例如：

```
{
  "took": 190,
  "timed_out": false,
  "hits": {
     "total": 1000001,
     "max_score": 1.0,
     "hits":
```

```
    [{
      "_id": "650612",
      "_source": {
         "name": "tom",
         "books": ["book1","book2","book3"]
      }
    },
    {
      "_id": "650616",
      "_source": {
         "name": "tom",
         "books": ["book1","book2","book3"]
      }
    },
   {
      "_id": "650618",
      "_source": {
         "name": "tom",
         "books": ["book1","book2","book3"]
      }
    }]
  }
}
```

如果想以如下格式写入到数据库，则需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON对象”，并且指定记录节点为“hits.hits”，然后在作业第二步进行字段匹配。

表 5-120 示例

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。

表 5-121 示例

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：

表 5-122 示例

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 5-77 添加字段

ii. 在添加的新字段后面，单击添加字段转换器。

图 5-78 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 5-79 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高、性能稳定，且不需要在作业第二步进行字段匹配。

- **文件传输的目录结构**

  CDM的文件传输，支持单文件，也支持一次传输目录下所有的文件。传输到目的端后，目录结构会保持原样。

- **增量迁移文件**

  使用CDM进行二进制传输文件时，目的端有一个参数“重复文件处理方式”，可以用作文件的增量迁移，具体请参见**文件增量迁移**。

  增量迁移文件的时候，选择“重复文件处理方式”为“跳过重复文件”，这样如果源端有新增的文件，或者是迁移过程中出现了失败，只需要再次运行任务，已经迁移过的文件就不会再次迁移。

- **写入到临时文件**

  二进制迁移文件时候，可以在目的端指定是否写入到临时文件。如果指定了该参数，在文件复制过程中，会将文件先写入到一个临时文件中，迁移成功后，再进行rename或move操作，在目的端恢复文件。

- **生成文件MD5值**

  对每个传输的文件都生成一个MD5值，并将该值记录在一个新文件中，新文件以“.md5”作为后缀，并且可以指定MD5值生成的目录。

文件格式的公共参数

- **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：

   - 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5-118**。

   - 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。

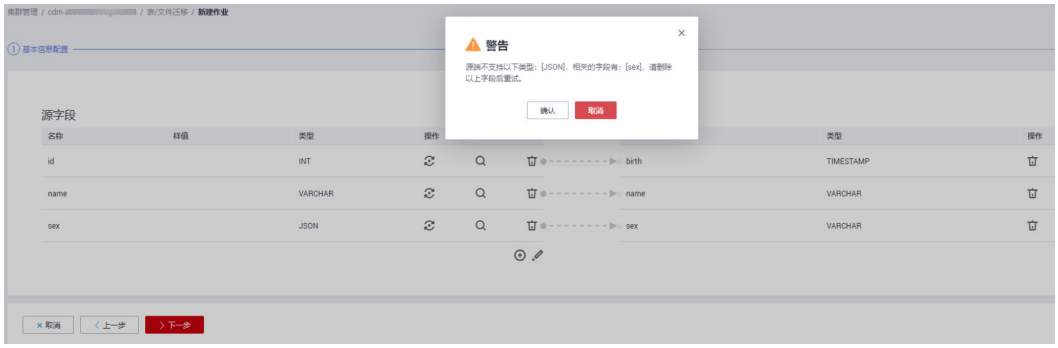
2. 数据库的数据包含换行符
   - 场景：使用CDM先将MySQL中的某张表（表的某个字段值中包含了换行符\n）导出到CSV格式的文件中，然后再使用CDM将导出的CSV文件导入到MRS HBase，发现导出的CSV文件中出现了数据被截断的情况。
   - 解决方法：指定换行符。

     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。

5.9.11 不支持数据类型转换规避指导

操作场景

CDM在配置字段映射时提示字段的数据类型不支持，要求删除该字段。如果需要使用该字段，可在源端作业配置中使用SQL语句对字段类型进行转换，转换成CDM支持的类型，达到迁移数据的目的。



操作步骤

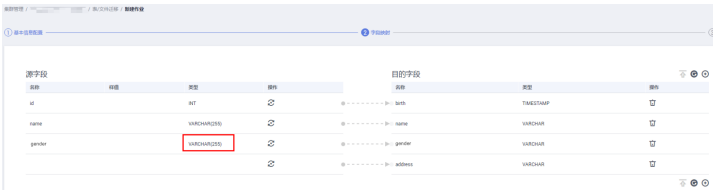
步骤1 修改CDM迁移作业，通过使用SQL语句的方式迁移。

说明

SQL语句格式为：“select id,cast(原字段名 as INT) as 新字段名可以和原字段名一样 from schemaName.tableName；”

例如：select `id`, `name`, cast(`gender` AS char(255)) AS `gender` from `test_1117869`.`test_no_support_type`;

步骤2 转换后的字段就转换为CDM支持的数据类型。



----结束

5.9.12 自动建表原理介绍

CDM将根据源端的字段类型进行默认规则转换成目的端字段类型，并在目的端建数据表。

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5-80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 5-80 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	SAP HANA	DWS
NUMBER(p,0) (p=3 or p=5)	SMALLINT,TINYINT	SMALLINT,TINYINT	SMALLINT	SMALLINT,TINYINT	SMALLINT
NUMBER(10,0)	INT	INT	INTEGER	INTEGER	INTEGER
NUMBER(19,0)	BIGINT	BIGINT	BIGINT	BIGINT	BIGINT
无	无	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(p,s) (0 < p <= 38)	NUMERIC(p,s) (p <= 1000)
RAW	BINARY	BINARY	BYTEA	BINARY	BYTEA
CHAR	CHAR	CHAR	CHAR	CHAR(p) (p <= 2000)	CHAR
NCHAR	NCHAR	NCHAR	NCHAR	NCHAR(p) (p <= 5000)	NCHAR
DATE	DATE	DATE	DATE	DATE	DATE
DATE	DATETIME	DATETIME2	TIMESTAMP	TIMESTAMP	TIMESTAMP
VARCHAR2(p) (p <= 4000)	VARCHAR	VARCHAR(p) (if p >= 8000 p=max)	VARCHAR(p) (p <= 10485760)	VARCHAR(p) (p <= 5000)	VARCHAR(p) (p <= 10485760)
FLOAT	DOUBLE	FLOAT	DOUBLE PRECISION	DOUBLE	DOUBLE PRECISION
FLOAT	REAL	FLOAT	REAL	REAL	REAL
CLOB	TEXT	TEXT	TEXT	CLOB	TEXT
DATE	无	TIME	TIME	TIME	TIME
BOOLEAN	无	无	BOOLEAN	BOOLEAN	BOOLEAN

CDM在Hive中自动建表时，Hive表与源表的字段类型映射关系参见表5-123、表5-124、表5-125及表5-126。例如使用CDM将MySQL整库迁移到Hive，CDM在Hive上自动建表，会将Oracle的**YEAR**字段映射到Hive的**DATE**。

说明

- 针对DECIMAL类型，源端数据源长度超过Hive长度可能导致精度丢失。
- Hive DECIMAL(P,S)类型1 ≤ precision ≤ 38,0 ≤ scale。源端p > 38位时，Hive按38位创建，s小于0时，按0创建，受Hive数据类型限制，此场景可能会导致数据写入后精度丢失。

表 5-123 MySQL->Hive 自动建表时的字段映射

数据类型（MySQL）	数据类型（Hive）	说明
数值类型		
tinyint(1)，bit(1)	BOOLEAN	-
TINYINT	SMALLINT	-
TINYINT UNSIGNED	SMALLINT	-

数据类型（MySQL）	数据类型（Hive）	说明
SMALLINT	SMALLINT	-
SMALLINT UNSIGNED	INTEGER	-
MEDIUMINT	INTEGER	-
MEDIUMINT UNSIGNED	BIGINT	-
INT	INTEGER	-
INT UNSIGNED	BIGINT	-
BIGINT	BIGINT	-
BIGINT UNSIGNED	DECIMAL(38,0)	-
DECIMAL(P,S)	DECIMAL(P,S)	MySQL最大位数为65位,Hive 1 <= precision <= 38,0 <= scale。MySQL p > 38位时，hive按38位创建，s 小于0时，按0创建。
FLOAT	FLOAT	-
FLOAT UNSIGNED	FLOAT	-
DOUBLE	DOUBLE	-
DOUBLE UNSIGNED	DOUBLE	-
时间类型		
DATE	DATE	-
YEAR	DATE	-
DATETIME	TIMESTAMP	-
TIMESTAMP	TIMESTAMP	-
TIME	STRING	-
字符类型		
CHAR(N)	CHAR(N*3)	（n*3<255 ）大于255(CHAR_MAX_LENGTH)时，创建为varchar(N*3), 大于65535(VARCHAR_MAX_LENGTH)时创建为String。
VARCHAR(N)	VARCHAR(N*3)	大于65535(VARCHAR_MAX_LENGTH)时创建为String。

数据类型（MySQL）	数据类型（Hive）	说明
BINARY	BINARY	-
VARBINARY	BINARY	-
TINYBLOB	BINARY	-
MEDIUMBLOB	BINARY	-
BLOB	BINARY	-
LONGBLOB	BINARY	-
TINYTEXT	VARCHAR(765)	-
MEDIUMTEXT	STRING	-
TEXT	STRING	-
LONGTEXT	STRING	-
其他类型	STRING	-

表 5-124 Oracle->Hive 自动建表时的字段映射

数据类型（Oracle）	数据类型（Hive）	说明
字符类型		
CHAR(N)	CHAR(N*3)	（n*3<255 ）大于255(CHAR_MAX_LENGTH)时，创建为varchar(N*3), 大于65535(VARCHAR_MAX_LENGTH)时创建为String。
VARCHAR(N)	VARCHAR(N*3)	大于65535(VARCHAR_MAX_LENGTH)时创建为String。
VARCHAR2	VARCHAR(N*3)	大于65535(VARCHAR_MAX_LENGTH)时创建为String。
NCHAR	CHAR(N*3)	-
NVARCHAR2	STRING	-
数值类型		
NUMBER	DECIMAL(P,S)	Hive 1 <= precision <= 38,0 <= scale。MySQL p > 38位时，hive按38位创建，s 小于0时，按0创建。
BINARY_FLOAT	FLOAT	-
BINARY_DOUBLE	DOUBLE	-

数据类型（Oracle）	数据类型（Hive）	说明
FLOAT	FLOAT	-
时间类型		
DATE	TIMESTAMP	-
TIMESTAMP	TIMESTAMP	-
TIMESTAMP WITH TIME ZONE	STRING	-
TIMESTAMP WITH LOCAL TIME ZONE	STRING	-
INTERVAL	STRING	-
二进制类型		
BLOB	BINARY	-
CLOB	STRING	-
NCLOB	STRING	-
LONG	STRING	-
LONG_RAW	BINARY	-
RAW	BINARY	-
其他类型	STRING	-

表 5-125 PostgreSQL、DWS->Hive 自动建表时的字段映射

数据类型（PostgreSQL、DWS）	数据类型（Hive）	说明
数值类型		
int2	SMALLINT	-
int4	INT	-
int8	BIGINT	-
real	FLOAT	-
float4	FLOAT	-
float8	DOUBLE	-
smallserial	SMALLINT	-

数据类型（PostgreSQL、DWS）	数据类型（Hive）	说明
serial	INT	-
bigserial	BIGINT	-
numeric(p,s)	DECIMAL(P,S)	Hive 1 <= precision <= 38,0 <= scale。MySQL p > 38位时，hive按38位创建，s小于0时，按0创建。
money	DOUBLE	-
bit(1)	TINYINT	-
varbit	STRING	-
字符类型		
varchar(n)	VARCHAR(N*3)	大于65535(VARCHAR_MAX_LENGTH)时创建为String。
bpchar(n)	CHAR(N*3)	（n*3<255 ）大于255(CHAR_MAX_LENGTH)时，创建为varchar(N*3), 大于65535(VARCHAR_MAX_LENGTH)时创建为String。
char(n)	CHAR(N*3)	（n*3<255 ）大于255(CHAR_MAX_LENGTH)时，创建为varchar(N*3), 大于65535(VARCHAR_MAX_LENGTH)时创建为String。
bytea	BINARY	-
text	STRING	-
时间类型		
interval	STRING	-
date	DATE	-
time	STRING	-
timetz	STRING	-
timestamp	TIMESTAMP	-
timestamptz	TIMESTAMP	-
布尔类型		
bool	BOOLEAN	-
其他类型	STRING	-

表 5-126 SQL Server->Hive 自动建表时的字段映射

数据类型（ SQL Server ）	数据类型（ Hive ）	说明
数值类型		
TINYINT	SMALLINT	-
SMALLINT	SMALLINT	-
INT	INT	-
BIGINT	BIGINT	-
DECIMAL	DECIMAL(P,S)	Hive 1 <= precision <= 38,0 <= scale。MySQL p > 38位时，hive按38位创建，s小于0时，按0创建。
NUMERIC	DECIMAL(P,S)	Hive 1 <= precision <= 38,0 <= scale。MySQL p > 38位时，hive按38位创建，s小于0时，按0创建。
FLOAT	DOUBLE	-
REAL	FLOAT	-
SMALLMONEY	DECIMAL(10,4)	-
MONEY	DECIMAL(19,4)	-
BIT(1)	TINYINT	-
时间类型		
DATE	DATE	-
DATETIME	TIMESTAMP	-
DATETIME2	TIMESTAMP	-
DATETIMEOFFSET	STRING	-
TIME(p)	STRING	-
TIMESTAMP	BINARY	-
字符类型		
CHAR(n)	CHAR(n*3)	（ n*3<255 ）大于255(CHAR_MAX_LENGTH)时，创建为varchar(N*3), 大于65535(VARCHAR_MAX_LENGTH)时创建为String。
VARCHAR(n)	VARCHAR(n*3)	（ n*3<255 ）大于255(CHAR_MAX_LENGTH)时，创建为varchar(N*3), 大于65536(VARCHAR_MAX_LENGTH)时创建为String。

数据类型（ SQL Server ）	数据类型（ Hive ）	说明
NCHAR(n)	VARCHAR(n*3)	（ n*3<255 ）大于255(CHAR_MAX_LENGTH)时，创建为varchar(N*3), 大于65537(VARCHAR_MAX_LENGTH)时创建为String。
NVARCHAR(n)	VARCHAR(n*3)	（ n*3<255 ）大于255(CHAR_MAX_LENGTH)时，创建为varchar(N*3), 大于65538(VARCHAR_MAX_LENGTH)时创建为String。
二进制类型		
BINARY	BINARY	-
VARBINARY	BINARY	-
TEXT	STRING	-
其他类型	STRING	-

5.10 使用教程

5.10.1 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。

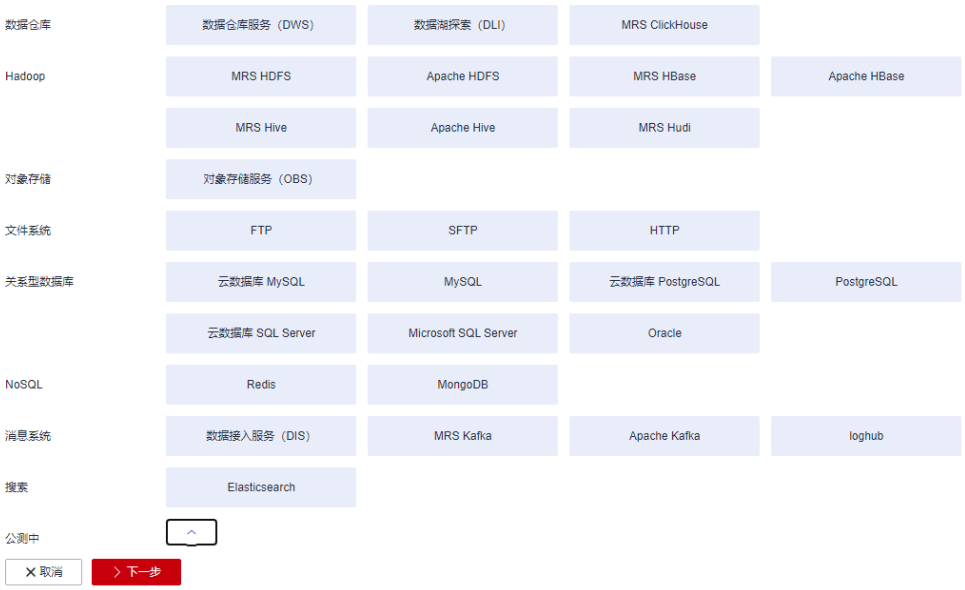
前提条件

- 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**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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

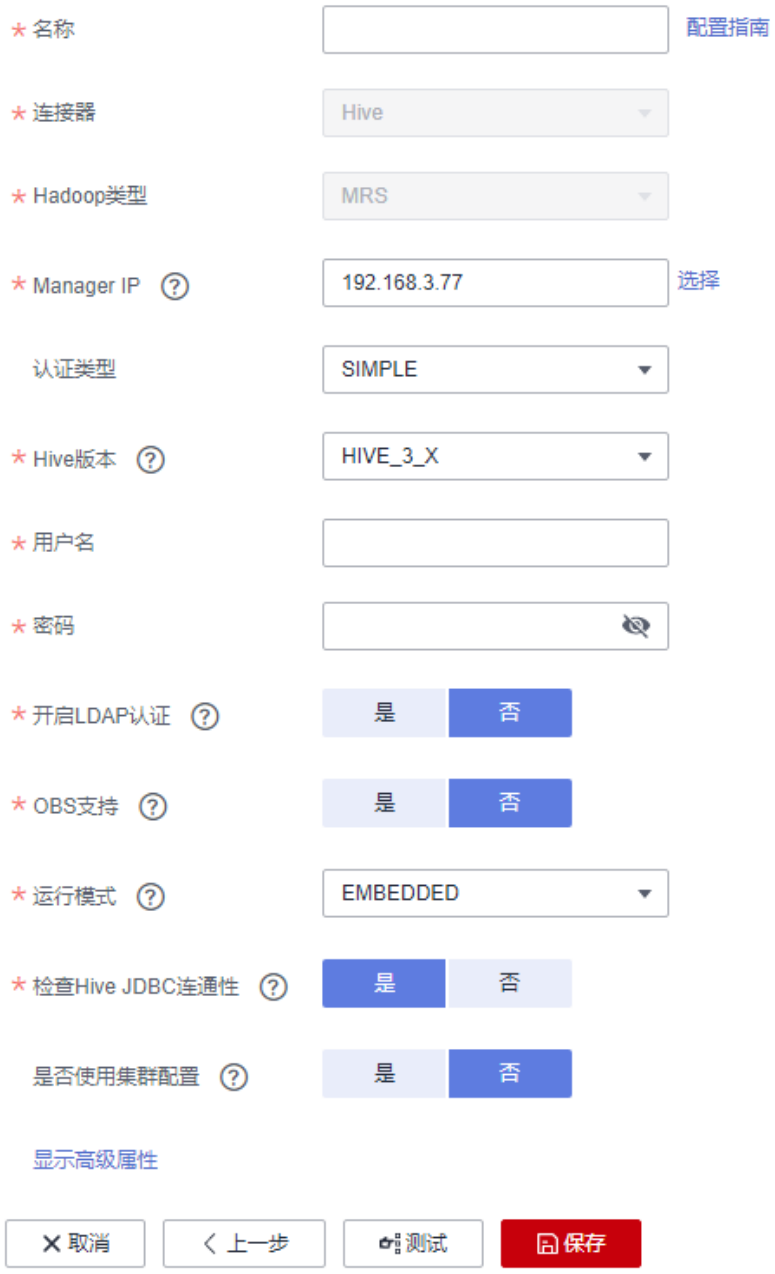
步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如图5-81所示。

图 5-81 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5-82所示。

图 5-82 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，必填参数如下表所示。

表 5-127 MRS Hive 连接参数

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

参数名	说明	取值样例
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。	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。	-
开启LDAP认证	通过代理连接的时候，此项可配置。 当MRS Hive对接外部LDAP开启了LDAP认证时，连接Hive时需要使用LDAP账号与密码进行认证，此时必须开启此参数，否则会连接失败。	否
LDAP用户名	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的用户名。	-

参数名	说明	取值样例
LDAP密码	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 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5-83所示。 **图 5-83** 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)		-

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

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

5.10.2 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。

步骤2 在“驱动管理”页面，单击MySQL驱动“建议版本”列中的资料链接，按照相应指导获取驱动文件。

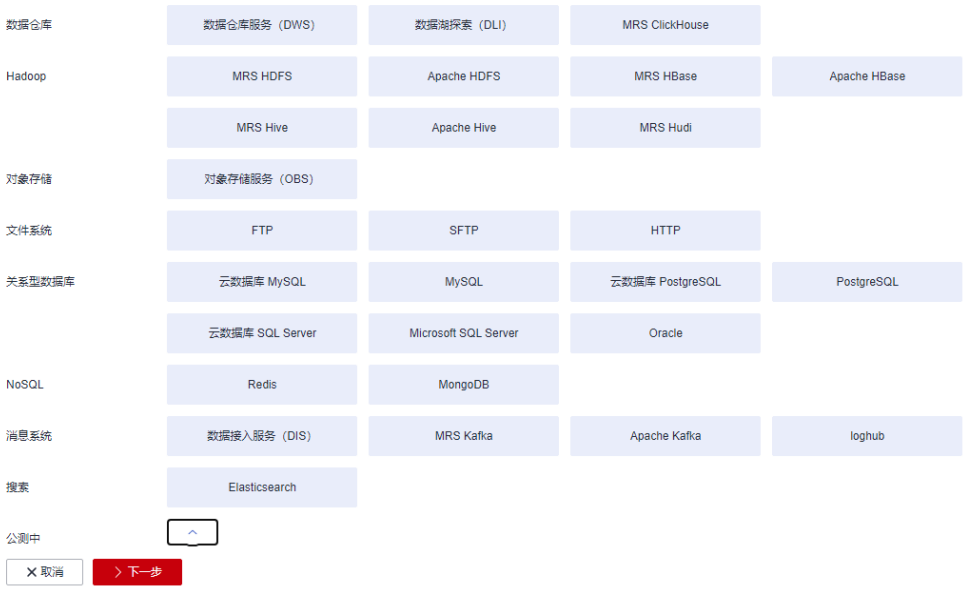
步骤3 在“驱动管理”页面中，选择以下方式上传MySQL驱动。

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5-84所示。

图 5-84 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。

表 5-128 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。	-

参数名	说明	取值样例
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	Agent功能待下线，无需配置。	-
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	Agent功能待下线，无需配置。	-
单次请求行数	指定每次请求获取的行数。	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。	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

5.10.3 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5-85所示。

图 5-85 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

- 已经购买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。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，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 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 5-86 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，如图5-87所示。

图 5-87 选择连接器类型

步骤2 选择“云数据库 MySQL”后单击“下一步”，配置云数据库 MySQL连接的参数。

图 5-88 创建 MySQL 连接

单击“显示高级属性”可查看更多可选参数，具体请参见**配置云数据库MySQL/MySQL数据库连接**。这里保持默认，必填参数如**表5-129**所示。

表 5-129 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。	-
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。	否
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	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。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

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

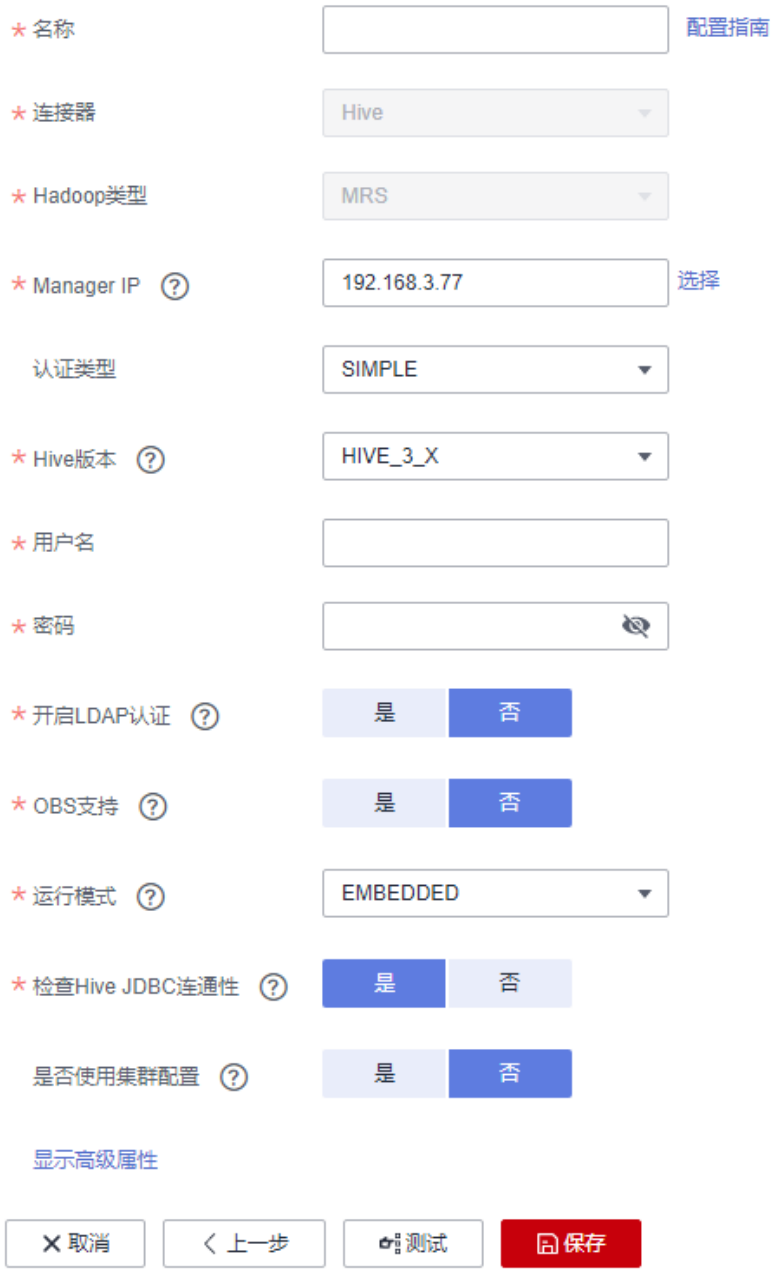
创建 Hive 连接

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，如**图5-89**所示。

图 5-89 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Hive连接参数，如图5-90所示。

图 5-90 创建 MRS Hive 连接



各参数说明如表5-130所示，需要您根据实际情况配置。

表 5-130 MRS Hive 连接参数

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

参数名	说明	取值样例
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。 **说明** 当前DataArts Studio不支持对接“Kerberos加密类型”为“aes256-sha2,aes128-sha2”的MRS集群。如需对接MRS集群，请注意“Kerberos加密类型”应为“aes256-sha1,aes128-sha1”。	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。	-
开启LDAP认证	通过代理连接的时候，此项可配置。 当MRS Hive对接外部LDAP开启了LDAP认证时，连接Hive时需要使用LDAP账号与密码进行认证，此时必须开启此参数，否则会连接失败。	否
LDAP用户名	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的用户名。	-

参数名	说明	取值样例
LDAP密码	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 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5-91所示。 **图 5-91** 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)		-

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

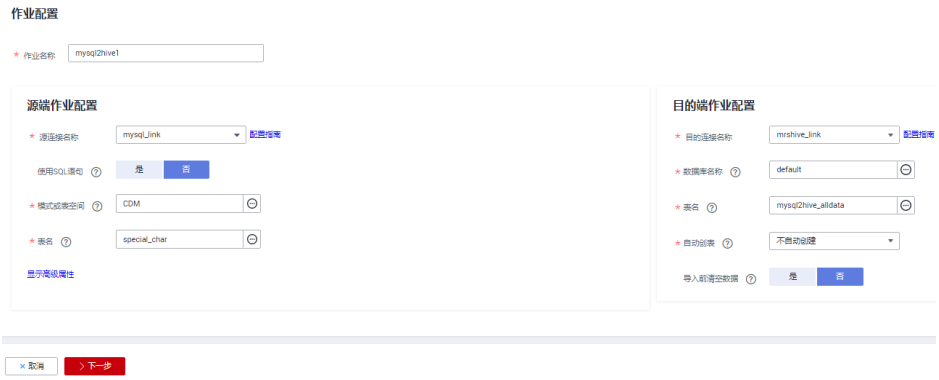
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建数据迁移任务，如**图5-92**所示。

图 5-92 创建 MySQL 到 Hive 的迁移任务



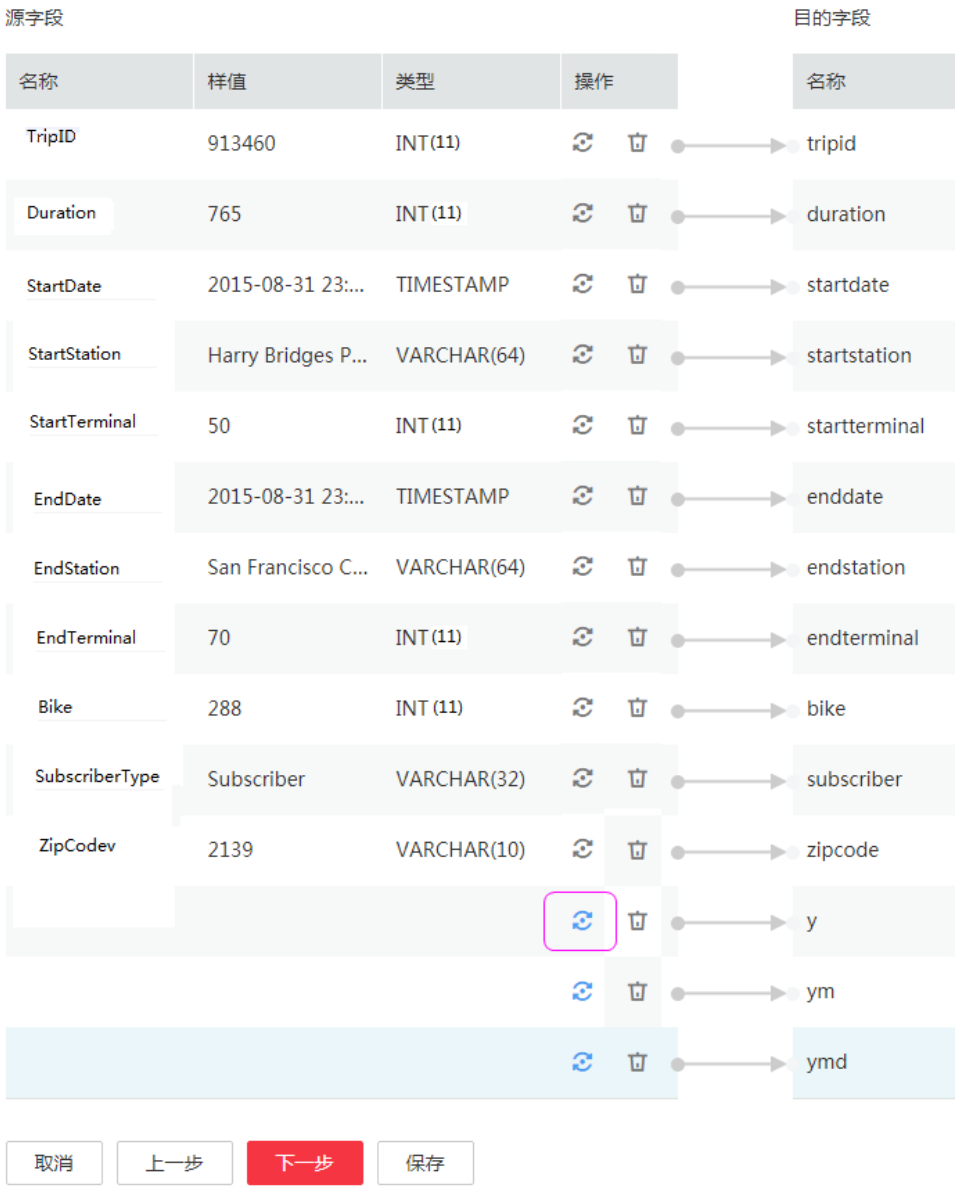
说明

“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5-93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 5-93 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5-94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 5-94 配置表达式

说明

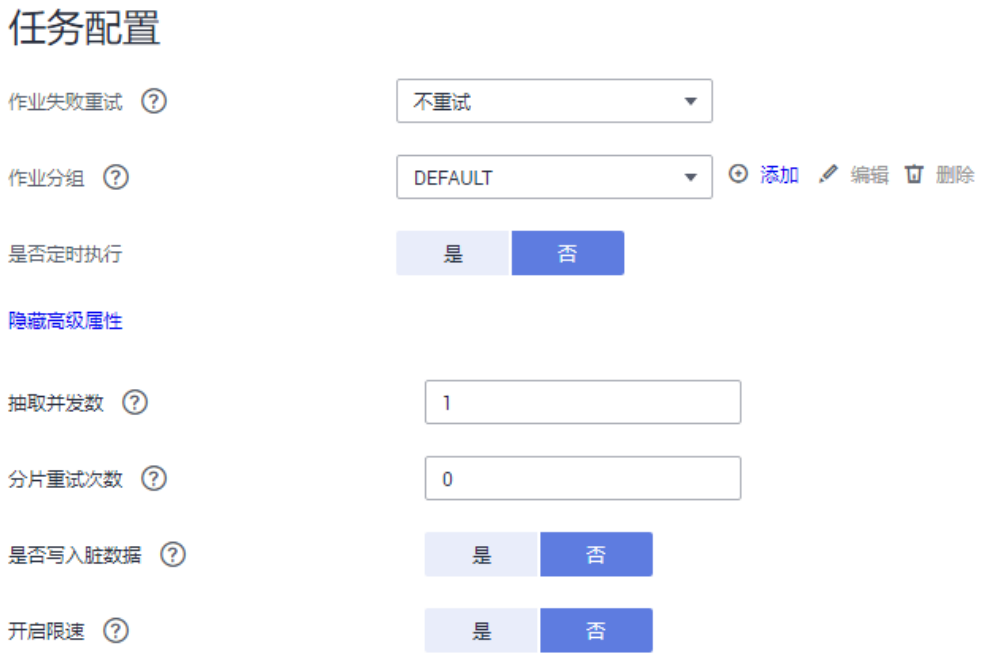
CDM的表达式已经预置常用字符串、日期、数值等类型的字段内容转换，详细请参见**字段转换**。

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可开启。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，适当的抽取并发数可以提升迁移效率，配置原则请参见**性能调优**。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，写入脏数据前需要在CDM先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。

图 5-95 任务配置



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

5.10.4 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。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了MySQL数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

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

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

说明

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

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图5-96**所示。

图 5-96 选择连接器类型

步骤2 选择“MySQL”后单击“下一步”，配置MySQL连接的参数。

单击“显示高级属性”可查看更多可选参数，具体请参见**配置云数据库MySQL/MySQL数据库连接**。这里保持默认，必填参数如**表5-131**所示。

表 5-131 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

参数名	说明	取值样例
数据库服务器	MySQL数据库的IP地址或域名。	-
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	Agent功能待下线，无需配置。	-
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	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。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

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，如**图5-97**所示。

图 5-97 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，如图5-99所示。

- 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。

  您可以通过如下方式获取访问密钥。

  a. 登录控制台，在用户名下拉列表中选择“我的凭证”。

  b. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图5-98所示。

  图 5-98 单击新增访问密钥

  c. 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。

图 5-99 创建 OBS 连接

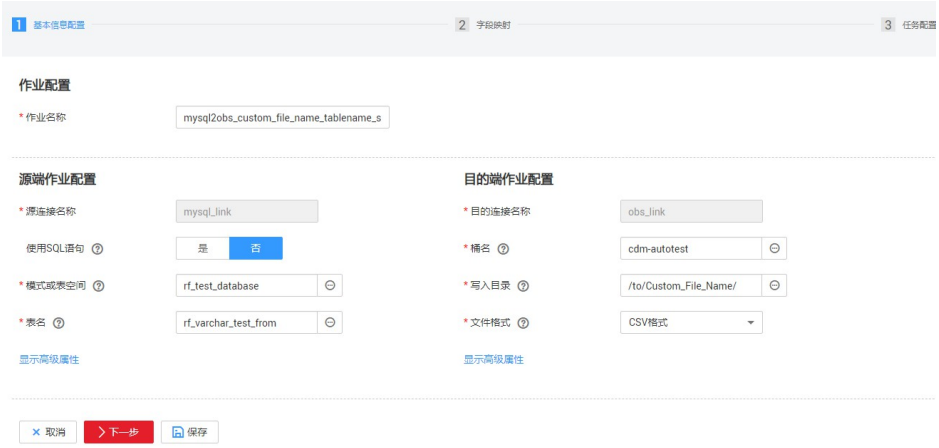
步骤3 单击“保存”回到连接管理界面。

----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 5-100 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。
- 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即可。

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5-101**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，详细请参见**字段转换**。

图 5-101 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，可打开此配置。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

5.10.5 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。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了MySQL数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

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

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

说明

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

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图5-102**所示。

图 5-102 选择连接器类型

步骤2 选择“MySQL”后单击“下一步”，配置MySQL连接的参数。

单击“显示高级属性”可查看更多可选参数，具体请参见**配置云数据库MySQL/MySQL数据库连接**。这里保持默认，必填参数如**表5-132**所示。

表 5-132 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。	-
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	Agent功能待下线，无需配置。	-
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8

参数名	说明	取值样例
驱动版本	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。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

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，如**图5-103**所示。

图 5-103 选择连接器类型

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5-133**所示，可选参数保持默认即可。

表 5-133 DWS 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	dwslink
数据库服务器	DWS数据库的IP地址或域名。	192.168.0.3
端口	DWS数据库的端口。	8000

参数名	说明	取值样例
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	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
导入模式	COPY模式：将源数据经过DWS管理节点后复制到数据节点。如果需要通过Internet访问DWS，只能使用COPY模式。	COPY

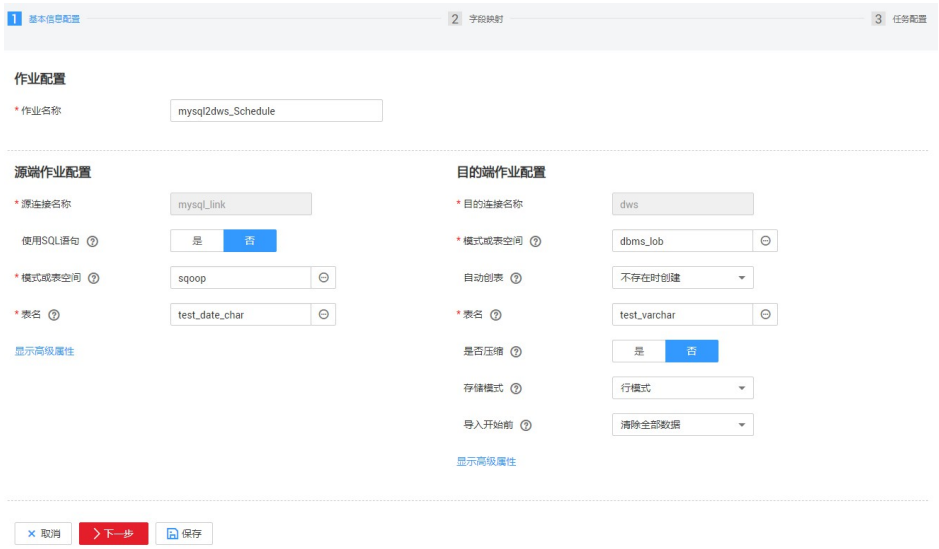
步骤3 单击“保存”完成创建连接。

----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 5-104 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。
- 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。更多压缩级别详细说明请参见**压缩级别**。
  - 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5-105**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，详细请参见**字段转换**。

图 5-105 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，可打开此配置。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

5.10.6 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。
- 用户已购买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。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了MySQL数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

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

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

图 5-106 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，如图5-107所示。

图 5-107 选择连接器类型

步骤2 选择“MySQL”后单击“下一步”，配置MySQL连接的参数。

单击“显示高级属性”可查看更多可选参数，具体请参见**配置云数据库MySQL/MySQL数据库连接**。这里保持默认，必填参数如**表5-134**所示。

表 5-134 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

参数名	说明	取值样例
数据库服务器	MySQL数据库的IP地址或域名。	-
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	Agent功能待下线，无需配置。	-
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	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。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

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，如**图5-108**所示。

图 5-108 选择连接器类型

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5-109所示。

图 5-109 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“mysqllink”。
  - 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。
  - 导入开始前：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。
  - 约束冲突处理：选择“insert into”，当迁移数据出现唯一约束冲突时的处理方式。
  - 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

5.10.7 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华为云的VPN通道或专线。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了Oracle数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群的VPC必须和云搜索服务集群所在VPC一致，且推荐子网、安全组也与云搜索服务一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

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

说明

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

----结束

创建云搜索服务连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图5-110**所示。

图 5-110 选择连接器类型

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。

图 5-111 创建云搜索服务连接

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，如图5-112所示。

图 5-112 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：

- 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。

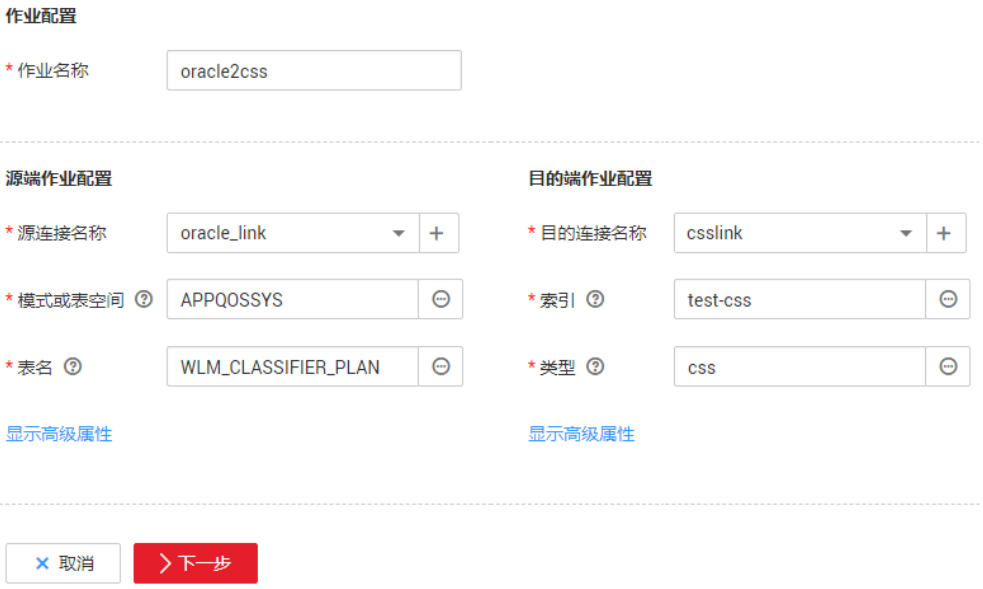
步骤3 单击“保存”回到连接管理界面。

----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 5-113 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即可。
- 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即可。

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5-114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，详细请参见**字段转换**。

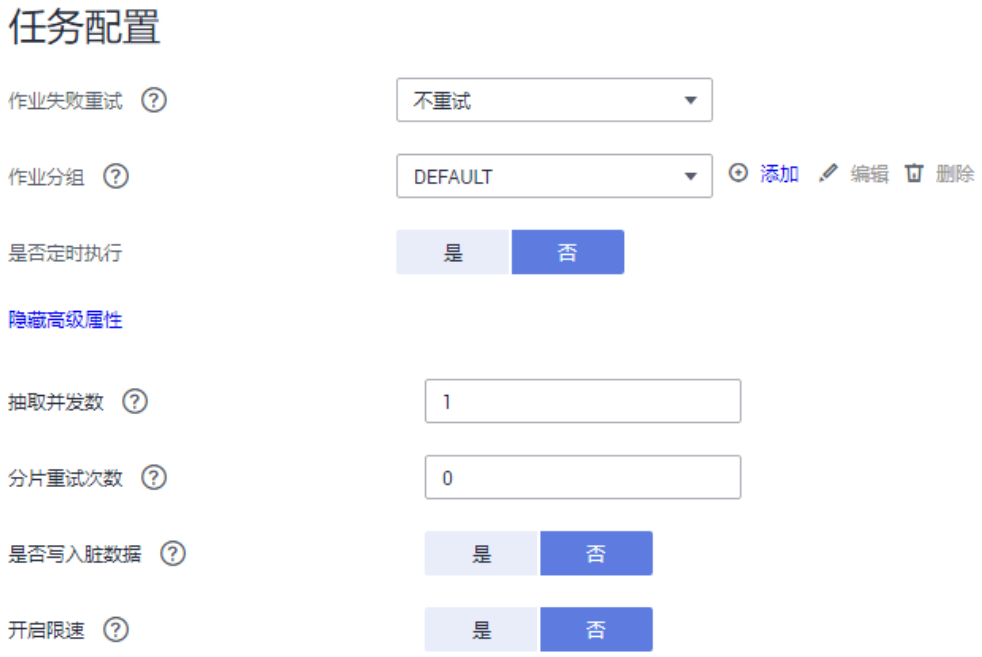
图 5-114 云搜索服务的字段映射

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可开启。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，适当的抽取并发数可以提升迁移效率，配置原则请参见**性能调优**。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，写入脏数据前需要在CDM先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。

图 5-115 任务配置



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

5.10.8 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华为云的VPN通道或专线。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了Oracle数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群所在VPC、子网、安全组，选择与DWS集群所在的网络一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

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

说明

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

----结束

创建 Oracle 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图5-116**所示。

图 5-116 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5-135**所示。

图 5-117 创建 Oracle 连接

表 5-135 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	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
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。	'

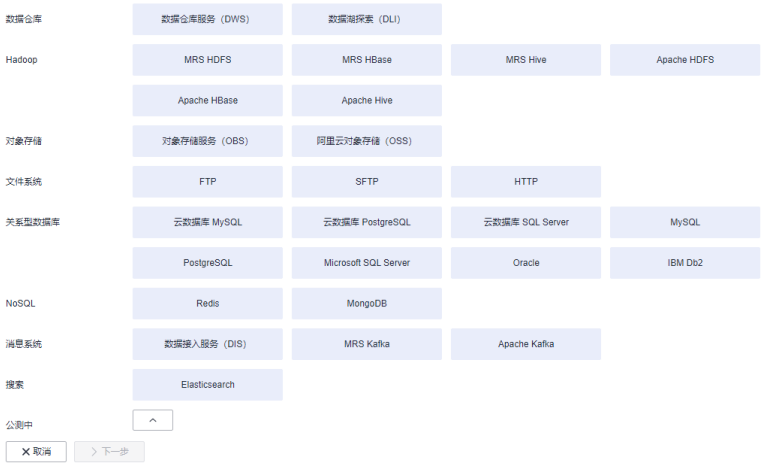
步骤3 单击“保存”回到连接管理界面。

----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，如图5-118所示。

图 5-118 选择连接器类型



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5-136**所示，可选参数保持默认即可。

表 5-136 DWS 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	dwslink
数据库服务器	DWS数据库的IP地址或域名。	192.168.0.3
端口	DWS数据库的端口。	8000
数据库名称	DWS数据库的名称。	db_demo
用户名	拥有DWS数据库的读、写和删除权限的用户。	dbadmin
密码	用户的密码。	-
使用Agent	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
导入模式	COPY模式：将源数据经过DWS管理节点后复制到数据节点。如果需要通过Internet访问DWS，只能使用COPY模式。	COPY

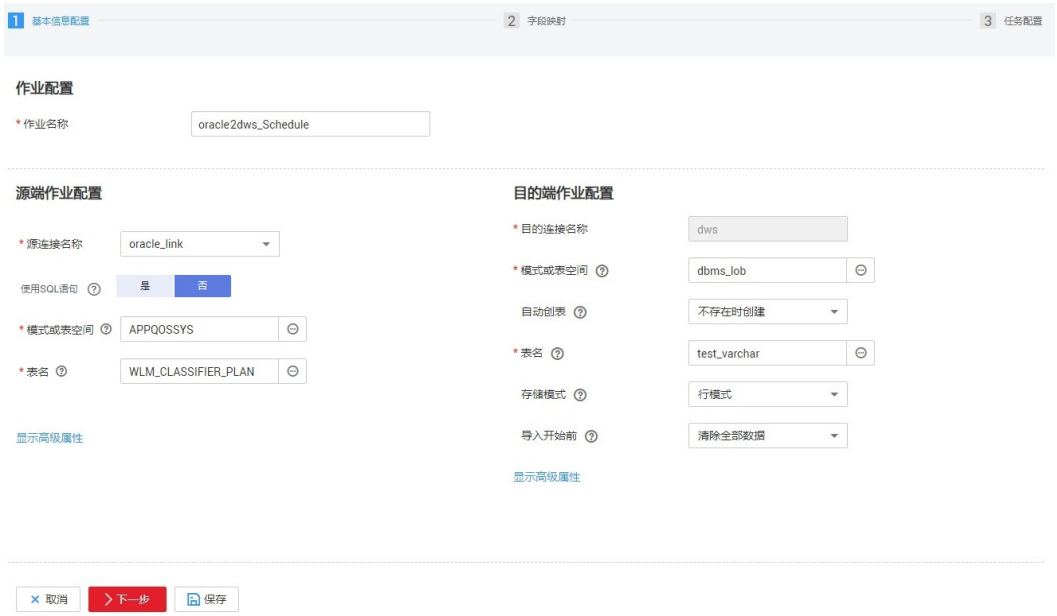
步骤3 单击“保存”完成创建连接。

----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 5-119 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即可。
- 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5-120**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，详细请参见**字段转换**。

图 5-120 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，可打开此配置。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。

5.10.9 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

如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群的VPC必须和云搜索服务集群所在VPC一致，且推荐子网、安全组也与云搜索服务一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

创建云搜索服务连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图5-121**所示。

图 5-121 选择连接器类型

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。

图 5-122 创建云搜索服务连接

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，如图5-123所示。

图 5-123 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，如图5-125所示。

- 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。

  您可以通过如下方式获取访问密钥。

  a. 登录控制台，在用户名下拉列表中选择“我的凭证”。

  b. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图5-124所示。

  图 5-124 单击新增访问密钥

  c. 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。

图 5-125 创建 OBS 连接

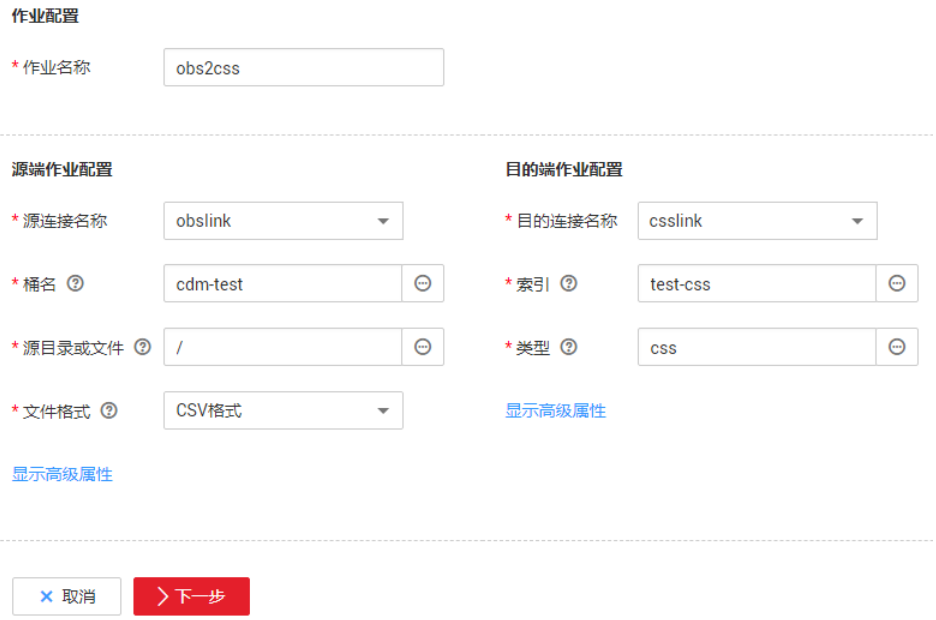
步骤3 单击“保存”回到连接管理界面。

----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 5-126 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即可。
- 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即可。

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5-127**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，详细请参见**字段转换**。

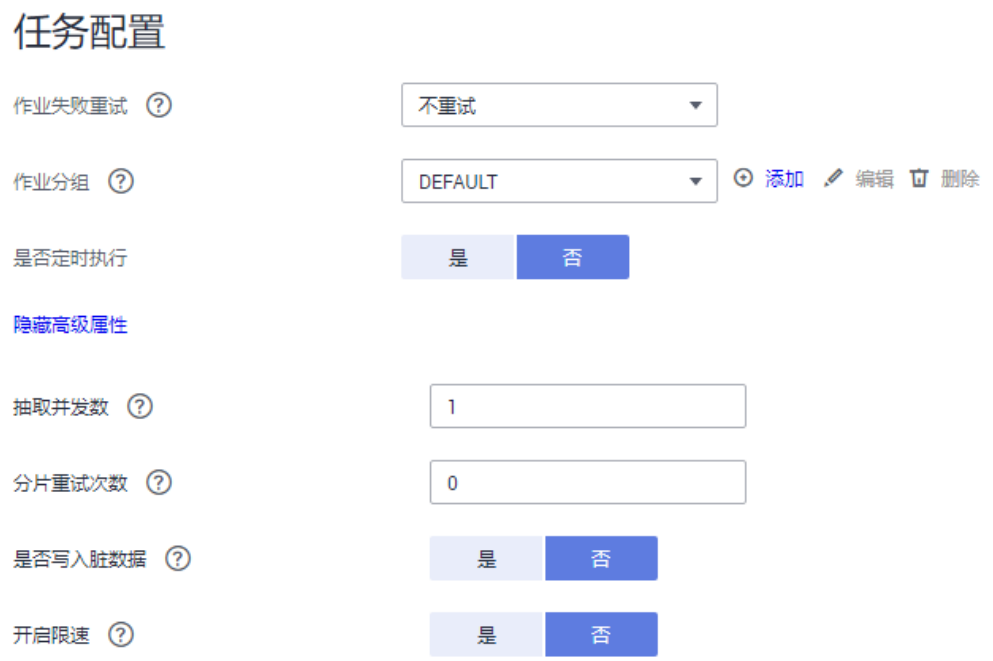
图 5-127 云搜索服务的字段映射

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可开启。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，适当的抽取并发数可以提升迁移效率，配置原则请参见**性能调优**。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，写入脏数据前需要在CDM先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。

图 5-128 任务配置



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

5.10.10 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

如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

该场景下，如果CDM集群只是用于迁移OBS数据到DLI，不需要迁移其他数据源，则CDM集群所在的VPC、子网、安全组选择任一个即可，没有要求，CDM通过内网访问DLI和OBS。主要是选择CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。

创建 DLI 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图5-129**所示。

图 5-129 选择连接器类型

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5-130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 5-130 创建 DLI 连接

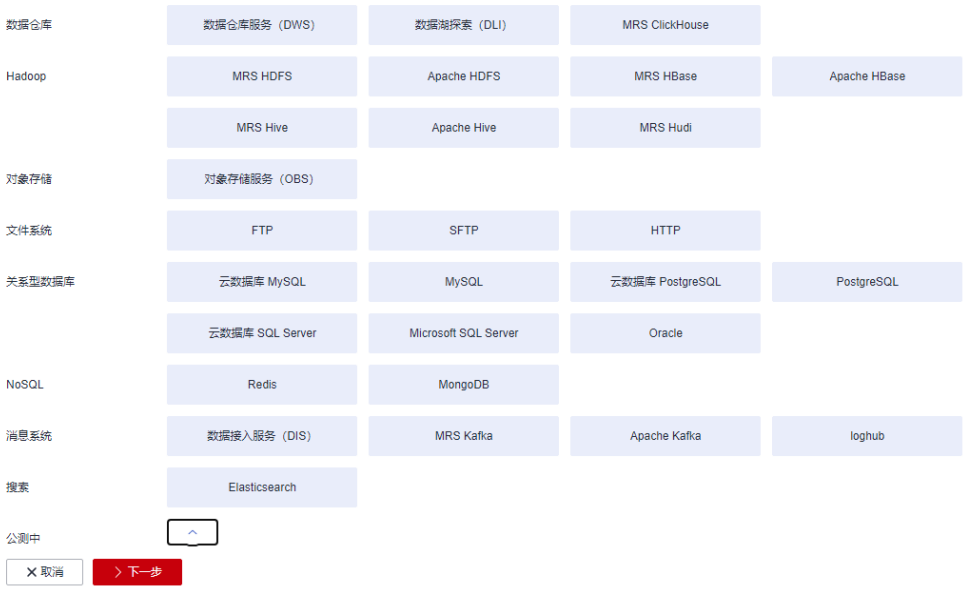
步骤3 单击“保存”回到连接管理界面。

----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，如图5-131所示。

图 5-131 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，如图5-133所示。

- 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。

  您可以通过如下方式获取访问密钥。

  a. 登录控制台，在用户名下拉列表中选择“我的凭证”。

  b. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图5-132所示。

  图 5-132 单击新增访问密钥

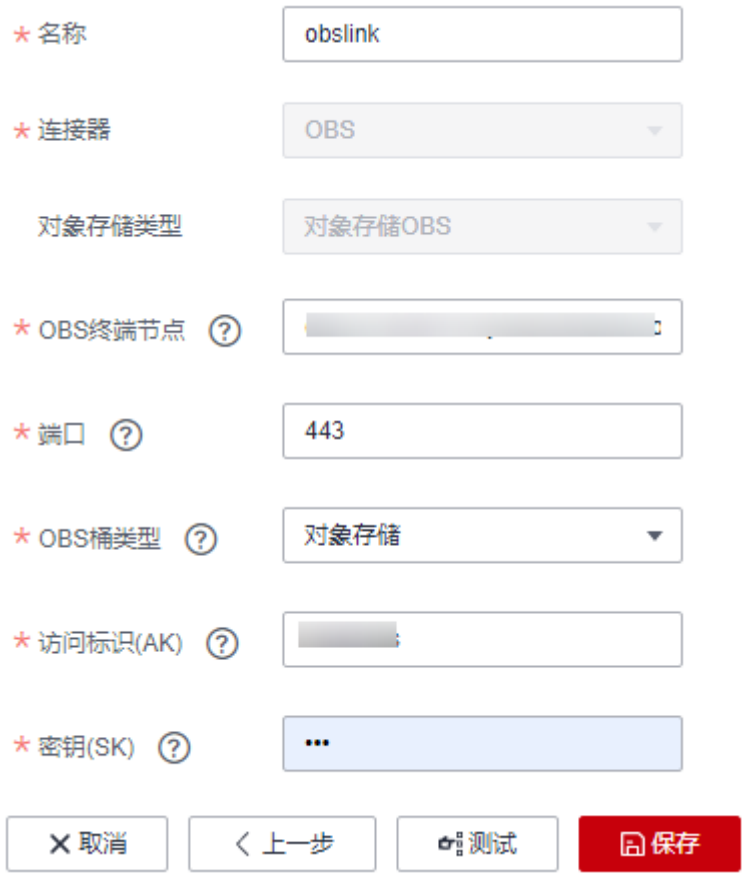
  

  c. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id和Secret Access Key）。

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。

图 5-133 创建 OBS 连接



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5-134所示。

图 5-134 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。
- 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，详细请参见**字段转换**。

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可开启。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，适当的抽取并发数可以提升迁移效率，配置原则请参见**性能调优**。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，写入脏数据前需要在CDM先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。

图 5-135 任务配置

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

5.10.11 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。
- 已经购买了MRS。
- 拥有EIP配额。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

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

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

说明

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

----结束

创建 MRS HDFS 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图5-136**所示。

图 5-136 选择连接器类型

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。

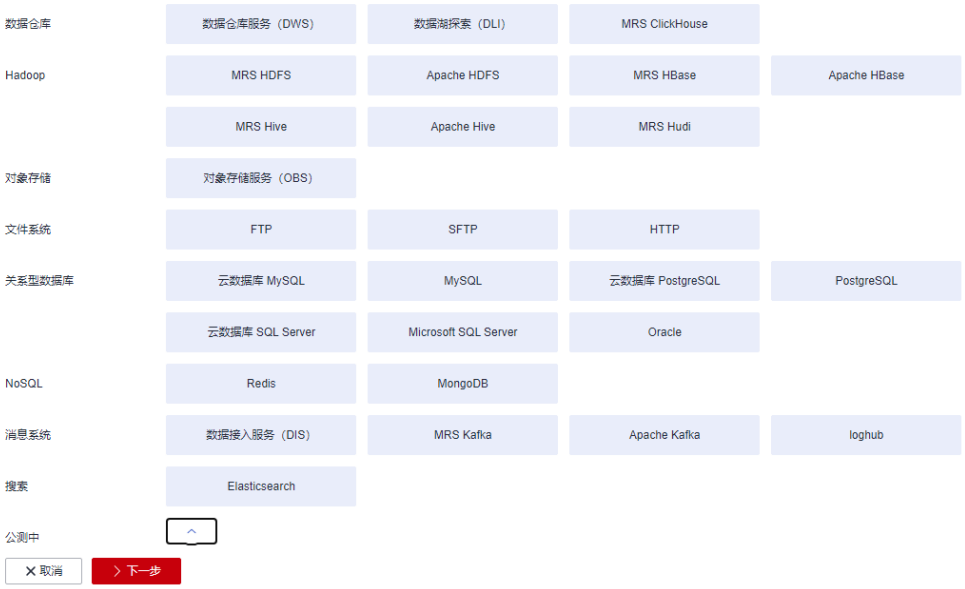
  从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。
- 密码：访问MRS Manager的用户密码。
- 认证类型：访问MRS的认证类型。
- 运行模式：选择HDFS连接的运行模式。

----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，如图5-137所示。

图 5-137 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，如图5-139所示。

- 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。

  您可以通过如下方式获取访问密钥。

  a. 登录控制台，在用户名下拉列表中选择“我的凭证”。

  b. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图5-138所示。

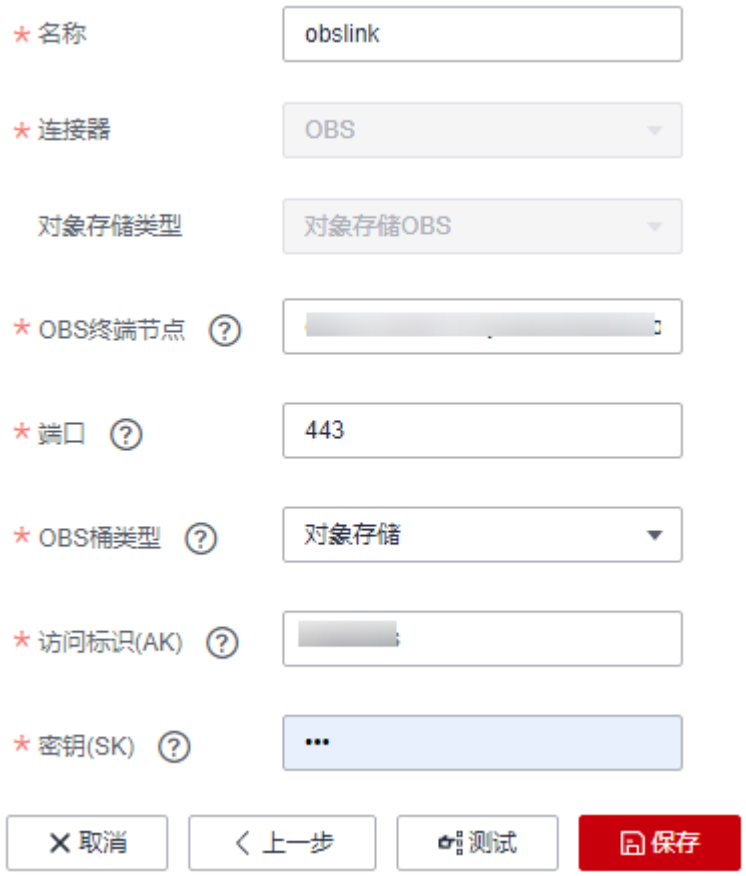
图 5-138 单击新增访问密钥

c. 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。

图 5-139 创建 OBS 连接



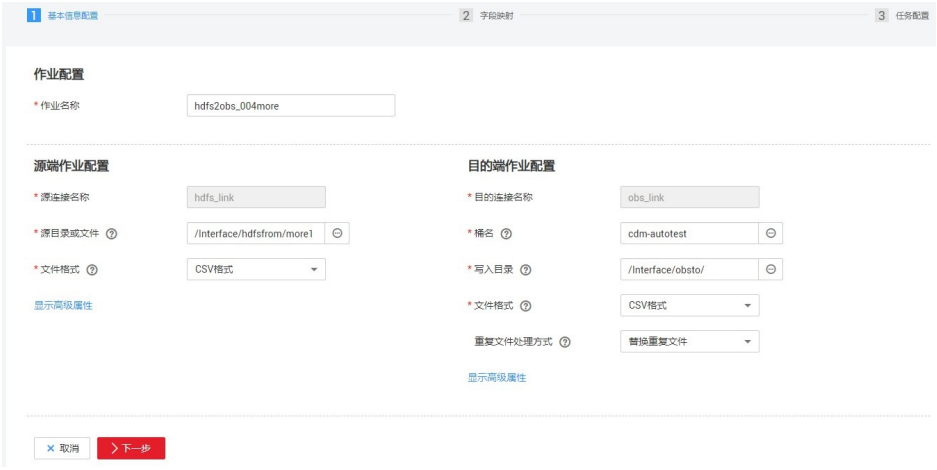
步骤3 单击“保存”回到连接管理界面。

----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 5-140 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。
- 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即可。

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，详细请参见**字段转换**。

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，可打开此配置。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

5.10.12 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华为云的VPN通道或专线。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群的VPC必须和云搜索服务集群所在VPC一致，且推荐子网、安全组也与云搜索服务一致。

- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

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

说明

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

----结束

创建云搜索服务连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如图5-141所示。

图 5-141 选择连接器类型

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。

图 5-142 创建云搜索服务连接

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，如图5-143所示。

图 5-143 选择连接器类型

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 5-144 创建 Elasticsearch 整库迁移作业

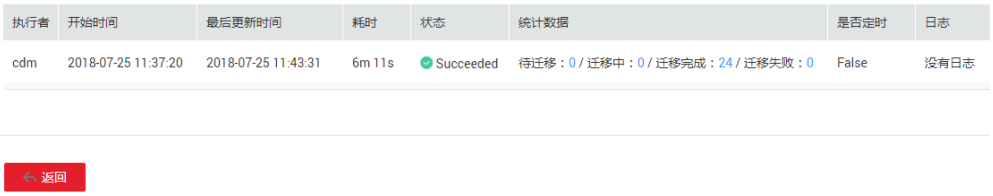
- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建Elasticsearch连接**中的“es_link”。
  - 索引：单击输入框后面的按钮，可选择本地Elasticsearch数据库中的一个索引，也可以手动输入索引名称，名称只能全部小写。需要一次迁移多个索引时，这里可配置为通配符，CDM会迁移所有符合通配符条件的索引。例如这里配置为cdm*时，CDM将迁移所有名称为cdm开头的索引：cdm01、cdmB3、cdm_45……
- 目的端作业配置
  - 目的连接名称：选择**创建云搜索服务连接**中的“csslink”。
  - 索引：待写入数据的索引，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 5-145 作业执行记录



----结束

5.11 常见错误码参考

如果操作请求在执行过程中出现异常导致未被处理，则会返回一条错误信息。错误信息中包括错误码和具体错误描述。表5-137列出了错误信息中的常见错误码。您可以通过表5-137中的处理建议进行下一步操作，处理相应的异常。

错误码说明

表 5-137 错误码说明

错误码	错误信息	处理建议
Cdm.0000	系统错误。	请联系客服或技术支持人员协助解决。
Cdm.0003	Kerberos登录失败。	检查keytab与principal配置文件是否正确。
Cdm.0009	*%s*不是整型数字或超出整型数的取值范围[0～2147483647]。	请根据错误提示将参数修改正确后请重试。
Cdm.0010	整数必须在区间[*%s*]。	请根据返回的详细错误信息，确认参数值是否合法，修改正确后请重试。
Cdm.0011	输入超过取值范围。	请根据返回的详细错误信息，确认参数值是否合法，修改正确后请重试。
Cdm.0012	没有匹配的数据库JDBC驱动。	请联系客服或技术支持人员协助解决。

错误码	错误信息	处理建议
Cdm.0013	Agent连接失败。	可能是由于网络不通、安全组或防火墙规则未放行等原因。若排除上述原因后仍无法解决，请联系客服或技术支持人员协助解决。
Cdm.0014	非法参数。	请确认参数值是否合法，修改正确后请重试。
Cdm.0015	解析文件内容出错。	请确认上传的文件内容或格式是否正确，修改正确后请重试。
Cdm.0016	上传文件不能为空。	请确认上传的文件是否为空，修改正确后请重试。
Cdm.0017	与MRS集群kerberos认证失败。	请确认kerberos认证用户和密码是否很强，修改正确后，请重试。
Cdm.0018	作业和连接内容不合法。	请联系客服或技术支持人员协助解决。
Cdm.0019	IP 和端口无效。	请稍后重试，或联系客服或技术支持人员协助解决。
Cdm.0020	必须包含子字符串： *%s*。	请根据错误提示将参数修改正确后，再重试。
Cdm.0021	不能连接服务器： *%s*。	请联系客服或技术支持人员协助解决。
Cdm.0023	写入数据失败.原因 :*%s*。	请联系客服或技术支持人员协助解决。
Cdm.0024	[*%s*]必须在区间[*%s*]。	请根据错误提示将参数修改正确后，再重试。
Cdm.0025	写入数据的长度超出表字段定义的长度，请参考数据库返回的错误消息: %s	请根据错误提示修改写入数据的长度，再重试。
Cdm.0026	主键重复，请参考数据库返回的错误消息: %s	请根据错误提示检查数据，解决主键冲突。
Cdm.0027	写入字符串的编码可能与表定义的编码不一致，请参考数据库返回的错误消息: %s	请根据错误提示修改字符串编码。
Cdm.0028	用户名或密码错误，请参考数据库返回的错误消息: %s	请修改用户名或者密码，再重试。
Cdm.0029	数据库名称不存在，请参考数据库返回的错误消息: %s	请选择正确的数据库， 再重试。
Cdm.0030	用户名或密码或数据库名称错误，请参考数据库返回的错误消息: %s	请根据错误提示修改为正确的用户名，密码、数据库名称后重试。

错误码	错误信息	处理建议
Cdm.0031	连接超时。	请检查IP、主机名、端口填写是否正确，检查网络安全组和防火墙配置是否正确。
Cdm.0032	用户名或密码错误，请参考服务端返回的错误消息: %s	请根据错误提示修改为正确的用户名和密码后重试。
Cdm.0033	不支持SIMPLE认证类型。	请尝试选择KERBEROS认证类型，再重试。
Cdm.0034	请重启CDM，重新加载MRS或者FusionInsight配置信息。	请重启CDM，重新加载MRS或者FusionInsight配置信息。
Cdm.0035	没有权限写文件，请参考详细消息: %s	请根据错误提示配置权限，再重试。
Cdm.0036	非法Datestamp或Date格式，请参考详细消息: %s	请根据错误提示配置Datestamp或Date格式，再重试。
Cdm.0037	非法参数。%s。	请根据错误提示修改为正确的参数，再重试。
Cdm.0038	连接超时。	请检查VPC和安全组规则。
Cdm.0039	连接名不允许修改。	不可修改连接名。
Cdm.0040	日志因为定期清理被删除。	请联系客服或技术支持人员协助解决。
Cdm.0041	不能更新或者删除已被使用的分组	请勿修改分组。
Cdm.0042	操作分组失败，请参考详细信息: %s	请根据错误提示选择正确的分组，再重试。
Cdm.0043	触发销毁抽取或加载失败. 原因 : %s"	请联系客服或技术支持人员协助解决。
Cdm.0051	无效的提交引擎： *%s*。	请指定正确的作业引擎后再重试。
Cdm.0052	作业*%s*正在运行。	作业正在运行，无法执行当前操作，请等待作业运行结束后再重试。
Cdm.0053	作业*%s*未运行。	请运行作业后再重试。
Cdm.0054	作业*%s*不存在。	请确认作业是否存在。
Cdm.0055	作业类型不支持。	请指定正确的作业类型后再重试。

错误码	错误信息	处理建议
Cdm.0056	不能提交作业。原因：*%s*。	请根据返回的详细错误信息，定位原因，修改正确后请重试。
Cdm.0057	无效的作业执行引擎：*%s*。	请指定正确的作业引擎后再重试。
Cdm.0058	提交和执行引擎组合不合法。	请指定正确的作业引擎后再重试。
Cdm.0059	作业*%s*已被禁用。不能提交作业。	当前作业无法提交，建议重新创建一个作业后再重试。或者，请联系客服或技术支持人员协助解决。
Cdm.0060	作业使用的连接*%s*已被禁用。不能提交作业。	请改为其他连接后，再重新提交作业。
Cdm.0061	连接器*%s*不支持此方向。不能提交作业。	该连接器不能作为作业的源端或目的端，请改为其他连接后，再重新提交作业。
Cdm.0062	二进制文件仅适合SFTP/FTP/HDFS/OBS连接器。	请指定正确的连接器后再重试。
Cdm.0063	创建表格错误。原因：*%s*。	请根据返回的详细错误信息定位原因，修改正确后请重试。
Cdm.0064	数据格式不匹配。	请根据返回的详细错误信息，确认数据格式是否正确，修改正确后请重试。
Cdm.0065	定时器启动失败，原因*%s*。	请联系客服或技术支持人员协助解决。
Cdm.0066	获取样值失败，原因：*%s*。	请联系客服或技术支持人员协助解决。
Cdm.0067	获取Schema失败，原因：*%s*。	请联系客服或技术支持人员协助解决。
Cdm.0068	清空表数据失败, 原因 : %s。	• 请确认当前账户是否有该表的操作权限。 • 请确认表是否被锁定。 • 若以上两种方案均不可行，请联系客服或技术支持人员协助解决。
Cdm.0070	运行任务%s失败, 原因 : 运行任务数目达到上限。	请联系客服或技术支持人员协助解决。
Cdm.0071	获取表数据失败，原因：%s。	请联系客服或技术支持人员协助解决。
Cdm.0074	修复表格失败。 原因 : %s。	请联系客服或技术支持人员协助解决。

错误码	错误信息	处理建议
Cdm.0075	删除表失败，原因：%s。	• 请确认当前账户是否有该表的操作权限。 • 请确认表是否被锁定。 • 若以上两种方案均不可行，请联系客服或技术支持人员协助解决。
Cdm.0080	无效的用户名 。	请根据错误提示修改为正确的用户名，再重试。
Cdm.0081	无效的证书。	请联系客服或技术支持人员协助解决。
Cdm.0082	证书不可读。	请联系客服或技术支持人员协助解决。
Cdm.0083	同一个进程不能配置多个证书，需要重启以使用新的证书。	请根据错误提示修改证书，再重启重试。
Cdm.0085	超过最大值。	请联系客服或技术支持人员协助解决。
Cdm.0088	XX配置项有误。	请根据错误提示修改配置项，再重试。
Cdm.0089	配置项XX不存在。	• 请根据错误提示修改配置项，再重试。 • 低版本CDM集群切换至高版本CDM集群时，创建数据连接或保存作业时会偶现配置项不存在情况，请手动清理缓存，再重试。
Cdm.0091	打补丁失败。	请联系客服或技术支持人员协助解决。
Cdm.0092	备份文件不存在。	请联系客服或技术支持人员协助解决。
Cdm.0093	无法加载krb5.conf。	请联系客服或技术支持人员协助解决。
Cdm.0094	名称为XX连接不存在。	请根据错误提示，确认XX连接是否存在，再重试。
Cdm.0095	名称为XX作业不存在。	请根据错误提示，确认XX作业是否存在，再重试。
Cdm.0100	作业[*%s*]不存在。	请指定正确的作业后再重试。
Cdm.0101	连接[*%s*]不存在。	请指定正确的连接后再重试。

错误码	错误信息	处理建议
Cdm.0102	连接器[*%s*]不存在。	请指定正确的连接器后再重试。
Cdm.0104	作业名已存在。	作业名已存在，请重新命名后，再重试。
Cdm.0105	表达式为空。	• 请参考帮助文档确认表达式是否有效。 • 若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0106	XX表达式运算失败。	• 请参考帮助文档确认表达式是否有效。 • 若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0107	任务执行中，请稍后再修改作业配置。	待任务执行完成后，再修改作业配置。
Cdm.0108	查询表记录失败。	• 自定义SQL，请首先确认正确性。 • 请确认查询未超时（小于60s）. • 若以上错误均无法规避，请联系客服或技术支持人员协助解决。
Cdm.0109	作业或连接名长度不能超过%s。	请根据错误提示修改作业或连接名称。
Cdm.0110	命名错误,只能以字符或数字开头，并且名字只能包含字符、数字、下划线、中划线、点符号。	请根据错误提示修改命名。
Cdm.0201	获取实例失败。	请联系客服或技术支持人员协助解决。
Cdm.0202	作业状态未知。	请稍后重试，或请联系客服或技术支持人员协助解决。
Cdm.0204	没有已创建的MRS连接。	当前没有MRS连接，您需要先前往集群的“连接管理”页面创建一个MRS连接，然后再重新执行当前的操作。
Cdm.0230	不能加载该类：*%s*。	请联系客服或技术支持人员协助解决。
Cdm.0231	不能初始化该类：*%s*。	请联系客服或技术支持人员协助解决。
Cdm.0232	数据写入失败。原因：*%s*。	请联系客服或技术支持人员协助解决。
Cdm.0233	提取数据过程异常。原因：*%s*。	请联系客服或技术支持人员协助解决。

错误码	错误信息	处理建议
Cdm.0234	载入数据过程异常。原因：*%s*。	请联系客服或技术支持人员协助解决。
Cdm.0235	数据已全部消费完毕。原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0236	从分区程序中检索到无效分区数。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0237	找不到连接器Jar包。	请联系客服或技术支持人员协助解决。
Cdm.0238	*%s*不能为空。	请根据错误提示将参数修改正确后再重试。
Cdm.0239	获取HDFS文件系统失败。原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0240	获取文件*%s*状态失败。	请联系客服或技术支持人员协助解决。
Cdm.0241	获取文件*%s*类型失败。	请联系客服或技术支持人员协助解决。
Cdm.0242	文件检查异常： *%s*。	请联系客服或技术支持人员协助解决。
Cdm.0243	重命名*%s*为*%s*失败。	可能是名称已存在，请重新命名后再重试。
Cdm.0244	创建文件*%s*失败。	请确认是否具有创建权限，或稍后重试。若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0245	删除文件*%s*失败。	请确认是否具有删除权限，或稍后重试。若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0246	创建目录*%s*失败。	请确认是否具有创建权限，或稍后重试。若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0247	操作HBase失败。原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0248	清空*%s*数据失败。原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0249	文件名*%s*无效。	请将文件名修改正确后，再重试。

错误码	错误信息	处理建议
Cdm.0250	不能操作该路径：*%s*。	请确认是否具有该路径的操作权限，或稍后重试。若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0251	向HBase加载数据失败。原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0307	无法释放连接，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0315	连接名*%s*已存在。	请指定其他连接名后再重试。
Cdm.0316	无法更新不存在的连接。	请指定正确的连接后再重试。
Cdm.0317	连接*%s*无效。	请指定正确的连接后再重试。
Cdm.0318	作业已存在，无法重复创建。	请指定其他作业名再重试。
Cdm.0319	无法更新不存在的作业。	请确认待更新的作业是否存在，作业名修改正确后再重试。
Cdm.0320	作业*%s*无效。	请联系客服或技术支持人员协助解决。
Cdm.0321	连接*%s*已被使用。	连接已被使用，无法执行当前的操作，请将连接释放后再重试。
Cdm.0322	作业*%s*已被使用。	请联系客服或技术支持人员协助解决。
Cdm.0323	该提交已存在，无法重复创建。	您已提交过相同操作的请求，请稍后再重试。
Cdm.0327	无效的连接或作业：*%s*。	请指定正确的连接或作业再重试。
Cdm.0411	连接到文件服务器时出错。	请联系客服或技术支持人员协助解决。
Cdm.0412	与文件服务器断开连接时出错。	请联系客服或技术支持人员协助解决。
Cdm.0413	向文件服务器传输数据时出错。	请联系客服或技术支持人员协助解决。
Cdm.0415	从文件服务器下载文件出错。	请联系客服或技术支持人员协助解决。
Cdm.0416	抽取数据时出错。	请联系客服或技术支持人员协助解决。

错误码	错误信息	处理建议
Cdm.0420	源文件或源目录不存在。	请确认源文件或源目录是否存在，修改正确后再重试。
Cdm.0423	目的路径存在重复文件。	请在目的路径中删除重复文件后再重试。
Cdm.0500	源目录或文件[*%s*]不存在。	请指定正确的源文件或目录后再重试。
Cdm.0501	无效的URI[*%s*]。	请指定正确的URI后，再重试。
Cdm.0518	连接HDFS失败。原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0523	用户权限不足导致连接超时。	新建一个业务用户，给对应的权限后，再重试。
Cdm.0600	无法连接FTP服务器。原因: %s。	可能是由于网络不通、安全组或防火墙规则未放行、FTP主机名无法解析、FTP用户名密码错误等原因。若排除上述原因后仍无法解决，请联系客服或技术支持人员协助解决。
Cdm.0700	无法连接SFTP服务器。原因: %s。	可能是由于网络不通、安全组或防火墙规则未放行、SFTP主机名无法解析、SFTP用户名密码错误等原因。若排除上述原因后仍无法解决，请联系客服或技术支持人员协助解决。
Cdm.0800	无法连接OBS服务器。原因: %s。	可能是由于OBS终端节点与当前区域不一致、AK/SK错误、AK/SK不是当前用户的AK/SK、安全组或防火墙规则未放行等原因。若排除上述原因后仍无法解决，请联系客服或技术支持人员协助解决。
Cdm.0801	OBS桶[*%s*]不存在。	指定的OBS桶可能不存在或不在当前区域，请指定正确的OBS桶后再重试。
Cdm.0900	表[*%s*]不存在。	请指定正确的表名后再重试。
Cdm.0901	无法连接数据库服务器。原因： *%s*。	请联系客服或技术支持人员协助解决。
Cdm.0902	SQL语句无法执行。原因*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。

错误码	错误信息	处理建议
Cdm.0903	元数据获取失败。原因：*%s*。	请确认在集群的“连接管理”页面创建连接时引用符号是否正确或查看数据库表是否存在。若仍无法解决，请联系客服或技术支持人员协助解决。
Cdm.0904	从结果中检索数据时发生错误。原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0905	未设置分区列。	请指定分区列后再重试。
Cdm.0906	分区列没有找到边界。	请联系客服或技术支持人员协助解决。
Cdm.0911	表名或SQL需要指定。	请指定表名或SQL后再重试。
Cdm.0912	表名和SQL不可以同时指定。	请确认表名和SQL是否同时指定，仅指定其中一项后，再重试。
Cdm.0913	Schema和SQL不可以同时指定。	请确认Schema和SQL是否同时指定，仅指定其中一项后，再重试。
Cdm.0914	基于查询的导入方式时必须提供分区字段。	请指定分区字段后，再重试。
Cdm.0915	基于SQL的导入方式和ColumnList不能同时使用。	请确认两种是否同时使用，仅使用其中一项后，再重试。
Cdm.0916	增量读取情况下必须指定上次的值。	请指定上次的值后再重试。
Cdm.0917	缺少字段检查将无法获得上次的值。	请联系客服或技术支持人员协助解决。
Cdm.0918	没有指定中转表的情况下不可以指定“shouldClearStageTable”。	请指定中转表后再重试。
Cdm.0921	不支持类型*%s*。	请指定正确的类型后再重试。
Cdm.0925	分区字段含有不支持的值。	请确认分区字段是否含有不支持的值，修改正确后再重试。
Cdm.0926	取不到Schema。原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.0927	中转表不为空。	请指定一个空的中转表后再重试。

错误码	错误信息	处理建议
Cdm.0928	中转表到目的表进行数据迁移时发生错误。	请联系客服或技术支持人员协助解决。
Cdm.0931	Schema字段大小[*%s*]与结果集的字段大小[*%s*]不匹配。	请将Schema字段大小和结果集中的字段大小改为一致后再重试。
Cdm.0932	找不到字段最大值。	请联系客服或技术支持人员协助解决。
Cdm.0934	不同Schema/Catalog下有重名表。	请联系客服或技术支持人员协助解决。
Cdm.0935	缺少主键。请指定分区字段。	请指定主键字段后再重试。
Cdm.0936	错误脏数据条数达到上限。	您可以编辑作业，在作业的任务配置中将错误脏数据条数增大。
Cdm.0940	表名准确匹配失败。	匹配不到表名，请指定正确的表名后再重试。
Cdm.0941	无法连接服务器。原因:[*%s*]	请检查IP、主机名、端口填写是否正确，检查网络安全组和防火墙配置是否正确，参考数据库返回消息进行定位。若仍无法解决，请联系客服或技术支持人员协助解决。
Cdm.0950	当前认证信息无法连接到数据库。	认证信息错误，请修改正确后再重试。
Cdm.0960	必须指定服务器列表。	请指定服务器列表后再重试。
Cdm.0961	服务器列表格式非法。	请修改正确的格式后再重试。
Cdm.0962	必须指定主机IP。	未指定主机IP，请指定主机IP后，再重试。
Cdm.0963	必须指定主机端口。	未指定主机端口，请指定主机端口后，再重试。
Cdm.0964	必须指定数据库。	未指定数据库，请指定数据库后，再重试。
Cdm.1000	Hive表[*%s*]不存在。	请输入正确的Hive表名后，再重试。

错误码	错误信息	处理建议
Cdm.1010	无效的URI *%s*。URI必须为null或有效的URI。	请输入正确的URI后，再重试。下面是一些URI示例： ● hdfs://example.com:8020/ ● hdfs://example.com/ ● file:/// ● file:///tmp ● file://localhost/tmp
Cdm.1011	连接Hive失败，原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1012	初始化hive客户端失败，原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1100	表[*%s*]不存在。	请确认表是否存在，输入正确的表名后再重试。
Cdm.1101	获取连接失败，原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1102	创表失败，原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1103	未设置Rowkey。	请设置Rowkey后再重试。
Cdm.1104	打开表格失败。原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1105	作业初始化失败。原因*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1111	表名不能为空。	请输入正确的表名后，再重试。
Cdm.1112	导入方式不能为空。	请设置导入方式后再重试。
Cdm.1113	导入前是否清空数据未设置。	请设置“导入前是否清空数据”参数后再重试。
Cdm.1114	Rowkey为空，请在字段映射步骤重新设置。	请按照错误提示进行处理。
Cdm.1115	Columns为空，请在字段映射步骤重新设置。	请按照错误提示进行处理。

错误码	错误信息	处理建议
Cdm.1116	列名重复，请在字段映射步骤重新设置。	请按照错误提示进行处理。
Cdm.1117	判断表格是否存在失败，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1118	表*%s*不包含列族*%s*。	请指定列族后再重试。
Cdm.1119	列族数*%s*和列数*%s*不等。	请将列族数和列数改为一致后再重试。
Cdm.1120	表中有数据，请清空表数据或重新设置导入前是否清空表数据配置项。	请按照错误提示进行处理。
Cdm.1121	关闭连接已失败。原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1201	不能连接到Redis服务器，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1202	不能用单机模式去连接Redis集群。	请改为其他模式连接Redis集群。
Cdm.1203	从Redis服务器抽取数据失败，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1205	Redis值前缀不能为空白符。	请去除Redis前缀前的空白符，然后再重试。
Cdm.1206	Redis值存储类型必须指定为“string”或“hash”。	请按照错误提示进行处理。
Cdm.1207	当值存储类型为“string”时，必须指定值分隔符。	请指定分隔符后再重试。
Cdm.1208	Redis存储字段列表必须指定。	请指定Redis存储字段列表后再重试。
Cdm.1209	Redis键分隔符不能为空白符。	请输入正确的分隔符后，再重试。
Cdm.1210	必须指定Redis主键字段列表。	请指定Redis主键字段列表后再重试。
Cdm.1211	Redis主键字段列表必须在字段列表中存在。	请指定Redis主键字段列表后再重试。
Cdm.1212	Redis数据库类型必须指定为“Original”或“DCS”。	请按照错误提示进行处理。

错误码	错误信息	处理建议
Cdm.1213	必须指定Redis服务器列表。	请指定Redis服务器列表后再重试。
Cdm.1301	不能连接到MongoDB服务器，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1302	从MongoDB服务器抽取数据失败，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1304	必须指定MongoDB服务器的集合。	未指定MongoDB服务器的集合，请指定后，再重试。
Cdm.1305	必须指定MongoDB服务列表。	未指定MongoDB服务列表，请指定后，再重试。
Cdm.1306	必须指定MongoDB服务的数据库名称。	未指定MongoDB服务的数据库名称，请指定数据库后，再重试。
Cdm.1307	必须指定MongoDB服务的字段列表。	未指定MongoDB服务的字段列表，请指定字段列表后，再重试。
Cdm.1400	无法连接NAS服务器。	请联系客服或技术支持人员协助解决。
Cdm.1401	无NAS服务器权限。	请申请NAS服务器权限后再重试。
Cdm.1501	不能连接到Elasticsearch服务器，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1502	向Elasticsearch服务器写入数据失败，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1503	关闭Elasticsearch连接失败，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1504	获取Elasticsearch索引错误，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1505	获取Elasticsearch类型错误，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1506	获取Elasticsearch文档字段错误，原因：*%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。

错误码	错误信息	处理建议
Cdm.1507	获取Elasticsearch采样数据错误，原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1508	必须指定Elasticsearch服务器主机名或IP地址。	未指定Elasticsearch服务器主机名或IP地址，请指定后，再重试。
Cdm.1509	必须指定Elasticsearch服务器端口。	未指定Elasticsearch服务器端口，请指定端口后，再重试。
Cdm.1510	必须指定Elasticsearch索引。	当前未指定Elasticsearch索引，请指定后再重试。
Cdm.1511	必须指定Elasticsearch类型。	当前未指定Elasticsearch类型，请指定后再重试。
Cdm.1512	必须指定Elasticsearch文档字段列表。	当前未指定Elasticsearch文档字段列表，请指定后再重试。
Cdm.1513	字段列表中必须包含字段类型定义。	请确认字段列表中是否包含字段类型定义，修改正确后再重试。
Cdm.1514	字段列表中必须包含主键字段。	当前未设置主键字段，请设置主键字段后再重试。
Cdm.1515	解析JSON字符串时错误，原因： *%s*。	请根据返回的详细错误信息，定位原因，修改正确后请重试。如仍无法解决，请联系客服或技术支持人员协助解决。
Cdm.1516	非法列名*%s*。	请确认列名是否合法，输入正确的列名后再重试。
Cdm.1517	获取文档数量产生错误。	请联系客服或技术支持人员协助解决。
Cdm.1518	分区失败。	请联系客服或技术支持人员协助解决。
Cdm.1519	抽取数据错误。	请联系客服或技术支持人员协助解决。
Cdm.1520	获取类型失败。原因： *%s*。	请根据错误提示进行定位，若无法解决，请联系客服或技术支持人员协助解决。
Cdm.1601	连接服务器失败。	请联系客服或技术支持人员协助解决。
Cdm.1603	获取topic *%s*的样值失败。	请联系客服或技术支持人员协助解决。
Cdm.1604	topic *%s*没有数据。	该topic中无数据，请排查无数据的原因。或者，请改为其他topic后再重试。

错误码	错误信息	处理建议
Cdm.1605	无效的brokerList。	请指定正确的brokerList后再重试。

6 数据集成（离线作业）

6.1 离线作业概述

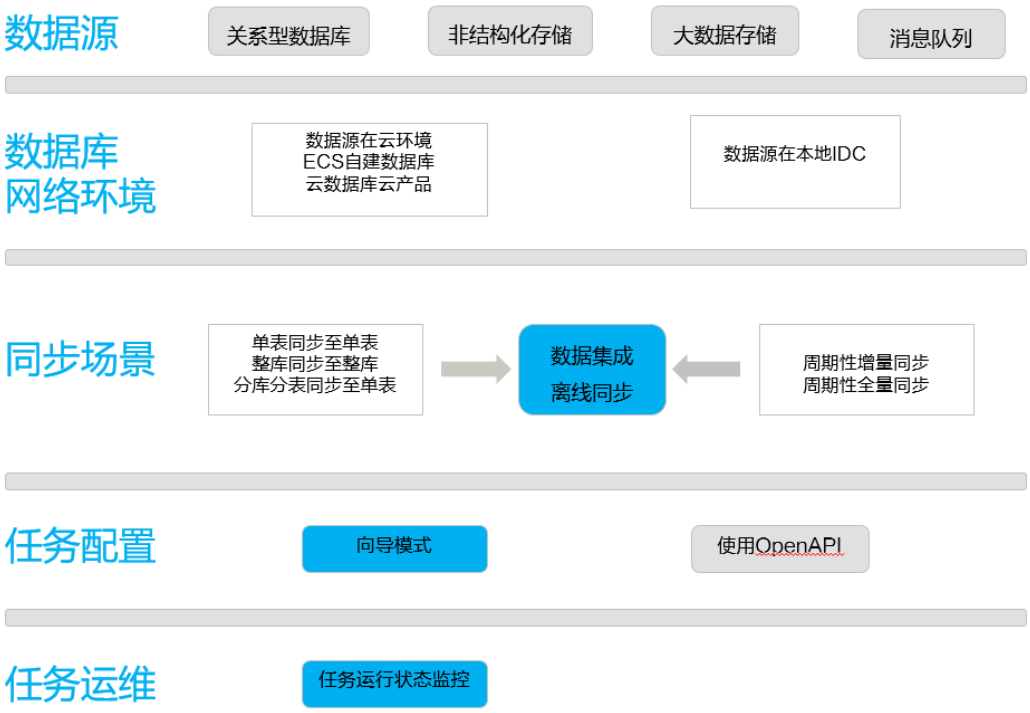
离线处理集成作业作为数据开发的一个作业类型，支持跨集群下发数据迁移作业，实现常用的批作业迁移能力。

相比于传统的依靠CDM集群进行生命周期管理CDM迁移作业，离线处理集成作业依靠数据开发组件的生命周期管理，由数据开发进行集成作业的统一调度和CDM集群资源的统一支配，作业运行可靠性更高、使用体验更佳。

说明

离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。

图 6-1 离线处理集成作业迁移原理



6.2 支持的数据源

数据集成离线同步支持单表同步至目标单表、分库分表同步至目标单表及整库同步至目标单表三种同步方式，不同的同步方式支持的数据源有所不同：

- 单表同步：适用于数据入湖和数据上云场景下，表或文件级别的数据同步，支持的数据源请参见**表/文件同步支持的数据源类型**。
- 分库分表同步：适用于数据入湖和数据上云场景下，多库多表同步场景，支持的数据源请参见**分库分表同步支持的数据源类型**。
- 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同步场景，支持的数据源请参见**整库同步支持的数据源类型**。

说明

因各版本集群支持的数据源有所差异，其他版本支持的数据源仅做参考。
不同CDM集群支持的数据源程度不一样，以实际为准。

表/文件同步支持的数据源类型

表/文件同步可以实现表或文件级别的数据同步。

支持单表同步的数据源如**表1 离线作业不同数据源读写能力说明**所示：

表 6-1 离线作业不同数据源读写能力说明

数据源分类	数据源	单表读	单表写	说明
数据仓库	DWS、DLI	支持	支持	不支持DWS物理机纳管模式。
Hadoop	MRS Hive、MRS Hudi、Doris、MRS ClickHouse、MRS HBase	支持	支持	• MRS ClickHouse建议使用的版本：21.3.4.X。 • 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、MRS Hudi暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X
对象存储	OBS	支持	支持	-
文件系统	FTP、SFTP	支持	不支持	-
关系型数据库	RDS（MySQL）、RDS（PostgreSQL）、RDS（SQL Server）、Oracle、RDS（SAP HANA）、GBASE8A **说明** 创建数据连接时也支持用户使用自建的数据库，如MySQL、PostgreSQL、SQL Server、达梦数据库DM、SAP HANA，在选择界面对应的RDS（MySQL）、RDS（PostgreSQL）、RDS（SQL Server）、RDS（达梦数据库DM）、RDS（SAP HANA）即可。	支持	支持	• SAP HANA仅支持2.00.050.00.1592305219版本。 • Apache HDFS建议使用的版本： – 2.8.X – 3.1.X

数据源分类	数据源	单表读	单表写	说明
	RDS（达梦数据库DM）	不支持	不支持	-
非关系型数据库	MongoDB、Redis	支持	支持	MongoDB建议使用的版本：4.2。
消息系统	Apache HDFS、DMS Kafka **说明** Apache HDFS目前仅支持作为源端数据源。	支持	支持	• Apache HDFS建议使用的版本： – 2.8.X – 3.1.X
	LTS	支持	不支持	-
	Apache RocketMq	不支持	支持	-
搜索	Elasticsearch	支持	支持	-
其他	Rest Client	支持	不支持	-
	OpenGauss（GaussDB）	支持	支持	-

分库分表同步支持的数据源类型

分库分表同步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，适用于多库多表同步场景。

支持分库分表同步的数据源如下所示：

源端为RDS（MySQL）时支持分库分表同步。

整库同步支持的数据源类型

整库同步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，适用于数据库离线同步场景，不适用于在线实时同步。

支持整库同步的数据源（已支持的数据源即可作为源端，又可作为目的端组成不同链路）如下所示：

- 读取能力：DWS、RDS（MySQL）、RDS（PostgreSQL）
- 写入能力：DWS、DLI

6.3 新建离线处理集成作业

约束限制

- 离线处理集成作业不支持在企业模式下运行。
- 离线处理集成作业功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 新建集成作业的方式有如下两种：

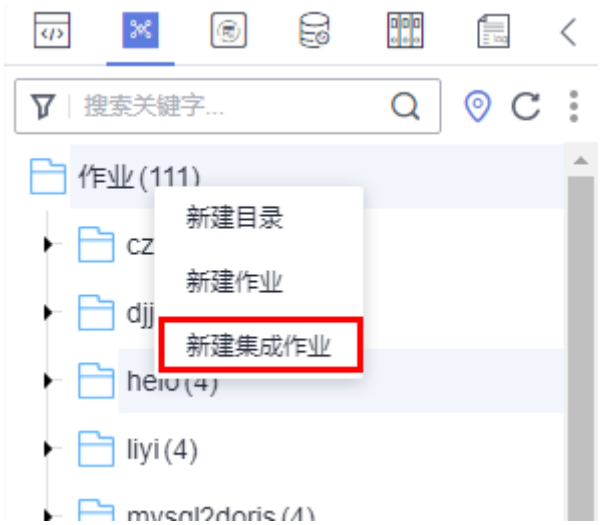
   方式一：在“作业开发”界面中，单击“新建集成作业”。

   图 6-2 新建集成作业（方式一）

   方式二：在作业目录中，右键单击目录名称，选择“新建集成作业”。

   图 6-3 新建集成作业（方式二）

5. 在弹出的“新建集成作业”页面，配置如**表6-2**所示的参数。

图 6-4 配置集成作业参数

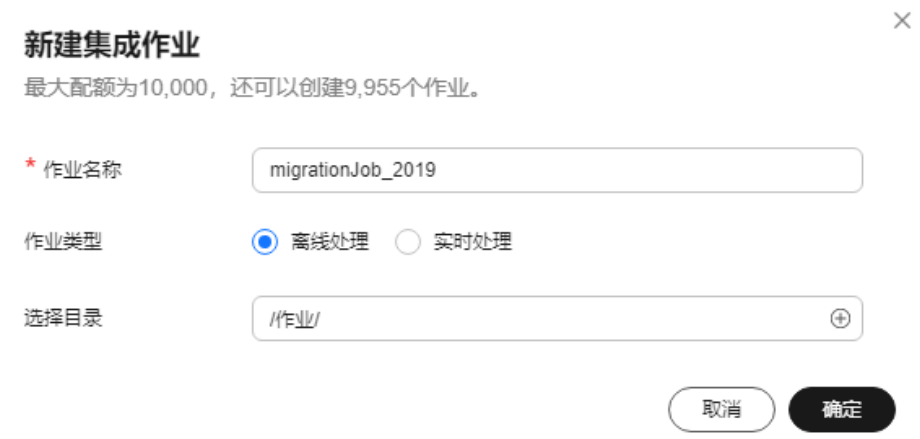


表 6-2 作业参数

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，须选择离线处理。 • 离线处理：对已收集的大量数据进行批量处理和分析，这些任务通常是在计算资源和存储资源方面经过优化，以确保高效的数据处理和分析。这些任务通常是定时（例如每天、每周）执行，主要处理大量历史数据，用于批量分析和数据仓库。 • 实时处理：对源源不断产生的新数据进行实时处理和分析，以满足业务对数据的即时性需求。这种处理方式要求数据在产生后能够立即被处理，并给出相应的结果或触发相应的操作。
选择目录	选择作业所属的目录，默认为根目录。

6. 单击“确定”，创建作业。

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6-3所示的参数。

表 6-3 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。

参数	说明
执行用户	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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。 **说明** 配置执行用户调度功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
作业委托	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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。
实例超时是否忽略等待时间	配置实例超时是否忽略等待时间。 如果勾选上，表示实例运行时等待时间不会被计入超时时间，可前往**默认项设置**修改此策略。 如果未选上，表示实例运行时等待时间会被计入超时时间。
自定义字段	配置自定义字段的参数名称和参数值。
作业标签	配置作业的标签，用以分类管理作业。 单击“新增”，可给作业重新添加一个标签。也可选择**管理作业标签**中已配置的标签。
节点状态轮询时间（秒）	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	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，该节点将会再次重试。
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。推荐值：不重试。 建议仅对文件类作业或启用了导入阶段表的数据库作业配置自动重试，避免自动重试重复写入数据导致数据不一致。 **说明** 如果通过DataArts Studio数据开发使用参数传递并调度CDM迁移作业时，不能在CDM迁移作业中配置“作业失败重试”参数，如有需要请在数据开发中的CDM节点配置“失败重试”参数。
当前节点失败后，后续节点处理策略	当前节点执行失败后，后续节点的处理策略： • 终止当前作业执行计划：终止当前作业运行，当前作业实例状态显示为“失败”。如果是周期调度作业，后续周期调度会正常运行。 • 忽略失败，作业结果设为成功：忽略当前节点失败，当前作业实例状态显示为“运行成功”。如果是周期调度作业，后续周期调度会正常运行。

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：

单击编辑器右侧的“参数”，展开配置页面，配置如表6-4所示的参数。

表 6-4 作业参数配置

功能	说明
变量	
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：${参数名称}
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见表达式概述。
修改	在参数名和参数值的文本框中直接修改。
掩码显示	在参数值为密钥等情况下，从安全角度，请单击  将参数值掩码显示。
删除	在参数值文本框后方，单击，删除作业参数。
常量	
新增	单击“新增”，在文本框中填写作业常量的名称和参数值。 ● 参数名称 名称只能包含字符：英文字母、数字、中划线和下划线。 ● 参数值 – 字符串类的参数直接填写字符串，例如：str1 – 数值类的参数直接填写数值或运算表达式。 参数配置完成后，在作业中的引用格式为：${参数名称}
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见表达式概述。
修改	在参数名和参数值的文本框中直接修改，修改完成后，请保存。
删除	在参数值文本框后方，单击，删除作业常量。

6.4 配置离线处理集成作业

数据集成支持创建离线作业，通过在界面勾选源端数据和目的端数据，并结合为其配置的参数，实现将源端单表、分库分表、整库的全量或增量数据周期性同步至目标数据表。

本文为您介绍离线同步任务的常规配置，各数据源配置存在一定差异，请以**配置作业源端参数**及**配置作业目的端参数**为准。

约束限制

需要源端和目的端字段类型及精度设置一致，否则可能导致作业运行失败。

说明

同步任务源端和目标端字段类型需要注意精度，如果目标端字段类型最大值小于源端最大值（或最小值大于源端最小值，或精度低于源端精度），可能会导致写入失败或精度被截断的风险。

前提条件

- 已完成数据连接的创建，且创建的连接必须已勾选数据集成选项，详情请参见**创建DataArts Studio数据连接** 。
- 存在正在运行的CDM集群，详情请参见**创建CDM集群**。

  说明

  DataArts Studio实例中已经包含一个CDM集群（试用版除外），如果该集群已经满足需求，您无需再购买数据集成增量包，可以跳过这部分内容。如果您需要再创建新的CDM集群，请参考**购买批量数据迁移增量包**，完成购买数据集成增量包的操作。
- CDM集群与待同步数据源可以正常通信。

  说明

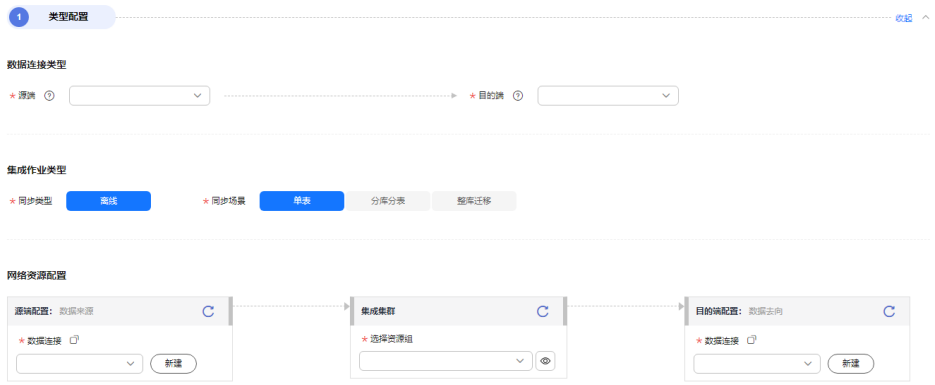
  - 当CDM集群与其他云服务所在的区域、VPC、子网、安全组一致时，可保证CDM集群与其他云服务内网互通，无需专门打通网络。
  - 当CDM集群与其他云服务所在的区域和VPC一致、但子网或安全组不一致时，需配置路由规则及安全组规则以打通网络。配置路由规则请参见**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**章节。
  - 当CDM集群与其他云服务所在的区域一致、但VPC不一致时，可以通过对等连接打通网络。配置对等连接请参见**如何配置对等连接**章节。

    注：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。
  - 当CDM集群与其他云服务所在的区域不一致时，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。
  - 另外，如果创建了企业项目，则企业项目也会影响CDM集群与其他云服务的网络互通，只有企业项目一致的云服务才能打通网络。

操作步骤

1. 参见**新建离线处理集成作业**创建一个离线处理集成作业。
2. 类型配置。

图 6-5 类型配置



a. 配置数据连接类型，包含配置源端数据类型和目的端数据类型，支持的数据类型请参见**支持的数据源**。

b. 选择集成作业类型。

   i. 同步类型：默认为离线，不可更改。

   ii. 同步场景：支持单表、分库分表和整库三种同步方式，具体支持的数据源请参见**支持的数据源**。

c. 设置网络资源配置。

   i. 选择已创建的源端数据连接，且创建的连接必须已勾选数据集成选项。连接不存在时可参见**创建DataArts Studio数据连接**创建所需连接。

   需要测试数据源端和资源组之间网络是否可用，不可用时根据界面提示修改。

   ii. 选择资源组，集群创建可参见**创建CDM集群**。

   选多个集群时系统会随机下发任务，故需要多个集群时版本规格建议选择集群版本一致的集群，否则可能因为集群版本不一致导致作业失败。

   iii. 选择已创建的目的端数据连接，且创建的连接必须已勾选数据集成选项。连接不存在时可参见**创建DataArts Studio数据连接**。

   需要测试数据连接是否可用，不可用时根据界面提示修改。

3. 配置源端数据参数。

   各数据源及各同步场景配置存在一定差异，选择源端配置后，请参见**配置作业源端参数**配置作业参数。

表 6-5 源端需要配置的作业参数

同步场景	源端需要配置参数	字段映射
单表	• 基本参数 • 高级属性	支持
分库分表	• 选择库表方式：精准匹配或正则匹配 • 高级属性	支持

同步场景	源端需要配置参数	字段映射
整库迁移	• 添加源数据，选择需要迁移的库表 • 高级属性	不支持

4. 配置目的端数据参数。

   各数据源及各同步场景配置存在一定差异，选择目的端配置后，请参见**配置作业目的端参数**配置作业参数。

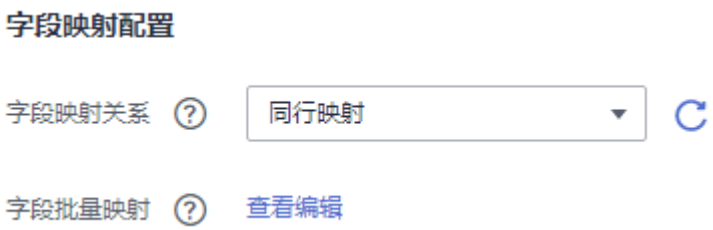
   表 6-6 目的端需要配置的作业参数

同步场景	目的端需要配置参数	字段映射
单表	• 基本参数 • 高级属性	支持
分库分表	• 基本参数 • 高级属性	支持
整库迁移	选择库匹配策略和表匹配策略	不支持

5. 配置字段映射关系。

   配置作业源端参数和目的端参数后，需要配置源端和目的端列的映射关系，配置字段映射关系后，任务将根据字段映射关系，将源端字段写入目标端对应类型的字段中。

   a. 字段映射配置：选择字段映射关系、设置字段批量映射规则。

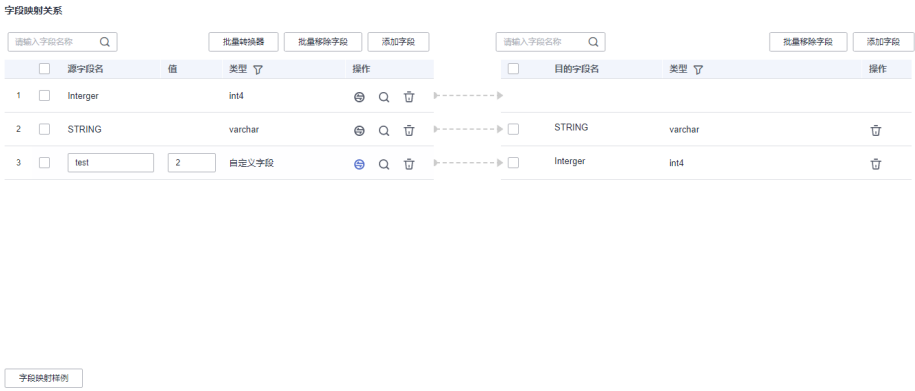
   

   - 字段映射关系
     - 同名映射：对字段名称相同的字段进行映射。使用已有数据进行相同列名的字段自动映射。
     - 同行映射：源表和目标表的字段名称不一致，但字段对应相同行的数据进行映射。查询源端和目的端的字段，再进行相同行的字段自动映射。
   - 字段批量映射：源端配置使用SQL语句为是时不显示该参数。

     批量输入字段映射数据，一行输入一个字段映射，等号左边为源表字段右边为目标表字段，例如：reader_column=writer_column。

     单击“查看编辑”，设置批量映射关系。

b. 字段映射关系：支持批量转换，添加字段，行移动等功能。



- 敏感信息检测：检测来源端数据是否包含敏感信息，存在敏感信息时无法进行数据迁移，需根据界面提示修改。
- 批量转换器：批量转换源字段名。

  勾选需要转换的字段名，单击“批量转换器”，在弹出的转换器列表对话框中根据提示新建转换器。

  批量移除字段：源端配置使用SQL语句为是时不显示该参数，勾选需要移除的字段名，单击“批量移除字段”。

  已移除的字段可以在添加字段里的“添加被移除的字段”中看到。
- 添加字段：源端配置使用SQL语句为是时不显示该参数。可以为源端和目的端添加新的字段。包含添加已被移除的字段和添加新字段。

  添加新字段支持以下类型：

  支持函数，例如mysql填写now()、curdate()、postgresql。

  支持填写now()、transaction_timestamp()。

  支持函数配合关键字，例如postgresql填写to_char(current_date,'yyyy-MM-dd')。

  支持填写固定值，例如：123、'123'，这两种填法都代表字符串：123。

  支持填写变量值，例如：${workDate}，workDate需要在作业变量中进行定义。

  JDBC支持填写固定变量，例如：DB_NAME_SRC（原始数据库名称）、TABLE_NAME_SRC（源端表名称）、DATASOURCE_NAME_SRC（源端数据源名称）。

  支持as语句，例如：'123' as test， now() as curTime。
- 行移动：源端配置使用SQL语句为是时，在设置字段映射关系阶段不支持该功能。鼠标拖住需要移动的字段所在行，可以任意移动上下位置。
- 查看转换器：（可选）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”。转换器使用详情请参见**字段转换器配置指导**。
- 查找目的端字段：CDM支持搜索查找目的端字段名并匹配字段，如果需要可单击操作列下，进入匹配目的字段对话框，通过搜索关键字或者直接单击目标进行匹配。

- 删除字段：CDM支持删除原有表的默认字段，如果需要可单击操作列下 删除字段，已移除的字段可以在添加字段里的“添加被移除的字段”中看到。
- 字段映射样例：源端配置使用SQL语句为是时不显示该参数，查看源端和目的端样例数据。

说明

- 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），没有字段映射这一步骤。

  整库迁移没有配置字段映射关系这一步骤。
- 迁移过程中可能存在源端与目标端字段类型不匹配，产生脏数据，导致数据无法正常写入目标端，迁移过程中关于脏数据的容忍条数，请参考下一步任务属性进行配置。
- 当源端某字段未与目标端字段进行映射时，源端该字段数据将不会同步到目标端。
- 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。
- 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（源端配置使用SQL语句为否时支持该功能）。
- 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，则可以单击 自定义新增字段，也可单击操作列下 创建字段转换器，确保导入到目的端的数据完整。
- 支持通过字段映射界面的 ，可自定义添加常量、变量及表达式。
- 列名仅支持源端为OBS数据源，迁移CSV文件时配置“解析首行为列名”参数为“是”时显示。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datatime）。
- 如果是导入到数据仓库服务（DWS），则还需在目的字段中选择分布列，建议按如下顺序选取分布列：
  1. 有主键可以使用主键作为分布列。
  2. 多个数据段联合做主键的场景，建议设置所有主键作为分布列。
  3. 在没有主键的场景下，如果没有选择分布列，DWS会默认第一列作为分布列，可能会有数据倾斜风险。

6. 配置任务属性。

   通过任务配置，控制数据同步过程的相关属性，具体请参见**表6-7**。

表 6-7 任务配置参数

参数	说明	取值样例
作业期望最大并发数	设置当前作业从源端并行读取或并行写入目标端的最大线程数，由于分片策略等原因，实际运行过程中的并发线程数可能小于此值。 其中，集群最大并发数的设置与CDM集群规格有关，并发数上限建议配置为vCPU核数*2。 例如8核16GB规格集群的最大抽取并发数上限为16。	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。 **说明** 目前仅对目的端为Hudi、DWS，导入模式为UPSERT生效，其他场景及配置分片重试次数不生效。	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，当脏数据过多时，会影响同步任务的整体同步速度。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
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。	obslink

参数	说明	取值样例
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。 单个分区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	是否开启同步限速。该速率代表CDM传输速率，而非网卡流量。 ● 限速：用户可以通过限速控制同步速率，可以保护读取端数据库，避免抽取速度过大，给源库造成太大的压力。限速最小配置为1MB/S。 ● 不限速：在不限速的情况下，任务将在所配置的并发数的限制基础上，提供现有硬件环境下最大的传输性能。 **说明** ● 支持对MRS Hive\DLI\关系数据库\OBS\Apache HDFS作为目的端的作业进行单并发限速。 ● 如果作业配置多并发则实际限制速率需要乘以并发数。	是
单并发速率上限（MB/s）	开启限速情况下设置的单并发速率上限值，如果配置多并发则实际速率限制需要乘以并发数。单位：MB/s。 **说明** 限制速率为大于1的整数。	10
单并发行数速率上限	设置单并发行数速率上限，单位：record/s。	100000

参数	说明	取值样例
中间队列缓存大小	数据写入时中间队列缓存大小，取值范围为1-500。 如果单行数据超过该值，可能会导致迁移失败。 如果该值设置过大时，可能会影响集群正常运行。 请酌情设置，无特殊场景请使用默认值。	64
实时检测作业敏感信息	是否开启了实时检测作业敏感信息。	否

7. 保存作业。

   作业配置完毕后，单击作业开发页面左上角“保存”按钮，保存作业的配置信息。

   作业如果开启了实时检测作业敏感信息，系统会自动检测来源端数据是否包含敏感信息，存在敏感信息时无法进行数据迁移，须根据界面提示修改。

   保存后，在右侧的版本里面，会自动生成一个保存版本，支持版本回滚。保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。

8. 测试运行作业。

   作业配置完毕后，单击作业开发页面左上角“测试运行”按钮，测试作业。如果测试未通过，请您查看作业节点的运行日志，进行定位处理。

   说明

   - 测试运行类似于单次运行，会对数据进行真实迁移。
   - 用户可以查看该作业的测试运行日志，单击“查看日志”可以进入查看日志界面查看日志的详细信息记录。
   - 作业未提交版本之前，进行手动测试运行，作业监控里面的作业运行实例版本显示是0。

9. 提交作业版本。

   若任务需要进行周期性调度运行，您需要将任务发布至生产环境。关于任务发布，详情请参见：**发布作业任务**。

10. 调度作业。
    对已编排好的作业设置调度方式。关于调度作业，详情请参见：**调度作业**。

6.5 配置作业源端参数

6.5.1 配置 MySQL 源端参数

支持从MySQL导出数据。

表 6-8 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
		• 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。
			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账号是否有元数据查询的权限。 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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 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。例如：表名配置为*user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
高级属性	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	Date类型值是否保留一位精度	Date类型值是否保留一位精度。	否
	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

参数类型	参数名	说明	取值样例
	抽取分片字段	“按表分区抽取”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片字段。 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片字段名。 **说明** 抽取分区字段支持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。	id
	分片字段是否允许空值	“按表分区抽取”选择“否”时，显示该参数，是否允许分片字段包含空值。 多并发抽取时，若确定分片字段不含Null，将该值设为“否”可提升性能，若不确定，请设为“是”，否则可能会丢数据。	是

6.5.2 配置 Hive 源端参数

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。

表 6-9 Hive 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	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
	数据库	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

参数类型	参数名	说明	取值样例
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_E
	使用SQL语句	“读取方式”选择“JDBC”时显示此参数。 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”时显示此参数，您可以在这里输入自定义的SQL语句，CDM将根据该语句导出数据。
		说明
		• SQL语句只能查询数据，支持join和嵌套写法，但不能有多条查询语句，比如 select * from table a; select * from table b。
		• 不支持with语句。
		• 不支持注释 ，比如 "--" ，“/*”。
		• 不支持增删改操作，包括但不限于以下操作：
		  • load data
		  • delete from
		  • alter table
		  • create table
		  • drop table
		  • into outfile
			select id,name from sqoop.user;

参数类型	参数名	说明	取值样例
	传输模式	支持记录迁移和文件迁移 默认为记录迁移。仅当源端为Hive2.x且数据存储在HDFS、目的端为Hive3.x且数据存在OBS并行文件系统时，才支持文件迁移。 当选择文件迁移时，需保证源端和目的端的表格式和属性需一致才能迁移成功。	• 记录迁移 • 文件迁移
	分区过滤条件	“读取方式”选择“HDFS”时显示此参数。 该参数表示抽取指定值的partition，属性名称为分区名称，属性值可以配置多个值（空格分隔），也可以配置为字段取值范围，接受时间宏函数。详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	• 单/多值过滤场景属性值：${dateformat(yyyyMMdd, -1, DAY)} ${dateformat(yyyyMMdd)} • 范围过滤场景属性值：${value} >= ${dateformat(yyyyMMdd, -7, DAY)} && ${value} < ${dateformat(yyyyMMdd)}

参数类型	参数名	说明	取值样例
高级属性	Where子句	“读取方式”选择“JDBC”，“使用SQL语句”选择“否”时显示此参数。 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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	age > 18 and age <= 60

6.5.3 配置 HDFS 源端参数

表 6-10 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。 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	/user/cdm/

参数类型	参数名	说明	取值样例
	文件格式	传输数据时所用的文件格式，可选择以下文件格式： ● CSV格式：以CSV格式解析源文件，用于迁移文件到数据表的场景。 ● 二进制格式：选择“二进制格式”时不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。 ● Parquet格式：以Parquet格式解析源文件，用于HDFS数据导到表的场景。	CSV格式
	列表文件	当“文件格式”选择为“二进制格式”时，才有该参数。 打开列表文件功能时，支持读取OBS桶中文件（如txt文件）的内容作为待迁移文件的列表。该文件中的内容应为待迁移文件的绝对路径（不支持目录），文件内容示例如下： /mrs/job-properties/application_1634891604621_0014/job.properties /mrs/job-properties/application_1634891604621_0029/job.properties	是
	列表文件源连接	当“列表文件”选择为“是”时，才有该参数。可选择列表文件所在的OBS连接。	OBS_test_link
	列表文件OBS桶	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶名。	01
	列表文件或目录	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶中的绝对路径或目录。	/0521/Lists.txt
高级属性	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n
	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。	,

参数类型	参数名	说明	取值样例
	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文件的前N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”。只有文本文件可以设置编码类型，当“文件格式”选择为“二进制格式”时，该参数值无效。	GBK
	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
	标识文件名	启动作业的标识文件名选择是显示该参数。输入文件名后，只有在源端路径下存在该文件的情况下才会执行迁移任务。标识文件不会被迁移。	ok.txt
	等待时间	启动作业的标识文件名选择是显示该参数，等待标识文件的时间，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。	60
	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。具体使用方法可参见**文件增量迁移**。	-
	目录过滤器	“过滤类型”选择“通配符”、“正则表达式”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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*input

参数类型	参数名	说明	取值样例
	文件过滤器	“过滤类型”选择“通配符”、“正则表达式”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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*.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等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**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-01 00:00:00
	终止时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-30 00:00:00

参数类型	参数名	说明	取值样例
	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。 详细使用方法请参见**迁移文件时加解密**。	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
	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，详细请参见**MD5校验文件一致性**。	.md5

6.5.4 配置 Hudi 源端参数

表 6-11 MRS Hudi 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	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Hudi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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_E
高级属性	Where子句	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	age > 18 and age <= 60

6.5.5 配置 PostgreSQL 源端参数

支持从云端的数据库服务导出数据。

这些非云服务的数据库，既可以是用户在本地数据中心自建的数据库，也可以是用户在ECS上部署的，还可以是第三方云上的数据库服务。

OpenGauss数据源与PostgreSQL一致，可参考本章节配置。

表 6-12 PostgreSQL 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
		● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。
			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账号是否有元数据查询的权限。 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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 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。例如：表名配置为*user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
高级参数	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	按表分区抽取	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

参数类型	参数名	说明	取值样例
	抽取分片字段	“按表分区抽取”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片字段。 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片字段名。 **说明** 抽取分区字段支持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。	id
	分片字段是否允许空值	“按表分区抽取”选择“否”时，显示该参数，是否允许分片字段包含空值。	是

6.5.6 配置 SQLServer 源端参数

支持从云端的数据库服务导出数据。

这些非云服务的数据库，既可以是用户在本地数据中心自建的数据库，也可以是用户在ECS上部署的，还可以是第三方云上的数据库服务。

表 6-13 SQLServer 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

参数类型	参数名	说明	取值样例
	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 ● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。	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账号是否有元数据查询的权限。 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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 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。例如：表名配置为 *user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
高级属性	抽取分片字段	“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片字段。 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片字段名。 **说明** 抽取分区字段支持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。	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	分片字段是否允许空值	“按表分区抽取”选择“否”时，显示该参数，是否允许分片字段包含空值。	是

6.5.7 配置 Oracle 源端参数

支持从Oracle导出数据。

表 6-14 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
		● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。
			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账号是否有元数据查询的权限。 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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 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。例如：表名配置为*user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
高级属性	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	Date类型值是否保留一位精度	Date类型值是否保留一位精度。	否
	按表分区抽取	“按表分区抽取”选择“否”时，显示该参数，表示从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

参数类型	参数名	说明	取值样例
	抽取分片字段	“按表分区抽取”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片字段。 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片字段名。 **说明** 抽取分区字段支持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。	id
	分片字段是否允许空值	“按表分区抽取”选择“否”时，显示该参数，是否允许分片字段包含空值。 多并发抽取时，若确定分片字段不含Null，将该值设为“否”可提升性能，若不确定，请设为“是”，否则可能会丢数据。	是

6.5.8 配置 DLI 源端参数

支持从DLI导出数据。

表 6-15 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	用于抽取分区的信息。是否支持配置以界面实际为准。	year=2020,location=sun

6.5.9 配置 OBS 源端参数

表 6-16 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

参数类型	参数名	说明	取值样例
	文件格式	传输数据时使用的格式。 • 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。 • 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。 • ORC格式：以ORC格式解析源文件，一般都是用于迁移文件到数据表的场景。 • PARQUET格式：以PARQUET格式解析源文件，一般都是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。适用于文件迁移场景，比如OBS到OBS。	CSV格式
	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，具体请参见**文件列表迁移**。 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	FROM/example.csv
	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

参数类型	参数名	说明	取值样例
	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。 此处建议选择为文件的绝对路径。当选择为目录时，也支持迁移子目录中的文件，但如果目录下文件量过大，可能会导致集群内存不足。	/0521/Lists.txt
	JSON类型	当“文件格式”选择为“JSON格式”时，才有该参数。JSON文件中存储的JSON对象的类型，可以选择“JSON对象”或“JSON数组”。	JSON对象
	记录节点	当“文件格式”选择为“JSON格式”并且“JSON类型”为“JSON对象”时，才有该参数。对该JSON节点下的数据进行解析，如果该节点对应的数据为JSON数组，那么系统会以同一模式从该数组中提取数据。多层嵌套的JSON节点以字符“.”分割。	data.list
高级属性	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n
	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
	使用转义符	选择“是”时，CSV数据行中的\作为转义符使用。 选择“否”时，CSV中的\作为数据不会进行转义。CSV只支持\作为转义符。	是
	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，正则表达式写法请参考**正则表达式分隔半结构化文本**。	^(\d.*\d)(\w*) \[(.*)\] ([\w\.]*)(\w.*).*

参数类型	参数名	说明	取值样例
	前N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文件的前N行数据作为标题行，不写入目的端的表。	否
	标题行数	“前N行为标题行”选择“是”时才有该参数。抽取数据时将被跳过的标题行数。 **说明** 标题行数不为空，取值为1-99之间的整数。	1
	解析首行为列名	“前N行为标题行”选择“是”时才有该参数。选择是否将标题的首行解析为列名，在配置字段映射时会在原字段中显示该列名。 **说明** ● 标题行数大于1时，当前仅支持解析标题的首行作为列名。 ● 列名不支持“&”字符，否则会导致作业迁移失败，需修改CSV文件“&”字符即可正常迁移。	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。	GBK
	压缩格式	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	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，否则就保持原样传输。当输入*或为空时，所有文件都会被解压。	*
	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
	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“|”。	|

参数类型	参数名	说明	取值样例
	标识文件名	选择开启作业标识文件的功能时，需要指定启动作业的标识文件名。指定文件后，只有在源端路径下存在该文件的情况下才会运行任务。该文件本身不会被迁移。	ok.txt
	等待时间	选择开启作业标识文件的功能时，如果源路径下不存在启动作业的标识文件，作业挂机等待的时长，当超时后任务会失败。 等待时间设置为0时，当源端路径下不存在标识文件，任务会立即失败。 单位：秒。	10
	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。具体使用方法可参见**文件增量迁移**。	通配符
	目录过滤器	“过滤类型”选择“通配符”、“正则表达式”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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*input
	文件过滤器	“过滤类型”选择“通配符”、“正则表达式”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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*.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

参数类型	参数名	说明	取值样例
	起始时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间大于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为（数据开发作业计划启动时间-偏移量），而不是（CDM作业实际启动时间-偏移量）。	2019-06-01 00:00:00
	终止时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-01 00:00:00
	忽略不存在原路径/文件	在迁移过程中发现文件在源路径下不存在的情况下是否报错。如果将其设为是，那么文件在源路径下不存在的情况下也能成功执行。	否
	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，详细请参见**MD5校验文件一致性**。	.md5

6.5.10 配置 SAP HANA 源端参数

表 6-17 SAP HANA 作源端时的作业参数

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账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： ● **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 ● ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 ● ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E

参数类型	参数名	说明	取值样例
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **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	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'

参数类型	参数名	说明	取值样例
	抽取分片字段	“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片字段。 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片字段名。 **说明** 抽取分区字段支持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。	id
	分片字段是否允许空值	多并发抽取时，若确定分片字段不含Null，将该值设为“否”可提升性能，若不确定，请设为“是”，否则可能会丢数据。	是

6.5.11 配置 Kafka 源端参数

表 6-18 Kafka 作为源端时的作业参数

参数类型	参数	说明	取值样例
基本参数	Topic	主题名称。支持单个topic。	cdm_topic
	数据格式	解析数据时使用的格式： ● JSON：以JSON格式解析源数据。 ● CSV格式：以CSV格式解析源数据。	JSON格式
	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
	消费记录策略	消费record策略。 ● 起止时间：根据kafka record元数据TIMESTAMP判断，抽取的record是否符合填入的起止时间范围。当消费到的record到达结束时间，则终止抽取任务。起止时间范围左闭右开:[起始时间，结束时间)。可配合调度任务使用。 ● 最早：表示从开始点位消费数据。 ● 最新：表示从最后点位消费数据。已提交：拉取已提交的数据。起止时间策略，等待时间，最大抽取时间相互独立。只要有任意一个条件符合，则kafka抽取结束。	起止时间

参数类型	参数	说明	取值样例
	起始时间	消费记录策略为起始时间时须设置起始时间。 格式为yyyy-MM-dd HH:mm:ss，支持配合DLF变量等方式设置。	2024-07-25 00:00:00
	结束时间	消费记录策略为起始时间时须设置结束时间。 格式为yyyy-MM-dd HH:mm:ss，支持配合DLF变量等方式设置。	2024-07-25 23:59:59
	等待时间	消费者获取数据返回值为空，持续X秒，任务停止。	30秒
	最大抽取时间	消费者最大抽取时间，单位min。 kafka抽取consumer端最大运行时间，当到达运行时间，抽取强制结束，如不填入，默认为30min。	1440
	字段分隔符	迁移时的字段分割符，默认为空格。	,
	记录分隔符	暂不支持@ || $特殊字符作为分隔符。	,

6.5.12 配置 Rest Client 源端参数

表 6-19 Rest Client 作为源端时的作业参数

参数	说明	取值样例
数据请求地址	数据请求的地址。	/data/query
请求方法	请求方法，GET/POST。	GET
请求体	请求方法为POST时显示该参数。 请求体，json格式。	是 {"namePrefix": "test"}
每次拉取的数量	每次拉取的数量。	1000
分页大小参数名称	分页大小参数名称，默认放到query参数中。如果参数名设置为page_size，也支持通过#page_size获取。	page_size
分页页码参数名称	分页参数名称，默认放到query参数中。如果参数名设置为page_index，也支持通过#page_index获取。	page_index
数据路径	数据在json中的位置，默认为根路径，不填则取默认。	student

参数	说明	取值样例
数据总数	数据总数，支持填写固定值，也支持从接口中获取。 1. 固定值，例如：100000。 2. 支持从返回的结果中，获取数据总量。例如：page.pageCount。 **说明** 如果接口不是分页接口，将数据总数的值设置为小于或者等于每次拉取的数据，则接口只会调用一次，否则调用多次可能导致数据重复。	100000

6.5.13 配置 DWS 源端参数

表 6-20 DWS 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
		• 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。
			select id,name from sqoop.user;

参数类型	参数名	说明	取值样例
	模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的账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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 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。例如：表名配置为*user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。 **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。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table

参数类型	参数名	说明	取值样例
高级属性	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。 指定抽取的是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，迁移失败。例如：age > 18 and age <= 60。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	抽取分片字段	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片字段。 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片字段名。 **说明** 抽取分区字段支持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。	id
	分片字段是否允许空值	是否允许分片字段包含空值。 多并发抽取时，若确定分片字段不含Null，将该值设为“否”可提升性能，若不确定，请设为“是”，否则可能会丢数据。	是

6.5.14 配置 FTP/SFTP 源端参数

表 6-21 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	要传输的目录或单个文件路径。	FROM_DIRECTORY/ or FROM_DIRECTORY/example.csv

参数类型	参数名	说明	取值样例
	文件格式	传输数据时使用的格式。 支持CSV格式，JSON格式及二进制格式。 其中CSV和JSON仅支持迁移到数据表场景，二进制格式适用于文件迁移场景。	CSV格式
	JSON类型	文件格式为JSON格式时支持此参数。 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	文件类型为JSON对象时支持此参数。 记录数据的根节点。该节点对应的数据为JSON数组，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	使用rfc4180解析器	文件格式为CSV格式时支持此参数。 是否使用rfc4180解析器解析CSV文件。	否
	换行符	文件格式为CSV格式时支持此参数。 文件中的换行符，默认自动识别"\n"、"\r"或"\r\n"。	\n
	使用包围符	文件格式为CSV格式时支持此参数。 使用包围符来括住字符串值。包围符内的字段分隔符被视为字符串值的一部分，目前只支持"作为包围符。	否
	使用转义符	文件格式为CSV格式时支持此参数。 CSV只支持\作为转义符。 选择是，CSV数据行中的\作为转义符使用。 选择否，CSV中的\作为数据不会进行转义。	是
	使用正则表达式分隔字段	文件格式为CSV格式时支持此参数。 是否使用正则表达式分隔字段。	是
	正则表达式	文件格式为CSV格式且使用正则表达式分隔字段为是时支持此参数。 分隔字段的正则表达式。	^(\d.*\d)(\w*) \[(.*)\] ([\w\.]*)(\w.*).*
	字段分隔符	文件格式为CSV格式且使用正则表达式分隔字段为否时支持此参数。 文件中的字段分隔符。	,
	首行为标题行	文件格式为CSV格式时支持此参数。 如果指定了该参数，程序在抽取数据时将读取第一行作为标题行。	否

参数类型	参数名	说明	取值样例
	编码类型	文件格式为CSV格式或JSON格式时支持此参数。 文件编码类型。 只有文本文件可以设置编码类型，否则设置无效。 支持的文件编码类型有UTF-8 、 GBK。	UTF-8
	压缩格式	压缩格式。 默认无。支持的压缩格式有GZIP，ZIP及TAR.GZ。	GZIP
	压缩文件后缀	压缩格式为GZIP，ZIP或TAR.GZ时支持此参数。 需要解压缩的文件的后缀名。 当一批文件中以该值为后缀时，才会执行解压缩操作，否则就保持原样传输。当输入"*"时或输入为空时，所有文件都会被解压。	tar.gz
	文件分隔符	多文件列表时指定的文件分隔符。	|
	启动作业标识文件	当源端路径下存在启动作业的标识文件时才启动任务，否则会挂起等待一段时间。	否
	标识文件名	启动作业标识文件为是时支持此参数。 启动作业的标识文件名。输入文件名后，只有在源端路径下存在该文件的情况下才会执行迁移任务。标识文件不会被迁移。	ok.txt
	等待时间	启动作业标识文件为是时支持此参数。 等待标识文件的时间。 超时后任务会失败，当等待时间设置为0且源端路径下不存在标识文件，任务会立即失败。单位：秒。	60
	标识文件类型	启动作业标识文件为是时支持此参数。 标识文件的类型。 ● MARK_DONE：只有在源端路径下存在标识文件的情况下才会执行迁移任务。 ● MARK_DOING：只有在源端路径下不存在标识文件的情况下才会执行迁移任务。	MARK_DONE
	过滤类型	传输满足过滤条件的文件。 支持的过滤条件有：无，通配符及正则表达式。	无

参数类型	参数名	说明	取值样例
	目录过滤器	过滤类型为通配符或正则表达式时支持此参数。 用于过滤输入路径下的一级或多级目录。	• 通配符使用input*/test* • 正则表达式使用intput.*/test.*
	文件过滤器	过滤类型为通配符或正则表达式时支持此参数。 用于过滤输入路径下的文件。	• 通配符使用*csv • 正则表达式使用.*\.csv
	时间过滤	用于过滤满足时间范围的文件。 • 文件的修改时间晚于输入的起始时间或早于输入的终止时间才会被传输。 • 同时输入起始时间和终止时间，文件的修改时间在这个区间内才会被传输。	否
	起始时间	时间过滤为是时支持此参数。 指定一个时间值，当文件的修改时间晚于该时间才会被传输。早于当前时间且不能晚于终止时间。时间格式为“yyyy-MM-dd HH:mm:ss”。	2018-01-01 00:00:00
	终止时间	时间过滤为是时支持此参数。 指定一个时间值，当文件的修改时间早于该时间才会被传输。早于当前时间且不能早于起始时间。时间格式为“yyyy-MM-dd HH:mm:ss”。	2018-01-01 00:00:00
	忽略不存在原路径/文件	在迁移过程中发现文件在源路径下不存在的情况下是否报错。如果将其设为是, 那么文件在源路径下不存在的情况下也能成功执行。	否
	是否跳过空行	文件格式为CSV格式时支持此参数。 如果某行数据为空，则跳过此行。	否
	null值	文件格式为CSV格式时支持此参数。 由于文本文件中无法用字符串定义null值，此配置项定义将何种字符串标识为null。 例如：如果配置为null，则数据中如果存在某行某列值为“null”，则会被解析为null值。	-
	MD5文件名后缀	文件格式为二进制格式时支持此参数。 校验CDM抽取的文件，是否与源文件一致。	.md5

6.5.15 配置 Doris 源端参数

Doris源端参数列表

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
		● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。
			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账号是否有元数据查询的权限。 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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 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。例如：表名配置为 *user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
高级属性	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	抽取分片字段	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片字段。 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片字段名。 **说明** 抽取分区字段支持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。	id
	分片字段含有空值	“按表分区抽取”选择“否”时，显示该参数，是否允许分片字段包含空值。	是

6.5.16 配置 HBase 源端参数

表 6-22 Hbase 作为源端时的作业参数

参数类型	参数名	说明	是否必须填	取值样例
基本参数	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	是	table
	整表迁移	源端和目的端都为HBase时显示该参数。 整表迁移通过二进制传输数据，表的所有信息都会传递。HBase->HBase整表迁移会传递列的timestamp信息，非整表迁移只传递列的value值。	是	否
	列族	导出数据的列族。 例如：CF1&CF2	是	CF1&CF2
高级属性	切分Rowkey	是否将选做Rowkey的数据同时写入HBase的列，默认否。	否	否
	Rowkey分隔符	切分Rowkey为是时显示该参数。 分隔符，用于切分Rowkey，若不设置则不切分。例如：|。	否	|

参数类型	参数名	说明	是否必须填	取值样例
	开始时间	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 ${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	否	2017-12-31 20:00:00" 或 "${dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00" 或 ${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}
	结束时间	终止时间（不包含该值）。格式为'yyyy-MM-dd hh:mm:ss', 支持dateformat时间宏变量函数。 例如："2018-01-01 20:00:00" 或 "${dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00" 或 "${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}"	否	"2018-01-01 20:00:00" 或 "${dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00" 或 "${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}"
	开始RowKey	填写需要查询的RowKey。	否	0001
	结束RowKey	填写需要结束的RowKey。	否	0100

6.5.17 配置 ClickHouse 源端参数

参数类型	参数名	说明	取值样例
基本参数	模式或表空间	“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账号是否有元数据查询的权限。 该参数支持配置正则表达式，实现导出满足规则的所有数据库。	SCHEMA_E
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。 单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
高级属性	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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'

6.5.18 配置 ElasticSearch 源端参数

表 6-23 ElasticSearch 作为源端时的作业参数

参数类型	参数名	说明	是否必须填	取值样例
基本参数	索引	类似关系数据库的schema或数据库名称，整库迁移多索引以逗号分隔。 支持输入索引别名。 支持输入通配符表达式(*)。如果选择了多个索引，索引的结构必须一致。 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	是	index_sample
	类型	类似关系数据库的schema或数据库名称，整库迁移多索引以逗号分隔。 支持输入索引别名。 支持输入通配符表达式(*)。如果选择了多个索引，索引的结构必须一致。	是	type_example
高级属性	拆分nested类型字段	是否将nested字段的json内容拆分,如 a:{ b:{ c:1, d:{ e:2, f:3 } } } 将拆成三个字段 [a.b.c], [a.b.d.e], [a.b.d.f]。	否	是
	过滤条件	对源数据进行过滤，使用ES查询的参数q语法。	否	last_name:Smith
	抽取元字段	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。	否	_index
	分页大小	分页大小。	否	1000
	ScrollId超时时间配置	ScrollId超时时间配置，默认5分钟。	否	5

参数类型	参数名	说明	是否必须填	取值样例
	重试次数	单次请求失败重试次数。最大限制重试次数10次。	否	3

6.5.19 配置 MongoDB 源端参数

表 6-24 MongoDB 作为源端时的作业参数

参数类型	参数名	说明	是否必须填	取值样例
基本参数	数据库	输入或选择数据库名称，单击输入框后面的按钮可进入集合的选择界面。	是	default
	集合名	输入或选择集合名，单击输入框后面的按钮可进入集合的选择界面。 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	是	table
高级属性	查询筛选	创建用于匹配文档的筛选器。 例如：{HTTPStatusCode:{$gt:"400",$lt:"500"},HTTPMethod:"GET"}。	否	{HTTPStatusCode:{$gt:"400",$lt:"500"},HTTPMethod:"GET"}

6.5.20 配置 RestApi 源端参数

表 6-25 RestApi 作为源端时的作业参数

参数类型	参数名	说明	是否必须填	取值样例
基本参数	数据请求地址	数据请求地址。	是	/api/getUsers
	请求方法	请求方法，支持GET/POST。	是	GET
	请求体	请求方法为POST时显示该参数。 请求体，json格式。	是	{"namePrefix":"test"}
	每次拉取的数量	每次拉取的数量。	是	1000
	分页大小参数名称	分页大小参数名称。 ● 默认会放到query参数中，它的值为每次拉取的数量。 ● 如果body参数中包含此参数名，则会将它的值替换为每次拉取的数量。	是	pageSize
	分页页码参数名称	分页页码参数名称。 ● 默认会放到query参数中，它的值为页码。 ● 如果body参数中包含此参数，则会将它的值替换为页面。	是	pageNumber
	数据路径	数据路径，指数据在响应json体中的位置，默认为根路径。	否	data.students
	数据总数	数据总数，可以支持填写固定值，也可以支持从接口中获取，支持spel表达式。 ● 固定值。 ● 从接口中获取：data.pageCount。 **说明** 如果接口不是分页接口，并且只想调用一次，则将数据总数的值设置的小于或者等于每次拉取的数据。	是	固定值时推荐1000

6.5.21 配置 GBase 源端参数

表 6-26 GBase 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，作业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
		● 如果SQL语句过长，会导致请求过长下发失败，继续创建作业系统会报错“错误请求”，此时您需要简化或清空SQL语句，再次尝试继续创建作业。
			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账号是否有元数据查询的权限。 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_EXAMPLE

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，或者对应连接里的账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，详细说明请参见**使用时间宏变量完成增量同步**。 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。例如：表名配置为*user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TABLE_EXAMPLE
高级属性	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。 如果要迁移的表中没有Where子句的字段，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，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	age > 18 and age <= 60
	Date类型值是否保留一位进度	Date类型值是否保留一位进度。 目的端为Hudi、Hive时显示该参数。	否

参数类型	参数名	说明	取值样例
	抽取分片字段	“按表分区抽取”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片字段。 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片字段名。 **说明** 抽取分片字段支持TINYINT、SMALLINT、INTEGER、BIGINT、FLOAT、DOUBLE、NUMERIC、DECIMAL、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。	id
	分片字段含有空值	“按表分区抽取”选择“否”时，显示该参数，是否允许分片字段包含空值。 多并发抽取时，若确定分片字段不含Null，将该值设为“否”可提升性能，若不确定，请设为“是”，否则可能会丢数据。	是

6.5.22 配置 Redis 源端参数

表 6-27 Redis 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	redis键前缀	对应关系数据库的表名。	TABLENAME
	值存储类型	存储类型分STRING、HASH。	STRING
高级属性	键分隔符	用来分隔关系数据库的表和列名。	_
	值分隔符	以STRING方式存储，列之间的分隔符。 存储类型为列表时字串分割成数组的字符。	;
	字段相同	“值存储类型”选择“HASH”时，显示该参数，是否允许哈希键内有相同的字段。	否

6.5.23 配置 LTS 源端参数

表 6-28 LTS 作为源端时的作业参数

参数名	说明	取值样例
源连接名称	对应关系数据库的表名。	TABLENAME
单次查询数据条数	一次从日志服务查询的数据条数。	128
日志分组	日志组是云日志服务进行日志管理的基本单位。	_
日志流	日志流是日志读写的基本单位。	-
数据消费开始时间	数据消费的开始时间位点，即日志数据到达LogHub（LTS）的时间，该参数为时间范围（左闭右开）的左边界。	20240701235959
数据消费结束时间	数据消费的结束时间位点，为时间范围（左闭右开）的右边界。	20240702235959

6.6 配置作业目的端参数

6.6.1 配置 PostgreSQL 目的端参数

OpenGauss数据源与PostgreSQL一致，可参考本章节配置。

表 6-29 PostgreSQL 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
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table

参数类型	参数名	说明	取值样例
	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
高级属性	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，具体请参见**事务模式迁移**。 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
	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
	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

6.6.2 配置 Oracle 目的端参数

表 6-30 Oracle 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
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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
高级属性	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，具体请参见**事务模式迁移**。 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
	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

参数类型	参数名	说明	取值样例
	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
	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

6.6.3 配置 MySQL 目的端参数

表 6-31 PostgreSQL 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
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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

参数类型	参数名	说明	取值样例
	约束冲突处理	导入数据到云数据库 MySQL且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
	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
高级属性	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，具体请参见**事务模式迁移**。 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
	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
	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

6.6.4 配置 SQLServer 目的端参数

表 6-32 SQL Server 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
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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
高级属性	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，具体请参见**事务模式迁移**。 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
	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

参数类型	参数名	说明	取值样例
	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
	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

6.6.5 配置 Hudi 目的端参数

表 6-33 MRS Hudi 作为目的端时的作业参数

类别	配置项	配置说明	推荐配置
基本参数	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。	dbadmin
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	cdm
	自动建表模式	是否自动创建Hudi表。 ● 一键建表：通过自动建表方式自动创建目的端表。 ● 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。	不存在时创建

类别	配置项	配置说明	推荐配置
	写入模式	数据写入模式。 ● TRUNCATE+LOAD：TRUNCATE方式会在导入前执行TRUNCATE语句清空填写的分区数据，再进行LOAD写入数据。 ● LOAD：写入前不做任何处理。 ● INSERT_OVERWRITE：对数据进行覆盖写入。	LOAD
	分区	分区信息，表为分区表的时候，写数据的时候，可以选择需要写入的分区数据。 例如：year=2020,location=sun。	-
高级属性	入库时间字段	将一个字段标记为入库时间字段，自动建表时将此字段自动加到建表语句中，写入Hudi时将把此字段的值替换为当前时间。所选字段必须为timestamp类型。	-
	写入参数	在执行Spark SQL往hudi插入数据前，通过set语法设置参数，从而控制spark的写入行为。	hoodie.combine.before.upsert

6.6.6 配置 Hive 目的端参数

支持快速导入数据到MRS的Hive。

表 6-34 Hive 作为目的端时的作业参数

类别	参数名	说明	取值样例
基本参数	数据库	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
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_X

类别	参数名	说明	取值样例
	hive写入模式	选择hive的写入模式。 ● TRUNCATE+LOAD：TRUNCATE模式只清理分区下的数据文件，不删除分区。 ● LOAD：写入前不做任何处理。 ● LOAD_OVERWRITE：将生成一个临时目录，目录名为:表名_UUID，使用hive的load overwrite语法将临时目录加载到hive表中。	LOAD_OVERWRITE
	分区过滤条件	TRUNCATE模式，支持多组分区，并在对应的输入框填的值即可。 LOAD_OVERWRITE模式，仅支持写入一组分区。	-
高级属	是否将null转换为“null”	配置null值的转换类型。 ● TO_NULL：null值不处理。 ● TO_EMPTY_STRRING：将null值转换为空字符串。 ● TO_NULL_STRING：将null值转换为"null"字符串。	TO_NULL
	换行符处理方	对于写入hive textfile格式表的数据中存在换行符的场景，指定对换行符的处理策略。 支持删除，替换为其它字符串及不处理三种方式。	删除
	换行符替换字符串	换行符处理方式设置为“替换为其他字符串”时，呈现此参数。 当换行符处理方式选择为替换时，指定替换的字符串。	-
	执行Analyze语句	数据全部写入完成后会异步执行ANALYZE TABLE语句，用于优化Hive表查询速度，执行的SQL如下： ● 非分区表：ANALYZE TABLE tablename COMPUTE STATISTICS ● 分区表：ANALYZE TABLE tablename PARTITION(partcol1[=val1], partcol2[=val2], ...) COMPUTE STATISTICS **说明** ● “执行Analyze语句”参数配置仅用于单表迁移场景。 ● 执行ANALYZE语句可能会对Hive造成压力。	是

6.6.7 配置 DLI 目的端参数

表 6-35 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。 新建队列操作请参考**创建队列**。	cdm
数据库名称	写入数据的数据库名称。	dli
表名	写入数据的表名。	car_detail
导入模式	选择导入模式。 • TRUNCATE方式：会在导入前执行。 • TRUNCATE方式：清空DLI表分区。 • INSERT_OVERWRITE方式：使用分区覆盖的方式写入数据。	INSERT_OVERWRITE
空字符串作为null	如果设置为true，空字符串将作为null。	否
自动建表模式	选择建表模式：一键建表，作业配置过程中一键建表，表生成后继续配置作业。	一键建表
分区	分区信息。在分区字段对应的框输入分区的值。	year=2020,location=sun

6.6.8 配置 ElasticSearch 目的端参数

表 6-36 Elasticsearch 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	索引	待写入数据的Elasticsearch的索引，类似关系数据库中的数据库名称。CDM支持自动创建索引和类型，索引和类型名称只能全部小写，不能有大写。	index
	类型	待写入数据的Elasticsearch的类型，类似关系数据库中的表名称。类型名称只能全部小写，不能有大写。 **说明** Elasticsearch搜索引擎7.x及以上版本不支持自定义类型，只能使用_doc类型。此处即使自定义也不会生效。	type

参数类型	参数名	说明	取值样例
	操作	操作类型。 ● INDEX：不指定主键，es内部生成id，使得每次写入都是不同id的新增数据文件。 ● CREATE：需要指定主键。如果主键已经存在，写入失败。 ● UPDATE：需要指定主键。如果主键已经存在，覆盖原有数据。 ● UPSERT：需要指定主键。如果主键已经存在，同UPDATE。如果主键不存在，则新建文档写入。	UPSERT
	主键取值方式	文档类型为UPSERT，UPDATE或CREATE时支持的主键取值方式。 ● 单主键：业务主键模式，选择主键，将其的值写入id。 ● 联合主键：联合主键模式，多选主键，将其的值用主键分隔符拼接写入id。 ● 无主键：仅操作类型为CREATE时支持，无需指定主键，目的端会自动生成id作为主键写入。	单主键
	导入前清空数据	定义当前任务在索引Index已经存在的情况是否需要删除数据。 ● 是：需要删除该索引下的数据。 ● 否：写入数据前保留数据。	否
	主键分隔符	主键取值方式为“联合主键”时，显示主键分隔符配置项，用于将多选的主键用主键分隔符拼接写入id。	_
高级属性	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
	开启路由	开启路由后，支持指定某一列的值作为路由写入Elasticsearch。 **说明** 开启路由前建议先建好目的端索引，可提高查询效率。	否
	路由字段	**“开启路由”**参数选择为“是”时配置，用于配置目的端路由字段。目的端索引存在但是获取不到字段信息时，支持手动填写字段。路由字段允许为空，为空时写入Elasticsearch不指定routing值。	value1

参数类型	参数名	说明	取值样例
	定时创索引	对于持续写入数据到Elasticsearch的流式作业，CDM支持在Elasticsearch中定时创建新索引并写入数据，方便用户后期删除过期的数据。支持按以下周期创建新索引： ● 每小时：每小时整点创建新索引，新索引的命名格式为“索引名+年+月+日+小时”，例如“index2018121709”。 ● 每天：每天零点零分创建新索引，新索引的命名格式为“索引名+年+月+日”，例如“index20181217”。 ● 每周：每周周一的零点零分创建新索引，新索引的命名格式为“索引名+年+周”，例如“index201842”。 ● 每月：每月一号零点零分创建新索引，新索引的命名格式为“索引名+年+月”，例如“index201812”。 ● 不创建：选择此项表示不创建定时索引。 从文件类抽取数据时，必须配置单个抽取（“抽取并发数”参数配置为1），否则该参数无效。	每小时
	单行提交次数	配置需要单次提交的大小。	10000
	重试次数	单次请求失败重试次数，最大限制重试次数10次。	3

6.6.9 配置 DWS 目的端参数

表 6-37 DWS 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。 整库迁移时无该参数。	schema

参数类型	参数名	说明	取值样例
	表名	写入数据的目标表名，单击输入框后面的按钮可进入表的选择界面。 整库迁移时无该参数。 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
	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_UPSERT模式，使用DWS专有的高性能批量入库工具。	COPY
	导入开始前	导入数据前，选择是否清除目的表的数据： • 不清除：写入数据前不清除目标表中数据，数据追加写入。 • 清除全部数据：写入数据前会清除目标表中数据。 • 清除部分数据：需要配置“where条件”参数，CDM根据条件选择性删除目标表的数据。	清除部分数据
	where条件	**“导入开始前”**参数选择为“清除部分数据”时配置，配置后导入前根据where条件删除目的表的数据。	age > 18 and age <= 60
高级属性 整库迁移时无该参数。	先导入阶段表	如果选择“是”，则启用事务模式迁移，CDM会自动创建临时表，先将数据导入到该临时表，导入成功后再通过数据库的事务模式将数据迁移到目标表中，导入失败则将目的表回滚到作业开始之前的状态。 默认为“否”，CDM直接将数据导入到目的表。这种情况下如果作业执行失败，已经导入到目标表中的数据不会自动回滚。 **说明** 如果“导入开始前”选择“清除部分数据”或“清除全部数据”，CDM的事务模式不会回滚已经删除的数据。	否

参数类型	参数名	说明	取值样例
	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
	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** 并发场景下有如下限制：约束冲突处理策略不支持"replace into"或"on duplicate key update"。	1

6.6.10 配置 OBS 目的端参数

支持使用CSV、CarbonData或二进制格式批量传输大量文件到OBS。

表 6-38 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	directory/

参数类型	参数名	说明	取值样例
	文件格式	传输数据时使用的格式。其中CSV和JSON仅支持迁移到数据表场景，二进制格式适用于文件迁移场景。例如：CSV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。 ● Parquet格式：按Parquet格式写入，适用于数据表到文件的迁移。 ● ORC格式：按ORC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。 **说明** ● 当源端为MRS Hive数据源时，仅支持配置CSV格式。 ● 当源端为FTP/SFTP数据源时，仅支持配置二进制格式。	CSV格式
	重复文件处理方式	“文件格式”为“CSV”时不存在该参数。 ● （二进制格式时）对于Binary，CSV的文件迁移场景，判断条件为文件名相同，文件大小相同。 – REPLACE：替换重复文件。 – SKIP：跳过重复文件。 – ABANDON：停止任务。 ● 对于Parquet、ORC的结构化集成场景，判断条件为自定义文件名前缀匹配。 – REPLACE：写入前清理自定义文件名前缀匹配的所有文件。例如“自定义文件名”：“abc”，将清理所有abc开头的文件。 – APPEND：写入前不进行任何处理 – ABANDON：如果目录下存在自定义文件名前缀匹配的文件，直接报错。	REPLACE
高级属性	字段分隔符	文件中的字段分隔符。“文件格式”为“二进制格式”时该参数值无效。	,

参数类型	参数名	说明	取值样例
	写入文件大小	源端为数据库时该参数才显示，支持按大小分成多个文件存储，避免导出的文件过大，单位为MB。	1024
	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
	首行为标题行	从关系型数据库导出数据到OBS，“文件格式”为“CSV格式”时，才有该参数。 在迁移表到CSV文件时，CDM默认是不迁移表的标题行，如果该参数选择“是”，CDM在才会将表的标题行数据写入文件。	否
	校验MD5值	计算源文件的MD5值，并与OBS返回的MD5值进行校验。 “文件格式”为“二进制格式”时，才有该参数。 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。例如：否	否
	记录校验结果	“文件格式”为“二进制格式”时，才有该参数。 将MD5的校验结果写入到OBS。记录每个文件的校验结果。例如：否	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
	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
	自定义目录层次	选择“是”时，支持迁移后的文件按照自定义的目录存储。即只迁移文件，不迁移文件所归属的目录。	是

参数类型	参数名	说明	取值样例
	目录层次	自定义迁移后文件的存储路径，支持时间宏变量。 **说明** 源端为关系型数据库数据源时，目录层次为源端表名+自定义目录，其他场景下为自定义目录。	${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}
	压缩格式	“文件格式”为“CSV”时不存在该参数。 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。	无
	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。 详细使用方法请参见**迁移文件时加解密**。	KMS
	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
	复制Content-Type属性	“文件格式”为“二进制格式”时，才有该参数。 上传对象时复制源文件的“Content-Type”属性，主要用于静态网站的迁移场景。不支持写入到归档存储的桶。	否

参数类型	参数名	说明	取值样例
	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}”，则生成的文件名为源表名“sqltabname.csv”。 ● 版本宏，例如“${version}”，则生成的文件名为集群版本号“2.9.2.200.csv”。 ● 字符串和宏（时间宏/表名宏/版本宏）任意组合，例如“cdm#${timestamp()}_${version}”，则生成的文件名为“cdm#1554108737_2.9.2.200.csv”。	cdm
	Blob开关	从关系型数据库导出数据到OBS，才有该参数。 启用后将会以根目录-表名-数据类型-数据的文件夹模型生成文件。例如：raw_schema/tbl_student/datas/tbl_student_1.csv	否
	Blog文件扩展名	“文件夹模式”为“是”时，才有该参数。文件夹模式下自定义Blob/Clog数据的文件扩展名。	.dat/.jpg/.png

6.6.11 配置 SAP HANA 目的端参数

表 6-39 SAP HANA 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

参数类型	参数名	说明	取值样例
	表名	写入数据的目标表名，单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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	导入开始前选择清除部分数据时显示该参数。 导入前根据条件删除目的表部分数据。	age > 18 and age <= 60
	写入模式	● INSERT：可向表中插入一行或多行数据。 ● UPSERT：数据存在则更新，不存在则新增。	INSERT
高级属性	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，具体请参见**事务模式迁移**。 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
	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

参数类型	参数名	说明	取值样例
	loader线程数	每个loader内部启动的线程数，可以提升写入并发数。 **说明** 不支持“约束冲突处理”策略为“replace into”或“on duplicate key update”的并发场景。	1

6.6.12 配置 ClickHouse 目的端参数

表 6-40 ClickHouse 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
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
高级属性	单次写入行数	指定单次批量写入的行数（注意：一次事务提交100个批量的数据）。	10000
	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

6.6.13 配置 Doris 目的端参数

表 6-41 Doris 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
	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，详细说明请参见使用时间宏变量完成增量同步。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	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	导入开始前为清除部分数据时，显示该参数。 导入前根据条件删除目的表部分数据。	age > 18 and age <= 60
高级属性	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
	loader线程数	每个loader内部启动的线程数，可以提升写入并发数。并发场景下有如下限制：约束冲突处理策略不支持"replace into"或"on duplicate key update"。	1
	stream load 配置参数	stream load 配置参数。	max_filter_ratio=0

6.6.14 配置 HBase 目的端参数

表 6-42 HBase 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	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
	导入前清空数据	导入前是否清空表中的数据。 是：清空表中数据。 否：不清空。	否
高级属性	Rowkey冗余	是否将选做Rowkey的数据同时写入HBase的列。	否
	WAL开关	是否写WAL，不写WAL能提升性能，但如果HBase服务宕机可能会造成数据丢失。	是
	匹配数据类型	是否匹配类型，例如数据库的int类型列数据按照int类型转换为二进制写入HBase。	否

6.6.15 配置 MongoDB 目的端参数

表 6-43 MongoDB 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	数据库	输入或选择数据库名称。单击输入框后面的按钮可以进入数据库的选择界面。	default

参数类型	参数名	说明	取值样例
	集合名	写入数据的集合名，单击输入框后面的按钮可进入集合的选择界面。 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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
高级属性	迁移行为	选择写入目的端的迁移方式。 • 新增：将文件记录直接插入指定的集合。 • 有则替换，无则新增：以指定的过滤键作为查询条件。如果在集合中找到匹配的记录，则替换该记录。如果不存在，则添加新记录。 • 替换：使用指定的过滤键作为查询条件。如果在集合中找到匹配的记录，则替换该记录。如果没有，则不会添加新记录。	新增
	导入前准备语句	执行任务之前率先执行的SQL语句。目前仅允许执行一条SQL语句。	create table

6.6.16 配置 MRS Kafka 目的端参数

表 6-44 Kafka 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	Topic	输入Topic数据库名称。	default
	数据格式	写入目的端时使用的数据格式。 CSV：将列按照字段分隔符拼接。 JSON: 将所有列按照指定字段名称拼接为JSON字符串。	JSON格式
	字段分隔	数据格式为CSV显示该参数。 写入目的端时数据之间的字段分隔符。默认为空格。	,

参数类型	参数名	说明	取值样例
	keyIndex	数据格式为CSV显示该参数。 Kafka Writer中作为Key的那一列，填写后value不会记录此列。如字段列下标为0、1、2，keyIndex取值为0，则valueIndex为1、2。 keyIndex下标取值范围是从0开始的正整数，否则任务执行会报错。	-
	额外配置	数据格式为JSON显示该参数。 该参数指定不同的类型的控制写入数据格式或者指定配置参数。 使用该能力前必须配置参数configType，当前支持的值为COMBINE_DATA。 configType为COMBINE_DATA支持的搭配的参数如下: ● batchnum：将多条数据合并成一条，默认值为1。 ● featureTag：将每一条数据都打tag标签。 ● startEndMark：默认是为false。设置为true时，写入消息前将会同步一个开始消息和结束的消息。 ● columnAsKey：指定写入数据key值，也可以通过指定字段值作为key，通过配置@{column1}--@{column2}。 例如：目的端字段为id、name，需要使用这两个字段值，则配置成@{id}--@{name}。 ● schema：该参数会显示在写入的数据的消息体中，此处配置该参数时后续显示为设置的参数；如果没有配置，默认使用原表的schema值。 ● table：该参数会显示在写入的数据的消息体中，此处配置该参数时后续显示为设置的参数；如果没有配置，默认使用源端的表名。 ● acks：取值0，1、all。 ● jobId：默认值为0，页面配置该参数，则在消息里生成该字段的信息。	比如需要给数据打标签需配置两个参数，即configType：COMBINE_DATA，featureTag：group。

6.6.17 配置 GBase 目的端参数

表 6-45 GBase 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_EXAMPLE
	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，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TABLE_EXAMPLE
	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，作业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
高级属性	先导入阶段表	导入目的表之前是否把数据先导入阶段表，如果成功导入阶段表，则从阶段表导入到目的表，这样避免导入过程失败，在目的表遗留部分成功数据。 如果选择“是”，则启用事务模式迁移，作业会自动创建临时表，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，具体请参见**事务模式迁移**。 默认为“否”，作业直接将数据导入到目的表。这种情况下如果作业执行失败，已经导入到目标表中的数据不会自动回滚。 **说明** 如果“导入开始前”选择“清除部分数据”或“清除全部数据”，CDM的事务模式不会回滚已经删除的数据。	否

参数类型	参数名	说明	取值样例
	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

6.6.18 配置 Redis 目的端参数

表 6-46 Redis 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	redis键前缀	对应关系数据库的表名。	TABLENAME
	值存储类型	存储类型分STRING、hash、list、set和zset。	STRING
	写入前将相同的键删除	写入前是否将相同的键删除。	否
高级属性	键分隔符	用来分隔关系数据库的表和列名。	_
	值分隔符	以STRING方式存储，列之间的分隔符。 存储类型为列表时字串分割成数组的字符。	;
	key值有效期	设置统一的生存时间。单位：秒。	3600

6.6.19 配置 HDFS 目的端参数

表 6-47 HDFS 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	写入目录	写入数据到HDFS服务器的目录。	/user/cdm/output
	文件格式	传输数据时使用的格式。其中CSV和JSON仅支持迁移到数据表场景，二进制格式适用于文件迁移场景。	CSV格式

参数类型	参数名	说明	取值样例
	换行符处理方式	指定在写入文本文件表的数据包含换行符，特指(\n|\r|\r\n)的情况下处理换行符的策略。 ● 删除 ● 不处理 ● 替换为其他字符串	删除
	换行符替换字符串	当换行符处理方式选择为替换时，指定替换的字符串。	-
高级属性	写入到临时文件	文件格式为二进制格式时显示该参数。 将二进制文件先写入到临时文件。临时文件以".tmp"作为后缀。	否
	换行符	文件格式为CSV格式时显示该参数。 文件中的换行符，默认自动识别"\n"、"\r"或"\r\n"。手动配置特殊字符，如空格回车需使用URL编码后的值。或通过编辑作业json方式配置，无需URL编码。	\n
	字段分隔符	文件格式为CSV格式时显示该参数。 文件中的字段分隔符。配置特殊字符需先url编码。	,
	作业成功标识文件	标识文件名。 当作业成功时，在写入目录下生成标识文件。不输入文件名时不启用该功能。	finish.txt
	使用包围符	文件格式为CSV格式时显示该参数。 使用包围符来括住字符串值。包围符内的字段分隔符被视为字符串值的一部分，目前只支持"作为包围符。：	否
	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	否
	目录层次	自定义目录层次选择是时显示该参数。 指定文件的目录层次，支持时间宏（时间格式为yyyy/MM/dd）。源端为关系型数据库数据源时，目录层次为源端表名+自定义目录，其他场景下为自定义目录。	${dateformat(yyyy/MM/dd, -1, DAY)}
	文件名前缀	文件格式为CSV格式时显示该参数。 设置文件名前缀。 文件名格式：prefix-jobname-timestamp-index。	data

参数类型	参数名	说明	取值样例
	压缩格式	文件格式为CSV格式时显示该参数。 选择写入文件的压缩格式。 ● NONE ● DEFLATE ● GZIP ● BZIP2 ● SNAPPY	SNAPPY
	加密方式	文件格式为二进制格式时显示该参数。 对上传的数据进行加密。 ● 无 ● AES-256-GCM	无
	数据加密密钥	文件格式为二进制格式且选择加密方式时显示该参数。 数据加密密钥（Data Encryption Key），AES-256-GCM密钥由长度64的十六进制数组成。	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
	初始化向量	文件格式为二进制格式且选择加密方式时显示该参数。 设置初始化向量，由长度32的十六进制数组成。	5C91687BA886EDCD12ACBC3FF19A3C3F

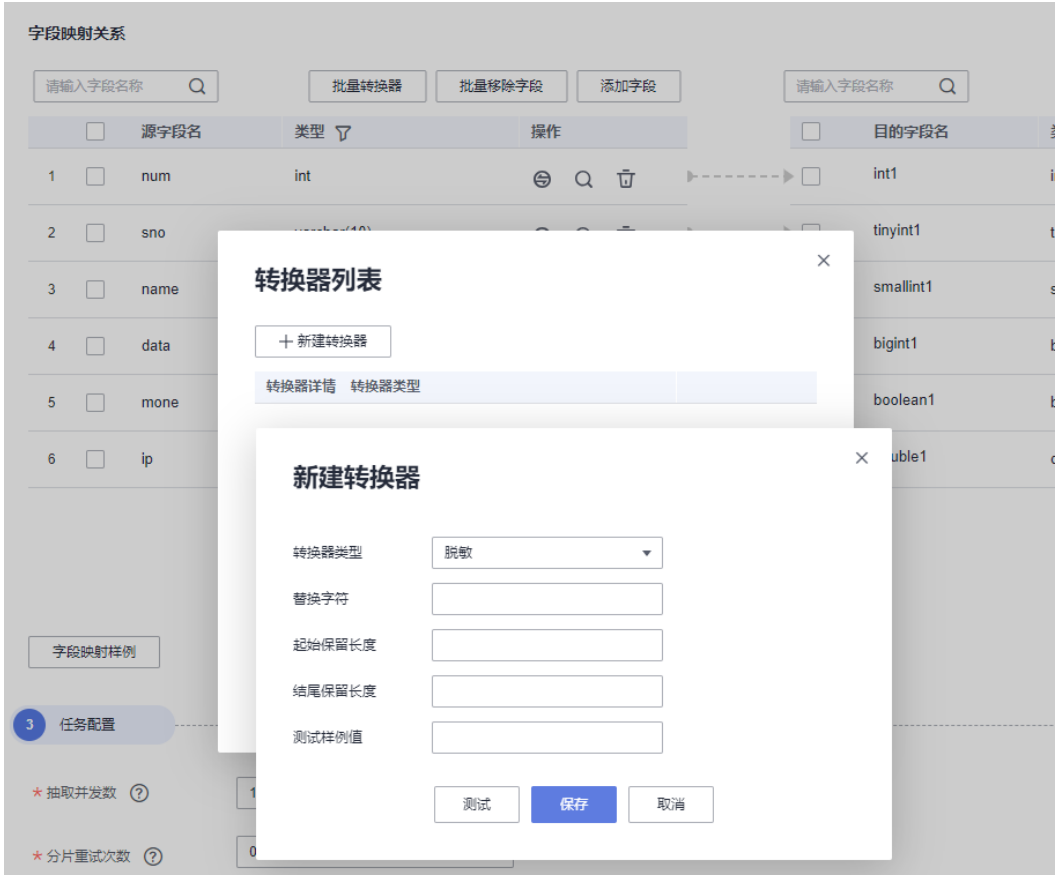
6.7 字段转换器配置指导

操作场景

- 作业参数配置完成后，将进行字段映射的配置，您可以单击操作列下 创建字段转换器。
- 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。

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6-6所示。

图 6-6 新建字段转换器



在迁移过程中可以对字段进行转换，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

约束限制

- 作业源端开启“使用SQL语句”参数时不支持配置转换器。
- 如果在字段映射界面，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。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datetime）。
- 当作业源端为OBS、迁移CSV文件时，并且配置“解析首行为列名”参数的场景下显示列名。

- 当使用二进制格式进行文件到文件的迁移时，没有配置字段转换器这一步。
- 自动创表场景下，需在目的端表中提前手动新增字段，再在字段映射里新增字段。
- 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，会将字段值直接写入目的端。
- 如果字段映射关系不正确，您可以通过拖拽字段、单击  对字段批量映射两种方式来调整字段映射关系。
- 创建表达式转换器时，表达式的功能是对该字段的数据进行处理，故不建议使用时间宏。
- 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：
  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。

脱敏

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

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

去前后空格

自动去字符串前后的空值，不需要配置参数。

字符串反转

自动反转字符串，例如将“ABC”转换为“CBA”，不需要配置参数。

字符串替换

替换字符串，需要用户配置被替换的对象，以及替换后的值。

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。

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

表达式：StringUtils.center(value,*4*)

o. 删除字符串末尾的一个换行符（包括“\n”、“\r”或者“\r\n”），例如将“abc\r\n\r\n”转换为“abc\r\n”。

表达式：StringUtils.chomp(value)

p. 如果字符串中包含指定的字符串，则返回布尔值true，否则返回false。例如“abc”中包含“a”，则返回true。

表达式：StringUtils.contains(value,"*a*")

q. 如果字符串中包含指定字符串的任一字符，则返回布尔值true，否则返回false。例如“zzabyycdxx”中包含“z”或“a”任意一个，则返回true。

表达式：StringUtils.containsAny(value,"*za*")

r. 如果字符串中不包含指定的所有字符，则返回布尔值true，包含任意一个字符则返回false。例如“abz”中包含“xyz”里的任意一个字符，则返回false。

表达式：StringUtils.containsNone(value,"*xyz*")

s. 如果当前字符串只包含指定字符串中的字符，则返回布尔值true，包含任意一个其它字符则返回false。例如“abab”只包含“abc”中的字符，则返回true。

表达式：StringUtils.containsOnly(value,"*abc*")

t. 如果字符串为空或null，则转换为指定的字符串，否则不转换。例如将空字符转换为null。

表达式：StringUtils.defaultIfEmpty(value,*null*)

u. 如果字符串以指定的后缀结尾（包括大小写），则返回布尔值true，否则返回false。例如“abcdef”后缀不为null，则返回false。

表达式：StringUtils.endsWith(value,*null*)

v. 如果字符串和指定的字符串完全一样（包括大小写），则返回布尔值true，否则返回false。例如比较字符串“abc”和“ABC”，则返回false。

表达式：StringUtils.equals(value,"*ABC*")

w. 从字符串中获取指定字符串的第一个索引，没有则返回整数-1。例如从“aabaabaa”中获取“ab”的第一个索引1。

表达式：StringUtils.indexOf(value,"*ab*")

x. 从字符串中获取指定字符串的最后一个索引，没有则返回整数-1。例如从“aFkyk”中获取“k”的最后一个索引4。

表达式：StringUtils.lastIndexOf(value,"*k*")

y. 从字符串中指定的位置往后查找，获取指定字符串的第一个索引，没有则转换为“-1”。例如“aabaabaa”中索引3的后面，第一个“b”的索引是5。

表达式：StringUtils.indexOf(value,"*b*",*3*)

z. 从字符串获取指定字符串中任一字符的第一个索引，没有则返回整数-1。例如从“zzabyycdxx”中获取“z”或“a”的第一个索引0。

表达式：StringUtils.indexOfAny(value,"*za*")

aa. 如果字符串仅包含Unicode字符，返回布尔值true，否则返回false。例如“ab2c”中包含非Unicode字符，返回false。

表达式：StringUtils.isAlpha(value)

ab. 如果字符串仅包含Unicode字符或数字，返回布尔值true，否则返回false。例如“ab2c”中仅包含Unicode字符和数字，返回true。

表达式：StringUtils.isAlphanumeric(value)

ac. 如果字符串仅包含Unicode字符、数字或空格，返回布尔值true，否则返回false。例如“ab2c”中仅包含Unicode字符和数字，返回true。

表达式：StringUtils.isAlphanumericSpace(value)

ad. 如果字符串仅包含Unicode字符或空格，返回布尔值true，否则返回false。例如“ab2c”中包含Unicode字符和数字，返回false。

表达式：StringUtils.isAlphaSpace(value)

ae. 如果字符串仅包含ASCII可打印字符，返回布尔值true，否则返回false。例如“!ab-c~”返回true。

表达式：StringUtils.isAsciiPrintable(value)

af. 如果字符串为空或null，返回布尔值true，否则返回false。

表达式：StringUtils.isEmpty(value)

ag. 如果字符串中仅包含Unicode数字，返回布尔值true，否则返回false。

表达式：StringUtils.isNumeric(value)

ah. 获取字符串最左端的指定长度的字符，例如获取“abc”最左端的2位字符“ab”。

表达式：StringUtils.left(value,*2*)

ai. 获取字符串最右端的指定长度的字符，例如获取“abc”最右端的2位字符“bc”。

表达式：StringUtils.right(value,*2*)

aj. 将指定字符串拼接至当前字符串的左侧，需同时指定拼接后的字符串长度，如果当前字符串长度不小于指定长度，则不转换。例如将“yz”拼接到“bat”左侧，拼接后长度为8，则转换后为“yzyzybat”。

表达式：StringUtils.leftPad(value,*8*,"*yz*")

ak. 将指定字符串拼接至当前字符串的右侧，需同时指定拼接后的字符串长度，如果当前字符串长度不小于指定长度，则不转换。例如将“yz”拼接到“bat”右侧，拼接后长度为8，则转换后为“batyzyzy”。

表达式：StringUtils.rightPad(value,*8*,"*yz*")

al. 如果当前字段为字符串类型，获取当前字符串的长度，如果该字符串为null，则返回0。

表达式：StringUtils.length(value)

am. 如果当前字段为字符串类型，删除其中所有的指定字符串，例如从“queued”中删除“ue”，转换后为“qd”。

表达式：StringUtils.remove(value,"*ue*")

an. 如果当前字段为字符串类型，移除当前字段末尾指定的子字符串。指定的子字符串若不在当前字段的末尾，则不转换，例如移除当前字段“www.domain.com”后的“.com”。

表达式：StringUtils.removeEnd(value,"*.com*")

ao. 如果当前字段为字符串类型，移除当前字段开头指定的子字符串。指定的子字符串若不在当前字段的开头，则不转换，例如移除当前字段“www.domain.com”前的“www.”。

表达式：StringUtils.removeStart(value,"*www.*")

ap. 如果当前字段为字符串类型，替换当前字段中所有的指定字符串，例如将“aba”中的“a”用“z”替换，转换后为“zbz”。

表达式：StringUtils.replace(value,"*a*","*z*")

aq. 如果当前字段为字符串类型，一次替换字符串中的多个字符，例如将字符串“hello”中的“h”用“j”替换，“o”用“y”替换，转换后为“jelly”。

表达式：StringUtils.replaceChars(value,"*ho*","*jy*")

ar. 如果字符串以指定的前缀开头（区分大小写），则返回布尔值true，否则返回false，例如当前字符串“abcdef”以“abc”开头，则返回true。

表达式：StringUtils.startsWith(value,"*abc*")

as. 如果当前字段为字符串类型，去除字段中首、尾处所有指定的字符，例如去除“abcyx”中首尾所有的“x”、“y”、“z”和“b”，转换后为“abc”。

表达式：StringUtils.strip(value,"*xyz*b")

at. 如果当前字段为字符串类型，去除字段末尾所有指定的字符，例如去除当前字段末尾的"abc"字符串。

表达式：StringUtils.stripEnd(value,*"abc"*)

au. 如果当前字段为字符串类型，去除字段开头所有指定的字符，例如去除当前字段开头的所有空格。

表达式：StringUtils.stripStart(value,*null*)

av. 如果当前字段为字符串类型，获取字符串指定位置后（索引从0开始，包括指定位置的字符）的子字符串，指定位置如果为负数，则从末尾往前计算位置，末尾第一位为-1。例如获取“abcde”第2个字符（即c）及之后的字符串，则转换后为“cde”。

表达式：StringUtils.substring(value,*2*)

aw. 如果当前字段为字符串类型，获取字符串指定区间（索引从0开始，区间起点包括指定位置的字符，区间终点不包含指定位置的字符）的子字符串，区间位置如果为负数，则从末尾往前计算位置，末尾第一位为-1。例如获取“abcde”第2个字符（即c）及之后、第4个字符（即e）之前的字符串，则转换后为“cd”。

表达式：StringUtils.substring(value,*2*,4)

ax. 如果当前字段为字符串类型，获取当前字段里第一个指定字符后的子字符串。例如获取“abcba”中第一个“b”之后的子字符串，转换后为“cba”。

表达式：StringUtils.substringAfter(value,"*b*")

ay. 如果当前字段为字符串类型，获取当前字段里最后一个指定字符后的子字符串。例如获取“abcba”中最后一个“b”之后的子字符串，转换后为“a”。

表达式：StringUtils.substringAfterLast(value,"*b*")

az. 如果当前字段为字符串类型，获取当前字段里第一个指定字符前的子字符串。例如获取“abcba”中第一个“b”之前的子字符串，转换后为“a”。

表达式：StringUtils.substringBefore(value,"*b*")

ba. 如果当前字段为字符串类型，获取当前字段里最后一个指定字符前的子字符串。例如获取“abcba”中最后一个“b”之前的子字符串，转换后为“abc”。

表达式：StringUtils.substringBeforeLast(value,"*b*")

bb. 如果当前字段为字符串类型，获取嵌套在指定字符串之间的子字符串，没有匹配的则返回null。例如获取“tagabctag”中“tag”之间的子字符串，转换后为“abc”。

表达式：StringUtils.substringBetween(value,"*tag*")

bc. 如果当前字段为字符串类型，删除当前字符串两端的控制字符（char≤32），例如删除字符串前后的空格。

表达式：StringUtils.trim(value)

bd. 将当前字符串转换为字节，如果转换失败，则返回0。

表达式：NumberUtils.toByte(value)

be. 将当前字符串转换为字节，如果转换失败，则返回指定值，例如指定值配置为1。

表达式：NumberUtils.toByte(value,*1*)

bf. 将当前字符串转换为Double数值，如果转换失败，则返回0.0d。

表达式：NumberUtils.toDouble(value)

bg. 将当前字符串转换为Double数值，如果转换失败，则返回指定值，例如指定值配置为1.1d。

表达式：NumberUtils.toDouble(value,*1.1d*)

bh. 将当前字符串转换为Float数值，如果转换失败，则返回0.0f。

表达式：NumberUtils.toFloat(value)

bi. 将当前字符串转换为Float数值，如果转换失败，则返回指定值，例如配置指定值为1.1f。

表达式：NumberUtils.toFloat(value,*1.1f*)

bj. 将当前字符串转换为Int数值，如果转换失败，则返回0。

表达式：NumberUtils.toInt(value)

bk. 将当前字符串转换为Int数值，如果转换失败，则返回指定值，例如配置指定值为1。

表达式：NumberUtils.toInt(value,*1*)

bl. 将字符串转换为Long数值，如果转换失败，则返回0。

表达式：NumberUtils.toLong(value)

bm. 将当前字符串转换为Long数值，如果转换失败，则返回指定值，例如配置指定值为1L。

表达式：NumberUtils.toLong(value,*1L*)

bn. 将字符串转换为Short数值，如果转换失败，则返回0。

表达式：NumberUtils.toShort(value)

bo. 将当前字符串转换为Short数值，如果转换失败，则返回指定值，例如配置指定值为1。

表达式：NumberUtils.toShort(value,*1*)

bp. 将当前IP字符串转换为Long数值，例如将“10.78.124.0”转换为LONG数值是“172915712”。

表达式：CommonUtils.ipToLong(value)

bq. 从网络读取一个IP与物理地址映射文件，并存放到Map集合，这里的URL是IP与地址映射文件存放地址，例如“http://10.114.205.45:21203/sqoop/IpList.csv”。

表达式：HttpsUtils.downloadMap("*url*")

br. 将IP与地址映射对象缓存起来并指定一个key值用于检索，例如“ipList”。

表达式：CommonUtils.setCache("*ipList*",HttpsUtils.downloadMap("*url*"))

bs. 取出缓存的IP与地址映射对象。

表达式：CommonUtils.getCache("*ipList*")

bt. 判断是否有IP与地址映射缓存。

表达式：CommonUtils.cacheExists("*ipList*")

bu. 根据指定的偏移类型（month/day/hour/minute/second）及偏移量（正数表示增加，负数表示减少），将指定格式的时间转换为一个新时间，例如将“2019-05-21 12:00:00”增加8个小时。

表达式：DateUtils.getCurrentTimeByZone("*yyyy-MM-dd HH:mm:ss*",value, "*hour*", *8*)

bv. 如果value值为空或者null时，则返回字符串"aaa"，否则返回value。

表达式：StringUtils.defaultIfEmpty(value,"aaa")

6.8 新增字段操作指导

操作场景

- 作业参数配置完成后，将进行字段映射的配置，您可以通过字段映射界面的可自定义新增字段。
- 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。

您可以单击字段映射界面的选择“添加新字段”自定义新增字段，通常用于标记数据库来源，以确保导入到目的端数据的完整性。

图 6-7 字段映射

目前支持以下类型自定义字段：

- **常量**

  常量参数即参数值是固定的参数，不需要重新配置值。例如“lable”=“friends”用来标识常量值。

- **变量**

  您可以使用时间宏、表名宏、版本宏等变量来标记数据库来源信息。变量的语法：${variable}，其中“variable”指的是变量。例如“input_time”=“${timestamp()}”用来标识当前时间的时间戳。

- **表达式**

  您可以使用表达式语言根据运行环境动态生成参数值。表达式的语法：#{expr}，其中“expr”指的是表达式。例如“time”=“#{DateUtil.now()}”用来标识当前日期字符串。

约束限制

- 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。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datetime）。
- 当作业源端为OBS、迁移CSV文件时，并且配置“解析首行为列名”参数的场景下显示列名。
- 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。
- 自动创表场景下，需在目的端表中提前手动新增字段，再在字段映射里新增字段。
- 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。
- 如果字段映射关系不正确，您可以通过拖拽字段、单击对字段批量映射两种方式来调整字段映射关系。
- 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：
  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CDM不支持源端迁移字段类型，请参见**不支持数据类型转换规避指导**将字段类型转换为CDM支持的类型。

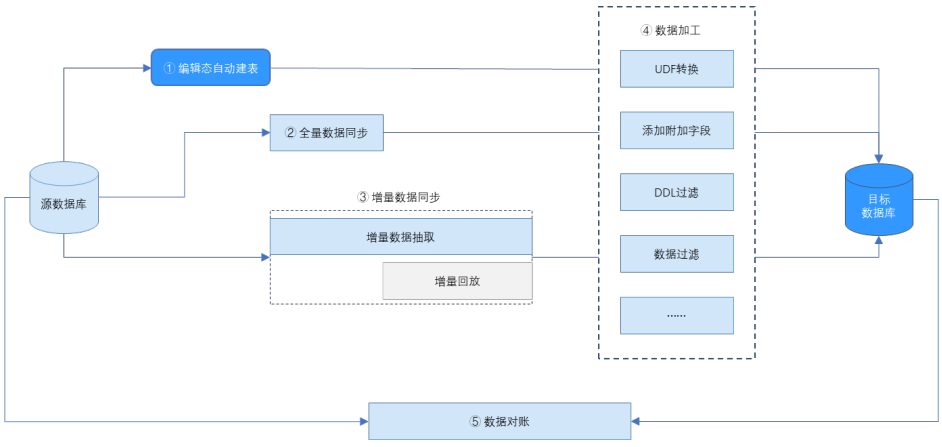
7 数据集成（实时作业）

7.1 实时作业概述

DataArts Studio的Migration服务提供了实时数据同步功能，可将数据通过同步技术从一个数据源复制到其他数据源，并保持一致，实现关键业务数据的实时流动。

- 常用场景：实时分析，报表系统，数仓环境等。
- 同步特点：实时同步功能聚焦于表和数据，并满足多种灵活性的需求，例如多对一、一对多，动态增减同步表，不同库表名之间同步数据等。

图 7-1 实时同步原理



说明

实时处理集成作业功能当前在北京四、上海一、广州、新加坡已上线（其他region后续会逐步放开，敬请期待！），但需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。

功能概述

实时集成作业支持多种数据源、多种场景下的实时数据同步，用户可根据自主需求，一次性全量加实时增量同步多个库表，功能总览如下图所示。

图 7-2 功能总览图

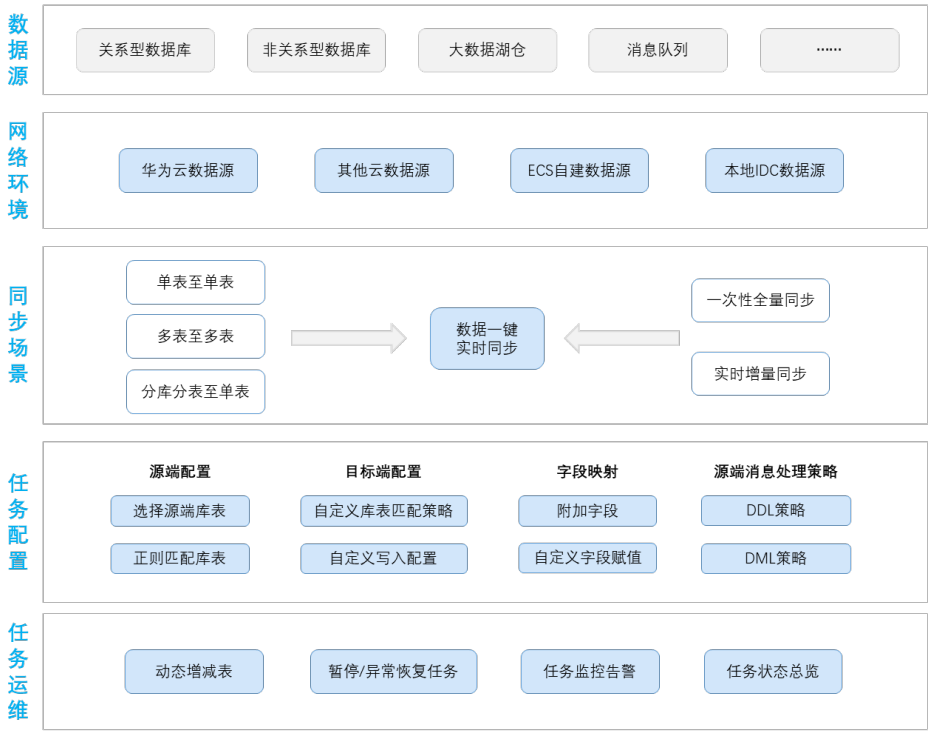


表 7-1 基本功能

功能	描述
多种数据源间的数据同步	支持多种数据源链路组合，您可以将多种输入及输出数据源搭配组成同步链路进行数据同步。详情请参见**支持的数据源**。
复杂网络环境下的数据同步	支持云数据库、本地IDC、ECS自建数据库等多种环境下的数据同步。在配置同步任务前，您可以根据数据库所在网络环境，选择合适的同步解决方案来确保数据集成资源组与您将同步的数据来源端与目标端网络环境已经连通，对应数据库环境与网络连通配置详情请参见：**网络打通**。
多类场景下的数据同步	支持单表、整库及分库分表实时增量数据同步。 • 单表同步：支持将源端一个实例下的单张表实时同步至目的端一个实例下的单张表。 • 整库同步：支持将源端一个实例下多个库的多张表批量实时同步到目的端一个实例下的多个库表，一个任务中最多支持200张目标表。 • 分库分表同步：支持将源端多个实例下多个分库的多张分表同步到目的端一个实例下的单个库表。

功能	描述
实时同步任务配置	支持通过简易的可视化配置完成实时数据同步。 • 数据源自定义参数配置。 • 图形化选择源端库表、正则匹配源端库表。 • 自定义源端与目的端库表匹配规则。 • 字段映射：附加字段、字段赋值（常量、变量、UDF）。 • 自动建表。 • 定义DDL消息处理策略。
实时同步任务运维	支持异常恢复、暂停恢复、动态增减表、配置告警、查看及导出任务日志等运维功能。

同步场景

Migration实时同步功能支持多种拓扑类型的同步场景，用户可根据自身需求进行规划，详细说明可参考以下内容。

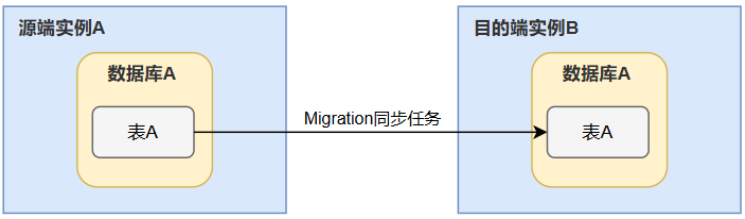
- **单表同步**

  支持将源端一个实例下的单张表实时同步至目的端一个实例下的单张表。

  单表同步支持以下链路：

  DMS Kafka > Hudi、DMS Kafka > OBS

  图 7-3 单表同步

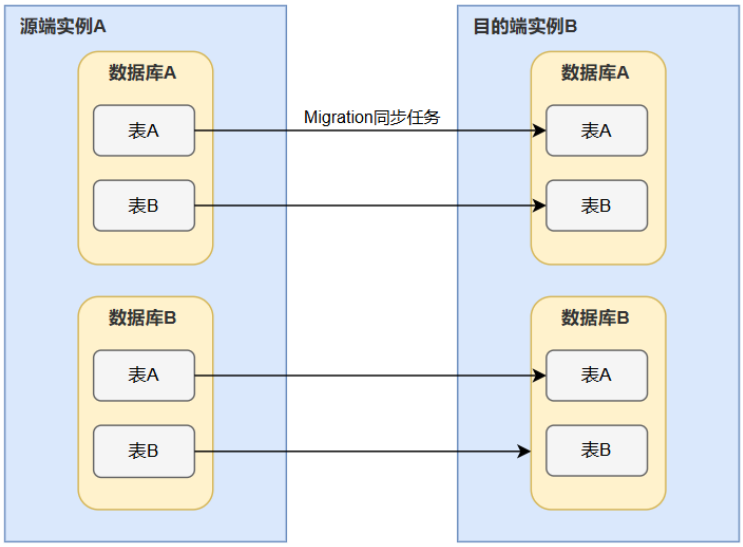
- **整库同步**

  支持将源端一个实例下多个库的多张表批量实时同步到目的端一个实例下的多个库表，一个任务中最多支持200张目标表。

  整库同步支持以下链路：

  - MySQL > MRS Hudi、MySQL > DWS、MySQL > Kafka
  - DMS Kafka > OBS 、Apache Kafka > MRS Kafka
  - SQLServer > MRS Hudi、SQLServer > DWS
  - PostgreSQL > DWS、PostgreSQL > MRS Hudi、PostgreSQL > DMS Kafka
  - Oracle > DWS、Oracle > MRS Hudi、Oracle > DMS Kafka
  - MongoDB > DWS
  - GaussDB集中式/分布式 > DWS、GaussDB集中式/分布式 > MRS Hudi、GaussDB集中式/分布式 > DMS Kafka

图 7-4 整库同步



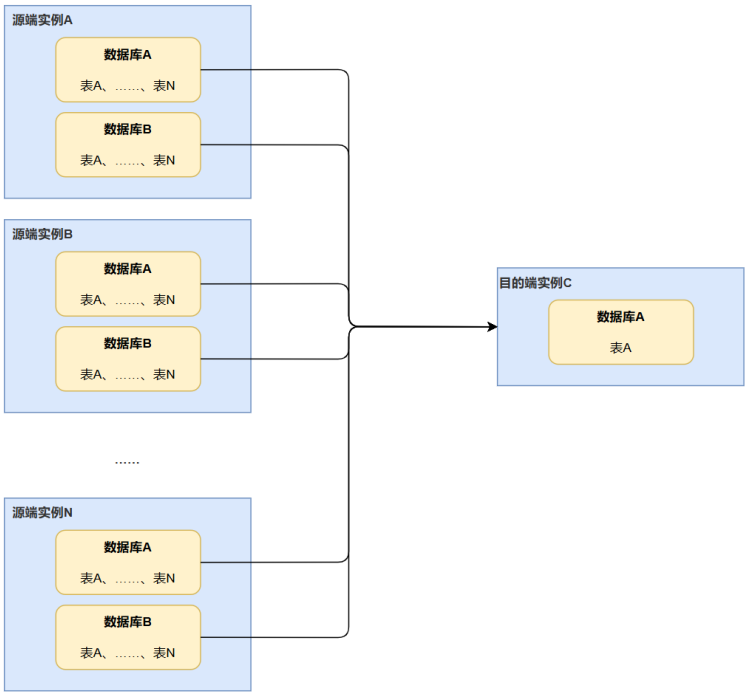
- **分库分表同步**

  支持将源端多个实例下多个分库的多张分表同步到目的端一个实例下的单个库表。

  分库分表同步支持以下链路：

  - MySQL > MRS Hudi、MySQL > DWS
  - PostgreSQL > DWS、MongoDB > DWS

图 7-5 分库分表同步



基本特性

实时数据集成为大数据开发提供了支撑，具有以下特性：

- 实时性：支持数据秒级同步。
- 可靠性：通过异常恢复，自动重试等多种机制确保数据的一致性和准确性。
- 多样性：
  - 数据源多样性：源端和目的端有多种数据源可供选择，为用户提供了多种选择。
  - 场景多样性：部分链路支持全量和增量同步，部分链路支持分库分表。
- 可维护性：支持作业监控和日志查看，方便运维人员进一步定位。
- 易用性：长界面更易操作，用户只需配置必要信息，学习成本减低。

7.2 支持的数据源

实时集成作业支持的数据源如表7-2所示。

表 7-2 实时集成作业支持的数据源

数据源分类	源端数据源	对应的目的端数据源	相关文档	说明
关系型数据	MySQL	Hadoop：MRS Hudi	MySQL同步到MRS Hudi作业配置	• MySQL数据库建议使用版本：5.6、5.7、8.x版本。 • Hudi建议使用版本：0.11.0。
		消息系统：DMS Kafka	MySQL同步到Kafka作业配置	• MySQL数据库建议使用版本：5.6、5.7、8.x版本。 • Kafka集群建议使用版本：2.7、3.x版本。
		数据仓库：DWS	MySQL同步到DWS作业配置	• MySQL数据库建议使用版本：5.6、5.7、8.x版本。 • DWS集群建议使用版本：8.1.3、8.2.0版本。

数据源分类	源端数据源	对应的目的端数据源	相关文档	说明
	SQLServer	Hadoop：MRS Hudi（公测中） **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	SQLServer同步到MRS Hudi作业配置	• SQLServer建议使用版本：企业版2016、2017、2019、2022版本，标准版2016 SP2及以上版本、2017、2019、2022版本。 • Hudi建议使用版本：0.11.0。
	PostgreSQL	数据仓库：DWS（公测中） **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	PostgreSQL同步到DWS作业配置	• PostgreSQL数据库建议使用版本：PostgreSQL 9.4、9.5、9.6、10、11、12、13、14版本。 • DWS集群建议使用版本：8.1.3、8.2.0版本。
	Oracle	数据仓库：DWS（公测中） **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	Oracle同步到DWS作业配置	• Oracle数据库建议使用版本：10、11、12、19版本。 • DWS集群建议使用版本：8.1.3、8.2.0版本。
		Hadoop：MRS Hudi（公测中） **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	Oracle同步到MRS Hudi作业配置	• Oracle数据库建议使用版本：10、11、12、19版本。 • Hudi建议使用版本：0.11.0。

数据源分类	源端数据源	对应的目的端数据源	相关文档	说明
消息系统	DMS Kafka	对象存储：OBS	DMS Kafka同步到OBS作业配置	Kafka集群建议使用版本：2.7、3.x版本。
	Apache Kafka	Hadoop：MRS Kafka（公测中） **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	Apache Kafka同步到MRS Kafka作业配置	Kafka集群建议使用版本：2.7、3.x版本。

7.3 使用前自检概览

当您在使用Migration服务创建实时同步任务前，需要预先检查是否做好了准备工作，以满足实时同步任务的环境要求。

表 7-3 自检项

自检项	说明	需要执行的准备工作
华为云账号及权限准备	准备华为账号，创建用户并授权使用Migration。	参考注册华为账号并开通华为云。 参考授权使用实时数据集成。
实时计算资源组准备	购买实时集成任务使用的计算资源，并关联到要使用的DataArts Studio工作空间。	参考购买数据集成资源组增量包。 参考实时集成资源组关联工作空间。
数据库准备	连接源和目标数据库以及对应连接账号权限准备。 **说明** • 建议创建单独用于Migration任务连接的数据库账号，避免因为账号修改导致的任务连接失败。 • 连接源和目标数据库的账号密码修改后，请尽快修改Migration任务中的连接信息，避免任务连接失败后的自动重试导致数据库账号被锁定，影响使用。	不同链路、数据库、权限要求不同，可参考以下链接，选择对应链路查看使用须知：使用教程。

自检项	说明	需要执行的准备工作
连接准备	准备DataArts Studio管理中心数据连接。 **说明** • 数据连接配置中必须勾选数据集成选项。 • 数据连接中使用的Agent代理实际为CDM集群，所用集群建议升级至较新版本（24.4.0B030版本以上），以满足功能特性需求，详情请联系客服或技术支持人员。	参考**创建DataArts Studio数据连接**。
网络准备	数据库部署在本地IDC	参考**数据库部署在本地IDC**进行网络准备。
	数据库部署在其他云	参考**数据库部署在其他云**进行网络准备。
	数据库部署在华为云	参考**数据库部署在华为云**进行网络准备。

7.4 网络打通

在配置实时同步任务前，您需要确保源端和目的端的数据库与运行实时同步任务的实时计算资源组之间网络连通，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。

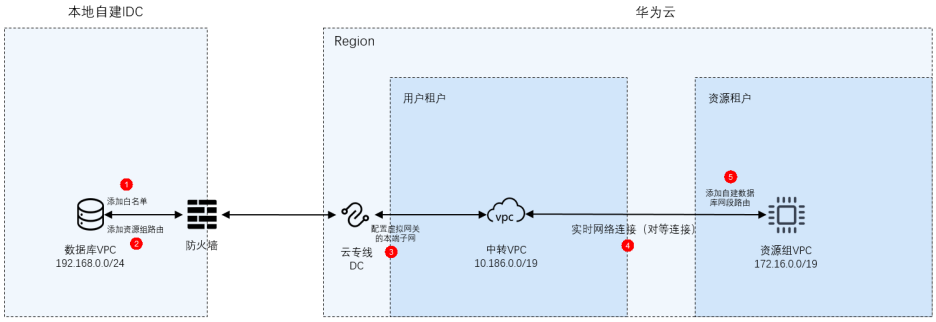
7.4.1 数据库部署在本地 IDC

7.4.1.1 通过云专线连通网络

在配置实时同步任务前，您需要确保源端和目的端的数据库与运行实时同步任务的实时计算资源组之间网络连通，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。

本章节主要为您介绍数据库部署在本地IDC场景下，通过云专线打通网络的方案。

图 7-6 网络示意图



约束限制

- 资源组为私网网段，不能与数据源网段重叠，否则会导致网络无法打通。
- 资源组不具有公网网段，因此本方案仅能与数据源的私网连通。

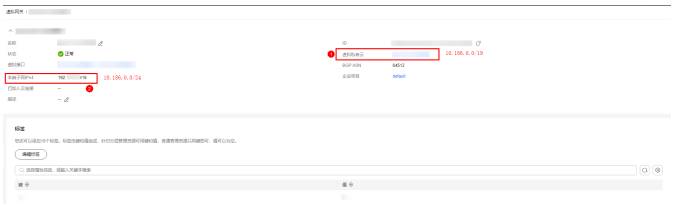
前提条件

- 已购买资源组，详情请参见**购买数据集成资源组**。
- 已购买并配置云专线，与云上的至少一个虚拟私有云VPC连通。若未开通云专线请参考**通过云专线实现云下IDC访问云上VPC**进行配置。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-4 资源网段规划

资源名称	说明	私网网段示例
数据源网段	本地IDC的数据源所属私网网段，请用户根据实际情况自行获取。	192.168.0.0/24
中转VPC及其子网	用于连通数据源和资源组网络的中间桥梁，本方案中需要使用云专线虚拟网关所配置的虚拟私有云和对应配置的子网。 **说明** 查看方式：登录云专线控制台，在左侧导航栏，选择“云专线 > 虚拟网关”，在列表中找到连通本地IDC所使用的虚拟网关，单击虚拟网关名称，查看关联的虚拟私有云和本端子网。 **图 7-7** 查看虚拟网关 	VPC：10.186.0.0/19 子网：10.186.0.0/24

资源名称	说明	私网网段示例
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 **说明** 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-8** 查询资源组网段  	172.16.0.0/19

网络配置流程

步骤1 本地IDC自建数据库添加白名单。

本地IDC自建数据库需要添加Migration资源组VPC网段（例如172.16.0.0/19）访问数据库的权限。各类型数据库添加白名单的方法不同，具体方法请参考各数据库官方文档进行操作。

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤2 （可选）本地IDC添加路由。

本地IDC需要添加路由，目的地址指向Migration资源组VPC网段（例如172.16.0.0/19），导向华为侧网关，添加路由可参考**配置本地路由**。

步骤3 云专线本端子网添加资源组网段。

为了允许云专线访问资源组网段，需登录云专线控制台，在左侧导航栏，选择“云专线 > 虚拟网关”，在列表中找到连通本地IDC所使用的虚拟网关，单击右侧操作栏中的“修改”按钮，在弹出框中的“本端子网”输入框里添加Migration资源组VPC网段（例如172.16.0.0/19）。

图 7-9 添加本端子网

步骤4 创建Migration实时网络连接（对等连接）。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-10 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-5 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多”>“绑定资源组”进行选择。
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。

参数	说明
子网	中转VPC的子网（例如10.186.0.0/24）。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤5 实时网络连接（对等连接）添加数据源网段路由。

单击步骤4所创建实时网络连接的“路由信息”，单击“添加路由”，输入本地IDC自建数据库的私有网络地址（例如192.168.0.0/24）。

图 7-11 添加路由 1

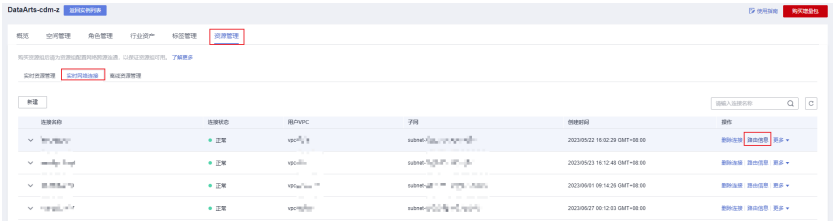


图 7-12 添加路由 2

步骤6 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

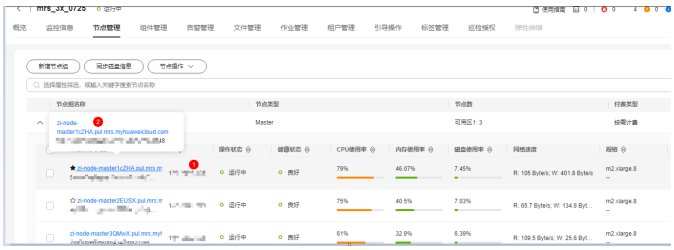
图 7-13 修改主机信息

说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。
  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-14 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤7 测试网络连接。

在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

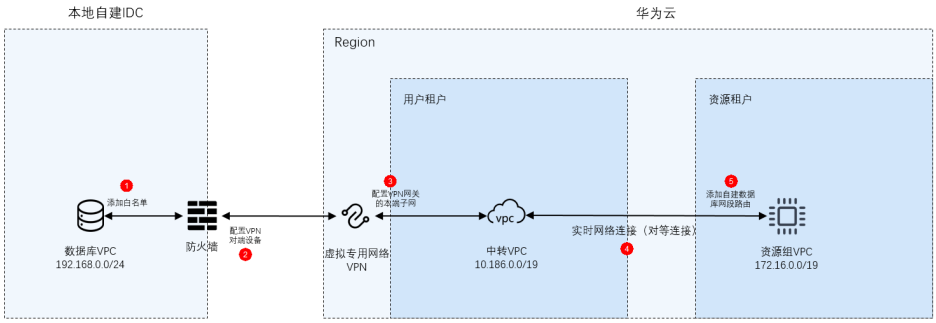
----结束

7.4.1.2 通过 VPN 连通网络

在配置实时同步任务前，您需要确保源端和目的端的数据库与运行实时同步任务的实时计算资源组之间网络连通，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。

本章节主要为您介绍数据库部署在本地IDC场景下，通过VPN打通网络的方案。

图 7-15 网络示意图



约束限制

- 资源组为私网网段，不能与数据源网段重叠，否则会导致网络无法打通。
- 资源组不具有公网网段，因此本方案仅能与数据源的私网连通。

前提条件

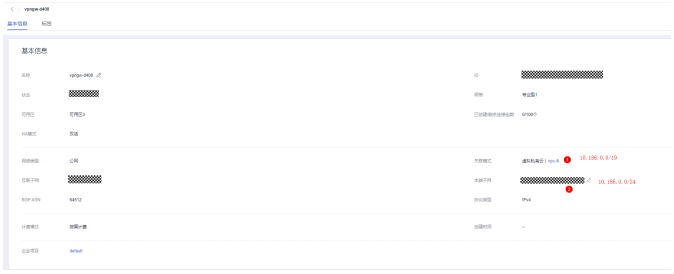
- 已购买资源组，详情请参见**购买数据集成资源组**。

- 已购买并配置虚拟专用网络VPN，与云上的至少一个虚拟私有云VPC连通。若未开通虚拟专用网络VPN请参考**经典版VPN购买流程**进行配置。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-6 资源网段规划

资源名称	说明	私网网段示例
数据源网段	本地IDC的数据源所属私网网段，请用户根据实际情况自行获取。	192.168.0.0/24
中转VPC及其子网	用于连通数据源和资源组网络的中间桥梁，本方案中需要使用虚拟专用网络VPN网关所配置的虚拟私有云和对应的子网。 查看方式： 登录虚拟专用网络控制台，在左侧导航栏，选择“虚拟专用网络 > VPN网关”，在列表中找到连通本地IDC所使用的VPN网关，单击VPN网关名称，查看关联的虚拟私有云和本端子网。 **图 7-16** 查看虚拟网关 	VPC：10.186.0.0/19 子网：10.186.0.0/24

资源名称	说明	私网网段示例
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-17** 查询资源组网段  	172.16.0.0/19

网络配置流程

步骤1 本地IDC自建数据库添加白名单。

本地IDC自建数据库需要添加Migration资源组VPC网段（例如172.16.0.0/19）访问数据库的权限。各类型数据库添加白名单的方法不同，具体方法请参考各数据库官方文档进行操作。

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

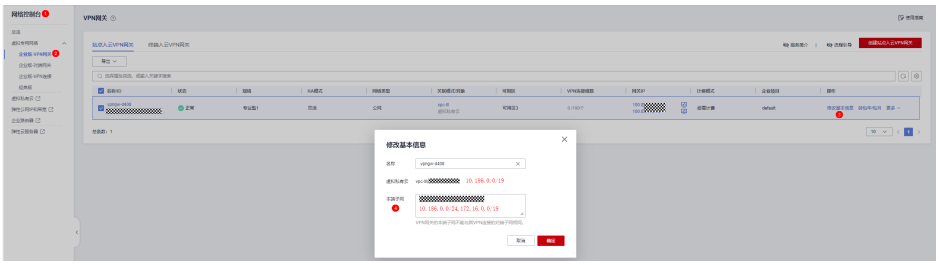
步骤2 （可选）本地IDC配置VPN对端网关设备。

本地IDC网络采用不同类型的防火墙或主机，可参考《虚拟专用网络快速入门》中的**配置对端设备**章节实现本地IDC数据库所在网络和华为云Migration资源组VPC网段（例如172.16.0.0/19）的互通。

步骤3 VPN本端子网添加资源组网段。

为了允许VPN访问资源组网段，请登录虚拟专用网络控制台，在左侧导航栏，选择“虚拟专用网络 > VPN网关”，在列表中找到连通本地IDC所使用的VPN网关，单击右侧操作栏中的“修改基本信息”按钮，在弹出框中的“本端子网”输入框里添加Migration资源组VPC网段（例如172.16.0.0/19）。

图 7-18 添加本端子网



步骤4 创建Migration实时网络连接（对等连接）。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-19 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-7 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多”>“绑定资源组”进行选择。
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。
子网	中转VPC的子网（例如10.186.0.0/24）。

参数	说明
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤5 实时网络连接（对等连接）添加数据源网段路由。

单击步骤4所创建实时网络连接的“路由信息”，单击“添加路由”，输入本地IDC自建数据库的私有网络地址（例如192.168.0.0/24）。

图 7-20 添加路由 1

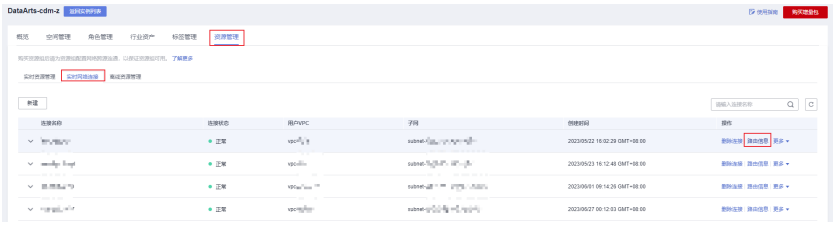


图 7-21 添加路由 2

步骤6 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

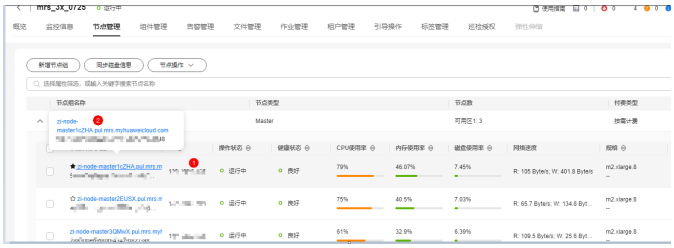
图 7-22 修改主机信息

说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。
  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-23 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤7 测试网络连接。

在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

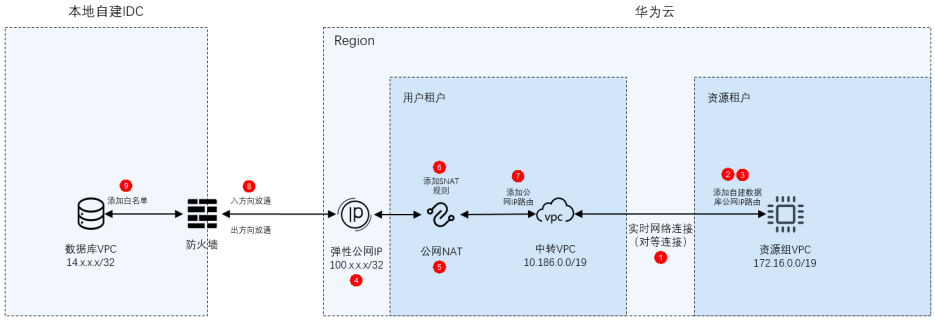
----结束

7.4.1.3 通过公网连通网络

在配置实时同步任务前，您需要确保源端和目的端的数据库与运行实时同步任务的实时计算资源组之间网络连通，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。

本章节主要为您介绍数据库部署在本地IDC场景下，通过公网打通网络的方案。

图 7-24 网络示意图



约束限制

资源组不具有公网网段，只能通过公网NAT转换成固定的弹性公网IP访问公网，且该IP不能与数据源公网IP重复。

前提条件

已购买资源组，详情请参见**购买数据集成资源组**。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-8 资源网段规划

资源名称	说明	私网网段示例
数据源公网IP	本地IDC数据源的公网IP，请用户根据实际情况自行获取。	14.x.x.x/32
弹性公网IP	资源组不具有公网网段，只能通过公网NAT转换成固定的弹性公网IP以访问公网。若未开通弹性公网IP，请登录弹性公网IP控制台，单击“购买弹性公网IP”，参考**通过VPC和EIP快速搭建可访问公网的网络**进行配置。	100.x.x.x/32
中转VPC及其子网	用于连通数据源和资源组网络的中间桥梁，本方案中需要使用当前租户下的一个虚拟私有云。若未开通VPC请参考**创建虚拟私有云**进行配置。	VPC：10.186.0.0/19 子网：10.186.0.0/24

资源名称	说明	私网网段示例
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-25** 查询资源组网段  	172.16.0.0/19

网络配置流程

步骤1 创建Migration实时网络连接（对等连接）。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-26 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-9 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多”>“绑定资源组”进行选择。
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。
子网	中转VPC的子网（例如10.186.0.0/24）。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤2 实时网络连接（对等连接）添加数据源网段路由。

单击步骤1所创建实时网络连接的“路由信息”，单击“添加路由”，输入本地IDC自建数据库的公网IP（例如14.x.x.x/32）。

图 7-27 添加路由 1

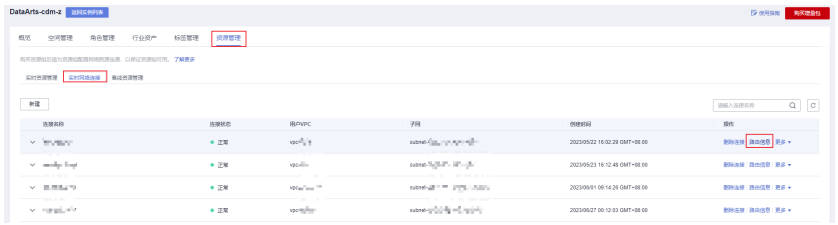
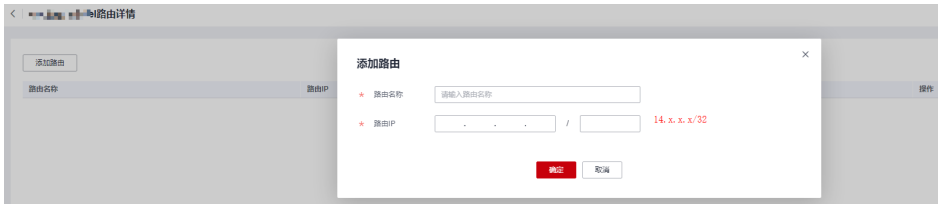


图 7-28 添加路由 2



步骤3 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

图 7-29 修改主机信息

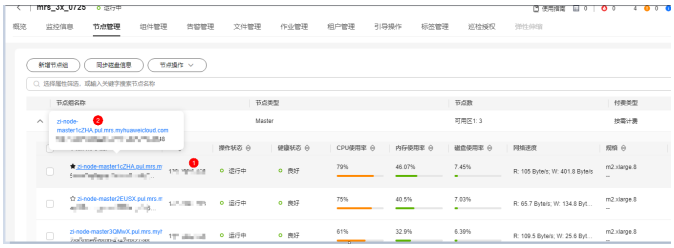
说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。

  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-30 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤4 购买弹性公网IP。

登录弹性公网IP控制台，单击“购买弹性公网IP”，根据界面提示配置参数，详情请参考**通过VPC和EIP快速搭建可访问公网的网络**。

步骤5 新建公网NAT网关。

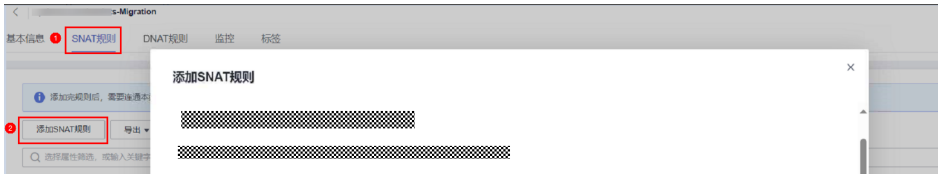
1. 登录NAT网关控制台，在左侧导航栏中选择“NAT网关 > 公网NAT网关”，单击“购买公网NAT网关”。
2. 配置NAT网关时，区域选择Migration所在Region，虚拟私有云选择中转VPC（例如10.186.0.0/19），子网选择中转VPC的子网（例如10.186.0.0/24），其余参数可参考**购买公网NAT网关**。

图 7-31 配置公网 NAT 网关

步骤6 为公网NAT网关添加SNAT规则。

为新建的NAT网关添加SNAT规则，才能实现资源组网段下的主机与Internet互相访问。请单击新建的公网NAT网关名称进入配置界面，选择“SNAT规则”页签，单击“添加SNAT规则”。

图 7-32 添加 SNAT 规则 1



其中，使用场景选择“云专线/云连接”，输入资源组VPC网段（例如172.16.0.0/19），然后绑定步骤3所购买的弹性公网IP（100.x.x.x/32）。

图 7-33 添加 SNAT 规则 2

步骤7 中转VPC子网网络添加路由。

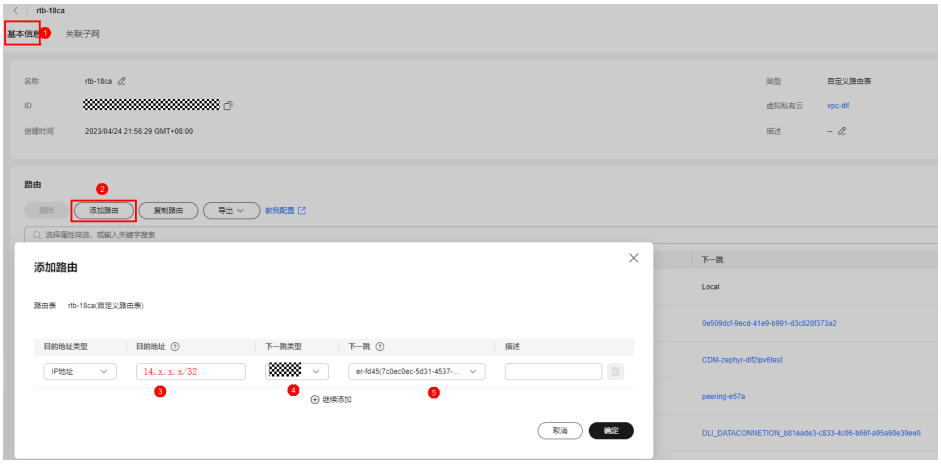
中转VPC子网的路由表中需要添加路由，目的地址指向本地IDC数据库的公网IP（例如14.x.x.x/32），下一跳跳至上面配置的NAT网关。

1. 登录虚拟私有云控制台，在左侧导航栏选择“虚拟私有云 > 子网”，找到中转VPC的子网并单击对应路由表名称进入配置界面。

图 7-34 查找路由表

2. 然后在路由表界面中选择“基本信息”页签，单击“添加路由”，目的地址指向本地IDC数据库的公网IP（例如14.x.x.x/32），下一跳跳至上面配置的NAT网关。

图 7-35 路由表添加路由



步骤8 本地IDC的防火墙设置。

本地IDC的防火墙需要放通弹性公网IP（例如100.x.x.x/32）的访问，使得Migration可以正常访问本地IDC自建数据库。

- 入方向放行：放通弹性公网IP到数据库监听端口的访问。
- 出方向放行：放通数据库监听端口到弹性公网IP的数据传输。

步骤9 本地IDC自建数据库添加白名单。

本地IDC自建数据库需要添加弹性公网IP（例如100.x.x.x/32）访问数据库的权限。各类型数据库添加白名单的方法不同，具体方法请参考各数据库官方文档进行操作。

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤10 测试网络连接。

在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

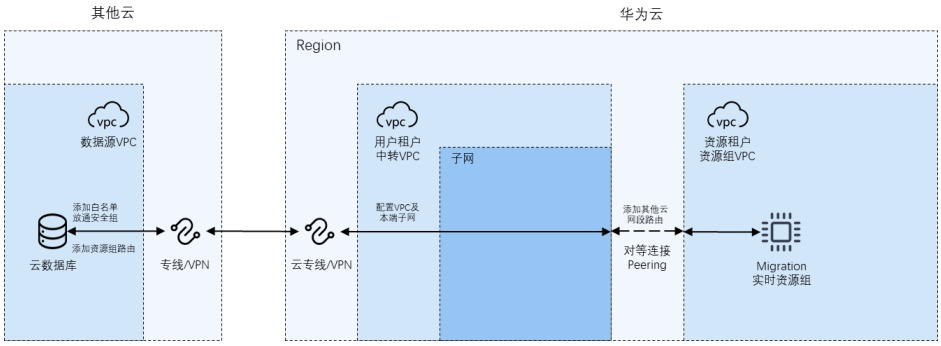
----结束

7.4.2 数据库部署在其他云

在配置实时同步任务前，您需要确保源端和目的端的数据库与运行实时同步任务的实时计算资源组之间网络连通，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。

本章节主要为您介绍数据库部署在其他云厂商场景下的网络打通方案。

图 7-36 网络示意图



约束限制

- 资源组为私网网段，不能与数据源网段重叠，否则会导致网络无法打通。
- 资源组不具有公网网段，因此本方案仅能与数据源的私网连通。

前提条件

已购买资源组，详情请参见**购买数据集成资源组**。

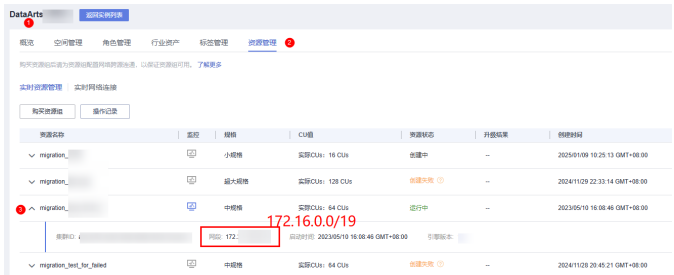
网络配置流程

步骤1 查询实时资源组VPC网段。

Migration实时资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，不占用用户账户的VPC网段。

可以登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。

图 7-37 查询资源组网段



步骤2 创建并配置中转VPC及其子网。

在本用户账户下创建虚拟私有云和子网，作为中转VPC，详情请参见**创建虚拟私有云和子网**。如当前账户已有可用VPC，可以跳过本步骤。

步骤3 在华为云购买并配置云专线或VPN虚拟专用网络。

为了连通其他云计算环境与华为云计算环境，可以通过开通云专线或虚拟专用网络来实现。

- 购买和配置云专线DC的相关操作，可以参考**通过云专线实现云下IDC访问云上VPC**。其中在创建虚拟网关时，虚拟私有云选择步骤2所创建的中转VPC，本端子网除了需要添加中转VPC的子网之外，还需要添加实时资源组的VPC网段。

- 购买和配置虚拟专用网络VPN的相关操作，可以参考**通过企业版站点入云VPN实现数据中心和VPC互通**。其中在创建VPN网关时，虚拟私有云选择步骤2所创建的中转VPC，本端子网除了需要添加中转VPC的子网之外，还需要添加实时资源组的VPC网段。

步骤4 在其他云购买专线或VPN虚拟专用网络

具体操作请参考其他云对应官网资料进行专线或VPN购买和对接。

步骤5 其他云数据库所在网络添加路由。

其他云数据库所属网络的路由表中需要添加路由，目的地址指向Migration资源组VPC网段，下一跳跳至步骤3中创建的云专线物理连接或VPN对端网关设备。

步骤6 其他云数据库添加白名单及安全组规则。

- 其他云数据库需要添加Migration资源组VPC网段访问数据库的权限。各厂商云数据库添加白名单的方法不同，请参考各数据库官方文档进行操作。
- 同时，其他云数据库若配置了安全组，则还需要增加入方向规则，放通Migration资源组VPC网段，使其可以访问数据库监听端口。

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤7 创建Migration实时网络连接。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-38 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-10 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多”>“绑定资源组”进行选择。
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC和子网。
子网	中转VPC的子网。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤8 为实时网络连接（对等连接）添加数据源网段路由。

单击步骤7所创建实时网络连接的“路由信息”，单击“添加路由”，输入本地IDC自建数据库的私有网络地址。

图 7-39 添加路由 1

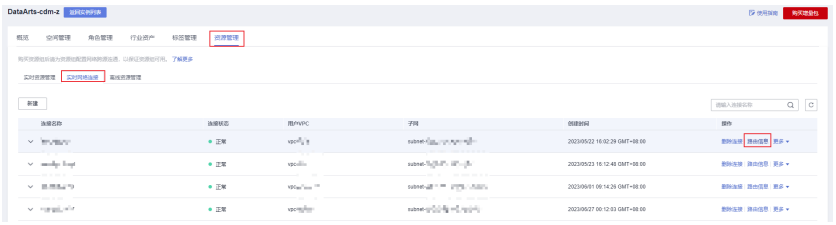


图 7-40 添加路由 2

步骤9 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

图 7-41 修改主机信息

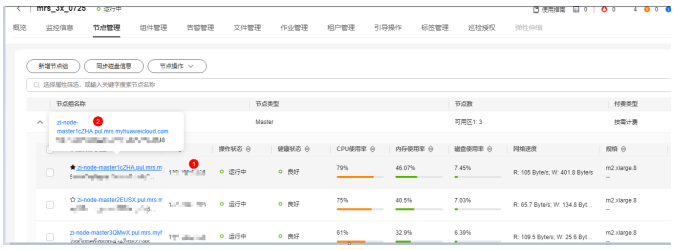
说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。

  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-42 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤10 测试网络连接。

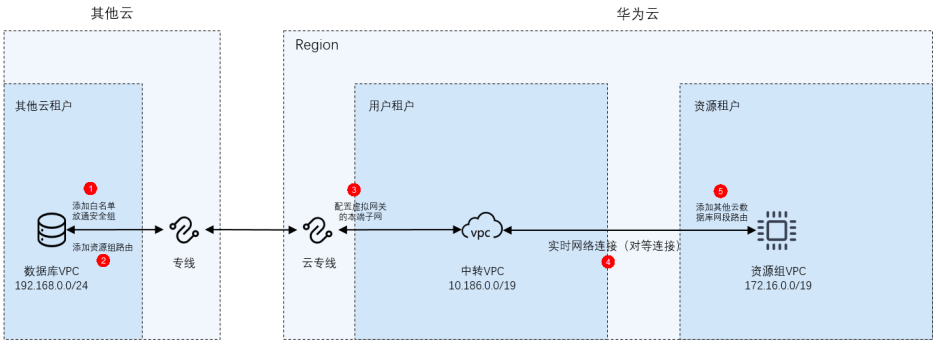
在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

----结束

7.4.2.1 通过云专线连通网络

本章节主要为您介绍数据库部署在其他云厂商场景下，通过云专线打通网络的方案。

图 7-43 网络示意图



约束限制

- 资源组为私网网段，不能与数据源网段重叠，否则会导致网络无法打通。
- 资源组不具有公网网段，因此本方案仅能与数据源的私网连通。

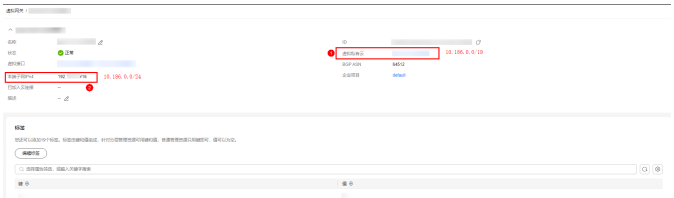
前提条件

- 已购买资源组，详情请参见**购买数据集成资源组**。
- 已购买并配置云专线，与云上的至少一个虚拟私有云VPC连通。若未开通云专线请参考**通过云专线实现云下IDC访问云上VPC**和其他云对应官网资料进行配置。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-11 资源网段规划

资源名称	说明	私网网段示例
数据源网段	其他云上数据源所属私网网段，请用户根据实际情况自行获取。	192.168.0.0/24
中转VPC及其子网	用于连通数据源和资源组网络的中间桥梁，本方案中需要使用云专线虚拟网关所配置的虚拟私有云和对应配置的子网。 查看方式： 登录云专线控制台，在左侧导航栏，选择“云专线 > 虚拟网关”，在列表中找到连通其他云所使用的虚拟网关，单击虚拟网关名称，查看关联的虚拟私有云和本端子网。 **图 7-44** 查看虚拟网关 	VPC：10.186.0.0/19 子网：10.186.0.0/24

资源名称	说明	私网网段示例
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-45** 查询资源组网段  	172.16.0.0/19

网络配置流程

步骤1 其他云数据库添加白名单及安全组规则。

- 其他云数据库需要添加Migration资源组VPC网段（例如172.16.0.0/19）访问数据库的权限。各类型数据库添加白名单的方法不同，具体方法请参考各数据库官方文档进行操作。
- 数据库若配置了安全组，则还需要增加入方向规则，放通Migration资源组VPC网段（例如172.16.0.0/19），使其可以访问数据库监听端口。

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤2 （可选）其他云数据库所在网络添加路由，专线添加远端子网。

- 必要时，其他云数据库所属网络需要添加路由，目的地址指向Migration资源组VPC网段（例如172.16.0.0/19），下一跳跳至云专线的物理连接，建议参考其他云官网资料进行配置。
- 必要时，其他云的专线需要将Migration资源组VPC网段（例如172.16.0.0/19添加到专线的远端子网中，确保路由通畅，建议参考其他云官网资料进行配置。

步骤3 云专线本端子网添加资源组网段。

为了允许云专线访问资源组网段，请登录云专线控制台，在左侧导航栏，选择“云专线 > 虚拟网关”，在列表中找到连通其他云所使用的虚拟网关，单击右侧操作栏中的

“修改”按钮，在弹出框中的“本端子网”输入框里添加Migration资源组VPC网段（例如172.16.0.0/19）。

图 7-46 添加本端子网

步骤4 创建Migration实时网络连接（对等连接）。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-47 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-12 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多” >“绑定资源组”进行选择。

参数	说明
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。
子网	中转VPC的子网（例如10.186.0.0/24）。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤5 实时网络连接（对等连接）添加数据源网段路由。

单击步骤4所创建实时网络连接的“路由信息”，单击“添加路由”，输入本地IDC自建数据库的私有网络地址（例如192.168.0.0/24）。

图 7-48 添加路由 1

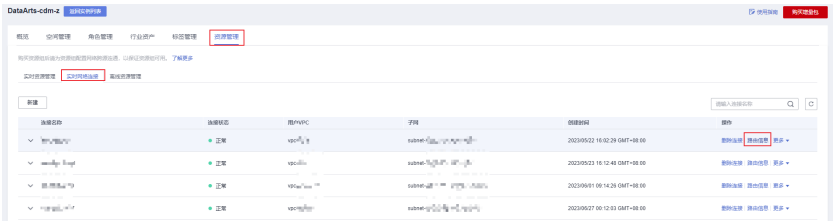


图 7-49 添加路由 2

步骤6 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

图 7-50 修改主机信息

说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。
  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-51 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤7 测试网络连接。

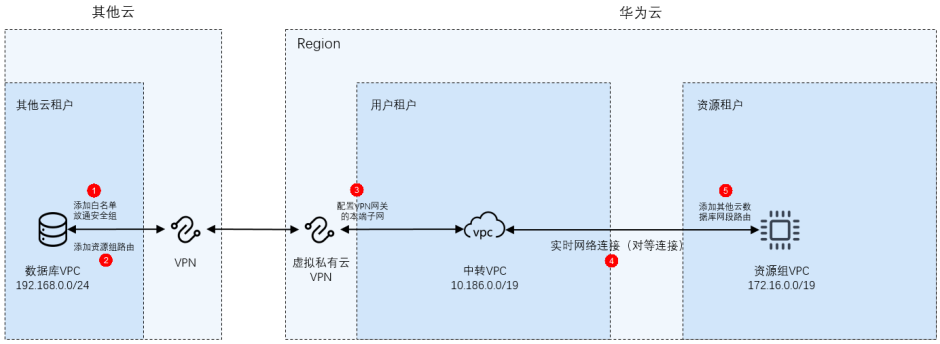
在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

----结束

7.4.2.2 通过 VPN 连通网络

本章节主要为您介绍数据库部署在其他云厂商场景下的网络打通方案。

图 7-52 网络示意图



约束限制

- 资源组为私网网段，不能与数据源网段重叠，否则会导致网络无法打通。
- 资源组不具有公网网段，因此本方案仅能与数据源的私网连通。

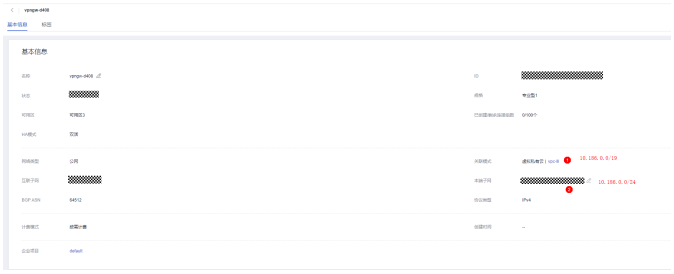
前提条件

- 已购买资源组，详情请参见**购买数据集成资源组**。
- 已购买并配置虚拟专用网络VPN，与云上的至少一个虚拟私有云VPC连通。若未开通虚拟专用网络VPN请参考**通过企业版站点入云VPN实现数据中心和VPC互通**进行配置。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-13 资源网段规划

资源名称	说明	私网网段示例
数据源网段	其他云上数据源所属私网网段，请用户根据实际情况自行获取。	192.168.0.0/24
中转VPC及其子网	用于连通数据源和资源组网络的中间桥梁，本方案中需要使用虚拟专用网络VPN网关所配置的虚拟私有云和对应的子网。 查看方式： 登录虚拟专用网络控制台，在左侧导航栏，选择“虚拟专用网络 > VPN网关”，在列表中找到连通其他云所使用的VPN网关，单击VPN网关名称，查看关联的虚拟私有云和本端子网。 **图 7-53** 查看虚拟网关 	VPC：10.186.0.0/19 子网：10.186.0.0/24

资源名称	说明	私网网段示例
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-54** 查询资源组网段  	172.16.0.0/19

网络配置流程

步骤1 其他云数据库添加白名单及安全组规则。

- 其他云数据库需要添加Migration资源组VPC网段（例如172.16.0.0/19）访问数据库的权限。各类型数据库添加白名单的方法不同，具体方法请参考各数据库官方文档进行操作。
- 数据库若配置了安全组，则还需要增加入方向规则，放通Migration资源组VPC网段（例如172.16.0.0/19），使其可以访问数据库监听端口。

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤2 （可选）其他云数据库所在网络及VPN网关添加路由。

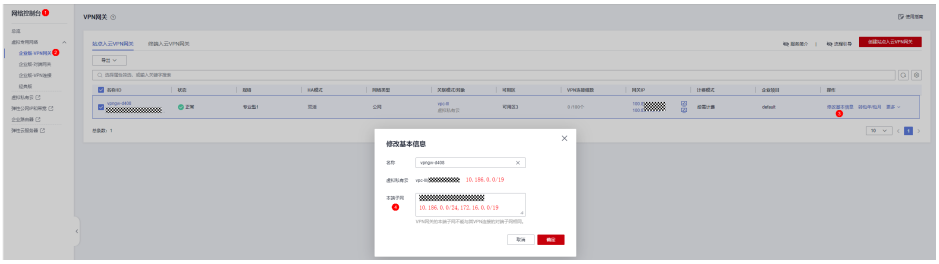
- 必要时，其他云数据库所在网络需要添加路由，目的地址指向Migration资源组VPC网段（例如172.16.0.0/19），下一跳跳至其他云的VPN网关，建议参考其他云VPN官网资料进行配置。
- 必要时，其他云上的VPN网关路由表中也需要添加路由，目的地址指向Migration资源组VPC网段（例如172.16.0.0/19），下一跳跳至其他云的VPN连接/网关，建议参考其他云VPN官网资料进行配置。

步骤3 虚拟专用网络本端子网添加资源组网段。

为了允许VPN访问资源组网段，请登录虚拟专用网络控制台，在左侧导航栏，选择“虚拟专用网络 > VPN网关”，在列表中找到连通其他云所使用的VPN网关，单击右

侧操作栏中的“修改基本信息”按钮，在弹出框中的“本端子网”输入框里添加Migration资源组VPC网段（例如172.16.0.0/19）。

图 7-55 添加本端子网



步骤4 创建Migration实时网络连接（对等连接）。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-56 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-14 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多”>“绑定资源组”进行选择。
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。

参数	说明
子网	中转VPC的子网（例如10.186.0.0/24）。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤5 实时网络连接（对等连接）添加数据源网段路由。

单击步骤4所创建实时网络连接的“路由信息”，单击“添加路由”，输入本地IDC自建数据库的私有网络地址（例如192.168.0.0/24）。

图 7-57 添加路由 1

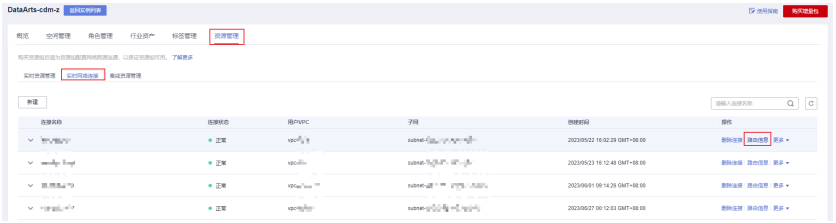


图 7-58 添加路由 2

步骤6 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

图 7-59 修改主机信息

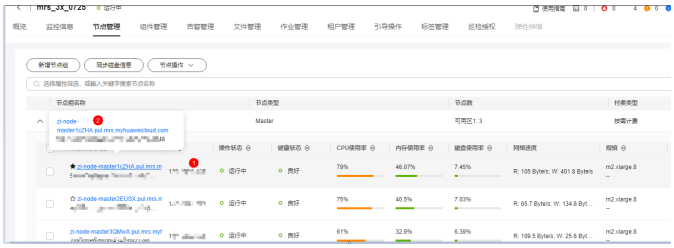
说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。

  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-60 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤7 测试网络连接。

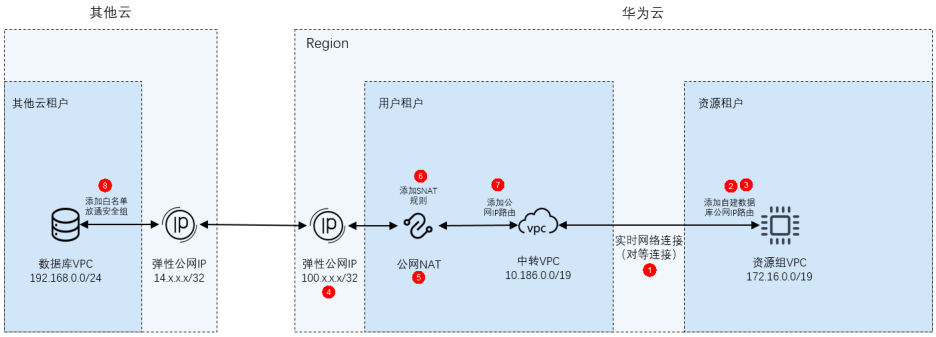
在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

----结束

7.4.2.3 通过公网连通网络

本章节主要为您介绍数据库部署在其他云场景下，通过公网打通网络的方案。

图 7-61 网络示意图



约束限制

资源组不具有公网网段，只能通过公网NAT转换成固定的弹性公网IP访问公网，且该IP不能与数据源公网IP重复。

前提条件

已购买资源组，详情请参见**购买数据集成资源组**。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-15 资源网段规划

资源名称	说明	私网网段示例
数据源公网IP	其他云数据源的公网IP，请用户根据实际情况自行获取。	14.x.x.x/32
弹性公网IP	资源组不具有公网网段，只能通过公网NAT转换成固定的弹性公网IP以访问公网。若未开通弹性公网IP，请登录弹性公网IP控制台，单击“购买弹性公网IP”，参考**通过VPC和EIP快速搭建可访问公网的网络**进行配置。	100.x.x.x/32
中转VPC及其子网	用于连通数据源和资源组网络的中间桥梁，本方案中需要使用当前租户下的一个虚拟私有云。若未开通VPC请参考**创建虚拟私有云**进行配置。	VPC：10.186.0.0/19 子网：10.186.0.0/24

资源名称	说明	私网网段示例
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-62** 查询资源组网段  	172.16.0.0/19

网络配置流程

步骤1 创建Migration实时网络连接（对等连接）。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-63 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-16 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多”>“绑定资源组”进行选择。
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。
子网	中转VPC的子网（例如10.186.0.0/24）。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤2 实时网络连接（对等连接）添加数据源网段路由。

单击步骤1所创建实时网络连接的“路由信息”，单击“添加路由”，输入本地IDC自建数据库的公网IP（例如14.x.x.x/32）。

图 7-64 添加路由 1

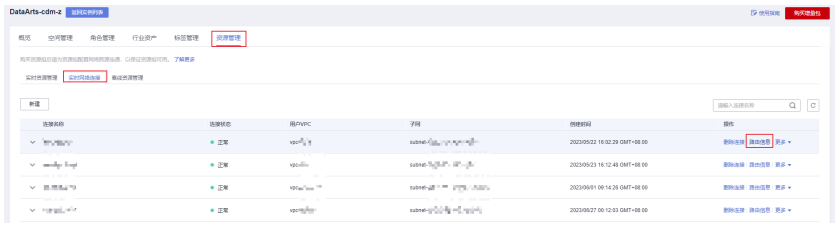
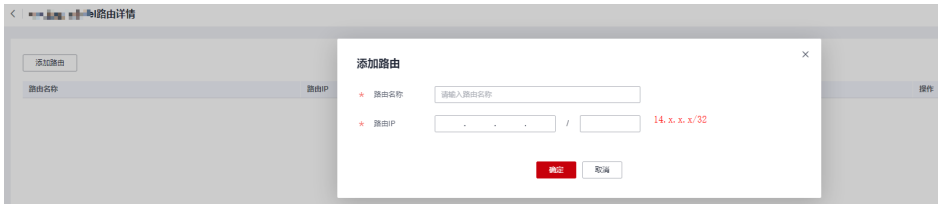


图 7-65 添加路由 2



步骤3 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

图 7-66 修改主机信息

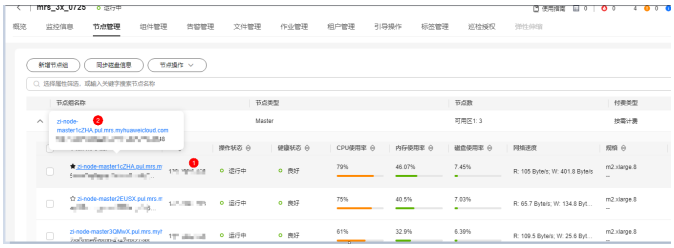
说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。

  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-67 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤4 购买弹性公网IP。

登录弹性公网IP控制台，单击“购买弹性公网IP”，根据界面提示配置参数，详情请参考**通过VPC和EIP快速搭建可访问公网的网络**。

步骤5 新建公网NAT网关。

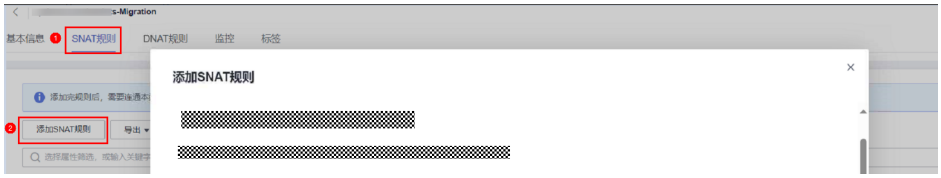
1. 登录NAT网关控制台，在左侧导航栏中选择“NAT网关 > 公网NAT网关”，单击“购买公网NAT网关”。
2. 配置NAT网关时，区域选择Migration所在Region，虚拟私有云选择中转VPC（例如10.186.0.0/19），子网选择中转VPC的子网（例如10.186.0.0/24），其余参数可参考**购买公网NAT网关**。

图 7-68 配置公网 NAT 网关

步骤6 为公网NAT网关添加SNAT规则。

为新建的NAT网关添加SNAT规则，才能实现资源组网段下的主机与Internet互相访问。请单击新建的公网NAT网关名称进入配置界面，选择“SNAT规则”页签，单击“添加SNAT规则”。

图 7-69 添加 SNAT 规则 1



其中，使用场景选择“云专线/云连接”，输入资源组VPC网段（例如172.16.0.0/19），然后绑定步骤3所购买的弹性公网IP（100.x.x.x/32）。

图 7-70 添加 SNAT 规则 2

步骤7 中转VPC子网网络添加路由。

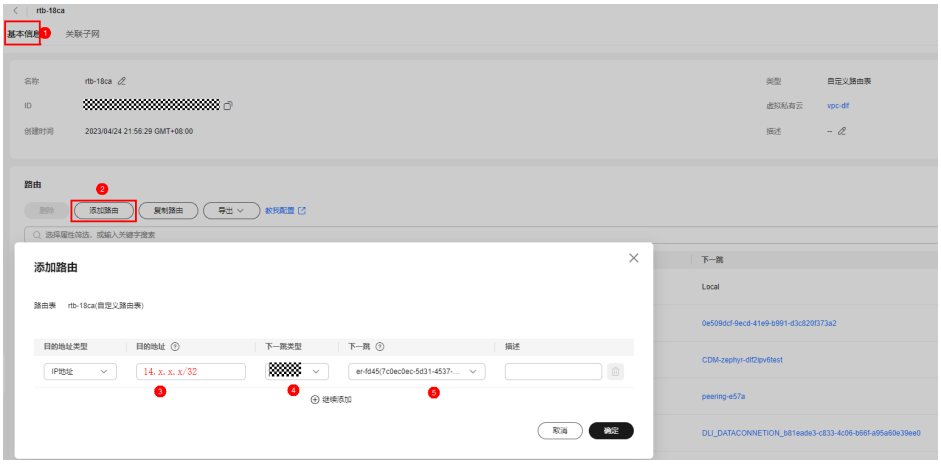
中转VPC子网的路由表中需要添加路由，目的地址指向本地IDC数据库的公网IP（例如14.x.x.x/32），下一跳跳至上面配置的NAT网关。

1. 登录虚拟私有云控制台，在左侧导航栏选择“虚拟私有云 > 子网”，找到中转VPC的子网并单击对应路由表名称进入配置界面。

图 7-71 查找路由表

2. 然后在路由表界面中选择“基本信息”页签，单击“添加路由”，目的地址指向本地IDC数据库的公网IP（例如14.x.x.x/32），下一跳跳至上面配置的NAT网关。

图 7-72 路由表添加路由



步骤8 其他云数据库添加白名单及安全组规则。

- 其他云数据库需要添加Migration资源组VPC网段（例如100.x.x.x/32）访问数据库的权限。各类型数据库添加白名单的方法不同，具体方法请参考各数据库官方文档进行操作。
- 数据库若配置了安全组，则还需要增加入方向规则，放通Migration资源组VPC网段（例如100.x.x.x/32），使其可以访问数据库监听端口。

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤9 测试网络连接。

在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

----结束

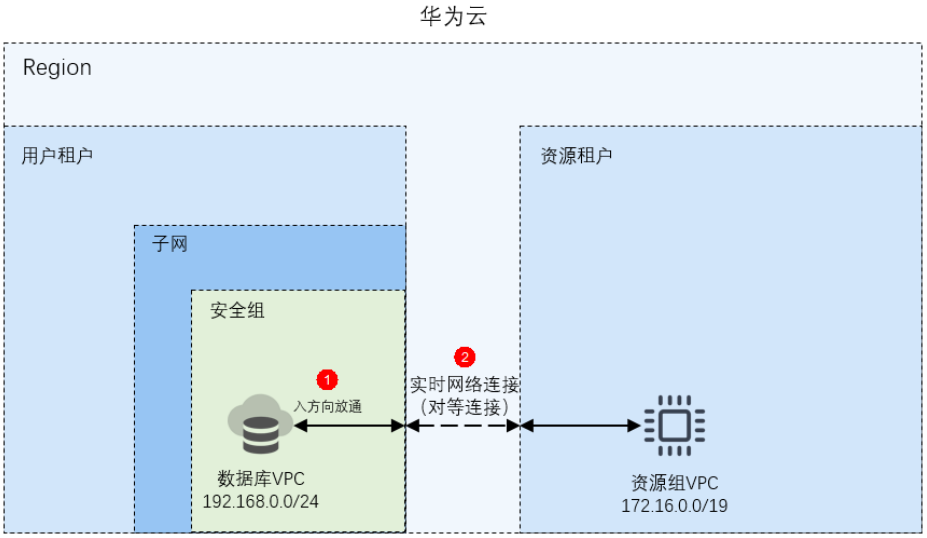
7.4.3 数据库部署在华为云

7.4.3.1 同 Region 同租户直接连通网络

在配置实时同步任务前，您需要确保源端和目的端的数据库与运行实时同步任务的实时计算资源组之间网络连通，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。

本章节主要为您介绍数据库部署在华为云，且与Migration资源组同Region同租户场景下的网络打通方案。

图 7-73 网络示意图



约束限制

- 资源组为私网网段，不能与数据源网段重叠，否则会导致网络无法打通。
- 资源组不具有公网网段，因此本方案仅能与数据源的私网连通。

前提条件

已购买资源组，详情请参见**购买数据集成资源组**。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-17 资源网段规划

资源名称	说明	私网网段示例
数据源VPC	华为云数据源所属的VPC，各数据源VPC查看方式不同，具体方法请参考数据源官方文档。	192.168.0.0/24

资源名称	说明	私网网段示例
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-74** 查询资源组网段  	172.16.0.0/19

网络配置流程

步骤1 配置华为云数据库所在安全组规则。

华为云数据库所在安全组需要增加入方向规则，放通Migration资源组VPC网段（例如172.16.0.0/19），使其可以访问数据库监听端口。

通用添加安全组规则方法：打开数据源服务界面，进入用户集群，找到网络部分，单击安全组，跳转到安全组编辑页面，单击入方向规则，添加规则。可参考如下示例放通资源组网段。

优先级	策略	类型	协议端口	源地址
1	允许	IPv4	全部协议	IP地址：资源组网段

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤2 创建Migration实时网络连接（对等连接）。

为了连通华为云数据库VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-75 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-18 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多” >“绑定资源组”进行选择。
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。
子网	数据源集群或实例所在的子网（例如192.168.0.0/24）。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤3 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

图 7-76 修改主机信息

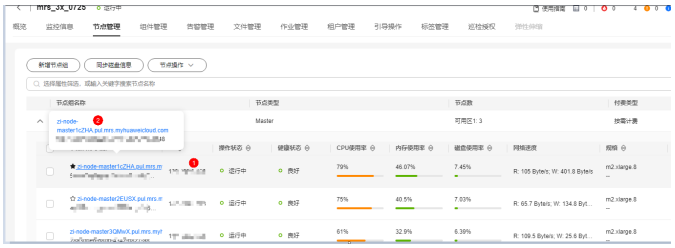
说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。

  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-77 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤4 测试网络连接。

在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

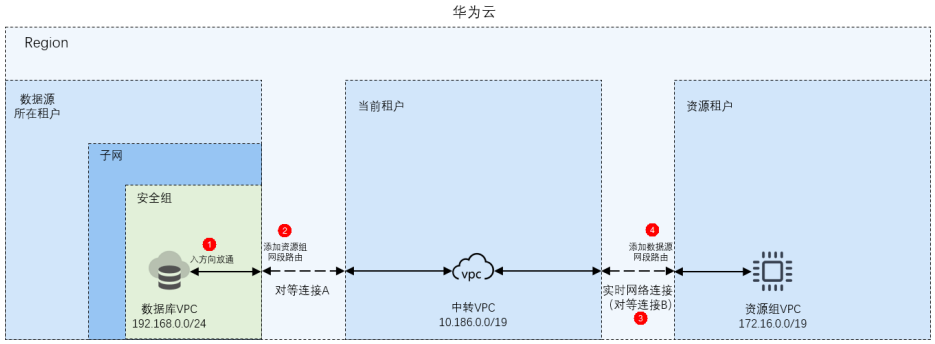
----结束

7.4.3.2 同 Region 不同租户通过对等连接连通网络

在配置实时同步任务前，您需要确保源端和目的端的数据库与运行实时同步任务的实时计算资源组之间网络连通，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。

本章节主要为您介绍数据库部署在华为云，且与Migration资源组同Region不同租户的场景下，通过对等连接打通网络的方案。

图 7-78 网络示意图



约束限制

- 资源组为私网网段，不能与数据源网段重叠，否则会导致网络无法打通。
- 资源组不具有公网网段，因此本方案仅能与数据源的私网连通。

前提条件

- 已购买资源组，详情请参见**购买数据集成资源组**。
- 已创建对等连接使数据源所在租户下的数据源VPC与当前租户下的一个VPC互联互通。若未创建对等连接请参考**创建不同账户下的对等连接**进行配置。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-19 资源网段规划

资源名称	说明	私网网段示例
数据源VPC	华为云数据源所属的VPC，各数据源VPC查看方式不同，具体方法请参考数据源官方文档。	192.168.0.0/24

资源名称	说明	私网网段示例
中转VPC	用于连通数据源和资源组网络的中间桥梁，本方案中需要使用当前租户下与数据源所在租户打通了对等连接的虚拟私有云。 查看方式： 在当前租户下，登录虚拟私有云控制台，在左侧导航栏，选择“虚拟私有云 > 对等连接”，在列表中查找“对端VPC网段”为数据源VPC的对等连接，它的“本端VPC”即可作为中转VPC。 **图 7-79** 查看对等连接 	VPC：10.186.0.0/19
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-80** 查询资源组网段 	172.16.0.0/19

网络配置流程

步骤1 配置华为云数据库所在安全组规则。

华为云数据库所在安全组需要增加入方向规则，放通Migration资源组VPC网段（例如172.16.0.0/19），使其可以访问数据库监听端口。

通用添加安全组规则方法：打开数据源服务界面，进入用户集群，找到网络部分，单击安全组，跳转到安全组编辑页面，单击入方向规则，添加规则。可参考如下示例放通资源组网段。

优先级	策略	类型	协议端口	源地址
1	允许	IPv4	全部协议	IP地址：资源组网段

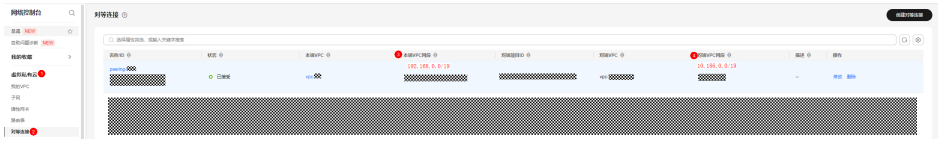
说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤2 为数据源VPC和中转VPC之间的对等连接添加路由。

在数据源所属租户下，登录虚拟私有云控制台，在左侧导航栏，选择“虚拟私有云 > 对等连接”，在列表中查找“本端VPC”为数据源VPC（例如192.168.0.0/19），且“对端VPC”为中转VPC（例如10.186.0.0/19）的对等连接，单击对等连接名称进入配置界面。

图 7-81 数据源所在租户下查找对等连接



单击“添加路由”，虚拟私有云默认为数据源VPC，路由表根据实际情况自行选择（建议使用与中转VPC路由同一路由表），目的地址添加资源组VPC（例如172.16.0.0/19），下一跳默认跳至该对等连接。

图 7-82 跨账户对等连接添加资源组网段路由

步骤3 创建Migration实时网络连接（对等连接）。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-83 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-20 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多” >“绑定资源组”进行选择。
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。
子网	中转VPC的子网（例如10.186.0.0/24）。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤4 实时网络连接（对等连接）添加数据源网段路由。

单击步骤3所创建实时网络连接的“路由信息”，单击“添加路由”，输入华为云数据源VPC子网网段（例如192.168.0.0/24）。

图 7-84 添加路由 1

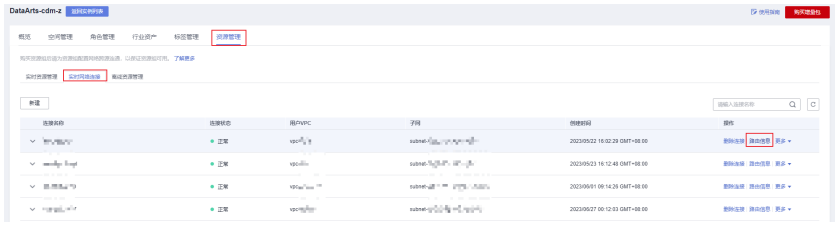


图 7-85 添加路由 2

步骤5 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

图 7-86 修改主机信息

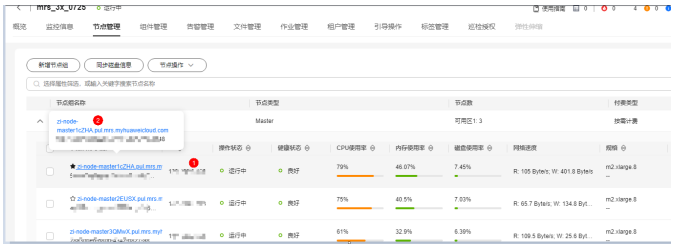
说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。

  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-87 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤6 测试网络连接。

在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

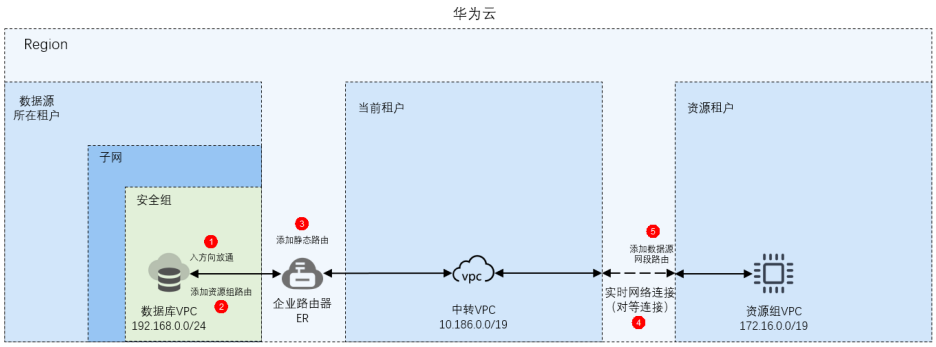
----结束

7.4.3.3 同 Region 不同租户通过企业路由器连通网络

在配置实时同步任务前，您需要确保源端和目的端的数据库与运行实时同步任务的实时计算资源组之间网络连通，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。

本章节主要为您介绍数据库部署在华为云，且与Migration资源组同Region不同租户场景下，通过企业路由器打通网络的方案。

图 7-88 网络示意图



约束限制

- 资源组为私网网段，不能与数据源网段重叠，否则会导致网络无法打通。
- 资源组不具有公网网段，因此本方案仅能与数据源的私网连通。

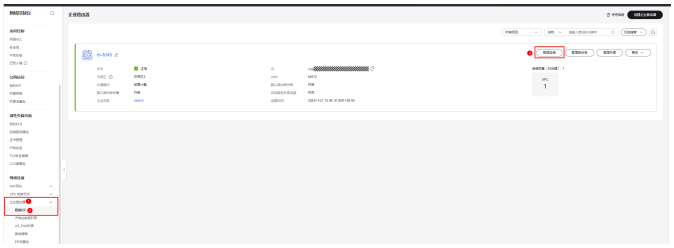
前提条件

- 已购买资源组，详情请参见**购买数据集成资源组**。
- 已购买并配置企业路由器，使数据源所在租户下的数据源VPC与当前租户下的一个VPC互联互通。若未开通企业路由器请参考**同区域VPC互通方案概述**和**共享企业路由器**进行配置。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-21 资源网段规划

资源名称	说明	私网网段示例
数据源网段	华为云数据源所属的VPC，各数据源VPC查看方式不同，具体方法请参考数据源官方文档。	192.168.0.0/24
中转VPC	用于连通数据源和资源组网络的中间桥梁，本方案中需要使用企业路由器中配置的当前租户下的虚拟私有云。 查看方式： 在当前租户下，登录企业路由器控制台，在左侧导航栏，选择“企业路由器 > 我的ER”，在列表中查找所用的企业路由器，单击“管理连接”进入配置界面，在“连接”页签中找到属于当前租户的VPC即可作为中转VPC。 **图 7-89** 查看企业路由器连接  **图 7-90** 确定中转 VPC 	VPC：10.186.0.0/19

资源名称	说明	私网网段示例
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-91** 查询资源组网段  	172.16.0.0/19

网络配置流程

步骤1 配置华为云数据库所在安全组规则。

华为云数据库所在安全组需要增加入方向规则，放通Migration资源组VPC网段（例如172.16.0.0/19），使其可以访问数据库监听端口。

通用添加安全组规则方法：打开数据源服务界面，进入用户集群，找到网络部分，单击安全组，跳转到安全组编辑页面，单击入方向规则，添加规则。可参考如下示例放通资源组网段。

优先级	策略	类型	协议端口	源地址
1	允许	IPv4	全部协议	IP地址：资源组网段

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤2 华为云数据库所在网络添加路由。

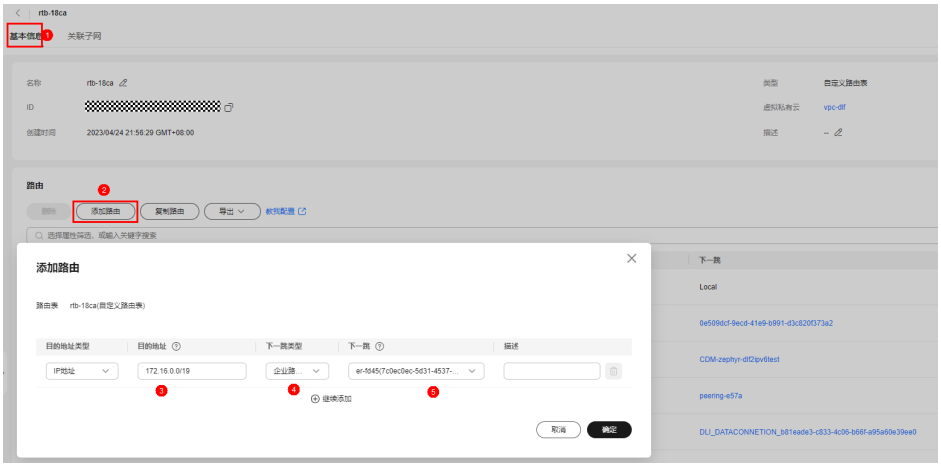
华为云数据库所属VPC子网的路由表中需要添加路由，目的地址指向Migration资源组VPC网段（例如172.16.0.0/19），下一跳跳至此前已配置好的企业路由器。

1. 登录数据源所在租户的虚拟私有云控制台，在左侧导航栏选择“虚拟私有云 > 子网”，找到数据源所在的子网并单击对应路由表名称进入配置界面。

图 7-92 查找数据源路由表

2. 在路由表界面中选择“基本信息”页签，单击“添加路由”，目的地址指向Migration资源组VPC网段（例如172.16.0.0/19），下一跳跳至此前已配置好的企业路由器。

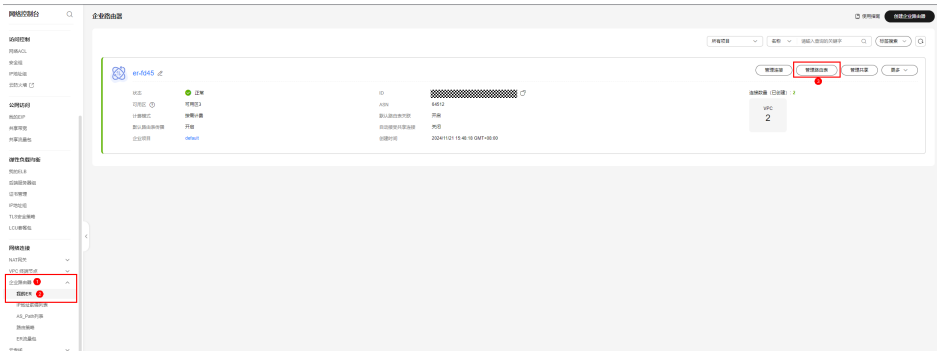
图 7-93 数据源路由表添加路由



步骤3 企业路由器中添加路由。

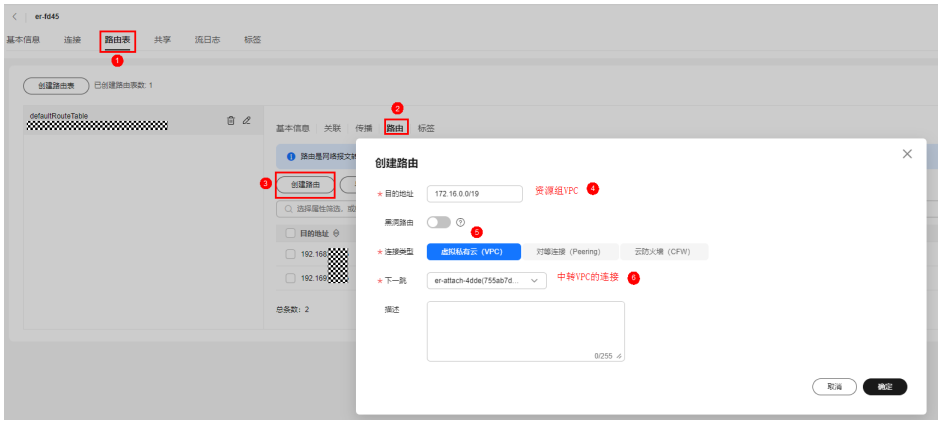
登录企业路由器控制台，在左侧导航栏，选择“企业路由器 > 我的ER”，在列表中查找所用的企业路由器，单击“管理路由表”进入配置界面。

图 7-94 查看企业路由器路由表



在“路由”页签中单击“创建路由”，目的地址填写实时资源组的VPC网段（例如172.16.0.0/19），连接类型选择虚拟私有云（VPC），下一跳选择中转VPC对应的连接。

图 7-95 创建路由



步骤4 创建Migration实时网络连接。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-96 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-22 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多” >“绑定资源组”进行选择。

参数	说明
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。
子网	中转VPC的子网（例如10.186.0.0/24）。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤5 为实时网络连接（对等连接）添加数据源网段路由。

单击步骤4所创建实时网络连接的“路由信息”，单击“添加路由”，输入华为云数据库的VPC子网网段（例如192.168.0.0/24）。

图 7-97 添加路由 1

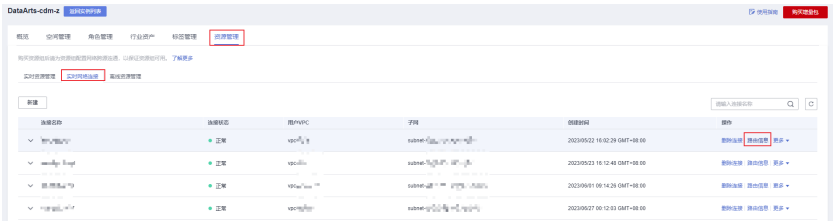


图 7-98 添加路由 2

步骤6 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

图 7-99 修改主机信息

说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。

  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-100 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤7 测试网络连接。

在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

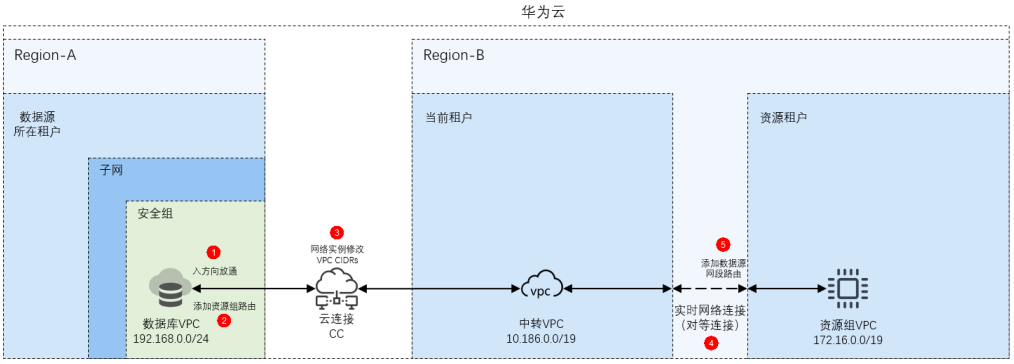
----结束

7.4.3.4 跨 Region 通过云连接连通网络

在配置实时同步任务前，您需要确保源端和目的端的数据库与运行实时同步任务的实时计算资源组之间网络连通，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。

本章节主要为您介绍数据库部署在华为云，且与Migration资源组不同Region场景下，通过云连接打通网络的方案。

图 7-101 网络示意图



约束限制

- 资源组为私网网段，不能与数据源网段重叠，否则会导致网络无法打通。
- 资源组不具有公网网段，因此本方案仅能与数据源的私网连通。

前提条件

- 已购买资源组，详情请参见**购买数据集成资源组**。
- 已购买并配置云连接，使数据源所在租户下的数据源VPC与当前租户下的一个VPC互联互通。若未开通云连接请参考**通过云连接实例实现跨区域VPC互通**进行配置。

准备工作

查询打通网络过程中所涉及到对象的网段（包含数据源、中转VPC、资源组），为便于理解，本章节将举例为您进行介绍。

表 7-23 资源网段规划

资源名称	说明	私网网段示例
数据源网段	华为云数据源所属的VPC，各数据源VPC查看方式不同，具体方法请参考数据源官方文档。	192.168.0.0/24
中转VPC	用于连通数据源和资源组网络的中间桥梁，本方案中需要使用云连接中配置的与Migration同region同租户的虚拟私有云。 查看方式： 登录云连接控制台，在左侧导航栏，选择“云连接 > 云连接实例”，在列表中查找所用的云连接，单击名称进入配置界面，在“网络实例”页签中找到与Migration同region同租户的VPC即可作为中转VPC。 **图 7-102** 查看云连接  **图 7-103** 确定中转 VPC 	VPC：10.186.0.0/19

资源名称	说明	私网网段示例
资源组VPC	Migration实时计算资源组所属VPC，由于资源组创建在用户账户下属的资源租户，使用资源租户的VPC网段，因此不占用用户账户的VPC网段。 查看方式： 登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面，在“实时资源管理”中单击指定资源组的下拉框，查看该资源组的VPC网段。 **图 7-104** 查询资源组网段  	172.16.0.0/19

网络配置流程

步骤1 配置华为云数据库所在安全组规则。

华为云数据库所在安全组需要增加入方向规则，放通Migration资源组VPC网段（例如172.16.0.0/19），使其可以访问数据库监听端口。

通用添加安全组规则方法：打开数据源服务界面，进入用户集群，找到网络部分，单击安全组，跳转到安全组编辑页面，单击入方向规则，添加规则。可参考如下示例放通资源组网段。

优先级	策略	类型	协议端口	源地址
1	允许	IPv4	全部协议	IP地址：资源组网段

说明

各数据源所用端口不尽相同，可参考**数据源安全组应放通哪些端口可满足Migration访问?** 进行安全组规则端口配置。

步骤2 华为云数据库所在网络添加路由。

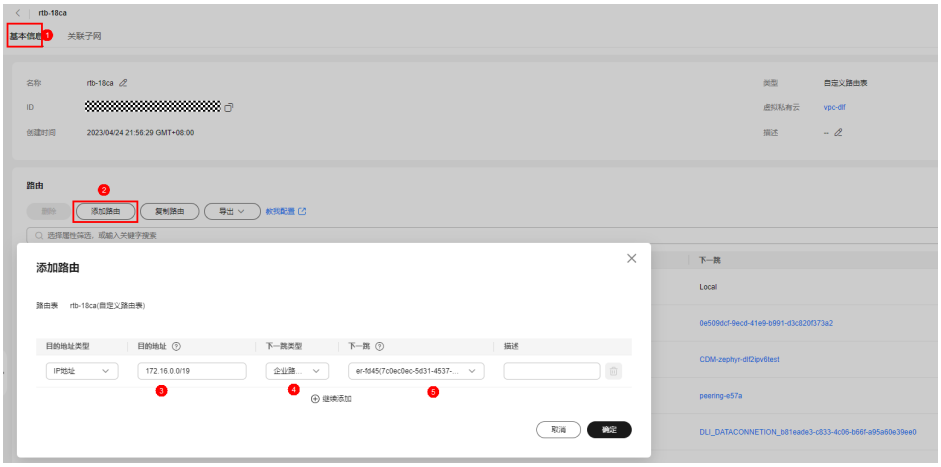
华为云数据库所属VPC子网的路由表中需要添加路由，目的地址指向Migration资源组VPC网段（例如172.16.0.0/19），下一跳跳至此前已配置好的云连接。

1. 登录数据源所在租户的虚拟私有云控制台，在左侧导航栏选择“虚拟私有云 > 子网”，找到数据源所在的子网并单击对应路由表名称进入配置界面。

图 7-105 查找数据源路由表

2. 在路由表界面中选择“基本信息”页签，单击“添加路由”，目的地址指向Migration资源组VPC网段（例如172.16.0.0/19），下一跳跳至此前已配置好的云连接。

图 7-106 数据源路由表添加路由



步骤3 云连接中转VPC的网络实例修改VPC CIDRs。

登录云连接控制台，在左侧导航栏，选择“云连接 > 云连接实例”，在列表中查找所用的云连接，单击名称进入配置界面，在“网络实例”页签中找到中转VPC的实例，单击右侧的“修改VPC CIDRs”按钮，在其他网段输入Migration资源组VPC网段（例如172.16.0.0/19）。

图 7-107 查看云连接

图 7-108 云连接网络实例

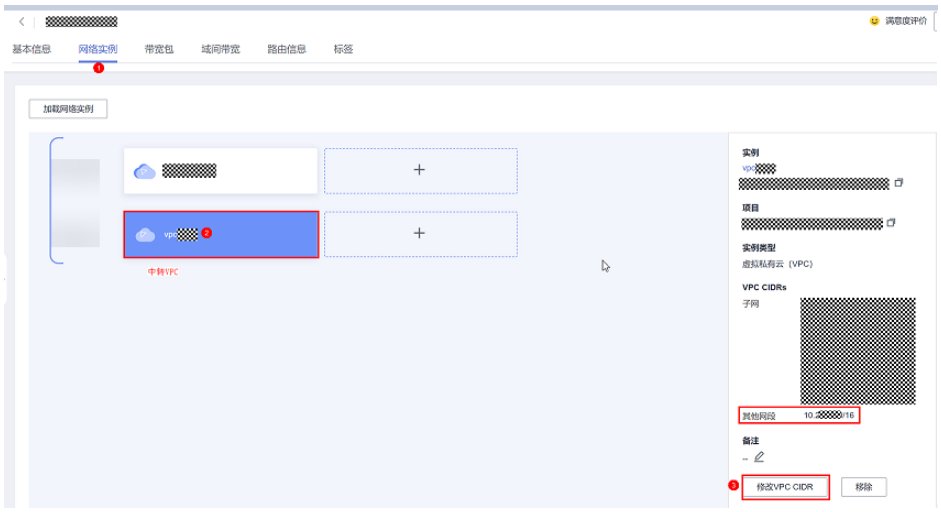


图 7-109 云连接网络实例修改 VPC CIDRs

步骤4 创建Migration实时网络连接。

为了连通中转VPC和实时资源组VPC网络，可以通过DataArts Studio资源管理功能来创建两个VPC间的对等连接。

登录DataArts Studio控制台，进入实例，单击“资源管理”进入资源管理页面。

图 7-110 新建网络连接

在“实时网络连接”页签中单击“新建”，在弹出的“新建网络连接”对话框输入对应参数，配置参数如下表所示：

表 7-24 新建网络连接参数

参数	说明
连接名称	填写待创建的网络连接名称。 只能包含字母、数字和下划线。
资源组	需要和指定VPC进行网络打通的资源组。 如果创建时未选择资源组，可以在网络连接创建好后再绑定资源组。支持绑定多个资源组，可以通过单击“更多” >“绑定资源组”进行选择。
虚拟私有云（VPC）	选择需要和资源组进行网络打通的虚拟私有云。 本方案中，资源组网段与中转VPC之间通过对等连接连通网络，因此必须选择中转VPC（例如10.186.0.0/19）。
子网	中转VPC的子网（例如10.186.0.0/24）。
路由表	子网实际关联的路由表，绑定资源组时会在此路由表中添加资源组的路由信息。本参数无需配置。 为网络连接绑定资源组，实际上是通过资源组网段与中转VPC之间的对等连接连通网络，因此绑定资源组时会在此路由表中添加一条指向资源组VPC网段的路由。

步骤5 为实时网络连接（对等连接）添加数据源网段路由。

单击步骤4所创建实时网络连接的“路由信息”，单击“添加路由”，输入华为云数据库的VPC子网网段（例如192.168.0.0/24）。

图 7-111 添加路由 1

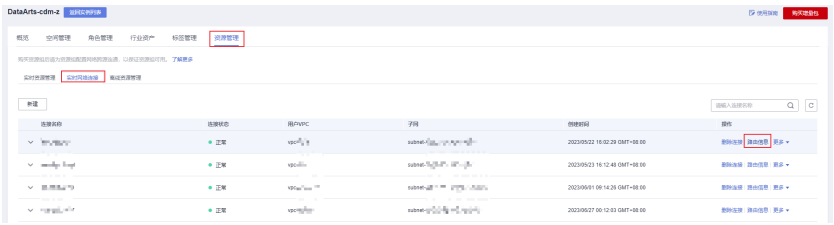


图 7-112 添加路由 2

步骤6 （可选）MRS类型数据源还需要进行以下操作打通网络。

实时网络连接创建完成并绑定资源组后，单击右侧“更多 > 修改主机信息”，按照输入框提示的格式填写MRS集群所有节点的IP和域名。

图 7-113 修改主机信息

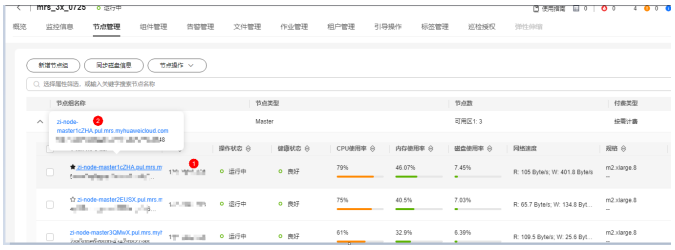
说明

查看MRS集群节点IP和域名的方式：

- 打开MRS页面，进入用户的MRS集群，单击“节点管理”页签，展开所有节点组，可以看到各节点IP、节点名称即是域名。

  须添加所有节点IP（图中序号1）、域名信息（图中序号2），用回车分割。

  图 7-114 查看 MRS 集群节点 IP 和域名

- 登录MRS集群节点，详情请参见**登录MRS集群节点**，执行命令**cat /etc/hosts**，可以列出所有节点的IP和域名。

步骤7 测试网络连接。

在DataArts Studio工作空间下创建数据连接，并创建实时集成作业，选择对应数据连接和资源组进行连通性测试，详情请参考**创建实时集成作业**。

----结束

7.5 新建实时集成作业

前提条件

作业在每个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。

操作步骤

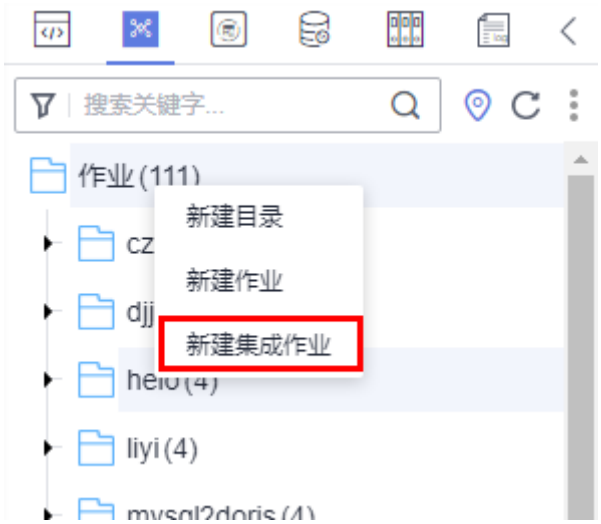
1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 新建集成作业的方式有如下两种：

   方式一：在“作业开发”界面中，单击“新建集成作业”。

图 7-115 新建集成作业（方式一）

方式二：在作业目录中，右键单击目录名称，选择“新建集成作业”。

图 7-116 新建集成作业（方式二）



5. 在弹出的“新建集成作业”页面，配置如表7-25所示的参数。

表 7-25 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、中划线和下划线。
作业类型	选择作业的类型，须选择实时处理。 • 离线处理：对已收集的大量数据进行批量处理和分析，这些任务通常是在计算资源和存储资源方面经过优化，以确保高效的数据处理和分析。这些任务通常是定时（例如每天、每周）执行，主要处理大量历史数据，用于批量分析和数据仓库。 • 实时处理：对源源不断产生的新数据进行实时处理和分析，以满足业务对数据的即时性需求。这种处理方式要求数据在产生后能够立即被处理，并给出相应的结果或触发相应的操作。
选择目录	选择作业所属的目录，默认为根目录。

参数	说明
日志路径	选择作业日志存放路径，默认为obs://dlf-log-...../。 勾选“我确认OBS桶obs://dlf-log-...../将被创建，该桶仅用于存储DLF的作业运行日志”选项，若要修改日志路径，请前往DataArts Studio空间管理进行编辑操作，详情请参考**（可选）修改作业日志存储路径**。
作业描述	自定义作业的描述信息。

6. 单击“确定”，创建作业。

7.6 配置实时集成作业

完成数据连接、网络、资源组等准备工作的配置后，您可创建并配置实时集成作业，将多种输入及输出数据源搭配组成同步链路，进行数据的实时同步。

前提条件

- 已**授权使用实时数据集成**。
- 已购买资源组，详情请参见**购买数据集成资源组**。
- 已准备数据源，对应连接账号具备权限，详情请参考**使用前自检概览**中对应数据库账号权限要求。
- 已创建数据连接，且创建的连接必须已勾选数据集成选项，详情请参见**创建DataArts Studio数据连接** 。
- 数据集成资源组与数据源网络已打通，详情请参见**网络打通**。

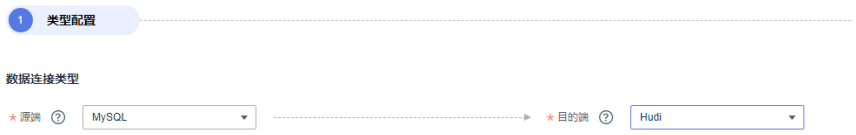
操作步骤

步骤1 参见**新建实时集成作业**创建一个实时处理集成作业。

步骤2 配置数据连接类型。

选择源端和目的端的数据类型，支持的源端与目的端请参见**新建实时集成作业**。

图 7-117 选择数据连接类型



步骤3 选择集成作业类型。

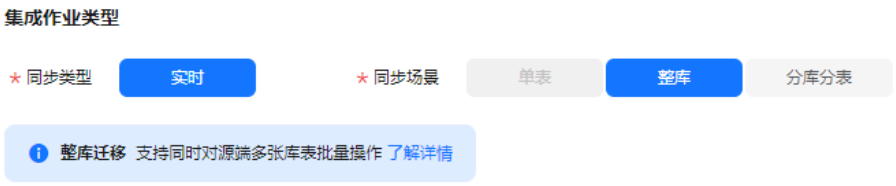
1. 同步类型：默认为实时，不可更改。
2. 同步场景：包含单表、整库、分库分表场景，各数据源支持的场景不一，详情请参见**使用教程**。

   不同场景介绍如**表7-26**所示。

表 7-26 同步场景参数说明

场景类型	说明
单表	支持将源端一个实例下的单张表实时同步至目的端一个实例下的单张表。
整库	支持将源端一个实例下多个库的多张表批量实时同步到目的端一个实例下的多个库表，一个任务中最多支持200张目标表。
分库分表	支持将源端多个实例下多个分库的多张分表同步到目的端一个实例下的单个库表。

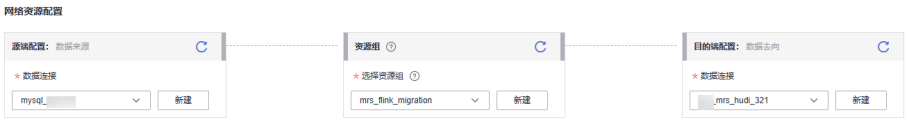
图 7-118 选择集成作业类型



步骤4 配置网络资源。

选择已创建的源端、目的端数据连接及已配置好网络连接的资源组。

图 7-119 选择数据连接及资源组



说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买数据集成资源组增量包**进行配置。

步骤5 检测网络连通性。

数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

说明

网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端、目标端参数。

各链路源端或目的端参数配置不同，详情请参见**使用教程**中对应的文档进行配置。

步骤7 刷新源表和目标表映射，检查映射关系是否正确，同时可根据需求修改表属性、添加附加字段。

步骤8 （可选）配置DDL消息处理规则。

实时集成作业除了能够同步对数据的增删改等DML操作外，也支持对部分表结构变化（DDL）进行同步。针对支持的DDL操作，用户可根据实际需求配置为正常处理/忽略/出错。

- 正常处理：Migration识别到源端库表出现该DDL动作时，作业自动同步到目的端执行该DDL操作。
- 忽略：Migration识别到源端库表出现该DDL动作时，作业忽略该DDL，不同步到目的端表中。
- 出错：Migration识别到源端库表出现该DDL动作时，作业抛出异常。

图 7-120 DDL 配置

步骤9 配置任务属性。

表 7-27 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒

参数	说明	默认值
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** ● 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 ● 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤10 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-121 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-122 启动配置

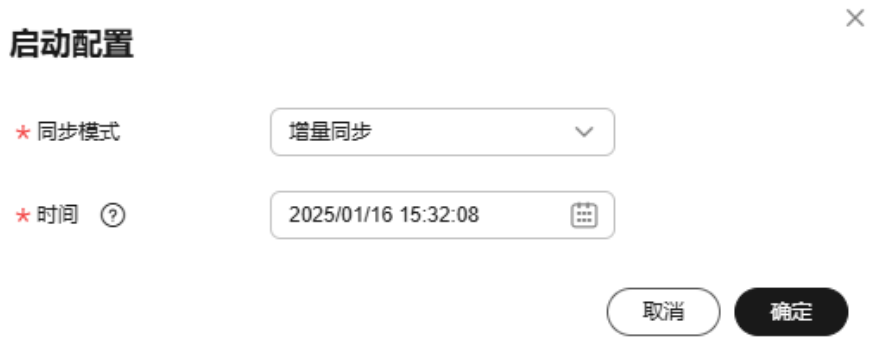


表 7-28 启动配置参数

参数	说明
同步模式	数据源通用同步模式： ● 增量同步：从指定时间位点开始同步增量数据。 ● 全量+增量：先同步全量数据，随后实时同步增量数据。
	Kafka数据源专用同步模式： ● 最早：从Kafka Topic最早偏移量开始消费数据。 ● 最新：从Kafka Topic最新偏移量开始消费数据。 ● 起止时间：根据时间获取Kafka Topic对应的偏移量，并从该偏移量开始消费数据。
时间	同步模式选择增量同步和起止时间时需要设置该参数，指示增量同步起始的时间位点。 **说明** ● 配置的位点时间早于数据源增量日志最早时间点时，默认会以日志最新时间点开始消费。 ● 配置的位点时间早于Kafka消息最早偏移量时，默认会从最早偏移量开始消费。

步骤11 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-123 前往监控

----结束

7.7 实时集成任务运维

7.7.1 查看监控指标

操作场景

当您启动了实时集成作业后，云监控服务会自动关联实时集成作业的监控指标，帮助您精确掌握作业的各项性能指标和运行情况。

说明

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。如果您的实时处理集成作业刚启动完成，请等待5～10分钟后查看监控数据。

前提条件

- 使用实时集成作业监控功能，需获取CES相关权限。
- 监控指标对应的实时集成作业需要正常运行，停止或异常的作业仅支持查看7天内的监控指标。
- 实时集成作业已正常运行一段时间（约10分钟）。

支持的监控指标

实时处理集成作业支持的监控指标如**表7-29**所示。

表 7-29 实时处理集成作业支持的监控指标

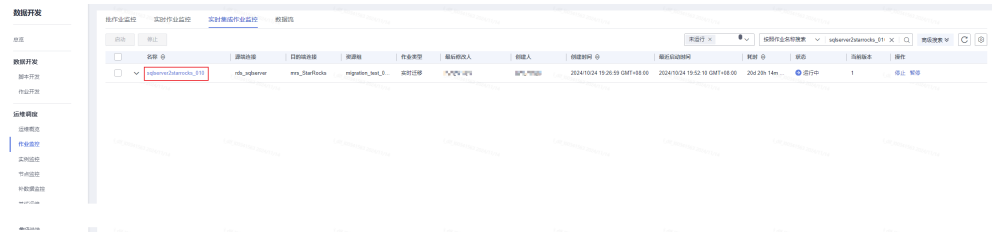
指标名称	指标含义	取值范围	测量对象	监控周期（原始指标）
源库WAL抽取时延	该指标用于统计当前从源库抽取WAL的时延	≥ 0ms	实时处理集成作业	1分钟
作业数据输入速率	展示用户Flink作业的数据输入速率，供监控和调试使用	≥ record/s	实时处理集成作业	1分钟

指标名称	指标含义	取值范围	测量对象	监控周期（原始指标）
作业数据输出速率	展示用户Flink作业的数据输出速率，供监控和调试使用	≥ record/s	实时处理集成作业	1分钟
作业数据输入总数	展示用户Flink作业的数据输入总数，供监控和调试使用	≥ records	实时处理集成作业	1分钟
作业数据输出总数	展示用户Flink作业的数据输出总数，供监控和调试使用	≥ records	实时处理集成作业	1分钟
作业字节输入速率	展示用户Flink作业每秒输入的字节数	≥ Byte/s	实时处理集成作业	1分钟
作业字节输出速率	展示用户Flink作业每秒输出的字节数	≥ Byte/s	实时处理集成作业	1分钟
作业字节输入总数	展示用户Flink作业字节的输入总数	≥ Byte	实时处理集成作业	1分钟
作业字节输出总数	展示用户Flink作业字节的输出总数	≥ Byte	实时处理集成作业	1分钟
作业CPU使用率	展示用户Flink作业的CPU使用率	≥ 0%	实时处理集成作业	1分钟
作业内存使用率	展示用户Flink作业的内存使用率	≥ 0%	实时处理集成作业	1分钟
作业最大算子时延	展示用户Flink作业的最大算子时延时间，单位ms	≥ 0ms	实时处理集成作业	1分钟
作业最大算子反压	展示用户Flink作业的最大算子反压值，数值从0-1，数值越大，反压越严重	≥ 0	实时处理集成作业	1分钟

操作步骤

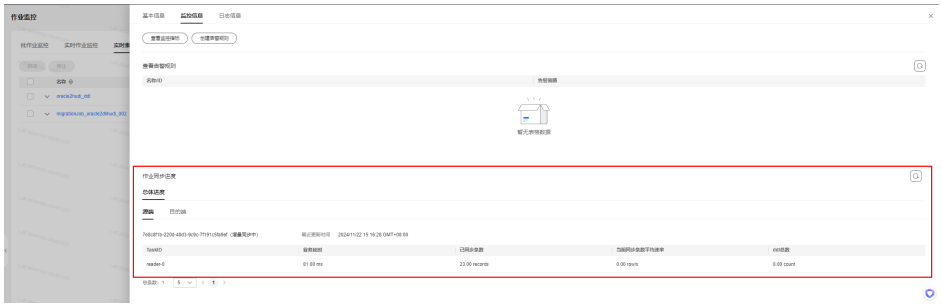
1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 选择“实时集成作业监控”页签，单击作业名称。

图 7-124 实时集成作业监控



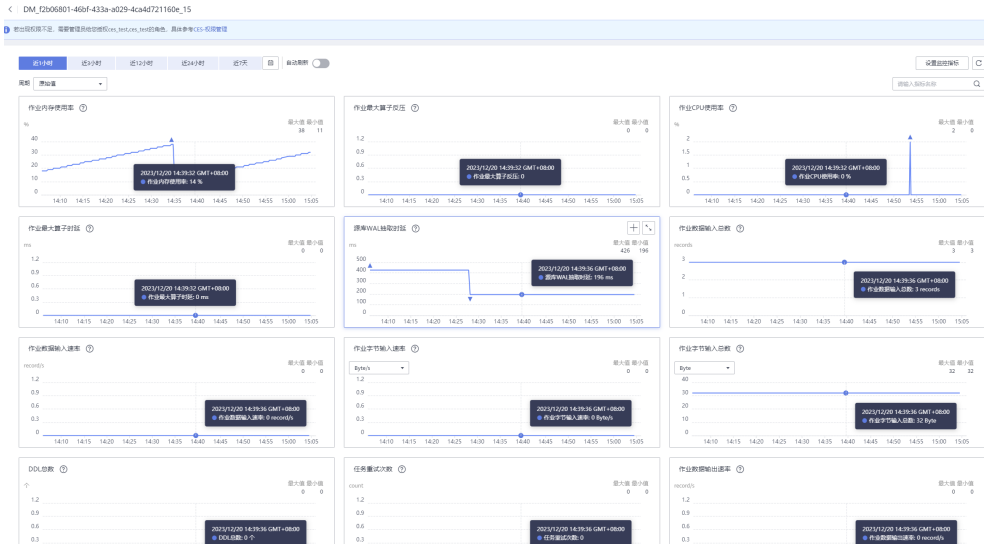
5. 在详情页面，选择“监控信息”，在页面最下方可直接查看作业的部分关键指标数据。

图 7-125 关键指标



6. 单击“查看监控指标”，跳转至云服务监控详情页面，查看图形化监控指标。

图 7-126 查看监控指标



说明

更多关于监控指标的信息，请参见云监控用户指南。

7.7.2 查看同步日志

Migration实时集成服务底层依托于Flink开发而来，同样对外开放了Flink的JobMamager和TaskManager日志，便于用户查看实时同步情况，并通过日志定位或排查异常问题。

前提条件

实时集成作业已正常运行一段时间（约5分钟）。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 选择“实时集成作业监控”页签，单击作业名称。
5. 在详情页面，选择“日志信息”，在左侧日志列表中单击具体日志文件，即可实时查看作业的运行日志。

图 7-127 日志信息 1

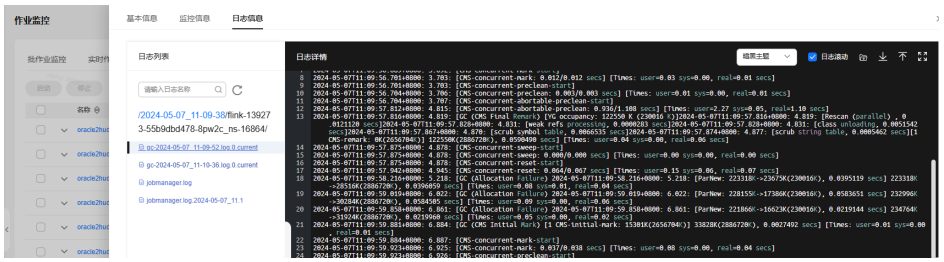
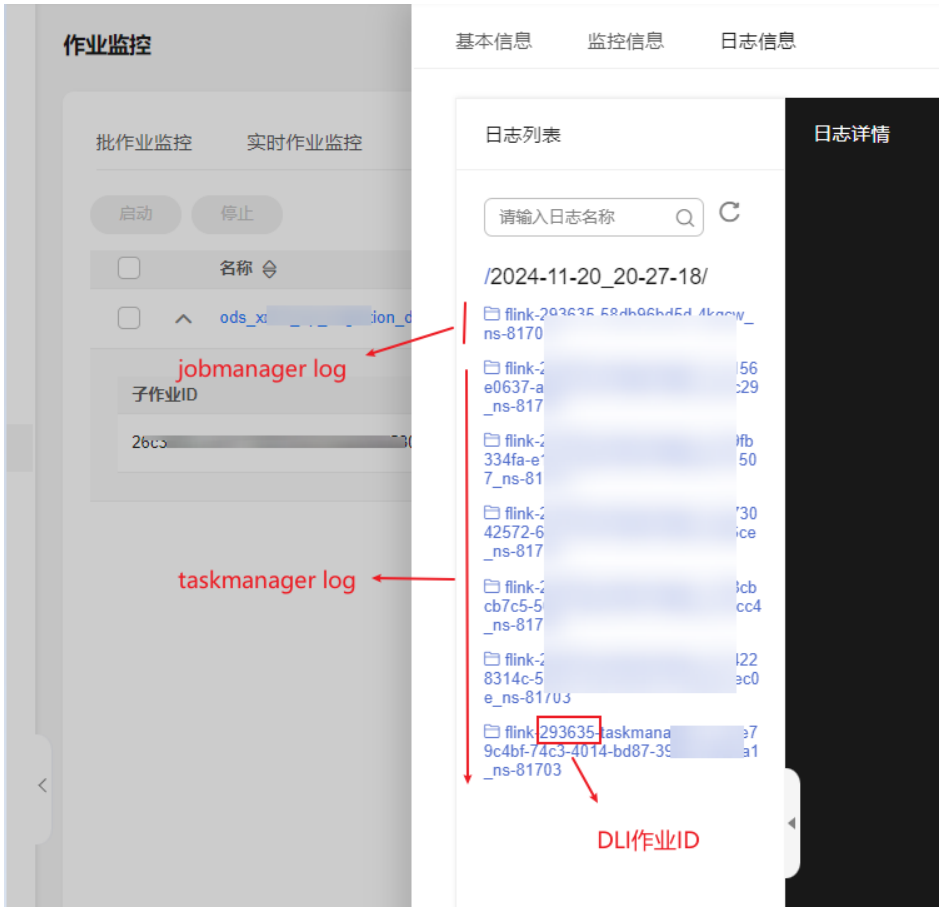


图 7-128 日志信息 2



说明

- 支持作业日志主题更换。
- 作业日志默认实时滚动更新，可在右上角去掉勾选“日志滚动”选项。
- 支持下载日志到本地，可单击右上角文件下载按钮进行下载。

7.7.3 配置告警规则

操作场景

通过设置实时集成作业的告警规则，用户可自定义监控目标与通知策略，及时了解作业状况，从而起到预警作用。

设置作业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。本节介绍了设置实时集成作业告警规则的具体方法。

配置一键告警

一键告警为您提供针对DataArts Studio服务下所有资源快速开启告警的能力，旨在帮助用户快速建立监控告警体系，在资源异常时可以及时获得通知。请参见**一键告警**打开“数据治理中心”一键告警开关。

配置所有资源告警

用户可以对实时处理集成作业的监控指标设置告警策略。当监控指标在一定周期内多次触发告警策略的阈值时，系统将向用户发送告警通知。具体操作请参见**设置实时处理集成作业告警规则**。

告警类型选择“指标”，云产品选择“数据治理中心-DataArts Studio作业”。

设置实时处理集成作业告警规则

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 选择“实时集成作业监控”页签，单击作业名称。
5. 在详情页面，选择“监控信息”，单击“创建告警规则”，进入云监控服务的创建告警规则界面，创建该作业的告警规则。

图 7-129 创建告警规则

6. 设置完成后，单击“立即创建”。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**。

7.7.4 动态修改任务配置

Migration实时集成任务拥有断点续传能力，支持用户通过“暂停 > 恢复”的方式动态加减表、修改任务配置、资源参数等，便于用户根据自身需求调整作业。

前提条件

实时集成作业正在运行中。

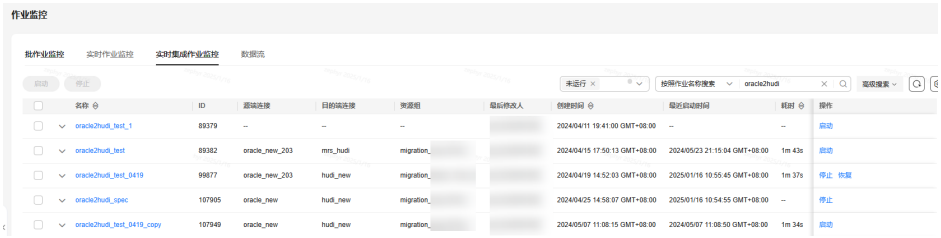
操作步骤

步骤1 暂停运行中的实时集成作业。

- 方式一：

  登录DataArts Studio控制台实例，进行所用空间的数据开发界面，单击左侧导航栏的“作业监控”，进入“实时集成作业监控界面”，搜索对应的实时集成作业，单击右侧操作栏中的“暂停”按钮。

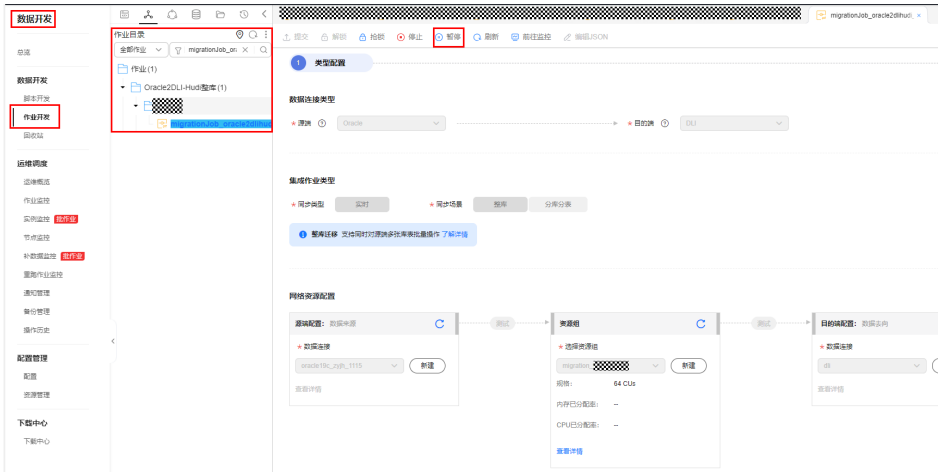
图 7-130 暂停作业 1



- 方式二：

  登录DataArts Studio控制台实例，进行所用空间的数据开发的“作业开发”界面，搜索并双击进入对应的实时集成作业配置界面，单击作业导航栏中的“暂停”按钮。

图 7-131 暂停作业 2



步骤2 修改任务配置。

根据实际需求，在实时集成作业配置界面中修改对应参数，随后保存提交作业。

步骤3 恢复实时集成作业。

- 方式一：登录DataArts Studio控制台实例，进行所用空间的数据开发界面，单击左侧导航栏的“作业监控”，进入“实时集成作业监控界面”，搜索对应的实时集成作业，单击右侧操作栏中的“恢复”按钮。

图 7-132 恢复作业 1

- 方式二：登录DataArts Studio控制台实例，进行所用空间的数据开发的“作业开发”界面，搜索并双击进入对应的实时集成作业配置界面，单击作业导航栏中的“恢复”按钮。

图 7-133 恢复作业 2

说明

动态加减表对于不同的启动模式有不同的操作效果，具体如下：

- 对于初始启动模式为“增量同步”的作业，暂停加表后恢复作业，新增的表将从暂停前的位点或用户重置的位点开始进行增量同步。
- 对于初始启动模式为“全量+增量”的作业，暂停加表后恢复作业，将对新增的表先进行全量同步，再从暂停前的位点开始进行增量同步。

----结束

7.8 字段类型映射关系

7.8.1 MySQL 与 MRS Hudi 字段类型映射

Migration会根据源端的字段类型按默认规则转换成目的端字段类型，并以此完成自动建表和实时同步。

字段类型映射规则

当源端为MySQL，目的端为Hudi时，支持的字段类型请参见下表，以确保数据完整同步到目的端。

表 7-30 MySQL > Hudi 支持的字段类型

类别	数据类型（MySQL）	数据类型（Hudi）	说明
字符串	CHAR(M)	STRING	-
	VARCHAR(M)	STRING	-
数值	BOOLEAN	BOOLEAN	-
	TINYINT	INT	TINYINT(1)默认会转成BOOLEAN类型，可通过在管理中心MySQL数据连接中添加“连接属性”使其仍保持转成TINYINT(1)：tinyInt1isBit = false。
	TINYINT UNSIGNED	INT	-
	SMALLINT	INT	-
	SMALLINT UNSIGNED	INT	-
	MEDIUMINT	INT	-
	MEDIUMINT UNSIGNED	BIGINT	-
	INT	INT	-
	INT UNSIGNED	BIGINT	-
	BIGINT	BIGINT	-
	BIGINT UNSIGNED	DECIMAL(20,0)	-
	REAL	不支持	-
	DECIMAL(M,D)	DECIMAL(38,10)	-
	NUMBERIC	不支持	-
	FLOAT(M,D)	FLOAT	-
	DOUBLE(M,D)	DOUBLE	-
	DOUBLE PRECISION	DOUBLE	-
位	BIT(M)	不支持	-
日期时间	DATE	DATE	-
	TIME	STRING	-
	DATETIME	TIMESTAMP	-
	TIMESTAMP	TIMESTAMP	-

类别	数据类型（MySQL）	数据类型（Hudi）	说明
	YEAR(M)	STRING	-
多媒体（二进制）	BINARY(M)	不支持	-
	VARBINARY(M)	不支持	-
	TEXT	STRING	-
	TINYTEXT	STRING	-
	MEDIUMTEXT	STRING	-
	LONGTEXT	STRING	-
	BLOB	不支持	-
	TINYBLOB	不支持	-
	MEDIUMBLOB	不支持	-
	LONGBLOB	不支持	-
特殊类型	SET	不支持	-
	JSON	STRING	-
	ENUM	不支持	-

7.8.2 PostgreSQL 与 DWS 字段类型映射

Migration会根据源端的字段类型按默认规则转换成目的端字段类型，并以此完成自动建表和实时同步。

字段类型映射规则

当源端为PostgreSQL，目的端为DWS时，支持的字段类型请参见下表，以确保数据完整同步到目的端。

表 7-31 PostgreSQL > DWS 支持的字段类型

类别	数据类型（PostgreSQL）	数据类型（DWS）	说明
字符串	CHAR(M)	CHAR(M)	固定长字符串，空格填充。
	VARCHAR(M)	VARCHAR(M)	有限制的变长字符串。
	TEXT	TEXT	无限制的变长字符串，类似没有长度声明词的VARCHAR。
数值	BOOLEAN	BOOL	逻辑布尔值（真/假）。

类别	数据类型（PostgreSQL）	数据类型（DWS）	说明
	SMALLINT	SMALLINT	即int2。
	INTEGER	INTEGER	即int/int4。
	BIGINT	BIGINT	即int8。
	DECIMAL(M,D)	DECIMAL(M,D)	可选择精度的精确数字。
	NUMERIC(M,D)	NUMERIC(M,D)	与NUMERIC等效。
	REAL	REAL	单精度浮点数（4字节）。
	DOUBLE	DOUBLE	即DOUBLE PRECISION，也可用没有精度的FLOAT表示，双精度浮点数（8字节）。
日期时间	DATE	TMESTAMP	源端为日期（没有一天中的时间），到目的端类型会变成日期+时间的timestamp。
	TIME(M)	TIME	一天中的时间（不带日期）。
	TIME(M) WITH TIME ZONE	TIMETZ	即TIMETZ，一天中的时间（不带日期），带有时区。
	TIMESTAMP(M)	TMESTAMP	包括日期和时间，无时区。
	TIMESTAMP(M) WITH TIME ZONE	TMESTAMPTZ	即TMESTAMPTZ，包括日期和时间，带有时区。
	INTERVAL	INTERVAL	时间间隔。
二进制	BYTEA	BYTEA	二进制数据（“字节数组”）。

7.9 任务性能调优

7.9.1 性能调优概述

实时处理集成作业各链路如果出现时延持续增长、反压持续处于高位或同步速率过慢（查看作业监控指标速率不符合实时集成作业提供的性能规格）等情况，需要考虑以下几点：

- 目的端写入过慢。
- 源端抽取过慢。
- 其他问题（请联系技术支持人员协助解决）。

因为目的端写入过慢会影响至源端，导致源端抽取速度下降，因此链路速度过慢请优先排查目的端写入速度，在排除目的端因素后再排查上游。

目的端写入慢

1. 检查目的端负载是否已达到目的端数据源上限。优先查看目的端数据源的监控指标，查看CPU、内存、IO等参数是否处于高负载状态。
2. 在排除目的端负载的情况下，加大作业并发，以提高写入速度。
3. 如果第2步也无法有效提升性能，请根据**源端抽取慢**排查源端的性能因素。
4. 如果排除了源端问题的情况下，请参考对应链路性能调优文档尝试进行参数优化。
5. 如果上述步骤仍然无法提升作业速度，请联系技术支持人员协助解决。

源端抽取慢

1. 检查源端负载是否已到达源端数据源上限。优先查看源端数据源的监控指标，查看CPU、内存、IO等参数是否处于高负载状态。
2. 在排除源端负载的情况下，如果源端是MySQL/Oracle/SQLServer/PostgreSQL/GaussDB等的全量+增量作业且作业处于全量抽取阶段，或者Kafka/Hudi等数据源抽取速度慢，请优先尝试加大作业并发数，以提高作业的并发抽取速率。

   MySQL/Oracle/SQLServer/PostgreSQL/GaussDB等关系型数据为保证事务有序，在增量阶段是单并发抽取，加大并发一般不会提升抽取性能。
3. 如果第2步也无法有效提升性能，请参考对应链路性能调优文档尝试进行参数优化。
4. 如果上述步骤仍然无法提升作业速度，请联系技术支持人员协助解决。

7.9.2 作业任务参数调优

概述

实时数据集成服务底层使用Flink流处理框架进行开发，因此包含了Flink系统中最重要的两个部分：JobManager和TaskManager。

作业任务配置中调整的处理器核数、并发数、执行内存参数等便是用来调整JobManager和TaskManager的，默认情况下单个作业使用2U8G资源，会对应创建出1个JobManager进程和1个TaskManager进程，且均使用1U4G资源。

作业调优

默认场景下，给定的1U4G规格可满足绝大部分使用场景，但Migration服务也提供修改JobManager和TaskManager规格的能力以应对极端情况。例如最常见的作业内存溢出，可以在实时集成作业的“任务配置”中添加自定义属性，根据实际情况调整JobManager和TaskManager的各类内存来适应同步场景。

图 7-134 添加自定义属性

表 7-32 作业任务参数一览表

参数名	参数类型	默认值	参数说明
jobmanager.memory.process.size	int	3586 MB	jobmanager的处理内存，直接影响堆内存大小。 **说明** 该配置会占用总体资源，影响新增其他作业，非必要不配置。
taskmanager.memory.process.size	int	3686 MB	taskmanager的处理内存，直接影响堆内存大小。 **说明** 该配置会占用总体资源，影响新增其他作业，非必要不配置。
taskmanager.memory.managed.fraction	int	0.2	taskmanager管理内存占比。
taskmanager.memory.network.max	int	128 MB	默认不需要配置，分库分表场景下如果实例数和表数过多，建议根据实际情况增加网络内存。
taskmanager.memory.network.fraction	int	0.1	默认不需要配置，分库分表场景下如果实例数和表数过多，建议根据实际情况增加网络内存。
checkpoint.interval	int	60000	Flink作业生成checkpoint的间隔，单位为毫秒。数据量大的作业建议调大，需要给更长时间进行数据Flush，但会增加时延。
checkpoint.timeout.ms	int	600000	Flink作业生成checkpoint的超时时间，单位为毫秒。

7.9.3 MySQL 到 MRS Hudi 参数调优

源端优化

MySQL抽取优化。

可通过在作业任务配置参数单击中“添加自定义属性”来新增MySQL同步参数。

图 7-135 添加自定义属性

可使用的调优参数具体如下：

表 7-33 全量阶段优化参数

参数名	类型	默认值	说明
scan.incremental.snapshot.backfill.skip	boolean	true	全量阶段是否跳过读取binlog数据，默认为true。跳过读取binlog数据可以有效降低内存使用。需要注意的是，跳过读取binlog功能只提供at-least-once保证。
scan.incremental.snapshot.chunk.size	int	50000	分片大小，决定了全量阶段单个分片最大数据的数据条数以及分片个数。分片大小越大，单个分片数据条数越多，分片个数越小。 当表的条数过多时，作业会划分较多的分片，从而占用过多的内存导致内存问题，请解决表的条数适当调整该值。 当scan.incremental.snapshot.backfill.skip为false时，实时处理集成作业会缓存单个分片的数据，此时分片越大，占用内存越多，引发内存溢出，在此场景下，可以考虑降低分片大小。
scan.snapshot.fetch.size	int	1024	全量阶段抽取数据时，从Mysql侧单次请求抽取数据的最大条数，适当增加请求条数可以减少对Mysql的请求次数提升性能。

参数名	类型	默认值	说明
debezium.max.queue.size	int	8192	数据缓存队列条数，默认为8192，当源表中单条数据过大时（如1MB），缓存过多数据会导致内存溢出，可以考虑减小该值。
debezium.max.queue.size.in.bytes	int	0	数据缓存队列大小，默认为0，即表示缓存队列不考虑数据大小，只按照数据条数计算。在debezium.max.queue.size无法有效限制内存占用时，考虑显式设置该值来限制缓存数据的大小。
jdbc.properties.socketTimeout	int	300000	全量阶段连接Mysql的socket超时时间，默认为5分钟。当Mysql负载较高，作业出现SocketTimeout异常时，考虑增大该值。
jdbc.properties.connectTimeout	int	60000	全量阶段连接Mysql的连接超时时间，默认为1分钟。当Mysq负载较高，作业出现ConnectTimeout异常时，考虑增大该值。

表 7-34 增量阶段优化参数

参数名	类型	默认值	说明
debezium.max.queue.size	int	8192	数据缓存队列条数，默认为8192，当源表中单条数据过大时（如1MB），缓存过多数据会导致内存溢出，可以考虑减小该值。
debezium.max.queue.size.in.bytes	int	0	数据缓存队列大小，默认为0，即表示缓存队列不考虑数据大小，只按照数据条数计算。在debezium.max.queue.size无法有效限制内存占用时，考虑显式设置该值来限制缓存数据的大小。

目的端优化

Hudi写入优化。

Hudi表写入性能慢，优先审视表设计是否合理，建议使用Hudi Bucket索引的MOR表，并根据实际数据量配置Bucket桶数，以达到Migration写入性能最佳。

说明

- 使用Bucket索引：通过在“Hudi表属性全局配置”或在映射后的单表“表属性编辑”中配置index.type和hoodie.bucket.index.num.buckets属性可进行配置。
- 判断使用分区表还是非分区表。

  根据表的使用场景一般将表分为事实表和维度表：
  - 事实表通常整表数据规模较大，以新增数据为主，更新数据占比小，且更新数据大多落在近一段时间范围内（年或月或天），下游读取该表进行ETL计算时通常会使用时间范围进行裁剪（例如最近一天、一月、一年），这种表通常可以通过数据的创建时间来做分区以保证最佳读写性能。
  - 维度表数据量一般整表数据规模较小，以更新数据为主，新增较少，表数据量比较稳定，且读取时通常需要全量读取做join之类的ETL计算，因此通常使用非分区表性能更好。
- 确认表内桶数。

  使用Hudi BUCKET表时需要设置Bucket桶数，桶数设置关系到表的性能，需要格外引起注意。
  - 非分区表桶数 = MAX（单表数据量大小（G）/2G*2，再向上取整，4）。
  - 分区表桶数 = MAX（单分区数据量大小（G）/2G*2，再后向上取整，1）。

    其中，要注意的是：
  - 需要使用的是表的总数据大小，而不是压缩以后的文件大小。
  - 桶的设置以偶数最佳，非分区表最小桶数请设置4个，分区表最小桶数请设置1个。

同时，可通过在Hudi的目的端配置中单击“Hudi表属性全局配置”或在映射后的单表“表属性编辑”中，添加优化参数。

图 7-136 添加自定义属性

表 7-35 Hudi 写入优化参数

参数名	类型	默认值	说明
hoodie.sink.flush.tasks	int	1	Hudi flush数据时的并发数，默认为1，即顺序写入。当Hud单次commit涉及FleGroup较多时（如源端表较多更新历史数据的场景），考虑增大该值。 已知单线程flush的FileGroup的数据 = 单次Commit的FileGroup数量 / 作业并发数。 单线程flush的FileGroup的数量 <= 5，推荐值2。 单线程flush的FileGroup的数量 <= 10，推荐值5。 单线程flush的FileGroup的数量 <= 25，推荐值10。 单线程flush的FileGroup的数量 <= 50，推荐值20。 单线程flush的FileGroup的数量 > 50，推荐值30。 flush的并发数越大，flush时内存会响应升高，请结合实时处理集成作业内存监控适当调整该值。
hoodie.context.flatmap.parallelism	int	1	Hudi在commit时，会进行分区扫描操作，默认是单并发操作，当Hudi单次commit涉及的分区较多时，考虑增大该值以提升commit速度。 单次Commit的分区数量 <= 10，推荐值5。 单次Commit的分区数量 <= 25，推荐值10。 单次Commit的分区数量 <= 50，推荐值20。 单次Commit的分区数量 > 50，推荐值30。
compaction.async.enabled	boolean	true	是否开启compaction，默认为true，即默认开启hudi的compaction操作。compaction操作一定程度会影响实时任务的写入性能，为了保证Migration作业的稳定性可以考虑设置为false关闭compaction操作，将Hudi Compaction单独拆成Spark作业交由MRS执行，具体可以参考**如何配置Hudi Compaction的Spark周期任务？**。
compaction.delta_commits	int	5	实时处理集成生成compaction request的频率，默认为5时，即每5次commit生成一个compaction request。compaction request生成频率降低可以使得compaction频率降低从而提升作业性能。如果hudi增量数据较小。可以考虑增大该值。

7.9.4 MySQL 到 DWS 参数调优

源端优化

MySQL抽取优化。

可通过在作业任务配置参数单击中“添加自定义属性”来新增MySQL同步参数。

图 7-137 添加自定义属性

可使用的调优参数具体如下：

表 7-36 全量阶段优化参数

参数名	类型	默认值	说明
scan.incremental.snapshot.backfill.skip	boolean	true	全量阶段是否跳过读取binlog数据，默认为true。跳过读取binlog数据可以有效降低内存使用。需要注意的是，跳过读取binlog功能只提供at-least-once保证。
scan.incremental.snapshot.chunk.size	int	50000	分片大小，决定了全量阶段单个分片最大数据的数据条数以及分片个数。分片大小越大，单个分片数据条数越多，分片个数越小。 当表的条数过多时，作业会划分较多的分片，从而占用过多的内存导致内存问题，请解决表的条数适当调整该值。 当scan.incremental.snapshot.backfill.skip为false时，实时处理集成作业会缓存单个分片的数据，此时分片越大，占用内存越多，引发内存溢出，在此场景下，可以考虑降低分片大小。
scan.snapshot.fetch.size	int	1024	全量阶段抽取数据时，从Mysql侧单次请求抽取数据的最大条数，适当增加请求条数可以减少对Mysql的请求次数提升性能。
debezium.max.queue.size	int	8192	数据缓存队列条数，默认为8192，当源表中单条数据过大时（如1MB），缓存过多数据会导致内存溢出，可以考虑减小该值。
debezium.max.queue.size.in.bytes	int	0	数据缓存队列大小，默认为0，即表示缓存队列不考虑数据大小，只按照数据条数计算。在debezium.max.queue.size无法有效限制内存占用时，考虑显式设置该值来限制缓存数据的大小。

参数名	类型	默认值	说明
jdbc.properties.socketTimeout	int	300000	全量阶段连接Mysql的socket超时时间，默认为5分钟。当Mysql负载较高，作业出现SocketTimeout异常时，考虑增大该值。
jdbc.properties.connectTimeout	int	60000	全量阶段连接Mysql的连接超时时间，默认为1分钟。当Mysq负载较高，作业出现ConnectTimeout异常时，考虑增大该值。

表 7-37 增量阶段优化参数

参数名	类型	默认值	说明
debezium.max.queue.size	int	8192	数据缓存队列条数，默认为8192，当源表中单条数据过大时（如1MB），缓存过多数据会导致内存溢出，可以考虑减小该值。
debezium.max.queue.size.in.bytes	int	0	数据缓存队列大小，默认为0，即表示缓存队列不考虑数据大小，只按照数据条数计算。在debezium.max.queue.size无法有效限制内存占用时，考虑显式设置该值来限制缓存数据的大小。

目的端优化

DWS写入优化。

可通过在DWS的目的端配置中修改写入相关配置，且可以通过单击高级配置的“查看编辑”按钮，添加高级属性。

图 7-138 添加高级属性

表 7-38 DWS 写入优化参数

参数名	类型	默认值	说明
写入模式	enum	UPSERT	DWS的写入模式，可在目的端配置中设置，实时处理集成作业推荐使用COPY MODE。 ● UPSERT：为批量更新入库模式。 ● COPY：为DWS专有的高性能批量入库模式。
批写最大数据量	int	50000	DWS单次写入的最大条数，可在目的端配置中设置。 当缓存的数据达到“批写最大数据量”和“定时批写时间间隔”之一的条件时，触发数据写入。 单次写入条数增大可以减少请求DWS的次数，但可能导致单次请求时长增加，同时也可能导致缓存的数据增加进而影响内存使用。请综合考虑DWS规格和负载， 适当调整该值。
定时批写时间间隔	int	3	DWS单次写入的时间间隔，可在目的端配置中设置。 当缓存的数据达到定时批写时间间隔的条件，触发数据写入。 增大该值有助于增加单次写入时缓存的数据条数，但由于写入频率降低，会提升DWS数据可见的时延。
sink.buffer-flush.max-size	int	512	DWS单次写入的数据大小，默认为512MB，可在目的端配置的高级配置中设置。 当缓存的数据达到数据大小限制时，触发数据写入。 与批写最大数据量类似，单次写入大小增大可以减少请求DWS的次数，但可能导致单次请求时长增加，同时也可能导致缓存的数据增加进而影响内存使用。请综合考虑DWS规格和负载， 适当调整该值。

7.9.5 MySQL 到 DMS Kafka 参数调优

源端优化

MySQL抽取优化。

可通过在作业任务配置参数单击中“添加自定义属性”来新增MySQL同步参数。

图 7-139 添加自定义属性

可使用的调优参数具体如下：

表 7-39 全量阶段优化参数

参数名	类型	默认值	说明
scan.incremental.snapshot.backfill.skip	boolean	true	全量阶段是否跳过读取binlog数据，默认为true。跳过读取binlog数据可以有效降低内存使用。需要注意的是，跳过读取binlog功能只提供at-least-once保证。
scan.incremental.snapshot.chunk.size	int	50000	分片大小，决定了全量阶段单个分片最大数据的数据条数以及分片个数。分片大小越大，单个分片数据条数越多，分片个数越小。 当表的条数过多时，作业会划分较多的分片，从而占用过多的内存导致内存问题，请解决表的条数适当调整该值。 当scan.incremental.snapshot.backfill.skip为false时，实时处理集成作业会缓存单个分片的数据，此时分片越大，占用内存越多，引发内存溢出，在此场景下，可以考虑降低分片大小。
scan.snapshot.fetch.size	int	1024	全量阶段抽取数据时，从Mysql侧单次请求抽取数据的最大条数，适当增加请求条数可以减少对Mysql的请求次数提升性能。
debezium.max.queue.size	int	8192	数据缓存队列条数，默认为8192，当源表中单条数据过大时（如1MB），缓存过多数据会导致内存溢出，可以考虑减小该值。
debezium.max.queue.size.in.bytes	int	0	数据缓存队列大小，默认为0，即表示缓存队列不考虑数据大小，只按照数据条数计算。在debezium.max.queue.size无法有效限制内存占用时，考虑显式设置该值来限制缓存数据的大小。

参数名	类型	默认值	说明
jdbc.properties.socketTimeout	int	300000	全量阶段连接Mysql的socket超时时间，默认为5分钟。当Mysql负载较高，作业出现SocketTimeout异常时，考虑增大该值。
jdbc.properties.connectTimeout	int	60000	全量阶段连接Mysql的连接超时时间，默认为1分钟。当Mysq负载较高，作业出现ConnectTimeout异常时，考虑增大该值。

表 7-40 增量阶段优化参数

参数名	类型	默认值	说明
debezium.max.queue.size	int	8192	数据缓存队列条数，默认为8192，当源表中单条数据过大时（如1MB），缓存过多数据会导致内存溢出，可以考虑减小该值。
debezium.max.queue.size.in.bytes	int	0	数据缓存队列大小，默认为0，即表示缓存队列不考虑数据大小，只按照数据条数计算。在debezium.max.queue.size无法有效限制内存占用时，考虑显式设置该值来限制缓存数据的大小。

目的端优化

Kafka写入优化。

Kafka写入通常速率极快，若有阻塞的场景请优先增加并发解决。

7.9.6 DMS Kafka 到 OBS 参数调优

源端优化

Kafka抽取优化。

可通过在源端配置中单击“Kafka源端属性配置”来添加Kafka优化配置。

图 7-140 添加自定义属性

可使用的调优参数具体如下：

表 7-41 全量阶段优化参数

参数名	类型	默认值	说明
properties.fetch.max.bytes	int	57671680	消费Kafka时每次fetch请求返回的最大字节数。Kafka单条消息大的场景，可以适当调高每次获取的数据量，以提高性能。
properties.max.partition.fetch.bytes	int	1048576	消费Kafka时服务器将返回的每个分区的最大字节数。Kafka单条消息大的场景，可以适当调高每次获取的数据量，以提高性能。
properties.max.poll.records	int	500	消费者每次poll时返回的最大消息条数。Kafka单条消息大的场景，可以适当调高每次获取的数据量，以提高性能。

目的端优化

OBS写入优化。

若开启了自动合并可尝试关闭，否则请优先增加并发解决。

7.9.7 Apache Kafka 到 MRS Kafka 参数调优

源端优化

Kafka抽取优化。

可通过在源端配置中单击“Kafka源端属性配置”来添加Kafka优化配置。

图 7-141 添加自定义属性

可使用的调优参数具体如下：

表 7-42 全量阶段优化参数

参数名	类型	默认值	说明
properties.fetch.max.bytes	int	57671680	消费Kafka时每次fetch请求返回的最大字节数。Kafka单条消息大的场景，可以适当调高每次获取的数据量，以提高性能。

参数名	类型	默认值	说明
properties.max.partition.fetch.bytes	int	1048576	消费Kafka时服务器将返回的每个分区的最大字节数。Kafka单条消息大的场景，可以适当调高每次获取的数据量，以提高性能。
properties.max.poll.records	int	500	消费者每次poll时返回的最大消息条数。Kafka单条消息大的场景，可以适当调高每次获取的数据量，以提高性能。

目的端优化

Kafka写入优化。

Kafka写入通常速率极快，若有阻塞的场景请优先增加并发解决。

7.9.8 SQLServer 到 MRS Hudi 参数调优

源端优化

SQLServer抽取优化。

可通过在作业任务配置参数单击中“添加自定义属性”来新增SQLServer同步参数。

图 7-142 添加自定义属性

可使用的调优参数具体如下：

表 7-43 全量阶段优化参数

参数名	类型	默认值	说明
scan.incremental.snapshot.backfill.skip	boolean	true	全量阶段是否跳过读取binlog数据，默认为true。跳过读取binlog数据可以有效降低内存使用。需要注意的是，跳过读取binlog功能只提供at-least-once保证。

表 7-44 增量阶段优化参数

参数名	类型	默认值	说明
debezium.max.iteration.transactions	int	1000	每张表在重演数据时每次抽取的数据条数，值较大时，会使得内存升高并阻塞增量同步任务。

目的端优化

Hudi写入优化。

Hudi表写入性能慢，优先审视表设计是否合理，建议使用Hudi Bucket索引的MOR表，并根据实际数据量配置Bucket桶数，以达到Migration写入性能最佳。

说明

- 使用Bucket索引：通过在“Hudi表属性全局配置”或在映射后的单表“表属性编辑”中配置index.type和hoodie.bucket.index.num.buckets属性可进行配置。
- 判断使用分区表还是非分区表。

  根据表的使用场景一般将表分为事实表和维度表：
  - 事实表通常整表数据规模较大，以新增数据为主，更新数据占比小，且更新数据大多落在近一段时间范围内（年或月或天），下游读取该表进行ETL计算时通常会使用时间范围进行裁剪（例如最近一天、一月、一年），这种表通常可以通过数据的创建时间来做分区以保证最佳读写性能。
  - 维度表数据量一般整表数据规模较小，以更新数据为主，新增较少，表数据量比较稳定，且读取时通常需要全量读取做join之类的ETL计算，因此通常使用非分区表性能更好。
- 确认表内桶数。

  使用Hudi BUCKET表时需要设置Bucket桶数，桶数设置关系到表的性能，需要格外引起注意。
  - 非分区表桶数 = MAX（单表数据量大小（G）/2G*2，再向上取整，4）。
  - 分区表桶数 = MAX（单分区数据量大小（G）/2G*2，再后向上取整，1）。

    其中，要注意的是：
  - 需要使用的是表的总数据大小，而不是压缩以后的文件大小。
  - 桶的设置以偶数最佳，非分区表最小桶数请设置4个，分区表最小桶数请设置1个。

同时，可通过在Hudi的目的端配置中单击“Hudi表属性全局配置”或在映射后的单表“表属性编辑”中，添加优化参数。

图 7-143 添加自定义属性

表 7-45 Hudi 写入优化参数

参数名	类型	默认值	说明
hoodie.sink.flush.tasks	int	1	Hudi flush数据时的并发数，默认为1，即顺序写入。当Hud单次commit涉及FleGroup较多时（如源端表较多更新历史数据的场景），考虑增大该值。 已知单线程flush的FileGroup的数据 = 单次Commit的FileGroup数量 / 作业并发数。 单线程flush的FileGroup的数量 <= 5，推荐值2。 单线程flush的FileGroup的数量 <= 10，推荐值5。 单线程flush的FileGroup的数量 <= 25，推荐值10。 单线程flush的FileGroup的数量 <= 50，推荐值20。 单线程flush的FileGroup的数量 > 50，推荐值30。 flush的并发数越大，flush时内存会响应升高，请结合实时处理集成作业内存监控适当调整该值。
hoodie.context.flatmap.parallelism	int	1	Hudi在commit时，会进行分区扫描操作，默认是单并发操作，当Hudi单次commit涉及的分区较多时，考虑增大该值以提升commit速度。 单次Commit的分区数量 <= 10，推荐值5。 单次Commit的分区数量 <= 25，推荐值10。 单次Commit的分区数量 <= 50，推荐值20。 单次Commit的分区数量 > 50，推荐值30。
compaction.async.enabled	boolean	true	是否开启compaction，默认为true，即默认开启hudi的compaction操作。compaction操作一定程度会影响实时任务的写入性能，为了保证Migration作业的稳定性可以考虑设置为false关闭compaction操作，将Hudi Compaction单独拆成Spark作业交由MRS执行，具体可以参考**如何配置Hudi Compaction的Spark周期任务?** 。
compaction.delta_commits	int	5	实时处理集成生成compaction request的频率，默认为5时，即每5次commit生成一个compaction request。compaction request生成频率降低可以使得compaction频率降低从而提升作业性能。如果hudi增量数据较小。可以考虑增大该值。

7.9.9 PostgreSQL 到 DWS 参数调优

源端优化

PostgreSQL抽取优化。

暂无优化配置项。

目的端优化

DWS写入优化。

可通过在DWS的目的端配置中修改写入相关配置，且可以通过单击高级配置的“查看编辑”按钮，添加高级属性。

图 7-144 添加高级属性

表 7-46 DWS 写入优化参数

参数名	类型	默认值	说明
写入模式	enum	UPSERT	DWS的写入模式，可在目的端配置中设置，实时处理集成作业推荐使用COPY MODE。 ● UPSERT：为批量更新入库模式。 ● COPY：为DWS专有的高性能批量入库模式。
批写最大数据量	int	50000	DWS单次写入的最大条数，可在目的端配置中设置。 当缓存的数据达到“批写最大数据量”和“定时批写时间间隔”之一的条件时，触发数据写入。 单次写入条数增大可以减少请求DWS的次数，但可能导致单次请求时长增加，同时也可能导致缓存的数据增加进而影响内存使用。请综合考虑DWS规格和负载， 适当调整该值。

参数名	类型	默认值	说明
定时批写时间间隔	int	3	DWS单次写入的时间间隔，可在目的端配置中设置。 当缓存的数据达到定时批写时间间隔的条件，触发数据写入。 增大该值有助于增加单次写入时缓存的数据条数，但由于写入频率降低，会提升DWS数据可见的时延。
sink.buffer-flush.max-size	int	512	DWS单次写入的数据大小，默认为512MB，可在目的端配置的高级配置中设置。 当缓存的数据达到数据大小限制时，触发数据写入。 与批写最大数据量类似，单次写入大小增大可以减少请求DWS的次数，但可能导致单次请求时长增加，同时也可能导致缓存的数据增加进而影响内存使用。请综合考虑DWS规格和负载， 适当调整该值。

7.9.10 Oracle 到 DWS 参数调优

源端优化

Oracle抽取优化。

暂无优化配置项。

目的端优化

DWS写入优化。

可通过在DWS的目的端配置中修改写入相关配置，且可以通过单击高级配置的“查看编辑”按钮，添加高级属性。

图 7-145 添加高级属性

表 7-47 DWS 写入优化参数

参数名	类型	默认值	说明
写入模式	enum	UPSERT	DWS的写入模式，可在目的端配置中设置，实时处理集成作业推荐使用COPY MODE。 ● UPSERT：为批量更新入库模式。 ● COPY：为DWS专有的高性能批量入库模式。
批写最大数据量	int	50000	DWS单次写入的最大条数，可在目的端配置中设置。 当缓存的数据达到“批写最大数据量”和“定时批写时间间隔”之一的条件时，触发数据写入。 单次写入条数增大可以减少请求DWS的次数，但可能导致单次请求时长增加，同时也可能导致缓存的数据增加进而影响内存使用。请综合考虑DWS规格和负载， 适当调整该值。
定时批写时间间隔	int	3	DWS单次写入的时间间隔，可在目的端配置中设置。 当缓存的数据达到定时批写时间间隔的条件，触发数据写入。 增大该值有助于增加单次写入时缓存的数据条数，但由于写入频率降低，会提升DWS数据可见的时延。
sink.buffer-flush.max-size	int	512	DWS单次写入的数据大小，默认为512MB，可在目的端配置的高级配置中设置。 当缓存的数据达到数据大小限制时，触发数据写入。 与批写最大数据量类似，单次写入大小增大可以减少请求DWS的次数，但可能导致单次请求时长增加，同时也可能导致缓存的数据增加进而影响内存使用。请综合考虑DWS规格和负载， 适当调整该值。

7.9.11 Oracle 到 MRS Hudi 参数调优

源端优化

Oracle抽取优化。

暂无优化配置项。

目的端优化

Hudi写入优化。

Hudi表写入性能慢，优先审视表设计是否合理，建议使用Hudi Bucket索引的MOR表，并根据实际数据量配置Bucket桶数，以达到Migration写入性能最佳。

说明

- 使用Bucket索引：通过在“Hudi表属性全局配置”或在映射后的单表“表属性编辑”中配置index.type和hoodie.bucket.index.num.buckets属性可进行配置。
- 判断使用分区表还是非分区表。

  根据表的使用场景一般将表分为事实表和维度表：
  - 事实表通常整表数据规模较大，以新增数据为主，更新数据占比小，且更新数据大多落在近一段时间范围内（年或月或天），下游读取该表进行ETL计算时通常会使用时间范围进行裁剪（例如最近一天、一月、一年），这种表通常可以通过数据的创建时间来做分区以保证最佳读写性能。
  - 维度表数据量一般整表数据规模较小，以更新数据为主，新增较少，表数据量比较稳定，且读取时通常需要全量读取做join之类的ETL计算，因此通常使用非分区表性能更好。
- 确认表内桶数。

  使用Hudi BUCKET表时需要设置Bucket桶数，桶数设置关系到表的性能，需要格外引起注意。
  - 非分区表桶数 = MAX（单表数据量大小（G）/2G*2，再向上取整，4）。
  - 分区表桶数 = MAX（单分区数据量大小（G）/2G*2，再后向上取整，1）。

    其中，要注意的是：
  - 需要使用的是表的总数据大小，而不是压缩以后的文件大小。
  - 桶的设置以偶数最佳，非分区表最小桶数请设置4个，分区表最小桶数请设置1个。

同时，可通过在Hudi的目的端配置中单击“Hudi表属性全局配置”或在映射后的单表“表属性编辑”中，添加优化参数。

图 7-146 添加自定义属性

表 7-48 Hudi 写入优化参数

参数名	类型	默认值	说明
hoodie.sink.flush.tasks	int	1	Hudi flush数据时的并发数，默认为1，即顺序写入。当Hud单次commit涉及FleGroup较多时（如源端表较多更新历史数据的场景），考虑增大该值。 已知单线程flush的FileGroup的数据 = 单次Commit的FileGroup数量 / 作业并发数。 单线程flush的FileGroup的数量 <= 5，推荐值2。 单线程flush的FileGroup的数量 <= 10，推荐值5。 单线程flush的FileGroup的数量 <= 25，推荐值10。 单线程flush的FileGroup的数量 <= 50，推荐值20。 单线程flush的FileGroup的数量 > 50，推荐值30。 flush的并发数越大，flush时内存会响应升高，请结合实时处理集成作业内存监控适当调整该值。
hoodie.context.flatmap.parallelism	int	1	Hudi在commit时，会进行分区扫描操作，默认是单并发操作，当Hudi单次commit涉及的分区较多时，考虑增大该值以提升commit速度。 单次Commit的分区数量 <= 10，推荐值5。 单次Commit的分区数量 <= 25，推荐值10。 单次Commit的分区数量 <= 50，推荐值20。 单次Commit的分区数量 > 50，推荐值30。
compaction.async.enabled	boolean	true	是否开启compaction，默认为true，即默认开启hudi的compaction操作。compaction操作一定程度会影响实时任务的写入性能，为了保证Migration作业的稳定性可以考虑设置为false关闭compaction操作，将Hudi Compaction单独拆成Spark作业交由MRS执行，具体可以参考**如何配置Hudi Compaction的Spark周期任务？**。
compaction.delta_commits	int	5	实时处理集成生成compaction request的频率，默认为5时，即每5次commit生成一个compaction request。compaction request生成频率降低可以使得compaction频率降低从而提升作业性能。如果hudi增量数据较小。可以考虑增大该值。

7.10 使用教程

7.10.1 概览

本章节汇总了基于Migration实时数据集成服务常见应用场景的操作指导，每个实践我们提供了详细的方案描述和操作指导，用于指导您快速实现数据库迁移和同步。

表 7-49 Migration 基础实践一览表

数据源分类	源端数据源	对应目的端数据源	相关文档
关系型数据	MySQL	Hadoop：MRS Hudi	MySQL同步到MRS Hudi作业配置
		消息系统：DMS Kafka	MySQL同步到Kafka作业配置
		数据仓库：DWS	MySQL同步到DWS作业配置
	SQLServer	Hadoop：MRS Hudi **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	SQLServer同步到MRS Hudi作业配置
	PostgreSQL	数据仓库：DWS **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	PostgreSQL同步到DWS作业配置
	Oracle	数据仓库：DWS **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	Oracle同步到DWS作业配置
		Hadoop：MRS Hudi **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	Oracle同步到MRS Hudi作业配置
消息系统	DMS Kafka	对象存储：OBS	DMS Kafka同步到OBS作业配置
	Apache Kafka	Hadoop：MRS Kafka **说明** 该链路目前需申请白名单后才能使用。如需使用该链路，请联系客服或技术支持人员。	Apache Kafka同步到MRS Kafka作业配置

7.10.2 DRS 任务切换到实时 Migration 作业配置

DRS（数据复制服务）任务迁移到Migration，完成作业切换和数据续传。

前提条件

- 已创建DRS实时同步任务，DRS相关操作请参见**数据复制服务**。
- 已按**使用前自检概览**准备好实时数据集成环境。

准备动作

- **Migration所需资源估算。**

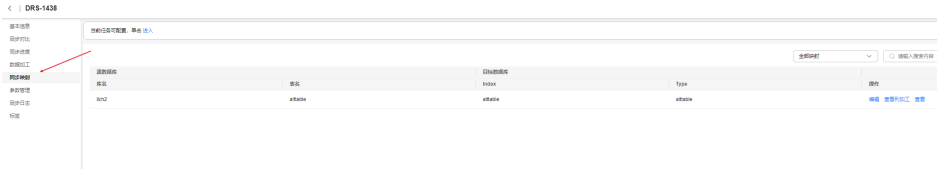
  根据业务实际情况估算DRS作业迁移到Migration后，Migration大概需要多少资源承载新作业，规划作业的拆分和创建。资源不够的情况下请购买新资源组。

  资源估算维度包括：

  - DRS任务表数量

    进入DRS任务，查看同步映射可以看到表数量。Migration单个作业表数量配置在50张以内性能最佳。

    图 7-147 查看 DRS 任务表数量

  - 同步流量查看

    进入DRS任务监控，查看监控指标，以“写目标库频率”为主要评估指标，同时观察DRS任务是否有时延。

    Migration在配置8CU的情况下可以支撑8000条/秒的同步速率。流量较大的表建议单独配置作业。

    图 7-148 查看监控指标

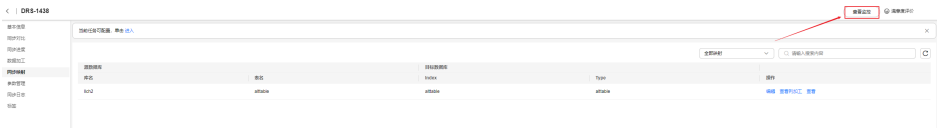
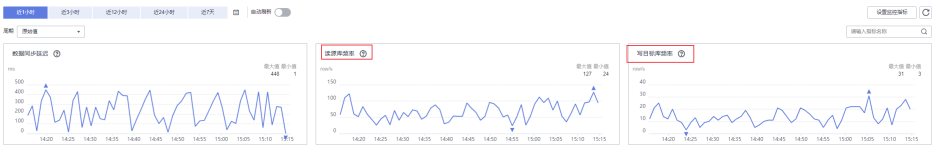
    

    图 7-149 查看指标详情

  - 参考客户建议，根据客户业务需求创建作业。

- **网络打通**

  Migration资源组需要打通数据源的网络连通。在DRS任务的基本信息中查看数据源配置，根据Migration网络打通教程完成网络打通。

图 7-150 查看数据源配置

Migration 作业创建与启动

步骤1 创建作业。

根据准备好的作业拆分方案创建Migration作业，暂不启动作业。

步骤2 获取DRS安全位点。

Migration作业需要根据DRS的同步位点启动作业，做到数据的续传、不漏数。

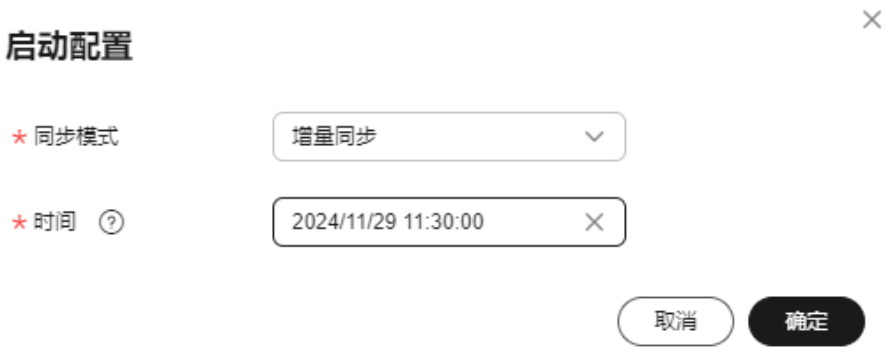
请联系DRS运维人员获取到DRS任务同步的安全位点（一个binlog文件名），联系MySQL数据库运维根据安全位点查询出当前DRS已同步binlog的时间戳，根据这个时间戳启动Migration作业。

步骤3 在启动Migration作业前务必将DRS作业暂停，避免造成写冲突。

根据查询到的安全位点时间启动Migration作业，设置Migration启动位点时可以比安全位点时间更早一点（建议30min左右），避免丢数。

例如，查询到的DRS安全位点时间戳为2024-11-29 12:00:00，启动Migration作业时可以将位点配置为2024-11-29 11:30:00。

图 7-151 设置 Migration 启动位点



Migration作业启动后，观察作业监控，确定Migration稳定后可以适时停止DRS作业。

----结束

7.10.3 MySQL 同步到 MRS Hudi 作业配置

支持的源端和目的端数据库版本

表 7-50 支持的数据库版本

源端数据库	目的端数据库
MySQL数据库（5.6、5.7、8.x版本）	• MRS集群（3.2.0-LTS.x、3.5.x） • Hudi版本（0.11.0）

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考**表7-51**进行赋权。

表 7-51 数据库账号权限

类型名称	权限要求
源数据库连接账号	需要具备如下最小权限：SELECT、SHOW DATABASES、REPLICATION SLAVE、REPLICATION CLIENT，即执行SQL： `GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '用户名'@'%';`
目标数据库连接账号	MRS用户需要拥有Hadoop和Hive组件的读写权限，建议参照下图所示角色及用户组配置MRS用户。 **图 7-152** MRS Hudi 最小化权限  具体MRS集群角色权限管理请参考《**MRS集群用户权限模型**》。

说明

- 建议创建单独用于Migration任务连接的数据库账号，避免因为数据库账号密码修改，导致的任务连接失败。
- 连接源和目标数据库的账号密码修改后，请同步修改管理中心对应的连接信息，避免任务连接失败后自动重试，导致数据库账号被锁定影响使用。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考表7-52。

表 7-52 同步对象范围

类型名称	使用须知
同步对象范围	• 支持同步DML：包括INSERT、UPDATE、DELETE。 • 支持同步的DDL：新增列。 • 仅支持同步主键表。 • 仅支持同步MyISAM和InnoDB表。 • 不支持同步视图、外键、存储过程、触发器、函数、事件、虚拟列、唯一约束和唯一索引。 • 自动建表支持同步表结构、普通索引、约束（主键、空、非空）、注释。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见表7-53。

表 7-53 注意事项

类型名称	使用和操作限制
数据库限制	目标数据库中的库名、表名、字段名仅支持数字、字母和下划线，且字段名必须以字母或下划线开头，建议尽量使用常规字符避免任务失败。

类型名称	使用和操作限制
使用限制	通用： ● 实时同步过程中，不支持IP、端口、账号、密码修改。 ● 不允许源数据库进行恢复操作。 ● 建议MySQL Binlog保留3天以上，不支持强制清理Binlog。 异常/暂停恢复作业时，记录的Binlog位点过期会导致作业恢复失败，需要关注作业异常/暂停时长及Binlog保留时长。 ● 实时同步过程中，不允许源数据库MySQL跨大版本升级，否则可能导致数据不一致或者同步任务失败（跨版本升级后数据、表结构、关键字等信息均可能会产生兼容性改变），建议在该场景下重建同步任务。 ● Hudi表使用Bucket索引的场景下不允许更新分区键，否则可能产生重复数据。 ● Hudi表使用Bucket索引的场景下仅保证单分区内主键唯一。 ● 本链路所使用的Hudi表需带有3个审计字段：cdc_last_update_date、logical_is_deleted、_hoodie_event_time，并会以_hoodie_event_time作为Hudi表的预聚合键。因此，若使用已存在的表，也需要携带这3个审计字段，否则可能导致任务异常。 – cdc_last_update_date：Migration任务处理CDC数据的时间。 – logical_is_deleted：逻辑删除标志。 – _hoodie_event_time：数据在MySQL Binlog中的时间戳。 全量同步阶段： ● 任务启动和全量数据同步阶段，请不要在源数据库执行DDL操作，否则可能导致任务异常。 ● 当前全量同步无法覆盖Hudi表中的存量数据，建议全量同步前先清空Hudi表。 增量同步阶段： ● 增量同步过程中，不支持指定位置加列的DDL操作（例如ALTER TABLE ddl_test ADD COLUMN c2 AFTER/FIRST c1;），Migration会删除AFTER/FIRST属性，可能会导致列顺序不一致。 ● 增量同步过程中，执行不幂等的DDL可能导致数据不一致（例如ALTER TABLE ddl_test ADD COLUMN c3 timestamp default now();），Migration会因数据库函数执行结果不幂等导致最终数据不一致。 ● 增量同步过程中，可识别的DDL类型有新建表、删除表、新增列、删除列、重命名表、重命名列、修改列类型、清空表，当前仅支持同步新增列操作到目的端Hudi，其余DDL可配置成忽略/异常。 – 分库分表场景下，执行新增列时，需保证每张表加列的类型一致，否则有可能导致任务失败。 – 新增列名时不能超出256字符，否则任务会失败。 ● 对于版本低于3.2.0-LTS1.5的MRS集群的Hudi MOR表，使用Migration（Flink）增量同步数据后，不支持直接使用CDM或Spark SQL写入数据，需要先进行Compaction后方可写入。 常见故障排查：

类型名称	使用和操作限制
	在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。
其他限制	支持目标数据库中的表比源数据库多列场景，但是需要避免以下场景可能导致的任务失败。 目标数据库多的列要求非空且没有默认值，源数据库insert数据，同步到目标数据库后多的列为null，不符合目标数据库要求。

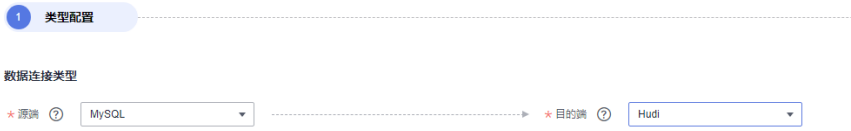
操作步骤

本小节以RDS for MySQL到MRS Hudi的实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

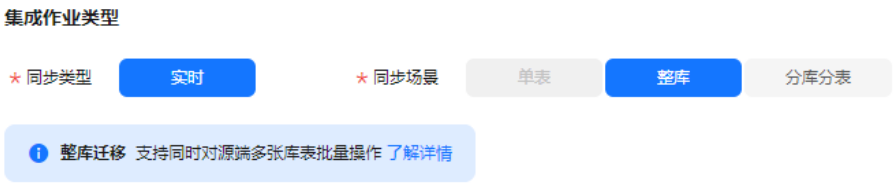
步骤2 选择数据连接类型：源端选MySQL，目的端选Hudi。

图 7-153 选择数据连接类型



步骤3 选择集成作业类型：同步类型默认为实时，同步场景包含整库和分库分表场景。

图 7-154 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的MySQL、MRS Hudi数据连接和已配置好网络连接的资源组。

图 7-155 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

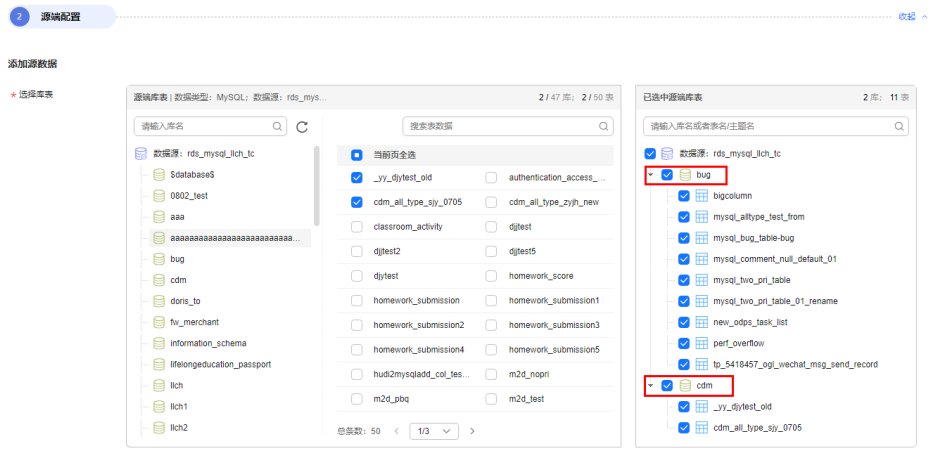
说明

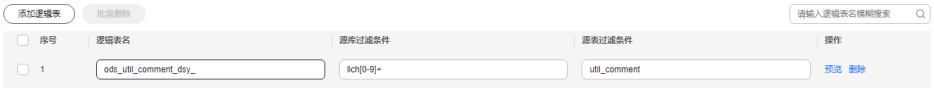
网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端参数。

各同步场景下选择需要同步库表的方式请参考**表7-54**。

表 7-54 选择需要同步的库表

同步场景	配置方式
整库	选择需要迁移的MySQL库表。 **图 7-156** 选择库表  库与表均支持自定义选择，即可选择一库一表，也可选择多库多表。

同步场景	配置方式
分库分表	添加逻辑表。 ● 逻辑表名：即最终写入到Hudi的表名。 ● 源库过滤条件：支持填入正则表达式，在所有MySQL实例中通过该正则表达式过滤出要写入目标端Hudi汇聚表的所有分库。 ● 源表过滤条件：支持填入正则表达式，在过滤出的源端分库中再次过滤出要写入目标端Hudi汇聚表的所有分表。 **图 7-157** 添加逻辑表  已添加的逻辑表支持预览表结构及来源库表，单击“操作”列的预览即可。预览逻辑表时，源表数量越多，等待时间可能越长，请耐心等待。 **图 7-158** 逻辑表预览 

步骤7 配置目的端参数。

- 源库表和目标匹配策略。
  各同步场景下源端库表和目标端库表的匹配策略请参考**表7-55**。

表 7-55 源库表和目标匹配策略

同步场景	配置方式
整库	– 库匹配策略。 ▪ 与来源库同名：数据将同步至与来源MySQL库名相同的Hudi库中。 ▪ 自定义：数据将同步至自行指定的Hudi库中。 – 表匹配策略。 ▪ 与来源表同名：数据将同步至与来源MySQL表名相同的Hudi表中。 ▪ 自定义：数据将同步至自行指定的Hudi表中。 **图 7-159** 整库场景下源库表和目标匹配策略  **说明** 自定义匹配策略时，支持用内置变量#{source_db_name}和#{source_table_name}标志来源的库名和表名，其中表匹配策略必须包含#{source_table_name}。
分库分表	– 目标端库名：数据将同步至指定的Hudi库中。 – 表匹配策略：默认与源端配置中填写的逻辑表同名。 **图 7-160** 分库分表场景下源库表和目标匹配策略 

- Hudi参数配置。
  其余Hudi目的端参数说明请参考表7-56。

图 7-161 Hudi 目的端配置项

表 7-56 Hudi 目的端配置项

配置项	默认值	单位	配置说明
数据存储路径	-	-	Hudi自动建表时的warehouse路径，每张表会在warehouse路径下创建子目录。支持填写HDFS和OBS路径，路径格式参考： – OBS路径：obs://bucket/warehouse。 – HDFS路径：/tmp/warehouse。
Hudi表属性全局配置	-	-	支持通过参数配置部分高级功能，参数详情可参考**Hudi高级配置一览表**。
Compaction作业	-	-	需要一个独立的SparkSql作业，不使用则由Flink执行compaction。

表 7-57 Hudi 高级配置一览表

参数名	参数类型	默认值	单位	参数说明
index.type	string	BLOOM	-	Hudi表索引类型。 支持BLOOM和BUCKET索引，数据量较大场景下强烈建议使用BUCKET索引性能更好。
hoodie.bucket.index.num.buckets	int	256	个	Hudi表单分区下Bucket桶数。 **说明** 使用Hudi BUCKET表时需要设置Bucket桶数，桶数设置关系到表的性能，需要格外引起注意。 – 非分区表桶数 = MAX（单表数据量大小（G）/2G*2，再向上取整，4）。 – 分区表桶数 = MAX（单分区数据量大小（G）/2G*2，再后向上取整，1）。 其中，要注意的是： – 需要使用的是表的总数据大小，而不是压缩以后的文件大小。 – 桶的设置以偶数最佳，非分区表最小桶数请设置4个，分区表最小桶数请设置1个。

参数名	参数类型	默认值	单位	参数说明
changelog.enabled	boolean	false	-	Hudi changelog功能开关，开启后Migration作业可输出DELETE和UPDATE BEFORE数据。
logical.delete.enabled	boolean	true	-	逻辑删除开关，changelog开启时必须关闭逻辑删除。
hoodie.write.liststatus.optimized	boolean	true	-	写log文件时是否开启liststatus优化。涉及到大表和分区数据量多的作业，在启动时list会非常耗时，可能导致作业启动超时，建议关闭。
hoodie.index.liststatus.optimized	boolean	false	-	定位数据时是否开启liststatus优化。涉及到大表和分区数据量多的作业，在启动时list会非常耗时，可能导致作业启动超时，建议关闭。
compaction.async.enabled	boolean	true	-	异步compaction开关。compaction操作一定程度会影响实时任务的写入性能，如果用户使用外置的compaction操作对hudi进行compaction，可以考虑设置为false关闭实时处理集成作业的compaction操作。
compaction.schedule.enabled	boolean	true	-	生成compaction计划的开关。compaction计划必须由本服务生成，计划的执行可以交给Spark。

参数名	参数类型	默认值	单位	参数说明
compaction.delta_commits	int	5	次	生成compaction request的频率。compaction request生成频率降低可以使得compaction频率降低从而提升作业性能。如果hudi增量数据较小。可以考虑增大该值。 **说明** 例如配置为40，即每40次commit生成一个compaction request，因为Migration每分钟生成1个commit，那么每个compaction request将间隔40分钟。
clean.async.enabled	boolean	true	-	做历史版本数据文件清理的开关。
clean.retain_commits	int	30	次	要保留的commit数。这些commit关联的数据文件版本将被保留 num_of_commits * time_between_commits 这么长的时间，建议配置为2倍的compaction.delta_commits。 **说明** 例如配置为80，因为Migration每分钟生成1个commit，那么超过80分钟后如果有旧版本数据文件 ，则会生成clean request，且在执行clean时保留最近80个commit。
hoodie.archive.automatic	boolean	true	-	Hudi commit文件老化开关。
archive.min_commits	int	40	次	将旧版commit归档到日志文件中时要保留不归档的最小commit数。 建议配置成clean.retain_commits + 1。 **说明** 例如配置成81，那么在触发归档动作时，将会保留最近81次commit文件。
archive.max_commits	int	50	次	触发归档动作的commit数。 建议配置成archive.min_commits + 20。 **说明** 例如配置成101，那么将在生成101个commit文件后触发归档commit文件动作。

说明

- 为了达到Migration作业性能最优，建议使用Hudi Bucket索引的MOR表，并根据实际数据量配置Bucket桶数。
- 为了保证Migration作业的稳定性，建议将Hudi Compaction单独拆成Spark作业交由MRS执行，在Migration任务里仅开启生成compaction计划，具体可以参考**如何配置Hudi Compaction的Spark周期任务?** 。

步骤8 刷新源表和目标表映射，检查映射关系是否正确，同时可根据需求修改表属性、添加附加字段，并通过“自动建表”能力在目的端Hudi数据库中建出相应的表。

图 7-162 源表与目标表映射

- 同步主键

  Hudi表必须设置“同步主键”，在源端为非主键表时，必须在字段映射阶段手动勾选主键。

- 表属性编辑

  单击操作列“表属性编辑”可配置Hudi表属性，包含表类型，分区类型及表自定义属性。

图 7-163 Hudi 表属性配置

- 表类型：MERGE_ON_READ、COPY_ON_WRITE。
- 分区类型：无分区、时间分区、自定义分区。

  说明

  其中时间分区需要用户指定一个源端表名，选择一个时间转换格式。

  比如时间分区用户指定一个源端表名src_col_1，选择一个时间转换格式，日（yyyyMMdd）、月（yyyyMM）、年（yyyy），自动建表时会在Hudi表默认创建一个cdc_partition_key的字段，系统会根据配置的时间转换格式将源端字段（src_col_1）的值格式化后写入cdc_partition_key中。

  - 表自定义属性：支持通过参数配置单表的部分高级功能，参数详情可参考Hudi高级配置一览表。

- 附加字段编辑：单击操作列“附加字段编辑”可为目的端的Hudi表中增加自定义字段，同时附加字段也会额外加入到Hudi表的建表中。用户可以在已有的源表字段基础上添加多个附加字段，并自定义字段名、选择字段类型、填写字段值。
  - 字段名称：目的端Hudi表新增字段的名称。
  - 字段类型：目的端Hudi表新增字段的类型。
  - 字段值：目的端Hudi表新增字段的取值来源。

  表 7-58 附加字段取值方式

类型	示例
常量	任意字符。
内置变量	▪ 源端host ip地址：source.host。 ▪ 源端schema名称：mgr.source.schema。 ▪ 源端table名称：mgr.source.table。 ▪ 目的端schema名称：mgr.target.schema。 ▪ 目的端table名称：mgr.target.table。
源表字段	源表中的任一字段。 配置附加字段的取值来源于源表字段时，请注意任务运行过程中不能修改对应源表字段的名称，否则可能导致作业异常。
udf方法	▪ substring(#col, pos[, len])： 截取源端col列的子串, 范围在[pos, pos+len)。 ▪ date_format(#col, time_format[, src_tz, dst_tz])： 将源端col列按time_format格式化, 可选转换时区。 ▪ now([tz])： 获取指定时区的当前时间。 ▪ if(cond_exp, str1, str2)： 满足条件表达式cond_exp时返回str1, 否则返回str2。 ▪ concat(#col[, #str, ...])： 拼接多个参数, 可为源端列或字符串。 ▪ from_unixtime(#col[, time_format])： 将unix时间戳按time_format格式化。 ▪ unix_timestamp(#col[, precision, time_format])： 将时间转成unix时间戳, 可显式定义时间格式及转换后精度。

- 自动建表：单击“自动建表”可按照已配置映射规则在目的端数据库自动建表，成功后表建立方式会显示为使用已有表。

图 7-164 自动建表

说明

- Migration仅支持自动建表，不支持自动建库和模式，需用户自行在目的端手动建出库和模式后再使用本功能建表。
- 自动建表时对应的字段类型映射关系请参见**字段映射关系**章节。
- 自动建出的Hudi表会带有3个审计字段，分别是cdc_last_update_date、logical_is_deleted、_hoodie_event_time，并会以_hoodie_event_time作为Hudi表的预聚合键。

步骤9 配置DDL消息处理规则。

实时集成作业除了能够同步对数据的增删改等DML操作外，也支持对部分表结构变化（DDL）进行同步。针对支持的DDL操作，用户可根据实际需求配置为正常处理/忽略/出错。

- 正常处理：Migration识别到源端库表出现该DDL动作时，作业自动同步到目的端执行该DDL操作。
- 忽略：Migration识别到源端库表出现该DDL动作时，作业忽略该DDL，不同步到目的端表中。
- 出错：Migration识别到源端库表出现该DDL动作时，作业抛出异常。

图 7-165 DDL 配置

步骤10 配置任务属性。

表 7-59 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2

参数	说明	默认值
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-

参数	说明	默认值
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** • 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 • 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤11 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-166 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-167 启动配置

表 7-60 启动配置参数

参数	说明
同步模式	• 增量同步：从指定时间位点开始同步增量数据。 • 全量+增量：先同步全量数据，随后实时同步增量数据。
时间	增量同步需要设置该参数，指示增量同步起始的时间位点。 **说明** 配置的位点时间早于Binlog日志最早时间点时，默认会以日志最早时间点开始消费。

步骤12 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-168 前往监控

----结束

性能调优

若链路同步速度过慢，可参考参见**任务性能调优**章节中对应链路文档进行排查及处理。

7.10.4 MySQL 同步到 DWS 作业配置

支持的源端和目的端数据库版本

表 7-61 支持的数据库版本

源端数据库	目的端数据库
MySQL数据库（5.6、5.7、8.x版本）	DWS集群（8.1.3、8.2.0版本）

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考**表7-62**进行赋权。

表 7-62 数据库账号权限

类型名称	权限要求
源数据库连接账号	需要具备如下最小权限：SELECT、SHOW DATABASES、REPLICATION SLAVE、REPLICATION CLIENT，即执行SQL： `GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '用户名'@'%';`
目标数据库连接账号	目标数据库的每张表必须具有如下权限：INSERT、SELECT、UPDATE、DELETE、CONNECT、CREATE。

说明

- 建议创建单独用于Migration任务连接的数据库账号，避免因为数据库账号密码修改，导致的任务连接失败。
- 连接源和目标数据库的账号密码修改后，请同步修改管理中心对应的连接信息，避免任务连接失败后自动重试，导致数据库账号被锁定影响使用。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考表7-63。

表 7-63 同步对象范围

类型名称	使用须知
同步对象范围	• 支持同步的DML：包括INSERT、UPDATE、DELETE。 • 支持同步的DDL：删除表、新增列、删除列、重命名表、重命名列、修改列类型、清空表。 • 仅支持同步有主键表。 • 仅支持同步MyISAM和InnoDB表。 • 不支持同步视图、外键、存储过程、触发器、函数、事件、虚拟列、唯一约束和唯一索引。 • 自动建表支持同步表结构、普通索引、约束（主键、空、非空）、注释。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见表7-64。

表 7-64 注意事项

类型名称	使用和操作限制
数据库限制	• 源端数据库中的库名、表名、字段名不能包含：.<'>/\"以及非ASCII字符，建议尽量使用常规字符避免任务失败。 • 目的端数据库中的对象名需要满足约束：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、$。

类型名称	使用和操作限制
使用限制	通用： ● 实时同步过程中，不支持IP、端口、账号、密码修改。 ● 不允许源数据库进行恢复操作。 ● 建议MySQL Binlog保留3天以上，不支持强制清理Binlog。异常/暂停恢复作业时，记录的Binlog位点过期会导致作业恢复失败，需要关注作业异常/暂停时长及Binlog保留时长。 ● 实时同步过程中，不允许源数据库MySQL跨大版本升级，否则可能导致数据不一致或者同步任务失败（跨版本升级后数据、表结构、关键字等信息均可能会产生兼容性改变），建议在该场景下重建同步任务。 全量同步阶段： 任务启动和全量数据同步阶段，请不要在源数据库执行DDL操作，否则可能导致任务异常。 增量同步阶段： ● 增量同步过程中，不支持指定位置加列的DDL操作（例如ALTER TABLE ddl_test ADD COLUMN c2 AFTER/FIRST c1;），Migration会删除AFTER/FIRST属性，可能会导致列顺序不一致。 ● 增量同步过程中，执行不幂等的DDL可能导致数据不一致（例如ALTER TABLE ddl_test ADD COLUMN c3 timestamp default now();），Migration会因数据库函数执行结果不幂等导致最终数据不一致。 ● 增量同步过程中，库级同步不支持Online DDL，表级同步目前只支持阿里云DMS产生的Online DDL。 ● 增量同步过程中，支持同步的DDL类型有新建表、删除表、新增列、删除列、重命名表、重命名列、修改列类型、清空表，用户可以根据自身需求选择需要同步的DDL类型。 – 分库分表场景下，执行重命名列操作，必须停业务操作，不然会有数据不一致的风险。 – 分库分表场景下，推荐只同步新增列DDL，其他的DDL同步可能会因为目标表被修改而导致任务失败或数据不一致。 – 分库分表场景下，执行新增列时，需保证每张表加列的类型一致，否则有可能导致任务失败。 – 新增和修改表名、列名时不能超出63字符，否则任务会失败。 – 增量阶段，源数据库执行CHANGE COLUMN修改列信息，如果该列在目标DWS数据库中为分布列，则该语句会可能导致异常，因为DWS不支持修改分布列。 常见故障排查： 在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。

类型名称	使用和操作限制
其他限制	• 支持目标数据库中的表比源数据库多列场景，但是需要避免以下场景可能导致的任务失败。 – 目标数据库多的列要求非空且没有默认值，源数据库insert数据，同步到目标数据库后多的列为null，不符合目标数据库要求。 – 目标数据库多的列设置固定默认值，且有唯一约束。源数据库insert多条数据后，同步到目标数据库后多的列为固定默认值，不符合目标数据库要求。 • Migration自动建表时，源库中char、varchar、nvarchar、enum、set字符类型长度在目标库会按照字节长自动扩大（因为DWS目标库为字节长）。 • 全量同步timestamp类型时，默认值中的on update current_timestamp语法将不会同步到目标库GaussDB（DWS）中。 • 重命名表仅支持rename后库表在同步范围中的DDL操作（例如：RENAME TABLE A TO B，B需要在同步范围内）。 不建议在分库分表同步场景下的进行rename操作，可能会导致任务失败或数据不一致。

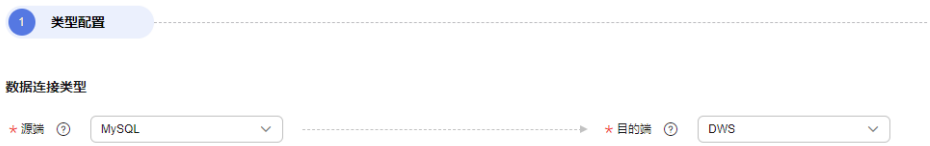
操作步骤

本小节以RDS for MySQL到DWS的实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

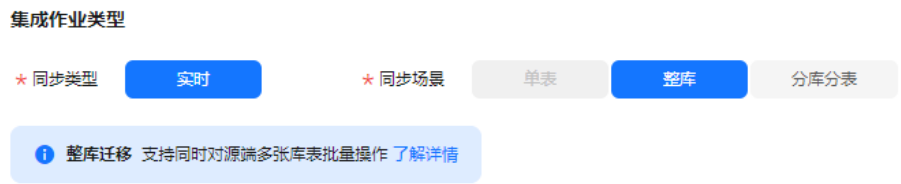
步骤2 选择数据连接类型：源端选MySQL，目的端选DWS。

图 7-169 选择数据连接类型



步骤3 选择集成作业类型：同步类型默认为实时，同步场景包含整库和分库分表场景。

图 7-170 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的MySQL、DWS数据连接和已配置好网络连接的资源组。

图 7-171 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

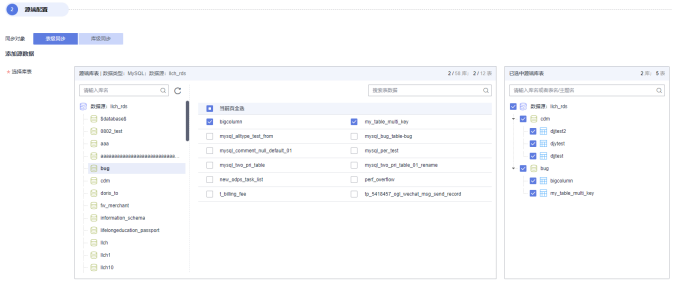
说明

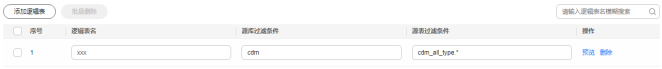
网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端参数。

各同步场景下选择需要同步库表的方式请参考下表。

表 7-65 选择需要同步的库表

同步场景	配置方式
整库	• 选择同步对象。 – 表级同步：支持选择Mysql实例下多个库中的多张表进行同步。 – 库级同步：支持选择Mysql实例下的多个库，对库中的所有表进行同步。 • 选择需要迁移的MySQL库表。 **图 7-172** 选择库表  库与表均支持自定义选择，即可选择一库一表，也可选择多库多表。

同步场景	配置方式
分库分表	添加逻辑表。 ● 逻辑表名：即最终写入到DWS的表名。 ● 源库过滤条件：支持填入正则表达式，在所有MySQL实例中通过该正则表达式过滤出要写入目标端DWS汇聚表的所有分库。 ● 源表过滤条件：支持填入正则表达式，在过滤出的源端分库中再次过滤出要写入目标端DWS汇聚表的所有分表。 **图 7-173** 添加逻辑表  已添加的逻辑表支持预览表结构及来源库表，单击“操作”列的预览即可。预览逻辑表时，源表数量越多，等待时间可能越长，请耐心等待。 **图 7-174** 逻辑表预览  逻辑表预览 表名： logical_table_test 表结构： 多表只展示其中一张表的字段 序号 | 字段名 | 类型 1 A1_INT INT 2 A2_varchar CHAR 3 A3_FLOAT DOUBLE 4 A4_DOUBLE DOUBLE 5 A5_DECIMAL DOUBLE 6 A6_BOOLEAN CHAR 7 A7_SMALLINT DOUBLE 8 A8_SHORT DOUBLE 9 A9_BIGINT DOUBLE 10 A10_LONG DOUBLE 11 A11_TIMESTAMP TIMESTAMP 12 A12_CHAR CHAR 13 A13_VARCHAR CHAR

步骤7 配置目的端参数。

- 源库表和目标匹配策略。
  各同步场景下源端库表和目标端库表的匹配策略请参考下表。

表 7-66 源库表和目标匹配策略

同步场景	配置方式
整库	– Schema匹配策略。 ■ 与来源库同名：数据将同步至与来源MySQL库名相同的DWS Schema中。 ■ 自定义：数据将同步至自行指定的DWS Schema中。 – 表匹配策略。 ■ 与来源表同名：数据将同步至与来源MySQL表名相同的DWS表中。 ■ 自定义：数据将同步至自行指定的DWS表中。 **图 7-175** 整库场景下源库表和目标匹配策略  **说明** 自定义匹配策略时，支持用内置变量#{source_db_name}和#{source_table_name}标志来源的库名和表名，其中表匹配策略必须包含#{source_table_name}。
分库分表	– 目标端库名：数据将同步至指定的DWS Schema中。 – 表匹配策略：默认与源端配置中填写的逻辑表同名。 **图 7-176** 分库分表场景下源库表和目标匹配策略 

- DWS参数配置。
  其余DWS目的端参数说明请参考下表。

图 7-177 DWS 配置项

表 7-67 DWS 配置项

配置项	默认值	单位	配置说明
写入模式	UPSERT MODE	-	– UPSERT MODE：批量更新入库模式。 – COPY MODE：DWS专有的高性能批量入库模式。
批写最大数据量	50000	条	单批次写入DWS数据的条数，可根据表数据大小和作业内存使用适当调整。
定时批写时间间隔	3	秒	支持配置每批次数据写入DWS的时间间隔。
高级配置	-	-	支持通过参数配置部分高级功能，参数详情可参考DWS高级配置一览表。

表 7-68 DWS 高级配置一览表

参数名	参数类型	默认值	单位	参数说明
sink.buffer-flush.max-size	int	512	MB	写入DWS时每批数据的最大字节数，可根据作业配置内存和数据大小适当调整。
sink.keyby.enable	boolean	true	-	数据分流开关，在多并发场景下开启数据分流可将数据按规则分配给不同的工作进程写入目的端，可提高写入性能。
sink.keyby.mode	string	table	-	数据分流模式，可选填写： – pk：按数据主键值进行分流。 – table：按表名进行分流。 **说明** ■ 多并发场景下，若开启DDL功能，只能按表名分流，否则可能导致数据不一致。 ■ 确保不会有DDL时，可以选择按主键分流，多并发场景下可提高写入性能。

参数名	参数类型	默认值	单位	参数说明
sink.field.name.case-sensitive	boolean	true	-	同步数据大小写敏感开关，开启后在同步数据时对库名、表名、字段名大小写均敏感。
sink.verify.column-number	boolean	false	-	校验数据列数的开关，链路默认以同名映射方式同步数据，不检验是否所有列均同步。 开启本开关后，若源端与目的端列数不同将认为是数据不一致的场景，导致作业异常。
sink.server.timezone	string	本地时区	-	连接目的端数据库时指定的session时区，支持时区标准写法，例如UTC+8等。
logical.delete.enabled	boolean	false	-	逻辑删除开关。
logical.delete.column	string	logical_is_deleted	-	逻辑删除标记列名称，默认为logical_is_deleted，支持用户自定义。

步骤8 刷新源表和目标表映射，检查映射关系是否正确，同时可根据需求修改表属性、添加附加字段，并通过“自动建表”能力在目的端DWS数据库中建出相应的表。

图 7-178 源表与目标表映射

- 附加字段编辑：单击操作列“附加字段编辑”可为目的端的DWS表中增加自定义字段，同时附加字段也会额外加入到DWS表的建表中。用户可以在已有的源表字段基础上添加多个附加字段，并自定义字段名、选择字段类型、填写字段值。
  - 字段名称：目的端DWS表新增字段的名称。

- 字段类型：目的端DWS表新增字段的类型。
- （可选）字段类型长度：目的端DWS表新增字段类型的长度。
- 字段值：目的端DWS表新增字段的取值来源。

表 7-69 附加字段取值方式

类型	示例
常量	任意字符。
内置变量	▪ 源端host ip地址：source.host。 ▪ 源端schema名称：mgr.source.schema。 ▪ 源端table名称：mgr.source.table。 ▪ 目的端schema名称：mgr.target.schema。 ▪ 目的端table名称：mgr.target.table。
源表字段	源表中的任一字段。 配置附加字段的取值来源于源表字段时，请注意任务运行过程中不能修改对应源表字段的名称，否则可能导致作业异常。
udf方法	▪ substring(#col, pos[, len])： 截取源端col列的子串, 范围在[pos, pos+len)。 ▪ date_format(#col, time_format[, src_tz, dst_tz])： 将源端col列按time_format格式化, 可选转换时区。 ▪ now([tz])： 获取指定时区的当前时间。 ▪ if(cond_exp, str1, str2)： 满足条件表达式cond_exp时返回str1, 否则返回str2。 ▪ concat(#col[, #str, ...])： 拼接多个参数, 可为源端列或字符串。 ▪ from_unixtime(#col[, time_format])： 将unix时间戳按time_format格式化。 ▪ unix_timestamp(#col[, precision, time_format])： 将时间转成unix时间戳, 可显式定义时间格式及转换后精度。

- 自动建表：单击“自动建表”可按照已配置映射规则在目的端数据库自动建表，成功后表建立方式会显示为使用已有表。

图 7-179 自动建表

说明

- Migration仅支持自动建表，不支持自动建库和模式，需用户自行在目的端手动建出库和模式后再使用本功能建表。
- 自动建表时对应的字段类型映射关系请参见**字段映射关系**章节。
- 自动建出的Hudi表会带有3个审计字段，分别是cdc_last_update_date、logical_is_deleted、_hoodie_event_time，并会以_hoodie_event_time作为Hudi表的预聚合键。

步骤9 配置DDL消息处理规则。

实时集成作业除了能够同步对数据的增删改等DML操作外，也支持对部分表结构变化（DDL）进行同步。针对支持的DDL操作，用户可根据实际需求配置为正常处理/忽略/出错。

- 正常处理：Migration识别到源端库表出现该DDL动作时，作业自动同步到目的端执行该DDL操作。
- 忽略：Migration识别到源端库表出现该DDL动作时，作业忽略该DDL，不同步到目的端表中。
- 出错：Migration识别到源端库表出现该DDL动作时，作业抛出异常。

图 7-180 DDL 配置

步骤10 配置任务属性。

表 7-70 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2

参数	说明	默认值
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-

参数	说明	默认值
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** • 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 • 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤11 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-181 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-182 启动配置

表 7-71 启动配置参数

参数	说明
同步模式	• 增量同步：从指定时间位点开始同步增量数据。 • 全量+增量：先同步全量数据，随后实时同步增量数据。
时间	增量同步需要设置该参数，指示增量同步起始的时间位点。 **说明** 配置的位点时间早于Binlog日志最早时间点时，默认会以日志最新时间点开始消费。

步骤12 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-183 前往监控

----结束

性能调优

若链路同步速度过慢，可参考参见**任务性能调优**章节中对应链路文档进行排查及处理。

7.10.5 MySQL 同步到 Kafka 作业配置

支持的源端和目的端数据库版本

表 7-72 支持的数据库版本

源端数据库	目的端数据库
MySQL数据库（5.6、5.7、8.x版本）	Kafka集群（2.7、3.x版本）

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考下表进行赋权。

表 7-73 数据库账号权限

类型名称	权限要求
源数据库连接账号	需要具备如下最小权限：SELECT、SHOW DATABASES、REPLICATION SLAVE、REPLICATION CLIENT，即执行SQL： `GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '用户名'@'%';`
目标数据库连接账号	MRS用户需要拥有Kafka对应Topic的读写权限，即必须属于kafka/kafkaadmin/kafkasuperuser用户组。 **说明** kafka普通用户需要被Kafka管理员用户授予特定Topic的读写权限，才能访问对应Topic。

说明

- 建议创建单独用于Migration任务连接的数据库账号，避免因为数据库账号密码修改，导致的任务连接失败。
- 连接源和目标数据库的账号密码修改后，请同步修改管理中心对应的连接信息，避免任务连接失败后自动重试，导致数据库账号被锁定影响使用。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考下表。

表 7-74 同步对象范围

类型名称	使用须知
同步对象范围	• 支持同步DML和DDL。 • 仅支持同步MyISAM和InnoDB表。 • 不支持同步视图、外键、存储过程、触发器、函数、事件、虚拟列、唯一约束和唯一索引。 • 不支持同步对象中存在包含CASCADE、SET NULL、SET DEFAULT之类引用操作的外键。这些关联操作会导致更新或删除父表中的行会影响子表对应的记录，并且子表的相关操作并不记录binlog。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见下表。

表 7-75 注意事项

类型名称	使用和操作限制
数据库限制	源端数据库中的库名、表名、字段名不能包含：.<'>/\"以及非ASCII字符，建议尽量使用常规字符避免任务失败。

类型名称	使用和操作限制
使用限制	通用： • 实时同步过程中，不支持IP、端口、账号、密码修改。 • 不允许源数据库进行恢复操作。 • 建议MySQL Binlog保留3天以上，不支持强制清理Binlog。异常/暂停恢复作业时，记录的Binlog位点过期会导致作业恢复失败，需要关注作业异常/暂停时长及Binlog保留时长。 • 实时同步过程中，不允许源数据库MySQL跨大版本升级，否则可能导致数据不一致或者同步任务失败（跨版本升级后数据、表结构、关键字等信息均可能会产生兼容性改变），建议在该场景下重建同步任务。 全量同步阶段： 任务启动和全量数据同步阶段，请不要在源数据库执行DDL操作，否则可能导致任务异常。 增量同步阶段： 增量同步过程中，分库分表场景下，在多个分表执行的DDL，会同步多条数据到Kafka的Topic中。 常见故障排查： 在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。
其他限制	重命名表仅支持rename后库表在同步范围中的DDL操作（例如：RENAME TABLE A TO B，B需要在同步范围内）。

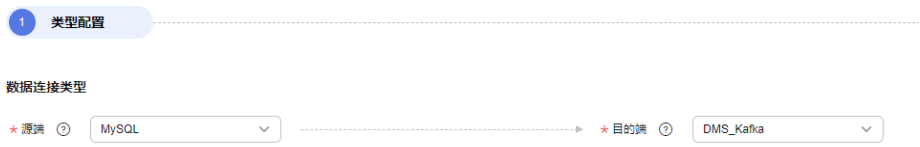
操作步骤

本小节以RDS for MySQL到DMS Kafka实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

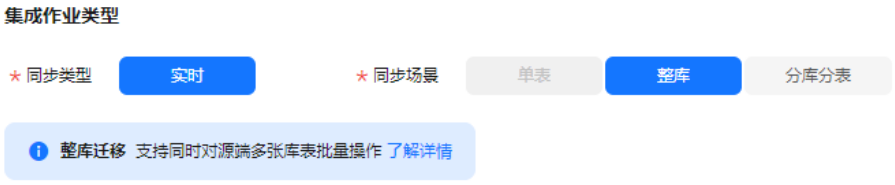
步骤2 选择数据连接类型：源端选MySQL，目的端选DMS Kafka。

图 7-184 选择数据连接类型



步骤3 选择集成作业类型：同步类型默认为实时，同步场景包含整库和分库分表场景。

图 7-185 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的MySQL、DMS Kafka数据连接和已配置好网络连接的资源组。

图 7-186 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

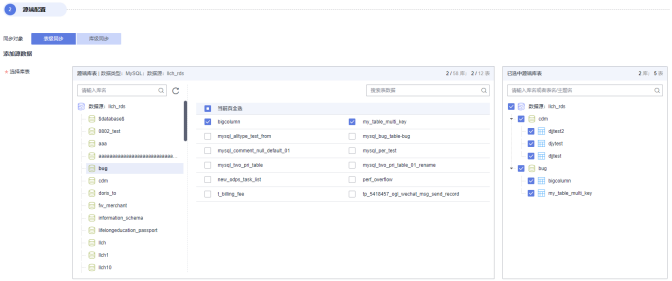
说明

网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端参数。

各同步场景下选择需要同步库表的方式请参考下表。

表 7-76 选择需要同步的库表

同步场景	配置方式
整库	选择需要迁移的MySQL库表。 **图 7-187** 选择库表  库与表均支持自定义选择，即可选择一库一表，也可选择多库多表。

同步场景	配置方式
分库分表	添加逻辑表。 • 逻辑表名：即最终写入到DMS Kafka的Topic名。 • 源库过滤条件：支持填入正则表达式，在所有MySQL实例中通过该正则表达式过滤出要抽取数据写入目标端Kafka Topic的所有分库。 • 源表过滤条件：支持填入正则表达式，在过滤出的源端分库中再次过滤出要抽取数据写入目标端Kafka Topic的所有分表。 **图 7-188** 添加逻辑表  已添加的逻辑表支持预览表结构及来源库表，单击“操作”列的预览即可。预览逻辑表时，源表数量越多，等待时间可能越长，请耐心等待。 **图 7-189** 逻辑表预览  逻辑表预览 表名： logical_table_test　　表结构： 多表只展示其中一张表的字段 序号 | 字段名 | 类型 1 A1_INT INT 2 A2_varchar CHAR 3 A3_FLOAT DOUBLE 4 A4_DOUBLE DOUBLE 5 A5_DECIMAL DOUBLE 6 A6_BOOLEAN CHAR 7 A7_SMALLINT DOUBLE 8 A8_SHORT DOUBLE 9 A9_BIGINT DOUBLE 10 A10_LONG DOUBLE 11 A11_TIMESTAMP TIMESTAMP 12 A12_CHAR CHAR 13 A13_VARCHAR CHAR

步骤7 配置目的端参数。

图 7-190 Kafka 目的端配置项

- 目标Topic名称规则。

配置源端MySQL库表与目的端Kafka Topic的映射规则

表 7-77 目标 Topic 名称规则

同步场景	配置方式
整库	配置源端MySQL库表与目的端Kafka Topic的映射规则，可指定为固定的一个Topic，也可使用内置变量做映射，将不同源表数据同步到不同的Topic中。 可以使用的内置变量有： – 源库名：#{source_db_name}。 – 源表名：#{source_table_name}。
分库分表	无该配置项，默认使用源端配置的逻辑表名作为目的端同步的Topic名。

- 同步kafka partition策略

  支持以下三种投递策略将源端的数据按规则同步到Kafka Topic的特定Partition：
  - 全部投递到Partition 0。
  - 按库名+表名的hash值投递到不同Partition。
  - 按表的主键值hash值投递到不同的Partition。

  说明

  源端无主键情况下，目的端默认投递到partition 0。
- 需要同步的数据库操作

  支持同步的数据库操作包括DDL和DML，可单选或多选，不选择的情况下默认同步所有操作。
- 投递到Kafka的数据格式

  选择投递到Kafka的数据组织格式，当前支持Debezium JSON和Canal JSON。
- 新建Topic的Partition数量

  设定目的端Kafka无对应Topic时，Migration自动建Topic的分区数量，默认为3。
- Kafka目标端属性配置

  支持设置Kafka的配置项，需要增加 properties. 前缀，作业将自动移除前缀并传入底层Kafka客户端，具体参数可参考**Apache Kafka官方文档**中的配置说明。
- 高级配置

  支持在作业“任务配置”中添加自定义属性来开启部分高级功能，参数详情可参考MySQL->Kafka高级参数一览表。

图 7-191 添加自定义属性

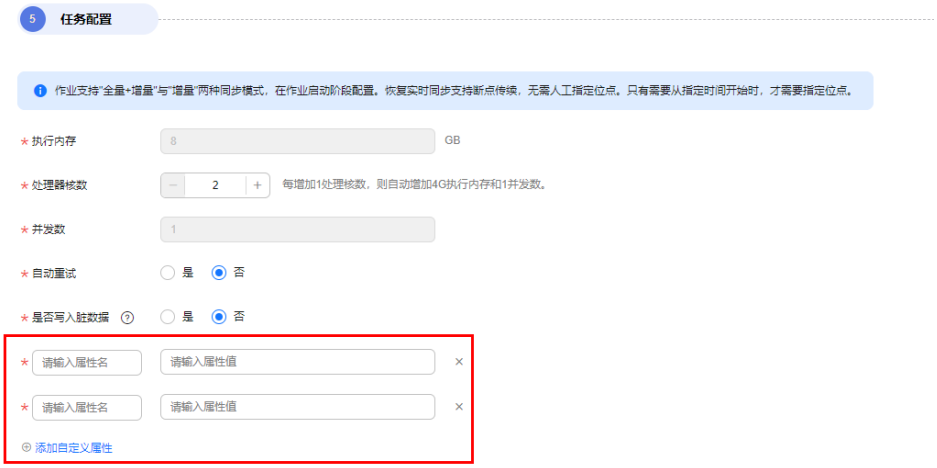


表 7-78 MySQL > Kafka 高级参数一览表

参数名	参数类型	默认值	单位	参数说明
source.server.timezone	string	本地时区	-	连接源端数据库时指定的session时区，支持时区标准写法，例如UTC+8等。
source.convert.timestampWithServerTimeZone	boolean	true	-	timestamp类型数据输出时转为按源端时区。
source.convert.bit1AsInt	boolean	true	-	是否将bit1输出成int类型。

参数名	参数类型	默认值	单位	参数说明
sink.delivery-guarantee	string	at-least-once	-	Flink写Kafka时的语义保证机制。 – at-least-once：在 checkpoint 时会等待 Kafka 缓冲区中的数据全部被 Kafka producer 确认。消息不会因 Kafka broker 端发生的事件而丢失，但可能会在 Flink 重启时重复，因为 Flink 会重新处理旧数据。 – exactly-once：该模式下，Kafka sink 会将所有数据通过在 checkpoint 时提交的事务写入。因此，如果 consumer 只读取已提交的数据，在 Flink 发生重启时不会发生数据重复。然而这会使数据在 checkpoint 完成时才会可见，因此请按需调整 checkpoint 的间隔。

步骤8 刷新源表和目标表映射，检查映射关系是否正确。

表 7-79 源表与目标表映射

同步场景	配置方式
整库	支持用户根据实际需求修改映射后的目的端Topic名称，可以配置为一对一、多对一的映射关系。 **图 7-192** 整库场景下源表与目标表映射 
分库分表	默认使用源端配置的逻辑表名作为目的端的Topic名称。 **图 7-193** 分库分表场景下源表与目标表映射 

步骤9 配置任务属性。

表 7-80 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB

参数	说明	默认值
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-

参数	说明	默认值
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** • 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 • 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤10 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-194 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-195 启动配置

表 7-81 启动配置参数

参数	说明
同步模式	• 增量同步：从指定时间位点开始同步增量数据。 • 全量+增量：先同步全量数据，随后实时同步增量数据。
时间	增量同步需要设置该参数，指示增量同步起始的时间位点。 **说明** 配置的位点时间早于Binlog日志最早时间点时，默认会以日志最新时间点开始消费。

步骤11 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-196 前往监控

----结束

性能调优

若链路同步速度过慢，可参考参见**任务性能调优**章节中对应链路文档进行排查及处理。

7.10.6 DMS Kafka 同步到 OBS 作业配置

支持的源端和目的端数据库版本

表 7-82 支持的数据库版本

源端数据库	目的端数据库
Kafka集群（2.7、3.x版本）	-

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考下表进行赋权。

表 7-83 数据库账号权限

类型名称	权限要求
源数据库连接账号	DMS Kafka开启密文接入场景下，所配置用户需要有发布和订阅Topic的权限，其余场景无特殊权限要求。
目标数据库连接账号	需要有目标OBS桶访问权限，且拥有在桶下读写对象的权限，详情可参考**OBS权限控制**。

说明

- 建议创建单独用于Migration任务连接的数据库账号，避免因为数据库账号密码修改，导致的任务连接失败。
- 连接源和目标数据库的账号密码修改后，请同步修改管理中心对应的连接信息，避免任务连接失败后自动重试，导致数据库账号被锁定影响使用。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考下表。

表 7-84 同步对象范围

类型名称	使用须知
同步对象范围	支持同步所有Kafka消息，其中支持对JSON或CSV格式的消息体进行解析。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见下表。

表 7-85 注意事项

类型名称	使用和操作限制
数据库限制	• 支持开启SASL_PLAINTEXT的Kafka实例，包括SCRAM-SHA-512及PLAIN认证机制。 • 不支持开启SASL_SSL的Kafka实例。
使用限制	通用： 实时同步过程中，不支持IP、端口、账号、密码修改。 增量同步阶段： 整库场景下需要根据同步的Topic分区数对应增加作业并发数，否则可能导致任务内存溢出。 常见故障排查： 在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。

类型名称	使用和操作限制
其他限制	• 支持目标数据库中的表比源数据库多列场景，但是需要避免以下场景可能导致的任务失败。 – 目标数据库多的列要求非空且没有默认值，源数据库insert数据，同步到目标数据库后多的列为null，不符合目标数据库要求。 – 目标数据库多的列设置固定默认值，且有唯一约束。源数据库insert多条数据后，同步到目标数据库后多的列为固定默认值，不符合目标数据库要求。 • Migration自动建表时，源库中char、varchar、nvarchar、enum、set字符类型长度在目标库会按照字节长自动扩大（因为DWS目标库为字节长）。 • 全量同步timestamp类型时，默认值中的on update current_timestamp语法将不会同步到目标库GaussDB(DWS)中。 • 重命名表仅支持rename后库表在同步范围中的DDL操作（例如：RENAME TABLE A TO B，B需要在同步范围内）。

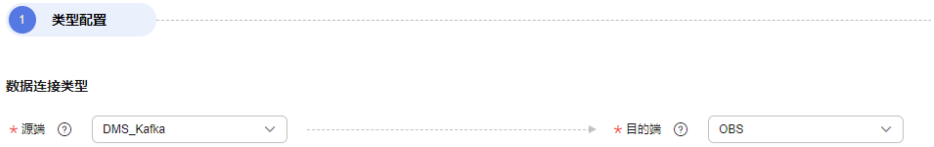
操作步骤

本小节以DMS Kafka到OBS的实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

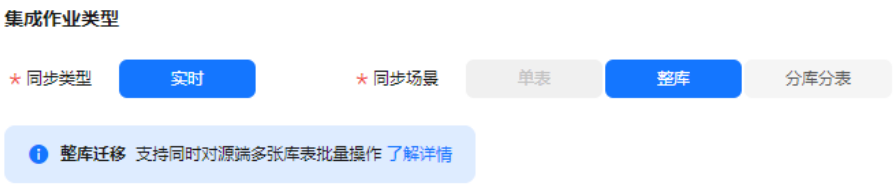
步骤2 选择数据连接类型：源端选DMS Kafka，目的端选OBS。

图 7-197 选择数据连接类型



步骤3 选择集成作业类型：同步类型默认为实时，同步场景包含单表、整库场景。

图 7-198 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的DMS Kafka、OBS数据连接和已配置好网络连接的资源组。

图 7-199 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

说明

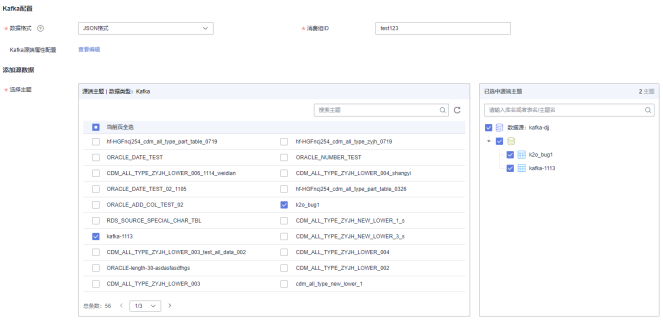
网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端参数。

- 选择需要同步的Kafka Topic，各同步场景下选择需要同步主题的方式请参考下表。

表 7-86 选择需要同步的主题

同步场景	配置方式
单表	输入一个需要迁移的Kafka Topic。 **图 7-200** 输入 Kafka Topic 

同步场景	配置方式
整库	选择需要迁移的Kafka Topic。 **图 7-201** 选择 Kafka Topic 

- 数据格式

  源端Kafka Topic中消息内容的格式，Migration当前支持对如下三种消息进行处理：

  - JSON格式：支持对消息内容以JSON的层级格式进行解析。
  - CSV格式：支持对消息内容以CSV格式指定分隔符进行解析。
  - TEXT格式：将整条消息内容作为文本直接同步。

- 消费组ID

  消费者是从Topic订阅消息的一方，消费组是由一个或多个消费者组成的。Migration支持指定本次消费动作所属的Kafka消费组。

- Kafka源端属性配置

  支持设置Kafka的配置项，需要增加 properties. 前缀，作业将自动移除前缀并传入底层Kafka客户端，具体参数可参考**Apache Kafka官方文档**中的配置说明。

步骤7 配置目的端参数。

图 7-202 目的端 OBS 配置

- 文件存储格式

  写入OBS的文件格式，当前支持Parquet、SequenceFile和TextFile。

- 文件压缩方式

  指定写入OBS文件的压缩方式，默认不进行压缩，支持以下列表：

  - Parquet格式：UNCOMPRESSED、SNAPPY。
  - SequenceFile格式：UNCOMPRESSED、SNAPPY、GZIP、LZ4、BZIP2。

- TextFile格式：UNCOMPRESSED。

- OBS存储路径

  指定OBS文件存储的路径，支持填写#{source_Topic_name}内置变量，可将源端不同的Topic的数据写入不同的路径下，例如：obs://bucket/dir/test.db/prefix_#{source_Topic_name}_suffix/

- 全局高级配置

  支持通过参数配置部分高级功能，详情请参考下表。

  表 7-87 OBS 高级配置一览表

参数名	参数类型	默认值	单位	参数说明
auto-compaction	boolean	false	-	文件自动合并开关。数据会先被写入临时文件，当checkpoint完成后，该配置控制检查点内产生的临时文件是否被合并。开启该配置部分场景下可减少小文件数量，但会较大降低同步速率。

步骤8 刷新源表和目标表映射，单击“目标字段编辑”检查要写入目的端的字段情况，并根据实际情况选择配置分区字段。

图 7-203 源表与目标表映射

- 分区字段

  支持配置分区字段，将在写入OBS时自动生成对应分区目录，目录名为“分区字段=分区值”。同时，字段选择顺序影响分区的层级，例如选择par1、par2作为分区字段，那么par1为一级分区，par2为二级分区，最多支持五级分区。

- 目标字段编辑

  Migration会根据选择的源端消息格式自动解析源端消息，生成对应的字段信息，用户可在此基础上进行编辑，自定义字段名、选择字段类型、填写字段值。

  - 字段名称：目的端OBS文件中写入字段的名称。字段名称至少包含一个字母，允许下划线、中划线，不支持纯数字。
  - 字段类型：目的端OBS文件中写入字段的类型。当前支持STRING、BOOLEAN、INTEGER、LONG、FLOAT、DOUBLE、SHORT、DECIMAL、DATE、TIMESTAMP类型。
  - 字段值：目的端OBS文件中写入字段的取值来源。

表 7-88 目标字段取值方式

类型	字段取值
手动赋值	任意字符。
内置变量	Kafka的元数据，包括__key__、__value__、__Topic__、__partition__、__offset__、__timestamp__共6个字段。

源表字段	从源端Kafka Topic消息解析出的任意字段。
	说明
	如果源端Kafka消息为嵌套JSON的形式，本链路支持解析不同层级的字段值（包含数组，数组下标索引从1开始）。
	例如，JSON的内容为：
	{
	"col1": "1",
	"col2": "2",
	"level1": {
	"level2": [
	{
	"level3": "test"
	}
	]
	}
	}
	则可以通过level1.level2[1].level3取到数据”test”作为目标端某一个字段的值。

udf方法	支持填写Flink的内置函数用于数据转换，例如： ■ CONCAT(CAST(NOW() as STRING), `col_name`) ■ DATE_FORMAT(NOW(), 'yy') 注意，其中的字段名要用反引号包围起来。Flink完整内置函数可参考**Flink官方文档**。

步骤9 配置任务属性。

表 7-89 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒

参数	说明	默认值
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** ● 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 ● 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤10 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-204 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-205 启动配置

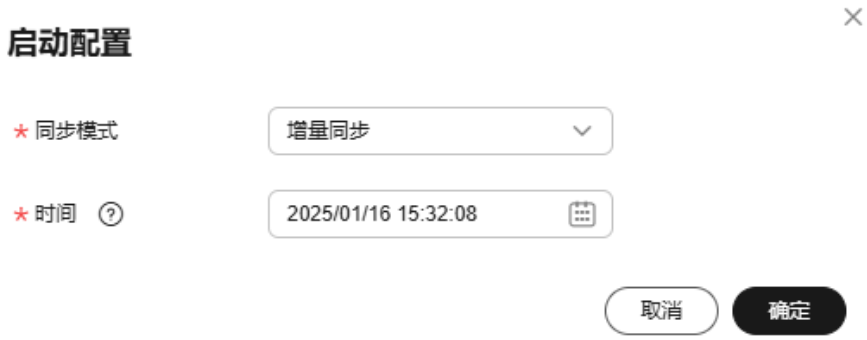


表 7-90 启动配置参数

参数	说明
偏移量参数	• 最早：从Kafka Topic最早偏移量开始消费数据。 • 最新：从Kafka Topic最新偏移量开始消费数据。 • 起止时间：根据时间获取Kafka Topic对应的偏移量，并从该偏移量开始消费数据。
时间	起止时间需要设置该参数，指示同步起始的时间位点。 **说明** 配置的位点时间早于Kafka消息最早偏移量时，默认会从最早偏移量开始消费。

步骤11 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-206 前往监控

----结束

性能调优

若链路同步速度过慢，可参考参见**任务性能调优**章节中对应链路文档进行排查及处理。

7.10.7 Apache Kafka 同步到 MRS Kafka 作业配置

支持的源端和目的端数据库版本

表 7-91 支持的数据库版本

源端数据库	目的端数据库
Kafka集群（2.7、3.x版本）	Kafka集群（2.7、3.x版本）

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考下表进行赋权。

表 7-92 数据库账号权限

类型名称	权限要求
源数据库连接账号	-
目标数据库连接账号	MRS用户需要拥有Kafka对应Topic的读写权限，即必须属于kafka/kafkaadmin/kafkasuperuser用户组。 **说明** kafka普通用户需要被Kafka管理员用户授予特定Topic的读写权限，才能访问对应Topic。

说明

- 建议创建单独用于Migration任务连接的数据库账号，避免因为数据库账号密码修改，导致的任务连接失败。
- 连接源和目标数据库的账号密码修改后，请同步修改管理中心对应的连接信息，避免任务连接失败后自动重试，导致数据库账号被锁定影响使用。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考下表。

表 7-93 同步对象范围

类型名称	使用须知
同步对象范围	支持完整同步Kafka Topic所有消息内容，但不支持对Kafka Topic消息进行解析重组后同步。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见下表。

表 7-94 注意事项

类型名称	使用和操作限制
数据库限制	• 支持开启/未开启Keberos认证的MRS集群Kafka实例。 • 不支持开启SASL_SSL的Kafka实例。
使用限制	通用： 实时同步过程中，不支持IP、端口、账号、密码修改。 增量同步阶段： 整库场景下需要根据同步的Topic分区数对应增加作业并发数，否则可能导致任务内存溢出。 常见故障排查： 在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。
其他限制	-

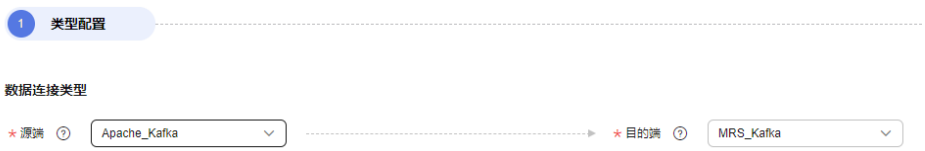
操作步骤

本小节以Apache Kafka到MRS Kafka实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

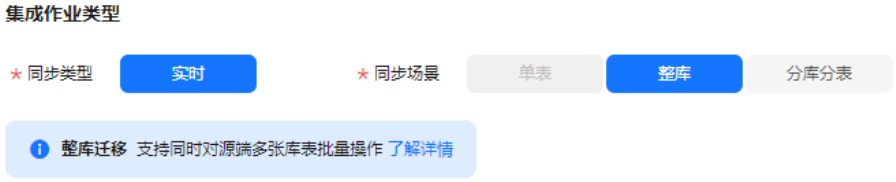
步骤2 选择数据连接类型：源端选Apache_Kafka，目的端选MRS_Kafka。

图 7-207 选择数据连接类型



步骤3 选择集成作业类型：同步类型默认为实时，同步场景仅支持整库场景。

图 7-208 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的DMS Kafka、OBS数据连接和已配置好网络连接的资源组。

图 7-209 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

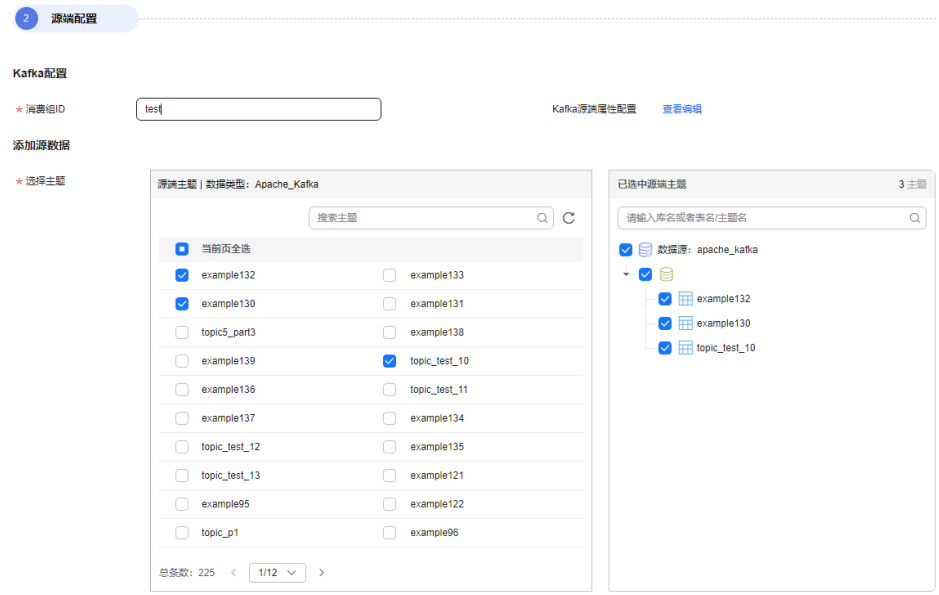
说明

网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端参数。

- 选择需要同步的Kafka Topic。

图 7-210 选择需要同步的 Kafka Topic



- 消费组ID

  消费者是从Topic订阅消息的一方，消费组是由一个或多个消费者组成的。Migration支持指定本次消费动作所属的Kafka消费组。

- Kafka源端属性配置

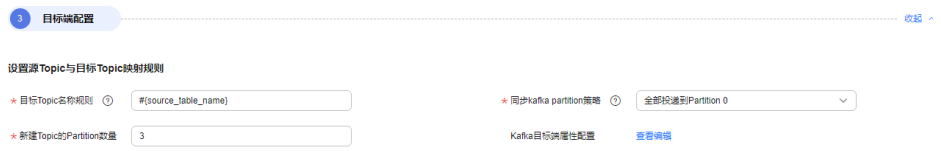
  支持设置Kafka的配置项，需要增加 properties. 前缀，作业将自动移除前缀并传入底层Kafka客户端，具体参数可参考

  Apache Kafka官方文档

  中的配置说明。

步骤7 配置目的端参数。

图 7-211 Kafka 目的端配置项



- 目标Topic名称规则。

  配置源端MySQL库表与目的端Kafka Topic的映射规则。可指定为固定的一个Topic，也可使用内置变量做映射，将不同源表数据同步到不同的Topic中。

  可以使用的内置变量有：源Topic名：#{source_Topic_name}

- 同步kafka partition策略

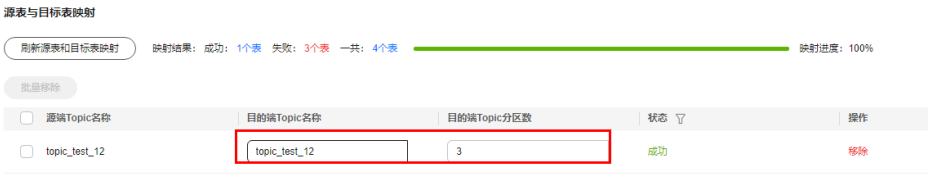
  支持以下三种投递策略将源端的数据按规则同步到Kafka Topic的特定Partition：

  - 全部投递到Partition 0。
  - 按源端分区投递到对应的Partition：源端消息在第n个分区，则投递到目的端的第n个分区，该策略可以保证消息顺序。
  - 按轮询模式投递到不同的Partition：采用Kafka粘性分区策略均匀的投递到目的端主题的所有分区，该策略无法保证消息顺序。

- 新建Topic的Partition数量
  设定目的端Kafka无对应Topic时，Migration自动建Topic的分区数量，默认为3。
- Kafka目标端属性配置
  支持设置Kafka的配置项，需要增加 properties. 前缀，作业将自动移除前缀并传入底层Kafka客户端，具体参数可参考Apache Kafka官方文档中的配置说明。

步骤8 刷新源表和目标表映射，检查源端Topic和目的端Topic映射关系是否正确，支持用户根据实际需求修改映射后的目的端Topic名称，可以配置为一对一、多对一的映射关系。

图 7-212 源表与目标表映射



步骤9 配置任务属性。

表 7-95 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒

参数	说明	默认值
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** ● 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 ● 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤10 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-213 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-214 启动配置

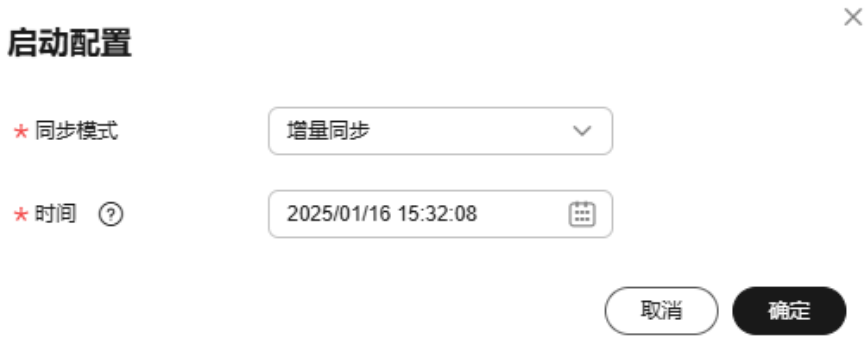


表 7-96 启动配置参数

参数	说明
同步模式	• 最早：从Kafka Topic最早偏移量开始消费数据。 • 最新：从Kafka Topic最新偏移量开始消费数据。 • 起止时间：根据时间获取Kafka Topic对应的偏移量，并从该偏移量开始消费数据。
时间	起止时间需要设置该参数，指示同步起始的时间位点。 **说明** 配置的位点时间早于Kafka消息最早偏移量时，默认会从最早偏移量开始消费。

步骤11 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-215 前往监控

----结束

性能调优

若链路同步速度过慢，可参考参见**任务性能调优**章节中对应链路文档进行排查及处理。

7.10.8 SQLServer 同步到 MRS Hudi 作业配置

支持的源端和目的端数据库版本

表 7-97 支持的数据库版本

源端数据库	目的端数据库
SQLServer数据库（企业版2016、2017、2019、2022版本，标准版2016 SP2及以上版本、2017、2019、2022版本）	• MRS集群（3.2.0-LTS.x、3.5.x） • Hudi版本（0.11.0）

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考下表进行赋权。

表 7-98 数据库账号权限

类型名称	权限要求

源数据库连接账号	需要具备sysadmin权限，或者view server state权限以及待同步数据库的db_datareader或db_owner权限。
	• 启动数据库及表的CDC能力。
	1. 启用数据库CDC。
	`USE YourDatabaseName;`
	`EXEC sys.sp_cdc_enable_db;`
	`-- 查看数据库是否启动CDC`
	`SELECT is_cdc_enabled, name FROM sys.databases WHERE name = 'YourDatabaseName'`
	2. 启用表CDC。
	`EXEC sys.sp_cdc_enable_table`
	`@source_schema = N'dbo', -- Schema`
	`@source_name = N'YourTable',-- 表名`
	`@role_name = NULL,-- 可选，CDC访问角色名称`
	`@supports_net_changes = 0;`
	`-- 查看表是否启动CDC`
	`SELECT name,is_tracked_by_cdc FROM sys.tables WHERE name = 'YourTable';`
	• 源端SQLServer需要给管理中心数据连接中配置的用户赋予以下全部权限。
	– 给用户添加数据库CONNECT, VIEW DATABASE STATE 权限。
	`USE YourDatabaseName;`
	`GRANT CONNECT, VIEW DATABASE STATE TO [YourUserName];`
	– 给用户添加CDC schema的SELECT 权限。
	`USE YourDatabaseName;`
	`GRANT SELECT ON SCHEMA::[cdc] TO [YourUserName];`
	– 给用户添加表的SELECT权限。
	`USE YourDatabaseName;`
	`GRANT SELECT ON OBJECT::[YourSchema].[YourTable] TO [YourUserName];`

目标数据库连接账号	MRS用户需要拥有Hadoop和Hive组件的读写权限，建议参照图1所示角色及用户组配置MRS用户。 **图 7-216** MRS Hudi 最小化权限  具体MRS集群角色权限管理请参考《**MRS集群用户权限模型**》。

说明

- 建议创建单独用于Migration任务连接的数据库账号，避免因为数据库账号密码修改，导致的任务连接失败。
- 连接源和目标数据库的账号密码修改后，请同步修改管理中心对应的连接信息，避免任务连接失败后自动重试，导致数据库账号被锁定影响使用。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考下表。

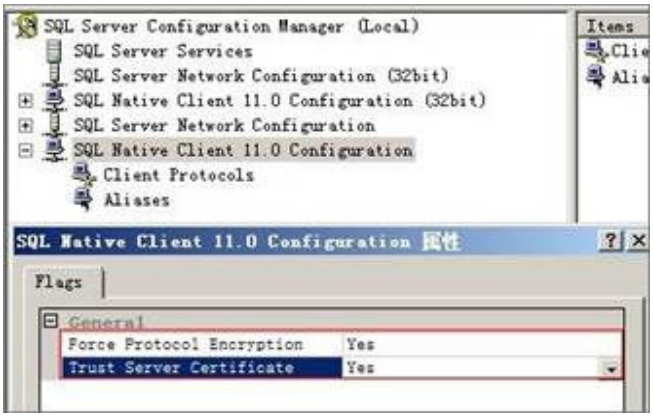
表 7-99 同步对象范围

类型名称	使用须知
同步对象范围	• 支持同步DML：包括INSERT、UPDATE、DELETE。 • 不支持同步DDL。 • 仅支持同步主键表。 • 不支持同步源数据库中开启TDE（Transparent Data Encryption）加密的数据库。 • 不支持列加密。 • 不支持同步自增属性列。 • 自动建表支持同步表结构、普通索引、约束（主键、空、非空）、注释。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见下表。

表 7-100 注意事项

类型名称	使用和操作限制
数据库限制	• 目标数据库中的库名、表名、字段名仅支持数字、字母和下划线，且字段名必须以字母或下划线开头，建议尽量使用常规字符避免任务失败。 • 源数据库如果开启客户端配置中的“强制协议加密（Force Protocol Encrypton）”，必须同时开启“信任服务器证书（trust server certificate）”，如下图所示： **图 7-217** 查看客户端属性 
使用限制	通用： • 实时同步过程中，不支持IP、端口、账号、密码修改。 • Hudi表使用Bucket索引的场景下不允许更新分区键，否则可能产生重复数据。 • Hudi表使用Bucket索引的场景下主键仅保证单分区内唯一。 • 本链路所使用的Hudi表需带有3个审计字段：cdc_last_update_date、logical_is_deleted、_hoodie_event_time，并会以_hoodie_event_time作为Hudi表的预聚合键。因此，若使用已存在的表，也需要携带这3个审计字段，否则可能导致任务异常。 – cdc_last_update_date：Migration任务处理CDC数据的时间。 – logical_is_deleted：逻辑删除标志。 – _hoodie_event_time：数据在SQLServer CDC中的时间戳。 全量同步阶段： 任务启动和全量数据同步阶段，请不要在源数据库执行DDL操作，否则可能导致任务异常。 增量同步阶段： • 支持DML：包括INSERT、UPDATE、DELETE。 • 不支持DDL操作，源数据库进行的DDL操作不会同步到目标数据库。 • 不支持大数据类型IMAGE、TEXT、NTEXT的删除操作。 常见故障排查： 在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。

类型名称	使用和操作限制
其他限制	• 支持目标数据库中的表比源数据库多列场景，但是需要避免以下场景可能导致的任务失败。 目标数据库多的列要求非空且没有默认值，源数据库insert数据，同步到目标数据库后多的列为null，不符合目标数据库要求。 • 不支持源数据库主备切换，源数据库主备切换会导致同步任务失败。 • 不支持源数据库Microsoft SQL Server为TLS 1.0、TLS 1.1协议的同步，如果需要同步，建议源库升级到TLS 1.2及以上版本。

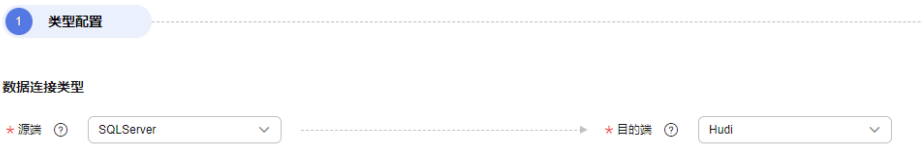
操作步骤

本小节以Microsoft SQL Server到MRS Hudi的实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

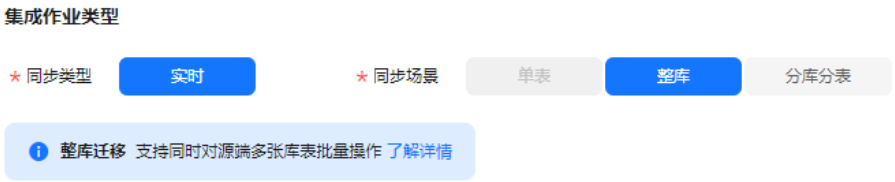
步骤2 选择数据连接类型：源端选SQLServer，目的端选Hudi。

图 7-218 选择数据连接类型



步骤3 选择集成作业类型：同步类型默认为实时，同步场景包含整库场景。

图 7-219 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的SQLServer、MRS Hudi数据连接和已配置好网络连接的资源组。

图 7-220 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

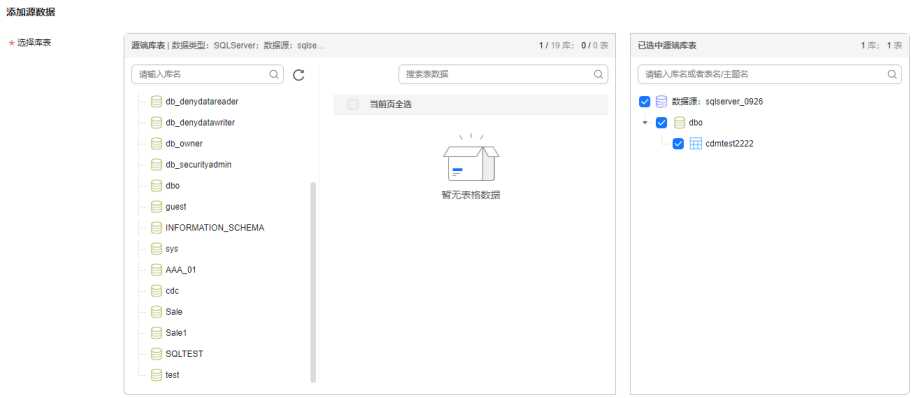
说明

网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端参数。

- 选择需要迁移的SQLServer库表。

图 7-221 选择库表



库与表均支持自定义选择，即可选择一库一表，也可选择多库多表。

步骤7 配置目的端参数。

- 源库表和目标匹配策略。

  各同步场景下源端库表和目标端库表的匹配策略请参考下表。

表 7-101 源库表和目标匹配策略

同步场景	配置方式
整库	– 库匹配策略。 ▪ 与来源库同名：数据将同步至与来源SQLServer Schema名相同的Hudi库中。 ▪ 自定义：数据将同步至自行指定的Hudi库中。 – 表匹配策略。 ▪ 与来源表同名：数据将同步至与来源SQLServer Schema名相同的Hudi表中。 ▪ 自定义：数据将同步至自行指定的Hudi表中。 **图 7-222** 整库场景下源库表和目标匹配策略  **说明** 自定义匹配策略时，支持用内置变量#{source_db_name}和#{source_table_name}标志来的源库名和表名，其中表匹配策略必须包含#{source_table_name}。

- Hudi参数配置。
  其余Hudi目的端参数说明请参考下表。

图 7-223 Hudi 目的端配置项

表 7-102 Hudi 目的端配置项

配置项	默认值	单位	配置说明
数据存储路径	-	-	Hudi自动建表时的warehouse路径，每张表会在warehouse路径下创建子目录。支持填写HDFS和OBS路径，路径格式参考： – OBS路径：obs://bucket/warehouse。 – HDFS路径：/tmp/warehouse。

配置项	默认值	单位	配置说明
Hudi表属性全局配置	-	-	支持通过参数配置部分高级功能，参数详情可参考Hudi高级配置一览表。
Compaction作业	-	-	需要一个独立的SparkSql作业，不使用则由Flink执行compaction。

表 7-103 Hudi 高级配置一览表

参数名	参数类型	默认值	单位	参数说明
index.type	string	BLOOM	-	Hudi表索引类型。 支持BLOOM和BUCKET索引，数据量较大场景下强烈建议使用BUCKET索引性能更好。
hoodie.bucket.index.num.buckets	int	256	个	Hudi表单分区下Bucket桶数。 **说明** 使用Hudi BUCKET表时需要设置Bucket桶数，桶数设置关系到表的性能，需要格外引起注意。 – 非分区表桶数 = MAX（单表数据量大小（G）/2G*2，再向上取整，4）。 – 分区表桶数 = MAX（单分区数据量大小（G）/2G*2，再后向上取整，1）。 其中，要注意的是： – 需要使用的是表的总数据大小，而不是压缩以后的文件大小。 – 桶的设置以偶数最佳，非分区表最小桶数请设置4个，分区表最小桶数请设置1个。
changelog.enabled	boolean	false	-	Hudi changelog功能开关，开启后Migration作业可输出DELETE和UPDATE BEFORE数据。
logical.delete.enabled	boolean	true	-	逻辑删除开关，changelog开启时必须关闭逻辑删除。

参数名	参数类型	默认值	单位	参数说明
hoodie.write.liststatus.optimized	boolean	true	-	写log文件时是否开启liststatus优化。涉及到大表和分区数据量多的作业，在启动时list会非常耗时，可能导致作业启动超时，建议关闭。
hoodie.index.liststatus.optimized	boolean	false	-	定位数据时是否开启liststatus优化。涉及到大表和分区数据量多的作业，在启动时list会非常耗时，可能导致作业启动超时，建议关闭。
compaction.async.enabled	boolean	true	-	异步compaction开关。compaction操作一定程度会影响实时任务的写入性能，如果用户使用外置的compaction操作对hudi进行compaction，可以考虑设置为false关闭实时处理集成作业的compaction操作。
compaction.schedule.enabled	boolean	true	-	生成compaction计划的开关。compaction计划必须由本服务生成，计划的执行可以交给Spark。
compaction.delta_commits	int	5	次	生成compaction request的频率。compaction request生成频率降低可以使得compaction频率降低从而提升作业性能。如果hudi增量数据较小。可以考虑增大该值。 **说明** 例如配置为40，即每40次commit生成一个compaction request，因为Migration每分钟生成1个commit，那么每个compaction request将间隔40分钟。
clean.async.enabled	boolean	true	-	做历史版本数据文件清理的开关。

参数名	参数类型	默认值	单位	参数说明
clean.retain_commits	int	30	次	要保留的commit数。这些commit关联的数据文件版本将被保留 num_of_commits * time_between_commits 这么长的时间，建议配置为2倍的compaction.delta_commits。 **说明** 例如配置为80，因为Migration每分钟生成1个commit，那么超过80分钟后如果有旧版本数据文件，则会生成clean request，且在执行clean时保留最近80个commit。
hoodie.archive.automatic	boolean	true	-	Hudi commit文件老化开关。
archive.min_commits	int	40	次	将旧版commit归档到日志文件中时要保留不归档的最小commit数。建议配置成clean.retain_commits + 1。 **说明** 例如配置成81，那么在触发归档动作时，将会保留最近81次commit文件。
archive.max_commits	int	50	次	触发归档动作的commit数。建议配置成archive.min_commits + 20。 **说明** 例如配置成101，那么将在生成101个commit文件后触发归档commit文件动作。

说明

- 为了达到Migration作业性能最优，建议使用Hudi Bucket索引的MOR表，并根据实际数据量配置Bucket桶数。
- 为了保证Migration作业的稳定性，建议将Hudi Compaction单独拆成Spark作业交由MRS执行，在Migration任务里仅开启生成compaction计划，具体可以参考**如何配置Hudi Compaction的Spark周期任务？**。

步骤8 刷新源表和目标表映射，检查映射关系是否正确，同时可根据需求修改表属性、添加附加字段，并通过“自动建表”能力在目的端Hudi数据库中建出相应的表。

图 7-224 源表与目标表映射

- 同步主键

  Hudi表必须设置“同步主键”，在源端为非主键表时，必须在字段映射阶段手动勾选主键。

- 表属性编辑

  单击操作列“表属性编辑”可配置Hudi表属性，包含表类型，分区类型及表自定义属性。

  图 7-225 Hudi 表属性配置

  - 表类型：Hudi的表类型，可选MERGE_ON_READ和COPY_ON_WRITE。
  - 分区类型：Hudi表分区类型，可选无分区、时间分区、自定义分区。

    说明

    其中时间分区需要用户指定一个源端表名，选择一个时间转换格式。

    比如时间分区用户指定一个源端表名src_col_1，选择一个时间转换格式，日（yyyyMMdd）、月（yyyyMM）、年（yyyy），自动建表时会在Hudi表默认创建一个cdc_partition_key的字段，系统会根据配置的时间转换格式将源端字段(src_col_1）的值格式化后写入cdc_partition_key中。
  - 表自定义属性：支持通过参数配置单表的部分高级功能，参数详情可参考Hudi高级配置一览表。

- 附加字段编辑：单击操作列“附加字段编辑”可为目的端的Hudi表中增加自定义字段，同时附加字段也会额外加入到Hudi表的建表中。用户可以在已有的源表字段基础上添加多个附加字段，并自定义字段名、选择字段类型、填写字段值。
  - 字段名称：目的端Hudi表新增字段的名称。
  - 字段类型：目的端Hudi表新增字段的类型。
  - 字段值：目的端Hudi表新增字段的取值来源。

    表 7-104 附加字段取值方式

类型	示例
常量	任意字符。

类型	示例
内置变量	■ 源端host ip地址：source.host。 ■ 源端schema名称：mgr.source.schema。 ■ 源端table名称：mgr.source.table。 ■ 目的端schema名称：mgr.target.schema。 ■ 目的端table名称：mgr.target.table。
源表字段	源表中的任一字段。 配置附加字段的取值来源于源表字段时，请注意任务运行过程中不能修改对应源表字段的名称，否则可能导致作业异常。
udf方法	■ substring(#col, pos[, len])： 截取源端col列的子串, 范围在[pos, pos+len)。 ■ date_format(#col, time_format[, src_tz, dst_tz])： 将源端col列按time_format格式化, 可选转换时区。 ■ now([tz])： 获取指定时区的当前时间。 ■ if(cond_exp, str1, str2)： 满足条件表达式cond_exp时返回str1, 否则返回str2。 ■ concat(#col[, #str, ...])： 拼接多个参数, 可为源端列或字符串。 ■ from_unixtime(#col[, time_format])： 将unix时间戳按time_format格式化。 ■ unix_timestamp(#col[, precision, time_format])： 将时间转成unix时间戳, 可显式定义时间格式及转换后精度。

- 自动建表：单击“自动建表”可按照已配置映射规则在目的端数据库自动建表，成功后表建立方式会显示为使用已有表。

图 7-226 自动建表

说明

- Migration仅支持自动建表，不支持自动建库和模式，需用户自行在目的端手动建出库和模式后再使用本功能建表。
- 自动建表时对应的字段类型映射关系请参见**字段映射关系**章节。
- 自动建出的Hudi表会带有3个审计字段，分别是cdc_last_update_date、logical_is_deleted、_hoodie_event_time，并会以_hoodie_event_time作为Hudi表的预聚合键。

步骤9 配置任务属性。

表 7-105 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒

参数	说明	默认值
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** ● 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 ● 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤10 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-227 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-228 启动配置

表 7-106 启动配置参数

参数	说明
同步模式	• 增量同步：从指定时间位点开始同步增量数据。 • 全量+增量：先同步全量数据，随后实时同步增量数据。
时间	增量同步需要设置该参数，指示增量同步起始的时间位点。 **说明** 配置的位点时间早于CDC日志最早时间点时，默认会以日志最新时间点开始消费。

步骤11 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-229 前往监控

----结束

性能调优

若链路同步速度过慢，可参考参见**任务性能调优**章节中对应链路文档进行排查及处理。

7.10.9 PostgreSQL 同步到 DWS 作业配置

支持的源端和目的端数据库版本

表 7-107 支持的数据库版本

源端数据库	目的端数据库
PostgreSQL数据库（PostgreSQL 9.4、9.5、9.6、10、11、12、13、14版本）	DWS集群（8.1.3、8.2.0版本）

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考下表进行赋权。

表 7-108 数据库账号权限

类型名称	权限要求
源数据库连接账号	数据库的CONNECT权限，模式的USAGE权限，表的SELECT权限，序列的SELECT权限，REPLICATION连接权限。 **说明** REPLICATION连接权限的添加方法： • 在源数据库的“pg_hba.conf”配置文件的所有配置前增加一行配置“host replication <src_user_name> <drs_instance_ip>/32 <认证方式>”；认证方式可参考PostgreSQL官方文档**pg_hba.conf文件配置**，常见的认证方式有scram-sha-256等。 • 在源库使用SUPERUSER用户执行语句“select pg_reload_conf();”生效，或重启数据库实例生效。
目标数据库连接账号	目标数据库的每张表必须具有如下权限：INSERT、SELECT、UPDATE、DELETE、CONNECT、CREATE。

说明

- 建议创建单独用于Migration任务连接的数据库账号，避免因为数据库账号密码修改，导致的任务连接失败。
- 连接源和目标数据库的账号密码修改后，请同步修改管理中心对应的连接信息，避免任务连接失败后自动重试，导致数据库账号被锁定影响使用。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考下表。

表 7-109 同步对象范围

类型名称	使用须知
同步对象范围	● 支持同步DML：包括INSERT、UPDATE、DELETE。 ● 不支持同步DDL。 ● 仅支持同步有主键表。 ● 不支持同步视图、外键、存储过程、触发器、函数、事件、虚拟列、唯一约束和唯一索引。 ● 不支持同步无日志表（UNLOGGED TABLE）、临时表、系统模式和系统表。 ● 自动建表支持同步表结构、普通索引、约束（主键、空、非空）、注释。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见下表。

表 7-110 注意事项

类型名称	使用和操作限制
数据库限制	● 库名不可以包含+"%'\<>，模式名和表名不可以包含".'<>，列名不可以包含"和'，列名不能为CTID、XMIN、CMIN、XMAX、CMAX、TABLEOID、XC_NODE_ID、TID等DWS禁止的字段。建议尽量使用常规字符避免任务失败。 ● 目的端数据库中的对象名需要满足约束：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、$。 ● 源数据库的分区表触发器不可以设置为disable。 ● 如果做增量同步：源数据库的“pg_hba.conf” 文件中包含如下的配置： host replication all 0.0.0.0/0 md5

类型名称	使用和操作限制
使用限制	通用： • 实时同步过程中，不支持IP、端口、账号、密码修改。 • PostgreSQL的WAL日志建议保留3天以上。 全量同步阶段： 任务启动和全量数据同步阶段，请不要在源数据库执行DDL操作，否则可能导致任务异常。 增量同步阶段： • 请勿修改源数据库表的主键或者唯一键（主键不存在时），否则可能导致增量数据不一致或任务失败。 • 请勿修改源数据库中表的replica identity属性，否则可能导致增量数据不一致或任务失败。 • Postgres数据源复制槽数达到上限时，无法执行新的作业，可以通过设置max_replication_slots的数值提高复制槽的使用上限或手动删除复制槽（Postgres数据源不支持自动删除复制槽）解决，手动删除请参见**PostgreSQL数据源如何手动删除复制槽?** 。 常见故障排查： 在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。
其他限制	• 目标数据库的block_size参数值必须大于源库中的对应参数值。 • 启动任务前，请确保源库中未启动长事务，源库启动长事务会阻塞逻辑复制槽的创建，进而引发任务失败。 • 任务启动后，不支持源库发生主备倒换。 • 支持目标数据库中的表比源数据库多列场景，但是需要避免以下场景可能导致的任务失败。 – 目标数据库多的列要求非空且没有默认值，源数据库insert数据，同步到目标数据库后多的列为null，不符合目标数据库要求。 – 目标数据库多的列设置固定默认值，且有唯一约束。源数据库insert多条数据后，同步到目标数据库后多的列为固定默认值，不符合目标数据库要求。 • Migration自动建表时，源库中char、varchar、nvarchar、enum、set字符类型长度在目标库会按照字节长自动扩大（因为DWS目标库为字节长）。

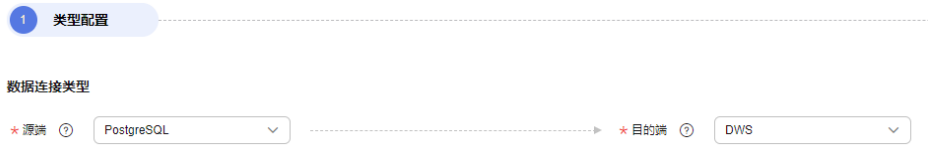
操作步骤

本小节以PostgreSQL到DWS的实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

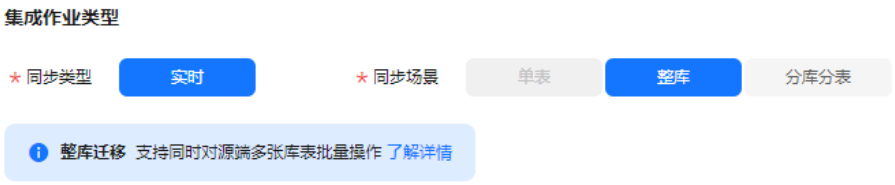
步骤2 选择数据连接类型：源端选PostgreSQL，目的端选DWS。

图 7-230 选择数据连接类型



步骤3 选择集成作业类型：同步类型默认为实时，同步场景包含整库和分库分表场景。

图 7-231 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的PostgreSQL、DWS数据连接和已配置好网络连接的资源组。

图 7-232 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

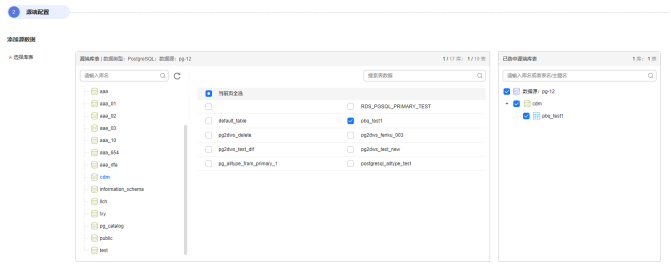
说明

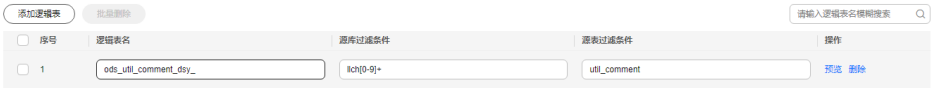
网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端参数。

各同步场景下选择需要同步库表的方式请参考下表。

表 7-111 选择需要同步的库表

同步场景	配置方式
整库	选择需要迁移的PostgreSQL库表。 **图 7-233** 选择库表  库与表均支持自定义选择，即可选择一库一表，也可选择多库多表。

同步场景	配置方式
分库分表	添加逻辑表。 ● 逻辑表名：即最终写入到DWS的表名。 ● 源库过滤条件：支持填入正则表达式，在所有PostgreSQL实例中通过该正则表达式过滤出要写入目标端DWS汇聚表的所有分库。 ● 源表过滤条件：支持填入正则表达式，在过滤出的源端分库中再次过滤出要写入目标端DWS汇聚表的所有分表。 **图 7-234** 添加逻辑表  已添加的逻辑表支持预览表结构及来源库表，单击“操作”列的预览即可。预览逻辑表时，源表数量越多，等待时间可能越长，请耐心等待。 **图 7-235** 逻辑表预览 

步骤7 配置目的端参数。

- 源库表和目标匹配策略。
  各同步场景下源端库表和目标端库表的匹配策略请参考下表。

表 7-112 源库表和目标匹配策略

同步场景	配置方式
整库	– Schema匹配策略。 ■ 与来源库同名：数据将同步至与来源PostgreSQL库名相同的DWS Schema中。 ■ 自定义：数据将同步至自行指定的DWS Schema中。 – 表匹配策略。 ■ 与来源表同名：数据将同步至与来源PostgreSQL表名相同的DWS表中。 ■ 自定义：数据将同步至自行指定的DWS表中。 **图 7-236** 整库场景下源库表和目标匹配策略  **说明** 自定义匹配策略时，支持用内置变量#{source_db_name}和#{source_table_name}标志来源的库名和表名，其中表匹配策略必须包含#{source_table_name}。
分库分表	– 目标端库名：数据将同步至指定的DWS Schema中。 – 表匹配策略：默认与源端配置中填写的逻辑表同名。 **图 7-237** 分库分表场景下源库表和目标匹配策略 

- DWS参数配置。
  其余DWS目的端参数说明请参考下表。

图 7-238 DWS 配置项

表 7-113 DWS 配置项

配置项	默认值	单位	配置说明
写入模式	UPSERT	-	– UPSERT MODE：批量更新入库模式。 – COPY MODE：DWS专有的高性能批量入库模式。
批写最大数据量	50000	条	单批次写入DWS数据的条数，可根据表数据大小和作业内存使用适当调整。
定时批写时间间隔	3	秒	支持配置每批次数据写入DWS的时间间隔。
高级配置	-	-	支持通过参数配置部分高级功能，参数详情可参考DWS高级配置一览表。

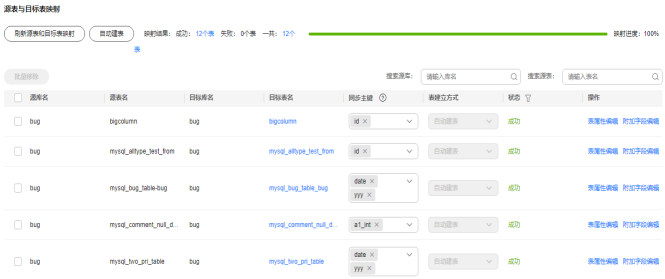
表 7-114 DWS 高级配置一览表

参数名	参数类型	默认值	单位	参数说明
sink.buffer-flush.max-size	int	512	MB	写入DWS时每批数据的最大字节数，可根据作业配置内存和数据大小适当调整。
sink.keyby.enable	boolean	true	-	数据分流开关，在多并发场景下开启数据分流可将数据按规则分配给不同的工作进程写入目的端，可提高写入性能。
sink.keyby.mode	string	table	-	数据分流模式，可选填写： – pk：按数据主键值进行分流。 – table：按表名进行分流。 **说明** ■ 多并发场景下，若开启DDL功能，只能按表名分流，否则可能导致数据不一致。 ■ 确保不会有DDL时，可以选择按主键分流，多并发场景下可提高写入性能。

参数名	参数类型	默认值	单位	参数说明
sink.field.name.case-sensitive	boolean	true	-	同步数据大小写敏感开关，开启后在同步数据时对库名、表名、字段名大小写均敏感。
sink.verify.column-number	boolean	false	-	校验数据列数的开关，链路默认以同名映射方式同步数据，不检验是否所有列均同步。 开启本开关后，若源端与目的端列数不同将认为是数据不一致的场景，导致作业异常。
sink.server.timezone	string	本地时区	-	连接目的端数据库时指定的session时区，支持时区标准写法，例如UTC+8等。
logical.delete.enabled	boolean	false	-	逻辑删除开关。
logical.delete.column	string	logical_is_deleted	-	逻辑删除标记列名称，默认为logical_is_deleted，支持用户自定义。

步骤8 刷新源表和目标表映射，检查映射关系是否正确，同时可根据需求修改表属性、添加附加字段，并通过“自动建表”能力在目的端DWS数据库中建出相应的表。

图 7-239 源表与目标表映射



- 附加字段编辑：单击操作列“附加字段编辑”可为目的端的DWS表中增加自定义字段，同时附加字段也会额外加入到DWS表的建表中。用户可以在已有的源表字段基础上添加多个附加字段，并自定义字段名、选择字段类型、填写字段值。

- 字段名称：目的端DWS表新增字段的名称。
- 字段类型：目的端DWS表新增字段的类型。
- （可选）字段类型长度：目的端DWS表新增字段类型的长度。
- 字段值：目的端DWS表新增字段的取值来源。

表 7-115 附加字段取值方式

类型	示例
常量	任意字符。
内置变量	■ 源端host ip地址：source.host。 ■ 源端schema名称：mgr.source.schema。 ■ 源端table名称：mgr.source.table。 ■ 目的端schema名称：mgr.target.schema。 ■ 目的端table名称：mgr.target.table。
源表字段	源表中的任一字段。 配置附加字段的取值来源于源表字段时，请注意任务运行过程中不能修改对应源表字段的名称，否则可能导致作业异常。
udf方法	■ substring(#col, pos[, len])： 截取源端col列的子串, 范围在[pos, pos+len)。 ■ date_format(#col, time_format[, src_tz, dst_tz])： 将源端col列按time_format格式化, 可选转换时区。 ■ now([tz])： 获取指定时区的当前时间。 ■ if(cond_exp, str1, str2)： 满足条件表达式cond_exp时返回str1, 否则返回str2。 ■ concat(#col[, #str, ...])： 拼接多个参数, 可为源端列或字符串。 ■ from_unixtime(#col[, time_format])： 将unix时间戳按time_format格式化。 ■ unix_timestamp(#col[, precision, time_format])： 将时间转成unix时间戳, 可显式定义时间格式及转换后精度。

- 自动建表：单击“自动建表”可按照已配置映射规则在目的端数据库自动建表，成功后表建立方式会显示为使用已有表。

图 7-240 自动建表

说明

- Migration仅支持自动建表，不支持自动建库和模式，需用户自行在目的端手动建出库和模式后再使用本功能建表。
- 自动建表时对应的字段类型映射关系请参见**字段映射关系**章节。

步骤9 配置任务属性。

表 7-116 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒

参数	说明	默认值
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** ● 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 ● 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤10 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-241 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-242 启动配置

表 7-117 启动配置参数

参数	说明
同步模式	• 增量同步：从指定时间位点开始同步增量数据。 • 全量+增量：先同步全量数据，随后实时同步增量数据。
时间	增量同步需要设置该参数，指示增量同步起始的时间位点。 **说明** 配置的位点时间早于Binlog日志最早时间点时，默认会以日志最新时间点开始消费。

步骤11 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-243 前往监控

----结束

性能调优

若链路同步速度过慢，可参考参见**任务性能调优**章节中对应链路文档进行排查及处理。

7.10.10 Oracle 同步到 DWS 作业配置

支持的源端和目的端数据库版本

表 7-118 支持的数据库版本

源端数据库	目的端数据库
Oracle数据库（10、11、12、19版本）	DWS集群（8.1.3、8.2.0版本）

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考下表进行赋权。

表 7-119 数据库账号权限

类型名称	权限要求
源数据库连接账号	Oracle 库需要开启归档日志，同时需表查询权限和日志解析权限，开通对应权限详情请参考**Oracle数据源如何开通归档日志、查询权限和日志解析权限？**。
目标数据库连接账号	目标数据库的每张表必须具有如下权限：INSERT、SELECT、UPDATE、DELETE、CONNECT、CREATE。

说明

- 建议创建单独用于Migration任务连接的数据库账号，避免因为数据库账号密码修改，导致的任务连接失败。
- 连接源和目标数据库的账号密码修改后，请同步修改管理中心对应的连接信息，避免任务连接失败后自动重试，导致数据库账号被锁定影响使用。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考下表。

表 7-120 同步对象范围

类型名称	使用须知
同步对象范围	● 支持同步的DML：包括INSERT、UPDATE、DELETE。 ● 支持同步的DDL：新增列。 ● 仅支持同步有主键表。 ● 不支持同步视图、外键、存储过程、触发器、函数、事件、虚拟列、唯一约束和唯一索引。 ● 自动建表支持同步表结构、普通索引、约束（主键、空、非空）、注释。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见下表。

表 7-121 注意事项

类型名称	使用和操作限制
数据库限制	● 源端数据库中的库名、表名、字段名不能包含：.-以及非ASCII字符，建议尽量使用常规字符避免任务失败。 ● 目的端数据库中的对象名需要满足约束：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、$。

类型名称	使用和操作限制
使用限制	通用： ● 实时同步过程中，不支持IP、端口、账号、密码修改。 ● Oracle归档日志建议保留3天以上。 ● 禁止对Oracle源库做resetlogs操作，否则会导致数据无法同步且任务无法恢复。 ● 不支持修改源数据库Oracle用户名（SCHEMA名），包括11.2.0.2之前版本通过修改USER$字典表方式及11.2.0.2之后通过ALTER USER username RENAME TO new_username修改SCHEMA名称的场景。 ● Oracle为源端时，暂不支持迁移CLOB、NCLOB和BLOB类型。 ● Oracle为源端时，暂不支持Oracle RAC集群。 全量同步阶段： 任务启动和全量数据同步阶段，请不要在源数据库执行DDL操作，否则可能导致任务异常。 增量同步阶段： ● 支持DML：包括INSERT、UPDATE、DELETE。 ● 支持的DDL：新增列。 ● 不支持混合分区表。混合分区表中的外部分区数据变更不产生DML日志，增量数据同步时无法获取变更信息，会存在数据不一致的风险。 ● 表名和列名长度限制为30个字符。Oracle日志读取采用Oracle logminer，logminer限制了表名和列名在30个字符以内，详情请参见**LogMiner分析日志相关介绍**。 常见故障排查： 在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。
其他限制	● 支持目标数据库中的表比源数据库多列场景，但是需要避免以下场景可能导致的任务失败。 – 目标数据库多的列要求非空且没有默认值，源数据库insert数据，同步到目标数据库后多的列为null，不符合目标数据库要求。 – 目标数据库多的列设置固定默认值，且有唯一约束。源数据库insert多条数据后，同步到目标数据库后多的列为固定默认值，不符合目标数据库要求。 ● Migration自动建表时，源库中char、varchar、nvarchar、enum、set字符类型长度在目标库会按照字节长自动扩大（因为DWS目标库为字节长）。 ● Oracle为源端时全量+增量或增量作业，如果需要同步PDB库中的表，Oracle连接中需要填写CDB库的用户名和密码，不能为PDB用户名和密码，因为Oracle日志统一在存储在CDB库中，同时Oracle logminer只能运行在CDB库中。

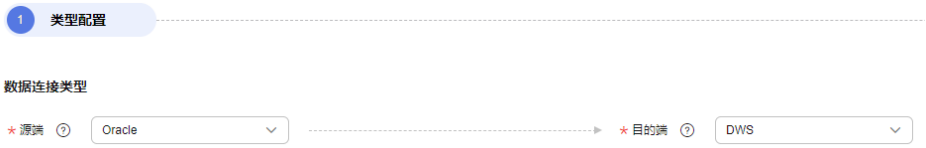
操作步骤

本小节以Oracle到DWS的实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

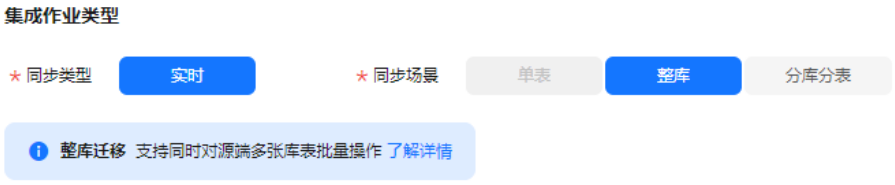
步骤2 选择数据连接类型：源端选Oracle，目的端选DWS。

图 7-244 选择数据连接类型



步骤3 选择集成作业类型：同步类型默认为实时，同步场景包含整库场景。

图 7-245 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的Oracle、DWS数据连接和已配置好网络连接的资源组。

图 7-246 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

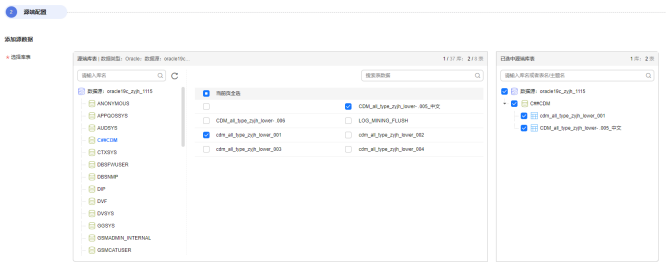
  说明

  网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端参数。

各同步场景下选择需要同步库表的方式请参考下表。

表 7-122 选择需要同步的库表

同步场景	配置方式
整库	选择需要迁移的Oracle库表。 **图 7-247** 选择库表  库与表均支持自定义选择，即可选择一库一表，也可选择多库多表。

步骤7 配置目的端参数。

- 源库表和目标匹配策略。

  各同步场景下源端库表和目标端库表的匹配策略请参考下表。

表 7-123 源库表和目标匹配策略

同步场景	配置方式
整库	– Schema匹配策略。 ■ 与来源库同名：数据将同步至与来源Oracle库名相同的DWS Schema中。 ■ 自定义：数据将同步至自行指定的DWS Schema中。 – 表匹配策略。 ■ 与来源表同名：数据将同步至与来源Oracle表名相同的DWS表中。 ■ 自定义：数据将同步至自行指定的DWS表中。 **图 7-248** 整库场景下源库表和目标匹配策略  **说明** 自定义匹配策略时，支持用内置变量#{source_db_name}和#{source_table_name}标志来源的库名和表名，其中表匹配策略必须包含#{source_table_name}。

- DWS参数配置。
  其余DWS目的端参数说明请参考下表。

图 7-249 DWS 配置项

表 7-124 DWS 配置项

配置项	默认值	单位	配置说明
写入模式	UPSERT	-	– UPSERT MODE：批量更新入库模式。 – COPY MODE：DWS专有的高性能批量入库模式。
批写最大数据量	50000	条	单批次写入DWS数据的条数，可根据表数据大小和作业内存使用适当调整。

配置项	默认值	单位	配置说明
定时批写时间间隔	3	秒	支持配置每批次数据写入DWS的时间间隔。
高级配置	-	-	支持通过参数配置部分高级功能，参数详情可参考DWS高级配置一览表。

表 7-125 DWS 高级配置一览表

参数名	参数类型	默认值	单位	参数说明
sink.buffer-flush.max-size	int	512	MB	写入DWS时每批数据的最大字节数，可根据作业配置内存和数据大小适当调整。
sink.keyby.enable	boolean	true	-	数据分流开关，在多并发场景下开启数据分流可将数据按规则分配给不同的工作进程写入目的端，可提高写入性能。
sink.keyby.mode	string	table	-	数据分流模式，可选填写： – pk：按数据主键值进行分流。 – table：按表名进行分流。 **说明** ■ 多并发场景下，若开启DDL功能，只能按表名分流，否则可能导致数据不一致。 ■ 确保不会有DDL时，可以选择按主键分流，多并发场景下可提高写入性能。
sink.field.name.case-sensitive	boolean	true	-	同步数据大小写敏感开关，开启后在同步数据时对库名、表名、字段名大小写均敏感。
sink.verify.column-number	boolean	false	-	校验数据列数的开关，链路默认以同名映射方式同步数据，不检验是否所有列均同步。 开启本开关后，若源端与目的端列数不同将认为是数据不一致的场景，导致作业异常。

参数名	参数类型	默认值	单位	参数说明
sink.server.timezone	string	本地时区	-	连接目的端数据库时指定的session时区，支持时区标准写法，例如UTC+8等。
logical.delete.enabled	boolean	false	-	逻辑删除开关。
logical.delete.column	string	logical_is_deleted	-	逻辑删除标记列名称，默认为logical_is_deleted，支持用户自定义。

步骤8 刷新源表和目标表映射，检查映射关系是否正确，同时可根据需求修改表属性、添加附加字段，并通过“自动建表”能力在目的端DWS数据库中建出相应的表。

图 7-250 源表与目标表映射

- 附加字段编辑：单击操作列“附加字段编辑”可为目的端的DWS表中增加自定义字段，同时附加字段也会额外加入到DWS表的建表中。用户可以在已有的源表字段基础上添加多个附加字段，并自定义字段名、选择字段类型、填写字段值。
  - 字段名称：目的端DWS表新增字段的名称。
  - 字段类型：目的端DWS表新增字段的类型。
  - 字段值：目的端DWS表新增字段的取值来源。

    表 7-126 附加字段取值方式

类型	示例
常量	任意字符。

- 自动建表：单击“自动建表”可按照已配置映射规则在目的端数据库自动建表，成功后表建立方式会显示为使用已有表。

图 7-251 自动建表

说明

- Migration仅支持自动建表，不支持自动建库和模式，需用户自行在目的端手动建出库和模式后再使用本功能建表。
- 自动建表时对应的字段类型映射关系请参见**字段映射关系**章节。

步骤9 配置DDL消息处理规则。

实时集成作业除了能够同步对数据的增删改等DML操作外，也支持对部分表结构变化（DDL）进行同步。针对支持的DDL操作，用户可根据实际需求配置为正常处理/忽略/出错。

- 正常处理：Migration识别到源端库表出现该DDL动作时，作业自动同步到目的端执行该DDL操作。
- 忽略：Migration识别到源端库表出现该DDL动作时，作业忽略该DDL，不同步到目的端表中。
- 出错：Migration识别到源端库表出现该DDL动作时，作业抛出异常。

图 7-252 DDL 配置

步骤10 配置任务属性。

表 7-127 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1

参数	说明	默认值
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-

参数	说明	默认值
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** • 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 • 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤11 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-253 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-254 启动配置

表 7-128 启动配置参数

参数	说明
同步模式	• 增量同步：从指定时间位点开始同步增量数据。 • 全量+增量：先同步全量数据，随后实时同步增量数据。
时间	增量同步需要设置该参数，指示增量同步起始的时间位点。 **说明** 配置的位点时间早于Binlog日志最早时间点时，默认会以日志最新时间点开始消费。

步骤12 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-255 前往监控

----结束

性能调优

若链路同步速度过慢，可参考参见**任务性能调优**章节中对应链路文档进行排查及处理。

7.10.11 Oracle 同步到 MRS Hudi 作业配置

支持的源端和目的端数据库版本

表 7-129 支持的数据库版本

源端数据库	目的端数据库
Oracle数据库（10、11、12、19版本）	• MRS集群（3.2.0-LTS.x、3.5.x） • Hudi版本（0.11.0）

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考下表进行赋权。

表 7-130 数据库账号权限

类型名称	权限要求
源数据库连接账号	Oracle 库需要开启归档日志，同时需表查询权限和日志解析权限，开通对应权限详情请参考**Oracle数据源如何开通归档日志、查询权限和日志解析权限？**。

类型名称	权限要求
目标数据库连接账号	MRS用户需要拥有Hadoop和Hive组件的读写权限，建议参照图1所示角色及用户组配置MRS用户。 **图 7-256** MRS Hudi 最小化权限  具体MRS集群角色权限管理请参考《MRS集群用户权限模型》。

说明

- 建议创建单独用于Migration任务连接的数据库账号，避免因为数据库账号密码修改，导致的任务连接失败。
- 连接源和目标数据库的账号密码修改后，请同步修改管理中心对应的连接信息，避免任务连接失败后自动重试，导致数据库账号被锁定影响使用。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考下表。

表 7-131 同步对象范围

类型名称	使用须知
同步对象范围	• 支持同步的DML：包括INSERT、UPDATE、DELETE。 • 支持同步的DDL：新增列。 • 仅支持同步主键表。 • 不支持视图、外键、存储过程、触发器、函数、事件、虚拟列、唯一约束、唯一索引、外键索引、Check约束的同步。 • 自动建表支持同步表结构、普通索引、约束（主键、空、非空）、注释。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见下表。

表 7-132 注意事项

类型名称	使用和操作限制
数据库限制	目标数据库中的库名、表名、字段名仅支持数字、字母和下划线，且字段名必须以字母或下划线开头，建议尽量使用常规字符避免任务失败。
使用限制	通用： • 实时同步过程中，不支持IP、端口、账号、密码修改。 • Oracle归档日志建议保留3天以上。 • 禁止对Oracle源库做resetlogs操作，否则会导致数据无法同步且任务无法恢复。 • 不支持修改源数据库Oracle用户名（SCHEMA名），包括11.2.0.2之前版本通过修改USER$字典表方式及11.2.0.2之后通过ALTER USER username RENAME TO new_username修改SCHEMA名称的场景。 • Oracle为源端时，暂不支持迁移CLOB、NCLOB和BLOB类型。 • Oracle为源端时，暂不支持Oracle RAC集群。 • Hudi表使用Bucket索引的场景下不允许更新分区键，否则可能产生重复数据。 • Hudi表使用Bucket索引的场景下主键仅保证单分区内唯一。 • 本链路所使用的Hudi表需带有3个审计字段：cdc_last_update_date、logical_is_deleted、_hoodie_event_time，并会以_hoodie_event_time作为Hudi表的预聚合键。因此，若使用已存在的表，也需要携带这3个审计字段，否则可能导致任务异常。 – cdc_last_update_date：Migration任务处理CDC数据的时间。 – logical_is_deleted：逻辑删除标志。 – _hoodie_event_time：数据在Oracle CDC中的时间戳。 全量同步阶段： 任务启动和全量数据同步阶段，请不要在源数据库执行DDL操作，否则可能导致任务异常。 增量同步阶段： • 支持DML：包括INSERT、UPDATE、DELETE。 • 支持的DDL：新增列。 • 不支持混合分区表。混合分区表中的外部分区数据变更不产生DML日志，增量数据同步时无法获取变更信息，会存在数据不一致的风险。 • 表名和列名长度限制为30个字符。Oracle日志读取采用Oracle logminer，logminer限制了表名和列名在30个字符以内，详情请参见**LogMiner分析日志相关介绍**。 常见故障排查： 在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。

类型名称	使用和操作限制
其他限制	• 支持目标数据库中的表比源数据库多列场景，但是需要避免以下场景可能导致的任务失败。 目标数据库多的列要求非空且没有默认值，源数据库insert数据，同步到目标数据库后多的列为null，不符合目标数据库要求。 • Oracle中表结构长度（所有列长字节数之和，char、varchar2等类型字节长度和编码有关）超过65535时，可能导致同步失败。 • 当使用PDB数据库同步时，由于Oracle LogMiner组件的限制，增量同步时必须打开全部PDB。 • Oralce 12.2及以上版本，由于Oracle LogMiner组件的限制，增量同步不支持表名或列名超过30个字符。 • Oracle为源端时全量+增量或增量作业，如果需要同步PDB库中的表，Oracle连接中需要填写CDB库的用户名和密码，不能为PDB用户名和密码，因为Oracle日志统一在存储在CDB库中，同时Oracle logminer只能运行在CDB库中。

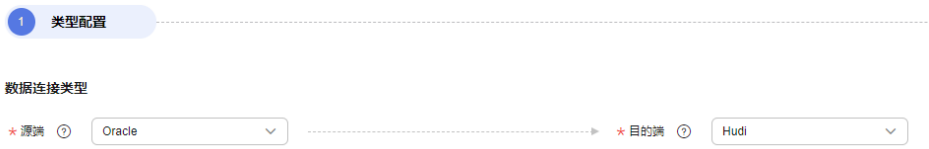
操作步骤

本小节以Oracle到MRS Hudi的实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

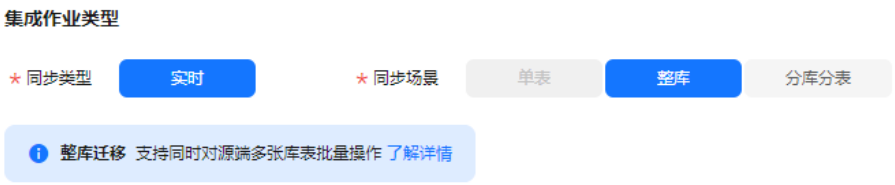
步骤2 选择数据连接类型：源端选Oracle，目的端选Hudi。

图 7-257 选择数据连接类型



步骤3 选择集成作业类型：同步类型默认为实时，同步场景包含整库场景。

图 7-258 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的Oracle、MRS Hudi数据连接和已配置好网络连接的资源组。

图 7-259 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

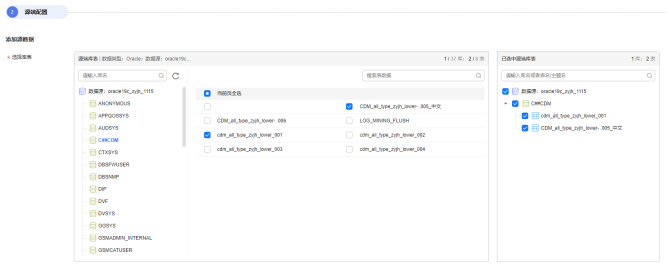
说明

网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？** 章节进行排查。

步骤6 配置源端参数。

选择需要迁移的Oracle库表。

图 7-260 选择库表



库与表均支持自定义选择，即可选择一库一表，也可选择多库多表。

步骤7 配置目的端参数。

- 源库表和目标匹配策略。

  各同步场景下源端库表和目标端库表的匹配策略请参考下表。

表 7-133 源库表和目标匹配策略

同步场景	配置方式
整库	– 库匹配策略。 ▪ 与来源库同名：数据将同步至与来源Oracle 库名相同的Hudi库中。 ▪ 自定义：数据将同步至自行指定的Hudi库中。 – 表匹配策略。 ▪ 与来源表同名：数据将同步至与来源Oracle 库名相同的Hudi表中。 ▪ 自定义：数据将同步至自行指定的Hudi表中。 **图 7-261** 整库场景下源库表和目标匹配策略  **说明** 自定义匹配策略时，支持用内置变量#{source_db_name}和#{source_table_name}标志来源的库名和表名，其中表匹配策略必须包含#{source_table_name}。

- Hudi参数配置。
  其余Hudi目的端参数说明请参考下表。

图 7-262 Hudi 目的端配置项

表 7-134 Hudi 目的端配置项

配置项	默认值	单位	配置说明
数据存储路径	-	-	Hudi自动建表时的warehouse路径，每张表会在warehouse路径下创建子目录。支持填写HDFS和OBS路径，路径格式参考： – OBS路径：obs://bucket/warehouse。 – HDFS路径：/tmp/warehouse。

配置项	默认值	单位	配置说明
Hudi表属性全局配置	-	-	支持通过参数配置部分高级功能，参数详情可参考Hudi高级配置一览表。

表 7-135 Hudi 高级配置一览表

参数名	参数类型	默认值	单位	参数说明
index.type	string	BLOOM	-	Hudi表索引类型。 支持BLOOM和BUCKET索引，数据量较大场景下强烈建议使用BUCKET索引性能更好。
hoodie.bucket.index.num.buckets	int	256	个	Hudi表单分区下Bucket桶数。 **说明** 使用Hudi BUCKET表时需要设置Bucket桶数，桶数设置关系到表的性能，需要格外引起注意。 – 非分区表桶数 = MAX（单表数据量大小（G）/2G*2，再向上取整，4）。 – 分区表桶数 = MAX（单分区数据量大小（G）/2G*2，再后向上取整，1）。 其中，要注意的是： – 需要使用的是表的总数据大小，而不是压缩以后的文件大小。 – 桶的设置以偶数最佳，非分区表最小桶数请设置4个，分区表最小桶数请设置1个。
changelog.enabled	boolean	false	-	Hudi changelog功能开关，开启后Migration作业可输出DELETE和UPDATE BEFORE数据。
logical.delete.enabled	boolean	true	-	逻辑删除开关，changelog开启时必须关闭逻辑删除。
hoodie.write.liststatus.optimized	boolean	true	-	写log文件时是否开启liststatus优化。涉及到大表和分区数据量多的作业，在启动时list会非常耗时，可能导致作业启动超时，建议关闭。

参数名	参数类型	默认值	单位	参数说明
hoodie.index.liststatus.optimized	boolean	false	-	定位数据时是否开启liststatus优化。涉及到大表和分区数据量多的作业，在启动时list会非常耗时，可能导致作业启动超时，建议关闭。
compaction.async.enabled	boolean	true	-	异步compaction开关。compaction操作一定程度会影响实时任务的写入性能，如果用户使用外置的compaction操作对hudi进行compaction，可以考虑设置为false关闭实时处理集成作业的compaction操作。
compaction.schedule.enabled	boolean	true	-	生成compaction计划的开关。compaction计划必须由本服务生成，计划的执行可以交给Spark。
compaction.delta_commits	int	5	次	生成compaction request的频率。compaction request生成频率降低可以使得compaction频率降低从而提升作业性能。如果hudi增量数据较小。可以考虑增大该值。 **说明** 例如配置为40，即每40次commit生成一个compaction request，因为Migration每分钟生成1个commit，那么每个compaction request将间隔40分钟。
clean.async.enabled	boolean	true	-	做历史版本数据文件清理的开关。
clean.retain_commits	int	30	次	要保留的commit数。这些commit关联的数据文件版本将被保留 num_of_commits * time_between_commits 这么长的时间，建议配置为2倍的compaction.delta_commits。 **说明** 例如配置为80，因为Migration每分钟生成1个commit，那么超过80分钟后如果有旧版本数据文件 ，则会生成clean request，且在执行clean时保留最近80个commit。

参数名	参数类型	默认值	单位	参数说明
hoodie.archive.automatic	boolean	true	-	Hudi commit文件老化开关。
archive.min_commits	int	40	次	将旧版commit归档到日志文件中时要保留不归档的最小commit数。建议配置成clean.retain_commits + 1。 **说明** 例如配置成81，那么在触发归档动作时，将会保留最近81次commit文件。
archive.max_commits	int	50	次	触发归档动作的commit数。建议配置成archive.min_commits + 20。 **说明** 例如配置成101，那么将在生成101个commit文件后触发归档commit文件动作。

说明

- 为了达到Migration作业性能最优，建议使用Hudi Bucket索引的MOR表，并根据实际数据量配置Bucket桶数。
- 为了保证Migration作业的稳定性，建议将Hudi Compaction单独拆成Spark作业交由MRS执行，在Migration任务里仅开启生成compaction计划，具体可以参考**如何配置Hudi Compaction的Spark周期任务?** 。

步骤8 刷新源表和目标表映射，检查映射关系是否正确，同时可根据需求修改表属性、添加附加字段，并通过“自动建表”能力在目的端Hudi数据库中建出相应的表。

图 7-263 源表与目标表映射

- 同步主键

  Hudi表必须设置“同步主键”，在源端为非主键表时，必须在字段映射阶段手动勾选主键。
- 表属性编辑

  单击操作列“表属性编辑”可配置Hudi表属性，包含表类型，分区类型及表自定义属性。

图 7-264 Hudi 表属性配置

- 表类型：Hudi的表类型，可选MERGE_ON_READ和COPY_ON_WRITE。
- 分区类型：Hudi表分区类型，可选无分区、时间分区、自定义分区。

  说明

  其中时间分区需要用户指定一个源端表名，选择一个时间转换格式。

  比如时间分区用户指定一个源端表名src_col_1，选择一个时间转换格式，日（yyyyMMdd）、月（yyyyMM）、年（yyyy），自动建表时会在Hudi表默认创建一个cdc_partition_key的字段，系统会根据配置的时间转换格式将源端字段(src_col_1)的值格式化后写入cdc_partition_key中。
- 表自定义属性：支持通过参数配置单表的部分高级功能，参数详情可参考Hudi高级配置一览表。

- 附加字段编辑：单击操作列“附加字段编辑”可为目的端的Hudi表中增加自定义字段，同时附加字段也会额外加入到Hudi表的建表中。用户可以在已有的源表字段基础上添加多个附加字段，并自定义字段名、选择字段类型、填写字段值。
  - 字段名称：目的端Hudi表新增字段的名称。
  - 字段类型：目的端Hudi表新增字段的类型。
  - 字段值：目的端Hudi表新增字段的取值来源。

  表 7-136 附加字段取值方式

类型	示例
常量	任意字符。
内置变量	■ 源端host ip地址：source.host。 ■ 源端schema名称：mgr.source.schema。 ■ 源端table名称：mgr.source.table。 ■ 目的端schema名称：mgr.target.schema。 ■ 目的端table名称：mgr.target.table。
源表字段	源表中的任一字段。 配置附加字段的取值来源于源表字段时，请注意任务运行过程中不能修改对应源表字段的名称，否则可能导致作业异常。

类型	示例
udf方法	■ substring(#col, pos[, len])： 截取源端col列的子串, 范围在[pos, pos+len)。 ■ date_format(#col, time_format[, src_tz, dst_tz])： 将源端col列按time_format格式化, 可选转换时区。 ■ now([tz])： 获取指定时区的当前时间。 ■ if(cond_exp, str1, str2)： 满足条件表达式cond_exp时返回str1, 否则返回str2。 ■ concat(#col[, #str, ...])： 拼接多个参数, 可为源端列或字符串。 ■ from_unixtime(#col[, time_format])： 将unix时间戳按time_format格式化。 ■ unix_timestamp(#col[, precision, time_format])： 将时间转成unix时间戳, 可显式定义时间格式及转换后精度。

- 自动建表：单击“自动建表”可按照已配置映射规则在目的端数据库自动建表，成功后表建立方式会显示为使用已有表。

  图 7-265 自动建表

  说明

  - Migration仅支持自动建表，不支持自动建库和模式，需用户自行在目的端手动建出库和模式后再使用本功能建表。
  - 自动建表时对应的字段类型映射关系请参见**字段映射关系**章节。
  - 自动建出的Hudi表会带有3个审计字段，分别是cdc_last_update_date、logical_is_deleted、_hoodie_event_time，并会以_hoodie_event_time作为Hudi表的预聚合键。

步骤9 配置任务属性。

表 7-137 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档

参数	说明	默认值
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** • 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 • 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤10 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-266 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-267 启动配置

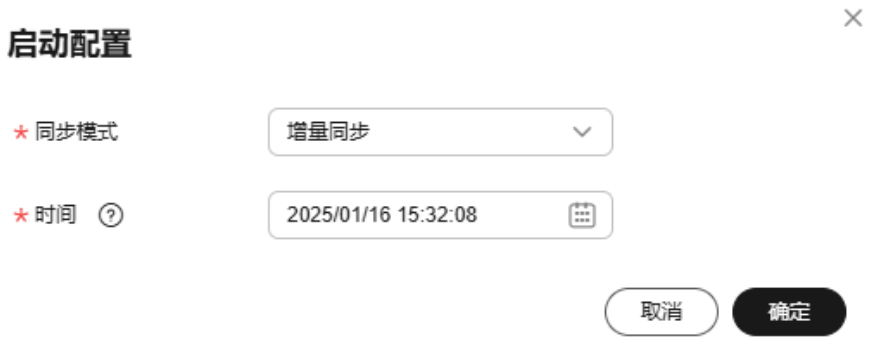


表 7-138 启动配置参数

参数	说明
同步模式	• 增量同步：从指定时间位点开始同步增量数据。 • 全量+增量：先同步全量数据，随后实时同步增量数据。
时间	增量同步需要设置该参数，指示增量同步起始的时间位点。 **说明** 配置的位点时间早于CDC日志最早时间点时，默认会以日志最新时间点开始消费。

步骤11 监控作业。

通过单击作业开发页面导航栏的“前往监控”按钮，可前往作业监控页面查看运行情况、监控日志等信息，并配置对应的告警规则，详情请参见**实时集成任务运维**。

图 7-268 前往监控

----结束

性能调优

若链路同步速度过慢，可参考参见**任务性能调优**章节中对应链路文档进行排查及处理。

7.10.12 MongoDB 同步到 DWS 作业配置

支持的源端和目的端数据库版本

表 7-139 支持的数据库版本

源端数据库	目的端数据库
MongoDB数据库（4.0.0及以上版本）	DWS集群（8.1.3、8.2.0版本）

数据库账号权限要求

在使用Migration进行同步时，源端和目的端所使用的数据库账号需要满足以下权限要求，才能启动实时同步任务。不同类型的同步任务，需要的账号权限也不同，详细可参考下表进行赋权。

表 7-140 数据库账号权限

类型名称	权限要求
源数据库连接账号	目标数据库用户的read/readwrite角色，具有对目标集合授予changeStream和find动作的权限。
目标数据库连接账号	目标数据库的每张表必须具有如下权限：INSERT、SELECT、UPDATE、DELETE、CONNECT、CREATE。

支持的同步对象范围

在使用Migration进行同步时，不同类型的链路，支持的同步对象范围不同，详细情况可参考下表。

表 7-141 同步对象范围

类型名称	使用须知
同步对象范围	● 支持同步的DML：包括INSERT、UPDATE、DELETE。 ● 不涉及且不支持同步的DDL：同步时需要指定好字段的映射。 ● 仅支持同步主键表：MongoDB主键默认为_id。 ● 不涉及且不支持视图、外键、存储过程、触发器、函数、事件、虚拟列、唯一约束、唯一索引、外键索引、Check约束的同步。 ● 不支持自动建表，目标端表需手动建立。

注意事项

除了数据源版本、连接账号权限及同步对象范围外，您还需要注意的事项请参见下表。

表 7-142 注意事项

类型名称	使用和操作限制
数据库限制	• 源端数据库命名规则遵循MongoDB开源规则。 – 数据库名约束： 请勿依靠大小写来区分数据库。例如，不能使用下面这两个名称相似的数据库：salesData 和SalesData。 在 MongoDB 中创建数据库后，在引用数据库时必须使用一致的大小写。例如，如果创建了salesData 数据库，引用时不要使用其他大小写，例如salesdata或SalesData。 在 Windows 上运行时，不得包含以下任意字符：/\. "$*<>:\|?。 在 Unix 和 Linux 系统上运行时，不得包含以下任何字符：/\. "$。 长度限制为 63 字节。 – 集合名称约束： 应以下划线或字母字符开头。 不能包含 null 字符或 $、不能为空字符串（例如 ""）。 不能以 system. 开头。 未分片集合和视图的命名空间长度限制为 255 字节，分片集合的命名空间长度限制为 235 字节。 – 字段名称限制： 长度限制为255字节，不能包含null字符或. $。 • 目的端数据库中的对象名需要满足约束：长度不超过63个字符，以字母或下划线开头，中间字符可以是字母、数字、下划线、$。
使用限制	通用： • 实时同步过程中，不支持IP、端口、账号、密码修改。 • MongoDB实时数据同步不支持单副本的数据源。 • 不支持在运行过程中修改MongoDB库名、集合名。 • 不支持自动建表，需要手动在DWS目标端建立接收表。 • 支持在设置字段映射时，选取extraColumn这个默认自带的源端字段，来确认目标端的某个字段，去接收所有在任务中未被定义过映射MongoDB源端字段。 • 支持提前自定义字段映射，Migration在同步时，若检测到同名字段则传输，若未检测到对应字段则置空。 • 支持DML：包括INSERT、UPDATE、DELETE。 常见故障排查： 在任务创建、启动、全量同步、增量同步、结束等过程中，如有遇到问题，可先参考**常见问题**章节进行排查。
其他限制	-

操作步骤

本小节以MongoDB到DWS的实时同步为示例，介绍如何配置Migration实时集成作业。配置作业前请务必阅读**使用前自检概览**， 确认已做好所有准备工作。

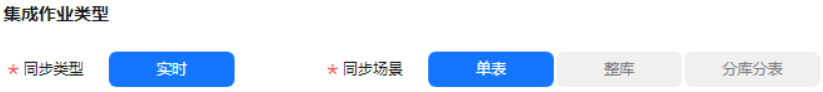
步骤1 参见**新建实时集成作业**创建一个实时集成作业并进入作业配置界面。

步骤2 选择数据连接类型：源端选MongoDB，目的端选DWS。

图 7-269 选择数据连接类型

步骤3 选择集成作业类型：同步类型默认为实时，同步场景包含整库和分库分表场景。

图 7-270 选择集成作业类型



说明

同步场景相关介绍请参见**同步场景**。

步骤4 配置网络资源：选择已创建的MongoDB、DWS数据连接和已配置好网络连接的资源组。

图 7-271 选择数据连接及资源组

说明

无可选数据连接时，可单击“新建”跳转至管理中心数据连接界面，单击“创建数据连接”创建数据连接，详情请参见**配置DataArts Studio数据连接参数**进行配置。

无可选资源组时，可单击“新建”跳转至购买资源组页面创建资源组配置，详情请参见**购买创建数据集成资源组增量包**进行配置。

步骤5 检测网络连通性：数据连接和资源组配置完成后需要测试整个迁移任务的网络连通性，可通过以下方式进行数据源和资源组之间的连通性测试。

- 单击展开“源端配置”触发连通性测试，会对整个迁移任务的连通性做校验。
- 单击源端和目的端数据源和资源组中的“测试”按钮进行检测。

  说明

  网络连通性检测异常可先参考**数据源和资源组网络不通如何排查？**章节进行排查。

步骤6 配置源端参数。

各同步场景下选择需要同步库表的方式请参考下表。

表 7-143 选择需要同步的库表

同步场景	配置方式
单表	选择需要迁移的MongoDB集合。 **图 7-272** 选择库表 

步骤7 配置目的端参数。

- 源库表和目标匹配策略。
  各同步场景下源端库表和目标端库表的匹配策略请参考下表。

表 7-144 源库表和目标匹配策略

同步场景	配置方式
单表	– Schema匹配策略。 ■ 与来源库同名：数据将同步至与来源MongoDB库名相同的DWS Schema中。 ■ 自定义：数据将同步至自行指定的DWS Schema中。 – 表匹配策略。 ■ 与来源表同名：数据将同步至与来源PostgreSQL表名相同的DWS表中。 ■ 自定义：数据将同步至自行指定的DWS表中。 **图 7-273** 整库场景下源库表和目标匹配策略  **说明** 自定义匹配策略时，支持用内置变量#{source_db_name}和#{source_table_name}标志来源MySQL的库名和表名，其中表匹配策略必须包含#{source_table_name}。

- DWS参数配置。
  其余DWS目的端参数说明请参考下表。

图 7-274 DWS 配置项

表 7-145 DWS 配置项

配置项	默认值	单位	配置说明
写入模式	UPSERT MODE	-	– UPSERT MODE：批量更新入库模式。 – COPY MODE：DWS专有的高性能批量入库模式。
批写最大数据量	50000	条	单批次写入DWS数据的条数，可根据表数据大小和作业内存使用适当调整。

配置项	默认值	单位	配置说明
定时批写时间间隔	3	秒	支持配置每批次数据写入DWS的时间间隔。
高级配置	-	-	支持通过参数配置部分高级功能，参数详情可参考**表8 DWS高级配置一览表**。

表 7-146 DWS 高级配置一览表

参数名	参数类型	默认值	单位	参数说明
sink.buffer-flush.max-size	int	512	MB	写入DWS时每批数据的最大字节数，可根据作业配置内存和数据大小适当调整。
sink.keyby.enable	boolean	true	-	数据分流开关，在多并发场景下开启数据分流可将数据按规则分配给不同的工作进程写入目的端，可提高写入性能。
sink.keyby.mode	string	table	-	数据分流模式，可选填写： – pk：按数据主键值进行分流。 – table：按表名进行分流。 **说明** ■ 多并发场景下，若开启DDL功能，只能按表名分流，否则可能导致数据不一致。 ■ 确保不会有DDL时，可以选择按主键分流，多并发场景下可提高写入性能。
sink.field.name.case-sensitive	boolean	true	-	同步数据大小写敏感开关，开启后在同步数据时对库名、表名、字段名大小写均敏感。
sink.verify.column-number	boolean	false	-	校验数据列数的开关，MySQL > DWS默认以同名映射方式同步数据，不检验是否所有列均同步。 开启本开关后，若源端与目的端列数不同将认为是数据不一致的场景，导致作业异常。

参数名	参数类型	默认值	单位	参数说明
sink.server.timezone	string	本地时区	-	连接目的端数据库时指定的session时区，支持时区标准写法，例如UTC+8等。

步骤8 刷新源表和目标表映射，检查映射关系是否正确，同时可根据需求修改表属性、添加附加字段，并通过“自动建表”能力在目的端DWS数据库中建出相应的表。

图 7-275 源表与目标表映射

- 目标字段赋值：单击操作列“目标字段赋值”可自定义MongoDB到DWS的字段映射情况。同时用户可以为所有DWS目标端的字段，设置对应的源表映射字段，或者设置手动赋值的字符串内容。
  - 列名：目的端DWS表字段的名称。
  - 类型：目的端DWS表字段的类型。
  - 字段值：目的端DWS表字段的取值来源。

  表 7-147 字段值的取值方式

类型	示例
手动赋值	任意字符。
源表字段	预设的源表字段：下拉选项中获取或者手动输入的，符合MongoDB字段限制的字段名（参考表4-数据库限制）。 extraColumns：自带的特殊字段名，使用该字段，会将所有没有设置过映射的MongoDB源端字段，写入到该字段中传输到DWS中。

  说明

  - Migration当前不支持MongoDB源端同步链路的自动建表。
  - 自动建表时对应的字段类型映射关系请参见**字段映射关系**章节。

步骤9 配置任务属性。

表 7-148 任务配置参数说明

参数	说明	默认值
执行内存	作业执行分配内存，跟随处理器核数变化而自动变化。	8GB
处理器核数	范围：2-32。 每增加1处理核数，则自动增加4G执行内存和1并发数。	2
并发数	作业执行支持并发数。该参数无需配置，跟随处理器核数变化而自动变化。	1
自动重试	作业失败时是否开启自动重试。	否
最大重试次数	“自动重试”为是时显示该参数。	1
重试间隔时间	“自动重试”为是时显示该参数。	120秒
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据，当脏数据过多时，会影响同步任务的整体同步速度。 链路是否支持写入脏数据，以实际界面为准。 ● 否：默认为否，不记录脏数据。 表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ● 是：允许脏数据，即任务产生脏数据时不影响任务执行。 允许脏数据并设置其阈值时： – 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 – 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据；单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。用户可以在同步任务配置时，配置同步过程中是否写入脏数据，配置脏数据条数（单个分片的最大错误记录数）保证任务运行，即当脏数据超过指定条数时，任务失败退出。	否
脏数据策略	“是否写入脏数据”为是时显示该参数，当前支持以下策略： ● 不归档：不对脏数据进行存储，仅记录到任务日志中。 ● 归档到OBS：将脏数据存储到OBS中，并打印到任务日志中。	不归档

参数	说明	默认值
脏数据写入连接	“脏数据策略”选择归档到OBS时显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	-
脏数据目录	脏数据写入的OBS目录。	-
脏数据阈值	是否写入脏数据为是时显示该参数。 用户根据实际设置脏数据阈值。 **说明** • 脏数据阈值仅针对每个并发生效。比如阈值为100，并发为3，则该作业可容忍的脏数据条数最多为300。 • 输入-1表示不限制脏数据条数。	100
添加自定义属性	支持通过自定义属性修改部分作业参数及开启部分高级功能，详情可参见**任务性能调优**章节。	-

步骤10 提交并运行任务。

作业配置完毕后，单击作业开发页面左上角“提交”，完成作业提交。

图 7-276 提交作业

提交成功后，单击作业开发页面“启动”按钮，在弹出的启动配置对话框按照实际情况配置同步位点参数，单击“确定”启动作业。

图 7-277 启动配置

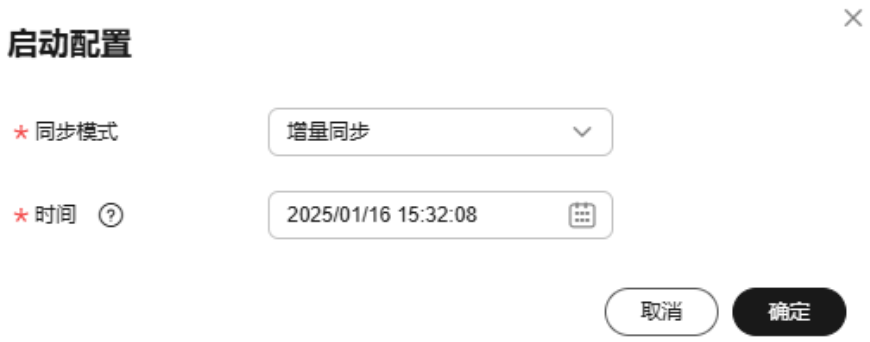


表 7-149 启动配置参数

参数	说明
同步模式	● 增量同步：从指定时间位点开始同步增量数据。 ● 全量+增量：先同步全量数据，随后实时同步增量数据。
时间	增量同步需要设置该参数，指示增量同步起始的时间位点。 **说明** 配置的位点时间早于Binlog日志最早时间点时，默认会以日志最新时间点开始消费。

步骤11 暂停同步任务并修改作业内容。

作业在运行过程中，允许用户通过暂停按钮，暂停该同步任务。用户可以在任务暂停后修改作业内容，提交作业版本后再恢复同步任务。

当前支持的修改操作有：

- 在目标字段赋值中增、减加字段的映射规则。
- 修改全局参数，如处理器核数、是否自动重试等内容。
- 修改数据源的高级属性。

当前不支持的修改操作有：

- 在目标字段赋值中，修改已有的字段映射规则。如，将原本是源表字段赋值的规则，修改成手动赋值的规则。
- 修改源端的集合或修改目标端的表。

----结束

8 数据架构

8.1 数据架构概述

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）；容易被人所理解；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三种：

- **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，可以从数仓规划去设计物理模型。

  - **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。

- **数据集市**

  又称为DM（Data Mart），DM面向展现层，数据有多级汇总，由一个特定的分析对象及其相关的统计指标组成的，向用户提供了以统计粒度为主题的所有统计数据。

在DataArts Studio数据架构中，维度建模是以维度建模理论为基础，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，通过数据集市构建出汇总模型。

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8-1所示。

图 8-1 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到该对象的管理页面。
- **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**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，包括逻辑实体、物理表、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”，可以新建逻辑实体、物理表、维度、事实表和汇总表。创建的过程可以参见**逻辑模型**、**关系建模**、**新建维度**、**新建事实表**、**数据集市**。

- **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。企业模式下，进行同步时，可以选择同步到生产环境或开发环境。系统默认同步到生产环境。

- **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，其中发布审核中，已发布和下线审核中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”，指发布审核后又重新编辑的数据。

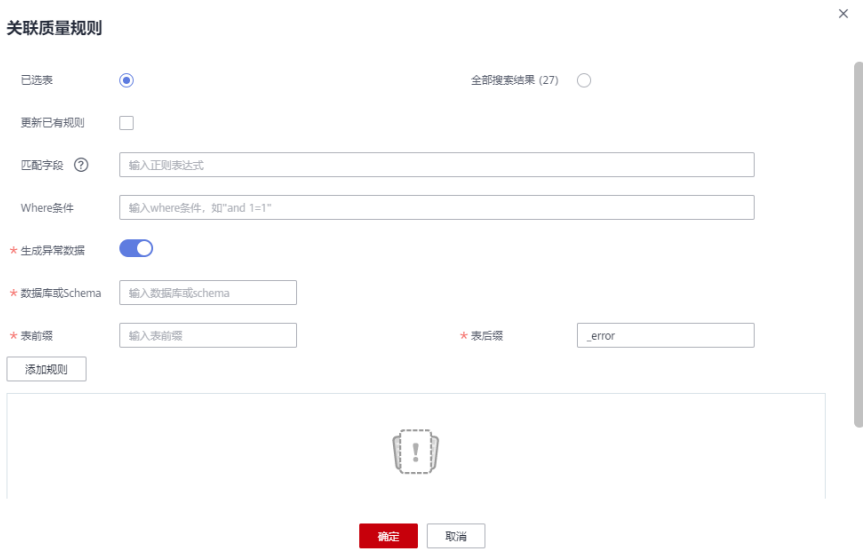
- **发布**

  单击“发布”，可发布数据表。发布审核中、下线审核中、已发布（不带下展）状态的数据表不支持发布。企业模式下，进行发布时，可以选择发布到生产环境或开发环境。系统默认发布到生产环境。

- **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**关联质量规则**。

图 8-2 关联质量规则

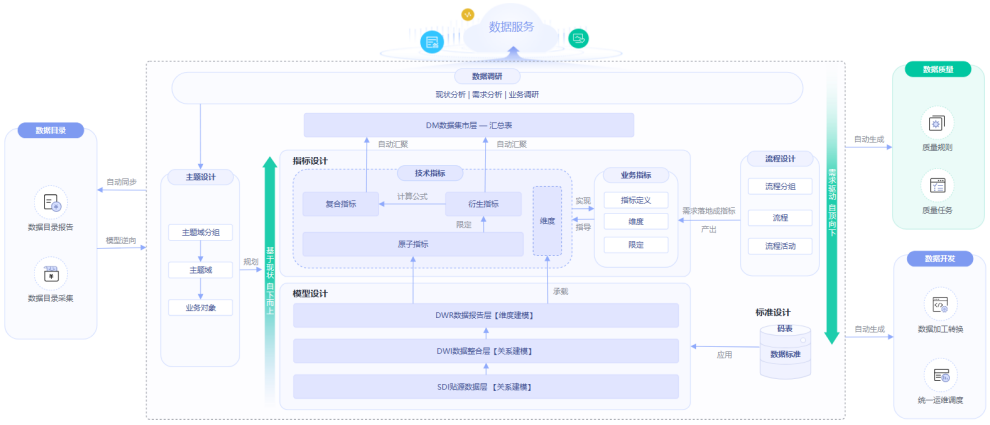


生成异常数据：打开此项，表示异常数据将按照配置的参数存储到规定的库中。

8.2 数据架构使用流程

DataArts Studio数据架构的流程如下：

图 8-3 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。
- **数仓规划**：对数仓分层以及数仓建模进行统一管理。支持用户自定义数仓分层。

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：**应用逻辑模型、关系建模、维度建模和数据集市的方法，进行分层建模。
   - **逻辑模型**：用于创建逻辑模型以及逻辑模型的修改和删除，转化为物理模型。同时，可以对逻辑实体进行创建及发布，进行逆向数据库等操作。
   - **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。
   - **数据集市**：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）。
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。
   - **业务指标**：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。

8.3 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 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。
3. 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。
4. 在“审核人管理”页面，单击“添加”按钮。
5. 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 8-4 添加审核人

说明

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 8-5 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。

8.4 数据调研

8.4.1 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，流程支持三层至七层，如需要修改，请参考**流程层级数**。

1. 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。
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 8-6 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 8-7 新建流程

表 8-1 新建流程参数说明

参数名	说明
流程名称	流程名称，只允许除/、\、<、>、%、"、'、;及换行符以外的字符。
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 8-8 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

“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。流程导出的是全部流程信息，并不是用户的勾选项。

----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 8-9 导入流程

表 8-2 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8-3**。 • **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8-3**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 8-3 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

8.4.2 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：

- **新建主题并发布**

  手动新建并发布主题。

- **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**主题流程配置**对主题层级进行自定义配置。

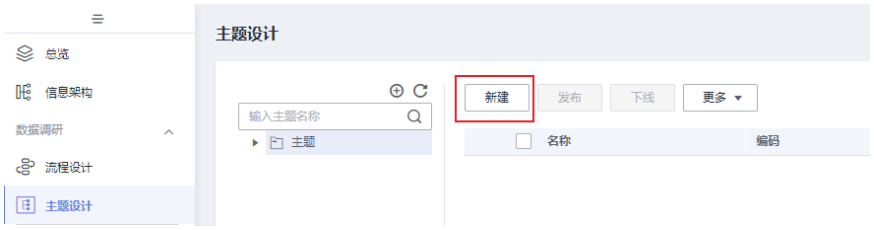
约束与限制

单工作空间允许创建的主题个数最多5000个。

新建主题并发布

1. 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。
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 8-10 主题设计

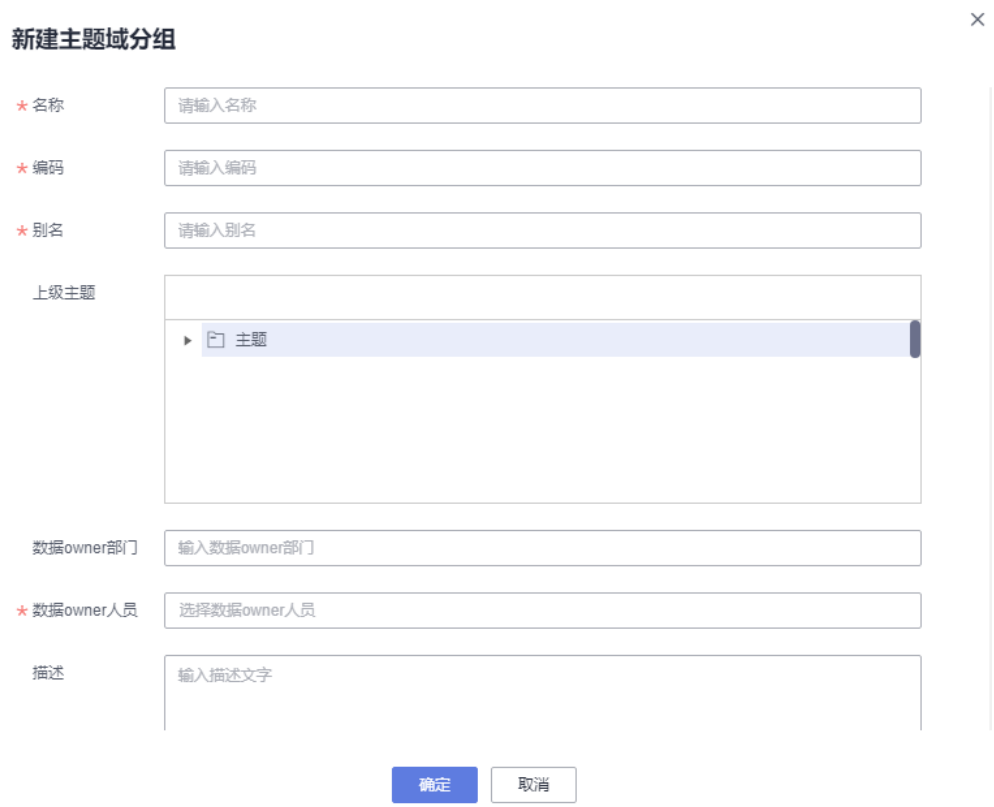


3. 在“新建主题域分组”对话框中，配置如下参数，然后单击“确定”完成主题域分组的创建。

表 8-4 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。只允许英文字母、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 8-11 新建主题

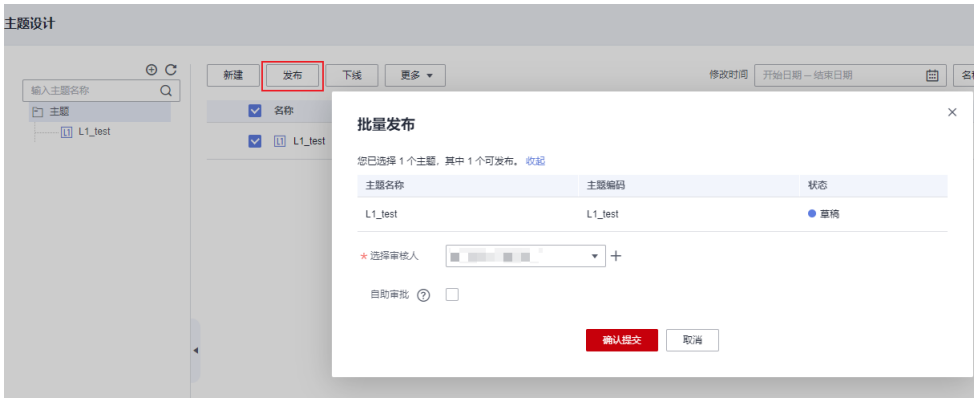


4. 选择新建的主题域分组，单击“发布”，在提交发布对话框中选择审核人，再单击“确认提交”提交审核。审核通过后，返回“主题设计”页面，在列表中可以查看已建好的主题域分组且状态显示为“已发布”，已发布的主题域分组才可被使用。

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，主题域分组状态显示为“已发布”。

图 8-12 发布主题



5. 在一个主题下，还可以新建多个主题。注意，多层主题发布时只能按层级由上至下发布，只有上层主题发布后，下层主题才被允许发布。

说明

创建主题时，如果主题创建到业务对象L3层级时，即创建主题层级出现“新建业务对象”，系统会自动显示“编码”参数，编码规则支持“自动生成”和“自定义”两种方式。

- 自动生成：按照配置中心的**编码规则**自动生成
- 自定义：输入自定义编码

主题设计中，不同L1层级下的业务对象支持重名。

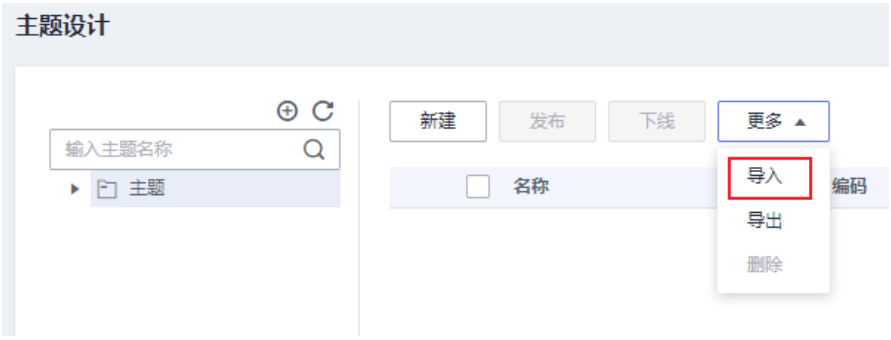
主题层级数目由用户在配置中心的**主题层级**中自定义，系统默认有三个层级，从上到下分别命名为主题域分组（L1）、主题域（L2）、业务对象（L3）。

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 单击上方的“更多 > 导入”按钮，弹出导入主题对话框。

图 8-13 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 8-14 导入配置

表 8-5 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8-6**。 • **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8-6**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 8-6 模板参数说明

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 8-15 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 单击上方的“更多 > 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

- 导出主题时，可以直接勾选自己想要导出的主题名称，单击“导出”按钮，系统会递归导出选中的主题及其所有子主题。
- 导出主题时，可以勾选目录树上的主题，如果右侧没有勾选主题名称，单击“导出”按钮，则会按照选中的目录树上的主题进行递归导出。如果右侧同时勾选了主题名称，则会递归导出选中的主题及其所有子主题。

----结束

管理主题设计

图 8-16 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，在公共空间下可查找所有。

- 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。已发布的主题在编辑后如果要生效，需要在下拉框中选择修改后的草稿，然后进行发布。

- 删除
  您可以在主题列表中，选择一个对象，单击上方“更多 > 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。

8.4.3 逻辑模型

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**
- **通过逆向数据库导入逻辑实体**

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，选择对应工作空间的“数据架构”模块，进入数据架构页面。
2. 在数据架构控制台，单击左侧导航树中的“数据调研 > 逻辑模型”。
3. 在“逻辑模型”页面，单击 按钮新建逻辑模型。

图 8-17 新建逻辑模型

4. 在弹出窗口中配置如下参数，然后单击“确定”。

图 8-18 配置逻辑模型

表 8-7 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。
前缀校验	只能包含英文字母、数字和下划线，且英文字母开头。 **说明** 模型校验前缀：针对关系建模里面的物理表（关系表）、维度建模里面的事实表、数据集市的汇总表的新建、修改、导入表时，会校验是否有前缀，没有的话会校验失败。进行逆向操作时，也会校验是否有前缀。
描述	逻辑模型的描述信息。

5. 更多操作如下。
   - 单击已新建的逻辑模型右侧的“编辑”，可以修改逻辑模型的参数信息。
   - 单击已新建的逻辑模型右侧的“删除”，可以删除逻辑模型。删除操作无法恢复，请谨慎操作。如果模型包含业务表，无法删除。
   - 单击已新建的逻辑模型右侧的“转化为物理模型”，可以将逻辑模型转化为物理模型。具体操作请参见**逻辑模型转换为物理模型**。
   - 单击已新建的逻辑模型的“逻辑实体”或“逻辑属性”或“标准覆盖率”可以跳转到逻辑实体列表页面，查看该逻辑模型的详细内容。

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“逻辑模型”进入逻辑模型页面。

步骤2 在逻辑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 8-19 基本配置

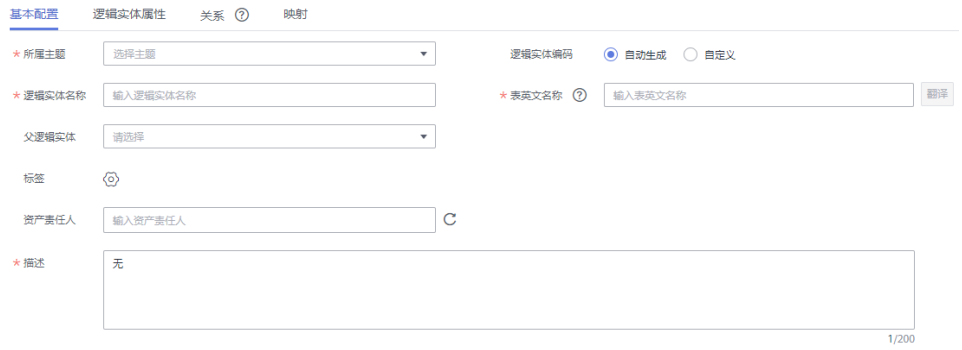


表 8-8 基本配置

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。 只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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。 系统支持通过翻译功能按照已配置的命名词典自动生成表英文名称。
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。

参数名称	说明
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**管理资产标签**，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。 关系建模的数据标签不支持热发布修改，修改标签需要先将表进行下线，待修改好后再进行上线即可。
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8-9**。

图 8-20 添加逻辑实体属性

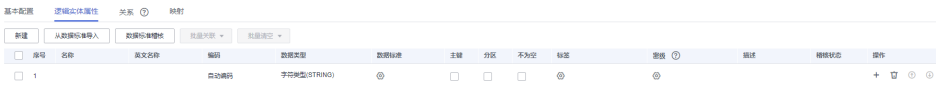


表 8-9 逻辑实体属性参数

参数名称	说明
*名称	只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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**添加数据类型。

参数名称	说明
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**新建数据标准**进行创建。 **说明** – 当逻辑实体发布上线后，如果修改数据标准的编码，需要手动将数据标准的维度表同步至数据目录，否则无法更新逻辑实体详情中的数据标准编码信息。
主键	选中时为主键。 **说明** 当逻辑模型需要转换为物理模型时，该参数有如下限制： 数据连接为MRS Spark连接（通过MRS Spark连接支持MRS Hudi数据源）时，由于Hudi的限制，必须存在字段主键才能数据落库成功，否则会导致表同步失败。
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**管理资产标签**。 – 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。
密级	单击 按钮可以为逻辑实体属性添加密级。 如果没有您想要的密级，可点击**跳转**到数据安全界面中创建需要的密级。 如不使用该功能，可在配置中心 > 模型设计中关闭该功能。
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 8-21 新建关系

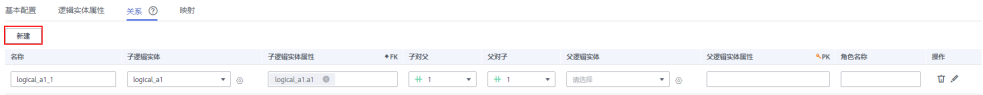


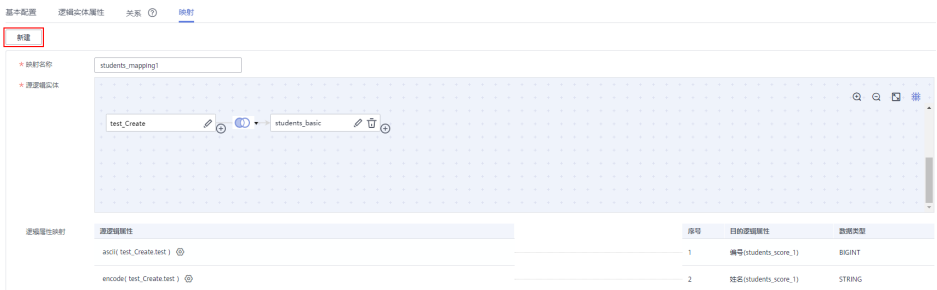
表 8-10 新建关系参数说明

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。

参数名称	说明
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 8-22 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 8-23 设置源表 JOIN 条件

表 8-11 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。

说明

企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

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”：

- 对于新建的逻辑实体，单击“发布”可直接将逻辑实体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实体，单击列表上方的“更多 > 同步”可将逻辑实体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

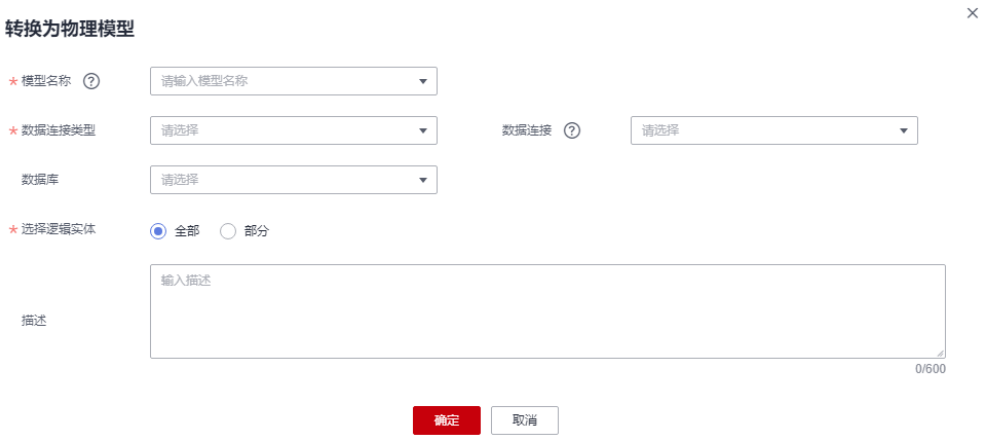
完成逻辑模型的创建后，您可以将逻辑模型转换为物理模型，支持转换为已有的物理模型。

1. 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。
2. 在数据架构控制台，单击左侧导航树中的“逻辑模型”。
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。逻辑模型只支持转换为关系建模的模型。

图 8-24 逻辑模型转化为物理模型

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 8-25 转换为物理模型



说明

逻辑模型转换为物理模型时，系统会先校验是否有前缀。

表 8-12 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。在下拉列表中选择一个已有的模型。

参数名称	说明
*更新已有表	当选择了模型名称后才显示该参数。 ● 不更新 ● 更新 如果选择更新已有表，则需要选择“物理表更新方式”。 – 不删除多余字段 – 删除多余字段
*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。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**配置DataArts Studio数据连接参数**。
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**新建数据库**。
选择逻辑实体	● 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。

通过逆向数据库导入逻辑实体

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逻辑实体目录中，使其变成逻辑实体。

步骤1 在数据架构控制台，单击左侧导航树中的“逻辑模型”，进入逻辑模型页面，选择一个逻辑模型进入逻辑实体列表页面。

步骤2 在逻辑实体列表上方，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 8-13 逆向数据库配置

参数名称	说明
*所属主题	在下拉列表中选择所属主题。
*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。

参数名称	说明
*数据连接	选择数据连接。 如需从其他数据源逆向数据库到逻辑实体目录中，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**配置DataArts Studio数据连接参数**。
*数据库	选择数据库。
*Schema	下拉选择Schema。该参数仅DWS和POSTGRESQL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，在逻辑实体中已存在同名的表，选择是否更新已有的逻辑实体。
名称来源	逆向后表名称/字段名称的来源，可以是描述或者是相应英文名，如表/字段未指定描述则固定使用英文名。 • 来自描述 • 来自英文名称 **说明** 进行逆向数据库配置时，如果逆向后表中文名称/字段中文名称的来源选择“来自描述”，则用中文名在进行描述时，表的字段注释不能重复。
*数据表	选择全部或部分需导入的数据表。

图 8-26 逆向配置

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 8-27 逆向结果

----结束

导入逻辑实体

导入EXCEL

1. 单击逻辑实体列表上方“导入”中的“导入EXCE”。在“导入表”对话框中，选择“导入配置”页签，单击“下载关系建模导入模板”。

   图 8-28 导入 EXCEL

2. 下载关系建模导入模板后，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”退出该页面。

导入LDM

说明

- 导入LDM模型时，请先选择一个主题。不选择则无法导入。
- 当前支持导入逻辑模型。
- 请准备好需要导入的**.ldm**格式的逻辑模型。该逻辑模型是从第三方系统Power Designer导出来的。
- 导入的LDM模型支持的版本：16.x

1. 单击逻辑实体列表上方“导入”中的“导入LDM”。在“导入表”对话框中，选择“导入配置”页签。

图 8-29 导入 LDM

2. 选择是否更新已有数据。
   - 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。
3. 单击“添加文件”，选择提前准备好的**.ldm**格式的逻辑模型。
4. 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。
5. 单击“关闭”退出该页面。

导出逻辑实体

1. 单击逻辑实体列表上方“导出”，进入“导出模型”对话框。
2. 选择“导出对象”。

   选择“表”或者“DDL”。

   当选择DDL时，需要**选择表**，选择“全部”或者“部分”的表。选择部分表示，需要勾选所要导出的表。
3. 单击“确定”。

逻辑实体更多操作

- 同步

  在逻辑实体列表中，选择需要同步的逻辑实体，单击列表上方的“同步”，单击“确定”，完成逻辑实体的同步。只有当表处于已发布状态时，才能执行此操作。

  说明

  逻辑实体关联了质量规则进行发布后，在数据质量作业目录上面单击“同步主题为目录”后，数据架构自动生成的质量作业，会按照主题结构同步到数据质量对应的目录下。

- 发布

  在逻辑实体列表中，选择需要发布的逻辑实体，单击列表上方的“发布”或者单击“操作”列的“发布”，选择审核人，再单击“确认提交”提交审核，审核通过后完成发布。

  说明

  企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

  选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

  如果勾选“自助审批”，勾选后审批单将自动处理，此功能为体验功能，不推荐在真实项目中使用。

- 下线

  在逻辑实体列表中，选择需要下线的逻辑实体，单击列表上方的“下线”或者单击“操作”列的“更多 > 下线”，进行逻辑实体下线。只有当表处于已发布状态时，才能执行此操作。

- 修改主题

  在逻辑实体列表中，选择需要修改主题的逻辑实体，单击列表上方的“修改主题”，可以修改逻辑实体的主题。

- 删除

  在逻辑实体列表中，选择需要删除的逻辑实体，单击列表上方的“删除”，可以删除逻辑实体。只有当表处于草稿/已驳回/已下线状态时，才能执行此操作。

- 标签

  在逻辑实体列表中，选择需要设置标签的逻辑实体，单击列表上方的“标签”，进入后添加标签，单击“确定”，完成逻辑实体的标签设置。

  说明

  输入文字并回车可临时添加标签，整页信息提交后才可新建标签。标签最多可添加20个。

  逻辑实体可以通过标签过滤进行模糊查询。

- 编辑

  在逻辑实体列表中，选择需要编辑的逻辑实体，单击“操作”列的“编辑”，进入编辑页面进行编辑。编辑逻辑实体时，支持关联质量规则。单击“关联质量规则”按钮，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”。

- 发布历史

  在逻辑实体列表中，选择需要查看发布历史的逻辑实体，单击“操作”列的“更多 > 发布历史”，进入后可查看逻辑实体的发布历史和版本对比。

- 浏览SQL

在逻辑实体列表中，选择需要预览SQL的逻辑实体，单击“操作”列的“更多 > 预览SQL”，进入后可预览逻辑实体的SQL信息。

8.5 标准设计

8.5.1 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，选择对应工作空间的“数据架构”模块，进入数据架构页面。
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 8-30 码表管理页面

4. 在弹出窗口中进行参数配置，单击“确定”。

图 8-31 新建码表目录

表 8-14 参数描述

参数名称	说明
*目录名称	只允许除/、\、<、>和.以外的字符。
*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 8-32 基础配置

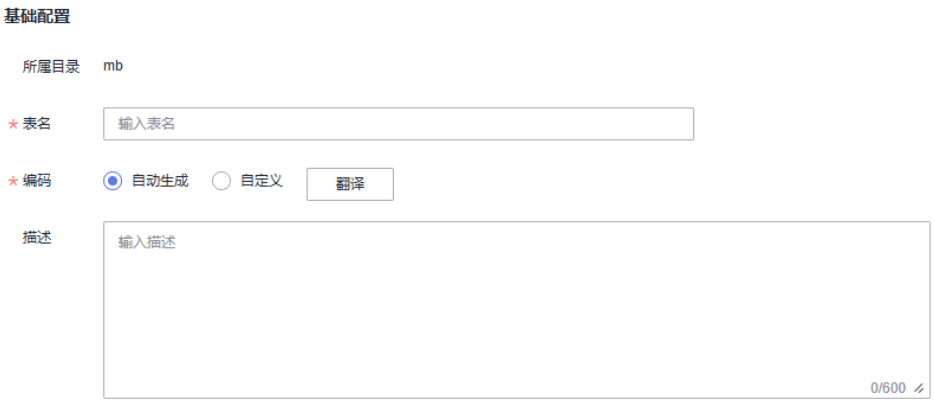
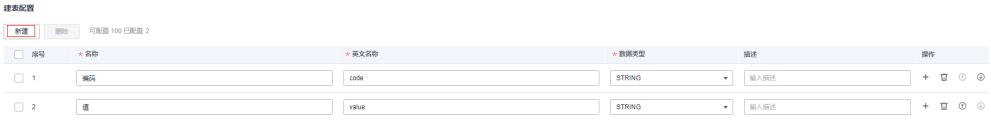


表 8-15 基础配置

参数名称	说明
*表名	码表名称。 只允许除\、<、>、%、''、'、;及换行符以外的字符。
*编码	码表的英文名称。支持自动生成码表，也可选择自定义手动输入。只能包含英文字母、数字和下划线，且以英文字母开头。 选择自定义时，系统支持通过翻译功能按照已配置的命名词典自动进行编码生成码表英文名称。
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 8-33 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”。

   选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

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 8-34 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

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 8-35 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 8-36 导入码表

表 8-16 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** 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8-17**。 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页的名称可以是码表名称或码表编码。 – 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 8-17 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。只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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。
*英文名称	字段英文名称。只能包含英文字母、数字、下划线，且以英文字母开头。
*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或码表编码的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。

说明

如果Sheet页的命名过长，系统会自动将超长的部分进行截断。

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。

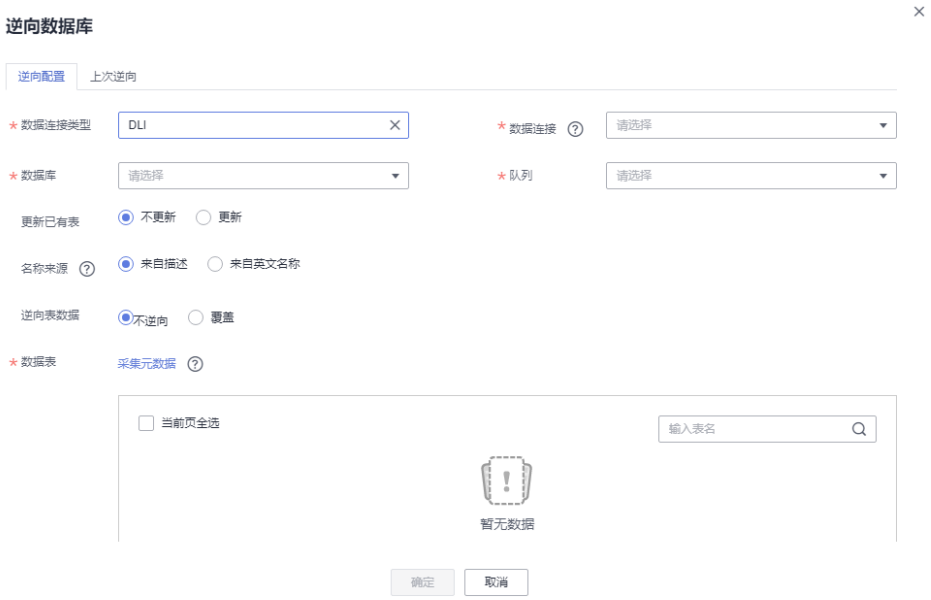
步骤2 在左侧的码表目录树中，选中一个目录，然后在码表列表上方，单击“逆向数据库”。

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 8-18 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**配置DataArts Studio数据连接参数**。
*数据库	选择数据库。
*Schema	下拉选择Schema。该参数仅DWS和POSTGRESQL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。
名称来源	逆向后表名称/字段名称的来源，可以是描述或者是相应英文名，如表/字段未指定描述则固定使用英文名。 • 来自描述 • 来自英文名称 **说明** 进行逆向数据库配置时，如果逆向后表中文名称/字段中文名称的来源选择“来自描述”，则用中文名在进行描述时，表的字段注释不能重复。
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，同时将表数据导入到该码表中。
*数据表	选择一个或多个需导入的数据表。

图 8-37 逆向配置



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 8-38 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。

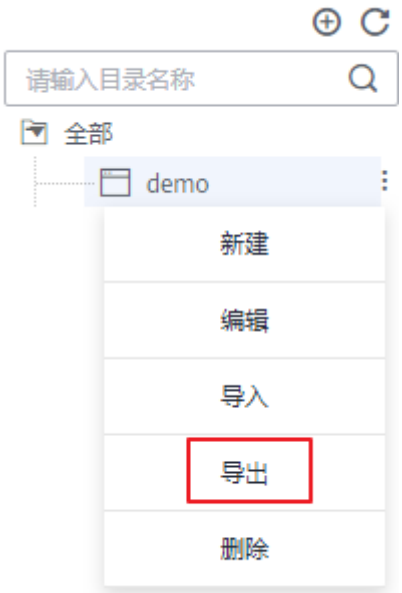
- **导出码表**

  在码表列表中，选中所需导出的码表，然后单击“更多 > 导出”。

图 8-39 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 8-40 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，如果码表为发布审核中、已发布或下线审核中状态，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

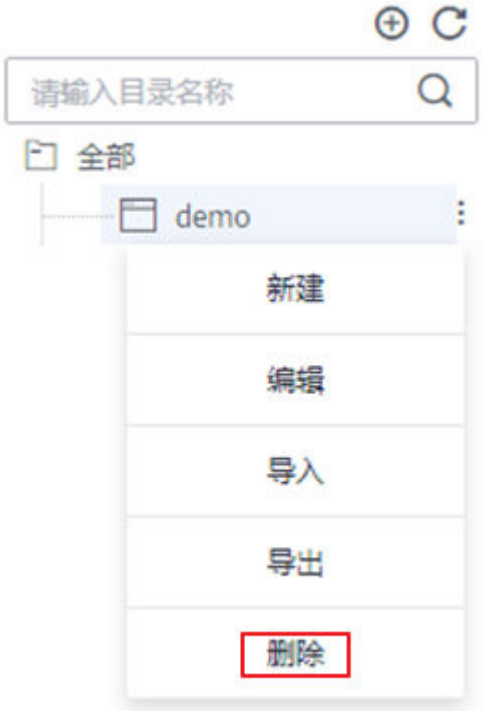
删除码表目录

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 8-41 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、下线或发布等操作。

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。

说明

- 普通空间均可查询到“公共层空间”目录下创建的码表，“公共层空间”无法反向查询到普通空间目录下创建的码表。
- 普通空间仅对本空间内创建的码表和目录有编辑权限，不支持对“公共层空间”的码表和其所属的目录进行操作，仅能查看引用。

图 8-42 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。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”。

- **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。

8.5.2 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。

约束与限制

单工作空间允许创建的数据标准目录最多500条，个数最多20000个。

新建数据标准目录

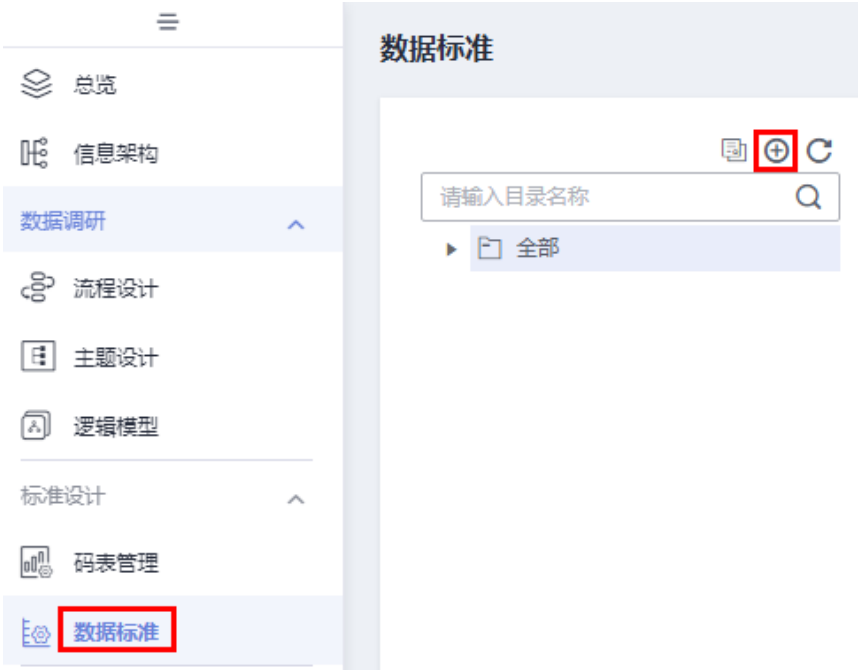
1. 在DataArts Studio控制台首页，选择对应工作空间的“数据架构”模块，进入数据架构页面。
2. 在数据架构控制台，单击左侧导航树中的“数据标准”。

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 8-43 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 8-44 新建数据标准目录

表 8-19 参数描述

参数名称	说明
*目录名称	只允许除/、\、<、>和.以外的字符。
*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。

单击按钮，可以刷新目录。

单击 按钮，可以刷新目录，可以同步主题目录到数据标准目录。

说明

- 同步目录前，请检查当前空间是否有已发布主题。如果没有已发布主题，同步时系统会报错提示。
- 同步目录时，最多同步五级主题到数据标准目录（目录层级不能超过5层），五级之后的主题不做处理。同步后的目录数量不能超过配额（一般是500），否则系统将报错提示并取消同步操作。每次同步之前系统会自动检测数据标准的目录是否有空目录（该目录及其子目录下没有数据标准），有空目录则进行删除。
- 由主题目录同步过来的目录显示为L1~L5图标，数据标准自有的目录显示原来的图标。

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。

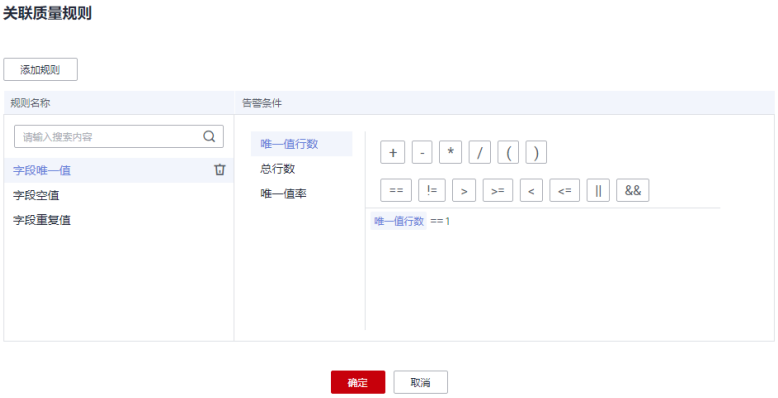
步骤2 在新建数据标准页面中，请参考**表8-20**配置参数。

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8-20**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 8-20 数据标准参数说明

参数名称	说明
*标准名称	只允许除\、<、>、%、"、'、;及换行符以外的字符。 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8-65**。
英文名称	在“配置中心”开启“英文名称”时，显示该参数。您可以手动进行配置数据标准英文名称。 系统支持通过翻译功能按照已配置的命名词典自动生成英文名称。

参数名称	说明
*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**添加数据类型。
英文名称	数据标准的英文名称。 只能包含英文字母、数字、左右括号、空格和下划线，且以英文字母开头。
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。 ● 选择 = 可以设置为具体的数值。输入1~10000之间的数值。 ● 选择 ≤ 可以设置为一个范围。输入数据范围的临界值，输入值范围1~10000。 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。
允许值	在“配置中心”开启“是否有允许值”后，由用户自行输入。输入一个值并按回车即可添加一个允许值，支持添加多个允许值，最多支持20个。
引用码表	在“配置中心”开启“引用码表”后，需要配置该字段。 ● 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**新建码表**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。 ● 如果已开启公共层空间，在普通空间选择码表时，需要手动选择引用码表来源为“选择公共层空间数据”或“选择本空间数据”。“选择公共层”开启后，可以将公共层空间的码表引用到普通空间。
码表字段	选择码表相应的字段。

参数名称	说明
质量规则	在“配置中心 > 标准模板管理”页面中的“质量规则”勾选的情况下，创建数据标准时，会显示质量规则选项。进行关联质量规则时，可以关联系统规则也可以关联自己创建的质量规则。 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 8-45** 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。
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。

步骤4 选中待发布的数据标准，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“数据标准”页面，在列表中可以查看已建好的数据标准且状态显示为“已发布”，已发布的数据标准才可被使用。

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，状态显示为“已发布”。

选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

----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 8-46 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8-21、表8-22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 8-21 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。 只允许除\、<、>、%、"、'、;及换行符以外的字符。

参数名称	说明
*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8-65**。
*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**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（1，20），输入值范围1~10000。 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“配置中心 > 标准模板管理”页面中的“质量规则”勾选的情况下，下载的导入模板中会显示“质量规则”Sheet页，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。

表 8-22 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
表达式	只有当“规则名称”配置为“表达式校验”或者“合法性校验”时，需要配置表达式。

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 8-47 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。

说明

- 普通空间均可查询到“公共层空间”目录下创建的数据标准，“公共层空间”无法反向查询到普通空间目录下创建的数据标准。
- 普通空间仅对本空间内创建的数据标准和目录有编辑权限，不支持对“公共层空间”的数据标准和其所属的目录进行操作，仅能查看引用。

图 8-48 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，其中发布审核中，已发布和下线审核中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“自助审批”。

说明

如果当前账号在审批人列表中，才会有“自助审批”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“自助审批”，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

8.6 模型设计

8.6.1 数仓规划

数仓规划，目前系统默认的数仓分层包含SDI、DWI、DWR、DM（Data Mart）等4层，支持用户自定义数仓分层。数仓规划对数仓分层以及数仓模型进行统一管理。

- 关系建模下包含SDI层和DWI层两层模型，物理模型归属于两层模型之一。
  - 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。

    说明

    物理模型设计时的考虑事项如下：
    - 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。
- 维度建模需要基于维度，新建DWR层模型，最终将数据汇总到DM层模型中。
  - 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。
- 数据集市，面向展现层，数据有多级汇总。
  - DM（Data Mart）：又称数据集市。DM面向展现层，数据有多级汇总。

系统默认的数仓分层的四层层级的名称支持由管理员自定义，单击层级名后的 编辑 即可重命名。重命名建议能够区分不同层级，规则为只能包含英文字母、中文、数字、下划线，且以英文字母或中文开头。

物理模型、维度模型、数据集市，都是模型，在数仓规划进行统一管理。

说明

数仓规划支持细粒度权限管控，在数据安全模块对数据架构模型目录权限管控策略进行配置。详细操作请参见**配置目录权限（高级特性）**。

新建数仓分层

数仓分层支持用户根据实际业务场景进行自定义。具体操作如下：

1. 进入数据架构主页面。
2. 在数据架构控制台，单击左侧导航树中的“模型设计 > 数仓规划”。
3. 单击一个数仓分层右侧的“新建”，选择“添加至前面”或“添加至后面”，进入“新建数仓分层”页面。

   说明

   “添加至前面”或“添加至后面”表示新建的数仓分层在当前数仓分层的前面或者后面。

图 8-49 自定义数仓分层

4. 配置数仓分层相关参数。

图 8-50 新建数仓分层

表 8-23 数仓分层参数说明

参数	说明
*分层名称	定义数仓分层名称。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。输入长度不能超过10个字符。

参数	说明
*分层类型	选择分层类型。分层类型选择以后不支持修改。 ● 关系建模 ● 维度建模 ● 数据集市 **说明** 1. 关系建模一般用于业务系统及数仓贴源层、整合层的建模。 2. 维度建模用于数仓公共层或数据报告层的建模。 3. 数据集市用于汇总表和应用表等数据应用表的建模。
描述	数仓分层描述信息。支持的长度0~200字符。
禁用自定义项	选择自定义项。如果没有自定义项，则表示没有可禁用的自定义项。

5. 单击“确定”。数仓分层新建完成。
6. 更多操作如下：
   - 单击已新建的数仓分层右侧的“编辑”，可以修改数仓分层的参数信息，分层类型不支持修改。
   - 单击已新建的数仓分层右侧的“删除”，可以删除数仓分层。该分层下有模型数据，不可删除。

新建模型

1. 进入数据架构主页面。
2. 在数据架构控制台，单击左侧导航树中的“模型设计 > 数仓规划”。
3. 单击一个数仓分层下面的“添加模型”，进入“新建模型”页面。
4. 配置模型相关参数。

图 8-51 新建模型

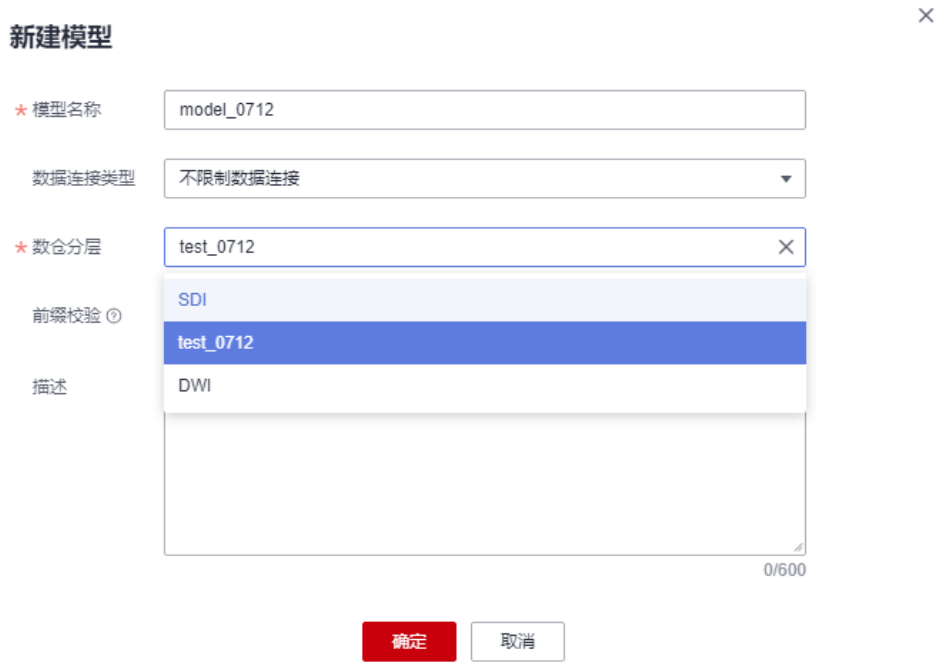


表 8-24 模型参数说明

参数	说明
*模型名称	定义模型名称。只能包含中文、英文字母、数字和下划线。
数据连接类型	选择数据连接类型。 ● 不限制数据连接 ● 选择数据连接

参数	说明
*数仓分层	● 如果是在DWI层、SDI层或者自定义关系建模数仓分层，此处支持选择DWI、SDI、自定义数仓分层。 **说明** ● **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。 ● 如果是在DWR层或者自定义维度建模数仓分层，此处仅可选择DWR、自定义数仓分层。 ● 如果是在DM层或者自定义数据集市数仓分层，此处仅可选择DM、自定义数仓分层。
前缀校验	输入检验前缀。只能包含英文字母、数字和下划线，且以英文字母开头。 **说明** 模型校验前缀，针对关系建模里面的物理表（关系表）、维度建模里面的事实表、数据集市的汇总表的新建、修改、导入表时，会校验是否有前缀，没有的话会校验失败。进行逆向操作时，也会校验是否有前缀。
描述	数仓模型描述信息。支持的长度0~600字符。

5. 单击“确定”。数仓模型新建完成。
6. 更多操作如下：
   - 单击已新建的数仓模型右侧的“编辑”，可以修改数仓模型的参数信息，数据连接类型不支持修改。
   - 单击已新建的数仓模型右侧的“删除”，可以删除数仓模型。删除操作无法恢复，请谨慎操作。如果模型包含业务表，无法删除。
   - 单击已新建的数仓模型的“数据表”或“字段”或“标准覆盖率”可以跳转到对应的数仓分层页面。比如，单击DWI数仓分层模型的“数据表”会跳转到“关系建模”页面。
   - 如果数仓模型比较多，可以单击“查看更多”和“收起更多”进行折叠展示。
   - “未分层”的数仓模型会在页面上方显示。支持编辑和删除。
     - 单击“编辑”，可以修改数仓模型的参数信息，可以给未分层的数仓模型配置数仓分层（此处支持选择DWI、SDI、自定义数仓分层）。数据连接类型不支持修改。

- 单击“删除”，可以修改数仓模型的参数信息，可以删除数仓模型。删除操作无法恢复，请谨慎操作。如果模型包含业务表，无法删除。

8.6.2 关系建模

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**
- **通过逆向数据库导入物理表**

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

数仓分层和模型管理相关功能已迁移至数仓规划页面。创建物理模型请参见**数仓规划**。

新建表并发布

当您在数仓规划中完成关系模型的创建之后，您就可以在关系建模中新建业务表（物理表）。

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

步骤2 在页面中间栏位的最上方的下拉列表中选择所需要建表的物理模型，或者在数仓规划中选择所需的物理模型，单击进入，然后单击列表上方的“新建”按钮新建一个表。

图 8-52 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 8-53 表基本配置

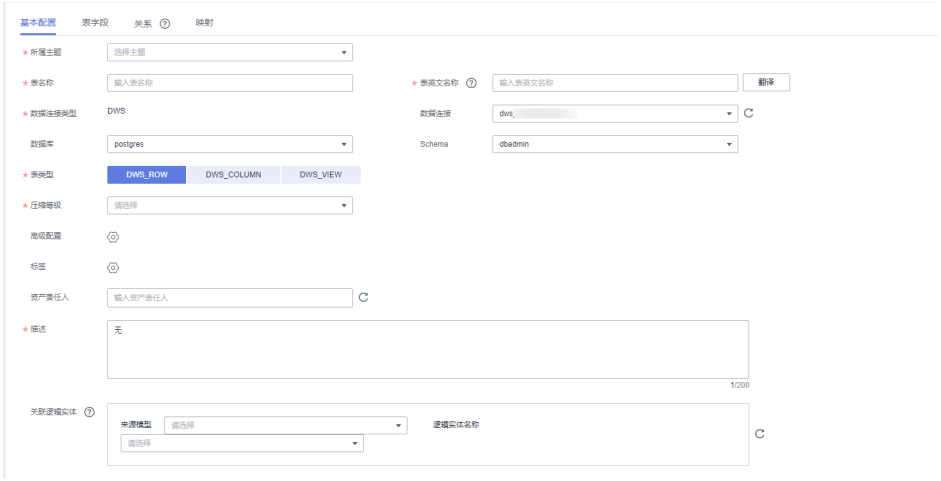


表 8-25 基本配置

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。 只允许除\、<、>、%、"、'、;及换行符以外的字符。 **说明** 物理模型表字段中文名长度不能超过200个字符。
*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。 系统支持通过翻译功能按照已配置的命名词典自动生成表英文名称。
*数据连接类型	系统默认为数仓分层中配置为数据连接类型。不可修改。
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**配置DataArts Studio数据连接参数**。
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。

参数名称	说明
*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模型支持HIVE_TABLE和HIVE_EXTERNAL_TABLE。 MRS_SPARK模型支持HUDI_COW和HUDI_MOR。 POSTGRESQL模型仅支持POSTGRESQL_TABLE。 MRS_CLICKHOUSE模型仅支持CLICKHOUSE_TABLE。 Oracle模型仅支持ORACLE_TABLE。 MySQL模型仅支持MYSQL_TABLE。 DORIS模型仅支持DORIS_TABLE。
路径	该参数仅数据源为MRS_HIVE且表类型选择HIVE_EXTERNAL_TABLE时有效。 只支持英文字母、数字、左斜杠(/)、英文句号(.)、中划线(-)、下划线(_)、冒号(:)。
压缩等级	当数据连接类型为DWS时，可选择压缩等级，以减少数据存储成本。 不同表类型可选以下压缩等级： – DWS_ROW：“NO”、“YES”。 – DWS_COLUMN：“NO”、“LOW”、“MIDDLE”、“HIGH”。 – DWS_VIEW：不支持设置压缩等级。
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**管理资产标签**，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。 关系建模的数据标签不支持热发布修改，修改标签需要先将表进行下线，待修改好后再进行上线即可。
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。
关联逻辑实体	在下拉列表中手动选择需要关联的逻辑实体以及逻辑实体所在的来源模型。 也可单击右侧的刷新按钮，由系统自动同步与物理表主题同名的来源模型以及和物理表英文名称同名的逻辑实体。同一逻辑实体可关联多个物理表。

2. 在“表字段”页面添加所需要的字段。

图 8-54 添加所需表字段

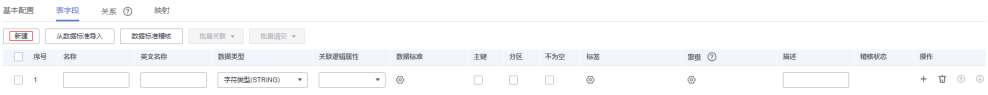


表 8-26 表字段参数

参数名称	说明
名称	只允许除\、<、>、%、"、'、;及换行符以外的字符。
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**添加数据类型。
关联逻辑属性	如果表配置已关联逻辑实体，则此处在下拉列表中手动选择需要关联的逻辑属性，可以将表字段与逻辑实体中的逻辑属性进行关联。

参数名称	说明
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**新建数据标准**进行创建。
主键	选中时为主键。 **说明** 数据连接为MRS Spark连接（通过MRS Spark连接支持MRS Hudi数据源）时，由于Hudi的限制，必须存在字段主键才能数据落库成功，否则会导致表同步失败。
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**管理资产标签**。 – 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。
密级	单击 按钮可以为逻辑实体属性添加密级。 如果没有您想要的密级，可点击**跳转**到数据安全界面中创建需要的密级。 如不使用该功能，可在配置中心 > 模型设计中关闭该功能。
描述	描述信息。
稽核状态	表示是否进行数据标准稽核。 单击“数据标准稽核”，进行数据标准稽核。
操作	相关操作按钮。

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 8-55 新建关系（可选）

表 8-27 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。

参数名称	说明
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”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8-56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8-28。

图 8-56 配置映射关系

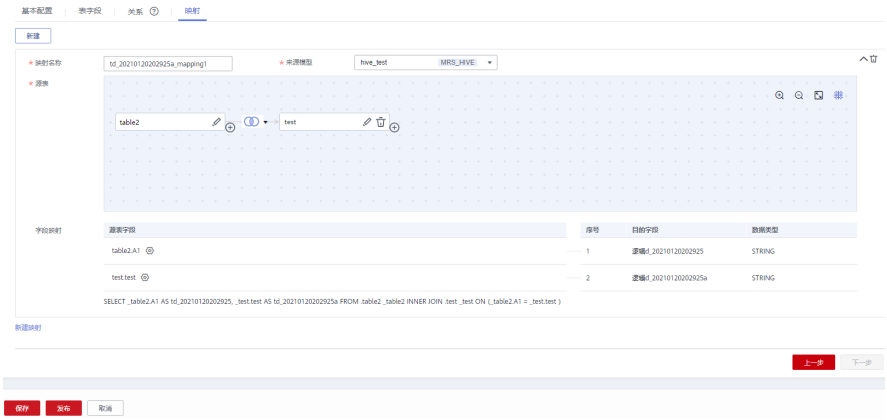


表 8-28 映射参数

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，请参见关系建模进行创建。

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 8-57** 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. （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 8-58 新建视图

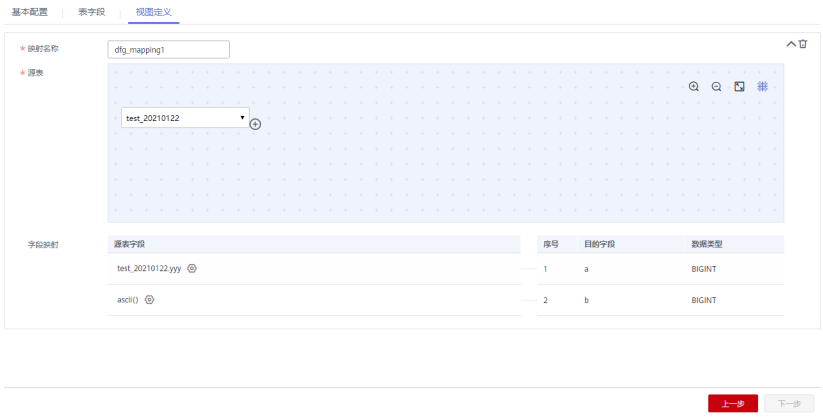


表 8-29 视图定义参数

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 8-59** JOIN 条件界面 
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。

说明

企业模式下，进行发布时，可以选择发布表到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

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。
- 在开启**“同步业务资产”**功能的前提下，若关闭了**“物理表同步业务资产”**按钮，鼠标移至**“同步状态”**中的同步业务资产图标之上，会显示“未同步”。

  说明

  企业模式下，进行同步时，可以选择同步表到生产环境或开发环境。默认同步到生产环境，不勾选则无法同步。

----结束

通过逆向数据库导入物理表

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物理表目录中，使其变成物理表。

步骤1 在数据架构控制台，单击左侧导航树中的“模型设计 > 关系建模”，进入关系建模页面，在页面中间栏位的最上方的下拉列表中选择一个物理模型，或者从“数仓规划”中选择一个物理模型，单击物理模型进入。

步骤2 在物理表的列表上方，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 8-30 逆向数据库配置

参数名称	说明
*所属主题	在下拉列表中选择所属主题。
*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物理表目录中，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**配置DataArts Studio数据连接参数**。
*数据库	选择数据库。
*Schema	下拉选择Schema。该参数仅DWS和POSTGRESQL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。

参数名称	说明
更新已有表	如果从其他数据源逆向过来的表，在物理表中已存在同名的表，选择是否更新已有的物理表。
名称来源	逆向后表名称/字段名称的来源，可以是描述或者是相应英文名，如表/字段未指定描述则固定使用英文名。 • 来自描述 • 来自英文名称 **说明** 进行逆向数据库配置时，如果逆向后表中文名称/字段中文名称的来源选择“来自描述”，则用中文名在进行描述时，表的字段注释不能重复。
*数据表	选择全部或部分需导入的数据表。

图 8-60 逆向配置

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 8-61 逆向结果

----结束

物理表更多操作

- 同步

  在物理表列表中，选择需要同步的物理表，单击列表上方的“更多 > 同步”，单击“确定”，完成物理表的同步。只有当表处于已发布状态时，才能执行此操作。

  说明

  物理表关联了质量规则进行发布后，在数据质量作业目录上面单击“同步主题为目录”后，数据架构自动生成的质量作业，会按照主题结构同步到数据质量对应的目录下。

- 发布

  在物理表列表中，选择需要发布的物理表，单击列表上方的“发布”或者单击“操作”列的“发布”，选择审核人，再单击“确认提交”提交审核，审核通过后完成发布。

  说明

  企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

  选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

  如果勾选“自助审批”，勾选后审批单将自动处理，此功能为体验功能，不推荐在真实项目中使用。

- 下线

  在物理表列表中，选择需要下线的物理表，单击列表上方的“更多 > 下线”或者单击“操作”列的“更多 > 下线”，进行物理表下线。只有当表处于已发布状态时，才能执行此操作。

- 修改主题

  在物理表列表中，选择需要修改主题的物理表，单击列表上方的“更多 > 修改主题”，可以修改物理表的主题。

- 删除

  在物理表列表中，选择需要删除的物理表，单击列表上方的“更多 > 删除”，可以删除物理表。只有当表处于草稿/已驳回/已下线状态时，才能执行此操作。

- 标签

  在物理表列表中，选择需要设置标签的物理表，单击列表上方的“标签”，进入后添加标签，单击“确定”，完成物理表的标签设置。

  说明

  输入文字并回车可临时添加标签，整页信息提交后才可新建标签。标签最多可添加20个。

  物理表可以通过标签过滤进行模糊查询。

- 导入

  导入EXCEL

  在物理表列表中，单击列表上方的“导入”，选择“导入EXCEL”，进入导入表页面后，选择是否更新已有表，添加并上传要导入的文件，上传成功后，单击“关闭”退出该页面。系统支持可以查看“上次导入”的结果。

图 8-62 导入 EXCEL

- 导出

  在物理表列表中，选择需要导出的表，单击列表上方的“导出”，进入导出模型页面后，选择导出对象（表或者DDL），如果选择DDL，需要选择表“全部”或者“部分”，“包含库名”默认勾选，单击“确定”可导出所选择的物理表。
- 编辑

  在物理表列表中，选择需要编辑的物理表，单击“操作”列的“编辑”，进入编辑页面进行编辑。
- 发布历史

  在物理表列表中，选择需要查看发布历史的物理表，单击“操作”列的“更多 > 发布历史”，进入后可查看物理表的发布历史和版本对比。
- 浏览SQL

  在物理表列表中，选择需要预览SQL的物理表，单击“操作”列的“更多 > 预览SQL”，进入后可预览物理表的SQL信息。

8.6.3 维度建模

8.6.3.1 新建维度

维度建模包含维度、维度表和事实表三个部分。

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）。

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，选择对应工作空间的“数据架构”模块，进入数据架构页面。

2. 在数据架构控制台，单击左侧导航树中的“模型设计 > 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”，或者直接单击“新建”按钮，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**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 8-63 配置参数

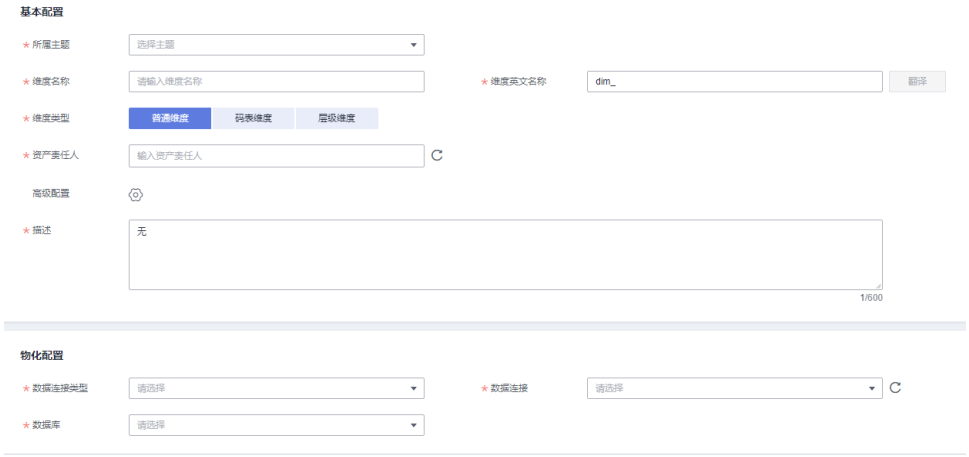


表 8-31 基本配置

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	只允许除\、<、>、%、"、'、;及换行符以外的字符。
*维度英文名称	只能包含英文字母、数字和下划线，且只能以**dim_**开头。 系统支持通过翻译功能按照已配置的命名词典自动生成维度英文名称。
*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。
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。
*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。

参数名称	说明
*描述	描述信息。支持的长度为1~600个字符。

表 8-32 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**配置DataArts Studio数据连接参数**。
*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**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	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模型支持HIVE_TABLE和HIVE_EXTERNAL_TABLE。 MRS_SPARK模型支持HUDI_COW和HUDI_MOR。 POSTGRESQL模型仅支持POSTGRESQL_TABLE。 MRS_CLICKHOUSE模型仅支持CLICKHOUSE_TABLE。 Oracle模型仅支持ORACLE_TABLE。 MySQL模型仅支持MYSQL_TABLE。 DORIS模型仅支持DORIS_TABLE。

参数名称	说明
压缩等级	当数据连接类型为DWS时，可选择压缩等级，以减少数据存储成本。 不同表类型可选以下压缩等级： ● DWS_ROW：“NO”、“YES”。 ● DWS_COLUMN：“NO”、“LOW”、“MIDDLE”、“HIGH”。 ● DWS_VIEW：不支持设置压缩等级。
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。
PreCombineField	该参数仅SPARK连接类型有效。
路径	该参数仅数据源为MRS_HIVE且表类型选择HIVE_EXTERNAL_TABLE时有效。 只支持英文字母、数字、左斜杠(/)、英文句号(.)、中划线(-)、下划线(_)、冒号(:)。

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 8-64 属性配置

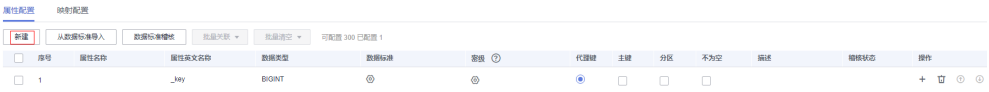


表 8-33 属性配置

参数名称	说明
属性名称	只允许除\、<、>、%、"、'、;及换行符以外的字符。
属性英文名称	只能包含英文字母、数字和下划线，且英文字母开头。
数据类型	根据原始数据定义数据类型。

参数名称	说明
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，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**新建数据标准**进行创建。
密级	单击  按钮可以为逻辑实体属性添加密级。 如果没有您想要的密级，可点击**跳转**到数据安全界面中创建需要的密级。 如不使用该功能，可在配置中心 > 模型设计中关闭该功能。
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。 **说明** 数据连接为MRS Spark连接（通过MRS Spark连接支持MRS Hudi数据源）时，由于Hudi的限制，必须存在字段主键才能数据落库成功，否则会导致表同步失败。
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。
稽核状态	表示是否进行数据标准稽核。 单击“数据标准稽核”，进行数据标准稽核。
操作	相关操作按钮。

在“映射配置”页签，单击“新建映射”，创建维度的映射（映射是指维度与物理模型源表的映射）。需配置如下参数：

图 8-65 映射配置

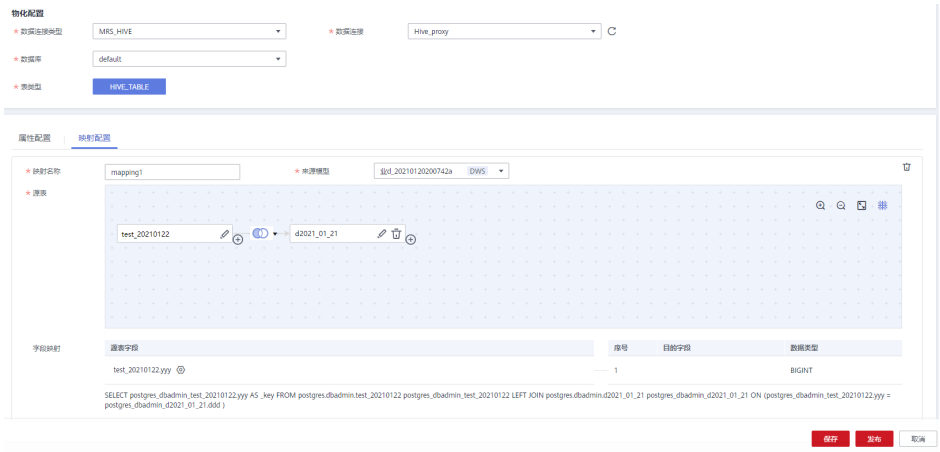


表 8-34 映射参数

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，请参见关系建模进行创建。
*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 8-66** 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. 配置完成后，单击“发布”。

   说明

   企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

6. 在弹出对话框中，选择审核人，单击“确认提交”，提交维度的发布审核。

   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，状态显示为“已发布”。

   选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

7. 可以参照步骤3~步骤6，完成其他维度的创建。
8. 完成所有维度的新建之后，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 8-67 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。

说明

企业模式下，进行同步时，可以选择同步到生产环境或开发环境。默认同步到生产环境，不勾选则无法同步。

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。

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。

说明

企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

----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”。

步骤3 在弹出对话框中，选择审核人，单击“确认提交”，审批完成后，完成维度的发布。

说明

企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

----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”。

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”，审批完成后，完成维度的发布。

说明

企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

注意，此处“数据质量作业调度时间”指的是维度发布后，自动创建质量作业的调度时间。

图 8-68 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”。

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

----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“更多 > 下线”。

步骤3 在弹出对话框中，选择审核人，单击“确认提交”，审批完成后，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。

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

通过逆向数据库导入维度

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维度目录中，使其变成维度。

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 在维度列表上方，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 8-35 逆向数据库配置

参数名称	说明
*所属主题	在下拉列表中选择所属主题。
*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维度目录中，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**配置DataArts Studio数据连接参数**。
*数据库	选择数据库。
*Schema	下拉选择Schema。该参数仅DWS和POSTGRESQL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，在维度中已存在同名的表，选择是否更新已有的维度。
名称来源	逆向后表名称/字段名称的来源，可以是描述或者是相应英文名，如表/字段未指定描述则固定使用英文名。 • 来自描述 • 来自英文名称 **说明** 进行逆向数据库配置时，如果逆向后表中文名称/字段中文名称的来源选择“来自描述”，则用中文名在进行描述时，表的字段注释不能重复。
*数据表	选择全部或部分需导入的数据表。

图 8-69 逆向配置

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 8-70 逆向结果

----结束

8.6.3.2 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。
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。
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**。
- 维度表关联了质量规则进行发布后，在数据质量作业目录上面单击“同步主题为目录”后，数据架构自动生成的质量作业，会按照主题结构同步到数据质量对应的目录下。

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。

   说明

   企业模式下，进行同步时，可以选择同步到生产环境或开发环境。默认同步到生产环境，不勾选则无法同步。

   图 8-71 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。您可以切换生产环境和开发环境查看同步结果。

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 8-72 关联维度表质量规则

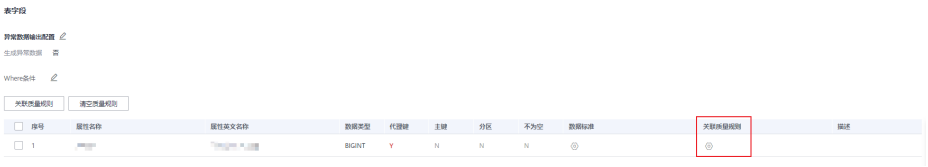
4. 在弹出的页面中配置关联质量规则参数。配置完成单击确定。
   - **更新已有规则**：若勾选此项，新添加的规则会覆盖旧规则。
   - **匹配字段**：此参数默认应用于所有字段，依据用户输入的正则表达式对字段进行过滤。
   - **Where条件**：可依据用户输入的where条件对字段进行过滤。
   - **生成异常数据**：开启此项，表示异常数据将按照配置的参数存储到规定的库中。
   - **数据库或Schema**：开启“生成异常数据”时显示此项，表示存储异常数据的数据库或Schema
   - **表前缀**：开启“生成异常数据”时显示此项，表示存储异常数据的表的前缀。
   - **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 8-73 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 8-74 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 8-75 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 8-76 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 8-77 异常数据输出开关

单击开关，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 8-78 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。

配置完成后单击 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 8-79 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发布审核中、已发布或下线审核中状态，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 8-80 删除维度表

4. 如果确认要删除，单击“是”。

查看维度表详情

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. 单击维度表名称，进入维度表详情页面。
4. 可以查看维度表基本信息和表字段信息。同时，您可以配置异常数据输出信息。
   a. 单击“编辑”按钮，并打开“生成异常数据”的开关。开启此项，表示异常数据将按照配置的参数存储到规定的库中。
   b. 输入数据库或Schema信息，表示存储异常数据的数据库或Schema。
   c. 设置异常表的表前缀和表后缀，表示存储异常数据的表前缀和后缀。

      说明

      异常表的前后和后缀只能包含英文字母、数字和下划线。

   d. 配置好以后，单击保存异常数据配置信息。
5. 系统支持配置where表达式，可依据用户输入的where条件对字段进行过滤。

8.6.3.3 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，选择对应工作空间的“数据架构”模块，进入数据架构页面。
2. 在数据架构控制台，单击左侧导航树中的“模型设计 > 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，或者直接单击“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 8-81 事实表基本配置

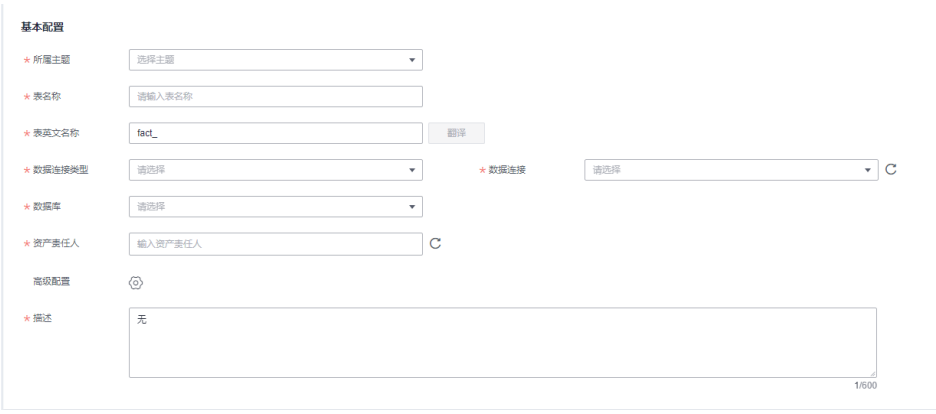


表 8-36 基本配置参数说明

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	只允许除\、<、>、%、"、'、;及换行符以外的字符。
*表英文名称	只能以**fact**_开头，支持英文字母、数字、下划线。 系统支持通过翻译功能按照已配置的命名词典自动生成表英文名称。
*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	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模型支持HIVE_TABLE和HIVE_EXTERNAL_TABLE。 MRS_SPARK模型支持HUDI_COW和HUDI_MOR。 POSTGRESQL模型仅支持POSTGRESQL_TABLE。 MRS_CLICKHOUSE模型仅支持CLICKHOUSE_TABLE。 Oracle模型仅支持ORACLE_TABLE。 MySQL模型仅支持MYSQL_TABLE。 DORIS模型仅支持DORIS_TABLE。

参数名称	说明
压缩等级	当数据连接类型为DWS时，可选择压缩等级，以减少数据存储成本。 不同表类型可选以下压缩等级： ● DWS_ROW：“NO”、“YES”。 ● DWS_COLUMN：“NO”、“LOW”、“MIDDLE”、“HIGH”。 ● DWS_VIEW：不支持设置压缩等级。
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。
PreCombineField	该参数仅SPARK连接类型有效。
路径	该参数仅数据源为MRS_HIVE且表类型选择HIVE_EXTERNAL_TABLE时有效。 只支持英文字母、数字、左斜杠(/)、英文句号(.)、中划线(-)、下划线(_)、冒号(:)。
*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。
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。
*描述	描述信息。支持的长度1~600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，会弹出“选择维度”页面。选择一个维度（选择公共层空间数据或者选择本空间数据），选择维度建模的模型，可以

勾选一个或多个已创建的维度表，单击“确定”后，会将所选维度的维度表及维度表的属性值字段添加到列表中。

- 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8-37**。字段配置完成后，单击字段后的↑或↓可以调整字段的顺序。

图 8-82 配置维度或度量字段

表 8-37 字段配置参数

参数名称	说明
类型	包含度量和维度两种类型。
字段名称	只允许除\、<、>、%、"、'、;及换行符以外的字符。 维度属性的字段会自动显示所添加的维度表及维度表的属性值字段，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。
主键	选中时表示该字段为主键。 **说明** 数据连接为MRS Spark连接（通过MRS Spark连接支持MRS Hudi数据源）时，由于Hudi的限制，必须存在字段主键才能数据落库成功，否则会导致表同步失败。
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。 或者单击“从数据标准导入”，可以选择一个数据标准与字段相关联。 如果您还未创建数据标准，请参见**新建数据标准**进行创建。

参数名称	说明
密级	单击  按钮可以为逻辑实体属性添加密级。 如果没有您想要的密级，可点击**跳转**到数据安全界面中创建需要的密级。 如不使用该功能，可在配置中心 > 模型设计中关闭该功能。
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及字段名称。单击  可以更换关联的维度。 若已开启公共层空间，支持选择公共层空间维度进行关联。
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。 当同一个维度的相同字段被添加多次时，需要设置不同的角色来加以区分。
描述	描述信息。
稽核状态	表示是否进行数据标准稽核。 单击“数据标准稽核”，进行数据标准稽核。
操作	相关操作按钮。

c. 在“映射配置”页签，单击“新建映射”，配置映射参数。

图 8-83 配置映射

表 8-38 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**关系建模**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 8-84** 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”，并在弹出框中，选择审核人，单击“确认提交”，提交事实表的发布审核。

   说明

   企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

   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，状态显示为“已发布”。

   选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

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。您可以切换生产环境和开发环境查看同步结果。

   - 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。

  说明

  企业模式下，进行同步时，可以选择同步到生产环境或开发环境。默认同步到生产环境，不勾选则无法同步。

  事实表关联了质量规则进行发布后，在数据质量作业目录上面单击“同步主题为目录”后，数据架构自动生成的质量作业，会按照主题结构同步到数据质量对应的目录下。

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 8-85 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。

  说明

  企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

- **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。

  说明

  企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

  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。

- **创建指标**

a. 在事实表列表中，找到所需要的事实表，在右侧单击“更多 > 创建指标”，进入新建衍生指标页面。

b. 新建衍生指标请参考**新建衍生指标并发布**。

- **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。当事实表处于发布审核中、已发布或下线审核中状态时，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”。

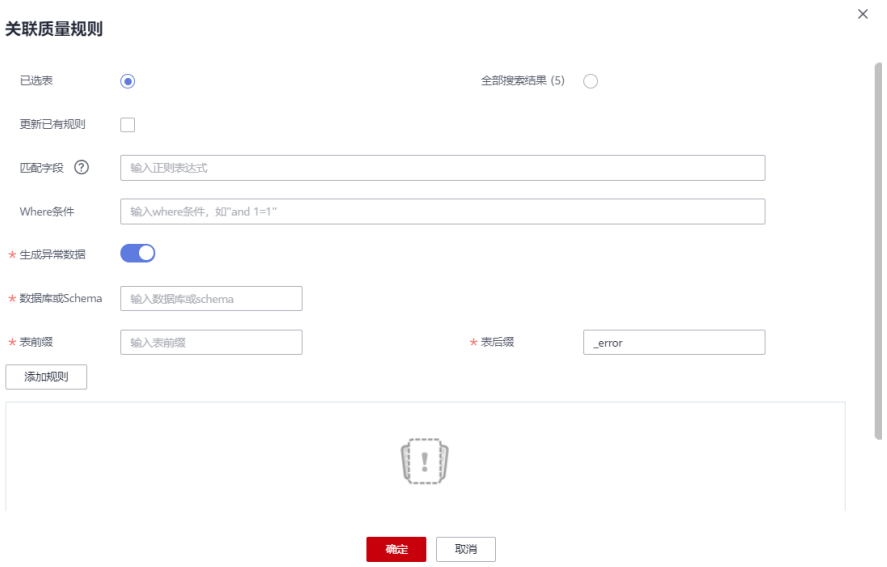
事实表关联质量规则

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 8-86 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，表示异常数据将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。

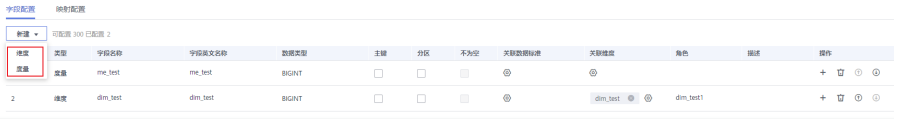
图 8-87 事实表关联质量规则



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 8-88 新建字段



5. 配置完成后，单击“确定”，完成事实表新建字段。

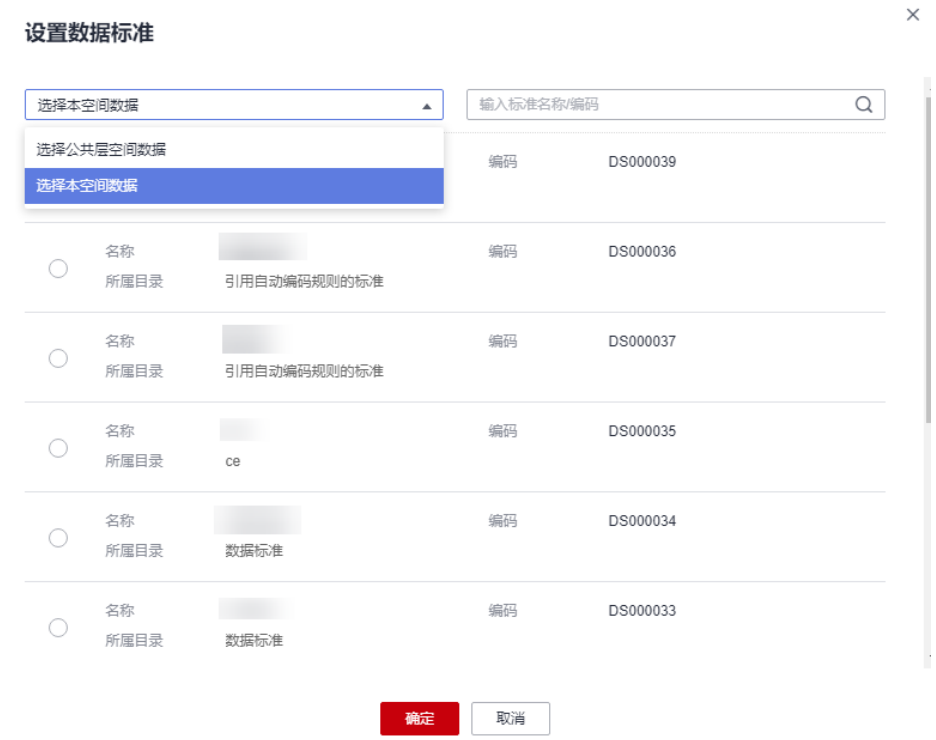
事实表字段关联数据标准

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 8-89 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。如果已开启公共层空间，在普通空间选择数据标准时，需要手动选择数据标准来源为“选择公共层”或“选择本空间”。“选择公共层”开启后，可以将公共层空间的数据标准引用到普通空间。

图 8-90 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 8-91 事实表单个字段关联质量规则

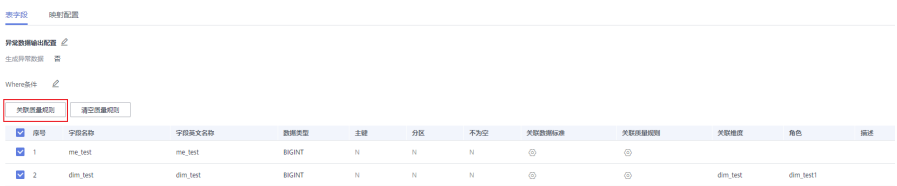
5. 配置完成后，单击“确定”，完成事实表字段关联质量规则。

图 8-92 添加事实表质量规则

事实表字段批量关联质量规则

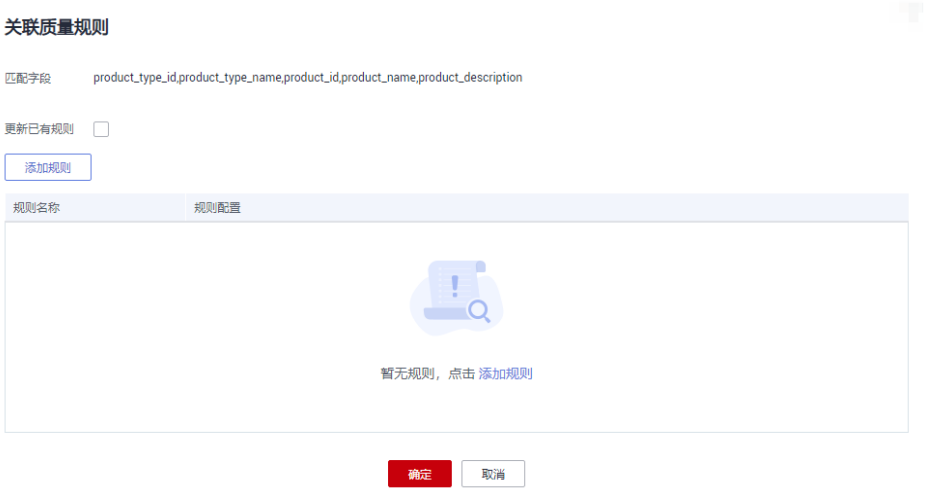
1. 在数据架构控制台，选择“模型设计 > 维度建模”，进入维度建模页面。
2. 单击“事实表”页签，进入事实表页面。
3. 在事实表列表中，单击需要关联质量规则的事实表名称。
4. 在事实表的详情页的表字段列表中，勾选需要关联质量规则的表字段，单击关联质量规则。

图 8-93 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 8-94 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。

通过逆向数据库导入事实表

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事实表目录中，使其变成事实表。

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 在事实表的列表上方，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 8-39 逆向数据库配置

参数名称	说明
*所属主题	在下拉列表中选择所属主题。
*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事实表目录中，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**配置DataArts Studio数据连接参数**。
*数据库	选择数据库。
*Schema	下拉选择Schema。该参数仅DWS和POSTGRESQL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，在事实表中已存在同名的表，选择是否更新已有的事实表。

参数名称	说明
名称来源	逆向后表名称/字段名称的来源，可以是描述或者是相应英文名，如表/字段未指定描述则固定使用英文名。 • 来自描述 • 来自英文名称 **说明** 进行逆向数据库配置时，如果逆向后表中文名称/字段中文名称的来源选择“来自描述”，则用中文名在进行描述时，表的字段注释不能重复。
*数据表	选择全部或部分需导入的数据表。

图 8-95 逆向配置

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 8-96 逆向结果

----结束

查看事实表详情

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. 单击事实表名称，进入事实表详情页面。
4. 可以查看事实表基本信息和表字段信息。同时，您可以配置异常数据输出信息。
   a. 单击“编辑”按钮，并打开“生成异常数据”的开关。开启此项，表示异常数据将按照配置的参数存储到规定的库中。
   b. 输入数据库或Schema信息，表示存储异常数据的数据库或Schema。
   c. 设置异常表的表前缀和表后缀，表示存储异常数据的表前缀和后缀。

      说明

      异常表的前后和后缀只能包含英文字母、数字和下划线。

   d. 配置好以后，单击保存异常数据配置信息。
5. 系统支持配置where表达式，可依据用户输入的where条件对字段进行过滤。

8.6.4 数据集市

数据集市，也称为DM模型。是汇总表的统称。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，选择对应工作空间的“数据架构”模块，进入数据架构页面。
2. 在数据架构控制台，单击左侧导航树中的“模型设计 > 数据集市”。
3. 在左侧主题目录中选中一个主题，然后单击“新建”按钮，或者直接单击“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 8-97 汇总表基本配置

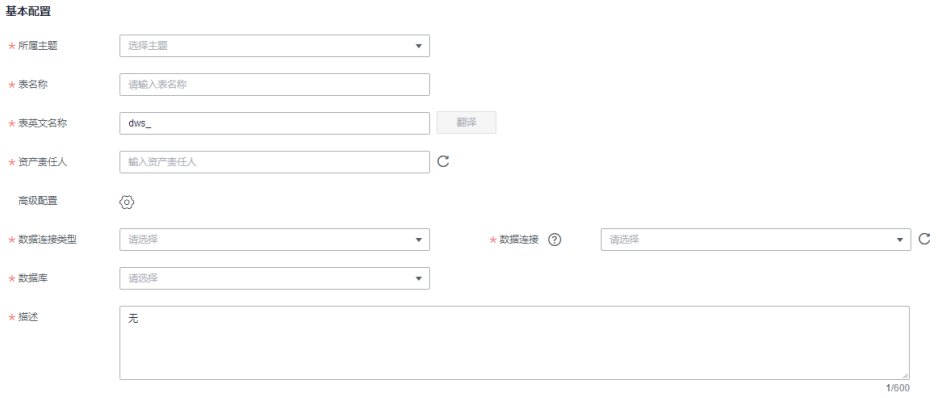


表 8-40 基本配置参数说明

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。只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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**dws_**开头。 系统支持通过翻译功能按照已配置的命名词典自动生成表英文名称。
*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。
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。
*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	数据集市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。

参数说明	说明
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模型支持HIVE_TABLE和HIVE_EXTERNAL_TABLE。 MRS_SPARK模型支持HUDI_COW和HUDI_MOR。 POSTGRESQL模型仅支持POSTGRESQL_TABLE。 MRS_CLICKHOUSE模型仅支持CLICKHOUSE_TABLE。 Oracle模型仅支持ORACLE_TABLE。 MySQL模型仅支持MYSQL_TABLE。 DORIS模型仅支持DORIS_TABLE。
压缩等级	当数据连接类型为DWS时，可选择压缩等级，以减少数据存储成本。 不同表类型可选以下压缩等级： ● DWS_ROW：“NO”、“YES”。 ● DWS_COLUMN：“NO”、“LOW”、“MIDDLE”、“HIGH”。 ● DWS_VIEW：不支持设置压缩等级。
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。

参数说明	说明
*描述	描述信息。支持的长度为1~600个字符。

b. 选择“属性配置”页签，配置汇总表的属性信息。

单击“添加”，可以添加一个或多个相关联的属性信息，例如衍生指标。

单击“导入字段”，可以选择“从指标导入”、“从维度属性导入”或“从数据指标导入”，可以导入所需的字段信息。

说明

从维度属性导入字段时，指标引用的维度属性，必须先关联指标/导入指标字段，再关联维度，才能够从维度属性导入字段。

从指标导入字段时，支持模糊搜索。

单击“数据标准稽查”，可以对汇总表的属性信息进行数据标准稽查，“稽查状态”为 ✔ 。

单击“批量关联”，可以对多个属性配置批量关联数据标准和密级。

单击“批量清空”，可以对多个属性配置批量清空数据标准和密级。

图 8-98 属性配置

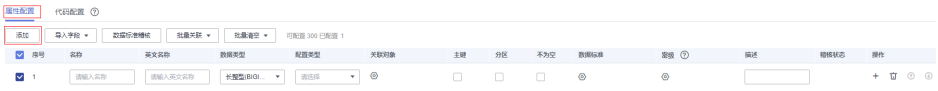


表 8-41 属性配置参数

参数名称	说明
名称	只允许除\、<、>、%、"、'、;及换行符以外的字符。 维度属性的字段会自动显示所添加的维度表及维度表的属性值字段，一般不需要修改。
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	显示该字段名称的数据类型。
配置类型	表示该字段名称对应的配置类型。比如衍生指标。
关联对象	表示该字段名称的配置类型对应的关联对象。比如衍生指标的名称。
主键	选中时表示该字段为主键。 **说明** 数据连接为MRS Spark连接（通过MRS Spark连接支持MRS Hudi数据源）时，由于Hudi的限制，必须存在字段主键才能数据落库成功，否则会导致表同步失败。
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。

参数名称	说明
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**新建数据标准**进行创建。
密级	单击 按钮可以为逻辑实体属性添加密级。 如果没有您想要的密级，可点击**跳转**到数据安全界面中创建需要的密级。 如不使用该功能，可在配置中心 > 模型设计中关闭该功能。
描述	描述信息。
稽核状态	表示是否进行数据标准稽核。 单击“数据标准稽核”，进行数据标准稽核。
操作	相关操作按钮。

c. 选择“代码配置”页签，可以查看系统生成的代码以及对指标代码进行格式化。

单击“生成代码”，可以对已经生成的代码进行刷新。单击“复制到指标代码”可以复制代码到下面的指标代码，单击“格式化”，可以对指标代码进行格式化。

5. 单击“发布”，并在弹出框中，选择审核人，单击“确认提交”，提交汇总表的发布审核。

说明

企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

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，状态显示为“已发布”。

选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

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“模型设计 > 数据集市 > 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。您可以切换生产环境和开发环境查看同步结果。
   - 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。

说明

企业模式下，进行同步时，可以选择同步到生产环境或开发环境。默认同步到生产环境，不勾选则无法同步。

汇总表关联了质量规则进行发布后，在数据质量作业目录上面单击“同步主题为目录”后，数据架构自动生成的质量作业，会按照主题结构同步到数据质量对应的目录下。

管理汇总表

1. 在数据架构控制台，单击左侧导航树中的“模型设计 > 数据集市”，选择“汇总表”页签，进入汇总表页面。

图 8-99 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...
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”。

说明

企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

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。

      说明

      企业模式下，进行发布时，可以选择发布到生产环境或开发环境。默认发布到生产环境，不勾选则无法发布。

   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. 在“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“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

      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，勾选所需要关联质量规则的汇总表，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 8-100 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

如果系统中已有的表英文名称和模板中的表英文名称相同，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

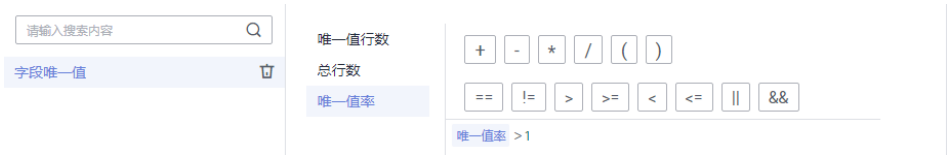
- 在左侧主题树中选中某个主题，可以导出该主题下的所有汇总表；
- 当该空间下不超过500条汇总表数据时可以全部导出。

汇总表关联质量规则

1. 在数据架构控制台，选择“模型设计 > 数据集市”，进入数据集市页面。
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”。

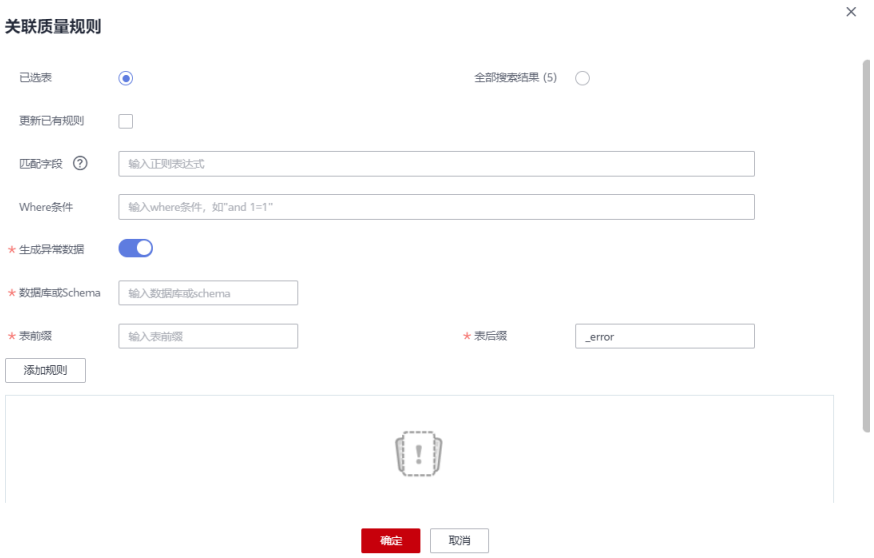
图 8-101 关联汇总表质量规则

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，表示异常数据将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 8-102 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 > 数据集市”，进入数据集市页面。
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 8-103 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 8-104 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 > 数据集市”，进入数据集市页面。
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 8-105 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 8-106 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 > 数据集市”，进入数据集市页面。
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 8-107 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 8-108 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。

查看汇总表详情

1. 在数据架构控制台，单击左侧导航树中的“模型设计 > 数据集市”进入数据集市页面。
2. 单击“汇总表”页签，进入汇总表页面。
3. 单击汇总表名称，进入汇总表详情页面。
4. 可以查看汇总表基本信息和表字段信息。同时，您可以配置异常数据输出信息。
   a. 单击“编辑”按钮，并打开“生成异常数据”的开关。开启此项，表示异常数据将按照配置的参数存储到规定的库中。

b. 输入数据库或Schema信息，表示存储异常数据的数据库或Schema。
c. 设置异常表的表前缀和表后缀，表示存储异常数据的表前缀和后缀。

说明

异常表的前后和后缀只能包含英文字母、数字和下划线。

d. 配置好以后，单击保存异常数据配置信息。

5. 系统支持配置where表达式，可依据用户输入的where条件对字段进行过滤。

8.7 指标设计

8.7.1 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，用于定义指标的设置目的、计算公式等，并不进行实际运算，可与技术指标进行关联。而技术指标是对业务指标的具体实现，定义了指标如何计算。

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**流程设计**。

新建业务指标并发布

1. 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。
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 8-109 新建业务指标

表 8-42 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。只允许除\、<、>、%、"、'、;及换行符以外的字符。
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**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。
自定义指标	如果在配置中心的指标配置页面设置了自定义指标，页面中会显示自定义指标参数。创建流程请参见**指标配置**。

b. 配置指标数据信息。

图 8-110 指标数据信息

表 8-43 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**新建维度**。
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、复合指标和原子指标。
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 8-111 管理信息

表 8-44 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. 在弹出框中，选择审核人，单击“确认提交”，提交审核。

   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，状态显示为“已发布”。

   选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

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。

   业务指标创建完成后，单击指标名称，可以查看该业务指标的详情、关系图、发布历史和审核历史。

   通过关系图，可以查看该业务指标的血缘图。

   通过发布历史，可以查看该业务指标的发布历史和不同发布版本之间的差异对比。

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 8-112 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 8-113 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 8-114 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 8-115 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。

导入/导出业务指标

导入指标：您可以通过导入功能，批量导入业务指标。

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。
2. 单击业务指标列表上方“更多”中的“导入”。在“导入业务指标”对话框中，单击“下载关系建模导入模板”。

图 8-116 导入业务指标

表 8-45 导入配置参数说明

参数名	说明
更新已有数据	如果所要导入的表已存在，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“下载业务指标导入模板”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。

3. 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“业务指标”Sheet页中，所需填写的参数，说明如下：

   表 8-46 业务指标 Sheet 页参数说明

参数名	参数说明
*流程架构	指标对应的一级流程。
*指标名称	指标的标准名称，需要保持唯一性。
指标编码	由系统自动生成。
指标别名	指标在具体应用场景（报表/报告）中习惯或者简化使用的名字。
*设置目的	简要描述通过此指标希望达到的管理目的。
*指标定义	准确描述指标含义，相关人员能够理解指标所度量的内容。
*计算公式	给出指标清晰的计算规则，可以根据公式计算得出指标数据。
数据来源	需要明确来自于哪个系统，如果可能，请标示出具体的数据表名、字段。
计量单位	指标数据统计的基本计量单位。
*统计周期	指标统计的周期颗粒度。
统计维度	常用的统计维度，维度一般存在层级关系。
*刷新频率	指标数据的刷新的最小频率。

参数名	参数说明
统计口径&修饰词	除统计周期和维度外，该指标常用的统计口径&修饰词，限制指标数据的范围。
指标应用场景	描述该指标重要的应用场景，包括在线报表、例行报告、汇报材料等。
备注	在指标描述之外还需要补充的信息，有助于正确理解和使用该指标。
度量对象	衡量该指标的度量字段，如果不涉及可以不填写。
*指标管理部门	指标管理的Owner，负责指标定义、维护和解释，并提供指标数据。
*指标责任人	填写指标解释人（华为账号名称）。
关联技术指标	当前业务指标在规范设计中的实现。

4. 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 8-117 上次导入

导出指标：您可以通过导出功能，导出已生成的业务指标。

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。
2. 在业务指标页面，找到需要导出的业务指标，然后单击“更多 -> 导出”。

图 8-118 导出业务指标

说明

如果在配置中心的指标配置页面新建了自定义指标，在导出的表格中会呈现该指标。

8.7.2 技术指标

8.7.2.1 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。

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以及与该度量相关的属性，旨在用于支撑指标的敏捷自助消费。敏捷自助消费指的是业务用户能够自主地、快速地访问和使用指标，而不依赖于IT部门或数据团队进行复杂的查询和计算。原子指标提供了非常基础且易于理解的度量，可以支持用户在需要时灵活地创建自己的报表、查询或分析。通过提供原子指标，用户可以在现有的基础数据上自由地组合和计算，更加敏捷地满足自己的需求。

背景信息

原子指标来源于事实表和维度表：

- 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，因此可以基于事实逻辑表明细数据表来创建，也可以基于维度表来创建。
- 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。

约束与限制

单工作空间允许创建的原子指标个数最多5000个。

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**新建事实表**。

新建原子指标并发布

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据架构”模块，进入数据架构页面。
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8-47**配置参数，然后单击“发布”。

图 8-119 新建原子指标

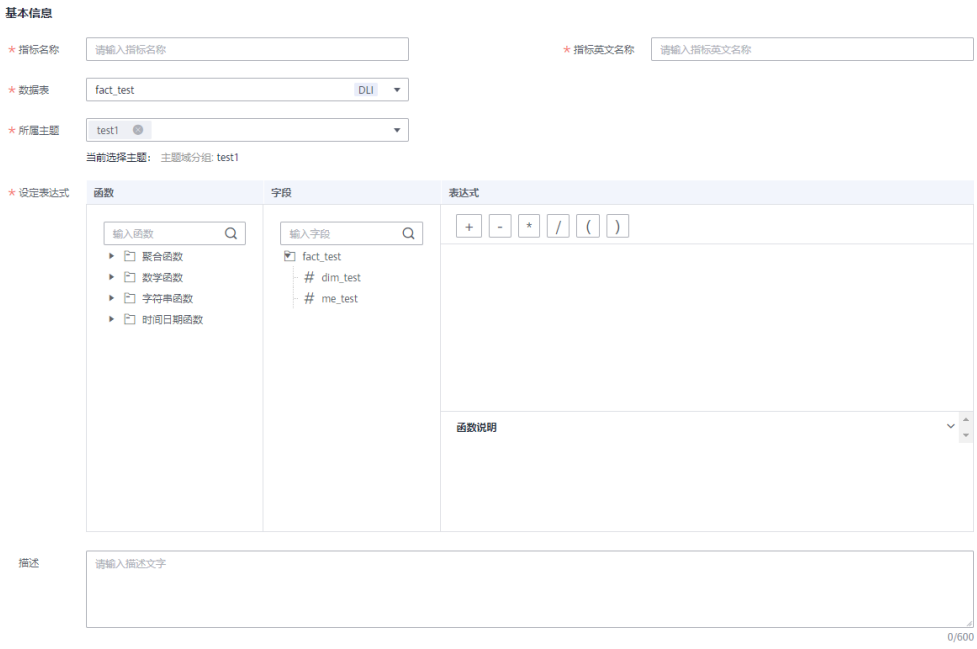


表 8-47 新建原子指标参数说明

参数名称	说明
*指标名称	只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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。

参数名称	说明
*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。函数列表及函数说明请参考**函数说明**。
描述	描述信息。支持的长度为0~600个字符。

5. 在弹出框中，选择审核人，单击“确认提交”，提交审核。

   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，状态显示为“已发布”。

   选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

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。

   原子指标创建完成后，单击指标名称，可以查看该原子指标的详情、关系图、发布历史和审核历史。

   通过关系图，可以查看该原子指标的血缘图。

   通过发布历史，可以查看该原子指标的发布历史和不同发布版本之间的差异对比。

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 8-120 管理原子指标

2. 您可以根据实际需要选择如下操作。

   表 8-48 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 > 删除”，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 8-121 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5000条原子指标数据时可以全部导出。

函数说明

新建原子指标时，需要按照函数设定表达式。以聚合函数的部分函数为例，函数说明如表8-49所示：

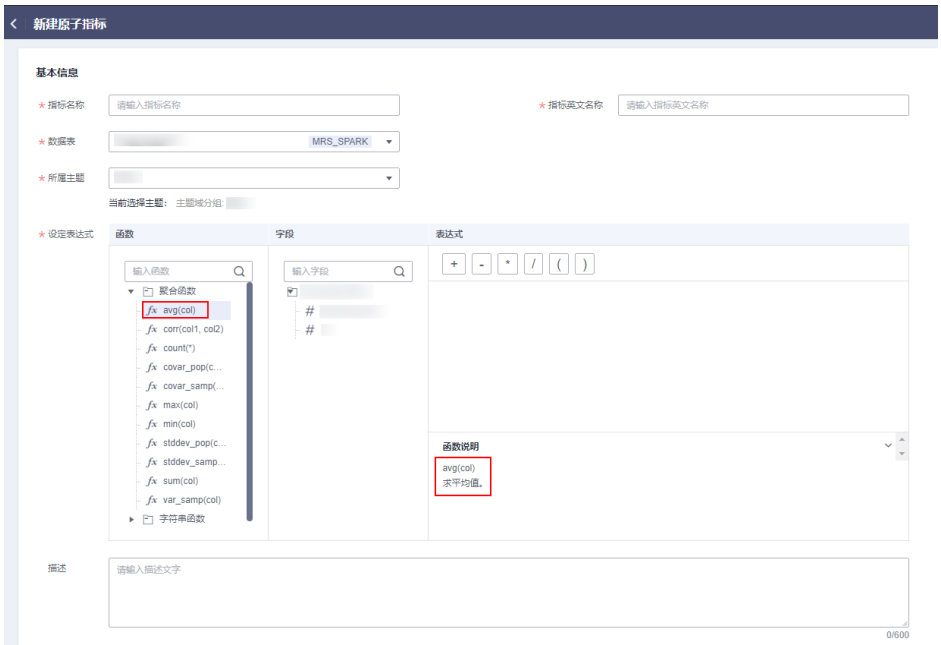
表 8-49 聚合函数说明

函数名	表达式	函数说明
avg(col)	avg()	求平均值。
corr(col1, col2)	corr()	返回两列数值的相关系数。

函数名	表达式	函数说明
count(*)	count()	返回记录条数。
covar_pop(col1, col2)	covar_pop()	返回两列数值协方差。
covar_samp(col1, col2)	covar_samp()	返回两列数值样本协方差。
max(col)	max()	返回最大值。
min(col)	min()	返回最小值。
stddev_pop(col)	stddev_pop()	返回指定列的偏差。
stddev_samp(col)	stddev_samp()	返回指定列的样本偏差。
sum(col)	sum()	求和。
var_samp(col)	var_samp()	返回指定列的样本方差。

如果想要查询更多函数的功能及说明，可以在新建原子指标页面的基本信息中的设定表达式项，单击对应函数，在页面右侧的函数说明框中会显示对应的函数说明。

图 8-122 函数说明



8.7.2.2 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。

约束与限制

单工作空间允许创建的衍生指标个数最多5000个。

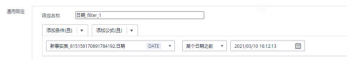
新建衍生指标并发布

1. 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。
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 8-123 新建衍生指标

表 8-50 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。

参数名称	说明
*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**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。只能包含中文、英文字母、数字和下划线，且只能以中文或英文字母开头。 如**图8-124**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。 当选择的字段是字符串类型（例如string、varchar）时，并且条件选择“属于”或“不属于”时，支持从码表中导入数据。单击“从码表导入”，在码表配置页面，选择“码表”和“码表字段”，单击“确定”。导入的码表值数量不能超过50。 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 8-124** 通用限定 
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。

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 8-51 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，选择审核人，单击“确认提交”，提交审核。

   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，状态显示为“已发布”。

   选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。
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。

    衍生指标创建完成后，单击指标名称，可以查看该衍生指标的详情、关系图、发布历史和审核历史。

    通过关系图，可以查看该衍生指标的血缘图。

    通过发布历史，可以查看该衍生指标的发布历史和不同发布版本之间的差异对比。

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 8-125 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 8-126 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5000条衍生指标数据时可以全部导出。

8.7.2.3 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。

约束与限制

单工作空间允许创建的复合指标个数最多5000个。

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**新建衍生指标**。

新建复合指标

1. 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。
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 8-127 新建复合指标

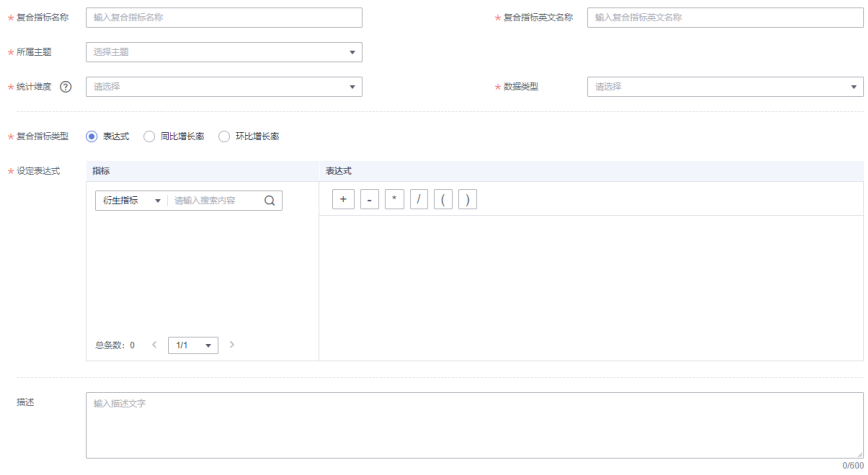


表 8-52 新建复合指标参数说明

参数名称	说明
*复合指标名称	只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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。
*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。
*复合指标类型	当前支持如下几种类型。 ● 表达式 ● 同比增长率 ● 环比增长率
描述	描述信息。支持的长度为0~600个字符。
表达式	
*设定表达式	选择所需要的衍生指标或复合指标，并根据实际需求在“表达式”中设置表达式。
同比增长率	
*同比配置	选择年同比、月同比或者周同比。
*设定衍生指标	选择所需要的衍生指标，此处仅展示有时间限定衍生指标。系统会根据同比配置，利用时间限定自动计算同比增长率。
环比增长率	
*设定衍生指标	选择所需要的衍生指标，此处仅展示有时间限定衍生指标。系统会利用时间限定自动计算环比增长率。

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，选择审核人，单击“确认提交”，提交审核。

   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，状态显示为“已发布”。

   选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。
8. 等待审核人员审核。

   审核通过后，复合指标创建完成。

复合指标创建完成后，单击指标名称，可以查看该复合指标的详情、关系图、发布历史和审核历史。

通过关系图，可以查看该复合指标的血缘图。

通过发布历史，可以查看该复合指标的发布历史和不同发布版本之间的差异对比。

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 8-128 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8-52**。
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。如果复合指标处于发布审核中、已发布或下线审核中状态，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 8-129 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5000条复合指标数据时可以全部导出。

8.7.2.4 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，选择对应工作空间的“数据架构”模块，进入数据架构页面。
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 8-130 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8-53**配置参数，然后单击“发布”。

图 8-131 时间限定

表 8-53 新建时间限定参数说明

参数名称	说明
*限定名称	只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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”。

参数名称	说明
描述	描述信息。支持的长度0~490字符。

6. 在弹出框中，选择审核人，单击“确认提交”，提交发布审核。

   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，状态显示为“已发布”。

   选择审核人时，系统支持选择多个审核人，全部审批通过后，状态才会显示为已发布。如果有任意一个人驳回，则状态为已驳回。

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 8-132 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”。

8.8 通用操作

8.8.1 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**配置元数据采集任务**。

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 在页面的中间栏位，从最上方的下拉框中选择一个物理模型；或者从“数仓规划”选择一个物理模型进入物理表列表页面。单击上方的“逆向数据库”。

图 8-133 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 8-134 配置逆向数据库参数

表 8-54 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。
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**配置DataArts Studio数据连接参数**。
数据库	选择数据库。

参数名称	说明
队列	仅限DLI连接类型，需选择DLI队列。
Schema	下拉选择Schema。该参数仅DWS和POSTGRESQL模型的表有效。
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。
名称来源	逆向后表名称/字段名称的来源，可以是描述或者是相应英文名，如表/字段未指定描述则固定使用英文名。 • 来自描述 • 来自英文名称 **说明** 进行逆向数据库配置时，如果逆向后表中文名称/字段中文名称的来源选择“来自描述”，则用中文名在进行描述时，表的字段注释不能重复。
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。
起始页	当数据表选择“全部”时，需要配置。

步骤4 单击“确定”开始执行逆向数据库操作。

----结束

8.8.2 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**配置元数据采集任务**。

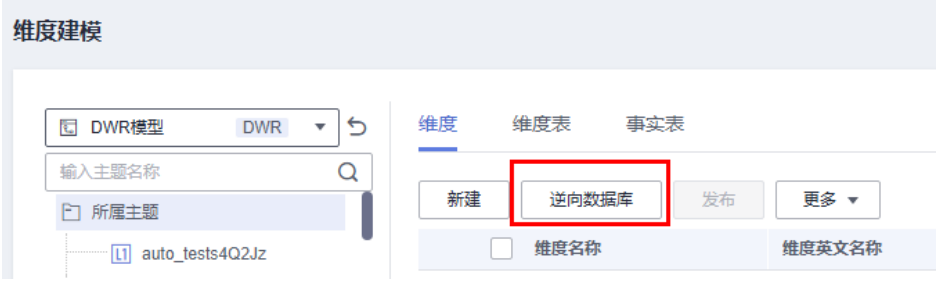
逆向数据库导入表到维度模型中

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

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，从下拉列表选择需要逆向数据库的维度或表，然后单击列表上方的“逆向数据库”。

图 8-135 选中对象



步骤4 在“逆向数据库”对话框中配置参数。

表 8-55 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**配置DataArts Studio数据连接参数**。
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。
名称来源	逆向后表名称/字段名称的来源，可以是描述或者是相应英文名，如表/字段未指定描述则固定使用英文名。 • 来自描述 • 来自英文名称 **说明** 进行逆向数据库配置时，如果逆向后表中文名称/字段中文名称的来源选择“来自描述”，则用中文名在进行描述时，表的字段注释不能重复。
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

8.8.3 导入导出

数据架构支持流程、主题、码表、数据标准、关系建模表（物理表）、逻辑实体、维度建模维度/事实表、业务指标、技术指标、数据集市汇总表的导入导出，暂不支持时间限定、审核中心和配置中心数据的导入导出。

本例中以导入和导出关系建模表为例说明如何进行导入导出，其他数据操作类似。如果您想了解其他数据如何导入导出以及使用场景等，请参考**数据架构数据搬迁**。

约束与限制

- 导入关系建模表、逻辑实体、维度建模维度/事实表、数据集市汇总表前请确保已创建管理中心连接，确保数据连接可用。
- 数据架构中的时间限定、审核中心和配置中心数据不支持导入导出。如有涉及，请您在其他数据迁移前，先进行手动配置同步。
- 数据架构支持最大导入文件大小为4Mb；支持最大导入指标个数为3000个；支持一次最大导出500张表。

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“逻辑模型”进入逻辑模型页面。

步骤2 在逻辑模型中，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 8-136 导入表

表 8-56 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 8-57 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**主题设计**进行新建。
*逻辑实体名称	表名称，只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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**管理资产标签**。
*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。

参数名	参数说明
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、REPLICATION2种方式分布。
*属性名称（CHN）	表中的属性字段的中文名称。只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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。
属性编码	表中的属性字段的编码，系统自动生成。
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**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。
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。如果未创建数据标准，请参见**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**管理资产标签**。
其他配置	填写“高级配置”中自定义项的名称与输入值。

步骤5 在模板的“关系”Sheet页中，所需填写的参数，说明如下：

表 8-58 关系 Sheet 页参数说明

参数名	参数说明
关系名称	关系的名称，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。

参数名	参数说明
*子表	输入关系中子表的英文名称。
*子表字段	输入关系中子表的字段英文名，该字段应为子表的外键，映射为父表的主键。
*子对父	子表对父表的映射关系，可以有以下四种取值： ● 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	父表对子表的映射关系，可以有以下四种取值： ● 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	输入关系中父表的英文名称。
*父字段表	输入关系中父表的字段英文名，该字段应为父表的主键，映射为子表的外键。
角色名称	自定义角色名称，用于标识该关系，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。

步骤6 在“关联质量规则”中填入关联的表名称和属性名称(ENG)。

在模板的“关联质量规则”Sheet页中，所需填写的参数，说明如下：

表 8-59 关联质量规则 Sheet 页参数说明

参数名	参数说明
*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。
*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。
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。

参数名	参数说明
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
表达式	只有当“规则名称”配置为“表达式校验”或者“合法性校验”时，需要配置表达式。

步骤7 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 8-137 上次导入

说明

- 当导入的逻辑实体关联的标准编码不存在或者未发布时，系统会自动弹出报错拦截及详细的编码名称，请修改后再重新上传。
- 当导入的数据不存在时，在“上次导入”页签中的备注中会出现格式为“表名称：属性名称”的报错提示。

----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 在页面的中间栏位，从最上方的下拉列表中找到所需要的物理模型，或者从“数仓规划”选择一个物理模型单击进入，在主题目录中选中一个对象，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 8-138 导入表

表 8-60 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 8-61 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**主题设计**进行新建。
*逻辑实体名称	表名称，只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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**管理资产标签**。
*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。 表类型为DLI_EXTERNAL的表支持的数据格式有：Parquet、Carbon、CSV、ORC、JSON、Avro。

参数名	参数说明（导入DLI/POSTGRESQL/DWS/MRS_HIVE类型的表）
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。
DWS表 DISTRIBUTE BY	仅DWS连接支持，支持HASH(属性名称)、REPLICATION2种方式分布。
HUDI表 PreCombineField	版本字段，仅Hudi表需要填写。
*属性名称（CHN）	表中的属性字段的中文名称。只允许除\、<、>、%、"、'、;及换行符以外的字符。
*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	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**。
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**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**管理资产标签**。

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

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

版本号	可选参数。
其他配置	填写“高级配置”中自定义项的名称与输入值。

步骤5 在模板的“关系”Sheet页中，所需填写的参数，说明如下：

表 8-62 关系 Sheet 页参数说明

参数名	参数说明
关系名称	关系的名称，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。
*子表	输入关系中子表的英文名称。
子表所属数据库	输入关系中子表所属数据库的名称。
*子表字段	输入关系中子表的字段英文名，该字段应为子表的外键，映射为父表的主键。
*子对父	子表对父表的映射关系，可以有以下四种取值： ● 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	父表对子表的映射关系，可以有以下四种取值： ● 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	输入关系中父表的英文名称。

参数名	参数说明
父表所属数据库	输入关系中父表所属数据库的名称。
*父字段表	输入关系中父表的字段英文名，该字段应为父表的主键，映射为子表的外键。
角色名称	自定义角色名称，用于标识该关系，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。

步骤6 在“关联质量规则”中填入关联的表名称和属性名称(ENG)。

在模板的“关联质量规则”Sheet页中，所需填写的参数，说明如下：

表 8-63 关联质量规则 Sheet 页参数说明

参数名	参数说明
*表名称	表英文名称，只能包含英文字母、数字、下划线、$、{、}，且不能以数字开头。
*属性名称（ENG）	表中的属性字段的英文名称。只能包含英文字母、数字和下划线，且以英文字母开头。
规则名称	填写已有的规则名称。在DataArts Studio控制台左上角的模块下拉列表中选择“数据质量”进入DataArts Studio数据质量控制台，然后您可以进入“规则模板”页面查看已有的规则名称。
告警配置	告警条件表达式，由告警参数和逻辑运算符组成。在质量作业运行时，系统将计算出告警条件表达式的结果，并根据表达式的结果是否为真来判断是否触发告警。如果表达式的结果为真则触发质量告警，结果为假则不触发质量告警。 在告警条件表达式中，告警参数以${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
表达式	只有当“规则名称”配置为“表达式校验”或者“合法性校验”时，需要配置表达式。

步骤7 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。

说明

- 当导入的关系表的标准编码不存在或者未发布时，系统会自动弹出报错拦截及详细的编码名称，请修改后再重新上传。
- 当导入的数据不存在时，在“上次导入”页签中的备注中会出现格式为“表名称：属性名称”的报错提示。

----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“逻辑模型”进入逻辑模型页面。

步骤2 在逻辑模型中，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 8-139 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 8-140 导出表

导出DDL时，会将所选表的DDL语句导出成txt文件。

图 8-141 导出 DDL

步骤4 单击“确定”。

----结束

导入/导出维度

- **导入维度**

  可通过导入的方式将维度批量快速的导入到系统中。

  a. 在维度页面，单击“更多 > 导入”，进入“导入配置”页签。

  图 8-142 导入表

  b. 下载维度导入模板，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”。

说明

当导入的维度关联的标准编码不存在或者未发布时，系统会自动弹出报错拦截及详细的编码名称，请修改后再重新上传。

- **导出维度**

  可通过导出的方式将维度导出到本地。

  在维度页面，单击“更多 > 导出”，即可将系统中的维度导出到本地。

导入/导出事实表

- **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 8-143 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”。

  说明

  当导入的事实表关联的标准编码不存在或者未发布时，系统会自动弹出报错拦截及详细的编码名称，请修改后再重新上传。

- **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。

8.8.4 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 8-144 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字段，然后单击“关联质量规则”按钮。

图 8-145 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 8-146 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**新建数据质量规则**新建规则。

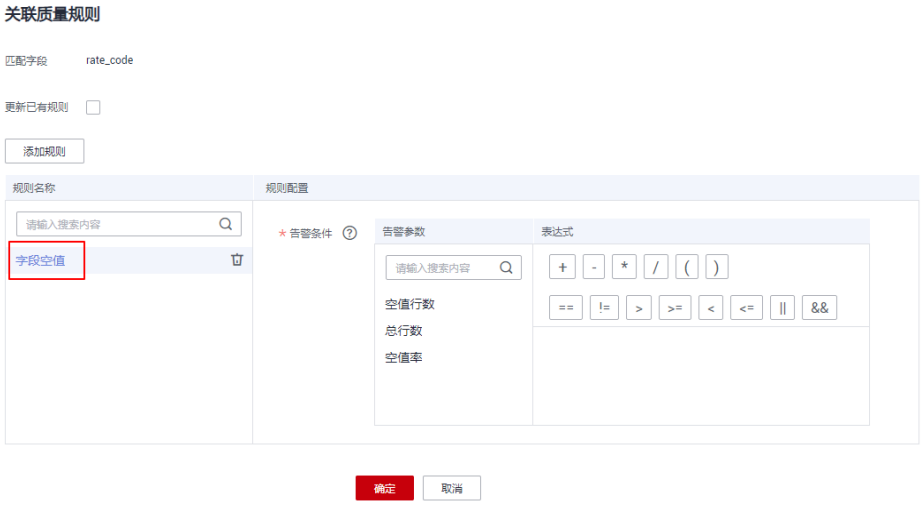
图 8-147 添加规则

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 8-148 设置告警条件

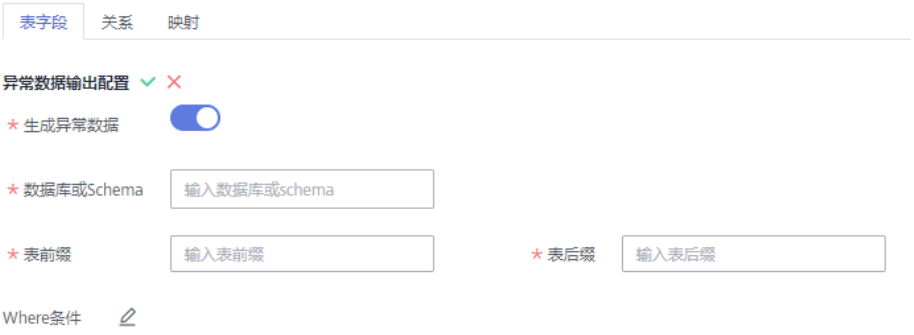


步骤6 （可选）如需要将质量作业中不符合设定规则的异常数据存储在异常表中，可以打开“异常数据输出配置”开关。

图 8-149 异常数据输出开关

单击开关，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 8-150 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。

配置完成后单击 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 8-151 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。

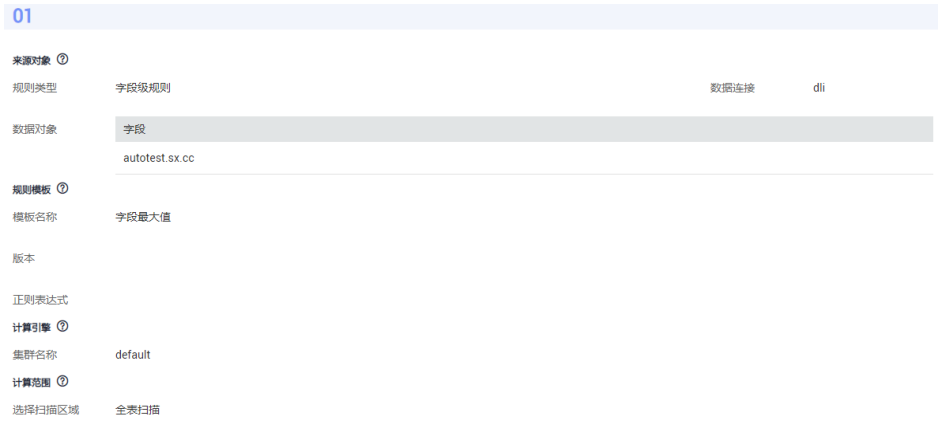
图 8-152 关联结果

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 8-153 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。

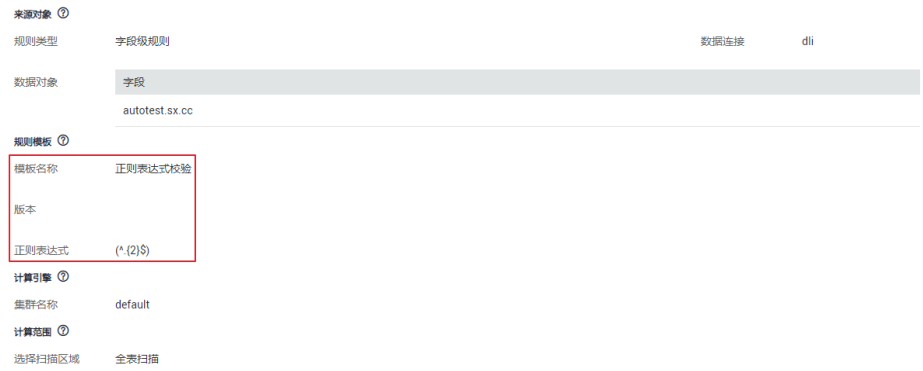
图 8-154 质量规则



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。

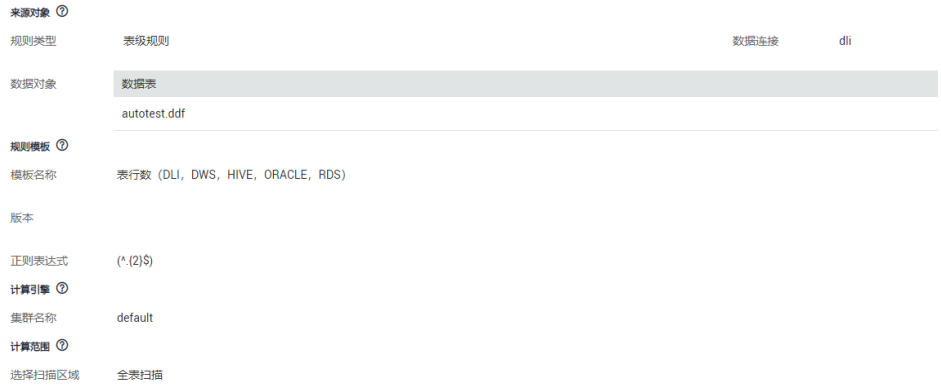
字段关联的数据标准生成的质量规则，示例如下：

图 8-155 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 8-156 码表的质量规则



----结束

8.8.5 查看表

对于关系建模中的表，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

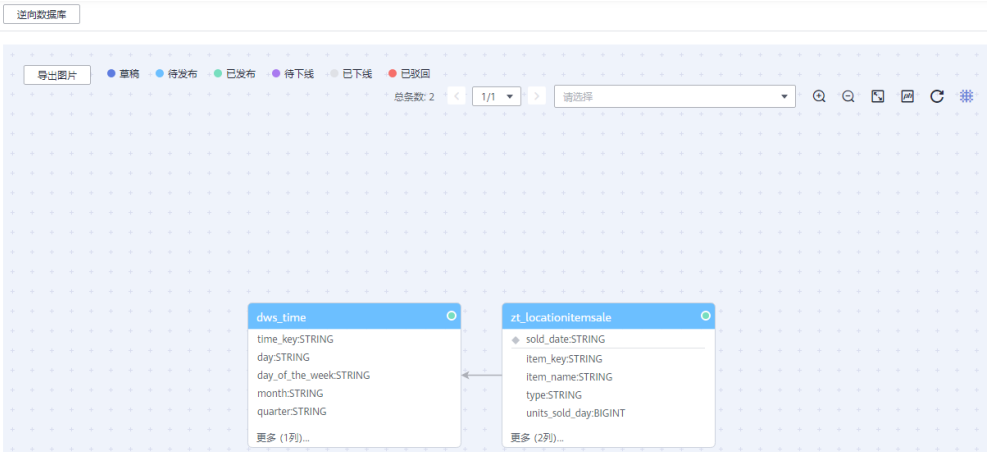
当您在关系模型中完成表的新建后，就可以通过列表视图和模型视图两种形式查看表模型。关系模型页面默认显示为列表视图，您可以切换为模型视图进行查看。

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 在页面的中间栏位，从最上方的下拉列表中找到所需要的物理模型，或者从“数仓规划”选择一个物理模型单击进入，在主题目录中选中一个对象。

步骤3 单击表名称进入后，选择“关系图”页签，查看模型视图。

图 8-157 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。

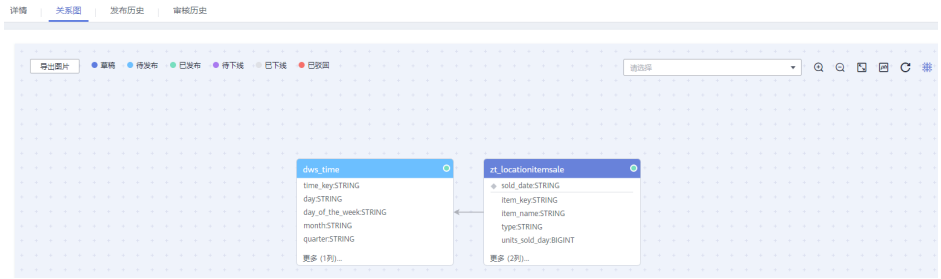
步骤2 在页面的中间栏位，从最上方的下拉列表中找到所需要的物理模型，或者从“数仓规划”选择一个物理模型单击进入，在主题目录中选中一个主题，右侧的列表中将显示该主题下所有的表。

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 8-158 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 8-159 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 在页面的中间栏位，从最上方的下拉列表中找到所需要的物理模型，或者从“数仓规划”选择一个物理模型单击进入，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 8-160 发布历史

----结束

8.8.6 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、逻辑模型、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。配置完成单击“确定”。

图 8-161 批量修改主题

----结束

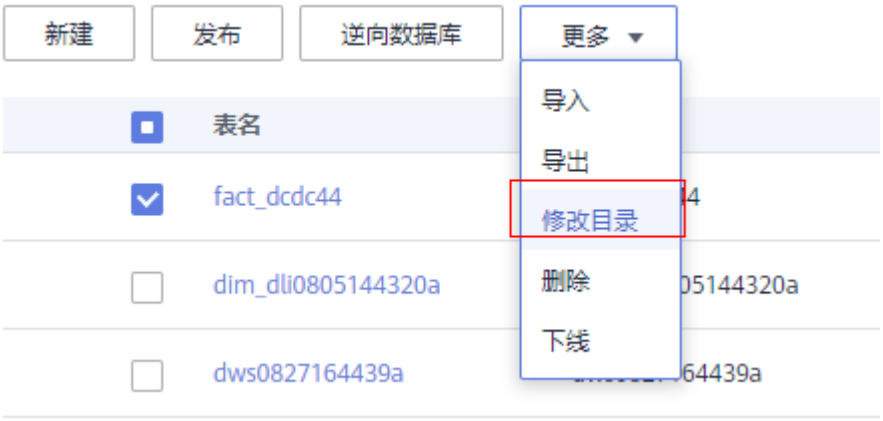
批量修改目录

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 8-162 批量修改目录（此处以码表管理模块为例）



----结束

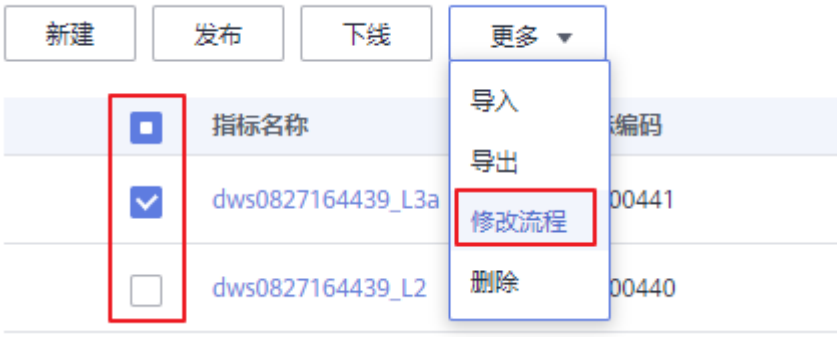
批量修改流程

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 8-163 批量修改流程



----结束

8.8.7 管理配置中心

约束与限制

配置中心中各类对象的自定义项配额如下：

- 主题自定义项10条。
- 表自定义项30条。
- 属性自定义项10条。

- 业务指标自定义项50条。

主题流程配置

主题流程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，选择对应工作空间的“数据架构”模块，进入数据架构页面。
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“主题流程配置”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. 在主题自定义项区域，可对属性进行增加、删除和编辑操作。
   - 在“主题自定义项”右侧，单击“新建”可新增一条自定义属性。主题自定义项属性的可选值支持一次性可输入多个值，可选值不可重复。
   - 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（中文）、属性名称（英文）、可选值，是否必填，描述，实现“编辑”操作。
5. 在流程层级数区域，可设置流程设计的层数，层级最小3级，最大7级。

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，选择对应工作空间的“数据架构”模块，进入数据架构页面。
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”、“可选值”。
   - 保存模板后，在新建数据标准时需要设置此处模板中所选中选项的参数值。
   - 首次进入数据架构的数据标准页面，可选项默认选取“数据长度”和“描述”，其他选项请按需求勾选。
   - 添加自定义项时，支持同时添加中文与英文的自定义项。

图 8-164 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，选择对应工作空间的“数据架构”模块，进入数据架构页面。
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 8-165 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。
- **物理表同步业务资产：**在开启了**“同步业务资产”**且没有创建逻辑实体的前提下，为了避免物理表发布会覆盖同名逻辑表的情况发生，可主动关闭该选项，物理表发布后不会同步业务资产，只会进行资产关联。数据资产关联前会进行业务资产查找，如没有查找到相应的业务资产则会报错并结束资产关联。

- **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。
- **导入数据标准时自动创建目录**：默认开启，打开后导入数据标准时可以自动创建目录。
- **是否启用公共层**：开关打开后，可将当前空间转化为公共层空间。公共层空间的码表和数据标准会共享给所有普通空间；普通空间可以查询、引用公共层空间的码表和数据标准，但无法进行新增、修改和删除的操作。

  说明

  - 当前空间转换为公共层空间后，不支持回退为普通空间，其他普通空间也不能再转换为公共层空间。请谨慎选择您的公共层空间。
  - 公共层空间无法反向查询、引用或操作普通空间的数据。
- **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 8-166 动态表达式

说明

如果第一次打开开关，需重置DDL模板中的衍生指标。如之前有修改过DDL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”。

说明

选择“字段一致性校验”的前提是码表在数据库中存在，通过以下方式生成的码表会在数据库中存在：

- 逆向数据库生成的码表。
- 新建维度时，通过码表维度发布的码表。

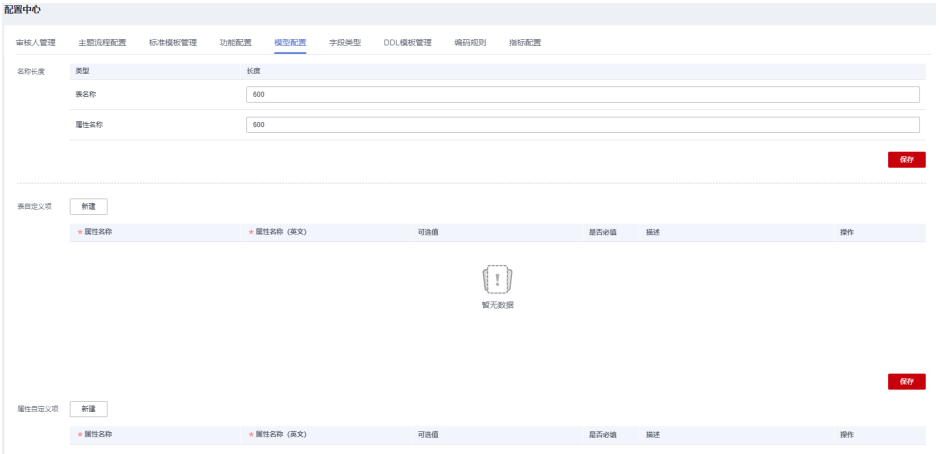
- **汇总表引用维度字段命名规则：**设置汇总表在新建、编辑、导入和生成时的命名规则，可选“维度表名_维度属性名”和“维度属性名”。
- **导出文件类型**：数据架构导出功能支持“xlsx”和“et”两种格式。逻辑模型、物理模型、维度（表）、事实表、汇总表以及其他导出均支持两种格式。
- **生成数据服务API**：包含“按汇总表整表生成单个API”和“按汇总表指标生成数个API”两种生成数据服务API的方式。

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。
- 启用密级。
- 设置长度。
- 增加表的自定义字段。
- 增加属性的自定义字段。

图 8-167 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。
- 启用密级。默认开启该字段。
- 名称长度：设置表名称和属性名称的长度。
- 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。
- 属性自定义项。在新建、编辑表字段时，可以在表字段中设置自定义的属性。会影响业务表、维度（维度表）、事实表和汇总表等。

字段类型

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。

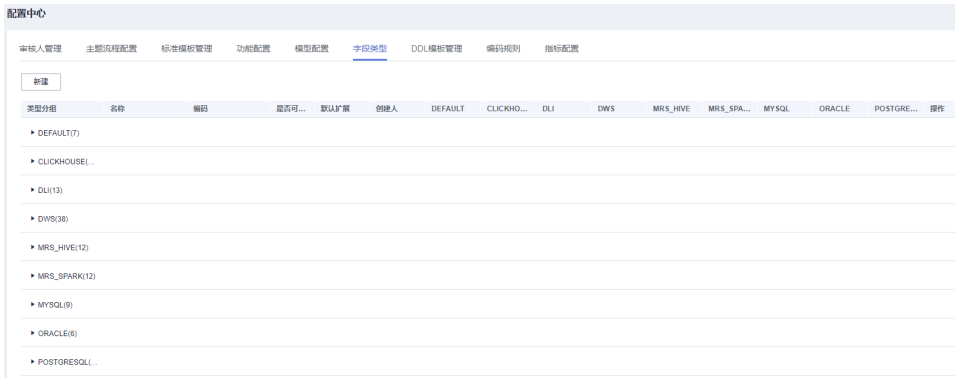
步骤1 在数据架构控制台，单击左侧导航树中的“配置中心”，进入相应页面后再单击“字段类型”页签。

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。
- MRS_SPARK：MRS_SPARK连接类型Hudi表的数据类型。
- POSTGRESQL：POSTGRESQL连接类型的表的数据类型。
- CLICKHOUSE：CLICKHOUSE连接类型的表的数据类型。
- MYSQL：MYSQL连接类型的表的数据类型。
- ORACLE：ORACLE连接类型的表的数据类型。
- DORIS：DORIS连接类型的表的数据类型。

图 8-168 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 8-169 新建类型

  表 8-64 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。只能包含大写字母，下划线，数字，且以大写字母或下划线开头。
所属域	选择新建类型所属的域。
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。
数仓对应类型	选择新建类型所映射连接的数据类型。
DEFAULT	选择新建类型所映射的DEFAULT连接的数据类型。
CLICKHOUSE	选择新建类型所映射的CLICKHOUSE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。
MRS_SPARK	选择新建类型所映射的MRS_SPARK连接的数据类型。
MYSQL	选择新建类型所映射的MYSQL连接的数据类型。

参数名称	说明
ORACLE	选择新建类型所映射的ORACLE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。
DORIS	选择新建类型所映射的DORIS连接的数据类型。

- **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8-64。

- **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

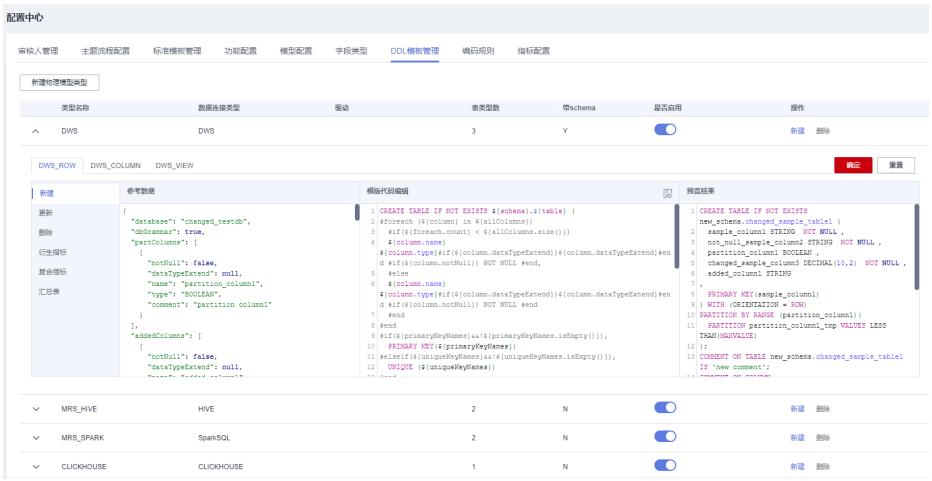
在DataArts Studio数据架构中，支持修改各种类型（例如DLI、POSTGRESQL、DWS、Hive、SPARK、DORIS）的表或DLI视图的DDL模板。如果您需要将已创建的某一类型的表生成其他数据源的DDL语句，您就可以根据目标数据源的DDL语法，修改该类型的表的DDL模板。

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8-170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 8-170 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 8-171 添加编码规则

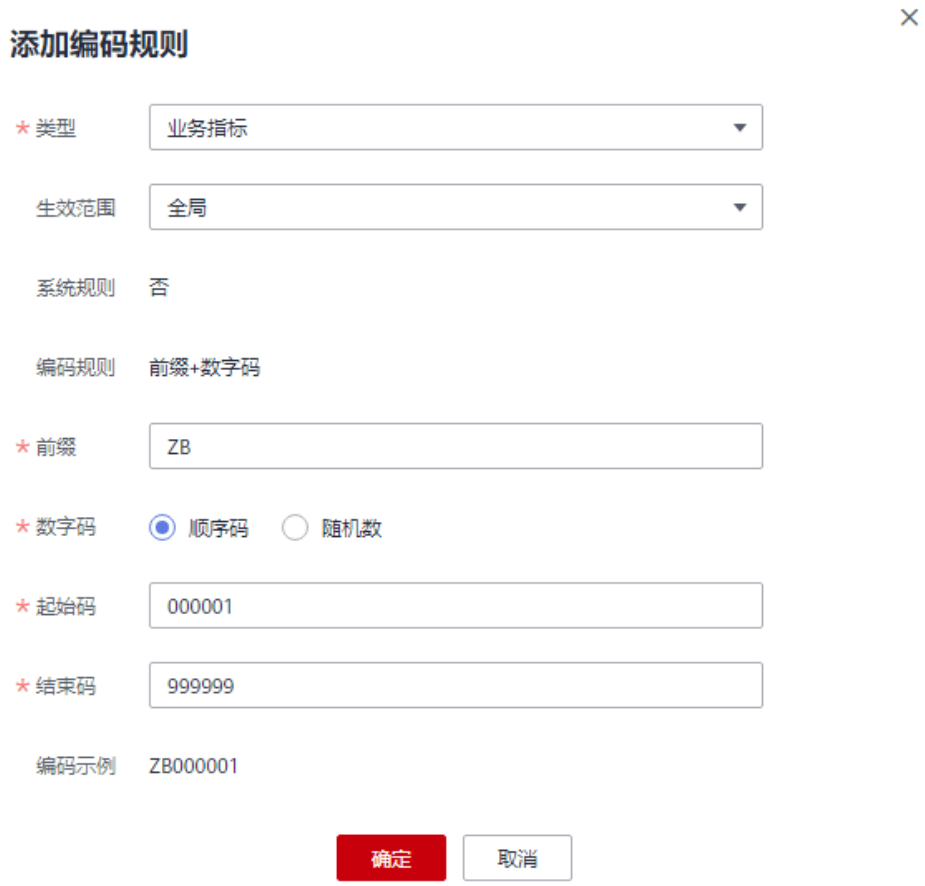


表 8-65 添加编码规则说明

参数名称	说明
类型	选择编码规则的类型，当前支持如下六种： 业务指标，逻辑实体，逻辑属性， 数据标准、码表、业务对象。
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。
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。

参数名称	说明
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

  系统预置的六个编码规则（逻辑实体、 数据标准、逻辑属性、业务指标、码表、业务对象），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”。

指标配置

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“指标配置”页签。
2. 管理业务指标

   a. 新建指标

   单击业务指标自定义项旁的新建按钮，或在已有指标的情况下，单击操作列的 + 图标新增指标。完成后配置如下参数，然后单击“保存”。

图 8-172 新建指标

表 8-66 新建指标参数说明

参数名称	说明
选项名称	自定义指标名称。不超过100字符。
选项名称（英文）	自定义指标英文名称。不超过100字符。
可选值	设置自定义指标在创建业务指标时的可选值。
是否必填	设置自定义指标在创建业务指标时是否为必填项
描述	自定义指标的描述。不超过200字符。

b. 调整指标排序

在有多个指标的情况下，可以通过操作列调整指标的排序。单击图标可进行指标的上移或者下移，双击图标可以输入序号将当前行移动到指定位置。

图 8-173 调整指标排序

图 8-174 移动到指定位置

c. 删除指标

如果需要删除自定义指标，单击操作列的 图标就可删除该指标。

图 8-175 删除指标

3. 完成自定义指标的设置后，在新建业务指标界面和完成发布的业务指标的基本信息界面，会显示已保存的自定义指标。

图 8-176 新建业务指标

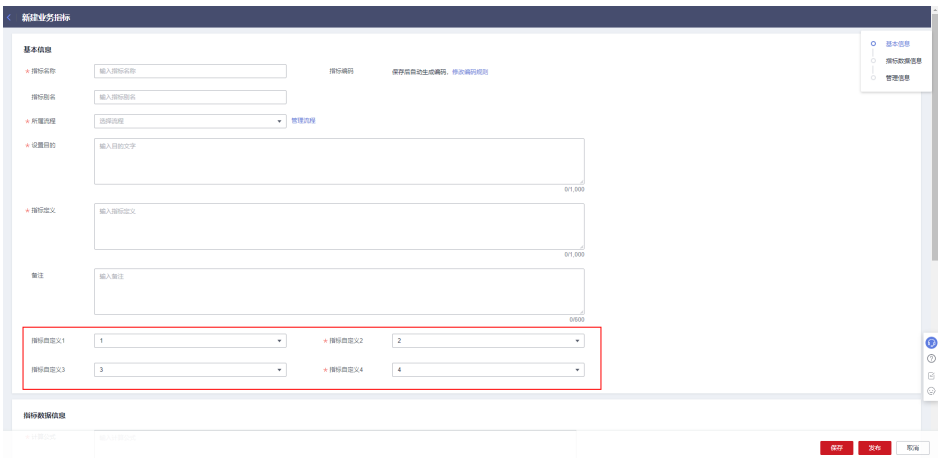


图 8-177 基本信息界面

8.8.8 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，请使用审核人员的账号参考以下步骤审核对象。

1. 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。
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。

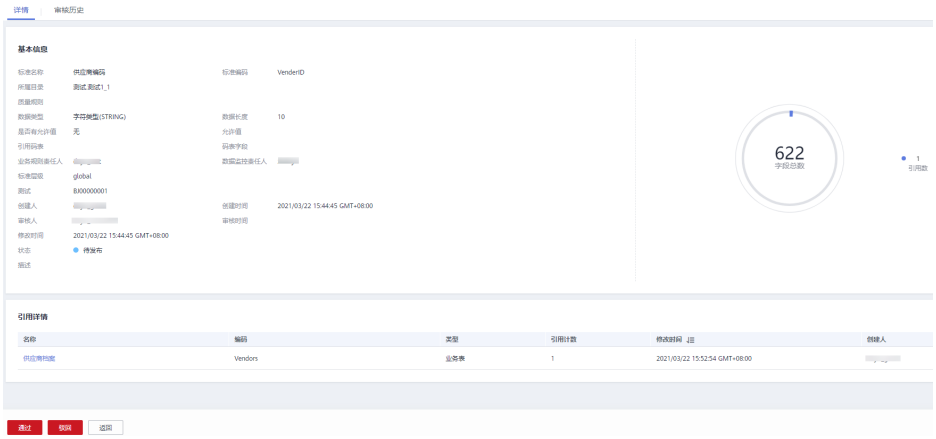
   图 8-178 审核

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 8-179 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，可以查看待审核的对象。

- **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，可以查看已通过审核的对象。

- **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核的对象。

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 8-180 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。 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。

功能区域	说明
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. 根据实际情况选择审核结果（通过或驳回），并输入有效的审核的意见。 3. 单击“确定”，完成审核。
3	•  ：通过该按钮过滤出修改时间段内的待审核信息。 •  ：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 8-181 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

消息通知

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“消息通知”，进入消息通知页面，如下图所示。

图 8-182 消息通知页面

您可以进行如下操作：

- 通过操作列“确认”，已确认知晓所选消息的相关变化。支持批量确认操作。
- 查询：支持通过属性筛选或者关键字模糊搜索消息通知信息。

----结束

8.9 使用教程

8.9.1 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 8-67 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

序号	字段名称	字段描述
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 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**：新建码表和数据标准。
   - **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。
   - **数仓规划：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. **数据集市：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，选择对应工作空间的“数据架构”模块，进入数据架构页面。
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、开发者、自定义角色审批），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账号为审核人，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 8-183 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，您可以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，如下图所示，设置“模型设计业务流程步骤”。

图 8-184 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8-68所示，说明如下：

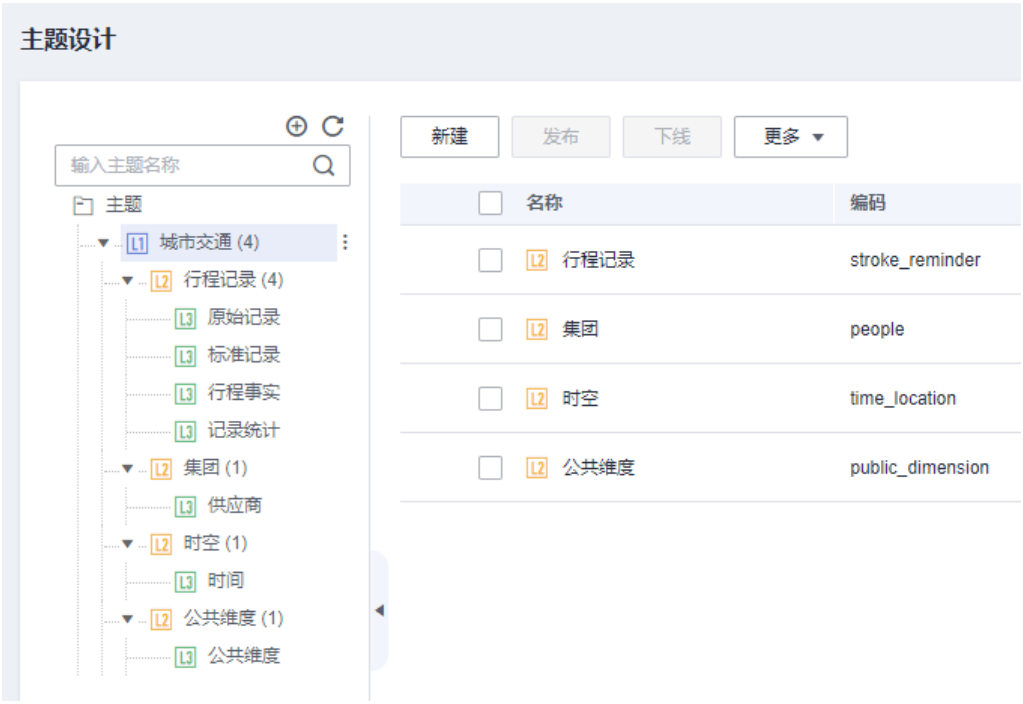
- 新建1个主题域分组：城市交通。
- 在主题域分组“城市交通”下，新建4个主题域：行程记录、集团、时空、公共维度。
- 在主题域“行程记录”下，新建4个业务对象：原始记录、标准记录、行程事实、记录统计。
- 在主题域“集团”下，新建1个业务对象：供应商。
- 在主题域“时空”下，新建1个业务对象：时间。
- 在主题域“公共维度”下，新建1个业务对象：公共维度。

表 8-68 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
				记录统计	stroke_statistic
		集团	people	供应商	vendor

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）
		时空	time_location	时间	date
		公共维度	public_dimension	公共维度	public_dimension

图 8-185 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。

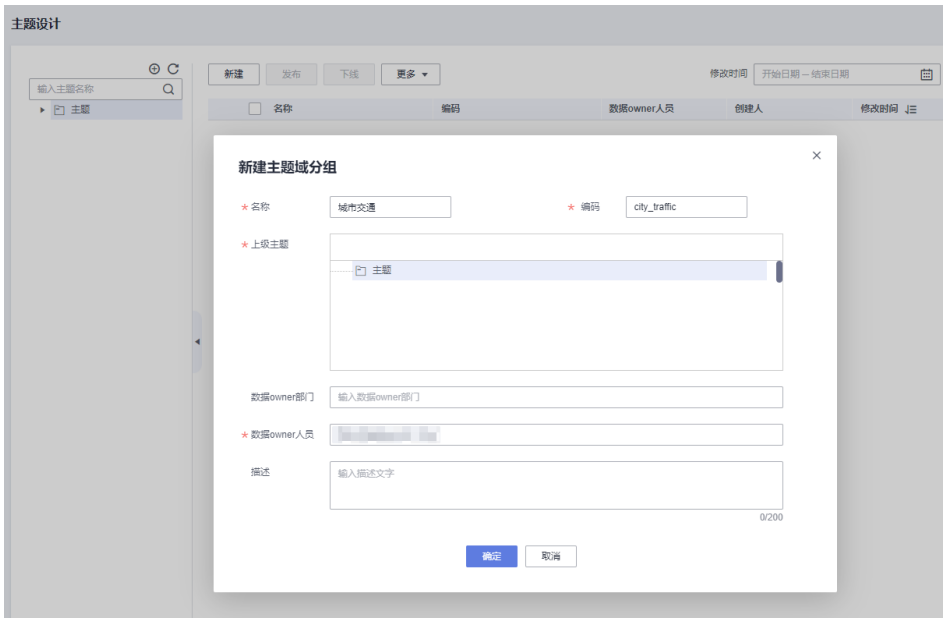
步骤3 在数据架构控制台，单击左侧菜单栏的“配置中心”。选择“ 主题流程配置”，使用默认的3层层级。

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 8-186 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 8-187 新建 L1 层主题



在弹出窗口中，按图8-187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 主题域分组创建完成后，您需要勾选主题域分组，并单击“发布”，发布主题域分组。在弹出的“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，主题域分组发布成功。如果当前账号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

图 8-188 发布主题域分组

步骤6 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。

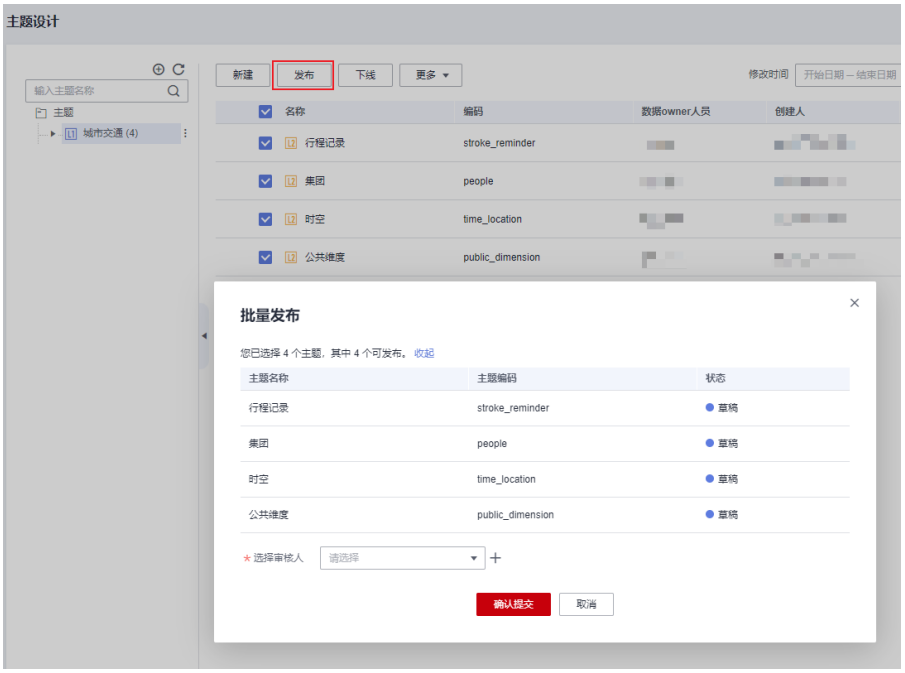
以主题域“行程记录”为例，新建主题域的步骤如下，其他主题域也请参照以下步骤进行添加：

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 8-189 创建 L2 层主题

2. 在弹出窗口中，“名称”和“编码”请参照表8-68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。
3. 主题域创建完成后，您需要勾选主题域，并单击“发布”，发布主题域。在弹出的“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，主题域发布成功。如果当前账号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

图 8-190 发布主题域



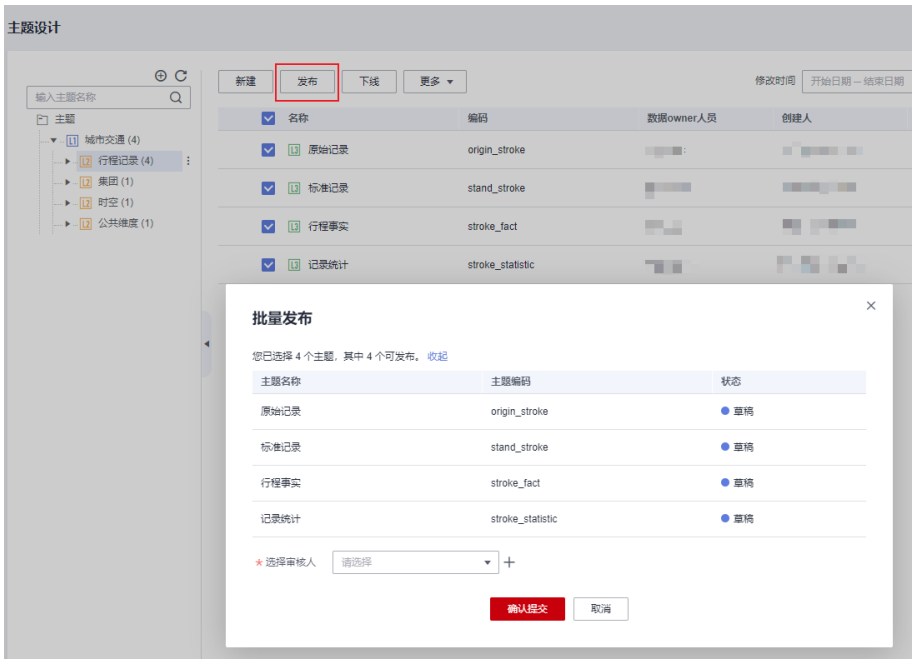
步骤7 新建业务对象。

- 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8-68**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。
3. 业务对象创建完成后，您需要勾选业务对象，并单击“发布”，发布业务对象。在弹出的“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，业务对象发布成功。如果当前账号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

图 8-191 发布业务对象



----结束

新建码表并发布

在本示例中，您需要新建如表8-69所示的3个码表：

表 8-69 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
付款方式	付款方式	payment_type	无	付款方式编码	payment_type_id	BIGINT	无
				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 8-192 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”。

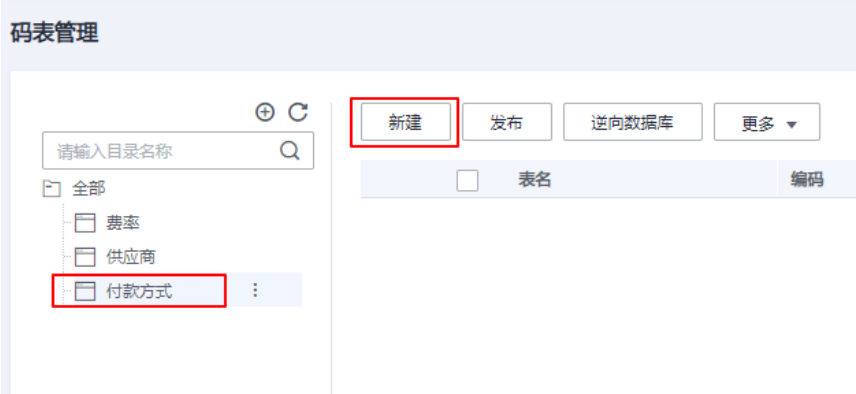
图 8-193 新建码表目录

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：

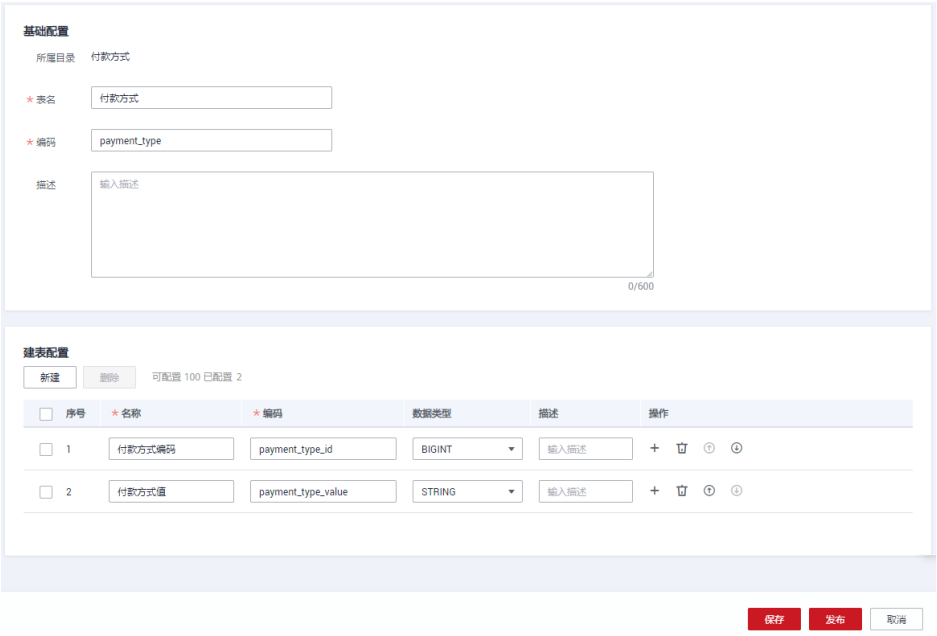
1. 在码表管理页面，在码表目录树中选择一个目录，然后在右侧单击“新建”按钮。

图 8-194 码表管理



2. 在新建码表页面中，请参考表8-69配置参数，然后单击“保存”。

图 8-195 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 8-196 供应商码表

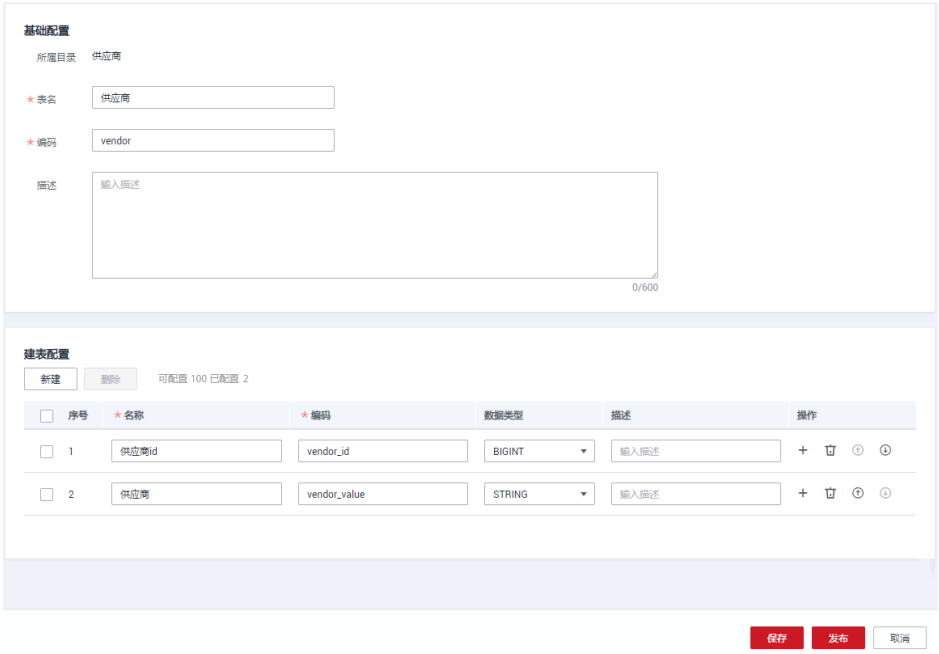
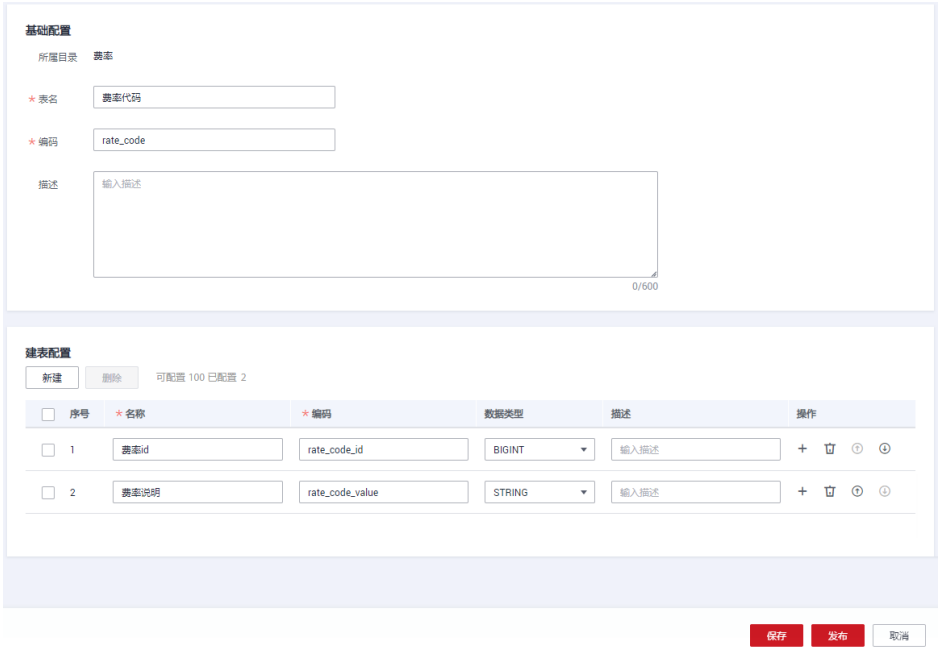


图 8-197 费率码表



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8-70**所示的数值。

表 8-70 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8-71所示的数值。

表 8-71 供应商码表的数值

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8-72所示的数值。

表 8-72 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。如果当前账号具备审核人权限，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8-73所示的3个数据标准：

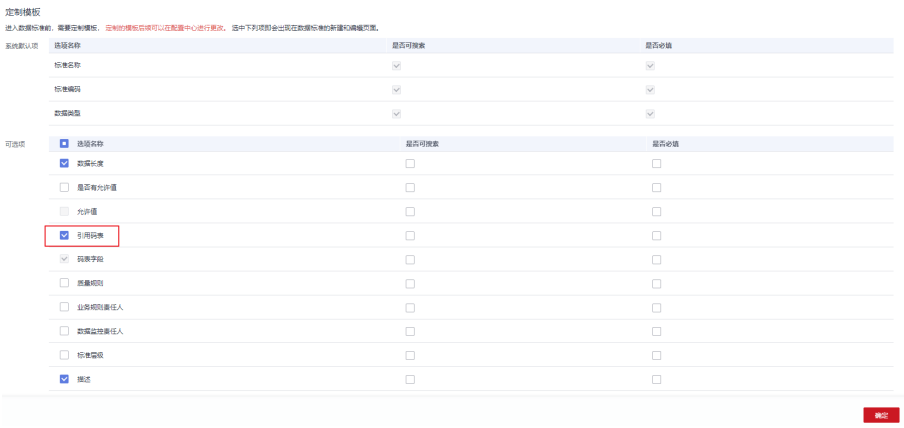
表 8-73 数据标准

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
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 8-198 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 8-199 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**配置中心**的“标准模板管理”定制数据标准模板。

图 8-200 数据标准-付款方式

图 8-201 数据标准-供应商

图 8-202 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。如果当前账号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

----结束

数仓规划：新建 SDI 层和 DWI 层两个模型

在数仓规划中，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“数仓规划”。

选择SDI层，单击“添加模型”，新建一个SDI层关系模型，命名为“sdi”，再选择DWI层，单击“添加模型”，新建一个DWI层关系模型，命名为“dwi”。单击“确定”即可。

图 8-203 添加 SDI 层关系模型

图 8-204 添加 DWI 层关系模型

1. 先新建一个SDI层关系模型，命名为“sdi”。在SDI层中，单击“添加模型”，进入新建模型页面，配置如下参数，单击“确定”。

图 8-205 新建 SDI 物理模型

2. 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“添加模型”，进入新建模型页面，配置如下参数，单击“确定”。

图 8-206 新建 DWI 模型

步骤2 在“数仓规划”页签中，单击新建的SDI关系模型，进入到“关系建模”页面，展开主题后，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。

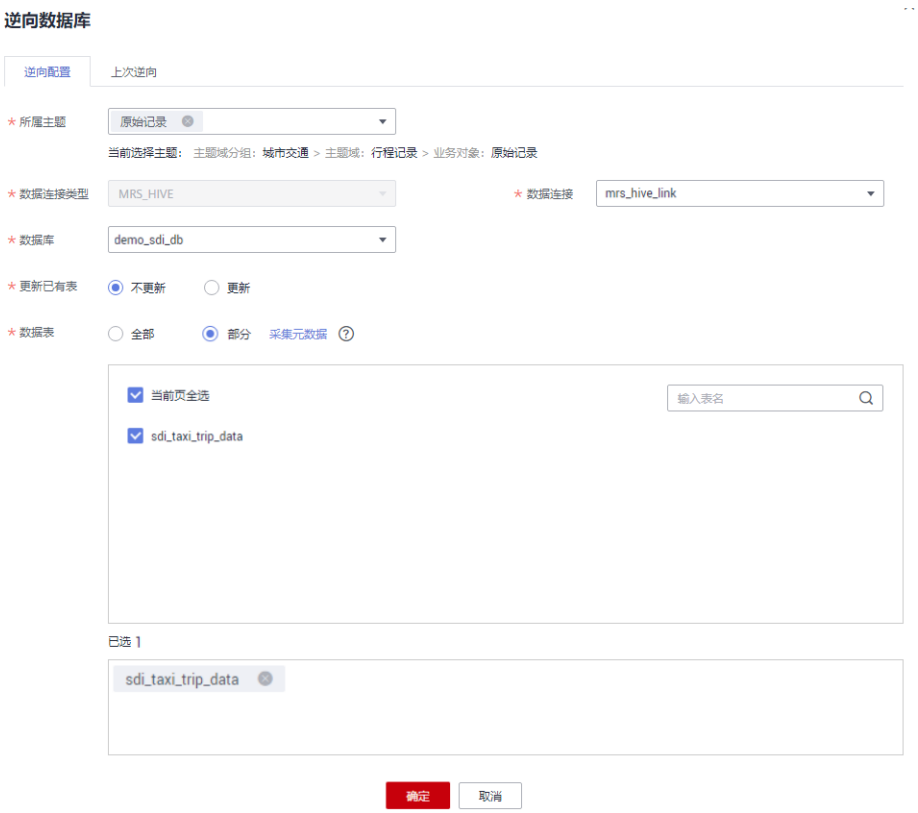
说明

逆向数据库依赖于数据资产采集，请您确保已对所需逆向的数据库完成数据资产采集。

图 8-207 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 8-208 逆向数据库



逆向数据库成功后，单击“关闭”。逆向后的表为草稿状态，在单击“发布”后，在列表中可查看导入并发布的表。

图 8-209 查看表

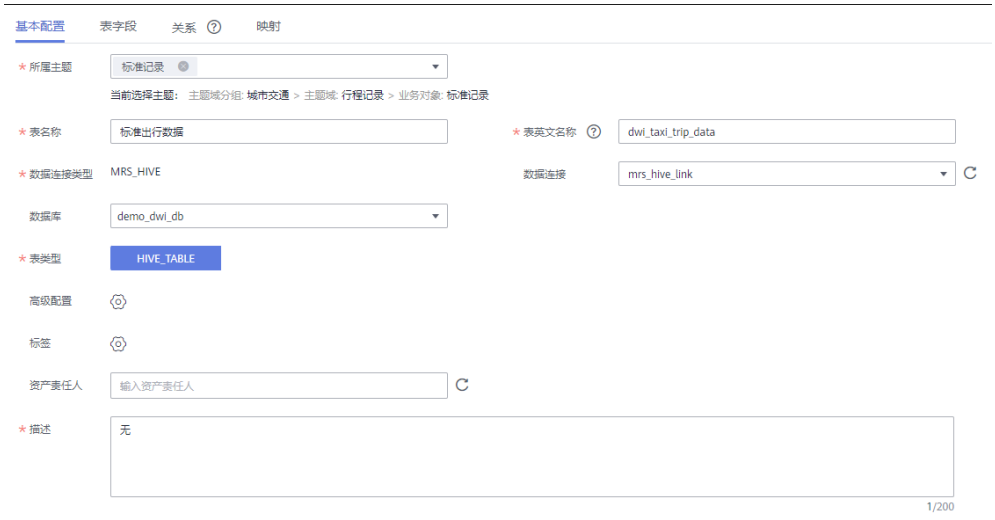
步骤3 请参照以下步骤，新建一个“标准出行数据”的标准化的业务表。

1. 在“数仓规划”页签中，单击新建的DWI关系模型，进入到“关系建模”页面，展开主题后，选中DWI模型中的业务对象“城市交通 > 行程记录 > 标准记录”，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：

表 8-74 标准出行数据表

*所属主题	*表名称	*表英文名称	*数据连接	数据库	*描述
标准记录	标准出行数据	dwi_taxi_trip_data	mrs_hive_link	demo_dwi_db	无

图 8-210 行程数据表基本配置



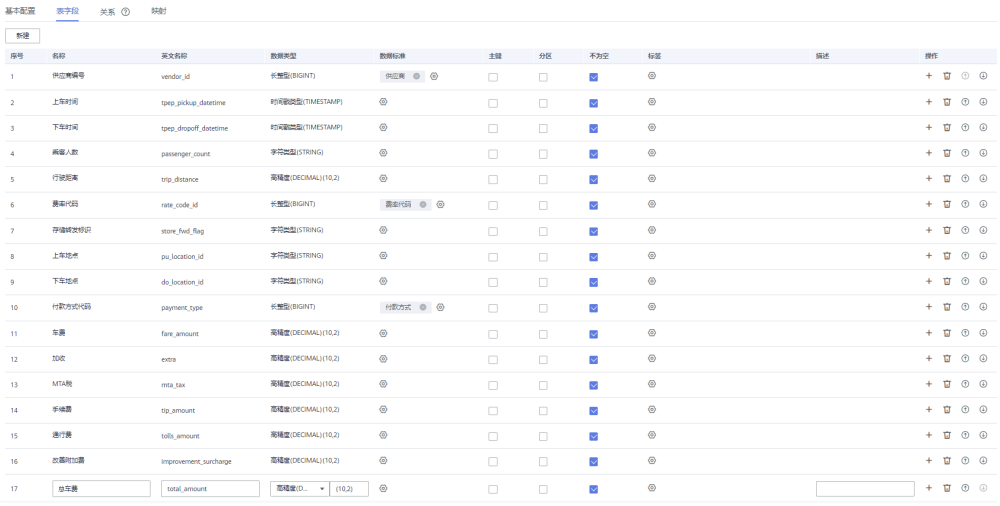
3. 单击“下一步”，进入“表字段”标签页。单击“新建”，在标准出行数据表中，依次添加如表8-75所示的字段，并单击字段供应商编号、费率代码、付款方式的“数据标准”列中的 按钮，分别关联数据标准“供应商”、“费率代码”和“付款方式”。添加完成后如图8-211所示。

表 8-75 标准出行数据表字段

序号	名称	英文名称	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	-
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	-

序号	名称	英文名称	数据类型	数据标准	主键	分区	不为空	标签
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	改善附加费	improvement_surcharge	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	-
17	总车费	total_amount	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	-

图 8-211 标准出行数据表字段



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**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。

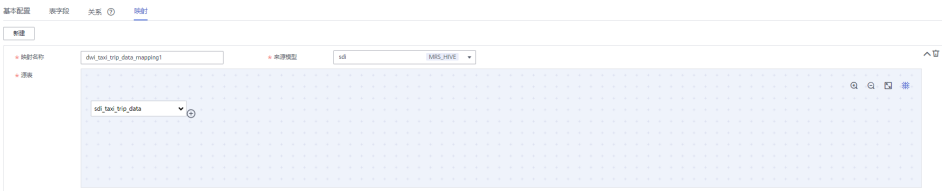
- **关联质量规则**

  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**关联质量规则**。

4. 单击“下一步”，进入“关系”标签页，本示例不涉及。
5. 继续单击“下一步”，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8-212**。
   - **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 8-212 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8-213**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”页面中勾选了“模型设计业务流程步骤 > 创建数据开发作业”（默认不勾选），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 8-213 字段映射

6. 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 8-214 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。如果当前账号具备审核人权限，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。如果当前账号具备审核人权限，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”页面选择对应模型，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 8-215 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 8-216 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8-76所示的3个码表维度。

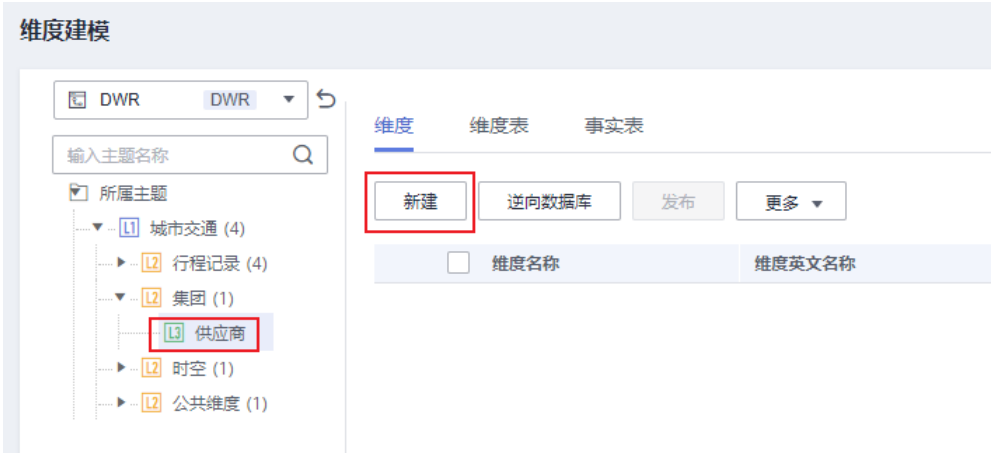
表 8-76 码表维度

*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

*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
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

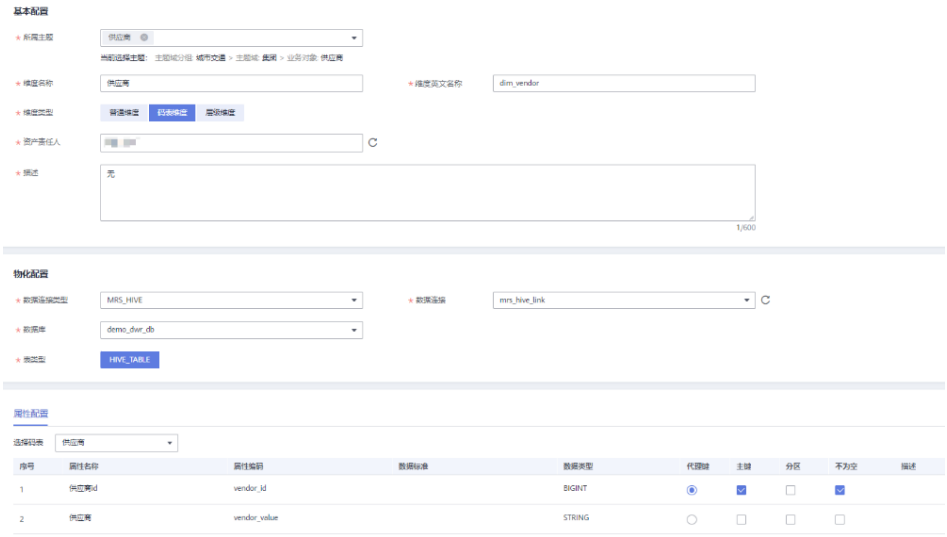
1. 在“维度建模”页面进入“维度”标签页，在主题树中选中“城市交通 > 集团 > 供应商”，然后单击“新建”新建供应商维度。

图 8-217 维度建模



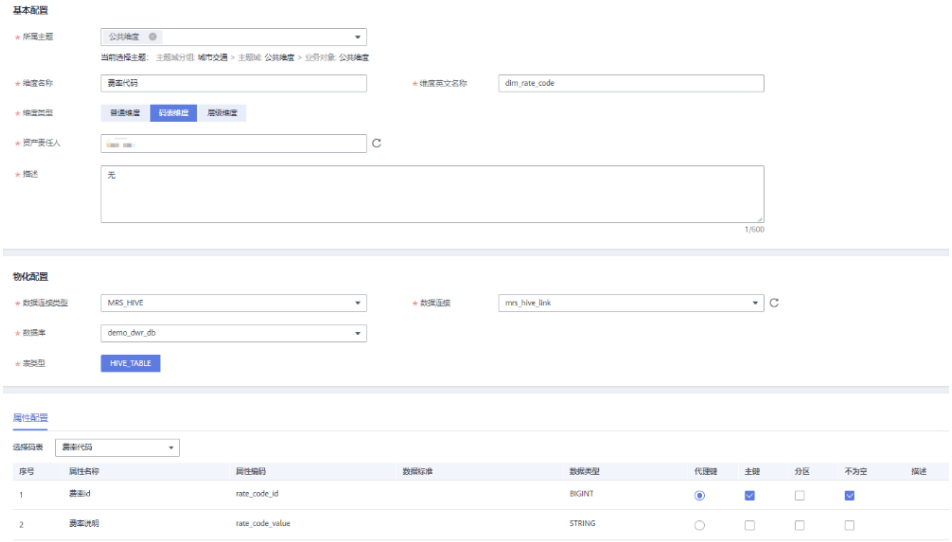
2. 在新建维度页面，如下图所示配置参数，然后单击“保存”完成维度的新建。

图 8-218 新建维度



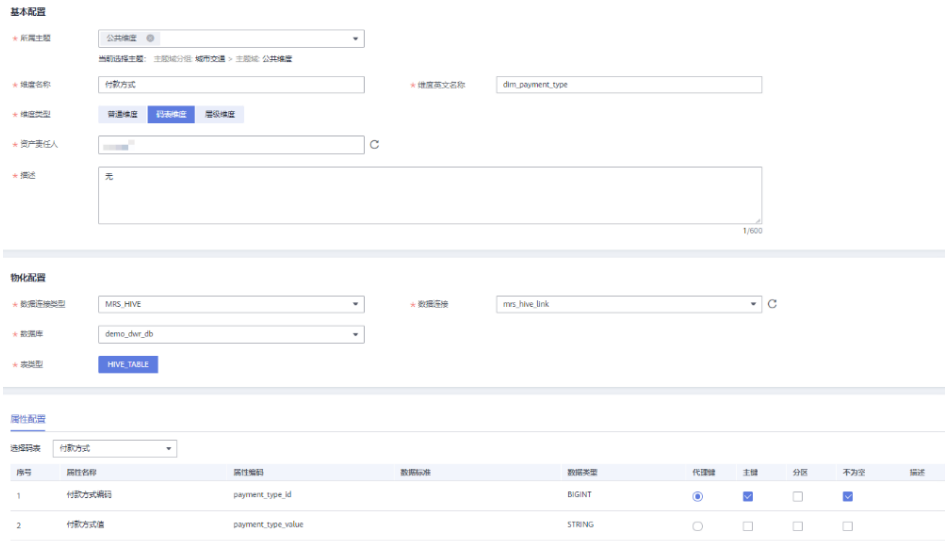
3. 在“维度建模”页面进入“维度”标签页，在主题树中选中“城市交通 > 公共维度 > 公共维度”，然后单击“新建”新建费率代码维度。在新建维度页面，配置如下，配置完成后单击“保存”。

图 8-219 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 8-220 付款方式维度



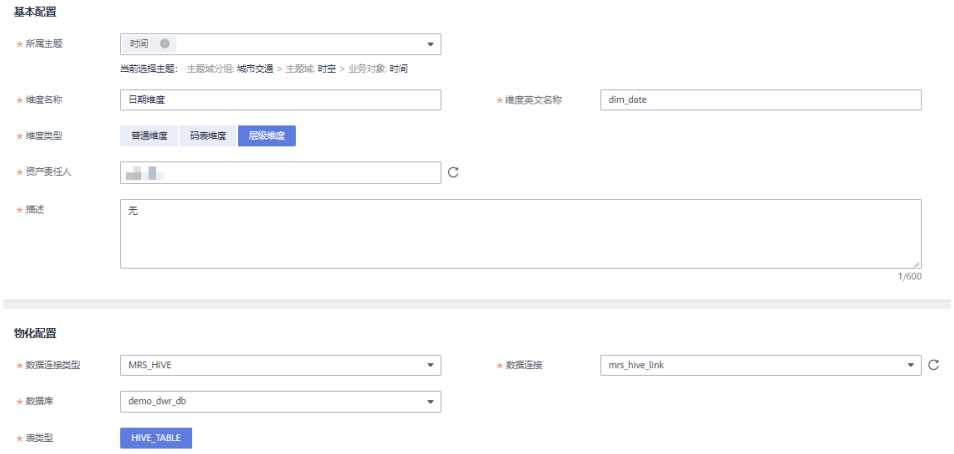
步骤3 新建一个层级维度“日期维度”。

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 8-77 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 8-221 日期维度



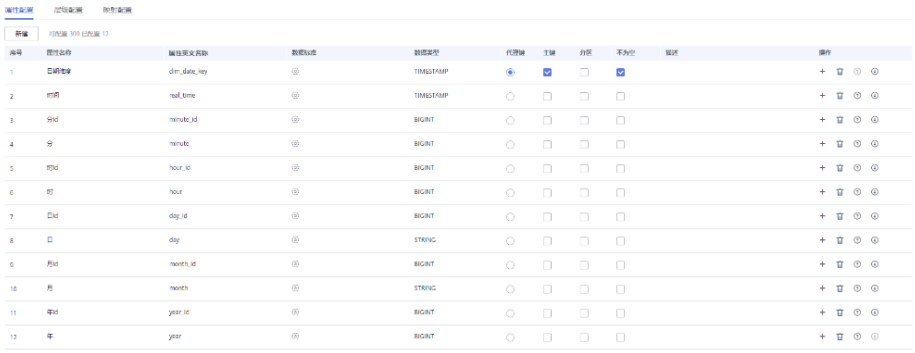
3. 属性配置如下：

表 8-78 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

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
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	不选	不选	不勾选	不勾选
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 8-222 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 8-223 层级 1

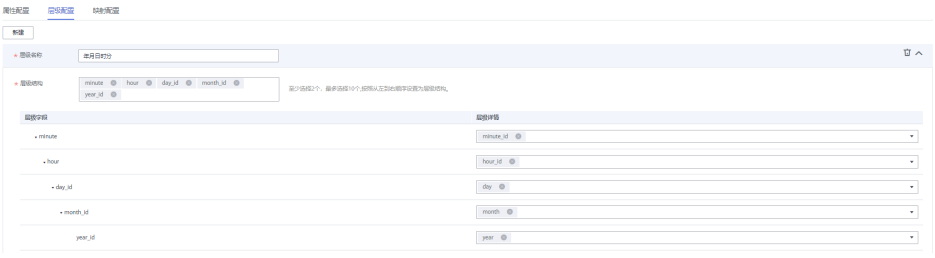


图 8-224 层级 2

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。如果当前账号具备审核人权限，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 8-225 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8-79所示。

表 8-79 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

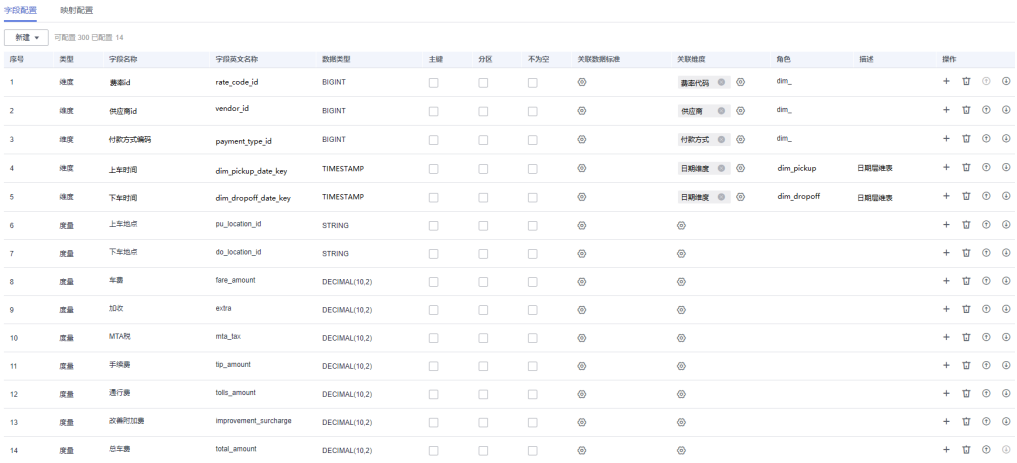
在“字段配置”区域，选择“新建 > 度量”，依次新建如表8-80所示的字段。

表 8-80 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
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	-

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
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 8-226 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。如果当前账号具备审核人权限，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8-81和表8-82所示的技术指标：

表 8-81 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 8-82 衍生指标

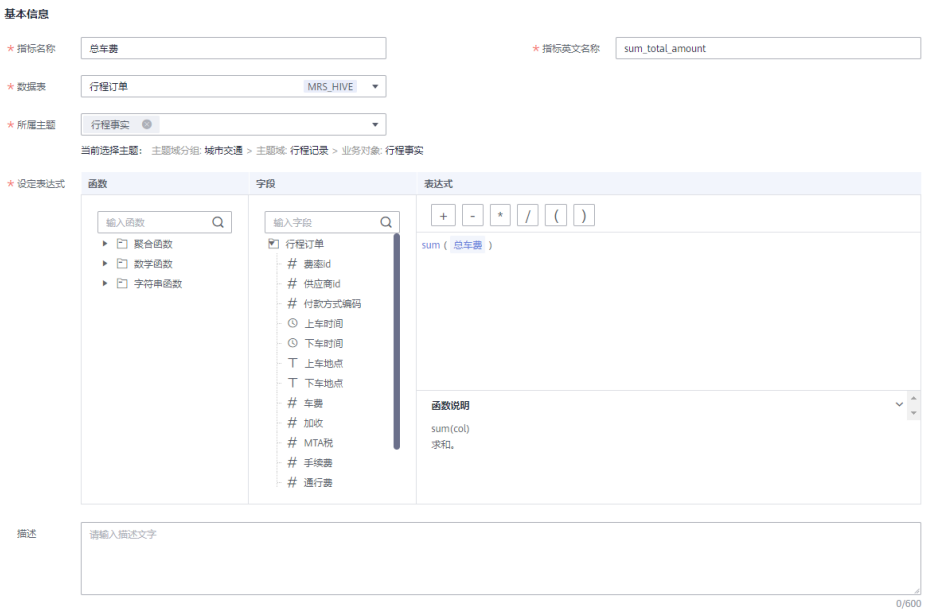
指标	*数据表	*所属主题	*原子指标	统计维度	时间限定	通用限定
基于付款方式维度统计总车费	行程订单	记录统计	总车费	付款方式	无	无
基于费率代码维度统计总车费	行程订单	记录统计	总车费	费率代码	无	无
基于供应商和下车时间维度统计总车费	行程订单	记录统计	总车费	供应商，行程订单.下车时间	无	无

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 8-227 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 8-228 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 8-229 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 8-230 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。如果当前账号具备审核人权限，也可以勾选“自助审批”，直接提交即可以审核通过。

----结束

数据集市：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8-83所示的汇总表。

表 8-83 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“数据集市”，进入数据集市页面。

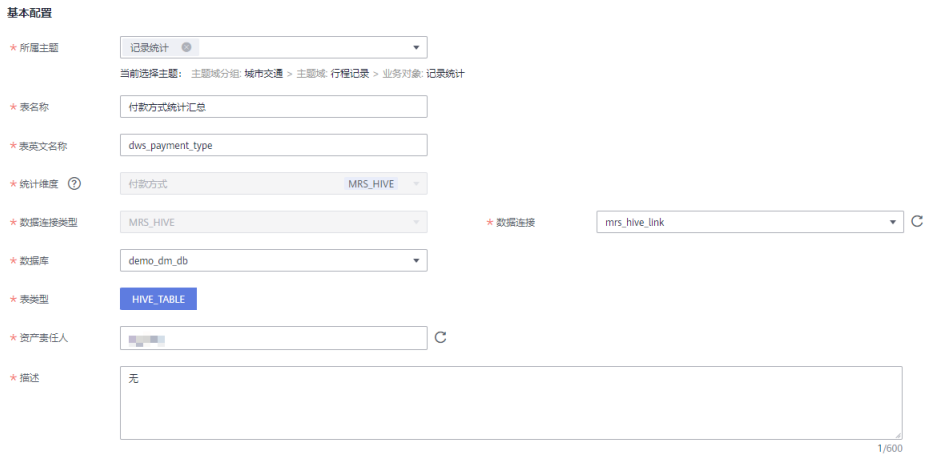
步骤2 单击“汇总表”页签，进入汇总表页面。

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 8-231 付款方式统计汇总



在“属性配置”区域，单击“添加”，输入时间周期字段名称以及选择数据类型。

图 8-232 属性配置 1

在“属性配置”区域，单击“添加”，添加衍生指标“总车费(付款方式)”，设置关联对象，选择对应的指标。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 8-233 属性配置 2

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 8-234 费率统计汇总-基本配置

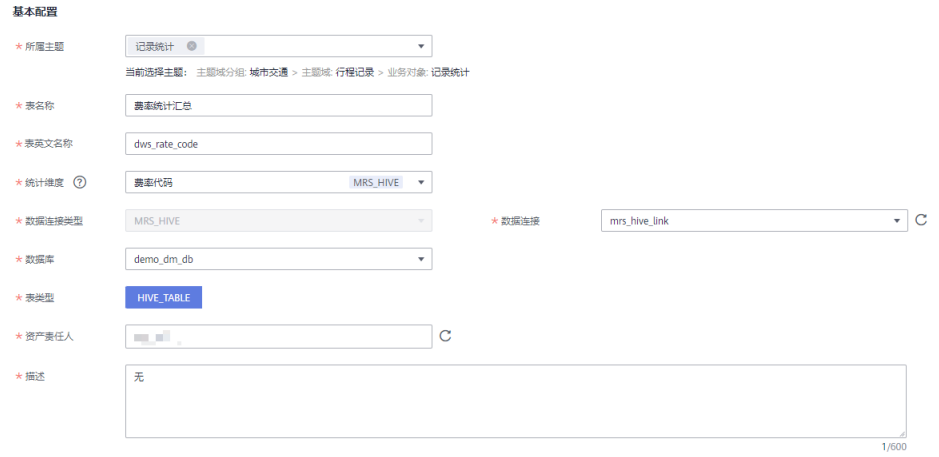
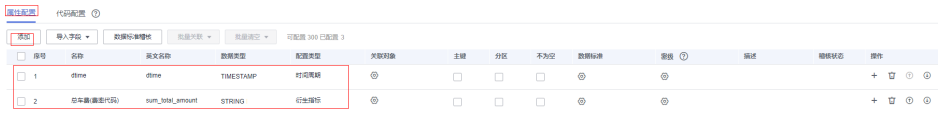


图 8-235 费率统计汇总-属性配置



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 8-236 供应商统计汇总-基本配置

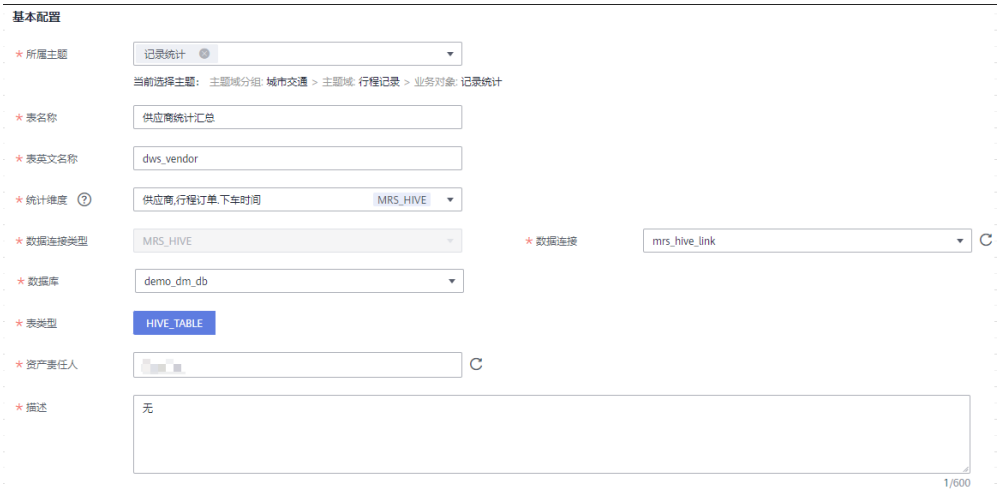
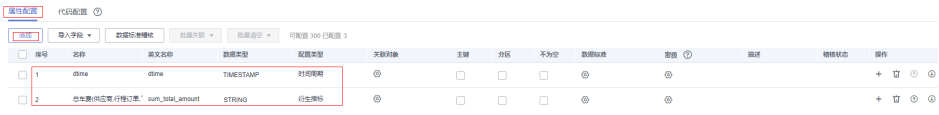


图 8-237 供应商统计汇总-属性配置



步骤4 返回数据集市页面的“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。如果当前账号具备审核人权限，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“数据集市 > 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

9 数据开发

9.1 数据开发概述

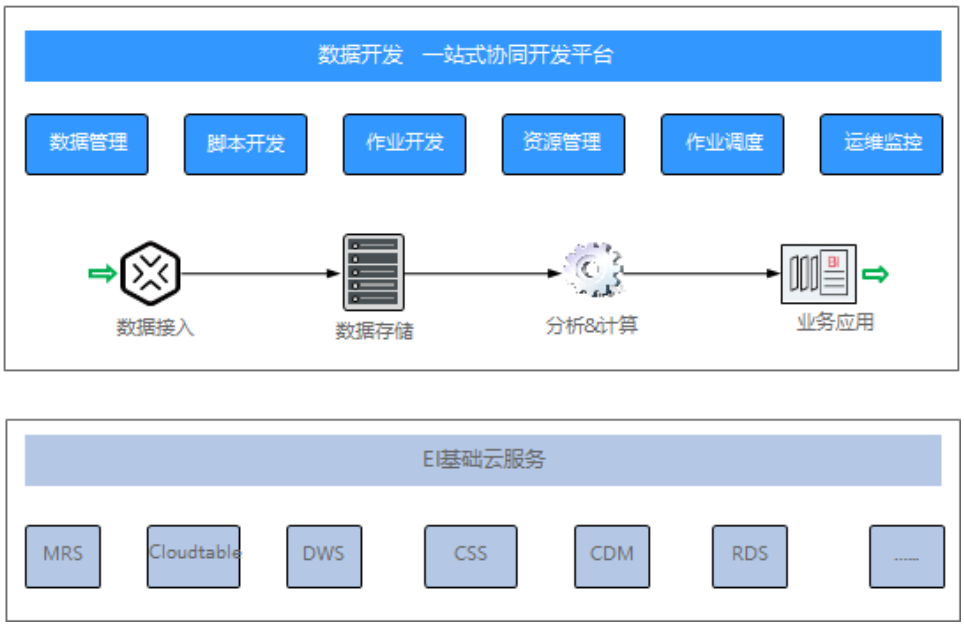
数据开发是一个一站式的大数据协同开发平台，提供全托管的大数据调度能力。它可管理多种大数据服务，极大降低用户使用大数据的门槛，帮助您快速构建大数据处理中心。

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 9-1 数据开发模块架构



数据开发的主要功能

表 9-1 数据开发的主要功能

支持的功能	说明
数据管理	● 支持管理DWS、DLI、MRS Hive等多种数据仓库。 ● 支持可视化和DDL方式管理数据库表。
脚本开发	● 提供在线脚本编辑器，支持多人协作进行SQL、Shell、Python脚本在线代码开发和调测。 ● 支持使用变量和函数。
作业开发	● 提供图形化设计器，支持拖拉拽方式快速构建数据处理工作流。 ● 预设数据集成、SQL、Shell等多种任务类型，通过任务间依赖完成复杂数据分析处理。 ● 支持导入和导出作业。
资源管理	支持统一管理在脚本开发和作业开发使用到的file、jar、archive类型的资源。
作业调度	支持单次调度、周期调度和事件驱动调度，周期调度支持分钟、小时、天、周、月多种调度周期。调度周期配置为小时，系统支持按间隔小时和离散小时配置调度周期。
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，执行作业可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。

9.2 数据管理

9.2.1 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、DLI等多种数据湖。
- 支持可视化和DDL方式管理数据库表。

  说明

  注意，在MRS API连接方式下，不支持通过可视化方式查看与管理该连接下的数据库、数据表和字段。

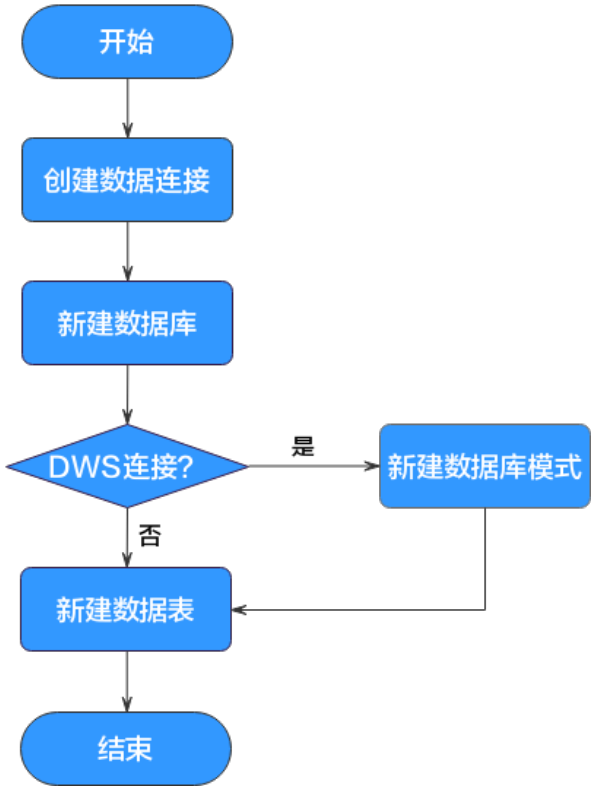
- 单击  可以查看数据连接目录树下的数据库、数据表以及字段信息。DWS SQL、DLI SQL、MRS Hive SQL代理模式均支持查看目录树，其他数据连接均不支持。

说明

如果您在使用数据开发前，已创建了数据连接和对应的数据库和数据表，则可跳过数据管理操作，直接进入**脚本开发**或**作业开发**。

数据管理的使用流程如下：

图 9-2 数据管理流程



1. 创建数据连接，连接相关数据湖底座服务。具体请参见**新建数据连接**。
2. 基于相应服务，新建数据库。具体请参见**新建数据库**。
3. 如果是DWS连接，则需要新建数据库模式；否则直接新建数据表。具体请参见**（可选）新建数据库模式**。
4. 新建数据表。具体请参见**新建数据表**。

9.2.2 新建数据连接

通过新建数据连接，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**配置DataArts Studio数据连接参数**。

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 单击  ，进入连接目录列表。
5. 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。
6. 在引用列表窗口，可以查看该连接被作业或脚本引用的情况。

图 9-3 引用列表

9.2.3 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。比如，MRS服务。
- 已新建数据连接，请参见**新建数据连接**。
- MRS API方式连接不支持通过可视化模式管理数据库，建议通过SQL脚本方式进行创建。
- 删除数据库时，请确保该数据库未被使用，且没有关联数据表。

新建数据库（可视化模式）

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本开发导航栏，选择，右键单击数据连接名称，选择“新建数据库”，配置如**表9-2**所示的参数。

   表 9-2 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。

5. 单击“确定”，新建数据库。

相关操作

- 修改数据库：在脚本开发导航栏，选择，展开下方的数据连接，右键单击数据库名称，选择“修改”后，在弹出的页面中修改数据库的信息。

- 删除数据库：在脚本开发导航栏，选择，展开下方的数据连接，右键单击数据库名称，选择“删除”后，在弹出的页面中单击“确定”完成删除。

  说明

  删除操作不可撤销，请谨慎操作。

9.2.4 （可选）新建数据库模式

DWS数据连接创建完成后，可以在右侧区域中管理DWS数据连接的数据库模式。

说明

如果已有的数据库模式满足您的使用需求，则您可以跳过本章节；否则，请您按照本章节描述新建数据库模式。

前提条件

- 已新建DWS数据连接，请参见**新建数据连接**。
- 已新建DWS数据库，请参见**新建数据库**。

新建数据库模式

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本开发导航栏，选择，展开DWS数据连接，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”。
5. 在弹出的“新建模式”页面，配置如**表9-3**所示的参数。

   表 9-3 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。

6. 单击“确定”，新建数据库模式。

相关操作

- 修改数据库模式：在脚本开发导航栏，选择，展开下方的数据连接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”后，在弹出的页面中修改数据库模式的信息。
- 删除数据库模式：在脚本开发导航栏，选择，展开下方的数据连接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”后，在弹出的页面中单击“确定”完成删除。

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。

9.2.5 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，使用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

- 已创建数据库及DWS数据库模式，请参见**新建数据库**和**（可选）新建数据库模式**。
- 已在数据开发模块中创建与数据表类型匹配的数据连接，请参见**新建数据连接**。

新建数据表（可视化模式）

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本开发导航栏，选择，展开数据连接层级至“tables”，右键单击“新建数据表”或者单击新建数据表。
5. 在弹出的对话框中，显示“配置基本属性”页面，参见**表9-4**配置相关参数。

   表 9-4 基本属性

数据连接类型	参数说明
DLI	请见**表9-8**的“基本属性”部分
DWS	请见**表9-9**的“基本属性”部分
MRS Hive	请见**表9-10**的“基本属性”部分

6. 单击“下一步”，在“配置表结构”页面配置如**表9-5**所示的参数。

表 9-5 表结构

数据连接类型	参数说明
DLI	请见表9-8的“表结构”部分
DWS	请见表9-9的“表结构”部分
MRS Hive	请见表9-10的“表结构”部分

7. 单击“保存”，新建数据表。

新建数据表（DDL 模式）

1. 参考访问DataArts Studio实例控制台，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本开发导航栏，选择，展开数据连接层级至“tables”，右键单击“新建数据表”或者单击新建数据表。
5. 单击“DDL模式建表”，如表9-6所示的参数，系统自动默认，并在下方的编辑器中输入SQL语句。例如：

```
CREATE TABLE userinfo ( id INT, name STRING);
```

说明

不同数据源的SQL语法有所差异，开发SQL语句前请预先了解各数据源的语法参考文档。

表 9-6 数据表参数

参数	说明
数据连接类型	数据表所属的数据连接类型。
数据连接	数据表所属的数据连接。
数据库	数据表所属的数据库。

6. 单击“保存”，新建数据表。

相关操作

- 查看表详情：在脚本开发导航栏，选择，展开下方的数据连接至数据表层级，右键单击表名称，选择“查看表详情”，可查看如表9-7所示的数据表信息。

表 9-7 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。

页签名称	说明
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。

- 删除表：在脚本开发导航栏，选择，展开下方的数据连接至数据表层级，右键单击表名称，选择“删除”后，在弹出的页面中单击“确定”完成删除。

  说明

  删除操作不可撤销，请谨慎操作。

参数说明

表 9-8 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。只能包含英文字母、数字、“_”，但不能为纯数字，且不能以“_”、数字开头。
别名	否	数据表的别名。只能包含中文、英文字母、数字、“_”，但不能为纯数字，且不能以“_”开头。
数据连接类型	是	数据表所属的数据连接类型。系统默认。
数据连接	是	数据表所属的数据连接。系统默认。
数据库	是	数据表所属的数据库。系统默认。
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。

参数	是否必选	说明
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。 如果OBS路径不存在或者OBS桶不存在，系统支持可以自动创建OBS目录。 **说明** 如果OBS桶创建超过上限，系统会自动提示“创建obs目录失败，错误原因：[Create OBS Bucket failed:TooManyBuckets:You have attempted to create more buckets than allowed]”。
表描述	否	数据表的描述信息。
表结构		
列类型	是	选择列类型。包含分区列和普通列。系统默认普通列。
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。 单击  ，删除列。

表 9-9 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、“_”，但不能为纯数字，且不能以“_”、数字开头。
别名	否	数据表的别名。只能包含中文、英文字母、数字、“_”，但不能为纯数字，且不能以“_”开头。
数据连接类型	是	数据表所属的数据连接类型。系统默认。
数据连接	是	数据表所属的数据连接。系统默认。
数据库	是	数据表所属的数据库。系统默认。
模式	是	选择数据库的模式。
表描述	否	数据表的描述信息。

参数	是否必选	说明
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。
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》。

参数	是否必选	说明
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。 单击，删除列。

表 9-10 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。只能包含英文字母、数字、“_”，但不能为纯数字，且不能以“_”、数字开头。
别名	否	数据表的别名。只能包含中文、英文字母、数字、“_”，但不能为纯数字，且不能以“_”开头。
数据连接类型	是	数据表所属的数据连接类型。系统默认。
数据连接	是	选择数据表所属的数据连接。系统默认。
数据库	是	选择数据表所属的数据库。系统默认。
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。

参数	是否必选	说明
数据分类	是	选择数据类型的类别： • 原始类型 • ARRAY • MAP • STRUCT • UNION
类型	是	选择数据类型，具体说明请参见 LanguageManual DDL。
列描述	否	填写列的描述信息。
操作	否	单击，增加列。 单击，删除列。

9.3 脚本开发

9.3.1 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，支持生成保存版本和提交版本。

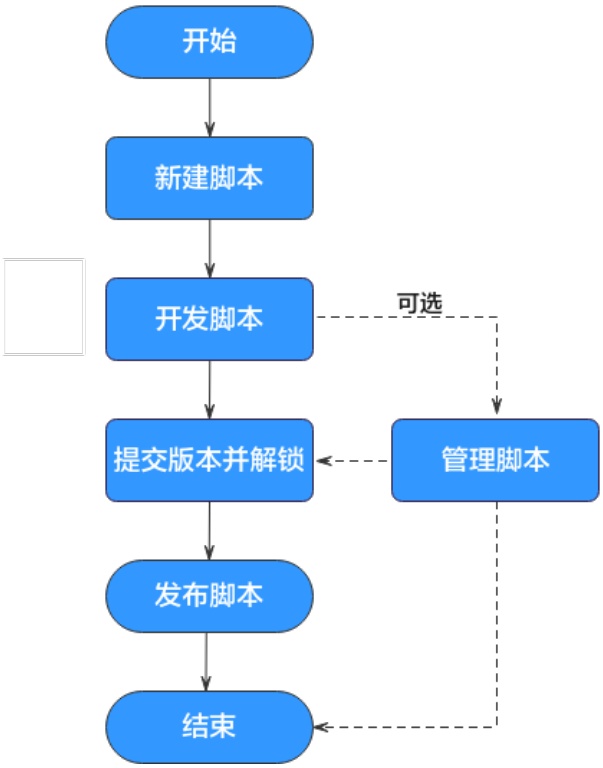
  说明

  保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。
- 支持单击右键，可快速复制脚本名称，同时可以快速的关闭已打开的脚本页签。
- 在MRS API连接模式下，MRS Spark SQL和MRS Hive SQL脚本运行完以后，在执行结果中查看运行日志，增加一键跳转MRS Yarn查看日志的链接。
- 企业模式下，开发脚本时，鼠标放置在上，单击“前往发布”跳转到任务发布页面。
- 支持对“已提交”和“未提交”的脚本进行筛选。未提交的脚本通过红色进行标识。
- 系统支持脚本参数以弹框的形式进行展示，参数名不能修改，参数值可以修改。你可以单击“测试参数”查看脚本中所引用的参数信息，同时可以查看环境中已配置的环境变量信息，不可修改，SQL语句中的参数可以按照参数名进行排序。

- 支持SQL编辑器风格配置。鼠标放置在☰上，单击“风格配置”，可以对编辑器、操作栏、注释模板进行配置、以及查询SQL脚本编辑器可使用的快捷键。
- SQL查询结果展示支持表格和列表两种展示方式。单击“风格配置”，在“编辑器配置”里面可以对SQL查询结果展示进行配置。
- 企业模式下，支持从脚本开发界面快速前往发布。鼠标放置在☰上，单击“前往发布”，进入待发布任务界面。
- 支持Hive SQL、DLI SQL、DWS SQL、RDS SQL和Impala SQL脚本可以查看右侧的数据表，单击表名前面的单选框，可以查看该数据的列名、字段类型和描述。
- 脚本开发支持细粒度权限管控，在数据安全模块对数据开发脚本目录权限管控策略进行配置。

脚本开发的使用流程如下：

图 9-4 脚本开发流程



1. 新建脚本：新建相应类型的脚本。具体请参见**新建脚本**。
2. 开发脚本：基于新建的脚本，进行脚本的在线开发、调试和执行。具体请参见**开发脚本**。
3. 提交版本并解锁：脚本开发完成后，您需要提交版本并解锁，提交版本并解锁后才能正式地被作业调度运行，便于其他开发者修改。具体请参见**提交版本**。
4. （可选）管理脚本：脚本开发完成后，您可以根据需要，进行脚本管理。具体请参见**（可选）管理脚本**。
5. 发布脚本。企业模式下需要发布脚本，具体请参见**发布脚本任务**。

9.3.2 新建脚本

数据开发模块的脚本开发功能支持新建、编辑、调试、执行各类SQL、Python和shell脚本，开发脚本前请先新建脚本。

前提条件

- 已完成**新建数据连接**和**新建数据库**等操作。
- 脚本在每个工作空间的最大配额为10000，脚本目录最多5000个，目录层级最多为10层。请确保当前数量未达到最大配额。

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本目录中，右键单击目录名称，选择“新建目录”。
5. 在弹出的“新建目录”页面，配置如**表9-11**所示的参数。

   表 9-11 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。

6. 单击“确定”，新建目录。

新建脚本

1. 在脚本目录中，右键单击目录名称，选择新建相应的脚本。
2. 进入脚本开发页面，具体操作请参见**开发SQL脚本**、**开发Shell脚本**、**开发Python脚本**。

   说明

   当前最多支持创建5个同类型的临时脚本。当关闭了临时未保存的脚本，再次新建同类型的脚本时，会打开上次未保存的临时脚本。

9.3.3 开发脚本

9.3.3.1 开发 SQL 脚本

数据开发支持对SQL脚本进行在线开发、调试和执行，开发完成的脚本可以在作业中调度运行（请参见**开发Pipeline作业**）。

数据开发模块支持如下类型SQL脚本。而不同数据源的SQL语法有所差异，开发SQL语句前请预先了解各数据源的语法规则。

- DLI SQL脚本：请参见**SQL语法参考**。
- Hive SQL脚本：请参见**SQL语法参考**。
- DWS SQL脚本：请参见**SQL语法参考**。
- Spark SQL脚本：请参见**SQL语法参考**。
- ClickHouse SQL脚本：请参见**SQL语法参考**。
- IMPALA SQL脚本：请参见**SQL语法参考**。
- Flink SQL脚本：请参见**SQL语法参考**。
- RDS SQL脚本：请参见**SQL语法参考**。
- Presto SQL脚本：请参见**SQL语法参考**。
- Spark Python脚本：请参见**SQL语法参考样例**。
- Doris SQL脚本：请参见**SQL语法参考**。

前提条件

- 已开通相应的云服务并在云服务中创建数据库。
- 已创建与脚本的数据连接类型匹配的数据连接，请参见**新建数据连接**。Flink SQL脚本不涉及该操作。
- 当前用户已锁定该脚本，否则需要通过“抢锁”锁定脚本后才能继续开发脚本。新建或导入脚本后默认被当前用户锁定，详情参见**编辑锁定功能**。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本目录中，双击脚本名称，进入脚本开发页面。
5. 在编辑器上方，选择如**表9-12**所示的属性。创建Flink SQL脚本时请跳过此步骤。

表 9-12 SQL 脚本属性

属性	说明
数据连接	选择数据连接。
DLI数据目录	选择DLI的数据目录。 • 在DLI默认的数据目录dli。 • 在DLI所绑定的LakeFormation已创建元数据catalog。
数据库	选择数据库。 DLI数据目录如果选择DLI默认的数据目录dli，表示为DLI的数据库和数据表。 DLI数据目录如果选择DLI所绑定的LakeFormation已创建元数据catalog，表示为LakeFormation的数据库和数据表。

属性	说明
资源队列	输入执行作业的资源队列。 Impala SQL、Hive SQL只能手动输入，不支持选择。 选择执行DLI作业的资源队列。当脚本为DLI SQL时，配置该参数。选择了资源队列以后，单击可以查看队列性能，系统支持查看DLI运行作业数和队列CU使用量，系统显示队列近24小时性能情况。 **说明** • 当队列选择为“default”时，会提示“暂不支持"default"队列性能展示”。 • 还未选择资源队列时，图标灰化不可查看队列性能。 如需新建资源队列，请参考以下方法： • 单击，进入DLI的“购买队列”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”，仅用于用户体验，用户间可能会出现抢占资源的情况，不能保证每次都可以得到资源执行相关操作。当遇到执行时间较长或无法执行的情况，建议您在业务低峰期再次重试，或选择自建队列运行业务。 另外，“default”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： **说明** • 环境变量配置项需要以"hoodie."或"dli.sql."或"dli.ext."或"dli.jobs."或"spark.sql."或"spark.scheduler.pool"开头。 • 环境变量为dli.sql.autoBroadcastJoinThreshold时，值只能为整数，环境变量为dli.sql.shuffle.partitions时，值只能为正整数。 • 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。 • dli.sql.autoBroadcastJoinThreshold（自动使用BroadcastJoin的数据量阈值） • dli.sql.shuffle.partitions（指定Shuffle过程中Partition的个数） • dli.sql.cbo.enabled（是否打开CBO优化策略） • dli.sql.cbo.joinReorder.enabled（开启CBO优化时，是否允许重新调整join的顺序） • dli.sql.multiLevelDir.enabled（OBS表的指定目录或OBS表分区表的分区目录下有子目录时，是否查询子目录的内容；默认不查询） • dli.sql.dynamicPartitionOverwrite.enabled（在动态分区模式时，只会重写查询中的数据涉及的分区，未涉及的分区不删除）

属性	说明
	说明 在非调度场景的DLI SQL脚本运行和DLI SQL单任务作业测试运行时，系统会默认开启以下四个配置参数： • spark.sql.adaptive.enabled（启用AQE，使Spark能够根据正在处理的数据的特征动态优化查询的执行计划，可以通过减少需要处理的数据量来提高性能。） • spark.sql.adaptive.join.enabled（启用AQE用于连接操作，可以通过根据正在处理的数据动态选择最佳连接算法来提高性能。） • spark.sql.adaptive.skewedJoin.enabled（启用AQE用于倾斜的连接操作，可以通过自动检测倾斜的数据并相应地优化连接算法来提高性能） • spark.sql.mergeSmallFiles.enabled（启用合并小文件功能，可以通过将小文件合并成较大的文件来提高性能，可以减少处理许多小文件的时间，并通过减少需要从远程存储中读取的文件数量来提高数据本地性。） 如果不使用的话，可以手动配置相关参数进行关闭，参数值设置为false。

6. 在编辑器中输入SQL语句，支持输入多条SQL语句。

   不同数据源的SQL语法有所差异，开发SQL语句前请预先了解各数据源的语法规则。

   - DLI SQL脚本：请参见**SQL语法参考**。
   - Hive SQL脚本：请参见**SQL语法参考**。
   - DWS SQL脚本：请参见**SQL语法参考**。
   - Spark SQL脚本：请参见**SQL语法参考**。
   - ClickHouse SQL脚本：请参见**SQL语法参考**。
   - IMPALA SQL脚本：请参见**SQL语法参考**。
   - Flink SQL脚本：请参见**SQL语法参考**。
   - RDS SQL脚本：请参见**SQL语法参考**。
   - Presto SQL脚本：请参见**SQL语法参考**。
   - Spark Python脚本：请参见**SQL语法参考样例**。
   - Doris SQL脚本：请参见**SQL语法参考**。

说明

- 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.
```
- RDS SQL当前不支持begin ... commit事务语法，若有需要，请使用start transaction ... commit事务语法。
- 脚本内容大小不能超过16MB。
- 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。
- 当前用户提交Spark SQL脚本到MRS时，默认提交至其绑定的租户队列（绑定队列即用户绑定的租户类型角色所对应的队列）中运行。当绑定多个队列时，系统会优先根据内部排序选择队列进行提交；如果需要给该用户使用固定一个队列进行提交， 可以登录FusionInsight Manager界面，在"租户资源 > 动态资源计划 > 全局用户策略"中给该用户配置默认队列，详细操作请参见**管理全局用户策略**。
- Flink SQL、Hive SQL、Spark SQL脚本支持语法检查。单击“语法检查”，SQL语句校验完成后，可以在下方查看语法校验结果。

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。

  - F8：运行
  - F9：停止
  - 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 + 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
  - Ctrl + D：删除一行
  - Shift + Ctrl + U：解锁

- Ctrl + Alt + K：同词选择
- Ctrl + B：格式化
- Ctrl + Shift + Z：重做
- Ctrl + Enter：执行所选行/选中内容
- Ctrl + Alt + F：标记
- Ctrl + Shift + K：查找上一个
- Ctrl + K：查找下一个
- Ctrl + Backspace：删除左侧单词
- Ctrl + Delete：删除右侧单词
- Alt + Backspace：删除至行首
- Alt + Delete：删除至行尾
- Alt + Shift-Left：选择行首
- Alt + Shift-Right：选择行尾

– 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 Python、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，在右下角单击“生成SQL语句”，生成的SQL语句需要您手动格式化。

– 支持脚本参数（当前仅Flink SQL不支持该功能）。

在SQL语句中直接写入脚本参数，调试脚本时可以在脚本编辑器下方输入参数值。如果脚本被作业引用，在作业开发页面可以配置参数值，参数值支持使用EL表达式（参见**表达式概述**）。

说明

SQL脚本中的参数如果涉及变量，变量的格式应该与**脚本变量定义**中设置的格式保持一致，如果不一致，变量将不会被识别。

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 9-5 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。

- 企业模式下，支持从脚本开发界面快速前往发布。标放置在☰上，单击“前往发布”，进入待发布任务界面。
- 在MRS API连接方式下，Spark SQL和HIve SQL脚本支持配置指定参数和参数值。代理连接不支持。

  说明

  单击右上角的⚙，设置相关脚本的环境变量。举例如下所示：

  设置Hive SQL脚本的环境变量：

  --hiveconf hive.merge.mapfiles=true;

  --hiveconf mapred.job.queue.name=queue1

  设置Spark SQL脚本的环境变量：

  --num-executors 1

  --executor-cores 4

  --queue queue2

  前者表示参数名，后者表示参数值。

  脚本运行完之后，到MRS管理面查看运行详情。

7. （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。
8. 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

- 执行SQL结果最多展示1000条，仅DLI SQL支持最多10000条。如需查看更多执行结果，请参考**下载或转储脚本执行结果**通过下载或转储获取。
- 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。
- 脚本执行历史结果可以进行权限管控，可设置为“仅自己可见”或“所有用户可见”，默认配置项请参见**脚本执行历史展示**。

9. 在编辑器上方，单击“保存”，保存脚本。

   如果脚本是新建且未保存过的，请配置如**表9-13**所示的参数。

表 9-13 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。

脚本保存后，在右侧的版本里面，会自动生成一个保存版本，支持版本回滚。保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。

下载或转储脚本执行结果

脚本运行成功后，支持下载和转储SQL脚本执行结果。系统默认支持所有用户都能下载和转储SQL脚本的执行结果。如果您不希望所有用户都有该操作权限，可参考**配置数据导出策略**进行配置。

- 脚本执行完成后在“执行结果”中，单击“下载”可以直接下载CSV格式的结果文件到本地。可以在**下载中心**查看下载记录。
- 脚本执行完成后在“执行结果”中，单击“转储”可以将脚本执行结果转储为CSV和JSON格式的结果文件到OBS中，详情请参见**表9-14**。

说明

- 转储功能依赖于OBS服务，如无OBS服务，则不支持该功能。
- 当前仅支持转储SQL脚本查询（query）类语句的结果。
- DataArts Studio的下载或转储的SQL结果中，如果存在英文逗号、换行符等这种特殊符号，可能会导致数据错乱、行数变多等的问题。

表 9-14 转储配置

参数	是否必选	说明
数据格式	是	目前支持导出CSV和JSON格式的结果文件。
资源队列	否	选择执行导出操作的DLI队列。当脚本为DLI SQL时，配置该参数。
压缩格式	否	选择压缩格式。当脚本为DLI SQL时，配置该参数。 ● none ● bzip2 ● deflate ● gzip
存储路径	是	设置结果文件的OBS存储路径。选择OBS路径后，您需要在选择的路径后方自定义一个文件夹名称，系统将在OBS路径下创建文件夹，用于存放结果文件。 您也可以到**下载中心**配置默认的OBS路径地址，配置好后在转储时会默认填写。
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。
是否导出列名	否	是：导出列名 否：不导出列名
字符集	否	● UTF-8：默认字符集。 ● GB2312：当导出数据中包含中文字符集时，推荐使用此字符集。 ● GBK：国家标准GB2312基础上扩容后兼容GB2312的标准。

参数	是否必选	说明
引用字符	否	仅在数据格式为csv格式时支持配置引用字符。 引用字符在导出作业结果时用于标识文本字段的开始和结束，即用于分割字段。 仅支持设置一个字符。默认值是英文双引号（"）。 主要用于处理包含空格、特殊字符或与分隔符相同字符的数据。 关于“引用字符”和“转义字符”的使用示例请参考**引用字符和转义字符使用示例**。
转义字符	否	仅在数据格式为csv格式时支持配置转义字符。 在导出结果中如果需要包含特殊字符，如引号本身，可以使用转义字符（反斜杠 \ ）来表示。 仅支持设置一个字符。默认值是英文反斜杠（\）。 常用转义字符的场景： • 假设两个引用字符之间的数据内容存在第三个引用字符，则在第三个引用字符前加上转义字符，从而避免字段内容被分割。 • 假设数据内容中原本就存在转义字符，则在这个原有的转义字符前再加一个转义字符，避免原来的那个字符起到转义作用。 关于“引用字符”和“转义字符”的使用示例请参考**引用字符和转义字符使用示例**。

相对于直接查看SQL脚本的执行结果，通过下载和转储能够支持获取更多的执行结果。各类SQL脚本查看、下载、转储支持的规格如**表9-15**所示。

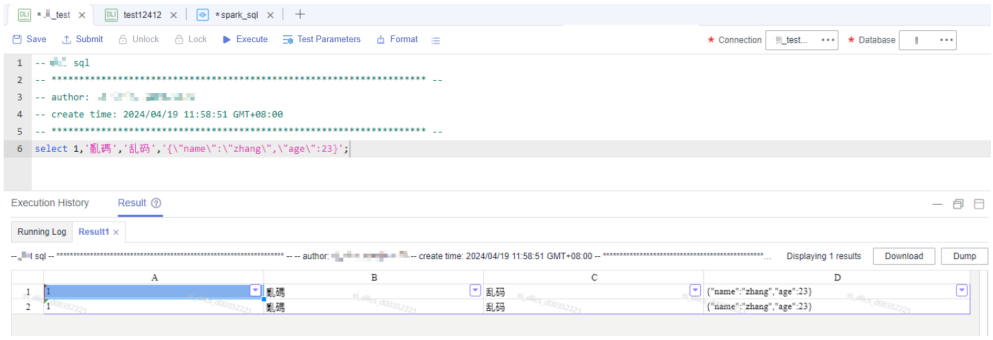
表 9-15 SQL 脚本支持查看/下载/转储规格

SQL类型	在线查看最大结果条数	下载最大结果	转储最大结果
DLI	10000	1000条且少于3MB	无限制
Hive	1000	1000条且少于3MB	10000条或3MB
DWS	1000	1000条且少于3MB	10000条或3MB
Spark	1000	1000条且少于3MB	10000条或3MB
RDS	1000	1000条且少于3MB	不支持
Presto	1000	下载结果直接转储至OBS，条数无限制。	无限制
ClickHouse	1000	1000条且少于3MB	10000条或3MB
HetuEngine	1000	1000条且少于3MB	10000条或3MB

SQL类型	在线查看最大结果条数	下载最大结果	转储最大结果
Impala	1000	1000条且少于3MB	10000条或3MB
Doris	1000	1000条且少于3MB	1000条或3MB

引用字符和转义字符使用示例

- 引用字符和转义字符使用说明：
  - 引用字符：用于识别分割字段，默认值：英文双引号（"）。
  - 转义字符：在导出结果中如果需要包含特殊字符，如引号本身，可以使用转义字符（反斜杠 \ ）来表示。默认值：英文反斜杠（\）。
    i. 假设两个quote_char之间的数据内容存在第三个quote_char，则在第三个quote_char前加上escape_char，从而避免字段内容被分割。
    ii. 假设数据内容中原本就存在escape_char，则在这个原有的escape_char前再加一个escape_char，避免原来的那个字符起到转义作用。
- 应用示例：



在进行转储时，如果引用字符和转义字符不填，如下图所示。

下载的.csv用excel打开以后如下图所示，是分成两行的。

D	E
{\name\":\"zhang\"	\"age\":23}"
{\name\":\"zhang\"	\"age\":23}"

在转储时，如果引用字符和转义字符都填写，比如，引用字符和转义字符都填英文双引号（"），则下载以后查看结果如下图所示。

D	E
{"name":"zhang","age":23}	
{"name":"zhang","age":23}	

9.3.3.2 开发 Shell 脚本

数据开发支持对Shell脚本进行在线开发、调试和执行，开发完成的脚本可以在作业中调度运行（请参见**开发Pipeline作业**）。

前提条件

- 已新增Shell脚本，请参见**新建脚本**。
- 已新建主机连接，该Linux主机用于执行Shell脚本，请参见**主机连接参数说明**。
- 连接主机的用户需要具有主机**/tmp**目录下文件的创建与执行权限。
- Shell或Python脚本可以在该ECS主机上运行的最大并发数由ECS主机的**/etc/ssh/sshd_config**文件中MaxSessions的配置值确定。请根据Shell或Python脚本的调度频率合理配置MaxSessions的值。
- 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本目录中，双击脚本名称，进入脚本开发页面。
5. 在编辑器上方，配置如**表9-16**所示的属性。

表 9-16 Shell 脚本属性

参数	说明
主机连接	选择执行Shell脚本的主机。

单击右侧的“输入参数”，可以输入执行Shell脚本的参数和交互式参数。

表 9-17 Shell 脚本参数

参数	说明
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`

参数	说明

交互式参数	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）。交互式参数之间以空格分隔，Shell脚本根据交互情况按顺序读取参数值。
	例如执行如下交互式shell脚本， 交互参数1、2、3 分别对应begin、end、exit。
	● 当交互参数输入1时，执行结果显示为“start something”。
	● 当交互参数输入2时，执行结果显示为“stop something”。
	● 当交互参数输入3时，执行结果显示为“exit”。
	`#!/bin/bash`
	`select Actions in "begin" "end" "exit"`
	`do`
	`case $Actions in`
	`"begin")`
	`echo "start something"`
	`break`
	`;;`
	`"end")`
	`echo "stop something"`
	`break`
	`;;`
	`"exit")`
	`echo "exit"`
	`break`
	`;;`
	`*)`
	`echo "Ignorant"`
	`;;`
	`esac`
	`done`
	read -p语法的使用示例：
	`read -p “输入参数1和参数2”变量1 变量2`

6. 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。
     - F8：运行
     - F9：停止
     - 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 + 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
- Ctrl + D：删除一行
- Shift + Ctrl + U：解锁
- Ctrl + Alt + K：同词选择
- Ctrl + B：格式化
- Ctrl + Shift + Z：重做
- Ctrl + Enter：执行所选行/选中内容
- Ctrl + Alt + F：标记
- Ctrl + Shift + K：查找上一个
- Ctrl + K：查找下一个
- Ctrl + Backspace：删除左侧单词
- Ctrl + Delete：删除右侧单词
- Alt + Backspace：删除至行首
- Alt + Delete：删除至行尾
- Alt + Shift-Left：选择行首
- Alt + Shift-Right：选择行尾

– 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。

- 脚本内容大小不能超过16MB。

- 企业模式下，支持从脚本开发界面快速前往发布。标放置在三上，单击“前往发布”，进入待发布任务界面。

7. 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。
   - Shell脚本运行的输出结果不能大于30M，大于30M会报错。
   - 脚本执行历史结果可以进行权限管控，可设置为“仅自己可见”或“所有用户可见”，默认配置项请参见**脚本执行历史展示**。

8. 在编辑器上方，单击“保存”，保存脚本。

   如果脚本是新建且未保存过的，请配置如**表9-18**所示的参数。

表 9-18 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。

脚本保存后，在右侧的版本里面，会自动生成一个保存版本，支持版本回滚。保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。

9.3.3.3 开发 Python 脚本

数据开发支持对Python脚本进行在线开发、调试和执行，开发完成的脚本可以在作业中调度运行（请参见**开发Pipeline作业**）。

Python脚本开发的样例教程请参见**开发一个Python脚本**。

前提条件

- 已新增Python脚本，请参见**新建脚本**。
- 已新建主机连接，该Linux主机配有用于执行Python脚本的环境。新建主机连接请参见**主机连接参数说明**。
- 连接主机的用户需要具有主机**/tmp**目录下文件的创建与执行权限。
- Shell或Python脚本可以在该ECS主机上运行的最大并发数由ECS主机的**/etc/ssh/sshd_config**文件中MaxSessions的配置值确定。请根据Shell或Python脚本的调度频率合理配置MaxSessions的值。
- 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本目录中，双击脚本名称，进入脚本开发页面。
5. 在编辑器上方，配置执行Python脚本的Python版本和主机连接。

   表 9-19 Python 脚本属性

参数	说明
Python版本	选择Python版本。 • Python2：Python版本为Python2 • Python3：Python版本为Python3
主机连接	选择执行Python脚本的主机。

   单击右侧的“输入参数”，可以输入执行Python脚本的参数和交互式参数。

   表 9-20 Python 脚本参数

参数	说明
参数	填写执行Python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Python脚本中引用，否则配置无效。
交互式参数	填写交互式参数，即执行Python脚本的过程中，需要用户输入的交互式信息（例如密码）。交互式参数之间以空格分隔，Python语句根据交互情况按顺序读取参数值。

6. 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。

- F8：运行
- F9：停止
- 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 + 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
- Ctrl + D：删除一行
- Shift + Ctrl + U：解锁
- Ctrl + Alt + K：同词选择
- Ctrl + B：格式化
- Ctrl + Shift + Z：重做
- Ctrl + Enter：执行所选行/选中内容
- Ctrl + Alt + F：标记
- Ctrl + Shift + K：查找上一个
- Ctrl + K：查找下一个
- Ctrl + Backspace：删除左侧单词
- Ctrl + Delete：删除右侧单词
- Alt + Backspace：删除至行首
- Alt + Delete：删除至行尾

- Alt + Shift-Left：选择行首
- Alt + Shift-Right：选择行尾

– 支持脚本参数功能，使用方法如下：

i. 在Python语句中直接写入脚本参数名称和参数值。当Python脚本被作业引用时，如果作业配置的参数名称与Python脚本的参数名称相同，Python脚本的参数值将被作业的参数值替换。

在脚本内部进行传参，脚本示例如下：

```
a=1
print (a)
或者
a= 'qqq'
print (a)
```

在脚本外部进行传参，比如，当Python脚本被Python作业引用时，Python脚本需要传递参数给Python作业，并且字符串参数需要用英文的单引号括起来，脚本示例如下：

```
a= 'zhang'
print (${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. 在右侧的“输入参数”中配置参数，在执行Python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Python脚本中引用，否则配置无效。

– 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。

– 脚本内容大小不能超过16MB。

– 企业模式下，支持从脚本开发界面快速前往发布。标放置在☰上，单击“前往发布”，进入待发布任务界面。

7. 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。
- Python脚本运行的输出结果不能大于30M，大于30M会报错。
- 脚本执行历史结果可以进行权限管控，可设置为“仅自己可见”或“所有用户可见”，默认配置项请参见**脚本执行历史展示**。

8. 在编辑器上方，单击“保存”，保存脚本。

如果脚本是新建且未保存过的，请配置如**表9-21**所示的参数。

表 9-21 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。

脚本保存后，在右侧的版本里面，会自动生成一个保存版本，支持版本回滚。保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。

9.3.4 提交版本

提交版本涉及到数据开发的版本管理功能。

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100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。

前提条件

已完成脚本开发任务。

提交脚本版本

“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。

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。

步骤4 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5 在脚本编辑器上方单击“提交”，提交版本。选择审批人，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，不勾选则无法单击确认。在提交版本时，单击“版本对比”可以查看当前提交版本与最近一个版本之间的差异对比。

图 9-6 提交

说明

- 如果在“审批中心”开启了提交审批的开关，则脚本提交审批后，需要审批人在“审批中心”的“待审批”页签进行审批，只有当审批通过后，脚本才能提交成功。具体操作请参见**审批配置**。如果开关是关闭状态，则不需要审批，直接提交新版本即可。

  如果要撤销已提交的审批流程，请您在“审批中心”的“我的申请”页签里进行撤销。修改完成后，可以重新提交审批。
- 开启了提交审批开关后，提交脚本、删除脚本以及导入“提交态”的脚本时，均需要进行审批。
- 关闭提交审批开关前，请确保当前工作空间已无待未审批的流程。
- 企业模式下，不支持提交审批。

----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100条版本信息）。单击“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. 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，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 9-7 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。

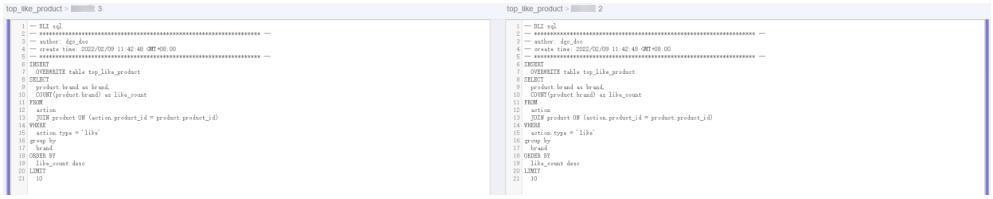
操作如下：

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 9-8 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 9-9 版本对比详情



9.3.5 发布脚本任务

在企业模式中，开发者提交脚本版本后，系统会对应产生一个脚本类型的发布任务。开发者确认发包后，待拥有管理员、部署者、DAYU Administrator、Tenant Administrator权限的用户审批通过，然后将修改后的脚本同步到生产环境。

须知

- 管理员导入脚本时，选择导入提交态，会生成对应的待发布项。
- 管理员导入脚本时，选择导入生产态，则不会生成待发布项。

前提条件

已提交版本，详情请参见**提交版本**。

操作步骤

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择开发环境，选择“数据开发 > 任务发布”。

步骤4 在待发布任务界面，会展示因提交版本而生成的待发布任务。您可以通过“查看”操作，查看当前任务相比上一版本的修改点，确认修改无误后，请通过“发布”操作，将任务进行发布。

支持通过“任务名称”和“提交人”进行发布项筛选。同时可以使用任务名称进行模糊查询。

说明

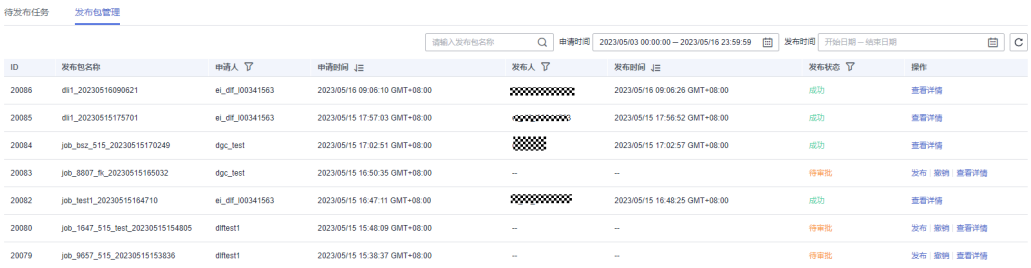
- 如果您只具备开发者权限，则需通过“发布”操作提交任务，由管理员或者部署者审批通过，才能将修改后的脚本同步到生产环境。
- 单击“发布”后，指定审批人，审批人必须是工作空间的管理员或部署者、拥有DAYU Administrator、Tenant Administrator权限的用户，至少指定一个审批人，不能指定自己为审批人。单击“审批人管理”可以跳转到“空间管理”页面，单击“编辑”按钮可以维护审批人信息。
- 可以进行批量发布。发布多个待发布项时，发布流程采用异步发布，可以看到发布任务的过程，最大的发布项个数为100。
- 对于暂时不发布的发布项，开发者、部署者和管理员可以进行撤销，支持批量撤销。

图 9-10 选择发布

步骤5 发布之后，您可以通过“发布包管理”查看任务的发布状态。待审批通过后，任务发布成功。

支持通过“申请人”、“申请时间”、“发布时间”、“发布人”和“发布状态”进行发布项筛选。同时可以使用发布包名称进行模糊查询。

图 9-11 查看任务状态



说明

对于暂时不发布的发布项，开发者、部署者和管理员可以进行撤销。

发布后，通过操作列的“查看详情”可以查看任务的发布状态和启动状态，在操作列的“版本对比”可以查看发布包不同版本间的内容差异。

图 9-12 查看发布详情

----结束

9.3.6 （可选）管理脚本

9.3.6.1 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，请参见**开发脚本**。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本目录中选择需要复制的脚本，右键单击脚本名称，选择“拷贝另存为”。

图 9-13 复制脚本

5. 在弹出的“另存为”页面，配置如表9-22所示的参数。

表 9-22 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。

6. 单击“确定”，复制脚本。

9.3.6.2 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，请参见开发脚本。

复制名称

1. 参考访问DataArts Studio实例控制台，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 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。

图 9-14 复制脚本名称

重命名脚本

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 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”。

图 9-15 重命名

说明

已经打开了的脚本文件不支持重命名。

5. 在弹出的“重命名脚本名称”页面，配置新脚本名称。

图 9-16 重命名脚本名称

6. 单击“确定”，重命名脚本。

9.3.6.3 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，请参见**开发脚本**。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 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”。

   图 9-17 移动

   b. 在弹出的“移动脚本”或“移动目录”页面，配置如表9-23所示的参数。

图 9-18 移动脚本

图 9-19 移动目录

表 9-23 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。

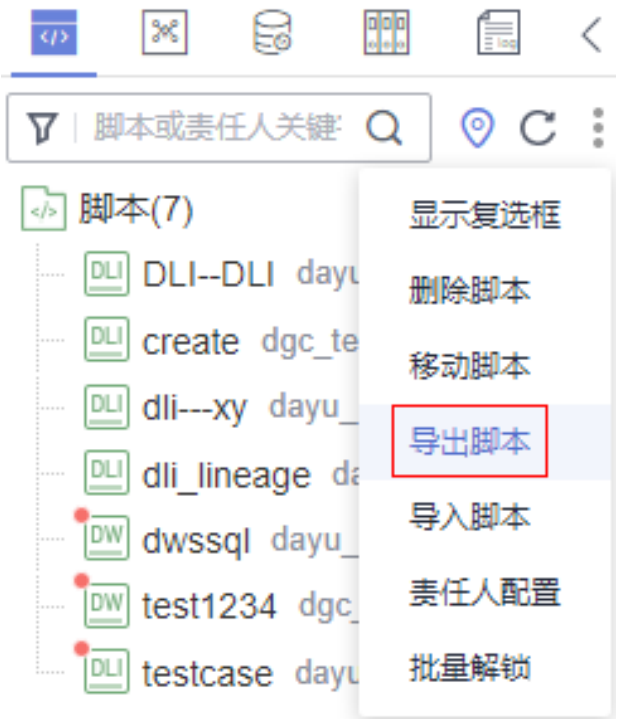
9.3.6.4 导出导入脚本

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 单击脚本目录中的，选择“显示复选框”。
5. 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 9-20 选择并导出脚本



6. 在弹出的“导出脚本”界面，选择需要导出的脚本的状态，单击“确定”。

图 9-21 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 单击脚本目录中的 > 导入脚本，选择待导入的脚本文件，并配置重名处理策略。

   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 9-22 导入脚本

5. 单击“下一步”，根据提示导入脚本。

9.3.6.5 查看脚本引用

当您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，请参见**开发脚本**。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 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。

5. 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。

图 9-23 某个脚本被引用列表

9.3.6.6 删除脚本

当您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。单击删除时，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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本目录中，右键单击脚本名称，选择“删除”。
5. 在弹出的“删除脚本”页面，单击“确认”，删除脚本。

批量删除

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 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。
5. 选择需要删除的脚本，再次单击，选择“删除脚本”。
6. 在弹出的“删除脚本”页面，单击“确认”，批量删除脚本。

9.3.6.7 解锁脚本

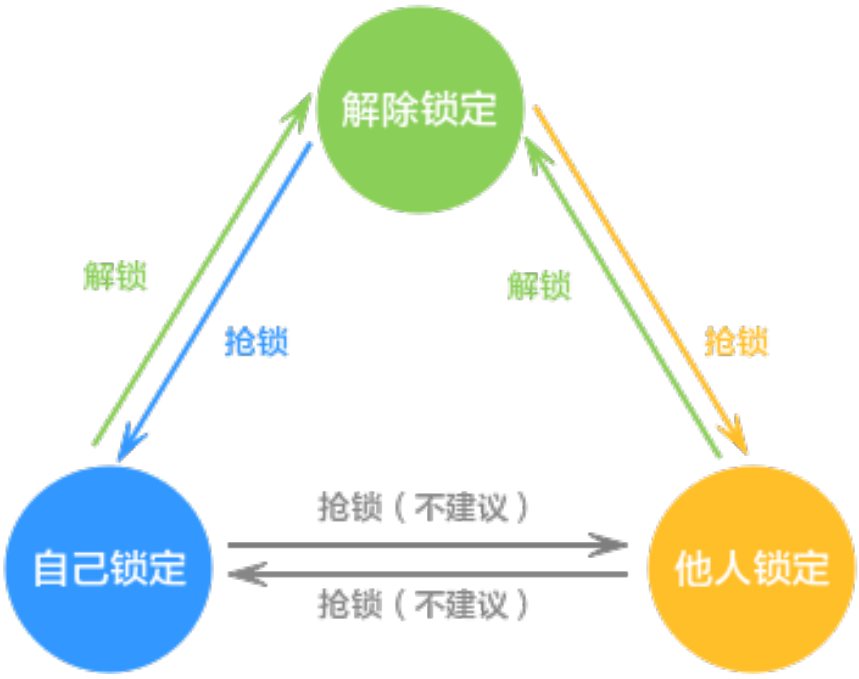
脚本/作业解锁涉及到数据开发的编辑锁定功能。

编辑锁定：用于避免多人协同开发脚本/作业时产生的冲突。新建或导入脚本/作业后，默认当前用户锁定脚本/作业，只有当前用户自己锁定的脚本/作业才可以直接编辑、保存或提交，通过“解锁”功能可解除锁定；处于解除锁定或他人锁定状态的脚本/作业，必须通过“抢锁”功能获取锁定后，才能继续编辑、保存或提交。

须知

- 当前脚本/作业的锁定状态可以通过脚本/作业的目录树查看。
- 对于已被他人锁定状态的脚本/作业，您需要通过重新打开该脚本/作业，查看最近的保存/提交时的内容。已打开的脚本/作业内容不会实时刷新。
- 在DataArts Studio更新编辑锁定功能前已经创建的脚本/作业，在更新后默认为解除锁定状态。您需要通过“抢锁”功能获取锁定后，才能继续编辑、保存或提交。
- 抢锁的操作依赖于软硬锁的处理策略。配置软硬锁的策略请参见**配置默认项**。
  - 软锁：忽略当前作业或脚本是否被他人锁定，可以进行抢锁或解锁。
  - 硬锁：若作业或脚本被他人锁定，则需锁定的用户解锁之后，当前使用人方可抢锁，空间管理员或DAYU Administrator可以任意抢锁或解锁。
- 不建议直接抢锁处于他人锁定状态的脚本/作业，这会导致他人的修改丢失。如果您有修改需求，请先联系锁定人将脚本/作业解锁，然后再抢锁。

图 9-24 锁定状态转换图



前提条件

已完成脚本开发任务。

解锁脚本

提交脚本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，建议您在提交脚本后通过“解锁”解除该脚本锁定。

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

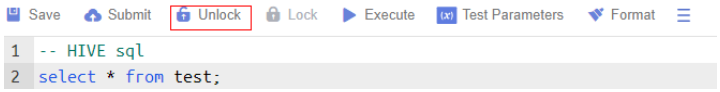
步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。

步骤4 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5 提交脚本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 9-25 解锁



----结束

9.3.6.8 转移脚本责任人

数据开发模块提供了转移脚本责任人的功能，您可以将责任人A的所有脚本一键转移到责任人B名下。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 在脚本目录顶部，单击 ，选择“责任人转移”。

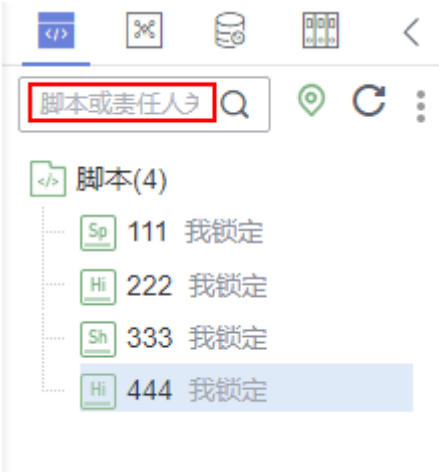
图 9-26 责任人转移

5. 分别设置“当前责任人”和“目标责任人”，单击“转移”。
6. 提示转移成功后，单击“关闭”。

相关操作

您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 9-27 根据脚本责任人筛选脚本



9.3.6.9 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 单击脚本目录中的，选择“显示复选框”。

图 9-28 显示脚本复选框

5. 勾选需要解锁的脚本，单击 > 批量解锁。弹出“解锁成功”提示。

图 9-29 批量解锁

9.4 作业开发

9.4.1 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，支持生成保存版本和提交版本。

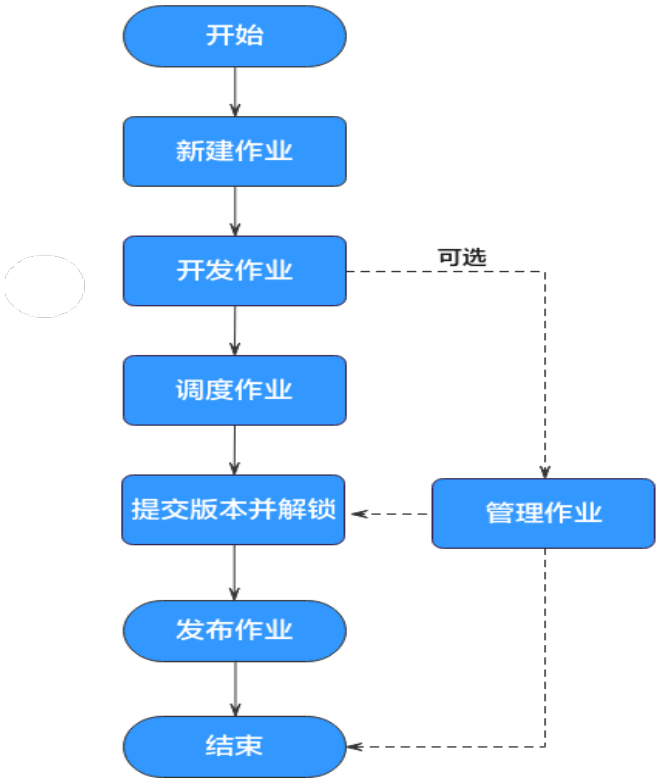
  说明

  保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。
- 支持单击右键，可快速复制作业名称，同时可以快速的关闭已打开的作业页签。
- 在MRS API连接模式下，单任务MRS Spark SQL和MRS Hive SQL运行完以后，在执行结果中查看运行日志，增加一键跳转MRS Yarn查看日志的链接。
- 企业模式下，开发作业时，单击页面上方的“前往发布”跳转到任务发布页面。
- 支持对“已提交”、“未提交”、“已调度”和“未调度”的作业进行筛选。同时未提交的作业通过红色进行标识，未调度的作业通过黄色进行标识。

- 单任务作业支持SQL编辑器风格配置。单击“风格配置”，可以对编辑器、操作栏、注释模板进行配置、以及查询SQL脚本编辑器可使用的快捷键。
- 单任务SQL查询结果展示支持表格和列表两种展示方式。单击“风格配置”，在“编辑器配置”里面可以对SQL查询结果展示进行配置。
- 作业开发支持细粒度权限管控，在数据安全模块对数据开发作业目录权限管控策略进行配置。
- 支持单击“基线链路”按钮，可以查看作业所属的基线链路信息。如果作业没有关联基线，则按钮置灰。

开发作业前，您可以通过**图9-30**了解数据开发模块作业开发的基本流程。

图 9-30 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，其中批处理作业还支持Pipeline和单节点作业两种模式，具体请参见**新建作业**。
2. 开发作业：基于新建的作业，进行作业开发，您可以进行编排、配置节点。具体请参见**开发Pipeline作业**。
3. 调度作业：配置作业调度任务。具体请参见**调度作业**。
   - 如果您的作业是批处理作业，您可以配置作业级别的调度任务，即以作业为一个整体进行调度，支持单次调度、周期调度、事件驱动调度三种调度方式。具体请参见**配置作业调度任务（批处理作业）**。
   - 如果您的作业是实时处理作业，您可以配置节点级别的调度任务，即每一个节点可以独立调度，支持单次调度、周期调度、事件驱动调度三种调度方式。具体请参见**配置节点调度任务（实时作业）**。
4. 提交版本并解锁：作业调度配置完成后，您需要提交版本并解锁，提交版本并解锁后才能用于调度运行，便于其他开发者修改。具体请参见**提交版本**。

5. （可选）管理作业：作业开发完成后，您可以根据需要，进行作业管理。具体请参见**（可选）管理作业**。
6. 发布作业。企业模式下需要发布作业，具体请参见**发布作业任务**。

9.4.2 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

作业在每个工作空间的最大配额为10000，作业目录最多5000个，目录层级最多为10层。请确保当前数量未达到最大配额。

新建普通目录

如果已存在可用的目录，则可以跳过当前操作。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中，右键单击目录名称，选择“新建目录”。
5. 在弹出的“新建目录”页面，配置如**表9-24**所示的参数。

   表 9-24 作业目录参数

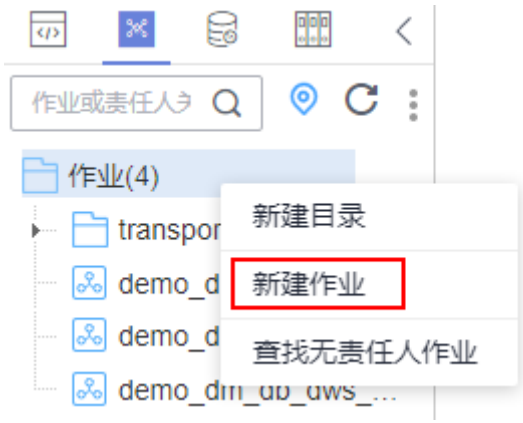
参数	说明
目录名称	作业目录的名称，只能包含英文字母、数字、中文字符、“_”、“-”，且长度为1~64个字符。
选择目录	选择该作业目录的父级目录，父级目录默认为根目录。

6. 单击“确定”，新建目录。

新建作业

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中，右键单击目录名称，选择“新建作业”。

图 9-31 新建作业



5. 在弹出的“新建作业”页面，配置如**表9-25**所示的参数。

表 9-25 作业参数

参数	说明
作业名称	自定义作业的名称，只能包含英文字母、数字、中文、“-”、“_”、“.”，且长度为1～128个字符。
作业类型	选择作业的类型。 • 批处理作业：按调度计划定期处理批量数据，主要用于实时性要求低的场景。批作业是由一个或多个节点组成的流水线，以流水线作为一个整体被调度。被调度触发后，任务执行一段时间必须结束，即任务不能无限时间持续运行。 批处理作业可以配置作业级别的调度任务，即以作业为一整体进行调度，具体请参见**配置作业调度任务（批处理作业）**。 • 实时处理作业：处理实时的连续数据，主要用于实时性要求高的场景。实时作业是由一个或多个节点组成的业务关系，每个节点可单独被配置调度策略，而且节点启动的任务可以永不下线。在实时作业里，带箭头的连线仅代表业务上的关系，而非任务执行流程，更不是数据流。 实时处理作业可以配置节点级别的调度任务，即每一个节点可以独立调度，具体请参见**配置节点调度任务（实时作业）**。
模式	• Pipeline：即传统的流水线式作业，作业通过画布编辑，可以拖入一个或多个节点组成作业，各节点依次被流水线式地执行。 **说明** 在企业模式下，实时处理作业类型不支持Pipeline模式，仅支持单任务模式。 • 单任务：单任务作业可以认为是有且只有一个节点的批处理作业，整个作业即为一个脚本节点。当前支持DLI SQL、DWS SQL、RDS SQL、MRS Hive SQL、MRS Spark SQL、DLI Spark、Flink SQL和Flink JAR类型的单任务作业，相比于先新建脚本再在作业中以节点引用脚本的开发方式，单任务作业可以直接在SQL编辑器中调测脚本并进行调度配置。 **说明** 单任务Flink SQL目前支持的MRS集群版本是MRS 3.2.0-LTS.1及以上版本。

参数	说明
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。 **说明** 作业优先级是作业的一个标签属性，不影响作业的实际调度执行的先后顺序。
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**配置公共委托**，则新建的作业默认使用该工作空间级委托。您也可参见**配置作业委托**，修改为作业级委托。 **说明** 作业级委托优先于工作空间级委托。
日志路径	选择作业日志的OBS存储路径。日志默认存储在以dlf-log-{Projectid}命名的桶中。 **说明** • 若您想自定义存储路径，请参见**（可选）修改作业日志存储路径**选择您已在OBS服务侧创建的桶。 • 请确保您已具备该参数所指定的OBS路径的读、写权限，否则系统将无法正常写日志或显示日志。
作业描述	作业的描述信息。

6. 单击“确定”，创建作业。

9.4.3 开发 Pipeline 作业

对已新建的作业进行开发和配置。

开发Pipeline模式的批处理作业和实时处理作业，请您参考**编排作业节点**、**配置作业基本信息**、**配置作业参数**和**调测并保存作业**章节。

前提条件

- 已创建作业，详情请参见**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

编排作业节点适用于Pipeline模式的批处理作业和实时处理作业。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 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。

5. 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 9-32 编排作业

6. 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9-26所示的功能。

   表 9-26 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。 **说明** 用户可以查看该作业节点的测试运行日志，单击“查看日志”可以进入查看日志界面查看日志的详细信息记录。
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，

功能	说明
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。
新建便签	为该节点添加便签，每个节点可以有多个便签。 在作业节点上新建便签、显示便签和隐藏便签只对该节点有效，在画布上方新建便签、显示便签和隐藏便签对整个作业有效。

7. （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。

     ```
     #{(Job.getNodeStatus("DQM")) == "success" ? "true" : "false"}
     ```


图 9-33 设置条件

关于EL表达式的语法，您可以查看**EL表达式参考**；关于IF条件的使用，您可以查看**IF条件判断教程**。

8. 配置节点属性。单击画布中的节点，在右侧显示“节点属性”页签，默认展开此配置页面，请参见**节点概述**配置具体节点的属性。

配置作业基本信息

为作业配置责任人、优先级信息后，用户可根据责任人、优先级来检索相应的作业。操作方法如下：

单击画布右侧“作业基本信息”页签，展开配置页面，配置如**表9-27**所示的参数。

表 9-27 作业基本信息

参数	说明
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	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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。 **说明** 配置执行用户调度功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。

参数	说明
作业委托	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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，作业不会中断，继续运行。
实例超时是否忽略等待时间	配置实例超时是否忽略等待时间。 如果勾选上，表示实例运行时等待时间不会被计入超时时间，可前往**默认项设置 > 实例超时是否忽略等待时间**修改此策略。 如果未选上，表示实例运行时等待时间会被计入超时时间。
自定义字段	配置自定义字段的参数名称和参数值。
作业标签	配置作业的标签，用以分类管理作业。 单击“新增”，可给作业重新添加一个标签。也可选择**管理作业标签**中已配置的标签。
作业描述	作业的描述信息。

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：

Pipeline模式的批处理作业和实时处理作业，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9-28**所示的参数。

表 9-28 作业参数配置

功能	说明
变量	
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}。 **说明** ● 如果作业中有两个节点，比如第一个Rest Client节点返回了body，第二个节点使用返回的data。如果这个data的长度超过1000000个字符，内容就会被截断。在配置作业参数时，作业的参数值的结果最大不超过1000000个字符。

功能	说明
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见**表达式概述**。
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}。
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见**表达式概述**。
修改	在参数名和参数值的文本框中直接修改，修改完成后，请保存。
删除	在参数值文本框后方，单击，删除作业常量。
工作空间环境变量	
查看工作空间已配置的变量和常量。	

单击“作业参数预览”页签，展开预览页面，配置如**表9-29**所示的参数。

说明

MRS Flink Job、DLI Flink Job、DLI SQL、DWS SQL、MRS HetuEngine、MRS ClickHouse SQL、MRS Hive SQL、MRS Impala SQL、MRS Presto SQL、MRS Spark SQL、RDS SQL、DORIS SQL的算子脚本参数支持参数预览。

表 9-29 作业参数预览

功能	说明
当前时间	仅单次调度才显示。系统默认为当前时间。

功能	说明
事件触发时间	仅事件驱动调度才显示。系统默认为事件触发时间。
周期调度	仅周期调度才显示。系统默认为调度周期。
具体时间	仅周期调度才显示。周期调度配置的具体运行时间。
起始日期	仅周期调度才显示。周期调度的生效时间。
后N个实例	作业运行调度的实例个数。 ● 单次调度场景默认为1。 ● 事件驱动调度场景默认为1。 ● 周期调度场景 当实例数大于10时，系统最多展示10个日期实例，系统会自动提示“当前参数预览最多支持查看10个实例”。

说明

在作业参数预览中，如果作业参数配置存在语法异常情况系统会给出提示信息。

如果参数配置了依赖作业实际运行时产生的数据，参数预览功能中无法模拟此类数据，则该数据不展示。

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，会弹出测试参数配置的弹框，自动显示作业的变量参数，单击“确定”，测试作业。如果测试未通过，请您查看作业节点的运行日志，进行定位处理。

说明

- 用户可以查看该作业的测试运行日志，单击“查看日志”可以进入查看日志界面查看日志的详细信息记录。
- 作业未提交版本之前，进行手动测试运行，作业监控里面的作业运行实例版本显示是0。
- 进行手动测试运行时，作业测试运行日志查看有权限管控，比如，用户A进行作业测试运行后，可以在“实例监控”页面查看测试运行日志，不允许用户B查看该测试运行日志。

步骤2 测试通过后，单击画布上方的“保存”，保存作业的配置信息。

保存后，在右侧的版本里面，会自动生成一个保存版本，支持版本回滚。保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。

----结束

实时处理作业

步骤1 单击画布上方的“保存”，保存作业的配置信息。

保存后，在右侧的版本里面，会自动生成一个保存版本，支持版本回滚。保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。

步骤2 **提交**作业版本后，单击画布上方的“启动”，运行作业。运行完以后，前往实时作业监控查看作业运行结果。

----结束

9.4.4 开发批处理单任务 SQL 作业

对已新建的作业进行开发和配置。

开发单任务模式的批处理作业，请您参考**开发SQL脚本**、**配置作业参数**、**质量监控**、**数据表**、**调测并保存作业**和**下载或转储脚本执行结果**章节。

前提条件

- 已创建作业，详情请参见**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**。

开发 SQL 脚本

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中，双击单任务模式作业名称，进入作业开发页面。
5. 在SQL编辑器右侧，单击“基本信息”，可以配置作业的基本信息、属性和高级信息等。单任务SQL作业的基本信息如**表9-30**所示，属性如**表9-31**所示，高级信息如**表9-32**所示。

表 9-30 作业基本信息

参数	说明
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	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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。 **说明** 配置执行用户调度功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
作业委托	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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。

参数	说明
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，作业不会中断，继续运行。
实例超时是否忽略等待时间	配置实例超时是否忽略等待时间。 如果勾选上，表示实例运行时等待时间不会被计入超时时间，可前往**默认项设置 > 实例超时是否忽略等待时间**修改此策略。 如果未选上，表示实例运行时等待时间会被计入超时时间。
自定义字段	配置自定义字段的参数名称和参数值。
作业标签	配置作业的标签，用以分类管理作业。 单击“新增”，可给作业重新添加一个标签。也可选择**管理作业标签**中已配置的标签。
作业描述	作业的描述信息。

表 9-31 批处理单任务 SQL 作业属性信息

属性	说明
DLI SQL属性	
DLI数据目录	选择DLI的数据目录。 • 在DLI默认的数据目录dli。 • 在DLI所绑定的LakeFormation已创建元数据catalog。
数据库名称	选择数据库。 DLI数据目录如果选择DLI默认的数据目录dli，表示为DLI的数据库和数据表。 DLI数据目录如果选择DLI所绑定的LakeFormation已创建元数据catalog，表示为LakeFormation的数据库和数据表。
队列名称	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。
是否记录脏数据	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

属性	说明
DLI环境变量	• 环境变量配置项需要以"hoodie."或"dli.sql."或"dli.ext."或"dli.jobs."或"spark.sql."或"spark.scheduler.pool"开头。 • 环境变量为dli.sql.autoBroadcastJoinThreshold时，值只能为整数，环境变量为dli.sql.shuffle.partitions时，值只能为正整数。 • 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。 **说明** 在非调度场景的DLI SQL脚本运行和DLI SQL单任务作业测试运行时，系统会默认开启以下四个配置参数： • spark.sql.adaptive.enabled（启用AQE，使Spark能够根据正在处理的数据的特征动态优化查询的执行计划，可以通过减少需要处理的数据量来提高性能。） • spark.sql.adaptive.join.enabled（启用AQE用于连接操作，可以通过根据正在处理的数据动态选择最佳连接算法来提高性能。） • spark.sql.adaptive.skewedJoin.enabled（启用AQE用于倾斜的连接操作，可以通过自动检测倾斜的数据并相应地优化连接算法来提高性能） • spark.sql.mergeSmallFiles.enabled（启用合并小文件功能，可以通过将小文件合并成较大的文件来提高性能，可以减少处理许多小文件的时间，并通过减少需要从远程存储中读取的文件数量来提高数据本地性。） 如果不使用的话，可以手动配置相关参数进行关闭，参数值设置为false。
DWS SQL属性	
数据连接	选择数据连接。
数据库	选择数据库。
脏数据表	SQL脚本中定义的脏数据表名称。 脏数据属性用户不能编辑，自动从SQL脚本内容中关联推荐。
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。
RDS SQL属性	
数据连接	选择数据连接。
数据库	选择数据库。
Spark SQL属性	

属性	说明
MRS作业名称	MRS的作业名称。系统同时支持按照**作业名称**自动填入。 未设置MRS作业名称且选择直连模式时，节点名称只能由英文字母、数字、中划线和下划线组成。最大只能输入64个字符，不能包含中文字符。 **说明** 如果选择MRS API连接方式的数据连接，不支持设置作业名称。
数据连接	选择数据连接。
MRS资源队列	选择已创建好的MRS资源队列。 **说明** 您需要先在数据安全服务队列权限功能中，配置对应的队列后，才能在此处选择到已配置的队列。当有多处同时配置了资源队列时，此处配置的资源队列为最高优先级。
数据库	选择数据库。MRS API连接方式下不支持选择数据库。
运行程序参数	配置运行参数。 举例如下： 参数配置为**--queue**，参数值配置为**default_cr**，该示例表示配置了MRS集群的指定队列。同时在MRS集群的作业管理下，在操作的“更多 > 查看详情”里面可以查看该作业的详细信息。 **说明** 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。Spark代理连接不支持该配置。 在MRS API连接模式下，单算子作业Spark SQL支持程序运行参数。
Hive SQL属性	
MRS作业名称	MRS的作业名称。系统同时支持按照**作业名称**自动填入。 未设置MRS作业名称且选择直连模式时，节点名称只能由英文字母、数字、中划线和下划线组成。最大只能输入64个字符，不能包含中文字符。
数据连接	选择数据连接。
数据库	选择数据库。
MRS资源队列	选择已创建好的MRS资源队列。

属性	说明
运行程序参数	配置运行参数。 举例如下： 参数配置为**--hiveconf**，参数值配置为**mapreduce.job.queuename=default_cr**，该示例表示配置了MRS集群的指定队列。同时在MRS集群的作业管理下，在操作的“更多 > 查看详情”里面可以查看该作业的详细信息。 **说明** 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。Hive代理连接不支持该配置。 在MRS API连接模式下，单算子作业Hive SQL支持程序运行参数。
Doris SQL属性	
数据连接	选择数据连接。
数据库	选择数据库。

表 9-32 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。 节点运行过程中，根据设置的节点状态轮询时间查询节点是否执行完成。
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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	当前节点执行失败后，后续节点的处理策略： ● 终止当前作业执行计划：终止当前作业运行，当前作业实例状态显示为“失败”。如果是周期调度作业，后续周期调度会正常运行。 ● 忽略失败，作业结果设为成功：忽略当前节点失败，当前作业实例状态显示为“运行成功”。如果是周期调度作业，后续周期调度会正常运行。

6. 在SQL编辑器中输入SQL语句，支持输入多条SQL语句。

   说明

   - 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.
     ```
   - RDS SQL当前不支持begin ... commit事务语法，若有需要，请使用start transaction ... commit事务语法。
   - 脚本内容大小不能超过16MB。
   - 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。
   - 当前用户提交Spark SQL脚本到MRS时，默认提交至其绑定的租户队列（绑定队列即用户绑定的租户类型角色所对应的队列）中运行。当绑定多个队列时，系统会优先根据内部排序选择队列进行提交；如果需要给该用户使用固定一个队列进行提交， 可以登录FusionInsight Manager界面，在"租户资源 > 动态资源计划 > 全局用户策略"中给该用户配置默认队列，详细操作请参见**管理全局用户策略**。
   - Spark SQL、Hive SQL脚本支持语法检查。单击“语法检查”，SQL语句校验完成后，可以在下方查看语法校验结果。

   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。

     - F8：运行
     - F9：停止
     - Ctrl + /：注释或解除注释光标所在行或代码块
     - Ctrl + Z：撤销
     - Ctrl + F：查找
     - Ctrl + Shift + R：替换
     - Ctrl + X：剪切
     - Ctrl + S：保存
     - 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
- Ctrl + D：删除一行
- Shift + Ctrl + U：解锁
- Ctrl + Alt + K：同词选择
- Ctrl + B：格式化
- Ctrl + Shift + Z：重做
- Ctrl + Enter：执行所选行/选中内容
- Ctrl + Alt + F：标记
- Ctrl + Shift + K：查找上一个
- Ctrl + K：查找下一个
- Ctrl + Backspace：删除左侧单词
- Ctrl + Delete：删除右侧单词
- Alt + Backspace：删除至行首
- Alt + Delete：删除至行尾
- Alt + Shift-Left：选择行首
- Alt + Shift-Right：选择行尾

– 支持系统函数功能。

  单击编辑器右侧的“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。

– 支持脚本参数。

  在SQL语句中直接写入脚本参数，然后在编辑器右侧的“参数”处选择“更新脚本参数”。也可以直接配置该作业脚本的参数与常量。

  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;
  ```

– 支持可视化读取数据表生成SQL语句功能。

  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，在右下角单击“生成SQL语句”，生成的SQL语句需要您手动格式化。

7. （可选）在编辑器上方，单击“格式化”，格式化SQL语句。
8. 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。

   说明

   - 您可以单击“查看日志”，进入查看日志界面查看日志的详细信息记录。
   - 脚本执行历史结果可以进行权限管控，可设置为“仅自己可见”或“所有用户可见”，默认配置项请参见**脚本执行历史展示**。
9. 在编辑器上方，单击“保存”，保存该作业。

配置作业参数

单击编辑器右侧的“参数”，展开配置页面，配置如**表9-33**所示的参数。

表 9-33 作业参数配置

功能	说明
变量	
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}。
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见**表达式概述**。
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	

功能	说明
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}。
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见**表达式概述**。
修改	在参数名和参数值的文本框中直接修改，修改完成后，请保存。
删除	在参数值文本框后方，单击，删除作业常量。
工作空间环境变量	
查看工作空间已配置的变量和常量。	

单击“作业参数预览”页签，展开预览页面，配置如**表9-34**所示的参数。

表 9-34 作业参数预览

功能	说明
当前时间	仅单次调度才显示。系统默认为当前时间。
事件触发时间	仅事件驱动调度才显示。系统默认为事件触发时间。
周期调度	仅周期调度才显示。系统默认为调度周期。
具体时间	仅周期调度才显示。周期调度配置的具体运行时间。
起始日期	仅周期调度才显示。周期调度的生效时间。
后N个实例	作业运行调度的实例个数。 ● 单次调度场景默认为1。 ● 事件驱动调度场景默认为1。 ● 周期调度场景 当实例数大于10时，系统最多展示10个日期实例，系统会自动提示“当前参数预览最多支持查看10个实例”。

说明

在作业参数预览中，如果作业参数配置存在语法异常情况系统会给出提示信息。

如果参数配置了依赖作业实际运行时产生的数据，参数预览功能中无法模拟此类数据，则该数据不展示。

质量监控

对已编排好的单任务作业关联质量作业，当前暂不支持集成作业和单任务的实时作业。

质量监控支持并行和串行两种方式。单击画布右侧“质量监控”页签，展开配置页面，配置如表9-35所示的参数。

表 9-35 质量监控配置

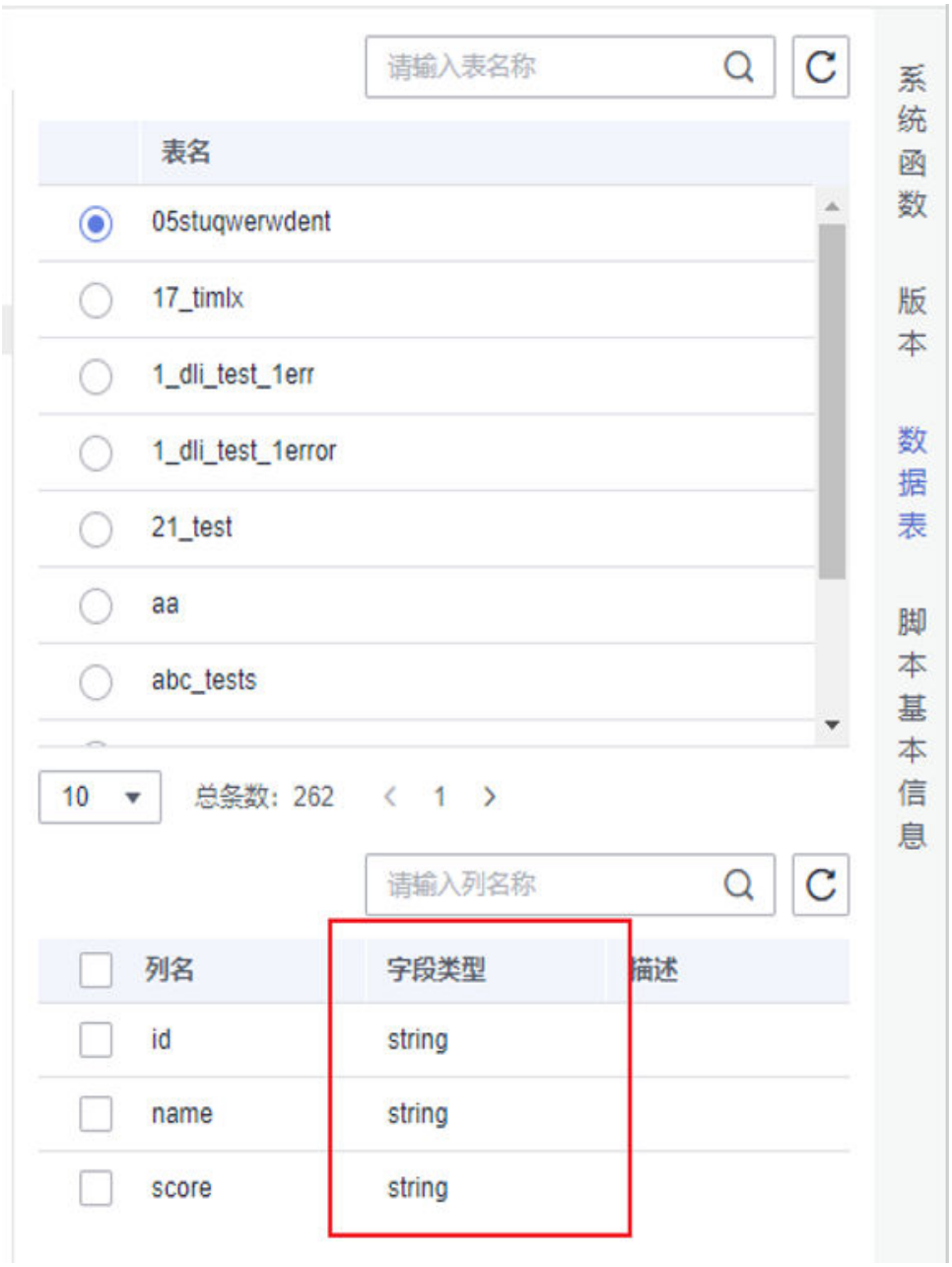
参数	说明
执行方式	选择质量监控的执行方式： • 并行：并行模式下，所有质量作业算子的上游都被设置为主算子。 • 串行：串行模式下，质量作业将依照配置面板由上至下的顺序依次串联，顶部的质量作业依赖于主算子。

参数	说明
质量作业	关联质量作业。 1. 单击“新增”，右侧自动弹出Data Quality Monitor算子的页面。 2. 节点名称可自定义。 3. DQC作业类型选择“质量作业”。 **说明** 对账作业目前不支持。 4. 选择需要关联的“质量作业名称”，其他参数根据实际业务需要配置。如果没有质量作业，请参考**新建数据质量作业**创建一个质量作业。 **说明** • 单击“新增”可以关联多个质量作业。 • 单击可以修改已关联的质量作业。 • 单击可以删除已关联的质量作业。 5. 是否忽略质量作业告警 是：质量作业告警可以忽略 否：质量作业告警不可忽略，产生告警时，上报告警。 6. 配置高级参数。 a. 配置节点执行的最长时间。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，该节点将会再次重试。 b. 失败重试。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。 超时重试 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。 c. 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”。 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”。 7. 单击“确定”，质量监控配置完成。

数据表

支持Hive SQL、Spark SQL、DLI SQL、DWS SQL、Doris SQL和RDS SQL单任务批处理作业可以查看右侧的数据表，单击表名前面的单选框，可以查看该数据的列名、字段类型和描述。

图 9-34 查看数据表



调测并保存作业

作业配置完成后，请执行以下操作：

步骤1 单击画布上方的运行按钮 ▶，运行作业。

说明

用户可以查看该作业的运行日志，单击“查看日志”可以进入查看日志界面查看日志的详细信息记录。

步骤2 运行完成后，单击画布上方的保存按钮，保存作业的配置信息。

保存后，在右侧的版本里面，会自动生成一个保存版本，支持版本回滚。保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。

----结束

下载或转储脚本执行结果

脚本运行成功后，支持下载和转储SQL脚本执行结果。系统默认支持所有用户都能下载和转储SQL脚本的执行结果。如果您不希望所有用户都有该操作权限，可参考**配置数据导出策略**进行配置。

- 脚本执行完成后在“执行结果”中，单击“下载”可以直接下载CSV格式的结果文件到本地。可以在**下载中心**查看下载记录。
- 脚本执行完成后在“执行结果”中，单击“转储”可以将脚本执行结果转储为CSV和JSON格式的结果文件到OBS中，详情请参见**表9-36**。

说明

- 转储功能依赖于OBS服务，如无OBS服务，则不支持该功能。
- 当前仅支持转储SQL脚本查询（query）类语句的结果。
- DataArts Studio的下载或转储的SQL结果中，如果存在英文逗号、换行符等这种特殊符号，可能会导致数据错乱、行数变多等的问题。

表 9-36 转储配置

参数	是否必选	说明
数据格式	是	目前支持导出CSV和JSON格式的结果文件。
资源队列	否	选择执行导出操作的DLI队列。当脚本为DLI SQL时，配置该参数。
压缩格式	否	选择压缩格式。当脚本为DLI SQL时，配置该参数。 • none • bzip2 • deflate • gzip
存储路径	是	设置结果文件的OBS存储路径。选择OBS路径后，您需要在选择的路径后方自定义一个文件夹名称，系统将在OBS路径下创建文件夹，用于存放结果文件。 您也可以到**下载中心**配置默认的OBS路径地址，配置好后在转储时会默认填写。

参数	是否必选	说明
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。
是否导出列名	否	是：导出列名 否：不导出列名
字符集	否	● UTF-8：默认字符集。 ● GB2312：当导出数据中包含中文字符集时，推荐使用此字符集。 ● GBK：国家标准GB2312基础上扩容后兼容GB2312的标准。
引用字符	否	仅在数据格式为csv格式时支持配置引用字符。 引用字符在导出作业结果时用于标识文本字段的开始和结束，即用于分割字段。 仅支持设置一个字符。默认值是英文双引号（"）。 主要用于处理包含空格、特殊字符或与分隔符相同字符的数据。 关于“引用字符”和“转义字符”的使用示例请参考**引用字符和转义字符使用示例**。
转义字符	否	仅在数据格式为csv格式时支持配置转义字符。 在导出结果中如果需要包含特殊字符，如引号本身，可以使用转义字符（反斜杠 \ ）来表示。 仅支持设置一个字符。默认值是英文反斜杠（\）。 常用转义字符的场景： ● 假设两个引用字符之间的数据内容存在第三个引用字符，则在第三个引用字符前加上转义字符，从而避免字段内容被分割。 ● 假设数据内容中原本就存在转义字符，则在这个原有的转义字符前再加一个转义字符，避免原来的那个字符起到转义作用。 关于“引用字符”和“转义字符”的使用示例请参考**引用字符和转义字符使用示例**。

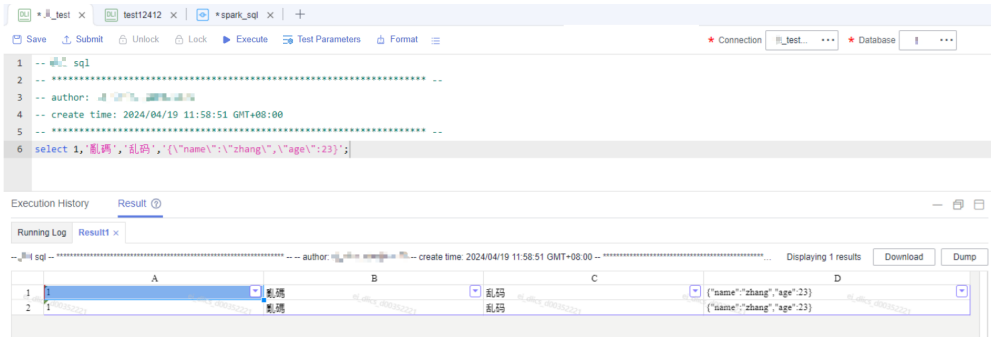
相对于直接查看SQL脚本的执行结果，通过下载和转储能够支持获取更多的执行结果。各类SQL脚本查看、下载、转储支持的规格如**表9-37**所示。

表 9-37 SQL 脚本支持查看/下载/转储规格

SQL类型	在线查看最大结果条数	下载最大结果	转储最大结果
DLI	1000	1000条且少于3MB	无限制
Hive	1000	1000条且少于3MB	10000条或3MB
DWS	1000	1000条且少于3MB	10000条或3MB
Spark	1000	1000条且少于3MB	10000条或3MB
RDS	1000	1000条且少于3MB	不支持
Doris	1000	1000条且少于3MB	1000条或3MB

引用字符和转义字符使用示例

- 引用字符和转义字符使用说明：
  - 引用字符：用于识别分割字段，默认值：英文双引号（"）。
  - 转义字符：在导出结果中如果需要包含特殊字符，如引号本身，可以使用转义字符（反斜杠 \ ）来表示。默认值：英文反斜杠（\）。
    i. 假设两个quote_char之间的数据内容存在第三个quote_char，则在第三个quote_char前加上escape_char，从而避免字段内容被分割。
    ii. 假设数据内容中原本就存在escape_char，则在这个原有的escape_char前再加一个escape_char，避免原来的那个字符起到转义作用。
- 应用示例：



在进行转储时，如果引用字符和转义字符不填，如下图所示。

下载的.csv用excel打开以后如下图所示，是分成两行的。

D	E
{\name\":\"zhang\"	\"age\":23}"
{\name\":\"zhang\"	\"age\":23}"

在转储时，如果引用字符和转义字符都填写，比如，引用字符和转义字符都填英文双引号（"），则下载以后查看结果如下图所示。

D	E
{"name":"zhang","age":23}	
{"name":"zhang","age":23}	

9.4.5 开发实时处理单任务 MRS Flink SQL 作业

对已新建的作业进行开发和配置。

开发单任务模式的实时处理Flink SQL作业，请您参考**开发SQL脚本**、**配置作业参数**、**保存作业**和**模板**章节。

前提条件

- 已**新建作业**。
- 当前用户已锁定该作业，否则需要通过“抢锁”锁定作业后才能继续开发作业。新建或导入作业后默认被当前用户锁定，详情参见**编辑锁定功能**。

开发 SQL 脚本

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中，双击单任务模式作业名称，进入作业开发页面。
5. 在SQL编辑器右侧，单击“基本信息”，可以配置作业的基本信息等。单任务MRS Flink SQL作业的基本信息如**表9-38**所示。

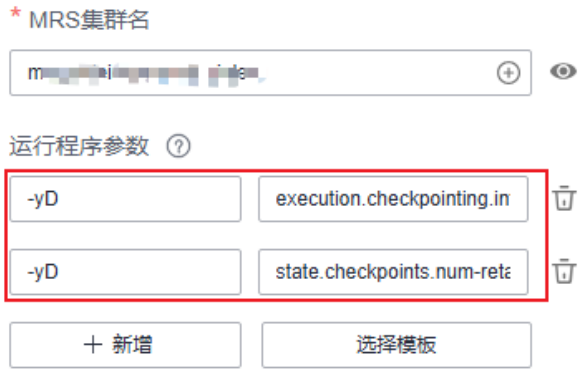
表 9-38 作业基本信息

参数	说明
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	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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。 **说明** 配置执行用户调度功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
作业委托	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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，作业不会中断，继续运行。
实例超时是否忽略等待时间	配置实例超时是否忽略等待时间。 如果勾选上，表示实例运行时等待时间不会被计入超时时间，可前往**默认项设置 > 实例超时是否忽略等待时间**修改此策略。 如果未选上，表示实例运行时等待时间会被计入超时时间。
自定义字段	配置自定义字段的参数名称和参数值。

参数	说明
作业标签	配置作业的标签，用以分类管理作业。 单击“新增”，可给作业重新添加一个标签。也可选择**管理作业标签**中已配置的标签。
作业描述	作业的描述信息。

表 9-39 实时处理单任务 MRS Flink SQL 作业属性信息

属性	说明
Flink SQL属性	
Flink作业名称	输入Flink作业名称。 系统支持Flink作业名称按照**工作空间-作业名称**格式自动填入。 **说明** 只能包含英文字母、数字、中划线和下划线。最大只能输入64个字符，不能包含中文字符。
MRS集群名	选择MRS集群名称。 **说明** 单任务Flink SQL目前支持的MRS集群版本是MRS 3.2.0-LTS.1及以上版本。

属性	说明
运行程序参数	配置作业运行参数。当选择了MRS集群名后，该参数才显示。 该参数为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 **注意** 系统支持实时Flink SQL作业运行前能够查询历史checkpoint，并选择从指定checkpoint启动。要使Flink Checkpoin生效，需要配置两个运行参数： **图 9-35** 配置运行程序参数  • 用来控制checkpoint间隔 **-yD：execution.checkpointing.interval=1000** • 用来控制保留的checkpoint数量 **-yD：state.checkpoints.num-retained=10** 查询checkpoint列表时，配置-s参数，鼠标单击参数值输入框，checkpoint列表参数值会自动弹出。 **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。 单击“选择模板”，选择已创建好的脚本模板，系统支持可以引用多个模板。创建模板的详细操作请参见**配置模板**。 MRS Flink作业的运行程序参数，请参见《MapReduce用户指南》中的**运行Flink作业**。
Flink作业执行参数	配置Flink作业执行参数。 Flink程序执行的关键参数，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。
MRS资源队列	选择已创建好的MRS资源队列。 需要先在数据安全服务队列权限功能中，配置对应的队列后，才能在此处选择到已配置的队列。当有多处同时配置了资源队列时，此处配置的资源队列为最高优先级。
重跑策略	• 从上一个检查点重跑 • 重新启动
输入数据路径	设置输入数据路径，系统支持从HDFS或OBS的目录路径进行配置。

属性	说明
输出数据路径	设置输出数据路径，系统支持从HDFS或OBS的目录路径进行配置。

表 9-40 高级参数

参数	是否必选	说明
作业状态轮询时间（秒）	是	设置轮询时间（30~60秒、120秒、180秒、240秒、300秒），每隔x秒查询一次作业是否执行完成。 作业运行过程中，根据设置的作业状态轮询时间查询作业运行状态。
最长等待时间	是	设置作业执行的超时时间，如果作业配置了重试，在超时时间内未执行完成，该作业将会再次重试。 **说明** 如果作业一直处于启动中状态，没有成功开始运行，超时后作业会被置为失败。
失败重试	是	作业执行失败后，是否重新执行作业。 ● 是：重新执行作业，请配置以下参数。 – 超时重试 – 最大重试次数 – 重试间隔时间（秒） ● 否：默认值，不重新执行作业。 **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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

6. 在SQL编辑器中输入SQL语句，支持输入多条SQL语句。

   说明

   - 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.
     ```
   - 脚本内容大小不能超过16MB。
   - 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。
   - Flink SQL作业支持语法检查。在编辑器上方，单击“语法检查”，可以对SQL语句进行语义校验。SQL语句校验完成后，可以在下方查看语法校验结果。
   - 脚本执行历史结果可以进行权限管控，可设置为“仅自己可见”或“所有用户可见”，默认配置项请参见**脚本执行历史展示**。

   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。
  - F8：运行
  - F9：停止
  - Ctrl + /：注释或解除注释光标所在行或代码块
  - Ctrl + Z：撤销
  - Ctrl + F：查找
  - Ctrl + Shift + R：替换
  - Ctrl + X：剪切
  - Ctrl + S：保存
  - 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
  - Ctrl + D：删除一行
  - Shift + Ctrl + U：解锁
  - Ctrl + Alt + K：同词选择
  - Ctrl + B：格式化
  - Ctrl + Shift + Z：重做
  - Ctrl + Enter：执行所选行/选中内容
  - Ctrl + Alt + F：标记
  - Ctrl + Shift + K：查找上一个
  - Ctrl + K：查找下一个
  - Ctrl + Backspace：删除左侧单词
  - Ctrl + Delete：删除右侧单词
  - Alt + Backspace：删除至行首

- Alt + Delete：删除至行尾
- Alt + Shift-Left：选择行首
- Alt + Shift-Right：选择行尾

– 支持脚本参数。

在SQL语句中直接写入脚本参数，然后在编辑器右侧的“参数”处选择“更新脚本参数”。也可以直接配置该作业脚本的参数与常量。

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;
```

7. （可选）在编辑器上方，单击“格式化”，格式化SQL语句。
8. 在编辑器上方，单击“保存”按钮，保存该作业并进行提交。

配置作业参数

单击编辑器右侧的“参数”，展开配置页面，配置如表9-41所示的参数。

表 9-41 作业参数配置

功能	说明
变量	
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}
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见表达式概述。
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	

功能	说明
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}
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见**表达式概述**。
修改	在参数名和参数值的文本框中直接修改，修改完成后，请保存。
删除	在参数值文本框后方，单击，删除作业常量。
工作空间环境变量	
查看工作空间已配置的变量和常量。	

单击“作业参数预览”页签，展开预览页面，配置如**表9-42**所示的参数。

表 9-42 作业参数预览

功能	说明
当前时间	仅单次调度才显示。系统默认为当前时间。
事件触发时间	仅事件驱动调度才显示。系统默认为事件触发时间。
周期调度	仅周期调度才显示。系统默认为调度周期。
具体时间	仅周期调度才显示。周期调度配置的具体运行时间。
起始日期	仅周期调度才显示。周期调度的生效时间。
后N个实例	作业运行调度的实例个数。 ● 单次调度场景默认为1。 ● 事件驱动调度场景默认为1。 ● 周期调度场景 当实例数大于10时，系统最多展示10个日期实例，系统会自动提示“当前参数预览最多支持查看10个实例”。

说明

在作业参数预览中，如果作业参数配置存在语法异常情况系统会给出提示信息。

如果参数配置了依赖作业实际运行时产生的数据，参数预览功能中无法模拟此类数据，则该数据不展示。

保存作业

作业配置完成后，请执行以下操作：

步骤1 单击画布上方的“启动”，运行作业。

说明

执行结果最多显示1000条数据；执行结果的大小不超过3MB，若超过3MB结果会被截断。

步骤2 运行完成后，单击画布上方的保存按钮，保存作业的配置信息。

保存后，在右侧的版本里面，会自动生成一个保存版本，支持版本回滚。保存版本时，一分钟内多次保存只记录一次版本。对于中间数据比较重要时，可以通过“新增版本”按钮手动增加保存版本。

----结束

模板

在开发Flink SQL单任务实时处理作业时，系统支持可以引用公共脚本模板。创建模板的详细操作请参见**配置模板**，脚本模板的使用场景指导请参见**引用脚本模板和参数模板的使用介绍**。

9.4.6 开发实时处理单任务 MRS Flink Jar 作业

前提条件

参见**新建作业**创建一个实时处理的单任务Flink Jar作业。

配置 MRS Flink Jar 作业

表 9-43 配置 MRS Flink Jar 作业属性参数

参数	是否必选	说明
Flink作业名称	是	输入Flink作业名称。 系统支持Flink作业名称按照**工作空间-作业名称**格式自动填入。 作业名称只能包含英文字母、数字、中划线和下划线，且长度为1~64个字符，不能包含中文字符。
MRS集群名	是	选择MRS集群名称。 **说明** 单任务Flink Jar目前支持的MRS集群版本是MRS 3.2.0-LTS.1及以上版本。

参数	是否必选	说明
运行程序参数	否	配置作业运行参数。当选择了MRS集群名后，该参数才显示。 该参数为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 **注意** 系统支持Flink Jar作业运行前能够查询历史checkpoint，并选择从指定checkpoint启动。要使Flink Checkpoin生效，需要配置两个运行参数： • 用来控制checkpoint间隔 **-yD：execution.checkpointing.interval=1000** • 用来控制保留的checkpoint数量 **-yD：state.checkpoints.num-retained=10** 查询checkpoint列表时，配置-s参数，鼠标单击参数值输入框，checkpoint列表参数值会自动弹出。 **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。 单击“选择模板”，选择已创建好的脚本模板，系统支持可以引用多个模板。创建模板的详细操作请参见**配置模板**。 MRS Flink作业的运行程序参数，请参见《MapReduce用户指南》中的**运行Flink作业**。
Flink作业执行参数	否	配置Flink作业执行参数。 Flink程序执行的关键参数，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。
MRS资源队列	否	选择已创建好的MRS资源队列。 需要先在数据安全服务队列权限功能中，配置对应的队列后，才能在此处选择到已配置的队列。当有多处同时配置了资源队列时，此处配置的资源队列为最高优先级。
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**。
重跑策略	否	• 从上一个检查点重跑 • 重新启动
输入数据路径	否	设置输入数据路径，系统支持从HDFS或OBS的目录路径进行配置。
输出数据路径	否	设置输出数据路径，系统支持从HDFS或OBS的目录路径进行配置。

表 9-44 配置高级参数

参数	是否必选	说明
作业状态轮询时间（秒）	是	设置轮询时间（30~60秒、120秒、180秒、240秒、300秒），每隔x秒查询一次作业是否执行完成。 作业运行过程中，根据设置的作业状态轮询时间查询作业运行状态。
最长等待时间	是	设置作业执行的超时时间，如果作业配置了重试，在超时时间内未执行完成，该作业将会再次重试。 **说明** 如果作业一直处于启动中状态，没有成功开始运行，超时后作业会被置为失败。
失败重试	否	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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数设置完成后，单击“保存”，并提交该作业。

单击“启动”，运行该作业。

配置作业基本信息

表 9-45 作业基本信息

参数	说明
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	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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。 **说明** 配置执行用户调度功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
作业委托	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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。

参数	说明
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，作业不会中断，继续运行。
实例超时是否忽略等待时间	配置实例超时是否忽略等待时间。 如果勾选上，表示实例运行时等待时间不会被计入超时时间，可前往**默认项设置 > 实例超时是否忽略等待时间**修改此策略。 如果未选上，表示实例运行时等待时间会被计入超时时间。
自定义字段	配置自定义字段的参数名称和参数值。
作业标签	配置作业的标签，用以分类管理作业。 单击“新增”，可给作业重新添加一个标签。也可选择**管理作业标签**中已配置的标签。
作业描述	作业的描述信息。

配置作业参数

单击编辑器右侧的“参数”，展开配置页面，配置如**表9-46**所示的参数。

表 9-46 作业参数配置

功能	说明
变量	
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}
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见**表达式概述**。
修改	在参数名和参数值的文本框中直接修改。
掩码显示	在参数值为密钥等情况下，从安全角度，请单击  将参数值掩码显示。
删除	在参数值文本框后方，单击，删除作业参数。

功能	说明
常量	
新增	单击“新增”，在文本框中填写作业常量的名称和参数值。 ● 参数名称 名称只能包含字符：英文字母、数字、中划线和下划线。 ● 参数值 – 字符串类的参数直接填写字符串，例如：str1 – 数值类的参数直接填写数值或运算表达式。 参数配置完成后，在作业中的引用格式为：${参数名称}
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见**表达式概述**。
修改	在参数名和参数值的文本框中直接修改，修改完成后，请保存。
删除	在参数值文本框后方，单击，删除作业常量。
工作空间环境变量	
查看工作空间已配置的变量和常量。	

单击“作业参数预览”页签，展开预览页面，配置如**表9-47**所示的参数。

表 9-47 作业参数预览

功能	说明
当前时间	仅单次调度才显示。系统默认为当前时间。
事件触发时间	仅事件驱动调度才显示。系统默认为事件触发时间。
周期调度	仅周期调度才显示。系统默认为调度周期。
具体时间	仅周期调度才显示。周期调度配置的具体运行时间。
起始日期	仅周期调度才显示。周期调度的生效时间。
后N个实例	作业运行调度的实例个数。 ● 单次调度场景默认为1。 ● 事件驱动调度场景默认为1。 ● 周期调度场景 当实例数大于10时，系统最多展示10个日期实例，系统会自动提示“当前参数预览最多支持查看10个实例”。

说明

在作业参数预览中，如果作业参数配置存在语法异常情况系统会给出提示信息。

如果参数配置了依赖作业实际运行时产生的数据，参数预览功能中无法模拟此类数据，则该数据不展示。

9.4.7 开发实时处理单任务 DLI Spark 作业

前提条件

参见**新建作业**创建一个实时处理的单任务DLI Spark作业。

配置 DLI Spark 作业

表 9-48 配置属性参数

参数	是否必选	说明
作业名称	是	输入DLI Spark作业名称。 作业名称只能包含英文字母、数字、下划线和中划线，且长度为1~64个字符。
DLI队列	是	选择DLI队列。
Spark版本	否	• 2.3.2 • 2.4.5 • 3.1.1
作业特性	否	用户作业使用的Spark镜像类型（当前支持基础型、AI增强型和自定义的Spark镜像）。 • 基础型 • AI增强型 • 自定义镜像 当选择“自定义镜像”时，请选择自定义的镜像名称，版本号系统自动展示。您可以前往**容器镜像服务**进行设置。
作业运行资源	否	• 8核32G内存 • 16核64G内存 • 32核128G内存
作业主类	否	该参数表示作业的Java/Scala主类。
Spark程序资源包	是	该参数表示Spark程序依赖的资源包。

参数	是否必选	说明
资源类型	是	• OBS路径 • DLI程序包 DLI程序包：作业执行前，会将资源包文件上传到DLI资源管理。 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。
分组设置	否	当“资源类型”选择“DLI程序包”时，才需要配置该参数。 将Spark程序资源包上传到指定的分组中，主Jar包和依赖包会上传到同一个分组中。 • 已有分组：选择已有的分组 • 创建新分组：创建新的分组，分组名称只能包含英文字母、数字、点号、中划线和下划线。 • 不分组
主类入口参数	否	配置该参数时，多个参数请以Enter键进行分隔。
Spark作业运行参数	否	配置该参数时，输入格式为key=value的参数，多个参数请以Enter键进行分隔。
Module名称	否	选择Module名称，支持选择多个。
访问元数据	否	访问元数据的开关。 如果需要在DLI Spark作业中访问由DLI SQL作业创建的OBS表，就要打开访问元数据开关。

表 9-49 配置高级参数

参数	是否必选	说明
作业状态轮询时间（秒）	是	设置轮询时间（30~60秒、120秒、180秒、240秒、300秒），每隔x秒查询一次作业是否执行完成。 作业运行过程中，根据设置的作业状态轮询时间查询作业运行状态。
最长等待时间	是	设置作业执行的超时时间，如果作业配置了重试，在超时时间内未执行完成，该作业将会再次重试。 **说明** 如果作业一直处于启动中状态，没有成功开始运行，超时后作业会被置为失败。

参数	是否必选	说明
失败重试	否	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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数设置完成后，单击“保存”，并提交该作业。

单击“启动”，运行该作业。

配置作业基本信息

表 9-50 作业基本信息

参数	说明
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	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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。 **说明** 配置执行用户调度功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。
作业委托	当“作业调度身份是否可配置”设置为“是”，该参数可见。 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，作业不会中断，继续运行。
实例超时是否忽略等待时间	配置实例超时是否忽略等待时间。 如果勾选上，表示实例运行时等待时间不会被计入超时时间，可前往**默认项设置 > 实例超时是否忽略等待时间**修改此策略。 如果未选上，表示实例运行时等待时间会被计入超时时间。

参数	说明
自定义字段	配置自定义字段的参数名称和参数值。
作业标签	配置作业的标签，用以分类管理作业。 单击“新增”，可给作业重新添加一个标签。也可选择**管理作业标签**中已配置的标签。
作业描述	作业的描述信息。

配置作业参数

单击编辑器右侧的“参数”，展开配置页面，配置如**表9-51**所示的参数。

表 9-51 作业参数配置

功能	说明
变量	
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}。
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见**表达式概述**。
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	

功能	说明
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}。
编辑参数表达式	在参数值文本框后方，单击  ，编辑参数表达式，更多表达式请参见**表达式概述**。
修改	在参数名和参数值的文本框中直接修改，修改完成后，请保存。
删除	在参数值文本框后方，单击，删除作业常量。
工作空间环境变量	
查看工作空间已配置的变量和常量。	

单击“作业参数预览”页签，展开预览页面，配置如**表9-52**所示的参数。

表 9-52 作业参数预览

功能	说明
当前时间	仅单次调度才显示。系统默认为当前时间。
事件触发时间	仅事件驱动调度才显示。系统默认为事件触发时间。
周期调度	仅周期调度才显示。系统默认为调度周期。
具体时间	仅周期调度才显示。周期调度配置的具体运行时间。
起始日期	仅周期调度才显示。周期调度的生效时间。
后N个实例	作业运行调度的实例个数。 ● 单次调度场景默认为1。 ● 事件驱动调度场景默认为1。 ● 周期调度场景 当实例数大于10时，系统最多展示10个日期实例，系统会自动提示“当前参数预览最多支持查看10个实例”。

说明

在作业参数预览中，如果作业参数配置存在语法异常情况系统会给出提示信息。

如果参数配置了依赖作业实际运行时产生的数据，参数预览功能中无法模拟此类数据，则该数据不展示。

9.4.8 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

- 已完成**开发Pipeline作业**或**开发批处理单任务SQL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9-53**所示的参数。

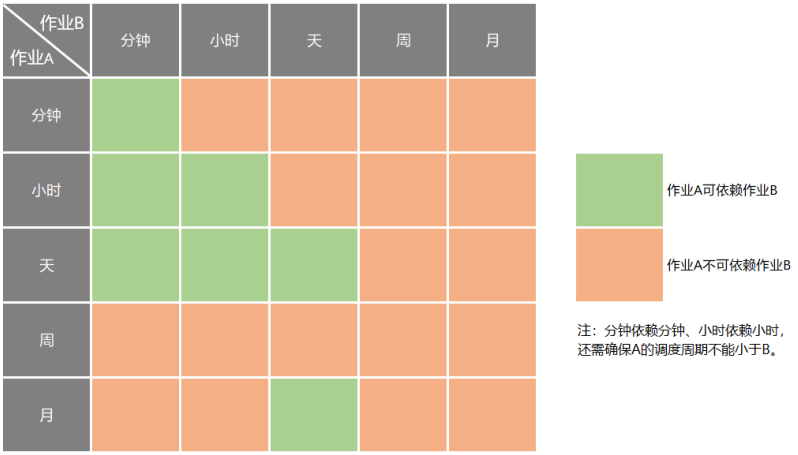
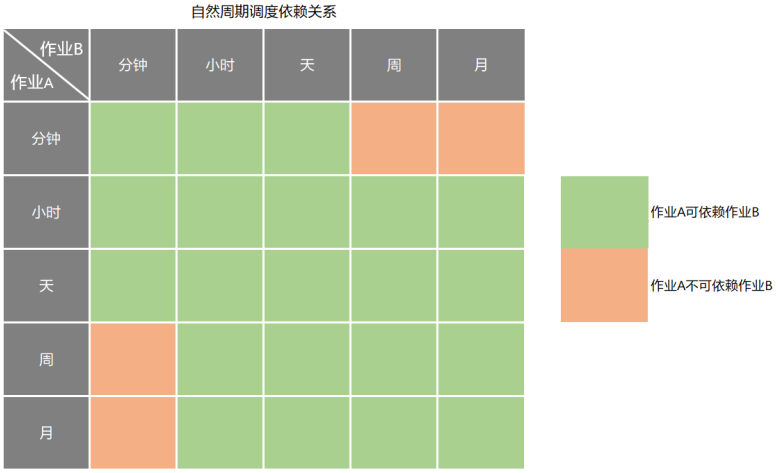
表 9-53 作业调度配置

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9-54**。 – 需要人工确认才执行：勾选后，需要人工确认执行后，作业实例才能够运行。如果不进行人工确认，影响后续作业实例执行。 **说明** **作业实例运行场景下**，在实例监控页面，作业实例运行状态显示为“待确认执行”，可以进行手动确认执行，单击“确认执行”后，作业实例运行状态显示为“等待运行”。 **重跑实例时**，作业实例运行状态显示为“待确认执行”，可以进行手动确认执行，单击“确认执行”后，作业实例运行状态显示为“等待运行”。 **补数据场景下**，在补数据监控页面，补数据作业实例运行状态显示为“待确认执行”，可以在实例监控页面进行手动确认执行，单击“确认执行”后，补数据作业实例运行状态显示为“等待运行”。 **批作业监控场景下**，在批作业监控页面，作业实例运行状态显示为“待确认执行”，可以进行手动确认执行，单击“确认执行”后，作业实例运行状态显示为“等待运行”。 ● 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9-55**。具体使用教程可参见**跨空间进行作业调度**。
是否空跑	如果勾选了空跑，任务不会实际执行，将直接返回成功。
任务组	选择已配置好的任务组。配置方法请参见**配置任务组**。 系统默认“不选择任务组”。 任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。 举例1：任务组里面最大并发数配置为2，作业节点有5个，当作业运行时，只有两个节点在运行中，其它节点在等待运行。 举例2：任务组里面最大并发数配置为2，补数据的并发周期数设置为5，当作业进行补数据时，有两个补数据生成的作业实例在运行中，其它的在等待运行。等待运行的实例，一段时间后，可以正常下发。 举例3：如果多个作业配置同一个任务组，任务组里面最大并发数配置为2，只有两个作业是运行中，其他作业在等待执行。如果多个作业节点上配置了任务组，任务组里面最大并发数配置为2，作业节点总共有5个，根据作业调度时间，只有两个节点在运行中，其它节点在等待运行。 **说明** 对于Pipeline作业，每个节点都可以配置一个任务组，也可以在作业里面统一配置任务组，如果配置了节点级任务组，则优先级高于作业级的任务组。

表 9-54 “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。 系统支持生效时间可以快速选到今天和明天。单击生效时间的时间框，在时间框界面单击“今天”或“明天”，可以快速选择当前日期。
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。 已经在运行中的作业，可以修改其调度周期。 ● 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。 **说明** 调度周期选择“分钟”时，系统不支持按照配置的时间间隔固定频率去运行，即不支持跨小时按照固定频率去运行。举例如下： ● 2024年6月19日14点20分配置了分钟调度，开始时间为0时30分，间隔时间为30分钟，结束时间为23时59分，则实际作业运行时间周期为2024-06-19 14:30:00、2024-06-19 15:30:00、2024-06-19 16:30:00、2024-06-19 17:30:00、2024-06-18 18:30:00等。 ● 2024年6月19日14点20分配置了分钟调度，开始时间为0时0分，间隔时间为50分钟，结束时间为23时59分，则实际作业运行时间周期为2024-06-19 14:50:00、2024-06-19 15:00:00、2024-06-19 15:50:00、2024-06-19 16:00:00、2024-06-19 16:50:00、2024-06-19 17:00:00、2024-06-19 17:50:00等。 ● 小时：支持按间隔小时配置调度周期，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。同时支持按离散小时进行调度周期配置，可以指定一天内的任意小时和分钟进行调度，离散小时调度仅支持自然周期调度。 ● 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。同时系统支持可以选择“每月最后一天”进行业务调度。 **说明** 因为DataArts Studio不支持底层服务（例如，以前的CDM、DLI等服务）的补数据实例和周期调度作业实例并发运行，为了保证补数据实例不影响周期调度作业实例运行，两种类型作业实例不会抢占并发，所以，作业的周期调度的日期与该作业补数据的业务日期不能重合，周期调度和补数据不能同时运行，避免出现运行异常问题。

参数	说明
调度日历	根据已配置的日历信息，选择所需的调度日历。系统默认不使用调度日历。配置调度日历的操作请参见**配置调度日历**。 ● 使用按日历进行自定义工作日期进行周期调度，如果非工作日，作业会进行空跑。例如作业周期调度、补数据。 ● 配置好的调度日历，如果工作日期进行变更，已经在执行的作业实例无法生效，还没生成的作业实例可以立即生效。
监听OBS	打开监听OBS开关后，系统会自动监听OBS路径是否有新作业文件。关闭开关后不再监听OBS路径。 配置参数如下： ● OBS文件，支持EL表达式。 ● 监听间隔，可设置为1-60之间，单位为分钟。 ● 超时时间，可设置为1-1440之间，单位为分钟。

参数	说明
依赖作业	此处可以选择不同工作空间的周期调度作业作为依赖作业，则仅当依赖的作业运行完成时，才开始执行当前作业。支持单击“自动解析依赖”对作业依赖关系进行自动识别。 **说明** 跨工作空间的作业依赖规则，请参见**作业依赖规则**。 数据开发当前支持两种调度依赖策略：传统周期调度依赖和自然周期调度依赖。这两种周期调度依赖只能选择其中一种。对于新的应用实例而言，默认使用自然周期调度作为DataArts Studio新实例默认选项。 **图 9-36** 传统周期调度作业依赖关系全景图  **图 9-37** 自然周期调度作业依赖关系全景图  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**周期调度依赖策略**。

参数	说明
依赖的作业失败后，当前作业处理策略	当依赖的作业在当前作业周期内存在运行失败实例后，选择当前作业的处理策略： ● 等待执行 等待执行当前作业，等待执行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。 ● 取消执行 取消执行当前作业，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取消执行，依赖的作业12个实例中只要有一个失败的，当前作业就取消执行。 ● 如果当前参数配置的是继续执行，只要依赖的作业12个实例跑完了，当前作业就继续执行。 **说明** 依赖的作业失败后，当前作业处理策略可通过配置默认项进行批量设置，无需每个作业单独设置。具体请参见**配置默认项**。该配置仅对新建作业有效。
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。
配置作业依赖时，可以对所依赖的作业是否在调度中进行过滤	配置作业依赖关系时，可以对所依赖的作业是否在调度中进行过滤，避免上游依赖的作业未开始调度，从而导致下游作业失败。 ● 全部作业 ● 调度中的作业
配置作业依赖时，支持选择依赖周期	● 同周期 ● 上N周期，输入值必须在1到30之间。

参数	说明
跨周期依赖	选择作业实例之间的依赖关系。 ● 不依赖上一调度周期。此处可以配置并发数，表示多个作业实例并行执行的个数。如果并发数配置为1，前一个批次执行完成后(包括成功、取消、或失败)，下一批次才开始执行。 ● 自依赖（上一调度周期的作业实例执行成功下一周期才会执行，否则处于等待运行状态。等待上一调度周期结束才能继续运行）。 ● 跳过等待的实例，运行最近的批次（如果勾选该参数，跳过的作业实例将被置为取消状态且不再执行，当单作业实例执行时间过长时，可能会造成后续多批次作业全部被跳过。当作业实例需要持续执行时，强行跳过可能会造成业务逻辑错误，如当输出为分区表时，跳过冗余作业实例可能会造成“分区空洞”，建议谨慎配置此选项）。 **说明** ● “跳过等待的实例，运行最近的批次”当前只支持分钟或小时调度的作业实例跳过。 ● 作业并发数配置比较低，实例未生成情况下，阻塞实例不会跳过。 ● 小周期依赖大周期时，可能会存在部分实例没有跳过，仍然执行。
是否清理超期等待运行的作业实例	● 不清理 ● 清理 若不配置该参数，默认会按照空间级配置来清理超期的等待运行作业实例。用户可根据实际场景设置清理或者不清理等待运行的作业实例。
是否空跑	如果勾选了空跑，任务不会实际执行，将直接返回成功。
任务组	选择已配置好的任务组。配置方法请参见**配置任务组**。 系统默认“不选择任务组”。 任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。 **说明** 对于Pipeline作业，每个节点都可以配置一个任务组，也可以在作业里面统一配置任务组，如果配置了节点级任务组，则优先级高于作业级的任务组。

表 9-55 “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。 ● **“DIS”** ● **“KAFKA”**
“DIS”触发事件类型的参数	
DIS通道名称	选择DIS通道，当指定的DIS通道有新消息时，数据开发模块将新消息传递给作业，触发该作业运行。

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128。
事件检测间隔	配置时间间隔，检测DIS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，读取下一个事件
是否空跑	如果勾选了空跑，任务不会实际执行，将直接返回成功。
任务组	选择已配置好的任务组。配置方法请参见**配置任务组**。 系统默认“不选择任务组”。 任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。 **说明** 对于Pipeline作业，每个节点都可以配置一个任务组，也可以在作业里面统一配置任务组，如果配置了节点级任务组，则优先级高于作业级的任务组。
“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，读取下一个事件
是否空跑	如果勾选了空跑，任务不会实际执行，将直接返回成功。

参数	说明
任务组	选择已配置好的任务组。配置方法请参见**配置任务组**。 系统默认“不选择任务组”。 任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。 **说明** 对于Pipeline作业，每个节点都可以配置一个任务组，也可以在作业里面统一配置任务组，如果配置了节点级任务组，则优先级高于作业级的任务组。
是否空跑	如果勾选了空跑，任务不会实际执行，将直接返回成功。
任务组	选择已配置好的任务组。配置方法请参见**配置任务组**。 系统默认“不选择任务组”。 任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。 **说明** 对于Pipeline作业，每个节点都可以配置一个任务组，也可以在作业里面统一配置任务组，如果配置了节点级任务组，则优先级高于作业级的任务组。

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9-56**所示的参数。

表 9-56 节点调度配置

参数	说明
调度方式	选择作业的调度方式： ● 单次调度：手动触发作业单次运行。 ● 周期调度：周期性自动运行作业。 ● 事件驱动调度：根据外部条件触发作业运行。
“周期调度”的参数	
生效时间	调度任务的生效时间段。

参数	说明
调度周期	选择调度任务的执行周期，并配置相关参数： ● 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。 已经在运行中的作业，可以修改其调度周期。
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 选择“并发数”。多个作业实例并行执行的个数。如果并发数配置为1，前一个批次执行完成后（包括成功、取消、或失败），下一批次才开始执行。 ● 自依赖（上一调度周期的作业实例执行成功下一周期才会执行，否则处于等待运行状态。）
“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。
DIS通道名称	选择DIS通道，当指定的DIS通道有新消息时，数据开发模块将新消息传递给作业，触发该作业运行。 当“触发事件类型”选择“DIS”时才需要配置。
连接名称	选择数据连接，需先在“管理中心”创建kafka数据连接。当“触发事件类型”选择“KAFKA”时才需要配置。
Topic	选择需要发往kafka的消息Topic。当“触发事件类型”选择“KAFKA”时才需要配置。
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。

参数	说明
事件检测间隔	配置时间间隔，检测DIS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	• 从上次位置读起 • 从最新位置读起 当“触发事件类型”选择“DIS”或“KAFKA”时才需要配置。
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

9.4.9 提交版本

提交版本涉及到数据开发的版本管理功能。

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100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。

前提条件

已完成作业开发任务。

提交作业版本

“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。

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。

步骤4 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5 在作业画布或编辑器上方单击“提交”，提交版本。选择审批人，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，不勾选则无法单击确认。在提交版本时，单击“版本对比”可以查看当前提交版本与最近一个版本之间的差异对比。

图 9-38 提交

说明

- 如果在“审批中心”开启了提交审批的开关，则作业提交审批后，需要审批人在“审批中心”的“待审批”页签进行审批，只有当审批通过后，作业才能提交成功。具体操作请参见**审批配置**。如果开关是关闭状态，则不需要审批，直接提交新版本即可。

  如果要撤销已提交的审批流程，请您在“审批中心”的“我的申请”页签里进行撤销。修改完成后，可以重新提交审批。
- 开启了提交审批开关后，提交作业、删除作业以及导入“提交态”的作业时，均需要进行审批。
- 关闭提交审批开关前，请确保当前工作空间已无待未审批的流程。
- 企业模式下，不支持提交审批。

----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100条版本信息）。单击“回滚”，可以回退到任意一个已提交的版本。

回滚内容包括：

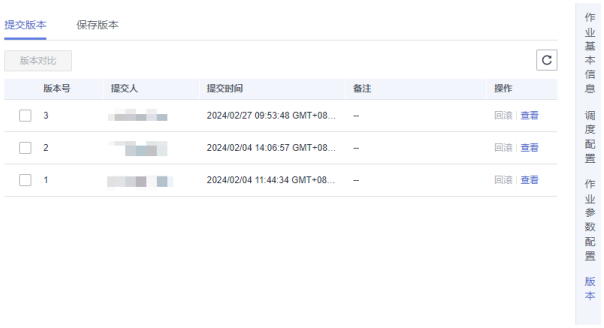
- 作业定义（算子属性、连线等）；
- 作业基本信息、作业调度配置、作业参数、血缘关系中的所有内容；

操作如下：

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。

4. 在作业目录中，双击作业名称，进入作业开发页面。
5. 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 9-39 版本回滚操作界面



版本详情查看

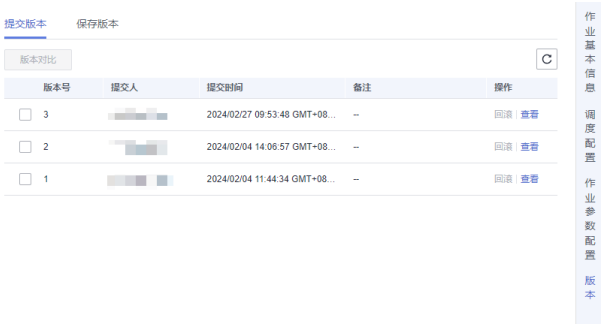
用户可以在版本列表中看到已经提交过的版本信息。

操作如下：

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中，双击作业名称，进入作业开发页面。
5. 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。

   单击查看，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 9-40 版本详情查看



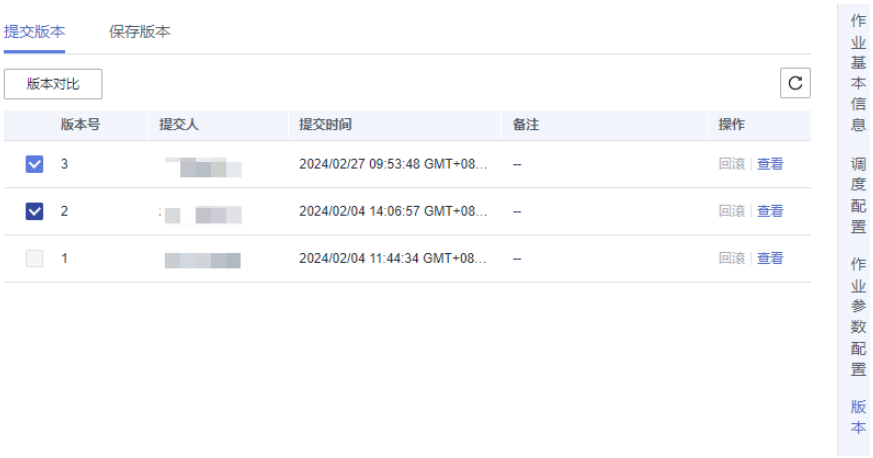
版本对比

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中，双击作业名称，进入作业开发页面。
5. 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 9-41 对比版本操作界面



9.4.10 发布作业任务

在企业模式中，开发者提交作业版本后，系统会对应产生一个作业类型的发布任务。开发者确认发布后，待拥有管理员、部署者、DAYU Administrator、Tenant Administrator权限的用户审批通过，然后将修改后的作业同步到生产环境。

须知

- 管理员导入作业时，选择导入提交态，会生成对应的待发布项。
- 管理员导入作业时，选择导入生产态，则不会生成待发布项。
- 开发者创建单任务的实时作业后，提交版本时，只生成当前作业的待发布项，不会生成子作业的待发布项。

前提条件

已提交版本，详情请参见**提交版本**。

操作步骤

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择开发环境，选择“数据开发 > 任务发布”。

步骤4 在待发布任务界面，会展示因提交版本而生成的待发布任务。您可以通过“查看”操作，查看当前任务相比上一版本的修改点，确认修改无误后，请通过“发布”操作，将任务进行发布。

支持通过“任务名称”和“提交人”进行发布项筛选。同时可以使用任务名称进行模糊查询。

说明

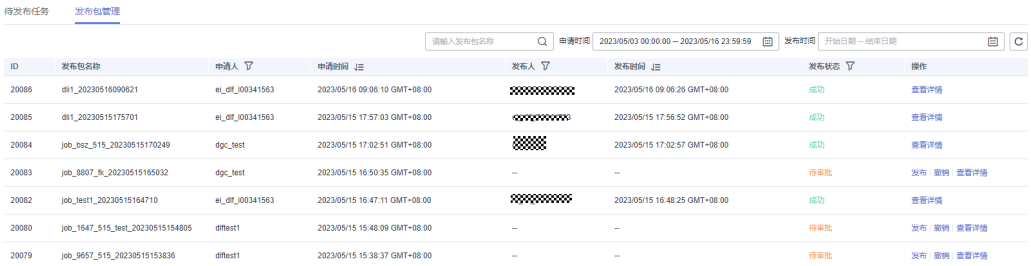
- 如果您只具备开发者权限，则需通过“发布”操作提交任务，由管理员或者部署者审批通过，才能将修改后的脚本同步到生产环境。
- 单击“发布”后，指定审批人，审批人必须是工作空间的管理员或部署者、拥有DAYU Administrator、Tenant Administrator权限的用户，至少指定一个审批人，不能指定自己为审批人。单击“审批人管理”可以跳转到“空间管理”页面，单击“编辑”按钮可以维护审批人信息。
- 可以进行批量发布。发布多个待发布项时，发布流程采用异步发布，可以看到发布任务的过程，最大的发布项个数为100。
- 单击发布后，系统会提示您“发布成功后，立即对发布包中的作业启动调度”。
- 对于暂时不发布的发布项，开发者、部署者和管理员可以进行撤销，支持批量撤销。

图 9-42 选择发布

步骤5 发布之后，您可以通过“发布包管理”查看任务的发布状态。待审批通过后，任务发布成功。

支持通过“申请人”、“申请时间”、“发布时间”、“发布人”和“发布状态”进行发布项筛选。同时可以使用发布包名称进行模糊查询。

图 9-43 查看任务状态



说明

对于暂时不发布的发布项，开发者、部署者和管理员可以进行撤销。

发布后，通过操作列的“查看详情”可以查看任务的发布状态和启动状态，在操作列的“版本对比”可以查看发布包不同版本间的内容差异。

图 9-44 查看发布详情

----结束

9.4.11 （可选）管理作业

9.4.11.1 复制作业

本章节主要介绍如何复制一份作业。

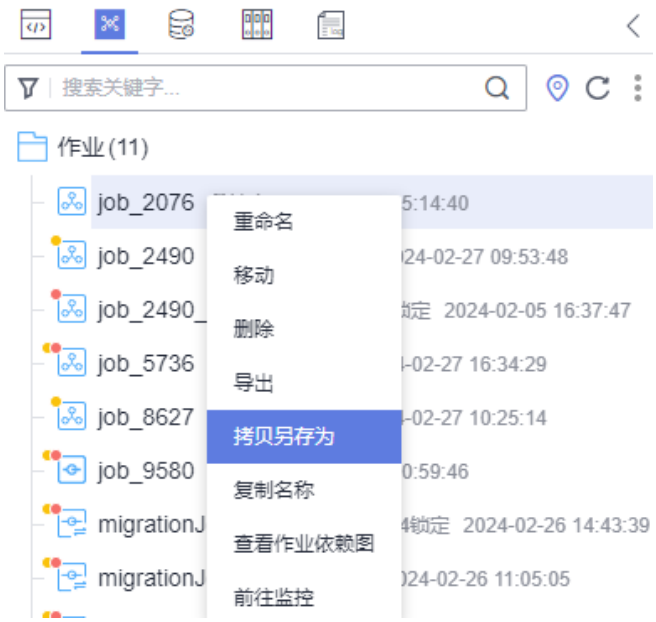
前提条件

已完成作业开发。如何开发作业，请参见**开发Pipeline作业**。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中选择需要复制的作业，右键单击作业名称，选择“拷贝另存为”。

图 9-45 复制作业



5. 在弹出的“另存为”页面，配置如表9-57所示的参数。

表 9-57 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。

6. 单击“确定”，复制作业。

9.4.11.2 复制名称和重命名作业

您可以通过复制名称功能复制当前作业名称，通过重命名功能修改当前作业名称。

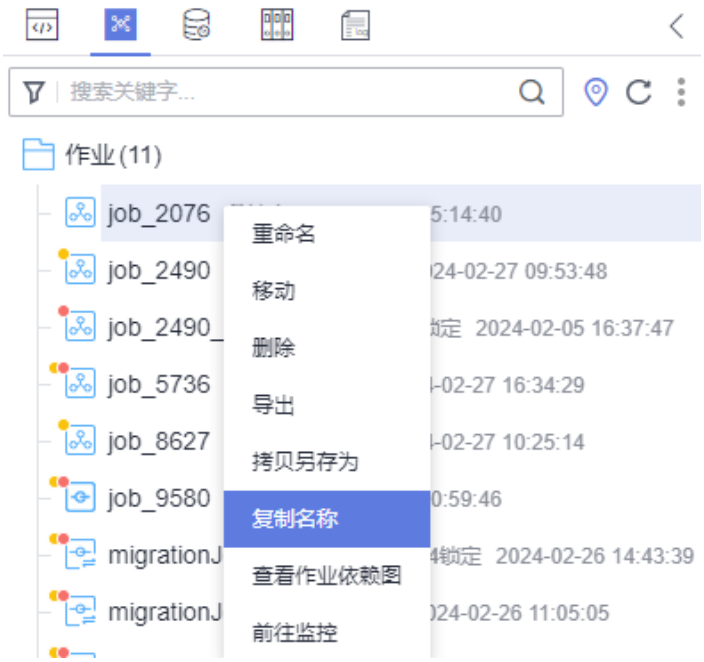
前提条件

已完成作业开发。如何开发作业，请参见开发Pipeline作业。

复制名称

1. 参考访问DataArts Studio实例控制台，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 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。

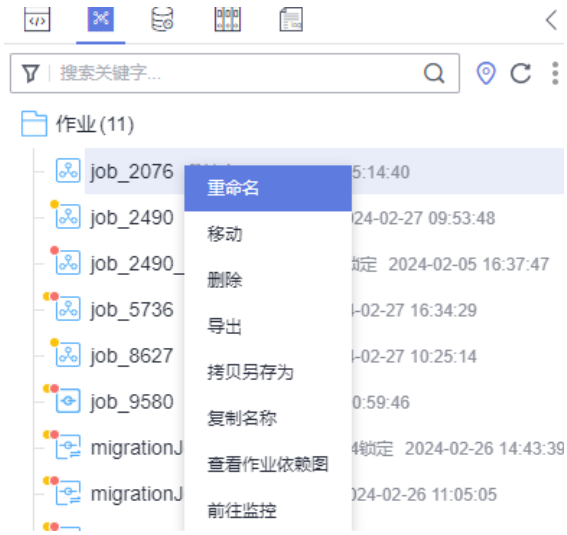
图 9-46 复制作业名称



重命名作业

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 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”。

图 9-47 重命名作业



5. 在弹出的“重命名作业名称”页面，配置新作业名。

图 9-48 重命名作业名称

表 9-58 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。

6. 单击“确定”，重命名作业。

9.4.11.3 移动作业/作业目录

您可以通过移动功能把作业文件或作业目录从当前目录移动到另一个目录。

前提条件

已完成作业开发。如何开发作业，请参见**开发Pipeline作业**。

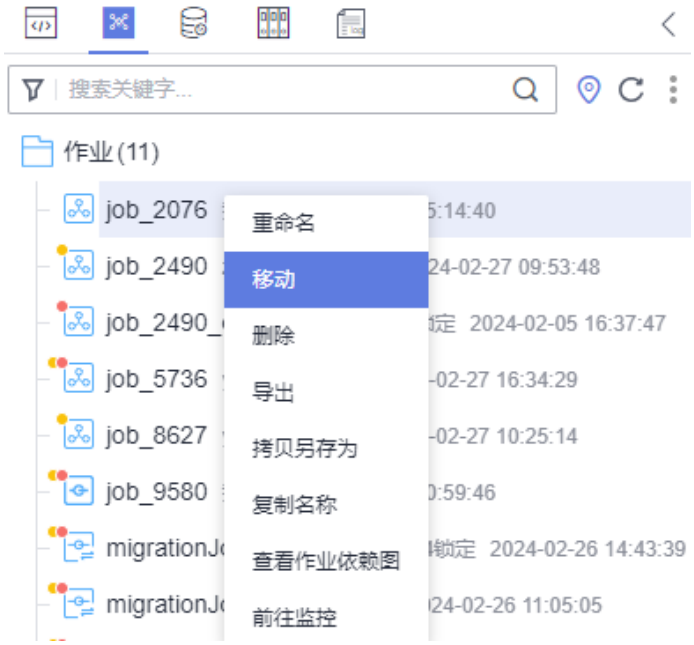
操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 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”。

   图 9-49 选择要移动的作业

   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。

图 9-50 移动作业

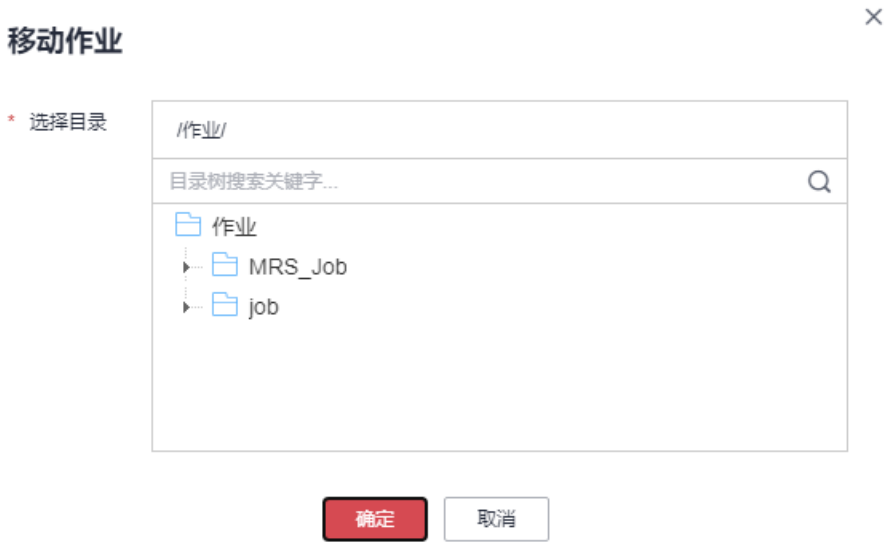


图 9-51 移动目录

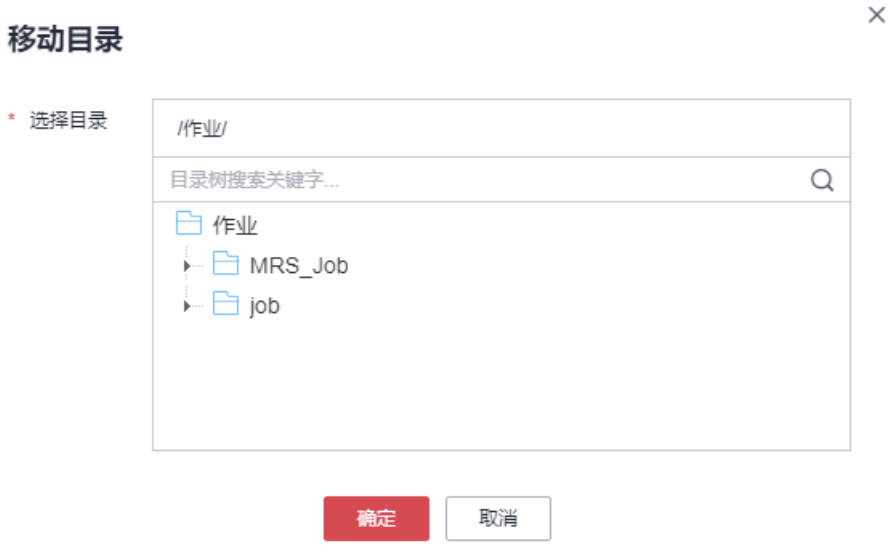


表 9-59 移动作业/作业目录参数

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。

9.4.11.4 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。

说明

数据开发在跨时区导出导入作业时，需要手动修改expressionTimeZone字段为目标时区。

导出作业

步骤1 参考访问DataArts Studio实例控制台，登录DataArts Studio管理控制台。

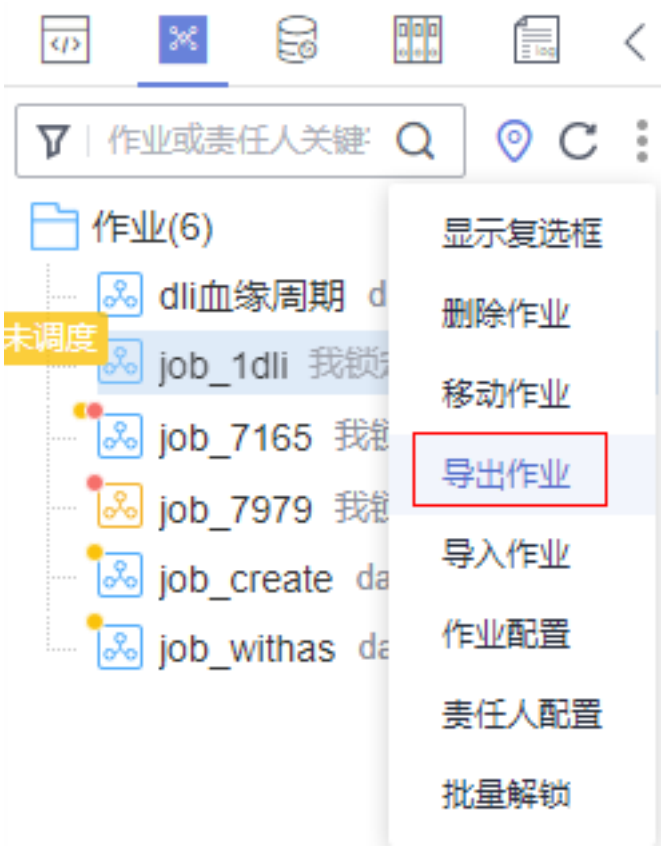
步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。

步骤4 单击作业目录中的，选择“显示复选框”。

步骤5 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 9-52 选择并导出作业



步骤6 在弹出的“导出作业”界面，选择需要导出的作业范围和状态，单击“确定”，可以在下载中心查看导入结果。

图 9-53 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。

说明

- 从OBS导入的作业文件，最大支持10Mb；从本地导入的作业文件，最大支持1Mb。从本地导入的作业文件， 解压后大小最大支持1Mb。
- 如果导入的作业在系统中有重名时，需要确保系统中该作业状态为“停止”时，才能导入成功。

在作业目录中导入一个或多个作业

步骤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”。

步骤2 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 9-54 导入作业定义及依赖

步骤3 单击“下一步”，根据提示导入作业。

说明

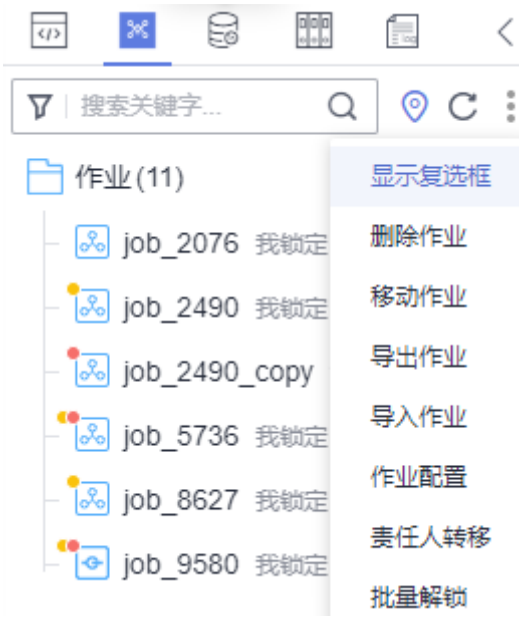
- 导入作业时，如果作业中存在“锁定”状态的标签，则作业导入会失败。
- 当作业导入失败需要自动生成标签时，如果标签已存在且被锁定，则导入失败的作业不会添加上该标签。
- 在导入作业过程中，若作业关联的数据连接、dis通道、dli队列、GES图等在数据开发模块系统中不存在时，系统会提示您重新选择。

----结束

9.4.11.5 批量配置作业

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 单击作业目录中的，选择“显示复选框”。

图 9-55 显示作业复选框



5. 勾选需要批量配置的作业，单击 > 作业配置。

图 9-56 作业配置菜单

6. 配置作业的通用项。

图 9-57 通用配置

表 9-60 通用配置

参数	说明
节点状态轮询时间	设置所选作业的所有节点轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。 如果设置为保持不变，则各节点保持原来的节点轮询时间。
节点执行的最长时间	设置所选作业的所有节点执行超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将会再次重试。 如果设置为保持不变，则各节点保持原来的节点执行最长时间。
作业委托	设置所选作业的委托，配置了作业委托后，作业执行过程中，以委托的身份与其他服务交互。 如果设置为保持不变，则各作业保持原来的委托配置。
失败重试	设置所选作业的所有节点执行失败后，是否重新执行节点。 如果设置为保持不变，则各节点保持原来的失败重试策略。
超时重试	当“失败重试”配置为“是”才显示此配置参数。 设置所选作业的所有节点执行超时后，是否重新执行节点。 如果设置为保持不变，则各节点保持原来的超时重试策略。

参数	说明
最大重试次数	当“失败重试”配置为“是”才显示此配置参数。 配置节点失败重试次数。 取值范围[1, 100]，默认值 : 1。
重试间隔时间	当“失败重试”配置为“是”才显示此配置参数。 配置失败重试的时间间隔。 取值范围[5, 600]，默认值 : 120，单位为秒。
当前节点失败后，后续节点处理策略	设置所选作业的所有节点执行失败后的操作。 如果设置为保持不变，则各节点保持原来的失败策略。
依赖的作业失败后，当前作业处理策略	设置所选作业的依赖作业执行失败后的操作。若作业未配置依赖关系，该配置不生效。 如果设置为保持不变，则当前作业保持原来的失败策略。
责任人	设置所选作业的责任人，只能从当前工作空间中的用户选择。 如果设置为保持不变，则各作业保持原来的责任人。
周期作业实例并发数	设置所选作业并行处理的数量。 如果设置为保持不变，则保持原来的周期作业实例并发数。
是否清理超期等待运行的作业实例	如果设置为取消运行，需要配置超期天数。当作业实例等待运行的时间超过了所配置的期限天数时，作业实例将取消执行，则会清理超期等待运行的作业实例。 如果设置为不取消，则不清理超期等待运行的作业实例。 如果设置为保持不变，则保持原来的作业实例运行等待超期规则。
超期天数	当“是否清理超期等待运行的作业实例”配置为“取消运行”时才显示此配置参数。 取值范围[2, 271]，默认值 : 60，单位为天。 超期天数，最小需配置2天，即至少需要等待2天，才可取消未运行的作业实例。
备注	输入备注信息。

7. 单击“CDM集群”，配置所选作业的CDM Job节点的CDM集群。

   在左侧下拉框中选择待修改的CDM集群名称，右侧下拉框中选择要设置的CDM集群名称。

说明

1. CDM集群迁移的前提是需要在新集群创建同名作业。
2. CDM作业同时配置两个CDM集群：
   - 如果原集群选择其中一个时，迁移只影响其中一个集群，对另一个集群无影响。
   - 如果原集群选择全部（两个集群）时，会将2个集群都迁移到目标集群中。

- 搜索：输入作业名称，单击 ,可筛选需要修改的含有CDM Job节点的作业。
- 刷新：单击，刷新含有CDM Job节点的作业列表。
- 下载：单击，下载该界面中勾选的作业列表。

图 9-58 CDM 集群

8. 单击“DLI队列”，配置所选作业的DLI SQL节点的DLI队列。

   在左侧下拉框中选择待修改的DLI队列名称，右侧下拉框中选择要设置的DLI队列名称。

   说明

   - 搜索：输入作业名称，单击 ,可筛选需要修改的含有DLI SQL节点的作业。
   - 刷新：单击，刷新含有DLI SQL节点的作业列表。
   - 下载：单击，下载该界面中勾选的作业列表。

图 9-59 DLI 队列

9. 单击“确定”，完成配置。

9.4.11.6 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。

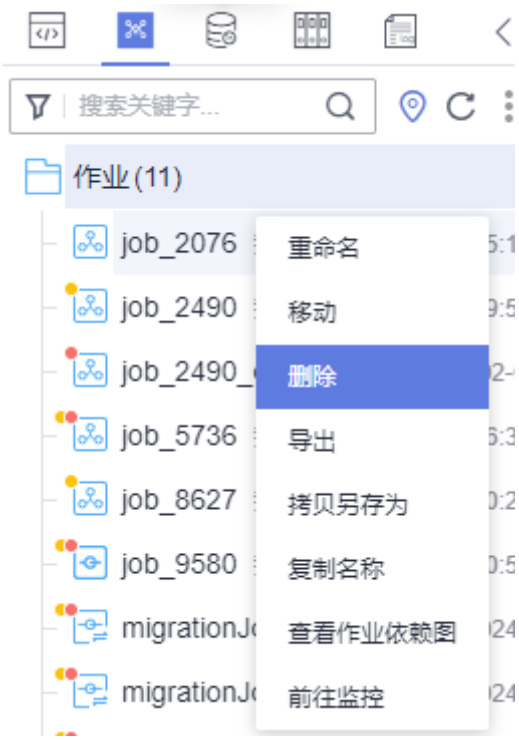
说明

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中，右键单击作业名称，选择“删除”。

图 9-60 删除作业



5. 在弹出的“删除作业”页面，单击“确定”，删除作业。

批量删除

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. 在作业目录顶部，单击，选择“显示复选框”，在作业目录前出现复选框。
3. 选择需要删除的作业，再次单击，选择“删除作业”。

图 9-61 批量删除作业

4. 在弹出的“删除作业”页面，单击“确定”，批量删除作业。

9.4.11.7 解锁作业

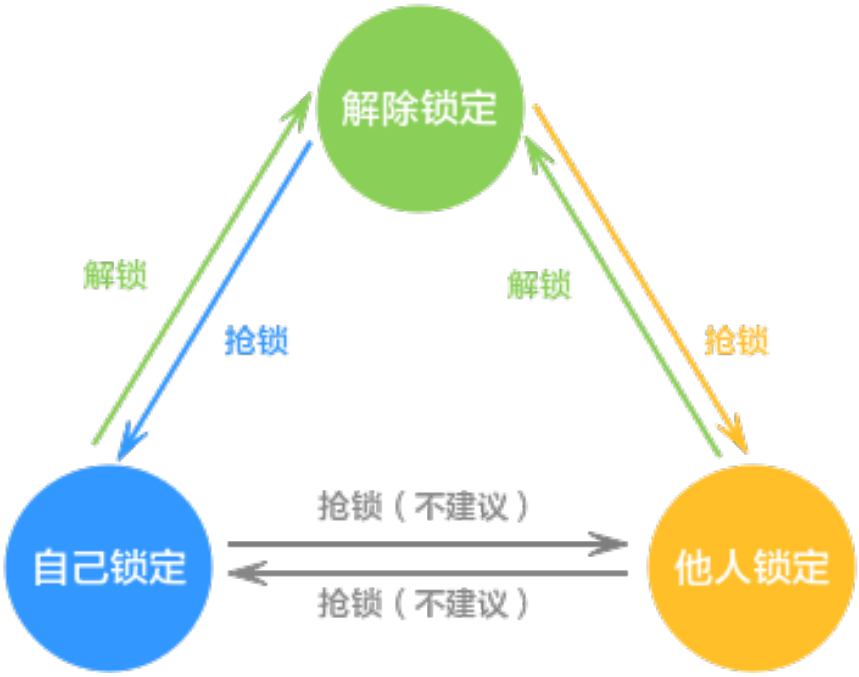
脚本/作业解锁涉及到数据开发的编辑锁定功能。

编辑锁定：用于避免多人协同开发脚本/作业时产生的冲突。新建或导入脚本/作业后，默认当前用户锁定脚本/作业，只有当前用户自己锁定的脚本/作业才可以直接编辑、保存或提交，通过“解锁”功能可解除锁定；处于解除锁定或他人锁定状态的脚本/作业，必须通过“抢锁”功能获取锁定后，才能继续编辑、保存或提交。

须知

- 当前脚本/作业的锁定状态可以通过脚本/作业的目录树查看。
- 对于已被他人锁定状态的脚本/作业，您需要通过重新打开该脚本/作业，查看最近的保存/提交时的内容。已打开的脚本/作业内容不会实时刷新。
- 在DataArts Studio更新编辑锁定功能前已经创建的脚本/作业，在更新后默认为解除锁定状态。您需要通过“抢锁”功能获取锁定后，才能继续编辑、保存或提交。
- 抢锁的操作依赖于软硬锁的处理策略。配置软硬锁的策略请参见**配置默认项**。
  - 软锁：忽略当前作业或脚本是否被他人锁定，可以进行抢锁或解锁。
  - 硬锁：若作业或脚本被他人锁定，则需锁定的用户解锁之后，当前使用人方可抢锁，空间管理员或DAYU Administrator可以任意抢锁或解锁。
- 不建议直接抢锁处于他人锁定状态的脚本/作业，这会导致他人的修改丢失。如果您有修改需求，请先联系锁定人将脚本/作业解锁，然后再抢锁。

图 9-62 锁定状态转换图



前提条件

已完成作业开发任务。

解锁作业

“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，建议您在提交作业后通过“解锁”解除该作业锁定。

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。

步骤4 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5 提交作业后在作业画布或编辑器上方单击“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 9-63 解锁

----结束

9.4.11.8 查看作业依赖关系图

您可以通过查看作业依赖关系视图，直观查看该作业关联的上下游作业。

前提条件

已经在**开发Pipeline作业**的作业调度配置中设置了依赖作业，否则视图中仅能展示当前作业节点，无法展示具备依赖关系的上下游作业节点。

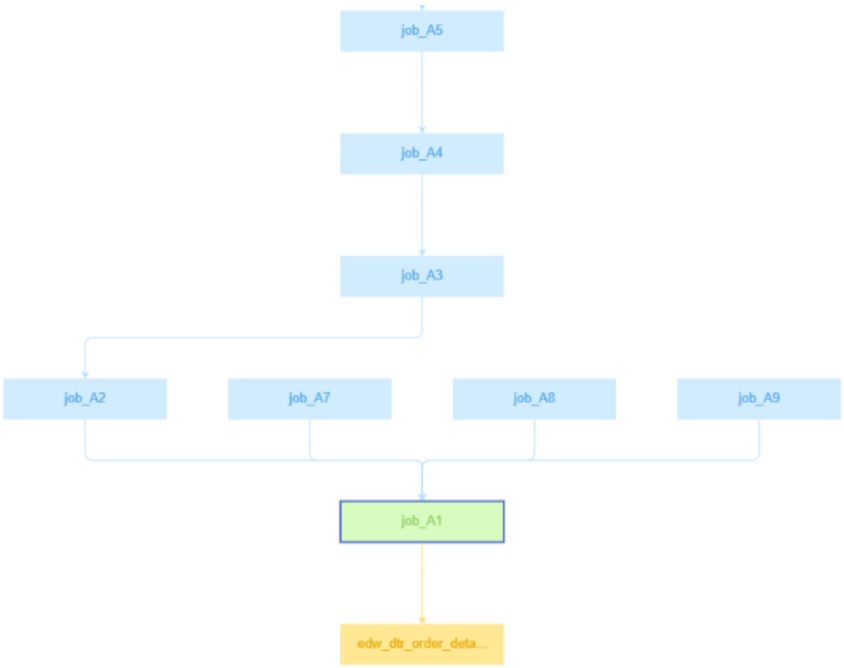
操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中选择需要查看的作业，右键单击作业名称，选择“查看作业依赖关系图”，界面弹出“作业依赖关系视图”页面。

图 9-64 作业依赖关系视图

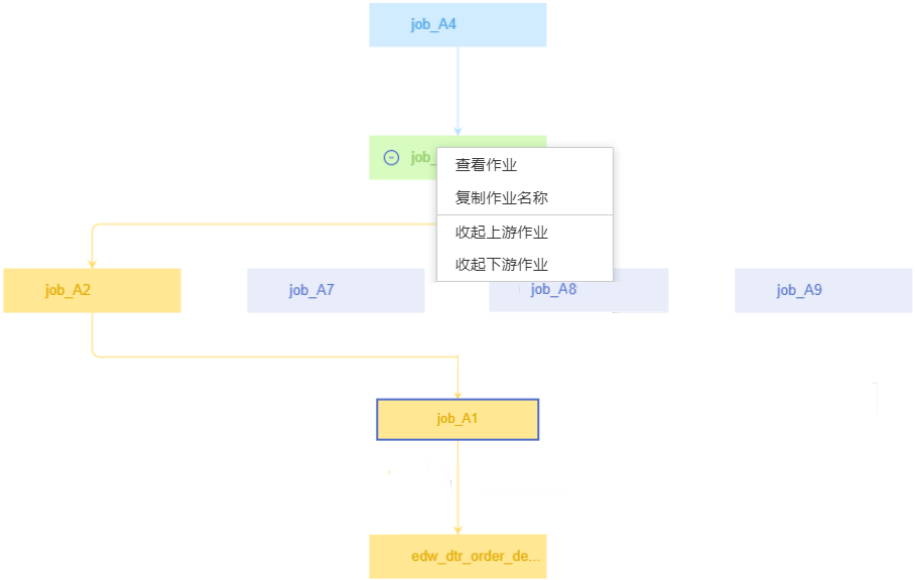
5. 在弹出的“作业依赖关系视图”页面，支持如下操作：
   - 视图右上角支持“显示完整依赖图”、“显示当前作业及其上下游”和“显示当前作业及其直接上下游”。
   - 视图右上角支持按节点名称进行搜索，搜索出来的作业节点高亮显示。
   - 单击下载按钮，可以下载作业的依赖关系文件。
   - 鼠标滚轮可放大、缩小关系图。
   - 鼠标按住空白处，可自由拖拽用以查看完整关系图。
   - 鼠标光标悬停在作业节点上，该作业节点会被标记为绿色，上游作业会被标记为蓝色，下游作业会被标记为黄色。

图 9-65 上下游作业节点标记



- 在作业节点上右键单击，可进行查看作业、复制作业名称、收起上/下游作业等操作。

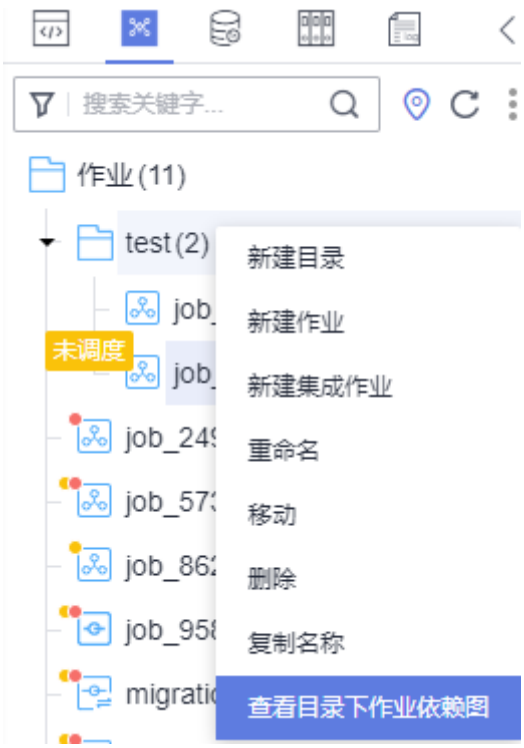
图 9-66 作业节点操作



通过作业树目录查看作业依赖关系图

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. 右键单击作业所在的目录，单击“查看目录下作业依赖关系图”进入该目录下作业依赖图查看界面。

图 9-67 在目录树上查看作业依赖关系图



3. 系统自动展示该目录下作业的所有依赖关系，您可以查看作业之间的相互依赖关系。系统支持通过作业名称进行查找并高亮显示。

   说明

   - 在依赖关系图中单击某节点，其上游作业会被标记为蓝色，下游作业会被标记为黄色。
   - 鼠标按住可自由拖拽以查看完整关系图。
   - 鼠标滚轮可缩放视图。

9.4.11.9 转移作业责任人

数据开发模块提供了转移作业责任人的功能，您可以将责任人A的所有作业一键转移到责任人B名下。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录顶部，单击，选择“责任人转移”。

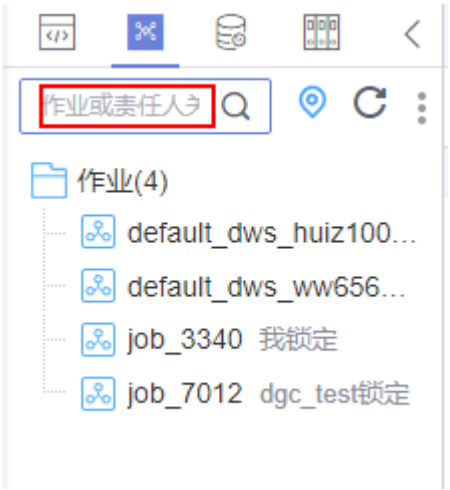
图 9-68 责任人转移

5. 分别设置“当前责任人”和“目标责任人”，单击“转移”。
6. 提示转移成功后，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 9-69 根据作业责任人筛选作业



9.4.11.10 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 单击作业目录中的，选择“显示复选框”。

5. 勾选需要解锁的作业，单击 > 批量解锁。弹出“解锁成功”提示。

图 9-70 批量解锁

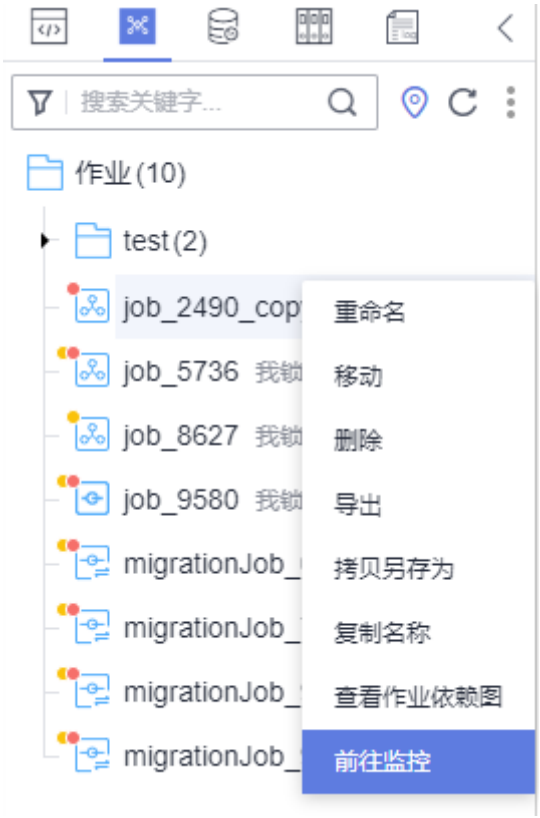
9.4.11.11 前往监控

您可以通过作业目录树，快速跳转到该作业的监控界面，查看该作业的监控详细信息。

操作步骤

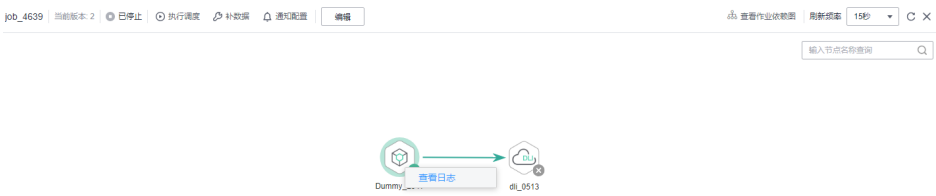
1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
4. 在作业目录中选择需要的作业，右键单击作业名称，选择“前往监控”，进入作业监控界面。

图 9-71 前往监控



5. 在监控界面，可以查看该作业节点的日志信息、版本信息、对该作业执行调度、单击编辑或者作业名称进行作业开发界面修改作业信息等。

图 9-72 作业监控界面



9.5 集成作业开发

集成作业包含离线处理作业和实时处理作业，操作入口在数据开发界面。

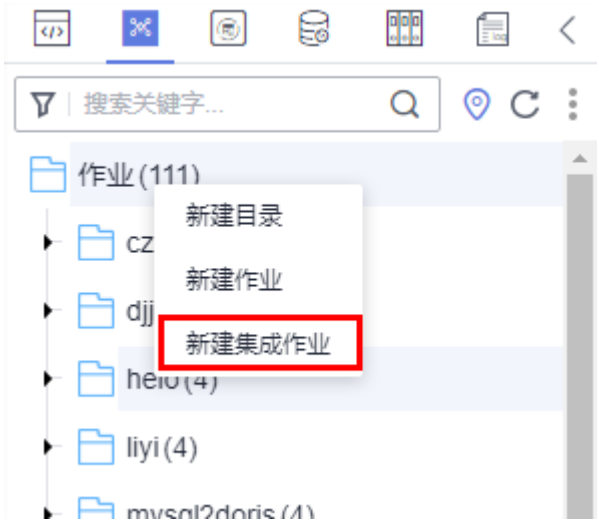
新建集成作业的方式有如下两种：

方式一：在“作业开发”界面中，单击“新建集成作业”。

图 9-73 新建集成作业（方式一）

方式二：在作业目录中，右键单击目录名称，选择“新建集成作业”。

图 9-74 新建集成作业（方式二）



离线处理作业的详细操作，请参见**数据集成（离线作业）**。

实时处理作业的详细操作，请参见**数据集成（实时作业）**。

9.6 Notebook 开发

9.6.1 Notebook 概述

说明

该功能为白名单功能，如需使用Notebook功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。

目前仅支持亚太-新加坡和拉美-圣地亚哥两个局点申请开放使用。

DataArts Studio的Notebook是一个交互式开发环境，提供全托管式JupyterLab云化版本，即开即用。帮助数据工程师及数据科学家轻松完成开发、调试、调度集群作业，并支持实时探索、处理和数据可视化。

Notebook是基于开源JupyterLab进行了深度优化的交互式数据分析挖掘模块，提供在线的开发和调试能力，用于编写和调测模型训练代码。完成DataArts Studio对接Notebook实例后，您可以基于Notebook提供的Web交互的开发环境同时完成代码的编写与作业的开发，使用Notebook灵活的进行数据分析与探索。

关于Jupyter Notebook的详细操作指导，请参见**Jupyter Notebook使用文档**。

使用Notebook实例提交DataArts Studio作业适用于在线开发调试场景下的作业需求，无需准备开发环境，一站式完成数据分析与探索。

在使用该功能前，需要先启用Notebook。如果还未启用Notebook，页面会显示“未启用Notebook”。启用Notebook的操作请参见**Notebook管理**。

说明

当前工作空间如果未启用Notebook，请联系管理员启用Notebook。或者，具有DAYU Administrator或者Tenant Administrator权限的用户也可以启用Notebook。

9.6.2 创建 Notebook 实例

本章节提供详细的创建Notebook实例的指导。

前提条件

登录用户需要授权DataArts Studio系统角色“DAYU User”。详细操作请参见**创建IAM用户并授予DataArts Studio权限**。

准备工作

- 已启用Notebook。如果还未启用Notebook，启用Notebook的操作请参见**Notebook管理**。
- 已创建好OBS桶。
- 已创建好VPC、子网、和安全组。请提前做好网络规划：
  - 如果仅对接DLI，则选择DLI增强型跨源所连接的用户的VPC、子网、和安全组。注意，安全组入方向规则需要放通DLI弹性资源池所在VPC的30000端口。
  - 请确保上述VPC子网已避开172.30.0.0/16和172.31.0.0/16网段。

约束条件

每个用户在每个空间仅能创建一个Notebook。

创建 Notebook 实例

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“数据开发 > Notebook”。

说明

- 如果当前工作空间已创建Notebook，单击“打开”按钮进入Notebook的开发界面。
- 如果当前工作空间未创建Notebook，系统会提示“当前工作空间xxxxx没有Notebook，您可以点击立即创建”。

图 9-75 立即创建

步骤4 单击“立即创建”，进入创建Notebook界面，配置如下参数。

表 9-61 创建 Notebook

配置项	参数	描述
基础配置	名称	Notebook名称。
	描述（选项）	Notebook描述。
OBS配置	选择OBS桶	选择的OBS桶用于保存用户的ipynb文件。 **说明** 当前登录用户需要有OBS上传文件的权限。
VCP配置	选择虚拟私有云	请选择MRS集群或DLI增强型跨源所在的VPC。
	选择子网	请选择MRS集群或DLI增强型跨源所在的子网。
	选择安全组	请选择MRS集群或DLI增强型跨源所在的安全组。

图 9-76 创建 notebook 实例

步骤5 单击右下角的“立即创建”，完成Notebook的创建。创建完成大概需要1分钟。

创建好的Notebook在页面下方进行展示，系统会自动显示该Notebook的创建人、创建时间、OBS以及网络信息等。

单击“打开”按钮进入Notebook的开发界面，详细内容参见后面的章节。

----结束

更多操作

- 创建好的Notebook实例可以进行删除。单击“删除”，在弹出的“删除Notebook”提示框中，单击“确定”，即可删除。
- 创建好的Notebook实例，若凭证过期，可以进行授权。单击“授权”，可以重置Notebook内用户授权，有效期24h。

  说明

  由于token会过期，导致无法获取资源。可通过“授权”正常使用。

9.6.3 开发任务

本章节详细介绍使用Notebook开发任务的操作指导。

前提条件

已启用Notebook并且已创建出Notebook实例。创建Notebook实例的操作请参见**创建Notebook实例**。

约束限制

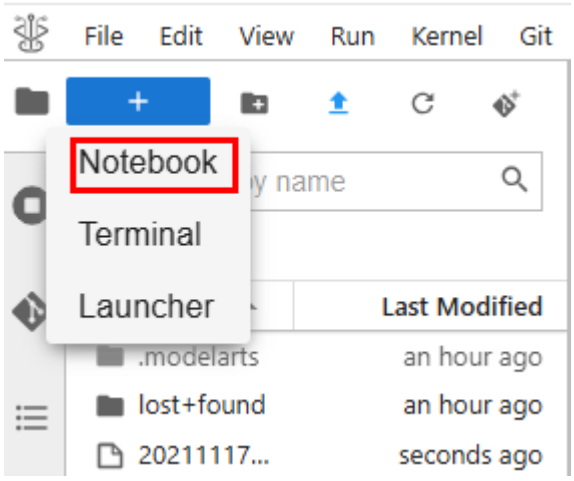
- 目前仅支持DLI数据源。

开发 Notebook

步骤1 在数据开发主界面的左侧导航栏，选择“数据开发 > Notebook”。

步骤2 单击“打开”按钮进入Notebook的开发界面。

步骤3 单击 + ，选择Notebook，新建Notebook文件，新增**.ipynb**格式的文件，实现新建Notebook。下一步开发Notebook。



步骤4 进入Notebook开发界面后，输入开发代码并进行调试。

说明

- 支持保存、下方插入、剪切、复制、粘贴、运行、中断等操作。
- Notebook支持对某行代码运行调试运行。单击代码行前面的，可以运行该行代码。
- 支持Code、Markdown、Raw三种格式的代码展示风格。
- 支持对代码行进行下方插入、上移、下移、删除等操作。

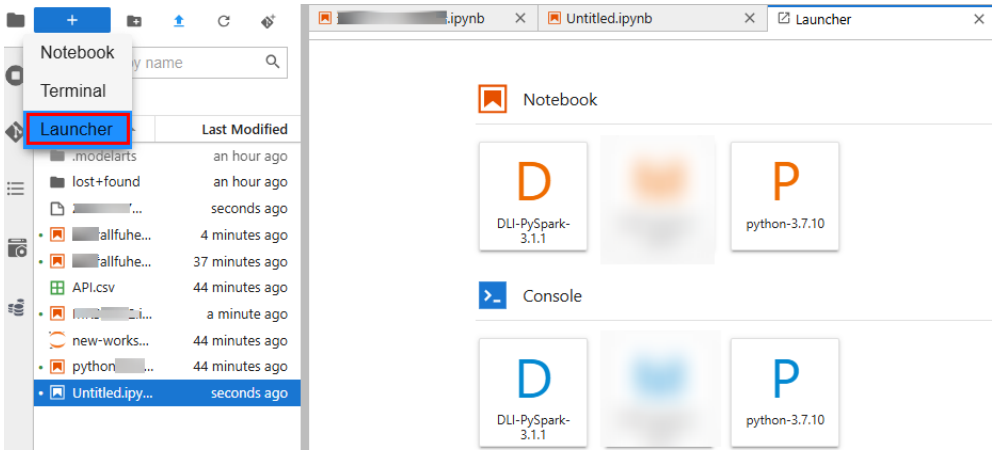
步骤5 单击运行按钮，运行代码。

步骤6 查看代码运行结果。

----结束

开发 launcher

步骤1 单击，选择Launcher，进入Launcher开发界面。目前仅支持DLI、Python两种任务开发。

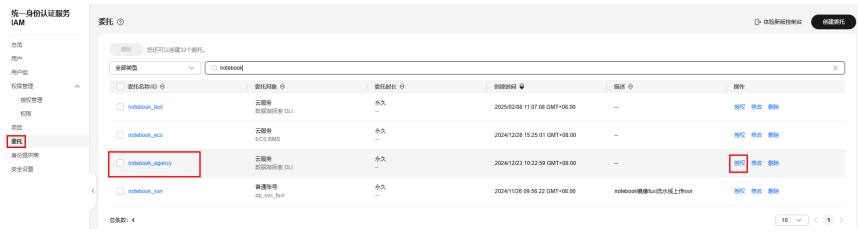


- **开发DLI任务**。

说明

开发DLI任务前，请先为DLI创建一个委托并进行授权，例如：委托名称notebook_agency，授权DLI FullAccess和OBS Administrator。操作方法请参考**创建DLI自定义委托**。

图 9-77 创建委托并授权



同时，在“权限管理 > 权限”里面配置如下自定义策略：


```
{
  "Version": "1.1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "iam:agencies:pass"
      ]
    }
  ]
}
```

a. 单击DLI-PySpark-3.1.1，进入DLI开发界面。

b. 在界面右上角，单击“connect”配置连接信息。

- 配置连接基本参数Pool和Queue（DLI的资源池和队列名称）。
- 配置高级参数（Advanced Settings）。

表 9-62 Spark Config

参数	描述
--conf	填写DLI任务运行的参数。例如配置DLI任务的代理名称。 spark.dli.job.agency.name=notebook_agency（必填） 其他参数根据实际业务需要进行配置。
--jars	填写文件的OBS路径，多个路径以Enter键分隔。（可选）
--py-files	填写文件的OBS路径，多个路径以Enter键分隔。（可选）
--files	填写文件的OBS路径，多个路径以Enter键分隔。（可选）

表 9-63 Resource Config

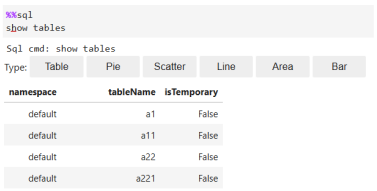
参数	描述
Driver Memory	设置Driver Memory的大小。大小在0-16之间。默认值为1，不能输入0。
Driver Cores	设置Driver Cores的大小。大小在0-4之间。默认值为1，不能输入0。
Executor Memory	设置Executor Memory的大小。大小在0-16之间。默认值为1，不能输入0。
Executor Cores	设置Executor Cores的大小。大小在0-4之间。默认值为1，不能输入0。
Executors	设置Executors的大小。大小在0-16之间。默认值为1，不能输入0。

- 单击“Connect”。配置好以后，会展示DLI队列信息和集群状态“cluster status: connected”。

c. 单击代码行，输入开发代码并进行调试。

d. 单击代码行前面的，可以运行该行代码。

表 9-64

示例代码	运行结果示意图
%%spark spark.read.parquet('obs://mytestbucket/demo/data.parquet').show()	
%%sql show tables	
%scala import org.apache.spark.sql.SparkSession val spark = SparkSession.builder().appName("demo").getOrCreate(); val inputFile = "obs://mytestbucket/demo/test.txt" val outputDir = "obs://mytestbucket/demo/test" val textFile = spark.read.textFile(inputFile) val wordCounts = textFile.flatMap(line => line.split(" ")).groupByKey(identity).count() wordCounts.write.format("csv").save(outputDir) wordCounts.show() spark.stop()	
%python from pyspark.sql import SparkSession spark = SparkSession.builder.appName("PySpark Example").getOrCreate() data = [("Alice", 34), ("Bob", 45), ("Charlie", 29), ("David", 35)] columns = ["Name", "Age"] df = spark.createDataFrame(data, columns) df.show() filtered_df = df.filter(df.Age > 30) filtered_df.show() average_age = df.groupBy().avg("Age").collect()[0][0] print(f"Average Age: {average_age}") spark.stop()	

- **开发Python任务**。
  a. 单击python-3.7.10，进入Python开发界面。
  b. 单击代码行，输入开发代码并进行调试。
  c. 单击代码行前面的，可以运行该行代码。

步骤2 代码开发完毕后，进行保存并运行。

步骤3 查看代码运行结果。

----结束

Notebook 更多操作

Notebook更多相关操作请参考**JupyterLab简介及常用操作**。

Cell 通用操作

表 9-65 Cell 通用操作

操作	说明
运行当前Cell	单击后，当前Cell开始运行。
停止运行当前Cell	单击后，正在运行的Cell停止运行。
清空当前Cell结果	鼠标悬浮至结果展示区，单击前方的选择清除单元格输出，即可清除当前Cell的运行结果。
插入Cell（在下方插入）	**下方插入**：鼠标单击已有Cell，单击右侧的向下插入Cell即可。
移动Cell顺序（向上方/向下方）	**单元格上移**：鼠标单击已有Cell，单击右侧的向上移动Cell。 **单元格下移**：鼠标单击已有Cell，单击右侧的向下移动Cell。
移除Cell	单击后，当前Cell将从Notebook中被移除。
复制/剪切/粘贴当前Cell	• 页面单击，可以用快捷键Ctrl + C。 • 页面单击，可以用快捷键Ctrl + X。 • 页面单击，可以用快捷键Ctrl + V。
支持Cell三种展示风格	单击 Code 的向下箭头，可以切换Cell展示风格 当前支持Code、Markdown、Raw三种展示风格。默认Code形式的展示。

9.6.4 常用操作按钮和功能菜单

操作按钮

表 9-66 操作按钮

按钮	说明
+	新增Notebook、Terminal、Launcher。

按钮	说明
	新建文件夹。 可以对文件夹进行删除、重命名等操作。 鼠标悬浮在文件夹名上右键选择New File可以创建.txt格式的任务。 鼠标悬浮在文件夹名上右键选择New Markdown File可以创建.md格式的任务。
	上传文件至JupyterLab。
	更新文件列表。
	Git克隆。

功能菜单

表 9-67 功能菜单

菜单	说明
	File Browser（文件浏览器） 可以通过名称过滤文件。支持通过模糊搜索查找文件。 可以对文件进行删除、重命名等操作。
	Running Terminals and Kernels 会展示**Open Tabs**、**Kernels**、**Terminals**。**Open Tabs**表示当前正在打开的文件。
	Git（Git存储库） **说明** You are not currently in a Git repository. To use Git, navigate to a local repository, initialize a repository here, or clone an existing repository. • Open the FileBrowser，单击该按钮后，自动跳转到File Browser页面。 • Initialize a Repository • Clone a Repository
	Table of Contents（目录）
	DataSource（数据源）表示Data Connections。

上传文件至 JupyterLab

步骤1 在数据开发主界面的左侧导航栏，选择“数据开发 > Notebook”。

步骤2 单击“打开”按钮进入Notebook的开发界面。

步骤3 单击进入“上传文件到Notebook”页面。上传文件的详细操作请参见**上传文件至JupyterLab**。

- 上传本地文件

  通过拖拽本地文件进行上传，文件夹请先压缩；或者单击“选择文件”，从本地选择所需文件进行上传。

- 上传Git文件

  输入GitHub开源仓库的URL，或者通过GitHub开源仓库克隆。

- 上传OBS文件

  输入OBS文件路径，或者从OBS File Browser中选择，支持模糊搜索。

  单击，可以设置OBS中转路径，单击“确定”设置完成。也可以使用OBS默认路径作为中转路径。

----结束

下载 JupyterLab 文件到本地

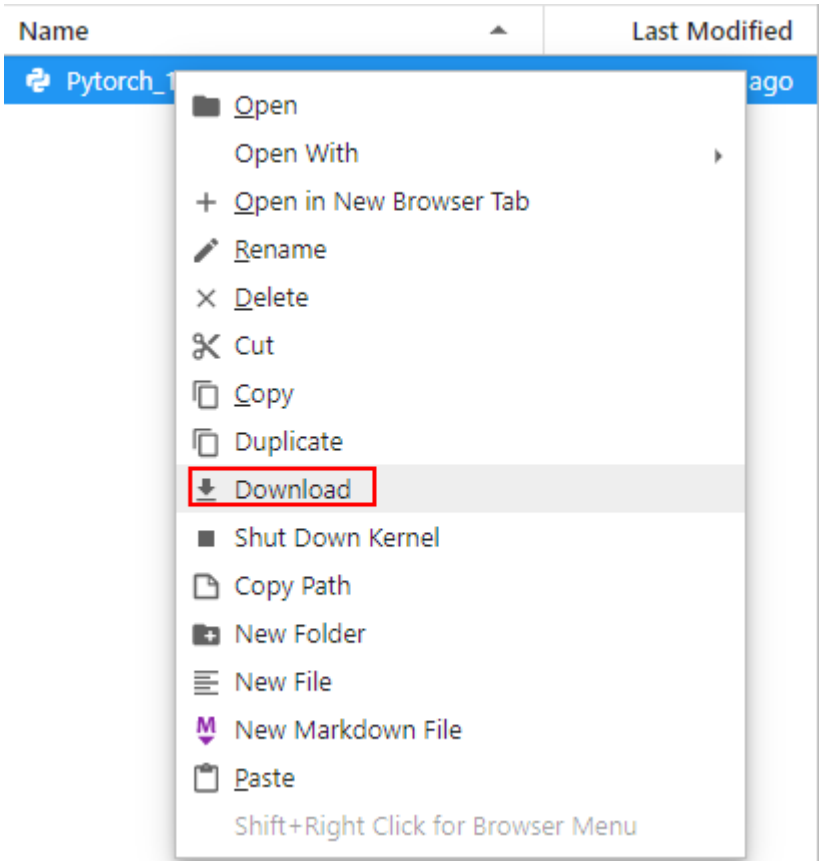
在JupyterLab中开发的文件，可以下载至本地。关于如何上传文件至JupyterLab，请参见**上传文件至JupyterLab**。

在JupyterLab中开发的文件，可以直接从JupyterLab中下载到本地。

在JupyterLab文件列表中，选择需要下载的文件，单击右键，在操作菜单中选择“Download”下载至本地。

下载的目的路径，为您本地浏览器设置的下载目录。

图 9-78 下载文件



9.7 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 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”。
4. 在左侧目录上方，单击解决方案图标，显示解决方案目录。
5. 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9-68**所示的参数。

图 9-79 新建解决方案

表 9-68 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。

6. 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。

9.8 运行历史

运行历史功能可支持查看脚本、作业和节点的一周（7天）内用户的运行记录。

前提条件

运行历史功能依赖于OBS桶，若要使用该功能，必须先配置OBS桶。请参考**配置OBS桶**进行配置。

脚本运行历史

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
4. 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。
5. 在过滤框中选择“脚本”，展示历史7天的脚本运行记录。
6. 单击某一条运行记录，可查看当时的脚本信息和运行结果。
7. 下载脚本历史运行结果。

说明

- 系统默认支持所有用户都能下载脚本的历史运行结果。
- 您可以在结果页签单击“下载结果”。
- 支持将CSV格式的结果文件下载到本地。查询结果和下载结果最大支持1000条。

作业运行历史

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
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。

9.9 运维调度

9.9.1 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，目前支持查看以下七种统计数据。

- 运行状态
  - 通过时间和责任人可以筛选出**今天**的**我的**或者**全部**责任人的作业实例调度运行状态概览
  - 通过时间和责任人可以筛选出**昨天**的**我的**或者**全部**责任人的作业实例调度运行状态概览
  - 通过时间和责任人可以筛选出**前天**的**我的**或者**全部**责任人的作业实例调度运行状态概览
  - 通过时间和责任人可以筛选出**近七天**的**我的**或者**全部**责任人的作业实例调度运行状态概览
  - 单击运行状态，可以跳转到实例监控界面，查看该运行状态的所有作业的详细信息。

    说明

    - 此处的统计数据包含实时作业的运行实例监控数据。单击运行状态后，实时作业不能跳转到实例监控页面，只能查看批作业的运行实例监控详情。
    - 系统默认查看**今天**的**全部**责任人的作业实例调度运行状态概览。
    - 支持查看通过条件筛选出来的实例总数，以及运行成功的实例总数及运行成功百分比。

    图 9-80 运行状态

- 任务完成情况

说明

只统计运行成功的实例，每小时统计一次今天的数据，任务表示作业中的算子。

- 支持指定开始日期并查看该日期的**前一天/选择天/7天历史平均**的运行成功的作业的**全部**节点算子的任务完成情况的曲线图。
- 支持指定开始日期并查看**前一天/选择天/7天历史平均**的运行成功的作业的不同类型节点算子的任务完成情况的曲线图。

- 任务数统计

说明

统计5分钟内启动执行的算子实例数，任务表示作业中的算子，可查看30天内的数据。

- 可以通过时间进行筛选，查看30天以内的每一天的启动执行的算子实例数据。
- 支持查看启动作业执行的**全部**节点算子实例数的曲线图。
- 支持查看启动作业执行的不同类型节点算子实例数的曲线图。

- DLI运行作业数/队列CU使用量

  支持通过**DLI队列**和**时间**筛选查看DLI运行作业数和队列CU使用量。

说明

- 系统默认支持查看七天内的数据。最多可查看一个月的数据。
- 仅支持查看非默认队列的数据。单击队列名称，可以将某个队列进行置顶。

- 作业数/任务日调度数

说明

统计较长周期总作业数量与日调度任务数量的变化趋势，任务表示作业中的算子。

作业数：所有批处理作业和实时作业的总数。

任务日调度数：按照当天调度成功的节点进行统计，不区分实时任务和离线任务。

- 系统默认查看一个月内的任务日调度数和作业数，支持通过时间段筛选进行查看。

- 任务类型分布

  可以直观地查看作业的任务节点类型分布图及数量。

说明

任务表示作业中的算子。

系统会统计已提交的所有作业节点数，含实时作业和批处理作业。

- 实例运行时长top100
  - 通过时间和责任人筛选出**我的**或者**全部**责任人的实例运行时长top100的数据。
  - 单击作业名称，可以跳转到实例监控界面，查看作业运行的详细信息。
  - 系统默认展示一个月的批处理作业实例运行时长数据。
- 实例运行失败top100
  - 通过时间和责任人筛选出**我的**或者**全部**责任人的实例运行失败top100的数据。
  - 单击作业名称，可以跳转到实例监控界面，查看作业运行的详细信息，查看作业实例运行失败的详细日志并分析原因。

- 系统默认展示一个月的批处理作业实例运行数据。

- 未来一周调度结束情况

  可以查看未来一周的作业调度结束的数据，包含作业名称、调度结束时间以及责任人。

  说明

  - 调度结束时间小于或等于2天，显示为红色。
  - 调度结束时间在3~5天，显示为橙色。
  - 调度结束时间在6~7天，显示为黑色。

9.9.2 作业监控

9.9.2.1 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、调度周期、调度开始时间等信息，以及进行如表9-69所示的操作。

图 9-81 批作业监控

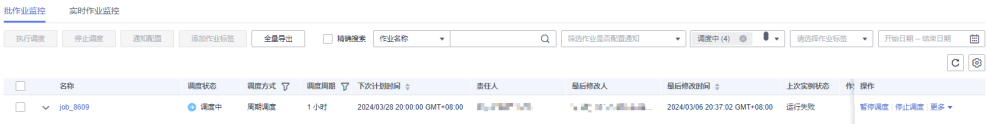


表 9-69 批作业监控支持的操作项

支持的操作项	说明
根据“作业名称”、“责任人”、“CDM作业”、“调度身份”或“节点类型”筛选作业	-
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	对于未配置通知的作业，系统支持可以通知类型（例如运行异常/失败）进行筛选，以便批量设置告警通知。
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。

支持的操作项	说明
查看作业实例状态	单击作业名称前方的，显示“最近的实例”信息，查看该作业最近的实例信息。 在最近的实例的“操作”列，可以查看作业实例的运行日志，重跑作业实例。 **说明** • 重跑的作业可能与正常调度的作业同时运行，需要确认作业是否支持并发执行；如果作业中节点个数或者名称发生变化，就会从第一个节点开始重跑。如果重跑成功状态的作业实例，就会从第一个节点开始重跑。 • 重跑作业实例时，需要选择“使用的作业参数”和“是否忽略OBS监听”。 使用的作业参数可设置为“使用原有作业参数重跑”或“使用最新提交作业参数重跑”。 是否忽略OBS监听默认为“是”。 • 企业模式下，开发者不能对作业实例进行重跑。
查看作业的节点信息	单击作业名称，在打开的页面中单击作业节点，查看该节点的相关关联作业/脚本与监控信息。 单击作业名称，在打开的页面中查看该作业的作业实例，详情请参见**批作业监控：作业实例**。
调度作业相关	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**。
通知配置	在作业的“操作”列，选择“更多 > 通知配置”，弹出“通知配置”页面，参考**表9-79**配置通知参数。
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。
调度配置	在作业的“操作”列，选择“更多 > 调度配置”，跳转到作业开发页面，查看该作业调度配置信息，可以对作业的调度信息进行配置。 **说明** 运行中的作业不支持配置调度操作。
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**。 只有配置为周期调度类型的作业才支持补数据功能。
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**。
查看作业依赖图	在作业的“操作”列，选择“更多 > 查看作业依赖图”，详情请参见**批作业监控：查看作业依赖图**。

支持的操作项	说明
全量导出	单击“全量导出”，进入到“导出全量数据”页面，单击“确认”。导出完成后，请到下载中心查看导出的内容。 如果没有配置默认存储路径，单击“批量导出”后，配置存储路径，可以将该存储路径设为OBS默认地址。 当前导出数据量最大为30M，超过30M系统会自动截断。 导出的作业实例与作业节点存在对应关系。目前不支持通过勾选作业名称导出所勾选的数据，可以通过筛选条件选择需要导出的数据。

单击作业名称，在打开的页面中查看该作业的作业参数、作业属性、作业实例。

单击作业的某个节点，可以查看节点属性、脚本内容、节点监控信息。

同时，您可以查看当前作业版本、作业调度状态、执行调度、停止调度、对运行中的作业暂停调度、补数据、通知配置、设置作业刷新频率等。

批作业监控：作业实例

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 单击“批作业监控”页签，进入批作业的监控页面。
5. 单击作业名称，在打开的页面中查看该作业的作业实例。您可以进行以下操作：
   - 当勾选上“显示尚未生成的实例”后，通过时间筛选未来时间内尚未生成的作业实例。

     说明

     勾选后进行筛选，能够显示未来时间内预计可能会生成的实例，显示的未生成实例数量不超过100个。
   - 对于未来时间内尚未生成的作业实例，可以进行“冻结”和“解冻”操作。您可以单击作业实例列表上面的“冻结”和“解冻”按钮，或者通过右侧操作列的“更多”中选择冻结和解冻进行冻结和解冻操作，支持批量操作。

     说明

     冻结：作业实例尚未生成或者作业实例是等待运行的状态，且实例未被冻结上，才能够进行冻结。

     已被冻结的作业实例，实例运行状态为冻结状态。

     作业被冻结后，会按照作业运行失败进行处理，下游依赖的相关作业，如果依赖设置的是挂起，则下游作业挂起；如果依赖设置的继续执行，则下游作业继续执行；如果依赖设置的取消，则下游作业取消执行。

     尚未生成的作业实例被冻结后，可以在批作业监控的作业实例中查看，也可以在实例监控中通过运行状态进行筛选去查看被冻结状态的实例。

     解冻：作业实例还未开始调度，且实例已被冻结，才能够进行解冻。
   - 对作业实例进行相关的其他操作，例如：停止、重跑、手工重试、继续执行、强制成功、查看作业等待实例，查看作业开发配置信息等操作。在查看

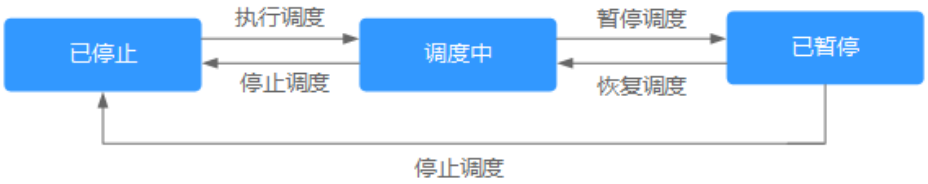
作业等待实例时，单击“操作 > 去除依赖”可以去除对上游单个实例的依赖关系。

- 手工确认执行场景下，在批作业监控页面，作业实例运行状态显示为“待确认执行”，可以进行手动确认执行，单击“确认执行”后，作业实例运行状态显示为“等待运行”。

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 9-82 调度作业



1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 单击“批作业监控”页签，进入批作业的监控页面。

   说明

   批作业监控支持按照调度方式和调度周期进行筛选，可以通过条件过滤查看所需要的作业调度实例。
5. 在作业的“操作”列，单击“执行调度”/“暂停调度”/“恢复调度”/“停止调度”。

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）**。

说明

如果该作业在基线任务链路上，暂停调度/停止调度时，系统会自动给出基线关联的弹窗提示。

如果该作业在基线任务链路上或者被其他作业依赖， 暂停调度/停止调度时，系统会自动给出弹窗提示。

图 9-83 启动作业

批作业监控：补数据

补数据是指作业执行一个调度任务，在过去某一段时间里生成一系列的实例。用户可以通过补数据，修正历史中出现数据错误的作业实例，或者构建更多的作业记录以便调试程序等。

只有配置了周期调度的作业，才支持使用该功能。如需查看补数据的执行情况，请参见**补数据监控**。

说明

当作业正在补数据时，请勿修改作业配置，否则会影响补数据过程中生成的作业实例。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 单击“批作业监控”页签，进入批作业的监控页面。
5. 在作业的“操作”列，选择“更多 > 补数据”。
6. 弹出“补数据”对话框，配置如**表9-70**所示的参数。

图 9-84 补数据参数

表 9-70 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。
调度时间方式	• 单段连续业务日期 补数据的时间是连续的业务日期时间段。 • 多段离散业务日期 补数据的时间是不连续的离散的业务日期时间段。
业务日期	**当“调度时间方式”选择为“单段连续业务日期”：** 选择需要补数据的时间段。业务日期不能大于当前时间，大于当前时间系统会默认显示当前时间。 **说明** 一个作业可进行多次补数据。但多次补数据的业务日期需要避免交叉重叠，否则可能导致数据重复或混乱，用户请谨慎操作。 如果勾选了“按日期倒序补数据”，则系统按照日期倒序补跑，每日内的补数顺序仍是正序。 **说明** • 该功能适合在各日数据不耦合的条件下使用。 • 为保证补数据可以倒序进行，补数据作业对更早日期作业实例的依赖关系将被忽略。 **当“调度时间方式”选择为“多段离散业务日期”：** 除了配置上面的业务日期参数，还需要配置以下补数据的参数： 单击“添加多段业务日期”可以添加多个离散的补数据的业务日期。您至少需要配置一个业务日期范围。 单击“删除”可以删除已添加的离散业务日期。 **说明** 因为DataArts Studio不支持底层服务（例如，以前的CDM、DLI等服务）的补数据实例和周期调度作业实例并发运行，为了保证补数据实例不影响周期调度作业实例运行，两种类型作业实例不会抢占并发，所以，作业的周期调度的日期与该作业补数据的业务日期不能重合，周期调度和补数据不能同时运行，避免出现运行异常问题。

参数	说明
是否设置周期补数据	• 是，补数据时会按照设置的周期进行补数据任务。 **第一个值**表示具体的值。 **第二个值**表示按指定周期补数据，例如：小时、天，周、月。 **说明** 设置周期后，将会按照周期进行补数据任务调度。对于调度周期为分钟，间隔小时以及天的任务，将按照新设置的周期去调度补数据任务，起始点为业务日期的第一个时间点。例如任务为每天1:00开始的小时任务，需要对2023/01/01 00:00 - 2023/02/01 00:00进行补数据操作, 周期为2天，则将调度2023/01/01 00:00，2023/01/03 00:00，2023/01/05 00:00……等任务。此外，当调度周期为月时候，如果第一个节点为月末最后一天，将默认调度每月最后一天。 • 否，补数据时不会按照周期进行补数据任务，默认原有的补数据规则进行补数据任务。
指定周期	当“调度时间方式”选择为“多段离散业务日期”时，需要配置此参数。 指定补数据的时间周期 通过“查看调度信息”可以查看当前时间段下任务实例执行时间。 **说明** 只有调度周期是小时调度和分钟调度时，进行离散补数据的时候才会有指定周期。
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。 如果补数据按照天粒度补数据，并行周期数就是在同一天内单个作业的实例并行数。 如果补数据不按照天粒度补数据，并行周期数就是按照调度周期内单个作业的实例并行数。 **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。
需要补数据的上下游作业	选择需要补数据的上下游作业（指依赖于当前作业的作业），支持多选。 此处系统会展示作业依赖关系视图，关于作业依赖关系视图的操作，请参考**批作业监控：查看作业依赖图**。 **说明** 周期补数据场景下，当前只允许针对调度周期相同的上下游作业进行补数据。

参数	说明
是否按天粒度补数据	如果选择了按天粒度去补数据，表示在同一天内单个作业补数据的实例可以并行去跑，不在同一天内的单个作业补数据的实例不能并行去跑。例如小时任务可以5点和6点的作业实例并行跑，而1号和2号的作业实例不能并行跑。 是：按天粒度补数据 否：不按天粒度补数据
失败后是否停止	如果“是否按天粒度补数据”选择“是”，需要配置此参数。 是：按天粒度补数据如果失败后，后面的补数据任务立即停止。 否：按天粒度补数据如果失败后，后面的补数据任务继续执行。 **说明** 按天粒度补数据，前面一天任务执行失败后，第二天补数据任务不再执行。系统仅支持按天维度的补数据状态进行判断，不支持同一天内小时任务的多批次场景。
优先级	选择补数据的优先级。通过**默认项配置**可以设置工作空间级的补作业的优先级。 **说明** 补数据的优先级高于工作空间的补数据优先级。 当前只支持对DLI SQL算子设置优先级。
是否忽略OBS监听	• 是，补数据场景下，系统会忽略OBS监听。 • 否，补数据场景下，系统会监听OBS路径。
是否设置运行时间段	设置补数据任务的运行时间段。 • 是 可以设置补数据任务每天运行的时间段。 • 否

7. 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 单击“批作业监控”页签，进入批作业的监控页面。
5. 在作业的“操作”列，选择“更多 > 添加作业标签”。
6. 弹出“添加作业标签”对话框，填写需要配置的作业标签。

图 9-85 添加作业标签参数

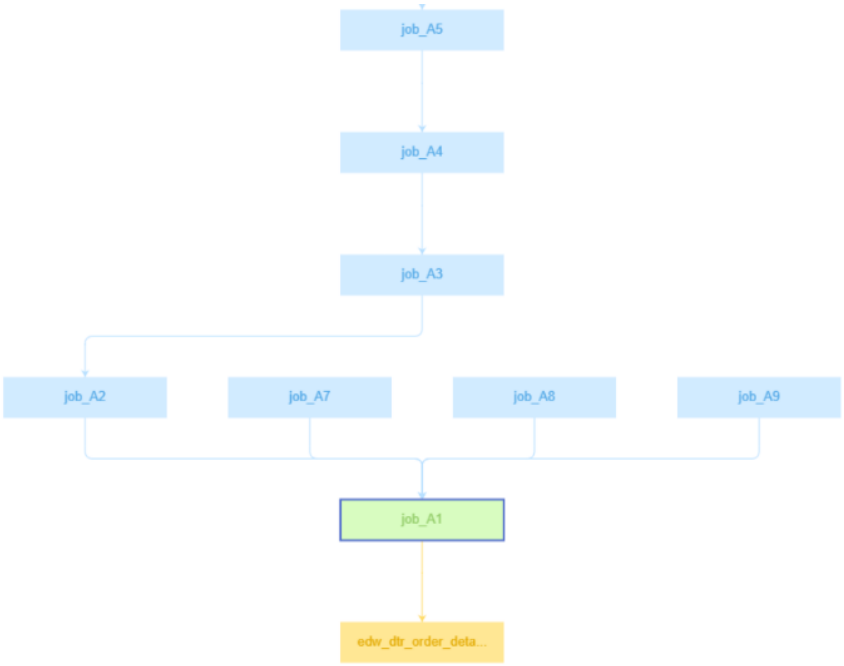
7. 填写完标签后，单击“确认”，完成作业标签的添加。

批作业监控：查看作业依赖图

作业依赖关系视图支持查看作业与其他作业的依赖关系。

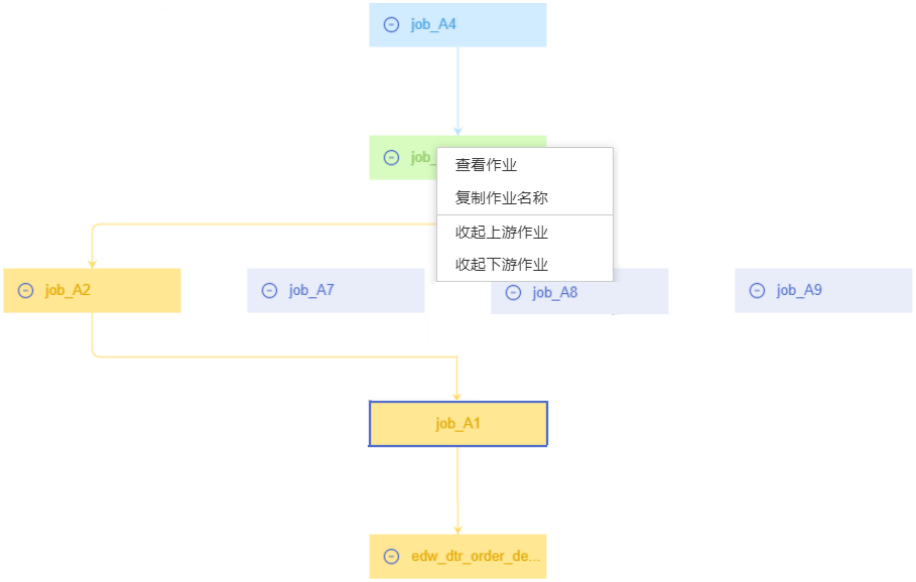
1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 单击“批作业监控”页签，进入批作业的监控页面。
5. 在作业的“操作”列，选择“更多 > 查看作业依赖关系图”。
6. 在弹出的“作业依赖关系视图”页面，支持如下操作：
   - 视图右上角支持“显示完整依赖图”、“显示当前作业及其上下游”和“显示当前作业及其直接上下游”。
   - 视图右上角支持按节点名称进行搜索，搜索出来的作业节点高亮显示。
   - 单击下载按钮，可以下载作业的依赖关系文件。
   - 鼠标滚轮可放大、缩小关系图。
   - 鼠标按住空白处，可自由拖拽用以查看完整关系图。
   - 鼠标光标悬停在作业节点上，该作业节点会被标记为绿色，上游作业会被标记为青蓝色，下游作业会被标记为橙黄色。

图 9-86 上下游作业节点标记



- 在作业节点上右键单击，可进行查看作业、复制作业名称、收起上/下游作业等操作。

图 9-87 作业节点操作



另外，作业的节点监控信息还可以通过作业详情查看。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。

4. 单击“批作业监控”页签，进入批作业的监控页面。
5. 单击作业名称，进入后单击作业节点。查看作业节点监控的详细信息。
   单击“编辑”，将进入该作业的开发页面。

9.9.2.2 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9-71所示的操作。

图 9-88 实时作业监控

表 9-71 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“作业名称”、“责任人”、“CDM作业”或“节点类型”筛选作业	-
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（启动、停止、添加作业标签）。
4	查看作业实例状态	单击作业名称前方的，显示“最近的实例”页面，查看该作业最近的实例信息。
5	作业状态相关	在作业的“操作”列，支持作业级别的启动、暂停、恢复、停止调度、重跑、添加作业标签等。
6	添加作业标签	单击“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。

序号	支持的操作项	说明
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	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。
12	查看日志	对于Flink SQL和Flink JAR两种实时处理的单任务作业，作业运行完成后，可以通过“更多 > 查看日志”一键跳转到日志查看页面查看Flink作业日志。 **说明** MRS集群版本为不支持时，界面不显示查看日志，则系统不支持通过一键跳转查看日志。

单击作业名称，在打开的页面中查看该作业的作业参数、作业属性、作业实例。

单击作业的某个节点，可以查看节点属性、脚本内容、节点监控等信息。在节点监控页签，可以查看实时作业的运行日志。

同时，您可以查看当前作业版本、作业运行状态、启动、重跑、作业开发、是否显示指标监控、设置作业刷新频率等。

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 选择“实时作业监控”页签，单击作业名称。
5. 进入“作业监控”详情页面后，右键单击节点，选择“禁用”。
6. 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9-72**所示的参数。

图 9-89 恢复操作

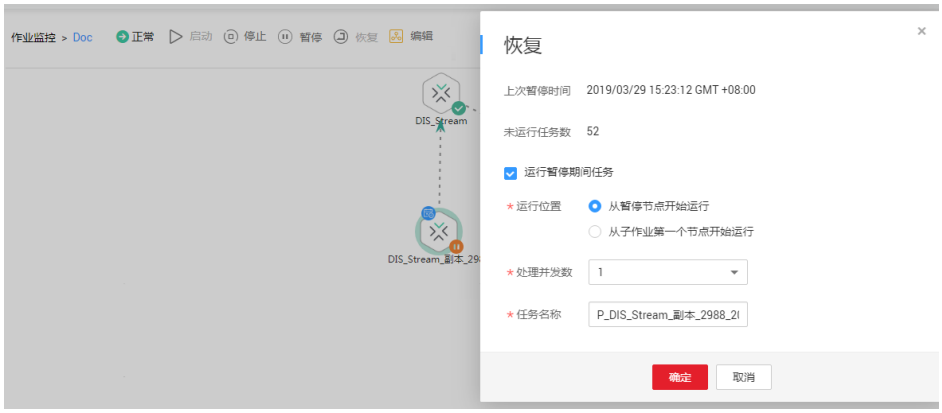


表 9-72 恢复参数说明

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
4. 选择“实时作业监控”页签，单击作业名称。
5. 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9-73**所示的参数。

图 9-90 调度配置

表 9-73 调度配策略参数说明

参数	说明
DIS通道名称	选择DIS通道，当指定的DIS通道有新消息时，数据开发模块将新消息传递给作业，触发该作业运行。
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

图 9-91 DIS 调度策略配置

9.9.2.3 实时集成作业监控

实时集成作业监控提供了对实时处理集成作业的状态进行监控的能力。

实时处理集成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 > 实时集成作业监控”页面查看实时处理作业的运行状态、运行时间、运行耗时等信息，以及进行如**表9-74**所示的操作。

图 9-92 实时集成作业监控

表 9-74 实时集成作业监控支持的操作项

序号	支持的操作项	说明
1	启动	支持批量启动作业。启动操作请参见**实时集成作业监控：启动**。
2	停止	支持批量停止作业。停止操作请参见**实时集成作业监控：停止**。
3	根据“状态”筛选作业	通过对作业运行状态进行筛选，查看不同运行状态下的集成作业。
4	按照作业名称搜索	通过作业名称搜索相关作业，支持模糊搜索。
5	操作项相关	在作业的“操作”列，支持作业级别的启动。启动操作请参见**实时集成作业监控：启动**。 在作业的“操作”列，支持作业级别的停止。停止操作请参见**实时集成作业监控：停止**。 在作业的“操作”列，支持作业级别的恢复。停止操作请参见**实时集成作业监控：恢复**。
6	查看作业实例状态	单击作业名称前方的，查看该作业下的子作业ID、源端数据源、目的端数据源、异常信息等内容。
7	查看作业详细信息	单击作业名称，支持查看该作业的基本信息、监控信息、日志信息。查看作业详细信息请参见**实时集成作业监控：查看作业详细信息**。

实时集成作业监控：启动

1. 单击“启动”，弹出“启动配置”界面。
2. 设置“同步模式”和“时间”。

说明

同步模式包含增量同步和全量同步。

时间表示配置的位点时间早于日志最早时间点时，会以日志最早时间点消费。当设置为“增量同步”时才显示时间参数。

3. 单击“确定”，启动该任务。

实时集成作业监控：停止

对于运行状态异常的实时集成作业，可以进行停止操作。

1. 单击“停止”，系统弹出停止任务的提示框。
2. 单击“确认”，停止该任务。

实时集成作业监控：恢复

对于运行状态异常的实时集成作业，可以进行恢复操作。

1. 单击“恢复”。
2. 系统提示“操作成功”，任务恢复成功。

实时集成作业监控：查看作业详细信息

单击作业名称，可以查看该作业的详细信息。

- 选择“基本信息”，查看该作业的基本信息。
- 选择“监控信息”，查看该作业的监控信息。
  - 单击“查看监控指标”，进入云监控服务界面查看该作业的相关监控指标。
  - 单击“创建告警规则”，进入云监控服务的创建告警规则界面，创建该作业的告警规则。
  - 查看已创建的告警规则，包含“名称/ID”、“告警策略”。
  - 查看作业同步进度信息。
- 选择“日志信息”，查看该作业的详细日志信息，对日志进行下载。

9.9.3 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”、“责任人”、“CDM作业”、“节点类型”和“作业标签”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。同时，支持通过“运行状态”和“调度方式”进行筛选作业实例。

作业实例操作

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 实例监控”。

4. 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9-75**。

   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。

5. 在实例列表中，提供如**表9-75**所示的操作。

表 9-75 实例监控操作

操作项	说明
根据“作业名称”、“创建人”或“责任人”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”、“节点类型”或“作业标签”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**。 **说明** • 手动调度的作业任务不支持重跑。 • 企业模式下，开发者不能对作业实例进行重跑。 手工确认执行场景下，重跑实例时，作业实例运行状态显示为“待确认执行”，可以进行手动确认执行，单击“确认执行”后，作业实例运行状态显示为“等待运行”。
手工重试	对于实例的状态为“运行异常”时，支持批量进行手工重试。
继续执行	对于实例的状态为“运行异常”时，支持批量操作，继续运行实例中的后续节点。
强制成功	对于实例的状态为“运行异常”、“取消”、“失败”时，可以批量操作，将运行状态改为“成功”，实例状态显示为“强制成功”。
确认执行	对于实例的状态为“待确认执行”时，支持批量进行手工确认执行。
强制解除依赖执行	可以对有依赖关系的作业实例批量选中进行强制解除依赖执行。
更多 > 手工重试	对于实例的状态为“运行异常”时，支持进行手工重试。

操作项	说明
更多 > 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。单击“操作 > 去除依赖”可以去除对上游单个实例的依赖关系。
更多 > 确认执行	对于实例的状态为“待确认执行”时，支持进行手工确认执行。
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 > 强制解除依赖执行	可以对有依赖关系的作业实例进行强制解除依赖执行。
更多 > 查看	跳转至作业开发页面，查看作业信息。
更多 > 历史性能	可以查看作业实例监控的历史性能折线图。
更多 > 查看重跑历史	可以查看作业实例重跑的历史记录。 当重跑次数大于0时，才能查看作业实例重跑历史记录。
更多 > 强制优先执行	可以对作业实例进行强制优先执行。
DAG	弹出DAG图，便于直观查看作业实例之间的依赖关系，并且支持在DAG图上进行运维操作。 详细操作请参见**查看DAG图**。
全量导出	单击“全量导出”，进入到“导出全量数据”页面，单击“确认”。导出完成后，请到下载中心查看导出的内容。 如果没有配置默认存储路径，单击“批量导出”后，配置存储路径，可以将该存储路径设为OBS默认地址。 当前导出数据量最大为30M，超过30M系统会自动截断。 导出的作业实例与作业节点存在对应关系。目前不支持通过勾选作业名称导出所勾选的数据，可以通过筛选条件选择需要导出的数据。

6. 单击实例前方的，显示该实例所有节点的运行记录。
7. 在节点的“操作”列，提供如**表9-76**所示的操作。

表 9-76 操作（节点）

操作项	说明
查看日志	查看节点的日志信息。 进行作业手动测试运行时，作业测试运行日志查看有权限管控，比如，用户A进行作业测试运行后，可以在“实例监控”页面查看测试运行日志，不允许用户B查看该测试运行日志。
手工重试	节点的状态为“失败”时，支持重新运行节点。 节点的状态为“运行异常”时，支持进行手工重试。 **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	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。 **说明** 若实例为单节点实例，不支持跳过操作。为多节点实例支持跳过操作。
更多 > 暂停	作业的实例状态是运行中，节点的状态是等待运行的时候，支持暂停该节点，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。
更多 > 历史性能	可以查看作业节点的历史性能折线图。

重跑作业实例

说明

企业模式下，开发者不能对作业实例进行重跑。

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 实例监控”。
4. 在作业所在的“操作”列，单击“重跑”设置重跑当前作业实例；或单击作业名称左边的复选框，再选择页面上方的“重跑”按钮可以批量设置多个作业的实例重跑。

图 9-93 设置单个作业重跑

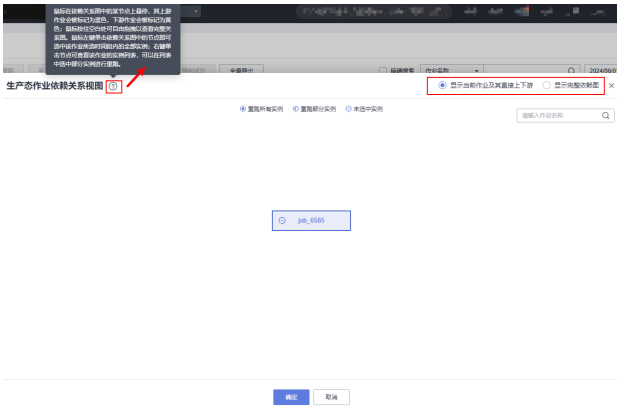
图 9-94 批量设置作业重跑

说明

批量设置多个作业实例重跑时，仅需要配置重跑开始位置、使用的作业参数、是否忽略OBS监听等参数。

表 9-77 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 ● 重跑当前作业及其上下游作业实例
开始时间	仅当“重跑类型”选择“重跑当前作业及其上下游作业实例”时，才需要配置。 设置好开始时间和结束时间，系统会重跑所设置的时间段内的作业实例。 **说明** 如果所选的时间段内没有可以重跑的作业实例，系统会报错“Job xxx have no instances to rerun”。

参数	说明
重跑作业实例列表	仅当“重跑类型”选择“重跑当前作业及其上下游作业实例”时，才需要配置。 作业依赖关系视图可设置为“显示当前作业及其直接上下游”或“显示完整依赖图”。 此处系统会展示作业依赖关系视图，支持输入作业名称进行查询。 **图 9-95** 作业依赖关系视图  选择需要重跑的当前作业及其上下游作业，支持多选。

参数	说明
	说明 鼠标放置于作业依赖关系视图右边的按钮上，会显示如下信息： • 鼠标在依赖关系图中的某节点上悬停，其上游作业会被标记为蓝色，下游作业会被标记为黄色。 • 鼠标按住空白处可自由拖拽以查看完整关系图。 • 鼠标左键单击依赖关系图中的节点即可选中该作业所选时间段内的全部实例，即重跑该作业的所有实例。 **图 9-96** 重跑所有实例  • 右键单击节点可查看该作业的实例列表，可以在列表中选中部分实例进行重跑，即重跑该作业的部分实例。 **图 9-97** 重跑部分实例  • 如果还未选中任何作业实例，系统会显示未选中实例。 **图 9-98** 未选中实例  

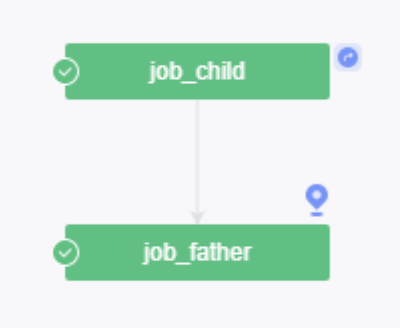
参数	说明
	关于作业依赖关系视图的详细操作，请参考**批作业监控：查看作业依赖图**。
重跑开始位置	选择作业实例重跑的开始位置。 • 从错误节点开始重跑：作业实例执行失败时，从实例执行失败的错误节点开始重跑。 • 从第一个节点开始重跑：从作业实例的第一个节点开始重跑。 • 从指定的节点开始重跑：从作业实例中指定的节点开始重跑。仅当“重跑类型”选择“重跑当前实例”时有此选项。 **说明** 以下两种情况，系统运行会从第一个节点开始重跑。 • 如果作业中节点个数或者名称发生变化，从第一个节点开始重跑。 • 如果重跑成功状态的作业实例，从第一个节点开始重跑。
使用的作业参数	• 使用原有作业参数重跑 • 使用最新提交作业参数重跑
处理并发数	仅当“重跑类型”选择“重跑当前作业及其上下游作业实例”时，才需要配置。 设置作业实例并行处理的数量，输入值不能小于1。默认值为1。
是否忽略OBS监听	系统默认为“是”。 • 是，重跑作业实例场景下，系统会忽略OBS监听。 • 否，重跑作业实例场景下，系统会监听OBS路径。 **说明** 若暂未使用该参数，可忽略。

查看 DAG 图

您可以查看作业实例之间的依赖关系，并且在DAG图上进行运维操作。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 实例监控”。
4. 选择作业名称，在作业的“操作”列，单击“DAG”，系统弹出DAG视图。

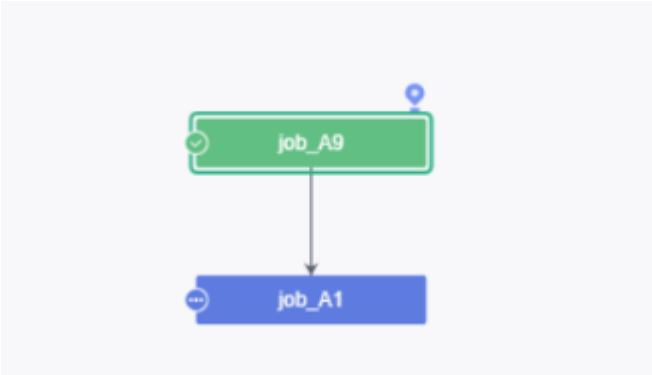
图 9-99 DAG 视图



DAG视图默认展示当前作业实例及上下游作业实例，并支持如下操作：

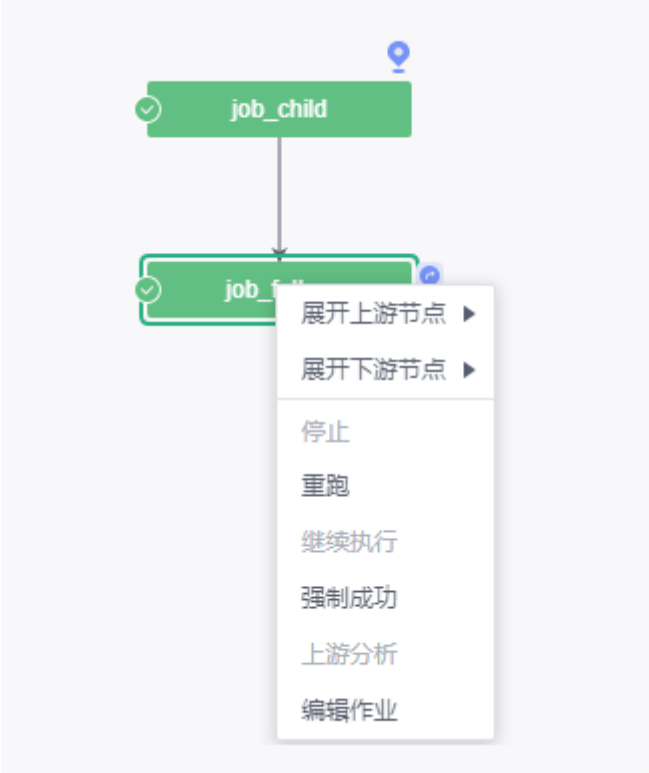
- DAG视图右上角表示恢复DAG图初始状态，表示关闭，单击可以关闭。左侧侧边图标可以拖动改变视图宽度。
- 单击可以选中某个作业实例：

  图 9-100 选中作业实例

  - 选中时，该作业实例及其上下游实例的背景颜色加深显示。
  - DAG视图右下角展示该实例的概要信息，且实例名称和实例ID支持直接复制。
  - 单击概览信息的“展开详情”打开详情面板，详情面板包含实例属性、作业参数、节点列表、历史实例等信息，支持调整高度并关闭详情面板。
  - 单击空白处，即可取消选中效果。
- 右键单击某个作业实例，可以展开该实例上下游的作业实例，并支持进行停止、重跑、继续执行、强制成功、上游分析、编辑作业等实例操作。

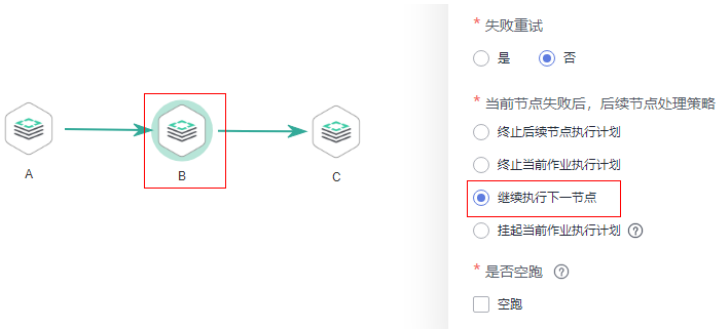
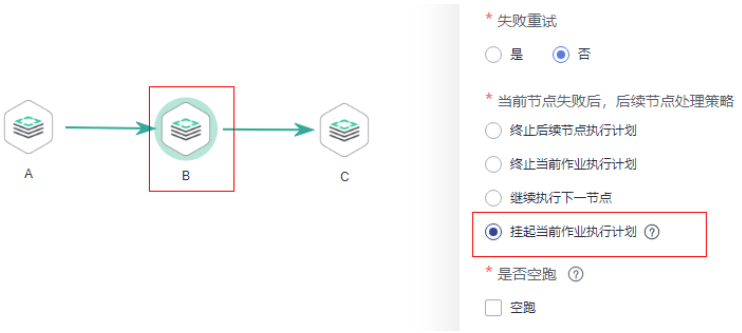
图 9-101 操作作业实例



作业实例运行状态

表 9-78 作业实例运行状态说明

运行状态	场景描述
等待运行	如果作业实例依赖的前置作业实例未最终完成（未最终完成的状态包括：未生成实例、等待运行、运行失败），该实例处于等待运行。
运行中	作业正常运行中。说明前置的依赖作业都已完成，该作业调度时间已到。
运行成功	作业真正成功执行了业务逻辑，并且最终成功（包含失败重试的成功）。 “运行成功”包括了“成功”、“强制成功”、“忽略失败”三种运行状态。
强制成功	作业实例处于失败或取消状态时，进行手动执行强制成功。

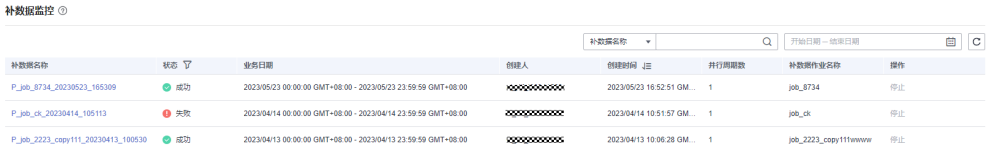
运行状态	场景描述
忽略失败成功	如下图所示，节点B设置了失败处理策略，当B执行失败了，会跳过B继续执行C，当存在这种节点运行失败，整个作业执行完成了就是忽略失败成功。 **图 9-102** 失败处理策略-继续执行下一节点 
运行异常	这种运行状态场景较少。如下图所示，节点B设置了失败处理策略，当B执行失败了，作业实例立即挂起，不会继续执行C，作业实例进入异常运行状态。 **图 9-103** 失败处理策略-挂起当前作业执行计划 
已暂停	这种运行状态场景较少。当某个作业的实例正在运行，测试人员在作业监控界面，手工暂停作业调度。此时，该作业正在运行的实例会进入已暂停状态。
已取消	• 等待运行状态的作业实例，进行手工停止，则实例处于已取消状态。 • 如果作业实例依赖的直接上游作业被停止调度了，该作业实例会自动进入已取消状态。作业A依赖作业B，作业B被停止调度，作业A实例生成后会自动取消。
冻结	对于未来时间内尚未生成的作业实例，进行冻结后，该作业实例会进入冻结状态。
失败	作业执行失败。执行失败的作业，可以查看失败原因，比如作业的哪个节点执行失败。

9.9.4 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。

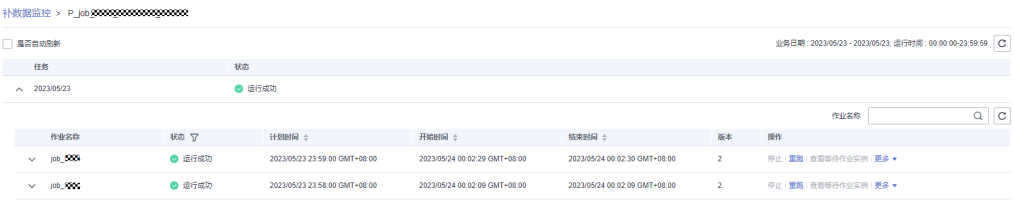
用户可以在图9-104的页面中，查看补数据的任务状态、业务日期、并行周期数、补数据作业名称、创建人、创建时间以及停止运行中的任务。系统支持按补数据名称、创建人、日期和状态进行筛选。

图 9-104 补数据监控主页



在图9-104的页面中，单击补数据名称，进入图9-105的页面。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**）。

图 9-105 补数据监控详情



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。
- 排序按钮单击顺序为：单击1下为升序，单击2下为降序，单击3下取消排序。
- 在查看作业等待实例时，单击“操作 > 去除依赖”可以去除对上游单个实例的依赖关系。
- 在补数据失败的情况下，单击“操作 > 停止”，补数据任务会停止。
- 补数据监控详情页面，每批补数据支持通过作业名称进行模糊筛选。
- 手工确认执行场景下，在进行补数据时，在补数据监控页面，补数据作业实例运行状态显示为“待确认执行”，可以进行手动确认执行，单击“确认执行”后，补数据作业实例运行状态显示为“等待运行”。

9.9.5 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。

9.9.5.1 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 通知管理”。
4. 在“通知管理”页签，单击“通知配置”，弹出“通知配置”页面，配置如**表9-79**所示的参数。

图 9-106 通知配置

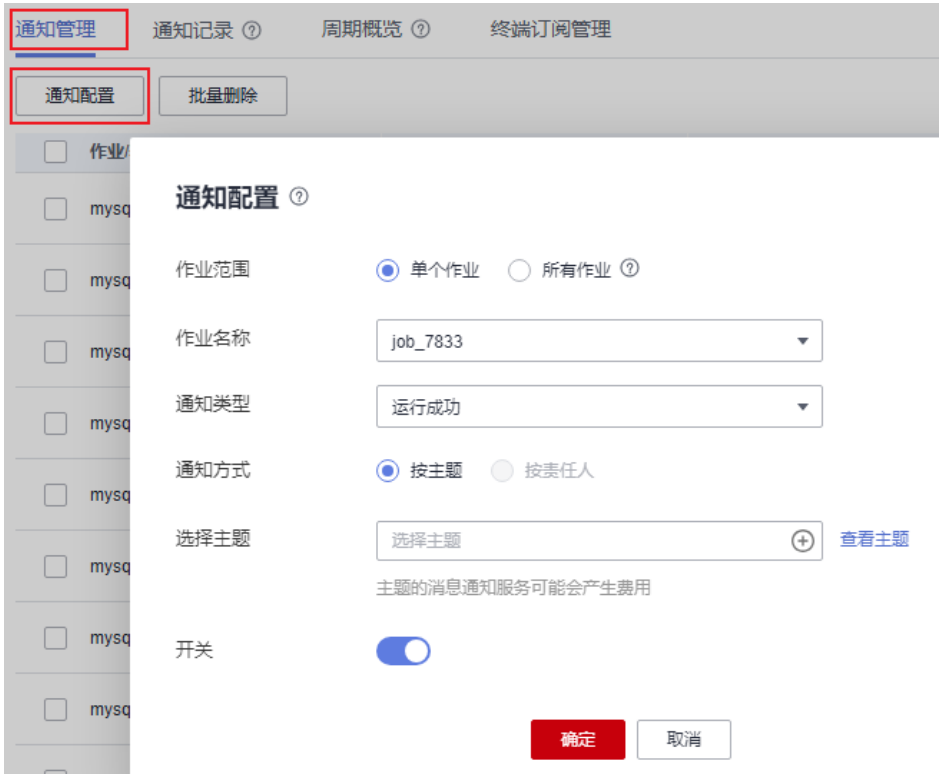


表 9-79 通知参数

参数	是否必选	说明
作业范围	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。所有作业指当前已有的作业和后续新创建的作业会使用这个通知配置。

参数	是否必选	说明
作业名称	是	选择作业。

参数	是否必选	说明
通知类型	是	选择通知类型： ● 单个作业： – 运行异常/失败：作业的状态为“运行异常”或“失败”时，发送通知。 系统支持配置“最大通知次数”和“最小通知间隔（分钟）”，作业运行异常或者失败后，在作业未修复前，可以设置间隔时间发送告警通知。 **说明** 最大通知次数可设置为1~50。默认为1时，最小通知间隔不显示。 最小通知间隔可设置为5~60。 – 运行成功：作业的状态为“成功”时，发送通知。 – 未完成：该功能仅支持按天调度的作业配置。如果作业执行时间超过设置的未完成时间，则发送通知。 – 运行取消：作业的状态为“已取消”时，则发送通知。 **说明** 调度中的作业手动停止调度时触发告警通知，运行中的作业实例手动停止时触发告警通知。 在作业执行人和作业取消人不一致的场景下，会触发作业取消告警通知。 – 失败作业重跑成功 **说明** 失败重跑后成功且失败的时候发送过失败告警才予以通知，如果作业失败不配置失败通知，则该当失败作业重跑成功后也不会进行通知。 – 作业改动 除了作业责任人外，其他人对作业进行改动（修改作业、删除作业、修改作业引用的脚本、删除作业引用的脚本）时，则发送通知。作业责任人为空时，作业改动也不会发送告警通知。 – 资源繁忙：如果执行作业时，DLI资源队列繁忙时，会遇到作业执行时间过长或无法执行的情况，从而发出告警，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。 系统支持配置“最大通知次数”和“最小通知间隔（分钟）”，作业运行异常或者失败后，在作业未修复前，可以设置间隔时间发送告警通知。

参数	是否必选	说明
		说明 最大通知次数可设置为1~50。默认为1时，最小通知间隔不显示。 最小通知间隔可设置为5~60。 – 运行取消：作业的状态为“已取消”时，则发送通知。 **说明** 调度中的作业手动停止调度时触发告警通知，运行中的作业实例手动停止时触发告警通知。 在作业执行人和作业取消人不一致的场景下，会触发作业取消告警通知。 – 失败作业重跑成功 **说明** 失败重跑后成功且失败的时候发送过失败告警才予以通知，如果作业失败不配置失败通知，则该当失败作业重跑成功后也不会进行通知。 – 作业改动 除了作业责任人外，其他人对作业进行改动（修改作业、删除作业、修改作业引用的脚本、删除作业引用的脚本）时，则发送通知。作业责任人为空时，作业改动也不会发送告警通知。 – 资源繁忙：如果执行作业时，DLI资源队列繁忙时，会遇到作业执行时间过长或无法执行的情况，从而发出告警，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。 ● 通常使用默认资源队列时，由于DLI的资源队列繁忙，用户间可能会出现抢占资源的情况，不能保证每次都可以得到资源执行相关操作。建议您在业务低峰期再次重试，或选择自建队列运行业务。 ● 作业运行成功时，在补数据、测试运行场景下不发送告警通知，避免邮件或短信轰炸。同时，补数据作业实例恢复时也不发送恢复通知。 ● 作业运行失败时，重跑作业并且作业运行成功后，会发送作业实例恢复通知。
通知方式	是	● 按主题 ● 按责任人

参数	是否必选	说明
选择主题	是	通知方式选择“按主题”时才需配置。 选择通知的消息主题。 单击“查看主题”，可以进入消息通知服务（SMN）界面查看消息主题信息。 **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。
终端协议	是	配置该参数前，请确保工作空间默认项设置中已配置**作业告警通知主题**。 通知方式选择“按责任人”时才需配置。 ● 短信 ● 邮件 ● 电话 单击“校验联系方式”，系统会自动校验作业责任人信息是否已配置。如果作业责任人信息未配置，请前往**终端订阅管理**界面进行配置。 单击“查看订阅信息”，会自动跳转到**终端订阅管理**界面查看已配置的终端订阅信息。 **说明** 终端协议为电话和短信时，依赖SMN服务给登录用户开通白名单，否则添加订阅会失败，可能会导致告警通知发送失败。
抄送人	是	通知方式选择“按责任人”时才需配置。 最多只能选择10个抄送人。
开关	是	是否开启通知，默认开启。

5. 单击“确定”，为作业配置通知。

   说明

   - 数据开发模块的通知管理功能是通过消息通知服务来发送消息，消息通知服务的使用可能会产生费用，具体请咨询消息通知服务。
   - 一个作业支持配置多个消息主题，当作业运行成功或失败，可同时向多个订阅了消息主题的终端发送通知。

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. 选择“通知管理”页签。
3. 在通知的“操作”列，单击“编辑”，弹出“通知配置”页面，参考**表9-79**修改通知的参数。

图 9-107 编辑通知

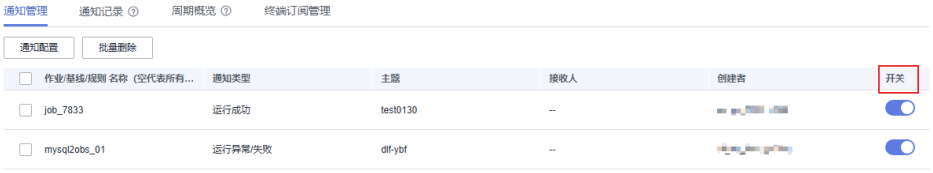
4. 单击“确定”，保存修改。

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. 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。

图 9-108 关闭通知



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. 选择“通知记录”页签，进入通知记录页面。系统只能查看最近30天的数据。

图 9-109 查看通知记录

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. 选择“通知管理”页签。
3. 支持如下两种方式删除通知。

图 9-110 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。

9.9.5.2 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 通知管理”。
4. 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9-80**所示的参数。

图 9-111 通知配置

表 9-80 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

参数	是否必选	说明
选择时间	是	当“调度周期”选择为“周”或者“月”时，才需要配置。 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。
选择主题	是	设置通知发送的主题。
选择OBS桶	是	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。

5. 单击“确定”。

   说明

   数据开发模块的通知管理功能是通过消息通知服务来发送消息，消息通知服务的使用可能会产生费用，具体请咨询消息通知服务。

6. 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。

9.9.5.3 终端订阅管理

操作场景

系统支持按照责任人配置终端订阅信息（短信、邮件、电话），配置好订阅信息后，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已配置的责任人发送通知。

前提条件

已开通消息通知服务并配置主题。按照责任人配置订阅信息前，请确保已在工作空间配置了**作业告警通知主题**。

配置通知

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

3. 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。单击“默认项设置”，设置“作业告警通知主题”配置项。按责任人配置工作空间作业告警通知主题的详细操作请参见**作业告警通知主题**。如果已配置，请忽略。

图 9-112 配置作业告警通知主题

4. 在数据开发主界面的左侧导航栏，选择“运维调度 > 通知管理”。
5. 选择“终端订阅管理”页签，单击“添加订阅”，弹出“添加订阅”页面，配置如**表9-81**所示的参数。

图 9-113 添加订阅

表 9-81 添加订阅参数

参数	是否必选	说明
责任人	是	设置添加订阅的责任人。责任人是创建作业时所配置的责任人信息。
终端协议	是	• 短信 • 邮件 • 电话
终端信息	是	设置订阅的终端信息。

6. 单击“确定”。
7. 终端订阅配置完成后，您可以在通知的“操作”列进行如下操作。
   - 单击“请求订阅”，打开“请求订阅”页面，订阅状态为“未确认”，单击“确定”，确认订阅后，订阅状态为“已确认”。
   - 单击“删除”，打开“删除订阅”页面，选择“确定”，删除通知。

     说明

     终端订阅管理页面支持批量删除和批量请求订阅，不支持编辑。
8. 以上操作完成后，请在**管理通知**界面按照责任人配置作业运行告警通知。

9.9.6 备份管理

通过备份功能，您可定时备份系统中的所有作业、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。
- 当前备份内容不会自动老化删除，您需要定期手动清理备份文件。

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“备份管理”。
4. 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。

图 9-114 备份管理

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 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤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 9-115 还原资产

步骤4 单击“确定”。

----结束

9.9.7 操作历史

通过操作历史可以查看数据开发的历史操作数据。系统最多保存最近三个月的历史数据，同时会自动清理三个月之前的更老的数据记录。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“运维调度 > 操作历史”。
4. 查看操作历史记录相关数据。
   - 可以通过时间筛选，查看指定操作时间段内的历史操作数据。
   - 可以对“涉及对象”进行过滤，查看作业名称或节点名称相关的历史操作数据。
   - 可以通过模糊查询，查看相关的历史操作数据。
   - 可以对“操作对象”、“操作类型”、“操作人”和“状态”进行过滤，查看相关的历史操作数据。

9.10 配置管理

9.10.1 配置

9.10.1.1 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。

说明

简单模式和企业模式下，配置工作空间的环境变量的角色有所不同：

简单模式：工作空间的环境变量开发者和管理员都能创建或编辑环境变量。简单模式不区分开发和生产环境，环境变量是共用的，允许开发者修改。

企业模式：工作空间的环境变量只有管理员才能创建或编辑环境变量。

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 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 单击“环境变量”，在“环境变量配置”页面，选择“导入”。

步骤5 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 9-116 导入环境变量

----结束

导出环境变量

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 单击“环境变量”，在“环境变量配置”页面，选择“导出”，可将环境变量导出到本地。

----结束

配置方法

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 单击“环境变量”，在“环境变量配置”页面，配置如**表9-82**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。

图 9-117 环境变量配置

表 9-82 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。 参数名称需根据**脚本变量定义**中设置的格式来命名。例如，脚本变量定义中设置的格式为${dlf.}，参数名称需要设置为dlf.*xxx*。
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；如果参数是字符串类型需要加上英文的双引号（""），如"05"。 关于EL表达式的使用，请参见**表达式概述**。
描述	否	参数说明。

配置完一个环境变量后，您还可以进行新增、修改、删除、重置等操作。

- 新增：单击“新增”配置新的环境变量。新增变量时，界面会提示新增变量。

- 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”。修改变量时，界面会提示修改变量。
- 删除：在参数值文本框后方，单击操作列的“删除”，删除环境变量。删除变量时，界面会提示删除变量。
- 重置：在修改或者删除配置时，单击操作列的“重置”，可以将参数值重置到之前的配置。

----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. #{Env.get(“环境变量名”)}

操作示例

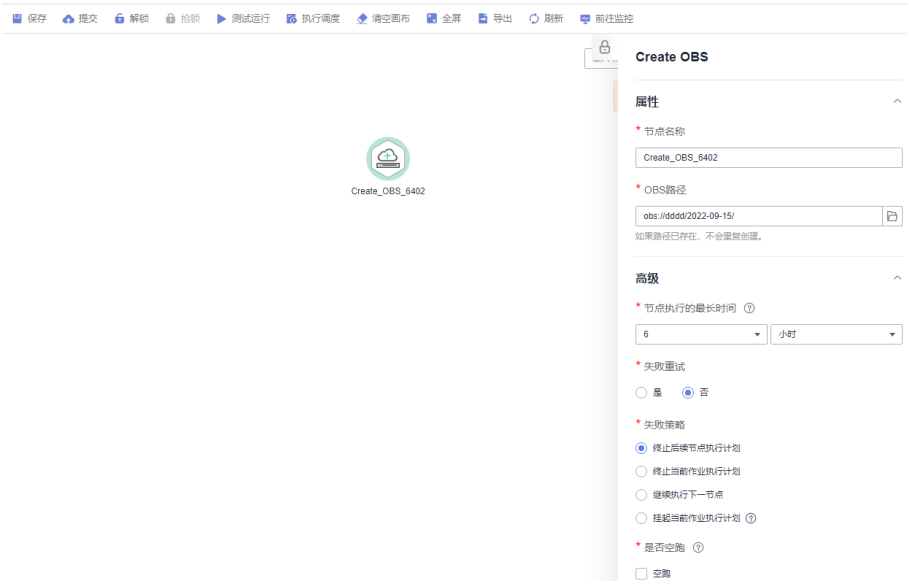
背景信息：

- 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 9-118 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

9.10.1.2 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。

约束限制

OBS路径仅支持OBS桶，不支持并行文件系统。

配置方法

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 选择“OBS桶”。

步骤5 配置OBS桶的信息。

图 9-119 配置 OBS 桶

步骤6 单击“保存”，完成配置。

----结束

9.10.1.3 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、删除、导入、导出等操作。

新建作业标签

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 选择“作业标签”，进入“作业标签管理”页面。

步骤5 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

删除作业标签

步骤1 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤2 选择“作业标签”，进入“作业标签管理”页面。

步骤3 单击作业标签名对应的“删除”，弹出删除确认对话框，单击“确认”即可删除该标签。

说明

当作业标签处于“锁定”状态时，无法删除该作业标签。如需解锁作业标签，请参考**锁定与解锁作业标签**。

----结束

监控某个作业标签下的作业

步骤1 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤2 选择“作业标签”，进入“作业标签管理”页面。

步骤3 单击作业标签名对应的“前往监控”，可进入作业监控界面，该界面展示具有此标签的所有作业。

----结束

锁定与解锁作业标签

具有**DAYU Administrator**或**Tenant Administrator**账号、空间管理员权限的用户才能锁定或解锁作业标签。

步骤1 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤2 选择“作业标签”，进入“作业标签管理”页面。

步骤3 单击作业标签列表中，标签名称所在行对应的“锁定”或“解锁”按钮，可对作业标签进行锁定或解锁操作。

说明

- 当作业标签是“锁定”状态时，不能被删除。
- 当作业标签是“锁定”状态时，导入该标签会失败。
- 当作业标签是“锁定”状态时，作业也不能添加或移除该标签。
- 导入作业时，如果作业中存在“锁定”状态的标签，则作业导入会失败。
- 当作业导入失败需要自动生成标签时，如果标签已存在且被锁定，则导入失败的作业不会添加上该标签。

----结束

导入作业标签

具有 **Administrator**或**Tenant Administrator**账号、空间管理员权限的用户才能导入作业标签。

步骤1 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤2 选择“作业标签”，进入“作业标签管理”页面。

步骤3 单击“导入作业标签”，弹出“导入作业标签”对话框。

步骤4 配置导入信息。

- 文件位置：支持从本地导入和从OBS导入两种方式。
- 选择文件：本地导入的文件选择本地路径；OBS导入的文件选择OBS桶路径。

  说明

  - 建议通过导出标签功能获取导入文件，导入文件的第一行为标签名，第一列为作业名。某作业具有某一标签，记录为1，否则记录为0。如果某单元格为空，导入时系统会按0标记。
  - 导入的文件大小最大支持10Mb。
  - 如果导入的标签名有重复，且标签标识一个为0，一个为1，系统会按1处理。
  - 如果导入的作业名有重复，系统会按后面一列来识别，标签标识按照该行来处理。
- 添加方式：支持追加和覆盖两种。
  - 追加：若该作业已设置了作业标签，新添加的标签不会覆盖原来的标签。
  - 覆盖：若该作业已设置了作业标签，新添加的标签将会直接覆盖原来的标签。

步骤5 单击“确定”，完成导入。

----结束

导出作业标签

步骤1 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤2 选择“作业标签”，进入“作业标签管理”页面。

步骤3 导出标签。

- 导出全部标签：单击标签列表上方的“导出全部标签”，可将所有标签导出。
- 导出选中标签：勾选本页需要导出的标签，单击标签列表上方的“导出选中标签”，可将本页选中的标签导出。

导出的作业标签如下图所示：

图 9-120 导出作业标签

	A	B	C	D	E	F
1	jobName	DWS_TRANSFORM	Invalid clust	The MRS cluster na	The cluster associated w	The cluster associat
2	job_foreach	1	0	1	0	1
3	job_real	0	0	1	1	0
4	job_weituo	0	1	0	0	1
5	job_subjob	1	1	0	1	0
6	job_ETL_dli2dws	0	0	1	1	1
7	job_foreach_copy	1	0	0	1	1
8	job_ETL_copy	0	0	1	0	0
9	guowangTest	1	1	0	0	1
10	guowangTest_qjxtest	1	0	0	0	0
11	qjxForeach	1	0	1	1	1
12	job_timlx_1	0	1	1	0	0
13	job_timlx_2	0	0	0	0	1
14	guowangTest_copy_hdfs2hiv	0	0	1	0	0

说明

- 导出的作业标签表格中，第一行为标签名称，第一列为作业名称。某作业具有某一标签，记录为1，否则记录为0。
- 导出的文件第一列将该空间下所有的作业名都展示出来，包括实时作业的节点、Foreach子作业、Subjob子作业。

----结束

9.10.1.4 配置调度身份

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账号在调度周期内被停用或删除后，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，则可配置作业调度身份。配置作业调度身份后，作业执行过程中，以配置的调度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。

说明

在作业进行周期调度时，该作业的默认用户被删除后，如果使用其他用户对该作业进行版本提交并执行调度，那该作业的执行用户就默认为提交版本的用户。

调度身份的分类

调度身份分为委托和IAM账户两大类。

- 委托：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。

  委托可以分为：

  - 公共委托：工作空间级别的全局委托。适用于该空间内的所有作业。配置公共委托请参考**配置公共委托**。
  - 作业委托：适用于单个作业级别。配置作业委托请参考**配置作业委托**。
- IAM账号：通过用户组统一配置，权限管理相对于委托来说，流程简便；并且使用IAM账号的兼容性更好，可支持MRS相关的节点（MRS Presto SQL、MRS Spark、MRS Spark Python、MRS Flink Job、 MRS MapReduce），通过直连方式的（MRS Spark SQL、MRS Hive SQL）节点，以及目标端为DWS的ETL Job节点，解决部分MRS集群和部分ETL Job节点不支持委托方式提交作业的问题。

  IAM账户可分为：

  - 公共IAM账户：工作空间级别的全局IAM账户。适用于该空间内的所有作业。配置公共IAM账户请参考**配置公共IAM账号**。
  - 执行用户：作业级别的IAM账户，适用于单个作业级别。配置执行用户请参考**配置执行用户**。

    说明

    配置执行用户调度功能当前需申请白名单后才能使用。如需使用该特性，请联系客服或技术支持人员。

调度身份的优先级

系统按照作业委托>公共委托>执行用户>公共IAM账号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，您可通过配置调度身份来解决该问题。

约束限制

- 创建或修改委托需要用户具有Security Administrator权限。
- 配置工作空间级的调度身份，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。

配置公共委托

⚠ 注意

公共委托是工作空间级别的，会委托影响该空间下所有的作业，请慎重配置。 特别是部分作业中包含 MRS相关的节点。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
3. 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。
4. 单击“调度身份配置”，公共调度身份选择“公共委托”。
5. 单击右边的“+”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**参考：创建委托**和**参考：配置委托权限**。

图 9-121 配置工作空间级委托

6. 单击“确定”，回到调度身份配置页面，再单击，完成公共委托配置。

   说明

   公共委托配置后的生效条件：批处理作业下一周期生效，实时作业需要手动重启一下生效。

配置作业委托

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. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
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。创建委托和配置权限，请参见**参考：创建委托**和**参考：配置委托权限**。

图 9-122 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。

配置公共 IAM 账号

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
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“调度身份配置”，公共调度身份选择“公共IAM账号”。
4. 在右边的文本框中输入已创建好的公共IAM账号。
5. 单击，完成公共IAM账号的设置。

配置执行用户

配置作业的执行用户

1. 在作业目录处，双击选中已有作业。
2. 单击画布右侧“作业基本信息”页签，展开配置页面，可设置作业的执行用户。

参考：创建委托

1. 登录IAM服务控制台。
2. 选择“委托”，单击“创建委托”。
3. 设置“委托名称”。例如：DGC_agency。
4. 在创建委托页面，委托类型选择“云服务”，云服务选择“数据湖治理中心 DGC”，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。

图 9-123 创建委托

5. 单击“下一步”，进入授权页面。
6. 在授权页面中搜索“Tenant Administrator”策略，勾选“Tenant Administrator”策略并单击“下一步”。
   - 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**参考：配置委托权限**。

   图 9-124 配置权限

7. 单击“确定”完成委托创建。

参考：配置委托权限

将账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9-83**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9-84**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

- 当满足如下条件之一时，MRS集群才支持委托方式提交作业。
  - 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。
- 当MRS集群不支持委托方式提交作业时，如下节点相关作业不能配置委托。

  MRS相关的节点（MRS Presto SQL、MRS Spark、MRS Spark Python、MRS Flink Job、 MRS MapReduce），以及通过API方式连接的（MRS Spark SQL、MRS Hive SQL）节点。

- 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 9-83 配置相关节点的 admin 权限

节点名称	系统权限	权限描述
CDM Job、DIS Stream、DIS Dump、DIS Client	DAYU Administrator	数据治理中心服务的所有执行权限。
Import GES	GES Administrator	图引擎服务的所有执行权限。该角色有依赖，需要在同项目中勾选依赖的角色：Tenant Guest、Server Administrator。
• MRS Presto SQL、MRS Spark、MRS Spark Python、MRS Flink Job、MRS MapReduce • MRS Spark SQL、MRS Hive SQL（通过MRS API方式连接MRS集群的）	MRS Administrator MRS Fullaccess KMS Administrator	MRS Administrator：RBAC策略下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。 MRS Fullaccess：细粒度策略下MRS管理员权限，拥有该权限的用户可以拥有MRS所有权限。 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**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、DIS Stream、DIS Dump、DIS Client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 9-84 自定义策略

节点名称	授权项
Import GES	• ges:graph:access • ges:graph:getDetail • ges:jobs:getDetail
• MRS Presto SQL、MRS Spark、MRS Spark Python、MRS Flink Job、MRS MapReduce • MRS Spark SQL、MRS Hive SQL（通过MRS API方式连接MRS集群的）	• mrs:job:delete • mrs:job:stop • mrs:job:submit • mrs:cluster:get • mrs:cluster:list • mrs:job:get • mrs:job:list • kms:dek:crypto • kms:cmk:get
MRS Spark SQL、MRS Hive SQL、MRS Kafka、Kafka Client（通过代理方式连接集群）	• kms:dek:crypto • kms:cmk:get • DataArts Studio Administarator(角色)

节点名称	授权项
DLI Flink Job、DLI SQL、DLI Spark	• dli:jobs:get • dli:jobs:update • dli:jobs:create • dli:queue:submit_job • dli:jobs:list • dli:jobs:list_all
DWS SQL、Shell、RDS SQL（通过代理方式连接数据源）	• kms:dek:crypto • kms:cmk:get • DataArts Studio Administarator(角色)
Create OBS、Delete OBS、OBS Manager	• obs:bucket:GetBucketLocation • obs:bucket:ListBucketVersions • obs:object:GetObject • obs:bucket:CreateBucket • obs:bucket:DeleteBucket • obs:object:DeleteObject • obs:object:PutObject • obs:bucket:ListAllMyBuckets • obs:bucket:ListBucket

9.10.1.5 配置节点并发数

本章节主要介绍如何配置当前作业空间同一时间允许正在运行的作业节点的最大并发数。

约束限制

工作空间的节点并发数不能大于实例的节点并发数上限。

配置方法

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 选择“节点并发数”。

步骤5 配置工作空间的节点并发数，工作空间的节点并发数不能大于DataArts Studio实例的并行节点并发数上限。

DataArts Studio实例的节点并发数上限可通过表9-85获取。其中的作业节点调度次数/天配额可通过DataArts Studio实例卡片上的“更多 > 配额使用量”入口查看，其中的“作业节点调度次数/天”总量即为当前实例配额。

表 9-85 DataArts Studio 实例并行节点数上限

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

图 9-125 配置节点并发数

步骤6 单击“保存”，完成配置。

----结束

查看历史节点并发数

步骤1 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤2 选择“节点并发数”。

步骤3 在历史节点并发数界面，选择历史时间段。

步骤4 单击“确定”。

说明

查看历史节点并发数的时间区间最大为24小时。

----结束

9.10.1.6 配置模板

本章节主要介绍如何创建并使用模板。用户在编写业务代码时，对于重复的业务逻辑，可以直接引用SQL模板，同时在配置作业运行参数的时候，可以直接使用作业参数模板，不用再进行重复配置。

约束限制

该功能适用于以下场景：

- Flink SQL脚本可以引用脚本模板。
- 在pipeline作业开发中，MRS Flink Job节点可以使用引入了脚本模板的Flink SQL脚本，同时在MRS Flink Job节点的“运行程序参数”里面可以引用参数模板。
- 在Flink SQL单任务作业中引用脚本模板。
- 在Flink Jar单任务作业中使用参数模板。
- Spark SQL和Hive SQL脚本及单任务作业支持引用参数模板。模板配置好之后，请到**配置默认项**去使用该模板。

配置方法

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

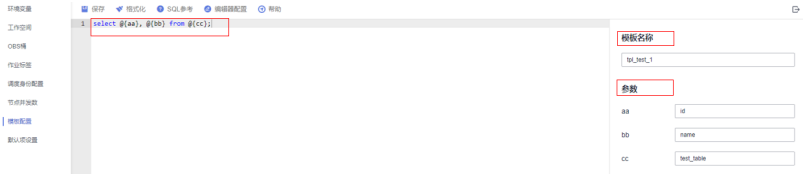
步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 选择“模板配置”。

- 配置脚本模板信息。
  a. 单击“新建”进入脚本模板配置界面。
  b. 输入“模板名称”。
  c. 在界面上输入SQL语句，并引入脚本参数。
  d. 配置脚本模板参数。参数名称不可修改，参数值可以进行修改。

  图 9-126 配置脚本模板

  e. 单击“保存”。
  您可以对已创建的脚本模板进行查看、修改和删除。
- 配置参数模板信息。
  a. 单击“新建”进入参数模板配置界面。
  b. 输入“模板名称”。
  c. 单击“添加参数”。配置参数值和参数名称。可以对配置的参数进行修改和删除。

图 9-127 配置参数模板

d. 单击“确定”。

您可以对已创建的参数模板进行查看、修改和删除。

脚本模板和参数模板的应用场景请参见**引用脚本模板和参数模板的使用介绍**。

----结束

9.10.1.7 配置调度日历

- 作业调度支持按照日历配置自定义工作日期进行周期调度。
- 调度日历配置完成后，在作业开发界面，在“调度配置”页签，选择周期调度，选择调度日历，即可按照调度日历所定义的工作日期进行调度。**如果作业不在日历范围内是空跑，在日历范围内是正常执行。**

  说明

  使用调度日历功能后，在作业正常调度和补数据时，作业实例在执行时，系统会检查计划执行时间，是否是工作日。

  - 如果实例的计划执行时间，是日历中的工作日，则实例正常执行。
  - 如果实例的计划执行时间，是日历中的非工作日，则实例空跑。

约束限制

按照日历配置自定义工作日期进行调度，不支持于实时处理作业，只支持批处理作业。

配置方法

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 选择“调度日历配置”。

步骤5 单击“新建”，进入创建调度日历页面。

图 9-128 创建调度日历

步骤6 配置调度日历相关参数。

输入“日历名称”、选择“默认工作日”以及对调度日历进行备注。

默认工作日可以选择“周一到周五”或“周一到周日”。系统默认周一到周五，生成对应的日历信息。

步骤7 单击“确认”，调度日历配置完成。

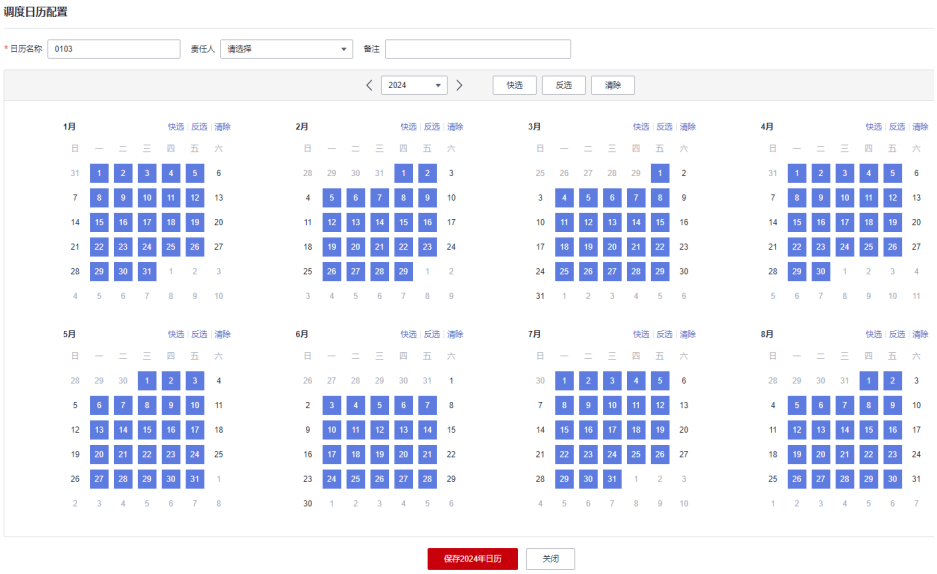
配置完成后，请前往作业开发界面，在所需作业画布右侧“调度配置”页签，选择周期调度，选择调度日历，即可按照调度日历所定义的工作日期进行调度。如果作业不在日历范围内是空跑，在日历范围内是正常执行。

----结束

更多操作

- 修改：单击操作列的“修改”，可以修改已配置好的日历。
  - 快选：快速选中本月的周一到周五
  - 反选：对已选择的工作日进行反选
  - 清除：对已选择的工作日进行清除

图 9-129 修改调度日历



- 删除：单击操作列的“删除”，可以删除已配置好的日历。

9.10.1.8 配置默认项

本章节主要介绍默认项的配置。当前只有具备**DAYU Administrator**或**Tenant Administrator**账号权限的用户才有默认配置项的相关操作权限。

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。

表 9-86 配置项列表

配置项	影响模块	主要用途
配置周期调度	作业调度	• 当前作业所依赖的作业执行失败后，**当前作业**的处理策略。
配置多IF策略	作业调度	节点执行依赖多个IF条件的处理策略。
配置软硬锁策略	脚本/作业开发	作业或脚本的抢锁操作依赖于软硬锁处理策略。
脚本变量定义	脚本开发	脚本变量的格式定义。SQL脚本的变量格式有${}和${dlf.}两种。
配置数据导出策略	脚本/作业开发	对SQL执行结果框中的数据配置下载或转储的策略。 • 所有用户都可以 • 所有用户都不能 • 仅工作空间管理员可以

配置项	影响模块	主要用途
禁用作业节点名称同步变化	作业开发	DataArts Studio作业中的节点关联脚本或者其他服务的作业时，节点名称不会同步变化。
是否使用简易变量集	作业开发	简易变量集提供了一系列自定义的变量，实现在任务调度时间内参数的动态替换。
忽略失败的通知策略	运维调度	对于运行状态为忽略失败的作业，支持发送的通知类型。
节点超时是否重试	作业运行	作业节点运行超时导致的失败也会重试。
实例超时是否忽略等待时间	作业运行	实例运行时超时计算将忽略等待时间。
MRS jar包参数拆分规则	作业开发	MRS MapReduce算子和MRS Spark算子jar包参数中字符串参数（使用""括起来的参数）拆分规则。
等待运行实例同步作业版本策略	运维调度	已生成的等待运行的作业实例，此时发布新的作业版本后，实例是否会使用最新的作业版本运行。
Hive SQL及Spark SQL执行方式	脚本/作业开发	• SQL语句放置在OBS中：将OBS路径返回给MRS。 • SQL语句放置在请求的消息体中：将脚本内容返回给MRS。
补数据优先级设置	运维调度-补数据	设置补数据作业的优先级。当系统资源不充足时，可以优先满足优先级较高的作业的计算资源，数字越大优先级越高，当前只支持对DLI SQL算子设置优先级。
历史作业实例取消策略	运维调度	配置等待运行作业实例的超期天数。当作业实例等待运行的时间，超过了所配置的期限天数时，作业实例将取消执行。超期天数，最小需配置2天，即至少需要等待2天，才可取消未运行的作业实例。超期天数默认为60天，单位：天。
历史作业实例告警策略	运维调度	配置“通知管理”中通知告警能监控的天数范围。 通知管理中配置的告警通知能监控的作业实例天数范围，默认配置为7天，即对7天内满足触发条件的作业实例都能正常上报通知告警，但7天之前的作业实例不会再上报告警。
作业告警通知主题	通知配置	按责任人发送通知时所使用的主题。

配置项	影响模块	主要用途
作业算子失败重试默认策略	运维调度	设置作业算子失败重试默认策略。
作业每次重试失败即告警	运维调度	当作业配置失败告警的时候，该配置项会触发作业每次重试失败即告警，可作用于全部作业、实时作业和批作业。 若选择不支持，则作业达到最大失败重试次数时才触发失败告警。
作业运行自动传递脚本名称	作业开发（作业运行）	开关打开后，系统自动传参将生效：将对当前空间内作业运行时，将Hive SQL脚本set mapreduce.job.name=脚本名称，自动传递至MRS。
作业依赖规则	作业调度	作业能被其他空间作业依赖，需要该空间作业列表的查询权限。工作空间内的默认角色均有该权限，自定义角色需要在有数据开发下的作业查询权限。
脚本执行历史展示	脚本/作业开发	对脚本执行历史结果进行权限管控。 • 仅自己可见：脚本执行历史只显示本用户的执行历史。 • 所有用户可见：脚本执行历史显示所有用户的执行历史。
作业测试运行使用的身份	作业开发（作业测试运行）	配置作业测试运行使用的身份。 • 公共委托或IAM账号：使用配置的公共委托或公共IAM账号身份执行作业。 • 个人账号：使用点击测试作业用户的身份执行作业。
Spark SQL作业/脚本默认模板配置	Spark SQL脚本/作业开发	Spark SQL作业/脚本配置运行，是否允许用户设置任意参数。
Hive SQL作业/脚本默认模板配置	Hive SQL脚本/作业开发	Hive SQL作业/脚本配置运行，是否允许用户设置任意参数。
作业/脚本变更管理	作业/脚本的导入和导出	工作空间是否开启作业/脚本变更管理。 • 是：表示作业/脚本变化时记录变更事件，支持根据时间点增量导出和导入所有变化的作业/脚本。 • 否：表示作业/脚本变化时不记录变更事件，只支持选定作业/脚本的导出和导入。

配置周期调度

- 当前作业所依赖的作业执行失败后，**当前作业**的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。

a. 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。
b. 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“取消执行”。

- 等待执行：当被依赖的作业执行失败后，当前作业会等待执行。
- 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。
- 取消执行：当被依赖的作业执行失败后，当前作业会取消执行。

c. 单击“保存”，对设置的配置项进行保存。该配置仅对新建作业有效。

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

脚本变量定义

SQL脚本的变量格式有${}和${dlf.}两种，支持用户根据实际情况进行配置。配置的变量格式会作用于SQL脚本、作业中SQL语句、单节点作业，环境变量。

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“脚本变量定义”配置项。

说明

系统默认脚本变量格式为${}。

- ${}格式：识别脚本中${}格式的定义，解析其中的字段为变量名，如${xxx}，识别为变量名：xxx。
- ${dlf.}格式：识别脚本中${dlf.}格式的定义，解析其中的dlf.字段为变量名，其他${}格式定义不再识别为变量，如${dlf.xxx}，识别为变量名：dlf.xxx。

步骤3 单击“保存”，对设置的配置项进行保存。

----结束

配置数据导出策略

系统默认支持所有用户都能下载和转储SQL脚本的执行结果。如果您不希望所有用户都有该操作权限，可参考下如下步骤对数据导出策略进行配置。

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“数据导出策略”配置项。

说明

系统默认的数据导出策略是“所有用户都可以”。

- 所有用户都可以：所有用户都能对SQL执行结果做“下载”或“转储”操作。
- 所有用户都不能：所有用户都不能对SQL执行结果做“下载”或“转储”操作。
- 仅工作空间管理员可以：只有工作空间管理员可以对SQL执行结果做“下载”或“转储”操作。

步骤3 单击“保存”，对设置的配置项进行保存。

----结束

禁用作业节点名称同步变化

在作业开发界面，系统默认选择脚本或关联其他云服务的功能时会同步更新节点名称，使之与脚本或功能名称一致。当前支持配置作业节点名称是否同步变更。

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”，可勾选“禁用作业节点名称同步”的节点。

说明

- 当前系统支持对CDM Job、DIS Stream、DLI SQL、DWS SQL、MRS Spark SQL、MRS Hive SQL、MRS Presto SQL、MRS HetuEngine、MRS ClickHouse、MRS Impala SQL、Shell、DORIS SQL、RDS SQL、Python、Subjob、For Each节点的名称是否同步为脚本或功能名称做配置。
- 系统默认为不勾选，即选择脚本或功能时会同步更新节点名称。
- 如果勾选了节点，在选择脚本或功能时，不会同步更新节点的名称。

步骤3 单击“保存”，对设置的配置项进行保存。

----结束

是否使用简易变量集

简易变量集提供了一系列自定义的变量，实现在任务调度时间内参数的动态替换。

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“是否使用简易变量集”配置项。

说明

- 是：支持使用简易变量集。通过简易变量集提供的一系列自定义的变量，自定义参数会根据任务调度的业务日期、计划时间及参数的取值格式自动替换为具体的值，实现在任务调度时间内参数的动态替换。
- 否：不支持使用简易变量集。

步骤3 单击“保存”，对设置的配置项进行保存。

----结束

忽略失败的通知策略

对于运行状态为忽略失败的作业，可选择支持发送的通知类型。

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“忽略失败的通知策略”配置项。

步骤3 选择忽略失败状态节点的通知类型。

说明

- 在作业基本信息界面，对于“当前节点失败后，后续节点处理策略 > 继续执行下一节点”的作业可理解为忽略失败的作业，系统默认运行结果为成功。
- 对于运行状态为忽略失败的作业，支持发送的通知类型如下：

  运行异常/失败：对于运行状态为忽略失败的作业，支持发送的通知类型为“运行异常/失败”。

  运行成功：对于运行状态为忽略失败的作业，支持发送的通知类型为“运行成功”，系统默认策略为运行成功。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

节点超时是否重试

对于作业节点运行超时导致失败的作业，可选择是否支持重试。

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“节点超时是否重试”配置项。

步骤3 配置节点运行超时是否重试。

说明

- 否：作业节点运行超时导致失败后，不重新执行节点。
- 是，作业节点运行超时导致失败后，可以重新执行节点。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

实例超时是否忽略等待时间

对于作业实例运行超时以后，可以配置实例超时是否忽略等待时间。

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“实例超时是否忽略等待时间”配置项。

步骤3 配置实例超时是否忽略等待时间。

说明

是：实例运行时超时计算将忽略等待时间。

否：实例运行时超时计算将等待时间会包含进去。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

MRS jar 包参数拆分规则

对MRS MapReduce算子和MRS Spark算子jar包参数中字符串参数（使用""括起来的参数）拆分规则进行配置。

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“MRS jar包参数拆分规则”配置项。

步骤3 配置MRS jar包参数拆分规则。

说明

按空格拆分字符串参数：如**"select * from table"**会按空格被拆分成四个参数，分别为select、*、from、table。

不拆分字符串参数：如**"select * from table"**会被当成一个完整的参数，不进行拆分。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

等待运行实例同步作业版本策略

已生成的等待运行的作业实例，此时发布新的作业版本后，实例是否会使用最新的作业版本运行。

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“等待运行实例同步作业版本策略”配置项。

步骤3 配置等待运行实例同步作业版本策略。

说明

是：等待运行的作业实例，当发布新的作业版本后，作业实例在运行时，会使用最新版本作业运行。

否：等待运行的作业实例，当发布新的作业版本后，作业实例在运行时，依旧使用当前版本的作业运行。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

Hive SQL 及 Spark SQL 执行方式

执行Hive SQL及Spark SQL语句时，DataArts Studio支持把SQL语句放在OBS中，同时还支持把SQL语句放在请求的消息体中。

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“Hive sql及Spark sql执行方式”配置项。

步骤3 配置Hive SQL及Spark SQL的执行方式。

说明

SQL语句放置在OBS中：执行Hive SQL及Spark SQL语句时，把SQL语句放在OBS中，将OBS路径返回给MRS。

SQL语句放置在请求的消息体中：执行Hive SQL及Spark SQL语句时，把SQL语句放在请求的消息体中，将脚本内容返回给MRS。

步骤4 单击“保存”，对设置的配置项进行保存。

说明

Hive SQL和Spark SQL脚本、Pipeline作业以及单任务作业支持Hive SQL及Spark SQL执行方式的配置。

----结束

补数据优先级设置

设置补数据作业的优先级。当系统资源不充足时，可以优先满足优先级较高的作业的计算资源，数字越大优先级越高。当前只支持对DLI SQL算子设置优先级。

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“补数据优先级设置”配置项。

步骤3 配置补数据的优先级策略。

步骤4 单击“保存”，对设置的配置项进行保存。

说明

补数据优先级设置和DLI的spark.sql.dli.job.priority优先级的映射关系如下：
补数据的优先级设置为1时，映射到DLI优先级spark.sql.dli.job.priority=1；
补数据的优先级设置为2时，映射到DLI优先级spark.sql.dli.job.priority=3；
补数据的优先级设置为3时，映射到DLI优先级spark.sql.dli.job.priority=5；
补数据的优先级设置为4时，映射到DLI优先级spark.sql.dli.job.priority=8；
补数据的优先级设置为5时，映射到DLI优先级spark.sql.dli.job.priority=10。

----结束

历史作业实例取消策略

配置等待运行作业实例的超期天数。当作业实例等待运行的时间超过了所配置的期限天数时，作业实例将取消执行。超期天数最小需要配置2天，即至少需要等待2天，才可以取消未运行的作业实例，超期天数默认为60天，单位为天。

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“历史作业实例取消策略”配置项。

步骤3 配置等待运行作业实例的超期天数。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

配置实例超期取消是否发送告警。若选择“是”，当历史作业实例被超期取消，且作业配置运行取消通知类型时，将会发送告警通知。如果选择“否”，将不会发送告警通知。

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“历史作业实例取消策略”配置项。

步骤3 配置实例超期取消是否发送告警。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

历史作业实例告警策略

通知管理中配置的告警通知能监控的作业实例天数范围，配置默认为7天，即对7天内满足触发条件的作业实例都能正常上报告警通知，但7天之前的作业实例不会再次上报告警。

例如：告警监控天数配置为2天时，昨天和今天的作业实例触发监控时会告警，但是前天以及3天前的作业实例，即使满足触发条件也不会再次发送通知告警。

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“历史作业实例告警策略”配置项。

步骤3 配置“通知管理”中通知告警能监控的天数范围。

说明

告警监控天数配置默认为7天，最小为1天，最大为270天。

告警监控天数配置好以后，告警通知只提示告警设置以后的作业实例，不再展示历史上的异常记录。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

作业告警通知主题

配置作业告警通知主题，此处所配置的主题是按责任人发送通知时所使用的主题。

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“作业告警通知主题”配置项。

步骤3 配置作业告警通知主题。单击“查看主题”可以跳转到消息通知服务界面查看已创建的主题。

说明

此处选择的主题需要在消息通知服务SMN界面新配置一个主题（防止与之前所配置的按主题发送通知的主题重复），只能由空间管理员配置。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

作业算子失败重试默认策略

设置作业算子失败重试默认策略后，仅对当前工作空间作业新增的作业算子生效，历史作业算子默认值不受影响。系统初始默认值为否。

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“作业算子失败重试默认策略”配置项。

步骤3 配置作业算子失败重试默认策略。

说明

设置作业算子失败重试默认策略后，新增的作业算子最大重试次数默认为1，重试间隔时间默认120秒。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

作业每次重试失败即告警

配置作业每次重试失败即告警后，当作业配置了失败重试时，在第一次运行失败后就上报告警。

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“作业每次重试失败即告警”配置项。

步骤3 配置作业每次重试失败即告警。

说明

- 当作业配置失败告警的时候，该配置项会触发作业每次重试失败即告警，可作用于全部作业、实时作业和批作业。
- 若选择不支持，则作业达到最大失败重试次数时才触发失败告警。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

作业运行自动传递脚本名称

作业运行自动传递脚本名称开关打开后，系统自动传参将生效，将对当前空间内作业运行时，将Hive sql脚本set mapreduce.job.name=“脚本名称”自动传递至MRS。

说明

仅对脚本中未设置上述参数值的情况下生效，如脚本中已设置此参数值，则优先以读取人工设置的值传递至MRS。特别提醒：如MRS集群是安全模式，则不支持此设置方式，需提前将集群设置为非安全模式。

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“作业运行自动传递脚本名称”配置项。

步骤3 配置作业运行自动传递脚本名称。

说明

- 是：作业运行时系统会自动传递Hive sql脚本名称到MRS。
- 否：作业运行时系统不会自动传递Hive sql脚本名称到MRS。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

作业依赖规则

作业能被其他空间作业依赖，需要该空间作业列表的查询权限。工作空间内的默认角色均有该权限，自定义角色需要在有数据开发下的作业查询权限。

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“作业依赖规则”配置项。

步骤3 配置作业依赖规则。

说明

- 作业不能被其他工作空间依赖：该空间的作业不能被其他空间作业依赖。
- 作业能被其他工作空间依赖：该空间的作业能被其他空间作业依赖，不需要为该用户配置所依赖空间的权限。
- 作业能被其他空间作业依赖（需要该空间作业列表的查询权限）：该空间的作业能被其他空间作业依赖，需要为该用户配置所依赖空间的权限。如果没有为该用户配置权限的话，在跨空间配置作业依赖关系时，系统会提示“当前用户没有工作空间xxx的获取作业列表的权限”。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

脚本执行历史展示

配置脚本执行历史展示后，可以对脚本执行历史的查看进行权限管控。

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“脚本执行历史展示”配置项。

步骤3 配置脚本执行历史展示。

说明

- 仅自己可见：脚本执行历史只显示本用户的操作历史。
- 所有用户可见：脚本执行历史显示所有用户的操作历史。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

作业测试运行使用的身份

配置作业测试运行使用的身份后，在作业测试运行时，可以对作业测试运行的身份进行指定。

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“作业测试运行使用的身份”配置项。

步骤3 配置作业测试运行使用的身份。

说明

- 公共委托或IAM账号：使用配置的公共委托或公共IAM账号身份执行作业。
- 个人账号：使用点击测试作业用户的身份执行作业。
  如果没有配置工作空间委托或IAM账号，作业测试运行统一使用个人账号身份
  如果是联邦账户，必须配置为公共调度身份，即配置为公共委托或IAM账号。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

Spark SQL 作业/脚本默认模板配置

Spark SQL作业/脚本配置运行时，通过默认参数模板去管控是否允许用户去设置任意参数覆盖模板设置的默认参数。

在MRS API连接方式下，Spark SQL脚本支持配置默认运行参数。代理连接不支持。

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“Spark SQL作业/脚本默认模板配置”配置项。

步骤3 配置Spark SQL作业/脚本运行时，是否允许用户设置任意参数。

说明

- 是：表示不配置这种参数，作业/脚本随便设置参数。
- 否：表示必须选择一个模板给这类作业/脚本绑定好，并且在作业/脚本配置中这些参数不允许被覆盖。选择“否”时，设置已经配置好的默认参数模板。配置模板请参见**配置模板**。

  设置好以后，请到**Spark SQL作业界面的基本信息或Spark SQL脚本界面，单击右上角的**去查看所配置的默认运行程序参数，预置的默认参数置灰，不能修改。

  用户根据需要也可以自定义运行程序参数，最终Spark SQL作业/脚本运行时，设置的模板参数可以允许作业/脚本参数进行覆盖。

----结束

Hive SQL 作业/脚本默认模板配置

Hive SQL作业/脚本配置运行时，通过默认参数模板去管控是否允许用户去设置参数覆盖模板设置的默认参数。

在MRS API连接方式下，Hive SQL脚本支持配置默认运行参数。代理连接不支持。

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“Hive SQL作业/脚本默认模板配置”配置项。

步骤3 配置Hive SQL作业/脚本运行时，是否允许用户设置任意参数。

说明

- 是：表示不配置这种参数，作业/脚本随便设置参数。
- 否：表示必须选择一个模板给这类作业/脚本绑定好，并且在作业/脚本配置中这些参数不允许被覆盖。选择“否”时，设置已经配置好的默认参数模板。配置模板请参见**配置模板**。

  设置好以后，请到**Hive SQL作业界面的基本信息或Hive SQL脚本界面，单击右上角的**去查看所配置的默认运行程序参数，预置的默认参数置灰，不能修改。

  用户根据需要也可以继续添加自定义运行程序参数，最终Hive SQL作业/脚本运行时，设置的模板参数可以允许作业/脚本参数进行覆盖。

步骤4 单击“保存”，对设置的配置项进行保存。

----结束

作业/脚本变更管理

在工作空间配置作业/脚本变更管理后，可以将工作空间A的作业/脚本的变更记录（新增、修改、删除）导出来，同时可以将导出的zip包（作业/脚本的变更记录）导入到工作空间B。

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“作业/脚本变更管理”配置项。

步骤3 配置工作空间是否开启作业/脚本变更管理。

说明

- 是：表示作业/脚本变化时记录变更事件，支持根据时间点增量导出和导入所有变化的作业/脚本。
- 否：表示作业/脚本变化时不记录变更事件，只支持选定作业/脚本的导出和导入。

步骤4 单击“保存”，对设置的配置项进行保存。

说明

当“作业/脚本变更管理”配置项开启后，才能够在作业/脚本的进行工作空间的导出和导入。

----结束

9.10.1.9 配置任务组

通过配置任务组，可以更细粒度的进行当前任务组中的作业节点的并发数控制。

约束限制

- 该功能不支持实时处理作业，只支持批处理作业。
- 任务组不能跨工作空间去使用。
- 对于Pipeline作业，每个节点都可以配置一个任务组，也可以在作业里面统一配置任务组，如果配置了节点级任务组，则优先级高于作业级的任务组。

配置方法

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 选择“任务组”。

步骤5 单击“创建任务组”进入“创建任务组”的界面。

步骤6 配置相关参数。

表 9-87 创建任务组参数

参数	说明
任务组名称	任务组的名称。任务组名称不能重名。
最大并发数	当前任务组作业节点最大并发数。 最大节点并发数即为当前DataArts Studio实例的并发数。 当前DataArts Studio实例的节点并发数上限为1000，请不要超过该上限。 最大并发数与DataArts Studio实例规格有关，不同规格的实例的节点并发数上限值，则各有不同。
描述	描述信息。

步骤7 单击“确定”，任务组创建完成。

配置完成后，请前往作业开发界面，在所需作业画布右侧“调度配置”页签，选择任务组，即可按照设置好的任务组更细粒度的进行当前任务组中的作业节点的并发数控制。

----结束

后续操作

修改：单击“修改”，可以修改已配置好的任务组。任务组的修改是实时生效的。

删除：单击“删除”，可以删除已配置好的任务组。如果任务组被作业引用，无法删除。

查看引用：单击“查看引用”，可以查看该任务组被引用的详细信息。

9.10.1.10 Notebook 管理

启用Notebook功能后，您可以通过Notebook松完成开发、调试、调度集群作业，并支持实时探索、处理和数据可视化。同时，您可以对该工作空间的Notebook使用配额进行配置。

前提条件

登录用户需要授权DataArts Studio系统角色“DAYU User”。详细操作请参见**创建IAM用户并授予DataArts Studio权限**。

约束限制

当前工作空间的管理员可以启用Notebook。或者，具有DAYU Administrator或者Tenant Administrator权限的用户也可以启用Notebook。

开启 Notebook

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 配置”。

步骤4 选择“Notebook管理”。

步骤5 打开“启用Notebook”的开关。启用中，页面会显示“当前Notebook功能正在初始化，大概需要20分钟左右，请稍后...”。

说明

当前工作空间的管理员可以启用Notebook。或者，具有DAYU Administrator或者Tenant Administrator权限的用户也可以启用Notebook。

用户所创建的Notebook会在页面底下列表进行展示，可以对不再使用的Notebook进行删除。开启后，在数据开发的“Notebook”下就可以使用此功能了。

----结束

配额管理

您可以针对单空间设置允许开通的Notebook数量，所有空间配额总和不超过DataArts实例的配额上限。

步骤1 单击“配额管理”进入配额管理页面。

步骤2 配置Notebook的数量。

图 9-130 配额配置

说明

Notebook数量的可调区间为大于已有的Notebook数量，小于空间允许开通的上限。比如，当前空间最多可开通6个Notebook，配置的数量最大为6个。

步骤3 单击“确定”，完成配额配置。

----结束

9.10.2 管理资源

用户可以通过资源管理功能，上传自定义代码或文本文件作为资源，在节点运行时调用。可调用资源的节点包含DLI Spark、MRS Spark、MRS MapReduce和DLI Flink Job。

创建资源后，配置资源关联的文件。在作业中可以直接引用资源。当资源文件变更，只需要修改资源引用的位置即可，不需要修改作业配置。关于资源的使用样例请参见**开发一个DLI Spark作业**。

约束限制

该功能依赖于OBS服务或MRS HDFS服务。

新建目录（可选）

如果已存在可用的目录，可以不用新建目录。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

2. 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。
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”。
4. 在资源目录中，单击，弹出“新建目录”页面，配置如表9-88所示的参数。

表 9-88 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。

5. 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. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
2. 单击“新建资源”，弹出“新建资源”页面，配置如表9-89所示的参数。单击“确定”，新建资源。

表 9-89 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。支持的文件后缀名为：zip、tgz、tar.gz、tar、jar。
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。
文件路径	是	当“资源位置”选择OBS时，文件路径选择OBS文件路径。 当“资源位置”选择HDFS时，文件路径选择MRS集群名称。
依赖包	否	当前只支持DLI Spark节点。 选择已上传到OBS中的依赖Jar包。“类型”为“jar”或“pyFile”时，配置该参数。

参数	是否必选	说明
选择目录	是	选择资源所属的目录，默认为根目录。
描述	否	资源的描述信息。

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. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
2. 在资源的“操作”列，单击“编辑”，弹出“编辑资源”页面，参考表9-89修改资源的参数。
3. 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。

须知

删除资源的时候，如果资源被作业引用，单击删除时，会自动弹出“删除资源”的界面，单击“确定”，会自动弹出“引用列表”的界面，可以查看资源被哪些作业所引用，单击“查看”可跳转到作业页面进行查看详情。

1. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
2. 在资源的“操作”列，单击“删除”，弹出“删除资源”页面。
3. 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. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
2. 在资源目录中，单击，选择“导入资源”，弹出“导入资源”页面。
3. 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. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
2. 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. 在数据开发主界面的左侧导航栏，选择“配置管理 > 资源管理”。
2. 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。
3. 在引用列表窗口，可以查看该资源被引用的情况。

图 9-131 引用列表

9.11 审批中心

对于简单模式工作空间，当前支持开发者在提交脚本和作业时，由指定审核人进行审批。审批中心可以对单据审批进行统一管理，对审批人进行工作空间级的配置和维护。

约束与限制

- 仅当前工作空间的管理员或者拥有DAYU Administrator、Tenant Administrator权限的用户，可新建、修改和删除审批人。
- 审批人必须为当前工作空间管理员或者拥有DAYU Administrator、Tenant Administrator权限的用户。
- 当前工作空间为企业模式时，通过任务发布方式进行单据的审批操作，不支持提交脚本或者作业进行审批。
- 开启审批功能时，相关API的请求体需要增加审批人属性，具体见**作业开发API**。
- 审批开关的配置、作业和脚本的审批只能在前台界面进行操作。
- 有实时Pipeline作业的情况下，不允许打开审批开关。
- 审批开关打开后，审批中心功能对审批人和单据提交人都可见。审批开关关闭时，仅当前工作空间的管理员或者拥有DAYU Administrator、Tenant Administrator权限的用户可以看到审批中心功能，其他用户不可见。

审批配置

仅当前工作空间的管理员或者拥有DAYU Administrator、Tenant Administrator权限的用户可以进行审批配置。打开审批开关后，可以对作业或者脚本进行审批配置。

图 9-132 配置审批

1. 选择“数据开发 > 审批中心”，单击“审批配置”页签。
2. 开启审批开关。仅当前工作空间管理员或者拥有DAYU Administrator、Tenant Administrator权限的用户，可以开启或者关闭审批开关。

   说明

   - 该开关开启后，在提交作业或脚本时，都需要指定审批人。开关关闭后，所有作业/脚本都将不再需要审批。
   - 如果当前工作空间还有未完成审批的流程，不可以关闭开关。
3. 配置审批时，系统支持三种不同场景的审批配置。
   - **全部作业/脚本**：工作空间内的所有作业和脚本都开启审批。
   - **自定义作业/脚本**：自定义添加需要审批的作业/脚本。

     选择“审批作业”页签，单击“添加”，进入“请选择要审批的作业”界面，选择需要审批的作业。 添加作业时会自动添加其关联的脚本。当选中目录时，仅将当前目录下的作业添加到需要审批的作业中。若在该目录下新创建作业，则需要再次添加。单击“确定”。

     作业添加后，可以删除，支持批量删除。

     选择“审批脚本”页签，单击“添加”，进入“请选择要审批的脚本”界面，选择需要审批的脚本。当选中目录时，仅将当前目录下的脚本添加到需要审批的脚本中。若在该目录下新创建脚本，则需要再次添加。单击“确定”。

     脚本添加后，可以删除，支持批量删除。
   - **基线任务作业**：从基线任务添加需要审批的作业。

     选择“审批作业”页签，单击“添加”，进入“从基线添加要审批的作业”界面，选择基线任务的优先级作业，所选基线对应的作业，将被指定为需要审批的作业，单击“确定”。基线任务上游的作业也需要审批。

     选择“审批脚本”页签，选择了基线对应的作业，作业关联的脚本会同步显示在该页面。
4. 在“审批人信息”，当前工作空间管理员或者拥有DAYU Administrator、Tenant Administrator权限的用户，可以配置审批人信息。
   a. 单击“审批人管理”，进入管理控制台，选择“工作空间”进入。
   b. 单击当前工作空间进入空间信息界面。
   c. 配置“空间成员”信息。单击“添加”进入添加成员界面。
   d. 搜索“成员账号”并设置管理员角色。
   e. 单击“确定”。配置好的审批人信息会自动显示出来便于查看。

审批管理

- 单据提交人可在审批中心页面，查看自己提交的申请及审批进度。

a. 选择“数据开发 > 审批中心”，单击“我的申请”页签。
b. 单击操作栏中的“查看”，可以查看申请单的详细信息。
c. 单击操作栏中的“撤回”，可以撤回申请单，修改后可重新提交审批。

- 仅审批人可以查看待自己审批的申请单。
  a. 选择“数据开发 > 审批中心”，单击“待审批”页签。在此页面查看当前需要审批的申请单。
  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单。
- 仅审批人可以查看自己已审批的历史记录。
  a. 选择“数据开发 > 审批中心”，单击“已审批”页签。在此页面查看已审批的申请单。
  b. 单击操作栏中的“查看”，审批人可以查看申请单的审批记录和申请内容等详细信息。

9.12 下载中心

数据开发模块对于SQL脚本执行的结果支持直接下载和转储。SQL执行结果进行下载和转储后，可以通过下载中心查看下载和转储的结果。

约束与限制

仅SQL脚本和单任务SQL作业运行完成并且返回结果后，执行下载和转储，在下载中心生成记录，可以查看下载和转储的结果。

下载中心

说明

- 下载中心的下载记录会定期老化，老化时下载中心记录和已转储的OBS数据会同时被删除。
- 操作者只能看到自己操作的下载记录，工作空间的管理员可以看到当前空间的所有下载记录。

通过下载中心，对SQL脚本执行的结果进行统一管理。对下载的结果可以进行查看和删除，对转储的结果可以进行查看、下载和删除。

图 9-133 下载中心

- 配置默认的OBS路径地址

  说明

  工作空间的管理员可以配置当前工作空间的默认OBS转储路径。

  a. 选择“数据开发 > 下载中心”进入。

b. 单击“配置OBS默认地址”进入“配置OBS默认路径”页面。

c. 配置默认的OBS路径。

说明

此处配置的OBS路径，是脚本开发或者单任务作业开发时测试运行结果的默认转储OBS路径。配置成功后，后续转储运行结果时，将默认使用此次配置的OBS路径进行转储；已转储的运行结果路径不会改变，请以列表中返回路径为准。

d. 单击“确定”。

- 查看脚本执行的结果

a. 选择“数据开发 > 下载中心”进入“下载中心”页面。

b. 可以查看本地下载任务和异步转储任务的文件名、操作人，操作时间，操作类型，任务状态，OBS路径。

对于转储任务下载失败的记录可以查看记录。

c. 单击“操作”列的，可以从OBS路径下载数据。

d. 单击“操作”列的，可以删除已下载和转储的记录。

单击删除按钮，系统提示“清除后将不能下载该数据，是否清除？”，单击“确认”进行删除。

- 对搜索条件进行过滤

支持通过操作时间、作业名称、OBS路径、操作人、操作类型、任务状态进行过滤筛选。可以输入关键字进行模糊查找。

9.13 节点参考

9.13.1 节点概述

节点定义对数据执行的操作。数据开发模块提供数据集成、计算&分析、数据库操作、资源管理等类型的节点，您可以根据业务模型选择所需的节点。

- 节点的参数支持使用EL表达式，EL表达式的使用方法详见**表达式概述**。
- 节点间的连接方式支持串行和并行。

串行连接：按顺序逐个执行节点，当A节点执行完成后，再执行B节点。

并行连接：A节点和B节点同时执行。

图 9-134 连接示意图

9.13.2 节点数据血缘

9.13.2.1 数据血缘方案简介

什么是数据血缘

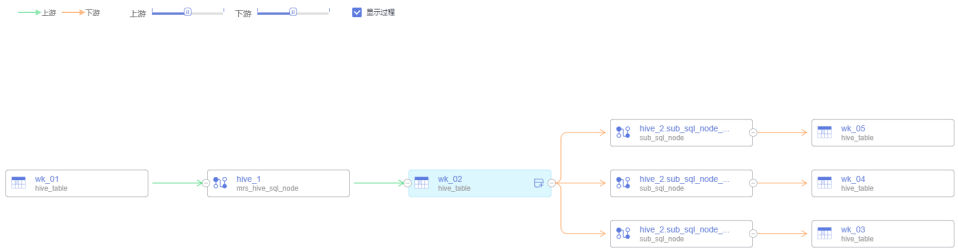
大数据时代，数据爆发性增长，海量的、各种类型的数据在快速产生。这些庞大复杂的数据信息，通过联姻融合、转换变换、流转流通，又生成新的数据，汇聚成数据的海洋。

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。

DataArts Studio生成的血缘关系图如图9-135所示， 为数据表对象， 为作业节点对象，通过对象和箭头的编排表示血缘信息。从血缘关系图中可以看到，wk_02表数据是由wk_01表数据经过hive_1作业节点加工而生成的，wk_02表数据经由hive_2作业节点加工又分别生成了wk_03、wk_04和wk_05的表数据。

图 9-135 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：

  DataArts Studio数据血缘解析方案包含自动分析血缘和手动配置血缘两种方式。一般推荐使用自动血缘解析的方式，无需手动配置即可生成血缘关系，在不支持自动血缘解析的场景下，再手动配置血缘关系。

  - 自动血缘解析，是由系统解析数据开发作业中的数据处理和数据迁移类型节点后自动产生的，无需进行手动配置。支持自动血缘解析的节点类型和场景请参见**自动血缘解析**。
  - 手动配置血缘，是在数据开发作业节点中，自定义血缘关系的输入表和输出表。注意手动配置血缘时，此节点的自动血缘解析将不生效。支持手动配置血缘的节点类型请参见**手动配置血缘**。

- 数据血缘的展示：

  首先在数据目录组件完成元数据采集任务，当数据开发作业满足**自动血缘解析要求**或已**手动配置血缘**，然后成功完成作业调度后，则可以在数据目录模块可视化查看数据血缘关系。

9.13.2.2 配置数据血缘

DataArts Studio数据血缘解析方案包含自动分析血缘和手动配置血缘两种方式。一般推荐使用自动血缘解析的方式，无需手动配置即可生成血缘关系，在不支持自动血缘解析的场景下，再手动配置血缘关系。

- 自动血缘解析，是由系统解析数据开发作业中的数据处理和数据迁移类型节点后自动产生的，无需进行手动配置。支持自动血缘解析的节点类型和场景请参见**自动血缘解析**。
- 手动配置血缘，是在数据开发作业节点中，自定义血缘关系的输入表和输出表。注意手动配置血缘时，此节点的自动血缘解析将不生效。支持手动配置血缘的节点类型请参见**手动配置血缘**。

约束限制

手动配置血缘当前暂不支持字段级血缘解析。

自动血缘解析

自动血缘解析无需进行手动配置，当数据开发作业中包含如**表9-90**所示节点及场景时，系统支持自动解析血缘关系。

说明

解析SQL节点的血缘时，支持多SQL解析及列级血缘解析，单条SQL语句不支持SQL中含有分号的场景。

表 9-90 支持自动血缘解析的作业节点及场景

作业节点	支持场景
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。
DWS SQL	支持Insert into等DML操作产生的DWS表之间的血缘。
MRS Hive SQL	支持Insert into/overwrite等DML操作产生的MRS表之间的血缘。
MRS Spark SQL	支持Insert into/overwrite等DML操作产生的MRS表之间的血缘。
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。
ETL Job	支持DLI、OBS、MySQL以及DWS之间的ETL任务产生的血缘。

手动配置血缘

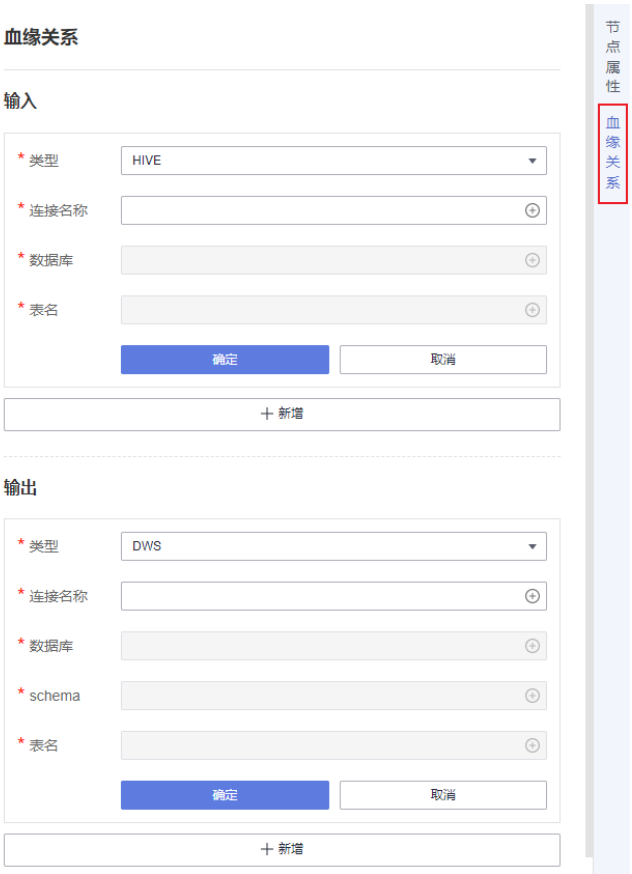
在DataArts Studio数据开发的作业中，您可以在数据开发作业节点中，自定义血缘关系的输入表和输出表。注意，当手动配置血缘时，此节点的自动血缘解析将不生效。

支持手动配置血缘的作业节点类型如下所示。

- CDM Job
- Rest Client
- DLI SQL
- DLI Spark
- DWS SQL
- MRS Spark SQL
- MRS Hive SQL
- MRS Presto SQL
- MRS Spark
- MRS Spark Python
- ETL Job
- OBS Manager

手动配置血缘时，在节点的“血缘关系”页签，配置血缘的输入和输出表。输入和输出表的所属数据源支持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。

图 9-136 手动配置血缘关系示例



例如，当需要配置数据开发Pipeline作业中MRS Spark节点的血缘关系时，由于MRS Spark节点不支持自动血缘解析，则需要手动配置MRS Spark节点的血缘关系。操作步骤如下：

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发组件，进入“数据开发 > 作业开发”页签，单击需要手动配置血缘关系的作业名，打开作业画布。

步骤4 单击作业画布中的MRS Spark节点，并切换到“血缘关系”页签。

图 9-137 进入血缘关系页签

步骤5 在MRS Spark节点的“血缘关系”页签，手动配置血缘的输入表。假如MRS Spark作业中的输入表为“hive”，则血缘输入配置如图9-138所示。

图 9-138 配置血缘输入

步骤6 完成血缘的输入表配置后，单击确定，继续配置血缘的输出表。假如MRS Spark作业中的输出表为“a”，则血缘输出配置如图9-139所示。

图 9-139 配置血缘输出

步骤7 完成血缘的输出表配置后，单击确认，则此MRS Spark节点的血缘关系手动配置成功。后续当需要查看血缘关系时，参考查看数据血缘完成元数据采集，并成功完成作业调度后，即可在数据目录组件查看手动配置的MRS Spark节点血缘关系。

----结束

9.13.2.3 查看数据血缘

首先在数据目录组件完成元数据采集任务，当数据开发作业满足**自动血缘解析要求**或已**手动配置血缘**，然后成功完成作业调度后，则可以在数据目录模块可视化查看数据血缘关系。

约束限制

- 数据血缘关系更新依赖于作业调度，数据血缘关系是基于最新的作业调度实例产生的。

  说明

  对于同一版本的数据开发作业，系统基于最新的作业调度实例生成数据血缘关系后，在冷却期（默认为48小时）内不会再次更新数据血缘关系。如需更新，需要等待冷却期结束或将数据开发作业再次提交版本后调度。

- 数据血缘关系删除需要通过删除作业或删除作业元数据的方式进行，仅将作业停止调度不会触发血缘关系的删除。

新建并运行元数据采集任务

请参见**配置元数据采集任务**，新建并运行元数据采集任务，注意任务中需要选择待查看血缘关系的数据表。

如果此前已创建并运行过待查看数据表的元数据采集任务，此操作可跳过。

启动作业调度

元数据采集完成后，系统基于最新的作业调度实例产生相关的数据血缘关系。

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

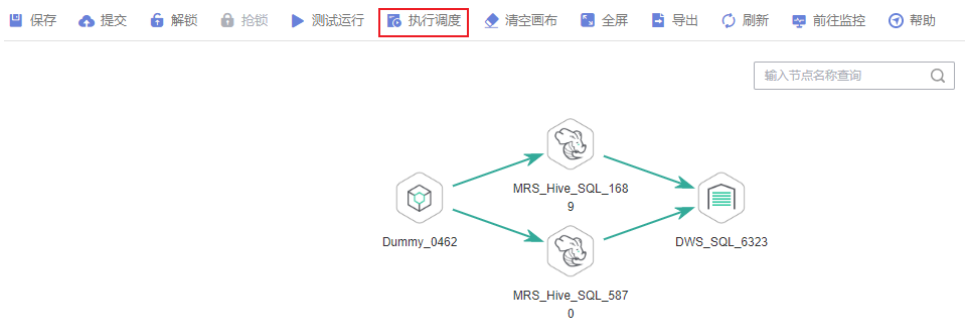
步骤3 在数据开发控制台，单击左侧导航栏中的作业开发按钮，进入作业开发页面后，打开已完成血缘配置的作业。

步骤4 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 9-140 作业调度



步骤5 待调度作业成功运行完成后，等待约1分钟左右，数据血缘关系即可生成成功。

----结束

查看数据血缘关系

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

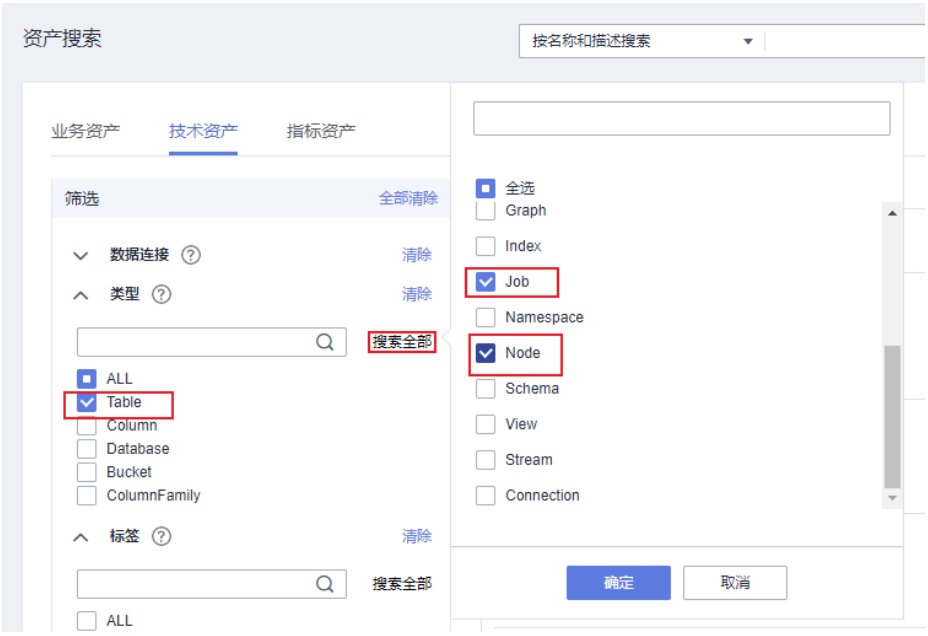
步骤2 在“数据目录 > 技术资产”页面，可以对数据开发的作业、节点、表进行查询。

在“类型”筛选区域，单击“搜索全部”按钮并在全部类型中勾选“Job”、“Node”和“Table”，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

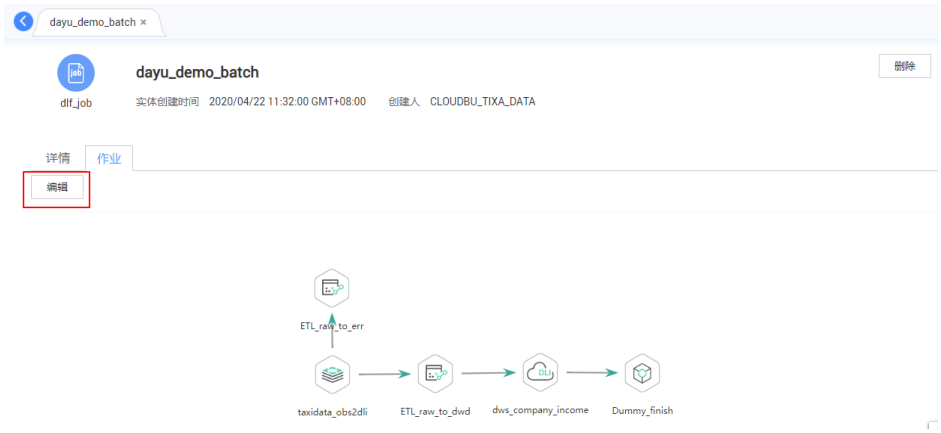
数据开发中的作业信息不属于任何一个数据连接，故如果在搜索条件中勾选数据连接，则查询不到结果。

图 9-141 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。

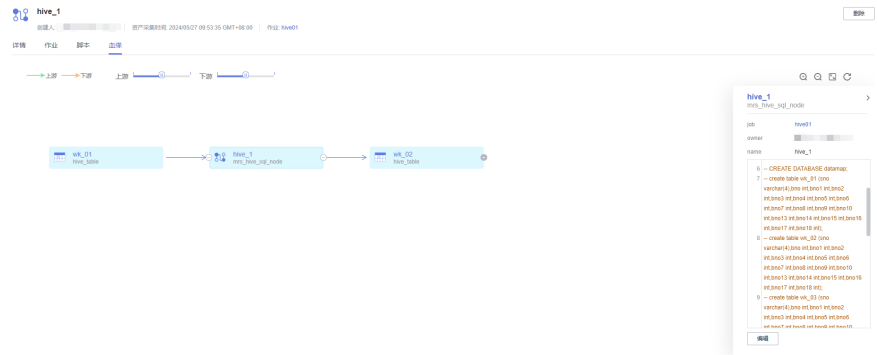
图 9-142 查看作业



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。

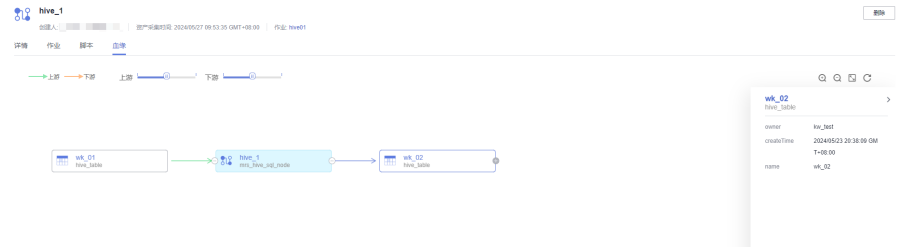
图 9-143 查看节点血缘



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 9-144 查看表血缘



----结束

9.13.3 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。

说明

如果CDM作业中配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为（数据开发作业计划启动时间-偏移量），而不是（CDM作业实际启动时间-偏移量）。

参数

用户可参考**表9-91**，**表9-92**和**表9-93**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 9-91 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。 • 勾选两个集群时，集群是随机下发，用于分担系统负荷。当其中一个集群状态异常后，会触发切换到另一个集群运行作业。 • 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”，且确保两个集群下分别存在该作业。您可以在其中一个集群新建CDM作业并导出，然后再导入作业到另一个集群，实现作业同步，具体操作方法请参见**导出导入CDM作业**。

参数	是否必选	说明

CDM作业类型	是	• 选择已存在的作业。
		• 创建新作业。
		说明
		• 如果作业类型为“选择已存在的作业”，当CDM作业有修改时，此处作业节点不会同步更新。如需更新此作业节点，需要重新保存该节点所在的作业，用于触发CDM作业更新。
		• 如果作业类型为“创建新作业”，节点运行时会检测是否有同名CDM作业。
		  • 如果CDM作业未运行，则按照请求体内容更新同名作业。
		  • 如果同名CDM作业正在运行中，则等待作业运行完成后更新该作业。在此期间该作业可能被其他任务启动，可能会导致数据抽取不符合预期（如作业配置未更新、运行时间宏未替换正确等），因此请注意不要创建多个同名作业。

CDM作业名称	否	仅当“作业类型”为“选择已存在的作业”时需要配置该参数。选择待执行的CDM作业。 如果此CDM作业使用了在数据开发时配置的**作业参数**或者**变量**，则后续在数据开发模块调度此节点，可以间接实现CDM作业根据参数变量进行数据迁移。
CDM作业消息体	否	仅当“作业类型”为“创建新作业”时需要配置该参数。此处需要填写CDM作业JSON。方便起见可以在CDM已有作业处选择操作“更多 > 查看作业JSON”，复制其中的JSON内容，在此处修改适配。 如果此CDM作业使用了在数据开发时配置的**作业参数**或者**变量**，则后续在数据开发模块调度此节点，可以间接实现CDM作业根据参数变量进行数据迁移。
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。 默认情况下，节点名称会与选择的CDM作业保持同步。若不需要节点名称和作业名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。

表 9-92 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，该节点将会再次重试。

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** ● 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。 ● 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。 ● 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，系统支持再重试。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-93 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。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。

参数	说明
编辑	单击，修改节点输入功能的参数配置，修改完成后，请保存。
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
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。
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。

9.13.4 Data Migration

功能

该节点用于执行一个集成作业，Data Migration节点支持离线处理集成作业和实时处理集成作业。

参数

用户可参考表9-94和表9-95配置Data Migration节点的参数。

表 9-94 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。
CDM集群名称	是	选择CDM集群。如需查看集群列表，请单击下拉框右侧的。系统最多允许选择16个集群。

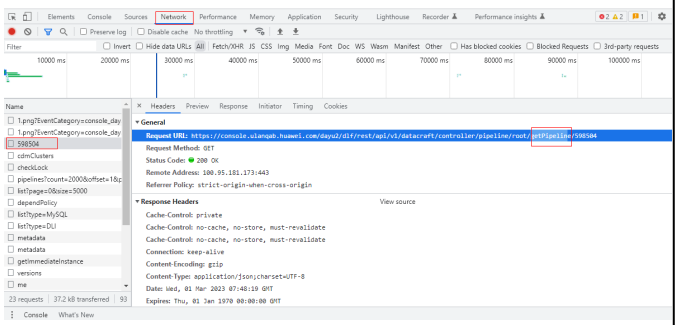
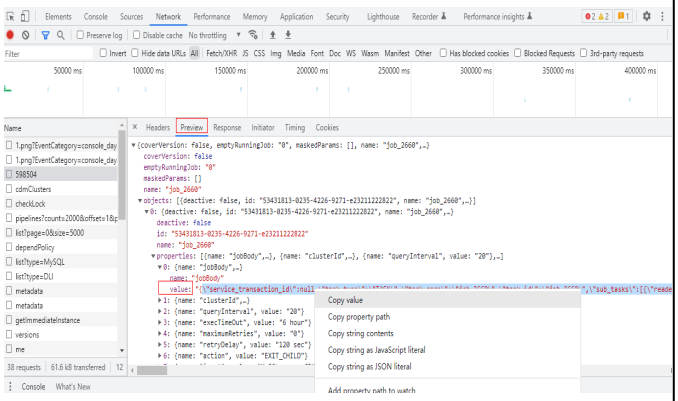
参数	是否必选	说明
CDM作业消息体	是	输入CDM作业消息体，作业消息体内容为JSON格式。 消息体JSON内容获取方法如下： 1. 参考**新建离线处理集成作业**创建一个单任务数据迁移作业。 2. 在键盘上按F12，打开创建好的单任务数据迁移作业，选择“network”页签。该任务请求方式为getPipeline。 **图 9-145** 请求方式 getPipeline  3. 在“Preview”的jobBody里面的“value”字段获取JSON消息体的内容。 **图 9-146** JSON 消息体内容  4. 将获取到的消息体内容复制到CDM作业消息体里面，可以对JSON消息体进行编辑。 5. 单击“保存”。

表 9-95 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，该节点将会再次重试。
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** ● 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。 ● 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，系统支持再重试。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.5 DIS Stream

功能

通过DIS Stream节点查询DIS通道的状态，如果DIS通道运行正常，继续执行后续的节点；如果DIS通道运行异常，DIS Stream将报错并退出，此时如果需要继续执行后续的节点，请配置“失败策略”为“继续执行下一节点”，请参见表9-97。

参数

用户可参考表9-96和表9-97配置DIS Stream节点的参数。

表 9-96 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。 默认情况下，节点名称会与选择的通道名称保持同步。若不需要节点名称和通道名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。
通道名称	是	选择或输入待查询的DIS通道，输入通道名称时支持引用作业参数和使用EL表达式（参见**表达式概述**）。 如需新建DIS通道，请参考以下方法： ● 单击，前往数据集成模块的“通道管理”页面新建DIS通道。 ● 前往DIS管理控制台进行新建。

表 9-97 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.6 DIS Dump

功能

通过DIS Dump节点配置DIS的数据转储任务。

参数

用户可参考**表9-98**和**表9-99**配置DIS Dump节点的参数。

表 9-98 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。
通道名称	是	选择或输入待配置的DIS通道，输入通道名称时支持引用作业参数和使用EL表达式（参见**表达式概述**）。 如需新建DIS通道，请参考以下方法： ● 单击，前往数据开发模块的“通道管理”“”页面新建DIS通道。 ● 前往DIS管理控制台进行新建。

参数	是否必选	说明
转储任务重名策略	是	选择重名策略。当“转储服务类型”配置的转储任务名称出现重名时，DIS Dump将根据重名策略进行下一步操作。 ● 忽略：不添加转储任务，并退出DIS Dump，DIS Dump的状态为“成功”。 ● 覆盖：继续添加转储任务，覆盖已存在的重名转储任务。
转储服务类型	是	1. 选择转储服务类型，目前支持转储至： ● OBS：通道里的流式数据存储在DIS中，并周期性导入对象存储服务OBS；通道里的实时文件数据传输完成后，导入OBS。 2. 单击，在弹出的对话框中配置转储任务的参数（参数说明请见《数据接入服务用户指南》的管理转储任务）。

表 9-99 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.7 DIS Client

功能

通过DIS Client节点可以给DIS通道发送消息。

您可以参考**跨空间进行作业调度**，获取DIS Client节点的使用案例。

参数

用户可参考**表9-100**配置DIS Client节点的参数。

表 9-100 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。
是否使用DIS数据通道连接	否	若使用数据通道连接，可以向其他账号的DIS通道发送消息；若不使用，仅能给本账号下所有region的通道发送消息。

参数	是否必选	说明
数据通道连接名称	否	仅当“是否使用DIS数据通道连接”选择为“是”时，需要配置此参数。 配置本参数前需在管理中心中组件**创建DIS连接**，然后在此处进行选择。 当“是否使用DIS数据通道连接”选择为“否”时，无需配置。
通道所属Region	否	使用DIS Client节点发送消息至目标DIS通道时，目标通道所在的Region。
通道名称	是	需要发送消息的DIS通道。可以直接输入DIS通道地址或选择DIS通道。
发送数据	是	发送到DIS通道的文本内容。可以直接输入文本或单击 使用EL表达式编辑。
相关作业	否	选择相关作业，您可以选择批作业或实时作业，最多只能选择10个作业。 相关作业参数用于节点运行后，方便跳转到对应作业的监控列表。选择完相关作业，单击“前往监控”在“作业监控”页面选择DIS Client节点时，单击页面下方的“查看相关作业”按钮可以查看相关作业。在“相关作业”页面，单击“查看”能跳转到对应的作业。

表 9-101 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.8 Rest Client

功能

通过Rest Client节点执行一个华为云内的RESTful请求。

Rest Client算子的具体使用教程，请参见**获取Rest Client算子返回值教程**。

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。您需要拥有弹性云服务器ECS，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。

Rest Client算子目前不支持大量的response返回体，目前代码限制30M。

参数

用户可参考**表9-102**，**表9-103**和**表9-104**配置Rest Client节点的参数。

表 9-102 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
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。 **说明** 代理集群可选多个，其中只要有一个集群可以正常连接即可。如果有多个集群可正常连接，则数据开发后台会随机选择一个用于连接。
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 ● DELETE
接口认证方式	是	● IAM认证：接口只允许云用户访问。DataArts Studio服务给接口发送消息的时候，会在请求消息头中带上当前用户的认证信息。 ● 无认证：接口不需要身份认证 ● 用户名密码认证：接口需要访问者输入账号和密码信息。DataArts Studio服务发送消息的时候，会在请求消息头中带上Authorization字段。 **说明** 如果使用用户名密码认证方式，您需要选择一个支持使用用户名密码进行认证的“数据连接”。
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。

参数	是否必选	说明
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"
		}
		其中“param4”属性的路径为“inner.inner.param4”。
		您也可以参考**获取Rest Client算子返回值教程**，获取本参数的配置案例。

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。

参数	是否必选	说明
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}。 **说明** 参数名（例如var4）作为作业变量，在本作业有唯一性约束。

表 9-103 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。

参数	是否必选	说明
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-104 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。
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。
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。

9.13.9 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

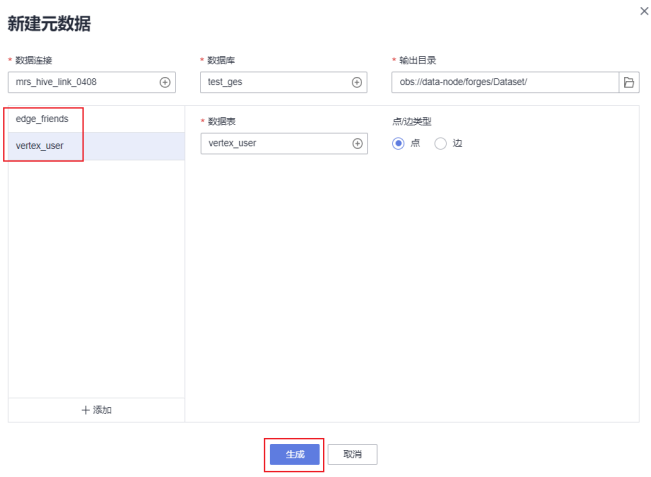
用户可参考表9-105和表9-106配置Import GES节点的参数。

表 9-105 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。
元数据来源	是	元数据来源支持以下两种方式： • 已有文件：从OBS桶中选择已有的xml格式元数据文件。 • 新建元数据：根据MRS Hive中的点表和边表，生成xml格式元数据文件到 OBS桶中。 **说明** 请至少输入元数据、边数据集与点数据集中的其中一个字段。

参数	是否必选	说明

元数据	否	根据“元数据来源”的选择，本参数有不同的填写方式。
		● 如果元数据来源为已有文件，单击输入框中的并选择对应的元数据文件 。
		● 如果元数据来源为新建元数据，单击输入框中的，进入新建元数据的界面，分别选择MRS Hive中的点表和边表，并填写元数据输出的OBS路径，单击生成元数据，系统会自动生成xml格式的元数据文件并回填到OBS路径。
		其中MRS Hive中的点表和边表，即为按GES图数据格式要求标准化后的边数据集和点数据集，需要与“边数据集”和“点数据集”参数所选的OBS桶中边数据集和点数据集保持一致。
		点数据集和边数据集应符合GES图数据格式要求。图数据格式要求简要介绍如下，详情可参见**一般图数据格式**。
		– 点数据集罗列了各个点的数据信息。一行为一个点的数据。格式如下所示，id是点数据的唯一标识。
		`id,label,property 1,property 2,property 3,…`
		– 边数据集罗列了各个边的数据信息，一行为一条边的数据。GES中图规格是以边的数量进行定义的，如一百万边。格式如下所示，id 1、id 2是一条边的两个端点的id。
		`id 1, id 2, label, property 1, property 2, …`
		说明
		选择新建元数据时，有如下注意事项：
		1. 生成元数据时，目前仅支持选择单标签（Label）场景的点表和边表。如果点表或边表中存在多个标签，则生成的元数据会存在缺失。
		2. 生成元数据xml文件是手动单击“生成元数据”触发的，如果在该节点在后续的作业调度运行中，点表和边表结构发生变化，元数据xml文件并不会随之更新，需要手动进入新建元数据窗口，再次单击“生成元数据”重新生成新的元数据xml文件。
		3. 生成的元数据xml文件，属性（Property）中的数据复合类型（Cardinality），目前仅支持填写为“single”类型，不支持自定义。
		4. 生成元数据功能本身，支持一次生成多对点表和边表的元数据xml文件。但考虑到Import GES节点的“边数据集”和“点数据集”参数，分别只能选择一张表，建议您在有多对点表和边表的情况下，分拆多个Import GES节点分别导入，以确保导入图数据时，元数据与每对点表和边表能够一一对应。

参数	是否必选	说明
		图 9-147 新建元数据 

边数据集	否	可以直接选择对应的OBS桶中的边数据集csv文件，也支持选择对应的边数据集的OBS路径。
		点数据集和边数据集应符合GES图数据格式要求。图数据格式要求简要介绍如下，详情可参见**一般图数据格式**。
		• 点数据集罗列了各个点的数据信息。一行为一个点的数据。格式如下所示，id是点数据的唯一标识。
		`id,label,property 1,property 2,property 3,…`
		• 边数据集罗列了各个边的数据信息，一行为一条边的数据。GES中图规格是以边的数量进行定义的，如一百万边。格式如下所示，id 1、id 2是一条边的两个端点的id。
		`id 1, id 2, label, property 1, property 2, …`

点数据集	否	可以直接选择对应的点数据集，也支持选择对应的点数据集的OBS路径。
		点数据集和边数据集应符合GES图数据格式要求。图数据格式要求简要介绍如下，详情可参见**一般图数据格式**。
		• 点数据集罗列了各个点的数据信息。一行为一个点的数据。格式如下所示，id是点数据的唯一标识。
		`id,label,property 1,property 2,property 3,…`
		• 边数据集罗列了各个边的数据信息，一行为一条边的数据。GES中图规格是以边的数量进行定义的，如一百万边。格式如下所示，id 1、id 2是一条边的两个端点的id。
		`id 1, id 2, label, property 1, property 2, …`

参数	是否必选	说明
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 9-106 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.10 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9-107**和**表9-108**配置MRS Kafka的参数。

表 9-107 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。具体操作请参见**从零开始使用Kafka**。
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 9-108 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.11 Kafka Client

功能

通过Kafka Client向Kafka的Topic中发送数据。

您可以参考**跨空间进行作业调度**，获取Kafka Client节点的使用案例。

参数

用户可参考**表9-109**配置Kafka Client节点的参数。

表 9-109 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 9-110 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。

参数	是否必选	说明
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.12 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 9-111 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。
实例所属Region	是	选择一个已存在的实例所属Region。
ROMA实例	是	选择一个已存在的ROMA实例。 DataArts Studio支持跨资源空间选择ROMA实例。
FDI任务	是	选择一个已存在的ROMA FDI任务。 DataArts Studio支持跨资源空间选择FDI任务。

表 9-112 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.13 DLI Flink Job

功能

DLI Flink Job节点用于创建和启动作业，或者查询DLI作业是否正在运行，实现实时流式大数据分析。

DLI Flink流式作业提交到DLI之后，若处于运行中的状态，则认为节点执行成功。若作业配置了周期调度，则会周期检查该Flink作业是否依然处于运行中的状态，如果处于运行状态，则认为节点执行成功。

参数

DLI Flink Job的参数配置，请参考以下内容：

- 属性参数：

  当作业类型为“Flink SQL作业”、“Flink OpenSource SQL作业”或“Flink自定义作业”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。

  - **选择已存在的Flink作业**：请参见表9-113。
  - **Flink SQL作业**：请参见表9-114。
  - **Flink OpenSource SQL作业**：请参见表9-115。
  - **Flink自定义作业**：请参见表9-116。
- 高级参数：表9-117

表 9-113 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 9-114 Flink SQL 作业-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。
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。
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。
脚本路径	是	选择需要执行的Flink SQL脚本。如果脚本未创建，请参考新建脚本和开发SQL脚本创建和开发Flink SQL脚本。

参数	是否必选	说明
脚本参数	否	关联的Flink SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的Flink SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
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。
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。 • 当前由于DLI的“default”队列默认Spark组件版本较低，可能会出现无法支持建表语句执行的报错，这种情况下建议您选择自建队列运行业务。如需“default”队列支持建表语句执行，可联系DLI服务客服或技术支持人员协助解决。 • DLI的“default”队列为共享队列，仅用于用户体验，用户间可能会出现抢占资源的情况，不能保证每次都可以得到资源执行相关操作。当遇到执行时间较长或无法执行的情况，建议您在业务低峰期再次重试，或选择自建队列运行业务。
CUs	是	CUs为DLI计费单位，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。
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。

表 9-115 Flink OpenSource SQL 作业-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。
作业类型	是	选择“Flink OpenSource SQL作业”。用户采用编写SQL语句来启动作业。

参数	是否必选	说明
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。
脚本路径	是	选择需要执行的Flink SQL脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**创建和开发Flink SQL脚本。
脚本参数	否	关联的Flink SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的Flink SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
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。
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。 • DLI的“default”队列为共享队列，仅用于用户体验，用户间可能会出现抢占资源的情况，不能保证每次都可以得到资源执行相关操作。当遇到执行时间较长或无法执行的情况，建议您在业务低峰期再次重试，或选择自建队列运行业务。
CUs	是	CUs为DLI计费单位，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。
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。

表 9-116 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”。

参数	是否必选	说明
jar包资源	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。 ● 当前由于DLI的“default”队列默认Spark组件版本较低，可能会出现无法支持建表语句执行的报错，这种情况下建议您选择自建队列运行业务。如需“default”队列支持建表语句执行，可联系DLI服务客服或技术支持人员协助解决。 ● DLI的“default”队列为共享队列，仅用于用户体验，用户间可能会出现抢占资源的情况，不能保证每次都可以得到资源执行相关操作。当遇到执行时间较长或无法执行的情况，建议您在业务低峰期再次重试，或选择自建队列运行业务。
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。关于自定义镜像的更多详情，请参见**自定义镜像**。
CUs	是	CUs为DLI计费单位，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。

参数	是否必选	说明
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 9-117 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.14 DLI SQL

功能

通过DLI SQL节点传递SQL语句到DLI SQL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9-118**，**表9-119**和**表9-120**配置DLI SQL节点的参数。

表 9-118 属性参数

参数	是否必选	说明
SQL或脚本	是	可以选择SQL语句或SQL脚本。 ● SQL语句 单击“SQL语句”参数下的文本框，在“SQL语句”页面输入需要执行的SQL语句。 ● SQL脚本 在“SQL脚本”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。 **说明** 若选择SQL语句方式，数据开发模块将无法解析您输入SQL语句中携带的参数。
DLI数据目录	否	选择DLI的数据目录。 ● 在DLI默认的数据目录dli。 ● 在DLI所绑定的LakeFormation已创建元数据catalog。
数据库名称	是	选择SQL脚本时： 默认选择SQL脚本中设置的数据库，支持修改。 选择SQL语句时： ● DLI数据目录如果选择DLI默认的数据目录dli，表示为DLI的数据库和数据表。 ● DLI数据目录如果选择DLI所绑定的LakeFormation已创建元数据catalog，表示为LakeFormation的数据库和数据表。

参数	是否必选	说明
DLI环境变量	否	• 环境变量配置项需要以"hoodie."或"dli.sql."或"dli.ext."或"dli.jobs."或"spark.sql."或"spark.scheduler.pool"开头。 • 环境变量的key为dli.sql.shuffle.partitions或dli.sql.autoBroadcastJoinThreshold时，不能包含>< 符号。 • 环境变量的key为dli.sql.autoBroadcastJoinThreshold的值只能为整数，环境变量的key为dli.sql.shuffle.partitions的值只能为正整数。 • 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。 **说明** 用户定义适用于此作业的配置参数。目前支持的配置项： • dli.sql.autoBroadcastJoinThreshold（自动使用BroadcastJoin的数据量阈值） • dli.sql.shuffle.partitions（指定Shuffle过程中Partition的个数） • dli.sql.cbo.enabled（是否打开CBO优化策略） • dli.sql.cbo.joinReorder.enabled（开启CBO优化时，是否允许重新调整join的顺序） • dli.sql.multiLevelDir.enabled（OBS表的指定目录或OBS表分区表的分区目录下有子目录时，是否查询子目录的内容；默认不查询） • dli.sql.dynamicPartitionOverwrite.enabled（在动态分区模式时，只会重写查询中的数据涉及的分区，未涉及的分区不删除）
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。 **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。 • 当前由于DLI的“default”队列默认Spark组件版本较低，可能会出现无法支持建表语句执行的报错，这种情况下建议您选择自建队列运行业务。如需“default”队列支持建表语句执行，可联系DLI服务客服或技术支持人员协助解决。 • DLI的“default”队列为共享队列，仅用于用户体验，用户间可能会出现抢占资源的情况，不能保证每次都可以得到资源执行相关操作。当遇到执行时间较长或无法执行的情况，建议您在业务低峰期再次重试，或选择自建队列运行业务。

参数	是否必选	说明
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。
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 **说明** 脏数据即Bad Records，由于数据类型不兼容、数据为空或者格式不兼容而导致无法加载到DLI中的记录归类为Bad Records。 选择记录脏数据后，Bad Records不会导入到目标表，而是导入到OBS脏数据路径中。 • 如果未配置工作空间中的DLI脏数据OBS路径，则默认会把DLI SQL执行过程中的脏数据写到dlf-log-{projectId}桶中。 • 若要自定义DLI脏数据日志路径，请前往空间管理进行编辑操作。详细操作请参见**配置OBS桶**。

表 9-119 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，该节点将会再次重试。

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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-120 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。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。
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。

9.13.15 DLI Spark

功能

通过DLI Spark节点执行一个预先定义的Spark作业。

DLI Spark节点的具体使用教程，请参见**开发一个DLI Spark作业**。

参数

用户可参考**表9-121**，**表9-122**和**表9-123**配置DLI Spark节点的参数。

表 9-121 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
DLI队列	是	下拉选择需要使用的队列。 **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。 • 当前由于DLI的“default”队列默认Spark组件版本较低，可能会出现无法支持建表语句执行的报错，这种情况下建议您选择自建队列运行业务。如需“default”队列支持建表语句执行，可联系DLI服务客服或技术支持人员协助解决。 • DLI的“default”队列为共享队列，仅用于用户体验，用户间可能会出现抢占资源的情况，不能保证每次都可以得到资源执行相关操作。当遇到执行时间较长或无法执行的情况，建议您在业务低峰期再次重试，或选择自建队列运行业务。
Spark版本	否	选定DLI队列后，下拉可选择作业使用Spark组件的版本号，使用时如无特定版本要求时使用默认版本号2.3.2，有特殊使用要求时选择对应的版本即可。
作业特性	否	作业使用的Spark镜像类型，当前支持基础型、AI增强型和自定义的Spark镜像。 自定义镜像需要选择自定义镜像名称和对应版本。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。
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**。

参数	是否必选	说明
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”。
组名称	否	当“资源类型”选择了“DLI程序包”时，需要设置。 ● 选择“已有分组”：可选择已有的分组。 ● 选择“创建新分组”：可输入自定义的组名称。 ● 选择“不分组”：不需要选择或输入组名称。
主类入口参数	否	用户自定义参数，多个参数请以Enter键分隔。 应用程序参数支持全局变量替换。例如，在“全局配置”>“全局变量”中新增全局变量key为batch_num，可以使用{{batch_num}}，在提交作业之后进行变量替换。
Spark作业运行参数	否	以“key/value”的形式设置提交Spark作业的属性，多个参数以Enter键分隔。具体参数请参见**Spark Configuration**。 Spark参数value支持全局变量替换。例如，在“全局配置”>“全局变量”中新增全局变量key为custom_class，可以使用"spark.sql.catalog"={{custom_class}}，在提交作业之后进行变量替换。 **说明** Spark作业不支持自定义设置jvm垃圾回收算法。

参数	是否必选	说明
Module名称	否	DLI系统提供的用于执行跨源作业的依赖模块，访问各个不同的服务，选择不同的模块： • CloudTable/MRS HBase: sys.datasource.hbase • DDS：sys.datasource.mongo • CloudTable/MRS OpenTSDB: sys.datasource.opentsdb • DWS: sys.datasource.dws • RDS MySQL: sys.datasource.rds • RDS PostGre: sys.datasource.rds • DCS: sys.datasource.redis • CSS: sys.datasource.css DLI内部相关模块： • sys.res.dli-v2 • sys.res.dli • sys.datasource.dli-inner-table
访问元数据	是	是否通过Spark作业访问元数据。具体请参考**使用Spark作业访问DLI元数据**。

表 9-122 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-123 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。
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。
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。

参数	说明
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。

9.13.16 DWS SQL

功能

通过DWS SQL节点传递SQL语句到DWS中执行。

DWS SQL算子的具体使用教程，请参见**开发一个DWS SQL脚本作业**。

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9-124**，**表9-125**和**表9-126**配置DWS SQL节点的参数。

表 9-124 属性参数

参数	是否必选	说明
SQL或脚本	是	可以选择SQL语句或SQL脚本。 • SQL语句 单击“SQL语句”参数下的文本框，在“SQL语句”页面输入需要执行的SQL语句。 • SQL脚本 在“SQL脚本”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。 **说明** 若选择SQL语句方式，数据开发模块将无法解析您输入SQL语句中携带的参数。
数据连接	是	默认选择SQL脚本中设置的数据连接，支持修改。
数据库	是	默认选择SQL脚本中设置的数据库，支持修改。

参数	是否必选	说明
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
脏数据表	否	SQL脚本中定义的脏数据表名称。 脏数据属性用户不能编辑，自动从SQL脚本内容中关联推荐。 DWS脏数据表的语法：**with** table_name或**log into** table_name
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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。

表 9-125 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，该节点将会再次重试。

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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-126 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。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。
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。

9.13.17 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9-127**，**表9-128**和**表9-129**配置MRS Spark SQL节点的参数。

表 9-127 属性参数

参数	是否必选	说明
MRS作业名称	否	MRS的作业名称。 如果未设置MRS作业名称且选择直连模式时，节点名称只能由英文字母、数字、中划线和下划线组成，长度不能超过64个字符，不能包含中文字符。 系统支持MRS作业名称按照**作业名称_节点名称**格式自动填入。
SQL脚本	是	选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。
数据连接	是	默认选择SQL脚本中设置的数据连接，支持修改。

参数	是否必选	说明
MRS资源队列	否	选择已创建好的MRS资源队列。 **说明** • 数据连接为MRS API连接时支持为Spark SQL作业独立配置需要的资源（例如线程、内存、CPU核数并指定MRS资源队列等）。代理连接时不支持配置。 • 您需要先在数据安全组件中配置对应的队列（参考**配置队列权限**，）后，才能在此处选择到已配置的队列。当有多处同时配置了资源队列时，此处配置的资源队列为最高优先级。
数据库	是	默认选择SQL脚本中设置的数据库，支持修改。 MRS API连接方式下不支持选择数据库。
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
运行程序参数	否	为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 **说明** • 数据连接为MRS API连接时支持为Spark SQL作业独立配置需要的资源（例如线程、内存、CPU核数并指定MRS资源队列等）。代理连接时不支持配置。 • 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“**运行SparkSql作业** > 表2 运行程序参数”。
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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。

表 9-128 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-129 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。

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。
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。

9.13.18 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。

MRS Hive SQL节点的具体使用教程，请参见**开发一个Hive SQL作业**。

说明

MRS Hive SQL节点不支持Hive的事务表。

参数

用户可参考**表9-130**，**表9-131**和**表9-132**配置MRS Hive SQL节点的参数。

表 9-130 属性参数

参数	是否必选	说明
MRS作业名称	否	MRS的作业名称。 如果未设置MRS作业名称且选择直连模式时，节点名称只能由英文字母、数字、中划线和下划线组成，长度不能超过64个字符，不能包含中文字符。 系统支持MRS作业名称按照**作业名称_节点名称**格式自动填入。
SQL脚本	是	选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。
数据连接	是	默认选择SQL脚本中设置的数据连接，支持修改。
数据库	是	默认选择SQL脚本中设置的数据库，支持修改。
MRS资源队列	否	选择已创建好的MRS资源队列。 **说明** 需要先在数据安全服务队列权限功能中，配置对应的队列后，才能在此处选择到已配置的队列。当有多处同时配置了资源队列时，此处配置的资源队列为最高优先级。
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮 同步。
运行程序参数	否	为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 **说明** 若集群为MRS 1.8.7版本或MRS 2.0.1之后版本，需要配置此参数。 MRS Hive SQL作业的运行程序参数，请参见《MapReduce用户指南》中的“**运行HiveSql作业** > 表2 运行程序参数”。
节点名称	是	默认显示为SQL脚本的名称，支持修改。规则如下： 节点名称只能由字母、数字、中划线和下划线组成，并且长度为1～64个字符。 **说明** 节点名称不得包含中文字符、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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。

表 9-131 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-132 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。

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。
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。

9.13.19 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9-133，表9-134和表9-135配置MRS Presto SQL节点的参数。

表 9-133 属性参数

参数	是否必选	说明
SQL或脚本	是	可以选择SQL语句或SQL脚本。 • SQL语句 单击“SQL语句”参数下的文本框，在“SQL语句”页面输入需要执行的SQL语句。 • SQL脚本 在“SQL脚本”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。 **说明** 若选择SQL语句方式，数据开发模块将无法解析您输入SQL语句中携带的参数。
数据连接	是	默认选择SQL脚本中设置的数据连接，支持修改。
模式	是	默认选择SQL脚本中设置的数据库，支持修改。
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。

表 9-134 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，该节点将会再次重试。

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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-135 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。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。
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。

9.13.20 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9-136**，**表9-137**和**表9-138**配置MRS Spark节点的参数。

表 9-136 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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。
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。

参数	是否必选	说明
MRS资源队列	否	选择已创建好的MRS资源队列。 **说明** 您需要先在数据安全服务队列权限功能中，配置对应的队列后，才能在此处选择到已配置的队列。当有多处同时配置了资源队列时，此处配置的资源队列为最高优先级。
Spark作业名称	是	MRS作业名称，只能由英文字母、数字、中划线和下划线组成，长度不能超过64个字符。 系统支持作业名称按照**作业名称_节点名称**格式自动填入。 **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。
运行模式	是	配置Spark作业的运行模式。 • 批处理：指Spark作业为批模式运行，节点会一直等待Spark作业执行完成才结束。 • 流处理：指Spark作业为流处理运行模式，节点执行时只要作业启动成功即执行成功。后续每次周期运行时检查任务是否处于运行状态，如果处于运行状态，则认为节点执行成功。 注意，此处不会为Spark增加对应的batch或streaming模式参数，您还需要为Spark作业指定对应参数。
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用户指南》中的**运行Spark作业**。
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 9-137 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-138 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。

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。
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。

9.13.21 MRS Spark Python

功能

通过MRS Spark Python节点实现在MRS中执行预先定义的Spark Python作业。

MRS Spark Python算子的具体使用教程，请参见**开发一个MRS Spark Python作业**。

参数

用户可参考**表9-139**，**表9-140**和**表9-141**配置MRS Spark Python节点的参数。

表 9-139 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
作业名称	是	MRS作业名称，只能由英文字母、数字、中划线和下划线组成，长度不能超过64个字符。 系统支持作业名称按照**作业名称_节点名称**格式自动填入。 **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。
脚本类型	是	• 离线脚本 • 在线脚本
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》中的**创建集群**章节。
MRS资源队列	否	选择已创建好的MRS资源队列。 **说明** 您需要先在数据安全服务队列权限功能中，配置对应的队列后，才能在此处选择到已配置的队列。当有多处同时配置了资源队列时，此处配置的资源队列为最高优先级。
SQL脚本	是	仅“脚本类型”配置为“在线脚本”时可以配置。 选择已创建的Spark Python脚本。
脚本参数	否	仅“脚本类型”配置为“在线脚本”时可以配置。 关联的Spark Python脚本中如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。
运行程序参数	否	仅“脚本类型”配置为“在线脚本”时可以配置。 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用户指南》中的**运行Spark作业**。
参数	是	仅“脚本类型”配置为“离线脚本”时可以配置。 输入参数信息，多个参数间使用Enter键分隔。

参数	是否必选	说明
执行程序参数	否	输入MRS的执行程序参数。 不同参数间用空格隔开，可通过在参数名前添加@的方式防止参数信息被明文存储。
属性	否	输入key=value格式的参数，多个参数间使用Enter键分割。

表 9-140 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-141 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。
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。
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。

9.13.22 MRS ClickHouse

功能

通过MRS ClickHouse节点执行数据开发模块中预先定义的ClickHouse SQL脚本。

参数

用户可参考表9-142，表9-143和表9-144配置MRS ClickHouse节点的参数。

表 9-142 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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。
SQL脚本	是	选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
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。
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。

表 9-143 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，该节点将会再次重试。

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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-144 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。
确定	单击“确认”，保存节点输入功能的参数配置。
取消	单击“取消”，取消节点输入功能的参数配置。
编辑	单击  ，修改节点输入功能的参数配置，修改完成后，请保存。

参数	说明
删除	单击，删除节点输入功能的参数配置。
查看表详情	单击，查看节点输入血缘关系创建数据表的详细信息。
输出	
新建	单击“新建”，在“类型”的下拉选项中选择要新建的类型。可以选择DWS，OBS，CSS，HIVE，CUSTOM和DLI类型。
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。

9.13.23 MRS Impala SQL

功能

通过MRS Impala SQL节点执行数据开发模块中预先定义的Impala SQL脚本。

参数

用户可参考**表9-145**和**表9-146**配置MRS Impala节点的参数。

表 9-145 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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。
SQL脚本	是	选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。
数据连接	是	默认选择SQL脚本中设置的数据连接，支持修改。
数据库	是	默认选择SQL脚本中设置的数据库，支持修改。

参数	是否必选	说明
资源队列	否	输入资源队列名称。
脚本参数	否	关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。

表 9-146 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。

参数	是否必选	说明
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-147 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。
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。
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。

9.13.24 MRS Flink Job

功能

通过MRS Flink Job节点执行数据开发模块中预先定义的Flink SQL脚本和Flink作业。

MRS Flink Job节点的具体使用教程，请参见**开发一个MRS Flink作业**。

参数

用户可参考表9-148和表9-149配置MRS Flink节点的参数。

表 9-148 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。
作业类型	是	选择 ● Flink SQL作业 ● Flink 自定义作业
脚本路径	是	选择Flink SQL作业时，可配置此参数。 选择需要执行的Flink SQL脚本。如果脚本未创建，请参考新建脚本和开发SQL脚本先创建和开发Flink SQL脚本。
脚本参数	否	选择Flink SQL作业时，可配置此参数。 关联的SQL脚本如果使用了参数，此处显示参数名称，请在参数名称后的输入框配置参数值。参数值支持使用EL表达式。 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步。
运行模式	是	配置Flink作业的运行模式。 ● 批处理：指Flink作业为批模式运行，节点会一直等待Flink作业执行完成才结束。 ● 流处理：指Flink作业为流处理运行模式，节点执行时只要作业启动成功即执行成功。后续每次周期运行时检查任务是否处于运行状态，如果处于运行状态，则认为节点执行成功。 注意，此处不会为Flink增加对应的batch或streaming模式参数，您还需要为Flink作业指定对应参数。
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。 **说明** MRS Flink Job目前支持的MRS集群版本是MRS 3.2.0-LTS.1及以上版本。

参数	是否必选	说明
Flink作业名称	是	MRS作业名称，只能由英文字母、数字、中划线和下划线组成，长度不能超过64个字符。 系统支持作业名称按照**作业名称_节点名称**格式自动填入。 **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。
MRS资源队列	否	选择已创建好的MRS资源队列。 **说明** 需要先在数据安全服务队列权限功能中，配置对应的队列后，才能在此处选择到已配置的队列。当有多处同时配置了资源队列时，此处配置的资源队列为最高优先级。
运行程序参数	否	为本次执行的作业配置相关优化参数（例如线程、内存、CPU核数等），用于优化资源使用效率，提升作业的执行性能。 **说明** 若集群为MRS 1.8.7版本或MRS 2.0.1之后版本，需要配置此参数。 MRS Flink作业的运行程序参数，请参见《MapReduce用户指南》中的**运行Flink作业**。
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 9-149 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，该节点将会再次重试。

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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.25 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9-150**和**表9-151**配置MRS MapReduce节点的参数。

表 9-150 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，只能由英文字母、数字、中划线和下划线组成，长度不能超过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 9-151 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，该节点将会再次重试。

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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.26 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9-152**和**表9-153**配置CSS节点的参数。

表 9-152 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。
集群或数据连接	是	选择集群或数据连接。 集群方式不支持开启安全模式的CloudSearch集群，请使用数据连接方式。
CloudSearch集群	是	选择“集群”时，才需要配置。 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“集群”时，才需要配置。 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。
数据连接	是	选择“数据连接”时，才需要配置。 选择已创建好数据连接。
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
请求消息体	否	Json格式的请求消息体。 仅当请求类型为POST、PUT和HEAD时，根据实际需要才需要配置请求消息体。
CloudSearch输出路径	否	选择输出数据的存储路径。

表 9-153 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.27 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9-154**和**表9-155**配置Shell节点的参数。

表 9-154 属性参数

参数	是否必选	说明
Shell或脚本	是	可以选择Shell语句或Shell脚本。 • Shell语句 单击“Shell语句”参数下的文本框，在“Shell语句”页面输入需要执行的Shell语句。 • Shell脚本 在“脚本路径”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发Shell脚本**先创建和开发脚本。 **说明** 若选择Shell语句方式，数据开发模块将无法解析您输入Shell语句中携带的参数。 Shell节点运行的输出结果不能大于30M，大于30M会报错。
主机连接	是	选择执行Shell脚本的主机。 **须知** • Shell或Python脚本可以在该ECS主机上运行的最大并发数由ECS主机的**/etc/ssh/sshd_config**文件中MaxSessions的配置值确定。请根据Shell或Python脚本的调度频率合理配置MaxSessions的值。 • 连接主机的用户需要具有主机**/tmp**目录下文件的创建与执行权限。 • Shell和Python脚本都是发往ECS主机的**/tmp**目录下去运行的，需要确保**/tmp**目录磁盘不被占满。
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）。交互式参数之间以空格分隔，Shell脚本根据交互情况按顺序读取参数值。 read -p语法的使用示例： `read -p “输入参数1和参数2” 变量1 变量2`

参数	是否必选	说明
节点名称	是	节点名称，只能包含英文字母、数字、中文字符、中划线、下划线、/、<>和点号，且长度小于等于128个字符。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。

表 9-155 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
重试条件	否	失败重试选择“是”时，支持设置重试条件。 打开重试条件的开关，设置返回码的范围。 Shell作业可以根据返回码判断作业节点执行失败是否重试。用户可以定义Shell的返回结果码中哪些返回码可以重跑。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.28 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9-156**和**表9-157**配置RDS SQL节点的参数。

表 9-156 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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。
数据连接	是	选择数据连接。
数据库	是	填写数据库名称，该数据库已创建好，建议不要使用默认数据库。

参数	是否必选	说明
SQL或脚本	是	可以选择SQL语句或SQL脚本。 ● SQL语句 单击“SQL语句”参数下的文本框，在“SQL语句”页面输入需要执行的SQL语句。 ● SQL脚本 在“SQL脚本”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。 **说明** 若选择SQL语句方式，数据开发模块将无法解析您输入SQL语句中携带的参数。

表 9-157 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.29 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。

说明

目标端是DWS的ETL Job节点，不支持使用委托进行调度，建议采用兼容性更佳的公共IAM账号方式进行调度，详见**配置调度身份**。

参数

用户可参考**表9-158**，**表9-159**和**表9-160**配置ETL Job节点的参数。

表 9-158 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
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 9-159 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-160 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。
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。
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。

参数	说明
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。

9.13.30 Python

须知

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。

功能

通过Python节点执行Python语句。

Python节点的具体使用教程，请参见**开发一个Python脚本**。

说明

Python节点支持脚本参数和作业参数。

参数

用户可以参考**表9-161**和**表9-162**配置Python节点的参数。

表 9-161 属性参数

参数	是否必选	说明
Python语句或脚本	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，选择Python脚本。 • Python脚本 在“Python脚本”参数后选择需要执行的Python脚本，系统自动默认显示Python版本，例如Python3。如果脚本未创建，请参考**新建脚本**和**开发Python脚本**先创建和开发脚本。 **说明** • 若选择Python语句方式，数据开发模块将无法解析您输入Python语句中携带的参数。 • 若选择Python脚本方式，系统自动默认显示的Python版本为创建Python脚本时所选择的Python版本。 • 对于原有的作业，默认使用Python2。 • Python节点运行的输出结果不能大于30M，大于30M会报错。
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。 **须知** • Shell或Python脚本可以在该ECS主机上运行的最大并发数由ECS主机的**/etc/ssh/sshd_config**文件中MaxSessions的配置值确定。请根据Shell或Python脚本的调度频率合理配置MaxSessions的值。 • 连接主机的用户需要具有主机**/tmp**目录下文件的创建与执行权限。 • Shell和Python脚本都是发往ECS主机的**/tmp**目录下去运行的，需要确保**/tmp**目录磁盘不被占满。
参数	否	填写执行Python语句时，向语句传递的参数，参数之间使用空格分隔，例如：a b c。此处的“参数”需要在Python语句中引用，否则配置无效。
交互式输入	否	填写交互式参数，即执行Python语句的过程中，需要用户输入的交互式信息（例如密码）。交互式参数之间以空格分隔，Python语句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。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。

表 9-162 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.31 DORIS SQL

功能

通过Doris SQL节点传递SQL语句到Doris中执行。

参数

用户可参考表9-163和表9-164配置Doris SQL节点的参数。

表 9-163 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。 默认情况下，节点名称会与选择的脚本名称保持同步。若不需要节点名称和脚本名称同步，请参考**禁用作业节点名称同步变化**禁用该功能。
SQL或脚本	是	• SQL语句 单击“SQL语句”参数下的文本框，在“SQL语句”页面输入需要执行的SQL语句。 • SQL脚本 在“SQL脚本”参数后选择需要执行的脚本。如果脚本未创建，请参考**新建脚本**和**开发SQL脚本**先创建和开发脚本。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。 **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	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。

表 9-164 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将会再次重试。

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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.32 ModelArts Train

功能

通过编排ModelArts Train算子，实现在DataArts Studio中调度ModelArts工作流。

参数

用户可参考**表9-165**和**表9-166**配置ModelArts Train节点的参数。

表 9-165 属性参数

参数	是否必选	说明
ModelArts工作空间	是	选择ModelArts工作空间。该工作空间必须与DataArts Studio在同一区域、同一Region。
工作流版本	是	选择ModelArts工作流版本。 ● V1 ● V2
ModelArts工作流	是	选择ModelArts工作流。该工作流必须是与DataArts Studio在同一区域，同一Region的ModelArts 工作流。
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 9-166 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.33 Create OBS

说明

OBS路径不支持s3a://开头的日志路径。

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9-167和表9-168配置Create OBS节点的参数。

表 9-167 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
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 9-168 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.34 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9-169和表9-170配置Delete OBS节点的参数。

表 9-169 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 9-170 高级参数

参数	是否必选	说明
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.35 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9-171**，**表9-172**和**表9-173**配置OBS Manager节点的参数。

表 9-171 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
操作类型	是	通过节点可以执行的操作： ● 移动文件：将源文件或目录，移动到新目录中。 ● 复制文件：复制源文件或目录。 ● 重命名文件：重命名文件仅支持最后一级目录或文件重命名。 如重命名目录时，源文件或目录：obs://test/a/b/c/，目的目录：obs://test/a/b/d/；重命名文件时，源文件或目录：obs://test/a/b/hello.txt，目的目录：obs://test/a/b/bye.txt ● 监测文件：监测文件或目录是否存在，如不存在则此节点运行失败，否则运行成功。 如果当前作业需要根据文件或目录是否存在，从而进行不同的处理，则可以根据本节点的执行状态设置IF条件判断，具体请参考**IF条件判断教程**章节。
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 9-172 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

表 9-173 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。
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。
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。

参数	说明
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。

9.13.36 Open/Close Resource

功能

通过Open/Close Resource节点按需开启或关闭华为云服务。

参数

用户可参考表9-174和表9-175配置Open/Close Resource节点的参数。

表 9-174 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
服务	是	选择需要开机/关机的服务： • ECS • CDM
开关机设置	是	选择开关机类型： • 开 • 关
开关机对象	是	选择需要开机/关机的具体对象，例如开启某个CDM集群。

表 9-175 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.37 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9-176和表9-177配置Data Quality Monitor节点的参数。

表 9-176 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。如何创建质量作业，请参见的**新建数据质量作业**。
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。如何创建对账作业，请参见的**新建数据对账作业**。
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 9-177 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，该节点将会再次重试。

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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.38 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9-178**和**表9-179**配置Sub Job节点的参数。

表 9-178 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
子作业名称	是	选择需要调用的子作业名称。 **说明** 您只能选择已存在的批处理作业名称，此批处理作业不能为作业本身，并且该批处理作业为不包含Sub Job节点的作业。
子作业参数名称	是/否	● 当节点属性中子作业参数配置为空时，子作业使用自身参数变量执行。父作业的“子作业参数名称”不显现。 ● 当节点属性中子作业参数配置了数据时，子作业将使用配置参数变量执行。此时父作业的“子作业参数名称”显现，并且节点属性中子作业参数配置的数据或者EL表达式，将根据父作业的环境变量读取替换。

表 9-179 高级参数

参数	是否必选	说明
节点状态轮询时间（秒）	是	设置轮询时间（1~60秒），每隔x秒查询一次节点是否执行完成。
节点执行的最长时间	是	设置节点执行的超时时间，如果节点配置了重试，在超时时间内未执行完成，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 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”。 • 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”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.39 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。

For Each节点的具体使用教程，请参见**For Each节点使用介绍**。

说明

For Each节点单次运行时，指定的子作业最多循环执行1000次。

如果DLI SQL作为前置节点，For Each节点最多支持100个子作业。

参数

用户可参考**表9-180**配置For Each节点的参数。

表 9-180 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。

参数	是否必选	说明
子作业参数	否	仅当循环执行的子作业配置了作业参数后，出现该参数。参数名即子作业中定义的变量，参数值按如下原则填写： • 当循环执行的子作业需要根据父作业的变量读取替换时，则本参数为可配置为EL表达式，一般配置为#{Loop.current[0]}或#{Loop.current[1]}等，表示循环中取遍历到的数据集二维数组当前行的第一个值或第二个值等，详见**Loop内嵌对象**；循环执行的子作业的作业参数名配置后，参数值无需配置可置为空。 • 当循环执行的子作业需要使用自身参数变量运行时，则本参数可置为空；循环执行的子作业的作业参数需配置参数值。
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，数据集应为二维数组，每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。如二维数组：[['001'],['002'],['003']] 。 **说明** • 如果要让“00”“01”当成数字类型作为参数传递，需要配置为[["00"],["01"]];[[00],[01]];[['00'],['01']]。 • 如果要让“00”“01”当成字符类型作为参数传递，需要加上转义字符，例如：[["\"00\""],["\"01\""]];[['\'00\''],['\'01\'']]
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。 **说明** 如果子作业中包含CDM Job节点，子作业并发数需要设置为1。
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。

表 9-181 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，该节点将会再次重试。

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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： ● 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”。 ● 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”。 ● 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.40 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9-182**和**表9-183**配置SMN节点的参数。

表 9-182 属性参数

参数	是否必选	说明
节点名称	是	节点名称，可以包含中文、英文字母、数字、“_”、“-”、“/”、“<”、“>”等各类特殊字符，长度为1～128个字符。
主题名称	是	选择消息的主题，该主题已在SMN服务中创建好。
消息标题	否	自定义消息的标题，长度必须少于512个字符。
消息类型	是	选择消息的发送格式。 ● 文本消息：按文本格式发送的消息。 ● JSON消息：按JSON格式发送的消息，用户可对不同的订阅者类型发送不同的消息。 – 手动输入JSON格式的消息：在“消息内容”直接输入。 – 通过工具自动生成JSON格式的消息：单击“生成JSON消息”，在弹出的对话框中填写“消息”和选择“协议”。 ● 模板消息：按模板格式发送的消息，即固定格式的消息，可以通过tag的方式来处理变量的部分。 – 手动输入模板格式的消息：在“消息内容”直接输入。 – 通过工具自动生成模板格式的消息：单击“生成模板消息”，在弹出的对话框中，选择“模板名称”，并设置{tag}的值。

参数	是否必选	说明

消息内容	是	填写消息的内容，不同消息类型的填写要求如下：
		● 文本消息：大小不超过10KB。
		● JSON消息：JSON消息中必须有Default协议，大小不超过10KB。
		示例如下：
		{
		"default": "Dear Sir or Madam, this is a default message.",
		"email": "Dear Sir or Madam, this is an email message.",
		"http": "{'message':'Dear Sir or Madam, this is an HTTP message.'}",
		"https": "{'message':'Dear Sir or Madam, this is an HTTPS message.'}",
		"sms": "This is an SMS message."
		}
		● 模板消息：大小不超过10KB。
		示例如下：
		"message_template_name":"confirm_message",
		"tags":{
		"topic_urn":"urn:smn:regionId:xxxx:SMN_01"
		}
		其中，“message_template_name”为模板名称，“tags”为模板中所有的tag标签。
		如需了解更多SMN的配置说明，请参见《消息通知服务用户指南》。

表 9-183 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，该节点将会再次重试。
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。 – 超时重试 – 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，系统支持再重试。 当节点运行超时导致的失败不会重试时，您可前往“默认项设置”修改此策略。 当“失败重试”配置为“是”才显示“超时重试”。

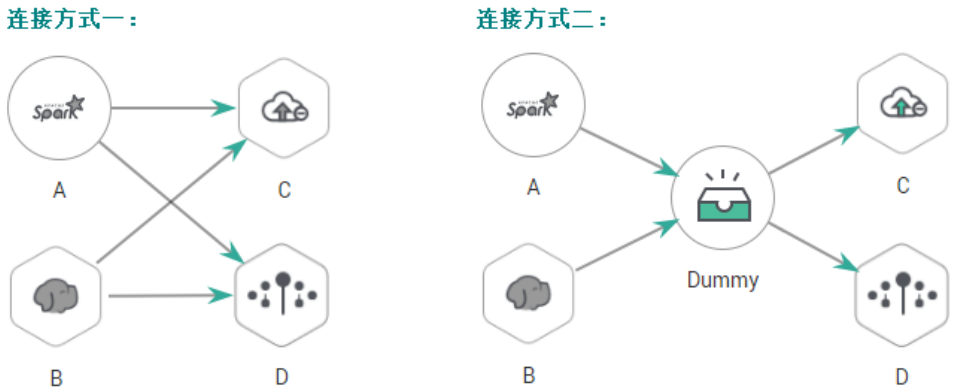
参数	是否必选	说明
当前节点失败后，后续节点处理策略	是	节点执行失败后的操作： • 终止后续节点执行计划：停止后续节点的运行，当前作业实例状态显示为“失败”。 • 终止当前作业执行计划：停止当前作业运行，当前作业实例状态显示为“失败”。 • 继续执行下一节点：忽略当前节点失败，当前作业实例状态显示为“忽略失败”。 • 挂起当前作业执行计划：当前作业实例的状态为运行异常，该节点的后续节点以及依赖于当前作业的后续作业实例都会处于等待运行状态。
是否空跑	否	如果勾选了空跑，该节点不会实际执行，将直接返回成功。
任务组	否	选择任务组。任务组配置好后，可以更细粒度的进行当前任务组中的作业节点的并发数控制，比如作业中包含多个节点、补数据、重跑等场景。

9.13.41 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9-148所示。

图 9-148 连接方式对比



参数

用户可参考表9-184配置Dummy节点的参数。

表 9-184 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。

9.14 EL 表达式参考

9.14.1 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")}”，如图9-149所示。

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。

图 9-149 表达式示例

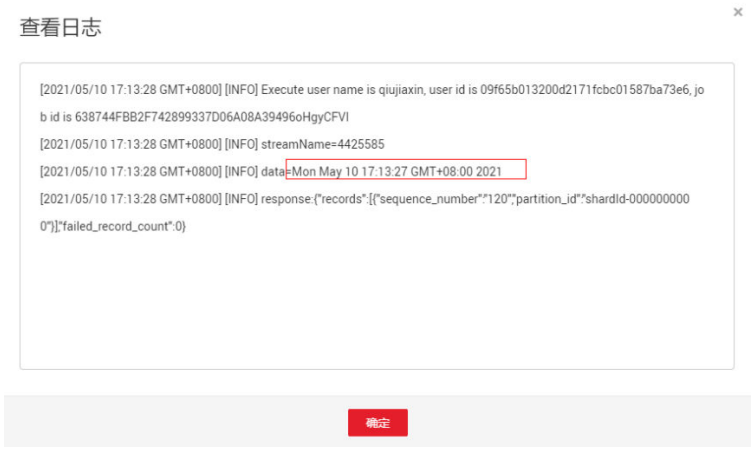
EL表达式在数据开发过程中被广泛应用，您可以参考**最佳实践**查看更多应用EL表达式的进阶实践。

调试方法介绍

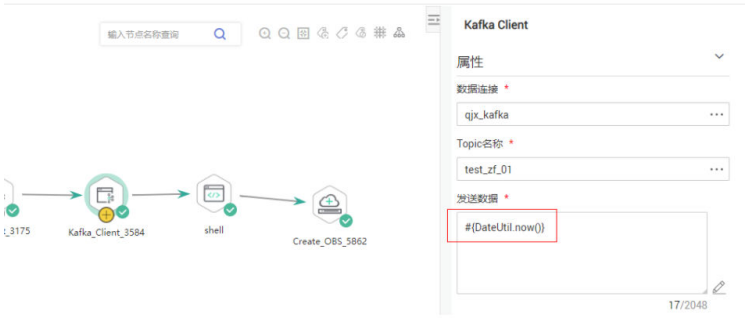
下面介绍几种EL表达式的调试方法，能够在调试过程中方便地看到替换结果。

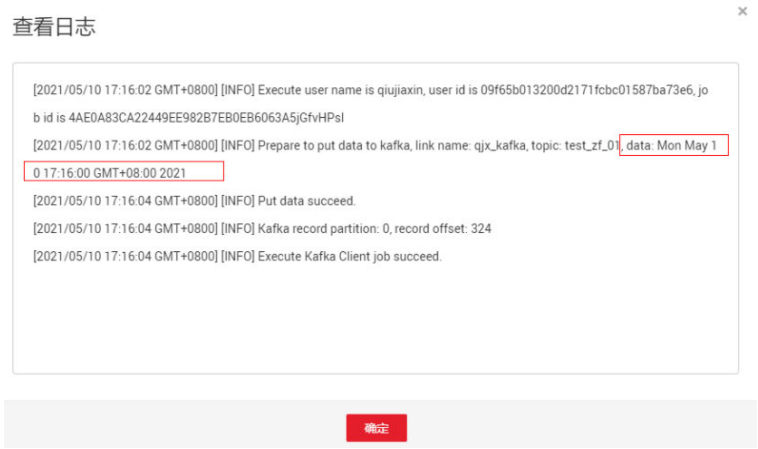
后文以#{DateUtil.now()}表达式为例进行介绍。

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，单击“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，单击“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。

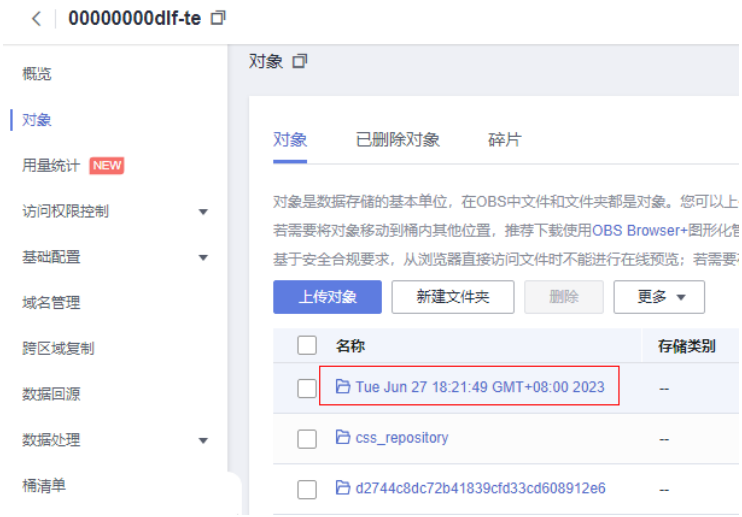




3. 使用Shell节点。
   - 前提：您需要具备弹性云服务器ECS。
   - 方法：创建一个主机连接，将EL表达式直接通过echo打印出来，单击“测试运行”之后查看日志，日志中会打印出EL表达式的值。

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，单击“测试运行”后，再去OBS界面查看创建出来的目录名称。



9.14.2 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 9-185 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

操作符	描述
|| 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}
```

9.14.3 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9-186所示。

表 9-186 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

字母	描述	示例
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

说明

日期和时间模式一般在DateUtil内嵌对象和Job内嵌对象中使用，更多日期和时间模式的使用举例，请参见**DateUtil内嵌对象**和**Job内嵌对象**。

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")}
```

9.14.4 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 9-187 方法说明

方法	描述	示例
String get(String name)	获取指定名称环境变量值。	获取环境变量名称为**test**的参数值： #{Env.get("test")}

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")}
```

9.14.5 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 9-188 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 9-189 方法说明

方法	描述	示例
String getNodeStatus(String nodeName)	获取指定节点运行状态，成功状态返回success，失败状态返回fail。 例如，判断节点是否运行成功，可以使用如下判断条件，其中test为节点名称： #{(Job.getNodeStatus("test")) == "success" }	获取test节点运行状态。 #{Job.getNodeStatus("test")}

方法	描述	示例
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。	• 获取test节点输出。#{Job.getNodeOutput("test")} • 当前一节点执行无结果时，输出结果为“null”。 • 当前一节点的输出结果是一个字段时，输出结果形如[["000"]]所示。此时可通过EL表达式分割字符串结果，获取前一节点输出的字段值，但注意输出结果类型为String。需要输出原数据类型时，仍需通过For Each节点及其支持的Loop内嵌对象EL表达式获取。#{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("前一节点名"),"]")[0],"[")[0],"\\"")[0]} • 当前一节点的输出结果是多个（两个及以上）字段时，输出结果形如[["000"],["001"]]所示。此时需要结合For Each节点及其支持的Loop内嵌对象EL表达式如#{Loop.current[0]}，循环获取输出结果。
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。	获取参数test的值： #{Job.getParam("test")}
String getPlanTime(String pattern)	获取指定pattern的计划时间字符串，pattern为日期、时间模式，请参考**日期和时间模式**。	获取作业调度计划时间，具体到毫秒： #{Job.getPlanTime("yyyy-MM-dd HH:mm:ss:SSS")}

方法	描述	示例
String getYesterday(String pattern)	获取执行pattern的计划时间前一天的时间字符串，pattern为日期、时间模式，请参考**日期和时间模式**。	获取作业调度计划时间的前一天的时间，具体到日期： #{Job.getYesterday("yyyy-MM-dd HH:mm:ss:SSS")}
String getLastHour(String pattern)	获取执行pattern的计划时间前一小时的时间字符串，pattern为日期、时间模式，请参考**日期和时间模式**。	获取作业调度计划时间前一小时的时间，具体到小时： #{Job.getLastHour("yyyy-MM-dd HH:mm:ss:SSS")}
String getRunningData(String nodeName)	获取指定节点运行中记录的数据，当前只支持获取DLI SQL节点SQL语句运行的作业id。此方法只能获取前面依赖节点的输出。 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]")}。	获取指定DLI SQL节点test中第三条语句的job ID： #{JSONUtil.path(Job.getRunningData("test"),"jobIds[2]")}
String getInsertJobId(String nodeName)	返回指定DLI SQL或Transform Load节点第一个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。	获取DLI SQL节点test中第一个Insert SQL语句的job ID： #{Job.getInsertJobId("test")}
String getPreviousWorkday(Integer num, String pattern)	按照指定的pattern返回计划时间前第num个工作日的时间字符串，num只可为正整数。若没有获取到符合条件的结果则返回null值。 该EL表达式适用于按照日历选择自定义日期进行周期调度。	获取作业调度前五天的工作日的日期。 #{Job.getPreviousWorkday(5, "yyyyMMdd")}
String getPreviousNonWorkday(Integer num, String pattern)	按照指定的pattern返回计划时间前第num个非工作日的时间字符串，num只可为正整数。若没有获取到符合条件的结果则返回null值。 该EL表达式适用于按照日历选择自定义日期进行周期调度。	获取作业调度前一天的非工作日的日期。 #{Job.getPreviousNonWorkday(1, "yyyyMMdd")}

举例 1

获取作业中节点名称为**test**的输出结果，EL表达式如下：

```
#{Job.getNodeOutput("test")}
```

9.14.6 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**apache commons文档**。

举例 1

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,".")}
```

举例 2

假设变量b为字符串No,0020，返回“,”后面的子字符串，EL表达式如下：

```
#{StringUtil.split(b,',')[1]}
```

举例 3

当前一节点的输出结果是一个字段时，输出结果如[["000"]]所示。第二个节点引用第一个节点的输出，此时可通过EL表达式分割字符串结果，获取前一节点输出的字段值。

```
#{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("前一节点名"),"]")[0],"[")[0],"\\"")[0]}
```

举例 4

如果前一个SQL节点的输出结果为[["11"]]。若要获取该值"11"，EL表达式可以写成：

```
#{StringUtil.getDigits(Job.getNodeOutput("nodeName"))}
```

举例 5

提取字符串中的数字，拼接后返回。

String getDigits(String str)

举例：str为"1123~45"，则返回"112345"；str为"abc"，则返回""；str为"12345"，则返回"12345"。

9.14.7 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 9-190 方法说明

方法	描述	示例
String format(Date date, String pattern)	将Date类型时间按指定pattern格式为字符串。	将作业调度计划的时间，转换为毫秒格式。 #{DateUtil.format(Job.planTime,"yyyy-MM-dd HH:mm:ss:SSS")} 将作业调度计划减一天的时间，转换为周格式。 ● #{DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyyw")} Job.planTime为2024年1月7日时，返回值为20241。 ● #{DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyyww")} Job.planTime为2024年1月7日时，返回值为202401。
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。	将作业调度计划减一个月的时间，转换为月份格式。 #{DateUtil.format(DateUtil.addMonths(Job.planTime,-1),"yyyy-MM")}
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。	将作业调度计划减一天的时间，转换为年月日格式。 #{DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyy-MM-dd")} 将作业调度计划减一天的时间，转换为周格式。 ● #{DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyyw")} Job.planTime为2024年1月7日时，返回值为20241。 ● #{DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyyww")} Job.planTime为2024年1月7日时，返回值为202401。
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。	将作业调度计划减一小时的时间，转换为小时格式。 #{DateUtil.format(DateUtil.addHours(Job.planTime,-1),"yyyy-MM-dd HH")}

方法	描述	示例
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。	将作业调度计划减一分钟的时间，转换为分钟格式。 #{DateUtil.format(DateUtil.addMinutes(Job.planTime,-1),"yyyy-MM-dd HH:mm")}
int getDay(Date date)	从date获取天，例如：date为2018-09-14，则返回14。	从作业调度计划获取具体的天。 #{DateUtil.getDay(Job.planTime)}
int getMonth(Date date)	从date获取月，例如：date为2018-09-14，则返回9。	从日期获取具体的月。 #{DateUtil.getMonth(Job.planTime)}
int getQuarter(Date date)	从date获取季度，例如：date为2018-09-14，则返回3。	从日期获取具体的季度。 #{DateUtil.getQuarter(Job.planTime)}
int getYear(Date date)	从date获取年，例如：date为2018-09-14，则返回2018。	从日期获取具体的年。 #{DateUtil.getYear(Job.planTime)}
Date now()	返回当前时间。	以秒格式返回当前的时间。 #{DateUtil.format(DateUtil.now(),"yyyy-MM-dd HH:mm:ss")}
long getTime(Date date)	将Date类型时间转换为long类型时间戳。	将作业调度计划时间转换为时间戳。 #{DateUtil.getTime(Job.planTime)}
Date parseDate(String str, String pattern)	字符串按pattern转换为Date类型，pattern为日期、时间模式，请参考**日期和时间模式**。	将字符串类型的作业启动时间转换为秒格式。 #{DateUtil.parseDate(Job.getPlanTime("yyyy-MM-dd HH:mm:ss:SSS"),"yyyy-MM-dd HH:mm:ss")}

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")}
```

9.14.8 JSONUtil 内嵌对象

JSONUtil内嵌对象提供了JSON对象方法。

方法

表 9-191 方法说明

方法	描述	示例
Object parse(String jsonStr)	将json字符串转换为对象。	假设变量a为JSON字符串，将json字符串转换为对象，EL表达式如下： #{JSONUtil.parse(a)}
String toString(Object jsonObject)	将对象转换为json字符串。	假设变量b为对象，将对象转换为json字符串，EL表达式如下： #{JSONUtil.toString(b)}

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。	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")}

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")}
```

9.14.9 Loop 内嵌对象

使用Loop内嵌对象可获取For Each节点数据集中的数据。

属性

表 9-192 属性说明

属性	类型	描述	示例
dataArray	String	Loop.dataArray表示For Each节点“数据集”中定义的二维数组。 一般定义格式为#{Loop.dataArray[0][0]}、#{Loop.dataArray[0][1]}等类似样式。其中[0][0]表示数组中第一行的第一个值，[0][1]表示第一行的第二个值，以此类推。	作为For Each节点的“子作业参数”取值，表示For Each循环中，始终取“数据集”中二维数组的第二行的第一个值。 #{Loop.dataArray[1][0]}
current	String	For Each节点在处理数据集的时候，是一行一行进行处理的。Loop.current表示当前遍历到的For Each节点“数据集”中定义的二维数组的某一行，该数据行为一维数组。 一般定义格式为#{Loop.current[0]}、#{Loop.current[1]}或其他。其中[0]表示遍历到的当前行的第一个值，[1]表示遍历到的当前行的第二个值，以此类推。	作为For Each节点的“子作业参数”取值，表示For Each循环遍历中，取“数据集”中二维数组的当前遍历行的第二个值。 #{Loop.current[1]}
offset	Int	For循环当前的偏移量，从0开始。 Loop.dataArray[Loop.offset] = Loop.current。	获取For Each循环当前的偏移量，即遍历次数，从0开始。 #{Loop.offset}

举例

For Each节点的子作业参数，获取当前处理到的某行数据的第2个值，EL表达式如下：

```
#{Loop.current[1]}
```

9.14.10 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 9-193 方法说明

方法	说明	示例
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。	• 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")}

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")}

9.14.11 常用 EL 表达式样例合集

本章节介绍常用的EL表达式及示例。

表 9-194 常用的 EL 表达式

方法	描述	示例
String getNodeStatus(String nodeName)	获取指定节点运行状态，成功状态返回success，失败状态返回fail。 例如，判断节点是否运行成功，可以使用如下判断条件，其中test为节点名称： #{(Job.getNodeStatus("test")) == "success" }	获取test节点运行状态。 #{Job.getNodeStatus("test")}

方法	描述	示例
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。	• 获取test节点输出。#{Job.getNodeOutput("test")} • 当前一节点执行无结果时，输出结果为“null”。 • 当前一节点的输出结果是一个字段时，输出结果形如[["000"]]所示。此时可通过EL表达式分割字符串结果，获取前一节点输出的字段值，但注意输出结果类型为String。需要输出原数据类型时，仍需通过For Each节点及其支持的Loop内嵌对象EL表达式获取。#{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("前一节点名"),"]")[0],"[")[0],"\\"")[0]} • 当前一节点的输出结果是多个（两个及以上）字段时，输出结果形如[["000"],["001"]]所示。此时需要结合For Each节点及其支持的Loop内嵌对象EL表达式如#{Loop.current[0]}，循环获取输出结果。
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。	获取参数test的值： #{Job.getParam("test")}
String getPlanTime(String pattern)	获取指定pattern的计划时间字符串，pattern为日期、时间模式，请参考**日期和时间模式**。	获取作业调度计划时间，具体到毫秒： #{Job.getPlanTime("yyyy-MM-dd HH:mm:ss:SSS")}

方法	描述	示例
String getYesterday(String pattern)	获取执行pattern的计划时间前一天的时间字符串，pattern为日期、时间模式，请参考**日期和时间模式**。	获取作业调度计划时间的前一天的时间，具体到日期： #{Job.getYesterday("yyyy-MM-dd HH:mm:ss:SSS")}
String getLastHour(String pattern)	获取执行pattern的计划时间前一小时的时间字符串，pattern为日期、时间模式，请参考**日期和时间模式**。	获取作业调度计划时间前一小时的时间，具体到小时： #{Job.getLastHour("yyyy-MM-dd HH:mm:ss:SSS")}
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。	将作业调度计划减一天的时间，转换为年月日格式。 #{DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyy-MM-dd")}
int getDay(Date date)	从date获取天，例如：date为2018-09-14，则返回14。	从作业调度计划获取具体的天。 #{DateUtil.getDay(Job.planTime)}
Date now()	返回当前时间。	以秒格式返回当前的时间。 #{DateUtil.format(DateUtil.now(),"yyyy-MM-dd HH:mm:ss")}

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。	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")}

方法	描述	示例
current	For Each节点在处理数据集的时候，是一行一行进行处理的。Loop.current表示当前遍历到的For Each节点“数据集”中定义的二维数组的某一行，该数据行为一维数组。 一般定义格式为#{Loop.current[0]}、#{Loop.current[1]}或其他。其中[0]表示遍历到的当前行的第一个值，[1]表示遍历到的当前行的第二个值，以此类推。	作为For Each节点的“子作业参数”取值，表示For Each循环遍历中，取“数据集”中二维数组的当前遍历行的第二个值。 #{Loop.current[1]}

9.14.12 EL 表达式使用实例

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**配置DataArts Studio数据连接参数**进行操作。
- 已在DLI中创建数据库，以“dli_db”数据库为例。
  如未创建，请参考**新建数据库**进行操作。
- 已在“dli_db”数据库中创建数据表trade_log和trade_report。
  如未创建，请参考**新建数据表**进行操作。

操作步骤

步骤1 新建和开发SQL脚本。

1. 在数据开发模块控制台的左侧导航栏，进入“数据开发 > 脚本开发”，选择“新建DLI SQL脚本”。

2. 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”。

图 9-150 脚本属性

3. 在脚本编辑器中输入以下SQL语句。

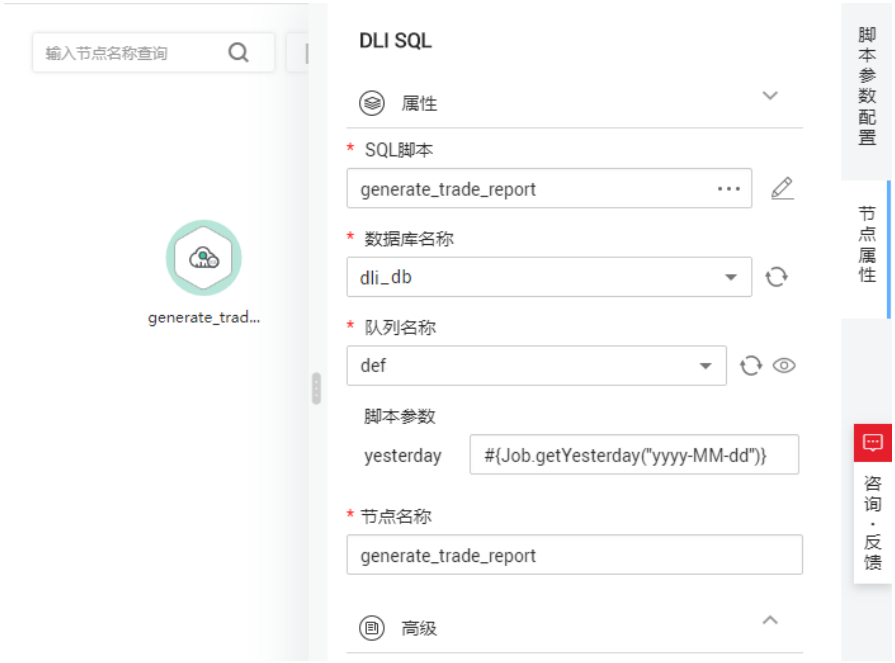
```
INSERT OVERWRITE TABLE trade_report
SELECT
  sum(trade_count),
  '${yesterday}'
FROM
  trade_log
where
  date_format(trade_time, 'yyyy-MM-dd') = '${yesterday}'
```

4. 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，进入“数据开发 > 作业开发”，选择“新建作业”，新建一个名称为“job”的空作业。
2. 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”。

图 9-151 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'
  ```

3. 单击  ，测试运行作业。
4. 作业测试无问题后，单击  ，保存作业配置。

----结束

更多案例

EL表达式在数据开发过程中被广泛应用，您可以参考**最佳实践**查看更多应用EL表达式的进阶实践。

9.15 简易变量集参考

简易变量集提供了一系列自定义的变量，自定义参数会根据任务调度的业务日期、计划时间及参数的取值格式自动替换为具体的值，实现在任务调度时间内参数的动态替换。

当前系统支持业务日期、计划时间、业务id三类参数的自定义。

- 业务日期是指在调度时间内，任务预期调度运行时间的前一天（即昨天），精确到天。业务日期可通过${yyyymmdd}获取。通常，业务日期为计划时间所在日期-1。
- 计划时间是指在调度时间内，任务预期调度运行的时间点（即当天），精确到秒。计划时间可通过$[yyyymmddhh24miss]获取。
- 业务ID参数包括作业ID和作业生成的实例ID两种，通过$job_id和$instance_id获取。

须知

使用简易变量集时，需要打开简易变量集的开关，功能才能生效。开关打开的方法请参见**配置默认项 > 是否使用简易变量集**。

业务日期参数

业务日期是指在调度时间内，任务预期调度运行时间的前一天（即昨天）。例如，调度日期为2023年1月1日，那么业务日期就是2022年12月31日。该参数是通过yyyy、yy、mm和dd自定义组合而生成的时间参数，其格式可自定义。例如，${yyyy}、${yyyymm}、${yyyymmdd}和${yyyy-mm-dd}等。

- yyyy：表示4位的年份，取值为业务日期的年份。
- yy：表示2位的年份，取值为业务日期的年份后两位。
- mm：表示月份，取值为业务日期的月份。
- dd：表示天，取值为业务日期的天。

取N年前、N月前、N天前的时间数据请参考**表9-195**，参数只能精确到年月日，不支持小时、分钟、秒的写法。

表 9-195 业务日期参数获取说明

业务日期场景	获取方法
前/后N年	${yyyy±N}
前/后N月	${yyyymm±N}
前/后N周	${yyyymmdd±7*N}
前/后N天	${yyyymmdd±N}
前/后N年（yy格式）	${yy±N}

计划时间参数

计划时间是指在调度时间内，任务预期调度运行的时间点（即当天）。该参数是通过yyyy、yy、mm、dd、hh24、mi和ss自定义组合而生成的时间参数，其格式可自定义。例如，$[yyyymmdd]、$[yyyy-mm-dd]、$[hh24miss]、$[hh24:mi:ss]和$[yyyymmddhh24miss]等。

- yyyy：表示4位的年份，取值为计划时间的年份。
- yy：表示2位的年份，取值为计划时间的年份后两位。
- mm：表示月份，取值为计划时间的月份。
- dd：表示天，取值为计划时间的天。
- hh：表示12小时制，取值为计划时间的小时。
- hh24：表示24小时制，取值为计划时间的小时。
- mi：表示分钟，取值为计划时间的分钟。
- ss：表示秒，取值为计划时间的秒。

取N小时前、N分钟前的时间数据请参考**表9-196**，该参数不支持通过$[yyyy-N]、$[mm-N]等直接获取多少年前、多少月前的时间数据。

表 9-196 计划时间参数获取说明

计划时间场景	获取方法
后N年	$[add_months(yyyymmdd,12*N)]
前N年	$[add_months(yyyymmdd,-12*N)]
后N月	$[add_months(yyyymmdd,N)]
前N月	$[add_months(yyyymmdd,-N)]
前/后N周	$[yyyymmdd±7*N]
前/后N天	$[yyyymmdd±N]
前/后N小时	获取该时间数据包含如下两种方式： ● $[hh24miss±N/24] ● $[自定义时间格式±N/24]。 例如，取前一个小时的不同时间格式： – 取月：$[mm-1/24]。 – 取年：$[yyyy-1/24]。 – 取年月：$[yyyymm-1/24]。 – 取年月日：$[yyyymmdd-1/24]。 – 取前一天且前一小时：$[yyyymmdd-1-1/24]
前/后N分钟	获取该时间数据包含如下四种方式： ● $[hh24miss±N/24/60] ● $[yyyymmddhh24miss±N/24/60] ● $[mi±N/24/60] ● $[自定义时间格式±N/24/60] 例如，取计划时间15分钟前的不同时间格式： – 取年：$[yyyy-15/24/60] – 取年月：$[yyyymm-15/24/60] – 取年月日：$[yyyymmdd-15/24/60] – 取小时：$[hh24-15/24/60] – 取分钟：$[mi-15/24/60]

说明

- 调度参数替换值在实例生成时已经确定，所以调度参数的替换值不会随着实例实际运行时间的改变而改变。
- 当调度参数取小时、分钟时，参数替换值由实例的计划时间决定，即由节点调度配置的计划调度时间决定。举例如下：
  - 如果当前节点为日调度节点，并且设置计划调度时间为01：00，则小时的参数取值为01。
  - 如果当前节点为小时调度节点，并且设置计划调度时间为00：00~23：59，每小时调度一次，则：第一个小时实例计划时间为0点，小时的参数取值为00，第二个小时实例计划时间为1点，小时的参数取值为01，以此类推。

业务 ID 参数

业务ID会替换成当前业务的实际ID，包括作业ID和作业生成的实例ID。

表 9-197 业务 ID 参数获取说明

方法	说明
$job_id	数据开发作业id。获取该ID请参考**查询作业详情**。
$instance_id	作业实例id（单节点作业测试运行不生成实例id，不支持）。获取该ID请参考**查询作业实例列表**。

9.16 使用教程

9.16.1 脚本及作业中引用参数使用介绍

该章节介绍如何在脚本及作业中引用参数，以及引用后的生效范围、是否支持EL表达式和简易变量集等，让您更加清晰地了解工作空间级和脚本、作业级配置参数的使用方法。

说明

工作空间环境变量参数、作业参数、脚本参数均可以配置参数，但作用范围不同；另外如果工作空间环境变量参数、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 > 工作空间环境变量参数 > 脚本参数**。

表 9-198 参数的使用方法

类别	场景	生效范围	调用方法
环境变量/环境常量	配置作业参数时，当某参数隶属于多个作业时，可将此参数提取出来作为环境变量。	当前工作空间	${*环境变量*} ${*环境常量*} 配置方法请参考：**环境变量**

类别	场景	生效范围	调用方法
作业变量/作业常量	作业参数为作业级的参数，可用于作业中的任意节点。	当前作业	${*作业变量*} ${*作业常量*} 配置方法请参考：**配置作业参数**
脚本参数	配置自定义字段的参数名称和参数值。	当前脚本	${*脚本参数*} 配置方法请参考：**脚本参数**

说明

SQL脚本的变量格式有${}和${dlf.}两种，支持用户根据实际情况进行配置。配置的变量格式会作用于SQL脚本、作业中SQL语句、单节点作业，环境变量。配置脚本变量格式的操作请参见**脚本变量定义**。

系统默认脚本变量格式为${}。

环境变量

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 9-152 环境变量

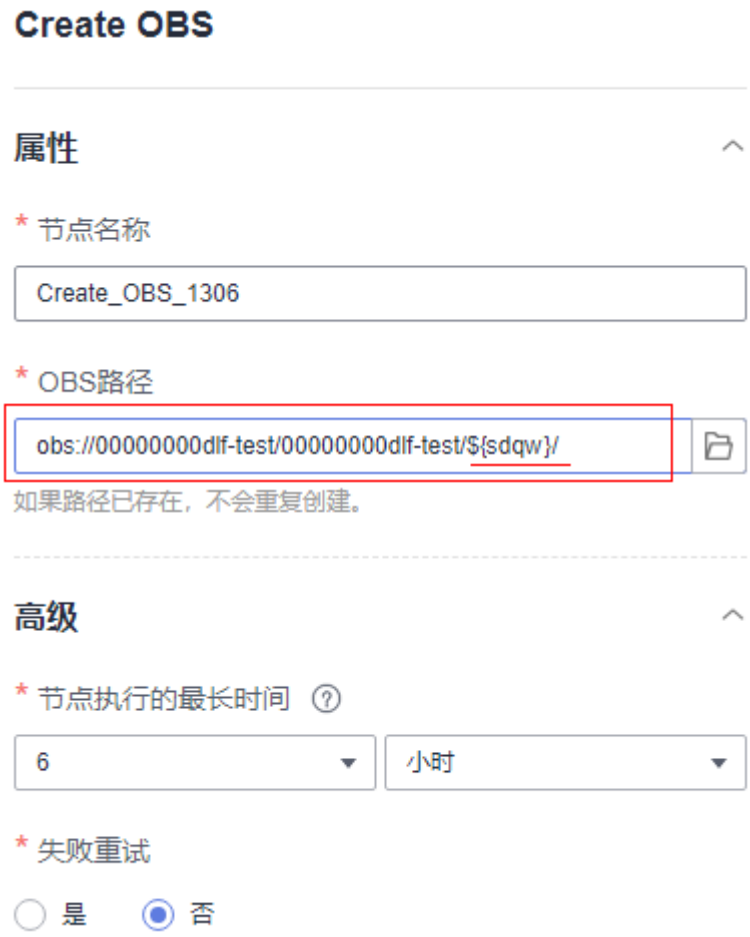
具体应用如下：

在环境变量中已新增一个变量，“参数名”为sdqw，“参数值”为wqewqewqe。

步骤1 打开一个已创建好的作业，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 9-153 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

配置作业参数

作业参数中支持定义变量和常量，作业参数的作用范围为当前作业。

- 变量是指不同的作业下取值不同，需要重新配置值。
- 常量是指在不同的作业下都是一样的，不需要重新配置值。

图 9-154 作业参数

作业参数定义好之后，可以在作业节点里面引用该参数。

图 9-155 引用作业参数

脚本参数

- 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），通过作业调度时可通过节点属性进行赋值，如2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脚本支持参数传递功能。
- 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9-156所示。

图 9-156 独立执行时的脚本参数

1. 开发一个脚本。开发脚本时，脚本表达式里面必须包含变量（例如，SQL中变量是DATE，脚本中就写${DATE}）。在作业参数配置里面，您可以在2中编写脚本参数DATE的语句表达式。

   在“脚本开发”界面，在编辑器中输入开发语句，如下图所示。

   ```
   INSERT INTO B FROM (SELECT * FROM A WHERE DATE = ${DATE})
   ```

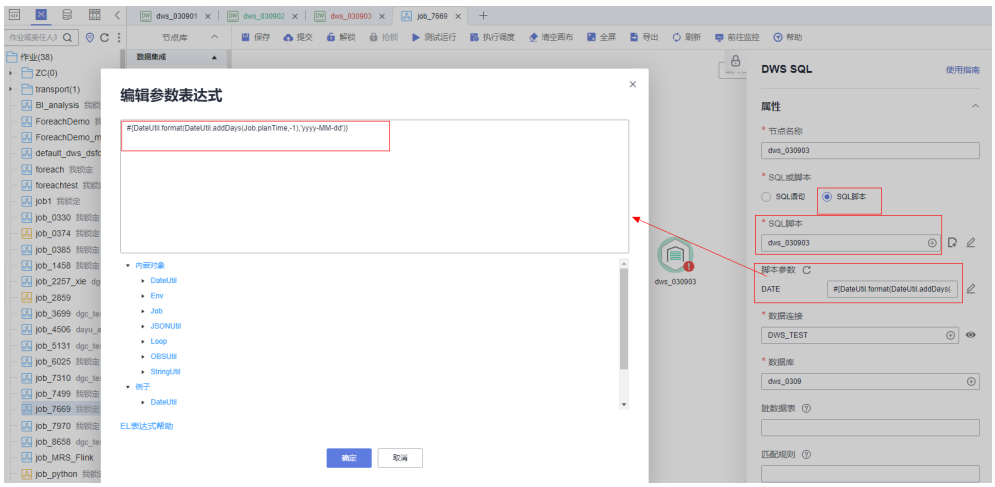
图 9-157 开发脚本

脚本dws_030903编写完成后，保存并提交此脚本的最新版本。

2. 开发一个批处理作业。开发作业时，您需要配置节点属性参数。

   在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。

图 9-158 作业调度时的脚本参数



说明

- 如果作业所关联的SQL脚本如果使用了参数，此处显示脚本参数名称（例如DATA），请在参数名称后的输入框配置参数值。参数值支持使用**EL表达式**。
- 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步，也可以单击进行编辑。
- 涉及脚本的节点均可以采用此方式引用脚本变量，如SQL脚本、Shell脚本和Python脚本。

简易变量集

简易变量集提供了一系列自定义的变量，自定义参数会根据任务调度的业务日期、计划时间及参数的取值格式自动替换为具体的值，实现在任务调度时间内参数的动态替换。简易变量集的详细内容请参见**简易变量集参考**。

9.16.2 作业调度支持每月最后一天

场景描述

在配置作业调度时，可以选择每个月的最后一天执行。如果您需要配置作业的调度时间为每月最后一天，请参考下面两种方法。

表 9-199 配置每月最后一天进行调度

配置方法	优势	如何配置
调度周期配置为天，通过条件表达式进行判断是否为每月最后一天	可以灵活适用多种场景。只需要编写条件表达式就可以灵活调度作业去运行。例如，每月最后一天，每月七号等。	**方法1**
调度周期配置为月，勾选每月最后一天	通过配置调度周期来执行任务调度。不用编写开发语句，通过勾选需要调度的时间去执行任务。	**方法2**

方法 1

在DataArts Studio中配置一个每天调度执行的作业，然后在作业里面新增一个Dummy节点（空节点，不处理实际的业务数据），在Dummy节点与后续执行任务的节点的连线上，您可以配置条件表达式，判断当前是否为每个月的最后一天。如果是最后一天，则执行后续节点，否则跳过后续节点。

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. 任务配置为天调度，如下图：

图 9-159 调度周期配置为天

3. 在节点的连线上，单击右键，选择设置条件，配置条件表达式。通过表达式来判断，是否执行后续的业务节点。

图 9-160 设置条件表达式

4. 表达式配置方法如下所示。

```
#{DateUtil.getDay(DateUtil.addDays(Job.planTime,1)) == 1 ? "true" : "false"}
```

表达式的含义是：获取当前的时间点，往后推一天，判断是不是1号，如果是，则表明当前是每个月的最后一天，执行后续节点。如果不是，则跳过后续的业务节点。

图 9-161 条件表达式

如果用户的作业是每个月的最后一天执行，可以按照上面的方法进行配置。

如果用户的作业是每月7号执行，可以按照下面的方法进行配置。

判断是否为7号，表达式配置方法如下所示。

```
#{DateUtil.getDay(Job.planTime) == 7 ? "true" : "false"}
```

方法 2

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. 单击作业画布右侧“调度配置”页签，进入调度配置页面。
3. 调度方式选择“周期调度”，调度周期选择“月”，选择时间为“每月最后一天”，如下图所示。

图 9-162 调度时间为每月最后一天

调度时间配置好之后，在每个月的最后一天，所配置的作业会按照调度时间去自动运行。

9.16.3 配置作业调度为年调度

场景描述

在配置作业配置调度时，可以选择一年中的某个时间进行调度。如果您需要配置作业的调度时间为年调度，请参考下面的方法进行配置。

配置方法

在DataArts Studio中配置一个按月调度执行的作业，然后在作业里面新增一个Dummy节点（空节点，不处理实际的业务数据），在Dummy节点与后续执行任务的节点的连线上，您可以配置条件表达式，判断当前的调度时间是否为一年中的指定的某一天进行调度（比如2023年6月29号）。如果是，则执行后续节点，否则跳过后续节点。

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. 任务配置为月调度，如下图：

图 9-163 调度周期配置为月

3. 在节点的连线上，单击右键，选择设置条件，配置条件表达式。通过表达式来判断，是否执行后续的业务节点。

图 9-164 设置条件表达式

4. 表达式配置方法如下所示。

```
#{DateUtil.getMonth(Job.planTime) == 6 ? "true" : "false"}
```

表达式的含义是：获取当前的时间点，判断是不是6月，如果是，则表明当前是6月，执行后续节点。如果不是，则跳过后续的业务节点。

图 9-165 条件表达式

9.16.4 补数据场景使用介绍

适用场景

在某项目搬迁场景下，当您需要补充以前时间段内的历史业务数据，需要查看历史数据的详细信息时，可以使用补数据特性。

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。

说明

- 补数据作业除了支持SQL脚本，其他节点也支持。
- 如果SQL脚本的内容有变化，补数据作业运行的是最新版本的脚本。
- 使用补数据功能时，如SQL中变量是DATE，脚本中就写${DATE}，在作业参数中会自动增加脚本参数DATE，脚本参数DATE的值支持使用EL表达式。如果是变量时间的话，需要使用DateUtil内嵌对象的表达式，平台会自动转换成历史日期。EL表达式用法可参考**EL表达式**。
- 补数据作业除了支持作业参数，脚本参数或者全局环境变量也支持。

约束条件

只有数据开发作业配置了周期调度，才支持使用补数据功能。

使用案例

案例场景

在某企业的产品数据表中，有一个记录产品销售额的源数据表A，现在需要把产品销售额的历史数据导入的目的表B里面，需要您配置补数据作业的相关操作。

需要导入的列表情况如表1所示。

表 9-200 需要导入的列表情况

源数据表名	目的表名
A	B

配置方法

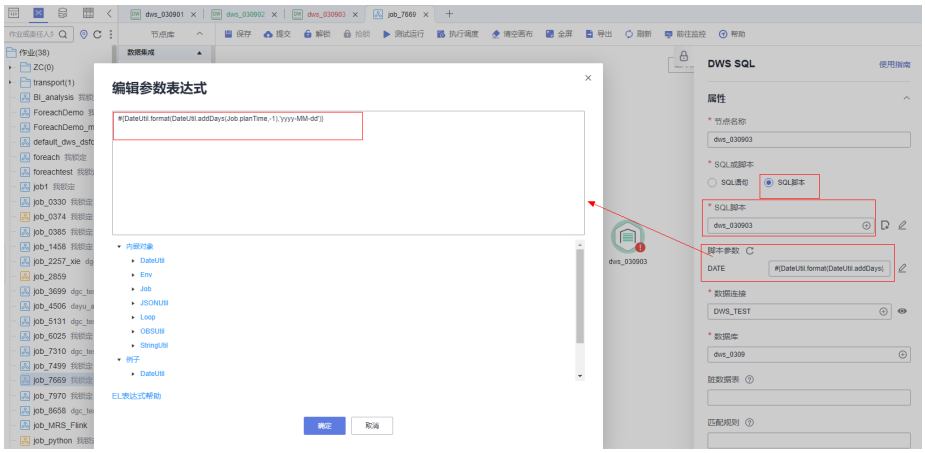
1. 准备源表和目的表。为了便于后续作业运行验证，需要先创建DWS源数据表和目的表，并给源数据表插入数据。
   a. 创建DWS表。您可以在DataArts Studio数据开发中，新建DWS SQL脚本执行以下SQL命令：
   ```
   /* 创建数据表 */
   CREATE TABLE A (PRODUCT_ID INT, SALES INT, DATE DATE);
   CREATE TABLE B (PRODUCT_ID INT, SALES INT, DATE DATE);
   ```
   b. 给源数据表插入示例数据。您可以在DataArts Studio数据开发模块中，新建DWS SQL脚本执行以下SQL命令：
   ```
   /* 源数据表插入示例历史数据 */
   INSERT INTO A VALUES ('1','60', '2022-03-01');
   INSERT INTO A VALUES ('2','80', '2022-03-01');
   INSERT INTO A VALUES ('1','50', '2022-02-28');
   INSERT INTO A VALUES ('2','55', '2022-02-28');
   INSERT INTO A VALUES ('1','60', '2022-02-27');
   INSERT INTO A VALUES ('2','45', '2022-02-27');
   ```
2. 开发一个补数据的脚本。开发脚本时，脚本表达式里面必须包含时间变量（例如，SQL中变量是DATE，脚本中就写${DATE}）。在作业参数配置里面，您可以在3中编写脚本参数DATE的语句表达式。
   在“脚本开发”界面，在编辑器中输入开发语句。
   ```
   INSERT INTO B (SELECT * FROM A WHERE DATE = ${DATE})
   ```

   图 9-166 开发脚本

   ```
   -- DWS sql
   -- ********************************************************************** --
   -- author: [redacted]
   -- create time: 2023/05/23 17:03:02 GMT+08:00
   -- ********************************************************************** --
   INSERT INTO B (SELECT * FROM A WHERE DATE = ${DATE})
   ```

   脚本编写完成后，保存并提交此脚本的最新版本。
3. 开发一个补数据的批处理作业。开发作业时，您需要配置节点属性参数和调度周期。
   在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。

图 9-167 节点参数



说明

- 如果作业所关联的SQL脚本使用了参数，此处显示脚本参数名称（例如DATE），请在参数名称后的输入框配置参数值。参数值支持使用EL表达式，EL表达式用法可参考**EL表达式**。

  如果参数是时间的话，请您查看下DateUtil内嵌对象的表达式例子，平台会自动替换成补数据的历史日期（由补数据的业务日期所决定）。

  您也可以直接编写SQL语句，编写SQL表达式。

- 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步，也可以单击进行编辑。
- 脚本参数的举例如下所示。

  例如：#{DateUtil.format(DateUtil.addDays(Job.planTime,-1),'yyyy-MM-dd')}

  - Job.planTime表示作业计划时间，yyyy-MM-dd表示时间格式。
  - 如果作业计划时间三月二号，减去一天就是三月一号。补数据时，配置的补数据的业务日期就会替换作业计划时间。
  - Job.planTime会把作业计划时间通过表达式转化为yyyy-MM-dd格式的时间。

配置补数据作业的调度周期。单击界面右侧的调度配置，配置补数据作业的调度周期，该使用指导配置周期设置为天。

图 9-168 配置调度周期

说明

- 作业调度周期设置为天，每天会进行作业调度，并生成一个调度实例。您可以在“实例监控”页面中，查看补数据实例的运行状态。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。
- 该作业调度时间从2023/03/09开始生效，每天2点调度一次作业。
- 执行以下SQL命令，查询目的表B里面是否存在源表A的数据。

  SELECT * FROM B

参数配置完成后，保存并提交此作业的最新版本，测试运行该作业。

单击“执行调度”，让该作业运行起来。

4. 创建补数据。

   您在创建了一个周期调度作业后，用户需要为该任务进行补数据的操作。

   a. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。

   b. 单击“批作业监控”页签，进入批作业的监控页面。在该作业的“操作”列，选择“更多 > 补数据”。进入“补数据”页面。

      如果您需要补充**2023-02-27至2023-03-01**之间的历史数据，补数据的业务日期需要设置为**2023-02-28至2023-03-02**，该业务日期系统会自动传给作业计划时间，脚本时间变量DATE的表达式中，定义的时间为作业计划时间减去一天，即作业计划时间的前一天时间为补数据的时间范围（**2023-02-27至2023-03-01**）。

      图 9-169 补数据

表 9-201 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	系统自动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。这个业务日期会传递给作业的计划时间。作业运行时，作业计划时间就会被补数据里面的业务时间替换掉。 **说明** 一个作业可进行多次补数据。但多次补数据的业务日期需要避免交叉重叠，否则可能导致数据重复或混乱，用户请谨慎操作。 如果勾选了“按日期倒序补数据”，则系统按照日期倒序补跑，每日内的补数顺序仍是正序。 **说明** • 该功能适合在各日数据不耦合的条件下使用。 • 为保证补数可以倒序进行，补数据作业对更早日期作业实例的依赖关系将被忽略。
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。
需要补数据的上下游作业	可选。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。

c. 单击“确定”，系统会根据作业的调度周期开始补数据。

d. 在“补数据监控”页面中，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，同时您可以查看补数据的详细日志信息。

图 9-170 补数据详细信息

e. 执行以下SQL命令，查询目的表B里面是否存在源表A的历史数据。

```
SELECT * FROM B
```


9.16.5 获取 SQL 节点的输出结果值

当您在数据开发模块进行作业开发，需要获取SQL节点的输出结果值，并将结果应用于后续作业节点或判断时，可参考本教程获取SQL节点的输出结果。

场景说明

使用EL表达式**#{Job.getNodeOutput("前一节点名")}**获取的前一节点的输出结果时，输出结果为二维数组形式 ，形如[["Dean",...,"08"],...,["Smith",...,"53"]]所示。为获取其中的值，本案例提供了如**表9-202**所示的两个常见方法示例。

表 9-202 获取结果值常见方法

方法	关键配置	适用场景要求
通过StringUtil提取输出结果值	当SQL节点的输出结果只有一个字段，形如[["11"]]所示时，可以通过StringUtil内嵌对象EL表达式分割二维数组，获取前一节点输出的字段值： #{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("***前一节点名***"),"]")[0],"[")[0],"\\"")[0]}	通过StringUtil提取输出结果值配置简单，但对适用场景有如下要求： • 前一SQL节点的输出结果只有一个字段，形如[["11"]]所示。 • 输出结果值数据类型为String，需要应用场景支持String数据类型。例如当需要使用IF条件判断输出结果值的数值大小时，不支持String类型，则不能使用本方法。
通过For Each节点提取输出结果值	通过For Each节点，循环获取数据集中二维数组的值： • For Each节点数据集：#{Job.getNodeOutput('***前一节点名***')} • For Each节点子作业参数：#{Loop.current[***索引***]}	通过For Each节点输出结果值适用场景更广泛，但需将作业拆分为主作业和子作业。

通过 StringUtil 提取输出结果值

场景说明

通过StringUtil内嵌对象EL表达式分割二维数组结果，获取前一节点输出的字段值，输出结果类型为String。

本例中，MRS Hive SQL节点返回单字段二维数组，Kafka Client节点发送的数据定义为StringUtil内嵌对象EL表达式，通过此表达式即可分割二维数组，获取MRS Hive SQL节点输出的字段值。

说明

为便于查看最终获得的结果值，本例选择Kafka Client节点进行演示。在实际使用中，您可以根据您的业务需求选择后续节点类型，在节点任务中应用StringUtil内嵌对象EL表达式，即可获取前一节点返回的数据值。

图 9-171 作业样例

其中，Kafka Client节点的关键配置为“发送数据”参数，取值如下：

```
#{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("count95"),"]")[0],"[")[0],"\\"")[0]}
```

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 构造原始表格student_score。新建临时Hive SQL脚本，选择Hive连接和数据库后，粘贴如下SQL语句并运行，运行成功后即可删除此脚本。

```
CREATE TABLE `student_score` (`name` String COMMENT '', `score` INT COMMENT '');
INSERT INTO
    student_score
VALUES
    ('ZHAO', '90'),
    ('QIAN', '88'),
    ('SUN', '93'),
    ('LI', '94'),
    ('ZHOU', '85'),
    ('WU', '79'),
    ('ZHENG', '87'),
    ('WANG', '97'),
    ('FENG', '83'),
    ('CEHN', '99');
```

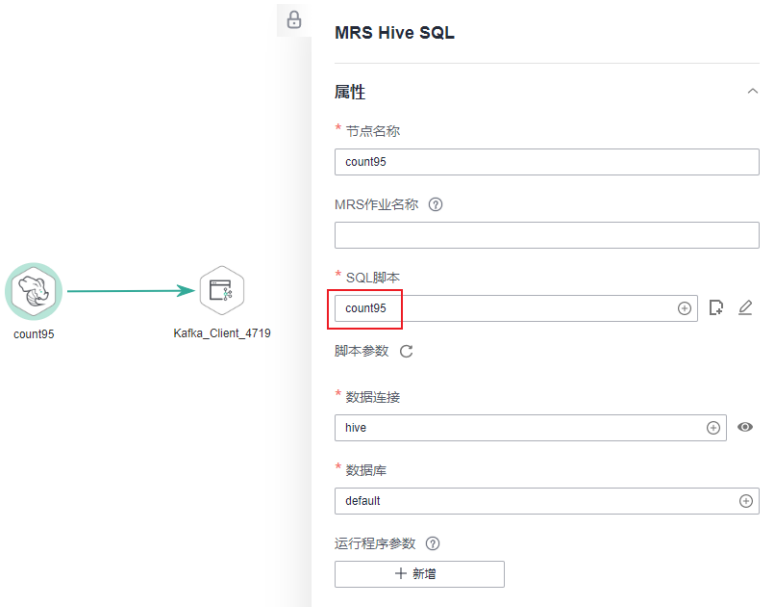
步骤4 新建MRS Hive SQL节点调用的Hive SQL脚本。新建Hive SQL脚本，选择Hive连接和数据库后，粘贴如下SQL语句并提交版本，脚本命名为count95。

```
--从student_score表中统计成绩在95分以上的人数--
SELECT count(*) FROM student_score WHERE score> "95" ;
```

步骤5 在“作业开发”页面，新建数据开发作业。选择一个MRS Hive SQL节点和一个Kafka Client节点，选中连线图标并拖动，编排如图9-171所示的作业。

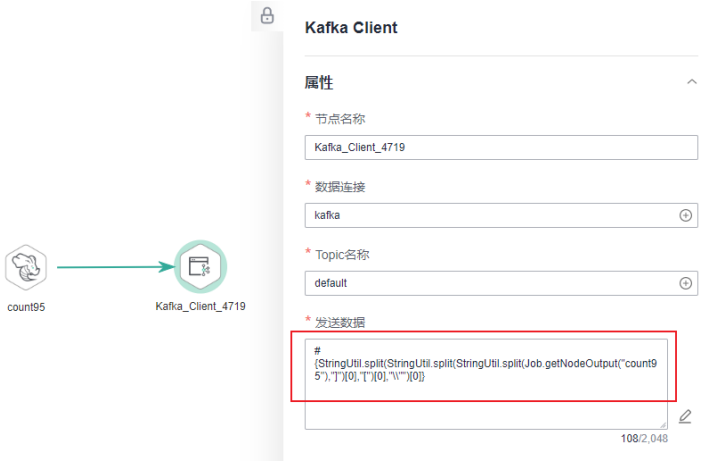
步骤6 配置MRS Hive SQL节点参数。SQL脚本选择步骤4中提交的脚本count95，选择Hive连接和数据库。

图 9-172 配置 MRS Hive SQL 节点参数



步骤7 配置Kafka Client节点参数。发送数据定义为：#{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("count95"),"]")[0],"[")[0],"\\"")[0]}，选择Kafka连接和Topic名称。

图 9-173 配置 Kafka Client 节点参数



步骤8 作业节点配置完成后，选择测试运行。待作业测试运行成功后，在Kafka Client节点上右键查看日志，可以发现MRS Hive SQL节点返回的二维数组**[["2"]]**已被清洗为**2**。

说明

您可以将Kafka Client节点中的发送数据定义为**#{Job.getNodeOutput("count95")}**，然后作业运行后查看Kafka Client节点日志，则可以验证MRS Hive SQL节点返回的结果为二维数组**[["2"]]**。

图 9-174 查看 Kafka Client 节点日志

----结束

通过 For Each 节点提取输出结果值

场景说明

结合For Each节点及其支持的Loop内嵌对象EL表达式#{Loop.current[0]}，循环获取前一节点输出的结果值。

本例中，MRS Hive SQL节点返回多字段的二维数组，选择For Each节点和EL表达式**#{Loop.current[]}**，再通过For Each循环调用Kafka Client节点子作业，Kafka Client节点发送的数据也定义为#{Loop.current[]}，通过此配置即可获取MRS Hive SQL节点输出的结果值。

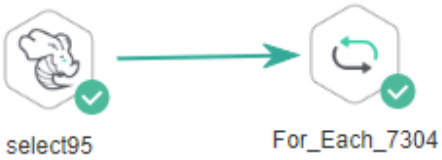
说明

为便于查看最终获得的结果值，本例中For Each节点子作业选择Kafka Client节点进行演示。在实际使用中，您可以根据您的业务需求选择子作业节点类型，在节点任务中应用Loop内嵌对象EL表达式，即可获取For Each前一节点返回的结果值。

For Each节点主作业编排如图9-175所示。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("select95")}**，其中**select95**为前一个节点的名称。
- 子作业参数：子作业参数是子作业中定义的参数名，然后在主作业中定义的参数值，传递到子作业以供使用。此处子作业参数名定义为**name**和**score**，其值为分别为数据集中的第一列和第二列数值，使用EL表达式**#{Loop.current[0]}**和**#{Loop.current[1]}**。

图 9-175 主作业样例



而For Each节点中所选的子作业，则需要定义For Each节点中的子作业参数名，以便让主作业识别参数定义，作业如图9-176所示。

图 9-176 子作业样例

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**EL_test_slave**。选择一个Kafka Client节点，并配置作业参数，编排图9-176所示的作业。

此处需将参数名填写为**name**和**score**，仅用于主作业的For Each节点识别子作业参数；参数值无需填写。

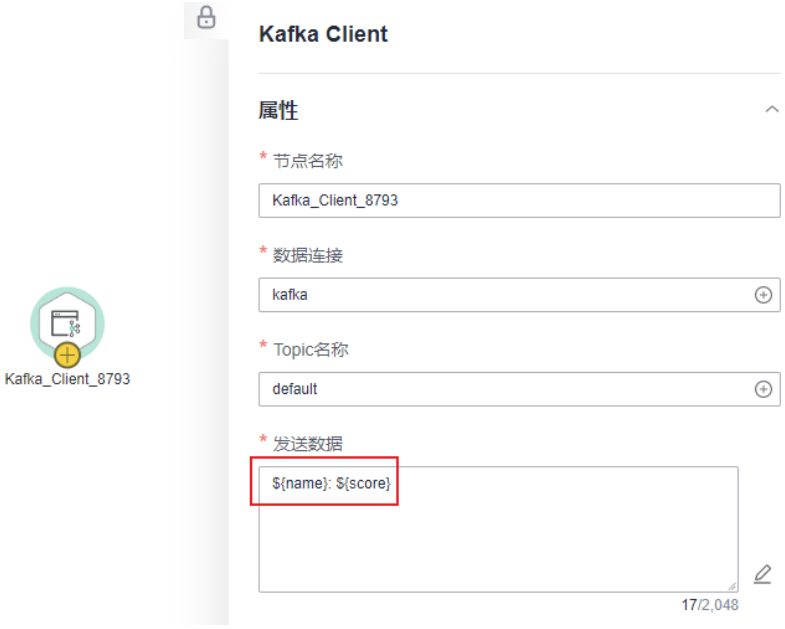
步骤4 配置Kafka Client节点参数。发送数据定义为：**${name}: ${score}**，选择Kafka连接和Topic名称。

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 9-177 配置 Kafka Client 节点参数



步骤5 配置完成后提交子作业。

----结束

开发主作业

步骤1 在“作业开发”主页面，进入脚本开发。

步骤2 构造原始表格student_score。新建临时Hive SQL脚本，选择Hive连接和数据库后，粘贴如下SQL语句并运行，运行成功后即可删除此脚本。

```
CREATE TABLE `student_score` (`name` String COMMENT '', `score` INT COMMENT '');
INSERT INTO
   student_score
VALUES
   ('ZHAO', '90'),
   ('QIAN', '88'),
   ('SUN', '93'),
   ('LI', '94'),
   ('ZHOU', '85'),
   ('WU', '79'),
   ('ZHENG', '87'),
   ('WANG', '97'),
   ('FENG', '83'),
   ('CEHN', '99');
```

步骤3 新建MRS Hive SQL节点调用的Hive SQL脚本。新建Hive SQL脚本，选择Hive连接和数据库后，粘贴如下SQL语句并提交版本，脚本命名为select95。

```
--从student_score表中展示成绩在95分以上的姓名和成绩--
SELECT * FROM student_score WHERE score> "95" ;
```

步骤4 在“作业开发”页面，新建数据开发主作业**EL_test_master**。选择一个HIVE SQL节点和一个For Each节点，选中连线图标并拖动，编排图9-175所示的作业。

步骤5 配置MRS Hive SQL节点参数。SQL脚本选择步骤3中提交的脚本select95，选择Hive连接和数据库。

图 9-178 配置 MRS Hive SQL 节点参数

步骤6 配置For Each节点属性，如图9-179所示。

- 子作业：子作业选择已经开发完成的子作业**EL_test_slave**。
- 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("select95")}**，其中**select95**为前一个节点的名称。
- 子作业参数：子作业参数是子作业中定义的参数名，然后在主作业中定义的参数值，传递到子作业以供使用。此处子作业参数名定义为**name**和**score**，其值为分别为数据集中的第一列和第二列数值，使用EL表达式**#{Loop.current[0]}**和**#{Loop.current[1]}**。

图 9-179 配置 For Each 节点参数

步骤7 配置完成后保存作业。

----结束

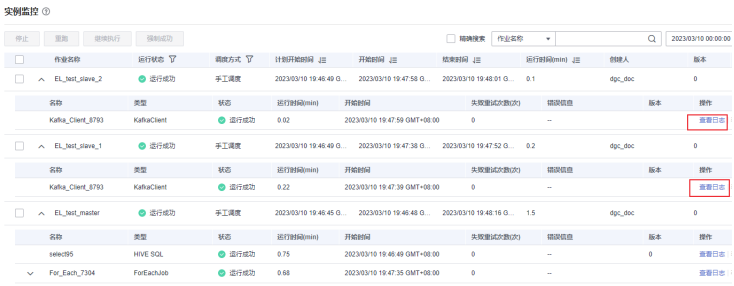
测试运行主作业

步骤1 单击主作业**EL_test_master**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节点循环调用运行子作业**EL_test_slave**。

步骤2 单击左侧导航栏中的“实例监控”，进入实例监控中查看作业运行结果。

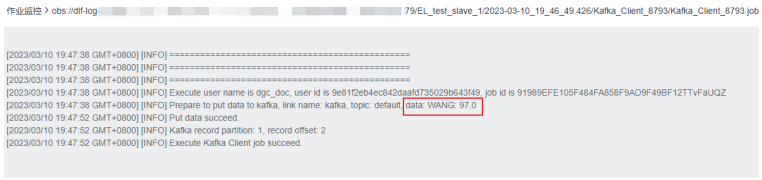
步骤3 待作业运行完成后，从实例监控中找到子作业**EL_test_slave**的循环运行结果，如**图9-180**所示。

图 9-180 子作业运行结果



步骤4 查看子作业**EL_test_slave**在循环运行中的结果日志，从日志中可以看到，结合For Each节点及其支持的Loop内嵌对象EL表达式，成功获取For Each前一节点输出的结果值。

图 9-181 查看日志



----结束

9.16.6 查询 SQL 获取 max 值传递给 CDM 作业

场景描述

通过查询SQL语句，将获取到的最大时间的max值传递给CDM作业。在CDM作业的高级属性里面，通过where子句判断最大时间范围，获取所需要的迁移数据，从而完成数据迁移任务，最终完成增量迁移任务。

约束条件

1. 已完成**新建数据连接**的操作。
2. 已完成**新建数据库**的操作。

使用案例

创建SQL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. 创建一个SQL脚本。本案例以MRS SPARK SQL为例。
3. 选择已创建好的数据连接和数据库。
4. 编写SQL脚本，从源表table1这张数据表里面获取最大时间值数据。
   ```
   select max(time) from table1
   ```
5. 保存并提交版本。脚本**maxtime**创建完成。

创建一个Pipeline子作业

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. 选择CDM Job节点，并配置节点属性参数。

图 9-182 配置 CDM Job 节点属性参数

选择CDM集群名称、关联已存在的CDM作业。

配置该作业的参数，引入作业参数名称maxtime，如下图所示。

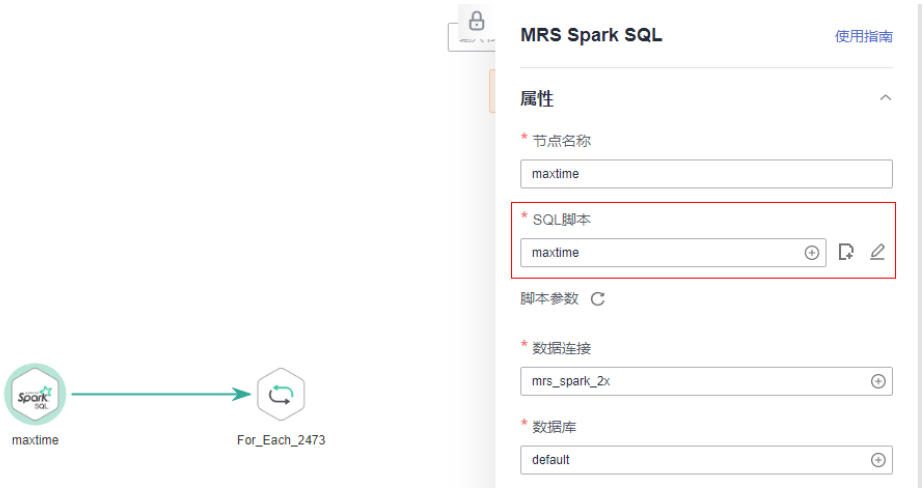
图 9-183 配置作业参数

3. 保存并提交版本。子作业**sub**创建完成。

创建一个Pipeline作业

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. 选择MRS Spark SQL节点和For Each循环执行的节点，让CDM子作业循环执行，并配置节点属性参数。
3. 配置MRS Spark SQL节点的属性参数，并关联已创建的脚本maxtime。

图 9-184 配置 MRS Spark SQL 节点属性参数



4. 配置For Each节点的属性参数，并关联已创建的CDM子作业。

图 9-185 配置 For Each 节点参数

关联已创建的子作业sub后，编写参数表达式。

```
#{Loop.current[0]}
```

配置数据集，支持EL表达式。

```
#{Job.getNodeOutput("maxtime")}
```

5. 保存并提交版本。作业创建完成。

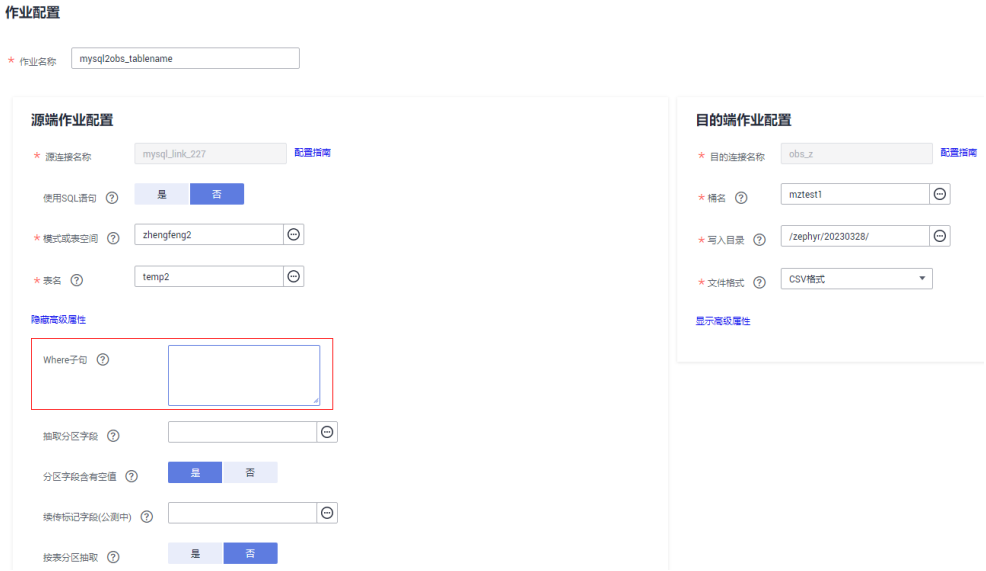
在CDM作业中通过where子句配置获取最大时间值数据并传递给目的端作业

1. 打开已创建的子作业。
2. 单击CDM作业名称后面的跳转到CDM作业配置界面。

图 9-186 编辑 CDM 作业

3. 在源端作业配置的高级属性里面，通过配置where子句获取迁移所需的数据，作业运行时，将从源端获取到的迁移数据复制导出并导入目的端。

图 9-187 配置 where 子句



where子句配置如下：

```
dt > '${maxtime}'
```

9.16.7 IF 条件判断教程

当您在数据开发模块进行作业开发编排时，想要实现通过设置条件，选择不同的执行路径，可使用IF条件判断。

本教程包含以下三个常见场景举例。

- **根据前一个节点的执行状态进行IF条件判断**
- **根据前一个节点的输出结果进行IF条件判断**
- **多IF条件下当前节点的执行策略**

IF条件的数据来源于EL表达式，通过EL表达式，根据具体的场景选择不同的EL表达式来达到目的。您可以参考本教程，根据您的实际业务需要，开发您自己的作业。

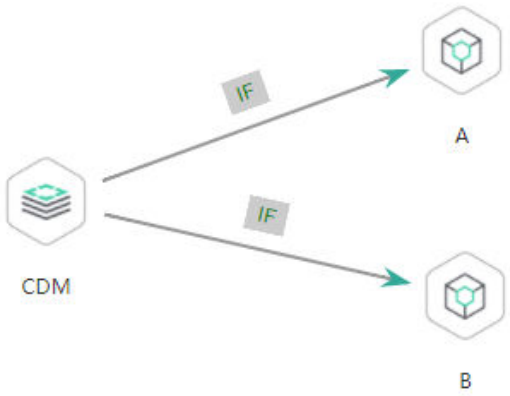
EL表达式用法可参考**EL表达式**。

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9-188**的样例，说明如何设置IF条件。

图 9-188 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9-188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”。

图 9-189 配置 CDM 节点的失败策略

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。

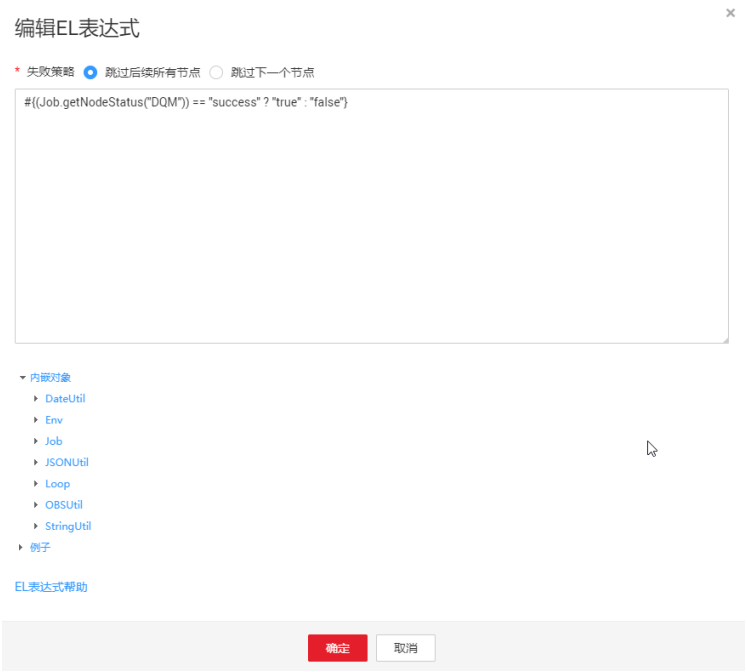
每一个条件分支都需要填写IF条件，IF条件为通过EL表达式语法填写三元表达式。当三元表达式结果为true的时候，才会执行连线后面的节点，否则后续节点将被跳过。

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。配置完成后单击确定，保存作业。

图 9-190 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9-191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 9-191 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：通过HIVE SQL统计成绩在85分以上的人数，并将执行结果作为参数传递到下一个节点，通过与人数通过标准进行数值比较，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语句执行结果为单字段的二维数组，因此为获取二维数组中的值，EL表达式**#{Loop.dataArray[][]}**或**#{Loop.current[]}**均可以实现，且当前只有For Each节点支持Loop表达式，所以HIVE SQL节点后面需要连接一个For Each节点。

说明

此场景下不能使用StringUtil表达式#{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("前一节点名"),"]")[0],"[")[0],"\\"")[0]}替代Loop表达式，因为StringUtil表达式最终获取的数据类型为String，无法与标准数据Int比较大小。

作业编排如图9-192所示：

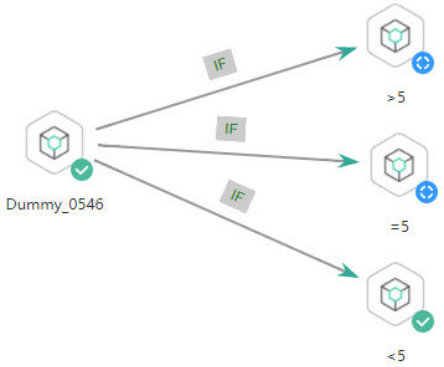
图 9-192 主作业样例

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。
- 子作业参数：子作业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]}**或**#{Loop.current[]}**，本例以**{Loop.dataArray[0][0]}**为例进行说明。

而For Each节点中所选的子作业，需要根据For Each节点传过来的子作业参数，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9-193所示。

图 9-193 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For Each。选择四个Dummy节点，选中连线图标并拖动，编排图9-193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。

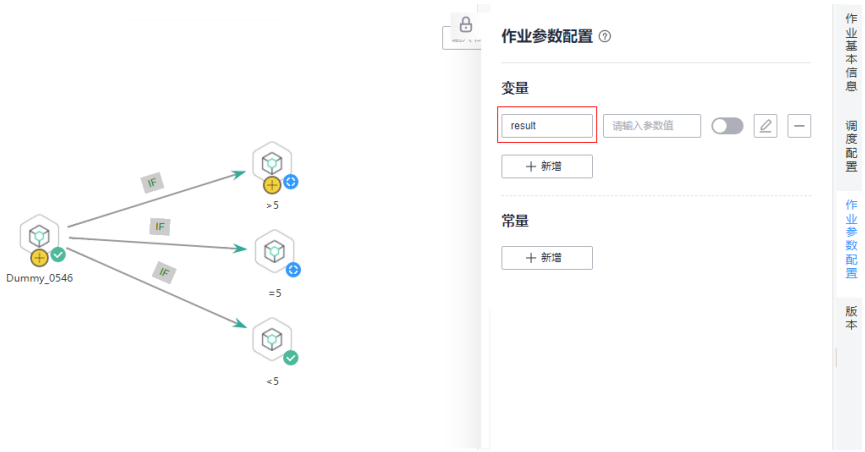
说明

表达式中包含多条件的场景下，可以通过“||”联合多个条件。例如：

#{(${result} >= 19 || ${result} <=9) ? "true" : "false"}

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 9-194 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9-192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。

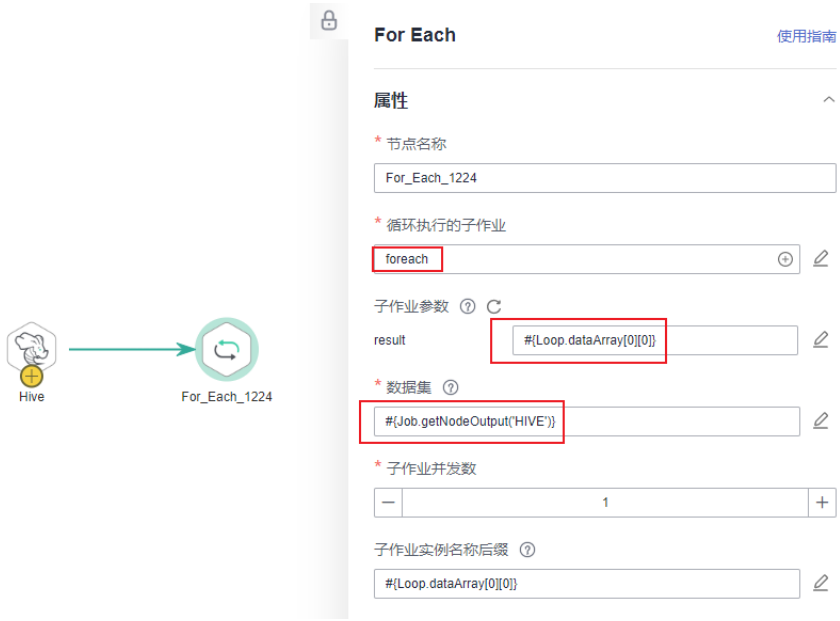
```
--从student_score表中统计成绩在85分以上的人数--
SELECT count(*) FROM student_score WHERE score> "85" ;
```

图 9-195 HIVE SQL 脚本执行结果

步骤3 配置For Each节点属性，如图9-196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。
- 子作业参数：子作业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 9-196 For Each 节点属性



步骤4 配置完成后保存作业。

----结束

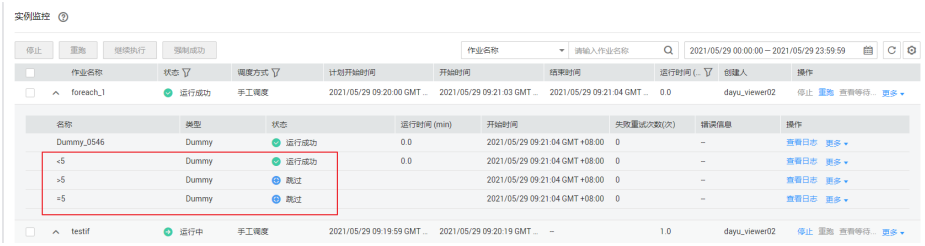
测试运行主作业

步骤1 单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 单击左侧导航栏中的“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9-197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4，所以>5和=5的分支被跳过，执行<5这个分支成功。

图 9-197 子作业运行结果



----结束

多 IF 条件下当前节点的执行策略

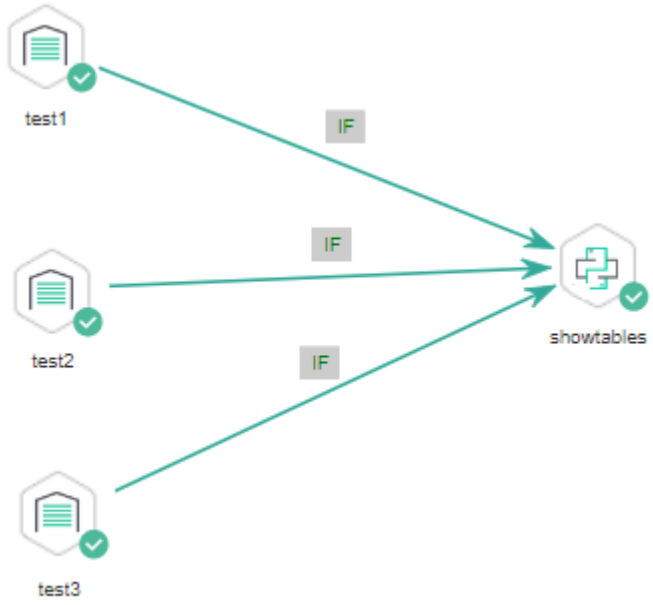
如果当前节点的执行依赖多个IF条件的节点，执行的策略包含逻辑或和逻辑与两种。

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 9-198 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9-198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。

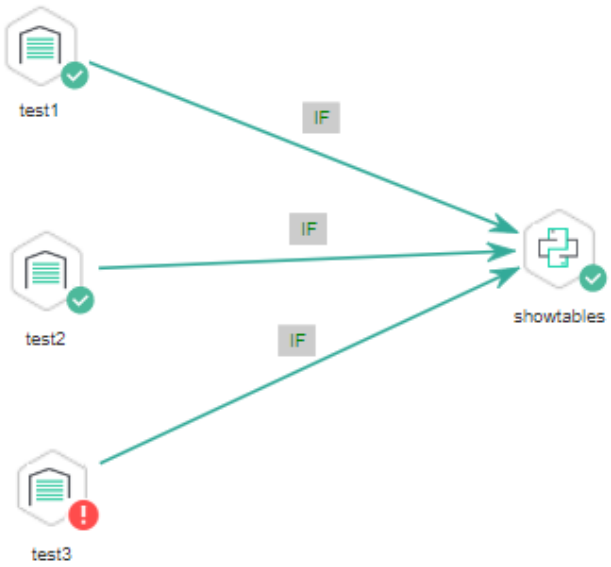
test1运行成功，则对应的IF条件为true；

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 9-199 配置为“逻辑或”的作业运行情况

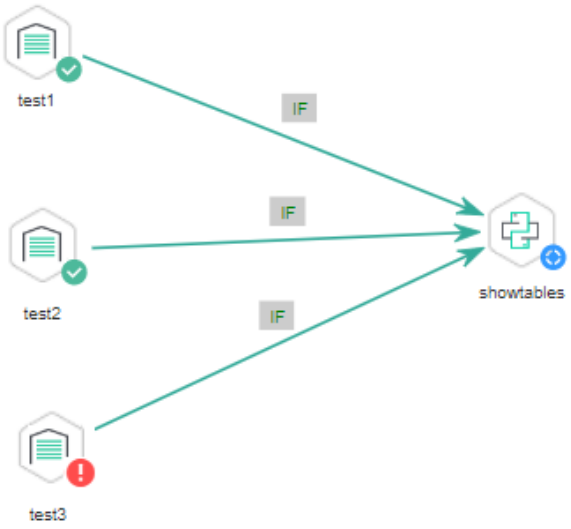


测试运行日志

[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 9-200 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

9.16.8 获取 Rest Client 节点返回值教程

Rest Client节点可以执行华为云内的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9-201所示，第一个Rest Client调用了MRS服务查询集群列表的API，图9-202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。

  说明

  响应消息体解析为参数传递定义时，传递的参数名（例如clusterId）在该作业的所有节点参数中需要保持唯一性，避免和其他参数同名。

图 9-201 Rest Clien 作业样例 1

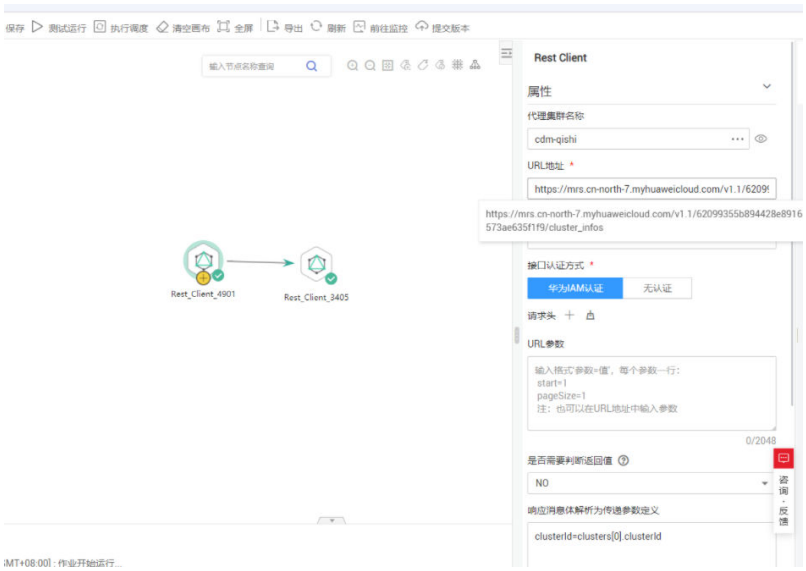


图 9-202 JSON 消息体


```
{
    "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":"cerxth-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 data sets, including core components such as HDFS, YARN, and MapReduce.",
                    "componentDescEn":null,
                    "multi_service_name":null
                },
```

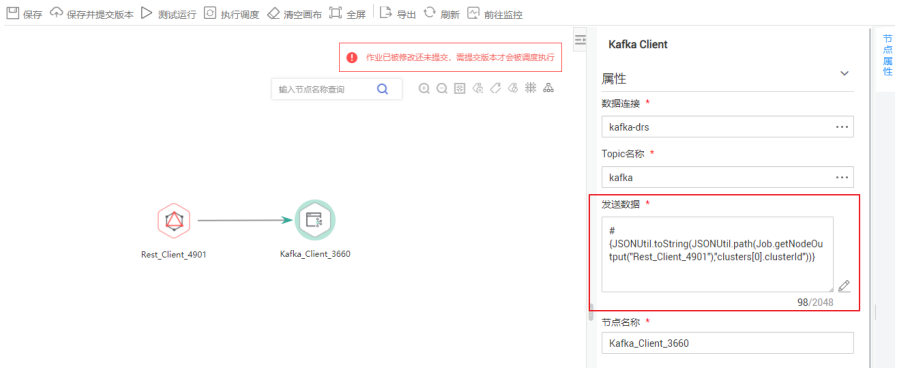

通过 EL 表达式获取返回值

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**EL表达式**。

如**图9-203**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 9-203 Rest Client 作业样例 2



9.16.9 For Each 节点使用介绍

适用场景

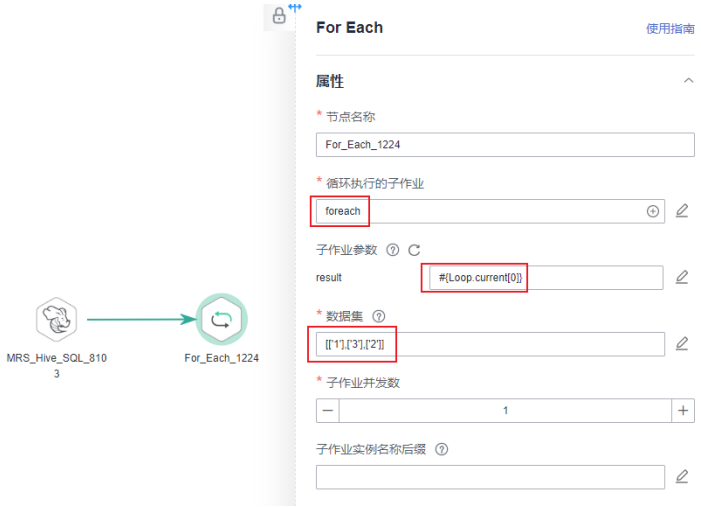
当您进行作业开发时，如果某些任务的参数有差异、但处理逻辑全部一致，在这种情况下您可以通过For Each节点避免重复开发作业。

For Each节点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。
- 子作业参数：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]}，即将数据集中每行数据的第一个数值遍历传递给子作业。

For Each节点举例如**图9-204**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 9-204 for each 节点



For Each 节点与 EL 表达式

要想使用好For Each节点，您必须对EL表达式有所了解。EL表达式用法请参考**EL表达式**。

下面为您展示For Each节点常用的一些EL表达式。

- #{Loop.dataArray}：For循环节点输入的数据集，是一个二维数组。
- #{Loop.current}：由于For循环节点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]}或其他，0表示遍历到当前行的第一个值。
- #{Loop.offset}：For循环节点在处理数据集时当前的偏移量，从0开始。
- #{Job.getNodeOutput('preNodeName')}：获取前面节点的输出。

使用案例

案例场景

因数据规整要求，需要周期性地将多组DLI源数据表数据导入到对应的DLI目的表，如**表1**所示。

表 9-203 需要导入的列表情况

源数据表名	目的表名
a_new	a
b_2	b
c_3	c
d_1	d
c_5	e
b_1	f

如果通过SQL节点分别执行导入脚本，需要开发大量脚本和节点，导致重复性工作。在这种情况下，我们可以使用For Each节点进行循环作业，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：

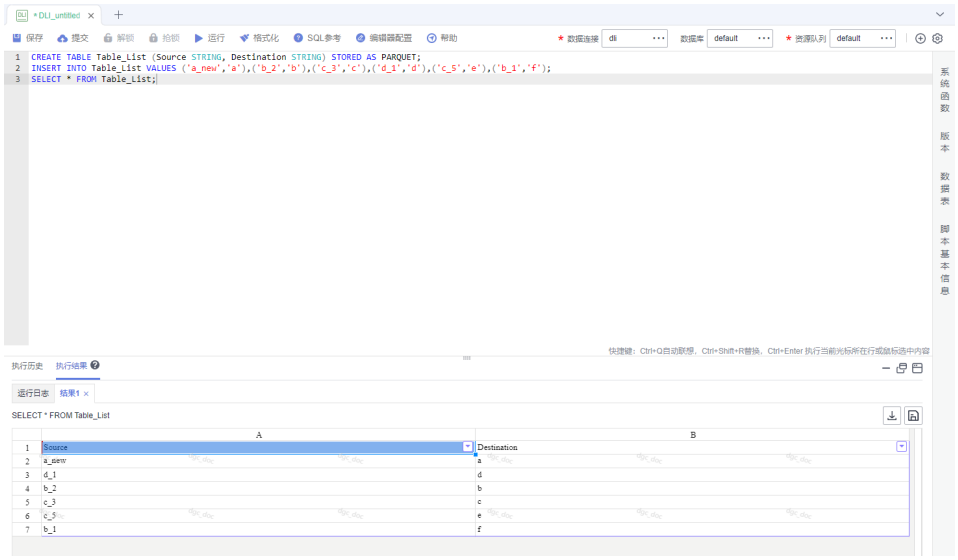
1. 您可以将**表1**数据导入到DLI表中，然后将SQL脚本读取的结果作为数据集。
2. 您可以将**表1**数据保存在OBS的CSV文件中，然后通过DLI SQL或DWS SQL创建OBS外表关联这个CSV文件，然后将OBS外表查询的结果作为数据集。DLI创建外表请参见**OBS输入流**，DWS创建外表请参见**创建外表**。
3. 您可以将**表1**数据保存在HDFS的CSV文件中，然后通过HIVE SQL创建Hive外表关联这个CSV文件，然后将HIVE外表查询的结果作为数据集。MRS创建外表请参见**创建表**。

本例以方式1进行说明，将**表1**中的数据导入到DLI表（Table_List）中。您可以在DataArts Studio数据开发模块中，新建DLI SQL脚本执行以下SQL命令导入数据，也可以在数据湖探索（DLI）服务控制台中的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 9-205 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9-206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：

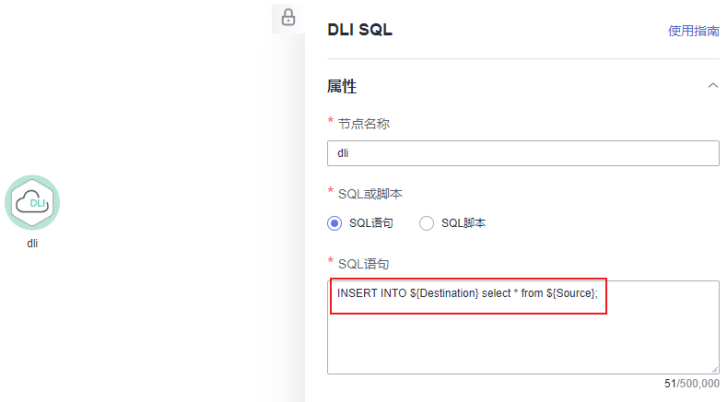
   ```
   INSERT INTO ${Destination} select * from ${Source};
   ```

   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 9-206 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 9-207 配置子作业参数

3. 配置完成后保存作业。

步骤4 创建For Each节点所在的主作业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9-208所示的作业。

图 9-208 编排作业

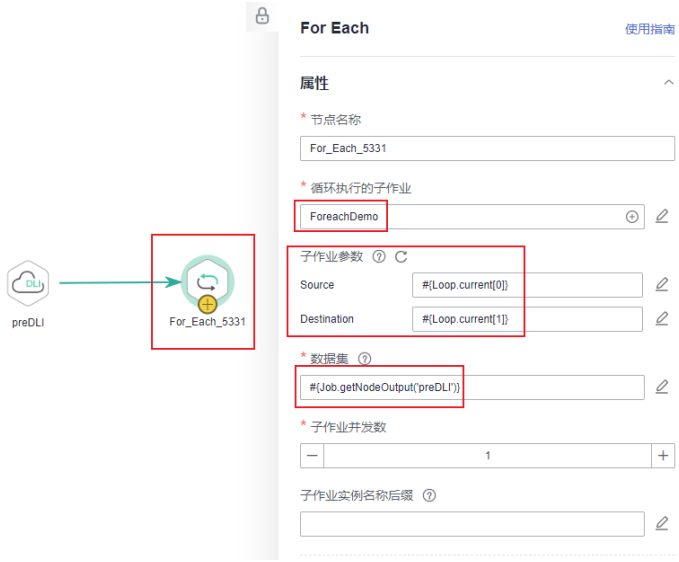
2. 配置DLI SQL节点属性，此处配置为SQL语句，语句内容如下所示。DLI SQL节点负责读取DLI表Table_List中的内容作为数据集。

```
SELECT * FROM Table_List;
```

图 9-209 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。
   - 子作业参数：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 9-210 配置 For Each 节点

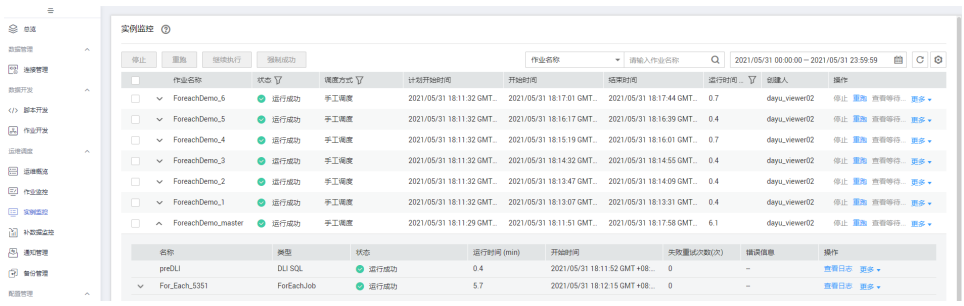


4. 配置完成后保存作业。

步骤5 测试运行主作业。

1. 单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. 单击左侧导航栏中的“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 9-211 查看作业实例

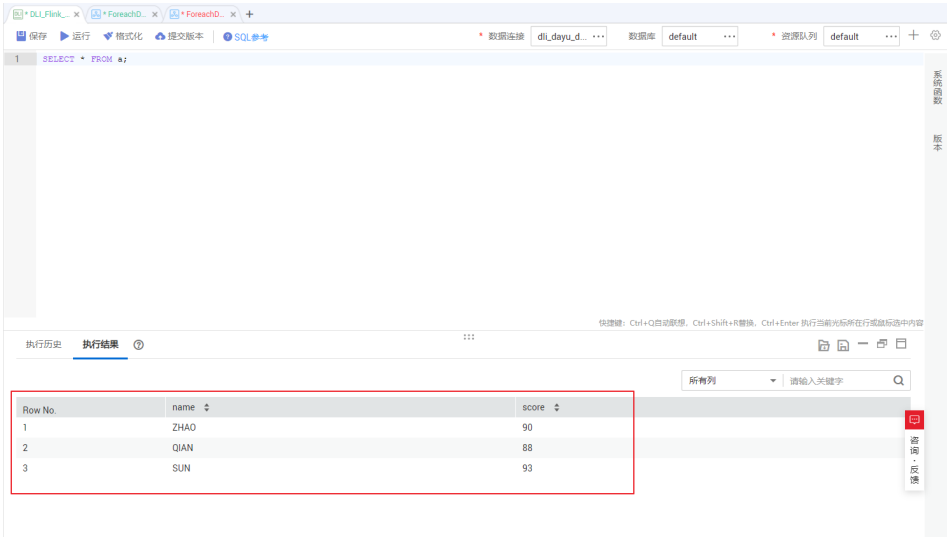


3. 查看对应的6个DLI目的表中是否已被插入预期的数据。您可以在DataArts Studio数据开发模块中，新建DLI SQL脚本执行以下SQL命令导入数据，也可以在数据湖探索（DLI）服务控制台中的SQL编辑器中执行以下SQL命令：

```
/* 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 9-212 目的表数据



----结束

更多案例参考

For Each节点可与其他节点配合，实现更丰富的功能。您可以参考以下案例，了解For Each节点的更多用法。

- **通过CDM节点批量创建分表迁移作业**
- **根据前一个节点的输出结果进行IF条件判断**

9.16.10 引用脚本模板和参数模板的使用介绍

使用场景

该功能适用于以下场景：

- Flink SQL脚本可以引用脚本模板。
- 在pipeline作业开发中，MRS Flink Job节点可以使用引入了脚本模板的Flink SQL脚本，同时在MRS Flink Job节点的“运行程序参数”里面可以引用参数模板。
- 在Flink SQL单任务作业中引用脚本模板。
- 在Flink Jar单任务作业中使用参数模板。

说明

在脚本中引用脚本模板时，SQL语句的写法为**@@{脚本模板}**。

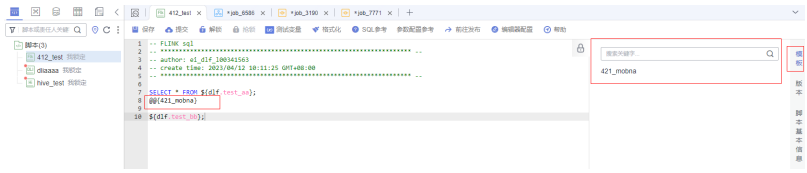
前提条件

已创建模板。如果模板还未创建，请参见**配置模板**进行创建。

引用模板案例

- Flink SQL脚本可以引用脚本模板。
  a. 在数据开发主界面的左侧导航栏，选择“数据开发 > 脚本开发”。
  b. 右键单击脚本，选择“新建Flink SQL脚本”进入。
  c. 单击右侧的“模板”，选择刚才创建好的脚本模板，例如412_mobna，系统支持可以引用多个模板。

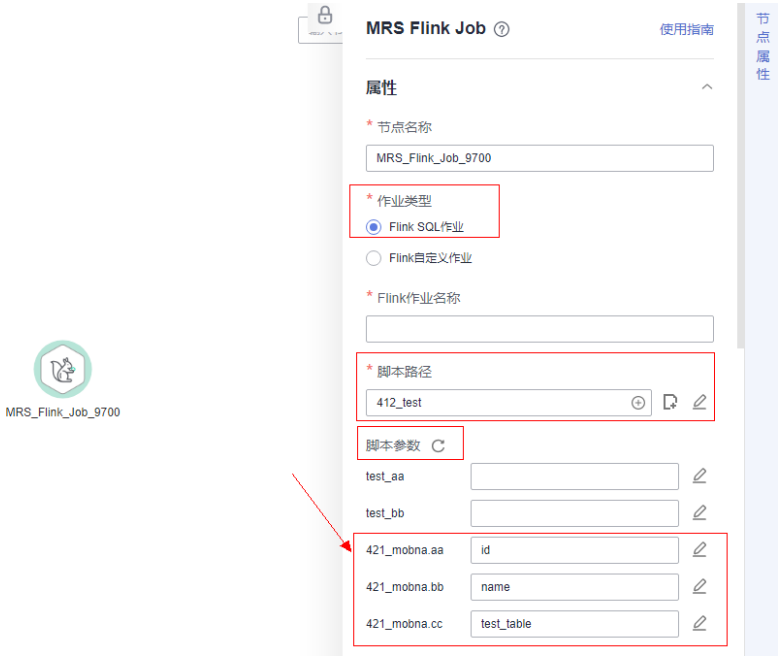
  图 9-213 引用脚本模板

  d. 脚本创建完成后，单击“保存”，脚本412_test创建完成。
- 在pipeline作业开发中，MRS Flink Job节点可以使用引入了脚本模板的Flink SQL脚本。
  a. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
  b. 右键单击作业，创建一个pipeline模式的批处理作业，进入作业开发界面。
  c. 选择“MRS_Flink_Job”节点。
  d. “作业类型”选择“Flink SQL作业”，“脚本路径”选择刚创建的Flink SQL脚本。

  选择脚本后，脚本里面引用的脚本模板参数及参数值会自动展示出来，如下图所示。

  图 9-214 引用 fink sql 脚本

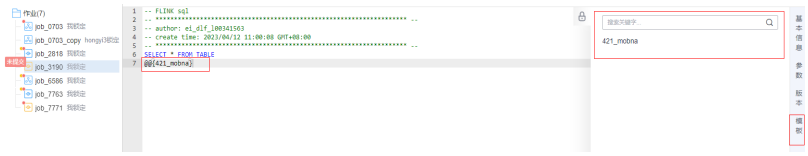
  

- 在pipeline作业开发中，在MRS Flink Job节点的“运行程序参数”里面引用参数模板。
  a. 选择MRS集群名。
  b. 运行程序参数会自动展示出来。单击“选择模板”进入后，选择已创建的参数模板，系统支持可以引用多个模板。
     参数名称及参数值会自动展示出来，如下图所示。

     图 9-215 运行程序参数引用参数模板

- 在Flink SQL单任务作业中引用脚本模板。
  a. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
  b. 右键单击作业，创建一个单任务模式的实时处理作业Flink SQL，进入作业开发界面。
  c. 单击右侧的“模板”，选择刚才创建好的脚本模板，例如412_mobna，系统支持可以引用多个模板。

     图 9-216 单任务 Flink sql 引用脚本模板



- 在Flink Jar单任务作业中使用参数模板。
  a. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
  b. 右键单击作业，创建一个单任务模式的实时处理作业Flink Jar，进入作业开发界面。
  c. 选择MRS集群名。
  d. 运行程序参数会自动展示出来。单击“选择模板”进入后，选择已创建的参数模板，系统支持可以引用多个模板。
     参数名称及参数值会自动展示出来，如下图所示。

图 9-217 单任务 Flink Jar 引用参数模板

9.16.11 开发一个 Python 作业

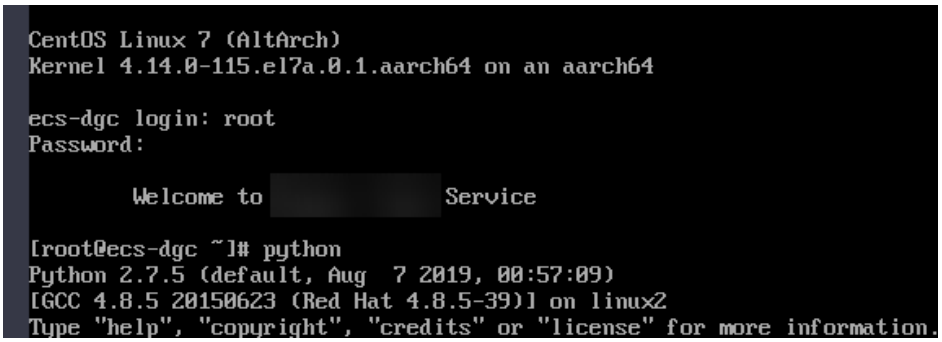
本章节介绍如何在数据开发模块上开发并执行Python作业示例。

环境准备

- 已开通弹性云服务器，并创建ECS，ECS主机名为“ecs-dgc”。

  说明

  本示例主机选择“CentOS 8.0 64bit with ARM(40GB)”的公共镜像，并且使用ECS自带的Python环境，您可登录主机后使用**python**命令确认服务器的Python环境。

- 已开通数据集成增量包，CDM集群名为“cdm-dlfpyhthon”，提供数据开发模块与ECS主机通信的代理。
- 请确保ECS主机与CDM集群网络互通，互通需满足如下条件：
  - CDM集群与ECS主机同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**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节点支持脚本参数和作业参数。
- 本示例以Python3为例。

建立主机数据连接

开发Python脚本前，我们需要建立一个到弹性云服务器ECS的连接。

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“管理中心”模块，进入管理中心页面。

步骤3 在管理中心页面，单击“数据连接”，进入数据连接页面并单击“创建数据连接”。

步骤4 参见**表9-204**配置相关参数，创建主机连接名称为“ecs”的数据连接，如**图9-218**所示。

表 9-204 主机连接

参数	是否必选	说明
数据连接类型	是	主机连接固定选择为主机连接。
数据连接名称	是	数据连接的名称，只能包含字母、数字、下划线和中划线，且长度不超过100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。
适用组件	是	选择此连接适用的组件。勾选组件后，才能在相应组件内使用本连接。
基础与网络连通配置		
主机地址	是	Linux操作系统主机的IP地址。 请参考**查看云服务器详细信息**获取。
绑定Agent	是	选择CDM集群，CDM集群提供Agent。如果没有可用的CDM集群，请参考**创建CDM集群**进行创建。 **说明** • CDM集群作为管理中心数据连接Agent时，单集群的并发活动线程最大为200。即当多个数据连接共用同一Agent时，通过这些数据连接提交SQL脚本、Shell脚本、Python脚本等任务的同时运行上限为200，超出的任务将排队等待。建议您按照业务量情况规划多个Agent分担压力。 • 在调度Shell、Python脚本时，Agent会访问ECS主机，如果Shell、Python脚本的调度频率很高，ECS主机会将Agent的内网IP加入黑名单。为了保障作业的正常调度，强烈建议您使用ECS主机的**root**用户将绑定Agent（即CDM集群）的内网IP加到**/etc/hosts.allow**文件里面。 CDM集群的内网IP获取方式请参见**查看并修改CDM集群配置**。

参数	是否必选	说明
端口	是	主机的SSH端口号。 Linux操作系统主机的默认登录端口为22，如有修改可通过主机路径“/etc/ssh/sshd_config”文件中的port字段确认端口号。
KMS密钥	是	通过KMS加解密数据源认证信息，选择KMS中的任一默认密钥或自定义密钥即可。 **说明** 第一次通过DataArts Studio或KPS使用KMS加密时，会自动生成默认密钥dlf/default或kps/default。关于默认密钥的更多信息，请参见**什么是默认密钥**。
数据源认证及其他功能配置		
用户名	是	主机的登录用户名。
登录方式	是	选择主机的登录方式： ● 密钥对 ● 密码
密钥对	是	“登录方式”为“密钥对”时，显示该配置项。 主机的登录方式为密钥对时，您需要获取并上传其私钥文件至OBS，在此处选择对应的OBS路径（OBS路径中不能存在中文字符）。 **说明** 此处上传的私钥文件应和主机上配置的公钥是一个密钥对，详情请参见**密钥对使用场景介绍**。
密钥对密码	是	如果密钥对未设置密码，则不需要填写该配置项。
密码	是	“登录方式”为“密码”时，显示该配置项。 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。

图 9-218 新建主机连接

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”，创建数据连接。

----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”。

图 9-219 创建 Python 脚本

步骤2 选择Python版本（以Python3为例），并选择主机连接，根据实际需要输入参数。

说明

配置的参数是指执行Python脚本时，向脚本传递的参数，参数之间使用空格分隔，例如：Microsoft Oracle。此处的“参数”需要在Python脚本中引用，否则配置无效。

步骤3 在编辑器中编辑Python语句。

本示例定义一个保存公司信息的字符串模板，然后应用该模板输出公司的信息。

```
import sys
Company_Name1=sys.argv[1]
Company_Name2=sys.argv[2]
template='No.{:0>9s} \t CompanyName: {:s} \t Website: https://www.{:s}.com'
context1=template.format('1',Company_Name1,Company_Name1.lower())
context2=template.format('2',Company_Name2,Company_Name2.lower())
print(context1)
print(context2)
```

说明

- 图9-220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。
- 主机连接：建立主机数据连接中创建的连接。

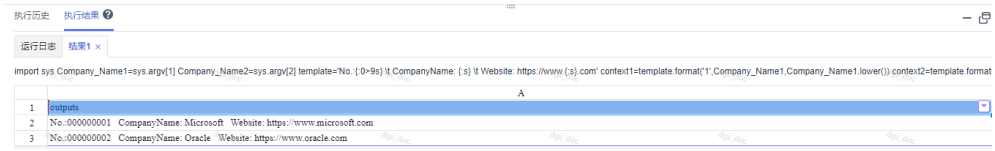
图 9-220 编辑 Python 语句

步骤4 单击“保存”，并提交版本。

步骤5 单击“运行”执行Python语句。

步骤6 查看脚本运行结果。

图 9-221 查看脚本运行结果



----结束

在作业中引用 Python 脚本

步骤1 创建一个作业。

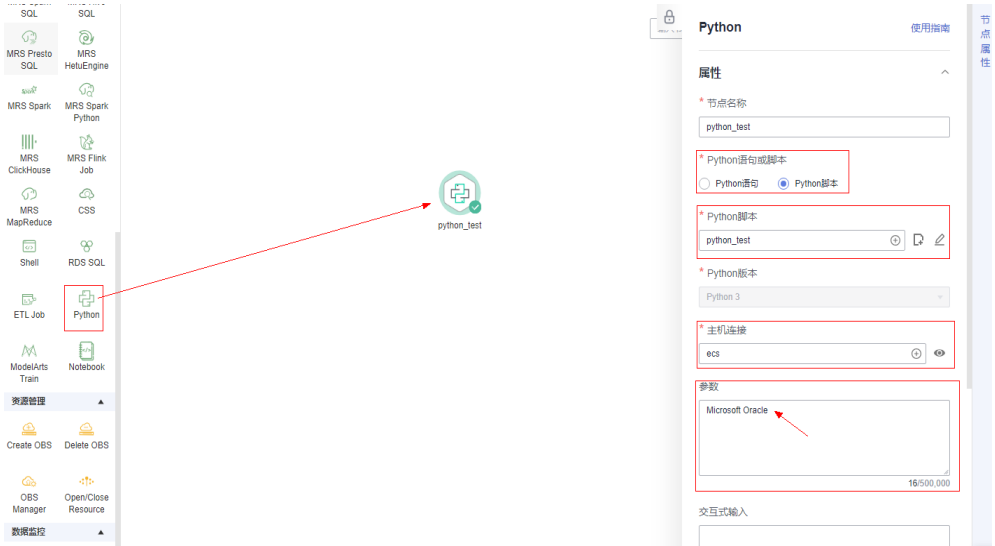
步骤2 选择Python节点，并配置节点属性。

选择已创建好的Python脚本，配置相关节点参数。在“参数”里面可以配置脚本参数，例如：

说明

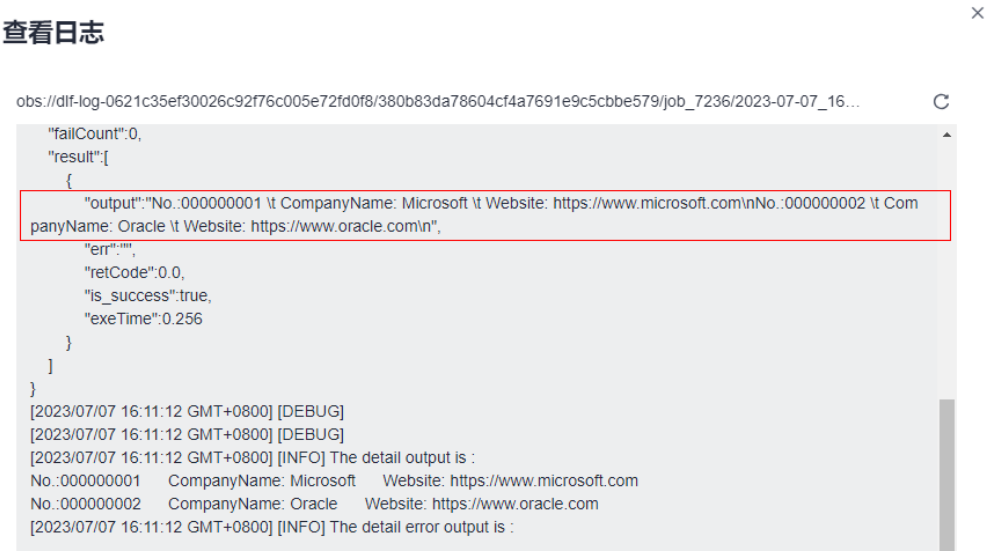
配置的参数是指执行Python语句时，向语句传递的参数，参数之间使用空格分隔，例如：Microsoft Oracle。此处的“参数”需要在Python语句中引用，否则配置无效。

图 9-222 配置 Python 节点属性



步骤3 单击“测试运行”，查看该作业的运行结果。

图 9-223 查看作业运行结果



步骤4 单击“保存”，作业配置信息创建完成。

步骤5 单击“提交”，提交版本后对该作业进行调度。

----结束

9.16.12 开发一个 DWS SQL 作业

介绍如何在数据开发模块上通过DWS SQL节点进行作业开发。

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”。创建DWS连接的操作请参见**DWS数据连接参数说明**。

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。创建数据库的详情请参考**新建数据库**进行操作。

创建数据表

在“gaussdb”数据库中创建数据表trade_log和trade_report。详情请参考如下建表脚本。

```
create schema store_sales;
set current_schema= store_sales;
drop table if exists trade_log;
CREATE TABLE trade_log
(
      sn          VARCHAR(16),
      trade_time   DATE,
      trade_count   INTEGER(8)
);
set current_schema= store_sales;
drop table if exists trade_report;
CREATE TABLE trade_report
(
      rq   DATE,
      trade_total   INTEGER(8)
);
```

开发 DWS SQL 脚本

在“数据开发 > 脚本开发”模块中创建一个DWS SQL脚本，脚本名称为“dws_sql”。在编辑器中输入SQL语句，通过SQL语句来实现统计前一天的销售额。

图 9-224 开发脚本

关键说明：

- **图9-224**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。

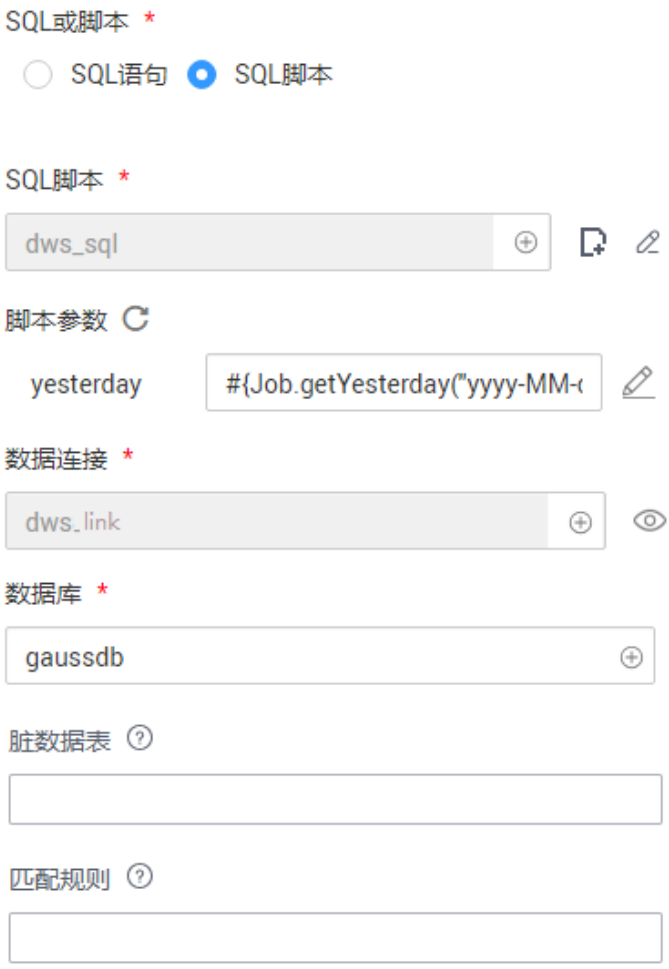
开发 DWS SQL 作业

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 创建一个批处理作业，作业名称为“job_dws_sql”。

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 9-225 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 9-226 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

9.16.13 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”。创建MRS Hive连接的操作请参见**MRS Hive数据连接参数说明**。

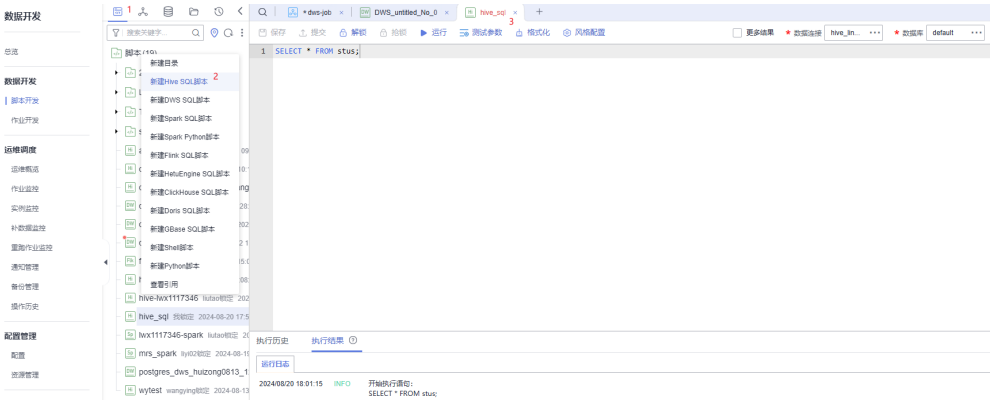
关键参数说明：

- 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 9-227 开发脚本



关键说明：

- **图9-227**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

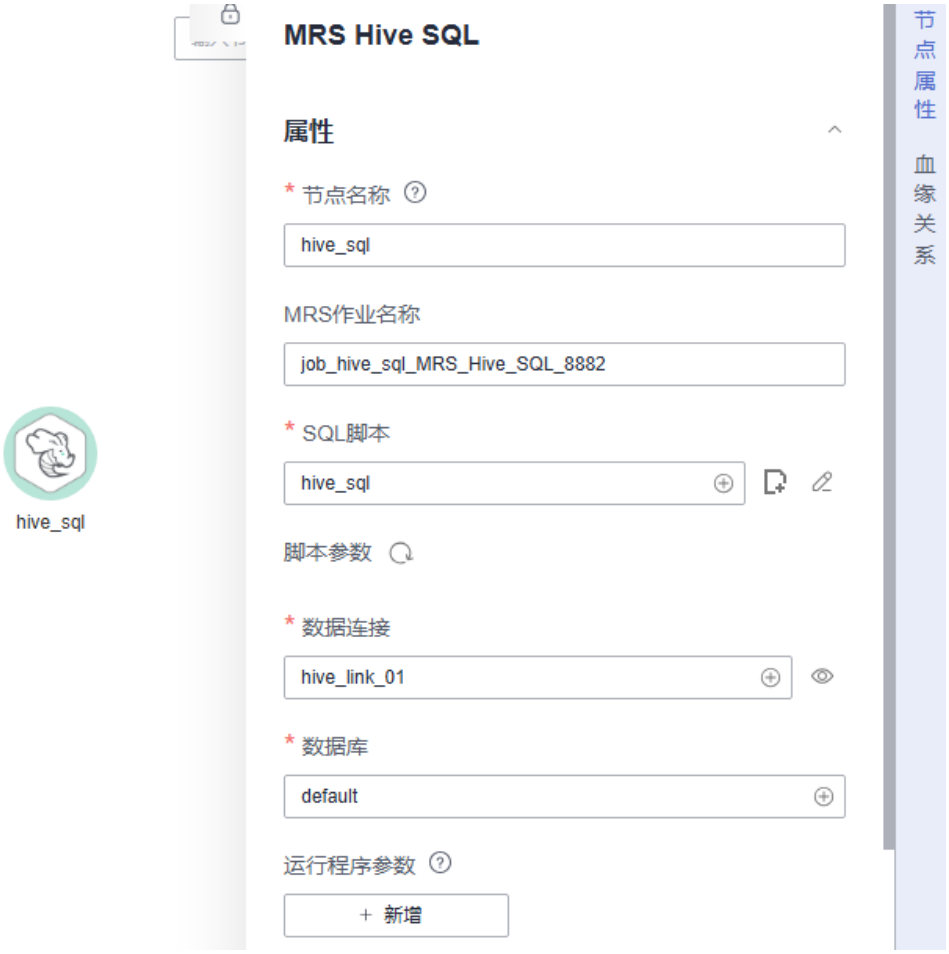
Hive SQL脚本开发完成后，我们为Hive SQL脚本构建一个周期执行的作业，使得该脚本能定期执行。

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 9-228 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 9-229 配置 MRS Hive SQL 节点属性



关键属性说明：

- 节点名称：默认显示为SQL脚本“hive_sql”的名称，支持修改。
- 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。

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 9-230 配置调度方式

说明

该作业调度时间在2021/01/01至2021/01/25，每天2点调度一次作业。

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

9.16.14 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

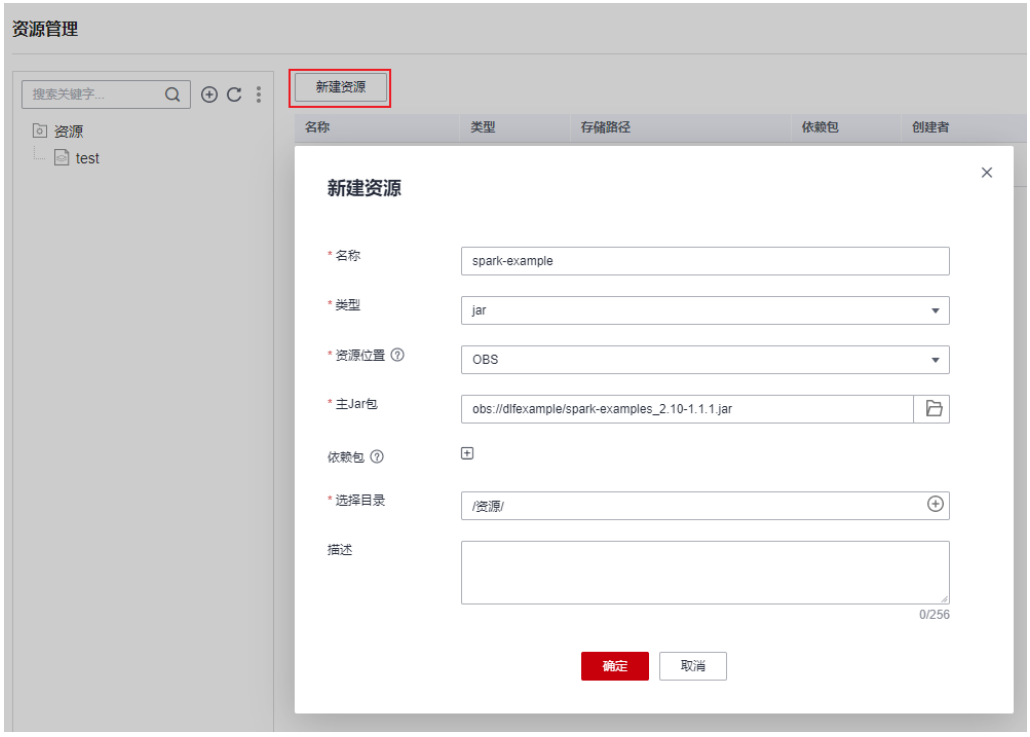
本示例使用的Spark作业代码来自maven库（下载地址：**https://repo.maven.apache.org/maven2/org/apache/spark/spark-examples_2.10/1.1.1/spark-examples_2.10-1.1.1.jar**），此Spark作业是计算π的近似值。

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 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发主界面的左侧导航栏，选择“配置管理 > 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 9-231 创建资源



----结束

提交 Spark 作业

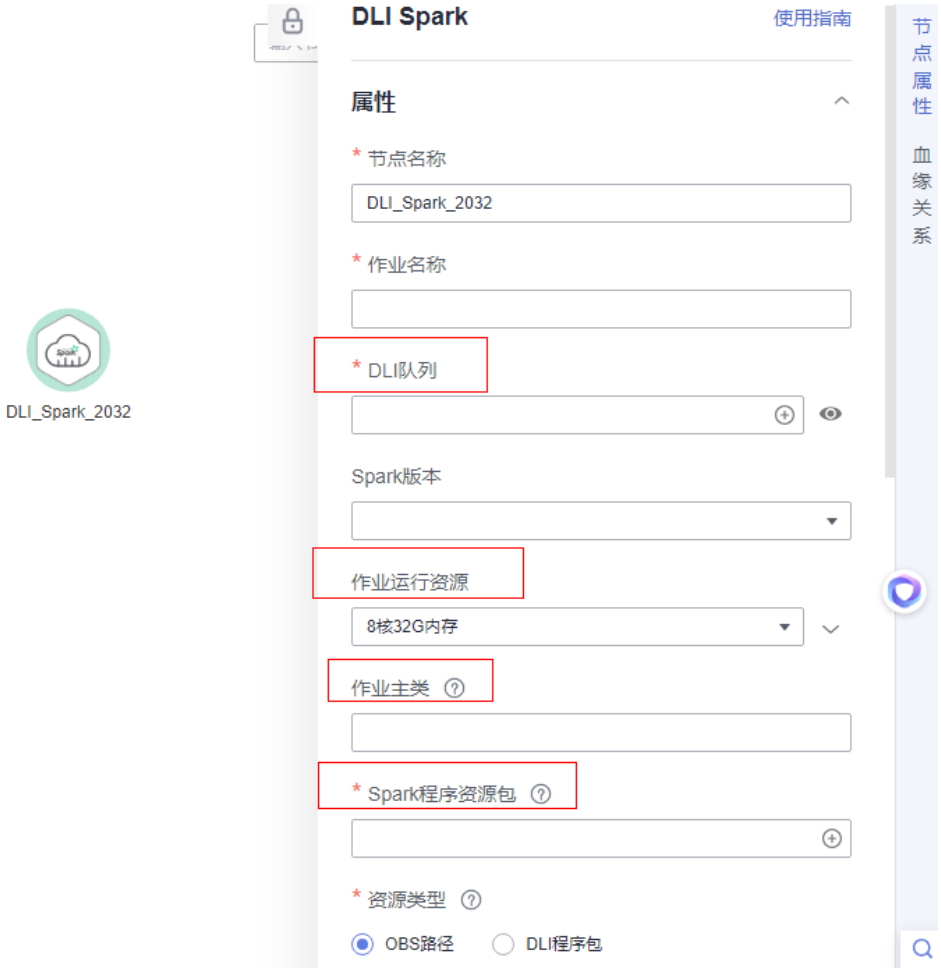
用户需要在数据开发模块中创建一个作业，通过作业的DLI Spark节点提交Spark作业。

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 9-232 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 9-233 配置节点属性

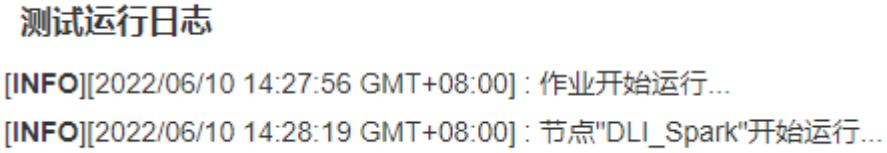


关键属性说明：

- DLI队列：DLI中创建的DLI队列。
- 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”。
- Spark程序资源包：**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 9-234 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

9.16.15 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 Flink作业开发。

场景说明

本教程通过开发一个MRS Flink作业来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/huaweicloudDocs/dgc/blob/master/WordCount.jar**

  下载的Flink作业资源包需要进行JAR包完整性校验。Windows操作系统下，打开本地命令提示符框，输入如下命令，在本地生成已下载JAR包的SHA256值，其中，“D:\wordcount.jar”为JAR包的本地存放路径和JAR包名，请根据实际情况修改。

  ```
  certutil -hashfile D:|wordcount.jar SHA256
  ```

  命令执行结果示例，如下所示：

  ```
  SHA256 的 D:\wordcount.jar 哈希:
  0859965cb007c51f0d9ddaf7c964604eb27c39e2f1f56e082acb20c8eb05ccc4
  CertUtil: -hashfile 命令成功完成。
  ```

  对比所下载JAR包的SHA256值和下面JAR包的SHA256值。如果一致，则表示下载过程不存在篡改和丢包。

  SHA256值：
  0859965cb007c51f0d9ddaf7c964604eb27c39e2f1f56e082acb20c8eb05ccc4

- 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

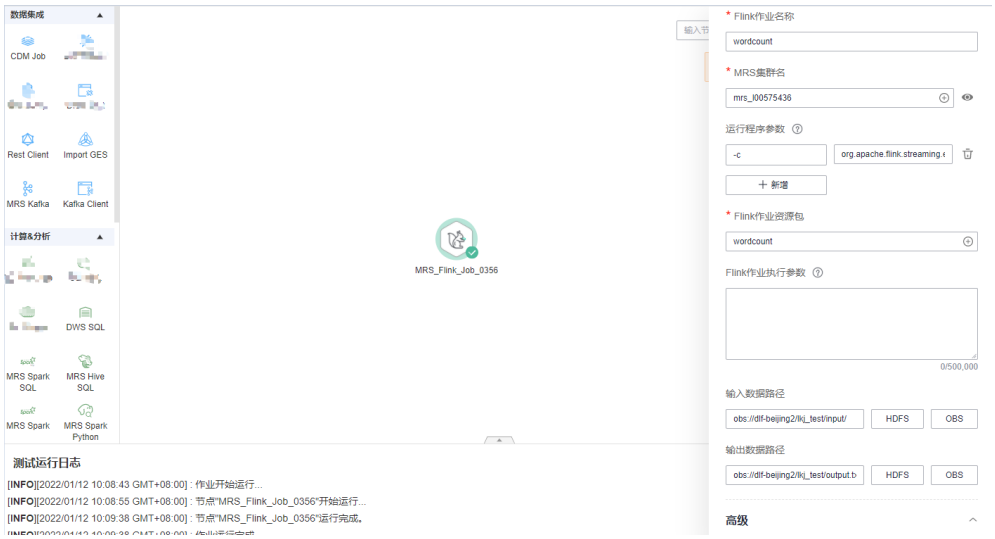
本例中，**WordCount.jar**文件上传路径为：lkj_test/WordCount.jar；**word.txt** 文件上传路径为：lkj_test/input/word.txt。

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 9-235 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 9-236 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-test/lkj_test/input/word.txt
--输出数据路径
obs://dlf-test/lkj_test/output.txt
```

其中：

obs://dlf-test/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-test/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

9.16.16 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：

开发者具有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 9-237 上传文件至 OBS 桶

说明

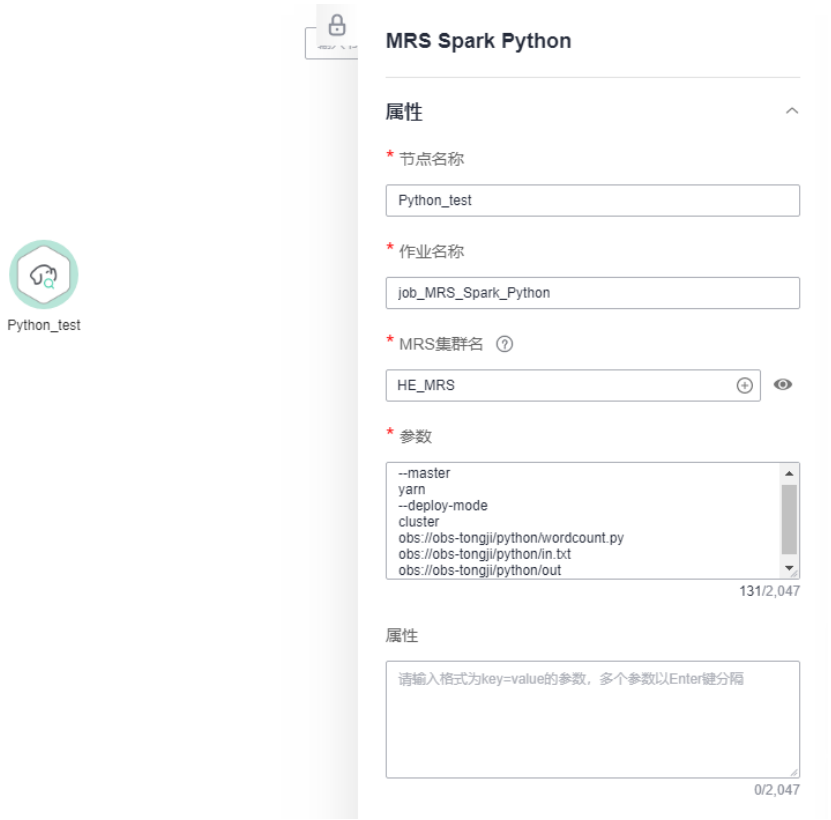
本例中，**wordcount.py**和**in.txt**文件上传路径为：obs://obs-tongji/python/

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 9-238 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 9-239 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；

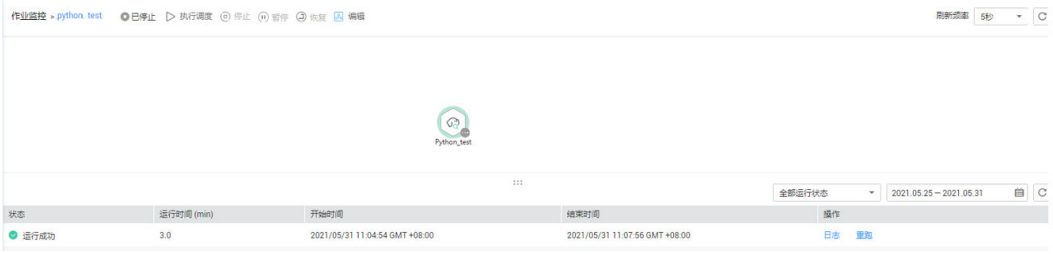
obs://obs-tongji/python/out为输出参数文件夹的路径，并且会在OBS桶中自动创建该目录（如已存在out目录，会报错）。

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 9-240 查看作业执行结果



作业日志中显示已运行成功

图 9-241 作业运行日志

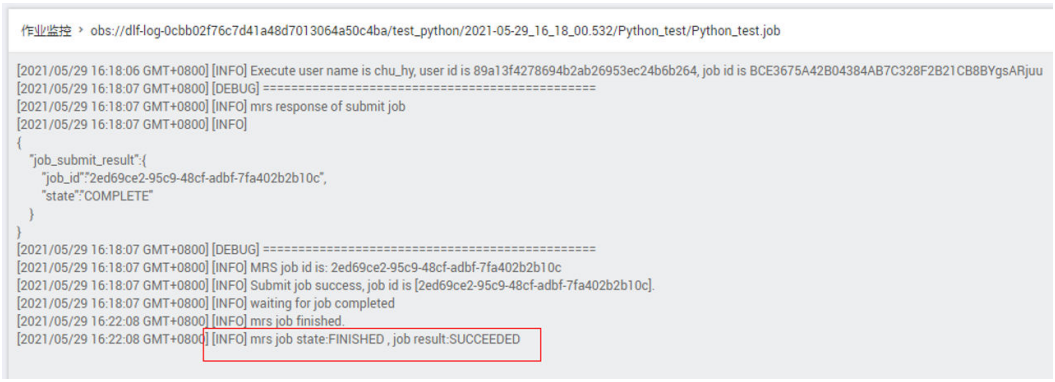


图 9-242 作业运行状态

步骤7 查看OBS桶中返回的记录。（没设置返回可跳过）

图 9-243 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：

开发者具有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
```

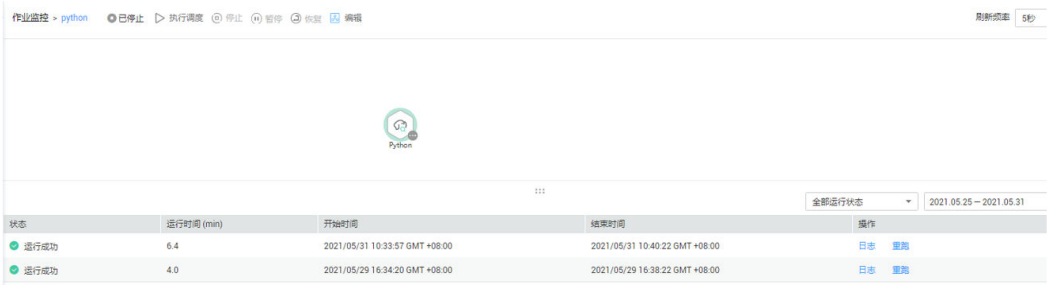
其中：zt_test_sparkPython1.py 为脚本所在路径

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 9-244 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 9-245 查看 YARN 上日志

----结束

10 数据质量

10.1 业务指标监控（待下线）

10.1.1 业务指标监控简介

须知

当前，数据架构有完善的指标设计和管理能力，建议您后续使用数据架构的业务指标功能，数据质量的业务指标监控模块即将下线。

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。

功能	说明
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。

10.1.2 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

已在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10-1创建归属目录。

图 10-1 新建指标的归属目录

表 10-1 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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据质量”模块，进入数据质量页面。
3. 选择“业务指标监控 > 指标管理”。
4. 单击“新建”，在弹出的对话框中，参见**表10-2**配置相关参数。

表 10-2 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、MRS Hive、DLI、MRS ClickHouse、DORIS。 • 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10-1**。
来源类型	支持“自定义”。 用户自定义SQL语句，定义指标的来源。

10.1.3 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

已在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10-2**创建归属目录。

图 10-2 新建规则的归属目录

表 10-3 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，选择对应工作空间的“数据质量”模块，进入数据质量页面。
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10-4**配置相关参数，新建规则。

表 10-4 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10-2**。
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。

10.1.4 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

已在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10-3创建归属目录。

图 10-3 新建业务场景的归属目录

表 10-5 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，选择对应工作空间的“数据质量”模块，进入数据质量页面。
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10-6配置相关参数，新建场景。

表 10-6 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。

参数名	说明
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10-3。
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。
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。 **说明** 当前仅支持“短信”、“邮件”这两种协议的订阅终端订阅主题。

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10-7。

表 10-7 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。

10.1.5 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“业务指标监控 > 运维管理”页面中的区域和按键功能。

图 10-4 运维管理页面

表 10-8 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。 ● 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 10-9 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。

菜单/按键	说明
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 > 处理日志	可查看历史处理记录。

10.2 数据质量监控

10.2.1 数据质量监控简介

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、跨源和跨表的分析。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。

功能	说明
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。

10.2.2 新建数据质量规则

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34种，分为库级规则、表级规则、字段级规则和跨字段级规则、跨源级规则等规则类型，如表10-10所示。

表 10-10 系统内置的规则模板一览表

规则类型	维度	模板名称	适用引擎	说明
库级	完整性	数据库空值扫描	DLI、DWS、HIVE、SparkSQL、CLICKHOUSE、ORACLE、RDS、DORIS	计算数据库每个表中每个字段的空值字段行数，结果以字段为维度呈现。
表级	准确性	表行数	DLI、DWS、HIVE、SparkSQL、CLICKHOUSE、HETUENGINE、ORACLE、RDS、DORIS	计算数据表的总行数。
	完整性	数据表空值扫描	DLI、DWS、HIVE、SparkSQL、CLICKHOUSE、HETUENGINE、ORACLE、RDS、DORIS	计算数据表中每个字段的空值行数，结果以字段为维度呈现。
	有效性	近1天波动率	DLI、DWS、HIVE、SparkSQL、CLICKHOUSE、HETUENGINE、ORACLE、RDS、DORIS	计算数据表的单表大小、字段分组、相关波动率近一天的规则波动监控。
		近7天波动率		计算数据表的单表大小、字段分组、相关波动率近七天的规则波动监控。

规则类型	维度	模板名称	适用引擎	说明
		近30天波动率		计算数据表的单表大小、字段分组、相关波动率近三十天的规则波动监控。
字段级	唯一性	字段唯一值	DLI、DWS、HIVE、SparkSQL、CLICKHOUSE、HETUENGINE、ORACLE、RDS、DORIS	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（当有多个不同的重复值时，以所有重复值个数的和作为该字段的重复值行数）。
		多字段唯一性校验	HIVE、SparkSQL、DLI、DWS、HETUENGINE	校验数据表中多个字段的组合是否唯一，最多支持10个字段的组合。
		多字段唯一性校验忽略Null		校验数据表中多个字段的组合是否唯一，最多支持10个字段的组合，Null值被统计在有效行中。
	完整性	字段空值	DLI、DWS、HIVE、SparkSQL、CLICKHOUSE、HETUENGINE、ORACLE、RDS、DORIS	计算数据表中指定字段的空值行数。
	准确性	字段平均值	DLI、DWS、HIVE、SparkSQL、CLICKHOUSE、HETUENGINE、ORACLE、RDS、DORIS	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。
	有效性	身份证校验	DLI、DWS、HIVE、SparkSQL、CLICKHOUSE、HETUENGINE、ORACLE、RDS、DORIS	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（如果数据为空，则视为非法字段）。
		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。

规则类型	维度	模板名称	适用引擎	说明
		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。
		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。
		字段长度校验	DLI、DWS、HETUENGINE	通过输入字段长度范围，校验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。
		枚举值校验忽略Null		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，Null值被统计在有效行中。
		正则表达式校验忽略Null		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，Null值被统计在有效行中。

规则类型	维度	模板名称	适用引擎	说明
		枚举值校验忽略大小写敏感	DLI、DWS、HETUENGINE	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，大小写敏感值被统计在有效行中。
		枚举值校验忽略Null忽略大小写敏感		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，Null值和大小写敏感值被统计在有效行中。
跨字段级	一致性	字段一致性校验	DLI、DWS、HIVE、SparkSQL、CLICKHOUSE、HETUENGINE、ORACLE、RDS、DORIS	针对相同数据源的不同字段，校验数据表中指定字段的值是否与参考字段所在表中的值一致。
	准确性	跨字段时间校验	DLI、DWS、HETUENGINE	针对相同数据源的不同字段，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。
跨源级	一致性	跨源字段一致性校验	HETUENGINE	基于Hetu连接，针对不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。

系统内置的规则模板不可编辑和查看发布历史。

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自定义模板和自定义规则：

说明

自定义规则模板是很多用户可能都要使用的数据，不能随意进行修改，开发者只有查询权限，如果要修改规则模板，请联系管理员进行修改。

- 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被划分为对应的规则类型（表级、字段级、跨字段级和多表多字段），模板类型显示为自定义模板。新建质量/对账作业应用自定义模板与其他内置模板选择方式相同，规则类型选择为“表级规则”、“字段级规则”、“跨字段级规则”或“多表多字段规则”后即可选择自定义模板，支持进行异常数据输出，不支持质量评分。
- 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。

说明

SQL语句可以包含同一数据库下的多张表，但不同数据库的表无法共存。

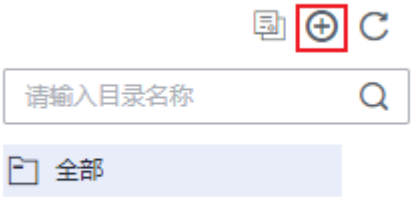
本文以新建自定义模板为例，说明如何创建规则。如果您需要新建自定义规则，请直接参考**新建数据质量作业**进行自定义规则质量作业的创建。

步骤1 （可选）选择“数据质量监控 > 规则模板”，新建目录。如果已存在可用的目录，可以不用新建目录。注意，规则模板、质量作业和对账作业的目录为同一目录，择一操作即可。

当前系统支持“新建目录”和“同步主题为目录”两种方式：

选择“新建目录”时，直接在目录处单击，输入目录名称，即可完成目录新建。直接新建目录的最大深度拓展为7层。

图 10-5 新建目录

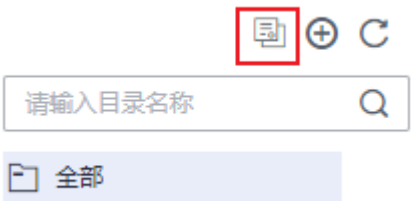


选择“同步主题为目录”时，在目录处单击，即可将**数据架构处的主题**同步到目录中（仅支持同步“已发布”状态的主题）。同步后的主题目录与数据架构发布后主题一致，按照主题层级如L1、L2等进行展示。

说明

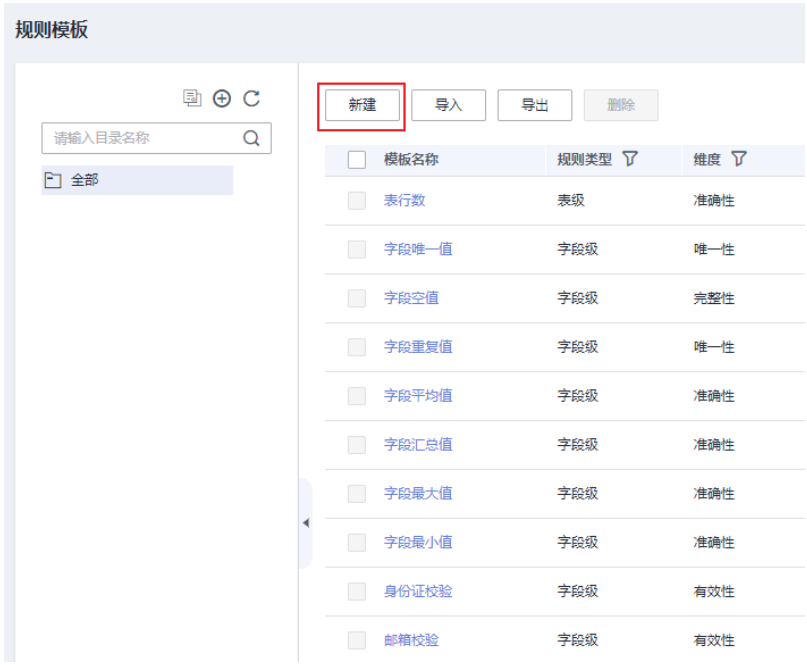
1. 直接新建的普通目录不受同步影响。但当普通目录名称与主题名称冲突时：
   - 主题首次同步名称冲突时，会将普通目录修改为主题目录属性，带上主题层级如L1、L2等进行展示。
   - 主题修改后再次同步名称冲突时，会出现同步失败的情况。
2. 不支持变更自动同步。即数据架构处的主题或主题层级变更并发布后，需要手动再次单击才能同步到主题目录。
   特殊的，数据架构处的主题或主题层级删除后，手动同步后目录不会删除，仅去除主题目录属性，作为普通目录进行展示。
3. 完成同步后，若有同步失败的主题，系统会自动弹出同步结果明细，可查看同步失败的主题名称。

图 10-6 同步主题为目录



步骤2 在“规则模板”页面，单击“新建”，在弹出的新建规则模板页面中进行配置。

图 10-7 新建规则模板



步骤3 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。
- 所属目录：选择该规则模板所在的目录。
- 标签：选择所需的标签。标签是在数据地图组件中定义的标签。如果未使用数据地图组件，则标签功能不生效。
- 描述：对此自定义模板进行简单说明。
- 定义关系：输入SQL语句，实现对数据的查找。其中，${Schema_Table1}表示质量/对账作业中所选的表，${Column1}为${Schema_Table1}中所选的字段，${Schema_Table2}仅当定义跨字段级规则时存在，表示质量作业中所选的参考表，${Column2}为${Schema_Table2}中所选的字段。系统支持对定义关系进行语义校验。

  说明

  在自定义规则模板时，在定义关系时，如果出现非数字时，只能输出运行结果，不能进行四则运算、逻辑运算和绝对值。

  自定义规则模板的定义关系目前最多支持10张表20个字段。

  自定义的SQL表达式有如下要求：

  a. 关系表达式中最多支持五列输出。

  b. 支持最多两张表的入参和两个字段的入参。注：${Column1}为 ${Schema_Table1}的入参 ，${Column2}为${Schema_Table2}的入参，内置逻辑指定。

c. 如果结果查到多行，只使用第一行数据。

d. 不支持使用.连接表和字段，如${Schema_Table2}.${Column1}.${Input_String1}。

e. 非多表多字段表达式中参数只能使用：${Schema_Table1}、${Schema_Table2}、${Column1}、${Column2}，且不要使用表别名。

f. 多表多字段表达式中参数可以使用：${Schema_Table1}、${Schema_Table2}、${Schema_Table3}、...${Schema_Table5}、${Column1}、${Column2}，${Column3}、...${Column20}、${Input_String1}、${Input_String2}、...${Input_String5}，且不要使用表别名。

例如统计表行数，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。

单击  ，开启“添加输入参数”，可以在SQL语句中灵活配置输入参数。

例如字段匹配配置表中的行数，输入select count(1) from ${Schema_Table1} where ${Column1} regexp ${Input_String1}。其中${Column1}通过单击“添加字段参数”生成，${Schema_Table1} 通过单击“添加库表参数”生成，${Input_String1}通过单击“添加输入参数”生成。

说明

配置多表多字段规则模板时，目前仅支持最多5个库表、20个字段、5个输入参数。

- 输出结果说明：对SQL获得结果的每一列进行说明，与关系定义的输出结果顺序一一对应，列说明之间用英文逗号进行分隔。

  例如当定义关系设置为：select max(${Column1}),min(${Column2}) from ${Schema_Table1}，则输出结果说明可写为“最大值，最小值”，注意输入顺序。

- 评分公式：此处输入评分公式。自定义模板在此处输入评分公式后，可以参与质量评分，在质量报告中显示评分和规则。

  示例：${1}/${2}，其中${1}和${2}分别表示第1列输出结果和第2列输出结果；公式的返回值范围是[0-1]。

- 异常表模板：此处需输入完整的SQL语句，指定输出哪些数据是异常数据。其中，${Schema_Table1}通过单击“添加库表参数”生成，表示异常表的表名；${Column1}通过单击“添加字段参数”生成，表示异常表中所选的字段；${Output_Columns}通过单击“添加输出参数”生成，表示异常表中指定输出的异常数据。系统支持对异常表模板进行语义校验。

  说明

  开启“多表多字段”开关后，“异常表模板”参数不显示，不支持配置。

例如，有一张涉及金额的表，表中“is_test”字段用于标识该条数据是否为测试数据（0为正式数据，1为测试数据）。期望计算正式数据的金额最小值，最大值，平均值以及总和。则自定义模板可设置如下：

- 维度：准确性。
- 所属目录：/全部/。
- 描述：计算正式数据的金额最小值，最大值，平均值以及总和。

- 定义关系：输入如下SQL语句，计算正式数据的金额最小值，最大值，平均值以及总和。其中${Schema_Table1}表示质量作业中所选的表，${Column1}表示${Schema_Table1}中所选的字段。

```
select
    min(${Column1}),
    max(${Column1}),
    ROUND(avg(${Column1}),2),
    sum(${Column1})
from ${Schema_Table1}
    where is_test='0'
```

- 输出结果说明：最小值，最大值，平均值，总和。
- 异常表模板：输入如下SQL语句，将正式数据中金额小于10对应的${Output_Columns}列作为异常表数据输出。其中${Output_Columns}表示质量作业中异常表参数所选的字段。

```
select ${Output_Columns} from ${Schema_Table1} where ${Column1}<10 and is_test='0'
```

图 10-8 自定义规则模板关键参数

步骤4 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

编辑规则模板

说明

自定义规则模板是很多用户可能都要使用的数据，不能随意进行修改，开发者只有查询权限，如果要修改规则模板，请联系管理员进行修改。

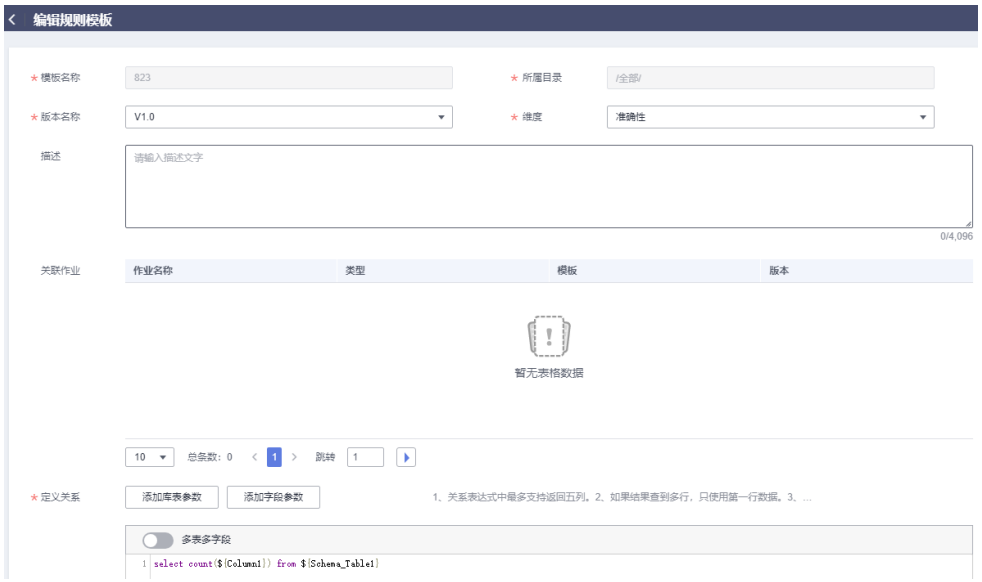
自定义规则模板支持直接修改规则模板内容并进行发布。同时，可以选择下线历史版本且将待下线历史版本关联的作业迁移到新版本上。具体请参见如下操作。

说明

编辑规则模板时，编辑界面增加了“版本名称”和“关联作业”两个参数。

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“编辑”进入编辑规则模板界面。

图 10-9 编辑规则模板



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”，在提交发布对话框中，选择发布的版本类型，重新设置版本名称，并确认发布。

说明

- 发布新版本时，需要修改版本名称。
- 发布当前版本时，可以不修改版本名称，之前老的版本会自动生成为历史版本。

图 10-10 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，支持查看版本变化信息、修改版本名称、下线对应版本等。

图 10-11 发布历史界面

步骤5 如需下线历史版本，单击历史版本最右侧的“下线”按钮。

- 如果该版本没有关联作业，单击确认即可下线。
- 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，单击确认才能完成下线。

图 10-12 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 10-13 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4MB数据的文件。

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。

图 10-14 导入规则模板

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。

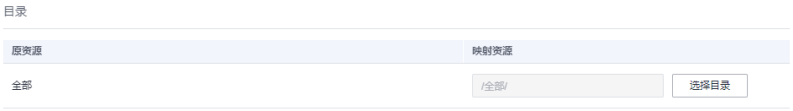
说明

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。如不选择，默认使用原映射资源信息。

图 10-15 配置映射资源信息



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

10.2.3 新建数据质量作业

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。

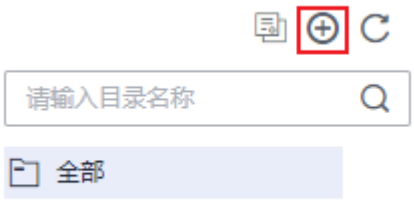
配置流程

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据质量”模块，进入数据质量页面。
2. （可选）选择“数据质量监控 > 质量作业”，新建目录。如果已存在可用的目录，可以不用新建目录。注意，规则模板、质量作业和对账作业的目录为同一目录，择一操作即可。

   当前系统支持“新建目录”和“同步主题为目录”两种方式：

   选择“新建目录”时，直接在目录处单击，输入目录名称，即可完成目录新建。直接新建目录的最大深度拓展为7层。

   图 10-16 新建目录

   选择“同步主题为目录”时，在目录处单击，即可将**数据架构处的主题**同步到目录中（仅支持同步“已发布”状态的主题）。同步后的主题目录与数据架构发布后主题一致，按照主题层级如L1、L2等进行展示。

   说明

   1. 直接新建的普通目录不受同步影响。但当普通目录名称与主题名称冲突时：
      - 主题首次同步名称冲突时，会将普通目录修改为主题目录属性，带上主题层级如L1、L2等进行展示。
      - 主题修改后再次同步名称冲突时，会出现同步失败的情况。
   2. 不支持变更自动同步。即数据架构处的主题或主题层级变更并发布后，需要手动再次单击才能同步到主题目录。

      特殊的，数据架构处的主题或主题层级删除后，手动同步后目录不会删除，仅去除主题目录属性，作为普通目录进行展示。
   3. 完成同步后，若有同步失败的主题，系统会自动弹出同步结果明细，可查看同步失败的主题名称。

   图 10-17 同步主题为目录

3. 在“质量作业”页面单击“新建”，在弹出的对话框中，参见**表10-11**配置相关参数。

表 10-11 配置作业参数

参数名	说明
*作业名称	质量作业的名称。
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1024个字符。
标签	选择所需的标签。标签是在数据地图组件中定义的标签。如果未使用数据地图组件，则标签功能不生效。
*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（可选）新建目录。
*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。
问题处理人	选择质量作业的问题处理人。
超时时间	输入超时时间。输入值必须在5到1440之间。单位为分钟。该参数为空或者默认1440分钟时，超时时间系统默认为24小时，支持修改。

4. 单击“下一步”，进入规则配置页面，每个规则卡片对应一个子作业。您需要单击规则卡片中的  ，然后参见表10-12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 10-18 打开质量作业规则配置

表 10-12 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。描述信息长度不能超过1024个字符。

添加方式	配置	说明
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、跨源级规则、多表多字段和自定义规则，自定义规则可针对表中的具体字段配置监控规则。 **说明** • 选择跨字段级规则时，需要在计算范围中同时配置数据表和参考表。 • 跨源级规则目前只支持基于Hetu连接的MRS Hive和DWS之间的字段对比作业。 • 配置跨源级规则前，需要在MRS Hetu中创建MRS Hive数据源和GaussDB数据源。详情请参考配置Hive数据源和配置GaussDB数据源。

	数据连接	来源对象/目的对象支持的数据源类型：DWS、MRS Hive、MRS Spark、DLI、ORACLE、RDS（MySQL、PostgreSQL）、Hetu、MRS Spark（Hudi）、MRS ClickHouse、DORIS。
		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。
		• 数据质量当前不支持MRS Hive组件的严格模式。

	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”或“表级规则”，数据对象选择对应的数据库即可。 • 当“规则类型”选择“自定义规则”，选择对应的数据库即可。
	scheme	如果数据源有scheme，才需要配置。没有则不显示该参数。
	数据表	选择配置的数据质量规则所应用到的表。 **说明** • 数据表与数据库强相关，基于已选择的数据库。 • 当“规则类型”选择“表级规则”，数据对象选择对应的数据表。 • 当“规则类型”选择“自定义规则”，表名选择对应的数据表。

添加方式	配置	说明
	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。 支持对SQL语句进行语义校验，语义校验结果仅供参考。
	参数默认值	自定义SQL可设置入参用于执行，SQL入参需要与参数默认值顺序匹配（数据质量单点执行时）。 **说明** 当通过数据开发任务调度质量算子的时候，优先使用数据开发中定义的参数值。
	字段名	仅用于异常表。配置字段名称用于异常表数据。 示例：column1,column2,column3
	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。
	数据对象	当“规则类型”选择“自定义规则”时，不需要配置该参数。其他规则类型均需要配置该参数，此处选择参考的数据字段。 选择表名时，搜索框支持大小写敏感。
	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。 选择表名时，搜索框支持大小写敏感。
	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。
	输出结果说明	当“规则类型”选择“自定义规则”时，需要配置该参数。 ● 对SQL获得结果的每一列进行说明，与SQL关系定义的输出结果顺序一一对应，输出结果说明字段个数与SQL的输出参数个数不相等时，会保存失败并提示报错信息。 ● 输出结果说明只能包含中文，英文字母、数字、下划线、中划线、空格等字符。 ● 例如SQL设置为：select max(${Column1}),min(${Column2}) from ${Schema_Table1}，则输出结果说明可写为“最大值，最小值”，注意输入顺序。输出结果说明中有多个字段时，用英文逗号进行分隔。如果输出结果说明中使用中文逗号，在保存时会自动替换成英文逗号。

添加方式	配置	说明
质量规则	入参	当“规则类型”选择“多表多字段”，需要配置该参数。 比如，入参为Input_String1，根据实际业务需要进行参数值配置。 **说明** 当“规则类型”选择“多表多字段”时，选择了规则模板对应的版本号以后，SQL语句会自动显示。SQL语句中包含的参数与入参的数量是一致的。如果SQL语句中没有包含参数，则无需配置入参。
计算引擎	队列名称	选择运行质量作业的引擎。仅数据连接为DLI、Hive或Hetu类型时，此参数有效，输入队列名称。 当连接类型为Hetu，规则类型为除了库级以外的所有系统模板，自定义模板，自定义规则时，队列名称指的是Hetu引擎的资源队列名称。查看Hetu引擎的资源队列名称，需要登录MRS的FusionInsight Manager系统，单击左侧导航的HetuEngine，在基本信息区域，单击HSConsole WebUI链接，在计算实例列表中查看Hetu引擎的资源队列名称。
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10-10**。除去系统内置规则模板外，您也可关联在**新建数据质量规则**中新建的自定义模板。 当“规则类型”选择“字段级规则”，规则模板名称选择“正则表达式校验”或“正则表达式校验忽略Null”时，正则表达式的规则长度最大支持1024个字符。
	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。
	SQL	选择了模板名称和版本后，SQL自动显示。
	规则权重	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。 数据质量作业支持传参，可以将环境变量参数传递给数据质量作业。 系统支持对多个表配置规则时，不同表的数据范围可支持独立设置。当“数据对象”和“参考数据对象”的字段名称都配置时，需要配置所对应扫描区域的数据扫描范围。

添加方式	配置	说明

	where条件	输入where子句，系统会选择符合条件的数据进行扫描。
		说明
		配置where条件语句时，最前面需要加and，因为在SQL生成中需要进行语法的校验，否则会报语法错误。
		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：
		`and 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：
		`and time >= (date_trunc('hour', now()) - interval '24 h') and time <= (date_trunc('hour', now()))`
		数据质量支持传递参数，可以输入条件表达式，例如环境变量传参，可设置如下的内容：
		`and p_date=${target_date}`
		数据质量支持从数据开发传递参数给数据质量，并且会主动获取数据开发的参数，自定义规则模板和系统模板都支持。

	参数默认值	当选择“条件扫描”时可填写。 请按照输入的where条件文本框中出现的参数名，依次填写默认参数值。 **说明** 参数默认值优先由数据开发传递，为空时可能会造成质量作业运行出错。 数据开发传递参数给数据质量后，作业运行完以后，通过“查看SQL”可以查看所传递的数据开发的参数以及参数值。

添加方式	配置	说明
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，单击告警参数“空值行数”，在“告警表达式”输入框会显示为“${1}”。

添加方式	配置	说明
	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。 开启质量评分开关后，选择“schema”（如果数据源有）和“表名”后，运行结果的“名称”显示为 **<数据库名>.<表名>**，不参与质量评分的子作业，“名称”显示 **<数据库名>.custom-sql**。
	评分公式	当“规则类型”选择“自定义规则”时，需要配置该参数。 输入评分公式。可以对评分公式进行试跑。 示例：${1}/${2}，参数和告警表达式保持一致，公式的返回值范围是[0-1]。
	权重规则	当“规则类型”选择“自定义规则”时，需要配置该参数。 设置规则的权重。权重范围：【1-9】，整数。默认值为5。

添加方式	配置	说明
	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** • 系统内置模板中，“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”、“字段最小值”模板，“多表多字段规则”模板均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。 • 当规则类型选择“跨源级规则”，必须开启“生成异常数据”。
	异常表	单击选择数据库和schema，可以自定义配置输出数据表名的前后缀、选择已有表和异常字段。如果不配置异常字段，默认会输出异常表的所有字段。 **说明** 自定义异常表包含四种方式：添加表前后缀、添加表前缀、添加表后缀、选择已有表。表前缀以英文字母和下划线开头，且只能包含英文字母、数字和下划线。表后缀只能包含英文字母、数字和下划线。 当单击“选择已有表”时，需要选择表名，数据库和schema系统默认，如果未选择表名，则显示**数据库名.scheme.undefined**。 设置异常表时，系统会默认添加表后缀err。
	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。 • 清理异常数据：勾选，清理异常数据会清除当前子规则历史异常数据，请谨慎操作。数据质量作业在重跑时，会清空异常表中的历史数据。
	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。系统支持对异常表SQL进行语义校验。
	查看SQL	单击后可以查看异常表SQL语句。在查看异常表的SQL时，支持查看所创建的SQL和插入的SQL。
	查看相同规则	单击后可查看如下相同规则： • 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。

5. 单击“下一步”以后，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。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服务主题。

   通知类型包含“触发告警”和“运行成功”两种，当前仅支持“短信”、“邮件”这两种协议的订阅终端订阅主题。

   说明

   开启订阅配置后，每个满足通知类型的子作业都会发送通知。如果开启告警，失败告警通知不需要单独配置，任务运行失败后会发送告警。

   打开“通知抑制”后，告警上报的通知策略可以进行配置，在最近N分钟以内，连续N次告警，则发送告警通知。最近时间可支持配置1~360分钟，连续次数可支持配置1~10次。

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10-13**。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 10-13 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。

质量作业创建完成后，可以在作业里面进行查看，系统支持通过作业名称、创建人、责任人、表名、最近运行时间进行筛选。同时，系统支持模糊搜索。

质量作业创建完成后，可以对质量作业进行编辑、删除、运行、启动调度、停止调度等操作。

说明

单次调度模式不支持启动调度。

运行单个质量作业

系统支持运行单个质量作业。

步骤1 选择“数据质量监控 > 质量作业”，选择要运行的质量作业。

步骤2 单击“操作”列的“运行”。

步骤3 企业模式下，选择运行环境，系统支持可选择“开发环境”或“生产环境”。

步骤4 单击“确定”。

----结束

导出质量作业

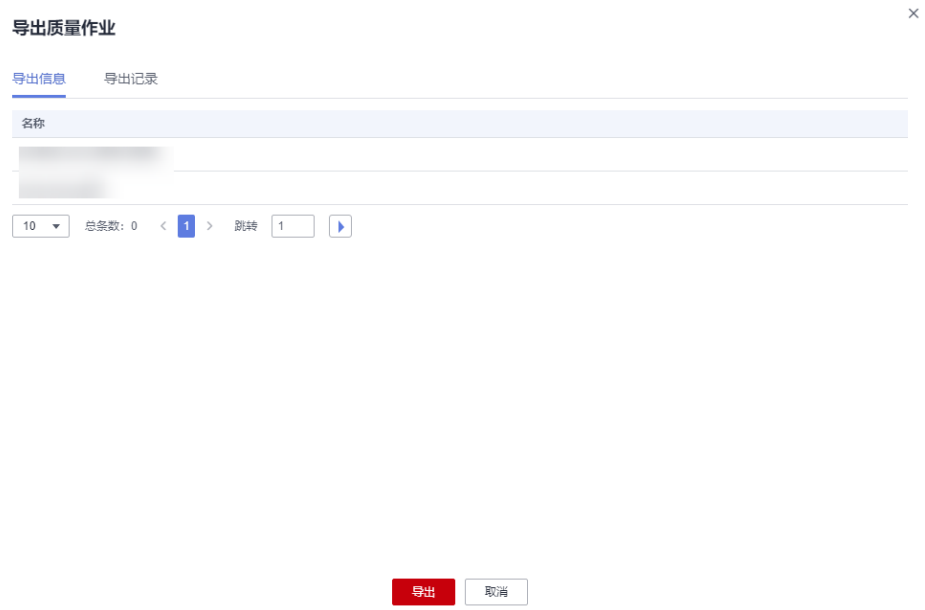
系统支持批量导出质量作业，一次最多可导出200个质量作业。导出作业时，导出的单元格内容最大长度支持65534个字符。

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。

图 10-19 导出

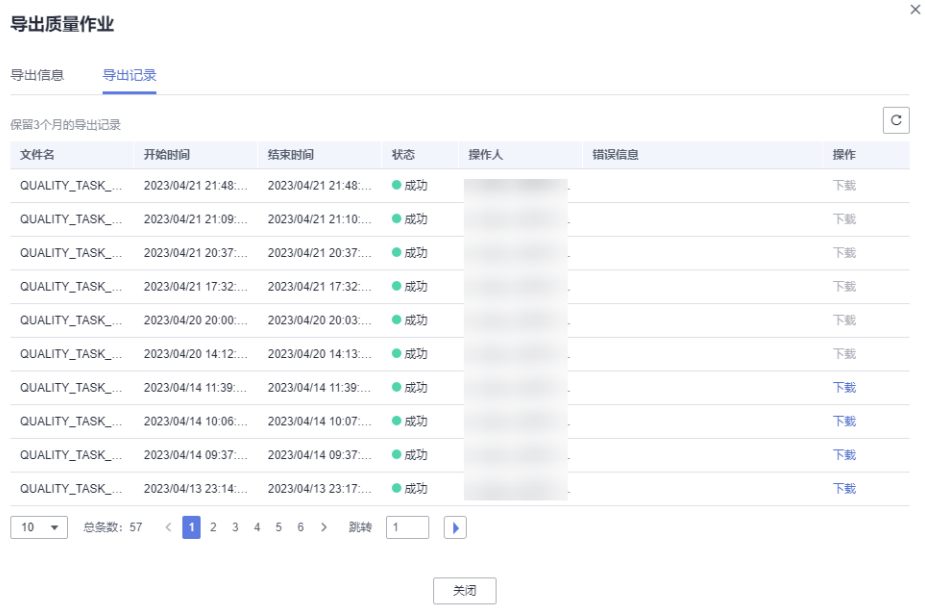
步骤2 单击“导出”，弹出“导出质量作业”对话框。

图 10-20 导出质量作业



步骤3 切换到“导出记录”页签，可查看当前任务的导出结果。

图 10-21 导出记录



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

导出全部质量作业

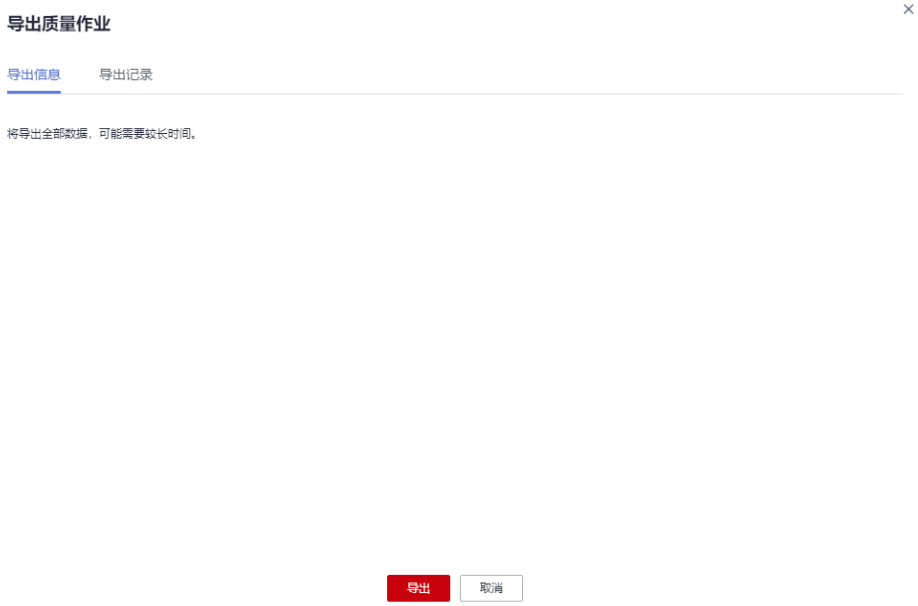
系统支持导出全部质量作业。导出作业时，导出的单元格内容最大长度支持65534个字符。

步骤1 选择“数据质量监控 > 质量作业”，单击“全部导出”。

图 10-22 全部导出

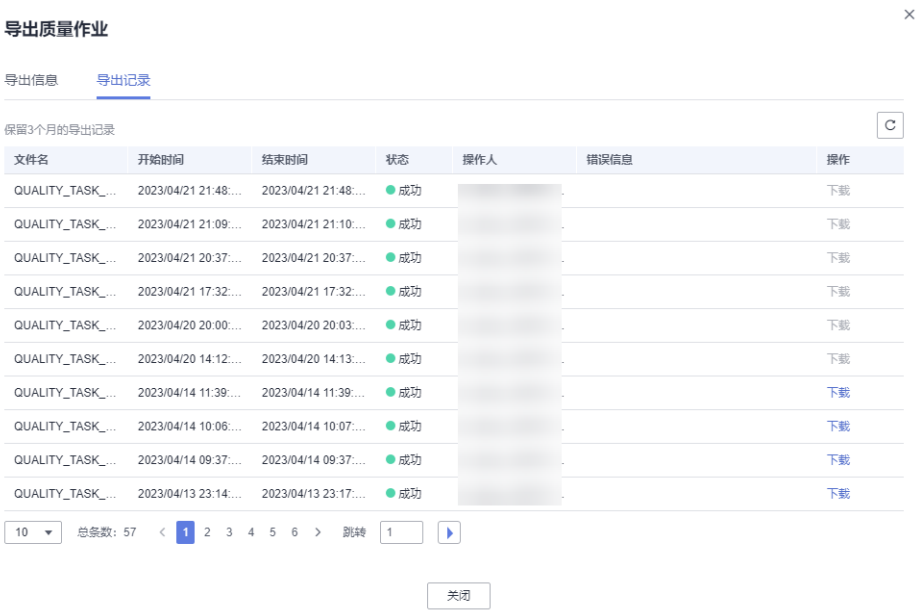
步骤2 在弹出“导出质量作业”对话框，单击“导出”。

图 10-23 导出全部质量作业



步骤3 切换到“导出记录”页签，可查看当前任务的导出结果。

图 10-24 导出记录



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4MB数据的文件。导入作业时，导入的单元格内容最大长度支持65534个字符。

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。

图 10-25 导入质量作业

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。

说明

如果选择覆盖，请在导入文件前，停止所有作业调度，否则调度中的作业会导致上传文件失败。

步骤3 单击“上传文件”，选择准备好的数据文件。

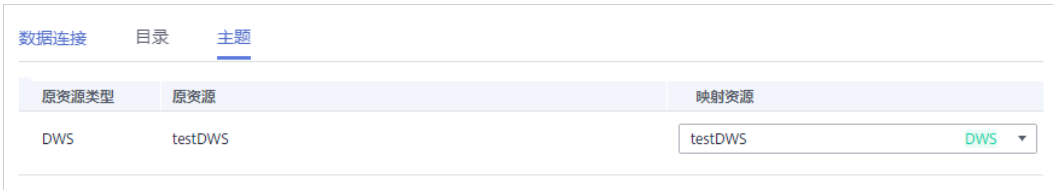
说明

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。如不选择，默认使用原映射资源信息。

图 10-26 配置映射资源信息



- 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

批量运行质量作业

系统支持批量运行质量作业，一次最多可批量运行200个质量作业。

步骤1 选择“数据质量监控 > 质量作业”，选择要批量运行的质量作业。

步骤2 单击“更多 > 批量运行”，即可完成质量作业的批量运行。

图 10-27 批量运行

步骤3 企业模式下，需要选择运行环境，系统支持可选择“开发环境”或“生产环境”。

步骤4 单击“确定”。

----结束

批量调度质量作业

系统支持批量调度质量作业，一次最多可批量调度200个质量作业。

步骤1 选择“数据质量监控 > 质量作业”，选择要批量调度的质量作业。

步骤2 单击“更多 > 启动调度”，即可完成质量作业的批量调度。

图 10-28 批量调度

----结束

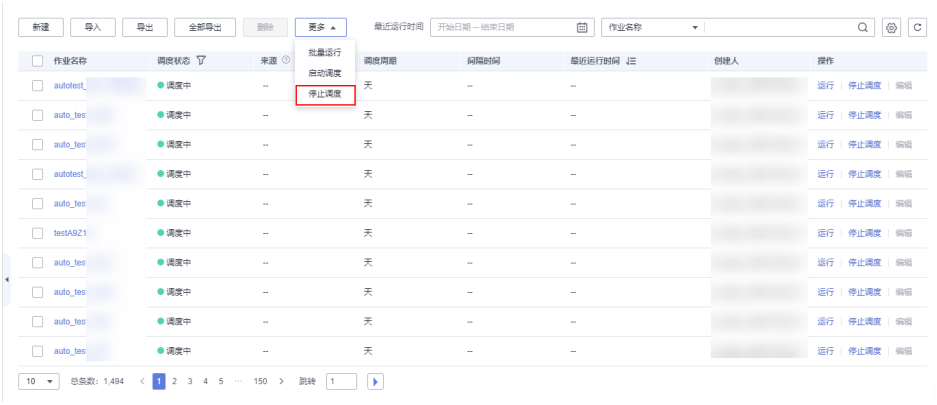
批量停止调度质量作业

系统支持批量停止调度质量作业，一次最多可批量停止200个质量作业。

步骤1 选择“数据质量监控 > 质量作业”，选择要批量停止调度的质量作业。

步骤2 单击“更多 > 停止调度”，即可完成质量作业的批量停止调度。

图 10-29 批量停止调度



----结束

批量停止运行质量作业

系统支持批量停止运行质量作业，一次最多可批量停止200个质量作业。

仅运行状态为“运行中”的质量作业可以停止。

步骤1 选择“数据质量监控 > 运维管理”，选择要批量停止的质量作业。

步骤2 单击“停止运行”，在弹出的“停止实例”界面中确认需要停止运行的质量作业实例，单击“是”，即可完成质量作业的批量停止运行。

图 10-30 批量停止运行

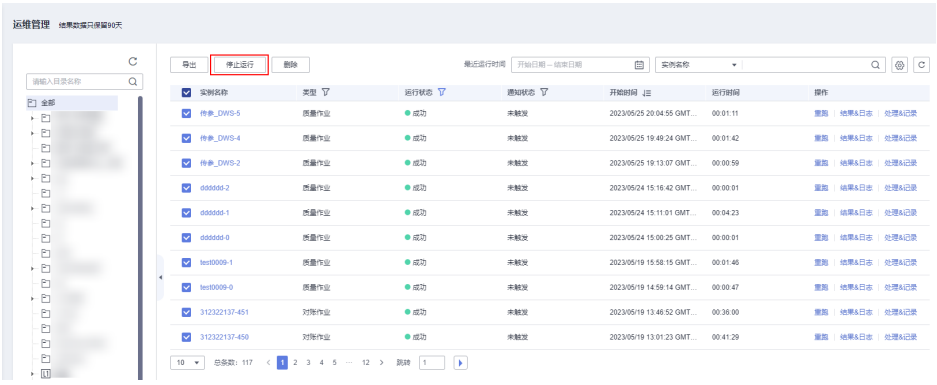


图 10-31 停止实例

----结束

10.2.4 新建数据对账作业

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。

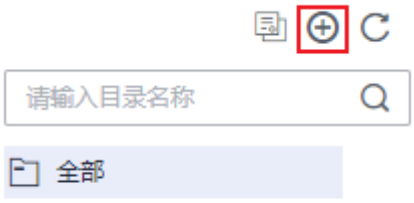
创建作业

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据质量”模块，进入数据质量页面。
2. （可选）选择“数据质量监控 > 对账作业”，新建目录。如果已存在可用的目录，可以不用新建目录。注意，规则模板、质量作业和对账作业的目录为同一目录，择一操作即可。

   当前系统支持“新建目录”和“同步主题为目录”两种方式：

   选择“新建目录”时，直接在目录处单击，输入目录名称，即可完成目录新建。直接新建目录的最大深度拓展为7层。

   图 10-32 新建目录

   选择“同步主题为目录”时，在目录处单击，即可将**数据架构处的主题**同步到目录中（仅支持同步“已发布”状态的主题）。同步后的主题目录与数据架构发布后主题一致，按照主题层级如L1、L2等进行展示。

   说明

   1. 直接新建的普通目录不受同步影响。但当普通目录名称与主题名称冲突时：
      - 主题首次同步名称冲突时，会将普通目录修改为主题目录属性，带上主题层级如L1、L2等进行展示。
      - 主题修改后再次同步名称冲突时，会出现同步失败的情况。
   2. 不支持变更自动同步。即数据架构处的主题或主题层级变更并发布后，需要手动再次单击才能同步到主题目录。

      特殊的，数据架构处的主题或主题层级删除后，手动同步后目录不会删除，仅去除主题目录属性，作为普通目录进行展示。
   3. 完成同步后，若有同步失败的主题，系统会自动弹出同步结果明细，可查看同步失败的主题名称。

图 10-33 同步主题为目录

3. 在“对账作业”页面，单击“新建”，在弹出的对话框中，参见**表10-14**配置相关参数。

表 10-14 配置作业参数

参数名	说明
作业名称	对账作业的名称。
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1024个字符。
标签	选择所需的标签。标签是在数据地图组件中定义的标签。如果未使用数据地图组件，则标签功能不生效。
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**（可选）新建目录**。
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。
超时时间	输入超时时间。输入值必须在5到1440之间。单位为分钟。该参数为空或者默认1440分钟时，超时时间系统默认为24小时，支持修改。

4. 单击“下一步”，进入规则配置页面。您需要单击规则卡片中的  ，然后参见**表10-15**配置数据对账规则。您也可选择添加对账规则。

图 10-34 打开对账作业规则配置

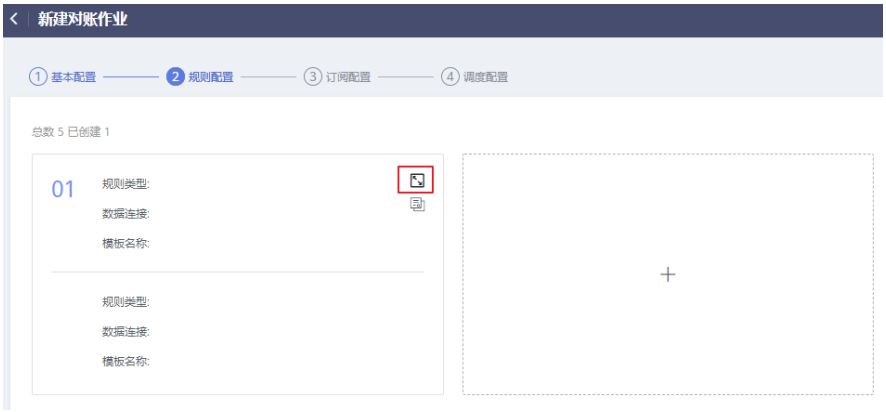


表 10-15 配置模板规则

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。描述信息长度不能超过1024个字符。
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、MRS Spark、DLI、ORACLE、RDS（MySQL、PostgreSQL）、Hetu、MRS Spark（Hudi）、MRS ClickHouse、DORIS。
		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。
		● 数据质量当前不支持MRS hive组件的严格模式 。

	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“自定义规则”，数据对象选择对应的数据库即可。
	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。

模块	参数名	说明
	默认参数值	自定义SQL可设置入参用于执行，SQL入参需要与参数默认值顺序匹配（数据质量单点执行时）。 **说明** 当通过数据开发任务调度质量算子的时候，优先使用数据开发中定义的参数值。
计算引擎	队列名称	选择运行对账作业的引擎。仅数据连接为DLI、Hive或Hetu类型时，此参数有效，输入队列名称。 当连接类型为Hetu，规则类型为除了库级以外的所有系统模板，自定义模板，自定义规则时，队列名称指的是Hetu引擎的资源队列名称。查看Hetu引擎的资源队列名称，需要登录MRS的FusionInsight Manager系统，单击左侧导航的HetuEngine，在基本信息区域，单击HSConsole WebUI链接，在计算实例列表中查看Hetu引擎的资源队列名称。
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10-10**。除去系统内置规则模板外，您也可关联在**新建数据质量规则**中新建的自定义模板。 当“规则类型”选择“字段级规则”，规则模板名称选择“正则表达式校验”或“正则表达式校验忽略Null”时，正则表达式的规则长度最大支持1024个字符。
	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对账作业时，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。
		说明
		配置where条件语句时，最前面需要加and，因为在SQL生成中需要进行语法的校验，否则会报语法错误。
		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：
		`and 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：
		`and time >= (date_trunc('hour', now()) - interval '24 h') and time <= (date_trunc('hour', now()))`

模块	参数名	说明
	参数默认值	当选择“条件扫描”时可填写。 请按照输入的where条件文本框中出现的参数名，依次填写默认参数值。 **说明** 参数默认值优先由数据开发传递，为空时可能会造成质量作业运行出错。 数据开发传递参数给数据质量后，作业运行完以后，通过“查看SQL”可以查看所传递的数据开发的参数以及参数值。

模块	参数名	说明
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 ● abs：绝对值 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“${2_1}”分别为通过告警参数配置的来源侧表和目的侧表的“总行数”，“\|\|”表示满足两个条件之一即会告警。 ● 需要配置来源侧表行数减去目的侧表行数的绝对值在除以来源侧表行数大于0.1时告警，则此处可设置为“abs(${1_1}-${2_1})/${1_1}>0.1”，其中“${1_1}”为通过告警参数配置的来源侧表“总行

模块	参数名	说明
		数”，“${2_1}”为通过告警参数配置的目的侧表“总行数”。
	告警参数	此参数来源于规则模板的输出结果。您可以单击界面显示的参数从而输入告警表达式中的告警参数，单击后系统会在“告警表达式”输入框给出参数的表达式。 例如“规则模板”为“表行数”时，单击告警参数“总行数”，在“告警表达式”输入框会显示为“${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 ● abs：绝对值 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服务主题，如**图10-35**。

图 10-35 订阅配置

说明

开启订阅配置后，每个满足通知类型的子作业都会发送通知。

如果开启告警，失败告警通知不需要单独配置，任务运行失败后会发送告警。

当前仅支持“短信”、“邮件”这两种协议的订阅终端订阅主题。

通知类型包含“触发告警”和“运行成功”两种。

打开“通知抑制”后，告警上报的通知策略可以进行配置，在最近N分钟以内，连续N次告警，则发送告警通知。最近时间可支持配置1~360分钟，连续次数可支持配置1~10次。

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10-16。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 10-16 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。

参数名	说明
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。

对账作业创建完成后，可以在作业里面进行查看，系统支持通过作业名称、创建人、最近运行时间进行筛选。同时，系统支持模糊搜索。

对账作业创建完成后，可以对该对账作业进行编辑、删除、运行、启动调度、停止调度等操作。

说明

单次调度模式不支持启动调度。

运行单个对账作业

系统支持运行单个对账作业。

步骤1 选择“数据质量监控 > 对账作业”，选择要运行的对账作业。

步骤2 单击“操作”列的“运行”。

步骤3 企业模式下，选择运行环境，系统支持可选择“开发环境”或“生产环境”。

步骤4 单击“确定”。

----结束

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。导出作业时，导出的单元格内容最大长度支持65534个字符。

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”，可将对账作业的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4M数据的文件。导入作业时，导出的单元格内容最大长度支持65534个字符。

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。

图 10-36 导入对账作业

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。

说明

如果选择覆盖，请在导入文件前，停止所有作业调度，否则调度中的作业会导致上传文件失败。

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、映射资源信息。如不选择，默认使用原映射资源信息。

图 10-37 配置映射资源信息

- 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”，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

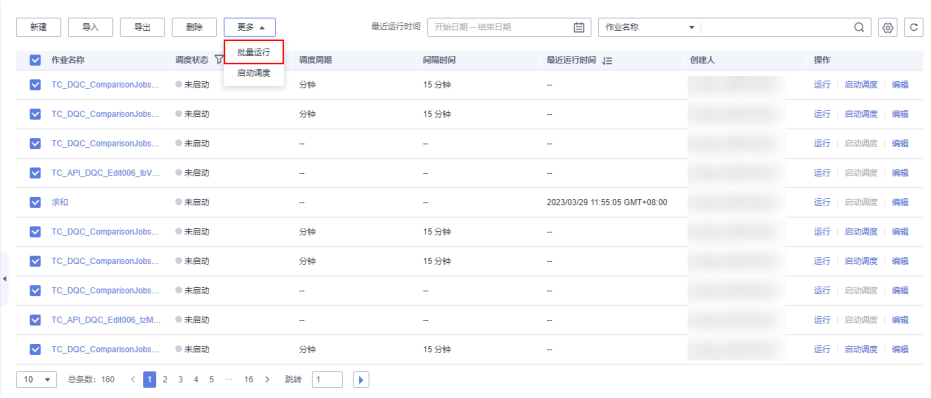
批量运行对账作业

系统支持批量运行对账作业，一次最多可批量运行200个对账作业。

步骤1 选择“数据质量监控 > 对账作业”，选择要批量运行的对账作业。

步骤2 单击“更多 > 批量运行”，即可完成对账作业的批量运行。

图 10-38 批量运行



步骤3 企业模式下，选择运行环境，系统支持可选择“开发环境”或“生产环境”。

步骤4 单击“确定”。

----结束

批量调度对账作业

系统支持批量调度对账作业，一次最多可批量调度200个对账作业。

步骤1 选择“数据质量监控 > 对账作业”，选择要批量调度的对账作业。

步骤2 单击“更多 > 启动调度”，即可完成对账作业的批量调度。

图 10-39 批量调度

----结束

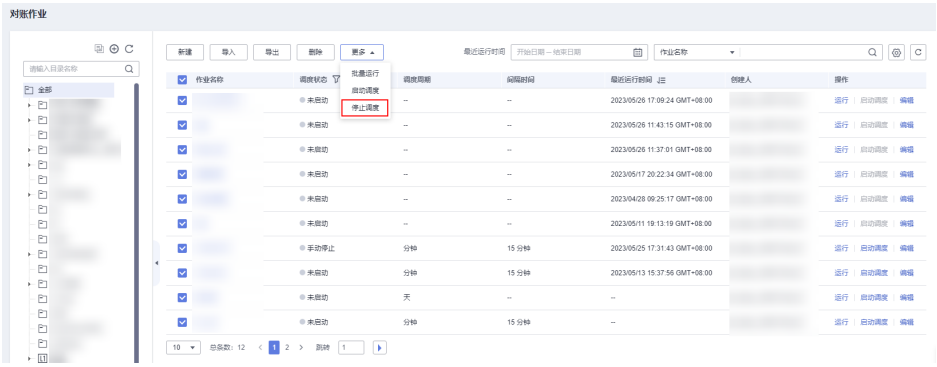
批量停止调度对账作业

系统支持批量停止调度对账作业，一次最多可批量停止200个对账作业。

步骤1 选择“数据质量监控 > 对账作业”，选择要批量停止调度的对账作业。

步骤2 单击“更多 > 停止调度”，即可完成对账作业的批量停止调度。

图 10-40 批量停止调度



----结束

批量停止运行对账作业

系统支持批量停止运行对账作业，一次最多可批量停止200个对账作业。

仅运行状态为“运行中”的对账作业可以停止。

步骤1 选择“数据质量监控 > 运维管理”，选择要批量停止的对账作业。

步骤2 单击“停止运行”，在弹出的“停止实例”界面中确认需要停止运行的对账作业实例，单击“是”，即可完成对账作业的批量停止运行。

图 10-41 批量停止运行

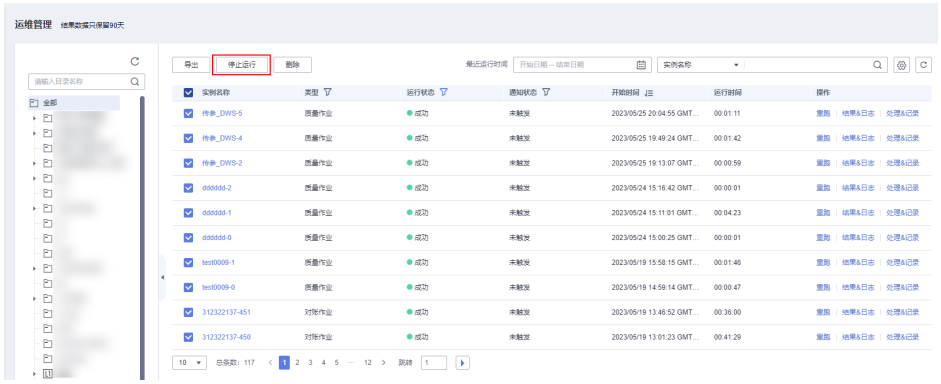


图 10-42 停止实例

----结束

10.2.5 查看作业实例

界面说明

介绍“数据质量监控 > 运维管理”页面中的区域和按键功能。

图 10-43 页面区域说明

表 10-17 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。

序号	区域	描述
2	规则实例列表	展示实例名称、类型、运行状态、运行结果等信息。
3	管理区域	可以对所选实例进行导出、删除、停止运行的操作。
4	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、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。
5	SQL并发数配置	单击SQL框，进入“单连接SQL并发数配置”页面，配置SQL并发数。输入值必须在10到1000之间。单击“确定”，完成配置。 **说明** 并发数是指单个数据连接下的SQL并发数，如果超出则等待排队执行。

表 10-18 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。 • 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。 • 超时：表示实例运行超时，状态显示为失败。
通知状态	展示实例通知状态，包含成功、失败和未触发。
操作人	展示实例的操作人。
创建时间	展示实例的创建时间。
开始时间	展示实例开始运行的时间。开始时间支持按照升序和降序进行排序。
运行时间	展示实例的运行时长。
结束时间	展示实例结束运行的时间。结束时间支持按照升序和降序进行排序。
处理人	展示实例的处理人。
重跑	再次运行规则实例。

菜单/按键	说明
结果&日志	详细展示作业实例的运行结果和日志。 ● 质量作业结果 质量作业运行结果中，支持查询每条规则的运行状态（包括正常和告警）。如果质量作业状态为告警，可查看该告警是由哪条规则触发的。 质量作业运行结果中，支持显示子作业运行状态，支持通过子作业名称和子作业运行状态进行过滤。 自定义SQL的运行结果展示最多最多300条数据，超出部分会自动截断。最多导出10000条数据。 ● 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。
更多 > 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。
更多 > 刷新作业状态	可以刷新作业的运行状态。

更多操作

- 导出

  勾选需要导出的作业实例名称，单击“导出”，弹出“导出实例运行结果”页面，再次单击“导出”，可以在“导出记录”页签查看导出实例的结果是否成功，可以下载导出成功的作业实例。
- 删除

  勾选需要删除的作业实例名称，单击“删除”，可以批量删除作业实例。
- 停止运行

  勾选需要停止运行的作业实例名称，单击“停止运行”，可以批量停止运行中的作业实例。
- 重跑

  选择需要重跑的作业实例名称，单击作业实例右侧“操作”列的“重跑”，可以重跑该作业实例。

10.2.6 查看数据质量报告

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。

说明

查看质量报告包含技术报告和业务报告。

技术报告的统计范围是依据质量作业的运行结果，包含数据连接、数据库、表名、评分等信息。

业务报告的统计范围是依据数据架构主题关联匹配的质量作业运行的结果，包含主题域、主题域分组、业务对象、表名、评分等信息。

查看技术报告数据质量评分

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10-19**。

表 10-19 对象评分计算公式

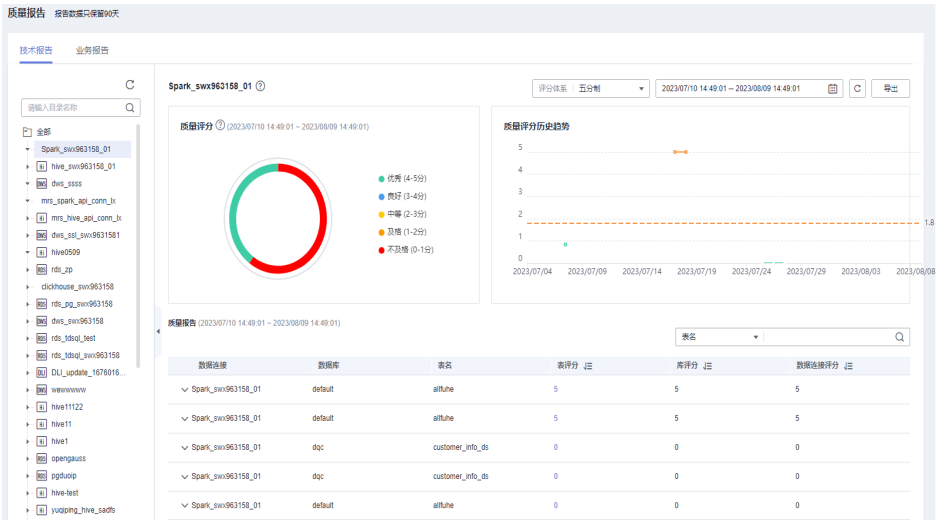
对象	评分计算公式
规则	创建质量作业时，作业关联的规则中结果说明列包含“比率”、“值率”的系统内置规则及用户自定义规则可以生成质量评分报告。 • 包含“比率”、“值率”的规则可以分为正向规则及反向规则，正向规则即比值越高，代表数据质量越好；反向规则即比值越高，则数据质量越差。 正向规则包含唯一值率、重复值率、合法比率规则，反向规则包含空值率规则。 • 正向规则评分=满足规则的数据行数/数据总行数*满分（5，10，100）。 • 反向规则评分=（1-满足规则的数据行数/数据总行数）*满分（5，10，100）。
数据表	表评分计算公式：∑(表关联的所有规则评分*规则权重)/∑规则权重
数据库	数据库下所有数据表评分的加权求平均值，即：∑数据库下所有数据表评分/表的数量。
数据连接	数据连接下所有数据库评分的加权平均值，即：∑数据连接下所有数据库的评分/数据库的数量。

步骤1 在DataArts Studio控制台首页，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 选择“数据质量监控 > 质量报告”。

步骤3 在“技术报告”页签，选择数据连接及时间段，系统支持查询报告的最大时间范围限制为30天，如**图10-44**所示。

图 10-44 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，如图10-45所示。

图 10-45 查看规则评分

说明

规则名称为运行实例名称，如果作业被运行多次，取最新时间运行实例的结果。如果同一运行实例中，有多个子实例检验该表，则每个子实例一条记录。

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10-46所示。

图 10-46 表关联规则评分界面

----结束

查看业务报告业务质量评分

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10-20**。

表 10-20 对象评分计算公式

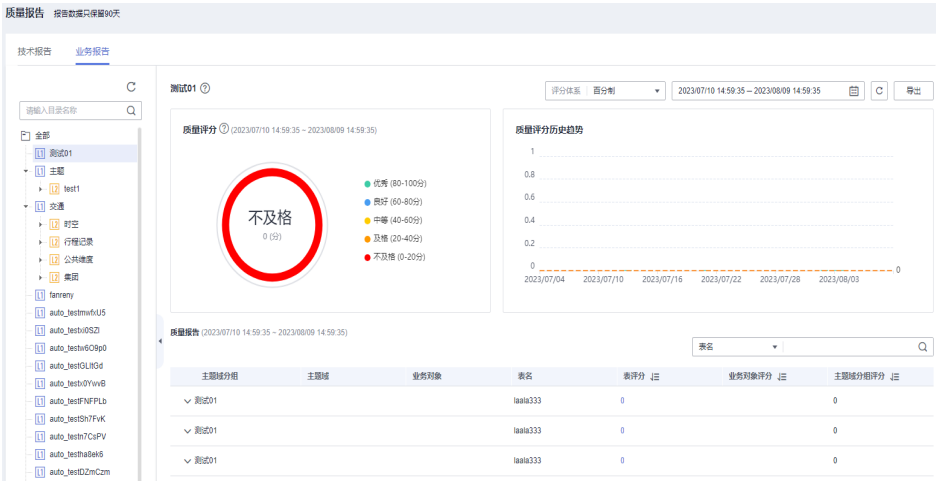
对象	评分计算公式
规则	创建质量作业时，包含“比率”、“值率”的系统内置规则及用户自定义规则可以生成质量评分报告。 • 包含“比率”、“值率”的规则可以分为正向规则及反向规则，正向规则即比值越高，代表数据质量越好；反向规则即比值越高，则数据质量越差。 正向规则包含唯一值率、重复值率、合法比率规则，反向规则包含空值率规则。 • 正向规则评分=满足规则的数据行数/数据总行数*满分（5，10，100）。 • 反向规则评分=（1-满足规则的数据行数/数据总行数）*满分（5，10，100）。 • 当表为空，即总行数为0时，正向规则评分固定为满分，反向评分固定为0分。
表	表评分计算公式：∑(表关联的所有规则评分*规则权重)/∑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：∑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：∑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：∑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，选择对应工作空间的“数据质量”模块，进入数据质量页面。

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10-47所示。

图 10-47 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10-48所示。

图 10-48 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

导出质量报告

您可以通过以下两种方式导出质量报告：

- 若使用局点有OBS服务，系统默认导出到关联的OBS桶中。

  说明

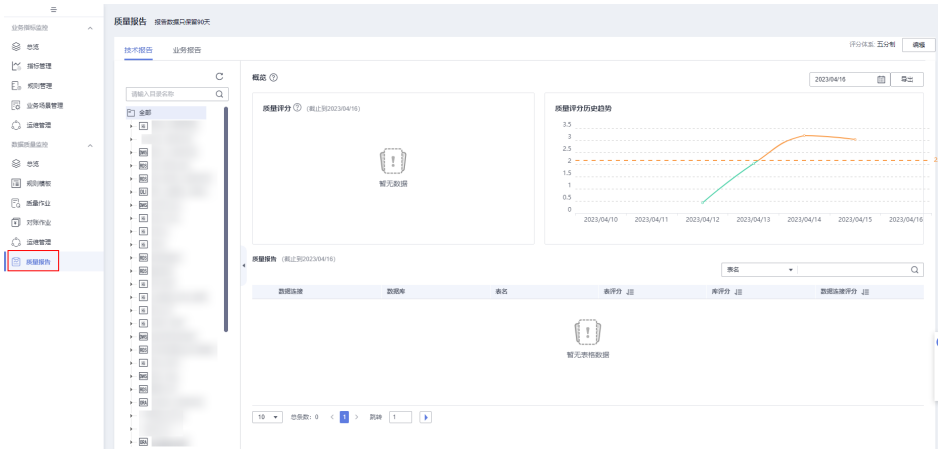
  - 由于质量报告数据量较大，单个导出文件字段条数最多为2000条，因此OBS桶里或许会有多个导出文件。
  - 导出的报告仅限当前工作空间内。

- 若使用局点没有OBS服务，系统默认导出到本地。

步骤1 在DataArts Studio控制台首页，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 选择“数据质量监控 > 质量报告”。

图 10-49 质量报告页面



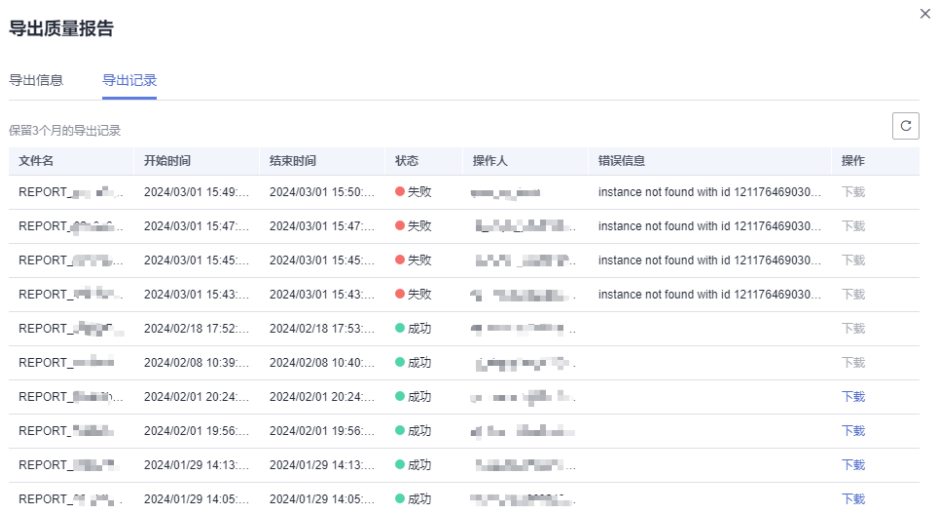
步骤3 单击页面右上角的“导出”按钮，将质量报告导出。

图 10-50 导出

图 10-51 导出到 OBS 桶

步骤4 可在导出记录中查看导出结果，单击“下载”可以下载数据质量报告。如果导出的报告文件过大，系统也支持直接下载大文件。

图 10-52 导出记录



----结束

立即刷新

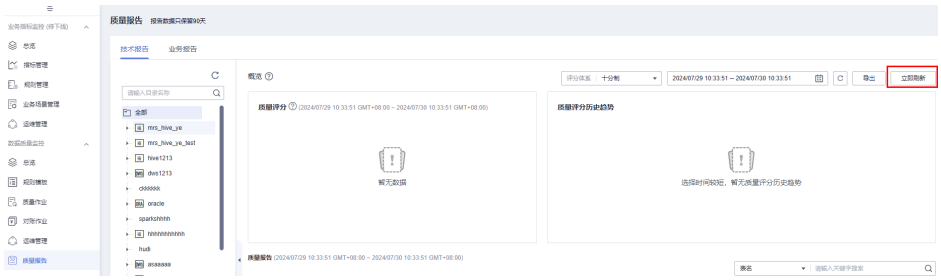
在质量作业和对账作业运行完毕后，通过立即刷新功能，用户可以立即获得零点到当前时间的数据质量报告临时数据。到第二天凌晨，质量报告的调度任务开始执行，此时生成的数据是前一天的全量数据质量报告。

步骤1 在DataArts Studio控制台首页，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 选择“数据质量监控 > 质量报告”。

步骤3 单击页面右上角的“立即刷新”按钮，页面将展示零点到当前时间的临时数据，用户可以立即获得当天的数据质量报告数据。

图 10-53 立即刷新



----结束

10.3 使用教程

10.3.1 新建一个业务场景

场景说明

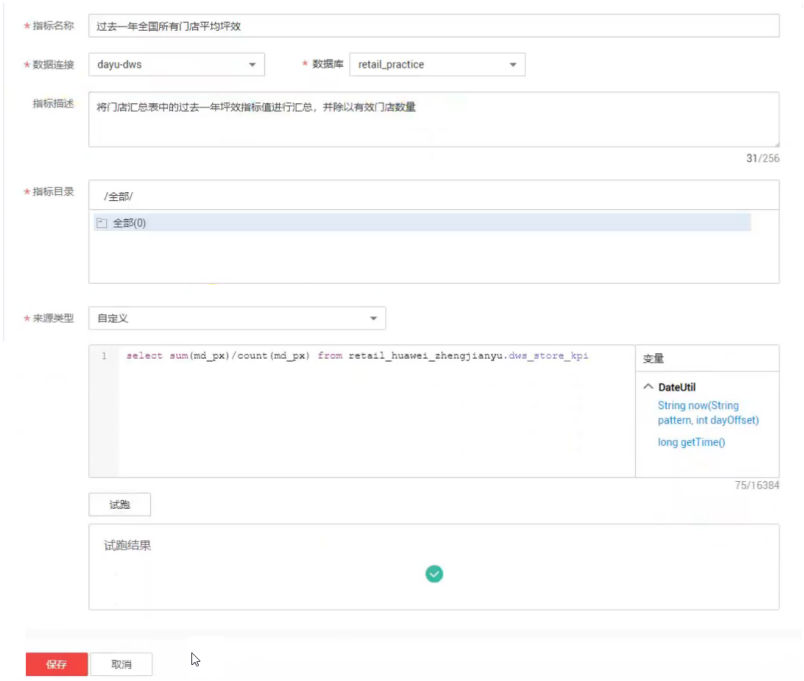
业务场景用于监控业务指标。本例以新建一个业务场景为例，介绍如何使用业务指标监控功能。

操作步骤

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据质量”模块，进入数据质量页面。

步骤2 新建业务指标。

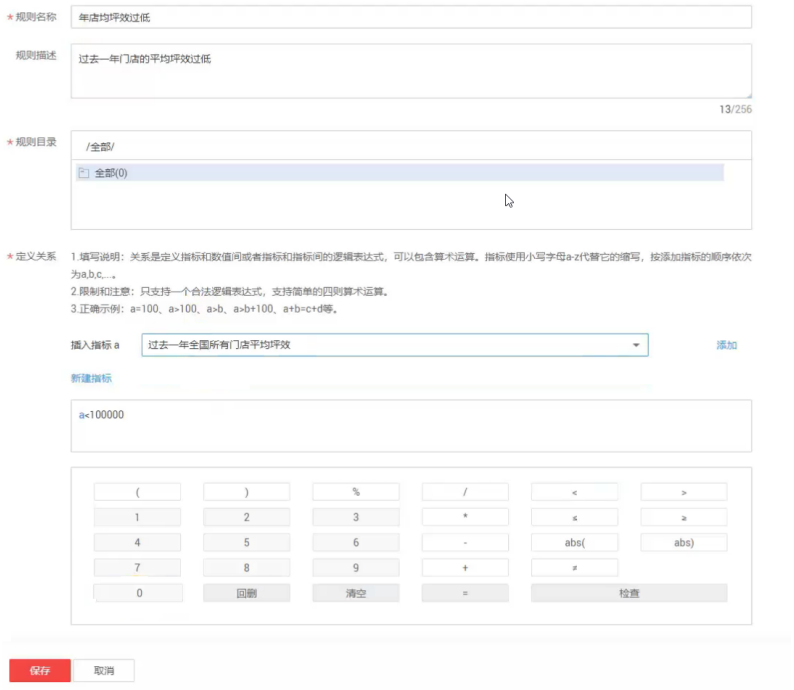
1. 单击左侧导航“指标管理”。
2. 单击页面上方的“新建”，如下图所示。



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。



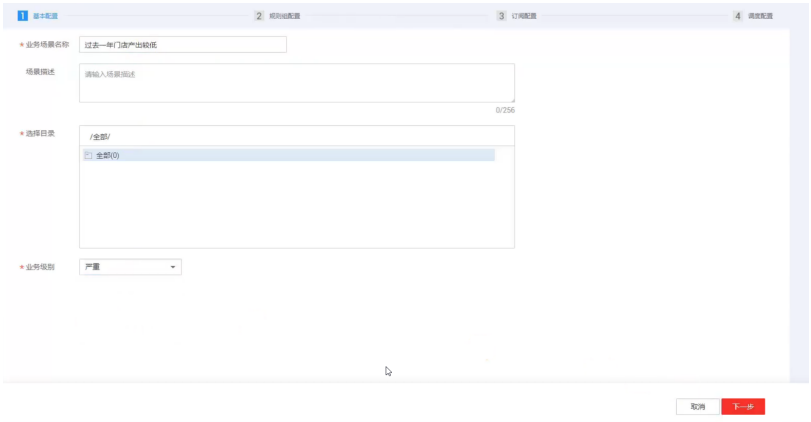
4. 单击“保存”。

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。

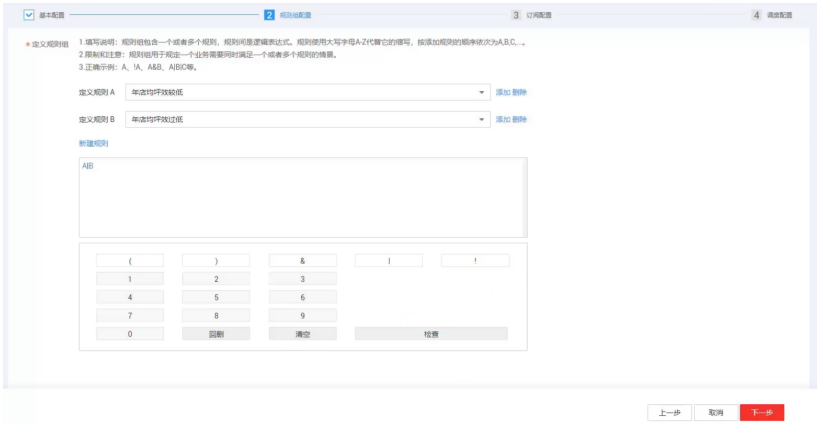
7. 单击“保存”。

步骤4 新建业务场景。

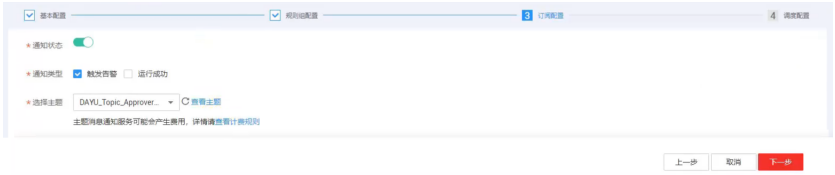
1. 单击左侧导航“业务场景管理”。
2. 单击页面上方的“新建”，输入场景的基本配置参数，如下图所示。



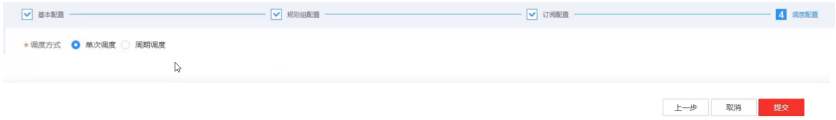
3. 单击“下一步”，输入规则组的配置参数，如下图所示。



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

10.3.2 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。

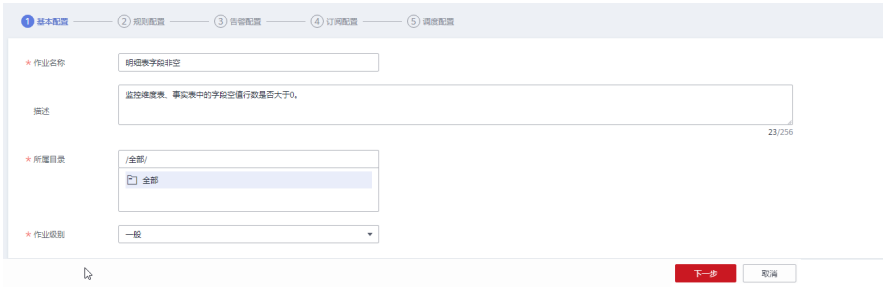
1. 单击左侧导航“规则模板”，默认展示系统自定义的规则。数据质量的规则包含6个维度，分别是：完整性、唯一性、及时性、有效性、准确性、一致性。
2. **可选:** 单击“新建”，可自定义创建规则。

说明

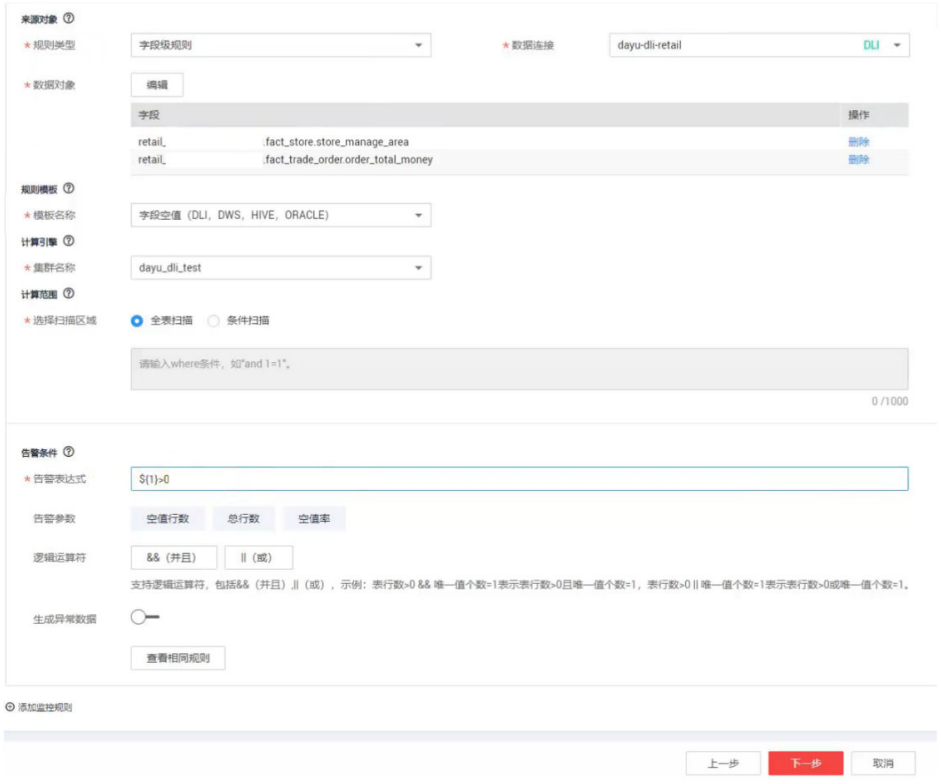
本例使用系统自定义的规则即可。

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。

图 10-54 技术报告

3. 单击“业务报告”页签，查看业务报告，如下图所示。

图 10-55 业务报告

----结束

10.3.3 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”。

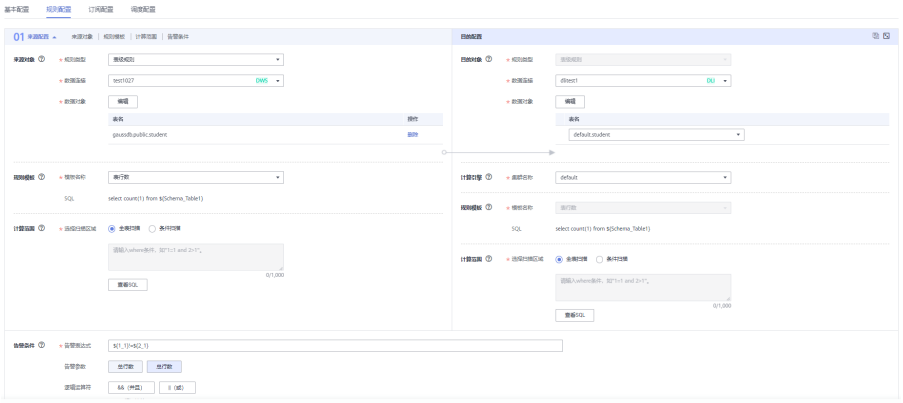
2. 创建DWS数据连接。在DataArts Studio管理中心模块，单击创建数据连接，数据连接类型选择“数据仓库服务（DWS）”，输入数据连接名称，设置其他参数，如下图所示，单击“测试”，提示连接成功，单击“确定”。

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。

图 10-56 配置基本信息

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

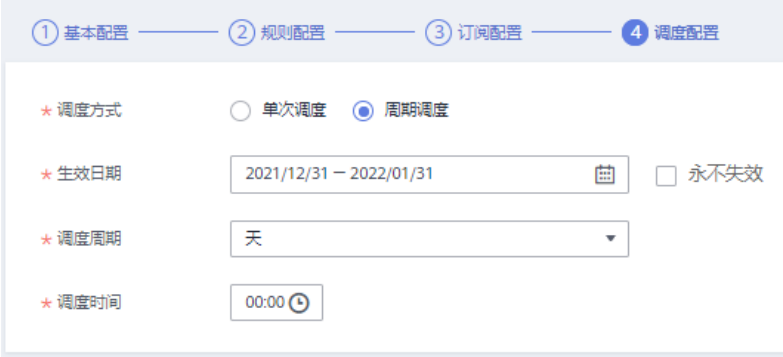
- 需要分别配置源端和目的端的信息。配置源端连接请参见**DWS数据连接参数说明**，配置目的端连接请参见**DLI数据连接参数说明**。
- 配置告警条件，其中单击左侧的表行数（${1_1}）表示左侧源端选中表的行数，单击右侧表行数（${2_1}）表示目的端表行数。此处配置告警条件为${1_1}!=${2_1}，表示当左侧表行数与右侧表行数不一致时，触发报警并显示报警状态。

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。

11 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。

11.1 查看工作空间数据地图

11.1.1 查看工作空间内的数据资产

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。

11.1.2 查看资产总览

通过总览，可以查看资产总览及资产报告。

- 资产总览可展示业务资产、技术资产和指标资产的情况。
  - 业务资产来自于数据架构组件中定义并发布过的逻辑实体与数据表，资产总览展示业务对象、逻辑实体、业务属性的数量及其详情。
  - 技术资产来自于数据连接和元数据采集任务，资产总览展示数据库、数据表、数据量的数量及其详情。
  - 指标资产来自于数据架构组件中定义并发布过的业务指标，资产总览展示业务指标及其详情。
- 资产报告可展示逻辑实体、数据表、资产关联、资产容量、标签、密级、以及TOP100的表容量、表行数、桶容量等内容。

约束限制

- 业务资产和指标资产来自于数据架构组件，会随数据架构同步的数据更新，但不支持随之删除。如需删除需要在数据目录中定位到资产后手动删除。
- 技术资产中的数据连接信息来自于管理中心的数据连接，会随管理中心同步的数据更新，但不支持随之删除。如需删除需要在数据目录中定位到资产后手动删除。
- 技术资产中的库表列等信息来自于元数据采集任务，是否更新和自动删除取决于元数据采集任务的参数配置，详情请参见**配置元数据采集任务**。
- 技术资产中的数据血缘关系更新依赖于作业调度，数据血缘关系是基于最新的作业调度实例产生的。需要注意的是，数据血缘关系删除需要通过删除作业或删除作业元数据的方式进行，仅将作业停止调度不会触发血缘关系的删除。

前提条件

- 已在数据架构组件中定义并发布过的逻辑实体与数据表、业务指标。
- 已配置元数据采集任务并成功运行，如何创建采集任务请参见**创建采集任务**。

资产总览

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据目录”模块，进入数据目录页面。
3. 选择“数据地图 > 总览”，默认进入“资产总览”页面。

   图 11-1 资产总览

4. 单击“业务资产”，查看业务资产情况。

   业务资产来自于数据架构组件中定义并发布过的逻辑实体与数据表，资产总览展示业务对象、逻辑实体、业务属性的数量及其详情。
5. 单击“技术资产”，查看技术资产情况。

   技术资产来自于数据连接和元数据采集任务，资产总览展示数据库、数据表、数据量的数量及其详情。
6. 单击“指标资产”，查看指标资产情况。

   指标资产来自于数据架构组件中定义并发布过的业务指标，资产总览展示业务指标及其详情。

资产报告

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据目录”模块，进入数据目录页面。
2. 选择“数据地图 > 总览”，单击并进入“资产报告”页面。

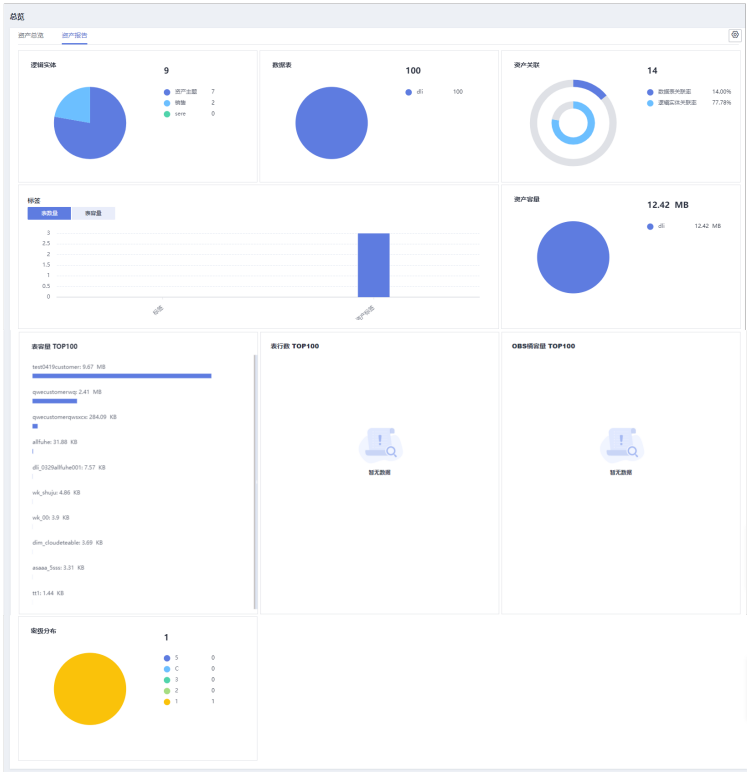
3. 首次进入“资产报告”页面，需要配置资产报告任务。单击右上方的配置图标，弹出配置窗口。

   依次选择生效时间、调度周期和调度具体时间，系统将按配置的时间调度运行资产报告任务，更新资产报告内容。

图 11-2 配置资产报告任务

4. 系统调度运行资产报告任务后，重新进入“资产报告”页面，可查看逻辑实体、数据表、资产关联、资产容量、标签、密级、以及TOP100的表容量、表行数、桶容量等资产内容。

图 11-3 资产报告



11.1.3 查看数据资产

通过数据目录可以对各类资产进行搜索、过滤、查看详情等操作。

- 业务资产来自于数据架构组件中定义并发布过的逻辑实体与数据表。
- 技术资产来自于数据连接和元数据采集任务，其中的数据连接来源于管理中心的数据连接，库表列等来源于数据目录的元数据采集任务。
- 指标资产来自于数据架构组件中定义并发布过的业务指标。

约束限制

- 业务资产和指标资产来自于数据架构组件，会随数据架构同步的数据更新，但不支持随之删除。如需删除需要在数据目录中定位到资产后手动删除。
- 技术资产中的数据连接信息来自于管理中心的数据连接，会随管理中心同步的数据更新，但不支持随之删除。如需删除需要在数据目录中定位到资产后手动删除。
- 技术资产中的库表列等信息来自于元数据采集任务，是否更新和自动删除取决于元数据采集任务的参数配置，详情请参见**配置元数据采集任务**。
- 技术资产中的数据血缘关系更新依赖于作业调度，数据血缘关系是基于最新的作业调度实例产生的。需要注意的是，数据血缘关系删除需要通过删除作业或删除作业元数据的方式进行，仅将作业停止调度不会触发血缘关系的删除。

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，选择对应工作空间的“数据目录”模块，进入数据目录页面。
2. 选择“数据地图 > 数据目录”，并根据需要进入“业务资产”、“技术资产”或“指标资产”页签。
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，搜索范围限定在“业务资产”、“技术资产”或“指标资产”页签内，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

对于技术资产搜索结果，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。
- 标签：数据资产所包含的标签，标签来自于数据目录中配置的标签数据，详见**管理资产标签**。
- 分类：数据资产所属分类，分类来自于数据目录中的分类数据。

  在已上线数据安全组件的区域，数据目录中的数据地图能力由数据地图组件提供，数据安全及数据权限能力由数据安全组件提供，数据目录中的相关能力不再演进。如果已具备数据安全和数据地图组件，数据目录中的相关能力会随之下

线，不再支持在数据目录中新建分类和为资产配置分类。在此情况下，您可以通过数据安全和数据地图组件新建分类并为资产配置分类，详见**定义数据分类**。

- 密级：数据资产所属密级，密级来自于数据安全组件中的密级数据。

  在已上线数据安全组件的区域，数据目录中的数据地图能力由数据地图组件提供，数据安全及数据权限能力由数据安全组件提供，数据目录中的相关能力不再演进。如果已具备数据安全和数据地图组件，数据目录中的相关能力会随之下线，不再支持在数据目录中新建密级和为资产配置密级。在此情况下，您可以通过数据安全和数据地图组件新建密级并为资产配置密级，详见**定义数据密级**。

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

本文以查看技术资产中的数据表详情为例进行说明。

步骤1 在技术资产搜索结果列表，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。

说明

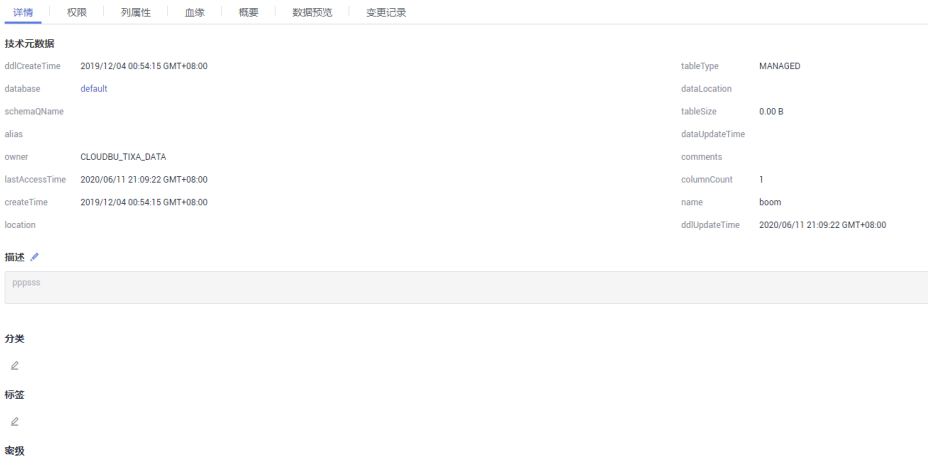
标签、分类和密级的来源分别如下：

- 标签：数据资产所包含的标签，标签来自于数据目录中配置的标签数据，详见**管理资产标签**。
- 分类：数据资产所属分类，分类来自于数据目录中的分类数据。

  在已上线数据安全组件的区域，数据目录中的数据地图能力由数据地图组件提供，数据安全及数据权限能力由数据安全组件提供，数据目录中的相关能力不再演进。如果已具备数据安全和数据地图组件，数据目录中的相关能力会随之下线，不再支持在数据目录中新建分类和为资产配置分类。在此情况下，您可以通过数据安全和数据地图组件新建分类并为资产配置分类，详见**定义数据分类**。
- 密级：数据资产所属密级，密级来自于数据安全组件中的密级数据。

  在已上线数据安全组件的区域，数据目录中的数据地图能力由数据地图组件提供，数据安全及数据权限能力由数据安全组件提供，数据目录中的相关能力不再演进。如果已具备数据安全和数据地图组件，数据目录中的相关能力会随之下线，不再支持在数据目录中新建密级和为资产配置密级。在此情况下，您可以通过数据安全和数据地图组件新建密级并为资产配置密级，详见**定义数据密级**。

图 11-4 查看详情



步骤3 在“权限”页签，可申请数据表权限或给其他用户授权。

图 11-5 权限页签详情

步骤4 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 11-6 管理列属性

步骤5 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**通过数据目录查看数据血缘关系**。数据开发作业配置了支持自动血缘的节点或手动配置节点的血缘关系后，作业执行时可以自动解析，在数据目录中展示数据血缘。

步骤6 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，概要采样方式以**元数据采集**任务配置为准）。

单击“更新”，可更新概要信息。

步骤7 在“数据预览”页签，预览当前表的业务数据。根据列的分类信息，支持对预览数据根据**配置脱敏策略（待下线）**的设置进行实时脱敏。

- 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。

步骤8 在“变更记录”页签，查看数据表变更详情。

----结束

11.1.4 管理资产标签

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。

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 11-1 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
来源	请参见**新建数据分类（待下线）**	请参见**管理资产标签**

管理标签

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据目录”模块，进入数据目录页面。
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，选择对应工作空间的“数据目录”模块，进入数据目录页面。
2. 选择“数据地图 > 数据目录”，并进入“技术资产”页签。
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 11-7 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**。

11.2 配置数据访问权限（待下线）

11.2.1 数据权限简介（待下线）

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。

须知

在已上线数据安全组件的区域，数据目录中的数据权限功能已由数据安全组件提供，不再作为数据目录组件能力。当前数据目录中的数据权限功能仅限于存量用户使用。

数据安全组件当前在中国-香港、亚太-新加坡、亚太-曼谷、亚太-雅加达、拉美-圣地亚哥、拉美-圣保罗一、非洲-约翰内斯堡和土耳其-伊斯坦布尔区域部署上线。

数据权限模块包含数据目录权限、数据表权限和审批中心三大子模块。具备的功能如下所示：

- 权限自助申请：用户可以选择自己需要权限的数据表，在线上快速发起申请。
- 权限审计：管理员可以快速方便地查看数据库表权限对应人员，进行审计管理。
- 权限回收/交还：管理员可以通过用户权限管理及时回收用户权限，用户也可以主动交还不再需要的权限。
- 权限审批管理：提供可视化、流程化的管理授权机制，以及对审批流程进行事后追溯。

11.2.2 配置数据目录权限（待下线）

本章节主要介绍数据目录权限管理。

须知

在已上线数据安全组件的区域，数据目录中的数据权限功能已由数据安全组件提供，不再作为数据目录组件能力。当前数据目录中的数据权限功能仅限于存量用户使用。

数据安全组件当前在中国-香港、亚太-新加坡、亚太-曼谷、亚太-雅加达、拉美-圣地亚哥、拉美-圣保罗一、非洲-约翰内斯堡和土耳其-伊斯坦布尔区域部署上线。

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。

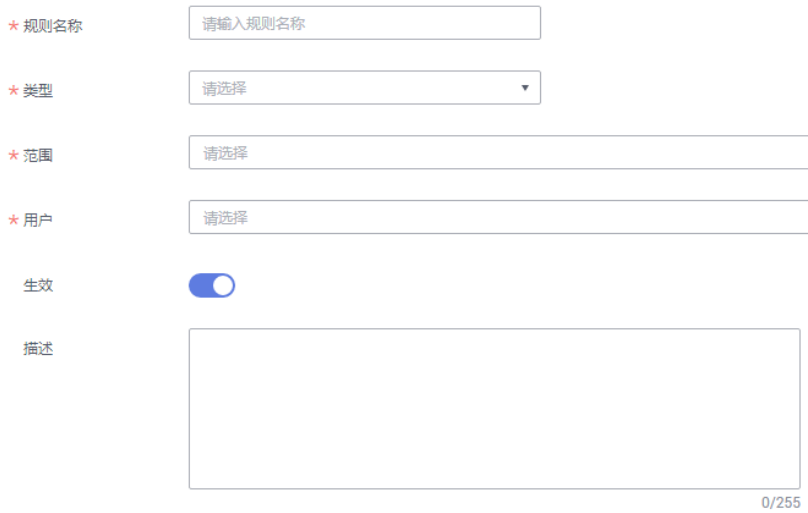
管理数据目录权限规则

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据目录”模块，进入数据目录页面。
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 11-8 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。

11.2.3 配置数据表权限（待下线）

须知

在已上线数据安全组件的区域，数据目录中的数据权限功能已由数据安全组件提供，不再作为数据目录组件能力。当前数据目录中的数据权限功能仅限于存量用户使用。

数据安全组件当前在中国-香港、亚太-新加坡、亚太-曼谷、亚太-雅加达、拉美-圣地亚哥、拉美-圣保罗一、非洲-约翰内斯堡和土耳其-伊斯坦布尔区域部署上线。

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

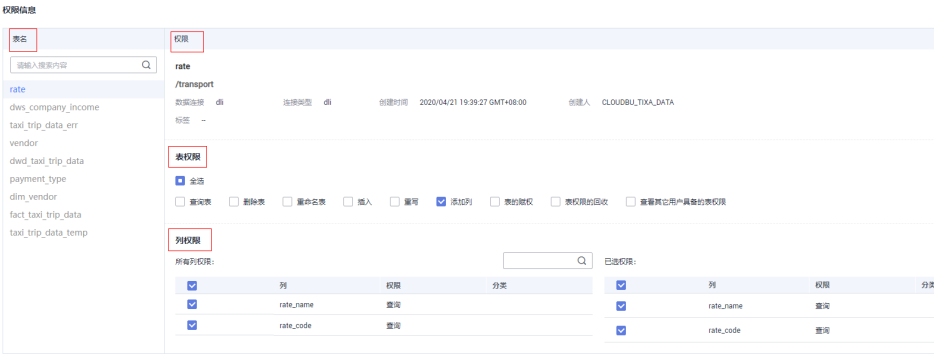
说明

- 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，选择对应工作空间的“数据目录”模块，进入数据目录页面。
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 11-9 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，选择对应工作空间的“数据目录”模块，进入数据目录页面。
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 11-10 管理表权限

审计用户权限

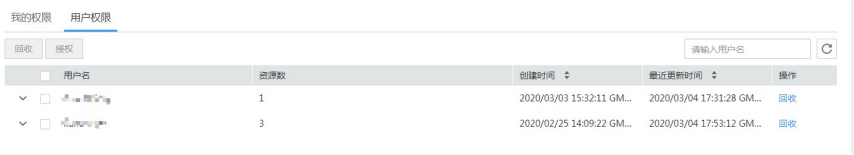
管理员可在“用户权限”页面查看同一工作空间内，分别有哪些账号拥有表和字段的权限，并可回收不必要的表和字段的权限，也可对用户进行批量授权。

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。

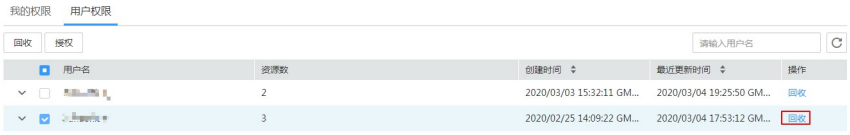
- 查看拥有表权限的账号和对应的资产列表
  选择“数据表权限 > 用户权限”，查看同一工作空间内，已申请表权限的账号。

图 11-11 查看拥有表权限的账号



- 回收用户的资产权限
  - 选择“数据表权限 > 用户权限”，单击账号后操作列的“回收”，可回收该账号所有的资产权限。
  - 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 11-12 回收用户的资产权限



- 对用户授权

图 11-13 授权

- 在资产上管理用户的权限

  选择“数据表权限 > 用户权限”，单击账号前的下拉列表，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 11-14 基于资产管理用户权限

11.2.4 管理审批中心（待下线）

须知

在已上线数据安全组件的区域，数据目录中的数据权限功能已由数据安全组件提供，不再作为数据目录组件能力。当前数据目录中的数据权限功能仅限于存量用户使用。

数据安全组件当前在中国-香港、亚太-新加坡、亚太-曼谷、亚太-雅加达、拉美-圣地亚哥、拉美-圣保罗一、非洲-约翰内斯堡和土耳其-伊斯坦布尔区域部署上线。

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11-15。审批人数据来源于工作空间中添加的人。

  图 11-15 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。

11.3 配置数据安全策略（待下线）

11.3.1 数据安全简介（待下线）

须知

在已上线数据安全组件的区域，数据目录中的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。当前数据目录中的数据安全功能仅限于存量用户使用。

数据安全组件当前在中国-香港、亚太-新加坡、亚太-曼谷、亚太-雅加达、拉美-圣地亚哥、拉美-圣保罗一、非洲-约翰内斯堡和土耳其-伊斯坦布尔区域部署上线。

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。

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。

11.3.2 新建数据密级（待下线）

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。

须知

在已上线数据安全组件的区域，数据目录中的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。当前数据目录中的数据安全功能仅限于存量用户使用。

数据安全组件当前在中国-香港、亚太-新加坡、亚太-曼谷、亚太-雅加达、拉美-圣地亚哥、拉美-圣保罗一、非洲-约翰内斯堡和土耳其-伊斯坦布尔区域部署上线。

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，单击“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。
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）。

11.3.3 新建数据分类（待下线）

本章主要介绍如何创建数据分类规则。

须知

在已上线数据安全组件的区域，数据目录中的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。当前数据目录中的数据安全功能仅限于存量用户使用。

数据安全组件当前在中国-香港、亚太-新加坡、亚太-曼谷、亚太-雅加达、拉美-圣地亚哥、拉美-圣保罗一、非洲-约翰内斯堡和土耳其-伊斯坦布尔区域部署上线。

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**新建数据密级（待下线）**。

新建分类规则

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“数据目录”模块，进入数据目录页面。
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 11-16 配置分类规则

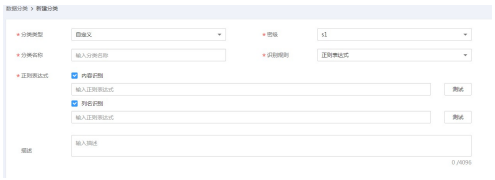


表 11-2 配置分类规则参数说明

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**新建数据密级（待下线）**。
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** 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。

配置	说明
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，单击“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。
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11-3**，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 11-3 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。
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。

11.3.4 配置脱敏策略（待下线）

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。

须知

在已上线数据安全组件的区域，数据目录中的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。当前数据目录中的数据安全功能仅限于存量用户使用。

数据安全组件当前在中国-香港、亚太-新加坡、亚太-曼谷、亚太-雅加达、拉美-圣地亚哥、拉美-圣保罗一、非洲-约翰内斯堡和土耳其-伊斯坦布尔区域部署上线。

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**新建数据分类（待下线）**。
- 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

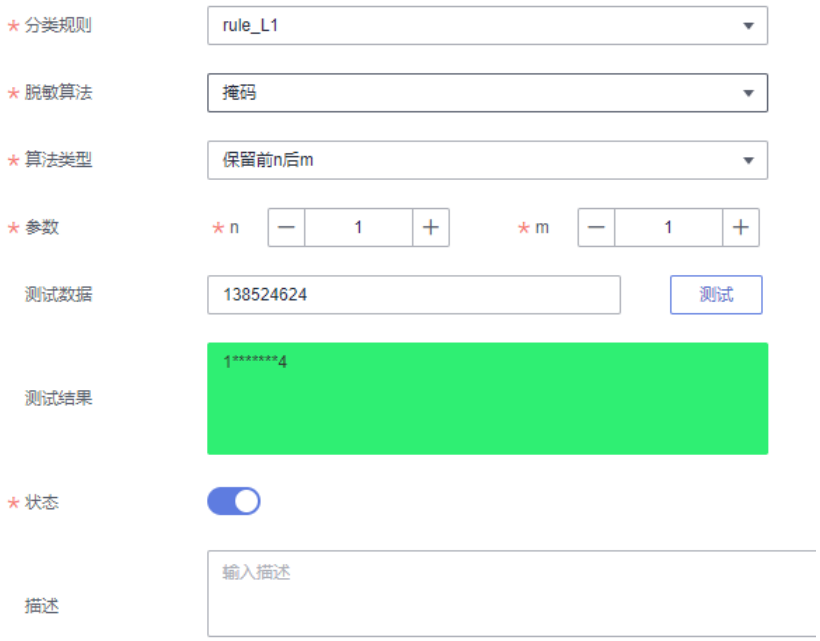
1. 在DataArts Studio控制台首页，选择实例，单击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

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 11-17 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，单击“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。
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。

11.4 采集数据源的元数据

11.4.1 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其所关心的数据，用于指导其进行数据管理和开发工作，提高工作效率。

在DataArts Studio中，元数据是数据的描述数据，可以为数据说明其属性（数据连接、类型、名称、大小等），或其相关数据（位于拥有者、标签、分类、密级等）。

元数据按用途的不同，可以分为两类：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。在DataArts Studio中，技术元数据即为技术资产，显示数据库、数据表、数据量的数量及其详情。
- 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。在DataArts Studio中，业务元数据包含业务资产和指标资产，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，指标资产显示业务指标及其详情。

DataArts Studio中的技术元数据来源于元数据采集任务，您需要在创建并运行元数据采集任务后才能在数据地图中查看元数据。

11.4.2 配置元数据采集任务

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。

约束与限制

- 当元数据采集任务未指定采集范围时，默认采集该数据连接下的所有数据表/文件。采集任务运行完成后，如果该数据连接下有新增数据表/文件，则需再次运行元数据采集任务，才能采集到新增数据表/文件的元数据。
- Oracle元数据采集前，需要确保数据连接中的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。详见**ORACLE数据连接参数说明**中的用户授权指导。
- 受MRS集群限制，默认情况下元数据采集任务无法直接采集到Hive分区表的元数据。

  如果需要采集Hive分区表的元数据，需要在MRS集群内的**HiveServer（角色）-> 自定义**下的“hive.server.customized.configs”参数值中新增名称**hive-ext.display.desc.statistic.stats**，且值为**true**。详情请参见**配置MRS集群Hive分区表支持元数据采集**。

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和ORACLE类型的数据源，首先需要在管理中心创建数据连接。如需采集其他数据源（如OBS、CSS、GES等）元数据，无需在管理中心创建数据连接。
- 采集Hudi元数据前，需要先在Hudi表开启“**同步hive表配置**”，然后才能通过采集MRS Hive元数据的方式采集Hudi表的元数据。
- 如果需要采集Hive分区表的元数据，需要在MRS集群内的**HiveServer（角色）-> 自定义**下的“hive.server.customized.configs”参数值中新增名称**hive-**

ext.display.desc.statistic.stats，且值为**true**。详情请参见**配置MRS集群Hive分区表支持元数据采集**。

新增采集任务

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据目录”模块，进入数据目录页面。
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11-18**创建进行。

图 11-18 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11-19**所示的两个入口。

图 11-19 新建采集任务入口

   a. 配置基本参数，参考**表11-4**。

表 11-4 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。

参数名	说明
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11-18。

b. 配置数据源信息，参考表11-5。

表 11-5 数据源信息参数说明

参数名		说明
数据连接类型		从下拉列表中选择数据连接类型。 **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和ORACLE类型的数据源，首先需要在管理中心创建数据连接。如需采集其他数据源（如OBS、CSS、GES等）元数据，无需在管理中心创建数据连接。
• DWS • DLI • MRS HBase • MRS Hive • ORACLE • RDS	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”，创建新的数据连接。
	数据库（或数据库和schema、命名空间）	呈现待采集的数据库（或数据库和schema、命名空间）和数据表。 • 单击数据库（或数据库和schema、命名空间）后的“设置”，设置采集任务扫描的数据库（或数据库和schema、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数据库（或数据库和schema、命名空间）。 • 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（或数据库和schema、命名空间）下的所有数据表。 • 当数据库（或数据库和schema、命名空间）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（或数据库和schema、命名空间）、数据表进行修改。
	数据表	
CSS	选择集群	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“刷新”，选择新建的CSS集群即可。
	绑定Agent	请选择由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。

参数名		说明
	索引	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。
GES	选择图	选择存储了以“关系”为基础的结构数据的图。
	绑定Agent	请选择由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。
OBS连接	OBS桶	选择待采集数据归属的OBS桶。
	OBS路径	选择待采集数据在OBS桶中的存储路径。
	采集范围	选择待采集数据的采集范围。 ● 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。 ● 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。
DIS	是否采集转储任务	勾选“采集”表示采集转储任务。
	采集通道	DIS服务的实例即通道。此参数表示选择通道，进行采集。

c. 元数据采集参数配置，参考**表11-6**。

说明

仅当数据连接类型为DWS、DLI、MRS HBase、MRS Hive、ORACLE、RDS时，支持配置元数据采集参数。

表 11-6 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11-7**。

说明

- 仅当数据连接类型为DWS、DLI时，支持配置数据概要。
- 如无特殊需求时，建议您无需开启数据概要。开启数据概要后会对数据源端产生较大的SQL执行压力，导致元数据采集任务时间超出预期。

表 11-7 数据概要参数说明

参数名	说明
基于全量数据	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。

参数名	说明
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。

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**新建数据分类（待下线）**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“数据地图 > 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11-8**。

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 11-8 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

参数名	说明
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，选择对应工作空间的“数据目录”模块，进入数据目录页面。
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 11-9 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。

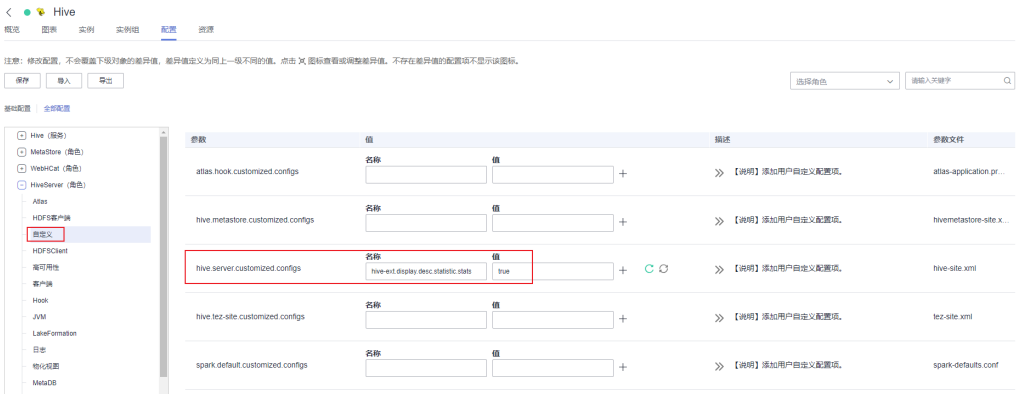
参数名	说明
操作	对已创建的采集任务可进行如下操作： • 编辑：支持对采集任务（状态为已启动、未启动、运行失败）的采集策略强相关参数进行修改，不支持修改数据源类型。 • 运行：单击“运行”，可单次运行此采集任务，并可在“任务监控”页面查看其状态和相关日志信息。 • 启动调度：当其状态为“已停止”，则可按照所配置的调度方式启动调度运行。 • 停止调度：当调度状态为“调度中”，则可停止调度。

配置 MRS 集群 Hive 分区表支持元数据采集

步骤1 使用admin账户登录MRS服务的Manager页面。

步骤2 在Manager页面选择“集群 > 服务 > Hive > 配置 > 全部配置”，选择**HiveServer（角色）->自定义**，在“hive.server.customized.configs”参数值中新增**hive-ext.display.desc.statistic.stats**名称，值为**true**，如图11-20所示。

图 11-20 新增自定义参数



步骤3 自定义参数配置完成后，单击左上角的“保存”，在弹窗中单击“确定”保存配置。

图 11-21 保存配置

步骤4 保存成功后，切换到实例页签，选择配置已过期的实例后，单击“更多 > 滚动重启实例”，使配置生效。

图 11-22 滚动重启实例

----结束

11.4.3 查看任务监控

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11-10**。

表 11-10 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。

参数名	说明
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复选框，单击“同步”，即可将分类和密级属性手动同步到资产。

11.5 数据目录典型场景教程

11.5.1 配置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”，进入任务管理页面。

步骤3 在任务管理页面单击“新建”，新建一个元数据采集任务。

步骤4 配置任务信息，如下图所示。

图 11-23 配置任务信息

步骤5 单击“下一步”，配置调度属性如下图所示。

图 11-24 配置调度属性

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”，进入任务管理页面。

步骤3 在任务管理页面单击“新建”，新建一个元数据采集任务。

步骤4 配置任务信息，如下图所示。

图 11-25 配置任务信息

步骤5 单击“下一步”，配置调度属性如下图所示。

图 11-26 配置调度属性

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”，进入任务管理页面。

步骤3 在任务管理页面单击“新建”，新建一个元数据采集任务。

步骤4 配置任务信息，如下图所示。

图 11-27 配置任务信息

步骤5 单击“下一步”，配置调度属性如下图所示。

图 11-28 配置调度属性

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”，进入任务管理页面。

步骤3 在任务管理页面单击“新建”，新建一个元数据采集任务。

步骤4 配置任务信息，如下图所示。

图 11-29 配置任务信息

步骤5 单击“下一步”，配置调度属性如下图所示。

图 11-30 配置调度属性

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

11.5.2 通过数据目录查看数据血缘关系

11.5.2.1 数据血缘方案简介

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。

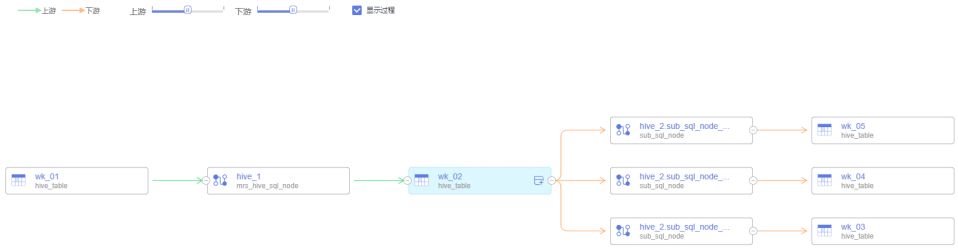
数据的产生、加工融合、流转流通，到最终消亡，数据之间自然会形成一种关系。我们借鉴人类社会中类似的一种关系来表达数据之间的这种关系，称之为数据的血缘关系。与人类社会中的血缘关系不同，数据的血缘关系还包含了一些特有的特征：

- **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。

DataArts Studio生成的血缘关系图如**图11-31**所示， 为数据表对象， 为作业节点对象，通过对象和箭头的编排表示血缘信息。从血缘关系图中可以看到，wk_02表数据是由wk_01表数据经过hive_1作业节点加工而生成的，wk_02表数据经由hive_2作业节点加工又分别生成了wk_03、wk_04和wk_05的表数据。

图 11-31 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：

  DataArts Studio数据血缘解析方案包含自动分析血缘和手动配置血缘两种方式。一般推荐使用自动血缘解析的方式，无需手动配置即可生成血缘关系，在不支持自动血缘解析的场景下，再手动配置血缘关系。

  - 自动血缘解析，是由系统解析数据开发作业中的数据处理和数据迁移类型节点后自动产生的，无需进行手动配置。支持自动血缘解析的节点类型和场景请参见**自动血缘解析**。
  - 手动配置血缘，是在数据开发作业节点中，自定义血缘关系的输入表和输出表。注意手动配置血缘时，此节点的自动血缘解析将不生效。支持手动配置血缘的节点类型请参见**手动配置血缘**。

- 数据血缘的展示：

  首先在数据目录组件完成元数据采集任务，当数据开发作业满足**自动血缘解析要求**或已**手动配置血缘**，然后成功完成作业调度后，则可以在数据目录模块可视化查看数据血缘关系。

11.5.2.2 配置数据血缘

DataArts Studio数据血缘解析方案包含自动分析血缘和手动配置血缘两种方式。一般推荐使用自动血缘解析的方式，无需手动配置即可生成血缘关系，在不支持自动血缘解析的场景下，再手动配置血缘关系。

- 自动血缘解析，是由系统解析数据开发作业中的数据处理和数据迁移类型节点后自动产生的，无需进行手动配置。支持自动血缘解析的节点类型和场景请参见**自动血缘解析**。
- 手动配置血缘，是在数据开发作业节点中，自定义血缘关系的输入表和输出表。注意手动配置血缘时，此节点的自动血缘解析将不生效。支持手动配置血缘的节点类型请参见**手动配置血缘**。

约束限制

手动配置血缘当前暂不支持字段级血缘解析。

自动血缘解析

自动血缘解析无需进行手动配置，当数据开发作业中包含如表11-11所示节点及场景时，系统支持自动解析血缘关系。

说明

解析SQL节点的血缘时，支持多SQL解析及列级血缘解析，单条SQL语句不支持SQL中含有分号的场景。

表 11-11 支持自动血缘解析的作业节点及场景

作业节点	支持场景
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。
DWS SQL	支持Insert into等DML操作产生的DWS表之间的血缘。
MRS Hive SQL	支持Insert into/overwrite等DML操作产生的MRS表之间的血缘。
MRS Spark SQL	支持Insert into/overwrite等DML操作产生的MRS表之间的血缘。
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。
ETL Job	支持DLI、OBS、MySQL以及DWS之间的ETL任务产生的血缘。

手动配置血缘

在DataArts Studio数据开发的作业中，您可以在数据开发作业节点中，自定义血缘关系的输入表和输出表。注意，当手动配置血缘时，此节点的自动血缘解析将不生效。

支持手动配置血缘的作业节点类型如下所示。

- CDM Job
- Rest Client
- DLI SQL
- DLI Spark
- DWS SQL
- MRS Spark SQL
- MRS Hive SQL
- MRS Presto SQL
- MRS Spark

- MRS Spark Python
- ETL Job
- OBS Manager

手动配置血缘时，在节点的“血缘关系”页签，配置血缘的输入和输出表。输入和输出表的所属数据源支持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。

图 11-32 手动配置血缘关系示例

例如，当需要配置数据开发Pipeline作业中MRS Spark节点的血缘关系时，由于MRS Spark节点不支持自动血缘解析，则需要手动配置MRS Spark节点的血缘关系。操作步骤如下：

步骤1 参考访问DataArts Studio实例控制台，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

步骤3 在数据开发组件，进入“数据开发 > 作业开发”页签，单击需要手动配置血缘关系的作业名，打开作业画布。

步骤4 单击作业画布中的MRS Spark节点，并切换到“血缘关系”页签。

图 11-33 进入血缘关系页签

步骤5 在MRS Spark节点的“血缘关系”页签，手动配置血缘的输入表。假如MRS Spark作业中的输入表为“hive”，则血缘输入配置如**图11-34**所示。

图 11-34 配置血缘输入

步骤6 完成血缘的输入表配置后，单击确定，继续配置血缘的输出表。假如MRS Spark作业中的输出表为“a”，则血缘输出配置如**图11-35**所示。

图 11-35 配置血缘输出

步骤7 完成血缘的输出表配置后，单击确认，则此MRS Spark节点的血缘关系手动配置成功。后续当需要查看血缘关系时，参考**查看数据血缘**完成元数据采集，并成功完成作业调度后，即可在数据目录组件查看手动配置的MRS Spark节点血缘关系。

----结束

11.5.2.3 查看数据血缘

首先在数据目录组件完成元数据采集任务，当数据开发作业满足**自动血缘解析要求**或已**手动配置血缘**，然后成功完成作业调度后，则可以在数据目录模块可视化查看数据血缘关系。

约束限制

- 数据血缘关系更新依赖于作业调度，数据血缘关系是基于最新的作业调度实例产生的。

  说明

  对于同一版本的数据开发作业，系统基于最新的作业调度实例生成数据血缘关系后，在冷却期（默认为48小时）内不会再次更新数据血缘关系。如需更新，需要等待冷却期结束或将数据开发作业再次提交版本后调度。

- 数据血缘关系删除需要通过删除作业或删除作业元数据的方式进行，仅将作业停止调度不会触发血缘关系的删除。

新建并运行元数据采集任务

请参见**配置元数据采集任务**，新建并运行元数据采集任务，注意任务中需要选择待查看血缘关系的数据表。

如果此前已创建并运行过待查看数据表的元数据采集任务，此操作可跳过。

启动作业调度

元数据采集完成后，系统基于最新的作业调度实例产生相关的数据血缘关系。

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据开发”模块，进入数据开发页面。

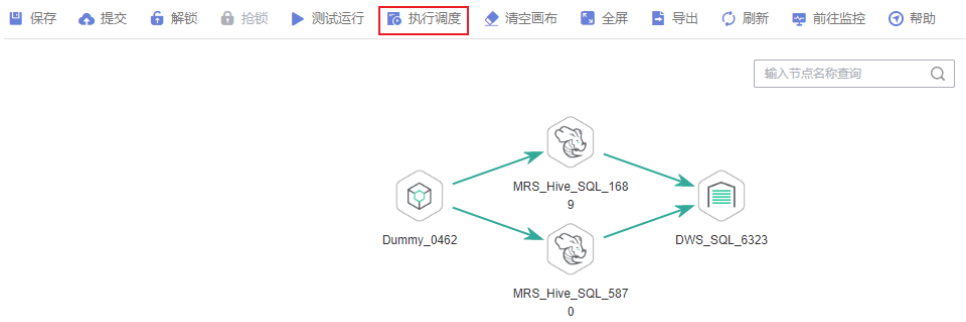
步骤3 在数据开发控制台，单击左侧导航栏中的作业开发按钮，进入作业开发页面后，打开已完成血缘配置的作业。

步骤4 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 11-36 作业调度



步骤5 待调度作业成功运行完成后，等待约1分钟左右，数据血缘关系即可生成成功。

----结束

查看数据血缘关系

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

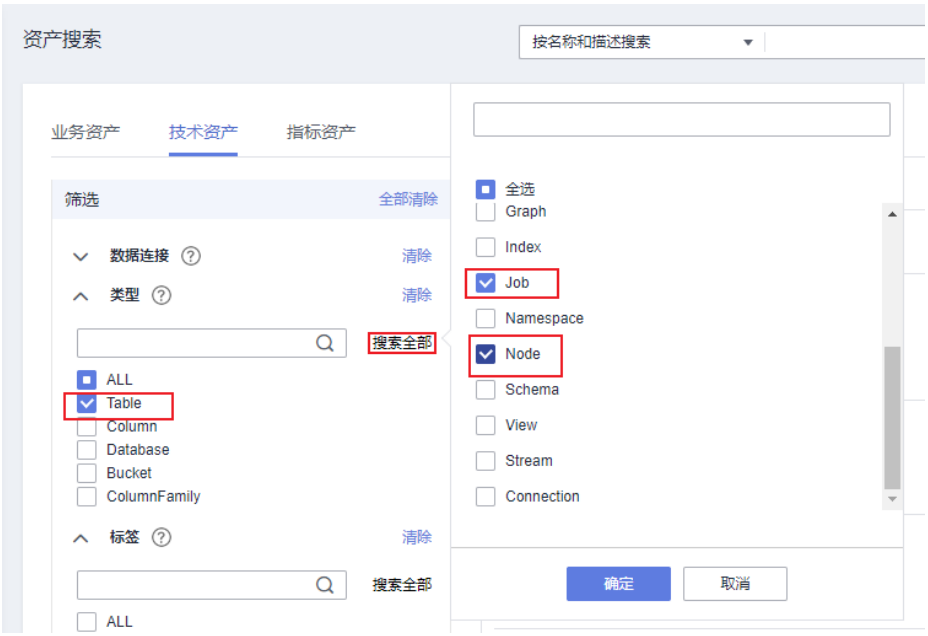
步骤2 在“数据目录 > 技术资产”页面，可以对数据开发的作业、节点、表进行查询。

在“类型”筛选区域，单击“搜索全部”按钮并在全部类型中勾选“Job”、“Node”和“Table”，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

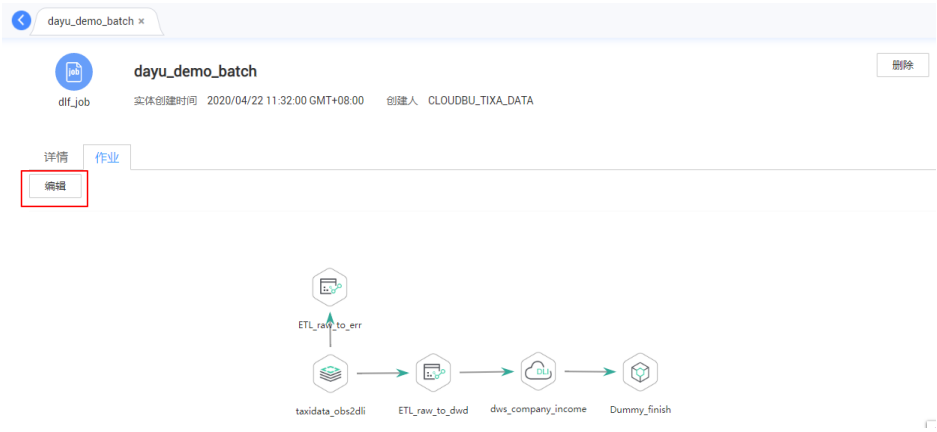
数据开发中的作业信息不属于任何一个数据连接，故如果在搜索条件中勾选数据连接，则查询不到结果。

图 11-37 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。

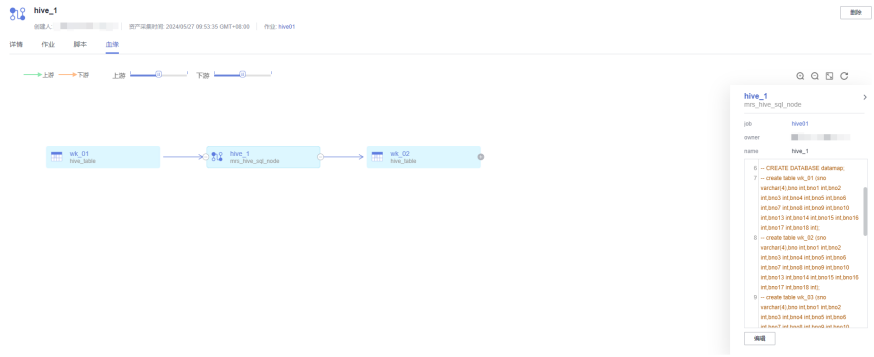
图 11-38 查看作业



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。

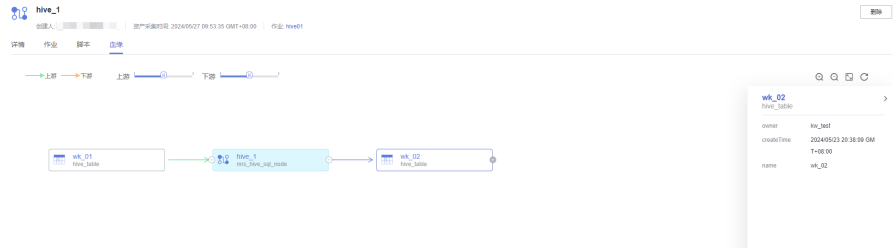
图 11-39 查看节点血缘



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 11-40 查看表血缘



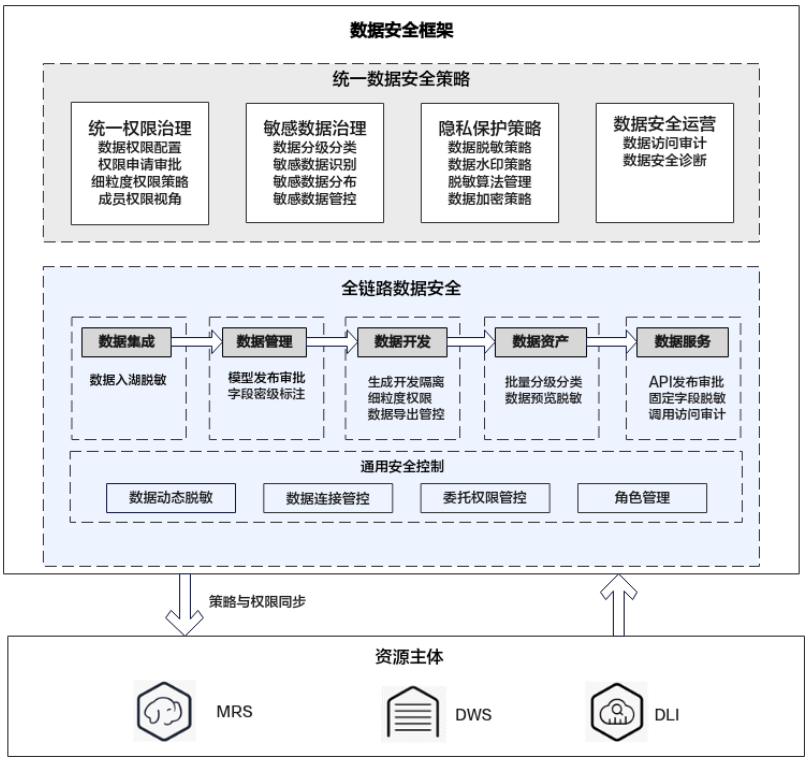
----结束

12 数据安全

12.1 数据安全概述

数据安全以数据为中心，基于数据动态流动场景，构建全链路数据湖安全的解决方案，全方位保障数据湖安全，以此满足不同角色（如数据开发工程师，数据安全管理员，数据安全审计员和数据安全运营人员）对数据安全和数据治理的诉求。

图 12-1 DataArts Studio 数据安全框架



- **资源主体**：即华为云数据湖中的库表字段及计算引擎队列资源。库表字段支持大数据MRS Hive/Spark，云数据仓库DWS，数据湖探索DLI等数据湖，计算引擎队列包含大数据MRS YARN计算队列和数据湖探索计算队列。

- **全链路数据安全**：DataArts Studio数据治理全链路包含数据集成、数据管理（架构设计、指标设计、数据质量管理）、数据开发、数据资产管理和数据服务等不同阶段。在数据动态流动场景下，可通过数据访问控制、数据脱敏等安全防护措施保障数据全链路、全生命周期安全能力。例如：数据入湖阶段，支持对敏感字段进行脱敏设置，支持对数据源连接进行管控，控制对数据源的访问权限；分析师查询数据时，支持通过动态脱敏策略或字段访问权限来保护敏感数据。
- **统一数据安全策略**：包括统一权限治理、敏感数据治理、隐私保护策略和数据安全运营四大能力。

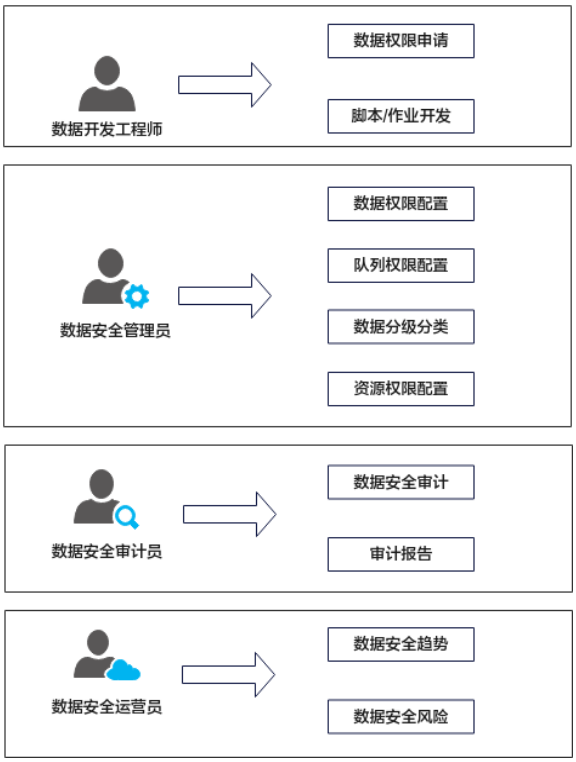
说明

数据安全组件当前在中国-香港、亚太-新加坡、亚太-曼谷、亚太-雅加达、拉美-圣地亚哥、拉美-圣保罗一、非洲-约翰内斯堡和土耳其-伊斯坦布尔区域部署上线。

使用场景

数据安全可满足不同角色（如数据开发工程师，数据安全管理员，数据安全审计员和数据安全运营人员）对数据安全和数据治理的诉求，不同角色用户使用数据安全的场景如图12-2所示。

图 12-2 用户使用场景图



特点优势

- 数据安全融合了不同的大数据服务进行统一入口管理，包括MRS、DWS、DLI，统一的权限配置入口能力，提高了易用性和可维护性。
- 数据安全以数据为中心，提供了围绕数据全链路的数据安全能力，如统一权限治理、敏感数据治理、隐私保护策略管理。

- 统一权限治理支持按照项目空间分配空间权限集（每个项目空间可以管理的库表权限范围），空间内按照角色给不同用户、用户组进行权限分配，跨空间依赖支持灵活按需的权限申请审批能力。
- 敏感数据管理支持敏感数据的分级分类，自动识别发现，以及基于敏感数据等级的安全管控策略能力。
- 隐私保护管理提供了静态与动态的数据脱敏能力、数据水印能力，满足业务需求同时保证数据安全。

功能介绍

数据安全包括如下功能：

- **统一权限治理**

  统一权限治理基于MRS、DWS、DLI服务，提供数据权限管理能力。您可以创建空间权限集、权限集或角色，并通过这些权限配置模型实现MRS、DWS、DLI数据的访问控制，按需为用户、用户组分配最小权限，从而降低企业数据信息安全风险。

- **敏感数据治理**

  敏感数据识别通过用户创建或内置的数据识别规则和规则组自动发现敏感数据并进行数据分级分类标注。

- **隐私保护管理**

  隐私保护管理可以通过数据静态脱敏、动态脱敏、数据水印、文件水印和动态水印等方式来防止敏感数据遭到有意或无意的误用、泄漏或盗窃，从而帮助企业采取合理措施来保护其敏感数据的机密性和完整性、可用性。

- **数据安全运营**

  提供数据安全诊断能力、数据湖访问审计日志查询能力，方便用户更好的做到安全管控。

12.2 数据安全总览页面

数据安全总览页面为您提供了配置数据安全管理员功能和数据安全的空间汇总信息，包括查看敏感表总数、敏感表密级分布饼图、敏感字段密级分布饼图、脱敏和水印任务数量趋势图。

配置安全管理员

安全管理员由具有DAYU Administrator系统角色权限的账号指定，在DataArts Studio实例内所有工作空间的数据安全组件内，拥有最高权限。数据安全组件中，仅安全管理员和DAYU Administrator系统角色有权限进行如下操作：

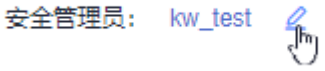
- 配置空间权限集
- 配置行级权限访问控制
- 同步用户
- 配置空间资源权限
- 配置细粒度认证
- 配置队列权限

如需配置安全管理员，则需要以具有DAYU Administrator系统角色权限的账号登录数据安全总览页面，选择某个IAM子用户或者用户组（选择用户组时，则该用户组中的所有用户均为安全管理员）作为安全管理员。

说明

- 配置安全管理员，必须由DAYU Administrator操作，安全管理员本身不可操作。
- 安全管理员的权限当且仅当在数据安全组件生效，对于周边组件和其他服务，此身份无效。

图 12-3 配置安全管理员



查看数据概况

在总览页，用户可以根据日期，根据不同数据源类型。例如查看数据仓库服务（DWS）、数据湖探索（DLI）或MapReduce服务（MRS Hive）类型的下所包含的数据库中的敏感数据，包括敏感表总数、敏感字段总数、脱敏表数、嵌入水印表数、水印溯源数。

图 12-4 数据概况

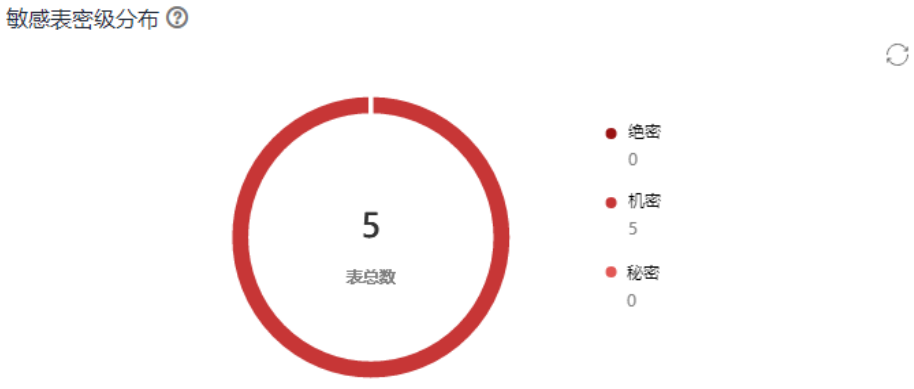
数据分析报表

- 敏感表密级分布图

  展示敏感发现任务识别出的表的密级分布，密级和用户定义的一致。右侧显示用户定义的密级及其关联的敏感表数目。

  敏感数据识别任务的创建和运行，参考**创建敏感数据发现任务**。

  图 12-5 敏感表密级分布图

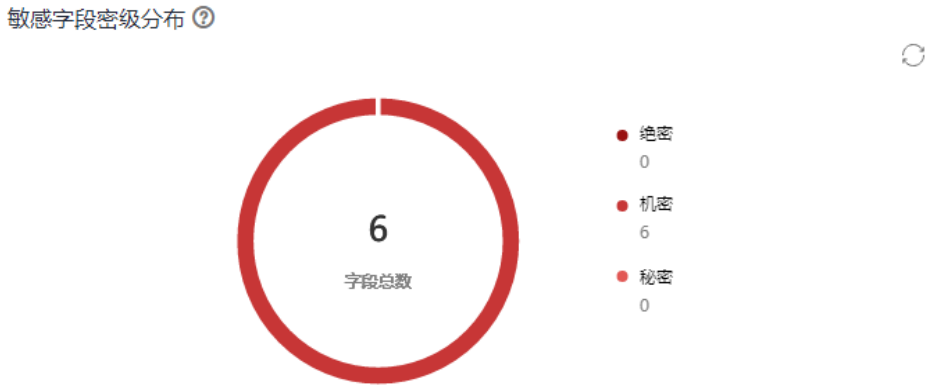


- 敏感字段密级分布图

  展示敏感发现任务识别出的表敏感字段，密级和用户定义的一致。右侧显示用户定义的密级及其关联的敏感字段数目。

  敏感数据识别任务的创建和运行，参考**创建敏感数据发现任务**。

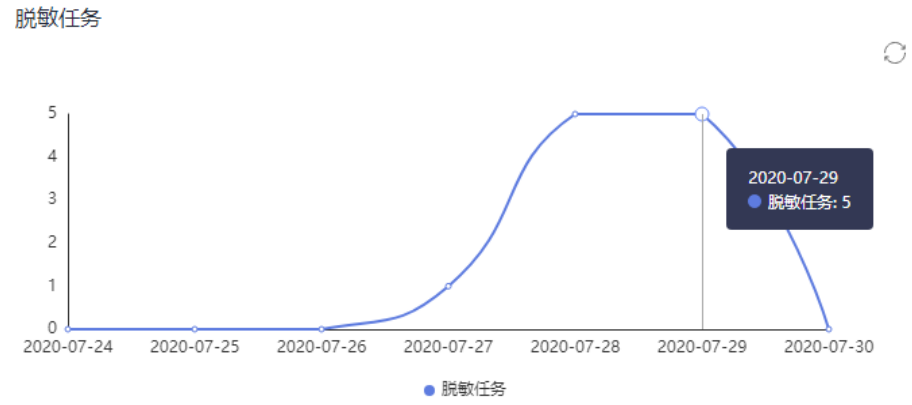
  图 12-6 敏感字段密级分布图

- 脱敏任务趋势图

  敏感任务趋势图展示七天内运行的脱敏任务数来反映任务趋势变化。数据脱敏任务创建和运行，参考**创建静态脱敏任务**。

  图 12-7 脱敏任务趋势图

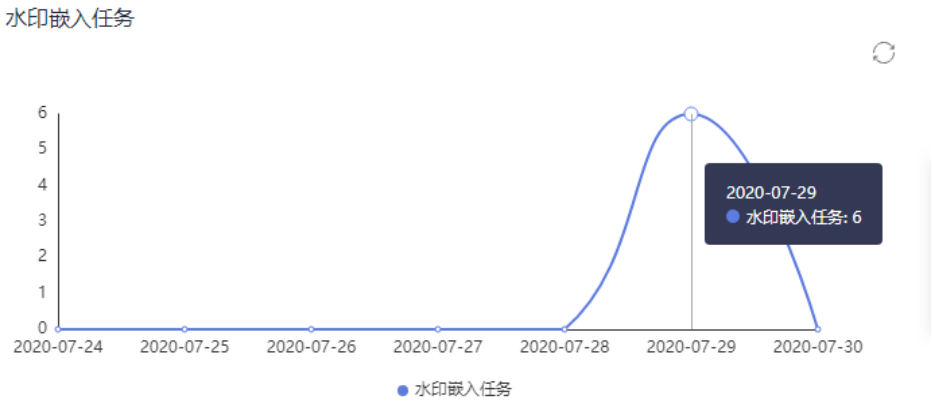
  

- 水印嵌入趋势图

  水印嵌入任务趋势图展示七天内运行的水印嵌入任务数来反映趋势变化。

  水印嵌入任务创建和运行，参考**创建数据水印嵌入任务**。

图 12-8 水印嵌入任务趋势图



12.3 统一权限治理

12.3.1 权限治理使用流程

统一权限治理提供了对于MRS、DWS、DLI数据湖仓中的库、表、字段的访问权限配置的核心能力，包含如下特性：

- 集中式访问控制

  融合了不同的大数据服务进行统一权限管理，如MRS、DWS、DLI等，给用户带来了统一的权限配置入口，极大的提高了易用性和可维护性。

- 多级权限配置模型

  权限模型分级管理，权责分明。空间权限集限定了整个工作空间用户可访问的权限范围，权限集或角色将空间权限集定义的权限范围进行拆分细化，将用户与权限关联进行权限管控。

- 精细化权限管控能力

  支持界面化的RBAC数据权限精细化配置能力，按不同角色、用户、用户组进行权限分配。并支持按需高效的权限申请审批能力，权限审批后即刻生效。

- 多维度权限视角展示
  - 成员视角支持以工作空间成员视角，来展示每个用户（组）所申请的数据表权限，以及展示、配置或收回每个用户（组）的权限集关系。
  - 数据视角支持以库、表、字段的数据视角，来展示、配置数据在当前权限集下的权限关系。
  - 权限视角支持以权限策略视角，来展示、配置或收回数据在当前权限集下的权限策略关系。

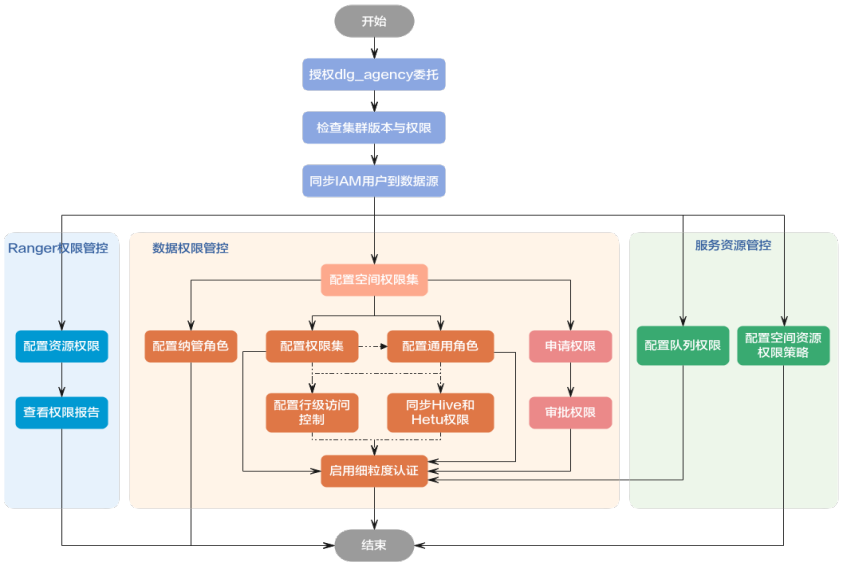
- 支持空间资源管控

  除了数据权限管控外，还支持对空间资源进行管控，例如数据连接、委托等资源。

使用流程

您可通过图12-9了解统一权限治理的使用流程。

图 12-9 统一权限治理使用流程图



统一权限治理支持**数据权限管控**、**服务资源管控**和**Ranger权限管理**，流程介绍如下：

数据权限管控流程

1. **授权dlg_agency委托**

   由于数据安全使用委托时，所需的云服务权限更高。因此在使用数据安全前，需要提前为dlg_agency委托授予相关权限。

2. **检查集群版本与权限**

   统一权限治理对数据连接Agent、数据源版本和用户权限等均有相应的要求。在使用前，您应先检查并准备相关配置。

3. **同步IAM用户到数据源**

   将IAM上的用户信息同步到数据源，以实现不同用户访问数据源时，能够根据其自身用户信息管控用户访问数据的权限。

4. **配置空间权限集**

   空间权限集作为DataArts Studio工作空间内的最大权限集合，主要用于确定整个工作空间用户可访问的权限范围。

5. **配置权限集**

   权限集将用户与权限直接关联，可以新建多个用于给不同使用场景的用户关联不同的权限，可通过权限同步进行权限管控（实际使用时，更推荐通过权限集关联角色进行权限管控）。

6. **配置通用角色**

   配置通用角色即在数据源上创建新角色，用于承载用户和权限之间的关联关系，可以更加直观地管理权限关系、进行权限管控。

7. **配置纳管角色**

   配置纳管角色即为纳管MRS数据源上已有的角色，并继承已有角色的MRS数据源权限。

8. **配置行级访问控制**

   数据安全支持成员管理视图，支持查看当前工作空间内成员的权限，并进行角色/权限集管理。

9. **同步MRS Hive和Hetu权限**

   数据安全支持成员管理视图，支持查看当前工作空间内成员的权限，并进行角色/权限集管理。

10. **申请权限**

    进行访问权限管理时，除了可以自上而下地通过权限集/角色方式为用户授权外，也支持自下而上的用户权限申请、审批流程。

11. **审批权限**

    审批人来自权限集/角色的管理员，权限审批后即刻生效。

12. **启用细粒度认证**

    配置细粒度认证后，在DataArts Studio数据开发执行脚本、测试运行作业或调度作业时，数据源将不再使用数据连接上的账号，而是使用当前用户身份认证鉴权，从而做到实现不同用户具有不同的数据权限，使角色/权限集或队列权限中的权限管控生效。

服务资源管控流程

1. **配置队列权限**

   队列权限可以为当前工作空间分配可使用的MRS Yarn和DLI队列资源，并为用户组/用户配置对应的队列权限策略。

   - 当为工作空间分配队列资源后，在数据开发组件在为作业节点配置队列资源时，可选择的队列为当前空间下已分配的队列资源。
   - 当为用户组/用户配置队列权限策略后，授权对象将按照策略内容被授予相应权限。

2. **配置空间资源权限策略**

   数据安全支持对空间资源进行管控，例如数据连接、委托等资源。空间资源管控后，对于非授权对象的普通用户，则无权再查看并使用此资源。

Ranger权限管理流程

1. **配置资源权限**

   通过统一入口创建MRS各个组件的权限策略，由Ranger组件实现权限控制。

2. **查看权限报告**

   通过全面的权限报告，查看资源配置权限策略及其详情。

数据权限管控说明

当前数据权限管控为白名单机制，是在待授权用户原有权限的基础上增加允许操作条件，不会影响用户的原有权限。如果仅需要当前数据权限管控所赋予的权限生效，则需要您手动去除待授权用户的原有权限。默认情况下，DataArts Studio用户的原有数据权限如下：

- 对于DLI数据源，DAYU Administrator或DAYU User用户默认具备DLI Service Admin权限，因此待授权用户默认具备DLI库表的所有数据权限。如果需要去除授权用户的默认权限，则需要删除用户的DLI Service Admin权限。
- 对于DWS数据源，即使DAYU Administrator或DAYU User用户默认具备DWS Administrator权限，但是由于DWS的数据库权限跟控制台IAM权限相互分离，因此默认情况下，待授权用户不具备DWS库表的数据权限，仅当前数据权限管控所赋予的数据权限生效。
- 对于MRS数据源，DAYU Administrator或DAYU User用户默认具备MRS Administrator权限，该用户同步到MRS后会被赋予对应角色（详见**IAM用户同步**

MRS说明），然后由Ranger组件提供默认策略放通权限（详见配置组件权限策略），因此待授权用户默认具备MRS Hive库表的数据权限。如果需要去除授权用户的默认权限，则需要您在Ranger组件上去除系统默认策略中的**public**用户组，操作步骤如下：

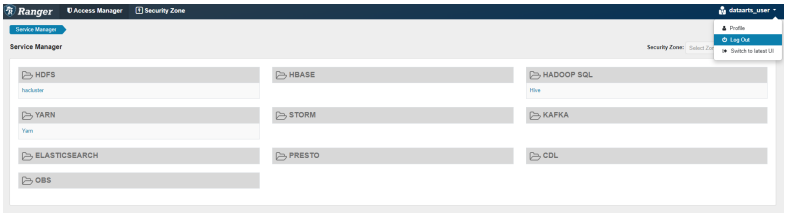
a. 使用admin账户登录MRS服务的Manager页面。

b. 在Manager页面选择“集群 > 服务 > Ranger”，进入Ranger概览页面，单击RangerAdmin进入Ranger WebUI。

图 12-10 进入 Ranger WebUI

c. 注销当前账号，使用Ranger管理员账号再次登录。注意，普通集群使用Manager页面的admin账号即可作为Ranger管理员账号，安全集群需要使用rangeradmin作为Ranger管理员账号，rangeradmin默认密码请参考用户账号一览表章节。

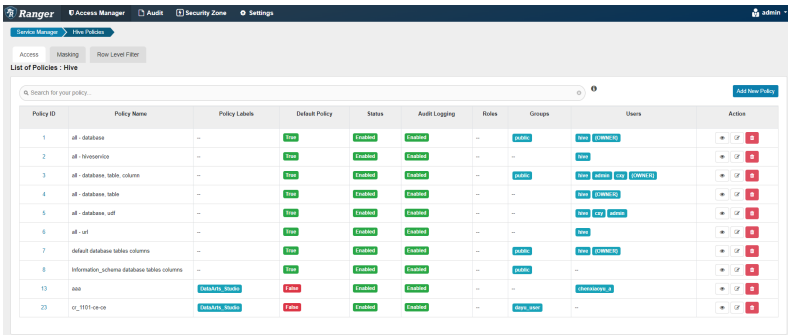
图 12-11 登出当前账号



d. 在首页中单击“HADOOP SQL”区域的组件插件名称如“Hive”。

e. 在“Access”页签，找到列表中**Groups**列包含**public**的默认策略（即**Default Policy**列为**True**的策略），然后分别进行编辑，移除其中的**public**用户组。

图 12-12 策略列表



12.3.2 授权 dlg_agency 委托

云服务委托可将相关云服务的操作权限委托给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。首次进入DataArts Studio控制台首页时，系统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。同意授权后，DataArts Studio会自动创建名为dlg_agency的委托。如果未同意授权，会在下次进入控制台首页时，再次弹出对话框。

由于数据安全使用委托时，所需的云服务权限更高。因此在使用数据安全前，需要提前为dlg_agency委托授予相关权限，所需权限如**表12-1**所示。

表 12-1 待授予权限合集

权限名称	配置目的	是否必选	授权项/系统权限（二者选其一配置即可）	
IAM权限	系统获取用户或用户组、创建角色时，需要该权限。 例如用户或权限同步时，如果无此权限会导致操作失败。	MRS/DWS/DLI权限管理时必选	• iam:users:listUsers • iam:groups:listGroups • iam:users:listUsersForGroup • iam:roles:createRole • iam:roles:deleteRole • iam:roles:updateRole • iam:permissions:grantRoleToGroup • iam:permissions:listRoleAssignments • iam:permissions:revokeRoleFromGroup **说明** 受IAM权限策略影响，暂无可授权项支持获取DLI用户组。如果涉及DLI用户组的权限管理场景，需要授权Security Administrator系统权限。	Security Administrator
MRS/DWS数据连接Agent权限	系统进行权限同步时，需要该权限。 例如权限集权限同步、角色权限同步或者审批权限申请时，如果无此权限会导致操作失败。	MRS/DWS权限管理时必选	任一CDM权限，例如：cdm:cluster:get	任一CDM权限，例如：CDM Administrator

权限名称	配置目的	是否必选	授权项/系统权限（二者选其一配置即可）	
MRS用户同步权限	MRS用户同步时，需要该权限。 例如MRS用户同步时，如果无此权限会导致用户同步失败。	MRS权限管理时必选	• mrs:cluster:syncUser	MRS FullAccess
DWS用户同步权限	DWS用户同步时，需要该权限。 例如DWS用户同步时，如果无此权限会导致用户同步失败。	DWS权限管理时必选	• dws:dbAuthority:syncIamUser • dws:dbAuthority:updateUser	DWS FullAccess
DLI权限同步权限	DLI权限同步时，需要该权限。 例如DLI权限同步时，如果无此权限会导致同步失败，系统提示权限不足。	DLI权限管理时必选	不支持授权项，需要配置系统权限DLI FullAccess	DLI FullAccess

前提条件

在进入DataArts Studio控制台首页时，已在弹出访问授权对话框中选择同意授权，以便系统自动创建名为dlg_agency的委托。

约束与限制

委托授权成功后，需要等待约15-30分钟待权限生效，然后可正常使用数据安全进行访问权限管理。

为 dlg_agency 委托授权

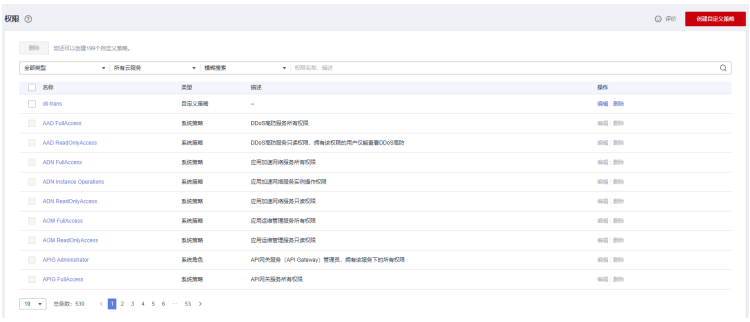
为dlg_agency委托授权时，应按照您的业务场景，从**表12-1**中选择所需的授权项或系统权限进行授权，其中授权项或系统权限二者选其一配置。

本例以MRS权限管理业务场景为例进行说明，则需要授予的权限为IAM权限、MRS/DWS数据连接Agent权限和MRS用户同步权限。本例中使用授权项配置最小权限，授权操作如下：

步骤1 登录统一身份认证服务IAM控制台。

步骤2 在IAM服务左侧导航窗格中，进入“权限管理 > 权限”，单击页面中的“创建自定义策略”。

图 12-13 单击创建自定义策略



步骤3 在弹出的创建自定义策略页面中，切换到JSON视图，填写MRS权限管理所需的IAM相关自定义策略后，单击“确定”完成IAM相关自定义策略创建。

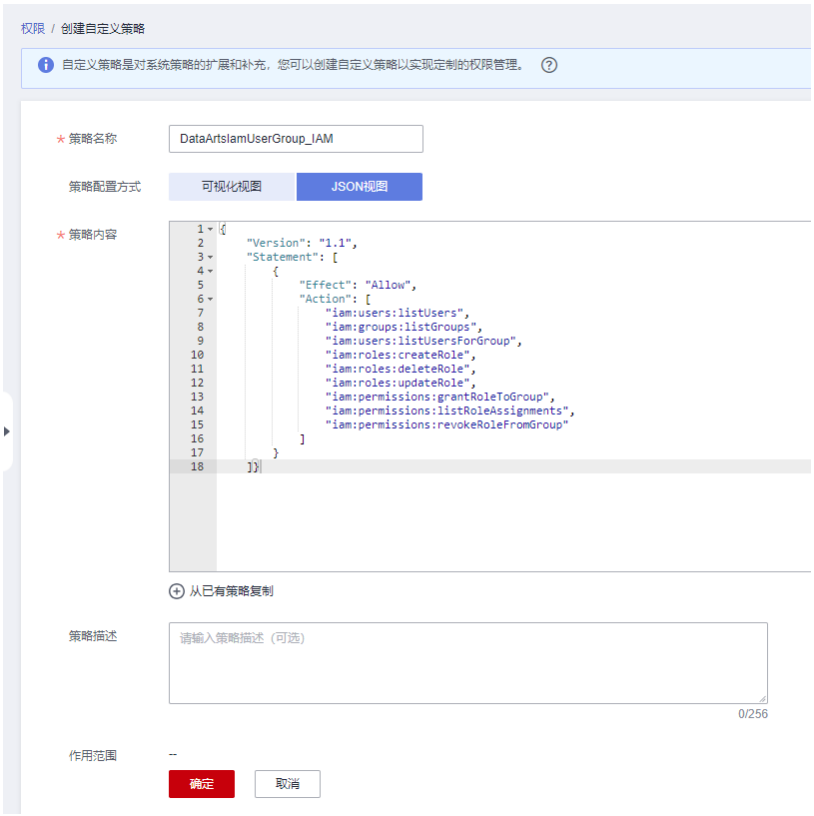
说明

创建自定义策略时，暂不支持同时选全局级云服务和项目级云服务，需要拆分为两条策略。因此本例先配置IAM相关策略，再配置MRS和CDM相关策略。

- 名称：DataArtsIamUserGroup_IAM
- 策略配置方式：单击“JSON视图”，切换到JSON视图。
- 策略内容：在JSON视图中，输入如下JSON代码，并单击“确认”。

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "iam:users:listUsers",
                "iam:groups:listGroups",
                "iam:users:listUsersForGroup",
                "iam:roles:createRole",
                "iam:roles:deleteRole",
                "iam:roles:updateRole",
                "iam:permissions:grantRoleToGroup",
                "iam:permissions:listRoleAssignments",
                "iam:permissions:revokeRoleFromGroup"
            ]
        }
    ]
}
```

图 12-14 创建 IAM 相关自定义策略



步骤4 再次单击“创建自定义策略”，在弹出的创建自定义策略页面中，切换到JSON视图，填写MRS权限管理所需的MRS和CDM相关自定义策略配置后，单击“确定”完成MRS和CDM相关自定义策略创建。

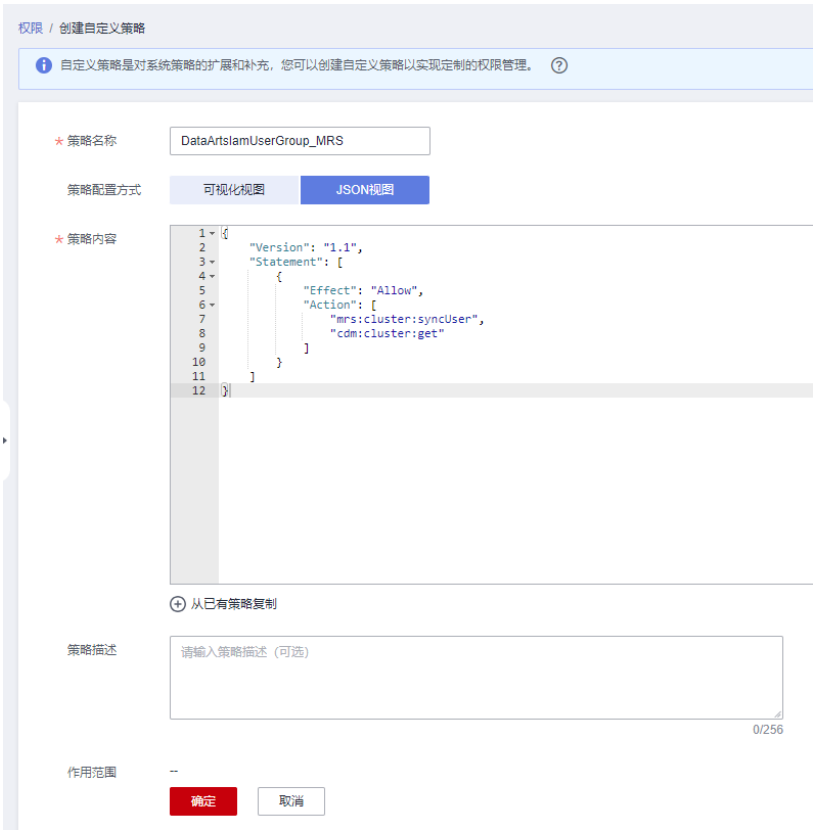
说明

创建自定义策略时，暂不支持同时选全局级云服务和项目级云服务，需要拆分为两条策略。因此本例先配置IAM相关策略，再配置MRS和CDM相关策略。

- 名称：DataArtsIamUserGroup_MRS
- 策略配置方式：单击“JSON视图”，切换到JSON视图。
- 策略内容：在JSON视图中，输入如下JSON代码。

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "mrs:cluster:syncUser",
                "cdm:cluster:get"
            ]
        }
    ]
}
```

图 12-15 创建 MRS 和 CDM 相关自定义策略

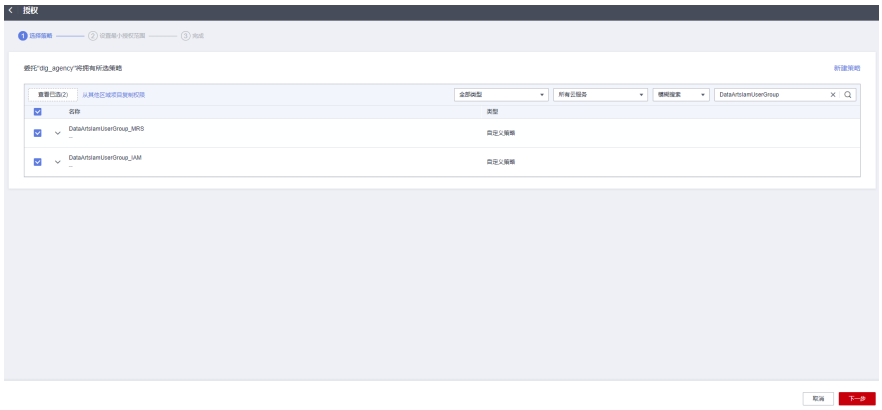


步骤5 在IAM服务左侧导航窗格中，进入“委托”，搜索“dlg_agency”，找到dlg_agency委托项，单击“授权”。

图 12-16 dlg_agency 授权

步骤6 在授权框中，找到并勾选之前创建的自定义策略DataArtsIamUserGroup_IAM和DataArtsIamUserGroup_MRS，单击“下一步”。

图 12-17 选择自定义策略



步骤7 单击“确定”，给委托完成授权。授权后，等待约15-30分钟，即可使用数据安全进行MRS访问权限管理。

----结束

12.3.3 检查集群版本与权限

统一权限治理对数据连接Agent、数据源版本和用户权限等均有相应的要求。在使用前，您应先按照**表12-2**，检查并准备相关配置。

说明

DLI权限管理仅涉及**授权dlg_agency委托**，不涉及检查集群版本与权限。

使用前检查 checklist

表 12-2 使用前检查 checklist

检查项	是否必选	检查内容	配置指导
数据连接Agent版本	MRS/DWS权限管理时必选	CDM集群为2.10.0.300及以上版本。	登录CDM管理控制台，进入“集群管理”，在集群列表中找到所需要的集群，然后单击集群名称，进入集群“基本信息”页面查看集群版本号。 如果非所需版本，请创建最新版本CDM集群或联系客服或技术支持人员。
Ranger组件配置	MRS权限管理时必选	MRS非安全集群Ranger组件开启同步ldap用户功能。	MRS非安全集群，由于Ranger组件默认同步unix用户，不会同步Manager上的用户/用户组/角色，因此需要切换用户同步策略。操作详情请参考**配置Ranger组件**。

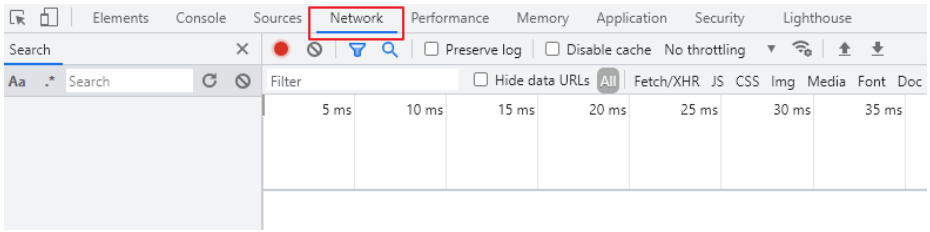
检查项	是否必选	检查内容	配置指导
Ranger连接用户权限		连接中的用户具备Ranger组件Admin权限。	Ranger连接中的用户需要具备Ranger组件Admin权限，操作详情请参考**准备Ranger Admin用户**。
DWS集群guest_agent版本	DWS权限管理时必选	DWS集群guest_agent版本为8.2.1，或在8.2.1以上、9.0.0以下	DWS集群guest agent版本号需要通过开发者调试工具查看，操作详情请参考**查看DWS集群guest agent版本**。
DWS连接用户权限		• 非三权分立模式，连接中的用户至少需具备数据库dbadmin权限， • 三权分立模式，连接中的用户需具备系统管理员权限。	• 非三权分立模式，参考**数据库用户**设置dbadmin管理员用户。 • 三权分立模式，参考**设置三权分立**设置系统管理员用户。

查看 DWS 集群 guest agent 版本

步骤1 登录GaussDB(DWS) 管理控制台，进入“集群管理”，在集群列表中找到所需要的集群。

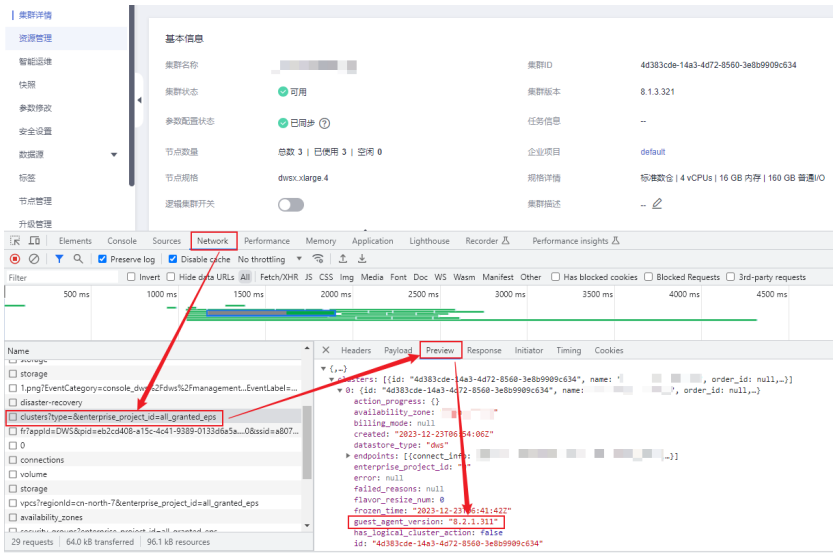
步骤2 按下F12，打开开发者调试工具，然后选择Network功能。

图 12-18 选择 Network



步骤3 在DWS控制台中，单击待查看的DWS集群名称，进入集群“基本信息”页面。然后在开发者调试工具的Network请求中，寻找Name形如“clusters?type=xxxxxx”的长字符串并单击，在右侧区域中选择“Preview”，依次展开字段，查找“guest_agent_version”字段，其值即为DWS集群的guest agent版本。

图 12-19 查找“guest_agent_version”字段



步骤4 如果非所需版本，请联系DWS服务客服或技术支持人员。

----结束

配置 Ranger 组件

对于MRS非安全集群，由于Ranger组件默认同步unix用户，不会同步FI Manager上的用户/用户组/角色，因此需要切换用户同步策略。操作步骤如下所示：

说明

MRS安全集群Ranger组件默认同步LDAP用户，默认情况下无需额外操作。如果默认配置被修改，也可以参考本章节切换用户同步策略。

步骤1 使用admin账户登录MRS服务的Manager页面。

步骤2 在Manager页面选择“集群 > 服务 > Ranger > 配置 > 基本配置”，在搜索框中搜索“ranger.usersync.config.expandor”参数，设置其名称为“ranger.usersync.sync.source”，对应值为“ldap”。

说明

MRS集群低版本（例如MRS 3.1.0）默认不开放此配置项，则需要联系MRS服务客服或技术支持人员协助处理。

图 12-20 配置 ranger.usersync.config.expandor 参数

步骤3 参数配置完成后，单击左上角的“保存”，在弹窗中单击“确定”保存配置。

步骤4 保存成功后，切换到实例页签，选择配置已过期的UserSync实例后，单击“更多 > 滚动重启实例”，使配置生效。

图 12-21 滚动重启实例

----结束

准备 Ranger Admin 用户

Ranger连接中的用户需要具备Ranger组件Admin权限，操作详情如下：

步骤1 使用admin账户登录MRS服务的Manager页面。

步骤2 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有人机用户作为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和hive，角色选择Manager_administrator，然后根据页面提示完成用户的创建。

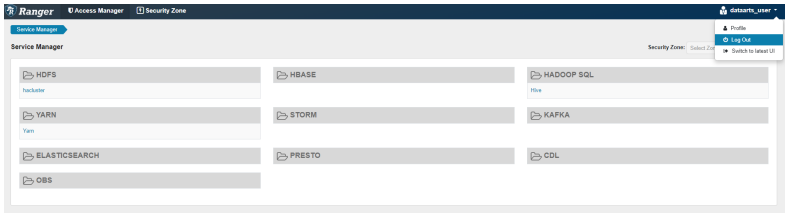
步骤3 使用新建的用户登录Manager页面，并更新初始密码。

步骤4 在Manager页面选择“集群 > 服务 > Ranger”，进入Ranger概览页面，单击RangerAdmin进入Ranger WebUI。

图 12-22 进入 Ranger WebUI

步骤5 注销当前账号，使用Ranger管理员账号再次登录。注意，普通集群使用Manager页面的admin账号即可作为Ranger管理员账号，安全集群需要使用rangeradmin作为Ranger管理员账号，rangeradmin默认密码请参考用户账号一览表章节。

图 12-23 登出当前账号

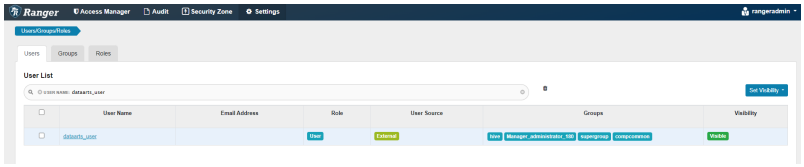


步骤6 修改新建用户的Ranger角色为Admin。在“Settings -> Users/Groups/Roles -> Users”下找到新建的用户名。

说明

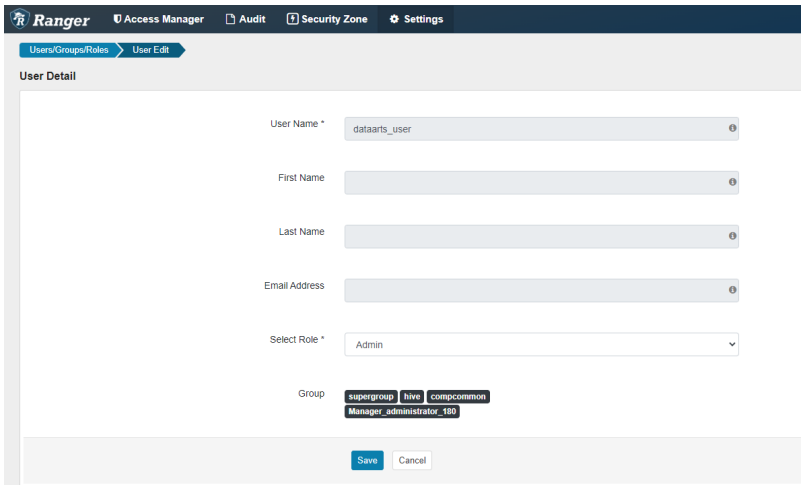
如果Ranger上未找到新建的用户，则需要等待约5分钟，直到Ranger组件自动触发并完成同步MRS集群角色。

图 12-24 查找用户名



步骤7 单击名称进入详情页，修改用户的角色为Admin，单击Save保存。

图 12-25 修改用户角色



----结束

12.3.4 同步 IAM 用户到数据源

默认情况下，用户通过DataArts Studio数据连接访问数据源（此处指MRS/DWS数据源）时，使用数据连接中的账号密码进行认证。为实现不同用户访问数据源时，能够根据其自身用户信息管控用户访问数据的权限，需要先将IAM上的用户信息同步到数

据源上，然后不同用户在数据源上才能有不同的身份，便于后续在数据权限管理中使用自身用户信息进行认证。

值得注意的是，DataArts Studio实例内对每个MRS/DWS集群只能有一个用户同步任务，因此用户同步任务为DataArts Studio实例级别配置，各工作空间之间数据互通。

前提条件

- 新建用户同步任务前，已在管理中心创建数据仓库服务（DWS）或MapReduce服务（MRS Ranger）类型的数据连接，请参考**创建DataArts Studio数据连接**。
- 新建用户同步任务前，已参考**授权dlg_agency委托**为dlg_agency委托配置权限。

约束与限制

- DataArts Studio实例内对每个MRS/DWS集群只能有一个用户同步任务。
- 用户同步任务如果持续运行超过半小时，则会因超时被停止；如果连续同步失败超过10次，则会终止调度。
- 由于联邦用户在只有用户组信息，因此联邦用户无法同步。
- 由于数据源只会同步自身租户的用户信息，因此对于通过IP连接等方式非当前租户的数据源集群无法同步。
- 当前用户同步仅支持MRS Hive和DWS数据源，DWS数据源必须进行用户同步，MRS数据源可以按自身需要创建IAM对应的MRS同名用户从而不进行用户同步。由于DLI数据源直接通过IAM用户进行鉴权，因此无需进行用户同步。
- MRS数据源的用户同步任务有如下约束：
  - 如果MRS数据源已有与待同步用户同名的人机用户，则会导致MRS用户同步任务失败。此失败暂无报错信息，建议可通过如下方式之一解决：
    - 使用MRS集群详情中“IAM用户同步”功能，不必再运行数据安全侧的用户同步任务。IAM用户同步与用户同步任务功能类似，但如果有用户同名时不会导致所有用户同步全部失败，而是只有重名用户才会同步失败。
    - 登录MRS服务Manager页面，选择“系统 > 权限 > 用户”，删除与待同步用户同名的人机用户。
    - 在IAM删除与MRS人机用户同名的待同步用户。
  - MRS数据源同步前，要求用户/用户组已至少配置如下任一的权限，否则不会进行同步。
    - Tenant Administrator
    - MRS FullAccess
    - MRS CommonOperations
    - MRS ReadOnlyAccess
    - MRS Administrator
    - MRS Admin
    - MRS User

    - MRS Viewer
    - Self Define（任意自定义策略）
- DWS数据源的用户同步任务有如下约束：
  - 仅当DWS集群guest_agent版本为8.2.1，或在8.2.1以上、9.0.0以下时，才支持用户同步。DWS集群guest_agent版本查看方法请参考**查看DWS集群guest agent版本**。
  - DWS数据源用户同步前，要求用户已至少配置DWS Database Access权限，否则会同步失败。
  - IAM用户同步到DWS，需要为dlg_agency委托配置如下权限，详见**授权dlg_agency委托**：
    - dws:dbAuthority:syncIamUse
    - iam:users:listUsers
    - iam:groups:listGroups
    - iam:users:listUsersForGroup
  - 由于DWS不支持用户组，因此IAM用户组同步到DWS时，会以“iam_group_***用户组id***”的命名格式在DWS上创建用户，并根据IAM上已删除的用户组在DWS上删除对应的“iam_group_***用户组id***”用户。因此DWS上应避免创建以“iam_group_”为前缀的用户，防止用户被误删。

新建用户同步任务

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“用户同步”，进入用户同步任务页面。

步骤3 在用户同步任务页面，单击“新建”，新建用户同步任务。

图 12-26 新建用户同步任务

步骤4 新建用户同步任务参数配置请参考**表12-3**，参数配置完成单击“确定”，即可新建用户同步任务。

图 12-27 配置用户同步任务

表 12-3 配置用户同步任务参数说明

配置	说明
*选择集群	选择DWS或Ranger数据连接中已连接的DWS或MRS集群。
*集群类型	无需选择，自动匹配集群类型。
*数据连接	无需选择，自动匹配集群数据连接中的数据源集群。
*调度时间	选择调度运行的时间段，左闭右开。 例如调度时间为00-05时，指的是在每天0点到5点时间段内，按照调度间隔进行调度，其他时间段不运行。0点会触发调度，但5点则不会触发。
*调度周期	支持按照小时或分钟进行调度。
*间隔	根据所选的调度周期，选择合理的调度间隔。调度间隔为距离上一次运行时间的间隔，手动同步也计入运行时间。 例如间隔为5分钟时，20:00开始调度，20:03手动执行，则下次调度时间为20:08分。
全量同步	当选择MRS集群时，支持配置是否全量用户同步，默认开启同步全量用户。 当您不需要同步全量用户时，可选择关闭此选项。
*用户/用户组	当关闭全量同步时可，支持指定待同步的用户/用户组，请至少选择一位用户或一个用户组。

步骤5 用户同步任务新建完成后，并不会直接运行。需要您手动同步或调度任务，任务同步成功后才能生效，详见**同步或调度任务**。

----结束

相关操作

- 同步或调度任务：在用户同步任务页面，单击对应任务操作栏中的“同步”或“更多 > 启动调度”，同步或调度任务。对于从未运行过的任务，首次调度如果满足调度时间范围，会立即触发运行。

  说明

  若任务运行发生失败，请参考如下方式处理：
  - 若报错信息为“权限不足”，请参考**授权dlg_agency委托**。
  - 若DWS任务报错信息为“下载DWS IAM凭证失败”，请确认当前用户是否具有至少DWS Database Access权限。
  - 若MRS任务报错信息为“Mrs sync failed, please check the failure cause on the MRS page”，可以登录MRS服务首页，在右侧导航栏中选择操作日志，查看问题原因。
  - 若MRS操作日志中没有报错信息，则一般为IAM用户名与MRS已有人机用户名称冲突，导致同步失败。请登录MRS服务Manager界面，删除与IAM用户同名的人机用户（区分方式：IAM同步用户，默认描述为“IAM Custom Policy User”，且不可被删除；MRS普通人机用户，可以被删除）。
  - 若为其他报错信息，请根据具体报错和日志信息处理。
- 查看任务运行日志：在用户同步任务页面，找到需要查看日志的任务，对应任务操作栏中的“详情”，即可查看运行日志。当前最多展示20条日志记录。

  运行失败可通过日志排查失败原因，问题修正后尝试重新运行。如果仍运行失败，请联系技术支持人员协助处理。
- 编辑任务：在用户同步任务页面，单击对应任务操作栏中的“更多 > 编辑”，即可编辑用户同步任务。
- 删除任务：在用户同步任务页面，单击对应任务操作栏中的“更多 > 删除”，即可删除任务。当需要批量删除时，可以在勾选任务后，在任务列表上方单击“删除”。

  说明

  删除操作无法撤销，请谨慎操作。

12.3.5 数据权限访问控制

12.3.5.1 配置空间权限集

在数据访问权限管理的实际场景下，通常会有一级部门、二级部门、三级部门等多级权限的划分。为此，数据安全组件提供了自上而下分层式的数据权限管理方式。您可以通过空间权限集配置工作空间内的最大权限，在此基础上，将其向下拆分出新的子权限集，提供进一步的细分权限管理。

空间权限集作为DataArts Studio工作空间内的最大权限集合，由DAYU Administrator、Tenant Administrator或者数据安全管理员创建，它限定了整个工作空间用户可访问的权限范围。在空间权限集之下定义的权限集，权限范围只能是空间权限集的子集。

空间权限集和权限集在配置上都是将用户与权限直接关联，二者使用上的区别在于：

- 空间权限集是没有父权限集的顶层权限集，一般每个工作空间下创建一个即可；而权限集必须关联一个空间权限集或其他权限集作为其父权限集，可以新建多个，用于给不同使用场景的用户关联不同的权限。
- 空间权限集主要用于确定工作空间权限范围，而权限集主要用于权限管控。即空间权限集一般无需进行权限同步，且不支持为空间权限集关联角色；而权限集可

通过权限同步进行权限管控（实际使用时，更推荐通过权限集关联角色进行权限管控）。

本章主要描述如何通过**创建空间权限集**和**配置空间权限集**定义工作空间权限范围。

前提条件

- 配置权限集前，已在管理中心创建数据仓库服务（DWS）、数据湖探索（DLI）、MapReduce服务（MRS Hive）和MapReduce服务（MRS Ranger）类型的数据连接，请参考**创建DataArts Studio数据连接**。
- 配置权限集前，已参考**授权dlg_agency委托**为dlg_agency委托配置权限。
- 配置权限集前，已参考**同步IAM用户到数据源**将IAM上的用户信息同步到数据源上。
- 如果希望在权限配置时能够展示数据连接中数据库、表以及字段等元数据提示信息，则需要在数据目录组件，对数据表成功进行过元数据采集，详见**元数据采集任务**。

约束与限制

- 仅DAYU Administrator、Tenant Administrator或者数据安全管理员可以创建、修改或同步空间权限集，权限集管理员支持同步空间权限集，其他普通用户无权限操作。
- 当前通过空间权限集定义权限时，仅支持DLI、MRS Hive和DWS数据源。
- 空间权限集配置完成后，权限管控并不会直接生效，而是需要将空间权限集手动同步到数据源后，权限管控才能生效。

  由于空间权限集主要用于确定工作空间权限范围，而非权限管控，因此一般无需同步空间权限集，实际使用中推荐通过**配置角色**进行权限管控。如果需要同步，则需注意以下限制：

  - 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。
  - DWS权限集授权时，如果给某一用户赋予了DWS数据源某个Schema下的全表权限（即将权限中的数据表配置为*号），则该用户具备对该Schema下的所有表的相应权限。但由于DWS自身权限特性限制，这些赋予的权限仅针对当前已有的表；而对于权限同步后再创建的新表（以下简称未来表），该用户依然没有权限，需要在角色/权限集中再次手动进行 权限同步后，才能确保该用户具备未来表的相应权限。

    为了解决未来表权限需要手动同步的问题，您可以通过未来表权限为指定Schema配置未来表的建表用户。当这些用户在指定Schema下创建未来表时，当前实例下所有对该Schema拥有全表权限的用户，将自动获得对所创建未来表的相应权限。
  - DLI权限集同步时会将权限由IAM创建自定义策略绑定到用户/用户组中。IAM最多可创建自定义策略200条，同步前请确保配额充足。
  - 进行权限同步时，需要为dlg_agency委托配置相关权限，请参考**授权dlg_agency委托**。
- 当前数据权限管控为白名单机制，是在待授权用户原有权限的基础上增加允许操作条件，不会影响用户的原有权限。如果仅需要当前数据权限管控所赋予的权限生效，则需要您手动去除待授权用户的原有权限。详见**数据权限管控说明**。
- 空间权限集删除后将被转移至回收站中，您可以在30天内进行还原，在回收站中超过30天的数据将被自动删除。详见**管理回收站**章节。

- 默认在DataArts Studio数据开发组件执行脚本、测试运行作业时，数据源（此处指MRS/DWS数据源）会使用数据连接上的账号进行认证鉴权。因此在数据开发时，权限管控依然无法生效。需要您启用细粒度认证，使得在数据开发执行脚本、测试运行作业时，使用当前用户身份认证鉴权，从而做到实现不同用户具有不同的数据权限，使角色/权限集中的权限管控生效。

创建空间权限集

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“空间权限集”，进入空间权限集页面。

步骤3 在“空间权限集”页面单击“新建”，创建权限集。

图 12-28 创建空间权限集

步骤4 新建空间权限集配置请参考表12-4，参数配置完成单击“确定”即可。

表 12-4 新建空间权限集参数设置

参数名	参数设置
*权限集名称	标识权限集，实例下唯一。 建议名称中包含含义，避免无意义的描述，以便于快速识别所需权限集。

参数名	参数设置
*管理员	选择管理员。当前权限集管理员支持最多选5个，且管理员类型必须同为用户或者用户组。 管理员为当前权限集的负责人，具有配置当前权限集内权限的能力。管理员职能范围： • 权限配置：为权限集分配数据源权限。 • 用户配置：将当前集合内权限分配给用户、用户组或工作空间角色。 • 创建权限集：基于当前权限集新建权限集和角色，新建权限集的权限不会大于当前权限集。
描述	为更好地识别权限集，此处加以描述信息。

图 12-29 创建空间权限集配置

----结束

配置空间权限集

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“空间权限集”，进入空间权限集页面。

步骤3 在“空间权限集”页面，找到需要配置的空间权限集，单击权限集名称进入详情页面。

图 12-30 进入空间权限集详情

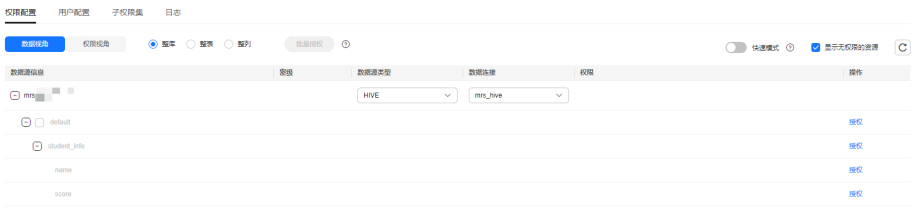
步骤4 **基本信息**：在空间权限集详情页面，基本信息区域可以查看空间权限集名称、ID、管理员等信息，详见图12-31。

图 12-31 空间权限集基本信息

步骤5 **权限配置**：在权限集详情页面，权限配置页签默认展示数据视角，可手动切换到权限视角。在这两种视角下，配置的权限数据是互通的，差异仅为展示视角的不同，推荐您使用权限视角进行批量授权。

- 数据视角：数据视角下，系统从数据的角度为您提供权限配置入口，当前仅支持MRS数据源。

图 12-32 数据视角权限配置



配置权限时，您可以选择“整库”、“整表”或“整列”等层级，然后在数据源信息中勾选对应层级，进行批量授权。另外，也可以在展开的导航树中，单击对应数据操作列中的“授权”，进行单一授权。

数据视图授权时，系统也提供了“快速模式”和“显示无权限的资源”功能。开启快速模式的情况下，库表列的元数据会从数据目录获取，否则会从数据源获取元数据。已完成元数据采集的场景下推荐开启快速模式。

说明

- 值得注意的是，库、表、列的权限是分层管理的，例如仅授予库权限后，则被授权用户对表和列依然是无权限的，如需对表或列授权，要再次按照对应层级进行授权。

  例如，选择数据库授权，当手动填写数据表的表名、或者填写“*”作为通配符时，此授权实际为对表进行授权；当手动填写数据列名、或者填写“*”作为通配符时，此授权实际为对列进行授权。
- 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。

图 12-33 数据视角授权

- 权限视角：权限视角下，系统从权限的角度为您提供权限配置入口。

  配置权限时，您需要直接单击“新建”，然后依次选择数据层级，进行权限配置。在权限视角下，同一层级（例如数据库、数据表或数据列）不允许选择多个对象进行批量授权。当前权限类型暂不支持选择为“禁止”。

说明

- 值得注意的是，库、表、列的权限是分层管理的，例如仅授予库权限后，则被授权用户对表和列依然是无权限的，如需对表或列授权，要再次按照对应层级进行授权。

  例如，选择数据库授权，当手动填写数据表的表名、或者填写“*”作为通配符时，此授权实际为对表进行授权；当手动填写数据列名、或者填写“*”作为通配符时，此授权实际为对列进行授权。
- 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。
- MRS Hive授权时，数据库可修改为URL，用于为存算分离场景下的OBS路径授权。存算分离场景下，使用Hive额外所需如下URL权限：
  - 创建库：write
  - 权限创建表/写入数据/删除表：read权限
- DWS授权时，数据库可修改为逻辑集群，用于为DWS数据源开启逻辑集群功能后的授权。逻辑集群场景下，使用DWS额外所需如下逻辑集群权限：
  - 允许在子集群中创建表对象：create
  - 允许访问子集群下的表对象：usage
  - 允许用户在具有compute权限的计算子集群上进行弹性计算：compute

配置权限后，在权限视角下支持您对所配置的权限进行编辑、同步或删除等操作。

图 12-34 权限视角权限配置

步骤6 用户配置：在权限集详情页面，单击“用户配置”进入用户配置页签。

用户配置的含义即为将权限配置中定义的数据权限，与此处的用户绑定起来。您可以单击“添加”，按照用户或用户组（当前暂不支持选择“工作空间角色”）的维度将用户添加到权限集中。其中的用户和用户组来自于当前工作空间中已添加的用户和用户组。

图 12-35 用户配置

步骤7 **子权限集**：在权限集详情页面，单击“子权限集”进入子权限集页签。

在子权限集页签，您可以查看到当前权限集下的子权限集。

图 12-36 查看子权限集

步骤8 **日志**：在权限集详情页面，单击“日志”进入日志页签。

在日志页签，当权限同步失败后，您可以查看到日志详情。系统每天0点定时删除30天前的日志。

图 12-37 查看日志

步骤9 权限集配置完成后，权限管控并不会直接生效。需要您手动将权限同步到数据源中，同步成功后权限管控才能生效，详见**同步权限集**。

实际上，由于空间权限集主要用于确定工作空间权限范围，而非权限管控，因此一般无需同步空间权限集，实际使用中推荐通过**配置角色**进行权限管控。

----结束

相关操作

- 同步空间权限集：空间权限集需要手动同步到数据源中权限管控才能生效。但由于空间权限集主要用于确定工作空间权限范围，而非权限管控，因此一般无需同步空间权限集，推荐通过**配置角色**进行权限管控。

  如需同步空间权限集，则在空间权限集页面，单击列表中对应权限集操作栏中的“同步”，即可将权限集同步至数据源。当需要批量同步时，可以在勾选权限集后，在列表上方单击“同步”。
- 编辑空间权限集：在空间权限集页面，单击列表中对应权限集操作栏中的“编辑”，即可修改权限集名称、管理员、描述信息。
- 删除空间权限集：在空间权限集页面，单击列表中对应权限集操作栏中的“删除”，在弹窗中再次确认后，即可删除权限集。当需要批量删除时，可以在勾选权限集后，在列表上方单击“删除”。

  注意，已配置权限、用户或有子权限集的空间权限集不可删除。如需删除应先清理空间权限集的相关配置。

  说明

  空间权限集删除后将被转移至回收站中，您可以在30天内进行还原，在回收站中超过30天的数据将被自动删除。详见**管理回收站**章节。

12.3.5.2 配置权限集

在数据访问权限管理的实际场景下，通常会有一级部门、二级部门、三级部门等多级权限的划分。为此，数据安全组件提供了自上而下分层式的数据权限管理方式。您可以通过空间权限集配置工作空间内的最大权限，在此基础上，将其向下拆分出新的子权限集，提供进一步的细分权限管理。

权限集本质上是将用户与权限直接关联。其中的空间权限集为没有父权限集的特殊权限集，限定了整个工作空间可访问的权限范围。在此之下定义的权限集均有其对应的父权限集，权限也为其父权限集的子集。

空间权限集和权限集在配置上都是将用户与权限直接关联，二者使用上的区别在于：

- 空间权限集是没有父权限集的顶层权限集，一般每个工作空间下创建一个即可；而权限集必须关联一个空间权限集或其他权限集作为其父权限集，可以新建多个，用于给不同使用场景的用户关联不同的权限。
- 空间权限集主要用于确定工作空间权限范围，而权限集主要用于权限管控。即空间权限集一般无需进行权限同步，且不支持为空间权限集关联角色；而权限集可通过权限同步进行权限管控（实际使用时，更推荐通过权限集关联角色进行权限管控）。

本章主要描述如何通过**创建权限集**和**配置权限集**进行权限管控，在实际使用中更加推荐您通过**配置角色**进行权限管控。

前提条件

- 配置权限集前，已完成空间权限集的配置，请参考**配置空间权限集**。
- 如果希望在权限配置时能够展示数据连接中数据库、表以及字段等元数据提示信息，则需要在数据目录组件，对数据表成功进行过元数据采集，详见**元数据采集任务**。

约束与限制

- DAYU Administrator、Tenant Administrator、数据安全管理员和父权限集管理员可以创建、修改或同步权限集，权限集管理员支持同步空间权限集，其他普通用户无权限操作。
- 当前通过权限集管控权限时，仅支持DLI、MRS Hive和DWS数据源。
- 权限集权限配置中，特殊情况下可能会出现子权限集权限超出父权限集范围的情况。例如当子权限集已配置某条权限记录后，父权限集中再删除此权限，会导致出现此情况，当前不支持级联删除权限。
- 权限集配置完成后，权限管控并不会直接生效，而是需要将权限集手动同步到数据源后，权限管控才能生效。

  由于角色管理基于权限集提供了更加直观、强大的权限管控能力，因此除DLI数据源外，一般无需同步权限集，实际使用中推荐通过**配置角色**进行权限管控。如果需要同步，则需注意以下限制：

  - 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。
  - DWS权限集授权时，如果给某一用户赋予了DWS数据源某个Schema下的全表权限（即将权限中的数据表配置为*号），则该用户具备对该Schema下的所有表的相应权限。但由于DWS自身权限特性限制，这些赋予的权限仅针对当前已有的表；而对于权限同步后再创建的新表（以下简称未来表），该用户依然没有权限，需要在角色/权限集中再次手动进行 权限同步后，才能确保该用户具备未来表的相应权限。

为了解决未来表权限需要手动同步的问题，您可以通过未来表权限为指定Schema配置未来表的建表用户。当这些用户在指定Schema下创建未来表时，当前实例下所有对该Schema拥有全表权限的用户，将自动获得对所创建未来表的相应权限。

- DLI权限集同步时会将权限由IAM创建自定义策略绑定到用户/用户组中。IAM最多可创建自定义策略200条，同步前请确保配额充足。
- 进行权限同步时，需要为dlg_agency委托配置相关权限，请参考**授权dlg_agency委托**。

- 当前数据权限管控为白名单机制，是在待授权用户原有权限的基础上增加允许操作条件，不会影响用户的原有权限。如果仅需要当前数据权限管控所赋予的权限生效，则需要您手动去除待授权用户的原有权限。详见**数据权限管控说明**。
- 默认在DataArts Studio数据开发组件执行脚本、测试运行作业时，数据源（此处指MRS/DWS数据源）会使用数据连接上的账号进行认证鉴权。因此在数据开发时，权限管控依然无法生效。需要您启用细粒度认证，使得在数据开发执行脚本、测试运行作业时，使用当前用户身份认证鉴权，从而做到实现不同用户具有不同的数据权限，使角色/权限集中的权限管控生效。

创建权限集

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“权限集”，进入权限集页面。

步骤3 在“权限集”页面单击“新建”，创建权限集。

图 12-38 创建权限集

步骤4 新建权限集配置请参考**表12-5**，参数配置完成单击“确定”即可。

表 12-5 参数设置

参数名	参数设置
*权限集名称	标识权限集，实例下唯一。 建议名称中包含含义，避免无意义的描述，以便于快速识别所需权限集。
*父权限集	选择对应的父权限集，父权限集可以是空间权限集或其他权限集。注意选择父权限集后，当前权限集的权限也为其父权限集的子集。
*管理员	管理员为当前权限集的负责人，具有配置当前权限集内权限的能力。 管理员职能范围： • 权限配置：为权限集分配数据源权限。 • 用户配置：将当前集合内权限分配给用户、用户组或工作空间角色。 • 创建权限集：基于当前权限集新建权限集和角色，新建权限集的权限不会大于当前权限集。
描述	为更好地识别权限集，此处加以描述信息。

图 12-39 创建权限集配置

----结束

配置权限集

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“权限集”，进入权限集页面。

步骤3 在“权限集”页面，找到需要配置的权限集，单击权限集名称进入详情页面。

图 12-40 进入权限集详情

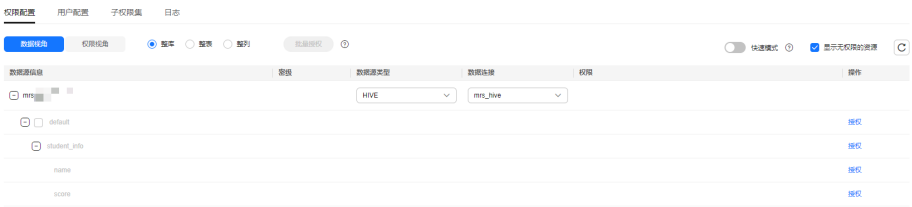
步骤4 **基本信息**：在权限集详情页面，基本信息区域可以查看权限集名称、ID、管理员等信息，详见图12-41。

图 12-41 权限集基本信息

步骤5 **权限配置**：在权限集详情页面，权限配置页签默认展示数据视角，可手动切换到权限视角。在这两种视角下，配置的权限数据是互通的，差异仅为展示视角的不同，推荐您使用权限视角进行批量授权。

- 数据视角：数据视角下，系统从数据的角度为您提供权限配置入口（当前仅支持MRS数据源）。在授权时可选的数据范围为父权限集中已授权的数据。

图 12-42 数据视角权限配置



配置权限时，您可以选择“整库”、“整表”或“整列”等层级，然后在数据源信息中勾选对应层级，进行批量授权。另外，也可以在展开的导航树中，单击对应数据操作列中的“授权”，进行单一授权。

数据视图授权时，系统也提供了“快速模式”和“显示无权限的资源”功能。开启快速模式的情况下，库表列的元数据会从数据目录获取，否则会从数据源获取元数据。已完成元数据采集的场景下推荐开启快速模式。

说明

- 值得注意的是，库、表、列的权限是分层管理的，例如仅授予库权限后，则被授权用户对表和列依然是无权限的，如需对表或列授权，要再次按照对应层级进行授权。

  例如，选择数据库授权，当手动填写数据表的表名、或者填写“*”作为通配符时，此授权实际为对表进行授权；当手动填写数据列名、或者填写“*”作为通配符时，此授权实际为对列进行授权。
- 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。

图 12-43 数据视角授权

- 权限视角：权限视角下，系统从权限的角度为您提供权限配置入口。在授权时可选的数据范围为父权限集中已授权的数据。

  配置权限时，您需要直接单击“新建”，然后依次选择数据层级，进行权限配置。在权限视角下，同一层级（例如数据库、数据表或数据列）不允许选择多个对象进行批量授权。当前权限类型暂不支持选择为“禁止”。

说明

- 值得注意的是，库、表、列的权限是分层管理的，例如仅授予库权限后，则被授权用户对表和列依然是无权限的，如需对表或列授权，要再次按照对应层级进行授权。

  例如，选择数据库授权，当手动填写数据表的表名、或者填写“*”作为通配符时，此授权实际为对表进行授权；当手动填写数据列名、或者填写“*”作为通配符时，此授权实际为对列进行授权。
- 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。
- MRS Hive授权时，数据库可修改为URL，用于为存算分离场景下的OBS路径授权。存算分离场景下，使用Hive额外所需如下URL权限：
  - 创建库：write
  - 权限创建表/写入数据/删除表：read权限
- DWS授权时，数据库可修改为逻辑集群，用于为DWS数据源开启逻辑集群功能后的授权。逻辑集群场景下，使用DWS额外所需如下逻辑集群权限：
  - 允许在子集群中创建表对象：create
  - 允许访问子集群下的表对象：usage
  - 允许用户在具有compute权限的计算子集群上进行弹性计算：compute

配置权限后，在权限视角下支持您对所配置的权限进行编辑、同步或删除等操作。

图 12-44 权限视角权限配置

步骤6 用户配置：在权限集详情页面，单击“用户配置”进入用户配置页签。

用户配置的含义即为将权限配置中定义的数据权限，与此处的用户绑定起来。您可以单击“添加”，按照用户或用户组（当前暂不支持选择“工作空间角色”）的维度将用户添加到权限集中。其中的用户和用户组来自于当前工作空间中已添加的用户和用户组。

图 12-45 用户配置

步骤7 **子权限集**：在权限集详情页面，单击“子权限集”进入子权限集页签。

在子权限集页签，您可以查看到当前权限集下的子权限集。

图 12-46 查看子权限集

步骤8 **日志**：在权限集详情页面，单击“日志”进入日志页签。

在日志页签，当权限同步失败后，您可以查看到日志详情。系统每天0点定时删除30天前的日志。

图 12-47 查看日志

步骤9 权限集配置完成后，并不会直接生效。需要您将权限集手动同步到数据源中，同步成功后权限管控才能生效，详见**同步权限集**。

但由于角色管理基于权限集提供了更加直观、强大的权限管控能力，因此一般无需同步空间权限集，实际使用中推荐通过**配置角色**进行权限管控。

----结束

相关操作

- 同步权限集：权限集需要同步到数据源中权限管控才能生效。但由于角色管理基于权限集提供了更加直观、强大的权限管控能力，因此一般无需同步权限集，实际使用中推荐通过**配置角色**进行权限管控。

  如需同步权限集，则在权限集页面，单击列表中对应权限集操作栏中的“同步”，即可将权限集同步至数据源。当需要批量同步时，可以在勾选权限集后，在列表上方单击“同步”。
- 编辑权限集：在权限集页面，单击列表中对应权限集操作栏中的“编辑”，即可修改权限集名称、管理员、描述信息。
- 删除权限集：在权限集页面，单击列表中对应权限集操作栏中的“删除”，在弹窗中再次确认后，即可删除权限集。当需要批量删除时，可以在勾选权限集后，在列表上方单击“删除”。

  注意，已配置权限、用户或有子权限集的权限集不可删除。如需删除应先清理权限集的相关配置。

  说明

  权限集删除后将被转移至回收站中，您可以在30天内进行还原，在回收站中超过30天的数据将被自动删除。详见**管理回收站**章节。

12.3.5.3 配置角色

数据安全中的角色管理，本质上是基于权限集提供的更加直观、强大的权限管控能力。角色与权限集的不同之处在于，权限集是将用户与权限直接关联，而角色是通过在数据源上创建或纳管一个角色，进而承载用户和权限之间的关联关系。

当您在角色管理页面，为权限集关联了角色之后，权限就不再同步到用户，而是只同步到角色。推荐您通过角色管理这种方式更加直观地管理权限关系、进行权限管控，角色管理还支持使用纳管角色管理已有的数据源权限。

- 通用角色：在数据源上创建新角色，用于承载用户和权限之间的关联关系。
- 纳管角色：纳管MRS数据源上已有的角色（可登录MRS FusionInsight Manager，选择“系统 > 权限 > 角色”查看），继承已有角色的MRS数据源权限。

本章主要描述如何**配置通用角色**，**配置纳管角色**以及**相关操作**。

前提条件

- 配置角色前，已完成空间权限集的配置，请参考**配置空间权限集**。
- MRS和DWS角色同步时，系统通过管理中心组件数据连接中的用户进行账号相关的增删改查等操作，因此对数据连接中的用户有以下权限要求：
  - MRS Ranger连接中的用户需具备Ranger组件Admin权限。
  - DWS连接中的数据库用户，在非三权分立模式下至少需具备数据库dbadmin权限，三权分立模式下需具备系统管理员权限。

  配置方法详见**检查集群版本与权限**。
- 如果希望在快速模式下权限配置时能够展示数据连接中数据库、表以及字段等元数据提示信息，则需要在数据目录组件，对数据表成功进行过元数据采集，详见**元数据采集任务**。

约束与限制

- 当前仅支持为MRS和DWS集群创建角色。
- 由于空间权限集主要用于确定工作空间权限范围，而非权限管控，因此不支持对空间权限集添加创建角色。
- 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。
- DWS权限集授权时，如果给某一用户赋予了DWS数据源某个Schema下的全表权限（即将权限中的数据表配置为*号），则该用户具备对该Schema下的所有表的相应权限。但由于DWS自身权限特性限制，这些赋予的权限仅针对当前已有的表；而对于权限同步后再创建的新表（以下简称未来表），该用户依然没有权限，需要在角色/权限集中再次手动进行 权限同步后，才能确保该用户具备未来表的相应权限。

  为了解决未来表权限需要手动同步的问题，您可以通过未来表权限为指定Schema配置未来表的建表用户。当这些用户在指定Schema下创建未来表时，当前实例下所有对该Schema拥有全表权限的用户，将自动获得对所创建未来表的相应权限。
- 当为权限集创建了角色之后，权限就不再同步到用户，而是只同步到角色。
- 仅当数据连接中的Agent选择的CDM集群为2.10.0.300及以上版本时，才支持角色管理。
- MRS和DWS角色同步时，系统通过管理中心组件数据连接中的用户进行账号相关的增删改查等操作，因此对数据连接中的用户有以下权限要求：

  - MRS Ranger连接中的用户需具备Ranger组件Admin权限。
  - DWS连接中的数据库用户，在非三权分立模式下至少需具备数据库dbadmin权限，三权分立模式下需具备系统管理员权限。

  配置方法详见**检查集群版本与权限**。
- 角色中的目录权限仅展示该空间下所指定角色在集群上的目录权限。
- 进行权限同步时，需要为dlg_agency委托配置相关权限，请参考**授权dlg_agency委托**。
- 当前数据权限管控为白名单机制，是在待授权用户原有权限的基础上增加允许操作条件，不会影响用户的原有权限。如果仅需要当前数据权限管控所赋予的权限生效，则需要您手动去除待授权用户的原有权限。详见**数据权限管控说明**。
- 默认在DataArts Studio数据开发组件执行脚本、测试运行作业时，数据源（此处指MRS/DWS数据源）会使用数据连接上的账号进行认证鉴权。因此在数据开发时，权限管控依然无法生效。需要您启用细粒度认证，使得在数据开发执行脚本、测试运行作业时，使用当前用户身份认证鉴权，从而做到实现不同用户具有不同的数据权限，使角色/权限集中的权限管控生效。

配置通用角色

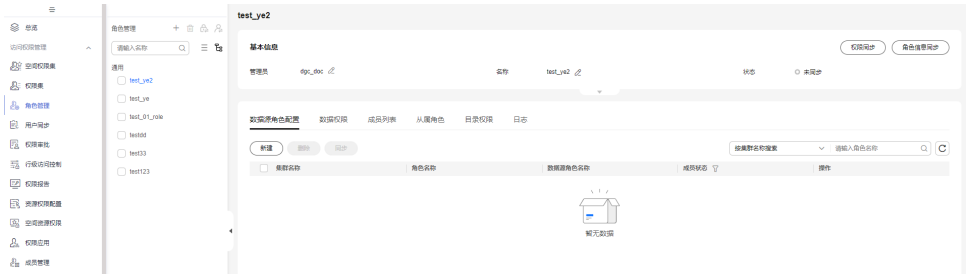
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“角色管理”，进入角色管理页面。

步骤3 您可以通过以下两种方式之一，进入配置通用角色入口。

- 已有角色：在“角色管理”页面，角色管理导航树上会默认展示已创建的权限集（详见**创建权限集**）作为通用角色。您可以单击角色名，进入角色详情配置页面。

  图 12-48 进入角色详情

- 新建角色：在“角色管理”页面，在角色管理导航树单击+，选择“创建通用角色”。参考**表12-6**完成通用角色创建，配置完成单击“确定”，系统默认进入新建的角色详情配置页面。

  表 12-6 参数设置

参数名	参数设置
*权限集名称	标识权限集，实例下唯一。 建议名称中包含含义，避免无意义的描述，以便于快速识别所需权限集。

参数名	参数设置
*父权限集	选择对应的父权限集，父权限集可以是空间权限集或其他权限集。注意选择父权限集后，当前权限集的权限也为其父权限集的子集。
*管理员	管理员为当前权限集的负责人，具有配置当前权限集内权限的能力。管理员职能范围： – 权限配置：为权限集分配数据源权限。 – 用户配置：将当前集合内权限分配给用户、用户组或工作空间角色。 – 创建权限集：基于当前权限集新建权限集和角色，新建权限集的权限不会大于当前权限集。
描述	为更好地识别权限集，此处加以描述信息。

图 12-49 创建通用角色

步骤4 **基本信息**：在角色详情页面，展开基本信息区域可以查看角色名称、ID、管理员等信息，详见图12-50。

另外，还可以在配置完角色和权限后，通过右上角的“权限同步”和“角色信息同步”进行同步。

图 12-50 角色基本信息

步骤5 **数据源角色配置**：在角色详情页面的数据源角色配置页签，可通过“新建”在数据源上创建新角色，用于承载用户和权限之间的关联关系。

图 12-51 数据源角色配置页签

单击“新建”，系统在弹出的窗口中展示数据源的信息，您需要勾选所需配置的数据源并填写“角色名”，然后单击“确定”，即可完成角色创建。

图 12-52 新建数据源角色

如果后续不再需要数据源角色，可以通过列表操作栏中的“删除”删除数据源中的角色。删除后权限同步就不再同步到角色，而是只同步到用户信息。

步骤6 **数据权限**：在角色详情页面，单击“数据权限”进入数据权限页签。数据权限页签默认展示数据视角，可手动切换到权限视角。在这两种视角下，配置的权限数据是互通的，差异仅为展示视角的不同，推荐您使用权限视角进行批量授权。

- 数据视角：数据视角下，系统从数据的角度为您提供权限配置入口，当前仅支持MRS数据源。

  图 12-53 数据视角权限配置

  配置权限时，您可以选择“整库”、“整表”或“整列”等层级，然后在数据源信息中勾选对应层级，进行批量授权。另外，也可以在展开的导航树中，单击对应数据操作列中的“授权”，进行单一授权。

  数据视图授权时，系统也提供了“快速模式”和“显示无权限的资源”功能。开启快速模式的情况下，库表列的元数据会从数据目录获取，否则会从数据源获取元数据。已完成元数据采集的场景下推荐开启快速模式。

说明

- 值得注意的是，库、表、列的权限是分层管理的，例如仅授予库权限后，则被授权用户对表和列依然是无权限的，如需对表或列授权，要再次按照对应层级进行授权。

  例如，选择数据库授权，当手动填写数据表的表名、或者填写“*”作为通配符时，此授权实际为对表进行授权；当手动填写数据列名、或者填写“*”作为通配符时，此授权实际为对列进行授权。
- 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。

图 12-54 数据视角授权

- 权限视角：权限视角下，系统从权限的角度为您提供权限配置入口。

  配置权限时，您需要直接单击“新建”，然后依次选择数据层级，进行权限配置。在权限视角下，同一层级（例如数据库、数据表或数据列）不允许选择多个对象进行批量授权。当前权限类型暂不支持选择为“禁止”。

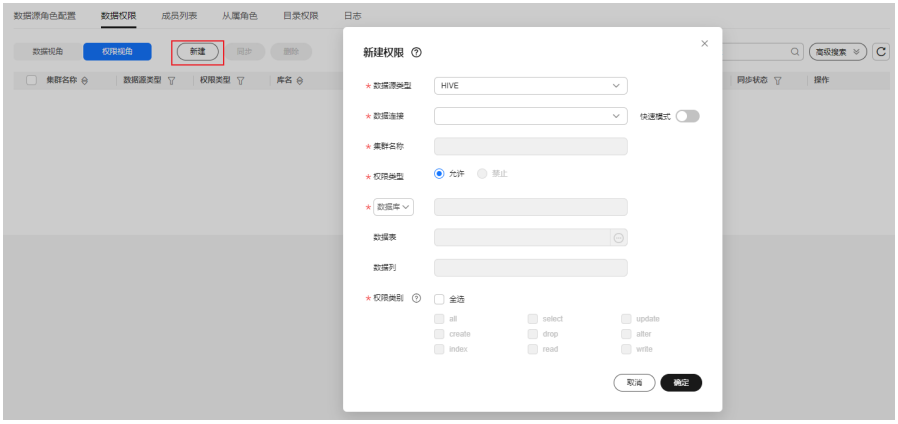
说明

- 值得注意的是，库、表、列的权限是分层管理的，例如仅授予库权限后，则被授权用户对表和列依然是无权限的，如需对表或列授权，要再次按照对应层级进行授权。

  例如，选择数据库授权，当手动填写数据表的表名、或者填写“*”作为通配符时，此授权实际为对表进行授权；当手动填写数据列名、或者填写“*”作为通配符时，此授权实际为对列进行授权。
- 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。
- MRS Hive授权时，数据库可修改为URL，用于为存算分离场景下的OBS路径授权。存算分离场景下，使用Hive额外所需如下URL权限：
  - 创建库：write
  - 权限创建表/写入数据/删除表：read权限
- DWS授权时，数据库可修改为逻辑集群，用于为DWS数据源开启逻辑集群功能后的授权。逻辑集群场景下，使用DWS额外所需如下逻辑集群权限：
  - 允许在子集群中创建表对象：create
  - 允许访问子集群下的表对象：usage
  - 允许用户在具有compute权限的计算子集群上进行弹性计算：compute

配置权限后，在权限视角下支持您对所配置的权限进行编辑、同步或删除等操作。

图 12-55 权限视角权限配置



步骤7 成员列表：在角色详情页面，单击“成员列表”进入成员列表页签。

成员列表的含义即为将数据源角色配置中的角色与此处的用户关联起来。您可以单击“添加”，按照用户、用户组或工作空间角色的维度将用户添加到角色中。其中的用户和用户组来自于当前工作空间中已添加的用户和用户组。

图 12-56 成员列表

步骤8 **从属角色**：在角色详情页面，单击“从属角色”进入从属角色页签。

在从属角色页签，您可以查看到当前角色的子角色。

图 12-57 查看从属角色

步骤9 **目录权限**：在角色详情页面，单击“目录权限”进入目录权限页签。

目录权限通过从Ranger组件获取对应角色的HDFS策略，从而显示该角色具有权限的HDFS路径，并支持查看对该路径有哪些操作权限。如果想查询某路径下的权限，则可以使用搜索功能进行查看，注意当前仅支持精确匹配。

图 12-58 查看目录权限

步骤10 **日志**：在角色详情页面，单击“日志”进入日志页签。

在日志页签，当权限同步失败后，您可以查看到日志详情。系统每天0点定时删除30天前的日志。

图 12-59 查看日志

步骤11 角色配置完成后，并不会直接生效。需要您将权限和角色手动同步到数据源中，同步成功后权限控制才能生效，详见**相关操作**。

----结束

配置纳管角色

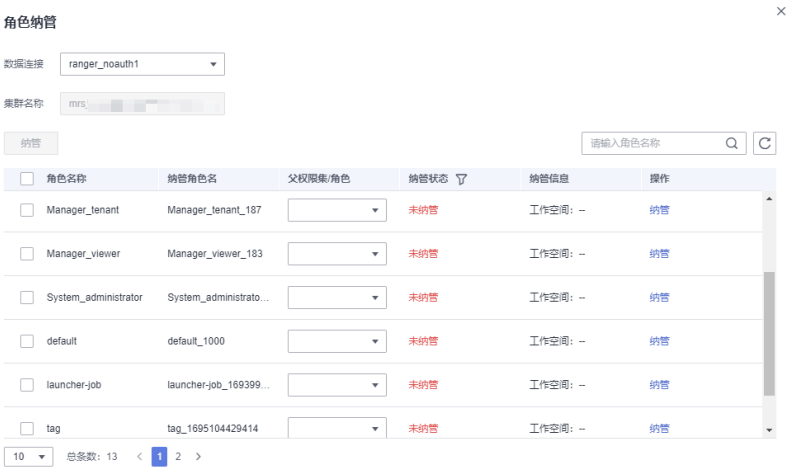
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“角色管理”，进入角色管理页面。

步骤3 在“角色管理”页面，在角色管理导航树单击+，选择“创建纳管角色”。在弹窗中选择已创建的Ranger连接，您需要在选择“父权限集/角色”后，单击所需纳管MRS角色操作栏中的“纳管”，完成纳管角色的创建。也可以在勾选多个所需纳管MRS角色后，单击列表上方“纳管”进行批量创建。

如果后续不再需要纳管角色，可以直接在角色管理导航树删除纳管角色，即可解除纳管角色。解除后权限同步就不再同步到角色，而是只同步到用户信息。

图 12-60 创建纳管角色



步骤4 关闭角色纳管弹窗，返回“角色管理”页面。在角色管理导航树上找到上一步中纳管的MRS角色，单击角色名，进入角色详情配置页面。

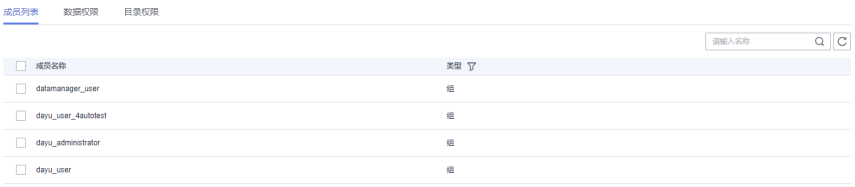
步骤5 **基本信息**：在角色详情页面，展开基本信息区域可以查看角色名称、ID、管理员等信息，详见图12-61。

另外，还可以在配置完角色和权限后，通过右上角的“权限同步”和“角色信息同步”进行同步。

图 12-61 角色基本信息

步骤6 **成员列表**：在角色详情页面的成员列表页签，可以查看当前MRS角色所关联的用户或用户组。纳管角色暂不支持在数据安全侧添加用户。

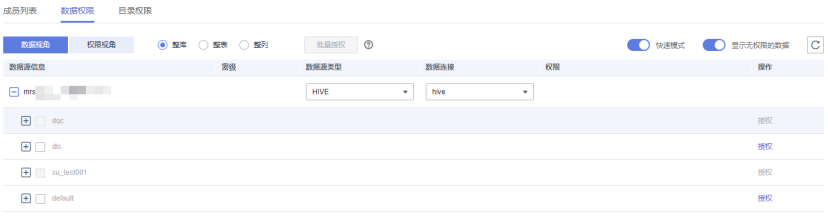
图 12-62 成员列表



步骤7 **数据权限**：在角色详情页面，单击“数据权限”进入数据权限页签。数据权限页签默认展示数据视角，可手动切换到权限视角。在这两种视角下，配置的权限数据是互通的，差异仅为展示视角的不同，推荐您使用权限视角进行批量授权。

- 数据视角：数据视角下，系统从数据的角度为您提供权限配置入口。如果已成功运行过元数据采集任务（详见**元数据采集任务**），则可以直接查看到数据源信息，单击可展开导航树。

图 12-63 数据视角权限配置



配置权限时，您可以选择“整库”、“整表”或“整列”等层级，然后在数据源信息中勾选对应层级，进行批量授权。另外，也可以在展开的导航树中，单击对应数据操作列中的“授权”，进行单一授权。

数据视图授权时，系统也提供了“快速模式”和“显示无权限的资源”功能。开启快速模式的情况下，库表列的元数据会从数据目录获取，否则会从数据源获取元数据。已完成元数据采集的场景下推荐开启快速模式。

说明

- 值得注意的是，库、表、列的权限是分层管理的，例如仅授予库权限后，则被授权用户对表和列依然是无权限的，如需对表或列授权，要再次按照对应层级进行授权。

  例如，选择数据库授权，当手动填写数据表的表名、或者填写“*”作为通配符时，此授权实际为对表进行授权；当手动填写数据列名、或者填写“*”作为通配符时，此授权实际为对列进行授权。
- 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。

图 12-64 数据视角授权

- 权限视角：权限视角下，系统从权限的角度为您提供权限配置入口。

  配置权限时，您需要直接单击“新建”，然后依次选择数据层级，进行权限配置。在权限视角下，同一层级（例如数据库、数据表或数据列）不允许选择多个对象进行批量授权。当前权限类型暂不支持选择为“禁止”。

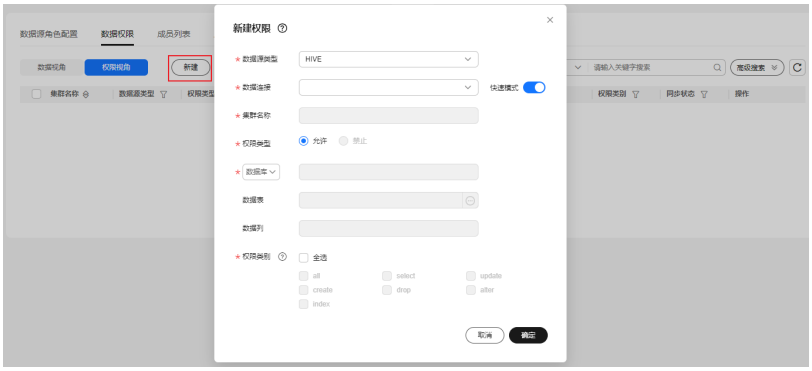
  说明

  - 值得注意的是，库、表、列的权限是分层管理的，例如仅授予库权限后，则被授权用户对表和列依然是无权限的，如需对表或列授权，要再次按照对应层级进行授权。

    例如，选择数据库授权，当手动填写数据表的表名、或者填写“*”作为通配符时，此授权实际为对表进行授权；当手动填写数据列名、或者填写“*”作为通配符时，此授权实际为对列进行授权。
  - 进行授权时，授权对象名（库表列名）当前仅支持包含数字、英文、下划线、中划线和通配符*，暂不支持中文以及其他特殊字符。
  - MRS Hive授权时，数据库可修改为URL，用于为存算分离场景下的OBS路径授权。存算分离场景下，使用Hive额外所需如下URL权限：
    - 创建库：write
    - 权限创建表/写入数据/删除表：read权限
  - DWS授权时，数据库可修改为逻辑集群，用于为DWS数据源开启逻辑集群功能后的授权。逻辑集群场景下，使用DWS额外所需如下逻辑集群权限：
    - 允许在子集群中创建表对象：create
    - 允许访问子集群下的表对象：usage
    - 允许用户在具有compute权限的计算子集群上进行弹性计算：compute

  配置权限后，在权限视角下支持您对所配置的权限进行编辑、同步或删除等操作。

图 12-65 权限视角权限配置



步骤8 **目录权限**：在角色详情页面，单击“目录权限”进入目录权限页签。

目录权限通过从Ranger组件获取对应角色的HDFS策略，从而显示该角色具有权限的HDFS路径，并支持查看对该路径有哪些操作权限。如果想查询某路径下的权限，则可以使用搜索功能进行查看，注意当前仅支持精确匹配。

图 12-66 查看目录权限

步骤9 纳管角色的权限配置完成后，并不会直接生效。需要您将权限手动同步到Ranger组件中，同步成功后权限控制才能生效，详见**同步权限**。

----结束

相关操作

- 同步权限：在角色管理中，配置数据权限后需要同步权限到数据源中权限管控才能生效。

  您可以在角色详情页面，单击基本信息区域右上角的“权限同步”进行同步。当需要批量同步时，可以在角色管理导航树上勾选角色后，在导航树上方单击进行权限同步。
- 同步角色：在通用角色管理（纳管角色无需同步角色）中，权限集配置角色后需要同步到数据源中权限管控才能生效。

  您可以在角色详情页面，单击基本信息区域右上角的“角色信息同步”，或数据源角色配置页签中列表操作栏的“同步”，进行角色信息同步。当需要批量同步时，可以在角色管理导航树上勾选角色后，在导航树上方单击进行角色信息同步。

说明

- 角色信息同步成功后，MRS数据源角色命名格式为“*角色名_时间戳*”，DWS据源角色命名格式为“dataarts_studio_role_*角色名*”。
- 同步角色到MRS集群的场景下，系统提示角色信息同步成功后，还需要等待约5分钟，直到Ranger组件自动触发并完成同步MRS集群角色后，权限管控才能生效。Ranger组件是否同步完成，可通过数据源角色配置页签中列表中的“数据源角色名称”确认：
  - 未完成同步的角色，数据源角色名称为：*角色名_10位时间戳*
  - 已完成同步的角色，数据源角色名称为：*角色名_13位时间戳*

- 删除角色：在角色管理导航树上勾选角色后，在导航树上方单击，在弹窗中再次确认后，即可删除权限集。

  注意，通用角色中已配置角色、权限、用户或有子权限集时不可删除，如需删除应先清理相关配置。纳管角色中已配置权限时不可删除，如需删除应先清理相关配置。

  说明

  通用角色删除后将被转移至回收站中，您可以在30天内进行还原，在回收站中超过30天的数据将被自动删除。详见**管理回收站**章节。

12.3.5.4 管理成员

数据安全支持成员管理视图，支持查看当前工作空间内成员的权限，并进行角色/权限集管理。

前提条件

- 为成员添加或删除所在的角色/权限集前，已完成权限集或角色的配置，请参考**配置权限集**或**配置角色**。

约束与限制

- 仅DAYU Administrator、Tenant Administrator、数据安全管理员或者角色/权限集管理员可以为成员添加或删除所在的角色/权限集。
- 为成员添加或删除所在的角色时，仅支持通用角色，暂不支持纳管角色。
- 成员的权限来自于角色/权限集，需要角色/权限集同步成功，成员的权限才会生效。

查看策略及详情

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“成员管理”，进入成员管理页面。

图 12-67 进入成员管理

步骤3 单击操作栏中的“编辑”，在弹窗中可以为成员添加或删除所在的角色/权限集，管理其权限。

图 12-68 编辑角色/权限集

步骤4 单击操作栏中的“查看权限”，可以查看当前成员的基本信息，以及拥有的权限以及权限来源。

----结束

12.3.5.5 配置行级访问控制

在业务开发过程中，存在多个开发者共同访问和维护同一张DWS表的场景，需要针对不同开发者设置不同行数据的访问权限。在这种场景下，您可以配置行级访问控制策略，为不同开发者按照行数据进行授权。

在数据安全组件新建行级访问控制策略后，通过策略同步，会将行级访问控制策略同步到DWS，并自动开启该DWS表的行访问控制开关使该策略生效。

值得注意的是，行级访问控制策略为DataArts Studio实例级别配置，各工作空间之间数据互通，全局可见并生效。

前提条件

- 新建DWS行级访问控制策略前，已在管理中心创建数据仓库服务（DWS）类型的数据连接，请参考**创建DataArts Studio数据连接**。DWS数据连接中的账户要具备待控制表的GRANT权限（数据库对象创建后，默认只有对象所有者或者系统管理员可以通过GRANT命令将对象的权限授予其他用户）。
- 行级访问控制为指定用户/用户组在数据源上关联策略，因此需要先将IAM上的用户信息同步到数据源上，详见**同步IAM用户到数据源**。
- 如果希望在DataArts Studio数据开发执行脚本、测试运行作业时，使用当前用户身份认证鉴权以实现行级访问控制策略生效，则需要**启用细粒度认证**。
- 为确保行级访问控制策略生效，须确保策略中指定的用户已具备待控制操作的表权限，同时需要将表所属模式的USAGE权限授予该用户。可通过如下命令为指定的user1，user2，user3授权。

```
GRANT USAGE ON SCHEMA schema_name TO user1,user2,user3;
GRANT SELECT,UPDATE,DELETE ON TABLE table_name TO user1,user2,user3;
```

约束与限制

- 仅DAYU Administrator、Tenant Administrator用户或者数据安全管理员可以创建、修改或删除行级访问控制策略，其他普通用户无权限操作。
- 当前行级访问控制策略仅支持DWS数据源，且不支持DWS逻辑集群。DWS数据连接中的账户要具备待控制表的GRANT权限（数据库对象创建后，默认只有对象所有者或者系统管理员可以通过GRANT命令将对象的权限授予其他用户）。
- 行级访问控制为指定用户/用户组在数据源上关联策略，因此需要先将IAM上的用户信息同步到数据源上，详见**同步IAM用户到数据源**。
- 当前行级访问控制支持影响数据表的读取操作（SELECT、UPDATE、DELETE、ALL，ALL表示会影响SELECT、UPDATE、DELETE三种操作），暂不支持影响数据表的写入操作（INSERT、MERGE INTO）。
- 行级访问控制策略名称是针对表的，同一个数据表上不能有同名的行访问控制策略；对不同的数据表，可以有同名的行访问控制策略。
- 支持对行存表、行存分区表、列存表、列存分区表、复制表、unlogged表、hash表定义行级访问控制策略，不支持HDFS表、外表、临时表定义行级访问控制策略。
- 不支持对视图定义行级访问控制策略。
- 同一张表上可以创建多个行级访问控制策略，一张表最多创建100个行访级问控制策略。
- 具有DWS管理员权限的用户和初始运维用户(Ruby)不受行访问控制影响，可以查看表的全量数据。
- 通过SQL语句、视图、函数、存储过程查询包含行级访问控制策略的表，都会受影响。
- 同步行访级问控制策略后，不支持对行访问控制策略依赖的列进行类型修改。

创建行级访问控制策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“行级访问控制”，进入行级访问控制页面。

图 12-69 进入行级访问控制页面

步骤3 单击“新建”，进入新建行级访问控制策略页面，参数配置参考**表12-7**。

图 12-70 新建行级访问控制策略参数配置

创建行级访问控制策略参数配置说明：

表 12-7 配置策略参数

参数名	参数说明
*策略名称	行级访问控制策略的标识，同一个数据表上不能有同名的行访问控制策略。 为便于策略管理，建议名称中标明要控制的对象和内容规则。
*数据源类型	当前仅支持DWS数据源。
*工作空间	从下拉列表中选择数据连接所在工作空间，支持跨空间选择数据连接。注意，行级访问控制策略与工作空间之间无关联关系，工作空间仅用于关联数据连接。
*数据连接	从下拉列表中选择所选工作空间中已创建的DWS数据连接，若未创建请参考**创建DataArts Studio数据连接**新建连接。
*集群名称	无需选择，自动匹配数据连接中的数据源集群。
*数据库	选择行数据所在的数据库。

参数名	参数说明
*数据表	选择行数据所在的数据表。选择后系统自动展示所选的表结构。
*SQL操作	选择需要控制的操作（SELECT、UPDATE、DELETE、ALL，ALL表示会影响SELECT、UPDATE、DELETE三种操作），暂不支持控制写入操作（INSERT、MERGE INTO）。 ● 当选择SELECT时，SELECT类操作受行访问控制的影响，所选用户组/用户只能查看到满足表达式条件的行数据，受影响的操作包括SELECT，UPDATE ... RETURNING，DELETE ... RETURNING。 ● 当选择UPDATE时，UPDATE类操作受行访问控制的影响，所选用户组/用户只能更新满足表达式条件的行数据，受影响的操作包括UPDATE，UPDATE ... RETURNING，SELECT ... FOR UPDATE/SHARE。 ● 当选择DELETE时，DELETE类操作受行访问控制的影响，所选用户组/用户只能删除满足表达式条件的行数据，受影响的操作包括DELETE，DELETE ... RETURNING。
*用户组/用户	指定当前工作空间成员中的用户或用户组。 指定的用户或用户组按照所选的“SQL操作”进行操作时，只能操作满足“表达式”条件的行级数据。 ● 当选择SELECT时，SELECT类操作受行访问控制的影响，所选用户组/用户只能查看到满足表达式条件的行数据，受影响的操作包括SELECT，UPDATE ... RETURNING，DELETE ... RETURNING。 ● 当选择UPDATE时，UPDATE类操作受行访问控制的影响，所选用户组/用户只能更新满足表达式条件的行数据，受影响的操作包括UPDATE，UPDATE ... RETURNING，SELECT ... FOR UPDATE/SHARE。 ● 当选择DELETE时，DELETE类操作受行访问控制的影响，所选用户组/用户只能删除满足表达式条件的行数据，受影响的操作包括DELETE，DELETE ... RETURNING。
*表达式	填写行数据的表达式。只有满足表达式的行数据，才允许被指定用户按照所选的“SQL操作”进行操作。格式如下： `*目标字段*="*操作值*" 建议表达式目标字段以反引号包裹，操作值以双引号包裹，需要匹配多个行数据时，可以用AND拼接，例如： `role`="test" AND `department`="sales"

步骤4 单击“提交”，完成行级访问控制策略创建。行级访问控制策略创建完成后，需要手动单击“同步”，将该策略同步到数据源中。

----结束

相关操作

- 同步策略：在行级访问控制页面，单击对应任务操作栏中的“同步”，即可将该策略同步到数据源中。当需要批量同步时，可以在勾选策略后，在列表上方单击“同步”。

只有处于“同步成功”状态的策略才能生效。如果策略同步失败，可通过**查看策略详情**查看策略运行日志，通过日志排查同步失败原因。待问题修复后请重新同步，如果仍同步失败，请联系技术支持人员协助处理。

- 编辑策略：在行级访问控制页面，单击对应任务操作栏中的“编辑”，即可编辑行级访问控制策略。
- 删除策略：在行级访问控制页面，单击对应任务操作栏中的“删除”，即可删除策略。当需要批量删除时，可以在勾选策略后，在列表上方单击“删除”。

  说明

  删除操作无法撤销，请谨慎操作。
- 查看策略详情：在行级访问控制页面，找到需要查看的策略，单击策略名即可查看策略详情。

图 12-71 查看策略详情

12.3.5.6 同步 MRS Hive 和 Hetu 权限

在MRS Hetu对接MRS hive数据源并使用Ranger权限管控的场景下，通过Hetu访问同集群的Hive数据源，会统一使用Hetu端的Ranger权限做鉴权，而不受Hive端的Ranger权限管控。

为了避免该场景下需要在Hetu端重复配置Hive数据源权限的问题，您可以参考本章节内容配置hetu权限同步策略，使Hive权限自动同步至Hetu端，增强权限管理一致性和易用性，无需重复配置。

值得注意的是，hetu权限同步策略为DataArts Studio实例级别配置，各工作空间之间数据互通，全局可见并生效。

前提条件

- MRS Hetu已开启Ranger权限管控，详情请参考**HetuEngine权限管理概述**。
- 配置hetu权限同步策略前，已在管理中心创建MapReduce服务（MRS Hive）和MapReduce服务（MRS Hetu）类型的数据连接，请参考**创建DataArts Studio数据连接**。

约束与限制

- 仅DAYU Administrator、Tenant Administrator用户或者数据安全管理员可以创建、修改或删除hetu权限同步策略，其他普通用户无权限操作。

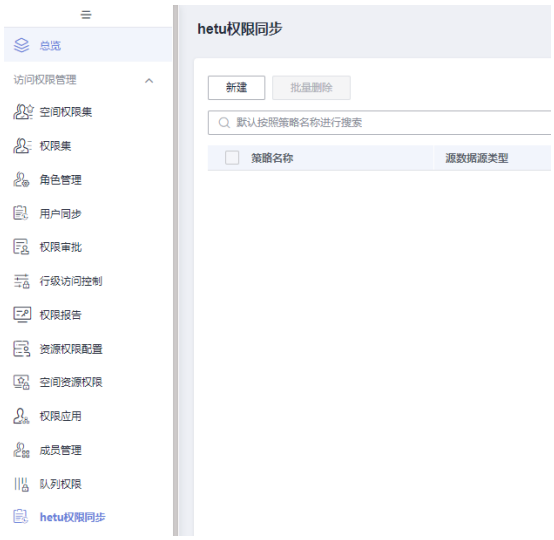
- 当前仅支持Hive权限同步至同一MRS集群的Hetu。
- Hetu权限同步策略需要配置Hive和Hetu catalog的对应关系。对于一个Hive源对接多个Hetu catalog场景，需要配置多个同步策略。
- Hetu权限同步策略创建后，不会自动将已有Hive权限同步至Hetu。仅当Hive权限同步触发后，才会同步权限至Hetu端，另外也会因此导致权限同步所需时间变长。
- 当Hive权限同步触发后，如果同步权限至Hetu端发生失败，Hive权限同步不受影响。
- Hetu权限同步策略删除后，不会回收已同步至Hetu的权限。
- 同步到Hetu端的Ranger的策略命名格式为“***catalog名_schema名+表名+列名***”。如果Hetu端的Ranger上已有相同资源、名称的策略，则会导致同步权限至Hetu端的失败，此时需要手动清理Hetu端的Ranger上资源、名称冲突的策略。

创建 hetu 权限同步策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“hetu权限同步”，进入hetu权限同步页面。

图 12-72 进入 hetu 权限同步页面



步骤3 单击“新建”，进入新建hetu权限同步策略页面，参数配置参考**表12-8**。

图 12-73 新建 hetu 权限同步策略参数配置

创建hetu权限同步策略参数配置说明：

表 12-8 配置策略参数

参数名	参数说明
*策略名称	hetu权限同步策略的标识，同一个数据表上不能有同名的hetu权限同步策略。 为便于策略管理，建议名称中标明要同步的集群名和Catalog名。
策略描述	为更好地识别hetu权限同步策略，此处加以描述信息，长度不能超过255个字符。
权限源端	
*数据源类型	当前仅支持MRS Hive数据源。
*数据连接	从下拉列表中选择数据连接类型中已创建的数据连接，若未创建请参考**创建DataArts Studio数据连接**新建连接。
集群名称	无需选择，自动匹配数据连接中的数据源集群。
权限目标端	
*数据源类型	当前仅支持MRS Hetu数据源。
*数据连接	从下拉列表中选择数据连接类型中已创建的数据连接，若未创建请参考**创建DataArts Studio数据连接**新建连接。 注意，所选择的Hetu连接所在的集群应与Hive连接所在的集群一致。
集群名称	无需选择，自动匹配数据连接中的数据源集群。

参数名	参数说明
*Catalog	Hetu上的数据源名称，本集群的Hive数据源名称默认为“hive”。由于Hetu支持多个Catalog对接同一个Hive，因此您也可以选择其他本集群的Catalog。

步骤4 单击“提交”，完成hetu权限同步策略创建。

步骤5 当Hive权限同步触发后，会同步权限至Hetu端Ranger，策略命名格式为“***catalog名_schema名+表名+列名***”。系统定义的Hive与Hetu间的策略映射关系如表12-9所示。

表 12-9 Hive 与 Hetu 的策略映射关系

Hive	Hetu
资源映射关系	
hive数据源	Hetu Catalog
hive数据库	Hetu Schema
hive表	Hetu表
hive列	Hetu列
权限映射关系	
select	select、use
update	insert、delete、update
create	create
drop	drop
alter	alter
all	all

----结束

相关操作

- 编辑策略：在hetu权限同步页面，单击对应任务操作栏中的“编辑”，即可编辑hetu权限同步策略。
- 删除策略：在hetu权限同步页面，单击对应任务操作栏中的“删除”，即可删除策略。当需要批量删除时，可以在勾选策略后，在列表上方单击“删除”。

  说明

  删除操作无法撤销，请谨慎操作。
- 查看策略详情：在hetu权限同步页面，找到需要查看的策略，单击对应任务操作栏中的“详情”，即可查看策略详情。

图 12-74 查看策略详情

12.3.5.7 申请与审批权限（部分高级特性）

进行访问权限管理时，除了可以自上而下地通过权限集/角色方式为用户授权外，也支持自下而上的用户权限申请、审批流程。

本章主要描述如何**配置审批策略**，申请者如何**申请权限**，以及审批者如何**审批权限**和**回收权限**。

说明

在新版本模式下仅当使用企业版时，才支持按字段粒度申请权限，以及设置权限有效期。旧版本模式使用基础版及更高版本时即可支持。

前提条件

- 权限申请前，已完成空间权限集的配置，请参考**配置空间权限集**。
- 权限申请前，需要在数据目录组件，对数据连接成功进行过元数据采集，详见**元数据采集任务**。

约束与限制

- 一个密级下只允许存在一条审批策略，不选密级也只允许存在一条审批策略。
- 创建基于密级的审批策略时，需要满足以下条件：
  - 已开启数据地图组件。
  - 已采集相关密级数据的元数据。
  - 已完成敏感数据发现任务，并将密级信息同步到数据地图。
- 当前仅支持按照数据表粒度，申请数据表的查询数据（SELECT）权限。因此权限申请前，请确保空间权限集已配置待申请数据表中所有列的SELECT权限。

  说明

  在新版本模式下仅当使用企业版时，才支持按字段粒度申请权限，以及设置权限有效期。旧版本模式使用基础版及更高版本时即可支持。
- 仅DAYU Administrator、Tenant Administrator、数据安全管理员和工作空间管理员可以回收其他用户权限。
- 单次申请多个数据表的权限，会拆成多个工单进行审批。

- 当前权限申请和审批模块，仅支持查看当前用户的权限申请与审批记录，不支持权限审计。
- DLI权限申请只支持为用户申请，不支持用户组。
- 进行权限同步时，需要为dlg_agency委托配置相关权限，请参考**授权dlg_agency委托**。
- 当前数据权限管控为白名单机制，是在待授权用户原有权限的基础上增加允许操作条件，不会影响用户的原有权限。如果仅需要当前数据权限管控所赋予的权限生效，则需要您手动去除待授权用户的原有权限。详见**数据权限管控说明**。
- 默认在DataArts Studio数据开发组件执行脚本、测试运行作业时，数据源（此处指MRS/DWS数据源）会使用数据连接上的账号进行认证鉴权。因此在数据开发时，权限管控依然无法生效。需要您启用细粒度认证，使得在数据开发执行脚本、测试运行作业时，使用当前用户身份认证鉴权，从而做到实现不同用户具有不同的数据权限，使角色/权限集中的权限管控生效。

配置审批策略

通过审批策略，您可以设置多级审批流程，或针对不同密级数据的权限申请设置不同的审批流程。

值得注意的是，审批策略为DataArts Studio实例级别配置，各工作空间之间数据互通，全局可见并生效。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“权限审批”，进入权限审批页面。

步骤3 在“权限审批”页面，单击“审批策略”进入审批策略页签。单击“新建”，新建一条审批策略。

图 12-75 创建审批策略

步骤4 在创建策略页面中，参考**表12-10**完成审批策略创建，审批节点可通过单击+进行新增。

图 12-76 配置审批策略

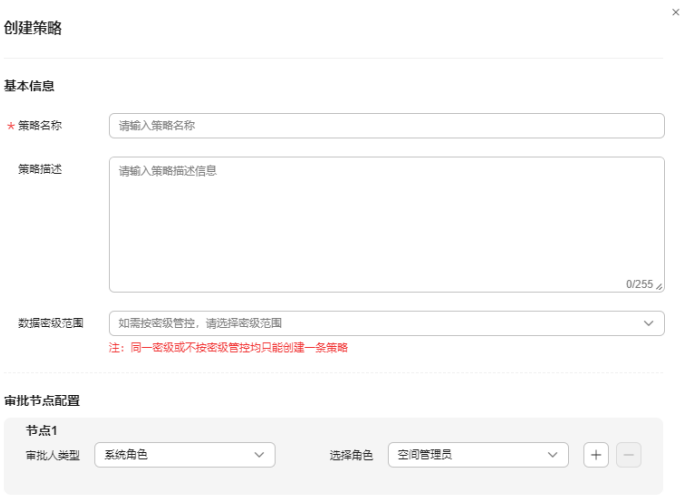
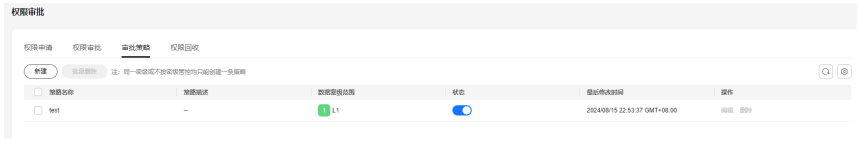


表 12-10 审批策略参数说明

配置项	说明
基本信息	
*策略名称	配置审批策略名。仅支持中英文、数字和下划线，长度不超过32个字符。
策略描述	为更好地识别审批策略，此处加以描述信息，长度不能超过255个字符。
数据密级范围	如果需要针对不同密级数据的权限申请设置不同的审批流程，此处需要选择数据密级。注意，一个密级下只允许存在一条审批策略，不选密级也只允许存在一条审批策略。 选择数据密级的条件为： ● 已开启数据地图组件。 ● 已采集相关密级数据的元数据。 ● 已完成敏感数据发现任务，并将密级信息同步到数据地图。
审批节点配置-系统角色/IAM用户/IAM用户组	
审批人类型	选择审批人类型。
选择角色	根据不同的审批人类型，选择对应的审批人角色。

步骤5 审批策略填写完成后，单击提交可新建一条审批策略。新建的审批策略默认为关闭状态，如需生效，请在审批策略列表处，单击  进行开启。

图 12-77 审批策略列表



----结束

申请权限

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“权限审批”，进入权限审批页面。

步骤3 在“权限审批”页面的权限申请页签，单击“创建权限申请”，创建权限申请工单。

图 12-78 创建权限申请

步骤4 在权限申请工单页面中，参考**表12-11**完成工单填写。

图 12-79 填写工单

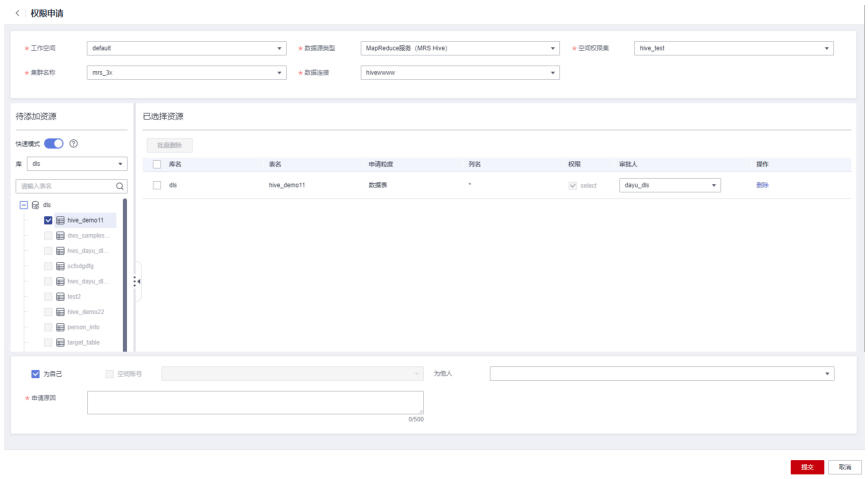


表 12-11 权限申请工单参数说明

配置项	说明
基本信息	
*工作空间	选择已配置空间权限集的工作空间。
*空间权限集	选择空间权限集，空间权限集权限范围应已包含所需资源权限。
*数据源类型	当前支持Hive、DWS、DLI。
*集群名称	选择要申请的资源所在的集群。
*数据连接	选择要申请的资源所在的数据连接。
资源选择	
*待添加资源	在导航树上选择数据库后，勾选所需的数据表，单次申请时支持选择不同数据库下的表。 **说明** • 当前仅支持按照数据表粒度，申请数据表的查询数据（SELECT）权限。因此权限申请前，请确保空间权限集已配置所选数据表中所有列的SELECT权限。 • 在新版本模式下仅当使用企业版时，才支持按字段粒度申请权限，以及设置权限有效期。旧版本模式使用基础版及更高版本时即可支持。 另外，导航树上的快速模式开启后，库表列的元数据会从数据目录获取，否则会从数据源获取元数据。推荐开启快速模式。
*已选择资源	在已选择资源列表中可查看所选的表、权限和审批人信息。 **说明** 审批人默认来自权限集/角色的管理员。例如，如果空间权限集、权限集A和角色B中均定义了所选数据表中所有列的SELECT权限，审批人可以选择为权限集A或角色B管理员；如果只有空间权限集定义了所选数据表中所有列的SELECT权限，审批人为空间权限集的管理员。
申请信息	
为自己	勾选为自己后，可为自己申请所选择的资源权限。
空间账号	当在数据开发组件配置调度的公共IAM账号后，可为空间账号申请所选择的资源权限。
为他人	可选择工作空间内的成员，为其申请所选择的资源权限。
*申请原因	填写申请原因，便于审批人审视是否应当审批。
有效期	选择权限有效期支持选择为固定时长（从申请之日开始计算），也可以自定义配置到期时间。不配置表示权限不存在超时时间。 **说明** 在新版本模式下仅当使用企业版时，才支持管理权限有效期。旧版本模式使用基础版及更高版本时即可支持。

步骤5 工单填写完成后，单击提交可生成一条待审批的工单记录。在工单列表处，可以查看工单ID、摘要、状态等信息，单击ID名查看工单详情，并支持撤回未审批的工单。

图 12-80 工单列表

----结束

审批权限

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“权限审批”，进入权限审批页面。

步骤3 在“权限审批”页面，审批人单击“权限审批”进入权限审批页签。

图 12-81 权限审批

步骤4 在权限审批页签中，工单列表默认展示待审批的工单。您可以查看工单ID、摘要、状态等信息，单击ID名查看工单详情。请从业务合理性和数据安全角度审视，确认“通过”或“驳回”该工单，同时也可以勾选工单后单击列表上方的“批量审批”批量“通过”或“驳回”工单。

步骤5 在权限审批页签中，单击“已审批”，可查看已经审批通过的工单。

图 12-82 已通过工单列表

----结束

回收权限

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“权限审批”，进入权限审批页面。

步骤3 在“权限审批”页面，单击“权限回收”，进入权限回收页签。

图 12-83 权限回收

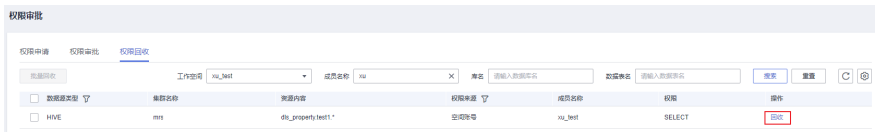
步骤4 在权限回收页签中，列表展示指定空间（默认当前空间）下的用户通过申请、审批获得的数据权限。您可以通过选择需要回收的权限所在的工作空间、成员名称或库表名，匹配权限记录（支持模糊匹配），然后通过操作栏中的“回收”，删除当前的权限记录。

仅DAYU Administrator、Tenant Administrator、空间管理员和数据安全管理员可以回收对应空间下用户的数据权限。

说明

在新版本模式下仅当使用企业版时（旧版本模式使用基础版及更高版本时即可支持），此处才支持进行“变更有效期”操作，详见**变更有效期**。

图 12-84 回收权限



----结束

相关操作

- 编辑审批策略：在审批策略页面，单击对应策略操作栏中的“编辑”，即可修改审批策略各项参数。
- 编辑审批策略状态：新增的审批策略默认为关闭状态。当审批策略为关闭状态时，表示该策略将不生效。

  需要修改审批策略状态时，在审批策略页面单击对应审批策略中的  或  ，即可启用或关闭审批策略。
- 删除审批策略：在审批策略页面，单击对应策略操作栏中的“删除”，即可删除策略。当需要批量删除时，可以在勾选审批策略后，在列表上方单击“批量删除”。

  说明

  删除操作无法撤销，请谨慎操作。

12.3.5.8 管理权限有效期（高级特性）

对于自下而上申请到的用户权限，如果仅短期需要，可以通过有效期进行管理，使权限仅在有效期内生效。

说明

在新版本模式下仅当使用企业版时，才支持管理权限有效期。旧版本模式使用基础版及更高版本时即可支持。

本章主要介绍申请者如何**申请短期权限**、**续期权限**、**订阅权限到期提醒**，以及管理员如何**变更有效期**、**配置权限到期提醒**。

约束与限制

- 仅DAYU Administrator、Tenant Administrator、数据安全管理员可以变更权限有效期、配置权限到期提醒、操作所有订阅提醒。非管理员用户只能操作自己的订阅信息，无法查看和操作其他用户订阅提醒信息。
- 权限到期提醒将在权限过期前7天开始提醒，该时间不支持修改。
- 有效期到期回收后，已失效权限会保留7天，用于及时审视续期，超期后清理。
- 配置权限到期提醒需要为dlg_agency委托配置SMN服务操作权限（SMN FullAccess）。
- 当到期提醒使用数据开发通知主题时，会由于在数据开发侧添加的订阅自带的订阅筛选策略，导致仅请求订阅还是不会收到通知。因此在使用数据开发通知主题的场景下，除了请求订阅外，还需要再进行关联订阅后才能收到权限到期提醒。
- 订阅列表中的订阅通过备注与用户关联，备注为用户名的订阅，视作相应用户的订阅。
- 订阅列表与主题绑定，切换主题后，需重新配置订阅提醒。
- 已配置的终端信息不支持编辑，如果手机号、邮箱等终端发生变化，需删除后重新添加并请求订阅。
- 用户组的权限到期提醒，如果用户的组信息发生过变化，需要刷新订阅策略，才可以及时接收到正确的组信息。
- 每天到期提醒会整合为一条消息通知。如果即将到期的权限过多，则优先展示最快到期的权限，最多提示100条或20w字节。
- 受限于并发控制以及smn性能等因素，smn消息通知可能会有数分钟的通知时延。

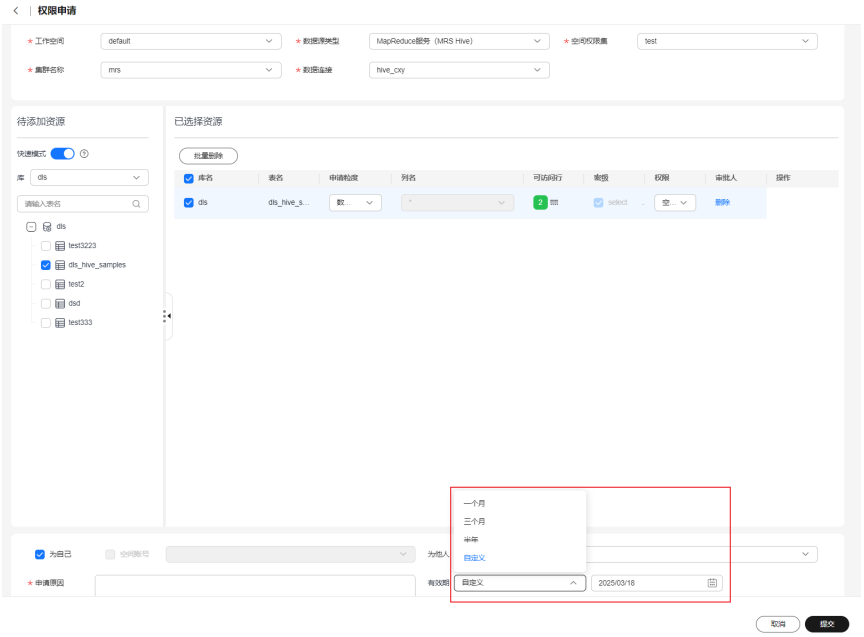
申请短期权限

在数据安全页面**申请权限**或在数据地图页面申请权限时，可以按需选择所申请权限的有效期。选择权限有效期支持选择为固定时长（从申请之日开始计算），也可以自定义配置到期时间（到期时间精确为当天晚上24点）。不配置表示权限不存在超时时间。

审批通过后，申请者在有效期内具有所申请的权限。

以数据安全页面申请权限为例，申请短期权限页面如**图12-85**所示。

图 12-85 填写工单



续期权限

对于即将到期的权限，如果有延长有效期的需要，申请者可以进行续期申请。申请审批通过后，有效期会延长至新的到期时间。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

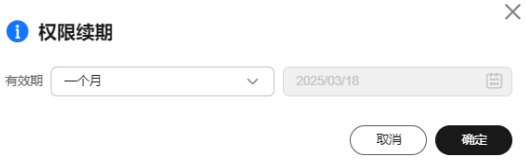
步骤2 单击左侧导航树中的“权限审批”，在“权限审批”页面，单击“我的权限”进入我的权限页签。

步骤3 在“我的权限页面”，查看我已申请到的权限。

图 12-86 我的权限

步骤4 在待续期权限的操作栏选择“续期”，或在选择待续期权限后选择列表上方的“批量续期”，在弹出的窗口中选择权限的有效期。选择权限有效期支持选择为固定时长（从申请之日开始计算），也可以自定义配置到期时间（到期时间精确为当天晚上24点）。

图 12-87 权限续期



步骤5 点击“确定”，完成续期申请。

----结束

变更有效期

对于用户已申请到的权限，管理员可以进行审视，并调整不适宜的权限有效期，以便权限及时更新或回收。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

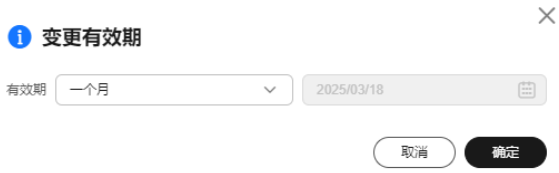
步骤2 单击左侧导航树中的"权限审批"，在“权限审批”页面，单击“权限回收”进入权限回收页签。

步骤3 在“权限回收”页签，查看当前空间下已审批通过的权限。

图 12-88 权限回收

步骤4 在待变更有效期权限的操作栏选择“变更有效期”，或在选择待变更有效期权限后选择列表上方的“批量变更有效期”，在弹出的窗口中选择权限的有效期。选择权限有效期支持选择为固定时长（从申请之日开始计算），也可以自定义配置到期时间（到期时间精确为当天晚上24点）。

图 12-89 变更有效期



步骤5 点击“确定”，完成权限有效期变更。

----结束

配置权限到期提醒

管理员可以配置权限到期提醒主题、通知时间等信息，已订阅的用户在权限过期前7天开始会收到权限到期通知信息。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的"权限审批"，在“权限审批”页面，单击“权限通知”进入权限通知页签。

步骤3 在"权限通知"页面，配置权限到期提醒。

- 权限到期通知主题：选择消息通知服务（SMN）中的消息主题。
- 权限到期通知时间(每天几点通知)：选择每日通知的整点时间。
- 权限到期通知空间账号管理角色：选择到期通知的工作空间角色。

----结束

订阅权限到期提醒

在完成**配置权限到期提醒**后，已订阅的用户在权限过期前7天开始会收到权限到期通知信息。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的"权限审批"，在“权限审批”页面，单击“权限通知”进入权限通知页签。

步骤3 在“权限通知”页面，系统会根据已配置的通知主题，将SMN主题相关的订阅终端信息展示在下方订阅列表中，列表中的用户名称对应订阅备注信息，申请者只能操作备注信息与用户名一致的订阅。

申请者选择希望接收消息的终端记录，进行请求订阅，订阅状态为“已确认 ”后即为订阅成功。

说明

当到期提醒使用数据开发通知主题时，会由于在数据开发侧添加的订阅自带的订阅筛选策略，导致仅请求订阅还是不会收到通知。因此在使用数据开发通知主题的场景下，除了请求订阅外，还需要再进行关联订阅后才能收到权限到期提醒。

图 12-90 订阅权限到期提醒

----结束

12.3.5.9 配置建库申请（高级特性）

为了管控存算分离MRS Hive数据源的创建数据库流程，数据安全支持通过申请审批的流程在数据源上创建数据库。

说明

在新版本模式下仅当使用企业版时，才支持建库申请和建库审批。旧版本模式使用基础版及更高版本时即可支持。

本章主要介绍管理员如何**配置数据库路径**，申请者如何**申请创建数据库**，以及审批者如何**审批建库申请**。

前提条件

- MRS集群已开启存算分离功能，已在管理中心创建MapReduce服务（MRS Hive）和MapReduce服务（MRS Ranger）类型的数据连接，请参考**创建DataArts Studio数据连接**。
- 使用建库申请功能前，已参考**授权dlg_agency委托**为dlg_agency委托配置权限。
- 仅DAYU Administrator、Tenant Administrator、数据安全管理员有权限配置数据库路径以及审批建库申请。
- 建库申请人应为空间管理员，且为任意空间权限集的管理员。

约束与限制

- 仅支持开启Ranger管控的MRS集群。
- 仅支持申请人对工单撤回、审批人对工单驳回，暂不支持创建库审批的回收功能。
- 创建数据库路径配置时，一个集群只允许配置一个路径策略。

配置数据库路径

由DAYU Administrator、Tenant Administrator、数据安全管理员配置数据库路径。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

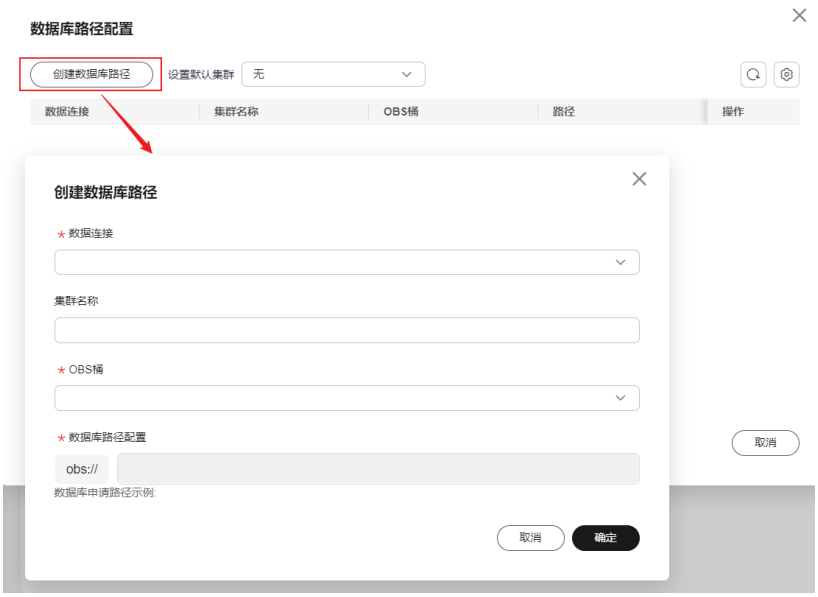
步骤2 单击数据安全左侧导航树中的“权限审批”，在“权限审批”页面，单击“建库审批”进入建库审批页签。

图 12-91 建库审批页面

步骤3 单击数据库路径配置，在弹出的数据库路径配置窗口中，单击创建数据库路径，配置数据库路径。

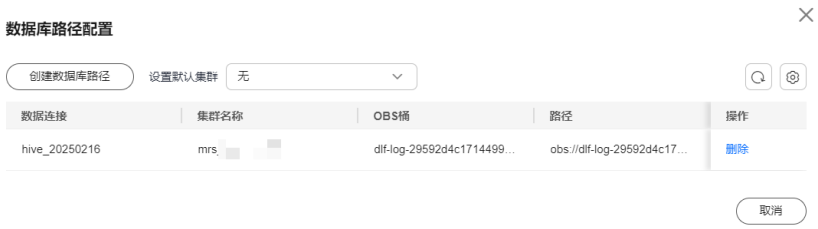
- 数据连接：选择已创建的MRS Hive数据连接。
- 集群名称：无需输入，自动从连接关联。
- OBS桶：选择OBS桶。
- 数据库路径配置：数据库路径根据OBS桶名称动态拼接，后缀部分支持内置关键字自动匹配，例如当输入{{?符号时，自动弹出所有选项{{?CURRENT_WORKSPACE}}, {{?OBS_BUCKET}}等。

图 12-92 配置数据库路径



步骤4 配置数据库路径完成后，单击确定，完成数据库路径的创建。

图 12-93 配置数据库路径完成



步骤5 在数据库路径配置窗口中，可通过选择默认集群，帮助申请人选择默认弹出的路径。配置完成后，关闭窗口，完成相关配置。

----结束

申请创建数据库

建库申请人应为空间管理员，且为任意空间权限集的管理员。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“权限审批”，在“权限审批”页面，单击“建库申请”进入建库申请页签。

图 12-94 建库申请页面

步骤3 单击创建数据库申请，选定管理员**配置数据库路径**中所配置MRS集群，输入待创建数据库名称，点击数据库路径后的“测试”，确认数据库为未创建的数据库（不允许创建重名数据库）。

测试通过后，选择空间权限集，填写数据库描述信息和申请原因，单击“确定”完成申请。

图 12-95 建库申请页面

步骤4 返回建库申请的工单列表后，工单创建成功即申请完成。

图 12-96 工单创建完成

----结束

审批建库申请

由DAYU Administrator、Tenant Administrator、数据安全管理员审批建库申请。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

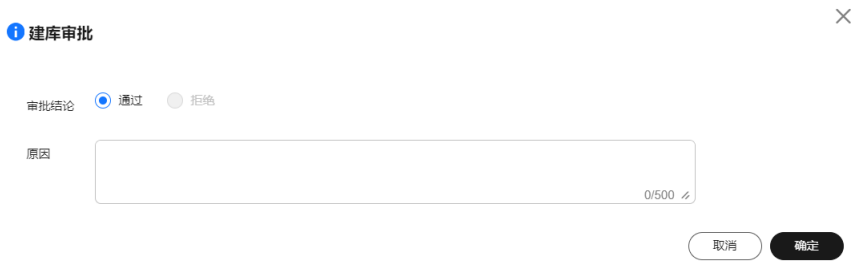
步骤2 单击数据安全左侧导航树中的“权限审批”，在“权限审批”页面，单击“建库审批”进入建库审批页签。

图 12-97 建库审批页面

步骤3 在建库审批页面，找到待审批的建库申请，点击操作栏选择通过或者拒绝。

- 通过：在指定数据源创建数据库，并在指定空间权限集下创建对应OBS路径权限策略。
- 拒绝：驳回申请单。

图 12-98 审批建库申请



----结束

12.3.5.10 启用细粒度认证

在未启用细粒度认证的情况下，当在DataArts Studio数据开发组件执行脚本、测试运行作业时，数据源会使用数据连接上的账号进行认证鉴权。因此，即使已通过配置角色/权限集对用户进行权限管控，当用户在进行数据开发时，权限管控依然无法生效。

而在启用细粒度认证后，在DataArts Studio数据开发执行脚本、测试运行作业或调度作业时，数据源将不再使用数据连接上的账号，而是使用当前用户身份认证鉴权，从而做到实现不同用户具有不同的数据权限，使角色/权限集中的权限管控生效。

细粒度认证开启状态对数据开发中的脚本、作业运行影响总结如下：

- 当关闭细粒度认证时，数据开发中的脚本执行、作业测试运行和作业调度使用数据连接上的账号进行认证鉴权。
- 当启用开发态细粒度认证后，数据开发中的脚本执行、作业测试运行使用当前用户身份认证鉴权，作业调度使用数据连接上的账号进行认证鉴权。
- 当启用调度态细粒度认证后，数据开发中的脚本执行、作业测试运行和作业调度使用当前用户身份认证鉴权。

前提条件

- 开启细粒度认证前，请确保已经为使用数据源的用户配置了业务所需的数据权限，避免开启后因用户无数据权限导致业务中断。配置权限详见**配置权限集**或**配置角色**。
- DWS联通性测试前，已完成用户同步，然后将当前登录账号切换为IAM子用户账号，且至少具有DWS Database Access权限。
- 已经为MRS Hive连接和MRS SPARK连接中的用户配置了代理权限，请参考**参考：为MRS数据连接用户配置代理权限**进行配置。
- MRS SPARK数据连接对应的SPARK2x组件为多主实例模式，否则请参考**配置多主实例与多租户模式切换**章节进行切换。

约束与限制

- 当前开发态细粒度认证仅支持DWS、代理模式的MRS Hive和MRS SPARK类型数据连接，调度态细粒度认证仅支持代理模式的MRS Hive类型数据连接。
- 仅DAYU Administrator、Tenant Administrator或者数据安全管理员有权限配置细粒度认证状态。
- 仅当数据连接中的Agent选择的CDM集群为2.10.0.300及以上版本时，才支持细粒度认证。
- 角色/权限集中配置的用户权限，需要在角色/权限集同步成功并启用细粒度认证后才能生效。
- DWS连接联通性测试约束如下：
  - 联通性测试时，系统会使用当前用户账号访问数据源，以确保正常访问。但由于DWS数据源不支持以华为账号直接访问，如果登录账号为华为账号，联通性测试会失败。因此，在DWS联通性测试前，需要先完成用户同步，再将当前登录账号切换为IAM子用户账号，且至少具有DWS Database Access权限。
  - 仅当DWS集群guest_agent版本为8.2.1，或在8.2.1以上、9.0.0以下时，才支持细粒度认证。DWS集群guest_agent版本查看方法请参考**查看DWS集群guest agent版本**。

- MRS Hive连接联通性测试约束如下：

  仅当MRS Hive数据连接中的用户配置了代理权限后，才支持细粒度认证。
- MRS SPARK连接联通性测试约束如下：
  - 仅当MRS SPARK数据连接中的用户配置了代理权限后，才支持细粒度认证。
  - 仅当MRS SPARK数据连接对应的SPARK2x组件为多主实例模式时才支持细粒度认证，为多租户模式时不支持。多租户模式切换多主实例模式请参考**配置多主实例与多租户模式切换**章节。

启用细粒度认证

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“权限应用”，进入权限应用页面。

步骤3 在“权限应用”页面，为希望启用细粒度认证的数据连接，进行联通性测试。联通性测试时，系统会使用当前用户账号访问数据源，以确保当前用户访问正常。

说明

- 由于DWS数据源不支持以华为账号直接访问，因此如果当前以华为账号登录，则会导致联通性测试失败。因此在DWS联通性测试前，需要先完成用户同步，再将当前登录账号切换为IAM子用户账号，且至少具有DWS Database Access权限。

图 12-99 联通性测试

如果联通性测试失败，可从以下方面进行排查：

1. 确保数据连接上的数据源可用。
2. 数据连接中的Agent选择的CDM集群应为2.10.0.300及以上版本。
3. 已完成用户同步，用户同步操作请参考**同步IAM用户到数据源**。
4. DWS连接：
   a. DWS连接中DWS集群guest_agent版本为8.2.1，或在8.2.1以上、9.0.0以下。DWS集群guest_agent版本查看方法请参考**查看DWS集群guest agent版本**。
   b. 已将当前登录账号切换为IAM子用户账号，且具有至少DWS Database Access权限。
5. MRS Hive连接：

   MRS Hive连接中的用户是否配置了代理权限，若没配置代理，可参考**参考：为MRS数据连接用户配置代理权限**。
6. MRS SPARK连接：
   a. MRS SPARK连接中的用户是否配置了代理权限，若没配置代理，可参考**参考：为MRS数据连接用户配置代理权限**。

b. MRS SPARK数据连接对应的SPARK2x组件是否为多主实例模式。多主实例模式时才支持细粒度认证，为多租户模式时不支持。多租户模式切换多主实例模式请参考**配置多主实例与多租户模式切换**章节。

步骤4 联通性测试成功后，在细粒度认证状态列，根据所需选择启用开发态或调度态的细粒度认证，然后单击下方的“提交”，即可开启细粒度认证。

图 12-100 开启细粒度认证

----结束

参考：为 MRS 数据连接用户配置代理权限

用户在DataArts Studio上通过MRS Hive或Spark数据连接访问数据源时，默认使用数据连接中配置的账号信息访问。而在为MRS Hive或Spark数据连接中的账号信息配置Hive或Spark代理权限后，用户在发起操作时，MRS支持切换为以用户自身身份执行，从而支持细粒度认证。具体配置方法详见**配置Hive代理权限**和**配置Spark代理权限**。

配置Hive代理权限

步骤1 登录MRS FusionInsight Manager。

步骤2 选择“集群 > 服务 > Hive > 配置 > 基础配置”，在搜索框中输入参数名“core.site.customized.configs”，配置相应参数，如**图12-101**所示。

表 12-12 配置参数

参数名	名称	值
core.site.customized.configs	hadoop.proxyuser.***数据连接上配置的用户名***.users	*
	hadoop.proxyuser.***数据连接上配置的用户名***.groups	*
	hadoop.proxyuser.***数据连接上配置的用户名***.hosts	*

图 12-101 配置 core.site.customized.configs 参数示例

步骤3 参数均配置完成后，单击左上角的“保存”，在弹窗中单击“确定”保存配置。

图 12-102 保存配置

步骤4 保存成功后，切换到实例页签，选择配置已过期的实例后，单击“更多 > 滚动重启实例”，使配置生效。

图 12-103 滚动重启实例

----结束

配置Spark代理权限

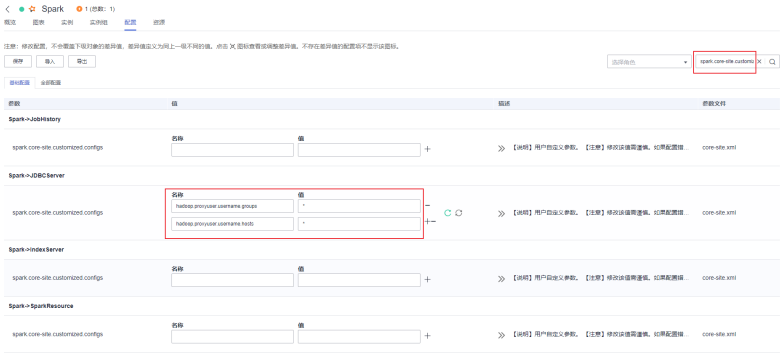
步骤1 登录MRS FusionInsight Manager。

步骤2 选择“集群 > 服务 > Spark> 配置 > 基础配置”或“集群 > 服务 > Spark2x> 配置 > 基础配置”，在搜索框中输入参数名“spark.core-site.customized.configs”，配置相应参数。后文以Spark组件为例进行说明，如图12-104所示。

表 12-13 配置参数

参数名		名称	值
Spark->JDBCServer 或 Spark2x->JDBCServer2x	core.site.customized.configs	hadoop.proxyuser.***数据连接上配置的用户名***.groups	*
		hadoop.proxyuser.***数据连接上配置的用户名***.hosts	*
		hadoop.proxyuser.***数据连接上配置的用户名***.groups	*
		hadoop.proxyuser.***数据连接上配置的用户名***.hosts	*

图 12-104 配置 spark.core-site.customized.configs 参数示例

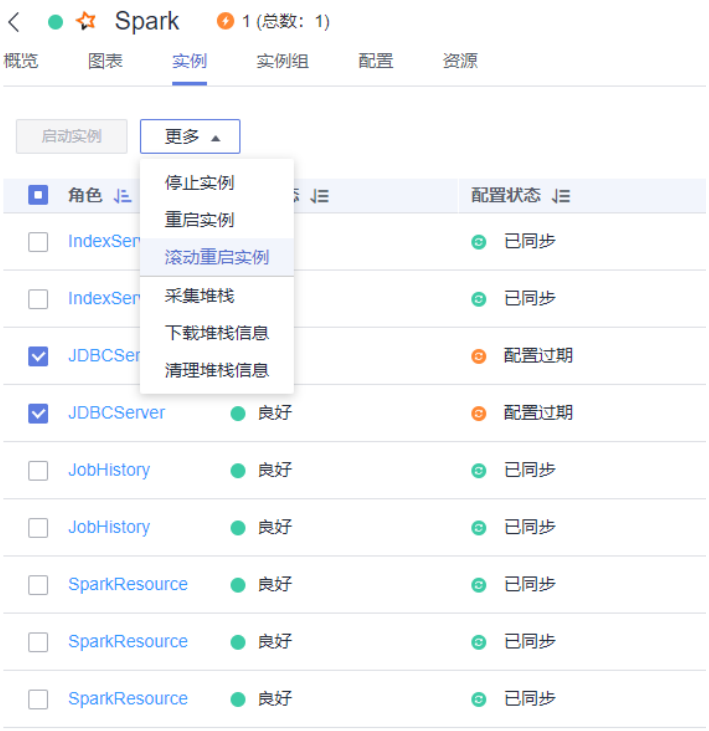


步骤3 参数均配置完成后，单击左上角的“保存”，在弹窗中单击“确定”保存配置。

图 12-105 保存配置

步骤4 保存成功后，切换到实例页签，选择配置已过期的实例后，单击“更多 > 滚动重启实例”，使配置生效。

图 12-106 滚动重启实例



----结束

12.3.5.11 启用账号映射（高级特性）

在未启用细粒度认证及账号映射策略的情况下，当在DataArts Studio数据开发组件执行脚本以及测试运行作业时，数据源默认会使用数据连接上的账号进行认证鉴权。因此，即使已通过配置角色/权限集对用户进行权限管控，当用户在进行数据开发时，权限管控依然无法生效。

而在启用账号映射策略后，在DataArts Studio数据开发执行脚本以及测试运行作业时，数据源将不再使用数据连接上的账号，而是将当前用户身份映射成MRS系统账号或ldap账号后进行认证鉴权，从而做到实现不同用户具有不同的数据权限。

说明

- **细粒度认证功能**与**账号映射功能**在使用场景上较为相似，不可同时配置，二者在各维度对比差异如**细粒度认证与账号映射功能差异**所示。在实际使用中，建议您根据您的需求选择其一进行配置。
- 在新版本模式下仅当使用企业版时，才支持账号映射功能。旧版本模式使用基础版及更高版本时即可支持。

前提条件

- 启用账号映射策略前，请确保已经为MRS系统账号或ldap账号配置了业务所需的数据权限，避免开启后因用户无数据权限导致业务中断。

约束与限制

- 当前开发态账号映射仅支持代理模式的MRS Hive、MRS SPARK，MRS Hetu、MRS Impala类型数据连接。当修改MRS Hive、MRS SPARK数据连接的连接方式时，例如将代理连接改成API直连，将会导致账号映射失效。
- 仅当数据连接中的Agent选择的CDM集群为2.10.0.300及以上版本时，才支持账号映射。
- 只有DAYU Administrator、Tenant Administrator或者数据安全管理员有权限创建账号映射策略，并配置账号映射。
- DAYU User用户修改账号映射策略时，仅支持修改自己账号的映射规则，不支持修改映射类型、默认访问身份等其他内容。
- 账号映射策略为DataArts Studio实例级别配置，各工作空间之间数据互通。因此一个集群的每种映射类型下只能创建一个账号映射策略。
- 在配置账号映射策略的映射列表时，不支持校验MRS系统账号或ldap账号是否存在以及密码是否正确，若配置错误的用户名密码，会导致账号映射失败。而对于默认访问身份的配置的账号密码，系统支持校验。

配置账号映射策略

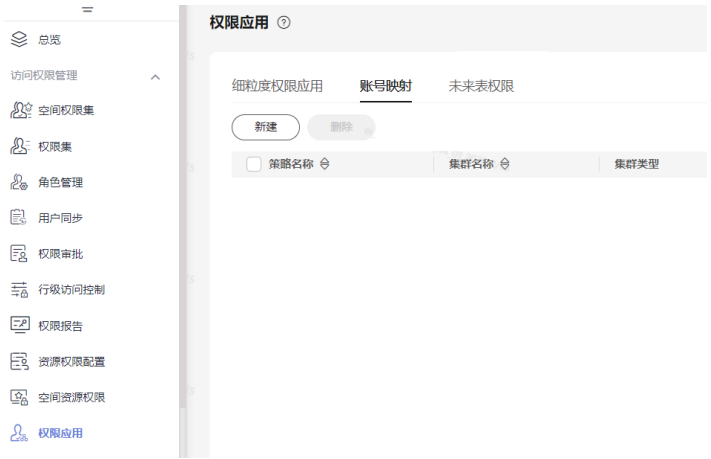
账号映射策略可以分为三部分：一为基本信息配置；二为默认访问身份，对于未在集群账号映射中配置的IAM账号，会使用默认访问身份执行；三为集群账号映射，对不同IAM用户设置对应的映射账号。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“权限应用”，在权限应用页面，进入“账号映射”页签。

步骤3 在“账号映射”页面，单击“新建”，创建账号映射策略。

图 12-107 配置账号映射策略



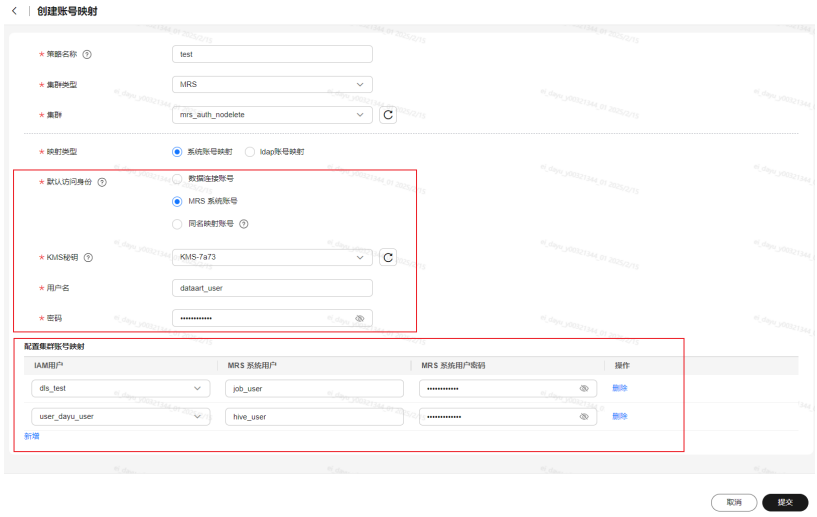
步骤4 新建账号映射策略配置请参考**表12-14**，参数配置完成单击“确定”即可。

表 12-14 参数设置

参数名	参数设置
*策略名称	标识账号映射策略，实例下唯一。 建议名称中包含含义，避免无意义的描述，以便于快速识别所需账号映射策略。
*集群类型	无需选择，当前仅支持MRS集群。
*集群	选择对应的MRS集群。注意每个集群的每种映射类型下只能创建一个账号映射策略。
*映射类型	选择账号映射类型。注意每个集群的每种映射类型下只能创建一个账号映射策略。 ● 系统账号映射：把当前IAM账号映射成MRS系统账号，默认展示。 ● ldap账号映射：把当前IAM账号映射成ldap账号。系统会根据MRS Hive、MRS Impala类型数据连接中的ldap相关配置的开启情况，自动选择是否展示该类型。
系统账号映射	
*默认访问身份	选择系统账号映射的默认映射账号类型。未配置账号映射的IAM账号将统一使用默认访问身份进行认证鉴权。 ● 数据连接账号：使用连接中的MRS系统账号进行认证鉴权，不做映射。 ● MRS系统账号：使用配置的通用MRS系统账号进行认证鉴权。 ● 同名映射账号：使用当前IAM账号同名的MRS系统账号进行认证鉴权。
KMS密钥	通过KMS加解密认证信息，选择KMS中的任一默认密钥或自定义密钥即可。
用户名 密码	选择“MRS系统账号”的默认访问身份时展示此选项。 配置为通用MRS系统账号，未配置账号映射的IAM账号将统一使用此 MRS系统账号进行认证鉴权。 注意系统不支持校验账号是否存在以及密码是否正确，若配置错误的用户名密码，会导致账号映射失败。
FusionInsight Manager账号 密码	当选择的MRS集群为安全集群，且选择“同名映射账号”的默认访问身份时展示此选项。 ● MRS Hive连接和MRS SPARK连接配置的FusionInsight账号需要拥有Manager User管理权限。 ● MRS Hetu连接配置的FusionInsight账号需要添加hetuadmin用户组权限，且要求集群为MRS 3.3.0-LTS以上版本。 注意系统不支持校验账号是否存在以及密码是否正确，若配置错误的用户名密码，会导致账号映射失败。
ldap账号映射	

参数名	参数设置
*默认访问身份	选择ldap账号映射的默认映射账号类型。未配置账号映射的IAM账号将统一使用默认访问身份进行认证鉴权。 ● 数据连接账号：使用连接中的ldap账号进行认证鉴权，不做映射。 ● ldap账号：使用配置的通用ldap账号进行认证鉴权。
KMS密钥	通过KMS加解密认证信息，选择KMS中的任一默认密钥或自定义密钥即可。
用户名	选择“ldap账号”的默认访问身份时展示此选项。 配置为通用ldap账号，未配置账号映射的IAM账号将统一使用此ldap账号进行认证鉴权。 注意系统不支持校验账号是否存在以及密码是否正确，若配置错误的用户名密码，会导致账号映射失败。
密码	
配置集群账号映射	
IAM用户	对单个IAM用户配置单独的账号映射规则。此处未配置的IAM账号将统一使用默认访问身份进行认证鉴权。
MRS系统用户/ldap账号	配置为单个IAM用户待映射的MRS系统用户或ldap账号及对应密码。
MRS系统用户密码/ldap密码	
操作	可通过“删除”移除一条映射规则，通过“新增”添加一条映射规则。 DAYU User用户修改账号映射策略时，仅支持修改自己账号的映射规则。

图 12-108 新建账号映射策略配置



----结束

启用账号映射

当成功配置账号映射策略后，需要在细粒度权限应用页签启用账号映射策略后，账号映射才能生效。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“权限应用”，在权限应用页面，进入“细粒度权限应用”页签。

步骤3 在“细粒度权限应用”页面，为希望启用账号映射的数据连接，进行账号映射联通性测试。

账号映射联通性测试时，系统会按照账号映射策略将当前用户身份映射成MRS系统账号或ldap账号后再访问数据源，以确保账号访问正常。

图 12-109 账号映射联通性测试

步骤4 账号映射联通性测试成功后，需要在细粒度认证状态处选择“开发态账号映射”，然后系统会自动根据数据连接的集群选择对应的账号映射策略，为对应连接开启账号映射。

图 12-110 启用账号映射

----结束

细粒度认证与账号映射功能差异

表 12-15 细粒度认证与账号映射策略差异

差异项	细粒度认证	账号映射
支持数据连接	• DWS • 代理模式的MRS Hive • 代理模式的MRS SPARK	• 代理模式的MRS Hive • 代理模式的MRS SPARK • MRS Hetu • MRS Impala
配置流程	测试联通性 > 启用细粒度认证	配置账号映射策略 > 测试账号映射联通性 > 启用账号映射
影响的操作	数据开发操作： • 开发态细粒度认证：脚本执行和作业测试运行 • 调度态细粒度认证：脚本执行、作业测试运行和作业调度	数据开发操作： • 开发态账号映射：脚本执行和作业测试运行

差异项	细粒度认证	账号映射
目标认证鉴权身份	以当前用户身份进行认证鉴权。 **说明** 建议提前通过配置角色/权限集，对用户进行权限管控。	以当前用户身份，在映射策略列表进行匹配： ● 匹配成功后，将当前用户身份映射成对应的MRS系统账号或ldap账号进行认证鉴权。 ● 匹配失败后，按照配置默认访问身份进行认证鉴权： 系统账号映射类型： – 数据连接账号：使用连接中的MRS系统账号进行认证鉴权，不做映射。 – MRS系统账号：使用配置的通用MRS系统账号进行认证鉴权。 – 同名映射账号：使用当前IAM账号同名的MRS系统账号进行认证鉴权。 ldap账号映射类型： – 数据连接账号：使用连接中的ldap账号进行认证鉴权，不做映射。 – ldap账号：是使用配置的通用ldap账号进行认证鉴权。 **说明** 建议提前在MRS集群中，对MRS系统账号或ldap账号进行权限管控。

12.3.5.12 配置未来表权限（高级特性）

在配置角色/权限集时，如果给某一用户赋予了DWS数据源某个Schema下的全表权限（即将权限中的数据表配置为*号），则该用户具备对该Schema下的所有表的相应权限。但由于DWS自身权限特性限制，这些赋予的权限仅针对当前已有的表；而对于权限同步后再创建的新表（以下简称未来表），该用户依然没有权限，需要在角色/权限集中再次手动进行 权限同步后，才能确保该用户具备未来表的相应权限。

为了解决未来表权限需要手动同步的问题，您可以通过未来表权限为指定Schema配置未来表的建表用户。当这些用户在指定Schema下创建未来表时，当前实例下所有对该Schema拥有全表权限的用户，将自动获得对所创建未来表的相应权限。

说明

在新版本模式下仅当使用企业版时，才支持配置未来表权限。旧版本模式使用基础版及更高版本时即可支持。

前提条件

- 配置权限集前，已在管理中心创建数据仓库服务（DWS）类型的数据连接，请参考**创建DataArts Studio数据连接**。

约束与限制

- 指定的未来表用户需要有对应Schema下的Create表权限。
- 单个数据库下，Schema视图最多为单个Schema配置200个未来表用户，未来表用户视图最多为单个未来表用户配置200个Schema。

为 Schema 配置未来表用户（schema 视图）

基于schema视图的配置，可以单次为一个schema配置多个未来表用户。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“权限应用”，在权限应用页面，进入“未来表权限 ”页签。

步骤3 在“未来表权限 ”页面，在数据连接区域选择需要配置未来表权限的DWS数据连接，并单击已选择的数据连接后的 添加数据库。

图 12-111 选择 DWS 数据连接

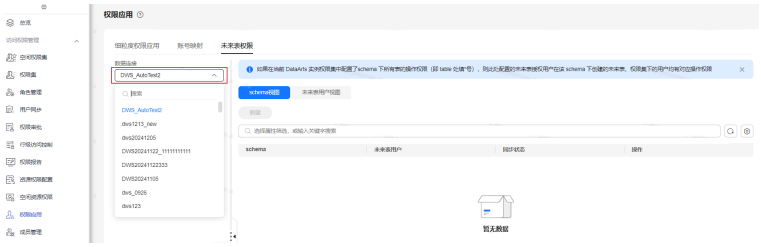


图 12-112 添加数据库

步骤4 单击待配置的数据库，在schema视图下单击“新建”，创建未来表配置。

图 12-113 创建未来表配置

步骤5 在弹出的窗口中，为指定Schema配置未来表的建表用户，单击确定完成配置。

图 12-114 未来表配置

步骤6 配置成功后，单击“同步”完成未来表配置。

同步完成后，当未来表用户在指定Schema下创建未来表时，当前实例下所有对该Schema拥有全表权限的用户，将自动获得对所创建未来表的相应权限。

图 12-115 同步未来表配置

----结束

为未来表用户配置 Schema（未来表用户视图）

基于未来表用户视图的配置，可以单次为一个未来表用户配置多个Schema。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“权限应用”，在权限应用页面，进入“未来表权限 ”页签。

步骤3 在“未来表权限 ”页面，在数据连接区域选择需要配置未来表权限的DWS数据连接，并单击已选择的数据连接后的 添加数据库。

图 12-116 选择 DWS 数据连接

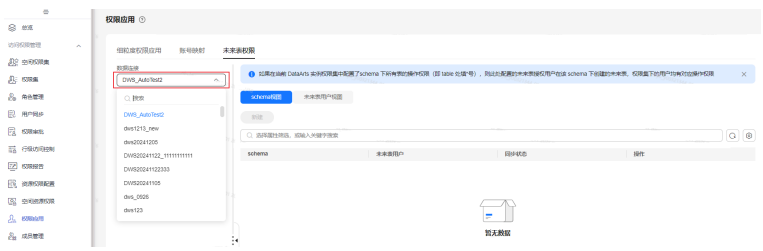
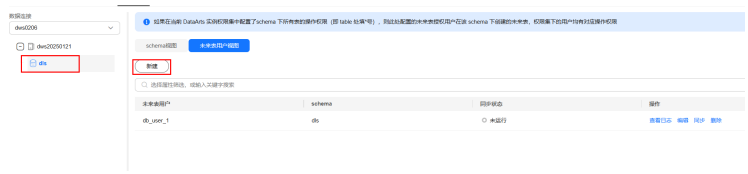


图 12-117 添加数据库

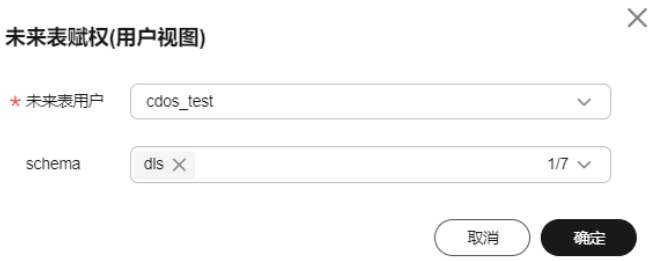
步骤4 单击待配置的数据库，在未来表用户视图下单击“新建”，创建未来表配置。

图 12-118 创建未来表配置



步骤5 在弹出的窗口中，为未来表的建表用户指定Schema，单击确定完成配置。

图 12-119 未来表配置



步骤6 配置完成后，单击“同步”完成未来表配置。

同步成功后，当未来表用户在指定Schema下创建未来表时，当前实例下所有对该Schema拥有全表权限的用户，将自动获得对所创建未来表的相应权限。

图 12-120 同步未来表配置

----结束

12.3.6 服务资源访问控制

12.3.6.1 配置队列权限

本章介绍如何通过队列权限管理，为当前工作空间分配可使用的MRS Yarn和DLI队列资源，并为用户组/用户配置对应的队列权限策略。

当前队列分配和队列权限管控均为白名单机制。即如果未分配队列，则无法选择队列；如果队列未对用户授权，则用户无法使用队列。

- 当为工作空间分配队列资源后，在数据开发组件在为作业节点配置队列资源时，可选择的队列为当前空间下已分配的队列资源。

  说明

  当前支持在选择MRS Yarn队列时，从已分配的队列资源获取队列列表。如果未分配队列资源，则只支持选择root.default队列。

- 当为用户组/用户配置队列权限后，MRS队列权限管控由MRS Ranger组件实现，DLI队列权限管控由DLI服务实现，仅被授权用户具备相应队列权限。

  说明

  需要说明的是，默认在DataArts Studio数据开发组件使用队列时，数据源会使用数据连接上的账号进行认证鉴权。因此当用户在数据开发时，队列权限管控依然无法生效。需要您启用细粒度认证，使得在数据开发使用队列时，使用当前用户身份认证鉴权，从而使队列权限管控生效。

前提条件

- 仅DAYU Administrator、Tenant Administrator或者数据安全管理员有权限给当前空间分配可用的队列资源、配置MRS队列属性（离线/实时）以及为指定的队列配置用户权限策略，另外工作空间管理员用户也可以为用户组/用户配置队列权限策略。
- 配置队列权限前，已在管理中心创建数据湖探索（DLI）和MapReduce服务（MRS Ranger）类型的数据连接，请参考**创建DataArts Studio数据连接**。
- 配置MRS Yarn队列权限前，需要参考**同步IAM用户到数据源**将IAM上的用户信息同步到数据源上。
- MRS Yarn队列权限的策略生效，需要配置YARN严格权限控制，即设置参数“yarn.acl.enable”为true，具体请参见**参考：配置Yarn严格权限控制**。

约束与限制

- 当前分配队列资源只支持MRS Yarn队列。队列权限管控只支持MRS Yarn和DLI队列，且由于DLI限制暂不支持为DLI default队列授权。
- 仅当数据连接中的Agent选择的CDM集群为2.10.0.300及以上版本时，才支持MRS Yarn队列权限管控。
- 仅DAYU Administrator、Tenant Administrator或者数据安全管理员有权限给当前空间分配可用的队列资源、配置MRS队列属性（离线/实时）以及为指定的队列配置用户权限策略，另外工作空间管理员用户也可以为用户组/用户配置队列权限策略。
- 当前工作空间分配的队列资源和配置的队列权限并无绑定关系，队列权限策略实际上落在数据源配置中。因此，当删除当前工作空间的队列资源后，已配置的队列权限策略依然生效；重新添加队列资源后，权限依然可见。
- 已配置的队列权限策略借由数据源的权限管控能力实现，因此也可以在数据源（如MRS Ranger策略和DLI队列管理）处查看已配置的策略。如果在数据源处删

掉队列策略，则在数据安全组件处不会自动删除，需要您手动在数据安全组件处清理该策略。

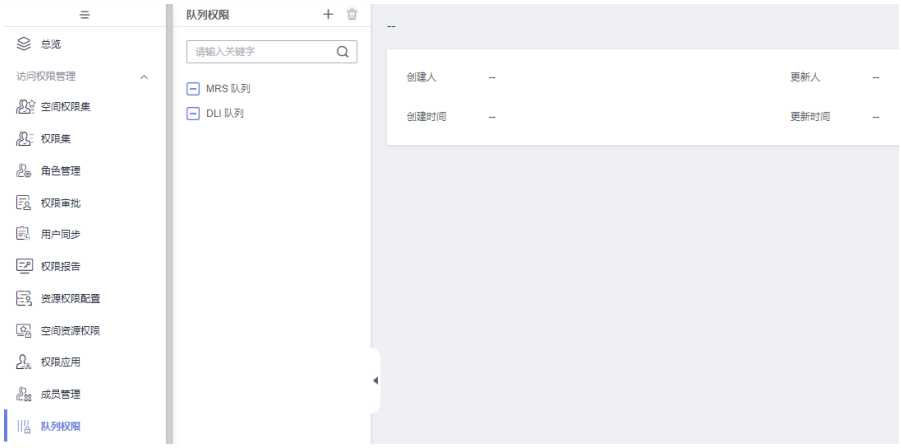
- 仅MRS Yarn队列支持配置队列属性（离线/实时），且同一队列在不同工作空间下支持指定为不同属性。
- 为DLI队列的授权时，当前由于DLI限制只支持授权给用户，不支持授权给用户组。

分配队列并授权

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“队列权限”，进入队列权限页面。

图 12-121 进入队列权限页面



步骤3 单击队列权限目录上方的+，为当前工作空间分配队列。在弹出的添加队列资源窗口配置相关参数，参考**表12-16**，配置完成单击“保存”，队列资源添加完成。

表 12-16 添加队列资源参数说明

参数名	参数描述
*资源类型	选择MRS队列或者DLI队列。
*数据连接	选择队列所在的数据连接。如需新建数据连接，请参考**创建DataArts Studio数据连接**。
*集群名称	仅当资源类型为MRS队列时显示，无需填写， 系统自动匹配数据连接对应的集群名称。

参数名	参数描述
*队列名称	选择需要授权的队列名称。 • 选择MRS队列时，可选队列来自于MRS集群的队列。可在MRS控制台的集群列表中，单击集群名进入集群详情后，在“租户管理 > 队列配置”下查看已有队列。 • 选择DLI队列时，可选队列来自于DLI服务中所购买的队列，可在DLI控制台中，在“资源管理 > 队列管理”下查看已有队列。另外，当前DLI队列分为SQL队列和通用队列两类，SQL队列用于运行SQL作业，通用队列用于运行Flink、Spark Jar作业。
描述	为更好地识别队列权限，此处加以描述信息。

图 12-122 添加队列资源

步骤4 单击队列权限目录中的队列，进入队列详情页面。

其中MRS Yarn队列可配置队列属性，主要应用于数据开发服务中的任务管理。实时队列用于运行实时作业，离线队列用于运行批处理作业，默认即不区分队列的作业类型。

图 12-123 MRS Yarn 队列详情

图 12-124 DLI 队列详情

步骤5 为分配的队列资源进行授权。

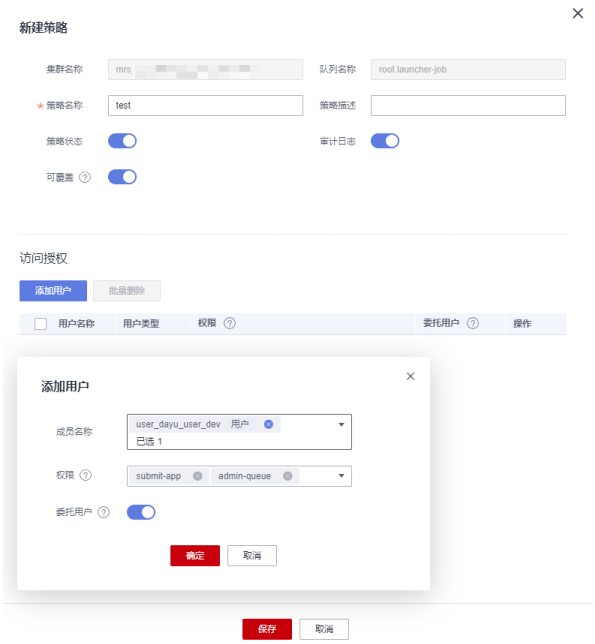
- **MRS Yarn队列**

  在MRS Yarn队列详情页面，单击“新建策略”，在弹出的窗口中参考**表12-17**填写策略相关参数，完成后单击保存，完成队列授权。

  表 12-17 MRS Yarn 队列策略参数说明

参数名	参数描述
集群名称	无需填写， 系统自动填写队列所在的集群名称。
队列名称	无需填写， 系统自动填写当前的队列名称。
*策略名称	用于标识MRS Yarn队列权限策略，为便于策略管理，建议名称中包含授权对象。
策略描述	为更好地识别策略，此处加以描述信息。
策略状态	开启后当前策略生效。
审计日志	开启后可记录当前队列的操作日志，需要在数据源侧查看对应的审计日志。
可覆盖	由于Ranger组件的限制，如果Ranger中已有此用户/用户组的队列权限策略，当前策略可能会被认为重复而添加失败。 开启可覆盖后，将尝试覆盖Ranger中此用户/用户组的队列权限策略。当覆盖失败时，需要您到Ranger组件中手动删除此用户/用户组的队列权限策略，然后再次重试添加策略。
***访问授权**（单击“添加用户”进入配置窗口）	
用户名称	选择需要授权的用户/用户组。用户/用户组列表来自于工作空间中已添加的用户/用户组。
权限	– submit-app：提交队列任务权限 – admin-queue：管理队列任务权限
委托用户	如果需要让待授权用户/用户组管理本条策略，可开启此选项，使这些用户成为当前策略管理员，当前策略管理员可以更新、删除本策略。

图 12-125 MRS Yarn 队列详情



- **DLI队列**

  在DLI队列详情页面，单击“授权”，在弹出的窗口中参考**表12-17**填写策略相关参数，完成后单击保存，完成队列授权。

表 12-18 DLI 队列授权参数说明

参数名	参数描述
用户名称	选择需要授权的用户。用户列表来自于工作空间中已添加的用户。 **说明** 为DLI队列的授权时，当前只支持用户，不支持用户组。
权限	- 提交作业：向此队列提交作业 - 取消作业：终止提交到此队列的作业 - 删除队列：删除此队列 - 赋权：当前用户可将队列的权限赋予其他用户 - 权限回收：当前用户可回收其他用户具备的该队列的权限，但不能回收该队列所有者的权限 - 查看其他用户具备的权限：当前用户可查看其他用户具备的该队列的权限 - 重启队列：重启此队列的权限 - 规格变更：修改队列规格的权限

图 12-126 DLI 队列详情

----结束

相关操作

- 删除队列资源：在队列权限目录页面，先单击选择需要删除的队列资源，然后单击目录上方的，即可删除队列资源。

  说明

  - 删除队列资源，不会将此队列资源从MRS/DLI中直接删除，而是不再将当前指定的队列资源分配给此工作空间。
  - 直接删除队列资源后，队列中的授权配置依然存在，权限继续生效。需要通过**删除策略**或**回收权限**，才能删除相应队列权限。
  - 数据安全无法删除当前正在数据开发中被使用的Yarn队列资源。
- 编辑策略：在MRS Yarn队列详情页面，单击对应策略操作栏中的“编辑”，即可编辑策略。
- 删除策略：在MRS Yarn队列详情页面，单击对应策略操作栏中的“删除”，即可删除策略。当需要批量删除时，可以在勾选策略后，在策略列表上方单击“批量删除”。

  说明

  删除操作无法撤销，请谨慎操作。
- 修改权限：在DLI队列详情页面，单击对应授权项操作栏中的“修改权限”，即可修改对该用户所授予的权限。
- 回收权限：在DLI队列详情页面，单击对应授权项操作栏中的“回收权限”，即可删除对该用户所授予的权限。

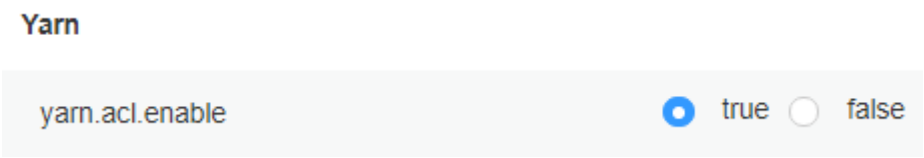
参考：配置 Yarn 严格权限控制

- 操作步骤为：

  a. 登录FusionInsight Manager界面，选择“集群 > 服务 > Yarn”。

  b. 选择“配置 > 全部配置”，搜索参数“yarn.acl.enable”，修改参数值为“true”。如果该参数值已经为“true”，则无需处理。

图 12-127 配置参数“yarn.acl.enable”



在配置Yarn队列权限前，需要启用Yarn队列的权限控制机制。

步骤1 登录MRS FusionInsight Manager。

步骤2 选择“集群 > 服务 > Yarn > 配置 > 基础配置”，在搜索框中输入参数名“yarn.acl.enable”，修改参数值为“true”。如果该参数值已经为“true”，则无需处理。如图12-128所示。

图 12-128 配置 yarn.acl.enable 参数示例

步骤3 参数配置完成后，单击左上角的“保存”，在弹窗中单击“确定”保存配置。

步骤4 保存成功后，切换到实例页签，选择配置已过期的实例后，单击“更多 > 滚动重启实例”，使配置生效。

图 12-129 滚动重启实例

----结束

12.3.6.2 配置空间资源权限策略

本章介绍如何通过空间资源权限策略，基于用户、用户组或角色，实现对管理中心所有数据连接和IAM委托（仅限于委托对象为“数据湖治理中心 DGC”的云服务委托）的精细权限控制。

- 当未配置某资源的空间资源权限策略时，所有用户默认可以查看并使用该资源。
- 当将某资源（例如某个连接或者某个委托）赋权给任一用户、用户组或角色后，对于非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户）而言，则无权再查看并使用此资源。

前提条件

仅DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户有权限新建、编辑或删除空间资源权限策略。

约束与限制

- 当前仅支持简单模式的工作空间资源管控，不支持企业模式。
- 如果未对某资源进行赋权，则默认该资源权限放开，不做权限管控。
- 当前仅数据开发组件支持空间资源权限策略，其他组件不受空间资源权限策略限制。在数据开发组件如下场景中，会根据空间资源权限策略进行鉴权。
  - 脚本开发或者作业开发中，选择连接或作业委托、公共委托。
  - 提交脚本或者作业。
- 对于历史版本中直接在数据开发组件创建的数据连接，暂不支持进行资源权限管理。
- 对于已有的空间资源权限策略，当已删除对应资源后，策略不会随之自动删除。

新建空间资源权限策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“空间资源权限”，进入空间资源权限页面。

图 12-130 进入空间资源权限页面

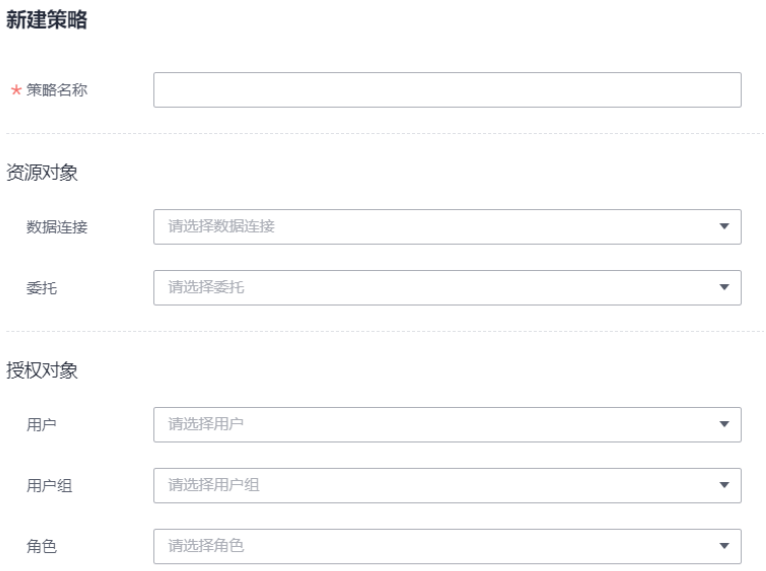
步骤3 单击空间资源权限页面的“新建”，在弹出的策略配置页参考**表12-19**配置相关参数，配置完成单击“保存”，策略配置完成。

表 12-19 配置空间资源权限策略参数说明

参数名	参数描述
*策略名称	标识空间资源权限策略，为便于策略管理，建议名称中包含资源对象和授权对象。
资源对象	
数据连接	选择需要授权的管理中心组件数据连接。如需新建数据连接，请参考**创建DataArts Studio数据连接**。 **说明** • 对于未选择的数据连接，则默认该连接权限放开，不做权限管控。 • 对于选择的数据连接，则非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户）将无权再查看并使用该连接。并且当查看或修改已使用该连接的作业时，数据连接及连接相关配置不可见。

参数名	参数描述
委托	选择需要授权的IAM委托，仅限于委托对象为“数据湖治理中心 DGC”的云服务类型委托。如需新建委托，请参考**参考：创建委托**。 **说明** • 对于未选择的委托，则默认该委托权限放开，不做权限管控。 • 对于选择的委托，则非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户）将无权再查看并使用该委托。
授权对象	
用户	选择需要授权的用户。用户列表来自于工作空间用户。
用户组	选择需要授权的用户组。用户组列表来自于工作空间用户组。
角色	选择需要授权的角色。角色列表来自于系统预置角色和自定义角色。

图 12-131 新建空间资源权限策略



----结束

相关操作

- 编辑策略：在空间资源权限页面，单击对应策略操作栏中的“编辑”，即可编辑策略。
- 删除策略：在空间资源权限页面，单击对应策略操作栏中的“删除”，即可删除策略。当需要批量删除时，可以在勾选策略后，在策略列表上方单击“批量删除”。

说明

删除操作无法撤销，请谨慎操作。

12.3.6.3 配置目录权限（高级特性）

本章介绍如何通过目录权限策略，基于用户、用户组或角色，对数据开发中脚本和作业的目录、数据服务专享版中API的目录以及数据架构中的物理模型和逻辑模型进行权限控制。

- 当工作空间内未配置数据开发、数据服务和数据架构的目录权限策略时，所有用户默认可以查看并操作数据开发、数据服务和数据架构的目录及其中的资源项。
- 当工作空间内已配置数据开发的脚本或作业目录权限策略时，对于非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户）而言，数据开发中的所有脚本和作业目录将由于无权限而置灰，具体影响包括：不能新建、编辑、查看、删除、导入导出目录下的作业或脚本，但是新建目录、作业关联脚本、选择依赖作业、配置全部作业告警、查看操作历史、备份作业、监控作业等操作不受限制。
- 当工作空间内已配置数据服务的API目录权限策略时，对于非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户）而言，数据服务中的所有API目录将由于无权限而置灰，具体影响包括：不能新建、编辑、查看、删除、导入导出目录下的API，但是新建目录、事件、日志、审核等操作不受限制。
- 当工作空间内已配置数据架构的物理模型和逻辑模型目录权限策略时，对于非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户）而言，数据架构中的模型将由于无权限而置灰，具体影响包括：不能新建、编辑、查看、删除、模型下的表，但是新建模型、审核等操作不受限制。

说明

在新版本模式下仅当使用企业版时，才支持配置目录权限。旧版本模式使用基础版及更高版本时即可支持。

前提条件

- 在配置数据开发、数据服务或数据架构的目录权限策略前，您应在数据开发、数据服务或数据架构组件已创建相关目录。
- 仅DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户有权限新建、编辑或删除目录权限策略。

约束与限制

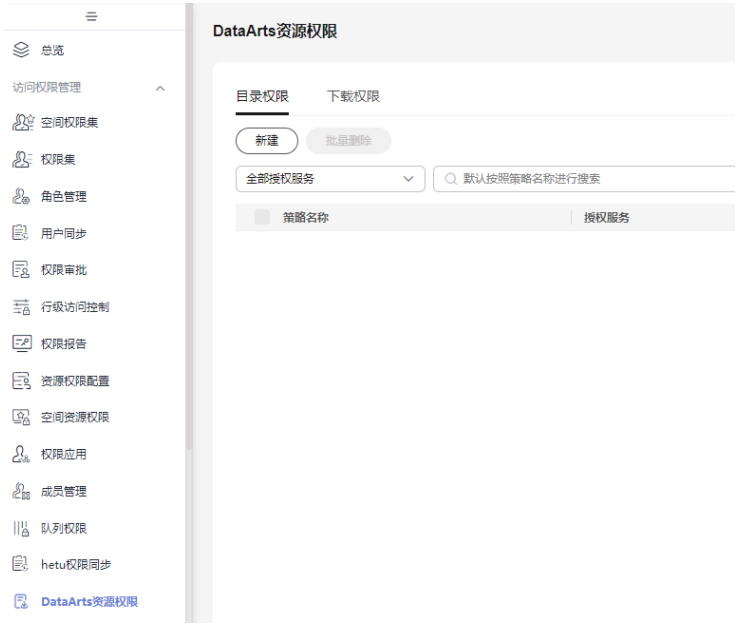
- 一旦在当前工作空间内配置了数据开发、数据服务或数据架构的目录权限策略，将导致非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户）无法再查看并操作数据开发、数据服务的目录及其中的资源，请您谨慎操作。
- 仅DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户有权限新建、编辑或删除目录权限策略。
- 目录权限策略中可以配置多个目录，但同一用户、用户组或角色仅能出现在一条策略中。

新建目录权限策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“DataArts资源权限”，进入目录权限页面。

图 12-132 进入目录权限页面



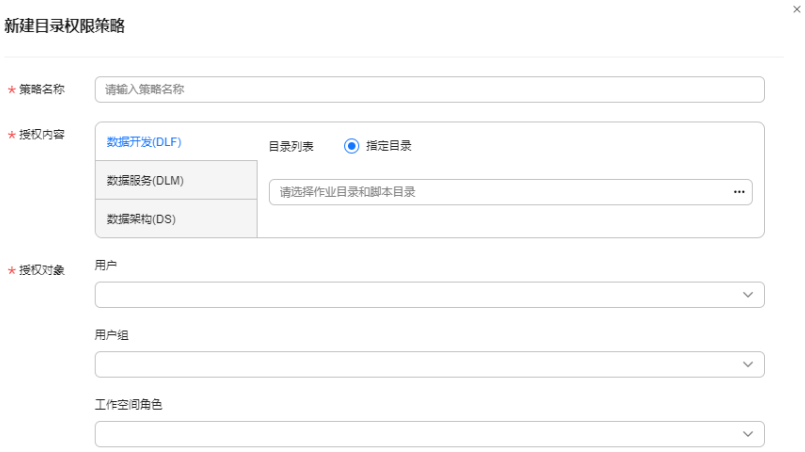
步骤3 单击目录权限页面的“新建”，在弹出的策略配置页参考**表12-20**配置相关参数，配置完成单击“提交”，策略配置完成。

表 12-20 配置目录权限策略参数说明

参数名	参数描述
*策略名称	标识目录权限策略，为便于策略管理，建议名称中包含资源对象和授权对象。
授权内容	
数据开发（DLF）	选择需要授权的数据开发脚本和作业的一级目录。 **说明** • 即使仅选择脚本目录或仅选择作业目录，策略配置后，对于非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户）而言，数据开发中的所有脚本和作业目录将由于无权限而置灰。 • 如果仅选择了数据开发的脚本或作业目录，则数据服务的目录权限不受此策略影响。
数据服务（DLM）	选择需要授权的数据服务API的一级目录。 **说明** 如果仅选择了数据服务的API目录，则数据开发的目录权限不受此策略影响。

参数名	参数描述
数据架构（DS）	选择需要授权的数据架构的物理模型或逻辑模型。 **说明** • 即使仅选择物理模型或仅选择逻辑模型，策略配置后，对于非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator、数据安全管理员或预置的工作空间管理员角色的用户）而言，数据架构中的所有物理模型和逻辑模型将由于无权限而置灰。 • 如果仅选择了数据架构的物理模型或逻辑模型目录，则数据开发或数据服务的目录权限不受此策略影响。
授权对象	
用户	选择需要授权的用户。用户列表来自于工作空间用户。
用户组	选择需要授权的用户组。用户组列表来自于工作空间用户组。
角色	选择需要授权的角色。角色列表来自于系统预置角色和自定义角色。

图 12-133 新建目录权限策略



----结束

相关操作

- 编辑策略：在目录权限页面，单击对应策略操作栏中的“编辑”，即可编辑策略。
- 删除策略：在目录权限页面，单击对应策略操作栏中的“删除”，即可删除策略。当需要批量删除时，可以在勾选策略后，在策略列表上方单击“批量删除”。

  说明

  删除操作无法撤销，请谨慎操作。

12.3.6.4 配置下载权限（高级特性）

本章介绍如何通过下载权限策略，基于用户或用户组，对数据开发中SQL脚本执行结果的转储以及在下载中心下载操作进行权限控制。

- DataArts Studio实例中默认具备命名为“SYSTEM_GENERATE_DEFAULT_DATA_DOWNLOAD_POLICY”的默认下载权限策略，放开所有用户的导出权限。默认策略允许修改和删除。
- 如果没有任何下载权限策略，则默认放开所有用户的导出权限。
- 如果存在一条或多条下载权限策略，用户的最终权限取决于多个策略中的用户权限并集。对于非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator或数据安全管理员）而言，数据开发中SQL脚本执行结果的转储以及在下载中心下载操作将会受限，如果试图操作则系统报错。

值得注意的是，下载权限策略为DataArts Studio实例级别配置，各工作空间之间数据互通。

说明

在新版本模式下仅当使用企业版时，才支持配置下载权限。旧版本模式使用基础版及更高版本时即可支持。

前提条件

- 仅DAYU Administrator、Tenant Administrator或数据安全管理员有权限新建、编辑或删除下载权限策略。
- 在配置下载权限策略前，应确保授权对象已具备在数据开发组件中SQL脚本执行结果的转储以及在下载中心下载操作权限（即已被授予DataArts Studio权限并被添加为对应工作空间角色，详见**授权用户使用DataArts Studio**），且已在数据开发中通过配置“数据导出策略”默认项允许授权对象进行数据导出（详见**配置默认项**）。否则，即使已在下载权限策略中为用户授权了转储以及在下载中心下载权限，用户依然无法进行相关操作。

约束与限制

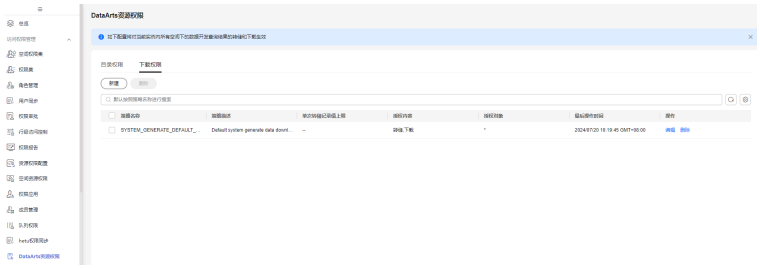
- 仅DAYU Administrator、Tenant Administrator或数据安全管理员有权限新建、编辑或删除下载权限策略。
- 通过下载权限策略为用户授权，前提是用户本身的具备相关操作权限并且数据开发中的“数据导出策略”配置项已授权，否则用户依然无法进行相关操作。
- 配置下载权限策略后，将导致非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator或数据安全管理员）无法再进行转储以及在下载中心下载操作，请您谨慎操作。
- 下载权限策略不支持直接对SQL脚本执行结果直接下载的操作进行管控，仅支持对转储以及在下载中心下载操作进行管控，并支持配置单次转储记录值上限。
- 每个用户或用户组只能有一条下载权限策略，但与全部成员的策略不冲突。

新建下载权限策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“DataArts资源权限”，在DataArts资源权限页面单击“下载权限”，进入下载权限页签。

图 12-134 进入下载权限页签



步骤3 单击下载权限页签的“新建”，在弹出的策略配置页参考**表12-21**配置相关参数，配置完成单击“提交”，策略配置完成。

表 12-21 配置下载权限策略参数说明

参数名	参数描述
*策略名称	标识下载权限策略，为便于策略管理，建议名称中包含授权对象。 策略名须以英文或中文开头，仅支持中英文、数字和下划线，最多64个字符。
策略描述	为更好地识别策略，此处加以描述信息。
*授权内容	授权对象默认为数据开发组件，需选择需要授权的操作，并支持配置单次转储记录值上限。 **说明** 数据开发组件中不同数据源的SQL脚本转储支持的单次最大记录值不同，详见**下载或转储脚本执行结果**。此处配置的记录值上限可参考此规格进行配置。
*授权对象	选择需要授权的用户。 ● 指定用户：可以配置为指定的用户以及用户组。 **说明** 每个用户或用户组只能有一条下载权限策略，但与全部成员的策略不冲突。 ● 全部成员（包含新增成员）：为全体成员配置操作策略。

图 12-135 新建下载权限策略

步骤4 （可选）在策略列表中，单击默认下载权限策略操作栏中的“删除”，删除默认策略。

为实现仅新建策略中的授权对象具备相关操作权限、非授权对象的普通用户（即非DAYU Administrator、Tenant Administrator或数据安全管理员）不具备转储以及在下载中心下载操作权限，您需要删除默认下载权限策略，否则所有用户依然具有转储以及在下载中心下载操作权限。

----结束

相关操作

- 编辑策略：在下载权限页签，单击对应策略操作栏中的“编辑”，即可编辑策略。
- 删除策略：在下载权限页签，单击对应策略操作栏中的“删除”，即可删除策略。当需要批量删除时，可以在勾选策略后，在策略列表上方单击“批量删除”。

  说明

  删除操作无法撤销，请谨慎操作。

12.3.7 Ranger 权限访问控制

12.3.7.1 配置资源权限

本章主要介绍如何通过资源权限创建权限策略到Ranger组件，实现MRS资源权限控制，从而降低企业数据信息安全风险。

当前支持创建的权限策略如下：

- **创建HDFS权限策略**
- **创建Hive访问权限策略**
- **创建Hive脱敏权限策略**
- **创建Hive行级过滤器权限策略**
- **创建HBase权限策略**
- **创建Yarn权限策略**
- **创建Kafka权限策略**
- **创建Storm权限策略**

前提条件

- 已在管理中心创建Ranger类型的数据连接，并确保已参考**MRS Ranger数据连接参数说明**填写正确的RangerAdmin业务IP和Ranger服务端口。

  说明

  在管理中心测试Ranger数据连接时，不会校验Ranger业务IP和服务端口，即使填写错误也不会提示，因此建议进行人工检查。

- 已开启对应MRS集群的Ranger鉴权功能，安全模式默认开启Ranger鉴权，普通模式默认关闭Ranger鉴权。详情请参考**启用Ranger鉴权**。

约束与限制

- 资源权限策略依赖于MRS集群的Ranger鉴权功能，当前仅支持对MRS资源进行权限控制。
- 权限策略配置完成后1分钟左右生效。

支持访问控制的 MRS 组件及权限列表

通过Ranger可以对MRS集群（MRS集群版本为3.0.0及以上）中的组件进行集成，实现组件的细粒度访问权限控制。目前已经支持的组件及相关权限如**表12-22**所示。具体权限解释可参考**MRS配置组件权限策略**。

表 12-22 支持的组件及权限列表

组件名	权限说明
HDFS	HDFS文件的权限： • Read：读权限 • Write：写权限 • Excute：执行权限

组件名	权限说明
Hive	Hive数据库、数据表、列的权限： ● Select：查询权限 ● Update：更新权限 ● Create：创建权限 ● Drop：drop操作权限 ● Alter：alter操作权限 ● All：所有执行权限 ● Temporary UDF Admin：临时UDF管理权限
Yarn	Yarn队列权限： ● submit-app：提交队列任务权限 ● admin-queue：管理队列任务权限
HBase	HBase列、列族的权限： ● Read：读权限 ● Write：写权限 ● Create：创建权限 ● Admin：管理员权限
Kafka	Kafka的Topic权限： ● Publish：生产权限 ● Consume：消费权限 ● Configure：topic扩容权限 ● Describe：查询权限 ● Create：创建主题权限 ● Delete：删除主题权限 ● Describe Configs：查询配置权限 ● Alter Configs：修改配置权限

组件名	权限说明
Storm	Storm的Topology权限： • Submit Topology：提交拓扑 • File Upload：上传文件 • File DownLoad：下载文件 • Kill Topology：删除拓扑 • Rebalance：Rebalance权限 • Activate：激活权限 • Deactivate：去激活权限 • Get Topology Conf：获取拓扑配置 • Get Topology：获取拓扑 • Get User Topology：获取用户拓扑 • Get Topology Info：获取拓扑信息 • Upload New Credential：上传新的凭证

创建 HDFS 权限策略

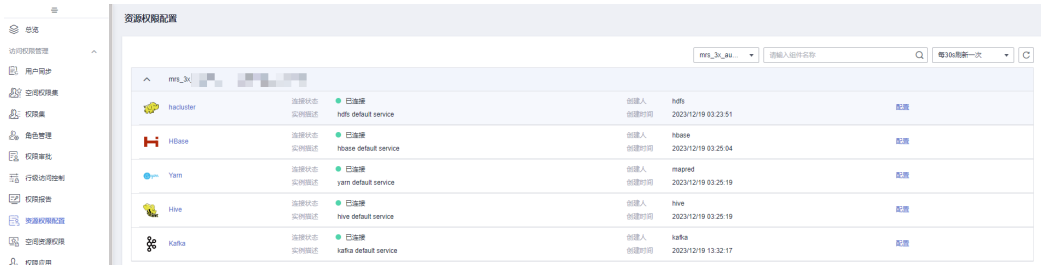
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“资源权限配置”，进入资源权限配置页面。

说明

如果报错“获取资源服务失败，由于[CDM返回为空：[404 NOT FOUND]]”，请在管理中心参考**MRS Ranger数据连接参数说明**，排查Ranger数据连接的RangerAdmin业务IP和Ranger服务端口是否正确。

图 12-136 资源权限配置页面



步骤3 单击待创建权限策略HDFS组件下“hacluster”的“配置”，进入配置界面单击“创建”，新建权限策略。

图 12-137 新建权限策略

步骤4 在弹出的策略配置页配置相关参数，配置完成单击“确定”，策略配置完成。

图 12-138 配置权限策略

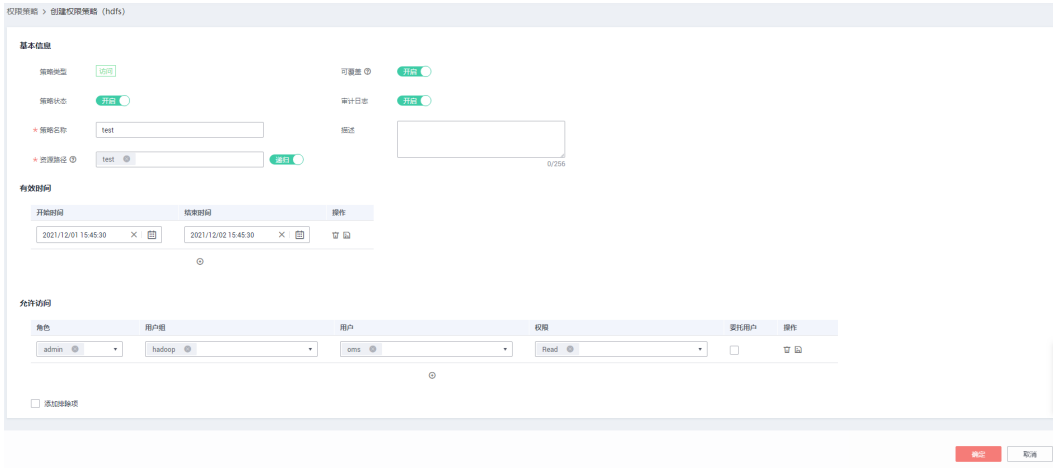


表 12-23 配置 HDFS 权限策略参数说明

参数名	参数描述
策略类型	根据用户所选服务组件自动生成。包括访问、脱敏、行过滤器，其中脱敏和行过滤器类型是Hive特有的。
策略状态	开启表示权限策略生效，关闭表示权限策略创建成功后不生效。默认开启。
可覆盖	开启可覆盖时，新创建的策略将覆盖当前策略（新策略生效而旧策略不生效）。默认开启。 当用户需要创建一个临时访问策略时，“可覆盖”可以配合“有效时间”一起使用，那么即使临时访问策略超过有效期失效后，也不影响原有的权限策略继续生效。
审计日志	开启表示记录日志，日志内容包括客户端访问时间、客户端IP、客户端用户、操作资源结果等信息。
策略名称	名称为必填项，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，且输入不能为空。
描述	对策略的描述信息，长度限制在256个字符以内。
资源路径	访问权限控制的HDFS路径。

参数名	参数描述
递归	开启表示资源路径为递归方式。关闭表示资源路径为非递归方式。默认开启。
有效时间	用户通过设置开始时间和结束时间来控制策略的生效时间段，可配置多条。
允许访问	定义允许访问的用户和用户组。 ● 用户：MRS服务的用户。 ● 角色：MRS服务的角色。 ● 用户组：MRS服务的用户组。 ● 权限：定义允许访问的用户拥有的权限。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 ● 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。
添加排除项	允许访问勾选“添加排除项”意思是在允许访问的用户组里添加禁止访问的用户。 禁止访问勾选“添加排除项”意思是在禁止访问的用户组里添加允许访问的用户。
拒绝所有其他访问	勾选此项表示只有策略中“允许访问”指定的用户或用户组可以访问，其他用户均禁止访问。
禁止访问	不勾选“拒绝所有其他访问”时显示此配置，该配置定义禁止访问的用户和用户组。 ● 用户：MRS服务的用户。 ● 角色：MRS服务的角色。 ● 用户组：MRS服务的用户组。 ● 权限：定义用户禁止的权限类型。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 ● 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。

----结束

创建 Hive 访问权限策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“资源权限配置”，进入资源权限配置页面。

说明

如果报错“获取资源服务失败，由于[CDM返回为空：[404 NOT FOUND]]”，请在管理中心参考**MRS Ranger数据连接参数说明**，排查Ranger数据连接的RangerAdmin业务IP和Ranger服务端口是否正确。

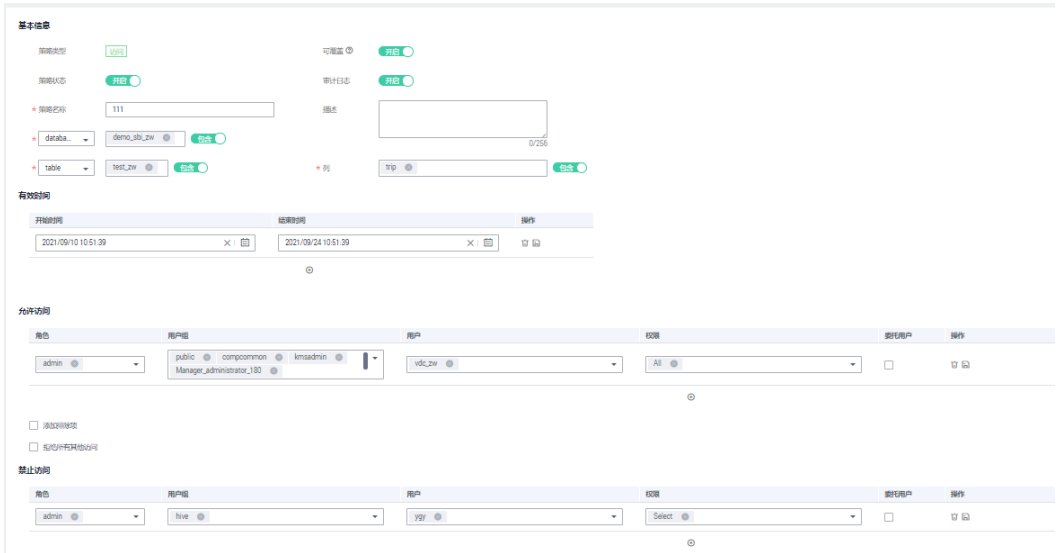
图 12-139 资源权限配置页面

步骤3 单击待创建权限策略Hive组件的“配置”，进入配置界面选择“访问”页签，单击“创建”，新建权限策略。

图 12-140 新建权限策略入口

步骤4 在弹出的策略配置页配置相关参数，配置完成单击“确定”，策略配置完成。

图 12-141 配置 Hive 权限策略



权限策略参数说明表：

表 12-24 Hive 权限策略参数说明表

参数名	参数描述
策略类型	根据用户所选服务组件自动生成。包括访问、脱敏、行过滤器，其中脱敏和行过滤器类型是Hive特有的。
策略状态	开启表示权限策略生效，关闭表示权限策略创建成功后不生效。默认开启。
可覆盖	开启可覆盖时，新创建的策略将覆盖当前策略（新策略生效而旧策略不生效）。默认开启。 当用户需要创建一个临时访问策略时，“可覆盖”可以配合“有效时间”一起使用，那么即使临时访问策略超过有效期失效后，也不影响原有的权限策略继续生效。
审计日志	开启表示记录日志，日志内容包括客户端访问时间、客户端IP、客户端用户、操作资源结果等信息。
策略名称	名称为必填项，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，且输入不能为空。
描述	对策略的描述信息，长度限制在256个字符以内。
数据库	必填项，此项表示需要进行权限控制的数据库，支持模糊搜索。
数据表	必填项，此项表示需要进行权限控制的数据表，支持模糊搜索。
列	必填项，此项表示需要进行权限控制的列，支持模糊搜索。
有效时间	用户通过设置开始时间和结束时间来控制策略的生效时间段，可配置多条。
允许访问	定义允许访问的用户和用户组。 ● 用户：MRS服务的用户。 ● 角色：MRS服务的角色。 ● 用户组：MRS服务的用户组。 ● 权限：定义允许访问的用户拥有的权限。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表 12-22**。 ● 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。
添加排除项	允许访问勾选“添加排除项”意思是在允许访问的用户组里添加禁止访问的用户。 禁止访问勾选“添加排除项”意思是在禁止访问的用户组里添加允许访问的用户。
拒绝所有其他访问	勾选此项表示只有策略中“允许访问”指定的用户或用户组可以访问，其他用户均禁止访问。

参数名	参数描述
禁止访问	不勾选“拒绝所有其他访问”时显示此配置，该配置定义禁止访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义用户禁止的权限类型。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考表12-22。 • 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。

----结束

创建 Hive 脱敏权限策略

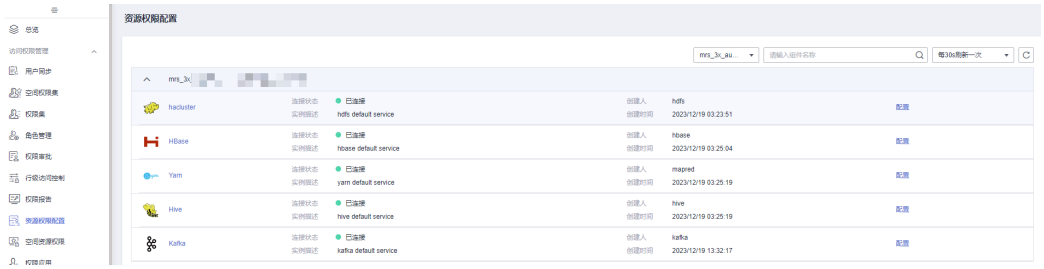
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“资源权限配置”，进入资源权限配置页面。

说明

如果报错“获取资源服务失败，由于[CDM返回为空：[404 NOT FOUND]]”，请在管理中心参考MRS Ranger数据连接参数说明，排查Ranger数据连接的RangerAdmin业务IP和Ranger服务端口是否正确。

图 12-142 资源权限配置页面



步骤3 单击待创建权限策略Hive组件的“配置”，进入配置界面选择“脱敏”页签，单击“创建”，新建权限策略。

图 12-143 新建权限策略界面

步骤4 在弹出的策略配置页配置相关参数，配置完成单击“确定”，策略配置完成。

图 12-144 配置 Hive 权限策略界面

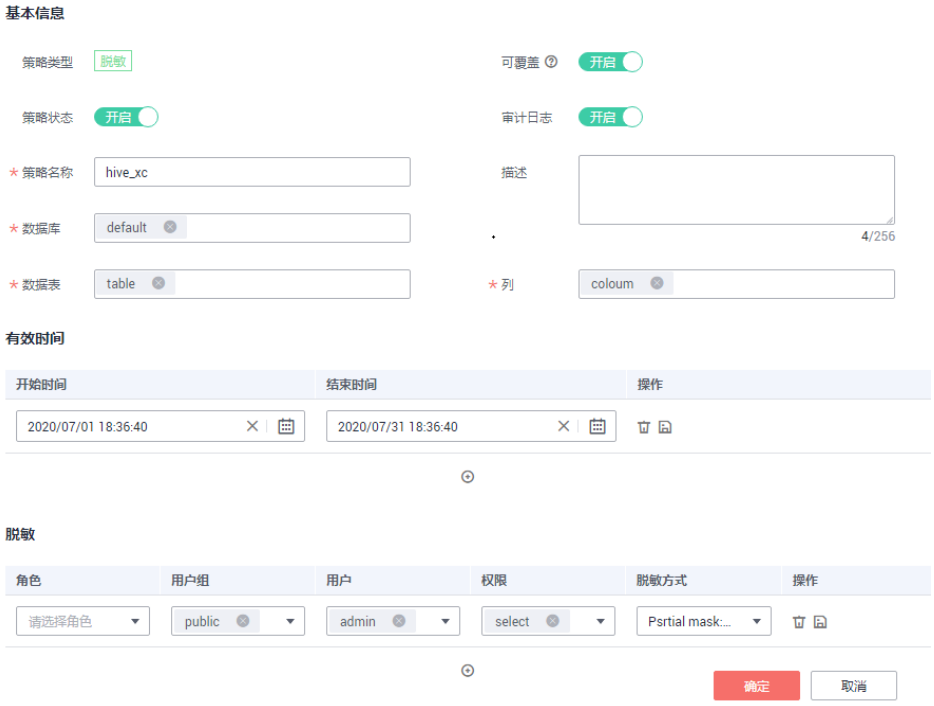


表 12-25 Hive 权限策略参数说明表

参数名	参数描述
策略类型	根据用户所选服务组件自动生成。包括访问、脱敏、行过滤器，其中脱敏和行过滤器类型是Hive特有的。
策略状态	开启表示权限策略生效，关闭表示权限策略创建成功后不生效。默认开启。

参数名	参数描述
可覆盖	开启可覆盖时，新创建的策略将覆盖当前策略（新策略生效而旧策略不生效）。默认开启。 当用户需要创建一个临时访问策略时，“可覆盖”可以配合“有效时间”一起使用，那么即使临时访问策略超过有效期失效后，也不影响原有的权限策略继续生效。
审计日志	开启表示记录日志，日志内容包括客户端访问时间、客户端IP、客户端用户、操作资源结果等信息。
策略名称	名称为必填项，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，且输入不能为空。
描述	对策略的描述信息，长度限制在256个字符以内。
数据库	必填项，此项表示需要进行权限控制的数据库，支持模糊搜索。
数据表	必填项，此项表示需要进行权限控制的数据表，支持模糊搜索。
列	必填项，此项表示需要进行权限控制的列，支持模糊搜索。
有效时间	用户通过设置开始时间和结束时间来控制策略的生效时间段，可配置多条。
脱敏	定义用户或用户组访问数据的脱敏方式。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义允许访问的用户拥有的权限。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 • 脱敏方式：按照该参数选定的值对Hive表中需要进行权限控制的列进行脱敏。

----结束

创建 Hive 行级过滤器权限策略

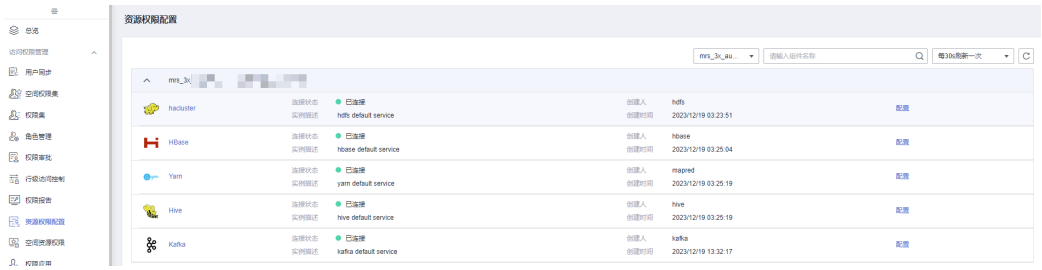
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“资源权限配置”，进入资源权限配置页面。

说明

如果报错“获取资源服务失败，由于[CDM返回为空：[404 NOT FOUND]]”，请在管理中心参考**MRS Ranger数据连接参数说明**，排查Ranger数据连接的RangerAdmin业务IP和Ranger服务端口是否正确。

图 12-145 资源权限配置页面



步骤3 单击待创建权限策略Hive组件的“配置”，进入配置界面选择“行级过滤器”页签，单击“创建”，新建权限策略。

图 12-146 创建 Hive 行级过滤器权限策略

步骤4 在弹出的策略配置页配置相关参数，配置完成单击“确定”，策略配置完成。

图 12-147 配置 Hive 权限策略参数

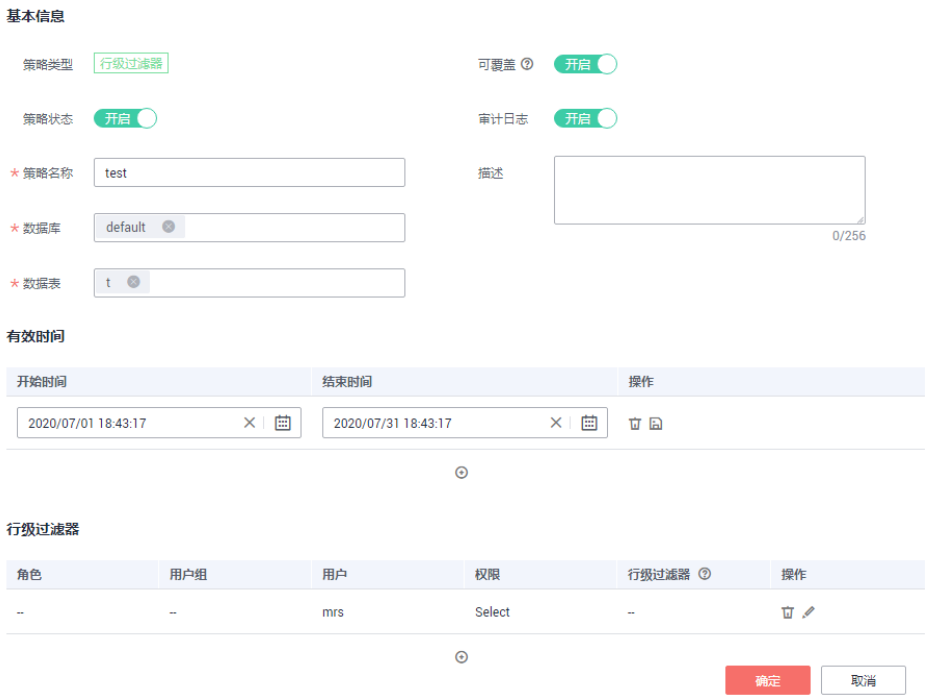


表 12-26 Hive 权限策略参数说明表

参数名	参数描述
策略类型	根据用户所选服务组件自动生成。包括访问、脱敏、行过滤器，其中脱敏和行过滤器类型是Hive特有的。
策略状态	开启表示权限策略生效，关闭表示权限策略创建成功后不生效。默认开启。
可覆盖	开启可覆盖时，新创建的策略将覆盖当前策略（新策略生效而旧策略不生效）。默认开启。 当用户需要创建一个临时访问策略时，“可覆盖”可以配合“有效时间”一起使用，那么即使临时访问策略超过有效期失效后，也不影响原有的权限策略继续生效。
审计日志	开启表示记录日志，日志内容包括客户端访问时间、客户端IP、客户端用户、操作资源结果等信息。
策略名称	名称为必填项，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，且输入不能为空。
描述	对策略的描述信息，长度限制在256个字符以内。
数据库	必填项，此项表示需要进行权限控制的数据库，支持模糊搜索。
数据表	必填项，此项表示需要进行权限控制的数据表，支持模糊搜索。
列	必填项，此项表示需要进行权限控制的列，支持模糊搜索。
有效时间	用户通过设置开始时间和结束时间来控制策略的生效时间段，可配置多条。
行级过滤器	定义允许访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义允许访问的用户拥有的权限。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表 12-22**。 • 行级过滤器：根据字段内容进行过滤，格式一般为：属性=属性值。例如：state=1。

----结束

创建 HBase 权限策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“资源权限配置”，进入资源权限配置页面。

说明

如果报错“获取资源服务失败，由于[CDM返回为空：[404 NOT FOUND]]”，请在管理中心参考**MRS Ranger数据连接参数说明**，排查Ranger数据连接的RangerAdmin业务IP和Ranger服务端口是否正确。

图 12-148 资源权限配置页面

步骤3 单击待创建权限策略HBase组件的“配置”，进入配置界面单击“创建”，新建权限策略。

图 12-149 创建 HBase 权限策略

步骤4 在弹出的策略配置页配置相关参数，配置完成单击“确定”，策略配置完成。

图 12-150 配置 HBase 权限策略

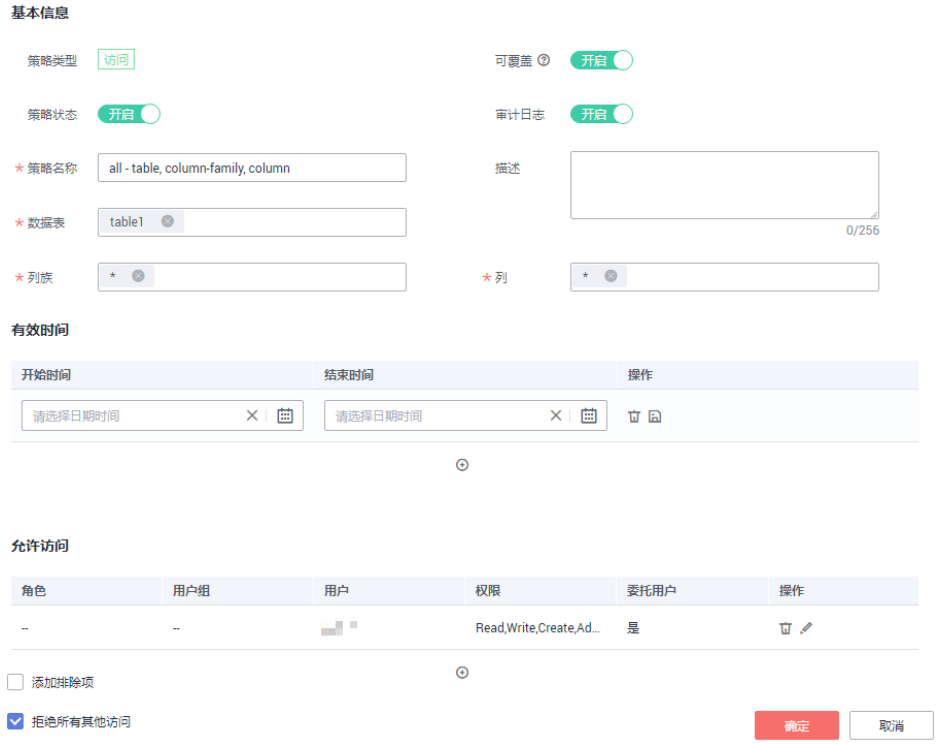


表 12-27 HBase 权限策略参数表

参数名	参数描述
策略类型	根据用户所选服务组件自动生成。包括访问、脱敏、行过滤器，其中脱敏和行过滤器类型是Hive特有的。
策略状态	开启表示权限策略生效，关闭表示权限策略创建成功后不生效。默认开启。
可覆盖	开启可覆盖时，新创建的策略将覆盖当前策略（新策略生效而旧策略不生效）。默认开启。 当用户需要创建一个临时访问策略时，“可覆盖”可以配合“有效时间”一起使用，那么即使临时访问策略超过有效期失效后，也不影响原有的权限策略继续生效。
审计日志	开启表示记录日志，日志内容包括客户端访问时间、客户端IP、客户端用户、操作资源结果等信息。
策略名称	名称为必填项，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，且输入不能为空。
描述	对策略的描述信息，长度限制在256个字符以内。
数据表	必填项，此项表示需要进行权限控制的数据表，支持模糊搜索。
列	必填项，此项表示需要进行权限控制的列，支持模糊搜索。
列族	必填项，此项表示HBase中Column Family，多列的集合。

参数名	参数描述
有效时间	用户通过设置开始时间和结束时间来控制策略的生效时间段，可配置多条。
允许访问	定义允许访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义允许访问的用户拥有的权限。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 • 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。
添加排除项	允许访问勾选“添加排除项”意思是在允许访问的用户组里添加禁止访问的用户。 禁止访问勾选“添加排除项”意思是在禁止访问的用户组里添加允许访问的用户。
拒绝所有其他访问	勾选此项表示只有策略中“允许访问”指定的用户或用户组可以访问，其他用户均禁止访问。
禁止访问	不勾选“拒绝所有其他访问”时显示此配置，该配置定义禁止访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义用户禁止的权限类型。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 • 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。

----结束

创建 Yarn 权限策略

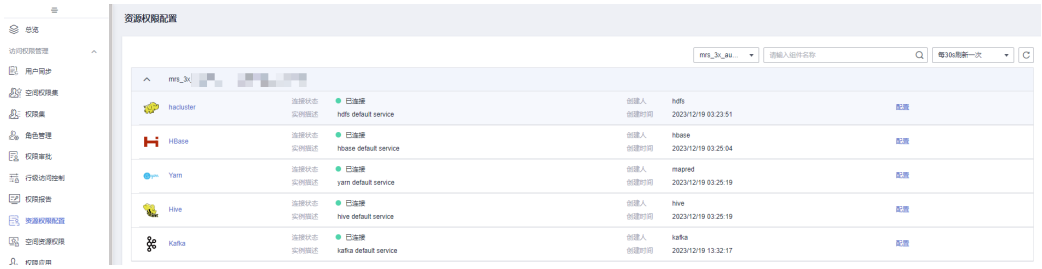
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“资源权限配置”，进入资源权限配置页面。

说明

如果报错“获取资源服务失败，由于[CDM返回为空：[404 NOT FOUND]]”，请在管理中心参考**MRS Ranger数据连接参数说明**，排查Ranger数据连接的RangerAdmin业务IP和Ranger服务端口是否正确。

图 12-151 资源权限配置页面



步骤3 单击待创建权限策略Yarn组件的“配置”，进入配置界面单击“创建”，新建权限策略。

图 12-152 新建 Yarn 权限策略

步骤4 在弹出的策略配置页配置相关参数，配置完成单击“确定”，完成策略配置。

图 12-153 配置 Yarn 权限策略

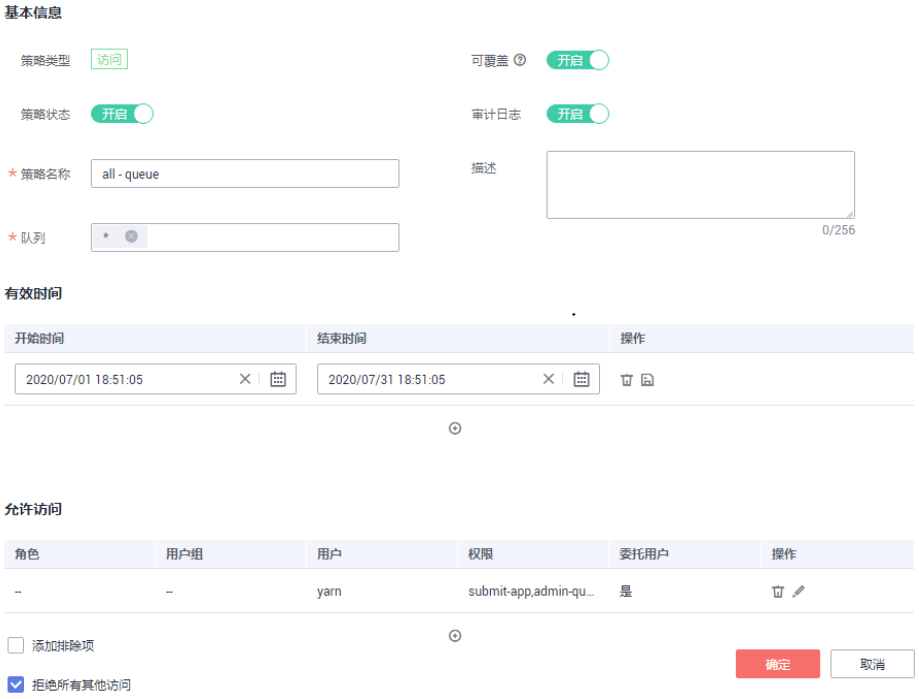


表 12-28 Yarn 权限策略参数表

参数名	参数描述
策略类型	根据用户所选服务组件自动生成。包括访问、脱敏、行过滤器，其中脱敏和行过滤器类型是Hive特有的。
策略状态	开启表示权限策略生效，关闭表示权限策略创建成功后不生效。默认开启。
可覆盖	开启可覆盖时，新创建的策略将覆盖当前策略（新策略生效而旧策略不生效）。默认开启。 当用户需要创建一个临时访问策略时，“可覆盖”可以配合“有效时间”一起使用，那么即使临时访问策略超过有效期失效后，也不影响原有的权限策略继续生效。
审计日志	开启表示记录日志，日志内容包括客户端访问时间、客户端IP、客户端用户、操作资源结果等信息。
策略名称	名称为必填项，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，且输入不能为空。
描述	对策略的描述信息，长度限制在256个字符以内。
队列	Yarn服务中的资源调度队列。
有效时间	用户通过设置开始时间和结束时间来控制策略的生效时间段，可配置多条。

参数名	参数描述
允许访问	定义允许访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义允许访问的用户拥有的权限。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 • 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。
添加排除项	允许访问勾选“添加排除项”意思是在允许访问的用户组里添加禁止访问的用户。 禁止访问勾选“添加排除项”意思是在禁止访问的用户组里添加允许访问的用户。
拒绝所有其他访问	勾选此项表示只有策略中“允许访问”指定的用户或用户组可以访问，其他用户均禁止访问。
禁止访问	不勾选“拒绝所有其他访问”时显示此配置，该配置定义禁止访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义用户禁止的权限类型。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 • 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。

----结束

创建 Kafka 权限策略

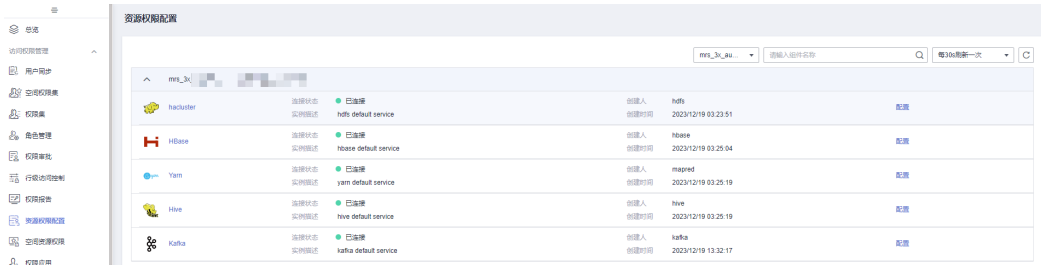
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“资源权限配置”，进入资源权限配置页面。

说明

如果报错“获取资源服务失败，由于[CDM返回为空：[404 NOT FOUND]]”，请在管理中心参考**MRS Ranger数据连接参数说明**，排查Ranger数据连接的RangerAdmin业务IP和Ranger服务端口是否正确。

图 12-154 资源权限配置页面



步骤3 单击待创建权限策略Kafka组件的“配置”，进入配置界面单击“创建”，新建权限策略。

图 12-155 新建 kafka 权限策略

步骤4 在弹出的策略配置页配置相关参数，配置完成单击“确定”，策略配置完成。

图 12-156 配置 Kafka 权限策略

表 12-29 Kafka 权限策略参数表

参数名	参数描述
策略类型	根据用户所选服务组件自动生成。包括访问、脱敏、行过滤器，其中脱敏和行过滤器类型是Hive特有的。
策略状态	开启表示权限策略生效，关闭表示权限策略创建成功后不生效。默认开启。
可覆盖	开启可覆盖时，新创建的策略将覆盖当前策略（新策略生效而旧策略不生效）。默认开启。 当用户需要创建一个临时访问策略时，“可覆盖”可以配合“有效时间”一起使用，那么即使临时访问策略超过有效期失效后，也不影响原有的权限策略继续生效。
审计日志	开启表示记录日志，日志内容包括客户端访问时间、客户端IP、客户端用户、操作资源结果等信息。
策略名称	名称为必填项，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，且输入不能为空。

参数名	参数描述
描述	对策略的描述信息，长度限制在256个字符以内。
策略条件	指定可访问Kafka主题的IP地址范围。
Topic	Kafka集群的消息主题。
有效时间	用户通过设置开始时间和结束时间来控制策略的生效时间段，可配置多条。
允许访问	定义允许访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义允许访问的用户拥有的权限。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 • 策略条件：指定可访问Kafka主题的IP地址范围。 • 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。
添加排除项	允许访问勾选“添加排除项”意思是在允许访问的用户组里添加禁止访问的用户。 禁止访问勾选“添加排除项”意思是在禁止访问的用户组里添加允许访问的用户。
禁止访问	不勾选“拒绝所有其他访问”时显示此配置，该配置定义禁止访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义用户禁止的权限类型。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 • 策略条件：指定可访问Kafka主题的IP地址范围。 • 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。

----结束

创建 Storm 权限策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“资源权限配置”，进入资源权限配置页面。

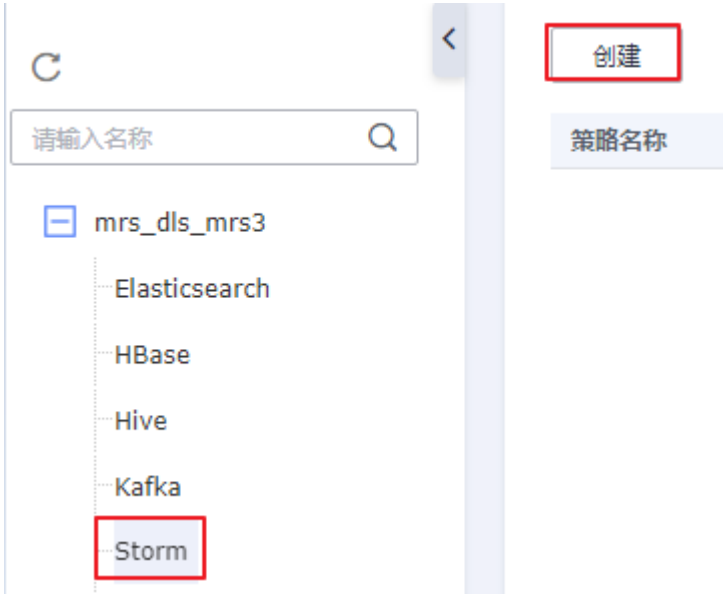
说明

如果报错“获取资源服务失败，由于[CDM返回为空：[404 NOT FOUND]]”，请在管理中心参考**MRS Ranger数据连接参数说明**，排查Ranger数据连接的RangerAdmin业务IP和Ranger服务端口是否正确。

图 12-157 资源权限配置页面

步骤3 单击待创建权限策略Storm组件的“配置”，进入配置界面单击“创建”，新建权限策略。

图 12-158 新建 Storm 权限策略



步骤4 在弹出的策略配置页配置相关参数，配置完成单击“确定”，策略配置完成。

图 12-159 配置 Storm 权限策略

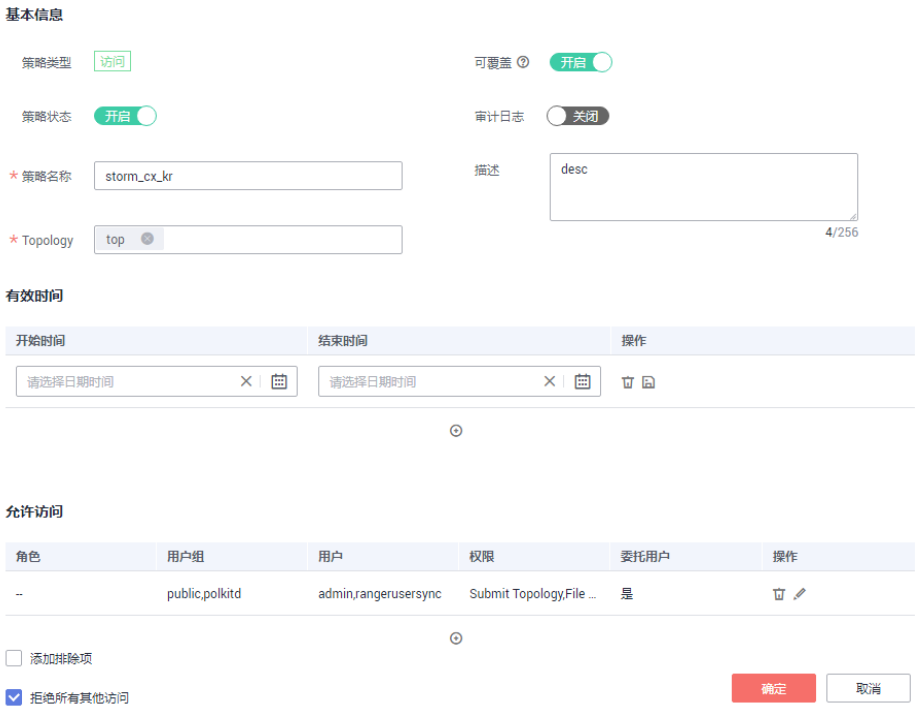


表 12-30 Storm 权限策略参数表

参数名	参数描述
策略类型	根据用户所选服务组件自动生成。包括访问、脱敏、行过滤器，其中脱敏和行过滤器类型是Hive特有的。
策略状态	开启表示权限策略生效，关闭表示权限策略创建成功后不生效。默认开启。
可覆盖	开启可覆盖时，新创建的策略将覆盖当前策略（新策略生效而旧策略不生效）。默认开启。 当用户需要创建一个临时访问策略时，“可覆盖”可以配合“有效时间”一起使用，那么即使临时访问策略超过有效期失效后，也不影响原有的权限策略继续生效。
审计日志	开启表示记录日志，日志内容包括客户端访问时间、客户端IP、客户端用户、操作资源结果等信息。
策略名称	名称为必填项，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，且输入不能为空。
描述	对策略的描述信息，长度限制在256个字符以内。
Topology	该参数表示Storm集群中的任务。
有效时间	用户通过设置开始时间和结束时间来控制策略的生效时间段，可配置多条。

参数名	参数描述
允许访问	定义允许访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义允许访问的用户拥有的权限。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 • 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。
添加排除项	允许访问勾选“添加排除项”意思是在允许访问的用户组里添加禁止访问的用户。 禁止访问勾选“添加排除项”意思是在禁止访问的用户组里添加允许访问的用户。
拒绝所有其他访问	勾选此项表示只有策略中“允许访问”指定的用户或用户组可以访问，其他用户均禁止访问。
禁止访问	不勾选“拒绝所有其他访问”时显示此配置，该配置定义禁止访问的用户和用户组。 • 用户：MRS服务的用户。 • 角色：MRS服务的角色。 • 用户组：MRS服务的用户组。 • 权限：定义用户禁止的权限类型。权限和用户允许同时为空值，或者同时不为空值。服务相关权限详情请参考**表12-22**。 • 委托用户：当勾选此项时，管理权限将分配给适用的用户和组。受委托的管理员可以更新和删除策略，还可以基于原始策略创建子策略。

----结束

12.3.7.2 查看权限报告

本章主要介绍如何查看资源配置权限策略及详情。

前提条件

已完成权限策略配置，未配置请参考**配置资源权限**。

查看策略及详情

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“权限报告”，进入权限报告页面。

步骤3 选择MRS集群（Ranger连接）> 服务查看该服务的策略及策略详情。

- 高级搜索功能：

  您在查看报告时，可以使用搜索操作，高级搜索提供了根据集群、策略名称、用户、用户组、策略类型、策略状态来搜索相关策略的功能。您只需单击权限报告页面右上角的“高级搜索”即可弹出搜索框。

  图 12-160 高级搜索

- 策略状态过滤：

  在服务的策略列表中，策略状态栏提供了过滤功能，您可以单击策略状态栏的 来过滤所需要查看的策略。

  图 12-161 策略状态过滤

----结束

12.4 敏感数据治理

12.4.1 敏感数据治理流程

敏感数据定义

敏感数据主要指未经个人或集团授权被他人使用，有可能给个人或集团带来严重损害的数据。

以《GBT 35273-2020 信息安全技术个人信息安全规范》为例，个人敏感数据有：

- 个人财产信息（存款、信贷、消费流水）
- 个人健康生理信息（体检信息、医疗记录）
- 个人生物识别信息（指纹、面部特征）
- 个人身份信息（身份证、社保卡、驾驶证）
- 其他信息（宗教信仰、精准定位）

敏感数据的保护方式

- **敏感数据识别与添加标签**

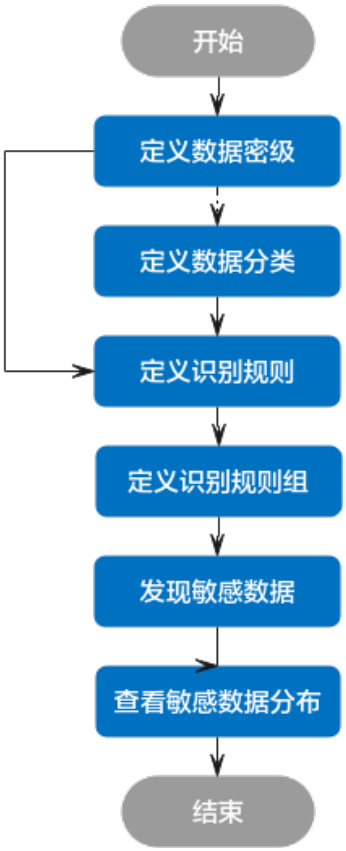
从海量数据中将数据进行分类分级，方便进行不同粒度和级别的安全管理。

- **数据泄露检测与防护**
  如果出现频繁访问敏感数据的异常行为，可以及时进行风险告警。
- **数据静态脱敏、数据水印**
  对于已标记特定安全级别的敏感数据，可在对外提供数据时进行脱敏或者加水印。
- **个人信息合规**
  精准区分和保护个人数据，避免产生合规问题。
- **满足GDPR要求**
  满足GDPR关于在海量数据中找到和保护敏感数据的要求，可对敏感数据的使用进行审计。
- **数据安全合规检查**
  通过对敏感数据的分析，制定数据安全合规管理制度，帮助企业建设以及改善信息安全合规管理体系。

敏感数据识别流程

在执行识别敏感数据任务之前，您可通过图12-162了解敏感数据识别流程。

图 12-162 敏感数据识别流程图



1. **定义数据密级**
   在对数据进行操作前，为数据定义密级，用以明确涉密的范围。

2. **定义数据分类**

   当数据密级已经无法满足大数据量下的数据分级分类诉求时，您可以进一步为不同价值的数据定义数据分类，以更好地管理和分组计量自己的数据。

3. **定义识别规则**

   定义敏感数据识别标准。

4. **定义识别规则组**

   通过定义敏感数据识别规则及规则组，来有效识别数据库内的敏感数据。

5. **敏感数据发现**

   创建并运行敏感数据识别任务。

6. **敏感数据分布**

   查看敏感数据识别任务识别出的敏感数据。

12.4.2 定义数据密级

为了方便对数据进行管理，在对数据进行操作前，需要您为数据定义密级，并对保密等级做相应的描述，例如明确涉密的范围。本章主要介绍如何定义数据密级并配置默认密级。

值得注意的是，数据密级、数据分类和识别规则，均为DataArts Studio实例级别配置，各工作空间之间数据互通。这样在数据地图组件中，就可以根据一套标准的分级分类管理对数据进行统一管理。

前提条件

配置默认密级前，请参考**创建密级**至少创建1个密级。

约束与限制

- 根据行业内的通用定义密级，约定密级数字越大表示保密等级越高。当前最多创建10层密级。
- 仅DAYU Administrator、Tenant Administrator或者数据安全管理员可以创建、修改或删除数据密级、分类和识别规则，其他普通用户无权限操作。
- 配置默认密级后，MRS Hive和DWS数据源中所有未被标记密级的数据表和字段（包括存量和增量数据）将被标记为默认密级，默认密级支持在数据地图组件中进行展示，并支持通过**管控敏感数据**进行数据预览时的权限管控。

  说明

  权限申请时的密级信息来源于数据地图组件，因此也会展示默认密级。除此之外的静态脱敏、动态脱敏时的密级信息来源于敏感数据发现任务，因此不会展示默认密级。

- 被引用的数据密级无法直接删除，需要先解除引用关系后才能删除。

创建密级

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“数据密级”，进入数据密级页面。

图 12-163 进入数据密级

步骤3 单击“新建”，参考表12-31输入数据密级信息。

图 12-164 新建数据密级

表 12-31 参数设置

参数名	参数设置
*密级名称	密级名称只能包含中文、英文字母、数字和下划线，创建完成后不支持“编辑”操作。
密级描述	密级描述支持所有字符输入，创建完成后支持通过“编辑”操作修改。

说明

新建密级时，系统默认按照安全程度由低到高的顺序依次创建。您可以在密级建立好后，按照安全程度高低，通过“上移”、“下移”操作来调整密级顺序。

----结束

配置默认密级

如果您需要统一为MRS Hive和DWS数据源中未被标记密级的资产标记密级，则您可以配置默认密级。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“数据密级”，进入数据密级页面。

图 12-165 进入数据密级

步骤3 单击密级列表右上方“默认密级”，在选择框中选择一个密级作为默认密级。

配置默认密级后，MRS Hive和DWS数据源中所有未被标记密级的数据表和字段（包括存量和增量数据）将被标记为默认密级，默认密级支持在数据地图组件中进行展示，并支持通过**管控敏感数据**进行数据预览时的权限管控。

说明

权限申请时的密级信息来源于数据地图组件，因此也会展示默认密级。除此之外的静态脱敏、动态脱敏时的密级信息来源于敏感数据发现任务，因此不会展示默认密级。

图 12-166 新建数据密级

----结束

相关操作

- 调整密级：在数据密级页面，单击对应密级操作栏中的“更多 > 上移”或“更多 > 下移”，即可调整该密级级别。
- 编辑密级：在数据密级页面，单击对应密级操作栏中的“编辑”，即可修改密级描述。
- 删除密级：在数据密级页面，单击对应密级操作栏中的“删除”，即可删除密级。当需要批量删除时，可以在勾选密级后，在列表上方单击“批量删除”。

  说明

  - 被引用的数据密级无法直接删除，需要先解除引用关系后才能删除。
  - 删除操作无法撤销，请谨慎操作。

12.4.3 定义数据分类

当数据密级已经无法满足大数据量下的数据分级分类诉求时，您可以进一步为不同价值的数据定义数据分类，以更好地管理和分组计量自己的数据，让各类各组之间属于并列、平等并且互相排斥的关系，使数据更清晰。本章主要介绍如何定义数据分类。

值得注意的是，数据密级、数据分类和识别规则，均为DataArts Studio实例级别配置，各工作空间之间数据互通。这样在数据地图组件中，就可以根据一套标准的分级分类管理对数据进行统一管理。

前提条件

导入预置数据分类前，请参考**定义数据密级**至少创建1个密级。

约束与限制

- 当前数据分类的最大层级数默认为5层，最大配额1000个。
- 仅DAYU Administrator、Tenant Administrator或者数据安全管理员可以创建、修改或删除数据密级、分类和识别规则，其他普通用户无权限操作。

- 当前支持在不同的父节点下创建同名的分类，但同一父节点下不能创建同名的分类。
- 导入预置数据分类时，需要先为所有的预置规则配置数据密级，才能导入预置数据分类。
- 导入预置数据分类时，会直接导入分类和对应的识别规则，与当前分类和规则同名的部分无法导入。
- 当父类下有子分类的时候，无法直接删除该父分类，需要先删除子分类。
- 被引用的数据分类无法直接删除，需要先解除引用关系后才能删除。

创建分类

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“数据分类”，进入数据分类页面。

图 12-167 进入数据分类

步骤3 首次新建分类时，需要通过分类目录上方的，至少新增一个根目录层级分类。后续再新建分类时，可通过或，新增同级或子级分类。

单击或后，在弹出的新建分类窗口中，参考**表12-32**填写数据分类信息。

图 12-168 新建数据分类

表 12-32 参数设置

参数名	参数设置
*分类名称	分类名称只能包含中文、英文字母、数字和下划线。
描述	分类描述支持所有字符输入。

----结束

导入预置分类

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击数据安全左侧导航树中的“数据分类”，进入数据分类页面。

图 12-169 进入数据分类

步骤3 如果还没有新建的分类，可以单击“导入预置数据分类”，进入导入窗口。已有新建分类时，可通过单击，进入导入窗口。

在弹出的导入预置数据分类窗口中，勾选需要导入的数据分类，为待导入规则逐一配置数据密级或批量设置密级后，单击“确定”完成预置数据分类和规则的导入。

图 12-170 导入预置数据分类

----结束

相关操作

- 编辑分类：在数据分类页面，先选择分类目录中需要修改的目录，然后单击分类目录上方的，即可修改分类名称和描述。
- 删除分类：在数据分类页面，先选择分类目录中需要删除的目录，然后单击分类目录上方的，即可删除分类。
  另外，也支持通过编辑数据分类目录的方式删除分类。您可以单击分类目录上方的，在“编辑数据分类目录”页面删除分类。

  说明

  - 当父类下有子分类的时候，无法直接删除该父分类，需要先删除子分类。
  - 被引用的数据分类无法直接删除，需要先解除引用关系后才能删除。
  - 删除操作无法撤销，请谨慎操作。
- 编辑数据分类目录：当需要整体编辑目录时，可以单击分类目录上方的，进入“编辑数据分类目录”页面。在“编辑数据分类目录”页面，支持新增子级分类，或删除分类。

  说明

  删除操作无法撤销，请谨慎操作。

12.4.4 定义识别规则（部分高级特性）

您可以通过定义敏感数据识别规则，来有效识别数据库内的敏感数据字段。当前识别规则支持使用内置规则和简单的正则表达式。

如果您对需要更强大的识别规则，数据还支持您使用组合规则。组合规则的多个子规则间可进行与或非逻辑判断，单个子规则支持Groovy脚本、正则表达式、等于、长度判断、内置规则等算法，匹配对象除了列内容识别外还支持列名、列注释、表名、表注释、数据库名等，能够满足您的各类识别需求。

说明

在新版本模式下仅当使用企业版时，才支持配置组合规则。旧版本模式使用基础版及更高版本时即可支持。

值得注意的是，数据密级、数据分类和识别规则，均为DataArts Studio实例级别配置，各工作空间之间数据互通。这样在数据地图组件中，就可以根据一套标准的分级分类管理对数据进行统一管理。

说明

识别规则定义后，默认为待确认状态，无法在静态脱敏任务中生效。需经如下操作后变更状态后，才能使识别规则状态生效：

敏感数据发现任务运行后，为使该识别规则在静态脱敏任务中生效，必须在“敏感数据分布>手工修正”页面对任务中的识别规则进行“确认”，使规则状态变更为“有效”。

前提条件

- （必须）数据密级定义已完成，请参见**定义数据密级**。
- （可选）数据分类定义已完成，请参见**定义数据分类**。

约束与限制

- 仅DAYU Administrator、Tenant Administrator或者数据安全管理员可以创建、修改或删除数据密级、分类和识别规则，其他普通用户无权限操作。
- 敏感数据识别过程中，如果规则为内容识别类型（即内置规则和内容识别类型的自定义规则），则仅当数据表中某字段匹配规则的记录数/总记录数>=指定阈值（默认80%）时，才认为该字段为敏感字段，并为之匹配相应密级和分类。
- 被引用的数据识别规则无法直接删除，需要先解除引用关系后才能删除。

创建数据识别规则

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 在数据安全控制台左侧的导航树中单击“数据识别规则”，进入数据识别规则页面。

步骤3 在“识别规则”页面单击“新建”，创建识别规则。

图 12-171 新建识别规则

步骤4 新建规则参数配置请参考**表12-33**，参数配置完成单击“确定”即可。

图 12-172 规则配置

表 12-33 配置识别规则参数说明

配置	说明
*规则类型	即规则所属分类，支持按模板添加内置规则和自定义规则。
*数据密级	对配置的数据进行等级划分。如果现有的分级不满足需求，请进入数据密级页面进行设置，详情请参见**定义数据密级**。
数据分类	对配置的数据进行分类划分。如果现有的分类不满足需求，请进入数据分类页面进行设置，详情请参见**定义数据分类**。
规则描述	对当前规则进行简单描述。
内置	

配置	说明
*规则模板	规则类型选择“内置”，呈现此参数。 系统内置了80+条敏感数据识别规则，可对个人敏感信息（银行卡、信用卡等）、个人基本资料（手机号码、电子邮箱等）、网络身份标识信息（IPv4地址、IPv6地址等）等敏感信息进行识别和脱敏。内置的敏感数据识别规则可在“内置规则模板”页签查看。 选择内置规则后，可输入测试数据，测试能否通过内置规则识别。
*规则名称	规则类型选择“内置”，规则名称自动关联分类模板生成。
自定义	
*规则名称	规则类型选择“自定义”，您可以自行填写分类名称，名称为必填项。建议包含规则含义，避免无意义的描述，以便于使用中能快速选择需要的规则。 **说明** 定义数据识别规则，名称必须唯一。
*识别规则	规则类型选择“自定义”，呈现此参数，支持正则表达式。 当选择“无”，表示关联了该规则的敏感数据发现任务不生效。无法自动为数据资产分类，需要您手动添加分类。
*正则表达式	识别规则选择“正则表达式”时，呈现此参数。 ● 内容识别：勾选此项后输入自定义正则表达式，该表达式将用于数据内容识别。内容识别正则表达式举例：“^男$|^女&”。 ● 列名识别：勾选此项后输入自定义正则表达式，该表达式将用于字段名精确匹配和模糊匹配两种方式，当前支持多个字段匹配。列名识别正则表达式举例：“age|years”。 ● 备注识别：勾选此项后输入自定义正则表达式，例如“.*comment.*”代表模糊匹配备注。

----结束

创建组合规则

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 在数据安全控制台左侧的导航树中单击“数据识别规则”，进入数据识别规则页面。

步骤3 在“识别规则”页面，单击“新建组合规则”，创建组合规则。

图 12-173 新建组合规则

步骤4 新建组合规则参数配置请参考**表12-33**，参数配置完成单击“确定”即可。

图 12-174 组合规则配置

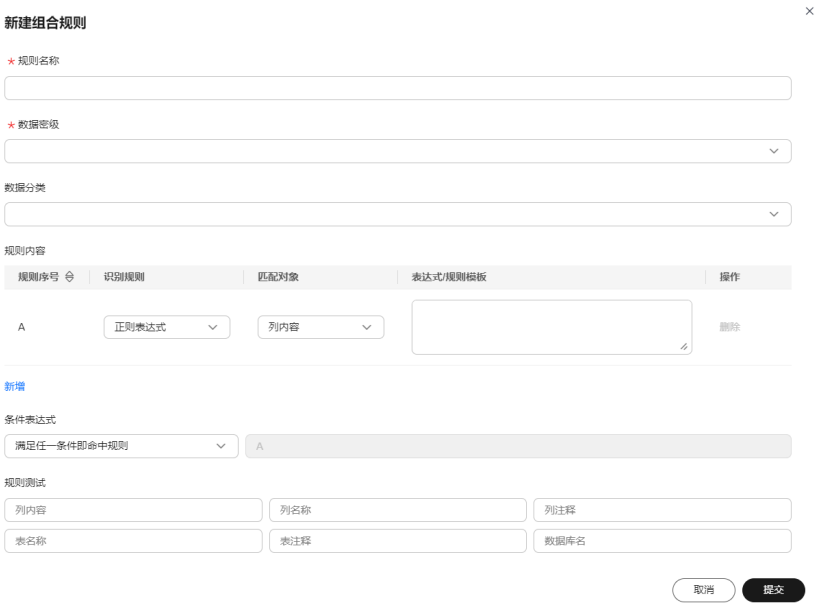


表 12-34 配置组合规则参数说明

配置	说明
*规则名称	您可以自行填写分类名称，名称为必填项。建议包含规则含义，避免无意义的描述，以便于使用中能快速选择需要的规则。 **说明** 定义数据识别规则，名称必须唯一。
*数据密级	对配置的数据进行等级划分。如果现有的分级不满足需求，请进入数据密级页面进行设置，详情请参见**定义数据密级**。
数据分类	对配置的数据进行分类划分。如果现有的分类不满足需求，请进入数据分类页面进行设置，详情请参见**定义数据分类**。

配置	说明
规则内容	定义组合规则中的一条子规则。 ● 规则序号：标识当前子规则，并在条件表达式中表示该子规则。 ● 识别规则：规则内容的类型，支持：正则表达式、GRROVY脚本、正则表达式（忽略大小写）、等于、长度等于、长度大于、长度小于、内置等类型。 正则表达式举例：“^男$|^女&”。 ● 匹配对象：规则识别的数据对象。包含表的列内容、列名称、列注释、表名称、表注释、数据库名等。 ● 表达式/规则模板：按照所选的识别规则填写规则表达式，该表达式将用于匹配对象的识别。 ● 操作：可删除此条子规则内容，或再新建一条子规则
*条件表达式	多个子规则间可进行与或非的逻辑判断。 ● 自定义：输入自定义正则表达式，用于对多个子规则进行与或非的逻辑判断。子规则用规则序号A-Z之间表示，逻辑运算符支持&&，||，!，(，)。 表达式举例：“A&&B”。 ● 满足所有条件即命中规则：勾选此项后，自动生成表达同时满足所有规则内容的逻辑表达式。 ● 满足任一条件即命中规则：勾选此项后，自动生成表达只需满足一条规则内容的逻辑表达式。
规则测试	通过输入测试数据，判断所写规则是否符合预期。
规则描述	对当前规则进行简单描述。

----结束

相关操作

- 编辑识别规则：在识别规则页面，单击对应识别规则操作栏中的“编辑”，即可修改识别规则关联的密级、分类和描述。如果为自定义规则，还支持修改识别规则和正则表达式。
- 编辑识别规则状态：新增的识别规则默认为启用状态。当识别规则为关闭状态时，表示该规则将不可被添加到识别规则组。

  需要修改识别规则状态时，在识别规则页面单击对应识别规则中的  或  ，即可启用或关闭对应规则。
- 删除识别规则：在识别规则页面，单击对应识别规则操作栏中的“删除”，即可删除识别规则。当需要批量删除时，可以在勾选识别规则后，在列表上方单击“批量删除”。

  说明

  - 被引用的数据识别规则无法直接删除，需要先解除引用关系后才能删除。
  - 删除操作无法撤销，请谨慎操作。

- 测试内置规则模板：在“内置规则模板”页签可查看所有内置规则模板，并且根据输入的自定义样例数据，测试验证内置规则模板的识别结果。

12.4.5 定义识别规则分组

定义敏感数据识别规则组，可以将多个零散的规则组合成为一个有业务逻辑的规则组，该操作是用户后续进行敏感数据发现任务操作的前提 。

前提条件

识别规则创建完成，请参考**定义识别规则（部分高级特性）**。

约束与限制

- 敏感数据识别过程中，当某个字段同时匹配到识别规则组中的多个识别规则时，此字段密级取多个识别规则的最高密级，字段分类允许有多个。
- 数据识别规则分组最多配置100条。
- 被引用的数据识别规则分组无法直接删除，需要先解除引用关系后才能删除。

创建数据识别规则组

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

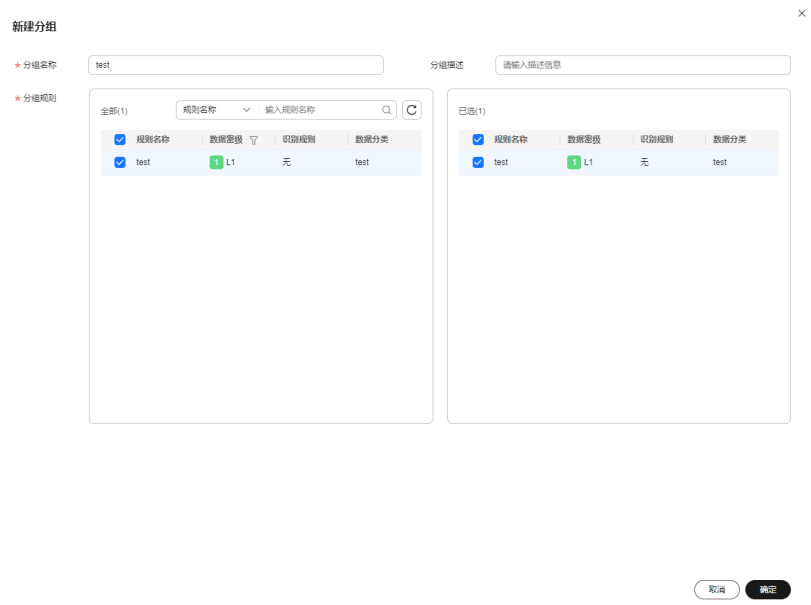
步骤2 单击左侧导航树中的“数据识别规则”，进入数据识别规则页面。

步骤3 单击页面上方“规则分组”页签，进入分组列表页面。

图 12-175 新建数据识别规则组

步骤4 单击“新建”，输入新建分组名称和描述，参数设置参考**表12-35**，并勾选左侧列表中的识别规则。配置完成后单击“确定”即可。

图 12-176 新建分组参数配置



您所勾选的规则将显示在右侧列表中，右侧已选列表中，已选规则可以通过单击操作来取消勾选。

表 12-35 参数配置表

配置	说明
*分组名称	规则组名称只能包含中文、英文字母、数字和下划线。 建议包含规则含义，避免无意义的描述，以便于使用中能快速选择需要的规则组。
分组描述	为更好地识别规则组，此处加以描述信息。

----结束

相关操作

- 编辑规则分组：在规则分组页面，单击对应规则分组操作栏中的“编辑”，即可修改规则分组的名称、描述和关联的识别规则。
- 删除规则分组：在规则分组页面，单击对应规则分组操作栏中的“删除”，即可删除识别规则。当需要批量删除时，可以在勾选规则分组后，在列表上方单击“批量删除”。

  说明

  - 被引用的数据识别规则分组无法直接删除，需要先解除引用关系后才能删除。
  - 删除操作无法撤销，请谨慎操作。

12.4.6 配置数据入湖检测规则（高级特性）

数据入湖检测规则可用于如下场景的实时敏感信息检测：

- 数据集成（离线作业）进行表数据迁移时的“敏感数据检测”，详见**配置离线处理集成作业**。
- 数据开发导入数据文件时的敏感数据自动实时检测。

说明

在新版本模式下仅当使用企业版时，才支持配置数据入湖检测规则。旧版本模式使用基础版及更高版本时即可支持。

前提条件

- 识别规则创建完成，请参考**定义识别规则（部分高级特性）**。

约束与限制

- 数据集成（离线作业）的表数据迁移和数据安全敏感数据发现中的推荐识别场景下，仅支持MRS Hive、DWS、DLI和RDS MySQL数据源。
- 数据集成（离线作业）的表敏感数据检测对表内容格式要求如下：
  - 表字段数量至多为500。
  - 对于字符串类型的表字段，仅会检测前1000个字符的敏感信息，超过1000字符部分会被截断。
- 识别规则仅支持规则类型为正则表达式的内置规则或自定义规则，识别规则至多可以配置50条。

配置识别规则

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

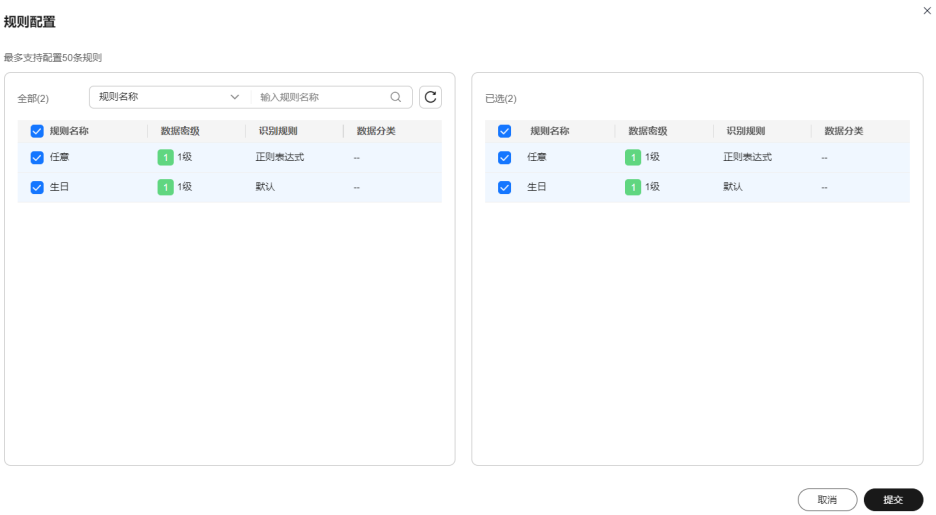
步骤2 单击左侧导航树中的“数据识别规则”，在数据识别规则页面中单击“数据入湖检测规则”，进入数据入湖检测规则页签。

步骤3 在“数据入湖检测规则”页面中，单击配置规则。

图 12-177 数据入湖检测规则页面

步骤4 在弹出的规则配置窗口中，选择所需的识别规则，单击“提交”完成规则配置。

图 12-178 配置规则



步骤5 （可选）如需将数据入湖检测规则应用到数据集成（离线作业）的表敏感数据实时检测中，则还需配置规则策略，各配置参数说明请参见**表12-36**。

图 12-179 规则策略配置参数

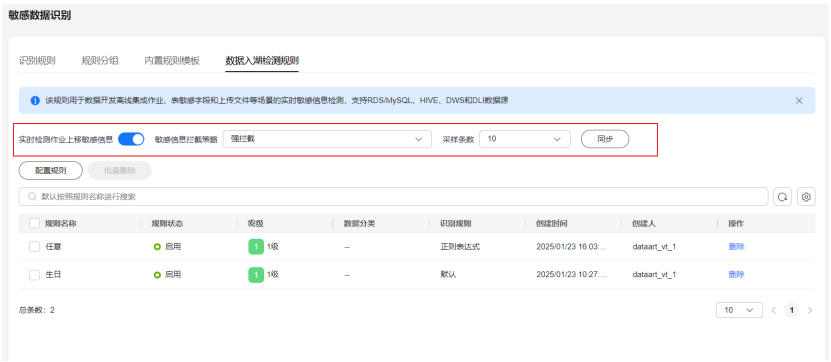


表 12-36 规则策略配置参数

参数名	参数说明
实时检测作业上移敏感信息	是否开启在数据集成（离线作业）进行表数据迁移时的表敏感数据实时检测。
敏感信息拦截策略	在数据集成（离线作业）作业中，如果识别到了敏感信息的拦截策略： ● 强拦截：只要检测到敏感信息，就无法保存作业。 ● 弱拦截：对配置了加解密或脱敏处理的敏感字段不做拦截。 ● 不拦截：无处理策略，不做任何拦截。
采样条数	在数据集成（离线作业）作业中，对表字段进行检测时采样的行数，至多为100行。

参数名	参数说明
同步	单击同步按钮，将策略同步至数据集成（离线作业）。此处的规则策略需要同步至数据集成（离线作业）中才会生效。

----结束

相关操作

- 删除数据入湖检测规则：在数据入湖检测规则页面，单击对应识别规则操作栏中的“删除”，即可删除识别规则。当需要批量删除时，可以在勾选识别规则后，在列表上方单击“批量删除”。

  说明

  删除操作无法撤销，请谨慎操作。

12.4.7 发现敏感数据

完成了敏感数据识别规则组定义后，就可以根据定义的规则来创建敏感数据识别任务，发现敏感数据，并将敏感数据同步到数据地图组件。

说明

敏感数据发现任务运行后，为使该识别规则在静态脱敏任务中生效，必须在“敏感数据分布>手工修正”页面对任务中的识别规则进行“确认”，使规则状态变更为“有效”。

前提条件

- 已完成敏感数据规则组定义，请参考**定义识别规则分组**。
- 已在管理中心创建数据仓库服务（DWS）、数据湖探索（DLI）、MapReduce服务（MRS Hive）类型的数据连接，请参考**创建DataArts Studio数据连接**。
- DLI敏感数据发现时，需要提前准备DLI通用队列，当前暂不支持Spark版本为3.3.1的通用队列。
- 如需将识别的敏感数据自动同步到数据地图组件，则必须由DAYU Administrator、Tenant Administrator或者数据安全管理员用户创建、运行或调度任务。
- 敏感数据同步到数据地图组件时，如需将敏感数据的分类同步成功，需要同时满足如下前提：
  - 已在数据目录组件，对数据表成功进行过元数据采集，详见**元数据采集任务**。
  - 管理中心组件对应的数据连接，已开启“元数据实时同步”功能，详见**创建DataArts Studio数据连接**。

约束与限制

- 当前仅支持对数据仓库服务（DWS）、数据湖探索（DLI）、MapReduce服务（MRS Hive）类型的数据源进行敏感数据识别，且仅支持标准数仓类型的DWS数据源。
- DLI敏感数据发现任务暂不支持Spark版本为3.3.1的通用队列。

- 当前仅DLI和DWS类型的敏感数据发现任务支持按照通配符匹配数据表或全部数据表进行敏感数据识别，仅DLI类型的敏感数据发现任务支持配置资源规格（如果配置资源大于可用资源，任务可能失败）。
- 仅DWS敏感数据发现任务支持断点续扫和日志展示任务进度。
- 敏感数据识别过程中，如果规则为内容识别类型（即内置规则和内容识别类型的自定义规则），则仅当数据表中某字段匹配规则的记录数/总记录数>=指定阈值（默认80%）时，才认为该字段为敏感字段，并为之匹配相应密级和分类。
- 敏感数据识别过程中，当某个字段同时匹配到识别规则组中的多个识别规则时，此字段密级取多个识别规则的最高密级，字段分类允许有多个。
- 敏感数据识别任务运行后，会为识别到的敏感字段生成相应密级和分类，默认不会生成数据表密级。在手动勾选任务中的“根据数据识别结果更新数据目录/数据地图中数据表密级”选项后，才会生成数据表密级，数据表密级取敏感字段的最高密级。
- 当前敏感数据同步仅支持同步到数据地图组件。不支持将识别到的敏感数据同步到数据目录组件，且数据目录组件也不再支持手动新增、编辑敏感数据的密级和分类信息。
- 敏感数据同步的权限要求较高，仅DAYU Administrator、Tenant Administrator用户或者数据安全管理员有权限将敏感数据通过自动或手动方式同步到数据地图组件。
  - 自动同步：创建敏感数据发现任务，默认不勾选任务中的“手动同步数据识别结果”参数时，会自动同步敏感数据到数据地图组件。
  - 手动同步：创建敏感数据发现任务，勾选任务中的“手动同步数据识别结果”参数时，表示取消敏感数据自动同步。待任务运行成功后，需要手动在“敏感数据分布>手工修正”页面单击“数据同步”将敏感数据同步到数据地图组件中。

  因此，非DAYU Administrator、Tenant Administrator或者数据安全管理员普通用户创建敏感数据发现任务时，必须勾选任务中的“手动同步数据识别结果”参数，才能创建成功。另外，当普通用户运行或调度未勾选“手动同步数据识别结果”参数的任务时，也会运行失败。

创建敏感数据发现任务

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“敏感数据发现”，进入敏感数据发现页面。

图 12-180 进入敏感数据发现页面

步骤3 单击“新建”，在弹出的窗口中新建发现任务页面，输入基本信息，参数配置参考**表 12-37**。

图 12-181 新建发现任务参数配置

创建敏感数据发现任务参数配置说明：

表 12-37 配置任务参数

参数名	参数说明
基本信息配置	
*任务名称	标识敏感数据发现任务，为便于任务管理，建议名称中包含要识别的数据表和使用的规则组。
任务描述	为更好地识别敏感数据发现任务，此处加以描述信息。
*数据源类型	从下拉列表中选择已创建的数据源类型。
*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。 若未创建请参考**创建DataArts Studio数据连接**新建连接。
*数据库	呈现待扫描的数据库。单击数据库后的“设置”，设置待扫描的数据库范围。单击“清除”，可对已选择的数据库进行修改。
*数据表	● 对于DLI和DWS类型的敏感数据发现任务，您需要设置选择表的方式，当前支持手动筛选、通配符匹配和全部三种方式。 – 手动筛选：即手动在数据表列表中选择需要进行敏感发现任务的表。手动筛选时，在表筛选窗口的搜索框中可以进行模糊匹配，如果需要全选表时仅支持分页全选。 手动筛选适用于需要敏感数据发现的目标表较少的情况。 – 通配符匹配：即通过输入匹配规则，按照通配符匹配目标表。单任务中匹配规则支持配置最多100条，以换行符分隔，每一行视作一条规则，规则中只能包含字母、数字、下划线（_）和通配符（*），例如 匹配规则为test_*时，表示匹配以“test_”开头的表。您也可以通过测试窗口，验证匹配规则是否符合预期。 通配符匹配适用于规则较多、结果表较多的情况。 – 全部：无需筛选或输入规则，直接选择当前数据库下的所有表作为任务目标表。 选择全部，适用于所选数据库下所有表的检索。 ● 对于MRS Hive类型的敏感数据发现任务，仅支持通过手动筛选方式选择目标表。手动筛选时，在表筛选窗口的搜索框中可以进行模糊匹配，如果需要全选表时仅支持分页全选。
采样条数	DWS类型的任务支持配置目标表的采样条数，最大支持10000条。
*计算队列	数据源类型为DLI时，需要选择通用队列。该参数表示执行DLI作业时的通用队列。 **说明** 暂不支持Spark版本为3.3.1的通用队列。
规则配置	

参数名	参数说明
*识别规则组	从下拉列表中选择数据识别规则组，若未定义请参考**定义识别规则分组**新建。 选择识别规则组后，会展示组内的识别规则详情，内置规则以及包含内容匹配的自定义规则支持配置规则阈值。阈值表示仅当数据表中某字段匹配规则的记录数/总记录数>=指定阈值（默认80%）时，才认为该字段为敏感字段。需要注意的是，不同规则组包含同一规则时，则需要该规则识别阈值相同。
手动同步数据识别结果	敏感数据同步的权限要求较高，仅DAYU Administrator、Tenant Administrator用户或者数据安全管理员有权限将敏感数据通过自动或手动方式同步到数据地图组件。 ● 自动同步：创建敏感数据发现任务，默认不勾选任务中的“手动同步数据识别结果”参数时，会自动同步敏感数据到数据地图组件。 ● 手动同步：创建敏感数据发现任务，勾选任务中的“手动同步数据识别结果”参数时，表示取消敏感数据自动同步。待任务运行成功后，需要手动在“敏感数据分布>手工修正”页面单击“数据同步”将敏感数据同步到数据地图组件中。 因此，非DAYU Administrator、Tenant Administrator或者数据安全管理员普通用户创建敏感数据发现任务时，必须勾选任务中的“手动同步数据识别结果”参数，才能创建成功。另外，当普通用户运行或调度未勾选“手动同步数据识别结果”参数的任务时，也会运行失败。
调度信息配置	
单次调度	选择单次调度时，敏感数据发现任务仅运行一次。
周期调度	选择周期调度时，敏感数据发现任务按照所选调度周期运行。 ● 调度日期： 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：选择调度开始时间和结束时间，配置间隔的分钟时长。 – 小时：选择调度开始时间和结束时间，配置间隔的小时时长。 – 天：配置每日调度时间。 – 周：选择星期几启动调度，配置调度具体时间。 – 月：选择几号启动调度，配置调度具体时间。 例如：选择调度周期是周，选择具体时间为15:52，时间选择为星期二。则在调度日期范围内，每周二的15点52分会执行任务。 ● 立即启动：勾选复选框，则表示立即启动此调度任务。
计算资源规格	

参数名	参数说明
资源规格	在DLI Spark资源较为充足的情况下，您可以通过配置Spark任务资源，加快敏感数据发现任务的执行速度。 系统提供3种默认资源规格供您选择，默认A第一种，您也可以自行调整。 **说明** 如果申请资源大于可用资源，任务可能会失败！ ● A（8核32G内存；Executor内存：4G，Executors个数：6个，Executor CPU数：1个，Driver CPU数：2个，Driver内存：7G） ● B（16核64G内存；Executor内存：8G，Executors个数：7个，Executor CPU数：2个，Driver CPU数：2个，Driver内存：7G） ● C（32核128G内存；Executor内存：8G，Executors个数：14个，Executor CPU数：2个，Driver CPU数：4个，Driver内存：15G） **说明** Spark资源并行度由Executor数量和Executor CPU核数共同决定。任务可并行执行的最大Task数量=Executor个数 * Executor CPU核数。您可以根据DLI队列资源合理规划计算资源规格。 需要注意的是，Spark任务执行需要driver、executor等多个角色共同调度完成，因此“Executor个数*Executor CPU核数”要小于队列的计算资源CU数，避免其他Spark任务角色无法启动。 Spark作业参数计算公式： ● CU数=driver CPU核数+Executor个数*Executor CPU核数 ● 内存数=driver内存+(Executor个数*Executor内存）
Executor内存	代表每个Executor的内存。通常建议Executor CPU核数：Executor内存=1：4。 GB输入值必须在0到16之间，MB输入值必须在0到16,384之间。注意，如申请资源大于可用资源，任务可能失败。
Executor CPU核数	用于设置作业申请的每个Executor的CPU核数，决定每个Executor并行执行Task的能力。 输入值必须在0到4之间。注意，如申请资源大于可用资源，任务可能失败。
Executor个数	用于设置作业申请的Executor的数量。输入值必须在0到100之间。注意，如申请资源大于可用资源，任务可能失败。
driver CPU核数	用于设置driver CPU核数。输入值必须在0到4之间。注意，如申请资源大于可用资源，任务可能失败。
driver内存	用于设置driver内存大小，通常建议即driver CPU核数：driver内存=1：4。GB输入值必须在0到16之间，MB输入值必须在0到16384之间。注意，如申请资源大于可用资源，任务可能失败。

步骤4 单击“确定”，完成创建敏感数据发现任务。

说明

如果敏感数据发现任务执行成功后，界面不显示执行结果，并且在查看运行日志时发现无匹配信息，这种情况下说明执行该任务时没有发现任何敏感数据。

----结束

相关操作

- 运行或调度任务：在敏感数据发现页面，单击对应任务操作栏中的“运行”或“更多 > 启动调度”，运行或调度任务。

  您可以通过调度周期区分该任务是单次调度还是周期调度任务。

  说明

  非DAYU Administrator、Tenant Administrator或者数据安全管理员的普通用户运行或调度未勾选“手动同步数据识别结果”参数的任务时，会运行失败。只有DAYU Administrator、Tenant Administrator或者数据安全管理员才能运行或调度未勾选“手动同步数据识别结果”参数的任务。

- 编辑任务：在敏感数据发现页面，单击对应任务操作栏中的“编辑”，即可编辑敏感数据发现任务。

  运行状态为正在“运行中”的任务不允许被编辑。

- 删除任务：在敏感数据发现页面，单击对应任务操作栏中的“更多 > 删除”，即可删除任务。当需要批量删除时，可以在勾选任务后，在任务列表上方单击“批量删除”。

  运行状态为正在“运行中”的任务不允许被删除。

  说明

  - 删除敏感数据发现任务会删除对应任务的识别结果，请谨慎操作。
  - 删除操作无法撤销，请谨慎操作。

- 查看运行实例日志：在敏感数据发现页面，找到需要查看实例的任务，单击 展开，即可找到运行实例。随后单击“操作 > 查看日志”，查看运行实例日志。

  运行失败可通过日志排查失败原因，问题修正后尝试重新运行。如果仍运行失败，请联系技术支持人员协助处理。

12.4.8 查看敏感数据分布

本章主要介绍如何查看敏感数据发现结果以及手工修正。

- 查看敏感数据发现结果：敏感数据识别任务完成后，需要查看任务的运行结果。
- 手工修正：发现敏感数据后，您必须根据具体情况进行手工修正，通过对任务中的识别规则进行“确认”，使规则状态变更为“有效”，才能使该识别规则在静态脱敏任务中生效。

  如果在敏感数据发现任务中勾选了“手动同步数据识别结果”，则还需要手动单击“数据同步”，才能将识别到的敏感数据同步到数据地图组件（同步数据前需确保已在数据目录中完成元数据采集任务，否则会同步失败）。

前提条件

- 完成敏感数据识别任务的创建和运行，如何创建和运行敏感数据识别任务请参见**创建敏感数据发现任务**。

- 敏感数据同步的权限要求较高，仅DAYU Administrator、Tenant Administrator用户或者数据安全管理员有权限将敏感同步到数据地图组件。
- 敏感数据同步前，需要在数据目录组件对数据连接成功进行过元数据采集，详见**元数据采集任务**。否则会导致同步失败，报错“数据连接不存在”。

约束与限制

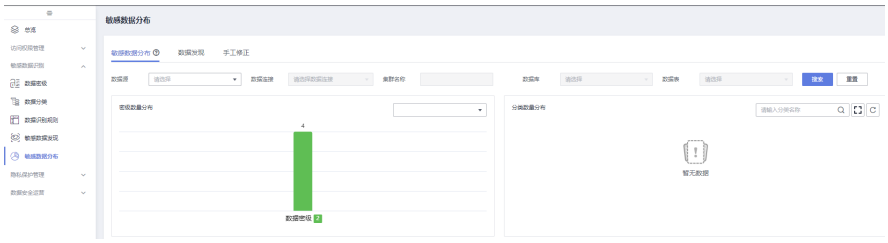
- 当前敏感数据同步仅支持同步到数据地图组件。不支持将识别到的敏感数据同步到数据目录组件，且数据目录组件也不再支持手动新增、编辑敏感数据的密级和分类信息。
- 敏感数据同步依赖于元数据采集任务。如果未对数据连接进行元数据采集，则无法找到数据连接。

发现敏感数据并手工修正

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“敏感数据分布”，进入敏感数据分布页面。

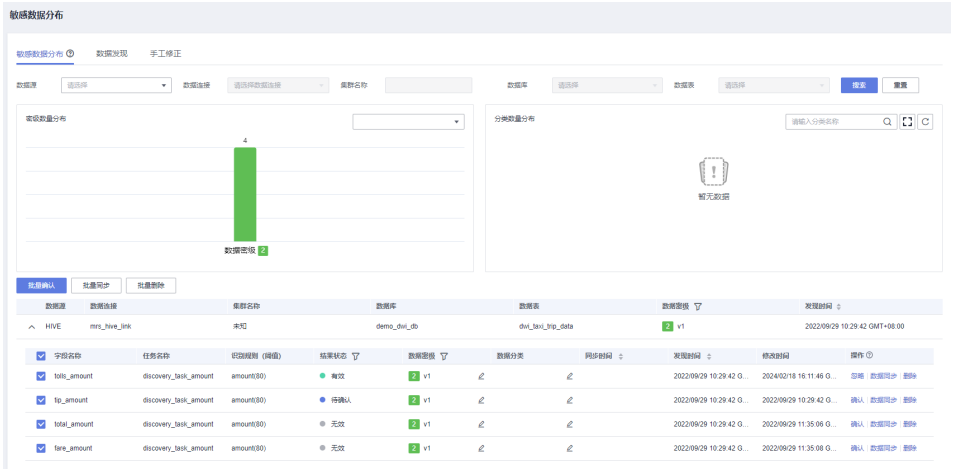
图 12-182 进入敏感数据分布页面



步骤3 在敏感数据分布页面，您可以通过如下两种方式之一来查看敏感数据发现结果并进行手工修正。推荐您使用方式1，相比于方式2，方式1支持修改数据密级、分类，无需切换页面即可完成敏感数据查看与修正，并支持批量操作。

- （推荐）方式1：在“敏感数据分布”页签，单击 展开数据源详情，查看敏感数据情况，并手工修正数据密级、分类以及数据状态。
  - 确认：确认该条识别结果为有效状态，“未确认”或“无效”状态的规则可以进行确认操作。静态脱敏任务可以基于有效状态的识别规则进行脱敏。
  - 忽略：确认该条识别结果为无效状态，“有效”状态的规则可以进行忽略操作。静态脱敏任务无法选择到未确认/无效状态的识别规则进行脱敏。
  - 数据同步：如果在敏感数据发现任务中勾选了“手动同步数据识别结果”，则还需要手动单击“数据同步”，才能将识别到的敏感数据同步到数据地图组件（同步数据前需确保已在数据目录中完成元数据采集任务，否则会同步失败）。
  - 删除：删除当前发现的字段结果。

图 12-183 查看敏感数据分布并手工修正



- 方式2：选择“数据发现”页签。然后通过搜索数据连接名称，找到待查看的敏感数据。最终单击“明细”查看敏感数据明细内容。

图 12-184 数据发现

图 12-185 查看明细内容

然后切换到“手工修正”页签，查找待修正的规则名称，单击“确认”、“忽略”或“数据同步”，手工修正数据状态。

- 确认：确认该条识别结果为有效状态，“未确认”或“无效”状态的规则可以进行确认操作。静态脱敏任务可以基于有效状态的识别规则进行脱敏。
- 忽略：确认该条识别结果为无效状态，“有效”状态的规则可以进行忽略操作。静态脱敏任务无法选择到未确认/无效状态的识别规则进行脱敏。
- 数据同步：如果在敏感数据发现任务中勾选了“手动同步数据识别结果”，则还需要手动单击“数据同步”，才能将识别到的敏感数据同步到数据地图组件（同步数据前需确保已在数据目录中完成元数据采集任务，否则会同步失败）。

图 12-186 修正敏感数据

----结束

12.4.9 管控敏感数据

数据安全支持对数据地图资产按照密级进行分级管控，控制不同用户对元数据的访问权限。通过敏感数据管控为指定用户/用户组配置指定密级后，则用户/用户组在数据预览时仅能访问资产密级小于等于指定密级的字段。

值得注意的是，密级权限管控策略为DataArts Studio实例级别配置，各工作空间之间数据互通，全局可见并生效。未配置密级权限管控策略时，数据安全会预置一条默认策略，该策略默认给所有用户最大的密级访问权限；在管理员将策略配置好后，可删除此默认策略。

前提条件

已通过敏感数据识别任务，自动或手动将敏感数据同步到数据地图组件，详见**发现敏感数据**或**查看敏感数据分布**。

约束与限制

- 仅DAYU Administrator、Tenant Administrator用户或者数据安全管理员可以创建、修改或删除密级权限管控策略，其他普通用户无权限操作。
- 密级权限管控仅支持对数据地图中已标记密级的字段在数据预览时进行权限管控，不支持对已标记密级的表进行权限管控。
- 用户/用户组和密级共同唯一标识一条密级权限管控策略，因此不支持创建同用户/用户组、同密级的策略。
- 同用户/用户组如果对应多个密级，则以最高密级为准进行密级权限管控。

创建敏感数据管控策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

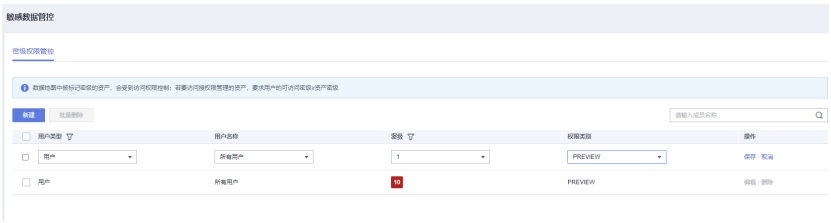
步骤2 单击左侧导航树中的“敏感数据管控”，进入敏感数据管控页面。

敏感数据管控页面的策略列表中，已有数据安全预置的默认策略，该策略默认给所有用户最大的密级访问权限。

图 12-187 进入敏感数据管控页面

步骤3 单击“新建”，新建密级权限管控策略页面，参数配置参考表12-38。

图 12-188 新建密级权限管控策略参数配置



创建密级权限管控策略参数配置说明：

表 12-38 配置策略参数

参数名	参数说明
*用户类型	选择为用户或用户组进行密级权限管控。
*用户名称	选择当前实例所有工作空间成员中的用户或用户组。
*密级	选择指定用户/用户组的指定密级，则指定用户/用户组仅能访问资产密级小于等于指定密级的资产。
*权限类别	当前仅支持数据地图中的数据预览权限。

步骤4 单击“保存”，完成密级权限管控策略创建。

说明

密级权限管控策略创建完成后，需要删除默认策略，以使新建的策略生效。

----结束

相关操作

- 编辑密级权限管控策略：在敏感数据管控页面，单击对应策略操作栏中的“编辑”，即可修改策略的用户/用户组、密级和权限类别。
- 删除密级权限管控策略：在敏感数据管控页面，单击对应策略操作栏中的“删除”，即可删除该策略。当需要批量删除时，可以在勾选策略后，在列表上方单击“批量删除”。

  说明

  删除操作无法撤销，请谨慎操作。

12.5 敏感数据保护

12.5.1 隐私数据保护简介

隐私数据保护是数据安全提供的一项用于敏感数据保护的功能。在隐私数据保护模块，您可以通过数据静态脱敏、动态脱敏、数据水印、文件水印和动态水印等方式来

防止敏感数据遭到有意或无意的误用、泄漏或盗窃，从而帮助企业采取合理措施来保护其敏感数据的机密性和完整性、可用性。

保护方式

隐私数据保护提供以下敏感数据保护方式：

- 静态脱敏

  数据静态脱敏，可以防止隐私数据在未经脱敏的情况下从企业流出。满足企业既要保护隐私数据，同时又保持监管合规，满足企业合规性。敏感数据通过静态脱敏，提供内置高效、丰富的脱敏算法，对原始数据中敏感数据进行掩码、截断、hash等，并将脱敏后的数据写入到目标端数据表。而目标表数据可以用来对外提供数据服务，为数据安全使用提供基础保障。

- 动态脱敏

  在数据安全组件创建动态脱敏策略后，系统会将动态脱敏策略同步到数据源服务，由数据源对数据列按照指定规则进行动态脱敏。当策略中指定的用户和用户组在访问敏感数据时，系统会直接返回由数据源动态脱敏后的数据，保护敏感数据不被泄露。

- 数据水印

  数据安全支持将水印标记嵌入到原始数据，保证数据的可用性。加入水印后的数据具有透明性、可用性、隐蔽性，不易被外部发现破解。数据泄漏后能够溯源水印标识，从而对安全事件精准定位追责。通过数据水印嵌入后的敏感数据一旦发生数据泄露，数据溯源可以通过导入泄露文件运行溯源任务提取水印标识，精准定位泄露单位及责任人。

- 文件水印

  文件水印支持如下两种场景，能够将水印注入数据文件中，实现对安全事件精准定位追责。

  - 对结构化数据文件（csv、xml和json）注入暗水印，水印内容不可见，需要进行水印提取。
  - 对非结构化数据文件（docx、pptx、xlsx和pdf）注入明水印，可在本地打开文件，查看水印内容。

- 动态水印

  在数据安全组件开启数据开发动态水印功能并创建动态水印策略后，当策略中指定的用户组或角色在数据开发组件中转储或下载敏感数据时，数据开发组件会为敏感数据注入暗水印，保护敏感数据不被泄露。

12.5.2 静态脱敏任务

12.5.2.1 管理脱敏算法

为了方便对数据进行脱敏，在创建脱敏策略前，需要您准备好脱敏算法。当前系统已内置20+脱敏算法，如果内置算法可以满足您的需求，您需要提前配置对应算法参数；否则，您可以新建脱敏算法。

本章主要介绍内置脱敏算法，和如何新建脱敏算法。

约束与限制

- 新建随机脱敏或字符替换类型的脱敏算法时，如果选择将敏感数据脱敏为样本库脱敏，则测试算法时限制样本文件大小不能超过10kb。注意，10kb仅为算法测试功能的限制，静态脱敏时并不限制样本文件大小不超过10kb。

内置脱敏算法介绍

数据安全提供了如下内置脱敏算法供您选择使用。建议您在选择算法之前，可以使用预先提供的内置算法配置和测试功能，以保证自己选择了合适的算法。

表 12-39 内置算法介绍

算法类型	内置算法名称	算法描述	是否支持配置
哈希	HMAC-SHA256哈希	使用HMAC-SHA256算法进行哈希处理。	支持配置盐值和密钥。 **说明** • 算法使用前必须先配置密钥，此算法才能正常使用。 • 算法盐值由您自行配置，而非系统给出的安全随机数，请关注相应使用风险。
	SHA-256哈希	使用SHA-256算法进行哈希处理。	支持配置盐值。 **说明** 算法盐值由您自行配置，而非系统给出的安全随机数，请关注相应使用风险。
截断	数值类型截断	保留小数点前x位，将小数点前第1到x-1位、小数点后的位数全部截断并填补为0。 例如x=3时，1234截断为1200，999.999截断为900，10.7截断为0。	支持配置保留小数点前几位。
	日期类型截断	截断日期指定位置。	支持配置日期格式和掩盖范围。
掩码	dws指定列全掩码	dws指定数据列全脱敏。 仅当静态脱敏任务中源端、目标端数据源同为DWS，且执行引擎为DWS时才可以选择此算法。	不支持。
	dws字符型掩码	从start到end的位置脱敏成指定的字符。 仅当静态脱敏任务中源端、目标端数据源同为DWS，且执行引擎为DWS时才可以选择此算法。	支持配置开始位置、结束位置和掩码标志。

算法类型	内置算法名称	算法描述	是否支持配置
	dws数值型掩码	从start到end的位置脱敏成指定的数字。 仅当静态脱敏任务中源端、目标端数据源同为DWS，且执行引擎为DWS时才可以选择此算法。	支持配置开始位置、结束位置和掩码标志。
	身份证号码掩码	掩码身份证号。	不支持。
	银行卡号掩码	掩码银行卡号。	不支持。
	Email掩码	掩码Email信息。	不支持。
	移动设备标识掩码	对设备码进行掩码，支持IMEI、MEDI、ESN。	支持配置类型。
	IPv6掩码	掩码IPv6地址。	不支持。
	IPv4掩码	掩码IPv4地址。	不支持。
	MAC地址掩码	掩码MAC地址。	不支持。
	电话号码掩码	掩码电话号码。	不支持。
	日期类型掩码	对指定日期格式进行掩码，支持ISO、EUR、USA格式。	支持配置日期格式和掩盖范围。
	掩码自x至y	掩码字符串第x至y位字符。	支持配置x和y。
	保留自x至y	保留字符串第x至y位字符。	支持配置x和y。
	掩码前n后m	掩码字符串前n后m位字符。	支持配置n和m。
	保留前n后m	保留字符串前n后m位字符。	支持配置n和m。

算法类型	内置算法名称	算法描述	是否支持配置
加密	dws列加密	调用GaussDB(DWS)提供的对称密码算法gs_encrypt_aes128(encryptstr,keystr)实现对DWS数据列的加密，此算法以keystr为密钥对encryptstr字符串进行加密，返回加密后的字符串。 算法注意事项如下： ● 仅当脱敏任务的目标源为DWS时，此算法才能正确生效。 ● 加密后执行SQL解密时，必须当所有的数据都解密成功时，才能正确返回解密结果，否则解密失败。	支持配置密钥，长度范围为1~16字节。 **说明** 算法使用前必须先配置密钥，此算法才能正常使用。
	hive列加密	调用MRS提供的Hive列加密功能来实现对Hive数据列的加解密，支持AES和SMS4两种加密算法。 算法注意事项如下： ● 仅当脱敏任务的目标源为Hive时，此算法才能正确生效。 ● 列加密只支持存储在HDFS上的TextFile和SequenceFile文件格式的表。 ● Hive列加密不支持视图以及Hive over HBase场景。	支持配置加密类型。

新建脱敏算法

如果内置算法不满足您的需求，您可以新建自定义脱敏算法，自定义脱敏算法支持掩码、截断、哈希、加密、置空、随机脱敏、字符替换、键值脱敏、数值区间变换、模糊脱敏等10余类算法类型。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

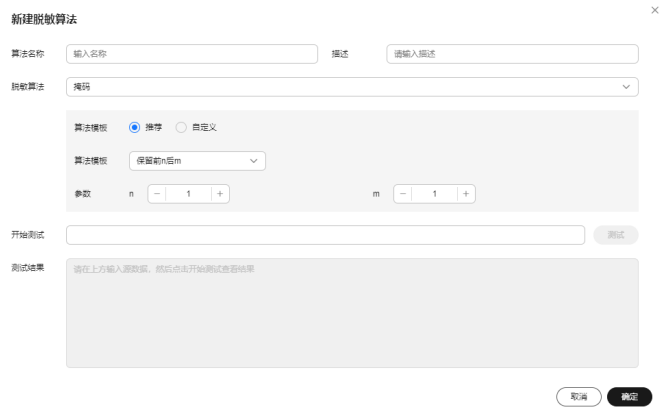
步骤2 在数据安全控制台左侧的导航树中单击“脱敏算法”，进入脱敏算法页面。

步骤3 单击“新建”，新建脱敏算法。

图 12-189 新建脱敏算法

步骤4 新建脱敏算法参数配置请参考**表12-40**，参数配置完成单击“确定”即可。

图 12-190 算法配置



脱敏算法参数说明表：

表 12-40 配置脱敏算法参数说明

配置	说明
*算法名称	用户自定义算法名称，长度不能超过64个字符。
描述	对当前算法进行简单描述，长度不能超过255个字符。

配置	说明
*脱敏算法	自定义脱敏算法支持掩码、截断、哈希、加密、置空、随机脱敏、字符替换、键值脱敏、数值区间变换、模糊脱敏等10余类算法类型，您可以根据脱敏需求自行选择。 • 掩码：支持字符型、数值型、日期型掩码，将指定位置的原始数据脱敏为固定值。 • 截断：支持日期类型和数值类型截断，将日期截断到月日小时分秒，将数值截断取整。 • 哈希：支持所有类型，使用所选的算法计算HASH值。 • 加密：支持所有类型，使用所选的数据源加密算法为对应数据源的数据进行加密。 • 置空：支持所有类型，将值设置为null。 • 随机脱敏：支持日期类型和数值类型随机脱敏，将日期或数值脱敏为指定区间范围之内或样本库中的值。新建样本库的请参考**管理样本库**章节。注意，选择样本库脱敏时，OBS样本文件只能用于DLI引擎的静态脱敏任务，HDFS样本文件只能用于MRS引擎的静态脱敏任务。静态脱敏场景与引擎之间的对应关系请参考**参考：静态脱敏场景介绍**。 随机脱敏支持配置“随机算法保持原数据关联性”参数，开启后不同数据库中的相同数据，经过相同的规则脱敏后，脱敏结果是一致的。注意此参数开启后会存在被破解的安全风险，如确需开启，建议配置随机盐值，用于抵抗字典攻击。 • 字符替换：支持数值类型和字符类型字符替换，将指定位置的字符替换为固定值或者样本库中样本文件的值；自定义替换位置时支持使用随机数值或随机小写英文字母替换，并支持身份证号末位计算（计算身份证末位时，位数只能选择1，且前面位数需要大于等于17）。 新建样本库的请参考**管理样本库**章节。注意，选择样本库替换时，OBS样本文件只能用于DLI引擎的静态脱敏任务，HDFS样本文件只能用于MRS引擎的静态脱敏任务。静态脱敏场景与引擎之间的对应关系请参考**参考：静态脱敏场景介绍**。 随机脱敏支持配置“随机算法保持原数据关联性”参数，开启后不同数据库中的相同数据，经过相同的规则脱敏后，脱敏结果是一致的。注意此参数开启后会存在被破解的安全风险，如确需开启，建议配置随机盐值，用于抵抗字典攻击。 • 键值脱敏：支持数值类型键值脱敏，根据自定义表达式，将数值脱敏为计算后的数值。填写表达式时，原始数据变量为X，支持对原始数据进行加（+)减(-)乘(*)除(/)、括号(())、取余(%)计算操作。例如表达式为“((X*4+3)%100)/2-1”时，数值3的脱敏结果为6.5。 • 数值区间变换：支持数值类型区间变换，将指定区间之内的数字变换为指定值。 • 模糊脱敏：支持数值类型模糊脱敏，支持在百分比或绝对值模糊的区间范围内随机取值。例如百分比模糊模式，百分比分别为-10%和20%时，数值10的模糊脱敏结果为[9,12]区间范围内随机取值。 随机脱敏支持配置“随机算法保持原数据关联性”参数，开启后不同数据库中的相同数据，经过相同的规则脱敏后，脱敏结

配置	说明
	果是一致的。注意此参数开启后会存在被破解的安全风险，如确需开启，建议配置随机盐值，用于抵抗字典攻击。
开始测试	输入待测试的数据后，单击“测试”，可在测试结果处查看脱敏结果。 **说明** 新建随机脱敏或字符替换类型的脱敏算法时，如果选择将敏感数据脱敏为样本库脱敏，则测试算法时限制样本文件大小不能超过10kb。
测试结果	

----结束

相关操作

- 编辑算法：在脱敏算法页面，单击对应算法操作栏中的“编辑”，即可修改算法参数。

  不同算法是否支持编辑和支持修改的参数因实际算法不同有所差异，请以操作界面为准。
- 测试算法：在脱敏算法页面，单击对应算法操作栏中的“测试”，即可测试该算法。

  说明

  建议您在使用算法之前，使用算法测试功能，以保证自己选择了合适的算法。

  不同算法是否支持测试因实际算法不同有所差异，请以操作界面为准。
- 删除算法：在脱敏算法页面，单击对应算法操作栏中的“删除”，即可删除算法。当需要批量删除时，可以在勾选算法后，在列表上方单击“批量删除”。

  注意，内置算法不支持删除，已在脱敏策略或指定列脱敏中引用的自定义算法无法删除。若要删除已引用的自定义算法，需要先修改引用关系，再进行删除操作。

  说明

  删除操作无法撤销，请谨慎操作。

12.5.2.2 管理样本库

数据安全支持将您提供的OBS或HDFS样本文件生成样本库。当新建随机脱敏或字符替换类型的脱敏算法时，可以选择将敏感数据脱敏为样本库文件中的值，详见**新建脱敏算法**。

本章主要介绍如何创建样本。

前提条件

已在OBS或HDFS中上传样本文件。样本文件只支持txt格式，大小建议不超过10MB，其中数据可通过换行“\n”、空格“ ”、英文逗号“,”、或分隔符“|”进行分隔。

约束与限制

- 新建随机脱敏或字符替换类型的脱敏算法时，如果选择将敏感数据脱敏为样本库脱敏，则测试算法时限制样本文件大小不能超过10kb。注意，10kb仅为算法测试功能的限制，静态脱敏时并不限制样本文件大小不超过10kb。
- 样本文件大小建议不超过10MB，否则运行需要解析样本文件的静态脱敏任务时，静态脱敏任务可能会失败。
- OBS样本文件只能用于DLI引擎的静态脱敏任务，HDFS样本文件只能用于MRS引擎的静态脱敏任务。静态脱敏场景与引擎之间的对应关系请参考**参考：静态脱敏场景介绍**。

新建样本

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“样本库”，进入样本库管理页面。

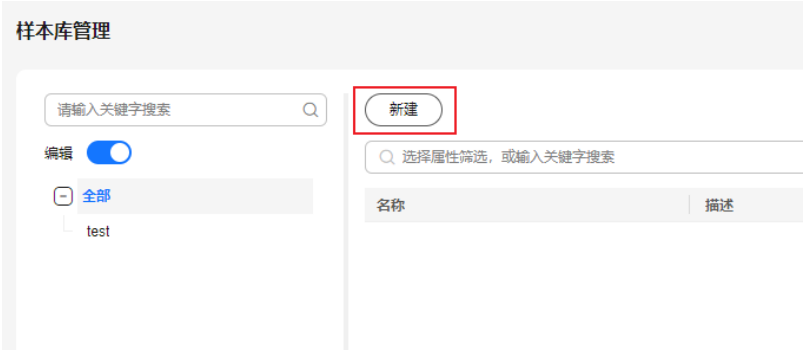
图 12-191 进入样本库管理页面

步骤3 在样本库管理页面，单击目录上的，然后在光标移动到目录上，单击后，输入分类名用于新增样本库分类。分类名称只能包含英文字母、数字、“_”，且长度不超过64个字符，超出部分将被截断。样本库分类最多支持10层（不包含“全部”层）。

图 12-192 新增样本库分类

步骤4 样本库分类创建完成后，在右侧点样本列表中单击“新建”，新建样本。新建样本时，默认填充分类为左侧选中的分类。

图 12-193 新建样本



步骤5 在弹出的新建窗口中填写样本信息，参考表12-41完成配置。配置完成后单击“确定”即可。

图 12-194 新建样本窗口

表 12-41 新建样本参数配置

参数	参数描述
*名称	样本名称，只能包含英文字母、数字、“_”，且长度不能超过64个字符，超出部分将被截断。
描述	为更好地识别样本，此处加以描述信息，长度不能超过1024个字符。
*分类	默认填充分类为左侧选中的样本分类，您也可以单击选择已有分类。

参数	参数描述
*选择样本	选择已上传至OBS或HDFS中的样本文件。样本文件只支持txt格式，大小建议不超过10MB，其中的数据可通过换行“\n”、空格“ ”、英文逗号“,”、或分隔符“|”进行分隔。 注意，OBS样本文件只能用于DLI引擎的静态脱敏任务，HDFS样本文件只能用于MRS引擎的静态脱敏任务。静态脱敏场景与引擎之间的对应关系请参考**参考：静态脱敏场景介绍**。
*分隔符	选择样本文件中数据的分隔符，可选择换行“\n”、空格“ ”、英文逗号“,”、或分隔符“|”。

----结束

相关操作

- 编辑样本库分类：在样本库管理页面，单击目录上的，然后在光标移动到待编辑的分类上，单击后，编辑分类名。
- 删除样本库分类：在样本库管理页面，单击目录上的，然后在光标移动到待编辑的分类上，单击后，删除分类。
  如果样本库分类下还存在样本，则不允许被删除。另外，“全部”根节点分类也不允许删除。

  说明

  删除操作无法撤销，请谨慎操作。
- 编辑样本：在样本库管理页面，单击对应样本操作栏中的“编辑”，即可修改样本的各项参数。
- 删除样本：在样本库管理页面，单击对应样本操作栏中的“删除”，即可删除样本。
  注意，被脱敏算法引用的样本不能被删除。若要删除已引用的样本，需要先修改引用关系，再进行删除操作。

  说明

  删除操作无法撤销，请谨慎操作。

12.5.2.3 管理脱敏策略

在实际生产中，会存在数据分析部门需要对数据进行数据分析，数据中存在敏感信息，但又不得不开放权限。此时就可以建立脱敏策略并对敏感数据进行脱敏，在满足业务需要的同时保证了数据的真实性不被泄露。

本章主要介绍如何创建脱敏策略。此处的脱敏策略仅适用于静态脱敏任务。

前提条件

- 已定义敏感数据识别规则，未定义请参考**定义识别规则（部分高级特性）**完成定义。
- 已配置内置脱敏算法或者已自定义脱敏算法，请参考**管理脱敏算法**进行配置和定义。

新建脱敏策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“脱敏策略”，进入脱敏策略页面，在页面上方单击“新建”，创建脱敏策略。

图 12-195 创建脱敏策略

步骤3 在弹出的新建脱敏页面中填写策略信息，参考表12-42完成配置。配置完成后单击“确定”即可。

图 12-196 创建脱敏策略界面

表 12-42 创建脱敏策略参数配置

参数	参数描述
*策略名称	用户自定义策略名称，只能包含英文字母、数字、“_”，且长度不能超过64个字符。
描述	为更好地识别脱敏策略，此处加以描述信息，长度不能超过255个字符。
*状态	开启状态表示该策略可供使用。关闭状态表示该策略不能被使用。
*识别规则和脱敏算法	选择敏感数据的识别规则，以及对应的脱敏算法。 ● *识别规则：选择已经定义的数据识别规则，详情请参考**定义识别规则（部分高级特性）**。 ● 规则描述：增加相应规则描述。 ● *算法类型：下拉选择算法类型，详情请参考**表12-39**。 ● *脱敏算法：下拉选择算法类型关联的算法，详情请参考**表12-39**。 **说明** 如下算法在使用前必须先在脱敏算法处配置密钥，才能正常使用。 ● 哈希算法中的“HMAC-SHA256哈希”算法。 ● 加密算法中的“dws列加密”算法。 不同脱敏算法的更多使用限制，请参考**管理脱敏算法**。

----结束

相关操作

- 编辑脱敏策略：在脱敏策略页面，单击对应策略操作栏中的“编辑”，即可修改脱敏策略各项参数。
- 编辑脱敏策略状态：新增的脱敏策略默认为启用状态。当脱敏策略为关闭状态时，表示该策略将不可被静态脱敏任务引用。

  需要修改脱敏策略状态时，在脱敏策略页面单击对应脱敏策略中的  或  ，即可启用或关闭脱敏策略。

  说明

  被静态脱敏任务引用的脱敏策略不能关闭。
- 删除脱敏策略：在脱敏策略页面，单击对应策略操作栏中的“删除”，即可删除策略。当需要批量删除时，可以在勾选脱敏策略后，在列表上方单击“批量删除”。

  注意，被静态脱敏任务引用的策略不能被删除。若要删除已引用的策略，需要先修改引用关系，再进行删除操作。

  说明

  删除操作无法撤销，请谨慎操作。

12.5.2.4 管理静态脱敏任务

本章主要介绍如何创建静态脱敏任务，静态脱敏支持的源端和目的端可通过**参考：静态脱敏场景介绍**查看。

数据静态脱敏，可以防止隐私数据在未经脱敏的情况下从企业流出。满足企业既要保护隐私数据，同时又保持监管合规，满足企业合规性。敏感数据通过静态脱敏，提供内置高效、丰富的脱敏算法，对原始数据中敏感数据进行掩码、截断、hash等，并将脱敏后的数据写入到目标端数据表。而目标表数据可以用来对外提供数据服务，为数据安全使用提供基础保障。

前提条件

- 静态脱敏任务需要根据脱敏策略来进行脱敏，相关前提条件如下：
  - 已配置内置脱敏算法或者已自定义脱敏算法，请参考**管理脱敏算法**进行配置和定义。
  - 已完成脱敏策略的创建，请参考**新建脱敏策略**。
  - 待脱敏的数据表已完成敏感数据发现任务，请参考**创建敏感数据发现任务**。
  - 已通过“敏感数据分布”，修正敏感数据字段的数据状态为“有效”，请参考**查看敏感数据分布**。
- DLI引擎静态脱敏任务，需要为dlg_agency委托授予如下OBS权限策略，授权方法可参考**授权dlg_agency委托**章节。

```
obs:bucket:HeadBucket
obs:bucket:CreateBucket
obs:object:PutObject
obs:object:DeleteObject
obs:bucket:ListBucket
obs:object:GetObject
obs:bucket:GetEncryptionConfiguration
obs:bucket:PutEncryptionConfiguration
```

约束与限制

- 静态脱敏时，请根据待脱敏数据的字段类型正确选择脱敏算法，否则可能会导致数据库数据异常。例如对date字段使用数值随机算法脱敏，会导致data类型将被强制脱敏为数值类型（Hive和DLI脱敏），或者写入失败报错（DWS脱敏）；对数值字段使用哈希算法脱敏，会导致数值类型被强制脱敏为哈希值字符串（Hive和DLI脱敏），或者写入失败报错（DWS脱敏）。
- 运行需要解析样本文件的静态脱敏任务时，样本文件大小建议不超过10MB，否则静态脱敏任务可能会失败。另外，OBS样本文件只能用于DLI引擎的静态脱敏任务，HDFS样本文件只能用于MRS引擎的静态脱敏任务。静态脱敏场景与引擎之间的对应关系请参考**参考：静态脱敏场景介绍**。
- DLI引擎的静态脱敏任务，运行参数需要存储在OBS桶中，任务运行完成或失败后会删除任务运行参数文件。
  - DLI引擎的同源静态脱敏任务，运行参数存储在工作空间日志桶中，默认以dlf-log-{Project id}命名。
  - DLI引擎的跨源静态脱敏任务，运行参数存储在自动创建的加密用户桶dls-dli-{projectId}中。

  因此DLI引擎静态脱敏前，还需要为dlg_agency委托授予如下OBS权限策略，授权方法可参考**授权dlg_agency委托**章节。

```
obs:bucket:HeadBucket
obs:bucket:CreateBucket
```

```
obs:object:PutObject
obs:object:DeleteObject
obs:bucket:ListBucket
obs:object:GetObject
obs:bucket:GetEncryptionConfiguration
obs:bucket:PutEncryptionConfiguration
```

- DLI引擎的静态脱敏任务，当源端或目的端为DWS时，请参考**配置DLI队列与内网数据源的网络联通**或**配置DLI队列与公网网络联通**打通DLI Spark通用队列与DWS的网络连接，否则会导致静态脱敏任务失败。
- 源端或目的端为DLI的静态脱敏任务，不支持对DLI中default数据库的数据表进行脱敏。
- MapReduce服务（MRS Hive）所在的MRS集群必须开启Kerberos认证，且必须安装Spark组件。
- MRS引擎的静态脱敏任务，当源端或目的端为DWS时，请参考**参考：授权并绑定委托**为MRS集群配置委托，并确保MRS集群安全组出方向规则满足如下要求，否则会导致静态脱敏任务失败。
  - 协议：TCP
  - 端口范围：80
  - 远端地址：169.254.0.0/16
- MRS引擎的静态脱敏任务，当源端或目的端仅一端为DWS时，支持的数据类型如下。如果有其他不支持的数据类型，将导致静态脱敏任务失败。
  - tinyint
  - smallint
  - int
  - bigint
  - decimal
  - double
  - float
  - boolean
  - string
  - timestamp
- DWS引擎的同源静态脱敏任务，不支持跨数据库脱敏，即DWS源端和目的端数据表所在的数据库必须相同。
- 静态脱敏任务的数据集范围选择为增量时，需选择时间字段类型Timestamp、Date字段类型来确定增量范围。

创建静态脱敏任务

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“静态脱敏”，进入静态脱敏页面，在页面上方单击“新建”，创建静态脱敏任务。

图 12-197 创建静态脱敏任务

步骤3 在弹出的创建任务页面中填写任务名称和描述，单击“下一步”。

图 12-198 基本信息配置

步骤4 进行脱敏任务源、目标端配置。参数配置参考**表12-43**。

图 12-199 配置脱敏任务

脱敏任务参数配置说明：

表 12-43 脱敏任务参数配置

参数名	参数描述
源端配置	
*数据源类型	目前支持数据湖探索（DLI）、数据仓库服务（DWS）和MapReduce服务（MRS Hive）。
*数据连接	选择已在管理中心组件创建的数据连接。若未创建请参考**创建DataArts Studio数据连接**新建连接。
*SQL队列	数据源类型为DLI时，需要选择DLI SQL队列。
*数据库	单击设置选择待脱敏的数据库。 不支持对DLI default数据库中的数据表进行脱敏。
*源表名	单击设置选择待脱敏的数据表。

参数名	参数描述
*是否指定列	支持指定列脱敏。开启后您可以对源表中的指定列配置脱敏算法，支持对多列分别配置不同的脱敏算法。 **说明** 注意该参数确定后，无法再通过编辑任务修改此选项。
*指定列	开启“是否指定列”时，此参数为必选项。 如果您需要对某列进行脱敏，则必须勾选对应列，然后选择脱敏算法才能生效。如果仅选择脱敏算法，则无法实现脱敏。 **说明** • 静态脱敏时，请根据待脱敏数据的字段类型正确选择脱敏算法，否则可能会导致数据库数据异常。例如对date字段使用数值随机算法脱敏，会导致data类型将被强制脱敏为数值类型（Hive和DLI脱敏），或者写入失败报错（DWS脱敏）；对数值字段使用哈希算法脱敏，会导致数值类型被强制脱敏为哈希值字符串（Hive和DLI脱敏），或者写入失败报错（DWS脱敏）。 • 如下算法在使用前必须先在脱敏算法处配置密钥，才能正常使用。 – 哈希算法中的“HMAC-SHA256哈希”算法。 – 加密算法中的“dws列加密”算法。 不同脱敏算法的更多使用限制，请参考**管理脱敏算法**。
*数据集范围	只有使用时间字段timestamp、Date来确定增量范围时，才可以选择增量模式。 一般而言，全量模式下脱敏任务使用单次调度，增量模式下脱敏任务使用周期调度。
*指定时间字段	增量模式下，选择时间字段timestamp、Date来确定增量范围。
脱敏策略配置	
*脱敏策略	仅当未指定列时可配置。 下拉选择您预先创建好的脱敏策略。 **说明** • 静态脱敏时，请根据待脱敏数据的字段类型正确选择脱敏算法，否则可能会导致数据库数据异常。例如对date字段使用数值随机算法脱敏，会导致data类型将被强制脱敏为数值类型（Hive和DLI脱敏），或者写入失败报错（DWS脱敏）；对数值字段使用哈希算法脱敏，会导致数值类型被强制脱敏为哈希值字符串（Hive和DLI脱敏），或者写入失败报错（DWS脱敏）。 • 如下算法在使用前必须先在脱敏算法处配置密钥，才能正常使用。 – 哈希算法中的“HMAC-SHA256哈希”算法。 – 加密算法中的“dws列加密”算法。 不同脱敏算法的更多使用限制，请参考**管理脱敏算法**。
目标端配置	
*数据源类型	选择存储脱敏后数据的数据源类型，支持的脱敏场景如**表12-45**所示。
*数据连接	选择已在管理中心组件创建的数据连接。若未创建请参考**创建DataArts Studio数据连接**新建连接。

参数名	参数描述
*SQL队列	数据源类型为DLI时，需要选择DLI SQL队列。
*数据库	单击设置选择存储已脱敏数据的数据库。 不支持对DLI default数据库中的数据表进行脱敏。
*目标表名	用户手动输入，不能与目标端数据库表名重复。当输入的表名不存在时会创建该表。 输入请单击“测试”，测试创建目标表并检测目标表是否可用，否则将无法进行下一步操作。
执行引擎	
*执行引擎	选择运行脱敏任务的引擎。不同脱敏场景下支持的引擎和注意事项如**表12-45**所示。
脱敏队列	
*脱敏队列	选择对应执行DLI或MRS引擎下的队列。 • 执行引擎为DLI时，脱敏队列选择为DLI Spark通用队列。DLI引擎的静态脱敏任务，当源端或目的端为DWS时，请参考**配置DLI队列与内网数据源的网络联通**或**配置DLI队列与公网网络联通**打通DLI Spark通用队列与DWS的网络连接，否则会导致静态脱敏任务失败。 • 执行引擎为MRS时，脱敏队列需要手动填写为MRS租户队列，可在MRS控制台集群列表中单击集群名进入集群详情，在“租户管理 > 队列配置”中查看可用队列。

步骤5 单击“下一步”，进行调度信息配置。

- 数据集范围为全量模式时，仅支持单次调度。
- 数据集范围为增量模式时，支持单次调度和周期调度。

当选择为周期调度时，参数配置参考**表12-44**。

表 12-44 配置周期调度参数

参数名	说明
*调度日期	调度任务的生效时间段。

参数名	说明
*调度周期	选择调度任务的执行周期，并配置相关参数。 ● 分：选择调度开始时间和结束时间，配置间隔的分钟时长。 ● 小时：选择调度开始时间和结束时间，配置间隔的小时时长。 ● 天：配置每日调度时间。 ● 周：选择星期几启动调度，配置调度具体时间。 ● 月：选择几号启动调度，配置调度具体时间。 例如：选择调度周期是周，选择具体时间为15:52，时间选择为星期二。则在调度日期范围内，每周二的15点52分会执行任务。
立即启动	勾选复选框，则表示立即启动此调度任务。

图 12-200 周期调度配置参数

步骤6 单击“确定”，完成创建静态脱敏任务。

----结束

相关操作

- 编辑任务：在静态脱敏页面，单击对应任务操作栏中的“编辑”，即可编辑静态脱敏任务。

  运行状态为正在“执行中”的任务不允许被编辑。
- 删除任务：在静态脱敏页面，单击对应任务操作栏中的“更多 > 删除”，即可删除任务。当需要批量删除时，可以在勾选任务后，在任务列表上方单击“批量删除”。

  运行状态为正在“执行中”的任务不允许被删除。

  说明

  删除操作无法撤销，请谨慎操作。
- 运行或调度任务：在静态脱敏页面，单击对应任务操作栏中的“运行”或“更多 > 启动调度”，运行或调度任务。

您可以通过调度周期区分该任务是单次调度还是周期调度任务。

- 查看运行实例日志：在静态脱敏页面，找到需要查看实例的任务，单击 展开，即可找到运行实例。随后单击“查看日志”，查看运行实例日志。

  运行失败可通过日志排查失败原因，问题修正后尝试重新运行。如果仍运行失败，请联系技术支持人员协助处理。

参考：授权并绑定委托

步骤1 登录IAM服务控制台。

步骤2 选择“委托”，在委托列表中查找MRS预置的**MRS_ECS_DEFAULT_AGENCY**委托，并单击“授权”。

说明

如果未找到MRS预置的**MRS_ECS_DEFAULT_AGENCY**委托，则可以通过自定义购买方式来购买MRS集群，在高级配置中选择绑定**MRS_ECS_DEFAULT_AGENCY**委托。MRS集群开始创建后，会自动生成**MRS_ECS_DEFAULT_AGENCY**委托。

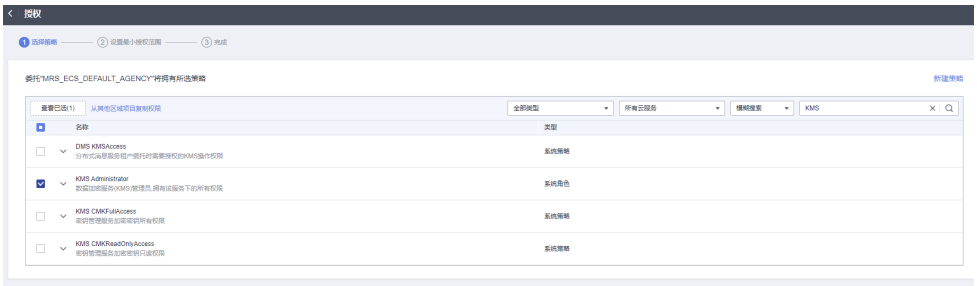
图 12-201 授权委托

步骤3 在授权页面中，在搜索框中输入“KMS”，勾选KMS Administrator策略。

说明

MRS_ECS_DEFAULT_AGENCY委托所需最小权限为“kms:cmk:decrypt”。除了直接授权KMS Administrator策略外，您也可以在IAM服务控制台创建一个自定义策略，包含KMS服务“kms:cmk:decrypt”权限，并将此策略授权给**MRS_ECS_DEFAULT_AGENCY**委托。

图 12-202 选择权限



步骤4 选择权限后，单击“下一步”设置授权范围。本例以默认选项为例，直接单击“确定”完成委托授权。

步骤5 在MRS服务控制台，选择“集群列表 > 现有集群”，单击集群名进入待配置集群的详情页面。

步骤6 在集群详情的概览页面，找到“运维管理”区域，确认集群已绑定**MRS_ECS_DEFAULT_AGENCY**委托。如果未绑定，需要手动选择**MRS_ECS_DEFAULT_AGENCY**委托并确认，完成绑定。

图 12-203 绑定委托

----结束

参考：静态脱敏场景介绍

隐私保护管理目前支持的静态脱敏场景如**表12-45**所示。

表 12-45 静态脱敏场景

源端数据源类型	目的端数据源类型	计算引擎	说明
数据湖探索（DLI）	数据湖探索（DLI）	使用DLI Spark通用队列	-
	数据仓库服务（DWS）	使用DLI Spark通用队列	• DLI引擎的静态脱敏任务，当源端或目的端为DWS时，请参考**配置DLI队列与内网数据源的网络联通**或**配置DLI队列与公网网络联通**打通DLI Spark通用队列与DWS的网络连接，否则会导致静态脱敏任务失败。

源端数据源类型	目的端数据源类型	计算引擎	说明
数据仓库服务（DWS）	数据仓库服务（DWS）	• 使用DWS集群 • 使用MRS集群 • 使用DLI Spark通用队列	**DWS引擎：** • DWS引擎的同源静态脱敏任务，不支持跨数据库脱敏，即DWS源端和目的端数据表所在的数据库必须相同。 **MRS引擎：** • MapReduce服务（MRS Hive）所在的MRS集群必须开启Kerberos认证，且必须安装Spark组件。 • MRS引擎的静态脱敏任务，当源端或目的端为DWS时，请参考**参考：授权并绑定委托**为MRS集群配置委托，并确保MRS集群安全组出方向规则满足如下要求，否则会导致静态脱敏任务失败。 – 协议：TCP – 端口范围：80 – 远端地址：169.254.0.0/16 **DLI引擎**： • DLI引擎的静态脱敏任务，当源端或目的端为DWS时，请参考**配置DLI队列与内网数据源的网络联通**或**配置DLI队列与公网网络联通**打通DLI Spark通用队列与DWS的网络连接，否则会导致静态脱敏任务失败。

源端数据源类型	目的端数据源类型	计算引擎	说明
	MapReduce服务（MRS Hive）	使用MRS Hive所在的MRS集群	• MapReduce服务（MRS Hive）所在的MRS集群必须开启Kerberos认证，且必须安装Spark组件。 • MRS引擎的静态脱敏任务，当源端或目的端为DWS时，请参考**参考：授权并绑定委托**为MRS集群配置委托，并确保MRS集群安全组出方向规则满足如下要求，否则会导致静态脱敏任务失败。 – 协议：TCP – 端口范围：80 – 远端地址：169.254.0.0/16 • MRS引擎的静态脱敏任务，当源端或目的端仅一端为DWS时，支持的数据类型如下。如果有其他不支持的数据类型，将导致静态脱敏任务失败。 – tinyint – smallint – int – bigint – decimal – double – float – boolean – string – timestamp
	数据湖探索（DLI）	使用DLI Spark通用队列	• DLI引擎的静态脱敏任务，当源端或目的端为DWS时，请参考**配置DLI队列与内网数据源的网络联通**或**配置DLI队列与公网网络联通**打通DLI Spark通用队列与DWS的网络连接，否则会导致静态脱敏任务失败。
MapReduce服务（MRS Hive）	MapReduce服务（MRS Hive）	使用源端MRS Hive所在的MRS集群	• MapReduce服务（MRS Hive）所在的MRS集群必须开启Kerberos认证，且必须安装Spark组件。

源端数据源类型	目的端数据源类型	计算引擎	说明
	数据仓库服务（DWS）	使用MRS Hive所在的MRS集群	• MapReduce服务（MRS Hive）所在的MRS集群必须开启Kerberos认证，且必须安装Spark组件。 • MRS引擎的静态脱敏任务，当源端或目的端为DWS时，请参考**参考：授权并绑定委托**为MRS集群配置委托，并确保MRS集群安全组出方向规则满足如下要求，否则会导致静态脱敏任务失败。 – 协议：TCP – 端口范围：80 – 远端地址：169.254.0.0/16 • MRS引擎的静态脱敏任务，当源端或目的端仅一端为DWS时，支持的数据类型如下。如果有其他不支持的数据类型，将导致静态脱敏任务失败。 – tinyint – smallint – int – bigint – decimal – double – float – boolean – string – timestamp

12.5.3 动态脱敏任务

12.5.3.1 管理动态脱敏策略

在数据安全组件创建动态脱敏策略后，系统会将动态脱敏策略同步到数据源服务，由数据源对数据列按照指定规则进行动态脱敏。当策略中指定的用户和用户组在访问敏感数据时，系统会直接返回由数据源动态脱敏后的数据，保护敏感数据不被泄露。

值得注意的是，动态脱敏策略为DataArts Studio实例级别配置，各工作空间之间数据互通，全局可见并生效。

前提条件

- 新建MRS Hive脱敏策略前，已完成如下操作：
  – 在管理中心创建MapReduce服务（MRS Ranger）类型的数据连接，请参考**创建DataArts Studio数据连接**。

  - 已完成用户同步，将IAM上的用户信息同步到数据源上，详见**同步IAM用户到数据源**。
- 新建DWS脱敏策略前，已完成如下操作：
  - 已在管理中心创建数据仓库服务（DWS）类型的数据连接，请参考**创建DataArts Studio数据连接**。
  - 已完成用户同步，将IAM上的用户信息同步到数据源上，详见**同步IAM用户到数据源**。
  - 已修改DWS集群“feature_support_options”参数的CN参数值和DN参数值均为“enable_data_redaction”，用于启用DWS动态脱敏能力，修改操作详见**修改数据库参数**。
  - 数据连接中的账户要具备待控制表的GRANT权限（数据库对象创建后，默认只有对象所有者或者系统管理员可以通过GRANT命令将对象的权限授予其他用户）。
- MRS Hive和DWS动态脱敏策略为指定用户/用户组在数据源上关联策略，因此需要如果希望在DataArts Studio数据开发执行脚本、测试运行作业时，使用当前用户身份认证鉴权以实现动态脱敏策略生效，则需要**启用细粒度认证**。
- 如果希望创建脱敏策略时能够查看哪些字段为敏感字段，则需要提前完成敏感数据发现任务，并通过“敏感数据分布”修正敏感数据字段的数据状态为“有效”。详情请参考**发现敏感数据**和**查看敏感数据分布**。

约束与限制

- 仅DAYU Administrator、Tenant Administrator用户或者数据安全管理员可以创建、修改或删除动态脱敏策略，其他普通用户无权限操作。
- MRS Hive和DWS动态脱敏策略为指定用户/用户组在数据源上关联策略，因此需要如果希望在DataArts Studio数据开发执行脚本、测试运行作业时，使用当前用户身份认证鉴权以实现动态脱敏策略生效，则需要**启用细粒度认证**。
- 当前动态脱敏策略仅支持MRS Hive和DWS数据源。
- 单条动态脱敏策略的配置维度为表级别，即一个表只允许绑定一个策略，一个策略也是只允许绑定一个表。只有处于“同步成功”状态的策略才能生效。
- MRS Hive动态脱敏时，MRS Ranger支持对同一列配置不同规则，按照配置的时间顺序先后匹配，因此可以配置多条同集群、同库表列的不同内容的脱敏策略。
- 当前MRS服务支持的脱敏规则如**表12-47**所示，但对中文字符仅支持NULL掩盖和哈希掩盖两种脱敏方式，如果选择其他脱敏方式则脱敏不生效。
- DWS动态脱敏不支持DWS逻辑集群，脱敏前需启用DWS动态脱敏能力（修改DWS集群“feature_support_options”参数的CN参数值和DN参数值均为“enable_data_redaction”，修改操作详见**修改数据库参数**），且DWS数据连接中的账户要具备待脱敏表的GRANT权限（数据库对象创建后，默认只有对象所有者或者系统管理员可以通过GRANT命令将对象的权限授予其他用户）。
- 当前DWS服务支持的脱敏规则如**表12-48**所示，不支持中文脱敏，如果对含有中文字符的数据进行脱敏则可能会出现乱码。

创建动态脱敏策略

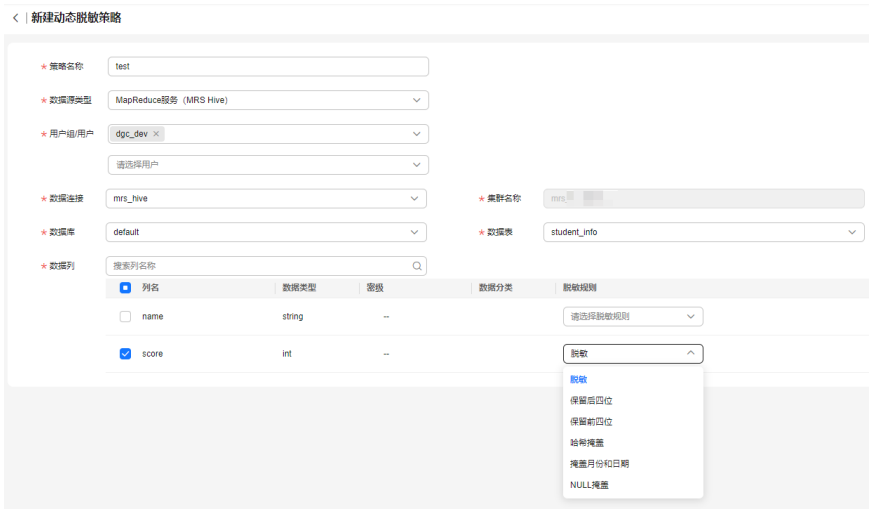
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“动态脱敏”，进入动态脱敏页面。

图 12-204 进入动态脱敏页面

步骤3 单击“新建”，进入新建动态脱敏策略页面，参数配置参考**表12-46**。

图 12-205 新建动态脱敏策略参数配置



创建动态脱敏策略参数配置说明：

表 12-46 配置策略参数

参数名	参数说明
*策略名称	动态脱敏策略的唯一标识，DataArts Studio实例内的名称唯一。 为便于策略管理，建议名称中标明要脱敏的对象和脱敏规则。
*数据源类型	当前支持MRS Hive、DWS数据源。
MRS Hive	
*用户组/用户	指定当前工作空间成员中的用户或用户组。当指定对象在数据开发组件中查询或导出敏感数据时，系统会对敏感数据进行动态脱敏，保护敏感数据不被泄露。

参数名	参数说明
*数据连接	从下拉列表中选择数据连接类型中已创建的数据连接，若未创建请参考**创建DataArts Studio数据连接**新建连接。
*集群名称	无需选择，自动匹配数据连接中的数据源集群。
*数据库	选择敏感数据所在的数据库。
*数据表	选择敏感数据所在的数据表。
*数据列	您需要勾选一个或多个待脱敏列，并根据不同数据列的数据类型，选择合适的脱敏规则。各类数据源中不同数据类型支持的脱敏规则不同，详见**参考：动态脱敏规则介绍**。 另外，如果选中的库表列有进行过敏感数据发现并且敏感数据字段的数据状态为“有效”，则将密级和数据分类显示在数据列区域中。
DWS	
*用户组/用户	指定当前工作空间成员中的用户或用户组。当指定对象在数据开发组件中查询或导出敏感数据时，系统会对敏感数据进行动态脱敏，保护敏感数据不被泄露。
*数据连接	从下拉列表中选择数据连接类型中已创建的数据连接，若未创建请参考**创建DataArts Studio数据连接**新建连接。
*集群名称	无需选择，自动匹配数据连接中的数据源集群。
*数据库	选择敏感数据所在的数据库。
*schema	选择敏感数据所在的schema。
*数据表	选择敏感数据所在的数据表。
*数据列	您需要勾选一个或多个待脱敏列，并根据不同数据列的数据类型，选择合适的脱敏规则。各类数据源中不同数据类型支持的脱敏规则不同，详见**参考：动态脱敏规则介绍**。 另外，如果选中的库表列有进行过敏感数据发现并且敏感数据字段的数据状态为“有效”，则将密级和数据分类显示在数据列区域中。

步骤4 单击“确定”，完成动态脱敏策略创建。动态脱敏策略创建完成后，需要手动单击“同步”，将该策略同步到数据源中。

----结束

相关操作

- 同步策略：在动态脱敏页面，单击对应任务操作栏中的“同步”，即可将该策略同步到数据源中。当需要批量同步时，可以在勾选策略后，在列表上方单击“同步”。

  只有处于“同步成功”状态的策略才能生效。如果策略同步失败，可通过**查看策略详情**查看策略运行日志，通过日志排查同步失败原因。待问题修复后请重新同步，如果仍同步失败，请联系技术支持人员协助处理。

- 编辑策略：在动态脱敏页面，单击对应任务操作栏中的“编辑”，即可编辑动态脱敏策略。
- 删除策略：在动态脱敏页面，单击对应任务操作栏中的“删除”，在弹窗中再次确认后，即可删除策略。当需要批量删除时，可以在勾选策略后，在列表上方单击“删除”。

  说明

  动态脱敏策略删除后将被转移至回收站中，您可以在30天内进行还原，在回收站中超过30天的数据将被自动删除。详见**管理回收站**章节。
- 查看策略详情：在动态脱敏页面，通过同步状态筛选策略或直接找到需要查看的策略，单击策略名即可查看策略详情。

图 12-206 查看策略详情

参考：动态脱敏规则介绍

- MRS Hive动态脱敏规则由MRS Ranger组件提供，当前支持的规则如**表12-47**所示。
- DWS动态脱敏规则由DWS提供，当前支持的规则如**表12-48**所示。

表 12-47 MRS 动态脱敏规则

数据类型	掩盖英文字符和数字	保留后四位	保留前四位	哈希掩盖	掩盖月份和日期	NULL掩盖
TINYINT	位数不变，将数值全部替换为1	无变化，最大值为127	无变化，最小值为-128	值变为NULL	位数不变，将数值全部替换为1	值变为NULL

数据类型	掩盖英文字符和数字	保留后四位	保留前四位	哈希掩盖	掩盖月份和日期	NULL掩盖
SMALLINT	位数不变，将数值全部替换为1	无变化，最大值为12767	无变化，最大值为-32768	值变为NULL	位数不变，将数值全部替换为1	值变为NULL
INT	位数不变，将数值全部替换为1	保留后四位	保留前四位	值变为NULL	位数不变，将数值全部替换为1	值变为NULL
BIGINT	位数不变，将数值全部替换为1	保留后四位	保留前四位	值变为NULL	位数不变，将数值全部替换为1	值变为NULL
BOOLEAN	值变为NULL	值变为NULL	值变为NULL	值变为NULL	值变为NULL	值变为NULL
FLOAT	值变为NULL	值变为NULL	值变为NULL	值变为NULL	值变为NULL	值变为NULL
DOUBLE	值变为NULL	值变为NULL	值变为NULL	值变为NULL	值变为NULL	值变为NULL
STRING	英文字母变为x，数字变为n	中文无变化，字母等变为X	中文无变化且占一位，字母等变为X	全部被hash到64长度	中文无变化且占一位，字母等变为X	值变为NULL
TIMESTAMP	值变为NULL	值变为NULL	值变为NULL	值变为NULL	值变为NULL	值变为NULL
CHAR	英文字母变为x，数字变为n	字母数字变为X，后面4位保留(定长有空格)	字母数字变为X，前面4位保留(定长有空格)	全部被hash到64长度	中文无变化且占一位，字母等变为X	值变为NULL
VARCHAR	英文字母变为x，数字变为n	后四位被保留(中文无变化且占一位)，字母等变为X	前四位被保留(中文无变化且占一位)字母等变为X	全部被hash到64长度	中文无变化且占一位，字母等变为X	值变为NULL
DATE	年月日变为0001-01-01	年月日变为0001-01-01	年月日变为0001-01-01	值变为NULL	year保留，其他数值变为01	值变为NULL

表 12-48 DWS 动态脱敏规则

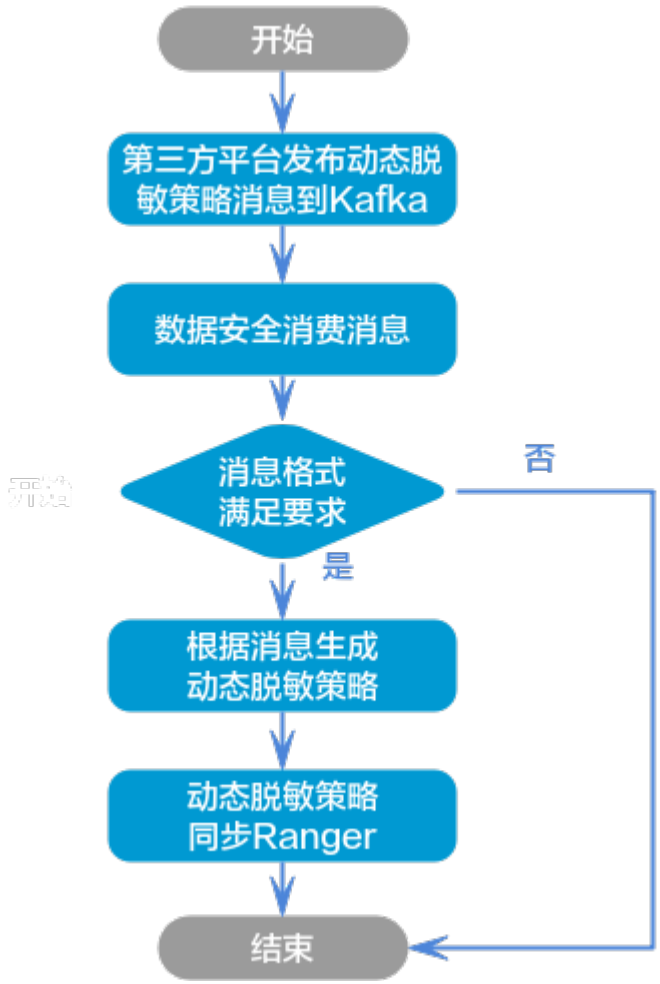
数据类型	全掩码	保留后4位，其余脱敏为‘*’	保留前2位，其余脱敏为‘*’	自定义
字符类型 bpchar、varchar、text、inet、macaddr、uuid、char、txt	全部脱敏为空。	后四位被保留，其余脱敏为“*”	前两位被保留，其余脱敏为“*”	自定义脱敏开始和结束位置，脱敏字符
数值类型 numeric、int2、int8、money、float8、float4、interval、decimal、double precision、real、integer、smallint、bigint	全部脱敏为“0”	不支持	不支持	自定义脱敏开始和结束位置，脱敏字符
时间类型 timestamp、time、timetz、timestamptz、date、time without time zone、timestamp without time zone、time without time zone、timestamp without time zone	全部脱敏为固定值	不支持	不支持	自定义勾选脱敏目标为年、月、日等
其他类型	全部脱敏为固定值	不支持	不支持	不支持

12.5.3.2 订阅动态脱敏策略

通过动态脱敏订阅，数据安全可以实现同步第三方平台的动态脱敏策略。

第三方平台的动态脱敏策略发布到Kafka消息队列后，数据安全进行订阅和消费。消息格式满足要求时，待消息消费成功后，数据安全会生成动态脱敏策略（策略名为Kafka消息中的策略名）并同步到MRS Ranger组件中生效。

图 12-207 动态脱敏订阅原理



值得注意的是，动态脱敏订阅为DataArts Studio实例级别配置，各工作空间之间数据互通，全局可见并生效。

前提条件

- 第三方平台的动态脱敏策略需要发布到Kafka消息队列，且消息格式满足要求，详见**参考：Kafka消息格式要求**。
- 已在管理中心创建MapReduce服务（MRS Kafka）类型的数据连接，请参考**创建DataArts Studio数据连接**。注意，Kafka应为第三方平台发布消息所在的Kafka，数据连接中的账户要具备kafkaadmin用户组的权限。

约束与限制

- 仅DAYU Administrator、Tenant Administrator用户或者数据安全管理员可以创建、编辑、启动、停止或同步动态脱敏订阅任务，其他普通用户无权限操作。
- 动态脱敏订阅仅支持订阅第三方平台中MRS Hive类型的动态脱敏策略，且动态脱敏策略中支持的脱敏规则仅限于数据安全中已支持的规则（暂不支持“自定义/保留前x后y”和“自定义/掩盖前x后y”两个自定义规则），详见**表12-47**。
- 通过订阅生成的动态脱敏策略名为Kafka消息中的策略名，由于数据安全不允许策略名重复，因此数据安全已有动态脱敏策略名需要避免与Kafka消息中的策略名重复。

- 订阅生成的动态脱敏策略名同步到Ranger后策略名为“dlsMasking-*库名-表名-列名*”，由于Ranger不允许策略名重复，因此Ranger已有策略名需要避免与生成的策略命名重复。
- 动态脱敏订阅时，数据安全通过订阅任务中的“MRS集群”+Kafka消息动态脱敏策略中的“库表列”来标识一条动态脱敏策略。当消息队列或数据安全中已存在同集群同库表列的动态脱敏策略时，则跳过不再重复生成。
- 数据安全消费Kafka消息时，需要消息的格式满足要求，详见**参考：Kafka消息格式要求**。
  - Kafka消息不满足消息格式：则记录同步失败消息日志，继续消费下一条消息，最终状态为部分失败或者同步失败。
  - Kafka消息合法，但是由于网络资源等原因消费失败：触发Kafka重试机制，重试3次，间隔分别为4、6、9s，如果依然失败，则记录日志，终止此次调度。
  - Kafka消息合法，正常消费，但是生成策略或同步Ranger时失败：记录同步失败消息日志，继续消费下一条，最终状态为部分失败或者同步失败。
  - 失败的kafka消息最多存储16M数据。

订阅动态脱敏策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“动态脱敏”，进入动态脱敏页面，然后切换到“动态脱敏订阅”页签。

图 12-208 进入动态脱敏订阅页签

步骤3 单击“创建订阅”，弹出创建订阅窗口，参数配置参考**表12-49**。

图 12-209 创建订阅参数配置

创建动态脱敏订阅任务参数配置说明：

表 12-49 配置任务参数

参数名	参数说明
连接配置	
*选择集群	选择需要同步第三方平台动态脱敏策略的集群。 当前暂不支持同步策略到多个集群。如果希望通过多个订阅任务分别同步到多个集群，则会由于生成的策略名重复导致Kafka消息消费失败。
集群类型	无需选择，自动根据选择的集群匹配集群类型。当前仅支持同步策略到MRS集群。
数据连接	无需选择，自动根据选择的集群匹配数据连接。
*kafka数据连接	选择在**前提条件**中已创建的MRS Kafka类型数据连接。注意，Kafka应为第三方平台发布消息所在的Kafka，Kafka数据连接中的账户要具备kafkaadmin用户组的权限。
*topic主题	选择第三方平台的动态脱敏策略发布Kafka消息的Topic主题。同一个MRS集群的一个Topic主题只能对应一个订阅任务。

参数名	参数说明
调度配置	
调度时间	选择每天调度生效的时间段。 建议消息量大小评估调度时间，目前消费一个数据加同步大约需要2秒。
调度周期	选择按小时还是按分钟调度。
调度间隔	选择调度间隔时间。

步骤4 单击“确定”，完成动态脱敏订阅任务的创建。动态脱敏策略创建完成后，需要手动单击“启动”，启动任务调度。

----结束

相关操作

- 启动/停止订阅任务：在动态脱敏订阅页签，单击对应任务操作栏中的“启动”或“停止”，即可启动或停止该任务的调度。
- 编辑订阅任务：在动态脱敏订阅页签，单击对应任务操作栏中的“更多 > 编辑”，即可编辑订阅任务。
- 删除订阅任务：在动态脱敏订阅页签，单击对应任务操作栏中的“更多 > 删除”，即可删除订阅任务。当需要批量删除时，可以在勾选订阅任务后，在列表上方单击“批量删除”。

  说明

  删除操作无法撤销，请谨慎操作。
- 同步订阅任务：在动态脱敏订阅页签，单击对应任务操作栏中的“更多 > 同步”，即可立即发起一次任务运行，数据安全开始订阅消费、启动消费、生成策略并同步Ranger。
- 查看订阅任务详情：在动态脱敏订阅页签，找到需要查看的任务，单击对应任务操作栏中的“详情”即可查看任务详情。

图 12-210 查看任务详情

参考：Kafka 消息格式要求

第三方平台的动态脱敏策略需要发布到Kafka消息队列，且消息格式满足要求，消息模板及参数说明如下所示。

```
{
  "mask_policy_template":
  {
      "create_time":1692839884000  //同步当前时间
      "name":" task1", //动态脱敏策略名，不能与当前已有动态脱敏策略名重复
      "database": "1", //数据库名
      "table": "1", //数据表名
      "column": "1", //字段名
      "column_type":"int", //字段类型
      "data_level": "1级", // 字段密级，非必填
      "algorithm_config": {
        "name": "MASK", //动态脱敏规则名称，支持范围为MASK、MASK_SHOW_LAST_4、
MASK_SHOW_FIRST_4、MASK_HASH、MASK_DATE_SHOW_YEAR、MASK_NULL
        "type": "MASK", //动态脱敏规则类型，均为MASK类型
        "description": "掩盖英文字符和数字", //动态脱敏规则描述
      },
      "datasource_type":"HIVE", //数据源类型，当前仅支持Hive
      "users":"aaa,bbb",  //指定脱敏用户
      "user_groups":"ggg"  //指定脱敏用户组
      "description":{
          "jdbc_url": "hive2://xxx" //自定义描述，用于在失败消息中返回携带
      }
  }
}
```

12.5.4 数据水印

12.5.4.1 嵌入数据水印

本章主要介绍如何进行数据水印嵌入。数据水印一般有如下场景：

- 规范数据外发流程

  实现对企业内部人员数据外发进行有效流程化管理，非授权用户在数据外发前需审批，审批通过后采取数据水印技术生成可外发数据文件。

- 数据版权保护

  通过在关系数据库中嵌入代表所有权的水印信息，可以将数据库与其拥有者联系起来，从而实现数据的版权保护。

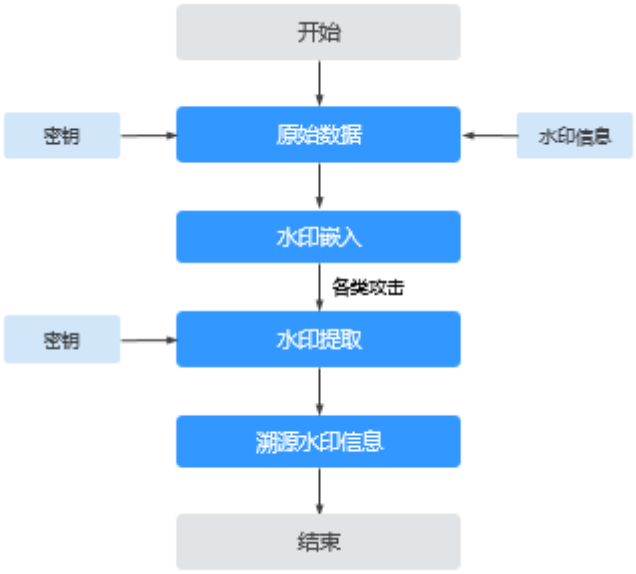
- 对泄露数据进行快速溯源

  通过对泄露数据文件解封，根据数据文件的完整度和水印信息痕迹来检测水印是否存在，快速识别水印标记信息（数据源地址、分发单位、负责人、分发时间等），从而对安全事件精准定位追责。

数据水印使用流程

您可以通过图12-211来了解。

图 12-211 水印使用流程



约束与限制

- 当前数据水印任务仅支持MRS Hive和MRS Doris数据源。
- 主键不支持嵌入水印。
- 数值整型字段嵌入水印可能会出现数据被修改的情况，请选择可以接受值发生改变的字段嵌入水印。
- 数据水印嵌入任务的数据集范围选择为增量时，需选择时间字段类型Timestamp、Date字段类型来确定增量范围。
- MRS Doris数据源仅支持在字符串类型字段嵌入水印，包含Varchar、Text、String等，请确保待嵌入水印的表中包含字符串类型字段
- MRS Doris数据水印任务除了需要MRS Doris数据源，还需要额外准备包含Hadoop、Spark和Yarn组件的MRS集群，用于运行数据水印任务。

前提条件

已创建源端数据源类型为MapReduce服务（MRS Hive）或MRS Doris的数据连接，请参考**创建DataArts Studio数据连接**。

创建数据水印嵌入任务

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“数据水印嵌入”，进入数据水印嵌入页面，在页面上方单击“新建”。

图 12-212 创建数据水印嵌入任务

步骤3 在弹出的创建任务页面输入基本信息，参数配置请参考**表12-50**。

表 12-50 基本信息参数配置

参数名	参数描述
*任务名称	嵌入水印任务的名称，只能包含英文字母、数字、中文字符、下划线或中划线，且长度为1~64个字符。 为便于水印嵌入任务管理，建议名称中标明要嵌入水印的对象和水印标识。
描述	为更好地识别嵌入水印任务，此处加以描述信息。
*水印标识	系统会将水印标识嵌入到数据表中，标识长度不超过16个字符即可。
*纠错等级	等级越高，水印信息编码位数越长，溯源时误码率越低。需注意高纠错等级需要更大的数据量来保证信息的嵌入完整性。默认为1。
*水印版本	V1版本：嵌入水印时依赖主键列，嵌入速度快。若主键遭受强攻击，溯源一定概率失败。 V2版本：嵌入水印时不依赖主键，只与嵌入列相关，嵌入速度慢，鲁棒性增强。

图 12-213 基本信息配置

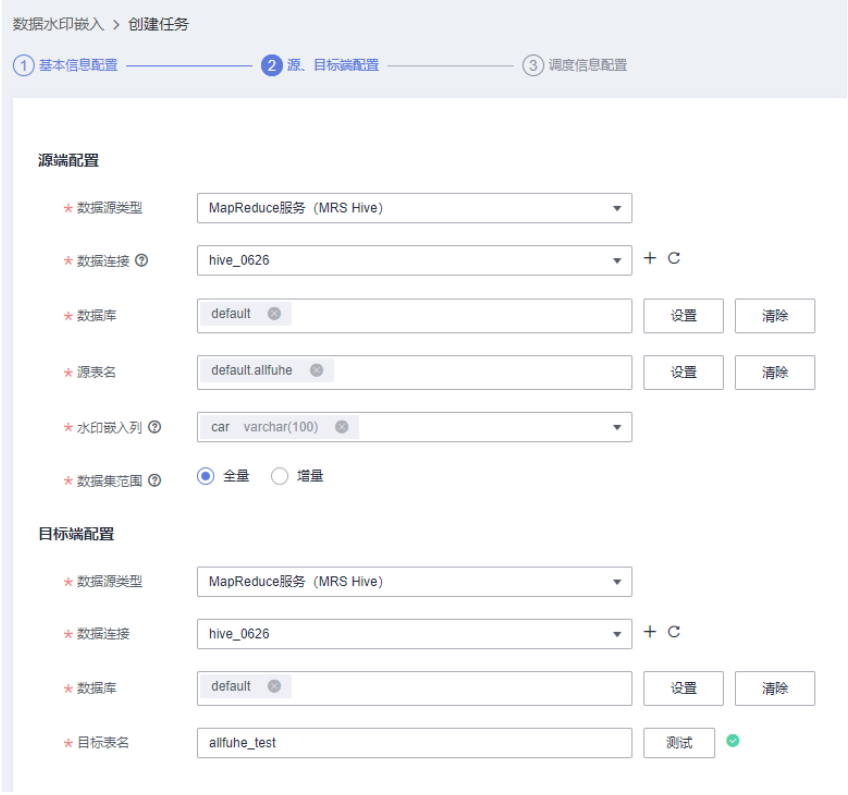
步骤4 单击“下一步”进行源、目标端配置，参数配置请参考**表12-51**。

表 12-51 源、目标端参数配置

参数名	参数描述
源端配置	
*数据源类型	目前只支持MapReduce服务（MRS Hive）。
*数据连接	选择已创建的数据连接。若未创建请参考**创建DataArts Studio数据连接**新建连接。
*Http/Https端口	MRS Doris数据源类型所需参数。获取方式：1.登录DORIS集群的MRS FusionInsight Manager。2.选择“集群 > 服务 > Doris > 配置。3.如果集群开启了Kerberos认证，则在此处填写https_port的值，否则填写http_port的值。
*数据库 *源表名	选择待嵌入水印的数据库和数据表。 • 单击数据库后的“设置”，设置待嵌入水印的数据库和数据表。 • 单击“清除”，可对已选择的数据库和数据表进行修改。
*水印嵌入列	下拉选择常见的字段类型作为嵌入列。如数值型、字符型。 注意：当选择水印版本为V1时，不支持选取主键列作为嵌入列。
*数据集范围	只有使用时间字段timestamp、Date来确定增量范围时，才可以选择增量模式 一般而言，全量模式下数据水印嵌入任务使用单次调度，增量模式下使用周期调度。
*指定时间字段	增量模式下，选择时间字段timestamp、Date来确定增量范围。

参数名	参数描述
目标端配置	
*数据源类型	目前只支持MapReduce服务（MRS Hive）。
*数据连接	选择已创建的数据连接。若未创建请参考**创建DataArts Studio数据连接**新建连接。
*Http/Https端口	MRS Doris数据源类型所需参数。获取方式：1.登录DORIS集群的MRS FusionInsight Manager。2.选择“集群 > 服务 > Doris > 配置。3.如果集群开启了Kerberos认证，则在此处填写https_port的值，否则填写http_port的值。
*数据库	下拉选择存放水印表的数据库。
*目标表名	用户手动输入，不能与目标端数据库表名重复。当输入的表名不存在时会创建该表。 输入请单击“测试”，否则将无法进行下一步操作。

图 12-214 源、目标端配置



步骤5 单击“下一步”，进行调度信息配置。

- 数据集范围为全量模式时，仅支持单次调度。
- 数据集范围为增量模式时，支持单次调度和周期调度。

当选择为周期调度时，参数配置参考**表12-52**。

表 12-52 配置周期调度参数

参数名	说明
*调度日期	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：选择调度开始时间和结束时间，配置间隔的分钟时长。 ● 小时：选择调度开始时间和结束时间，配置间隔的小时时长。 ● 天：配置每日调度时间。 ● 周：选择星期几启动调度，配置调度具体时间。 ● 月：选择几号启动调度，配置调度具体时间。 例如：选择调度周期是周，选择具体时间为15:52，时间选择为星期二。则在调度日期范围内，每周二的15点52分会执行任务。
立即启动	勾选复选框，则表示立即启动此调度任务。

图 12-215 调度信息配置

步骤6 单击“确定”，完成数据水印嵌入任务创建。

----结束

相关操作

- 编辑任务：在数据水印嵌入页面，单击对应任务操作栏中的“编辑”，即可编辑数据水印嵌入任务。

  运行状态为正在“执行中”的任务不允许被编辑。
- 删除任务：在数据水印嵌入页面，单击对应任务操作栏中的“更多 > 删除”，即可删除任务。当需要批量删除时，可以在勾选任务后，在任务列表上方单击“批量删除”。

  运行状态为正在“执行中”的任务不允许被删除。

说明

删除操作无法撤销，请谨慎操作。

- 运行或调度任务：在数据水印嵌入页面，单击对应任务操作栏中的“运行”或“更多 > 启动调度”，运行或调度任务。

  您可以通过调度周期区分该任务是单次调度还是周期调度任务。

- 查看运行实例日志：在数据水印嵌入页面，找到需要查看实例的任务，单击 展开，即可找到运行实例。随后单击“查看日志”，查看运行实例日志。

  运行失败可通过日志排查失败原因，问题修正后尝试重新运行。如果仍运行失败，请联系技术支持人员协助处理。

12.5.4.2 溯源数据水印

本章主要介绍如何利用泄露的数据文件进行水印溯源。

数据溯源主要用来对泄露数据进行快速溯源。通过对泄露数据文件的完整度和水印信息痕迹来检测水印是否存在，快速识别水印标记信息，从而对安全事件精准定位追责。

前提条件

- 用户获得泄露的数据文件后，生成字符分隔值（Comma-Separated Values，CSV）格式文件，文件大小不超过20M，并保存到本地。
- 已完成数据水印嵌入任务，请参考**嵌入数据水印**。

约束与限制

- 数据水印溯源的源文件大小不能超过20MB。
- 为实现准确溯源，请确保数据的完整性以及正确性：数据水印溯源的表数据文件第一列不允许为空，表数据记录数建议在5000以上。

创建数据水印溯源任务

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“数据水印溯源”，进入数据水印溯源页面，在页面上方单击“新建”。

图 12-216 创建数据水印溯源任务

步骤3 在弹出的创建任务页面输入信息，参数配置请参考**表12-53**。

图 12-217 创建数据水印溯源任务

表 12-53 水印溯源任务参数描述

参数名	参数描述
任务名称	嵌入水印任务的名称，只能包含英文字母、数字、中文字符、下划线或中划线，且长度为1~64个字符。
描述	为更好地识别嵌入水印任务，此处加以描述信息。长度不能超过1024个字符。

参数名	参数描述
源文件	得到泄露的数据文件后，利用其生成CSV格式文件，注意文件大小不超过20MB。
字段分隔符	根据上传的CSV文件，下拉选择分隔符，支持四种“,”、“Tab”、“|”、“;”。默认选择“,”。

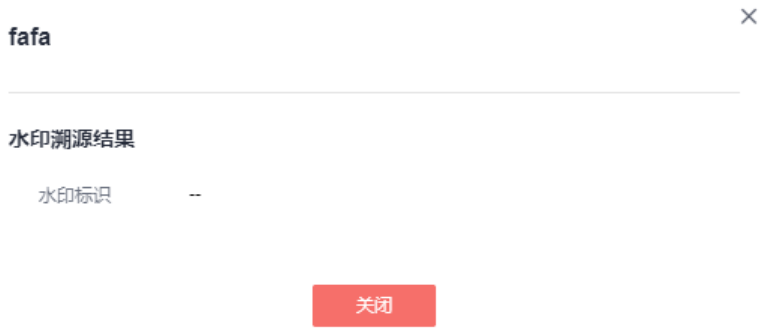
步骤4 单击“运行”，完成创建数据水印溯源任务。

----结束

相关操作

- 查看溯源结果：在数据水印溯源页面，找到需要查看溯源结果的任务，单击对应任务操作栏中的“查看结果”，即可查看溯源结果。注意，只有溯源成功的任务才会显示溯源信息。

  图 12-218 溯源信息

- 删除任务：在数据水印溯源页面，单击对应任务操作栏中的“删除”，即可删除任务。当需要批量删除时，可以在勾选任务后，在任务列表上方单击“批量删除”。

  运行状态为正在“执行中”的任务不允许被删除。

  说明

  删除操作无法撤销，请谨慎操作。

12.5.5 文件水印

本章主要介绍如何进行文件水印相关操作。

- 对结构化数据文件（csv、xml和json）注入暗水印，水印内容不可见，需要进行水印提取。
- 对非结构化数据文件（docx、pptx、xlsx和pdf）注入明水印，可在本地打开文件，查看水印内容。

约束与限制

- 结构化数据文件暗水印的注入和提取时，需限制文件大小在4MB之内。

- 非结构化数据文件明水印在注入时，需限制文件大小在20MB之内。
- 不支持为已注入水印的文件再次注入水印。
- 结构化数据文件水印嵌入的数据有以下要求：
  - 待嵌入水印的源数据需要大于等于5000行。小于5000行的源数据有可能因为特征不够导致提取水印失败。
  - 尽量选取数据取值比较多样的列嵌入水印，如果该列的值是可枚举穷尽的，则有可能因为特征不够导致提取失败。常见的适合嵌入水印的列如地址、姓名、UUID、金额、总数等。
  - 数值整型字段嵌入水印可能会出现数据被修改的情况，请选择可以接受值发生改变的字段插入水印。
- 结构化数据文件的水印提取与数据水印的水印溯源任务无关。仅支持同一账号下用户对已通过**水印注入**或**动态水印**注入水印后的结构化数据文件进行水印提取。

水印注入

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“文件水印”，默认进入水印注入页面。

图 12-219 进入水印注入页面

步骤3 在水印注入页面输入基本信息，参数配置请参考**表12-54**。

表 12-54 水印注入参数配置

参数名	参数描述
*数据类型	选择文件类型。 ● 结构化数据（csv、xml和json）。支持注入暗水印，水印内容不可见，需要进行水印提取。 ● 非结构化数据（docx、pptx、xlsx和pdf）。支持注入明水印，可在本地打开水印文件查看效果。
结构化数据	
*水印内容	系统会将水印标识嵌入到数据表中，标识长度不超过16个字符即可。
*数据选择	结构化数据仅支持csv、xml和json格式文件。
*选择注入字段	选择需要注入水印的字段。
非结构化数据	
*水印内容	系统会将水印标识嵌入到数据表中，标识长度不超过16个字符即可。
透明度	选择明文水印标识的透明度。
旋转角度	选择明文水印标识的旋转角度。
字体大小	选择明文水印标识的字体大小。
*数据选择	非结构化数据仅支持docx、pptx、xlsx和pdf格式文件。

步骤4 单击“注入水印”，完成文件水印注入，浏览器自动下载注入后的文件。

单击“重置”可重置配置参数至默认状态。

----结束

水印提取

当前仅支持对已通过**水印注入**注入暗水印的结构化数据文件（csv、xml和json）进行水印提取。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“文件水印”，然后选择“水印提取”，进入水印提取页面。

图 12-220 进入水印提取页面

步骤3 在水印提取页面输入基本信息，参数配置请参考**表12-55**。

表 12-55 水印提取参数配置

参数名	参数描述
*数据类型	选择文件类型，当前仅支持结构化数据（csv、xml和json）。 结构化数据文件类型支持注入暗水印，水印内容不可见，需要进行水印提取。
*水印内容	无需填写，执行提取水印后会显示提取到的水印信息。
*数据选择	选择已通过**水印注入**注入暗水印的结构化数据文件（csv、xml和json）。

步骤4 单击“提取水印”，完成文件水印提取，水印内容参数展示提取后的水印内容。

单击“重置”可重置配置参数至默认状态。

----结束

12.5.6 动态水印

动态水印指在数据的访问过程中，动态地在数据的查询访问请求返回结果集中注入水印的方式。本章主要介绍如何实现数据开发动态水印功能，最终在数据开发组件中转储或下载敏感数据时，系统动态注入数据水印。

在数据安全组件开启数据开发动态水印功能并创建动态水印策略后，当策略中指定的用户组或角色在数据开发组件中转储或下载敏感数据时，数据开发组件会为敏感数据注入暗水印，保护敏感数据不被泄露。

说明

暗水印内容为获取敏感数据用户的“IAM用户ID”前16位。用户ID可以参考如下步骤进行获取：

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“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。

值得注意的是，动态水印策略为DataArts Studio实例级别配置，各工作空间之间数据互通，全局可见并生效。

前提条件

已创建MRS Hive连接或MRS Spark连接。

约束与限制

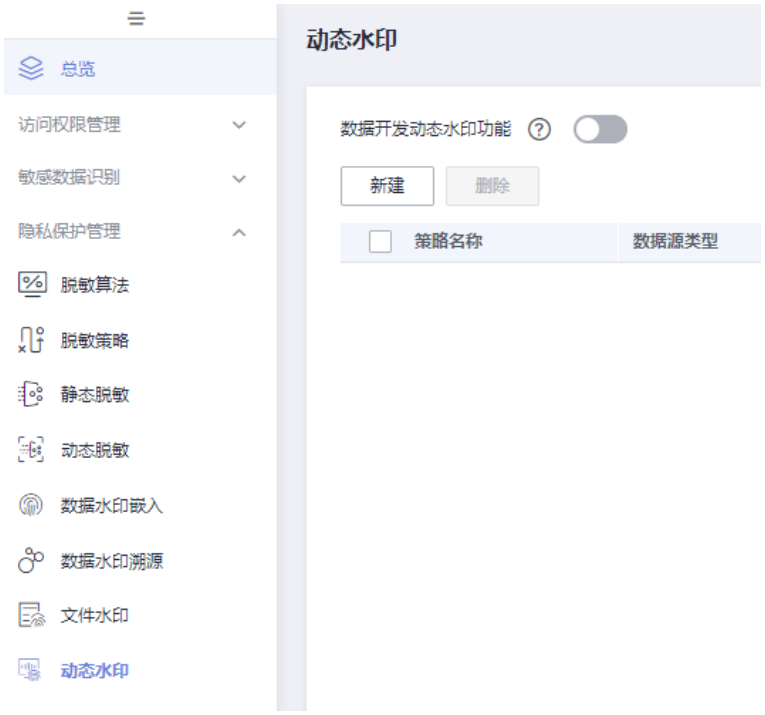
- 仅DAYU Administrator、Tenant Administrator用户或者数据安全管理员可以开启或关闭数据开发动态水印功能，至少为工作空间管理员角色才可以创建动态水印策略，其他普通用户无权限操作。
- 当前动态水印策略仅支持MRS Hive和MRS Spark数据源。
- 新增、删除或修改动态水印策略后，需要约5分钟后才能生效。
- 仅当转储或下载数据量大于500行时，系统才会进行水印嵌入。如果数量小于等于500行，即使嵌入水印后也难以溯源。

创建动态水印策略

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“动态水印”，进入动态水印页面。

图 12-221 进入动态水印页面



步骤3 单击  ，开启数据开发动态水印功能。然后单击“新建”，进入新建动态水印策略页面，参数配置参考**表12-56**。

图 12-222 新建动态水印策略参数配置

创建动态水印策略参数配置说明：

表 12-56 配置策略参数

参数名	参数说明
*策略名称	动态水印策略的唯一标识，DataArts Studio实例内的名称唯一。 为便于策略管理，建议名称中标明要添加水印的对象和水印内容。
*用户组/角色	指定当前工作空间成员中的用户、用户组或角色。当指定对象在数据开发组件中查询或导出敏感数据时，系统会对敏感数据添加动态水印，保护敏感数据不被泄露。
*数据源类型	从下拉列表中选择MRS Hive数据源或MRS Spark数据源。
*数据连接	从下拉列表中选择数据连接类型中已创建的数据连接，若未创建请参考**创建DataArts Studio数据连接**新建连接。
*集群名称	无需选择，自动匹配数据连接中的数据源集群。
*数据库	选择敏感数据所在的数据库。
*数据表	选择敏感数据所在的数据表。

步骤4 单击“确定”，完成动态水印策略创建。

----结束

相关操作

- 水印提取：获得从数据开发下载的动态水印CSV数据文件后，参考**水印提取**进行水印溯源。
- 配置策略：在动态水印页面，单击对应任务操作栏中的“配置”，即可配置动态水印策略。
- 编辑策略状态：新增的水印策略默认为启用状态。当水印策略为关闭状态时，表示该策略将不生效。

  需要修改水印策略状态时，在动态水印页面单击对应水印策略中的  或  ，即可启用或关闭水印策略。
- 删除策略：在动态水印页面，单击对应任务操作栏中的“删除”，即可删除策略。当需要批量删除时，可以在勾选策略后，在列表上方单击“删除”。

  说明

  删除操作无法撤销，请谨慎操作。
- 查看策略详情：在动态水印页面，找到需要查看的策略，单击策略名即可查看策略详情。

图 12-223 查看策略详情

12.6 数据安全运营

12.6.1 审计数据访问日志

数据安全提供DWS、HIVE和DLI数据源上详细的数据操作日志记录，包括时间、用户、操作对象、操作类型等信息。通过这些日志，可以快速进行数据操作审计，更好地做到数据安全管控。

前提条件

- 为实现MRS Hive数据源的数据访问审计，需要满足如下条件：
  - MRS Hive数据连接中选择Agent代理的CDM集群为2.10.0.300及以上版本。
  - MRS Hive数据连接中的用户账号需要同时满足如下条件：
    - 需要配置至少具备Cluster资源管理权限的角色（可直接配置为默认的Manager_operator角色）。
    - 需要配置hive用户组。
- 为实现DWS数据源的数据访问审计，需要满足如下条件：
  - 已开启DWS集群的审计功能开关audit_enabled。

    审计功能开关默认开启，如果已关闭则请参考**修改数据库参数**章节将audit_enabled设置为ON。
  - 已开启需要审计的审计项。

    DWS各类审计项及其开启方法，请参考**设置数据库审计日志**章节。
  - 对于DWS数据源，未开启三权分立时，默认拥有SYSADMIN属性的用户可以查看审计记录；如果开启了三权分立，则只有拥有AUDITADMIN属性的用户才可以查看审计记录。因此需要保证数据连接中的账号或当前用户账号拥有上述权限（未开启细粒度认证前，使用数据连接上的账号查看审计记录；如果开启了细粒度认证，则使用当前IAM用户身份查看审计记录）。

约束与限制

- 对于DWS数据源，数据访问审计需要手动开启DWS集群的审计功能开关和审计项。另外当未开启三权分立时，默认拥有SYSADMIN属性的用户可以查看审计记录；如果开启了三权分立，则只有拥有AUDITADMIN属性的用户才可以查看审计记录，因此需要保证数据连接中的账号或当前用户账号拥有上述权限（未开启细粒度认证前，使用数据连接上的账号查看审计记录；如果开启了细粒度认证，则使用当前IAM用户身份查看审计记录）。
- 对于MRS数据源，查看审计数据依赖于数据连接中Agent的版本，请确保CDM集群为2.10.0.300及以上版本。且MRS Hive数据连接中的用户账号需要同时满足如下条件：
  - 需要配置至少具备Cluster资源管理权限的角色（可直接配置为默认的Manager_operator角色）。
  - 需要配置hive用户组。

查看数据访问日志

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“数据访问审计”，进入审计日志页面。

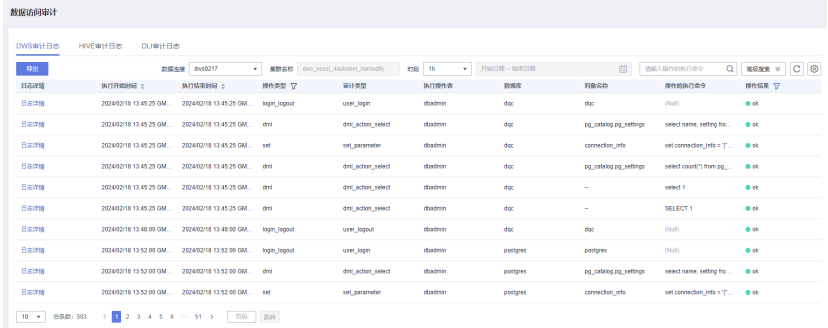
图 12-224 数据访问审计

步骤3 您可以通过切换页签，查看不同数据源的审计日志。日志范围默认1小时，支持自定义时段查询，自定义时段时的最大查询时间间隔为一个月。

- DWS审计日志：日志列表默认使用最新DWS数据连接。单击查看日志详情，可查看当前日志的全量信息。

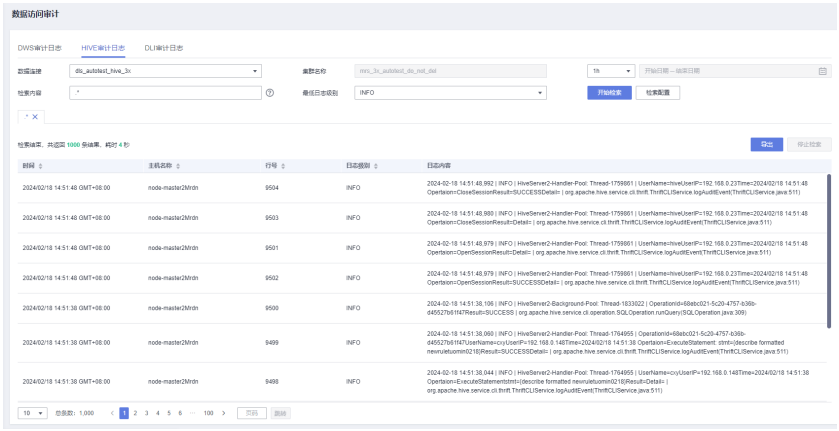
  DWS审计日志支持导出，单击“导出”后，会下载当前页的json数据。

图 12-225 DWS 审计日志列表



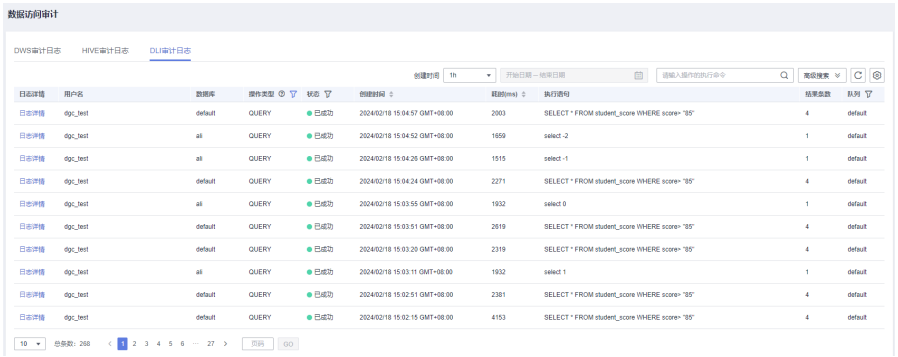
- MRS Hive审计日志：MRS Hive日志列表默认不展示日志内容，而是支持根据配置条件进行检索，检索结果按照页签呈现，支持展示最多5个检索结果页签。

图 12-226 MRS Hive 审计日志列表



- DLI审计日志：DLI日志列表默认展示日志信息。单击日志名查看日志详情，可查看当前日志的全量信息。

图 12-227 DLI 审计日志列表



----结束

12.6.2 诊断数据安全风险

数据安全诊断能够对数据安全能力进行全面诊断，并根据诊断结果，给出修复建议及解决方案。帮助您快速建立起基本数据安全体系，保障数据使用过程的安全可靠。

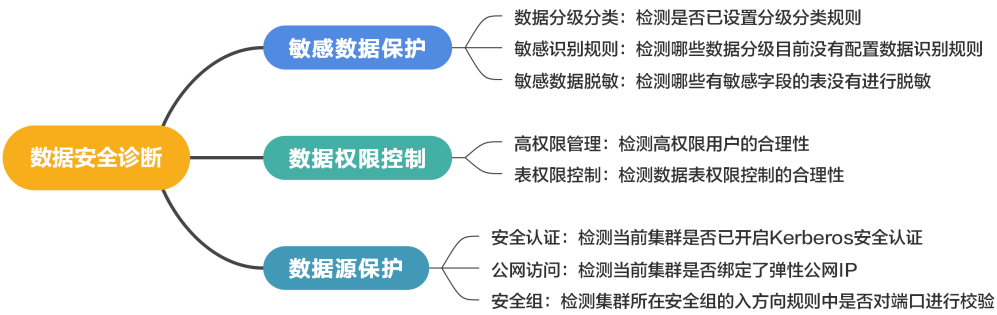
约束与限制

- 当前仅支持MRS数据源的安全诊断能力。
- 安全诊断的扫描任务超时时间为1小时。
- 数据权限控制诊断项，空间管理员与安全管理员仅统计用户，不统计用户组成员。

诊断数据安全风险

数据安全诊断当前支持敏感数据保护、数据权限控制和数据源保护三大诊断项，诊断详情如图12-228所示。

图 12-228 数据安全诊断详情

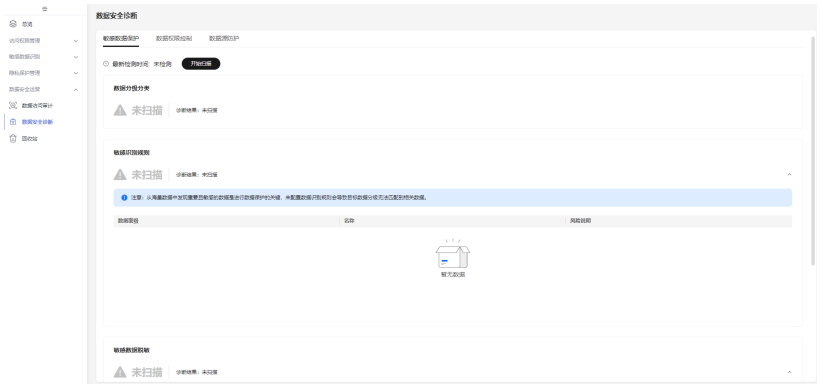


数据安全风险诊断的操作步骤如下，请您根据需要定期扫描处理，建议至少每月进行一次扫描，以保障数据使用过程的安全可靠。

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“数据安全诊断”，进入数据安全诊断页面。

图 12-229 数据安全诊断

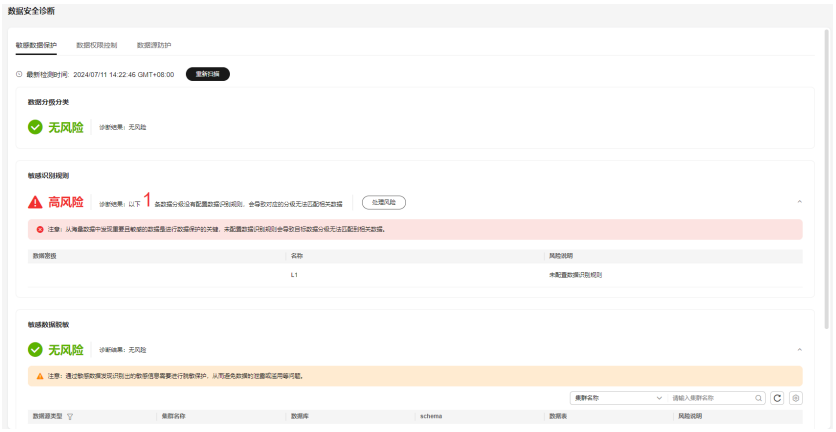


步骤3 根据需要，选择敏感数据保护、数据权限控制或数据源保护页签，单击“开始扫描”或“重新扫描”，进行安全诊断。

步骤4 扫描结束后，请您根据安全扫描结果和处理建议，识别风险项并单击“处理风险”进行优化，保障数据使用过程的安全可靠。

另外，中风险及高风险等级的风险问题属于潜在的安全隐患，建议您尽快处理。下图以敏感数据保护为例查看该检查项目前的风险等级及诊断结果。

图 12-230 安全诊断结果



----结束

12.6.3 查看表权限的拥有者（表权限视图）（高级特性）

数据安全支持权限清单查看，通过表名展示当前实例下拥有表权限的工作空间用户、用户组和角色（包含空间权限集、权限集和角色）。

说明

在新版本模式下仅当使用企业版时，才支持表权限视图。旧版本模式使用基础版及更高版本时即可支持。

约束与限制

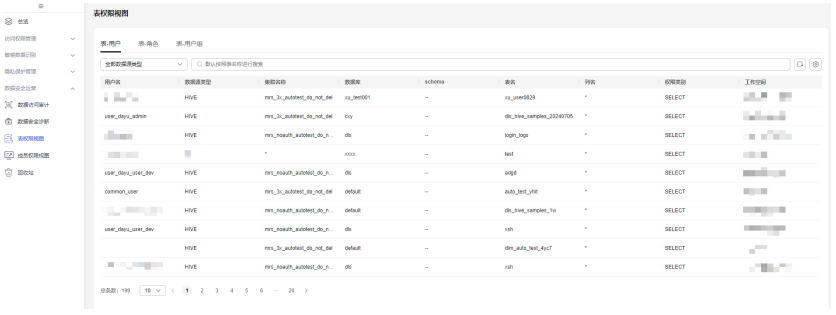
- 表-角色页签暂不支持展示存算分离MRS Hive的URL权限策略。
- 当前暂不支持在表权限视图页面直接对权限进行配置、回收。

查看表权限的拥有者

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“表权限视图”，进入表权限视图页面。

图 12-231 表权限视图

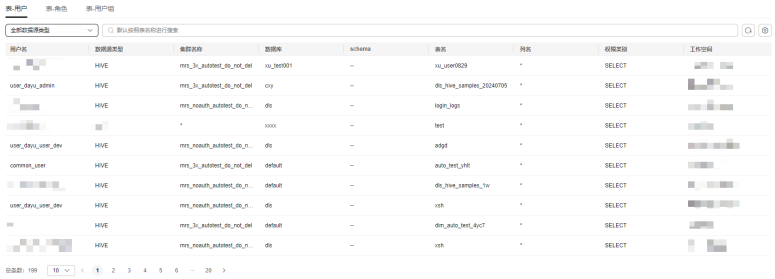


步骤3 在表权限视图页面，您可以通过切换页签，查看表权限的不同拥有对象：

- “表-用户”页签：默认展示当前实例下，通过授权对象为用户的权限申请和审批流程所获取的表权限。支持筛选不同的数据源类型，并通过用户名、集群名称、数据库或表名检索。

  权限申请和审批流程详见**申请与审批权限（部分高级特性）**。

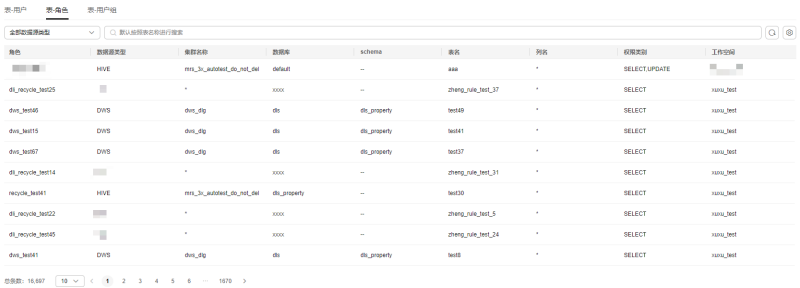
  图 12-232 表-用户

- “表-角色”页签：默认展示当前实例下，在角色（包含空间权限集、权限集和角色）中所授予的表权限。支持筛选不同的数据源类型，并通过角色、集群名称、数据库或表名检索。

  通过空间权限集、权限集或角色授权的流程详见**配置空间权限集**、**配置权限集**或**配置角色**。

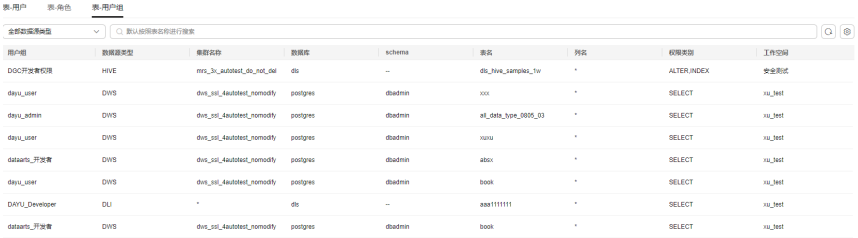
  图 12-233 表-角色

- “表-用户组”页签：默认展示当前实例下，通过授权对象为用户组的权限申请和审批流程所获取的表权限。支持筛选不同的数据源类型，并通过用户组、集群名称、数据库或表名检索。

  权限申请和审批流程详见**申请与审批权限（部分高级特性）**。

  图 12-234 表-用户组

----结束

12.6.4 查看用户的权限（成员权限视图）（高级特性）

数据安全支持权限清单查看，以当前实例下某工作空间的用户或用户组，查看其通过角色（包含空间权限集、权限集和角色）或权限申请和审批流程所获取的权限。

说明

在新版本模式下仅当使用企业版时，才支持成员权限视图。旧版本模式使用基础版及更高版本时即可支持。

约束与限制

- 查看用户权限时，不展示其继承自用户组的权限。
- 当前暂不支持在成员权限视图页面直接对权限进行配置、回收。

查看数据访问日志

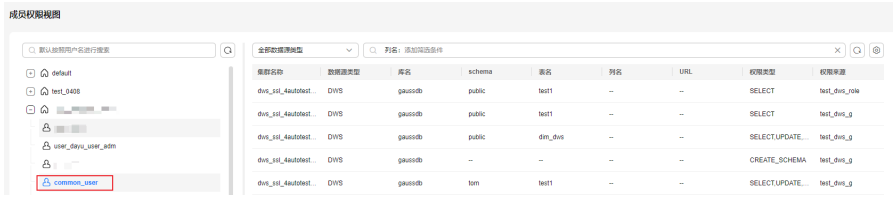
步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“成员权限视图”，进入成员权限视图页面。

图 12-235 成员权限视图

步骤3 在成员权限视图页面，通过在左侧单击选择某工作空间的用户或用户组（支持通过工作空间、用户或用户组进行筛选），系统默认展示其通过角色（包含空间权限集、权限集和角色）或权限申请和审批流程所获取的权限。在权限结果中，支持筛选不同的数据源类型，并通过集群名称、库名、schema、表名或列名检索。

图 12-236 查看用户权限



----结束

12.7 管理回收站

通过回收站功能，您可以恢复误删的数据安全关键数据。当前综合数据的重要程度、使用频次以及误删后恢复难易程度等各方面因素考虑，定义数据安全的关键数据为权限集（包含空间权限集、权限集以及通用角色）、动态脱敏策略和密钥。

前提条件

在30天内对权限集（包含空间权限集、权限集以及通用角色）、动态脱敏策略和密钥进行过删除操作。

约束与限制

- 仅DAYU Administrator、Tenant Administrator或者数据安全管理员可以执行还原操作，其他普通用户无权限操作。
- 由于MRS纳管角色是继承的MRS数据源已有角色，非定义的数据安全数据，因此删除的纳管角色数据不会进入回收站。
- 权限集和动态脱敏策略被删除进入回收站后，将同步状态将统一置为未同步，从回收站还原后也需要手动进行同步才能生效。
- 回收站中的数据最多保存30天，删除时间超过30天的数据将被自动清理。
- 单实例下回收站中的权限集、动态脱敏策略和密钥分别最多保存1000条数据，超过1000条后会自动清理更早删除的数据。
- 数据还原操作时，如果“同名处理方式”参数配置为“名称添加时间戳”，则如果同名会在还原数据的原名称后添加时间戳信息（***原名称_13位时间戳***）。如果添加时间戳后总长度超过64，会对原名称进行截断操作，确保总长度不会超出64的限制。
- 从回收站还原被误删的权限集时，会校验权限集之间的关联关系，若不满足则无法还原。例如某权限集的父权限集已被删除，则其无法直接还原，必须先还原父权限集。
- 批量还原时，每次最多还原20条数据。

还原回收站数据

步骤1 在DataArts Studio控制台首页，选择对应工作空间的“数据安全”模块，进入数据安全页面。

步骤2 单击左侧导航树中的“回收站”，进入回收站页面。

图 12-237 进入回收站页面

步骤3 在回收站页面，您可以通过切换页签，查看并还原已删除的权限集（包含空间权限集、权限集以及通用角色）、动态脱敏策略数据或密钥。

不同数据还原时的操作基本一致，后续步骤以还原权限集数据为例，为您介绍如何还原数据。

步骤4 在权限集页签，找到待还原的权限集，单击列表操作栏中的“还原”进行数据还原。或者勾选待还原的权限集，单击列表上方的“还原”，进行批量还原。

图 12-238 数据还原

步骤5 在弹出的确认窗口中，您需要选择同名处理方式，以避免还原的数据与现有数据冲突。然后单击“确定”，完成数据的还原。

- 报错：如果有重名数据，则报错且不还原重名的数据。
- 名称添加时间戳：如果有重名数据，则在还原数据的原名称后添加时间戳信息（***原名称_13位时间戳***）。如果添加时间戳后总长度超过64，会对原名称进行截断操作，确保总长度不会超出64的限制。

图 12-239 选择同名处理方式

步骤6 完成还原后，您可以在相应空间权限集、权限集、通用角色或动态脱敏策略的入口，检查还原后的数据，并手动进行同步，以确保还原后的数据生效。

----结束

13 数据服务

13.1 数据服务简介

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。

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. **购买并管理专享版集群**

   如果您需要使用数据服务，需要先购买专享版集群。

2. **新建数据服务审核人**

   在创建API前，需要新建数据服务审核人。

3. **创建API**

   创建API即**生成API**。其中，生成API支持两种方式（**配置方式生成API**和**脚本/MyBatis方式生成API**）。

4. **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。

5. **发布API**

   只有将API发布后，API才支持被调用。

6. **管理API**

   您可以根据您的需要，对已创建发布的API进行管理。

7. **编排API**

   编排API是将已经开发好的服务API接口，在无需编写复杂代码的情况下，根据特定的业务逻辑和流程进行可视化的重组和重构，从而实现在不影响原生接口的前提下进行简便的二次开发。

8. **（可选）配置流控策略**

   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。

9. **（可选）主动授权API**

   应用定义了一个API调用者的身份。对于使用APP或IAM认证方式的API，必须在API授权后，才能获得认证信息以用于API调用。

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. **申请API授权**

   对于API调用者而言，如果API开发者未授权APP或IAM认证方式的API，则需要自行申请API授权，等待审批通过后才能进行API调用。

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和APP的视角，统计了相关度量数据。

图 13-1 API 视角数据统计

表 13-1 API 视角数据统计

分类	指标	说明
数据总览区	已发布	统计已成功发布的API数量。
	开发中	统计开发中的API数量。
	申请者	统计已发布API所授权的应用数量。
	总调用	近7天（不含当天）所有集群下API的调用总次数。
	成功	统计API调用成功的次数。
	失败	统计API调用失败的次数。
	合法	统计API合法调用的总次数，合法调用指校验通过的调用。
	非法	统计API非法调用的总次数，非法调用指由于请求参数填写错误等原因导致的校验不通过的调用。
趋势图	调用趋势	展示所选时间维度下，集群维度的API调用次数曲线。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 集群维度：单集群，所有集群 ● 调用次数：调用总次数、成功次数/失败次数、合法次数/非法次数
	发布趋势	展示所选时间维度下，API发布次数曲线。 ● 时间维度：今日、本周、本月、今年。
TOP5统计	调用比率TOP5	统计所选时间维度下，按照集群维度的API调用比率，排序出TOP5 API。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 集群维度：单集群，所有集群 ● 比率：成功率、失败率、合法率、非法率
	调用时间TOP5	统计所选时间维度下，按照集群维度的API调用时长，排序出TOP5 API。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 集群维度：单集群，所有集群 ● 时长：时长平均总值、成功时长平均总值，失败时长平均总值
	总调用次数TOP5	统计所选时间维度下，按照集群维度的API调用次数（同一API授权不同应用则合并计数），排序出TOP5 API。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 集群维度：单集群，所有集群 ● 调用次数：调用总次数、成功次数、失败次数、合法次数和非法次数。

分类	指标	说明
	单API调用次数TOP5	统计所选时间维度下，按照集群维度的API调用次数（同一API授权不同应用则分开计数），排序出TOP5 API。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 集群维度：单集群，所有集群 ● 调用次数：调用总次数、成功次数、失败次数、合法次数和非法次数。

图 13-2 APP 视角数据统计

表 13-2 APP 视角数据统计

分类	指标	说明
数据总览区	已申请	统计所有API授权的APP数量。
	总调用	近7天（不含当天）所有集群下APP和IAM证方式API的调用总次数。
	成功	统计APP和IAM认证方式API调用成功的次数。
	失败	统计APP和IAM认证方式API调用失败的次数。
	合法	统计APP和IAM认证方式API合法调用的总次数，合法调用指校验通过的调用。
	非法	统计APP和IAM认证方式API非法调用的总次数，非法调用指由于请求参数填写错误等原因导致的校验不通过的调用。

分类	指标	说明
趋势图	调用趋势	展示所选时间维度下，所有集群的APP和IAM认证方式API调用次数曲线。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 调用次数：调用总次数、成功次数/失败次数、合法次数/非法次数
TOP5统计	调用比率TOP5	统计所选时间维度下，所有集群的APP和IAM认证方式API调用比率，排序出TOP5 API。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 比率：成功率、失败率、合法率、非法率
	调用时间TOP5	统计所选时间维度下，所有集群的APP和IAM认证方式API调用时长，排序出TOP5 API。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 时长：时长平均总值、成功时长平均总值，失败时长平均总值
	总调用次数TOP5	统计所选时间维度下，所有集群的APP和IAM认证方式API调用次数（同一API授权不同应用则合并计数），排序出TOP5 API。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 调用次数：调用总次数、成功次数、失败次数、合法次数和非法次数。
	单APP调用次数TOP5	统计所选时间维度下，所有集群的APP和IAM认证方式API调用次数（同一API授权不同应用则分别计数），排序出TOP5 API。 ● 时间维度：近12小时，近1天，近7天，近 30天 ● 调用次数：调用总次数、成功次数、失败次数、合法次数和非法次数。

13.2 规格说明

专享版规格

数据服务专享版的实例规格，如表13-3所示。

表 13-3 专享版实例规格说明

实例规格	最大支持发布的API数量	延时（单位：ms）
小规格	500	<20
中规格	1000	<15
大规格	2000	<10

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 13-4 API 的返回数据条数限制

API分类	使用场景	数据源	默认规格（条）
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

13.3 开发数据服务 API

13.3.1 购买并管理专享版集群

本小节指导您顺利购买专享版实例，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。

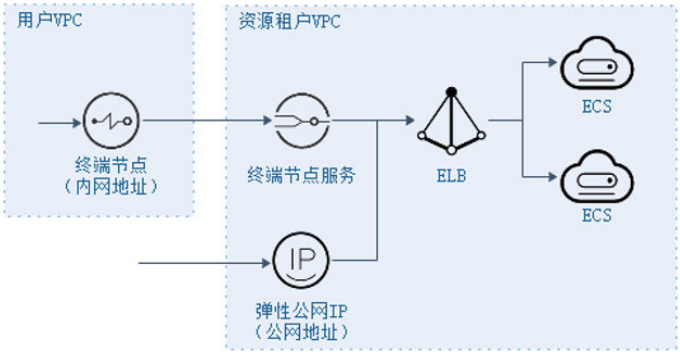
网络环境准备

如图13-3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。

专享版集群创建后，您可以通过如下途径访问集群API：

- 内网地址：内网地址为用户VPC内的终端节点IP地址，默认具备。
- 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。
- 内网域名（可选）：内网域名是指在VPC中生效的域名。创建集群后可通过“绑定内网域名”，输入自定义内网域名，数据服务调用DNS服务将内网域名与内网地址相关联。
- 公网域名（可选）：公网域名是在Internet中使用公网解析的域名。创建集群后可通过“绑定公网域名”，输入已完成注册的域名，数据服务调用DNS服务将公网域名与外网地址相关联。

图 13-3 专享版集群网络架构说明



因此，为了保证专享版集群API能够被用户访问，集群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。

  在购买时专享版实例时，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，单击配置路由项的“新建”按钮，新增物理机的IP地址，如图13-4所示。

  图 13-4 基本信息

操作步骤

购买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“购买增量包”。

步骤2 进入购买DataArts Studio增量包页面，参见表13-5进行配置。

表 13-5 购买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。
计费方式	实例收费方式，当前支持“包年包月”。
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。集群购买成功后，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。 如果需要在其他工作空间内使用该集群，您可以在集群创建成功后，参考管理集群共享将该集群共享给其他工作空间。

参数项	说明
可用区	选择数据服务专享集群所在的可用区。 支持单AZ和多AZ两种部署方式。推荐使用多AZ方式。 ● 单AZ：仅可以选择1个AZ，集群节点部署在同一AZ上。 ● 多AZ：可选择2-10个AZ，集群节点部署在不同AZ上，以提升集群的容灾能力。 详情请参见**什么是可用区**。
集群名称	集群名称必须以字母开头,可以包含字母、数字、中划线或者下划线,不能包含其他的特殊字符。输入长度不能小于5个字符。
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数量的支持能力不同。
公网入口	开启“公网入口”，创建集群时会为集群自动绑定一个新建的弹性公网IP，后续可以通过此公网IP地址调用专享版API。该功能新建的弹性公网IP不会计入收费项。 如果您存在需要本地调用或跨网调用API的使用场景，建议开启。如果在创建集群时未开启公网入口，后续则不再支持绑定EIP。
带宽大小	可配置公网带宽范围。
虚拟私有云 子网 安全组	DataArts Studio实例中的数据服务专享版集群所属的VPC、子网、安全组。 在相同VPC、子网、安全组中的云服务资源（如ECS），可以使用数据服务专享版实例的私有地址调用API。建议将专享版集群和您的其他关联业务配置一个相同的VPC、子网、安全组，确保网络安全的同时，方便网络配置。 VPC、子网、安全组的详细操作，请参见《**虚拟私有云用户指南**》。 **说明** ● 目前专享版集群创建完成后不支持切换VPC、子网、安全组，请谨慎选择。 ● 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。 ● 此处支持选择共享VPC子网，即由VPC的所有者将VPC内的子网共享给当前账号，由当前账号在购买数据服务专享版集群时选择共享VPC子网。通过共享VPC子网功能，可以简化网络配置，帮助您统一配置和运维多个账号下的资源，有助于提升资源的管控效率，降低运维成本。如何共享VPC子网，请参考《**共享VPC**》。
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**。
节点数量	-
购买时长	-

步骤3 单击“立即购买”，确认规格后提交。

----结束

管理集群共享

专享版集群创建成功后，默认仅能在绑定的工作空间内使用。如果您需要在其他工作空间使用此集群，则可以进行集群共享，共享后在其他工作空间可查看、使用但不能管理该集群，并能将API发布至该集群。

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

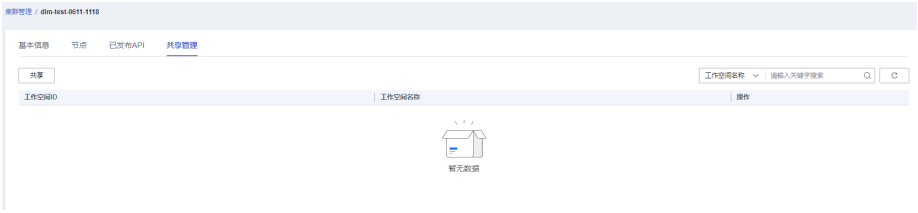
步骤2 在DataArts Studio控制台首页，选择已购买专享版集群的工作空间的“数据服务”模块，进入数据服务页面。

步骤3 在数据服务集群页面单击“集群”，进入集群列表页面。

步骤4 单击集群名称，进入集群详情页面。

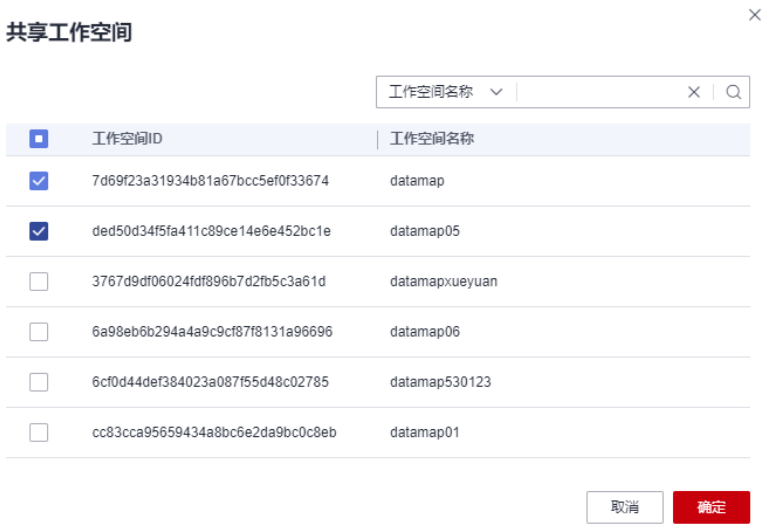
步骤5 在集群详情页面，单击“共享管理”页签，进入共享管理页面。

图 13-5 进入共享管理页面



步骤6 单击“共享”，在弹出的窗口中勾选需要共享的工作空间后，单击“确定”完成集群共享。

图 13-6 选择工作空间



步骤7 对于已共享集群的工作空间，您可以在该工作空间内，正常查看、使用该集群。

如后续需要取消该工作空间的集群共享，则需要先下线该工作空间已在集群上发布的API，再到绑定工作空间的数据服务集群详情页面，取消共享。

----结束

设置 API 分配配额

专享版集群创建成功后，需要为当前工作空间设置API分配配额。

DataArts Studio实例下数据服务专享版的API总分配配额默认为5000，当为工作空间分配配额之后，才能在工作空间下创建相应的API，配额分配参考如下步骤。

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

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

图 13-7 空间信息

步骤3 在“空间信息”中，单击“数据服务专享版API配额”中对应配额的“设置”按钮，对已分配配额进行配置。配置完成后单击“保存”，保存当前配置。

已分配配额表示分配给当前工作空间下可使用的配额。注意，已分配配额不能小于已使用配额，不能大于未分配配额（即总配额-总分配配额）。

说明

数据服务专享版在每个DataArts Studio实例下具有创建10个专享版API免费试用额度，超出试用配额后会产生数据服务专享版API的费用，所创建的超出试用配额API按每天每个进行收费。

图 13-8 设置已分配配额

步骤4 已分配配额设置完成后，单击“空间信息”中的“确定”，完成配置。

----结束

相关操作

- 设置集群日志转储：日志转储功能开启后，集群中当前工作空间下API的所有访问日志，会转储到工作空间指定的OBS桶或者LTS日志中。

  在集群页面单击集群名称，进入基本信息页签。选择打开日志转储功能选择转储方式：

  - 当选择OBS存储，当前工作空间中API的所有访问日志，会转储到工作空间指定的OBS桶。
  - 当选择LTS存储，在选择转储方式前，需要在LTS服务中提前新建日志组和日志流，如何新建日志组和日志流请参考**查看API访问日志**。选择后当前工作空间中API的所有访问日志，会转储到LTS服务新建的日志流中。

- 重启集群：重启集群将影响在该集群上发布的API，导致API无法调用，请谨慎操作！

  在集群页面单击“重启”，可进行重启操作。

- 删除集群：如果当前集群无法满足使用，可以删除集群。注意，删除集群后将无法恢复，请确保相关业务数据已导出备份，并谨慎操作！

  在集群页面单击“更多 > 删除”，可进行删除操作。

- 绑定内网域名：内网域名是指在VPC中生效的域名。绑定内网域名，可以将内网域名与内网地址相关联，然后在内网同一VPC中通过内网域名进行API调用。

  在集群页面单击“更多 > 绑定内网域名”，输入自定义的内网域名，数据服务调用DNS服务将内网域名与内网地址相关联。注意，每个租户在所有项目中支持添加的内网域名总配额为50个。

  自定义的内网域名支持各类域名级别，但需符合域名命名规范。

  - 由以点分割的字符串组成，单个字符串不超过63个字符。
  - 支持字母、数字以及中划线，中划线不能出现在域名的开头或末尾。
  - 域名总长度不超过254个字符。

- 绑定公网域名：公网域名是在Internet中使用公网解析的域名。绑定公网域名，可以将公网域名与外网地址相关联，然后在Internet中通过公网域名进行API调用。在集群页面单击“更多 > 绑定公网域名”，输入已完成域名注册的域名，数据服务调用DNS服务将公网域名与外网地址相关联。注意，绑定公网域名的前提是在集群创建时已开启“公网入口”绑定了弹性公网IP，否则将会绑定失败。另外，每个租户支持添加50个公网域名。

  公网域名支持添加主域名及主域名的子域名，即最多支持添加二级域名，例如abc.example.com。

13.3.2 新建数据服务审核人

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。

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。

说明

审核人支持管理员、开发者、运维者，访客无法添加为审核人。

工作空间管理员角色的用户，无论是否被添加为审核人，都默认具备审核人权限。

操作步骤

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
3. 在左侧导航选择服务版本（例如：专享版），进入总览页。
4. 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 13-9 新建审核人界面

5. 选择审核人（此处的账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。
6. 根据需要，可以添加多个审核人。

13.3.3 创建 API

13.3.3.1 配置方式生成 API

本节介绍如何通过配置方式生成API。

使用配置方式生成数据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。

约束与限制

API生成暂不支持Hive数据源的中文表和中文列场景。

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
3. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
4. 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。
5. 对于已成功创建的API目录，在API目录上右键单击，可选择编辑或删除API目录。

配置 API 基本信息

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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 13-6 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，指定当前API保存的位置，后续可以在指定的API目录中检索当前API。 API目录是数据服务中API的最小组织单元。您可以选择**新建API目录**已创建的目录。

配置	说明
请求Path	API访问路径，例如：/getUserInfo。 请求Path即完整的URL中，域名之后、查询参数之前的部分，如图13-10中的“/blogs/xxxx”。 **图 13-10** 统一资源定位符 URL 说明  在请求Path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，专享版支持HTTPS协议。 推荐选择HTTPS协议，HTTP安全性欠佳，可能会存在安全风险。 • 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。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	审核人拥有API的审核权限。可单击“添加”，进入“审核中心 > 审核人管理”页面，新建审核人。

配置	说明
安全认证	创建API时，有如下三种安全认证方式可选。三种方式的区别在于认证方式和调用方法不同，推荐使用安全性更高的APP认证。 ● APP认证：将APP认证方式的API授权给应用后，使用应用的密钥对（AppKey和AppSecret）进行安全认证，支持通过SDK或API调用工具调用，安全级别高，推荐使用。 ● IAM认证：将IAM认证方式的API授权给当前账号或其他账号后，借助从IAM服务获取的用户Token进行安全认证。支持通过API调用工具调用，安全级别中等。 ● 无认证：不需要认证，所有用户均可访问，建议仅在测试接口时使用，不推荐正式使用。使用无认证方式时，无需鉴权认证信息，安全级别低，通过API调用工具或浏览器即可直接调用。
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“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发布状态预留的最低期限，单位为小时，0表示不设限制。 如果需要停用/下线/解除授权，则停用/下线/解除授权时间必须选择在发布后的最低保留期限时间之后。选择时间后，停用/下线/解除授权会通知已授权用户。如果所有已授权用户均完成审核中心通知列表消息处理，或在应用中解绑与API的绑定关系，API就会停用/下线/解除授权；否则会以待停用/待下线/待解除授权状态，等待达到停用/下线/解除授权时间，再强制停用/下线/解除授权。 例如，最低保留期限设置为24小时，则此API发布后需要停用时，停用时间必须选择在发布24小时后，即发布第二天之后。如果期间内已授权用户已完成审核中心通知列表消息处理或解绑应用与API的绑定关系，则会直接停用；如果未完成，则会以待停用状态等待达到停用时间，强制停用。

配置	说明
入参定义	配置调用API需要输入的参数，此处定义后即为配置取数逻辑时的请求参数。 入参定义主要由参数位置、参数类型、是否必填、允许空值以及默认值等组成。 ● 参数位置主要包括Query、Header、Path、Body四大类，另外还支持Static静态参数。 – Query是位于URL后的查询参数内容，以“?”开始，通过“&”连接多个参数。 – Header参数是位于请求消息头中的参数，常用于传递当前信息。例如host，token等。 – Path是位于请求路径中的请求参数，如果定义了Path参数，则需要在请求Path中也添加此参数，作为请求Path的一部分。 – Body是位于请求体内的参数，一般使用json格式表示。 – Static是不随API调用者的传值变化的静态参数，仅当安全认证为APP认证方式时支持。Static参数数值由API授权时确定（如果授权时未配置参数值，则SDK调用时会使用API入参默认值，API工具调用时会导致缺少Static参数值的报错）。 ● 参数类型分为数值型Number与字符型String两大类。Number参数对应数据库中int、double、long等数值数据类型，String参数对应数据库中char、vachar、text等文本数据类型。 ● 是否必填、允许空值以及默认值。 – 如果设定为必填，则API在访问时，必须传入指定参数。 – 如果设定为非必填，则在API访问时，未传入的参数会使用默认值进行代替；如果未传入参数，也没有默认值，则允许空值时会使用null替换，不允许空值时忽略该参数条件。 **说明** 入参定义中，参数大小限制如下： ● Query+Path，URL最大32KB ● Header，最大128KB ● Body， 最大128KB 实际配置中，需要根据所设计的调用API时请求参数情况来设置入参。例如，在用户表中根据用户ID查询用户信息时，请求Path设置为：/getUserInfo。可按照如下不同场景来配置入参： ● API调用时请求参数为用户id，需要返回对应id的用户信息。 1. 单击“添加”，参数名配置为id。 2. 参数位置选择Query。 3. 类型设置为Number。 4. 是否必填选择必填。 5. 默认值保持默认，无需填写。

配置	说明
	• API调用时请求参数为用户id1和用户id2，需要返回id1-id2范围内的用户信息： 1. 单击“添加”，参数名配置为id1。 2. 参数位置选择Query。 3. 类型设置为Number。 4. 是否必填选择必填。 5. 默认值保持默认，无需填写。 6. 再次单击“添加”，按照id1参数的配置信息再配置id2。

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**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。

   另外，专享版数据服务支持返回总条数，开启后可返回取值脚本执行结果数据的总条数。

   图 13-11 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，选择为基本配置中定义的入参，是用户访问API时直接使用的参数。

- 绑定字段对外不可见，是所选的数据表中的字段，为API调用时实际访问的内容。
- 操作符则是用户访问API时，对绑定字段和绑定参数的处理方式。操作符左边为绑定字段，右边为绑定参数。当前支持的操作符及含义如下：

表 13-7 支持的操作符

操作符	描述
=	检查两个操作数的值是否相等。 如果绑定字段和绑定参数相等则条件为真。 **说明** 对于DWS数据库的FLOAT4、FLOAT8类型参数，不支持比较数值是否相等。
<>	检查两个操作数的值是否相等。 如果绑定字段和绑定参数不相等则条件为真。 **说明** 对于DWS数据库的FLOAT4、FLOAT8类型参数，不支持比较数值是否相等。
>	检查左操作数的值是否大于右操作数的值。 如果绑定字段大于绑定参数，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。 如果绑定字段大于等于绑定参数，则条件为真。
<	检查左操作数的值是否小于右操作数的值。 如果绑定字段小于绑定参数，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。 如果绑定字段小于等于绑定参数，则条件为真。
%like%	%like%表示忽略前后缀，进行字符匹配。 如果绑定字段忽略前后缀，能匹配绑定参数，则条件为真。
%like	%like表示忽略前缀，进行字符匹配。 如果绑定字段忽略前缀，能匹配绑定参数，则条件为真。
like%	like%表示忽略后缀，进行字符匹配。 如果绑定字段忽略后缀，能匹配绑定参数，则条件为真。
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。 如果绑定字段能匹配多个绑定参数中的值，则条件为真。

操作符	描述
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。 如果绑定字段无法匹配多个绑定参数中的值，则条件为真。

值得注意的是，请求参数可以通过复制并设置操作符，实现多个输入的绑定参数条件匹配绑定字段。

以图13-12为例，即为访问API时输入两个入参id1和id2，匹配id1和id2数值范围之间的x列值。

图 13-12 请求参数

4. 编辑返回参数信息。

   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，可自定义，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，是所选的数据表中的字段，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。

   图 13-13 返回参数

5. 编辑排序参数信息。

   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，支持多个排序参数。

   - 参数名可自定义，用于与字段名称关联。
   - 字段名称对外不可见，是所选的数据表中的字段，是API调用时实际访问的内容。
   - 是否可选决定了调用API时此排序参数是否必选，勾选则表示此参数可以不传，可以通过排序参数描述pre_order_by的值配置是否参与排序；不勾选则此参数必传，即使排序参数描述pre_order_by的值未配置此参数，依然会参与排序。
   - 排序方式表示了当前参数允许使用的排序形式，分为升序、降序以及自定义。自定义排序参数默认为升序排序，可通过排序参数描述pre_order_by的值进行调整；而升序或降序的排序参数，不支持通过pre_order_by的值调整排序方式，如果pre_order_by的值与此处设置排序方式不符，则会导致配置调试或调用报错。
   - 多个排序参数时，表示当第一个排序参数相等时，再逐一用后续排序参数去排序。参数的排序顺序不支持通过排序参数描述pre_order_by的值进行调整，如需调整可通过“参数设置”的“添加”进入添加参数界面，调整排序参数勾选顺序，可重新调整排序参数的顺序。

图 13-14 排序参数

6. 单击“下一步”，进行API测试页面。

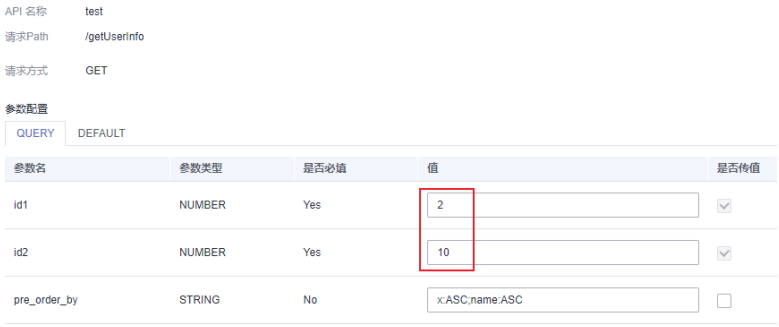
测试 API

1. 填写入参取值。

   如果单个参数需要传多个值时，写法如下：

   - 字符串：'a','b','c'
   - 数值：1,2
   - 字段：a,b,c

图 13-15 填写入参取值



2. （可选）调整排序参数描述pre_order_by的值。

   系统根据5中已配置的所有排序参数已给出pre_order_by的默认值，自定义排序默认为升序。排序参数描述pre_order_by的值填写形式为“***排序参数参数名***:ASC”或“***排序参数参数名***:DESC”，其中ASC表示升序，DESC表示降序，多个排序参数描述以“英文分号”进行分隔。勾选“是否传值”后，测试结果将按照pre_order_by的值排序。

   对于pre_order_by的值，您可以进行如下修改：

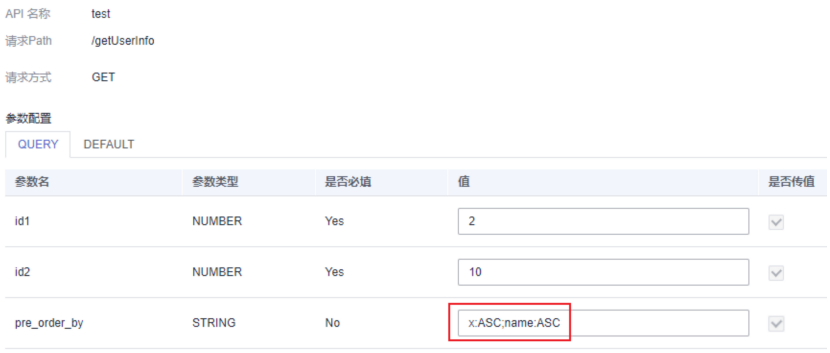
   - 删掉某可选的排序参数，则此排序参数不再参与排序。
   - 修改自定义排序方式的排序参数为升序或降序方式，则此排序参数按照修改后的排序方式排序。

   说明

   pre_order_by的值，不支持进行如下修改，否则会修改不生效或导致调用报错。

   - 删掉某必选的排序参数，则此排序参数依然会正常参与排序，删除不生效。
   - 调整排序参数的前后顺序，则排序依然以配置排序参数时的排序参数顺序为准。调整不生效。
   - 修改升序或降序的排序参数为其他排序方式，则会调用失败，不允许修改。

图 13-16 调整排序参数描述 pre_order_by 的值



3. （可选）调整分页参数值。

   系统会对返回数据进行分页，pageSize表示分页后的页面大小，pageNum表示页码。API调试时默认按100的大小分页，返回第1页数据。

   说明

   API调试时，page_size（系统默认）最大为100，当page_size值大于100时，默认查出的数据仍为100条。

图 13-17 调整分页参数值

4. 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。

   - 测试过程中，如果数据服务API查询及返回数据的总时长超过默认60秒，会报超时错误。
   - 如果测试失败，请查看错误提示并做相应的修改重新测试。

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，单击“编辑”按钮进行修改API的相关操作。

说明

API如果处于发布、下线、停用、恢复的待审核或待执行状态，则不支持编辑。

13.3.3.2 脚本/MyBatis 方式生成 API

本文将为您介绍如何通过脚本或MyBatis方式生成API。

为了满足高阶用户的个性化查询需求，数据服务提供了自定义SQL的脚本/MyBatis取数方式，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。

- 脚本方式：仅支持普通SQL语法。
- MyBatis方式：仅专享版数据服务支持此方式，此方式下脚本支持Mybatis标签语法。Mybatis方式下参数解析格式为#{parameter}，支持if、choose、when、foreach和where等标签语法，您可以借助标签语法来灵活实现空值校验、多值遍历、动态查表、动态排序及聚合等复杂查询逻辑。

前提条件

已在“管理中心 > 数据连接”页面，完成数据源的配置。

约束与限制

API生成暂不支持Hive数据源的中文表和中文列场景。

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. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
3. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
4. 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。
5. 对于已成功创建的API目录，在API目录上右键单击，可选择编辑或删除API目录。

配置 API 基本信息

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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 13-8 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，指定当前API保存的位置，后续可以在指定的API目录中检索当前API。 API目录是数据服务中API的最小组织单元。您可以选择**新建API目录**已创建的目录。
请求Path	API访问路径，例如：/getUserInfo。 请求Path即完整的URL中，域名之后、查询参数之前的部分，如**图13-18**中的“/blogs/xxxx”。 **图 13-18** 统一资源定位符 URL 说明  在请求Path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，专享版支持HTTPS协议。 推荐选择HTTPS协议，HTTP安全性欠佳，可能会存在安全风险。 • 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。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。

配置	说明
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	审核人拥有API的审核权限。可单击“添加”，进入“审核中心 > 审核人管理”页面，新建审核人。
安全认证	创建API时，有如下三种安全认证方式可选。三种方式的区别在于认证方式和调用方法不同，推荐使用安全性更高的APP认证。 ● APP认证：将APP认证方式的API授权给应用后，使用应用的密钥对（AppKey和AppSecret）进行安全认证，支持通过SDK或API调用工具调用，安全级别高，推荐使用。 ● IAM认证：将IAM认证方式的API授权给当前账号或其他账号后，借助从IAM服务获取的用户Token进行安全认证。支持通过API调用工具调用，安全级别中等。 ● 无认证：不需要认证，所有用户均可访问，建议仅在测试接口时使用，不推荐正式使用。使用无认证方式时，无需鉴权认证信息，安全级别低，通过API调用工具或浏览器即可直接调用。
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“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发布状态预留的最低期限，单位为小时，0表示不设限制。 如果需要停用/下线/解除授权，则停用/下线/解除授权时间必须选择在发布后的最低保留期限时间之后。选择时间后，停用/下线/解除授权会通知已授权用户。如果所有已授权用户均完成审核中心通知列表消息处理，或在应用中解绑与API的绑定关系，API就会停用/下线/解除授权；否则会以待停用/待下线/待解除授权状态，等待达到停用/下线/解除授权时间，再强制停用/下线/解除授权。 例如，最低保留期限设置为24小时，则此API发布后需要停用时，停用时间必须选择在发布24小时后，即发布第二天之后。如果期间内已授权用户已完成审核中心通知列表消息处理或解绑应用与API的绑定关系，则会直接停用；如果未完成，则会以待停用状态等待达到停用时间，强制停用。

配置	说明
入参定义	配置调用API需要输入的参数，此处定义后即为配置取数逻辑时的请求参数。 入参定义主要由参数位置、参数类型、是否必填、允许空值以及默认值等组成。 ● 参数位置主要包括Query、Header、Path、Body四大类，另外还支持Static静态参数。 – Query是位于URL后的查询参数内容，以“?”开始，通过“&”连接多个参数。 – Header参数是位于请求消息头中的参数，常用于传递当前信息。例如host，token等。 – Path是位于请求路径中的请求参数，如果定义了Path参数，则需要在请求Path中也添加此参数，作为请求Path的一部分。 – Body是位于请求体内的参数，一般使用json格式表示。 – Static是不随API调用者的传值变化的静态参数，仅当安全认证为APP认证方式时支持。Static参数数值由API授权时确定（如果授权时未配置参数值，则SDK调用时会使用API入参默认值，API工具调用时会导致缺少Static参数值的报错）。 ● 参数类型分为数值型Number与字符型String两大类。Number参数对应数据库中int、double、long等数值数据类型，String参数对应数据库中char、vachar、text等文本数据类型。 ● 是否必填、允许空值以及默认值。 – 如果设定为必填，则API在访问时，必须传入指定参数。 – 如果设定为非必填，则在API访问时，未传入的参数会使用默认值进行代替；如果未传入参数，也没有默认值，则允许空值时会使用null替换，不允许空值时忽略该参数条件。 **说明** 入参定义中，参数大小限制如下： ● Query+Path，URL最大32KB ● Header，最大128KB ● Body， 最大128KB 实际配置中，需要根据所设计的调用API时请求参数情况来设置入参。例如，在用户表中根据用户ID查询用户信息时，请求Path设置为：/getUserInfo。可按照如下不同场景来配置入参： ● API调用时请求参数为用户id，需要返回对应id的用户信息。 1. 单击“添加”，参数名配置为id。 2. 参数位置选择Query。 3. 类型设置为Number。 4. 是否必填选择必填。 5. 默认值保持默认，无需填写。

配置	说明
	• API调用时请求参数为用户id1和用户id2，需要返回id1-id2范围内的用户信息： 1. 单击“添加”，参数名配置为id1。 2. 参数位置选择Query。 3. 类型设置为Number。 4. 是否必填选择必填。 5. 默认值保持默认，无需填写。 6. 再次单击“添加”，按照id1参数的配置信息再配置id2。

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

说明

本例中以脚本方式说明如何配置API取数逻辑。Mybatis方式与之相比差异在于参数解析形式和支持的语法差异，在使用流程上没有区别。

如果使用Mybatis方式生成API，则需要将本章节脚本中的参数解析格式由***${parameter}***修改为***#{parameter}***形式，另外Mybatis方式支持的标签语法可在界面中单击脚本编辑处的，查看弹出的Mybatis脚本编辑提示。

“取数方式”选择“脚本方式”或“MyBatis方式”：

1. 选择数据源、数据连接、数据库等数据信息。

   说明

   数据服务仅支持部分数据源，详情请参见**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. 选择分页方式，推荐使用自定义分页方式。
   - 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 * FROM userinfo WHERE id=${userid}
     ```

     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 * FROM userinfo WHERE id=${userid}) LIMIT {limitValue} OFFSET {offsetValue}
     ```

     其中limitValue表示读取的数据条数，offsetValue表示跳过的数据条数（即偏移量），系统将默认赋值。
   - 自定义分页是指在创建API时，数据服务将不对SQL进行处理，分页逻辑需要在写SQL时由用户自定义。值得注意的是，为避免API查询数据量过大导致集群异常，自定义分页方式下必须在写SQL时添加分页逻辑。

     如果已知需要读取的数据条数limitValue和需要跳过的数据条数offsetValue，则分页逻辑可以写成以下脚本：

     ```
     SELECT * FROM userinfo WHERE id=${userid} LIMIT {limitValue} OFFSET {offsetValue}
     ```

     而在实际使用中，更多的是根据分页后的页面大小pageSize和页码pageNum定义分页逻辑，脚本样式如下：


```
SELECT * FROM userinfo WHERE id=${userid} LIMIT {pageSize} OFFSET {pageSize*(pageNum-1)}
```

说明

不同的数据源具有不同的语法风格，分页脚本应按照数据源语法要求调整。例如：

- DLI数据源不支持“LIMIT {limitValue} OFFSET {offsetValue}”的写法，仅支持“LIMIT {limitValue}” 。
- HETU数据源分页需要反转，不支持“LIMIT {limitValue} OFFSET {offsetValue}”的写法，仅支持“OFFSET {offsetValue} LIMIT {limitValue}”。

3. 编写API查询SQL。

在脚本编辑页面，单击脚本编辑处的，按照脚本编辑提示开发SQL查询语句。单击可将入参添加为SQL语句的API请求参数。另外，专享版数据服务支持返回总条数，开启后可返回取值脚本执行结果数据的总条数。

例如，需要在用户表中根据用户ID查询用户信息时，取值脚本可写为如下脚本。其中，“id”为userinfo表中的字段，“userid”为API中定义的入参。

```
SELECT * FROM userinfo WHERE id=${userid}
```

如果分页方式为自定义分页，页面大小pageSize为10、页码pageNum为2时，按照LIMIT {pageSize} OFFSET {pageSize*(pageNum-1)}转换方法，脚本可写为：

```
SELECT * FROM userinfo WHERE id=${userid} LIMIT 10 OFFSET 10
```

图 13-19 编写 API 查询 SQL

脚本编辑完成后，单击脚本编辑窗口下方的“测试SQL”，填写入参值，执行验证是否能返回预期结果。如果测试失败，可在“预览SQL”页签下查看实际运行的SQL语句是否符合预期，或者通过“日志”页签查看报错信息。

图 13-20 测试 SQL

说明

- SELECT查询的字段即为API返回参数，支持通过AS返回别名。
- WHERE条件中的参数为API请求参数，脚本方式下参数格式为${参数名}，MyBatis方式下参数格式为#{参数名}。
- 对于DWS数据库的FLOAT4、FLOAT8类型参数，不支持比较数值是否相等。
- 专享版数据服务支持返回总条数，开启后可返回取值脚本执行结果数据的总条数。
- 如果单个参数需要传多个值时，写法如下：
  - 字符串：'a','b','c'
  - 数值：1,2
  - 字段：a,b,c

4. 添加排序参数。

   在排序参数列表中，单击“新建”可设置排序字段。

   - 字段名称对外不可见，是所选的数据表中的字段，是API调用时实际访问的内容。在API查询SQL语句已编写完成且测试通过的前提下，可在“字段名称”输入框中选择排序字段。
   - 变量可自定义，用于与字段名称关联。在“变量”输入框中输入参数名称（一般填写为参数名称即可），系统会自动修改为变量形式。
   - 是否可选决定了调用API时此排序参数是否必选，勾选则表示此参数可以不传，可以通过排序参数描述pre_order_by的值配置是否参与排序；不勾选则此参数必传，即使排序参数描述pre_order_by的值未配置此参数，依然会参与排序。
   - 排序方式表示了当前参数允许使用的排序形式，分为升序、降序以及自定义。自定义排序参数默认为升序排序，可通过排序参数描述pre_order_by的值进行调整；而升序或降序的排序参数，不支持通过pre_order_by的值调整排序方式，如果pre_order_by的值与此处设置排序方式不符，则会导致配置调试或调用报错。
   - 多个排序参数时，表示当第一个排序参数相等时，再逐一用后续排序参数去排序。与配置方式不同的是，参数的排序顺序与添加排序字段的先后无关，而是需要通过SQL脚本自定义，并且不支持通过排序参数描述pre_order_by的值进行调整。

   注意，脚本/MyBatis API的排序字段必须要使用ORDER BY添加到SQL语句中，才能使该排序参数生效，单击⇒可将排序参数添加到SQL语句。添加ORDER BY参数时，关联字段名即可，多个排序字段的先后顺序由脚本定义，不支持在脚本中通过ASC或DESC设置顺序或降序方式。SQL语句中未添加的排序参数即使在排序参数描述pre_order_by的值中定义，排序也不会生效。

   例如，需要在用户表中根据用户ID查询用户信息，先后通过age和kk两个字段排序，页面大小pageSize为10、页码pageNum为2时，脚本样例如下。

```
SELECT * FROM userinfo WHERE id=${userid} order by ${age},${kk} LIMIT 10 OFFSET 10
```

图 13-21 添加排序参数

脚本编辑完成后，单击脚本编辑窗口下方的“测试SQL”，填写入参值和排序参数描述pre_order_by的值，执行验证是否能返回预期结果。

pre_order_by的默认值已由系统根据已配置的所有排序参数给出，自定义排序默认为升序。排序参数描述pre_order_by的值填写形式为“***排序参数参数名***:ASC”或“***排序参数参数名***:DESC”，其中ASC表示升序，DESC表示降序，多个排序参数描述以“英文分号”进行分隔。勾选“是否传值”后，测试结果将按照pre_order_by的值排序。

对于pre_order_by的值，您可以进行如下修改：

- 删掉某可选的排序参数，则此排序参数不再参与排序。
- 修改自定义排序方式的排序参数为升序或降序方式，则此排序参数按照修改后的排序方式排序。

说明

pre_order_by的值，不支持进行如下修改，否则会修改不生效或导致调用报错。

- 删掉某必选的排序参数，则此排序参数依然会正常参与排序，删除不生效。
- 调整排序参数的前后顺序，则排序依然以SQL中的排序参数顺序为准。调整不生效。
- 修改升序或降序的排序参数为其他排序方式，则会调用失败，不允许修改。

如果测试失败，可在“预览SQL”页签下查看实际运行的SQL语句是否符合预期，或者通过“日志”页签查看报错信息。

图 13-22 测试 SQL

5. 单击“下一步”，进行API测试页面。

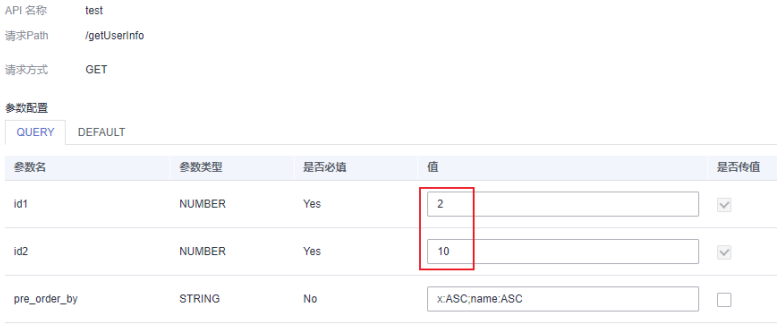
测试 API

1. 填写入参取值。

   如果单个参数需要传多个值时，写法如下：

   - 字符串：'a','b','c'
   - 数值：1,2
   - 字段：a,b,c

图 13-23 填写入参取值



2. （可选）调整排序参数描述pre_order_by的值。

   pre_order_by的默认值已由系统根据已配置的所有排序参数给出，自定义排序默认为升序。排序参数描述pre_order_by的值填写形式为“***排序参数参数名***:ASC”或“***排序参数参数名***:DESC”，其中ASC表示升序，DESC表示降序，多个排序参数描述以“英文分号”进行分隔。勾选“是否传值”后，测试结果将按照pre_order_by的值排序。

   对于pre_order_by的值，您可以进行如下修改：

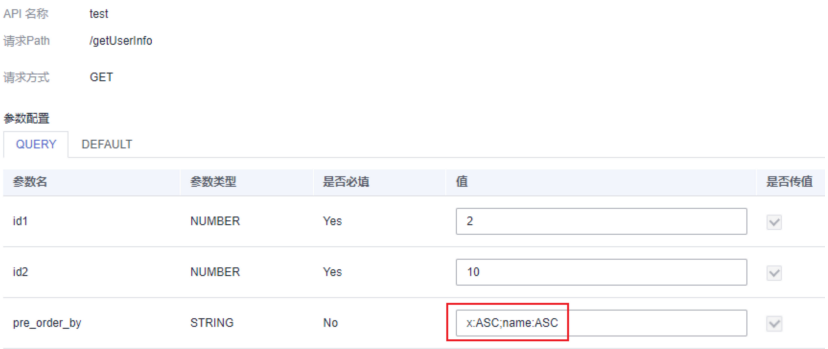
   - 删掉某可选的排序参数，则此排序参数不再参与排序。
   - 修改自定义排序方式的排序参数为升序或降序方式，则此排序参数按照修改后的排序方式排序。

   说明

   pre_order_by的值，不支持进行如下修改，否则会修改不生效或导致调用报错。

   - 删掉某必选的排序参数，则此排序参数依然会正常参与排序，删除不生效。
   - 调整排序参数的前后顺序，则排序依然以SQL中的排序参数顺序为准。调整不生效。
   - 修改升序或降序的排序参数为其他排序方式，则会调用失败，不允许修改。

图 13-24 调整排序参数描述 pre_order_by 的值



3. （可选）查看分页参数值。

   采用默认分页方式时，可以查看分页参数情况，其中pageSize表示分页后的页面大小，pageNum表示页码。默认按100的大小分页，返回第1页数据。

图 13-25 查看分页参数值

4. 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。

   - 测试过程中，如果数据服务API查询及返回数据的总时长超过默认60秒，会报超时错误。
   - 如果测试失败，请查看错误提示并做相应的修改重新测试。

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，单击“编辑”按钮进行修改API的相关操作。

说明

API如果处于发布、下线、停用、恢复的待审核或待执行状态，则不支持编辑。

13.3.4 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

已创建待调试的API。

操作步骤

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。

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13-9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 13-9 调试 API

参数名称	说明
API版本	仅专享版支持指定API版本调试。 当未指定API版本时，默认调试的是未发布的API。
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。
   - 调用超过默认60秒无结果时，会报超时错误。
   - 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。

相关操作

- 批量调试API：您可以在专享版的“开发API > API管理”页面，勾选需要调试的API后，依次单击API列表上方的“批量操作 > 批量调试”，然后在批量调试页面，导入修改后的API调试参数Excel，实现多个API的统一调试。

图 13-26 批量操作

- 发布API：API调试成功后，为方便API调用者调用，您可以将API发布，具体操作请参见**发布API**。

13.3.5 发布 API

本文将为您介绍如何发布数据服务中的API。

操作场景

为了安全起见，在数据服务中生成的API，都需要发布后才能对外提供服务。

前提条件

已调试成功待发布的API。

约束与限制

不支持单个或多个用户同时发布API到同一专享版集群，系统会提示“当前操作正在执行中，请稍后重试”。

操作步骤

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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，您可以选择集群进行发布。

   图 13-27 选择集群发布

   - 专享版默认发布到数据服务专享版集群上，支持按照API版本发布，发布成功后API调用者可以通过内网或公网调用该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。

   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。

说明

处于待审核状态的API无法修改数据连接，需要具有空间管理员角色的用户审批驳回才可进行修改。

审核人支持管理员、开发者、运维者，访客无法添加为审核人。

工作空间管理员角色的用户，无论是否被添加为审核人，都默认具备审核人权限。

6. 发布完成后，您可以进入到“服务目录”，查看已发布API信息。

相关操作

批量发布API：您可以在专享版的“开发API > API管理”页面，勾选需要发布的API后，依次单击API列表上方的“批量操作 > 批量发布”，实现多个API的统一发布。

图 13-28 批量操作

13.3.6 管理 API

13.3.6.1 API 版本管理

操作场景

数据服务专享版支持将API按照不同版本进行管理，可根据不同的API版本，分别进行调测、发布。

您也可以根据API版本追踪API的变更情况，支持版本对比。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。

前提条件

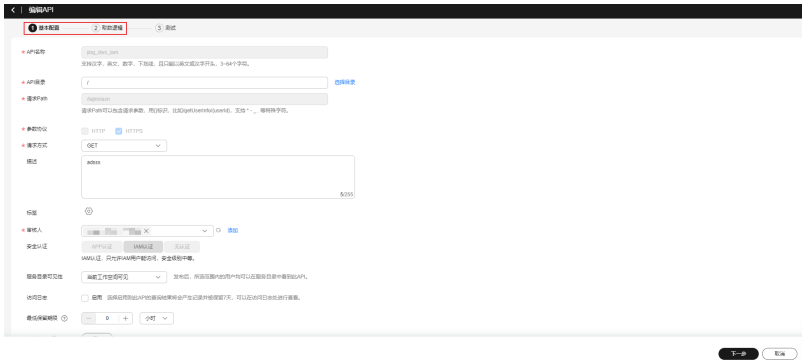
- 仅专享版支持API版本管理。
- API更新版本是通过对已发布的API进行编辑后再次发布实现的。API如果处于发布、下线、停用、恢复的待审核或待执行状态，则不支持编辑，因此无法更新版本。

更新 API 版本

API更新版本是通过对已发布的API进行编辑后，再次发布，从而实现版本更新。

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
2. 在左侧导航栏选择专享版，进入总览页。
3. 进入“开发API > API管理”页面。确保待更新版本的API为已发布状态后，在API服务列表操作列中，选择“编辑”。
4. 在API编辑界面，您可以修改API的基础配置或取数逻辑，例如API目录、描述、请求方式、入参、取数方式等，注意API名称、请求path、参数协议、安全认证不支持修改。

图 13-29 修改 API 的基础配置或取数逻辑



5. API修改完成后，单击“下一步”进入测试页面。填写相关参数后，进行API测试。

   左侧为API请求参数配置区域，参数说明如表13-10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 13-10 调试 API

参数名称	说明
API版本	仅专享版支持指定API版本调试。 当未指定API版本时，默认调试的是未发布的API。
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。

6. 测试完成后，单击“确定”返回API列表。已成功修改的API会在API名称后添加“已编辑”标签。

图 13-30 已编辑 API

7. 再次发布已编辑的API。在API服务列表操作列中，选择“更多 > 发布”，然后选择已调试通过的集群进行发布。

您可以将已编辑的API发布在上一次发布的集群上，该集群上的API信息将按编辑后的信息进行更新；您也可以将已编辑的API发布在其他集群上，则该API可以实现不同的版本发布在不同的集群上。

查看与对比版本

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
2. 在左侧导航栏选择专享版，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，单击API名称进入API详情页面。
4. 在API详情页面，单击“版本管理”，可查看当前保存的版本记录（最多保留最近10条）。

   您可以查看对应版本API的详细内容，也可以删除或发布对应版本。当勾选两个版本时，您也可以通过“版本对比”，对比两个版本之间的差异情况。

   图 13-31 API 版本管理

13.3.6.2 设置 API 可见

操作场景

当需要修改API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，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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，在待修改的API所在行，选择“更多 > 设置可见”。
4. 在弹出的窗口中单击添加，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。

   项目ID可以参考如下步骤进行获取：

   a. 注册并登录管理控制台。
   b. 在用户名的下拉列表中单击“我的凭证”。

c. 在“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。

图 13-32 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，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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，API如果处于发布、下线、停用、恢复的待审核或待执行状态，则不支持编辑。
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。

13.3.6.3 停用/恢复 API

操作场景

当已发布的API需要编辑、调试时，必须将API从相关环境中停用后才允许操作。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，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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。

说明

停用时间必须选择在API发布后的最低保留期限时间之后。选择停用时间后，停用操作会通知已授权用户。如果所有已授权用户均完成审核中心通知列表消息处理，或在应用中解绑与API的绑定关系，API就会直接停用；否则会以待停用状态，等待达到停用时间，再强制停用。

恢复 API

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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。

13.3.6.4 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。

下线 API

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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。

> **说明**
>
> 下线时间必须选择在API发布后的最低保留期限时间之后。选择下线时间后，下线操作会通知已授权用户。如果所有已授权用户均完成审核中心通知列表消息处理，或在应用中解绑与API的绑定关系，API就会直接下线；否则会以待下线状态，等待达到下线时间，再强制下线。

删除 API

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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”。

> **说明**
>
> - 只有未发布状态（如已创建、已下线）的API可以删除，已停用或发布状态不可删除。
> - 批量删除API最多同时删除1000个API。

4. 单击“确定”，完成API删除。

相关操作

批量下线API：您可以在专享版的“开发API > API管理”页面，勾选需要下线的已发布API后，依次单击API列表上方的“批量操作 > 批量下线”，实现多个API的批量下线。

图 13-33 批量操作

13.3.6.5 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，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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，单击确认即可完成API复制。

图 13-34 复制 API

13.3.6.6 同步 API

操作场景

您可以通过同步API功能，将专享版的API同步至数据地图。

前提条件

已创建API。

同步 API 到数据地图

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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. 勾选待同步的API所在行，在API列表上方，选择并单击“更多 > 同步至数据地图”。

图 13-35 同步至数据地图

5. 在同步结果页面，查看同步状态和详情，确认API同步结果。

图 13-36 同步结果

说明

- 仅已发布状态的API支持同步至数据地图。
- 仅以下数据源的API支持同步：DLI、DWS、HBase、Clickhouse。

13.3.6.7 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。

约束限制

- 全量导出必须具备DAYU Administrator或Tenant Administrator权限。
- 每个工作空间每分钟仅能全量导出一次，同时只能有一个全量导出任务执行。

全量导出 API

全量导出时会将全量API按照当前的筛选条件进行导出，须具备DAYU Administrator或Tenant Administrator权限。

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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

- 全量导出必须具备DAYU Administrator或Tenant Administrator权限。
- 每个工作空间每分钟仅能全量导出一次，同时只能有一个全量导出任务执行。

在导出窗口中单击“确认”导出全量API，单击确认即可以Excel文件的形式导出API。

图 13-37 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，单击下方页签可以切换查看并编辑。

导出 API

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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，单击确认即可以Excel文件的形式导出API。

图 13-38 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，单击下方页签可以切换查看并编辑。

导入 API

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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页面。
5. 在导入页面中配置导入参数后，单击“选择Excel文件”，选择待导入的API文件后单击导入，导入结果中可以展示导入状态。

表 13-11 导入参数配置说明

参数	说明
是否覆盖	配置导入的重名API是否需要更新，默认不更新。 ● 否：如果已存在同名API，则不导入该API。 ● 是：如果存在同名API，则按照导入的API更新API定义。
导入文件	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 13-39 导入 API

6. 导入成功后，即可在API列表中查看导入的API。

13.3.7 编排 API

13.3.7.1 编排 API 简介

数据服务API编排是指将已经开发好的服务API接口，在无需编写复杂代码的情况下，根据特定的业务逻辑和流程进行可视化的重组和重构，从而实现在不影响原生接口的前提下进行简便的二次开发。API编排为您提供拖拽式、可视化的API工作流程编排能力，您可以按照业务逻辑，以串行、并行等结构组合多个API为工作流，然后通过入口API调用API工作流，最终返回所需数据。

API编排使得业务流程的设计和优化变得更加直观和高效，同时也为二次开发提供了更便捷的方式。您可以在如下场景中可以使用API编排，简化开发工作：

- **对返回消息进行映射或格式转换**：通过API编排的方式能够灵活实现消息映射及格式转换。
- **数据请求依赖多个数据API**：使用API编排后，可以降低调用次数，减少集成成本，提升调用效率。

约束与限制

- 仅3.0.6及以上版本的数据服务专享版集群支持API编排。
- API工作流发布前，需确保其中的普通API均已处于已发布状态。

算子和工作流简介

在API工作流编排页面，您可以自由拖拽各类算子到画布中，然后基于特定的业务逻辑和流程通过连线编排工作流，最后配置算子，完成后即可保存、调试及发布工作流。

API编排支持五类可拖拽的算子，分别为：入口API、普通API、条件分支、并行处理和输出处理。其中，入口API位于最上游，输出处理位于最下游，中间部分可以是普通API、条件分支和并行处理这三类算子的任意组合。注意，编排工作流时需要满足如下要求：

- 有且只有一个入口API算子，并位于最上游，向下只能有一个分支。
- 至少有一个普通API算子，并位于中间层，上下游均有其他算子，向下只能有一个分支。
- 条件分支算子可选，位于中间层，必须至少有2个分支，最多支持20个分支，多个分支满足条件时仅执行第一个满足条件的分支。

  注意，条件分支的直接下游不能为输出处理算子，只能获取上级算子请求参数或结果集进行条件判断。
- 并行处理算子可选，位于中间层，必须至少有2个分支，最多支持20个分支，必须配置失败策略。

  注意，并行处理的直接下游不能为输出处理算子，只能支持同时执行多个分支逻辑，分支间互不影响。
- 有且只有一个输出处理算子，并位于最下游，直接上游必须为普通API算子，必须配置至少一个结果映射。
- API工作流不能有环状结构，不能有孤立算子，最多支持20层深度。

图 13-40 API 工作流编排页面

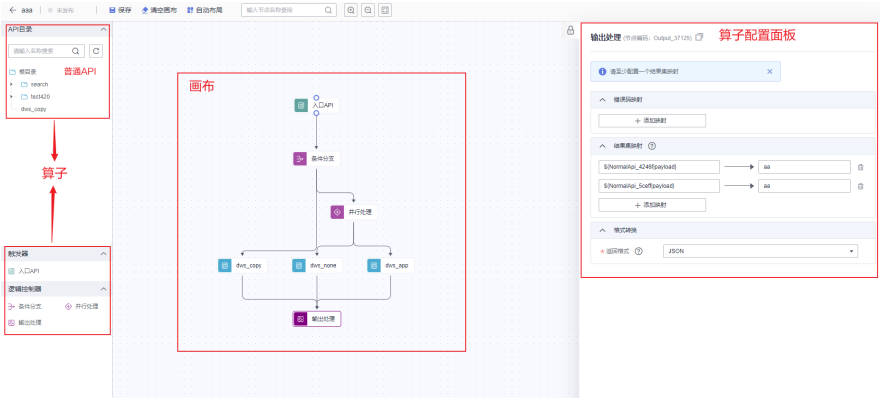


表 13-12 API 工作流算子介绍

配置入口	算子	是否必选	介绍
触发器	**入口API**	必选	入口API算子是API工作流的入口，工作流发布后可通过调用入口API来调用API工作流。在入口API算子内需定义API工作流的名称、URL、参数协议、请求方式、审核人、安全认证以及请求参数。 入口API算子的配置方法，详见**配置入口API算子**。
API目录	**普通API**	必选	普通API是执行数据查询操作的算子。普通API即已创建的数据API，编排API时您可以从API目录内拖拽一个普通API作为执行算子进行取数，并将请求参数或结果集作为变量传递下去。 普通API可参考**配置方式生成API**或**脚本/MyBatis方式生成API**进行创建。
逻辑控制器	**条件分支**	非必选	条件分支算子通过获取上游算子的请求参数或结果集进行条件判断，根据定义的表达式来确定下一步执行的分支。注意，多个分支满足条件时仅执行第一个满足条件的分支。 条件分支算子和表达式的配置方法，详见**配置条件分支算子**。
	并行处理	非必选	并行处理算子可以同时执行多个分支逻辑，分支间互不影响。 并行处理算子的配置方法，详见**配置并行处理算子**。
	输出处理	必选	输出处理算子负责对API工作流的执行结果进行错误码映射、结果集映射和格式转换，以确定最终返回的数据格式。 输出处理算子的配置方法，详见**配置输出处理算子**。

13.3.7.2 配置入口 API 算子

入口API算子是API工作流的入口，工作流发布后可通过调用入口API来调用API工作流。在入口API算子内需定义API工作流的名称、URL、参数协议、请求方式、审核人、安全认证以及请求参数。

表 13-13 入口 API 算子

参数	说明
API名称	入口API名称即API工作流名称。 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。

参数	说明
请求Path	入口API访问路径即API工作流访问路径，例如：/getUserInfo。 请求Path即完整的URL中，域名之后、查询参数之前的部分，如图13-41中的“/blogs/xxxx”。 **图 13-41** 统一资源定位符 URL 说明  在请求Path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，专享版支持HTTPS协议。 推荐选择HTTPS协议，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。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	审核人拥有API的审核权限。可单击“添加”，进入“审核中心 > 审核人管理”页面，新建审核人。

参数	说明
安全认证	创建API时，有如下三种安全认证方式可选。三种方式的区别在于认证方式和调用方法不同，推荐使用安全性更高的APP认证。 ● APP认证：将APP认证方式的API授权给应用后，使用应用的密钥对（AppKey和AppSecret）进行安全认证，支持通过SDK或API调用工具调用，安全级别高，推荐使用。 ● IAM认证：将IAM认证方式的API授权给当前账号或其他账号后，借助从IAM服务获取的用户Token进行安全认证。支持通过API调用工具调用，安全级别中等。 ● 无认证：不需要认证，所有用户均可访问，建议仅在测试接口时使用，不推荐正式使用。使用无认证方式时，无需鉴权认证信息，安全级别低，通过API调用工具或浏览器即可直接调用。
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“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发布状态预留的最低期限，单位为小时，0表示不设限制。 如果需要停用/下线/解除授权，则停用/下线/解除授权时间必须选择在发布后的最低保留期限时间之后。选择时间后，停用/下线/解除授权会通知已授权用户。如果所有已授权用户均完成审核中心通知列表消息处理，或在应用中解绑与API的绑定关系，API就会停用/下线/解除授权；否则会以待停用/待下线/待解除授权状态，等待达到停用/下线/解除授权时间，再强制停用/下线/解除授权。 例如，最低保留期限设置为24小时，则此API发布后需要停用时，停用时间必须选择在发布24小时后，即发布第二天之后。如果期间内已授权用户已完成审核中心通知列表消息处理或解绑应用与API的绑定关系，则会直接停用；如果未完成，则会以待停用状态等待达到停用时间，强制停用。

参数	说明
入参定义	配置调用API工作流需要输入的参数。 入参定义主要由参数位置、参数类型、是否必填、允许空值以及默认值等组成。 ● 参数位置主要包括Query、Header、Path、Body四大类，另外还支持Static静态参数。 – Query是位于URL后的查询参数内容，以“?”开始，通过“&”连接多个参数。 – Header参数是位于请求消息头中的参数，常用于传递当前信息。例如host，token等。 – Path是位于请求路径中的请求参数，如果定义了Path参数，则需要在请求Path中也添加此参数，作为请求Path的一部分。 – Body是位于请求体内的参数，一般使用json格式表示。 – Static是不随API调用者的传值变化的静态参数，仅当安全认证为APP认证方式时支持。Static参数数值由API授权时确定（如果授权时未配置参数值，则SDK调用时会使用API入参默认值，API工具调用时会导致缺少Static参数值的报错）。 ● 参数类型分为数值型Number与字符型String两大类。Number参数对应数据库中int、double、long等数值数据类型，String参数对应数据库中char、vachar、text等文本数据类型。 ● 是否必填、允许空值以及默认值。 – 如果设定为必填，则API在访问时，必须传入指定参数。 – 如果设定为非必填，则在API访问时，未传入的参数会使用默认值进行代替；如果未传入参数，也没有默认值，则允许空值时会使用null替换，不允许空值时忽略该参数条件。 **说明** 入参定义中，参数大小限制如下： ● Query+Path，URL最大32KB ● Header，最大128KB ● Body， 最大128KB 实际配置中，需要根据所设计的调用API工作流时请求参数情况来设置入参。例如，设计工作流在多张表中根据用户ID查询用户信息时，请求Path设置为：/getUserInfo。可按照如下不同场景来配置入参： ● API调用时请求参数为用户id，需要通过工作流返回对应id的用户信息。 1. 单击“添加”，参数名配置为id。 2. 参数位置选择Query。 3. 类型设置为Number。

参数	说明
	4. 是否必填选择必填。 5. 默认值保持默认，无需填写。 ● API调用时请求参数为用户id1和用户id2，需要通过工作流返回id1-id2范围内的用户信息： 1. 单击“添加”，参数名配置为id1。 2. 参数位置选择Query。 3. 类型设置为Number。 4. 是否必填选择必填。 5. 默认值保持默认，无需填写。 6. 再次单击“添加”，按照id1参数的配置信息再配置id2。

13.3.7.3 配置条件分支算子

条件分支算子通过获取上游算子的请求参数或结果集进行条件判断，根据定义的表达式来确定下一步执行的分支。注意，多个分支满足条件时仅执行第一个满足条件的分支。

表 13-14 条件分支算子

参数	说明
分支1	
条件类型	选择条件类型。 ● 满足当前条件时：表示传入“条件分支”的数据满足指定的表达式时，将执行该分支。 ● 不满足其他条件时：表示传入“条件分支”的数据不满足其他所有分支的条件时，将执行该分支。
表达式	当条件类型为“满足当前条件时”，需要根据表达式配置条件。 条件分支表达式由上游算子的节点编码和变量名组成，使用方法请参考**变量表达式定义方法**。
分支2	
条件类型	选择条件类型。 ● 满足当前条件时：表示传入“条件分支”的数据满足指定的表达式时，将执行该分支。 ● 不满足其他条件时：表示传入“条件分支”的数据不满足其他所有分支的条件时，将执行该分支。

参数	说明
表达式	当条件类型为“满足当前条件时”，需要根据表达式配置条件。 条件分支表达式由上游算子的节点编码和变量名组成，使用方法请参考**变量表达式定义方法**。
...	
分支n	
条件类型	选择条件类型。 • 满足当前条件时：表示传入“条件分支”的数据满足指定的表达式时，将执行该分支。 • 不满足其他条件时：表示传入“条件分支”的数据不满足其他所有分支的条件时，将执行该分支。
表达式	当条件类型为“满足当前条件时”，需要根据表达式配置条件。 条件分支表达式由上游算子的节点编码和变量名组成，使用方法请参考**变量表达式定义方法**。

表达式定义方法

在定义条件分支的表达式时，需要配置变量表达式。当前仅入口API和普通API支持定义变量，条件分支、并行处理和输出处理暂不支持。表达式标准写法为：${***节点编码|变量名***}，定义方法如表13-15所示。

- **节点编码**：由系统动态分配，不可改动。您可以在API编排的画布中，单击节点后在节点详情中查看节点编码，并支持通过复制节点编码。

  图 13-42 查看节点编码

- **变量名**：支持的变量包括请求参数值和结果集相关参数，详情请参见表13-15。

表 13-15 条件表达式定义方法

算子	变量表达式	样例
入口API	获取入口API的请求参数的值：${***节点编码*|*入参名***} **说明** 当入参位置为Query、Header、Path或入参位置为Body的POST请求时，支持此表达式。	入口API的节点编码为EntryApi_3909f，入参userId位置为Path，获取请求参数的值：${EntryApi_3909f|userId}
普通API	1. 获取普通API的请求参数的值：${***节点编码*\|*入参名***} **说明** 当入参位置为Query、Header、Path或入参位置为Body的POST请求时，支持此表达式。 2. 获取普通API的结果集及相关变量： ● 获取普通API的查询状态是否成功，结果为true或false：${***节点编码***|payload.success} ● 获取普通API查询结果集内的行数：${***节点编码***\|payload.rowSize} ● 获取普通API查询结果集内的列数：${***节点编码***|payload.columnSize} ● 获取普通API查询结果集内的列名：${***节点编码***\|payload.columnNames} ● 获取普通API查询结果集内的第n行、对应列名为id的值：${***节点编码***|payload.data[***n-1***].***id***}	● 普通API的节点编码为NormalApi_4246f，入参userId位置为Path，获取请求参数的值：${NormalApi_4246f\|userId} ● 普通API的节点编码为NormalApi_4246f，取值结果为多行单列的一维数组，获取结果集内第1行的值：${NormalApi_4246f|payload.data[0]} ● 普通API的节点编码为NormalApi_4246f，取值结果为多行多列的二维数组，获取结果集内第1行、列名为price的值：${NormalApi_4246f|payload.data[0].price}

例如，对于A（入口API）>B（普通API）>C（条件分支）这3个顺序节点，节点C需要取节点A的请求参数值和节点B的输出值：

- A节点编码为EntryApi_3909f，入参userId位置为Path。

  取A节点请求参数值：${EntryApi_3909f|userId}。
- B节点编码为NormalApi_4246f，取值结果为多行多列的二维数组，获取结果集内第1行、列名为name的值。

  取B节点输出：${NormalApi_4246f|payload.data[0].name}。

13.3.7.4 配置并行处理算子

并行处理算子可以同时执行多个分支逻辑，分支间互不影响。

表 13-16 并行处理算子

参数	说明
失败策略	当并行分支中存在失败情况时，配置API工作流的失败策略。 ● 任一分支失败则终止：表示当并行分支中存在失败情况时，则此API工作流置为失败状态，不再继续执行。 ● 分支失败继续执行：表示当并行分支中存在失败情况时，继续执行其他分支和后续算子。当所有分支均失败导致后续算子无法执行时，则此API工作流置为失败状态。
分支1	
超时时间（ms）	表示当前分支执行超过配置的超时时间后，则将此分支置为失败状态。默认为0无时间限制。
分支2	
超时时间（ms）	表示当前分支执行超过配置的超时时间后，则将此分支置为失败状态。默认为0无时间限制。
...	
分支n	
超时时间（ms）	表示当前分支执行超过配置的超时时间后，则将此分支置为失败状态。默认为0无时间限制。

13.3.7.5 配置输出处理算子

输出处理算子负责对API工作流的执行结果进行错误码映射、结果集映射和格式转换，以确定最终返回的数据格式。

表 13-17 输出处理算子

参数	是否必选	说明
错误码映射	否	针对数据服务返回的错误码，支持映射为自定义信息。 例如，将“DLM.0”错误码映射为“OK”。

参数	是否必选	说明
结果集映射	是	针对工作流中查询到的所有普通API结果集，支持对一个或多个节点的结果集名称进行映射，映射后的名称会作用到JSON或文件名中，未映射的结果集将不会输出到最终返回结果中。 节点映射表达式写法固定为“${***节点编码***\|payload}”，节点编码可通过在API编排的画布中，单击节点后在节点详情中查看，并支持通过复制。 **图 13-43** 查看节点编码  例如节点编码为NormalApi_5a256，则节点映射表达式为“${NormalApi_5a256|payload}”，结果集名称定义为“销售记录”。
格式转换	否	工作流默认按照JSON字符串格式输出结果，支持将已映射的结果集数据导出为CSV、TXT、Excel或XML文件，一个数据集一个文件，最终打包成ZIP压缩文件进行导出。 注意导出时不支持断点续传。

13.3.7.6 API 编排典型配置

API编排的典型使用场景如下：

- **对返回消息进行映射或格式转换**：通过API编排的方式能够灵活实现消息映射及格式转换。
- **数据请求依赖多个数据API**：使用API编排后，可以降低调用次数，减少集成成本，提升调用效率。

约束与限制

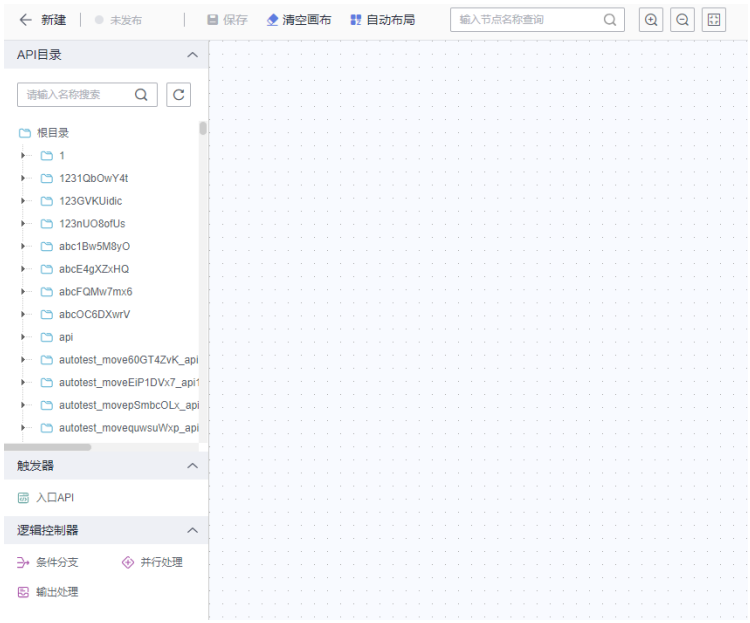
- 仅3.0.6及以上版本的数据服务专享版集群支持API编排。
- API工作流发布前，需确保其中的普通API均已处于已发布状态。

开发 API 工作流 1：对返回消息进行映射或格式转换

某API默认返回的是JSON数据，现需要将API调用结果转换为EXCEL格式。

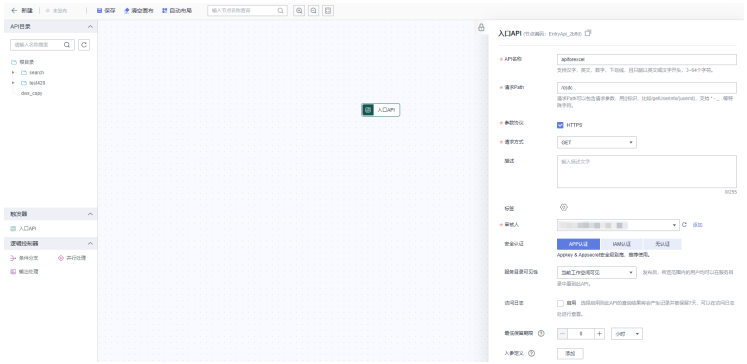
1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
3. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
4. 进入“开发API > > API编排”页面，单击新建，进入API编排页面。

图 13-44 进入 API 编排页面



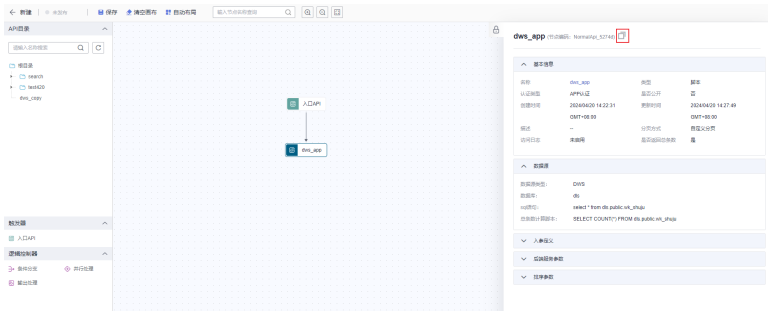
5. 拖拽“入口API”算子到画布，单击画布上的算子打开配置面板，配置入口API信息。

图 13-45 配置入口 API 算子



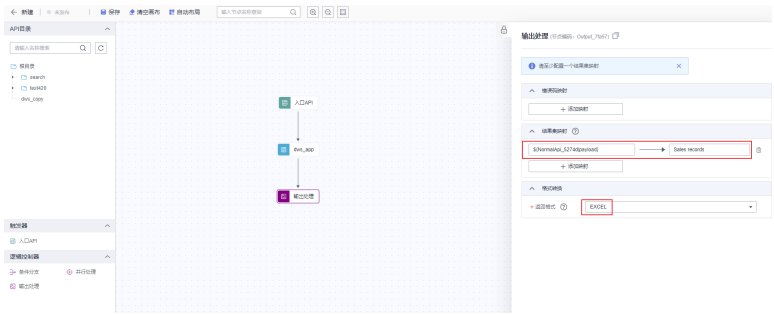
6. 在API目录找到需要转换的普通API并拖拽到画布，挂载到入口API下。单击画布上的普通API打开配置面板，复制节点编码例如：NormalApi_5274d。

图 13-46 复制节点编码



7. 拖拽“输出处理”算子到画布，挂载到普通API下。单击画布上的输出处理算子打开配置面板，配置输出处理算子：
   - 添加结果集映射。节点映射表达式取普通API的结果，如“${NormalApi_5274d|payload}”，结果集名称按照实际取值，此处以销售记录Sales records为例进行配置。
   - 格式转换选择返回格式为EXCEL。

图 13-47 配置输出处理算子



8. 保存API工作量，然后调试并发布到集群。则后续调用者就可以通过调用API工作流中的入口API，实现普通API取数结果保存在EXCEL文件中。

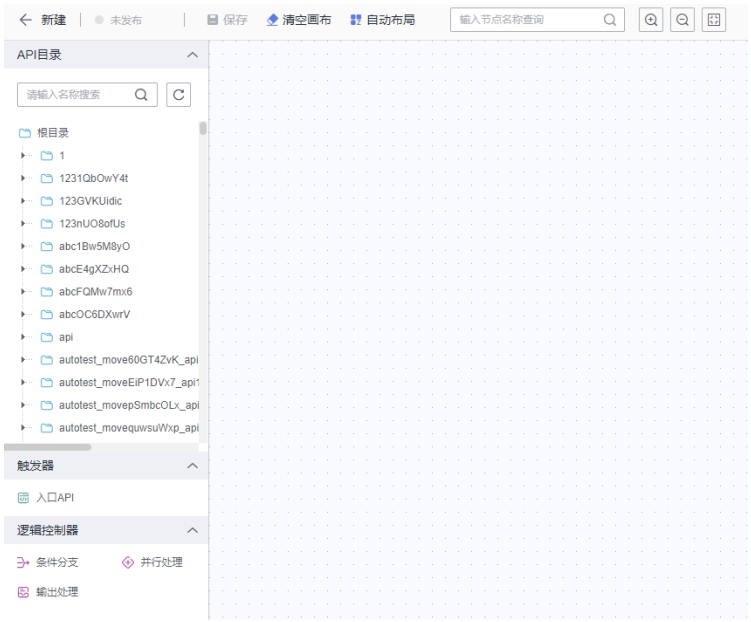
开发 API 工作流 2：数据请求依赖多个数据 API

在电子商务平台的场景中，某部门需要根据用户所在地区的不同，提供不同的信息和服务：如果用户位于area1地区，系统将提供供应商信息Supplier Information和销售评级数据Sales Rating；如果用户位于其他地区，系统则会返回零售商信息Retailer Information。

当前已有地区信息API是AreaInformation、供应商信息API是SupplierInformation、销售评级API是SalesRating、零售商信息API是RetailerInformation，则您可以参考后续步骤进行API工作流编排。

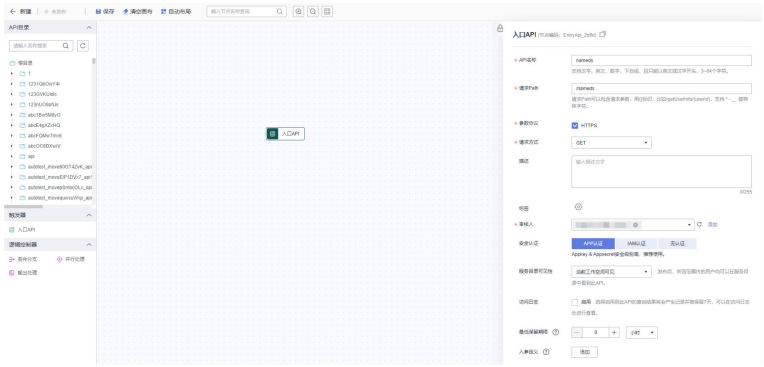
1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
3. 在左侧导航栏选择服务版本（例如：专享版），进入总览页。
4. 进入“开发API > > API编排”页面，单击新建，进入API编排页面。

图 13-48 进入 API 编排页面



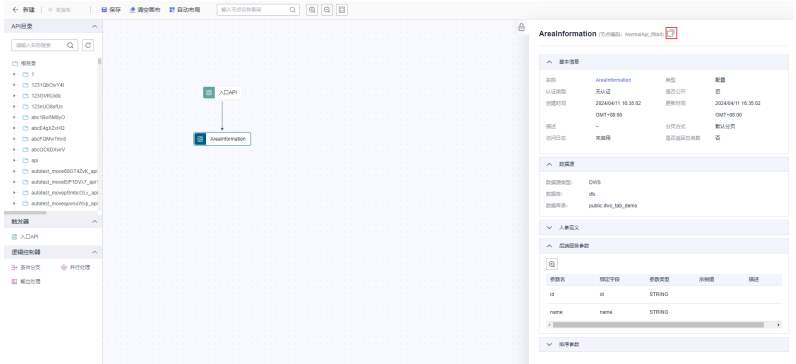
5. 拖拽“入口API”算子到画布，单击画布上的算子打开配置面板，配置入口API信息。

图 13-49 配置入口 API 算子



6. 在API目录找到地区信息AreaInformation API并拖拽到画布，挂载到入口API下。单击画布上的普通API打开配置面板，复制节点编码例如：NormalApi_ff8bd。

图 13-50 复制 AreaInformation 节点编码

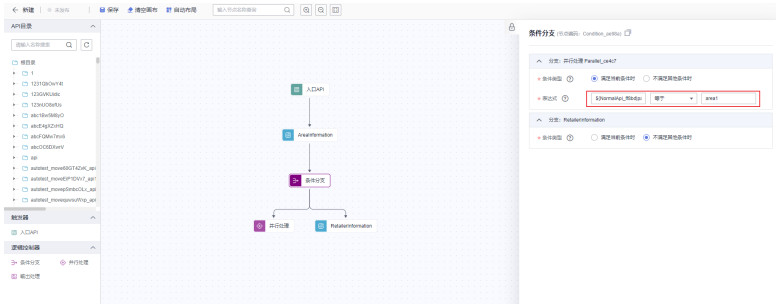


7. 拖拽“条件分支”算子到画布，挂载到AreaInformation API下，并在条件分支下挂载并行处理算子和零售商信息RetailerInformation API两个分支。其中零售商信息RetailerInformation普通API的节点编码为NormalApi_de62d。

   单击画布上的条件分支算子打开配置面板，配置条件分支算子：

   - 并行处理分支的条件类型配置为“满足条件时”，表达式配置为“${NormalApi_ff8bd|payload.data[0].area}”，该表达式含义是获取AreaInformation API的结果集内的第1行、对应列名为area的字段值。此处配置为如果该值等于“area1”，就执行并行处理分支。
   - 零售商信息RetailerInformation分支的条件类型配置为“不满足其他条件时”，表示如果不满足并行处理分支的条件，则执行零售商信息RetailerInformation分支。

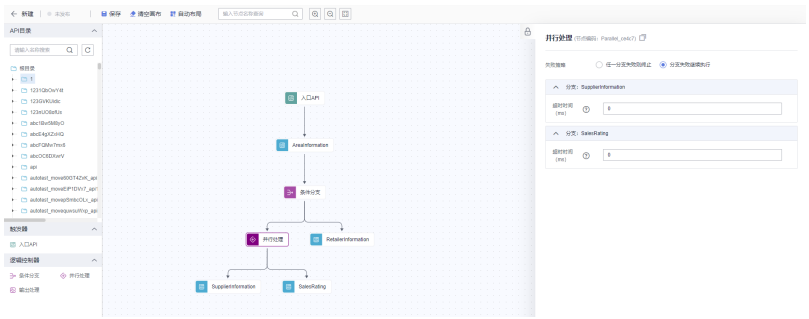
   图 13-51 配置条件分支算子

8. 在API目录找到供应商信息SupplierInformation API和销售评级SalesRating API并拖拽到画布，挂载到并行处理算子下。其中供应商信息SupplierInformation和销售评级SalesRating两个普通API的节点编码分别为NormalApi_3ad5c、NormalApi_01e7e。

   单击画布上的并行处理算子打开配置面板，可以配置失败策略及分支超时时间（此处无需特殊配置，报错默认即可）。当并行处理分支被执行时，SupplierInformation和SalesRating两个分支会被同时调度。

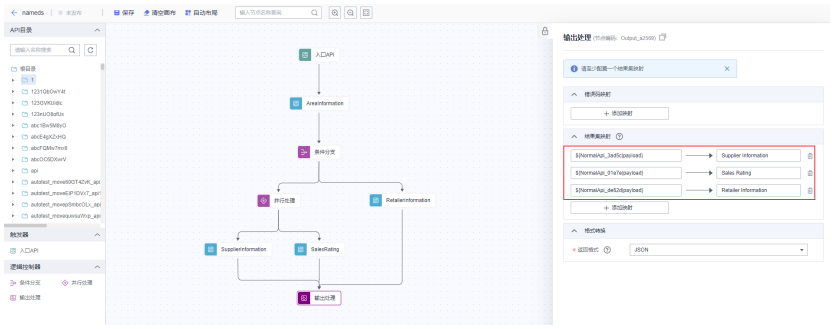
   图 13-52 配置并行处理算子

9. 拖拽“输出处理”算子到画布，挂载到3个普通API算子下。单击画布上的输出处理算子打开配置面板，添加结果集映射。

   为能够输出3个普通API算子的结果，需要配置3条映射记录。节点映射表达式取普通API的结果，如“${NormalApi_3ad5c|payload}”、“${NormalApi_01e7e|payload}”、“${NormalApi_de62d|payload}”，结果集名称按照实际取值。

图 13-53 配置输出处理算子



10. 保存API工作量，然后调试并发布到集群。则后续调用者就可以通过调用API工作流中的入口API，实现根据客户所在地区的不同，返回不同的信息。

相关操作

- 编辑API工作流：在API工作流列表页面，单击对应工作流操作栏中的“编辑”，即可进入API工作流编排页面，重新进行工作流编排或修改。
- 查看API工作流授权：在API工作流列表页面，单击对应工作流操作栏中的“查看授权”，即可进入API完整信息界面，并对工作流进行授权。

  注意，当入口API的安全认证方式为APP认证或IAM认证时，在调用API工作流前需要完成创建应用和将API授权给应用。工作流授权方式与API授权方式基本一致，可参考**授权API调用**或**申请API授权**。
- 调试API工作流：在API工作流列表页面，单击对应工作流操作栏中的“更多 > 调试”，即可进入API工作流调试页面。

  在添加请求参数后，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。调试流程与API调试流程基本一致，可参考**调试API**。
- 发布API工作流：在API工作流列表页面，单击对应工作流操作栏中的“更多 > 发布”，即可弹出API工作流发布窗口。

  API工作流需要发布后才能对外提供服务。发布流程与API发布流程基本一致，可参考**发布API**。
- 下线/删除API工作流：在API工作流列表页面，单击对应工作流操作栏中的“更多 > 下线”，即可弹出API工作流下线窗口；选择API工作流后，单击对应工作流操作栏上方的“删除”，即可弹出API工作流删除窗口。

  已发布的API工作流因为其他原因需要停止对外提供服务，可以从相关环境中下线，但需注意下线API工作流不会保留原有的授权信息。下线后的API工作流如果确认不再提供服务，可以进行删除，请注意删除操作无法撤销。下线/删除流程与API下线/删除流程基本一致，可参考**下线/删除API**。
- 停用/恢复API工作流：在API工作流列表页面，单击对应工作流操作栏中的“更多 > 停用”或“更多 > 恢复”，即可弹出API工作流停用/恢复窗口。

  已发布的API工作流需要编辑、调试时，必须将API工作流从相关环境中停用后才允许操作。停用API工作流会保留原有的授权信息，在停用期间您可以对API工作流进行编辑、调试等操作。停用后您可以通过恢复API工作流，使该API工作流继续对外提供服务。停用/恢复流程与API停用/恢复流程基本一致，可参考**停用/恢复API**。
- 设置API工作流可见：在API工作流列表页面，单击对应工作流操作栏中的“更多 > 设置可见”，即可弹出API工作流设置可见窗口。

设置API工作流可见可以修改API工作流在服务目录中的可见范围。设置可见流程与API设置可见流程基本一致，可参考**设置API可见**。

- 复制API工作流：在API工作流列表页面，单击对应工作流操作栏上方的“更多 > 复制”，即可弹出API工作流复制窗口。

  复制API工作流能够得到与原API工作流配置相同的API工作流。复制流程与API复制流程基本一致，可参考**复制API**。

- 同步API工作流至数据地图：在API工作流列表页面，单击对应工作流操作栏上方的“更多 > 同步至数据地图”，即可进入API工作流同步页面。

  同步API工作流至数据地图能够将API工作流资产同步到数据地图组件进行查看。同步流程与API同步流程基本一致，可参考**同步API到数据地图**。

13.3.8 配置 API 调用流控策略

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，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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13-18**所示信息。

图 13-54 创建流控策略

表 13-18 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。

信息项	描述
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，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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。

13.3.9 授权 API 调用

13.3.9.1 通过应用授权 APP 认证方式 API

应用定义了一个API调用者的身份。对于使用APP认证方式的API，必须在创建APP类型应用并将API授权给应用后，才能获得认证信息以用于API调用。

一个APP认证方式的API可以授权给多个APP类型的应用，多个APP认证方式的API也可以授权给同一个APP类型的应用。API授权后，在调用时就可以使用任意授权应用的密钥对（AppKey和AppSecret）进行安全认证，对调用者本身的用户身份无要求。

约束与限制

- 使用APP认证方式的API必须先通过应用授权才能调用。
- APP认证方式的API只能授权给APP类型的应用。
- 如果对无认证方式的API进行应用授权，则系统会忽略此操作。
- 仅DAYU Administrator、Tenant Administrator或者工作空间管理员支持重置APP类型应用的AppSecret。
- APPSecret限制一分钟内重置一次，重置记录可在事件管理内查看。
- 重置APPSecret会导致已授权的API调用失败，请谨慎操作。

创建 APP 类型的应用

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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13-19所示信息。

   表 13-19 应用信息

信息项	描述
应用名称	应用名称。
应用类型	选择APP应用类型，APP认证方式的API只能授权给APP类型的应用。 • IAM：IAM类型应用为IAM认证方式的API进行授权。IAM类型应用为实例级别配置，应用名称固定为华为账号，每个DataArts Studio实例下仅能创建一个，各工作空间之间均可见。 • APP：APP类型应用为APP认证方式的API进行授权。您可以将不同的APP认证方式API授权给不同的应用，提升数据安全性。
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”，进入应用详情页面可查看AppKey和AppSecret，并可以重置AppSecret。

说明

重置APPSecret会导致已授权的API调用失败，请谨慎操作。

图 13-55 应用详情

将 APP 认证方式的 API 授权给 APP 类型的应用

使用APP认证方式的API，必须将API授权给应用后，才能进行API调用。授权可以分为API开发者主动授权和API调用者申请授权，本文以API开发者主动授权为例进行介绍。

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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管理页面。
4. 在待绑定应用的APP认证方式API所在行，单击“更多 > 查看授权”，进入API完整信息界面。在“授权信息”页签中，单击“授权”。
5. 在添加授权页面，设置授权的截止时间、选择集群，勾选应用名称，然后单击“确认授权”，完成API的授权。

说明

如果生成API时设置入参位置为Static，则还需设置静态参数值。如果未配置Static参数值，则SDK调用时会使用API入参默认值，API工具调用时会导致缺少Static参数值的报错。

图 13-56 添加授权

6. 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”。

相关操作

批量授权应用：您可以在专享版的“开发API > API管理”页面，勾选需要授权应用的API后，依次单击API列表上方的“批量操作 > 批量授权应用”，实现多个API的批量授权应用。

说明

批量授权应用时，不支持同时授权多种认证类型API。

图 13-57 批量操作

13.3.9.2 通过应用授权 IAM 认证方式 API

IAM认证方式的API当前支持应用和白名单两种授权方式，通过IAM类型应用授权仅能授权给当前账号，而通过白名单授权可授权给任意账号，请您根据使用场景任选一种方式进行授权。

- 通过IAM类型应用授权。IAM类型应用本质上是当前的华为账号，每个DataArts Studio实例下仅能创建一个。因此，将IAM认证方式的API授权给IAM类型的应用，实际上是将API授权给了当前账号。因此在授权后，从IAM服务获取当前账号及其归属用户的Token，在调用API时才能通过安全认证，成功调用API。
- 通过白名单授权。IAM认证方式API支持添加华为账号白名单，将API授权给账号使用。添加白名单授权后，从IAM服务获取的授权账号及其归属用户的Token才能通过安全认证，成功调用API。

本章节为您介绍如何通过IAM类型应用授权将API授权给当前账号。

约束与限制

- 通过IAM应用授权的IAM认证方式API，仅支持通过当前账号及其归属用户的Token进行调用，不支持其他账号及其归属用户调用。如有需要可以通过白名单授权的方式授权给其他账号，详见**通过白名单授权IAM认证方式API**。
- IAM认证方式的API只能授权给IAM类型的应用。

- 如果对无认证方式的API进行应用授权，则系统会忽略此操作。
- IAM类型的应用每个DataArts Studio实例下仅能创建一个，名称固定为华为账号，且不支持修改。
- 专享版中使用IAM认证方式的API必须先通过应用或白名单授权才能调用。

创建 IAM 类型的应用

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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13-20所示信息。

   表 13-20 应用信息

信息项	描述
应用名称	应用名称，IAM应用类型固定为华为账号，且不支持修改。
应用类型	选择IAM应用类型，IAM认证方式的API只能授权给IAM类型的应用。 • IAM：IAM类型应用为IAM认证方式的API进行授权。IAM类型应用为实例级别配置，应用名称固定为华为账号，每个DataArts Studio实例下仅能创建一个，各工作空间之间均可见。 • APP：APP类型应用为APP认证方式的API进行授权。您可以将不同的APP认证方式API授权给不同的应用，提升数据安全性。
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。

将 IAM 认证方式的 API 授权给当前账号

使用IAM认证方式的API，必须将API授权后，才能进行API调用。授权可以分为API开发者主动授权和API调用者申请授权，本文以API开发者主动授权为例进行介绍。

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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管理页面。
4. 在待绑定应用的IAM认证方式API所在行，单击“更多 > 查看授权”，进入API完整信息界面。在“授权信息”页签中，单击“授权”。
5. 在添加授权页面，设置授权的截止时间、选择集群，勾选IAM应用名称，然后单击“确认授权”，完成API的授权。

图 13-58 添加授权

6. 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”。

相关操作

批量授权应用：您可以在专享版的“开发API > API管理”页面，勾选需要授权应用的API后，依次单击API列表上方的“批量操作 > 批量授权应用”，实现多个API的批量授权应用。

说明

批量授权应用时，不支持同时授权多种认证类型API。

图 13-59 批量操作

13.3.9.3 通过白名单授权 IAM 认证方式 API

IAM认证方式的API当前支持应用和白名单两种授权方式，通过IAM类型应用授权仅能授权给当前账号，而通过白名单授权可授权给任意账号，请您根据使用场景任选一种方式进行授权。

- 通过IAM类型应用授权。IAM类型应用本质上是当前的华为账号，每个DataArts Studio实例下仅能创建一个。因此，将IAM认证方式的API授权给IAM类型的应

用，实际上是将API授权给了当前账号。因此在授权后，从IAM服务获取当前账号及其归属用户的Token，在调用API时才能通过安全认证，成功调用API。

- 通过白名单授权。IAM认证方式API支持添加华为账号白名单，将API授权给账号使用。添加白名单授权后，从IAM服务获取的授权账号及其归属用户的Token才能通过安全认证，成功调用API。

本章节为您介绍如何通过白名单授权将API授权给账号。

约束与限制

- 专享版中使用IAM认证方式的API必须先通过应用或白名单授权才能调用。
- 只有IAM认证方式的API才支持白名单授权。

通过白名单授权将 API 授权给账号

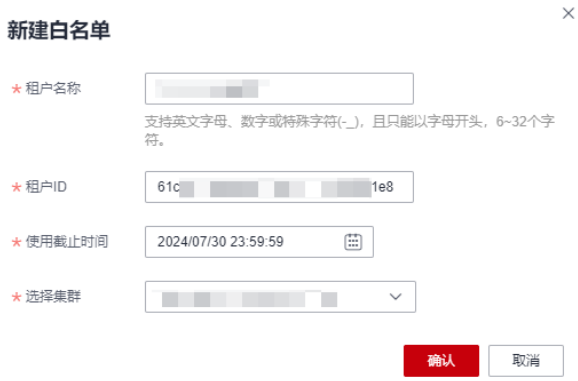
使用IAM认证方式的API，必须将API授权后，才能进行API调用。

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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管理页面。
4. 在待授权给其他华为账号的API所在行，单击“更多 > 查看授权”，进入API完整信息界面。
5. 单击“白名单信息”页签，在“白名单信息”页签中单击“新建”。
6. 在新建白名单窗口，设置需要授权的租户名称、租户ID、授权的截止时间、选择集群，然后单击“确认”，完成IAM认证方式的API针对其他华为账号的授权。

   租户名称和租户ID，需要登录到待授权的账号或其归属用户查看，可以参考如下步骤进行获取，租户名称和租户ID即账号名和账号ID：

   a. 注册并登录管理控制台。
   b. 在用户名的下拉列表中单击“我的凭证”。
   c. 在“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。

图 13-60 新建白名单



7. 授权成功后，则可以在“白名单信息”页签查看已授权的账号。

说明

如果不再需要授权给其他账号，在此租户名称所在行的操作列，单击“删除”，将无需授权的租户账号删除。

相关操作

批量添加白名单：您可以在专享版的“开发API > API管理”页面，勾选需要添加白名单的API后，依次单击API列表上方的“批量操作 > 批量添加白名单”，为多个IAM认证方式的API批量添加白名单。

说明

批量添加白名单时，仅支持IAM认证方式的API。

图 13-61 批量操作

13.4 调用数据服务 API

13.4.1 申请 API 授权

对于API调用者而言，如果API开发者未授权APP或IAM认证方式的API，则需要自行申请API授权，等待审批通过后才能进行API调用。

如果API开发者已完成授权APP或IAM认证方式的API给应用（详见**通过应用授权APP认证方式API**、**通过应用授权IAM认证方式API**或**通过白名单授权IAM认证方式API**），则无需再进行本章的相关操作。

约束与限制

- 专享版中使用IAM认证方式的API必须先通过应用或白名单授权才能调用。
- 申请API授权时，仅支持通过应用授权的方式，暂不支持白名单授权方式。
- APP认证方式的API只能授权给APP类型的应用。
- IAM认证方式的API只能授权给IAM类型的应用。

申请将 API 授权给应用

使用APP或IAM认证方式的API，在将API授权后，才能进行API调用。授权可以分为API开发者主动授权和API调用者申请授权，本文以API调用者申请授权为例进行介绍。

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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 > 服务目录”，可查看到所有已发布的API信息。
4. 单击待绑定应用的API名称，进入API信息页面。
5. 在“调用信息”页面，单击“申请权限”。
6. 在申请权限页面，设置使用截止时间、选择应用，然后单击“确认”，完成API的权限申请。

   说明

   如果生成API时设置入参位置为Static，则还需设置静态参数值。如果未配置Static参数值，则SDK调用时会使用API入参默认值，API工具调用时会导致缺少Static参数值的报错。

图 13-62 申请权限

7. 申请后，需要等待审核中心审核，方可授权成功。
8. 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”。

13.4.2 通过不同方式调用 API

13.4.2.1 调用 API 方式简介

创建API时，有三种认证方式可选，不同认证方式的API支持的调用方式也有所不同，详见表13-21。

表 13-21 API 认证与调用方式说明

认证方式	安全级别	授权与认证机制	支持的调用方式	调用方法示例	使用说明
（推荐）APP认证	高	通过APP应用将API授权给应用后，使用应用的密钥对（AppKey和AppSecret）进行安全认证。	• （推荐）SDK调用：支持Java、Go、Python、JavaScript、C#、PHP、C++、C、Android等多种语言。 • API工具调用：需要通过JavaScript SDK包中的demo.html手动生成签名后，再使用API工具调用。	• **（推荐）通过SDK调用APP认证方式的API** • **通过API工具调用APP认证方式的API**	推荐使用APP认证+SDK调用方式，帮助您简单、快速地通过数据API获取到开放数据。
IAM认证	中	通过IAM应用或白名单将API授权给账号后，借助从IAM服务获取的用户Token进行安全认证。	API工具调用：需要调用IAM服务的**获取用户Token**接口获取Token，再使用API工具调用。	**通过API工具调用IAM认证方式的API**	API工具调用场景可使用IAM认证方式。
无认证	低	无需授权，所有用户均可访问。	• API工具调用：直接调用，无需认证信息。 • 浏览器调用：当API入参位置在Query和Path时，支持浏览器调用。如果入参位置在Header或Body，由于无法传参因此不支持浏览器调用。	• **通过API工具调用无认证方式的API** • **通过浏览器调用无认证方式的API**	无认证方式建议仅在测试接口时使用，不推荐正式使用。若调用方为不可信任用户，则存在数据库安全风险（如数据泄露、数据库高并发访问导致宕机、SQL注入等风险）。

13.4.2.2 （推荐）通过 SDK 调用 APP 认证方式的 API

APP认证方式的API接口可以分别绑定不同的应用，安全级别最高。而APP认证方式的API使用SDK调用方式，支持Java、Go、Python、JavaScript、C#、PHP、C++、C、Android等多种语言，可帮助您简单、快速地通过数据API获取到开放数据。

本章以Java SDK为例，为您介绍如何使用SDK调用APP认证方式的API，主要包含如下几步：

1. **获取APP和API信息**：准备APP和API关键信息，用于API调用。
2. **获取SDK包**：下载SDK包并进行完整性校验。
3. **通过SDK调用API**：修改SDK代码并运行。

前提条件

- 已完成APP认证方式的API或API工作流的发布，在服务目录中可以查看已发布的API。
- 已完成创建应用并将API授权给应用，即API开发者已完成**通过应用授权APP认证方式API**，或API调用者已完成**申请API授权**。
- 本章以Java SDK为例，因此需要已安装Eclipse 3.6.0或以上版本，如果未安装，请至**Eclipse官方网站**下载。

约束与限制

- APP认证方式的API调用前必须先完成**通过应用授权APP认证方式API**或**申请API授权**操作。
- 如需在本地调用专享版API，则需在创建专享版集群时绑定一个弹性公网IP，作为实例的公网入口。
- 调用数据服务API时，如果查询及返回数据的总时长超过默认60秒则会报超时错误。此时可通过访问日志中的API调用时长信息，根据超时阶段进一步优化API配置。

```
____________Duration information_____________
duration: 60491ms //总耗时
url_duration: 0ms //URL匹配耗时
auth_duration: 70ms //鉴权耗时
befor_sql_duration: 402ms //执行SQL预处理耗时
sql_duration: 60001ms //SQL执行耗时
after_sql_duration:18ms //执行SQL后处理耗时
```

获取 APP 和 API 信息

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。

步骤3 在左侧导航栏选择服务版本（例如：专享版），进入总览页。

步骤4 获取API授权应用的AppKey和AppSecret（如已授权多个APP，获取其中一个APP信息即可）。

在左侧导航栏中进入应用管理，找到API授权的应用，并单击应用名称查看APP的完整信息，保存AppKey和AppSecret。

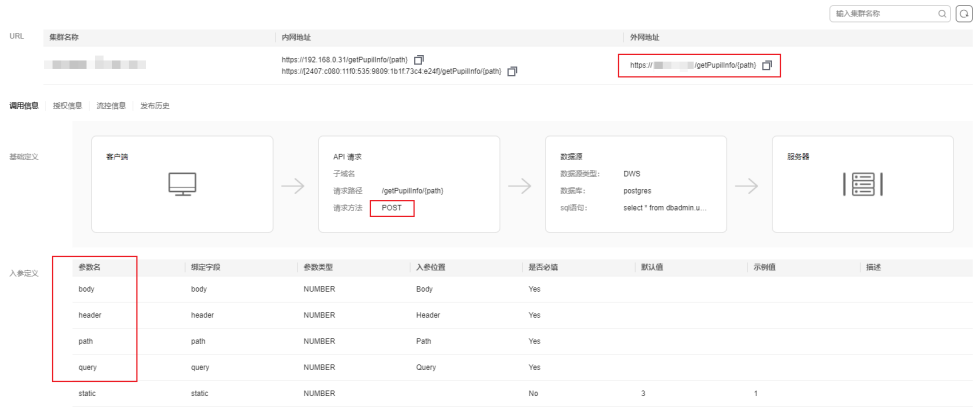
图 13-63 保存 AppKey 和 AppSecret 信息

步骤5 获取待调用API的调用地址、请求方法和入参信息。

在左侧导航栏中进入API管理，找到待调用的API，并单击API名称查看API的完整信息，保存调用地址、请求方法和入参信息。

- 调用地址：专享版支持内网地址和外网地址（外网地址需要您在创建集群时绑定弹性IP），如果需要在本地调用专享版API，需要使用外网地址，确保网络互通。
- 入参：本调用样例中创建了一个具备各类入参位置的API，以便为您介绍各类入参应如何在调用时输入。由于Static是不随API调用者的传值变化的静态参数，因此无需在调用时输入，不需要关注。

图 13-64 保存调用地址、请求方法和入参信息

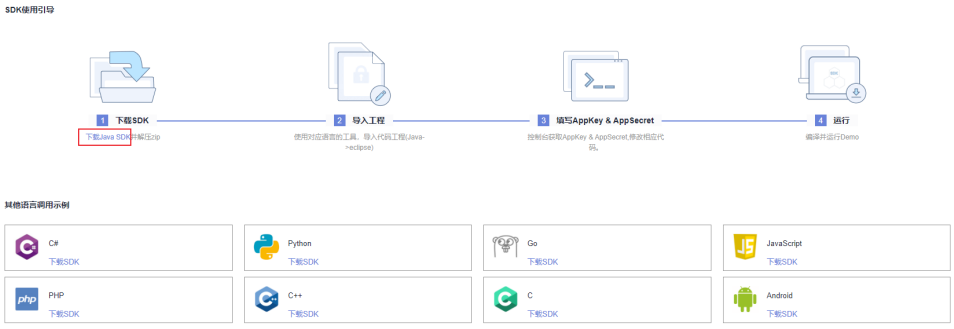


----结束

获取 SDK 包

步骤1 在数据服务页面，单击左侧导航栏的“SDK”，然后下载Java SDK。

图 13-65 下载 SDK



步骤2 进行SDK包完整性校验。Windows操作系统下，打开本地命令提示符框，输入如下命令，在本地生成已下载SDK包的SHA256值，其中，“D:\java-sdk.zip”为SDK包的本地存放路径和SDK包名，请根据实际情况修改。

```
certutil -hashfile D:\java-sdk.zip SHA256
```

命令执行结果示例，如下所示：


```
SHA256 的 D:\java-sdk.zip 哈希:
96fced412700cf9b863cb2d867e6f4edf76480bc679416efab88a9e1912503b9
CertUtil: -hashfile 命令成功完成。
```

对比所下载SDK包的SHA256值和下表中对应语言SDK包的SHA256值。如果一致，则表示下载过程不存在篡改和丢包。

表 13-22 SDK 包及对应的 SHA256 值

不同语言SDK包	SHA256值
Java	96fced412700cf9b863cb2d867e6f4edf76480bc679416efab88a9e1912503b9
Go	f448645da65b4f765d9569fc97ca45dc3e8f1ce4f79d70c5c43934318521d767
Python	54b4984d91db641d2b1b0e77064c162850cb2511a587f95e2f8b8340e7afa128
C#	b66caf856ffccb61fe758872aac08876aa33fb0cf5f4790e3bec163593b2cbae
JavaScript	43da0b54d6b04d1f5ed7f278c2918c2a63a1ddb8048e2d1c5db60baafb17663c
PHP	394c068420a3817f32d5d88b6c1632978f573f2a685e4a1d10c2f698e0f6786e
C++	abae5473d47594f88dcd5eaa0902dc12cd6f1e3bd63c0b82d9d1fab8b4351f54
C	a376573fe8aa3a636a6d123926ddc3dca11748b289b8c2c16a5056830a095acb
Android	c19175d736f05b1945dab4675df19311834ede0d9b1978b11b50c86687baf85c

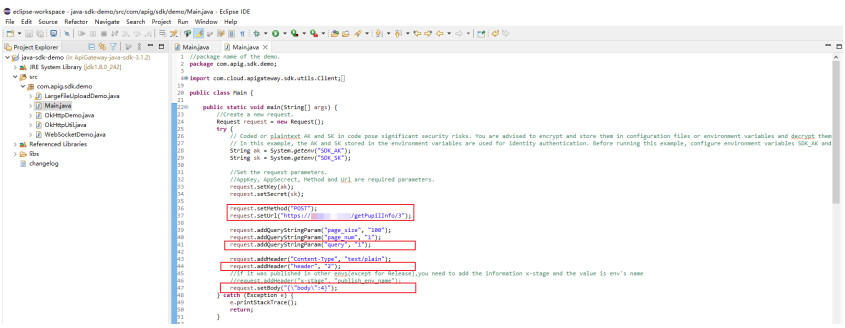
----结束

通过 SDK 调用 API

步骤1 解压**步骤1**中获取的Java SDK包，并在Eclipse中导入SDK工程。

步骤2 导入成功后，打开main.java文件，修改如下图红框所示的内容：

图 13-66 修改 main.java



- 如下参数设置API的请求方法和调用地址，可参考**步骤5**进行获取。

  注意如果入参中包含Path参数，则需要将调用地址中的**{path}**变量修改为具体取值，如下代码中将其取值为3。

  ```
  request.setMethod("POST");
  request.setUrl("https://xx.xx.xx.xx/getPupilInfo/3");
  ```

- 如下参数分别设置Query、Header和Body参数的取值。

  注意Body参数需要使用双引号和大括号**"{}"**将“"***Body参数名***":***Body参数值***”形式的字符串包围在内，且其内字符串中的双引号""需要使用\进行转义。

  ```
  request.addQueryStringParam("query", "1");
  request.addHeader("header", "2");
  request.setBody("{\"body\":4}");
  ```

- （可选）默认情况下，对于配置方式和默认分页的脚本/MyBatis方式API，系统将默认赋值返回量。如果想获取特定分页数据，可以修改如下参数设置分页，其中pageSize表示分页后的页面大小，pageNum表示页码。

  ```
  request.addQueryStringParam("page_size", "100");
  request.addQueryStringParam("page_num", "1");
  ```

  说明

  自定义分页的脚本/MyBatis方式API是在创建API时将分页逻辑写到取数SQL中，因此不支持在调用时修改分页设置。

- （可选）默认情况下，系统会根据排序参数信息给出默认排序情况，自定义排序默认为升序。如果需要修改排序情况，可以修改如下参数设置。其中排序参数描述pre_order_by的值填写形式为“***排序参数参数名***:ASC”或“***排序参数参数名***:DESC”，其中ASC表示升序，DESC表示降序，多个排序参数描述以“英文分号”进行分隔。

  ```
  request.addQueryStringParam("pre_order_by", "id:ASC;age:ASC;score:DESC");
  ```

  对于pre_order_by的值，您可以进行如下修改：

  - 删掉某可选的排序参数，则此排序参数不再参与排序。
  - 修改自定义排序方式的排序参数为升序或降序方式，则此排序参数按照修改后的排序方式排序。

  说明

  pre_order_by的值，不支持进行如下修改，否则会修改不生效或导致调用报错。

  - 删掉某必选的排序参数，则此排序参数依然会正常参与排序，删除不生效。
  - 调整排序参数的前后顺序，则排序依然以配置方式API配置排序参数时的排序参数顺序或脚本/MyBatis方式API SQL中的排序参数顺序为准，调整不生效。
  - 修改升序或降序的排序参数为其他排序方式，则会调用失败，不允许修改。

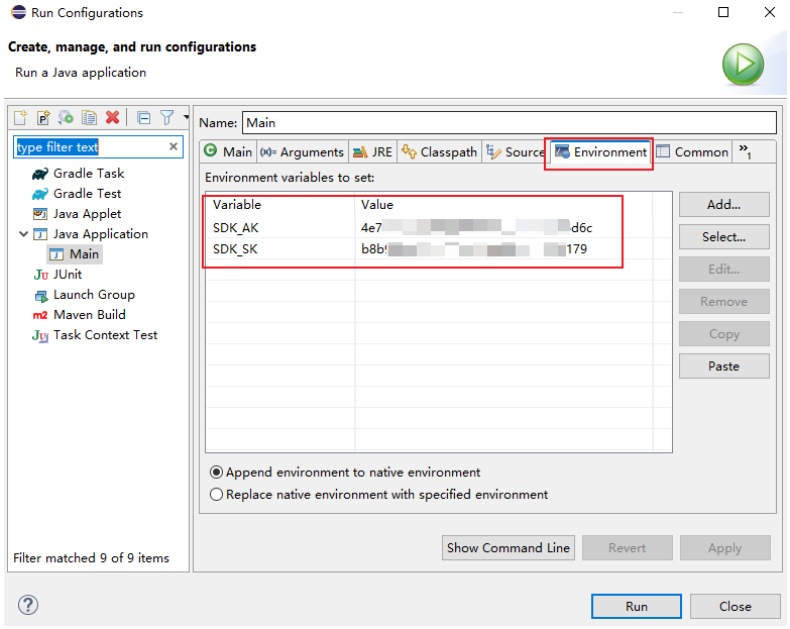
- （可选）在创建API时，如果已打开“返回总条数”开关，则当API对应的数据表数据量较大时，获取数据总条数将会比较耗时。此时，如果需要在调用时不计算并返回数据总条数，可以修改如下参数设置。其中的use_total_num参数用于控制是否计算并返回数据总条数，值为1返回数据总条数，值非1不返回数据总条数。

  ```
  request.addQueryStringParam("use_total_num", "0");
  ```

步骤3 配置AppKey和AppSecret。由于认证用的AppKey和AppSecret编码到代码中或者明文存储都有很大的安全风险，因此建议在配置文件或者环境变量中存放，确保安全，本示例从环境变量中获取。

在Eclipse中，单击菜单栏的“Run > Run Configurations”，在弹窗中选择“Environment”，分别新增变量名SDK_AK和SDK_SK，值分别对应**步骤4**中获取的AppKey和AppSecret。新增完成后，单击“Apply”后再单击“Run”，运行程序。

图 13-67 配置 AppKey 和 AppSecret



步骤4 运行程序后，查看API调用结果。200消息中的**"errCode":"DLM.0"**即表示API调用成功。如果失败，则请根据报错信息进行修复。

图 13-68 运行程序

----结束

13.4.2.3 通过 API 工具调用 APP 认证方式的 API

APP认证方式的API接口可以分别绑定不同的应用，安全级别最高。如果您需要API工具调用APP认证方式的API，则需要先通过JavaScript SDK包中的demo.html手动生成认证信息，再使用API工具调用。

本章节以Postman工具为例，为您介绍如何使用API工具调用APP认证方式的API，主要包含如下几步：

1. **获取APP和API信息**：准备APP和API关键信息，用于API调用。
2. **获取JavaScript SDK包**：下载JavaScript SDK包并进行完整性校验。
3. **生成认证信息**：通过JavaScript SDK包中的demo.html手动生成认证信息。
4. **调用API**：通过Postman工具调用API。

前提条件

- 已完成APP认证方式的API或API工作流的发布，在服务目录中可以查看已发布的API。
- 已完成创建应用并将API授权给应用，即API开发者已完成**通过应用授权APP认证方式API**，或API调用者已完成**申请API授权**。

- 如果API中入参定义了Static参数，则在API授权时已配置Static参数值。
- 本章以Postman工具为例，因此需要已安装Postman工具，如果未安装，请至**Postman官方网站**下载。

约束与限制

- APP认证方式的API调用前必须先完成**通过应用授权APP认证方式API**或**申请API授权**操作。
- 如果API中入参定义了Static参数，则在API授权时应配置Static参数值，否则API工具调用时会导致缺少Static参数值的报错。
- 如需在本地调用专享版API，则需在创建专享版集群时绑定一个弹性公网IP，作为实例的公网入口。
- 通过demo.html手动生成的认证信息，有效期为15分钟，超时则失效。
- 调用数据服务API时，如果查询及返回数据的总时长超过默认60秒则会报超时错误。此时可通过访问日志中的API调用时长信息，根据超时阶段进一步优化API配置。

```
_____Duration information_____
duration: 60491ms //总耗时
url_duration: 0ms //URL匹配耗时
auth_duration: 70ms //鉴权耗时
befor_sql_duration: 402ms //执行SQL预处理耗时
sql_duration: 60001ms //SQL执行耗时
after_sql_duration:18ms //执行SQL后处理耗时
```

获取 APP 和 API 信息

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。

步骤3 在左侧导航栏选择服务版本（例如：专享版），进入总览页。

步骤4 获取API授权应用的AppKey和AppSecret（如已授权多个APP，获取其中一个APP信息即可）。

在左侧导航栏中进入应用管理，找到API授权的应用，并单击应用名称查看APP的完整信息，保存AppKey和AppSecret。

图 13-69 保存 AppKey 和 AppSecret 信息

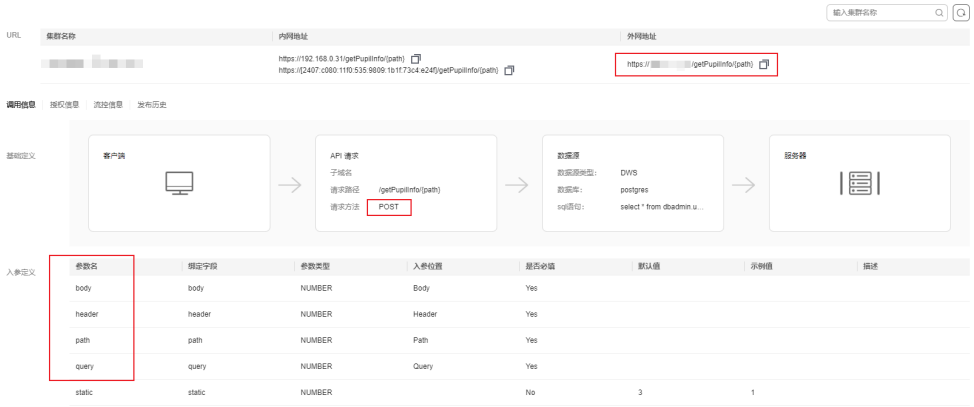
步骤5 获取待调用API的调用地址、请求方法和入参信息。

在左侧导航栏中进入API管理，找到待调用的API，并单击API名称查看API的完整信息，保存调用地址、请求方法和入参信息。

- 调用地址：专享版支持内网地址和外网地址（外网地址需要您在创建集群时绑定弹性IP），如果需要在本地调用专享版API，需要使用外网地址，确保网络互通。

- 入参：本调用样例中创建了一个具备各类入参位置的API，以便为您介绍各类入参应如何在调用时输入。由于Static是不随API调用者的传值变化的静态参数，因此无需在调用时输入，不需要关注。

图 13-70 保存调用地址、请求方法和入参信息

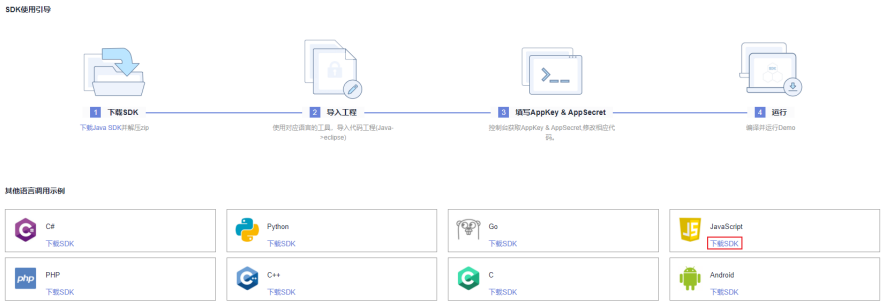


----结束

获取 JavaScript SDK 包

步骤1 在数据服务页面，单击左侧导航栏的“SDK”，然后下载JavaScript SDK。

图 13-71 下载 JavaScript SDK



步骤2 进行SDK包完整性校验。Windows操作系统下，打开本地命令提示符框，输入如下命令，在本地生成已下载SDK包的SHA256值，其中，“D:\javascript-sdk.zip”为SDK包的本地存放路径和SDK包名，请根据实际情况修改。

```
certutil -hashfile D:\javascript-sdk.zip SHA256
```

命令执行结果示例，如下所示：

```
SHA256 的 D:\javascript-sdk.zip 哈希:
43da0b54d6b04d1f5ed7f278c2918c2a63a1ddb8048e2d1c5db60baafb17663c
CertUtil: -hashfile 命令成功完成。
```

对比所下载SDK包的SHA256值和命令示例中的SHA256值。如果一致，则表示下载过程不存在篡改和丢包。

----结束

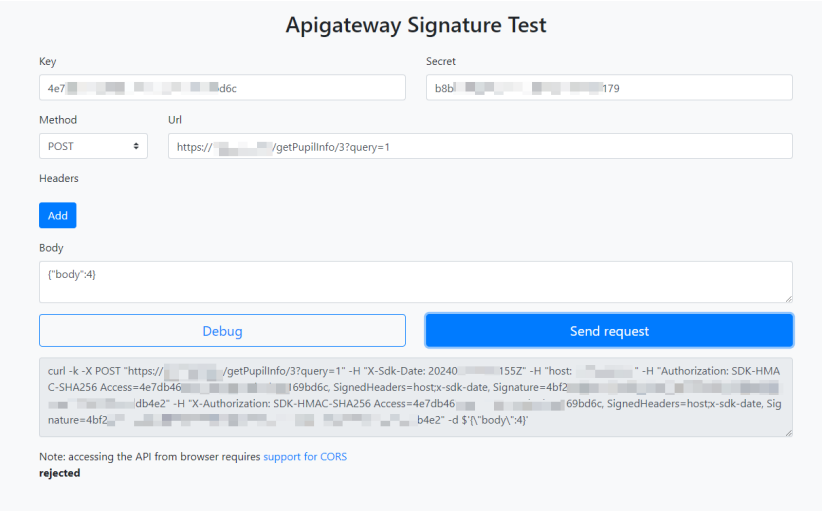
生成认证信息

步骤1 解压SDK包，双击打开其中的“demo.html”文件，输入如下参数后，单击“Send request”查看返回值。

- Key、Secret：API授权应用的AppKey和AppSecret，可参考**获取APP和API信息**获取。
- Method、Url：API的请求方法和调用地址，可参考**获取APP和API信息**获取。

  注意如果入参中包含Path和Query参数，则需要将调用地址中的**{path}**变量修改为Path参数具体取值，Query参数取值以“?***Query参数名*=***Query参数值***”的形式添加到调用地址的最后，如本例中为“?query=1”。
- Headers：Headers参数无需填写，即使已定义Header参数，此处也要保持为空。
- Body：使用大括号**{}**将“"***Body参数名***":***Body参数值***”形式的字符串包围在内，如本例中为“{"body":4}”。

图 13-72 手动生成认证信息



步骤2 从返回值中分别保存X-Sdk-Date、Authorization和X-Authorization的内容，例如本例中需要复制如下内容：

```
...
X-Sdk-Date: 202********55Z
...
Authorization: SDK-HMAC-SHA256 Access=4e7********d6c, SignedHeaders=host;x-sdk-date, Signature=4bf2********4e2
X-Authorization: SDK-HMAC-SHA256 Access=4e72********d6c, SignedHeaders=host;x-sdk-date, Signature=4bf2********4e2
...
```

----结束

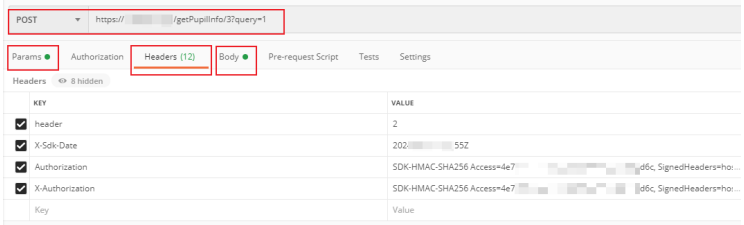
调用 API

步骤1 打开Postman工具，新增一个API请求。

步骤2 API请求配置如下。

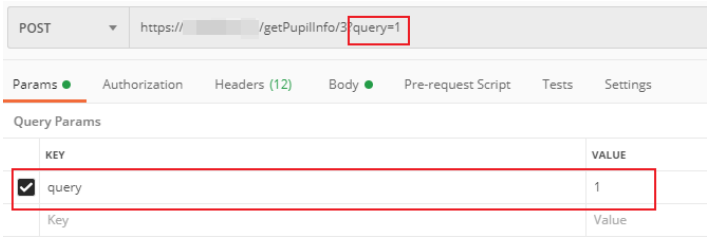
- 请求方法和调用地址：参考**获取APP和API信息**获取，与**生成认证信息**中的请求方法和调用地址保持一致。

图 13-73 请求方法和调用地址



- Params：如果Query参数已经以“?***Query参数名*=*Query参数值***”的形式添加到调用地址的最后，则此处会自动生成Query Params的值，否则就需要手动输入。

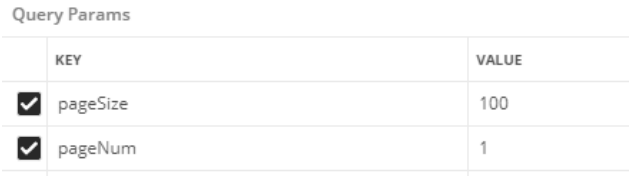
图 13-74 Params



如果您需要对调用结果进行自定义调整，则还可以配置如下Query参数：

– （可选）分页配置：默认情况下，对于配置方式和默认分页的脚本/MyBatis方式API，系统将默认赋值返回量。如果需要获取特定分页数据，您可以修改如下参数设置分页，其中pageSize表示分页后的页面大小，pageNum表示页码。

图 13-75 分页参数设置

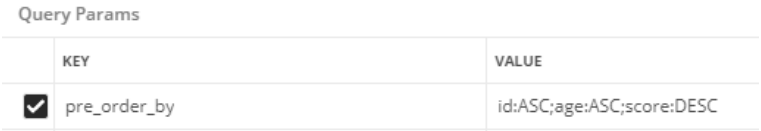


说明

自定义分页的脚本/MyBatis方式API是在创建API时将分页逻辑写到取数SQL中，因此不支持在调用时修改分页设置。

– （可选）排序配置：默认情况下，系统会根据排序参数信息给出默认排序情况，自定义排序默认为升序。如果需要修改排序情况，您可以修改pre_order_by参数。其中排序参数描述pre_order_by的值填写形式为“***排序参数参数名*:ASC**”或“***排序参数参数名*:DESC**”，其中ASC表示升序，DESC表示降序，多个排序参数描述以“英文分号”进行分隔。

图 13-76 排序参数设置



对于pre_order_by的值，您可以进行如下修改：

- 删掉某可选的排序参数，则此排序参数不再参与排序。
- 修改自定义排序方式的排序参数为升序或降序方式，则此排序参数按照修改后的排序方式排序。

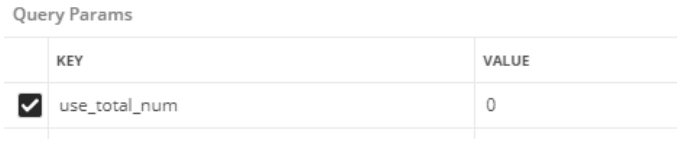
说明

pre_order_by的值，不支持进行如下修改，否则会修改不生效或导致调用报错。

- 删掉某必选的排序参数，则此排序参数依然会正常参与排序，删除不生效。
- 调整排序参数的前后顺序，则排序依然以配置方式API配置排序参数时的排序参数顺序或脚本/MyBatis方式API SQL中的排序参数顺序为准，调整不生效。
- 修改升序或降序的排序参数为其他排序方式，则会调用失败，不允许修改。

– （可选）“返回总条数”配置：在创建API时，如果已打开“返回总条数”开关，则当API对应的数据表数据量较大时，获取数据总条数将会比较耗时。此时，如果需要在调用时不计算并返回数据总条数，可以修改use_total_num参数。use_total_num参数用于控制是否计算并返回数据总条数，值为1返回数据总条数，值非1不返回数据总条数。

图 13-77 “返回总条数”参数配置

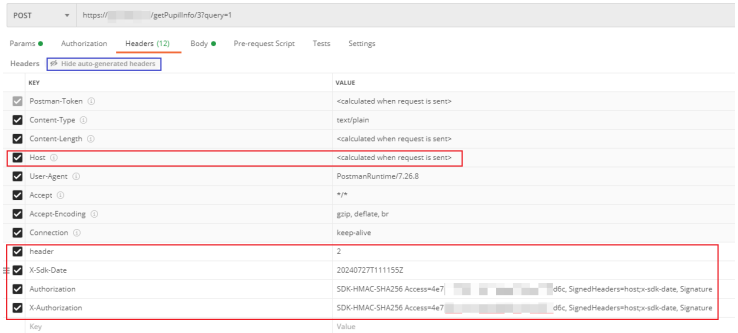


- Headers：将步骤2中保存的X-Sdk-Date、Authorization和X-Authorization及其值依次填入，并将Header参数的参数名和参数值填入其中。

说明

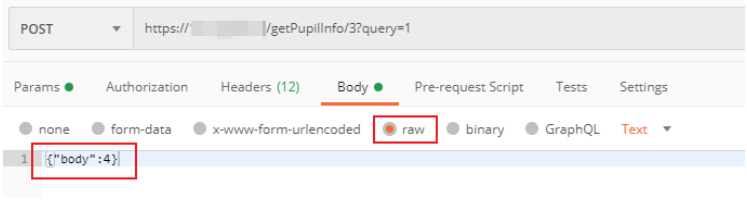
默认情况下，Postman工具会自动勾选Host并从URI中生成Host值，无需手动填写。

图 13-78 Headers



- Body：选择raw格式，使用大括号{}将“"*Body参数名*":*Body参数值*”形式的字符串包围在内，如本例中为“{"body":4}”。

图 13-79 Body



步骤3 API请求配置完成后，单击“Send”发送请求到服务端，然后查看返回结果。返回**"errCode":"DLM.0"**即表示API调用成功。如果失败，则请根据报错信息进行修复。

说明

如调用失败提示“Could not get any response”，可根据提示在Postman设置中关闭“SSL certificate verification”选项或关闭Proxy代理，然后再次尝试运行。

图 13-80 调用 API

----结束

13.4.2.4 通过 API 工具调用 IAM 认证方式的 API

IAM认证方式的API调用前，需要调用IAM服务的**获取用户Token**接口获取Token，然后通过Token进行安全认证。

本章节以Postman工具为例，为您介绍如何使用API工具调用IAM认证方式的API，主要包含如下几步：

1. **获取API信息**：准备API关键信息，用于API调用。
2. **获取Token**：调用IAM服务的**获取用户Token**接口获取Token。
3. **调用API**：通过Postman工具调用API。

前提条件

- 已完成IAM认证方式的API或API工作流的发布，在服务目录中可以查看已发布的API。
- 已完成API授权，即API开发者已完成**通过应用授权IAM认证方式API**或**通过白名单授权IAM认证方式API**，或者API调用者已完成**申请API授权**。

- 本章以Postman工具为例，因此需要已安装Postman工具，如果未安装，请至**Postman官方网站**下载。

约束与限制

- 通过IAM应用授权的IAM认证方式API，仅支持通过当前账号及其归属用户的Token进行调用，不支持其他账号及其归属用户调用。如有需要可以通过白名单授权的方式授权给其他账号，详见**通过白名单授权IAM认证方式API**。
- 如需在本地调用专享版API，则需在创建专享版集群时绑定一个弹性公网IP，作为实例的公网入口。
- Token的有效期为24小时，需要同一个Token鉴权时，可以先缓存起来，避免频繁调用。
- 调用数据服务API时，如果查询及返回数据的总时长超过默认60秒则会报超时错误。此时可通过访问日志中的API调用时长信息，根据超时阶段进一步优化API配置。

```
____________Duration information____________
duration: 60491ms //总耗时
url_duration: 0ms //URL匹配耗时
auth_duration: 70ms //鉴权耗时
befor_sql_duration: 402ms //执行SQL预处理耗时
sql_duration: 60001ms //SQL执行耗时
after_sql_duration:18ms //执行SQL后处理耗时
```

获取 API 信息

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。

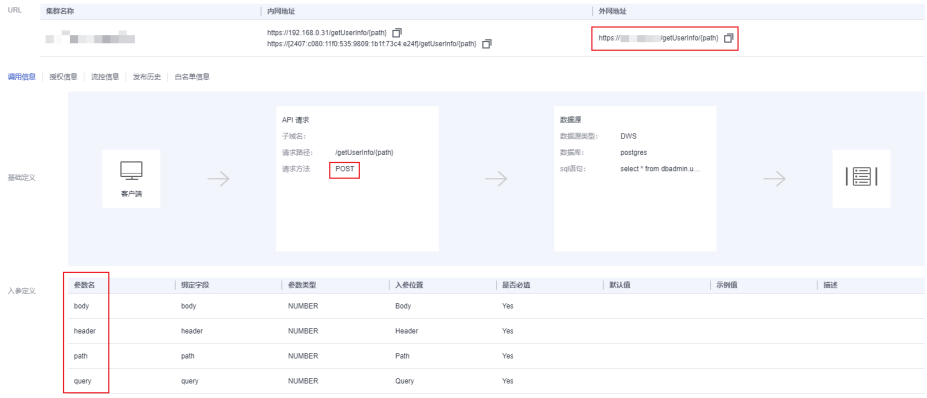
步骤3 在左侧导航栏选择服务版本（例如：专享版），进入总览页。

步骤4 获取待调用API的调用地址、请求方法和入参信息。

在左侧导航栏中进入API管理，找到待调用的API，并单击API名称查看API的完整信息，保存调用地址、请求方法和入参信息。

- 调用地址：专享版支持内网地址和外网地址（外网地址需要您在创建集群时绑定弹性IP），如果需要在本地调用专享版API，需要使用外网地址，确保网络互通。
- 入参：本调用样例中创建了一个具备各类入参位置的API，以便为您介绍各类入参应如何在调用时输入。

图 13-81 保存调用地址、请求方法和入参信息



----结束

获取 Token

步骤1 打开Postman工具，新增一个API请求。

步骤2 使用API工具调用接口获取Token。

Token可通过调用**获取用户Token**接口获取，调用本服务API需要project级别的Token，即调用**获取用户Token**接口时，请求body中auth.scope的取值需要选择project，如下所示。

说明

请求中的“POST https://***IAM endpoint***/v3/auth/tokens”为URL，“Content-Type: application/json”为消息头Header。{}内的内容为请求body体。

注意，请求中加粗的斜体字段需要根据实际值填写：

- ***IAM endpoint***为IAM服务的终端节点。

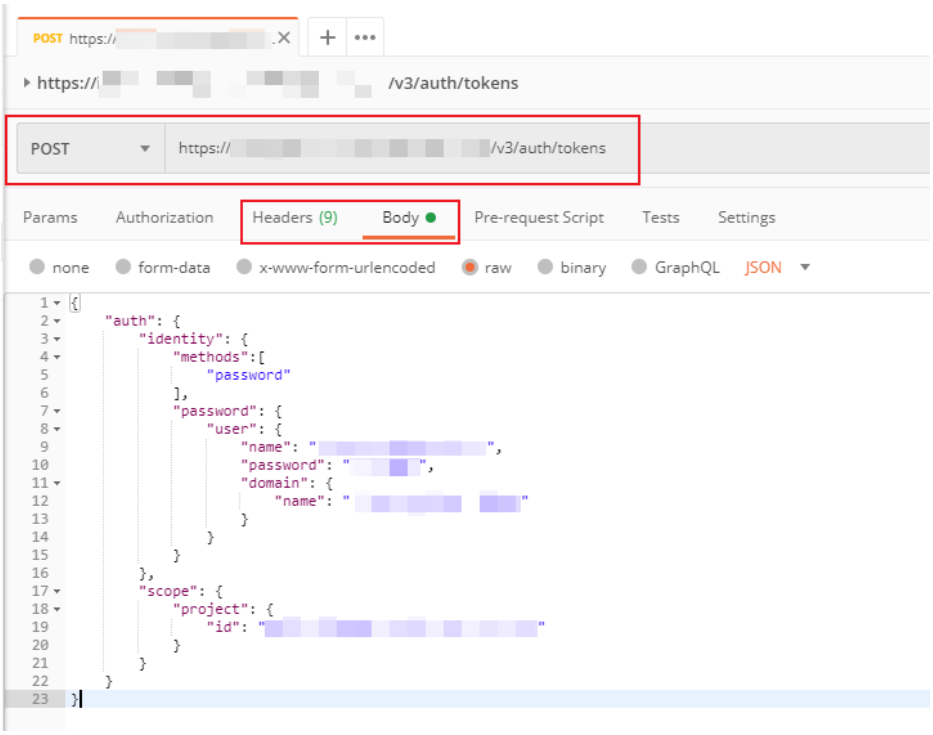
  终端节点（Endpoint）即调用API的**请求地址**，不同服务不同区域的终端节点不同。Endpoint您可以从**地区和终端节点**获取。

- ***username***为用户名，***domainname***为用户所属的账号名，*********为用户登录密码，***xxxxxxxxxxxxxxxxxx***为项目ID。用户名、账号名以及项目ID可以参考如下步骤进行获取：

  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“API凭证”页面，查看账号名和账号ID、IAM用户名和IAM用户ID，在项目列表中查看项目和项目ID。

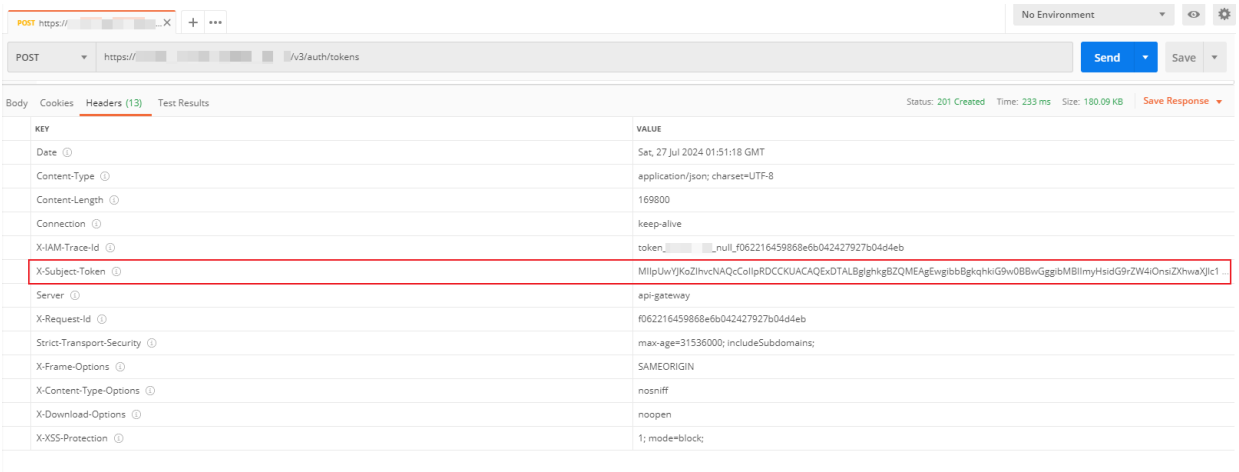
```
POST https://IAM endpoint/v3/auth/tokens
Content-Type: application/json
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username",
                    "password": "********",
                    "domain": {
                        "name": "domainname"
                    }
                }
            }
        },
        "scope": {
            "project": {
                "id": "xxxxxxxxxxxxxxxxxx"
            }
        }
    }
}
```

图 13-82 调用接口获取 Token



步骤3 获取返回的响应消息头Header中“x-subject-token”值，此即为用户Token。有了Token之后，您就可以在调用API的时候将Token加到请求消息头，从而通过身份认证，获得调用API的权限。

图 13-83 获取 Token



----结束

调用 API

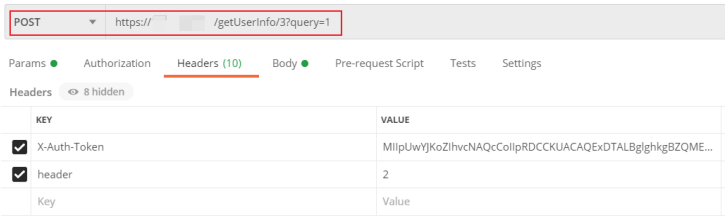
步骤1 打开Postman工具，新增一个API请求。

步骤2 API请求配置如下。

- 请求方法和调用地址：参考**获取API信息**获取，注意如果入参中包含Path和Query参数，则需要将调用地址中的**{path}**变量修改为Path参数具体取值，Query参数

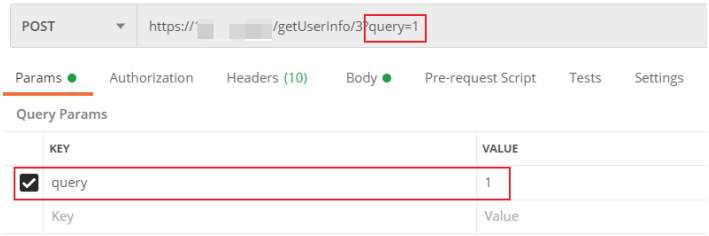
取值可以通过“?***Query参数名*=*Query参数值***”的形式添加到调用地址的最后，如本例中为“?query=1”。

图 13-84 请求方法和调用地址



- Params：如果Query参数已经以“?***Query参数名*=*Query参数值***”的形式添加到调用地址的最后，则此处会自动生成Query Params的值，否则就需要手动输入。

图 13-85 Params



如果您需要对调用结果进行自定义调整，则还可以配置如下Query参数：

– （可选）分页配置：默认情况下，对于配置方式和默认分页的脚本/MyBatis方式API，系统将默认赋值返回量。如果需要获取特定分页数据，您可以修改如下参数设置分页，其中pageSize表示分页后的页面大小，pageNum表示页码。

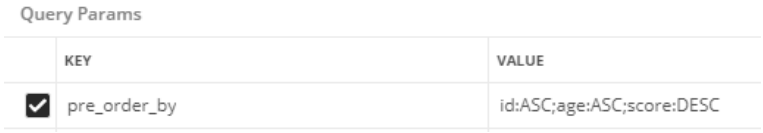
图 13-86 分页参数设置

说明

自定义分页的脚本/MyBatis方式API是在创建API时将分页逻辑写到取数SQL中，因此不支持在调用时修改分页设置。

– （可选）排序配置：默认情况下，系统会根据排序参数信息给出默认排序情况，自定义排序默认为升序。如果需要修改排序情况，您可以修改pre_order_by参数。其中排序参数描述pre_order_by的值填写形式为“***排序参数参数名*:ASC**”或“***排序参数参数名*:DESC**”，其中ASC表示升序，DESC表示降序，多个排序参数描述以“英文分号”进行分隔。

图 13-87 排序参数设置



对于pre_order_by的值，您可以进行如下修改：

- 删掉某可选的排序参数，则此排序参数不再参与排序。
- 修改自定义排序方式的排序参数为升序或降序方式，则此排序参数按照修改后的排序方式排序。

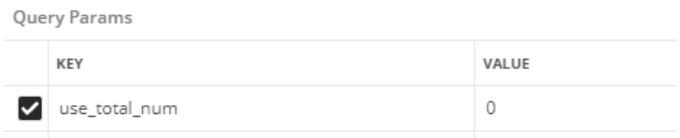
说明

pre_order_by的值，不支持进行如下修改，否则会修改不生效或导致调用报错。

- 删掉某必选的排序参数，则此排序参数依然会正常参与排序，删除不生效。
- 调整排序参数的前后顺序，则排序依然以配置方式API配置排序参数时的排序参数顺序或脚本/MyBatis方式API SQL中的排序参数顺序为准，调整不生效。
- 修改升序或降序的排序参数为其他排序方式，则会调用失败，不允许修改。

– （可选）“返回总条数”配置：在创建API时，如果已打开“返回总条数”开关，则当API对应的数据表数据量较大时，获取数据总条数将会比较耗时。此时，如果需要在调用时不计算并返回数据总条数，可以修改use_total_num参数。use_total_num参数用于控制是否计算并返回数据总条数，值为1返回数据总条数，值非1不返回数据总条数。

图 13-88 “返回总条数”参数配置

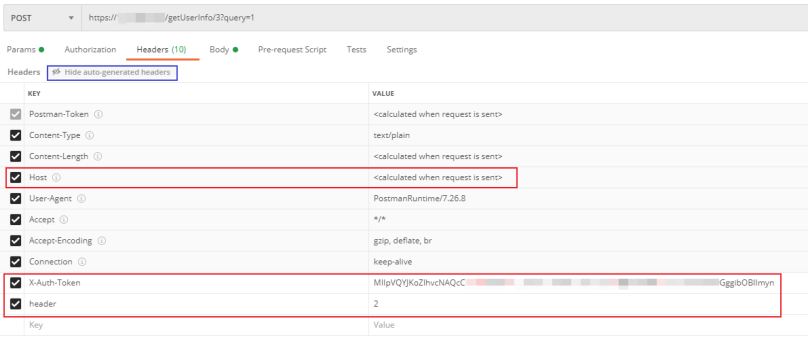


- Headers：将**获取Token**中保存的x-subject-token值填入X-Auth-Token的值中，并将Header参数的参数名和参数值填入其中。

说明

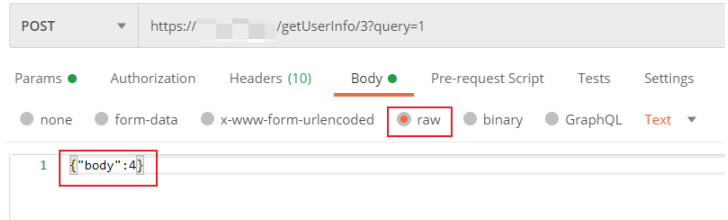
默认情况下，Postman工具会自动勾选Host并从URI中生成Host值，无需手动填写。

图 13-89 Headers



- Body：选择raw格式，使用大括号{}将“"*Body参数名*":*Body参数值*”形式的字符串包围在内，如本例中为“{"body":4}”。

图 13-90 Body



步骤3 API请求配置完成后，单击“Send”发送请求到服务端，然后查看返回结果。返回**"errCode":"DLM.0"**即表示API调用成功。如果失败，则请根据报错信息进行修复。

说明

如调用失败提示“Could not get any response”，可根据提示在Postman设置中关闭“SSL certificate verification”选项或关闭Proxy代理，然后再次尝试运行。

图 13-91 调用 API

----结束

13.4.2.5 通过 API 工具调用无认证方式的 API

无认证方式的API可以通过API工具直接调用，无需获取认证信息。

说明

无认证方式建议仅在测试接口时使用，不推荐正式使用。若调用方为不可信任用户，则存在数据库安全风险（如数据泄露、数据库高并发访问导致宕机、SQL注入等风险）。

本章节以Postman工具为例，为您介绍如何使用API工具调用无认证方式的API，主要包含如下几步：

1. **获取API信息**：准备API关键信息，用于API调用。
2. **调用API**：通过Postman工具调用API。

前提条件

- 已完成无认证方式的API或API工作流的发布，在服务目录中可以查看已发布的API。
- 本章以Postman工具为例，因此需要已安装Postman工具，如果未安装，请至**Postman官方网站**下载。

约束与限制

- 如需在本地调用专享版API，则需在创建专享版集群时绑定一个弹性公网IP，作为实例的公网入口。
- 调用数据服务API时，如果查询及返回数据的总时长超过默认60秒则会报超时错误。此时可通过访问日志中的API调用时长信息，根据超时阶段进一步优化API配置。

```
_____Duration information_____
duration: 60491ms //总耗时
url_duration: 0ms //URL匹配耗时
auth_duration: 70ms //鉴权耗时
befor_sql_duration: 402ms //执行SQL预处理耗时
sql_duration: 60001ms //SQL执行耗时
after_sql_duration:18ms //执行SQL后处理耗时
```

获取 API 信息

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。

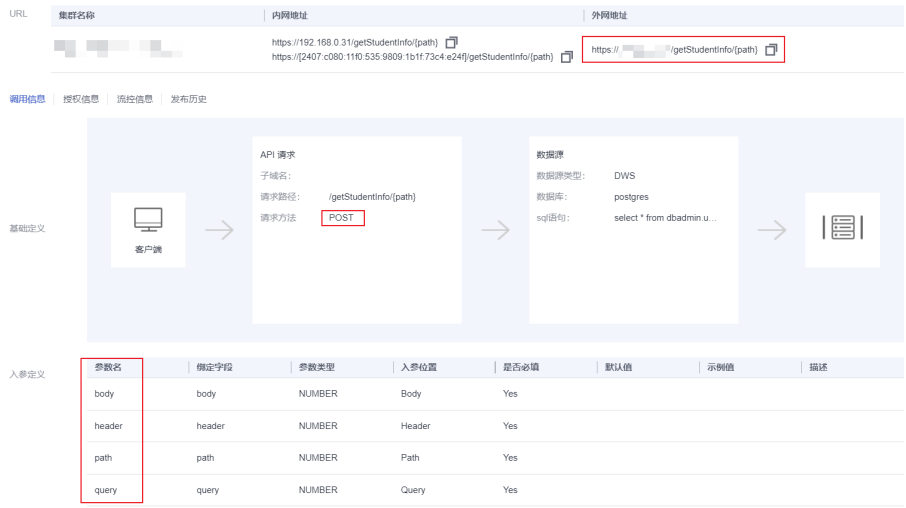
步骤3 在左侧导航栏选择服务版本（例如：专享版），进入总览页。

步骤4 获取待调用API的调用地址、请求方法和入参信息。

在左侧导航栏中进入API管理，找到待调用的API，并单击API名称查看API的完整信息，保存调用地址、请求方法和入参信息。

- 调用地址：专享版支持内网地址和外网地址（外网地址需要您在创建集群时绑定弹性IP），如果需要在本地调用专享版API，需要使用外网地址，确保网络互通。
- 入参：本调用样例中创建了一个具备各类入参位置的API，以便为您介绍各类入参应如何在调用时输入。

图 13-92 保存调用地址、请求方法和入参信息



----结束

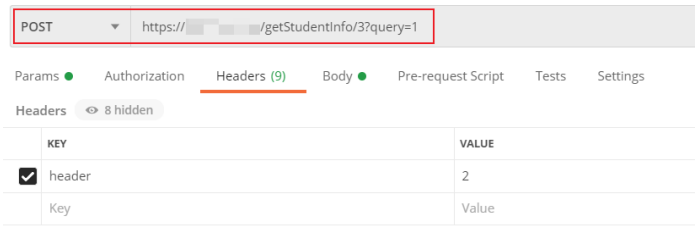
调用 API

步骤1 打开Postman工具，新增一个API请求。

步骤2 API请求配置如下。

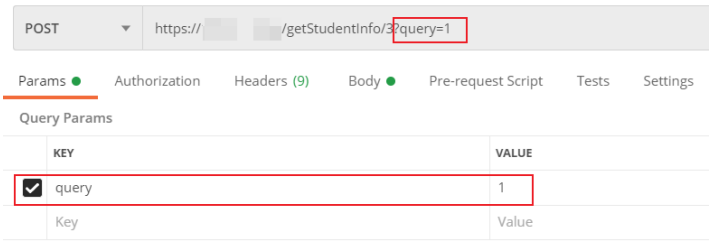
- 请求方法和调用地址：参考**获取API信息**获取，注意如果入参中包含Path和Query参数，则需要将调用地址中的**{path}**变量修改为Path参数具体取值，Query参数取值可以通过“?***Query参数名***=***Query参数值***”的形式添加到调用地址的最后，如本例中为“?query=1”。

  图 13-93 请求方法和调用地址



- Params：如果Query参数已经以“?***Query参数名***=***Query参数值***”的形式添加到调用地址的最后，则此处会自动生成Query Params的值，否则就需要手动输入。

  图 13-94 Params



  如果您需要对调用结果进行自定义调整，则还可以配置如下Query参数：

  – （可选）分页配置：默认情况下，对于配置方式和默认分页的脚本/MyBatis方式API，系统将默认赋值返回量。如果需要获取特定分页数据，您可以修改如下参数设置分页，其中pageSize表示分页后的页面大小，pageNum表示页码。

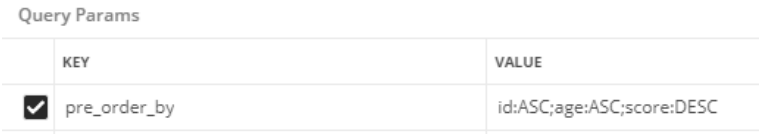
  图 13-95 分页参数设置

说明

自定义分页的脚本/MyBatis方式API是在创建API时将分页逻辑写到取数SQL中，因此不支持在调用时修改分页设置。

- （可选）排序配置：默认情况下，系统会根据排序参数信息给出默认排序情况，自定义排序默认为升序。如果需要修改排序情况，您可以修改pre_order_by参数。其中排序参数描述pre_order_by的值填写形式为“***排序参数参数名***:ASC”或“***排序参数参数名***:DESC”，其中ASC表示升序，DESC表示降序，多个排序参数描述以“英文分号”进行分隔。

图 13-96 排序参数设置



对于pre_order_by的值，您可以进行如下修改：

- 删掉某可选的排序参数，则此排序参数不再参与排序。
- 修改自定义排序方式的排序参数为升序或降序方式，则此排序参数按照修改后的排序方式排序。

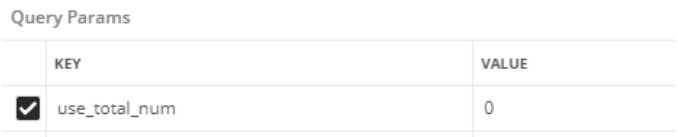
说明

pre_order_by的值，不支持进行如下修改，否则会修改不生效或导致调用报错。

- 删掉某必选的排序参数，则此排序参数依然会正常参与排序，删除不生效。
- 调整排序参数的前后顺序，则排序依然以配置方式API配置排序参数时的排序参数顺序或脚本/MyBatis方式API SQL中的排序参数顺序为准，调整不生效。
- 修改升序或降序的排序参数为其他排序方式，则会调用失败，不允许修改。

- （可选）“返回总条数”配置：在创建API时，如果已打开“返回总条数”开关，则当API对应的数据表数据量较大时，获取数据总条数将会比较耗时。此时，如果需要在调用时不计算并返回数据总条数，可以修改use_total_num参数。use_total_num参数用于控制是否计算并返回数据总条数，值为1返回数据总条数，值非1不返回数据总条数。

图 13-97 “返回总条数”参数配置

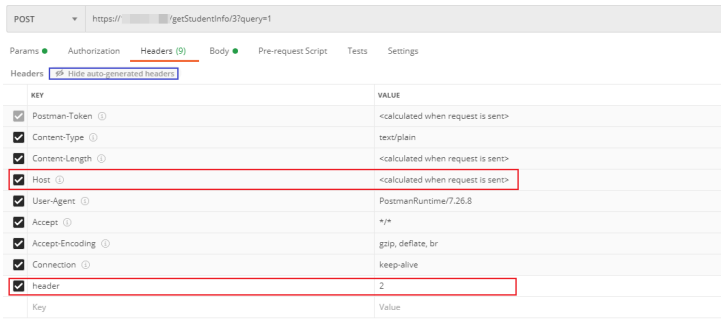


- Headers：将Header参数的参数名和参数值填入其中。

说明

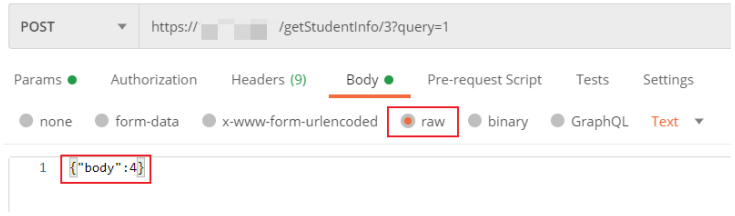
默认情况下，Postman工具会自动勾选Host并从URI中生成Host值，无需手动填写。

图 13-98 Headers



- Body：选择raw格式，使用大括号{}将“"***Body参数名***":***Body参数值***”形式的字符串包围在内，如本例中为“{"body":4}”。

图 13-99 Body



步骤3 API请求配置完成后，单击“Send”发送请求到服务端，然后查看返回结果。返回**"errCode":"DLM.0"**即表示API调用成功。如果失败，则请根据报错信息进行修复。

说明

如调用失败提示“Could not get any response”，可根据提示在Postman设置中关闭“SSL certificate verification”选项或关闭Proxy代理，然后再次尝试运行。

图 13-100 调用 API

----结束

13.4.2.6 通过浏览器调用无认证方式的 API

当无认证方式的API入参位置在Query或Path时，支持直接通过浏览器调用。

> **说明**
>
> 无认证方式建议仅在测试接口时使用，不推荐正式使用。若调用方为不可信任用户，则存在数据库安全风险（如数据泄露、数据库高并发访问导致宕机、SQL注入等风险）。

本章节以Chrome浏览器为例，为您介绍如何使用浏览器调用无认证方式的API，主要包含如下几步：

1. **获取API信息**：准备API关键信息，用于API调用。
2. **调用API**：通过Chrome浏览器调用API。

前提条件

- 已完成无认证方式的API或API工作流的发布，在服务目录中可以查看已发布的API。
- 本章以Chrome浏览器为例，因此需要已安装Chrome浏览器。

约束与限制

- 如需在本地调用专享版API，则需在创建专享版集群时绑定一个弹性公网IP，作为实例的公网入口。
- 调用数据服务API时，如果查询及返回数据的总时长超过默认60秒则会报超时错误。此时可通过访问日志中的API调用时长信息，根据超时阶段进一步优化API配置。

```
___________Duration information__________
duration: 60491ms //总耗时
url_duration: 0ms //URL匹配耗时
auth_duration: 70ms //鉴权耗时
befor_sql_duration: 402ms //执行SQL预处理耗时
sql_duration: 60001ms //SQL执行耗时
after_sql_duration:18ms //执行SQL后处理耗时
```

获取 API 信息

步骤1 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。

步骤2 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。

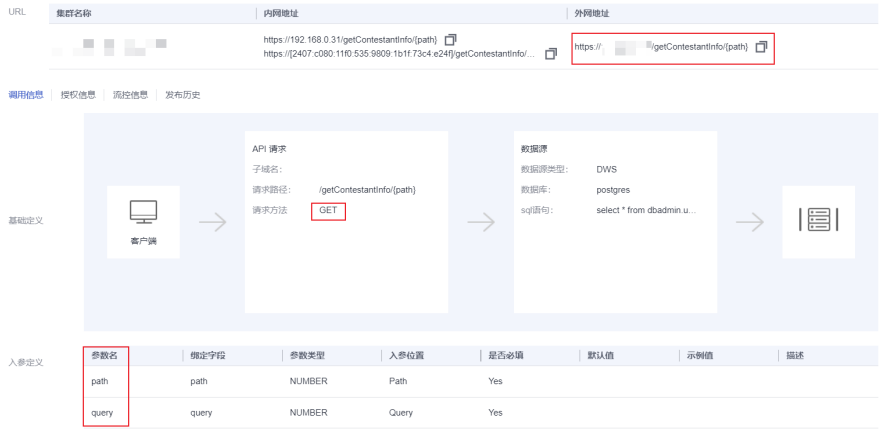
步骤3 在左侧导航栏选择服务版本（例如：专享版），进入总览页。

步骤4 获取待调用API的调用地址、请求方法和入参信息。

在左侧导航栏中进入API管理，找到待调用的API，并单击API名称查看API的完整信息，保存调用地址、请求方法和入参信息。

- 调用地址：专享版支持内网地址和外网地址（外网地址需要您在创建集群时绑定弹性IP），如果需要在本地调用专享版API，需要使用外网地址，确保网络互通。
- 入参：本调用样例中创建了一个具备Query和Path入参位置的API，以便为您介绍入参应如何在调用时输入。

图 13-101 保存调用地址、请求方法和入参信息



----结束

调用 API

步骤1 打开Chrome浏览器，新建一个空白页签。

步骤2 参考**获取API信息**，在浏览器中输入API调用地址并直接访问。注意如果入参中包含Path和Query参数，则需要将调用地址中的**{path}**变量修改为Path参数具体取值，Query参数取值可以通过“?***Query参数名***=***Query参数值***”的形式添加到调用地址的最后，如本例中为“?query=1”。

```
https://xx.xx.xx.xx/getContestantInfo/2?query=1
```

如果您需要对调用结果进行自定义调整，则还可以配置如下Query参数，通过“&”连接多个参数：

- （可选）分页配置：默认情况下，对于配置方式和默认分页的脚本/MyBatis方式API，系统将默认赋值返回量。如果需要获取特定分页数据，您可以添加如下参数设置分页，其中pageSize表示分页后的页面大小，pageNum表示页码。

  ```
  https://xx.xx.xx.xx/getContestantInfo/2?query=1&pageSize=100&pageNum=1
  ```

  说明

  自定义分页的脚本/MyBatis方式API是在创建API时将分页逻辑写到取数SQL中，因此不支持在调用时修改分页设置。

- （可选）排序配置：默认情况下，系统会根据排序参数信息给出默认排序情况，自定义排序默认为升序。如果需要修改排序情况，您可以修改pre_order_by参数。其中排序参数描述pre_order_by的值填写形式为“***排序参数参数名***:ASC”或“***排序参数参数名***:DESC”，其中ASC表示升序，DESC表示降序，多个排序参数描述以“英文分号”进行分隔。

  ```
  https://xx.xx.xx.xx/getContestantInfo/2?query=1&pre_order_by=id:ASC;age:ASC;score:DESC
  ```

  对于pre_order_by的值，您可以进行如下修改：

  - 删掉某可选的排序参数，则此排序参数不再参与排序。
  - 修改自定义排序方式的排序参数为升序或降序方式，则此排序参数按照修改后的排序方式排序。

说明

pre_order_by的值，不支持进行如下修改，否则会修改不生效或导致调用报错。

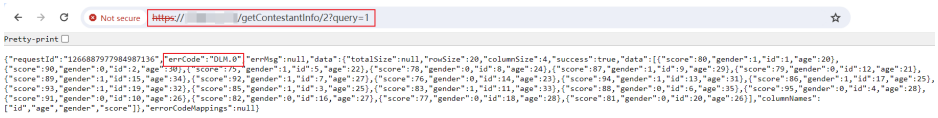
- 删掉某必选的排序参数，则此排序参数依然会正常参与排序，删除不生效。
- 调整排序参数的前后顺序，则排序依然以配置方式API配置排序参数时的排序参数顺序或脚本/MyBatis方式API SQL中的排序参数顺序为准，调整不生效。
- 修改升序或降序的排序参数为其他排序方式，则会调用失败，不允许修改。

- （可选）“返回总条数”配置：在创建API时，如果已打开“返回总条数”开关，则当API对应的数据表数据量较大时，获取数据总条数将会比较耗时。此时，如果需要在调用时不计算并返回数据总条数，可以修改use_total_num参数。use_total_num参数用于控制是否计算并返回数据总条数，值为1返回数据总条数，值非1不返回数据总条数。

```
https://xx.xx.xx.xx/getContestantInfo/2?query=1&use_total_num=0
```

步骤3 浏览器访问后可查看返回结果。返回**"errCode":"DLM.0"**即表示API调用成功。如果失败，则请根据报错信息进行修复。

图 13-102 使用浏览器调用 API



----结束

13.5 查看 API 访问日志

操作场景

支持专享版用户API调用日志的查询，包括请求路径，请求参数，请求响应结果等。

说明

当前数据服务日志仅支持专享版。

配置云日志基础服务

为了能够在云日志服务上面看到数据服务API的日志，需要配置云日志服务。关于云日志服务的详细配置和操作方法，请参见**《云日志服务用户指南》**。

步骤1 在“云日志服务”界面创建日志组。

1. 登录管理控制台。
2. 在管理控制台左上角单击 图标，选择区域和项目。
3. 选择“服务列表 > 管理与监管 > 云日志服务”。

图 13-103 进入云日志服务

4. 单击左侧导航栏“日志管理”。
5. 单击“创建日志组”，在弹出框内，输入日志组名称。
6. 单击“确定”，创建完成。

步骤2 在“云日志服务”界面创建日志流。

1. 选择已创建的日志组名称，进入该日志组页面。
2. 单击“创建日志流”，在弹出框内，输入日志流名称。
3. 单击“确定”，创建完成。

----结束

配置云服务访问日志转储

登录数据服务专享版页面，选择集群，选择日志转储，选择LTS云服务日志。

图 13-104 LTS 转储

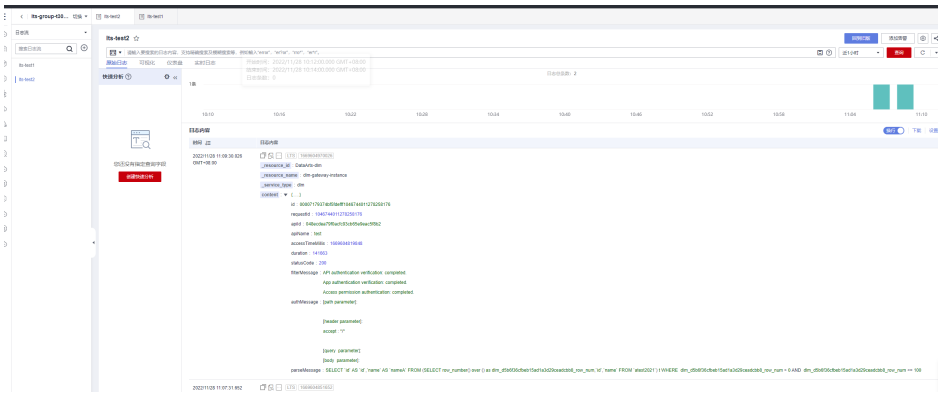
查看访问日志

当您配置了访问日志，可以查看访问日志的详细信息。

通过“云日志服务”控制台，进入日志主题界面，选择相应日志主题名称，在“原始日志”页签下，可查看访问日志。

日志显示格式如下，不支持修改日志格式。

图 13-105 日志显示格式



13.6 配置数据服务审核中心

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的发布等操作。

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。

待审核的API可在审核中心由发起者执行撤销操作。

说明

审核人支持管理员、开发者、运维者，访客无法添加为审核人。
工作空间管理员角色的用户，无论是否被添加为审核人，都默认具备审核人权限。

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，下面以新建审核人为例介绍如何操作。

1. 参考**访问DataArts Studio实例控制台**，登录DataArts Studio管理控制台。
2. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
3. 在左侧导航选择服务版本（例如：专享版），进入总览页。
4. 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 13-106 新建审核人界面

5. 选择审核人（此处的账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。
6. 根据需要，可以添加多个审核人。

审核 API 申请

1. 在DataArts Studio控制台首页，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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. 在待审核任务列表，可通过对应任务操作列的“审核”、或单击API名称进入API信息页面，逐一审核任务；也可以勾选多个审核任务后通列表上方的“批量审核”，统一审核任务。审核后申请立即生效。

图 13-107 审核按钮

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，选择对应工作空间的“数据服务”模块，进入数据服务页面。
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”。

14 审计日志

14.1 如何查看审计日志

概述

云审计服务（Cloud Trace Service，CTS）可以记录DataArts Studio相关的操作事件，用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。

在您开启了云审计服务后，系统开始记录DataArts Studio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。

前提条件

已开通云审计服务。开通方式请参见**开通云审计服务**。

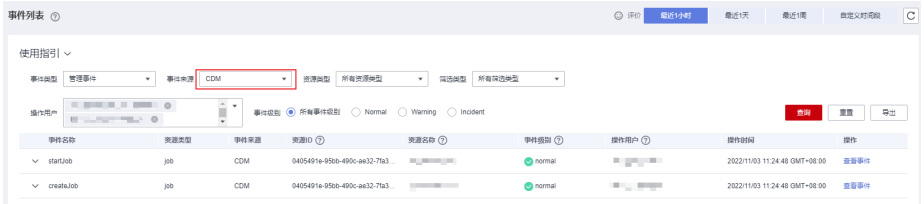
操作步骤

1. 登录管理控制台，在服务列表中选择“云审计服务 CTS”，进入云审计服务控制台。
2. 在云审计服务控制台，默认展示事件列表，您可以通过筛选来查询对应的操作事件。

   其中，DataArts Studio的相关事件在“事件来源”中包含如下分类：

   - CDM：数据集成组件的事件。
   - DLF：数据开发组件的事件。
   - DLG：管理中心、数据架构、数据质量、数据目录和数据服务组件的事件。

图 14-1 CDM 操作事件



3. 在需要查看的事件左侧，单击事件名称左侧箭头 ，展开该记录的详细信息。
4. 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**。

14.2 支持云审计的关键操作

14.2.1 管理中心操作列表

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，可以帮您记录相关的操作事件，便于日后的查询、审计和回溯。

表 14-1 支持云审计的关键操作列表

操作名称	资源类型	事件名称
创建数据连接	dataWarehouse	createDataWarehouse
编辑数据连接	dataWarehouse	updateDataWarehouse
删除数据连接	dataWarehouse	deleteDataWarehouse
创建工作空间	workspace	createWorkspaces
更新工作空间	workspace	updateWorkspaces
删除工作空间	workspace	deleteWorkspaces
冻结工作空间	workspace	frozenWorkspaces
解冻工作空间	workspace	unfrozenWorkspaces
添加工作空间用户	User	saveWorkspaceUser
编辑工作空间用户	User	updateWorkspaceUser
删除工作空间用户	User	deleteWorkspaceUser
下载文件	Config	downloadFile
创建导入导出任务	Config	createObsImportOrExportTask

14.2.2 数据集成操作列表

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，可以帮您记录相关的操作事件，便于日后的查询、审计和回溯。

表 14-2 支持云审计的关键操作列表

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	restartCluster
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
测试连接	link	verifyLink
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

14.2.3 数据架构操作列表

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，可以帮您记录相关的操作事件，便于日后的查询、审计和回溯。

表 14-3 支持云审计的关键操作列表

操作名称	资源类型	资源名称	事件名称
查看主题设计	DAYU_DS	dsSubject	getListSubject
创建主题设计	DAYU_DS	dsSubject	createSubject
更新主题设计	DAYU_DS	dsSubject	updateSubject
发布主题设计	DAYU_DS	dsSubject	publishedSubject
下线主题设计	DAYU_DS	dsSubject	offlineSubject

操作名称	资源类型	资源名称	事件名称
删除主题设计	DAYU_DS	dsSubject	deleteSubject
查看流程设计	DAYU_DS	dsBizCatalog	getListBizCatalog
创建流程设计	DAYU_DS	dsBizCatalog	createBizCatalog
更新流程设计	DAYU_DS	dsBizCatalog	updateBizCatalog
删除流程设计	DAYU_DS	dsBizCatalog	deleteBizCatalog
查看码表管理	DAYU_DS	dsCodeTable	getListCodeTable
创建码表管理	DAYU_DS	dsCodeTable	createCodeTable
更新码表管理	DAYU_DS	dsCodeTable	updateCodeTable
发布码表管理	DAYU_DS	dsCodeTable	publishedCodeTable
下线码表管理	DAYU_DS	dsCodeTable	offlineCodeTable
删除码表管理	DAYU_DS	dsCodeTable	deleteCodeTable
查看数据标准	DAYU_DS	dsStandardElement	getListStandardElement
创建数据标准	DAYU_DS	dsStandardElement	createStandardElement
更新数据标准	DAYU_DS	dsStandardElement	updateStandardElement
发布数据标准	DAYU_DS	dsStandardElement	publishedStandardElement
下线数据标准	DAYU_DS	dsStandardElement	offlineStandardElement
删除数据标准	DAYU_DS	dsStandardElement	deleteStandardElement
查看逻辑实体/物理表	DAYU_DS	dsTableModel	getListTableModel
创建逻辑实体/物理表	DAYU_DS	dsTableModel	createTableModel
更新逻辑实体/物理表	DAYU_DS	dsTableModel	updateTableModel
发布逻辑实体/物理表	DAYU_DS	dsTableModel	publishedTableModel
下线逻辑实体/物理表	DAYU_DS	dsTableModel	offlineTableModel
删除逻辑实体/物理表	DAYU_DS	dsTableModel	deleteTableModel
查看维度	DAYU_DS	dsDimension	getListDimension
创建维度	DAYU_DS	dsDimension	createDimension
更新维度	DAYU_DS	dsDimension	updateDimension

操作名称	资源类型	资源名称	事件名称
发布维度	DAYU_DS	dsDimension	publishedDimension
下线维度	DAYU_DS	dsDimension	offlineDimension
删除维度	DAYU_DS	dsDimension	deleteDimension
查看维度表	DAYU_DS	dsDimensionLogicTable	getListDimensionLogicTable
删除维度表	DAYU_DS	dsDimensionLogicTable	deleteDimensionLogicTable
查看事实表	DAYU_DS	dsFactLogicTable	getListFactLogicTable
创建事实表	DAYU_DS	dsFactLogicTable	createFactLogicTable
更新事实表	DAYU_DS	dsFactLogicTable	updateFactLogicTable
发布事实表	DAYU_DS	dsFactLogicTable	publishedFactLogicTable
下线事实表	DAYU_DS	dsFactLogicTable	offlineFactLogicTable
删除事实表	DAYU_DS	dsFactLogicTable	deleteFactLogicTable
查看汇总表	DAYU_DS	dsAggregationLogicTable	getListAggregationLogicTable
创建汇总表	DAYU_DS	dsAggregationLogicTable	createAggregationLogicTable
更新汇总表	DAYU_DS	dsAggregationLogicTable	updateAggregationLogicTable
发布汇总表	DAYU_DS	dsAggregationLogicTable	publishedAggregationLogicTable
下线汇总表	DAYU_DS	dsAggregationLogicTable	offlineAggregationLogicTable
删除汇总表	DAYU_DS	dsAggregationLogicTable	deleteAggregationLogicTable
查看业务指标	DAYU_DS	dsBizMetric	getListBizMetric
创建业务指标	DAYU_DS	dsBizMetric	createBizMetric
更新业务指标	DAYU_DS	dsBizMetric	updateBizMetric
发布业务指标	DAYU_DS	dsBizMetric	publishedBizMetric
下线业务指标	DAYU_DS	dsBizMetric	offlineBizMetric
删除业务指标	DAYU_DS	dsBizMetric	deleteBizMetric
查看原子指标	DAYU_DS	dsAtomicIndex	getListAtomicIndex
创建原子指标	DAYU_DS	dsAtomicIndex	createAtomicIndex

操作名称	资源类型	资源名称	事件名称
更新原子指标	DAYU_DS	dsAtomicIndex	updateAtomicIndex
发布原子指标	DAYU_DS	dsAtomicIndex	publishedAtomicIndex
下线原子指标	DAYU_DS	dsAtomicIndex	offlineAtomicIndex
删除原子指标	DAYU_DS	dsAtomicIndex	deleteAtomicIndex
查看衍生指标	DAYU_DS	dsDerivativeIndex	getListDerivativeIndex
创建衍生指标	DAYU_DS	dsDerivativeIndex	createDerivativeIndex
更新衍生指标	DAYU_DS	dsDerivativeIndex	updateDerivativeIndex
删除衍生指标	DAYU_DS	dsDerivativeIndex	deleteDerivativeIndex
发布衍生指标	DAYU_DS	dsDerivativeIndex	publishedDerivativeIndex
下线衍生指标	DAYU_DS	dsDerivativeIndex	offlineDerivativeIndex
查看复合指标	DAYU_DS	dsCompoundMetric	getListCompoundMetric
创建复合指标	DAYU_DS	dsCompoundMetric	createCompoundMetric
更新复合指标	DAYU_DS	dsCompoundMetric	updateCompoundMetric
删除复合指标	DAYU_DS	dsCompoundMetric	deleteCompoundMetric
发布复合指标	DAYU_DS	dsCompoundMetric	publishedCompoundMetric
下线复合指标	DAYU_DS	dsCompoundMetric	offlineCompoundMetric
查看时间限定	DAYU_DS	dsTimeCondition	getListTimeCondition
创建时间限定	DAYU_DS	dsTimeCondition	createTimeCondition
更新时间限定	DAYU_DS	dsTimeCondition	updateTimeCondition
发布时间限定	DAYU_DS	dsTimeCondition	publishedTimeCondition
下线时间限定	DAYU_DS	dsTimeCondition	offlineTimeCondition
删除时间限定	DAYU_DS	dsTimeCondition	deleteTimeCondition
查看目录	DAYU_DS	dsDirectory	getListDirectory
创建目录	DAYU_DS	dsDirectory	createDirectory
更新目录	DAYU_DS	dsDirectory	updateDirectory
删除目录	DAYU_DS	dsDirectory	deleteDirectory

操作名称	资源类型	资源名称	事件名称
查看模型	DAYU_DS	dsModel	getListModel
创建模型	DAYU_DS	dsModel	createModel
更新模型	DAYU_DS	dsModel	updateModel
删除模型	DAYU_DS	dsModel	deleteModel

14.2.4 数据开发操作列表

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，可以帮您记录相关的操作事件，便于日后的查询、审计和回溯。

表 14-4 支持云审计的关键操作列表

操作名称	资源类型	事件名称
创建作业	job	createJob(api)
修改作业	job	editJob(api)
保存作业	job	saveJob
删除作业	job	deleteJob
重命名作业	job	renameJob
导入作业	job	importPipeline/ importJob(api)
导出作业	job	exportPipeline/ exportJob(api)
批量导出作业	job	exportJobs(api)
提交作业版本	job	addNewVersion
抢作业锁	job	acquireEditLock
解作业锁	job	releaseLock
批量解作业锁	job	batchReleaseEditLock
测试运行	job	testRun
执行调度	job	startJob
执行调度	job	startJobByName
停止调度	job	stopJob
批量停止调度	job	stopJobs
暂停调度	job	pauseJob

操作名称	资源类型	事件名称
作业复制另存为	job	copyAndSaveJob
批量删除作业	job	deleteDirectoryList
移动作业	job	move
停止实例	task	stopTask/stop(api)
强制成功实例	task	forceTaskSuccess
继续执行实例	task	continueExecute
重跑实例	task	retryTask/restart(api)
节点暂停	task	pauseJob
节点恢复	task	resumeJob
节点手工重试	task	redoJobs
节点跳过	task	skipJob
节点强制成功	task	forceJobSuccess
新建脚本	script	addScript/createScript(api)
执行脚本	script	executeScript
修改脚本	script	saveScript/editScript(api)
导出脚本	script	exportScripts
导入脚本	script	importScript
脚本语法校验	script	checkSyntax
提交脚本版本	script	addNewVersion
抢脚本锁	script	acquireScriptLock
解脚本锁	script	releaseScriptLock
批量解脚本锁	script	batchReleaseScriptLock
批量删除脚本	script	deleteDirectoryList
移动脚本	script	move
创建目录	directory	createDirectory
修改目录	directory	modifyDirectory
删除目录	directory	deleteDirectoryByPath
移动目录	directory	move
批量删除目录	directory	deleteDirectoryList
创建数据连接	dataWarehouse	createDataWarehouse

操作名称	资源类型	事件名称
测试数据连接	dataWarehouse	testDataWarehouseConnectivity
更新数据连接	dataWarehouse	updateDataWarehouse
删除数据连接	dataWarehouse	deleteDataWarehouse
导出数据连接	dataWarehouse	exportConnection
导入数据连接	dataWarehouse	importConnection
创建数据库	dataWarehouse	createDatabase
更新数据库	dataWarehouse	updateDatabase
删除数据库	dataWarehouse	deleteDatabase
创建数据表	dataWarehouse	createDataTable
更新数据表	dataWarehouse	updateDataTable
删除数据表	dataWarehouse	deleteDataTable
创建schema	dataWarehouse	createSchema
删除schema	dataWarehouse	deleteSchema
更新schema	dataWarehouse	updateSchema
创建通知	alarmRule	createAlarmRules
创建并更新通知	alarmRule	createAndUpdateAlarmRules
删除通知	alarmRule	deleteAlarmRules
更新通知	alarmRule	updateAlarmRules
创建资源	dataResource	createResource
更新资源	dataResource	updateResource
删除资源	dataResource	deleteResources
导出资源	dataResource	exportResource
导入资源	dataResource	importResource
批量删除资源	dataResource	deleteDirectoryList
新建标签	tag	create
删除标签	tag	delete
导出标签	tag	exportJobTags
OBS导入标签	tag	importJobTag
本地导入标签	tag	importJobTag2

操作名称	资源类型	事件名称
保存环境变量	environmentVariable	saveEnvParams
删除环境变量	environmentVariable	deleteEnvParams
导出环境变量	environmentVariable	exportEnvParams
导入环境变量	environmentVariable	importEnvParams
更新空间配置项	workspaceConfig	updateWorkSpaceConfigs
上传文件	file	uploadFile
配置空间委托	agency	saveAgency
保存敏感变量	sensitiveParam	saveSensitiveParam
更新敏感变量	sensitiveParam	updateSensitiveParam
删除敏感变量	sensitiveParam	deleteSensitiveParam
新建cdm连接	createConnection	cdmConnection
更新cdm连接	updateConnection	cdmConnection
删除cdm连接	deleteConnection	cdmConnection
发送httpTrigger消息	sendMessage	httpTriggerMessage

14.2.5 数据质量操作列表

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，可以帮您记录相关的操作事件，便于日后的查询、审计和回溯。

表 14-5 支持云审计的关键操作列表

操作名称	资源类型	事件名称
创建目录	Category	createCategory
删除目录	Category	deleteCategory
更新目录	Category	updateCategory
批量停止	Instance	batchStop
批量删除	Instance	batchDeleteInstances
创建对账作业	ConsistencyTask	createConsistencyTask
批量删除对账作业	ConsistencyTask	batchDeleteConsistencyTask
编辑对战作业	ConsistencyTask	editConsistencyTask
启动调度对账作业	ConsistencyTask	startScheduleConsistencyTask

操作名称	资源类型	事件名称
停止对账作业	ConsistencyTask	stopScheduleConsistencyTask
运行对账作业	ConsistencyTask	runConsistencyTask
创建质量作业	Rule	createRuleTask
删除质量作业	Rule	deleteRule
更新质量作业	Rule	updateRule
运行质量作业	Rule	instanceScheduleOperation
批量运行质量作业	Rule	batchInstanceScheduleOperation
批量操作质量作业	Rule	batchOperateRules
创建规则模板	RuleTemplate	createTemplate
删除规则模板	RuleTemplate	deleteTemplate
查询规则模板列表	RuleTemplate	getRuleTemplateList
更新规则模板	RuleTemplate	updateTemplate
查询规则模板	RuleTemplate	getTemplate
获取依赖规则模板的质量作业和对账作业	RuleTemplate	getDependentTasks
批量更新作业的规则模板	RuleTemplate	batchUpdateDependentTasks

14.2.6 数据目录操作列表

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，可以帮您记录相关的操作事件，便于日后的查询、审计和回溯。

表 14-6 支持云审计的关键操作列表

操作名称	资源类型	事件名称
添加数据掩码	datamask	createDataMask
查询数据掩码列表	datamask	listDataMask
查询数据掩码	datamask	getDataMask
删除数据掩码	datamask	deleteDataMask
批量删除数据掩码	datamask	batchDeleteDataMask
修改数据掩码	datamask	updateDataMask

操作名称	资源类型	事件名称
配置采集任务并运行	bridgetask	createBridgeTask
查询采集任务列表	bridgetask	getBridgeTask
编辑采集任务	bridgetask	updateBridgeTask
批量删除采集任务	bridgetask	batchDeleteBridgeTask
数据资产添加标签	asset	addTagToAsset
添加标签	tag	createTag
批量添加标签	tag	batchCreateTag
批量删除标签	tag	batchDeleteTag
修改标签	tag	updateTag
查询标签列表	tag	getTags
删除标签	tag	deleteTag
新建任务目录	bridgetaskcategory	createBridgeTaskCategory
获取任务目录列表	bridgetaskcategory	getBridgeTaskCategoryTree
编辑任务目录	bridgetaskcategory	updateBridgeTaskCategory
删除任务目录	bridgetaskcategory	deleteBridgeTaskCategory
创建分类分组	classificationgroup	createClassificationGroup
查询分类分组列表	classificationgroup	listClassificationGroup
查询分类分组	classificationgroup	getClassificationGroup
批量删除分组	classificationgroup	batchDeleteClassificationGroup
修改分类分组	classificationgroup	updateClassificationGroup
创建分类规则	classificationrule	createClassificationRule
查询分类规则列表	classificationrule	listClassificationRule
查询分类规则	classificationrule	getClassificationRule
批量删除分类规则	classificationrule	batchDeleteClassificationRule
修改分类规则	classificationrule	updateClassificationRule
创建数据密级	secrecylevel	createSecrecyLevel
查询数据密级列表	secrecylevel	listSecrecyLevel
查询数据密级	secrecylevel	getSecrecyLevel
批量删除数据密级	secrecylevel	batchDeleteSecrecyLevel

操作名称	资源类型	事件名称
修改数据密级	secrecylevel	updateSecrecyLevel
创建采集任务	bridgetask	createBridgeTask
编辑采集任务	bridgetask	updateBridgeTask
删除采集任务	bridgetask	deleteBridgeTask
查询采集任务列表	bridgetask	getTasks

14.2.7 数据服务操作列表

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，可以帮您记录相关的操作事件，便于日后的查询、审计和回溯。

表 14-7 支持云审计的关键操作列表

操作名称	资源类型	事件名称
创建API	DLMApi	createApi
更新API	DLMApi	updateApi
查询API	DLMApi	getApi
查询API列表	DLMApi	getApiList(Api)
删除API	DLMApi	deleteApi
发布API	DLMApi	publishApi
下线API	DLMApi	unpublishApi
续约API	DLMApi	renewApi
停用API	DLMApi	stopApi
恢复API	DLMApi	recoverApi
复制API	DLMApi	copyApi
操作API	DLMApi	actionApi
创建APP	DLMApp	createApp
更新APP	DLMApp	updateApp
删除APP	DLMApp	deleteApp
查询APP	DLMApp	getApp
查询APP详情	DLMApp	getAppInfo
授权API	DLMRelation	authorizeApi

操作名称	资源类型	事件名称
查询已授权的应用	DLMRelation	getAuthorizeApp
取消授权	DLMRelation	cancelApprovalApi
查询未授权的应用	DLMRelation	getLeftApp
申请API	DLMApply	applyApi
取消申请	DLMApply	revokeApply
获取申请列表	DLMApply	getApplyList
获取申请详情	DLMApply	getApplyDetail
获取通知详情	DLMApply	getMessageDetail
创建申请	DLMApply	createApply
批量审核申请	DLMApply	batchApproveNewApply
发送通知	DLMApply	sendMesg
获取通知列表	DLMApply	getMessageList
获取发布趋势	DLMApply	getPublishTrend
创建流控策略	DLMFlowControl	createFlowControlStrategy
更新流控策略	DLMFlowControl	updateFlowControlStrategy
删除流控策略	DLMFlowControl	deleteFlowControlStrategy
查询流控策略	DLMFlowControl	getFlowControlStrategy
查询API列表（流控相关）	DLMFlowControlBindApi	getAllApiList
查询已绑定的API列表	DLMFlowControlBindApi	getBindingApiList
绑定API	DLMFlowControlBindApi	bindingApi
解绑API	DLMFlowControlBindApi	unBindingApi
查询统计用户相关的总览开发指标	DLMRequestRecord	getApisOverview
查询统计用户相关的总览调用指标	DLMRequestRecord	getAppsOverView
查询api 服务调用topN	DLMRequestRecord	getApisTop
查询app 服务使用topN	DLMRequestRecord	getAppsTop
查询api 统计数据详情	DLMRequestRecord	getApisDetail
查询app 统计数据详情	DLMRequestRecord	getAppsDetail
查询api 仪表板数据详情	DLMRequestRecord	getApisDashboard

操作名称	资源类型	事件名称
查询app 仪表板数据详情	DLMRequestRecord	getAppsDashboard
查询api 服务异常调用topN	DLMRequestRecord	getApisError
查询支持的数据源类型	DLMDataSourceType	getDatasources
查询数据连接	DLMDataSourceConnection	getDatasourceConnections
查询数据库	DLMDataSourceDatabase	getDatasourcedatabases
查询数据库表	DLMDataSourceTable	getDatasourcedatables
查询数据库表的字段	DLMDataSourceTableField	getDatasourceTableFields
查询数据源队列（DLI）	DLMDataSourceQueue	getQueue
查询有权成为审核人的用户	DLMAuthorizedUser	getAuthorizedUser
创建审核人	DLMApprover	createApprover
删除审核人	DLMApprover	deleteApprover
查询审核人	DLMApprover	getApproverList
查询服务目录下的所有内容	DLMServiceCatalog	getCatalogAllDetail
查询服务目录下的api	DLMServiceCatalog	getCatalogApis
查询服务目录下的目录	DLMServiceCatalog	getCatalogCatalogs
创建服务目录	DLMServiceCatalog	createCatalog
删除服务目录	DLMServiceCatalog	deleteCatalog
更新服务目录	DLMServiceCatalog	updateCatalog
查询服务目录详情	DLMServiceCatalog	getCatalogDetail
移动服务目录	DLMServiceCatalog	moveCatalog
移动API	DLMServiceCatalog	moveApi
获取标签列表	DLMTag	getTags
获取本地标签列表	DLMTag	getLocalTags
更新标签列表	DLMTag	updateTags

