数据治理中心

快速入门

文档版本 01

发布日期 2023-05-22

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 概述

数据治理中心DataArts Studio是具备数据全生命周期管理和智能数据管理能力的一站式数据治理平台，基于数据湖底座，提供数据集成、开发、治理等能力。针对不同的使用场景，我们提供了不同的使用方案示例：

表 1-1 快速入门案例定位

使用案例示例	所选数据湖底座	所需版本	业务能力	适用场景
入门使用者：基于电商BI报表的数据开发流程	DLI	• 免费版 • 初级版	数据开发	对数据全生命周期管理的流程诉求较低，需要全托管的大数据调度能力，适用于开发者试用、小规模验证等场景。
初级使用者：基于电影评分的数据集成与开发流程	DWS	• 免费版 • 初级版	数据集成+数据开发	用于大数据开发场景的数据ETL任务管理等场景，但不涉及数据治理，适用于开发者试用、小规模验证等场景。
高级使用者：基于出租车出行的数据治理流程	MRS Hive	• 企业版	数据集成+数据开发+数据治理	需求全功能数据治理能力，适用于有完善的数据管理团队和体系，要进行企业信息架构、数据标准、数据模型、数据指标的落地，匹配完整的DAYU数据治理方法论。

2 入门使用者：基于电商 BI 报表的数据开发流程

2.1 示例场景说明

本实践通过DataArts Studio服务的数据开发DLF组件和数据湖探索服务（DLI）对某电商商城的用户、商品、评论数据（脱敏后）进行分析，输出用户和商品的各种数据特征，可为营销决策、广告推荐、信用评级、品牌监控、用户行为预测提供高质量的信息。在此期间，您可以学习到数据开发模块脚本编辑、作业编辑、作业调度等功能，以及DLI的SQL基本语法。

说明

本入门示例涉及DataArts Studio服务的管理中心和数据开发模块，DataArts Studio的各版本均可以满足使用要求。

如果您从未使用过DataArts Studio，您可以选择**试用DataArts Studio**，按照本示例进行入门试用。

操作流程如下：

1. 准备工作，包括**使用DataArts Studio前的准备**、**数据源准备**和**数据湖准备**。
2. 数据开发，包含创建DLI SQL脚本和开发作业。
   - **分析10大用户关注最多的产品**
   - **分析10大用户评价最差的商品**
   - **开发并调度作业**，通过编排作业和配置作业调度策略，定期执行作业，使得用户可以每天获取到最新的数据分析结果。
3. **服务退订**，如果不再使用DataArts Studio相关服务，请及时进行退订和资源删除。

2.2 步骤 1：准备工作

使用 DataArts Studio 前的准备

如果您是第一次使用DataArts Studio，请参考**准备工作**章节完成注册华为账号、购买DataArts Studio实例（DataArts Studio企业版）、创建工作空间等一系列操作。然后进入到对应的工作空间，即可开始使用DataArts Studio。

数据源准备

本入门示例以某电商商城的BI报表数据为例，分析用户和商品的各种数据特征。

为方便演示，本示例提供了用于模拟原始数据的部分数据。为了方便将源数据集成到云上，我们需要先将样例数据存储为CSV文件，将CSV文件上传至OBS服务中。

步骤1 创建CSV文件（UTF-8无bom格式），文件名称为对应的数据表名，将后文提供的各样例数据分别复制粘贴到不同CSV文件中，然后保存CSV文件。

以下是Windows下生成.csv文件的办法之一：

1. 使用文本编辑工具（例如记事本等）新建一个txt文档，将后文提供的样例数据复制进文档中。注意复制后检查数据的行数及数据分行的正确性（注意，如果是从PDF文档中复制样例数据，单行的数据过长时会产生换行，需手动重新调整为单行）。
2. 单击“文件 > 另存为”，在弹出的对话框中，“保存类型”选择为“所有文件(*.*)”，在“文件名”处输入文件名和.csv后缀，选择“UTF-8”编码格式（不能带BOM），则能以CSV格式保存该文件。

步骤2 将源数据CSV文件上传到OBS服务。

1. 登录控制台，选择“存储 > 对象存储服务 OBS”，进入OBS控制台。
2. 单击“创建桶”，然后根据页面提示配置参数，创建一个名称为“fast-demo”的OBS桶。

   说明

   为保证网络互通，OBS桶区域请选择和DataArts Studio实例相同的区域。如果需要选择企业项目，也请选择与DataArts Studio实例相同的企业项目。

   使用OBS控制台创建桶的操作，请参见《对象存储服务控制台指南》中的**创建桶**。
3. 在名称为“fast-demo”的OBS桶中，创建user_data、product_data、comment_data和action_data的文件夹，分别将user_data.csv、product_data.csv、comment_data.csv和action_data.csv文件上传数据到对应文件夹中。

   说明

   由于DLI在关联CSV表格用于创建OBS外表时，不支持指定文件名、仅支持指定文件路径，因此需要将CSV表格分别放到不同的文件路径下，且确保文件路径下仅包含所需的CSV表格。

   使用OBS控制台上传文件的操作，请参见《对象存储服务控制台指南》中的**上传文件**。

----结束

本示例中涉及到4部分原始样例数据，分别为用户数据**user_data.csv**、商品数据**product_data.csv**、评价数据**comment_data.csv**和行为数据**action_data.csv**。具体数据和说明如下：

- user_data.csv：

```
user_id,age,sexuality,rank,register_time
100001,20,0,1,2021/1/1
100002,22,1,2,2021/1/2
100003,21,0,3,2021/1/3
100004,24,2,5,2021/1/4
100005,50,2,9,2021/1/5
100006,20,1,3,2021/1/6
100007,18,1,1,2021/1/7
100008,20,1,6,2021/1/8
100009,60,0,4,2021/1/9
100010,20,1,1,2021/1/10
100011,35,0,5,2021/1/11
100012,20,1,1,2021/1/12
100013,7,0,1,2021/1/13
100014,64,0,8,2021/1/14
100015,20,1,1,2021/1/15
100016,33,1,7,2021/1/16
100017,20,0,1,2021/1/17
100018,15,1,1,2021/1/18
100019,20,1,9,2021/1/19
100020,33,0,1,2021/1/20
100021,20,0,1,2021/1/21
100022,22,1,5,2021/1/22
100023,20,1,1,2021/1/23
100024,20,0,1,2021/1/24
100025,34,0,7,2021/1/25
100026,34,1,1,2021/1/26
100027,20,1,8,2021/1/27
100028,20,0,1,2021/1/28
100029,56,0,5,2021/1/29
100030,20,1,1,2021/1/30
100031,22,1,8,2021/1/31
100032,20,0,1,2021/2/1
100033,32,1,0,2021/2/2
100034,20,1,1,2021/2/3
100035,45,0,6,2021/2/4
100036,20,0,1,2021/2/5
100037,67,1,4,2021/2/6
100038,78,0,6,2021/2/7
100039,11,1,8,2021/2/8
100040,8,0,0,2021/2/9
```

数据说明如下：

表 2-1 用户数据说明

字段名称	字段类型	字段说明	字段取值
user_id	int	用户ID	脱敏
age	int	年龄段	-1表示未知
sexuality	int	性别	• 0表示男 • 1表示女 • 2表示保密
rank	Int	用户等级	有顺序的级别枚举，越高级别数字越大

字段名称	字段类型	字段说明	字段取值
register_time	string	用户注册日期	单位：天

- product_data.csv：

```
product_id,a1,a2,a3,category,brand
200001,1,1,1,300001,400001
200002,2,2,2,300002,400001
200003,3,3,3,300003,400001
200004,1,2,3,300004,400001
200005,3,2,1,300005,400002
200006,1,1,1,300006,400002
200007,2,2,2,300007,400002
200008,3,3,3,300008,400002
200009,1,2,3,300009,400003
200010,3,2,1,300010,400003
200011,1,1,1,300001,400003
200012,2,2,2,300002,400003
200013,3,3,3,300003,400004
200014,1,2,3,300004,400004
200015,3,2,1,300005,400004
200016,1,1,1,300006,400004
200017,2,2,2,300007,400005
200018,3,3,3,300008,400005
200019,1,2,3,300009,400005
200020,3,2,1,300010,400005
200021,1,1,1,300001,400006
200022,2,2,2,300002,400006
200023,3,3,3,300003,400006
200024,1,2,3,300004,400006
200025,3,2,1,300005,400007
200026,1,1,1,300006,400007
200027,2,2,2,300007,400007
200028,3,3,3,300008,400007
200029,1,2,3,300009,400008
200030,3,2,1,300010,400008
200031,1,1,1,300001,400008
200032,2,2,2,300002,400008
200033,3,3,3,300003,400009
200034,1,2,3,300004,400009
200035,3,2,1,300005,400009
200036,1,1,1,300006,400009
200037,2,2,2,300007,400010
200038,3,3,3,300008,400010
200039,1,2,3,300009,400010
200040,3,2,1,300010,400010
```

数据说明如下：

表 2-2 商品数据说明

字段名称	字段类型	字段说明	字段取值
product_id	int	商品编号	脱敏
a1	int	属性1	枚举，-1表示未知
a2	int	属性2	枚举，-1表示未知
a3	int	属性3	枚举，-1表示未知

字段名称	字段类型	字段说明	字段取值
category	int	品类ID	脱敏
brand	int	品牌ID	脱敏

- comment_data.csv：

```
deadline,product_id,comment_num,has_bad_comment,bad_comment_rate
2021/3/1,200001,4,0,0
2021/3/1,200002,1,0,0
2021/3/1,200003,2,2,0.1
2021/3/1,200004,3,3,0.05
2021/3/1,200005,1,0,0
2021/3/1,200006,2,0,0
2021/3/1,200007,3,2,0.01
2021/3/1,200008,4,1,0.001
2021/3/1,200009,4,0,0
2021/3/1,200010,1,0,0
2021/3/1,200011,2,2,0.2
2021/3/1,200012,3,3,0.04
2021/3/1,200013,1,0,0
2021/3/1,200014,2,2,0.2
2021/3/1,200015,3,2,0.05
2021/3/1,200016,4,1,0.003
2021/3/1,200017,4,0,0
2021/3/1,200018,1,0,0
2021/3/1,200019,2,2,0.3
2021/3/1,200020,3,3,0.03
2021/3/1,200021,1,0,0
2021/3/1,200022,2,5,1
2021/3/1,200023,3,2,0.07
2021/3/1,200024,4,1,0.006
2021/3/1,200025,4,0,0
2021/3/1,200026,1,0,0
2021/3/1,200027,2,2,0.4
2021/3/1,200028,3,3,0.03
2021/3/1,200029,1,0,0
2021/3/1,200030,2,5,1
2021/3/1,200031,3,2,0.02
2021/3/1,200032,4,1,0.003
2021/3/1,200033,4,0,0
2021/3/1,200034,1,0,0
2021/3/1,200035,2,2,0.5
2021/3/1,200036,3,3,0.06
2021/3/1,200037,1,0,0
2021/3/1,200038,2,1,0.01
2021/3/1,200039,3,2,0.01
2021/3/1,200040,4,1,0.009
```

数据说明如下：

表 2-3 评价数据说明

字段名称	字段类型	字段说明	字段取值
deadline	string	截止时间	单位：天
product_id	int	商品编号	脱敏

字段名称	字段类型	字段说明	字段取值
comment_num	int	累计评论数分段	• 0表示无评论 • 1表示有1条评论 • 2表示有2-10条评论 • 3表示有11-50条评论 • 4表示大于50条评论
has_bad_comment	int	是否有差评	0表示无，1表示有
bad_comment_rate	float	差评率	差评数占总评论数的比重

- action_data.csv：

```
user_id,product_id,time,model_id,type
100001,200001,2021/1/1,1,view
100001,200001,2021/1/1,1,add
100001,200001,2021/1/1,1,delete
100001,200002,2021/1/2,1,view
100001,200002,2021/1/2,1,add
100001,200002,2021/1/2,1,buy
100001,200002,2021/1/2,1,like
100002,200003,2021/1/1,1,view
100002,200003,2021/1/1,1,add
100002,200003,2021/1/1,1,delete
100002,200004,2021/1/2,1,view
100002,200004,2021/1/2,1,add
100002,200004,2021/1/2,1,buy
100002,200004,2021/1/2,1,like
100003,200001,2021/1/1,1,view
100003,200001,2021/1/1,1,add
100003,200001,2021/1/1,1,delete
100004,200002,2021/1/2,1,view
100005,200002,2021/1/2,1,add
100006,200002,2021/1/2,1,buy
100007,200002,2021/1/2,1,like
100001,200003,2021/1/1,1,view
100002,200003,2021/1/1,1,add
100003,200003,2021/1/1,1,delete
100004,200004,2021/1/2,1,view
100005,200004,2021/1/2,1,add
100006,200004,2021/1/2,1,buy
100007,200004,2021/1/2,1,like
100001,200005,2021/1/3,1,view
100001,200005,2021/1/3,1,add
100001,200005,2021/1/3,1,delete
100001,200006,2021/1/3,1,view
100001,200006,2021/1/4,1,add
100001,200006,2021/1/4,1,buy
100001,200006,2021/1/4,1,like
100010,200005,2021/1/3,1,view
100010,200005,2021/1/3,1,add
100010,200005,2021/1/3,1,delete
100010,200006,2021/1/3,1,view
100010,200006,2021/1/4,1,add
100010,200006,2021/1/4,1,buy
100010,200006,2021/1/4,1,like
100001,200007,2021/1/2,1,buy
100001,200007,2021/1/2,1,like
```

```
100002,200007,2021/1/1,1,view
100002,200007,2021/1/1,1,add
100002,200007,2021/1/1,1,delete
100002,200007,2021/1/2,1,view
100002,200007,2021/1/2,1,add
100002,200008,2021/1/2,1,like
100002,200008,2021/1/2,1,like
100003,200008,2021/1/1,1,view
100003,200008,2021/1/1,1,add
100003,200008,2021/1/1,1,delete
100004,200008,2021/1/2,1,view
100005,200009,2021/1/2,1,like
100006,200009,2021/1/2,1,buy
100007,200010,2021/1/2,1,like
100001,200010,2021/1/1,1,view
100002,200010,2021/1/1,1,add
100003,200010,2021/1/1,1,delete
100004,200010,2021/1/2,1,view
100005,200010,2021/1/2,1,like
100006,200010,2021/1/2,1,buy
100007,200010,2021/1/2,1,like
100001,200010,2021/1/3,1,view
100001,200010,2021/1/3,1,add
100001,200010,2021/1/3,1,delete
100001,200011,2021/1/3,1,view
100001,200011,2021/1/4,1,like
100001,200011,2021/1/4,1,buy
100001,200011,2021/1/4,1,like
100010,200012,2021/1/3,1,view
100011,200012,2021/1/3,1,like
100011,200012,2021/1/3,1,delete
100011,200013,2021/1/3,1,view
100011,200013,2021/1/4,1,like
100011,200014,2021/1/4,1,buy
100011,200014,2021/1/4,1,like
100007,200022,2021/1/2,1,like
100001,200022,2021/1/1,1,view
100002,200023,2021/1/1,1,add
100003,200023,2021/1/1,1,delete
100004,200023,2021/1/2,1,like
100005,200024,2021/1/2,1,add
100006,200024,2021/1/2,1,buy
100007,200025,2021/1/2,1,like
100001,200025,2021/1/3,1,view
100001,200026,2021/1/3,1,like
100001,200026,2021/1/3,1,delete
100001,200027,2021/1/3,1,view
100001,200027,2021/1/4,1,like
100001,200027,2021/1/4,1,buy
100001,200028,2021/1/4,1,like
100010,200029,2021/1/3,1,view
100011,200030,2021/1/3,1,like
100011,200031,2021/1/3,1,delete
100011,200032,2021/1/3,1,view
100011,200033,2021/1/4,1,like
100011,200034,2021/1/4,1,buy
100011,200035,2021/1/4,1,like
```

数据说明如下：

表 2-4 行为数据说明

字段名称	字段类型	字段说明	字段取值
user_id	int	用户编号	脱敏
product_id	int	商品编号	脱敏

字段名称	字段类型	字段说明	字段取值
time	string	行为时间	-
model_id	string	模块编号	脱敏
type	string	• 浏览view（指浏览商品详情页） • 加入购物车add • 购物车删除delete • 下单buy • 关注like	-

数据湖准备

在本示例中，选择数据湖探索（DLI）服务作为数据湖。为确保DataArts Studio与DLI网络互通，在创建DLI队列时区域和企业项目应与DataArts Studio实例保持一致。

说明

- 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。

开通DLI服务后，您需要在管理中心创建DLI连接，然后通过数据开发组件新建数据库，再执行SQL来创建OBS外表。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“管理中心”模块，进入管理中心页面。

步骤3 在“数据连接”页面，单击“创建数据连接”按钮。

图 2-1 数据连接

步骤4 创建一个到DLI的连接，数据连接类型选择“数据湖探索（DLI）”，数据连接名称设置为“dli”。

完成设置后，单击“测试”，测试成功后单击“确定”，完成DLI数据连接的创建。

图 2-2 创建数据连接

步骤5 DLI连接创建完成后，跳转到数据开发页面。

图 2-3 跳转到数据开发页面

步骤6 参见图2-4，在DLI连接上右键单击，创建一个数据库用于存放数据表，数据库名称为“BI”。

图 2-4 创建数据库

步骤7 创建一个DLI SQL脚本，以通过DLI SQL语句来创建数据表。

图 2-5 新建脚本

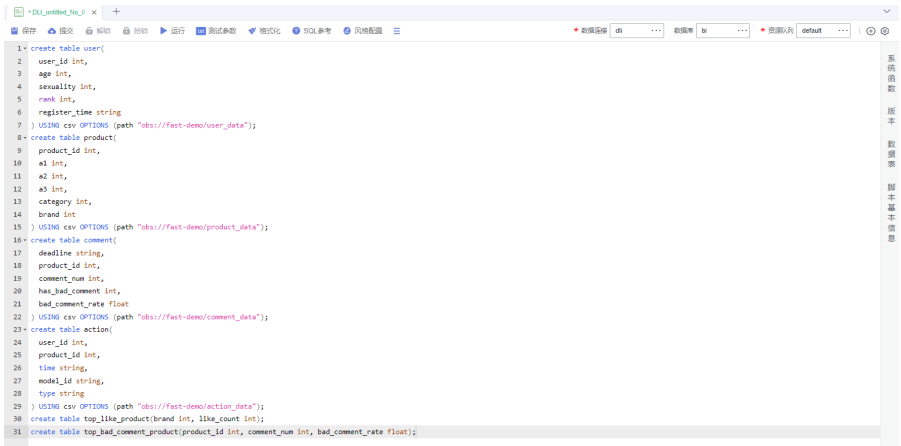
步骤8 在新建脚本弹出的SQL编辑器中输入如下SQL语句，并单击“运行”来创建数据表。其中，user、product、comment、action为OBS外表，使用指定OBS路径中的CSV文件来填充数据，用于存放原始数据；top_like_product和top_bad_comment_product为DLI表，用于存放分析结果。

```
create table user(
  user_id int,
```



```
  age int,
  sexuality int,
  rank int,
  register_time string
) USING csv OPTIONS (path "obs://fast-demo/user_data");
create table product(
  product_id int,
  a1 int,
  a2 int,
  a3 int,
  category int,
  brand int
) USING csv OPTIONS (path "obs://fast-demo/product_data");
create table comment(
  deadline string,
  product_id int,
  comment_num int,
  has_bad_comment int,
  bad_comment_rate float
) USING csv OPTIONS (path "obs://fast-demo/comment_data");
create table action(
  user_id int,
  product_id int,
  time string,
  model_id string,
  type string
) USING csv OPTIONS (path "obs://fast-demo/action_data");
create table top_like_product(brand int, like_count int);
create table top_bad_comment_product(product_id int, comment_num int, bad_comment_rate float);
```

图 2-6 创建数据表



关键参数说明：

- 数据连接：**步骤3**中创建的DLI数据连接。
- 数据库：**步骤5**中创建的数据库。
- 资源队列：可使用提供的默认资源队列“default”。

  说明

  - 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。

步骤9 脚本运行成功后，可以通过如下脚本检查数据表是否创建成功。

```
SHOW TABLES;
```

说明

确认数据表创建成功后，该脚本后续无需使用，可直接关闭。

----结束

2.3 步骤 2：数据开发

本步骤通过BI报表原始数据，分析10大用户关注最多的产品和10大用户评价最差的商品，然后通过作业定期调度执行并将结果每日导出到表中，以支撑信息分析。

分析 10 大用户关注最多的产品

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 创建一个DLI SQL脚本，以通过DLI SQL语句来创建数据表。

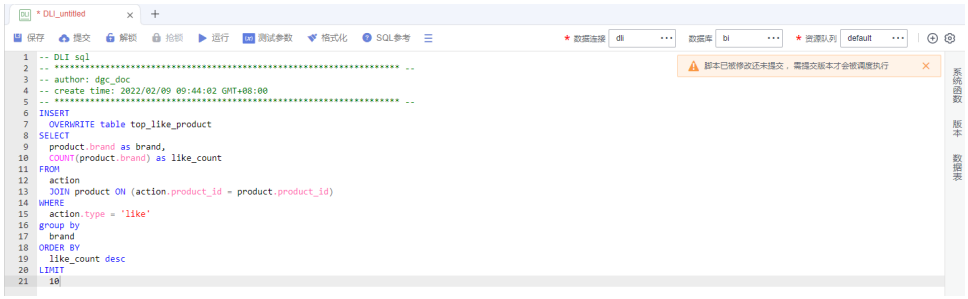
图 2-7 新建脚本

步骤3 在新建脚本弹出的SQL编辑器中输入如下SQL语句，单击“运行”，从OBS原始数据表中计算出10大用户关注最多的产品，将结果存放到top_like_product表。

```
INSERT
  OVERWRITE table top_like_product
SELECT
  product.brand as brand,
  COUNT(product.brand) as like_count
FROM
```

```
  action
  JOIN product ON (action.product_id = product.product_id)
WHERE
  action.type = 'like'
group by
  brand
ORDER BY
  like_count desc
LIMIT
  10
```

图 2-8 脚本（分析 10 大用户关注最多的产品）



关键参数说明：

- 数据连接：**步骤3**中创建的DLI数据连接。
- 数据库：**步骤5**中创建的数据库。
- 资源队列：可使用提供的默认资源队列“default”。

  说明

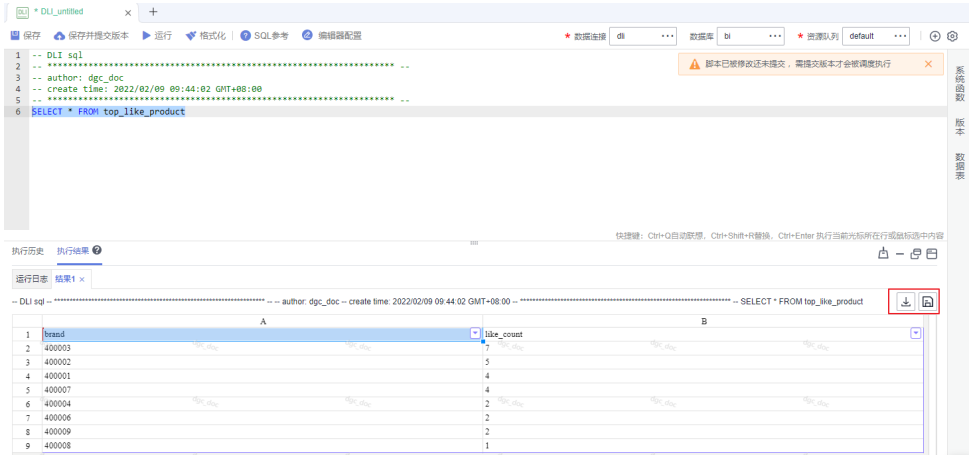
  - 当前由于DLI的“default”队列默认Spark组件版本较低，可能会出现无法支持建表语句执行的报错，这种情况下建议您选择自建队列运行业务。如需“default”队列支持建表语句执行，可联系DLI服务客服或技术支持人员协助解决。
  - DLI的“default”队列为共享队列，仅用于用户体验，用户间可能会出现抢占资源的情况，不能保证每次都可以得到资源执行相关操作。当遇到执行时间较长或无法执行的情况，建议您在业务低峰期再次重试，或选择自建队列运行业务。

步骤4 脚本调试无误后，单击“保存”保存该脚本，脚本名称为“top_like_product”。单击“提交”，提交脚本版本。在后续**开发并调度作业**会引用该脚本。

步骤5 脚本保存完成且运行成功后，您可通过如下SQL语句查看top_like_product表数据。您还可以参考**图2-9**，下载或转储表数据。

```
SELECT * FROM top_like_product
```

图 2-9 查看 top_like_product 表数据



----结束

分析 10 大用户评价最差的商品

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 创建一个DLI SQL脚本，以通过DLI SQL语句来创建数据表。

图 2-10 新建脚本

步骤3 在新建脚本弹出的SQL编辑器中输入如下SQL语句，单击“运行”，从OBS原始数据表中计算出10大用户评价最差的产品，将结果存放到top_bad_comment_product表。

```
INSERT
 OVERWRITE table top_bad_comment_product
```



```
SELECT
  DISTINCT product_id,
  comment_num,
  bad_comment_rate
FROM
  comment
WHERE
  comment_num > 3
ORDER BY
  bad_comment_rate desc
LIMIT
  10
```

图 2-11 脚本（分析 10 大用户评价最差的产品）

关键参数说明：

- 数据连接：**步骤3**中创建的DLI数据连接。
- 数据库：**步骤5**中创建的数据库。
- 资源队列：可使用提供的默认资源队列“default”。

  说明

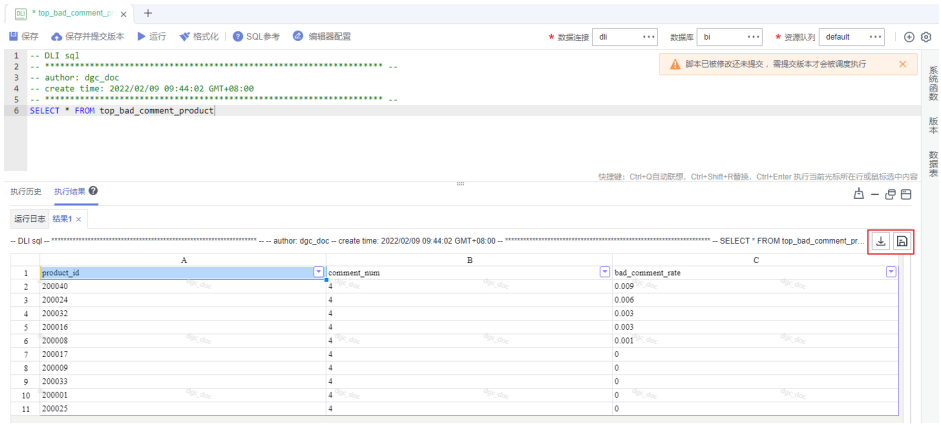
  - 当前由于DLI的“default”队列默认Spark组件版本较低，可能会出现无法支持建表语句执行的报错，这种情况下建议您选择自建队列运行业务。如需“default”队列支持建表语句执行，可联系DLI服务客服或技术支持人员协助解决。
  - DLI的“default”队列为共享队列，仅用于用户体验，用户间可能会出现抢占资源的情况，不能保证每次都可以得到资源执行相关操作。当遇到执行时间较长或无法执行的情况，建议您在业务低峰期再次重试，或选择自建队列运行业务。

步骤4 脚本调试无误后，单击“保存并提交版本”保存该脚本，脚本名称为“top_bad_comment_product”。在后续**开发并调度作业**会引用该脚本。

步骤5 脚本保存完成且运行成功后，您可通过如下SQL语句查看top_bad_comment_product表数据。您还可以参考**图2-12**，下载或转储表数据。

```
SELECT * FROM top_bad_comment_product
```


图 2-12 查看 top_bad_comment_product 表数据



----结束

开发并调度作业

假设在OBS中原始BI报表是每日更新的，我们希望每天更新分析结果，那么这里可以使用DLF作业编排和作业调度功能。

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 创建一个数据开发批处理作业，作业名称为“BI_analysis”。

图 2-13 新建作业

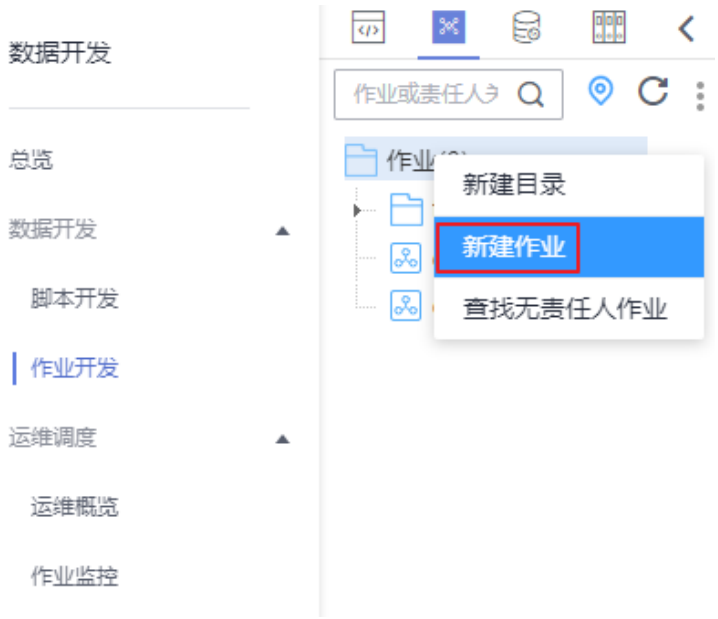
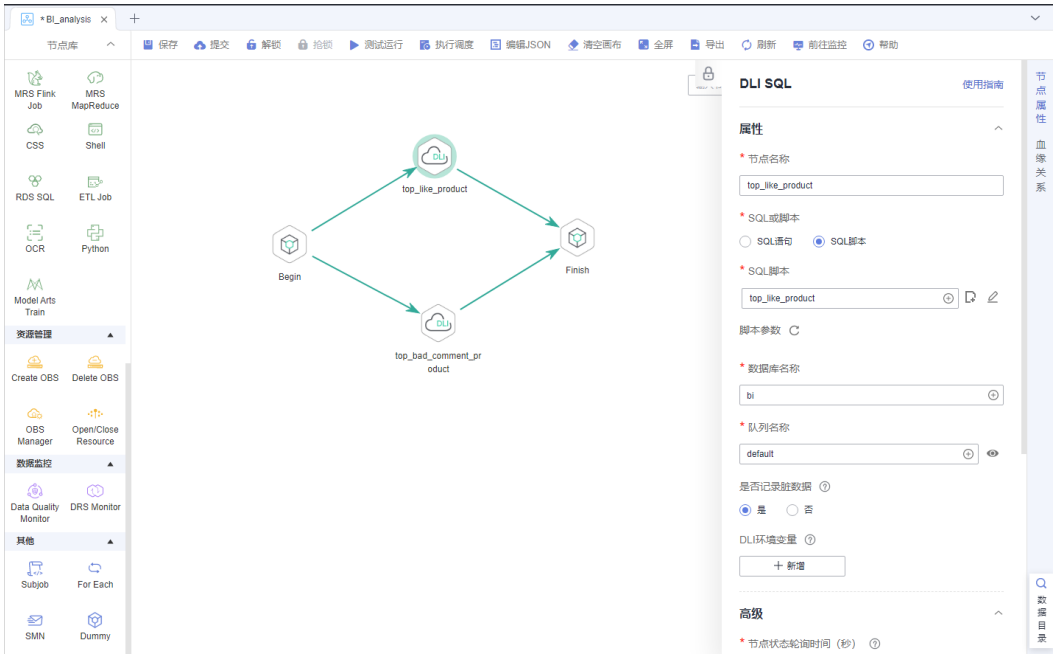


图 2-14 配置作业

步骤3 然后进入到作业开发页面，拖动两个Dummy节点和两个DLI SQL节点到画布中，选中连线图标并拖动，编排图2-15所示的作业。

图 2-15 连接和配置节点属性



关键节点说明：

- Begin（Dummy节点）：不执行任何操作，只作为起始点的标识。
- top_like_product（DLI SQL节点）：在节点属性中，关联**分析10大用户关注最多的产品**中开发完成的DLI SQL脚本“top_like_product”。
- top_bad_comment_product（DLI SQL节点）：在节点属性中，关联**分析10大用户评价最差的商品**中开发完成的DLI SQL脚本“top_bad_comment_product”。
- Finish（Dummy节点）：不执行任何操作，只作为结束点的标识。

步骤4 作业编排完成后，单击，测试运行作业。

步骤5 如果作业测试运行正常，单击右侧的“调度配置”，配置作业的调度策略。

图 2-16 调度配置

说明：

- 调度方式：本示例中选择“周期调度”。
- 调度属性：2022/02/09至2022/02/28，每天1点执行一次作业。
- 依赖属性：可以配置为依赖其他作业运行，本例不涉及，无需配置。
- 跨周期依赖：可以选择配置为自依赖或者不依赖上一周期，此处配置为不依赖上一周期即可。

步骤6 最后保存并提交版本（单击  ），执行调度作业（单击  ）。实现作业每天自动运行，BI报表分析结果自动保存到“top_like_product”和“top_bad_comment_product”表。

步骤7 您如果需要及时了解作业的执行结果是成功还是失败，可以通过数据开发的运维调度界面进行查看，如图2-17所示。

图 2-17 查看作业执行情况

----结束

数据开发还支持配置通知管理，可以选择配置当作业运行异常/失败后，进行短信、邮件等多种方式提醒，此处不再展开描述。

至此，基于电商BI报表的数据开发流程示例完成。此外，您还可以根据原始BI报表数据，分析用户的年龄分布、性别比例、商品评价情况、购买情况、浏览情况等，为营销决策、广告推荐、信用评级、品牌监控、用户行为预测等提供高质量的信息。

2.4 步骤 3：服务退订（可选）

本开发场景中，DataArts Studio、OBS和DLI服务均会产生相关费用。在使用过程中，如果您额外进行了通知配置，可能还会产生以下相关服务的费用：

- SMN服务：如果您在使用DataArts Studio各组件过程中开启了消息通知功能，则会产生消息通知服务费用，收费标准请参见**SMN价格详情**。

在场景开发完成后，如果您不再使用DataArts Studio及相关服务，请及时进行退订和资源删除，避免持续产生费用。

表 2-5 相关服务退订方式

服务	计费说明	退订方式
DataArts Studio	**DataArts Studio计费说明**	DataArts Studio实例仅支持包周期计费。您可以根据需要参考**云服务退订**退订DataArts Studio包年包月套餐。
OBS	**OBS计费说明**	OBS服务支持按需和包周期计费，套餐包暂不支持退订。本例中使用按需计费，完成后删除新建的存储桶即可；另外，DataArts Studio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，在退订DataArts Studio后可以一并删除。
DLI	**DLI计费说明**	DLI服务未购买专属队列时，涉及存储收费和扫描量计费。扫描量收费是在使用默认default队列提交作业时计费的，后续不使用队列不收费；存储收费需要您在DLI服务数据管理中删除相关数据。
SMN	**SMN计费说明**	SMN服务按实际用量付费，退订DataArts Studio服务后不会再产生通知，您也可以直接删除SMN服务已产生的主题和订阅。

3 初级使用者：基于电影评分的数据集成与开发流程

3.1 示例场景说明

本实践通过DataArts Studio服务的数据集成CDM组件、数据开发DLF组件和数据仓库服务（DWS）对电影评分原始数据进行分析，输出评分最高和最活跃Top10电影。您可以学习到数据集成模块的数据迁移和数据开发模块的脚本开发、作业开发、作业调度等功能，以及DWS SQL基本语法。

说明

本入门示例涉及DataArts Studio数据集成、管理中心和数据开发模块，DataArts Studio各版本均可以满足使用要求。

操作流程如下：

1. 准备工作，包括**使用DataArts Studio前的准备**、**数据源准备**、**数据湖准备**和**认证数据准备**。
2. 创建数据迁移作业，将**OBS数据迁移到DWS**。
3. 数据开发，包含创建DWS SQL脚本和开发作业。
   - **创建DWS SQL脚本top_rating_movie（用于存放评分最高的Top10电影）**
   - **创建DWS SQL脚本top_active_movie（用于存放最活跃的Top10电影）**
   - **开发并调度作业**，通过编排作业和配置作业调度策略，定期执行作业，使得用户可以每天获取到最新的Top10电影结果。
4. **服务退订**，如果不再使用DataArts Studio及相关服务，请及时进行退订和资源删除。

3.2 步骤 1：准备工作

使用 DataArts Studio 前的准备

如果您是第一次使用DataArts Studio，请参考**准备工作**章节完成注册华为账号、购买DataArts Studio实例（DataArts Studio企业版）、创建工作空间等一系列操作。然后进入到对应的工作空间，即可开始使用DataArts Studio。

数据源准备

本示例演示数据来自：**https://grouplens.org/datasets/movielens/100k/**，即1000名用户对1700部电影的100,000个评分数据。获取链接中的zip数据包并解压，其中的“u.item”和“u.data”文件分别为电影信息和评分信息。

为方便演示，本示例提供了用于模拟原始数据的部分数据。为了方便将源数据集成到云上，我们需要先将样例数据存储为CSV文件，将CSV文件上传至OBS服务中。

步骤1 创建CSV文件（UTF-8无bom格式），文件名称为对应的数据表名，将后文提供的各样例数据分别复制粘贴到不同CSV文件中，然后保存CSV文件。

以下是Windows下生成.csv文件的办法之一：

1. 使用文本编辑工具（例如记事本等）新建一个txt文档，将后文提供的样例数据复制进文档中。注意复制后检查数据的行数及数据分行的正确性（注意，如果是从PDF文档中复制样例数据，单行的数据过长时会产生换行，需手动重新调整为单行）。
2. 单击“文件 > 另存为”，在弹出的对话框中，“保存类型”选择为“所有文件(*.*)”，在“文件名”处输入文件名和.csv后缀，选择“UTF-8”编码格式（不能带BOM），则能以CSV格式保存该文件。

步骤2 将源数据CSV文件上传到OBS服务。

1. 登录控制台，选择“存储 > 对象存储服务 OBS”，进入OBS控制台。
2. 单击“创建桶”，然后根据页面提示配置参数，创建一个名称为“fast-demo”的OBS桶。

   说明

   为保证网络互通，OBS桶区域请选择和DataArts Studio实例相同的区域。如果需要选择企业项目，也请选择与DataArts Studio实例相同的企业项目。

   使用OBS控制台创建桶的操作，请参见《对象存储服务控制台指南》中的**创建桶**。
3. 上传数据到名称为“fast-demo”的OBS桶中。

   使用OBS控制台上传文件的操作，请参见《对象存储服务控制台指南》中的**上传文件**。

----结束

本示例中涉及到两部分样例数据，分别为电影数据**movies.csv**和评分数据**ratings.csv**。具体数据和说明如下：

- movies.csv：

```
movieId,movieTitle,videoReleaseDate,IMDbURL,unknown,Action,Adventure,Animation,Childrens,Comedy,Crime,Documentary,Drama,Fantasy,FilmNoir,Horror,Musical,Mystery,Romance,SciFi,Thriller,War,Western
1,Toy Story (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Toy%20Story%20(1995),0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0
2,GoldenEye (1995),1-Jan-95,http://us.imdb.com/M/title-exact?GoldenEye%20(1995),0,1,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0
3,Four Rooms (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Four%20Rooms%20(1995),0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0
4,Get Shorty (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Get%20Shorty%20(1995),0,1,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0
5,Copycat (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Copycat%20(1995),0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,1,0,0
6,Shanghai Triad (Yao a yao yao dao waipo qiao) (1995),1-Jan-95,http://us.imdb.com/Title?Yao+a+yao+yao+dao+waipo+qiao+(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
7,Twelve Monkeys (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Twelve%20Monkeys
```


%20(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0
8,Babe (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Babe
%20(1995),0,0,0,0,1,1,0,0,1,0,0,0,0,0,0,0,0,0,0
9,Dead Man Walking (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Dead%20Man%20Walking
%20(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
10,Richard III (1995),22-Jan-96,http://us.imdb.com/M/title-exact?Richard%20III
%20(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0
11,Seven (Se7en) (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Se7en
%20(1995),0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0
12,"Usual Suspects, The (1995)",14-Aug-95,"http://us.imdb.com/M/title-exact?Usual
%20Suspects,%20The%20(1995)",0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0
13,Mighty Aphrodite (1995),30-Oct-95,http://us.imdb.com/M/title-exact?Mighty%20Aphrodite
%20(1995),0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
14,"Postino, Il (1994)",1-Jan-94,"http://us.imdb.com/M/title-exact?Postino,%20Il
%20(1994)",0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0
15,Mr. Holland's Opus (1995),29-Jan-96,http://us.imdb.com/M/title-exact?Mr.%20Holland's%20Opus
%20(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
16,French Twist (Gazon maudit) (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Gazon%20maudit
%20(1995),0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0
17,From Dusk Till Dawn (1996),5-Feb-96,http://us.imdb.com/M/title-exact?From%20Dusk%20Till
%20Dawn%20(1996),0,1,0,0,0,1,1,0,0,0,0,1,0,0,0,0,1,0,0
18,"White Balloon, The (1995)",1-Jan-95,http://us.imdb.com/M/title-exact?Badkonake%20Sefid
%20(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
19,Antonia's Line (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Antonia
%20(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
20,Angels and Insects (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Angels%20and%20Insects
%20(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0
21,Muppet Treasure Island (1996),16-Feb-96,http://us.imdb.com/M/title-exact?Muppet%20Treasure
%20Island%20(1996),0,1,1,0,0,1,0,0,0,0,0,0,1,0,0,0,1,0,0
22,Braveheart (1995),16-Feb-96,http://us.imdb.com/M/title-exact?Braveheart
%20(1995),0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0
23,Taxi Driver (1976),16-Feb-96,http://us.imdb.com/M/title-exact?Taxi%20Driver
%20(1976),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0
24,Rumble in the Bronx (1995),23-Feb-96,http://us.imdb.com/M/title-exact?Hong%20Faan%20Kui
%20(1995),0,1,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
25,"Birdcage, The (1996)",8-Mar-96,"http://us.imdb.com/M/title-exact?Birdcage,%20The
%20(1996)",0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
26,"Brothers McMullen, The (1995)",1-Jan-95,"http://us.imdb.com/M/title-exact?Brothers
%20McMullen,%20The%20(1995)",0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
27,Bad Boys (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Bad%20Boys
%20(1995),0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
28,Apollo 13 (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Apollo
%2013%20(1995),0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0
29,Batman Forever (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Batman%20Forever
%20(1995),0,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0
30,Belle de jour (1967),1-Jan-67,http://us.imdb.com/M/title-exact?Belle%20de%20jour
%20(1967),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
31,Crimson Tide (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Crimson%20Tide
%20(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,1,0
32,Crumb (1994),1-Jan-94,http://us.imdb.com/M/title-exact?Crumb
%20(1994),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
33,Desperado (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Desperado
%20(1995),0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0
34,"Doom Generation, The (1995)",1-Jan-95,"http://us.imdb.com/M/title-exact?Doom
%20Generation,%20The%20(1995)",0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0
35,Free Willy 2: The Adventure Home (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Free%20Willy
%202:%20The%20Adventure%20Home%20(1995),0,0,1,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0
36,Mad Love (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Mad%20Love
%20(1995),0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0
37,Nadja (1994),1-Jan-94,http://us.imdb.com/M/title-exact?Nadja
%20(1994),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
38,"Net, The (1995)",1-Jan-95,"http://us.imdb.com/M/title-exact?Net,%20The
%20(1995)",0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0
39,Strange Days (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Strange%20Days
%20(1995),0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0
40,"To Wong Foo, Thanks for Everything! Julie Newmar (1995)",1-Jan-95,"http://us.imdb.com/M/title-
exact?To%20Wong%20Foo,%20Thanks%20for%20Everything!%20Julie%20Newmar
%20(1995)",0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
41,Billy Madison (1995),1-Jan-95,http://us.imdb.com/M/title-exact?Billy%20Madison


```
%20(1995),0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
42,Clerks (1994),1-Jan-94,http://us.imdb.com/M/title-exact?Clerks
%20(1994),0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
43,Disclosure (1994),1-Jan-94,http://us.imdb.com/M/title-exact?Disclosure
%20(1994),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0
44,Dolores Claiborne (1994),1-Jan-94,http://us.imdb.com/M/title-exact?Dolores%20Claiborne
%20(1994),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0
45,Eat Drink Man Woman (1994),1-Jan-94,http://us.imdb.com/M/title-exact?Yinshi%20Nan%20Nu
%20(1994),0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0
46,Exotica (1994),1-Jan-94,http://us.imdb.com/M/title-exact?Exotica
%20(1994),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
47,Ed Wood (1994),1-Jan-94,http://us.imdb.com/M/title-exact?Ed%20Wood
%20(1994),0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0
48,Hoop Dreams (1994),1-Jan-94,http://us.imdb.com/M/title-exact?Hoop%20Dreams
%20(1994),0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0
49,I.Q. (1994),1-Jan-94,http://us.imdb.com/M/title-exact?
I.Q.%20(1994),0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0
50,Star Wars (1977),1-Jan-77,http://us.imdb.com/M/title-exact?Star%20Wars
%20(1977),0,1,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0
```

数据说明如下：

表 3-1 电影数据说明

字段名称	字段类型	字段说明
movieId	INT	电影ID
movieTitle	VARCHAR	电影名
videoReleaseDate	VARCHAR	电影发行日期
IMDbURL	VARCHAR	IMDb链接
unknown	INT	未知，是则为1，否为0
Action	INT	动作，是则为1，否为0
Adventure	INT	冒险，是则为1，否为0
Animation	INT	动画，是则为1，否为0
Childrens	INT	儿童，是则为1，否为0
Comedy	INT	喜剧，是则为1，否为0
Crime	INT	犯罪，是则为1，否为0
Documentary	INT	纪录片，是则为1，否为0
Drama	INT	戏剧，是则为1，否为0
Fantasy	INT	幻想，是则为1，否为0
FilmNoir	INT	黑色，是则为1，否为0
Horror	INT	恐怖，是则为1，否为0
Musical	INT	音乐，是则为1，否为0
Mystery	INT	神秘，是则为1，否为0
Romance	INT	浪漫，是则为1，否为0

字段名称	字段类型	字段说明
SciFi	INT	科幻，是则为1，否为0
Thriller	INT	惊悚，是则为1，否为0
War	INT	战争，是则为1，否为0
Western	INT	西部，是则为1，否为0

- ratings.csv：

```
userId,movieId,rating,timestamp
210,40,3,891035994
224,29,3,888104457
308,1,4,887736532
7,32,4,891350932
10,16,4,877888877
99,4,5,886519097
115,20,3,881171009
138,26,5,879024232
243,15,3,879987440
293,5,3,888906576
162,25,4,877635573
135,23,4,879857765
62,21,3,879373460
59,23,5,888205300
43,14,2,883955745
19,4,4,885412840
5,2,3,875636053
72,48,4,880036718
224,26,3,888104153
299,14,4,877877775
151,10,5,879524921
6,14,5,883599249
250,7,4,878089716
268,2,2,875744173
292,11,5,881104093
181,3,2,878963441
145,15,2,875270655
1,33,4,878542699
276,2,4,874792436
18,26,4,880129731
87,40,3,879876917
272,12,5,879455254
296,20,5,884196921
5,17,4,875636198
128,15,4,879968827
287,1,5,875334088
65,47,2,879216672
1,20,4,887431883
290,50,5,880473582
45,25,4,881014015
109,8,3,880572642
157,25,3,886890787
301,33,4,882078228
62,12,4,879373613
276,40,3,874791871
269,22,1,891448072
10,7,4,877892210
244,17,2,880607205
222,26,3,878183043
185,23,4,883524249
207,13,3,875506839
8,22,5,879362183
222,49,3,878183512
200,11,5,884129542
```



```
90,25,5,891384789
15,25,3,879456204
234,10,3,891227851
295,39,4,879518279
217,2,3,889069782
189,20,5,893264466
42,44,3,881108548
268,21,3,875742822
262,28,3,879792220
90,22,4,891384357
270,25,5,876954456
194,23,4,879522819
161,48,1,891170745
58,9,4,884304328
79,50,4,891271545
221,48,5,875245462
223,11,3,891550649
292,9,4,881104148
16,8,5,877722736
17,13,3,885272654
148,1,4,877019411
280,1,4,891700426
110,38,3,886988574
90,12,5,891383241
239,9,5,889180446
311,9,4,884963365
151,13,3,879542688
2,50,5,888552084
8,50,5,879362124
286,44,3,877532173
85,25,2,879452769
274,50,5,878944679
217,27,1,889070011
181,14,1,878962392
297,25,4,874954497
1,47,4,875072125
6,23,4,883601365
222,22,5,878183285
314,28,5,877888346
291,15,5,874833668
94,24,4,885873423
83,43,4,880308690
43,40,3,883956468
44,15,4,878341343
158,24,4,880134261
151,12,5,879524368
66,1,3,883601324
5,1,4,875635748
207,25,4,876079113
109,1,4,880563619
227,50,4,879035347
181,1,3,878962392
213,13,4,878955139
121,14,5,891390014
117,15,5,880125887
85,13,3,879452866
313,22,3,891014870
43,5,4,875981421
11,38,3,891905936
72,28,4,880036824
115,8,5,881171982
95,1,5,879197329
145,22,5,875273021
66,7,3,883601355
267,17,4,878971773
25,25,5,885853415
103,24,4,880415847
87,9,4,879877931
49,47,5,888068715
```

```
135,39,3,879857931
269,13,4,891446662
99,50,5,885679998
306,14,5,876503995
291,7,5,874834481
312,28,4,891698300
184,36,3,889910195
305,11,1,886323237
198,7,4,884205317
104,7,3,888465972
293,39,3,888906804
256,25,5,882150552
92,15,3,875640189
1,17,3,875073198
214,42,5,892668130
82,14,4,876311280
305,50,5,886321799
223,8,2,891550684
91,28,4,891439243
315,13,4,879821158
269,9,4,891446246
217,7,4,889069741
49,7,4,888067307
87,2,4,879876074
268,1,3,875742341
262,47,2,879794599
84,12,5,883452874
264,33,3,886122644
224,20,1,888104487
200,24,2,884127370
92,24,3,875640448
276,38,3,874792574
286,34,5,877534701
49,38,1,888068289
311,5,3,884365853
269,47,4,891448386
194,4,4,879521397
57,28,4,883698324
108,50,4,879879739
207,4,4,876198457
181,16,1,878962996
94,9,5,885872684
234,20,4,891227979
68,7,3,876974096
13,14,4,884538727
98,47,4,880498898
53,24,3,879442538
239,10,5,889180338
63,20,3,875748004
276,43,1,874791383
272,48,4,879455143
116,7,2,876453915
26,25,3,891373727
62,24,4,879372633
295,47,5,879518166
63,50,4,875747292
49,17,2,888068651
310,24,4,879436242
7,44,5,891351728
326,22,4,879874989
213,12,5,878955409
222,29,3,878184571
249,11,5,879640868
217,22,5,889069741
189,1,5,893264174
234,50,4,892079237
296,48,5,884197091
81,3,4,876592546
151,15,4,879524879
```

```
59,12,5,888204260
246,8,3,884921245
276,34,2,877934264
97,50,5,884239471
244,7,4,880602558
298,8,5,884182748
7,28,5,891352341
41,28,4,890687353
```

数据说明如下：

表 3-2 评分数据说明

字段名称	字段类型	字段说明
userId	INT	用户ID
movieId	INT	电影ID
rating	INT	评分，5分制
timestamp	VARCHAR	时间戳

数据湖准备

在本示例中，选择数据仓库服务（DWS）服务作为数据湖。

创建DWS集群的具体操作请参见**创建集群**。为确保DWS集群与DataArts Studio实例网络互通，DWS集群需满足如下要求：

- DataArts Studio实例（指DataArts Studio实例中的CDM集群）与DWS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。
- DataArts Studio实例（指DataArts Studio实例中的CDM集群）与DWS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DWS集群与DataArts Studio工作空间所属的企业项目必须相同，如果不同，您需要修改工作空间的企业项目。

创建DWS集群后，您需要在管理中心创建DWS连接，然后通过数据开发组件新建数据库、数据库模式，再执行SQL来创建DWS表。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“管理中心”模块，进入管理中心页面。

步骤3 在“数据连接”页面，单击“创建数据连接”按钮。

图 3-1 数据连接

步骤4 在弹出窗口中，配置数据连接参数，完成配置后，单击“确定”完成数据连接的创建。参数配置如图3-2所示。

- **数据连接类型**：数据仓库服务（DWS）
- **数据连接名称**：dws_link
- **手动**：关闭“手动”，“IP”和“端口”不需要手动填写。
- **集群名**：选择所创建的DWS集群。
- **用户名**：数据库的用户名，创建DWS集群时指定的用户名，默认为dbadmin。
- **密码**：数据库的访问密码，创建DWS集群时指定的密码。
- **KMS密钥**：选择一个KMS密钥，使用KMS密钥对敏感数据进行加密。如果未创建KMS密钥，请单击“访问KMS”进入KMS控制台创建一个密钥。
- **绑定Agent**：需选择一个数据集成集群作为连接代理，该集群和DWS集群必须网络互通。本示例可选择创建DataArts Studio实例时自动创建的数据集成集群。

图 3-2 DWS 连接配置参数

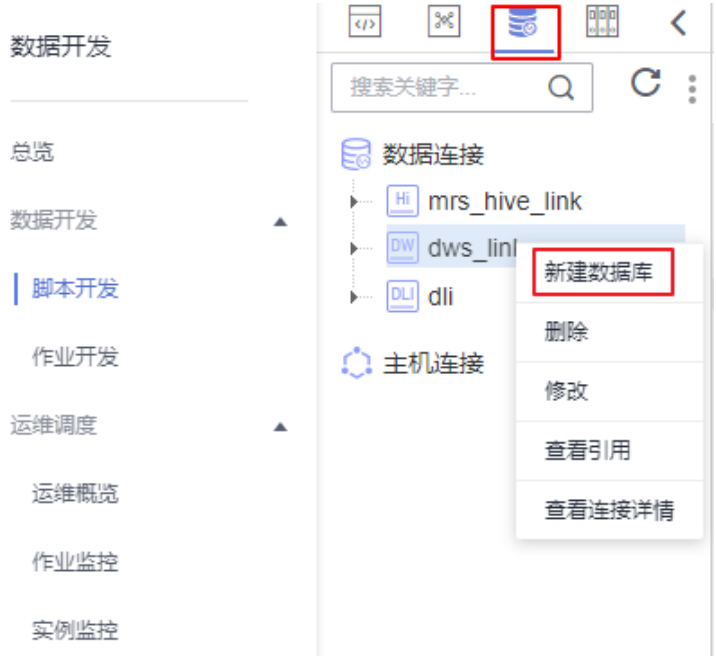
步骤5 DWS连接创建完成后，跳转到数据开发页面。

图 3-3 跳转到数据开发页面

步骤6 创建DWS数据库和数据库模式。

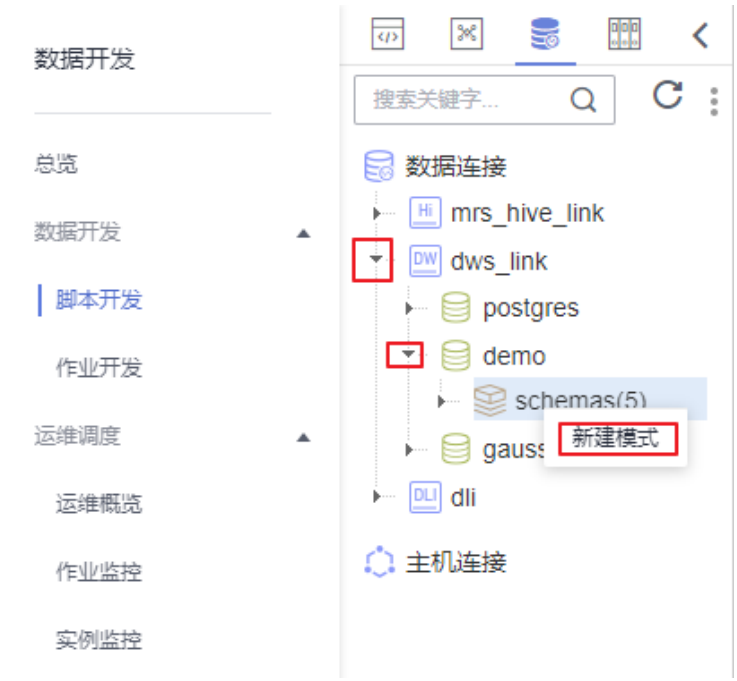
1. 在数据开发界面，在DWS连接上右键单击，选择“新建数据库”，创建一个数据库用于存放数据表，数据库名称为“demo”。

图 3-4 创建数据库



2. 展开DWS连接目录至demo数据库的数据库模式层级，然后再右键单击，选择“新建模式”，创建数据库模式用于存放数据表，数据库模式名称为“dgc”。

图 3-5 创建数据库模式



步骤7 创建一个DWS SQL脚本，以通过DWS SQL语句来创建数据表。

图 3-6 新建脚本

步骤8 在新建脚本弹出的SQL编辑器中输入如下SQL语句，并单击“运行”来创建数据表。其中，movies_item、ratings_item为原始数据表，具体数据将在之后通过CDM由OBS迁移到表中；top_rating_movie和top_active_movie为结果表，用于存放分析结果。

```
SET SEARCH_PATH TO dgc;
CREATE TABLE IF NOT EXISTS movies_item(
    movieId INT,
    movieTitle VARCHAR,
```



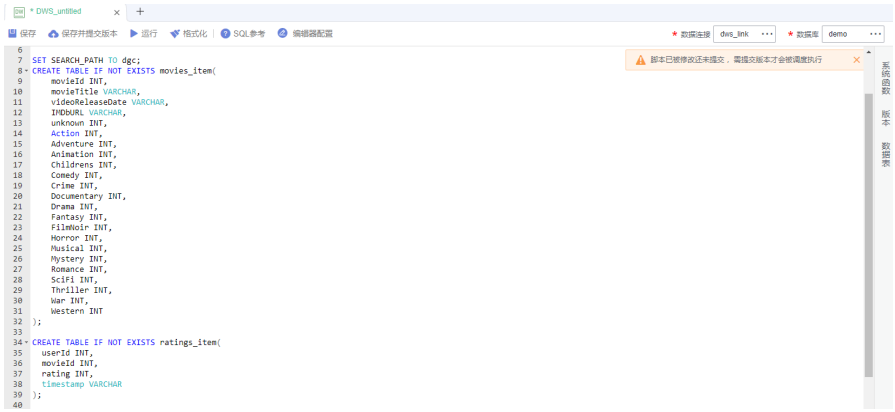
```
  videoReleaseDate VARCHAR,
  IMDbURL VARCHAR,
  unknown INT,
  Action INT,
  Adventure INT,
  Animation INT,
  Childrens INT,
  Comedy INT,
  Crime INT,
  Documentary INT,
  Drama INT,
  Fantasy INT,
  FilmNoir INT,
  Horror INT,
  Musical INT,
  Mystery INT,
  Romance INT,
  SciFi INT,
  Thriller INT,
  War INT,
  Western INT
);

CREATE TABLE IF NOT EXISTS ratings_item(
  userId INT,
  movieId INT,
  rating INT,
  timestamp VARCHAR
);

CREATE TABLE IF NOT EXISTS top_rating_movie(
  movieTitle VARCHAR,
  avg_rating float,
  rating_user_number int
);

CREATE TABLE IF NOT EXISTS top_active_movie(
  movieTitle VARCHAR,
  avg_rating float,
  rating_user_number int
);
```

图 3-7 创建数据表



关键参数说明：

- 数据连接：**步骤3**中创建的DWS数据连接。
- 数据库：**步骤5**中创建的数据库。

步骤9 脚本运行成功后，可以通过如下脚本检查数据表是否创建成功。确认数据表创建成功后，该脚本后续无需使用，可直接关闭。


```
SELECT * FROM pg_tables;
```

----结束

认证数据准备

当您需要通过CDM迁移OBS数据时，需要通过AK/SK认证方式进行认证鉴权，因此，我们必须先创建访问密钥（AK和SK）。

- Access Key Id（AK）：访问密钥ID。与私有访问密钥关联的唯一标识符；访问密钥ID和私有访问密钥一起使用，对请求进行加密签名。
- Secret Access Key（SK）：与访问密钥ID结合使用的密钥，对请求进行加密签名，可标识发送方，并防止请求被修改。

在创建访问密钥前，请确保登录控制台的账号已通过实名认证。

您可以通过如下方式获取访问密钥。

1. 登录控制台，在用户名下拉列表中选择“我的凭证”。
2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图3-8所示。

图 3-8 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。

3.3 步骤 2：数据集成

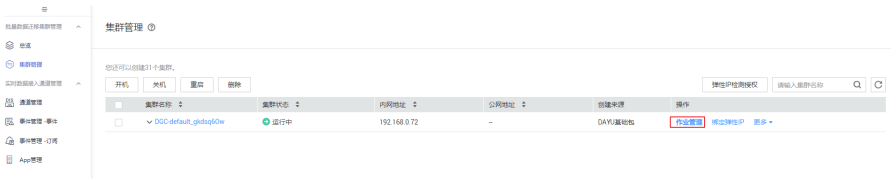
OBS 数据迁移到 DWS

步骤1 登录DataArts Studio控制台。选择实例，单击“进入控制台”，选择对应工作空间的“数据集成”模块，进入数据集成页面。

图 3-9 选择数据集成

步骤2 进入DataArts Studio数据集成主页面，单击操作列的“作业管理”。

图 3-10 作业管理



步骤3 在作业管理界面，选择“连接管理 - 新建连接”，进入创建连接页面。

步骤4 在创建连接页面，选择“对象存储服务（OBS）”，新建CDM到OBS的连接，数据连接名称为“obs_link”。

表 3-3 OBS 连接的参数

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

参数名	说明	取值样例
OBS终端节点	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。	-
端口	数据传输协议端口，https是443，http是80。	443
OBS桶类型	用户下拉选择即可，一般选择为“对象存储”。	对象存储
访问标识(AK)	AK和SK分别为登录OBS服务器的访问标识与密钥。 您需要先创建当前账号的访问密钥，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3-11所示。 **图 3-11** 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)		-

图 3-12 创建 OBS 连接

步骤5 在创建连接页面，选择“数据仓库服务（DWS）”，新建CDM到DWS的连接，数据连接名称为“dws_link”。

表 3-4 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	单击输入框后的“选择”，可获取用户的DWS实例列表。	-
端口	配置为要连接的数据库的端口。DWS数据库端口默认为8000。	8000
数据库名称	配置为要连接的数据库名称。	demo
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。	dbadmin
密码	用户密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	否

图 3-13 创建 DWS 连接

步骤6 CDM到OBS和DWS的连接创建成功后，单击“表/文件迁移”，再单击“新建作业”。

图 3-14 新建作业

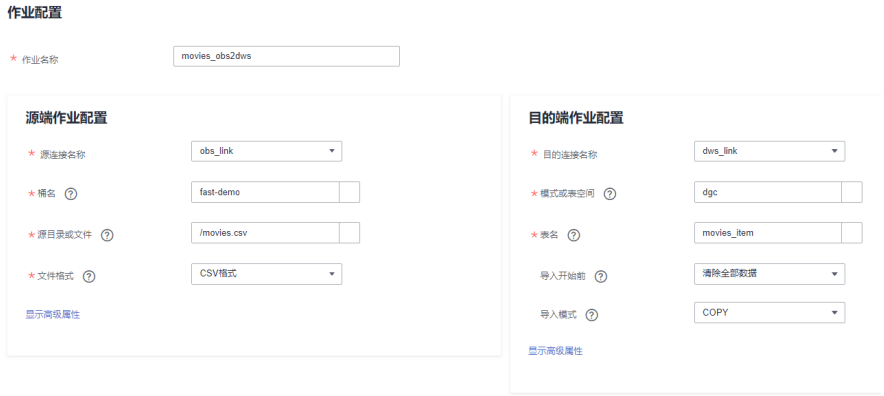
步骤7 按照如下步骤完成作业参数的配置。

1. 如图3-15所示，配置作业名为movies_obs2dws，配置源端作业参数，然后配置目的端作业参数。

   说明

   在本示例中，目的端作业参数“导入开始前”配置为“清除全部数据”，表示每次作业运行都会先清空数据再导入。在实际业务中，请视情况而定，需谨慎设置，以免造成数据丢失。

图 3-15 作业配置



2. 在源端、目的作业配置区域，单击“显示高级属性”，在“高级属性”中，系统提供了默认值，请根据实际业务数据的格式设置各项参数。

   例如，本例中根据**数据源准备**中的样例数据格式，源端高级属性需注意以下参数的设置，其他参数均保留默认值即可，如**图3-16**所示。目的端高级属性无需配置。

   - **字段分隔符**：默认值为逗号，本示例需要保留默认值。
   - **使用包围符**：由于IMDbURL有的原始数据中包含“,”，需要修改默认值为“是”。
   - **前N行为标题行**：默认值为“否”，本示例首行是标题行，修改默认值为“是”，**标题行数**配置为1。

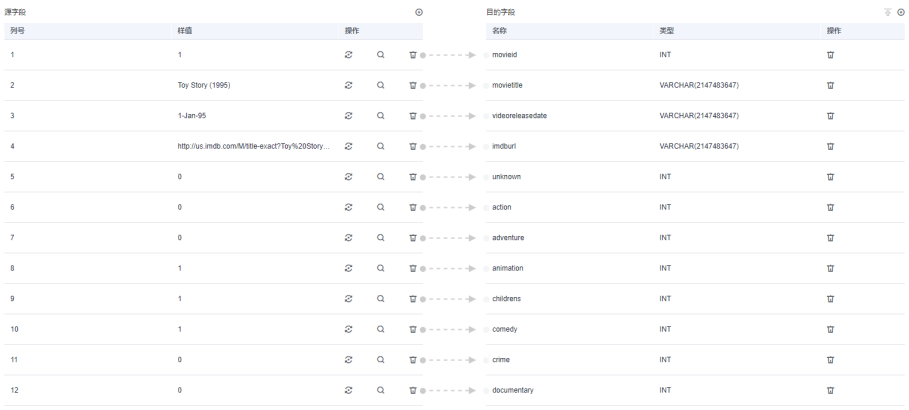
图 3-16 源端高级属性

3. 单击“下一步”后，请参考以下说明配置字段映射，如图3-17所示，配置完成后，单击“下一步”。

   字段映射：在本示例中，由于数据迁移的目标表字段顺序和原始数据的字段顺序是一样的，因此这里不需要调整字段映射的顺序。

   如果目标表字段顺序和原始数据不一致，请一一将源字段指向含义相同的目的字段。请将鼠标移至某一个字段的箭头起点，当光标显示为“+”的形状时，按住鼠标，将箭头指向相同含义的目的字段，然后松开鼠标。

图 3-17 字段映射



4. 根据需要配置任务的重试和定时执行、高级属性等。在本示例中仅需将“是否写入脏数据”设置为“是”，其他配置项保持默认即可。

图 3-18 任务配置

单击“显示高级属性”，可配置“抽取并发数”以及“是否写入脏数据”，如图3-18所示。

- **抽取并发数**：设置同时执行的抽取任务数。并发抽取数取值范围为1-1000，若配置过大，则以队列的形式进行排队。

  CDM迁移作业的抽取并发量，与集群规格和表大小有关。

  - 按集群规格建议每1CUs（1CUs=1核4G）配置为4。
  - 表每行数据大小为1MB以下的可以多并发抽取，超过1MB的建议单线程抽取数据。

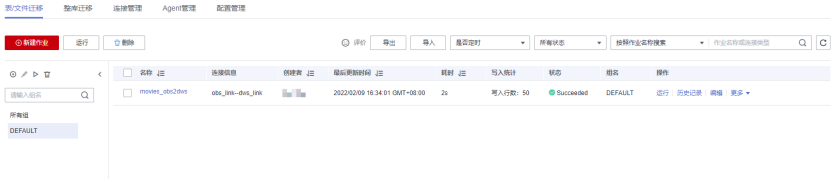
- **是否写入脏数据**：建议配置为“是”，然后参考图3-18配置相关参数。脏数据是指与目的端字段不匹的数据，该数据可以被记录到指定的OBS桶中。用户配置脏数据归档后，正常数据可以写入目的端，迁移作业不会因脏数据中断。

在本示例中，“OBS桶”配置为在**数据源准备**中创建的桶“fast-demo”，您需要前往OBS控制台，在桶中创建一个目录，例如err_data，然后再将“脏数据目录”配置为该目录。

步骤8 单击“保存并运行”完成作业的创建。

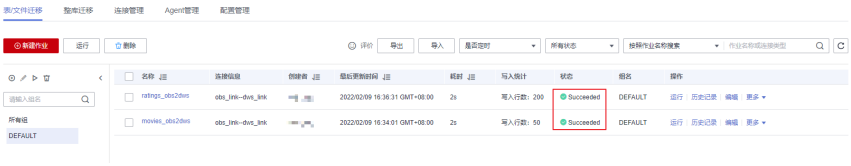
返回“表/文件迁移”页面后，可在作业列表中查看到新建的作业。

图 3-19 迁移作业运行结果



步骤9 参考**步骤6**~**步骤8**，再新建名为ratings_obs2dws的迁移作业，将ratings.csv数据迁移到DWS的ratings_item表中。待作业运行成功后，数据迁移结束。

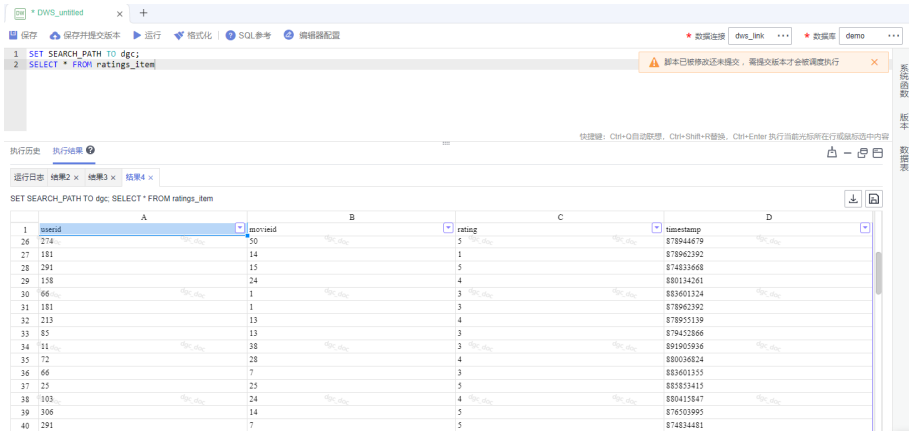
图 3-20 数据迁移结果



步骤10 数据迁移结束后，您也可以跳转到数据开发页面，新建一个DWS SQL脚本，并分别执行以下SQL语句检查DWS中的movies_item和ratings_item表数据是否符合预期。

- 查看movies_item表数据：
```
SET SEARCH_PATH TO dgc;
SELECT * FROM movies_item;
```
- 查看ratings_item表数据：
```
SET SEARCH_PATH TO dgc;
SELECT * FROM ratings_item;
```

图 3-21 查看 DWS 表数据



----结束

3.4 步骤 3：数据开发

本步骤通过电影信息和评分信息的原始数据，分析评分最高的Top10电影和最活跃的Top10电影，然后通过作业定期调度执行并将结果每日导出到表中，以支撑信息分析。

创建 DWS SQL 脚本 top_rating_movie（用于存放评分最高的 Top10 电影）

评分最高Top10电影的计算方法是：先计算出每部电影的总评分和参与评分的用户数，过滤掉参与评分的用户数小于3的记录，返回电影名称、平均评分和参与评分用户数。

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 创建一个DWS SQL脚本，以通过DWS SQL语句来创建数据表。

图 3-22 新建脚本

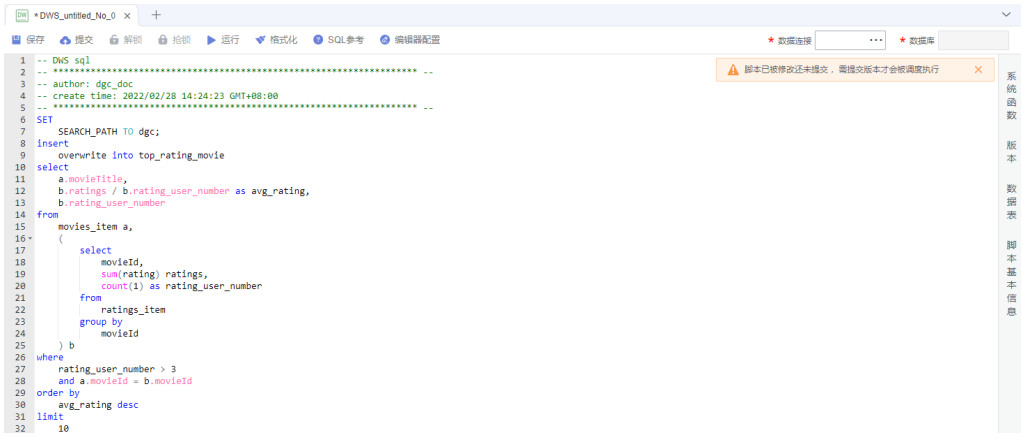
步骤3 在新建脚本弹出的SQL编辑器中输入如下SQL语句，单击“运行”，从movies_item和ratings_item表中计算出评分最高的Top10电影，将结果存放到top_rating_movie表。

```
SET
    SEARCH_PATH TO dgc;
insert
    overwrite into top_rating_movie
select
    a.movieTitle,
    b.ratings / b.rating_user_number as avg_rating,
    b.rating_user_number
from
    movies_item a,
    (
        select
            movieId,
            sum(rating) ratings,
            count(1) as rating_user_number
        from
```



```
        ratings_item
    group by
        movieId
) b
where
    rating_user_number > 3
    and a.movieId = b.movieId
order by
    avg_rating desc
limit
    10
```

图 3-23 脚本（top_rating_movie）



关键参数说明：

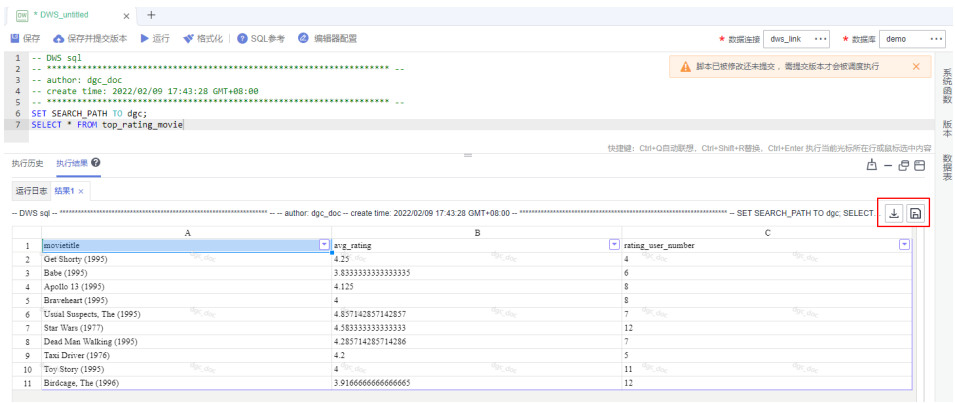
- 数据连接：**步骤3**中创建的DWS数据连接。
- 数据库：**步骤5**中创建的数据库。

步骤4 脚本调试无误后，单击“保存并提交版本”提交该脚本，脚本名称为“top_rating_movie”。在后续**开发并调度作业**引用该脚本。

步骤5 脚本保存完成且运行成功后，您可通过如下SQL语句查看top_rating_movie表数据。您还可以参考**图3-24**，下载或转储表数据。

```
SET SEARCH_PATH TO dgc;
SELECT * FROM top_rating_movie
```

图 3-24 查看 top_rating_movie 表数据



	A	B	C
1	movietitle	avg_rating	rating_user_number
2	Get Shorty (1995)	4.25	4
3	Babe (1995)	3.8333333333333335	6
4	Apollo 13 (1995)	4.125	8
5	Braveheart (1995)	4	8
6	Usual Suspects, The (1995)	4.857142857142857	7
7	Star Wars (1977)	4.583333333333333	12
8	Dead Man Walking (1995)	4.285714285714286	7
9	Taxi Driver (1976)	4.2	5
10	Toy Story (1995)	4	11
11	Birdcage, The (1996)	3.9166666666666665	12

----结束

创建 DWS SQL 脚本 top_active_movie（用于存放最活跃的 Top10 电影）

最活跃Top10电影的计算方法是：平均评分大于3.5的电影中用户评分数最多的10部电影。

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 创建一个DWS SQL脚本，以通过DWS SQL语句来创建数据表。

图 3-25 新建脚本

步骤3 在新建脚本弹出的SQL编辑器中输入如下SQL语句，单击“运行”，从movies_item和ratings_item表中计算出最活跃的Top10电影，将结果存放到top_active_movie表。

```
SET
    SEARCH_PATH TO dgc;
insert
    overwrite into top_active_movie
select
    *
from
    (
        select
            a.movieTitle,
            b.ratingSum / b.rating_user_number as avg_rating,
            b.rating_user_number
        from
            movies_item a,
            (
                select
                    movieId,
                    sum(rating) ratingSum,
                    count(1) as rating_user_number
                from
                    ratings_item
                group by
                    movieId
            ) b
        where
```



```
        a.movieId = b.movieId
    ) t
where
    t.avg_rating > 3.5
order by
    rating_user_number desc
limit
    10
```

图 3-26 脚本（top_active_movie）

关键参数说明：

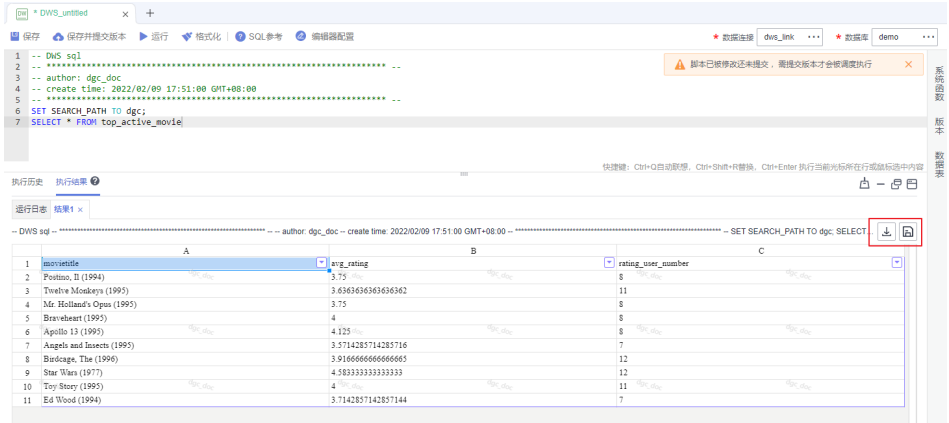
- 数据连接：**步骤3**中创建的DWS数据连接。
- 数据库：**步骤5**中创建的数据库。

步骤4 脚本调试无误后，单击“保存并提交版本”提交该脚本，脚本名称为“top_active_movie”。在后续**开发并调度作业**引用该脚本。

步骤5 脚本保存完成且运行成功后，您可通过如下SQL语句查看top_active_movie表数据。您还可以参考**图3-27**，下载或转储表数据。

```
SET SEARCH_PATH TO dgc;
SELECT * FROM top_active_movie
```

图 3-27 查看 top_active_movie 表数据



----结束

开发并调度作业

假设OBS中“movie”和“rating”表是每日更新的，我们希望每天更新Top10电影，那么这里可以使用DLF作业编排和作业调度功能。

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 创建一个DLF批处理作业，作业名称为“topmovie”。

图 3-28 新建作业

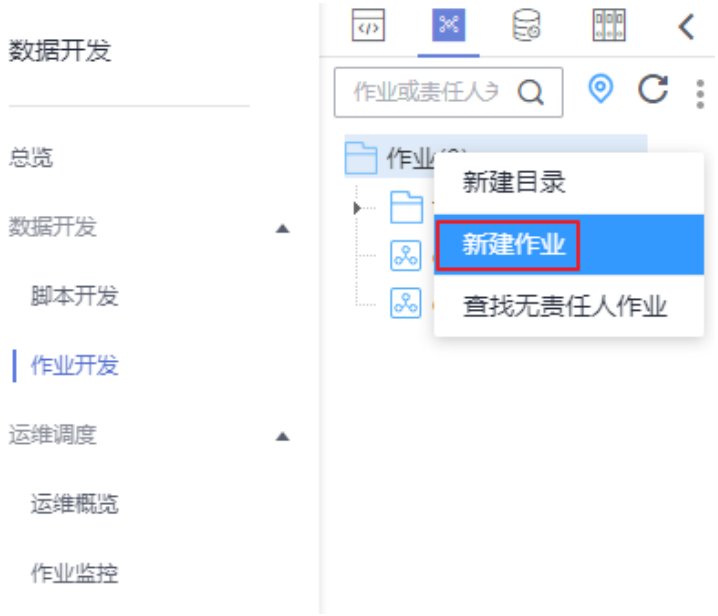


图 3-29 配置作业

步骤3 在作业开发页面，拖动2个CDM Job节点、3个Dummy节点和2个DWS SQL节点到画布中，选中连线图标并拖动，编排图3-30所示的作业。

图 3-30 连接和配置节点属性

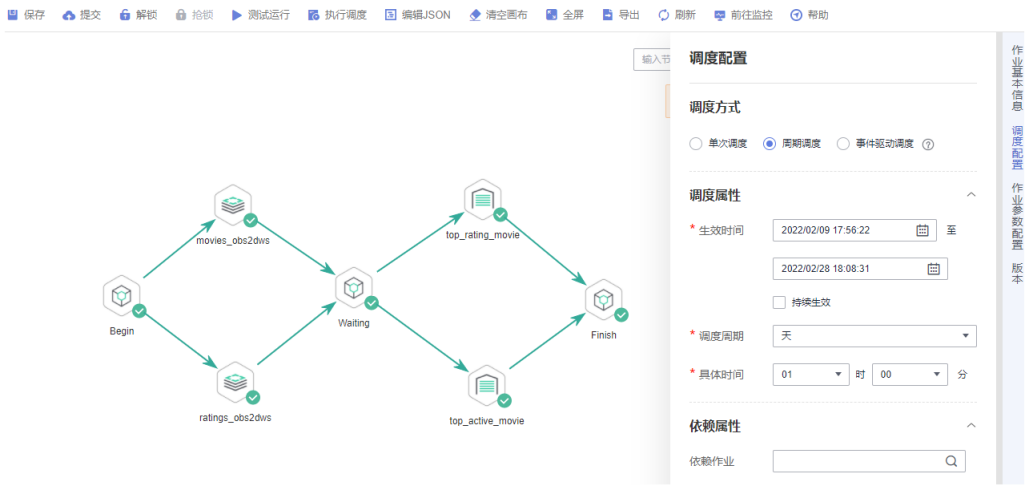
关键节点说明：

- Begin（Dummy节点）：不执行任何操作，只作为起始点的标识。
- movies_obs2dws（CDM Job节点）：在节点属性中，选择**步骤2：数据集成**中的CDM集群，并关联CDM作业“movies_obs2dws”。
- ratings_obs2dws（CDM Job节点）：在节点属性中，选择**步骤2：数据集成**中的CDM集群，并关联CDM作业“ratings_obs2dws”。
- Waiting（Dummy节点）：不执行任何操作，作为等待前侧节点执行结束的标识。
- top_rating_movie（DWS SQL节点）：在节点属性中，关联**创建DWS SQL脚本top_rating_movie**中开发完成的DWS SQL脚本“top_rating_movie”。
- top_active_movie（DWS SQL节点）：在节点属性中，关联**创建DWS SQL脚本top_active_movie**中开发完成的DWS SQL脚本“top_active_movie”。
- Finish（Dummy节点）：不执行任何操作，只作为结束点的标识。

步骤4 作业编排完成后，单击，测试运行作业。

步骤5 如果作业运行正常，单击“调度配置”，配置作业的调度策略。

图 3-31 调度配置



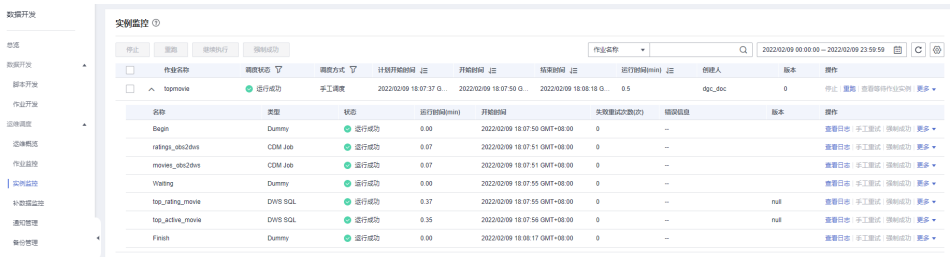
说明：

- 2022/02/09至2022/02/28，每天1点00分执行一次作业。
- 依赖属性：可以配置为依赖其他作业运行，本例不涉及，无需配置。
- 跨周期依赖：可以选择配置为依赖上一周期或者不依赖，此处配置为不依赖即可。

步骤6 最后保存并提交版本（单击 ），执行调度作业（单击 ）。实现作业每天自动运行，Top10电影的结果自动保存到“top_active_movie”和“top_rating_movie”表。

步骤7 您如果需要及时了解作业的执行结果是成功还是失败，可以通过数据开发的运维调度界面进行查看，如图3-32所示。

图 3-32 查看作业执行情况



----结束

数据开发还支持配置通知管理，可以选择配置当作业运行异常/失败后，进行短信、邮件、控制台等多种方式提醒，此处不再展开描述。

至此，基于电影评分的数据集成与开发流程示例完成。此外，您还可以根据原始数据，分析不同类型电影的评分、浏览情况等，为营销决策、广告推荐、用户行为预测等提供高质量的信息。

3.5 步骤 4：服务退订（可选）

本开发场景中，DataArts Studio、OBS和DWS服务均会产生相关费用。在使用过程中，如果您额外进行了通知配置，可能还会产生以下相关服务的费用：

- SMN服务：如果您在使用DataArts Studio各组件过程中开启了消息通知功能，则会产生消息通知服务费用，收费标准请参见**SMN价格详情**。
- EIP服务：如果您为数据集成集群开通了公网IP，则会产生弹性公网IP服务费用，收费标准请参见**EIP价格详情**。
- DEW服务：在数据集成或创建管理中心连接时，如果启用了KMS，则会产生密钥管理费用，收费标准请参见**DEW价格详情**。

在场景开发完成后，如果您不再使用DataArts Studio及相关服务，请及时进行退订和资源删除，避免持续产生费用。

表 3-5 相关服务退订方式

服务	计费说明	退订方式
DataArts Studio	**DataArts Studio计费说明**	DataArts Studio实例仅支持包周期计费。您可以根据需要参考**云服务退订**退订DataArts Studio包年包月套餐。
OBS	**OBS计费说明**	OBS服务支持按需和包周期计费，套餐包暂不支持退订。本例中使用按需计费，完成后删除新建的存储桶即可；另外，DataArts Studio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，在退订DataArts Studio后可以一并删除。
DWS	**DWS计费说明**	DWS服务支持按需和包周期计费。本例中使用按需计费，完成后删除DWS集群即可。如果使用包周期计费，您需要参考**云服务退订**退订包年包月套餐，并删除DWS集群。
SMN	**SMN计费说明**	SMN服务按实际用量付费，退订DataArts Studio服务后不会再产生通知，您也可以直接删除SMN服务已产生的主题和订阅。
EIP	**EIP计费说明**	EIP服务支持按需和包周期计费，本例中使用按需计费，完成后删除EIP即可。如果使用包周期计费，您需要参考**云服务退订**退订包年包月套餐，并删除EIP。
DEW	**DEW计费说明**	KMS密钥管理按密钥实例进行按需计费，您可以直接删除DEW服务已产生的KMS密钥。

4 高级使用者：基于出租车出行的数据治理流程

4.1 示例场景说明

本示例是一个DataArts Studio全流程入门教程，旨在介绍如何在DataArts Studio平台完成端到端的全流程数据运营。

本案例基于某市的出租车出行数据，选择MRS Hive作为数据湖底座，使用DataArts Studio实施全流程数据治理。期望通过实施数据治理达到以下目标：

- 数据标准化、模型标准化
- 统一统计口径，提供高质量数据报告
- 数据质量监控告警
- 统计每天收入
- 统计某月收入
- 统计不同支付类型收入占比

流程简介

本入门指导将参考如**表4-1**所示的流程，实现示例场景的数据治理。

表 4-1 DataArts Studio 数据治理流程

主流程	说明	子任务	操作指导
步骤1：流程设计	在使用DataArts Studio前，您需要进行业务调研和需求分析设计。	需求分析、业务调研与业务流程设计	**需求分析** **业务调研**
步骤2：准备工作	如果您是第一次使用DataArts Studio，请先完成创建DataArts Studio实例、创建工作空间等一系列操作。	使用DataArts Studio前的准备	**准备工作**

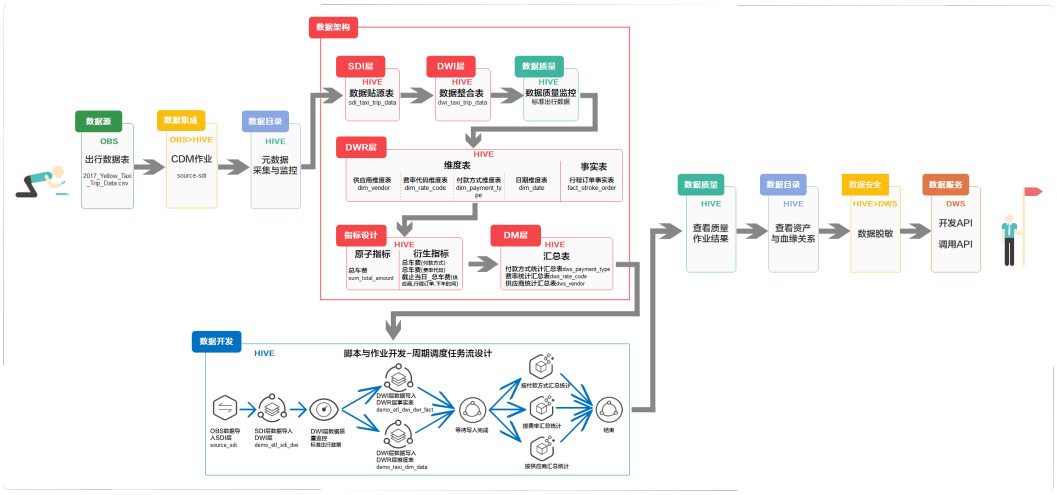
主流程	说明	子任务	操作指导
步骤3：数据集成	通过DataArts Studio平台将源数据上传或者接入到云上。 可以将离线或历史数据集成到云上。提供同构/异构数据源之间数据集成的服务，支持单表/文件迁移、整库迁移、增量集成，支持自建和云上的文件系统，关系数据库，数据仓库，NoSQL，大数据云服务，对象存储等数据源。	数据集成	创建集群 新建数据迁移的源连接、目的连接 新建表/文件迁移作业
步骤4：元数据采集	为了在DataArts Studio平台中对迁移到云上的原始数据层进行管理和监控，必须先对其元数据进行采集并监控。	元数据采集	采集并监控元数据
步骤5：数据架构	数据架构以关系建模、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。	准备工作	添加审核人 管理配置中心
		主题设计	主题设计
		标准管理	新建码表并发布 新建数据标准并发布
		关系建模	关系建模：新建SDI层和DWI层两个模型
		维度建模	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
		指标设计	指标设计：新建并发布技术指标
		数据集市建设	维度建模：在DM层新建并发布汇总表
步骤6：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、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
		脚本开发	脚本开发
		作业开发	作业开发
		运维调度	运维调度

主流程	说明	子任务	操作指导
步骤7：数据质量监控	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。	业务指标监控	监控业务指标
		数据质量监控	查看质量作业
步骤8：数据目录管理	在DataArts Studio数据目录模块中，您可以查看数据地图。	数据地图	查看业务资产和技术资产
步骤9：服务退订（可选）	进行服务退订，避免持续产生费用。	服务退订	服务退订（可选）

4.2 步骤 1：流程设计

本入门指南以某市出租车出行数据为例，统计某出租车供应商2017年度的运营数据。基于需求分析和业务调研，数据治理业务流程设计如图4-1所示，后续的数据治理操作均基于本业务流程完成。

图 4-1 流程设计



需求分析

通过需求分析，可以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。

在本示例场景下，当前面临的数据问题如下：

- 未建立标准化模型

- 数据字段命名不标准、不规范
- 数据内容不标准，数据质量不可控
- 统计口径不一致，困扰业务决策

通过DataArts Studio实施数据治理，期望能够达到以下目标：

- 数据标准化、模型标准化
- 统一统计口径，提供高质量数据报告
- 数据质量监控告警
- 统计每天收入
- 统计某月收入
- 统计不同支付类型收入占比

业务调研

在开始使用DataArts Studio前，您可以通过业务调研，明确业务过程中所需的DataArts Studio组件功能，并分析后续的业务负载情况。

表 4-2 业务调研表

序号	收集项	需收集信息描述	调研结果	填写说明
1	工作空间	企业大数据相关部门的组织和关系调查	本示例不涉及	用于合理规划工作空间，降低空间相互依赖的复杂度
		各组织部门之间对数据、资源的访问控制	本示例不涉及	涉及到用户的权限和资源权限控制
2	数据集成	有哪些数据源要迁移，数据源版本	CSV格式的数据，存储于OBS桶	-
		每种数据源的数据全量数据规模	2，114 字节	-
		每种数据源每天的增量数据规模	本示例不涉及	-
		迁移目的端数据源种类以及版本	迁移到MRS Hive3.1	-
		数据的迁移周期：天、小时、分钟还是实时迁移	天	-
		数据源与目的数据源之间的网络带宽	100MB	-
		数据源和集成工具之间的网络连通性描述	本示例不涉及	-

序号	收集项	需收集信息描述	调研结果	填写说明
		数据库类迁移，调研表的个数，最大表的规模	本示例不涉及，本示例需要从OBS文件迁移到数据库	了解数据库迁移的作业规模，了解最大表的迁移时间是否可接受
		文件类迁移，文件的个数，有没有单文件超过TB级文件	本示例的CSV文件仅1个，未超过TB级	-
3	数据开发	是否需要作业编排调度?	是	-
		编排调度会涉及哪些服务，例如MRS、DWS、CDM等?	本示例涉及DataArts Studio数据集成和数据质量、MRS Hive	了解作业的场景，用于进一步调查平台能力与客户场景匹配度
		作业数量规模是多少?	本示例作业数量在20个以内	大致了解作业的规模，通常用算子数来描述，可通过表的数量估计
		每日作业调度次数是多少?	没有特殊要求，次数不限	根据DataArts Studio各销售版本的调度限额，确定DataArts Studio的版本
		数据开发人员的数量是多少?	1个	-
4	数据架构	数据现状，有哪些数据源，多少张表?	本示例的CSV文件仅1个	原始端分析， 了解数据来源与整体概况
		业务需求，有哪些业务，有什么需求， 想要获得什么价值?	数据标准化、模型标准化，并灵活统计收入情况	目的端分析，了解数据治理以及数字化是为了什么
		数据调研，数据概况，数据标准程度，行业标准概况?	本示例不涉及	过程端分析，了解数据治理过程需要做到的标准与质量的遵从
5	数据质量	有哪些需求，需要获得什么价值?	监控数据质量	支持更多数据源和更多规则的监控
		作业数量规模是多少?	本示例仅1个	用户可手动创建几十个作业，也可以在数据架构中配置自动生成数据质量作业。如果调用数据质量监控的创建接口，则可创建超过100个质量作业

序号	收集项	需收集信息描述	调研结果	填写说明
		用户的使用场景?	对DWI层数据进行标准化清洗	一般在数据加工前后，对数据的质量通过六大维度的规则进行质量监控，当发现不符合规则的异常数据时向用户发送报警
6	数据目录	需要支持哪些数据源?	MRS Hive	-
		数据资产的数量规模有多大?	本示例表在百级以内	最大可支持100w数据表的管理
		元数据采集的调度频率是多少?	本示例不涉及	支持按照小时、天、周为周期运行采集任务
		元数据采集的重要指标包括什么?	本示例不涉及	表名称、字段名称、责任人、描述信息、创建时间等
		标签的使用场景是什么?	本示例不涉及	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，方便用户进行查询
7	数据安全	需要对哪些数据源进行访问管理?	本示例不涉及	仅支持对MRS中HDFS、Hive、HBase、Yarn、Kafka、Storm、Elasticsearch 七大主要服务提供权限访问控制
		需要识别哪些数据密级?	本示例不涉及	支持最大定义10个数据密级
		需要对哪些数据源数据进行脱敏?	本示例需要将MRS标准数据出行表脱敏至DWS	仅支持对DWS和MRS数据源数据进行脱敏。
		需要对哪些数据源数据进行水印管理?	本示例不涉及	仅支持对DWS和MRS数据源数据进行水印嵌入。
8	数据服务	需要开放哪些数据源数据?	收入汇总表	这些数据源一般存储的为数据仓库建设后的最后一层的表。这种表一般业务含金量比较高，记录条数比较少，可以直接展示
		每日数据调用量是多少?	本示例不涉及	若取数逻辑复杂造成数据库响应时间较长，调用量会下降
		每秒数据调用量峰值是多少?	本示例不涉及	根据不同规格和具体的取数逻辑有所增减

序号	收集项	需收集信息描述	调研结果	填写说明
		单次数据调用平均时延是多少?	本示例不涉及	数据库响应耗时与用户取数逻辑相关
		是否需要数据访问记录?	本示例不涉及	-
		数据访问方式，内网还是外网?	本示例不涉及	-
		数据服务开发人员数量是多少?	1	-

4.3 步骤 2：准备工作

使用 DataArts Studio 前的准备

如果您是第一次使用DataArts Studio，请参考**准备工作**章节完成注册华为账号、购买DataArts Studio实例（DataArts Studio企业版）、创建工作空间等一系列操作。然后进入到对应的工作空间，即可开始使用DataArts Studio。

本入门示例，为了演示DataArts Studio数据治理的全流程，华为账号需要具有DataArts Studio的所有执行权限。

准备数据源

本入门指南以某市出租车出行数据为例，统计某出租车供应商2017年度的运营数据。

说明

本示例演示的原始数据来自于**NYC开放数据平台**。

为方便演示，**您无需获取原始数据**，本示例提供了模拟原始数据的样例数据供您使用。您可以参考下文的样例数据准备方法，将样例数据存储为CSV文件，将CSV文件上传至OBS服务中，然后再使用DataArts Studio数据集成将样例数据集成到其他云服务中。

样例数据准备方法如下：

步骤1 创建一个CSV文件（UTF-8无bom格式），文件名称为“2017_Yellow_Taxi_Trip_Data.csv”，将后文提供的样例数据复制粘贴到CSV文件中，然后保存CSV文件。

以下是Windows下生成.csv文件的办法之一：

1. 使用文本编辑工具（例如记事本等）新建一个txt文档，将后文提供的样例数据复制进文档中。注意复制后检查数据的行数及数据分行的正确性（注意，如果是从PDF文档中复制样例数据，单行的数据过长时会产生换行，需手动重新调整为单行）。
2. 单击“文件 > 另存为”，在弹出的对话框中，“保存类型”选择为“所有文件(*.*)”，在“文件名”处输入文件名和.csv后缀，选择“UTF-8”编码格式（不能带BOM），则能以CSV格式保存该文件。

步骤2 将源数据CSV文件上传到OBS服务。

1. 登录控制台，选择“存储 > 对象存储服务 OBS”，进入OBS控制台。
2. 单击“创建桶”，然后根据页面提示配置参数，创建一个名称为“fast-demo”的OBS桶。

   说明

   为保证网络互通，OBS桶区域请选择和DataArts Studio实例相同的区域。如果需要选择企业项目，也请选择与DataArts Studio实例相同的企业项目。

   使用OBS控制台创建桶的操作，请参见《对象存储服务控制台指南》中的**创建桶**。
3. 上传数据到名称为“fast-demo”的OBS桶中。

   使用OBS控制台上传文件的操作，请参见《对象存储服务控制台指南》中的**上传文件**。

----结束

样例数据如下。

```
VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,RatecodeID,store_and_fwd_flag,PULocationID,DOLocationID,payment_type,fare_amount,extra,mta_tax,tip_amount,tolls_amount,improvement_surcharge,total_amount
2,02/14/2017 04:08:11 PM,02/14/2017 04:21:53 PM,1,0.91,1,N,237,163,2,9.5,1,0.5,0,0,0.3,11.3
2,02/14/2017 04:08:11 PM,02/14/2017 04:19:29 PM,2,1.03,1,N,237,229,1,8.5,1,0.5,2.06,0,0.3,12.36
1,02/14/2017 04:08:12 PM,02/14/2017 04:19:44 PM,1,1.6,1,N,186,163,2,9,1,0.5,0,0,0.3,10.8
1,02/14/2017 04:08:12 PM,02/14/2017 04:19:15 PM,1,1.2,1,N,48,48,2,8.5,1,0.5,0,0,0.3,10.3
2,02/14/2017 04:08:12 PM,02/14/2017 04:13:38 PM,5,0.61,1,N,161,162,1,5.5,1,0.5,2.19,0,0.3,9.49
2,02/14/2017 04:08:12 PM,02/14/2017 05:35:11 PM,1,19.31,2,N,152,132,1,52,4.5,0.5,12.57,5.54,0.3,75.41
1,02/14/2017 04:08:13 PM,02/14/2017 04:20:53 PM,1,1.9,1,N,236,143,1,10.5,1,0.5,1.85,0,0.3,14.15
2,02/14/2017 04:08:13 PM,02/14/2017 04:15:54 PM,1,0.61,1,N,48,164,1,6.5,1,0.5,1.66,0,0.3,9.96
2,02/14/2017 04:08:13 PM,02/14/2017 04:41:40 PM,1,6.04,1,N,244,262,1,25,1,0.5,6.7,0,0.3,33.5
2,02/14/2017 04:08:13 PM,02/14/2017 04:17:31 PM,1,1.39,1,N,170,234,1,8,1,0.5,1,0,0.3,10.8
2,02/14/2017 04:08:14 PM,02/14/2017 04:54:11 PM,2,10.12,1,N,140,189,1,37.5,1,0.5,7,0,0.3,46.3
2,02/14/2017 04:08:14 PM,02/14/2017 04:13:56 PM,1,0.71,1,N,179,7,2,5.5,1,0.5,0,0,0.3,7.3
2,02/14/2017 04:08:14 PM,02/14/2017 05:04:24 PM,1,18.1,2,N,263,132,1,52,4.5,0.5,15.71,5.54,0.3,78.55
2,02/14/2017 04:08:14 PM,02/14/2017 04:08:47 PM,1,0.02,1,N,231,231,2,2.5,1,0.5,0,0,0.3,4.3
2,02/14/2017 04:08:15 PM,02/14/2017 04:18:13 PM,1,1.34,1,N,100,162,1,8,1,0.5,1.2,0,0.3,11
1,02/14/2017 04:08:16 PM,02/14/2017 04:19:01 PM,1,1.8,1,N,239,151,1,9,1,0.5,2.15,0,0.3,12.95
2,02/14/2017 04:08:16 PM,02/14/2017 04:15:57 PM,1,1.06,1,N,68,170,1,6.5,1,0.5,1,0,0.3,9.3
2,02/14/2017 04:08:16 PM,02/14/2017 04:20:08 PM,2,1.5,1,N,161,142,1,9,1,0.5,2.16,0,0.3,12.96
2,02/14/2017 04:08:16 PM,02/14/2017 04:11:56 PM,1,0.62,1,N,87,88,2,4.5,1,0.5,0,0,0.3,6.3
2,02/14/2017 04:08:16 PM,02/14/2017 04:13:20 PM,1,0.88,1,N,262,236,2,5.5,1,0.5,0,0,0.3,7.3
```

数据说明如下：

表 4-3 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

序号	字段名称	字段描述
5	trip_distance	行驶距离
6	ratecodeid	费率代码 取值如下： 1=Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
7	store_fwd_flag	存储转发标识
8	PULocationID	上车地点
9	DOLocationID	下车地点
10	payment_type	付款方式代码 取值如下： 1=Credit card 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

准备数据湖

在使用DataArts Studio前，您需要根据业务场景选择符合需求的云服务或数据库作为数据湖底座，由数据湖底座提供存储和计算的能力，DataArts Studio基于数据湖底座进行一站式数据开发、治理和服务。

DataArts Studio平台支持对接如DLI、DWS、MRS Hive等云服务，也支持对接如MySQL、Oracle等传统数据库，支持程度各有不同，详情请参见**DataArts Studio支持的数据源**章节。

本示例选择MapReduce服务（MRS）的Hive组件作为DataArts Studio平台的数据湖底座。您需要先创建一个MRS安全集群（即开启“Kerberos认证”的MRS集群，安全性更强），具体操作请参见**创建集群**。

为确保MRS集群与DataArts Studio实例网络互通，MRS集群需满足如下要求：

- MRS集群必须包含Hive组件。
- 如需使用基于DataArts Studio数据架构的数据标准自动生成质量作业的功能，MRS集群版本必须是2.0.3及以上版本，集群必须包含Hive和Spark组件，集群总节点数至少4个。本示例需要使用该功能，因此必须满足这个条件。

  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。

在管理中心创建数据连接

准备好数据湖之后，在DataArts Studio管理中心模块中创建数据连接，以便连接作为数据湖的云服务。

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“数据连接”页面，单击“创建数据连接”按钮。

图 4-2 数据连接

步骤4 在弹出窗口中，配置数据连接参数，完成配置后，单击“确定”完成数据连接的创建。

此处创建MapReduce服务（MRS Hive）数据连接，参数配置如**图4-3**所示。

- **数据连接类型**：MapReduce服务（MRS Hive）。
- **数据连接名称**：mrs_hive_link。

- **标签**：可选参数。您可以输入新的标签名称，也可以在下拉列表中选择已有的标签。
- **集群名**：选择已有的MRS集群。
- **用户名**：新建的Kerberos认证用户。注意，MRS的策略中，admin用户是默认的管理页面用户，这个用户无法作为使用Kerberos认证集群的认证用户来使用。因此如果要为使用Kerberos认证的MRS集群创建连接，需要执行如下操作：

  a. 使用admin账户登录MRS服务的Manager页面。

  b. 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。

  c. 在Manager页面选择“系统 > 权限 > 用户”，单击“添加用户”，添加一个专有用户作为kerberos认证用户，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。

  d. 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。

  e. 同步IAM用户。

     i. 登录MRS管理控制台。

     ii. 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。

     iii. 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。

- **密码**：Kerberos认证用户对应的密码。
- **KMS密钥**：选择一个KMS密钥，使用KMS密钥对敏感数据进行加密。如果未创建KMS密钥，请单击“访问KMS”进入KMS控制台创建一个密钥。
- **连接方式**：通过代理连接。
- **绑定Agent**：需选择一个数据集成集群作为连接代理，该集群和MRS集群必须处于相同的区域、可用区、VPC和子网，并且安全组规则允许两者网络互通。本示例可选择创建DataArts Studio实例时自动创建的数据集成集群。

  如需连接MRS 2.x版本的集群，请选择2.x版本的数据集成集群作为Agent代理。

图 4-3 创建 MRS Hive 数据连接

----结束

创建数据库

根据数据湖治理落地流程，建议您在数据湖中为SDI层、DWI层、DWR层和DM层分别创建一个数据库，从而对数据进行分层分库。数据分层是后面在数据架构中将涉及到的概念，此处先简单了解即可，在数据架构时将深入了解与操作。

- 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。
- 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。

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

创建数据库的操作，一般您需要在数据湖产品中完成。

在本示例中，您可以参考以下任意一种方式在MRS Hive中创建数据库。

- 您可以在DataArts Studio数据开发模块中，可视化方式创建数据库，具体操作请参见**新建数据库**章节。
- 您可以通过在DataArts Studio数据开发模块或MRS客户端上，开发并执行用于创建数据库的SQL脚本，从而创建数据库。在DataArts Studio数据开发模块开发脚本的具体操作请参见**开发SQL脚本**章节；在MRS客户端开发脚本的具体操作请参见**从零开始使用Hive**章节。创建数据库的Hive SQL命令如下所示：

```
--创建SDI贴源层数据库
CREATE DATABASE demo_sdi_db;

--创建DWI多源整合层数据库
CREATE DATABASE demo_dwi_db;

--创建DWR明细数据层数据库
CREATE DATABASE demo_dwr_db;

--创建DM数据集市层数据库
CREATE DATABASE demo_dm_db;
```

创建数据表

基于样例数据，创建一个原始表，用于存储原始数据。从文件迁移到数据库的场景，您需要预先创建目标数据表。由于本示例的数据源源端为OBS上的CSV文件，而非数据库，在使用DataArts Studio数据集成将数据迁移上云时，不支持自动创建目标表的功能，因此，您需要在目的端（MRS服务）先建好表。

说明

在使用DataArts Studio进行数据集成时，关系型数据库之间的迁移和关系型数据库到Hive的迁移支持自动创建目标表。这种情况下可以不提前在目的端数据库中预先创建目标表。

执行如下SQL语句，在demo_sdi_db数据库中，创建一个原始表，用于存储原始数据。

在本示例中，您可以参考以下任意一种方式在MRS Hive中创建数据表。

- 您可以在DataArts Studio数据开发模块中，可视化方式创建数据表，具体操作请参见**新建数据表**章节。
- 您可以通过在DataArts Studio数据开发模块或MRS客户端上，开发并执行用于创建数据表的SQL脚本，从而创建数据表。在DataArts Studio数据开发模块开发脚本的具体操作请参见**开发SQL脚本**章节；在MRS客户端开发脚本的具体操作请参见**从零开始使用Hive** 在demo_sdi_db数据库中创建一个原始数据表的Hive SQL命令如下所示：

```
DROP TABLE IF EXISTS `sdi_taxi_trip_data`;

CREATE TABLE demo_sdi_db.`sdi_taxi_trip_data` (
  `VendorID` BIGINT COMMENT '',
  `tpep_pickup_datetime` TIMESTAMP COMMENT '',
  `tpep_dropoff_datetime` TIMESTAMP COMMENT '',
  `passenger_count` BIGINT COMMENT '',
  `trip_distance` DECIMAL(10,2) COMMENT '',
  `ratecodeid` BIGINT COMMENT '',
  `store_fwd_flag` STRING COMMENT '',
  `PULocationID` STRING COMMENT '',
  `DOLocationID` STRING COMMENT '',
  `payment_type` BIGINT COMMENT '',
  `fare_amount` DECIMAL(10,2) COMMENT '',
```

```
  `extra` DECIMAL(10,2) COMMENT '',
  `mta_tax` DECIMAL(10,2) COMMENT '',
  `tip_amount` DECIMAL(10,2) COMMENT '',
  `tolls_amount` DECIMAL(10,2) COMMENT '',
  `improvement_surcharge` DECIMAL(10,2) COMMENT '',
  `total_amount` DECIMAL(10,2) COMMENT ''
);
```

4.4 步骤 3：数据集成

本章节将介绍如何使用DataArts Studio数据集成将源数据批量迁移到云上。

创建集群

批量数据迁移集群提供数据上云和数据入湖的集成能力，全向导式配置和管理，支持单表、整库、增量、周期性数据集成。DataArts Studio基础包中已经包含一个数据集成的集群，如果无法满足业务需求，在购买DataArts Studio基础包实例后，您可以根据实际需求购买批量数据迁移增量包。

购买数据集成增量包的具体操作请参考**购买DataArts Studio增量包**章节。

新建数据迁移的源连接、目的连接

步骤1 登录DataArts Studio控制台。选择实例，单击“进入控制台”，选择对应工作空间的“数据集成”模块，进入数据集成页面。

图 4-4 选择数据集成

步骤2 在左侧导航栏中单击“集群管理”进入“集群管理”页面。然后，在集群列表中找到所需要的集群，单击“作业管理”。

图 4-5 集群管理

步骤3 进入作业管理后，选择“连接管理”。

图 4-6 连接管理

步骤4 创建两个连接，一个源连接OBS连接，用于读取存储在OBS上的原始数据，一个目的连接MRS Hive连接，用于将数据写入MRS Hive数据库中。

单击“新建连接”，进入相应页面后，选择连接器类型“对象存储服务（OBS）”，单击“下一步”，然后如下图所示配置连接参数，单击“保存”。

图 4-7 创建 OBS 连接

表 4-4 OBS 连接的参数

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

参数名	说明	取值样例
OBS终端节点	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。	-
端口	数据传输协议端口，https是443，http是80。	443
OBS桶类型	用户下拉选择即可，一般选择为“对象存储”。	对象存储
访问标识(AK)	AK和SK分别为登录OBS服务器的访问标识与密钥。 您需要先创建当前账号的访问密钥，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4-8所示。 **图 4-8** 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)		-

在“连接管理”页面，再次单击“新建连接”，进入相应页面后，选择连接器类型为“MRS Hive”，单击“下一步”，然后如下图所示配置连接参数，单击“保存”。

图 4-9 创建 MRS Hive 连接

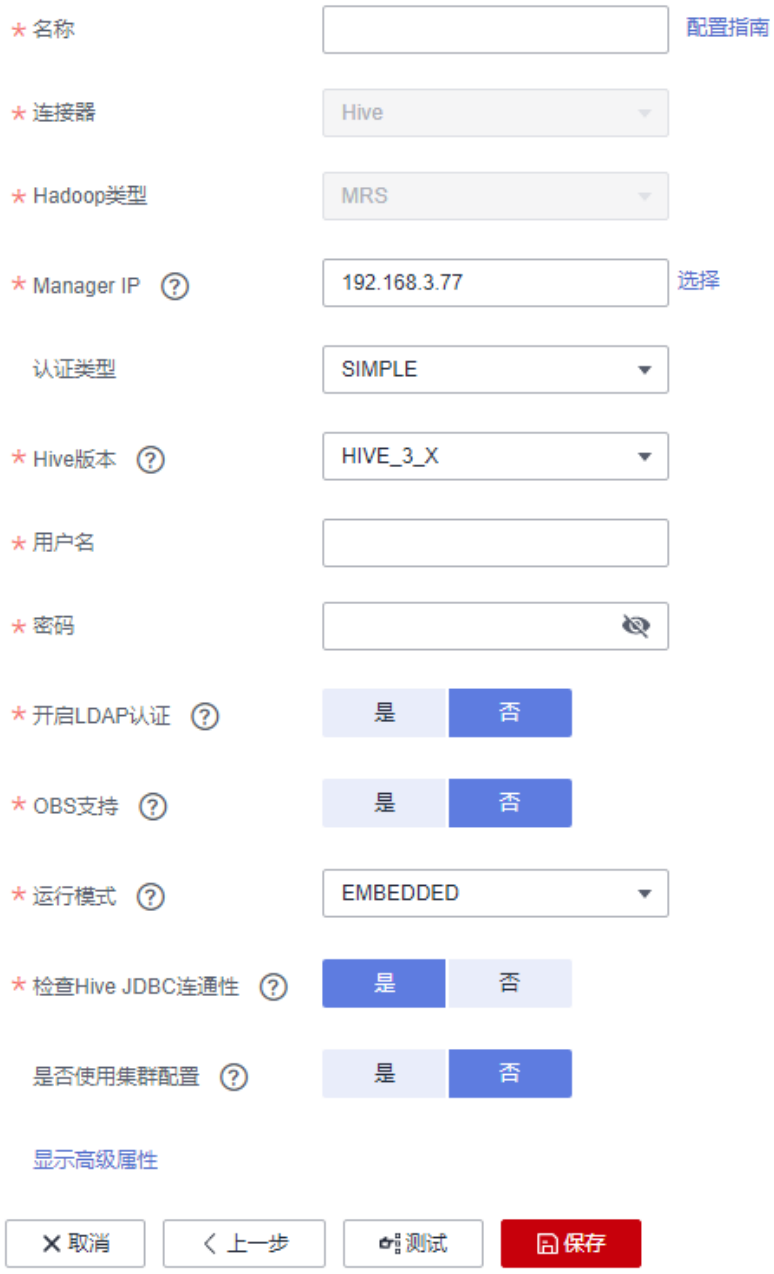


表 4-5 MRS Hive 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hivelink
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1

参数名	说明	取值样例
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。根据服务端Hive版本设置。	HIVE_3_X
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。 **说明** ● 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对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。	-
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

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4-10所示。 **图 4-10** 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
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

参数名	说明	取值样例
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hive_01

----结束

新建表/文件迁移作业

步骤1 在DataArts Studio数据集成控制台，进入“集群管理”页面，在集群列表中找到所需要的集群，单击“作业管理”。

步骤2 在“作业管理”页面，单击“表/文件迁移”，再单击“新建作业”。

图 4-11 表/文件迁移

步骤3 按照如下步骤完成作业参数的配置。

1. 如**图4-12**所示，配置作业名称、源端作业参数和目的端作业参数，然后单击“下一步”。
   - **作业名称**：source-sdi
   - **源端作业配置**
     - **源连接名称**：obs-link
     - **桶名**：fast-demo
     - **源目录或文件**：/2017_Yellow_Taxi_Trip_Data.csv
     - **文件格式**：CSV格式
     - **显示高级属性**：单击“显示高级属性”，在“高级属性”中，系统提供了默认值，请根据实际业务数据的格式设置各项参数。

       在本示例中，根据**准备数据源**中的样例数据格式，需注意以下参数的设置，其他参数经过一一确认均保留默认值即可。
       - **字段分隔符**：默认值为逗号，本示例保留默认值即可。
       - **前N行为标题行**：设置为“是”，本示例首行是标题行。
       - **标题行数**：配置为1。
       - **编码类型**：默认值为UTF-8，本示例保留默认值即可。
   - **目的端作业配置**
     - **目的连接名称**：mrs-link。

- **数据库名称**：demo_sdi_db。
- **表名**：sdi_taxi_trip_data。
- **导入前清空数据**：是。

  说明

  在本示例中，目的端作业参数“导入前清空数据”配置为“是”，表示每次作业运行都会先清空数据再导入。在实际业务中，请视情况而定，需谨慎设置，以免造成数据丢失。

图 4-12 作业配置

2. 在字段映射中，请参考以下说明配置字段映射以及日期字段的时间格式，如图4-13所示，配置完成后，单击“下一步”。
   - **字段映射**：在本示例中，由于数据迁移的目标表字段顺序和原始数据的字段顺序是一样的，因此这里不需要调整字段映射的顺序。

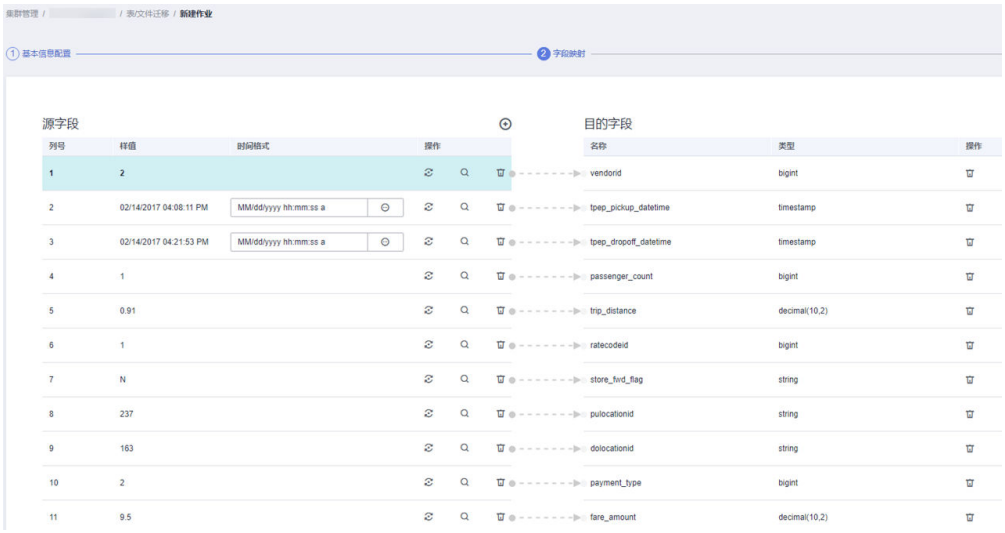
     如果目标表字段顺序和原始数据不一致，请一一将源字段指向含义相同的目的字段。请将鼠标移至某一个字段的箭头起点，当光标显示为“+”的形状时，按住鼠标，将箭头指向相同含义的目的字段，然后松开鼠标。
   - **时间格式**：样例数据中第2、第3个字段为时间字段，数据格式如“02/14/2017 04:08:11 PM”，因此此处设置这两个字段的时间格式为“MM/dd/yyyy hh:mm:ss a”，可以在输入框中手动输入该格式。

     时间格式请根据实际的数据格式进行设置，例如：

     yyyy/MM/dd HH:mm:ss 代表将时间转换为24小时制，例如2019/08/18 15:35:45。

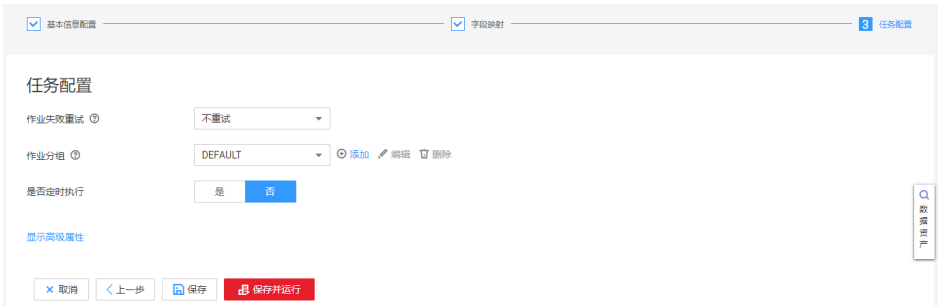
     yyyy/MM/dd hh:mm:ss a 代表将时间转换为12小时制，例如2019/06/27 03:24:21 PM。

图 4-13 字段映射



3. 根据需要配置任务的重试和定时执行。

图 4-14 任务配置



单击“显示高级属性”，可配置“抽取并发数”以及“是否写入脏数据”，如图4-15所示。

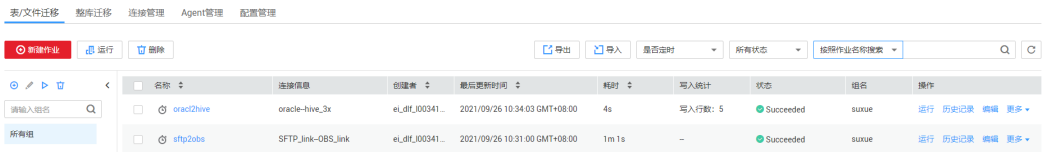
- **抽取并发数**：您可以根据业务量进行配置。数据源端如果是文件类型，当有多个文件时，增大并发数可以提升抽取速率。
- **是否写入脏数据**：建议配置为“是”，然后参考图4-15配置相关参数。脏数据是指与目的端字段不匹的数据，该数据可以被记录到指定的OBS桶中。用户配置脏数据归档后，正常数据可以写入目的端，迁移作业不会因脏数据中断。在本示例中，“OBS桶”配置为在准备数据源中的桶fast-demo，您需要前往OBS控制台，在fast-demo桶中单击“新建文件夹”创建一个目录，例如error-data，然后再将图4-15中的“脏数据目录”配置为该目录。

图 4-15 任务高级属性

步骤4 单击“保存”完成作业的创建。

返回“表/文件迁移”页面后，可在作业列表中查看到新建的作业。

图 4-16 迁移作业运行结果



----结束

4.5 步骤 4：元数据采集

为了在DataArts Studio平台中对迁移到云上的原始数据进行管理和监控，我们必须先在DataArts Studio数据目录模块中对SDI贴源层数据进行元数据采集并监控。

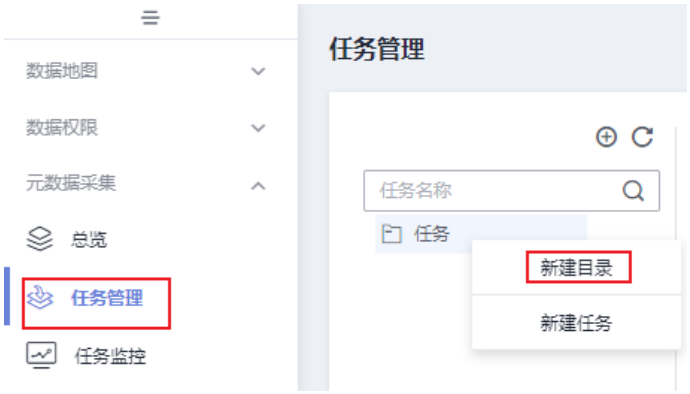
采集并监控元数据

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

图 4-17 选择数据目录

步骤2 在左侧导航栏选择“元数据采集 > 任务管理”，然后在任务树中选中一个目录并单击鼠标右键，选择菜单“新建目录”。在弹出框中输入目录名称，例如“transport”，选择目录，然后单击“确定”。

图 4-18 任务管理



步骤3 在任务树中选中transport目录，然后单击“新建”按钮，开始新建采集任务。

步骤4 按如下配置，新建采集任务transport_all。配置采集任务后，单击“下一步”。

图 4-19 新建采集任务-基本配置

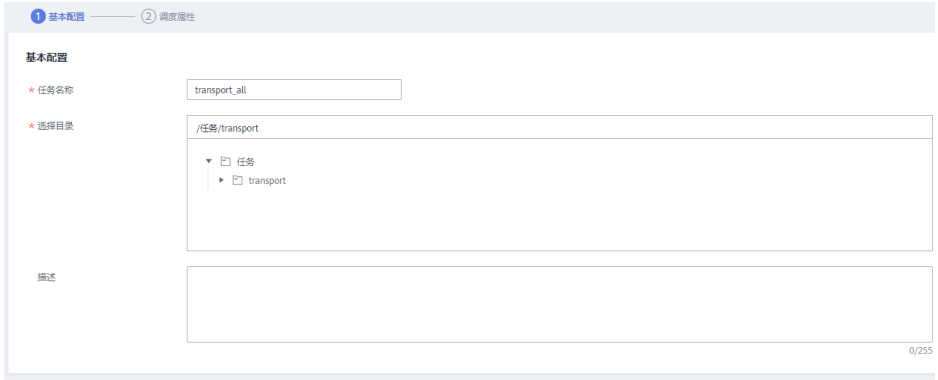
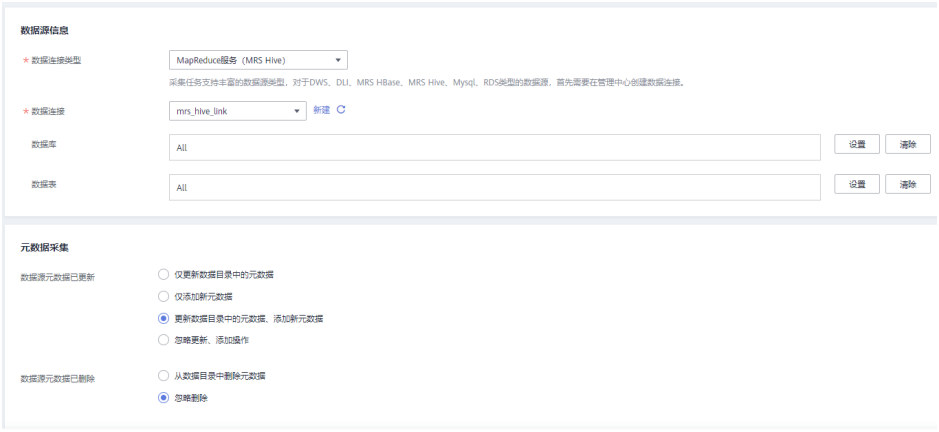


图 4-20 新建采集任务-元数据采集



步骤5 根据需要配置调度方式，配置完成后单击“提交”，完成采集任务的创建。

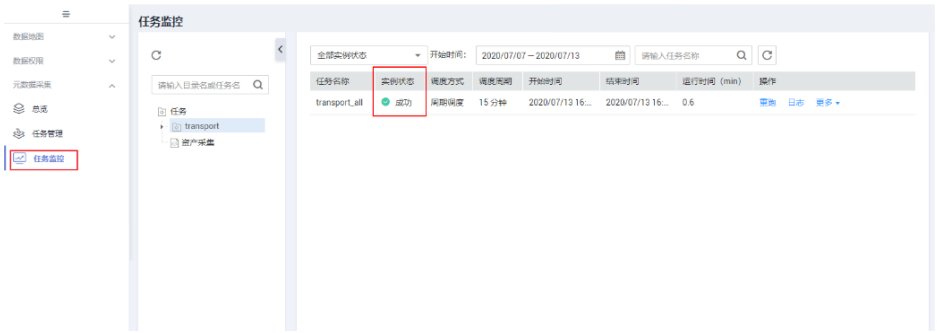
图 4-21 调度方式

步骤6 在采集任务列表中，找到刚才新建的采集任务，单击其所在行的“启动调度”按钮，启动周期采集任务。

图 4-22 启动调度

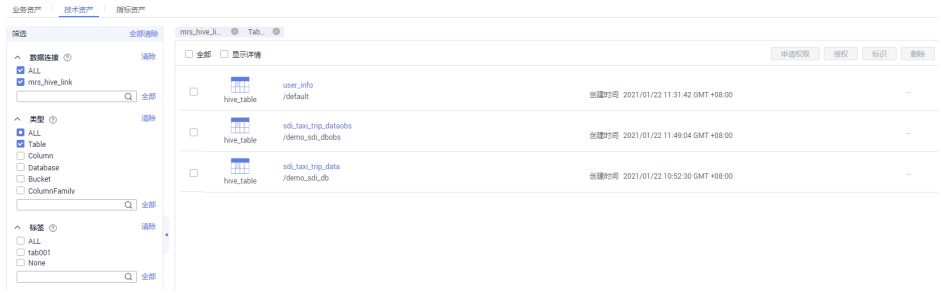
步骤7 在左侧导航树中，单击“任务监控”，查看采集任务是否成功。

图 4-23 查看监控任务



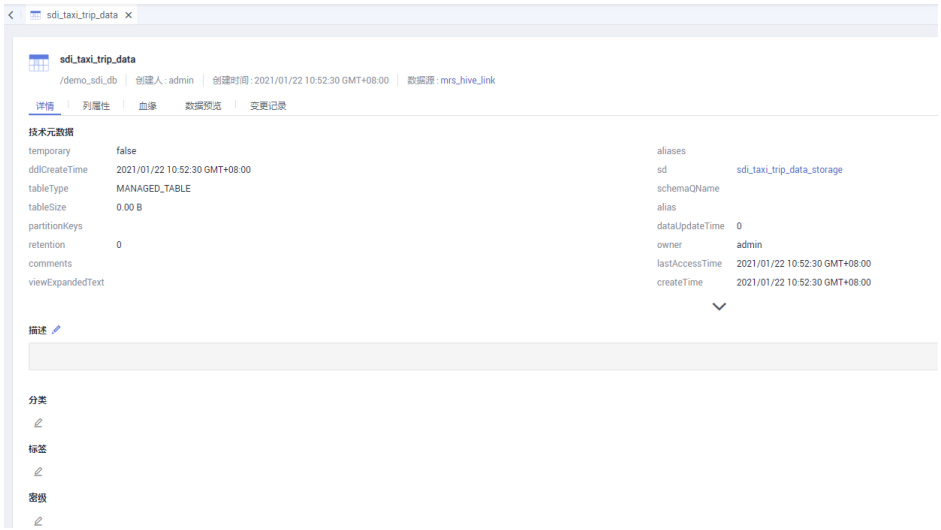
步骤8 当采集任务成功后，在左侧导航栏单击“数据目录”，选择“技术资产”页签，然后设置筛选条件，例如选中连接“mrs_hive_link”，以及选中“Table”，将显示符合条件的所有的表。

图 4-24 技术资产



步骤9 单击所需要的元数据名称，即可查看详情信息。

图 4-25 元数据详情



----结束

4.6 步骤 5：数据架构

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。

DataArts Studio数据架构建议的数据分层如下：

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

本章节为您介绍如何在DataArts Studio平台的“数据架构”模块中实现模型设计，流程如下。

添加审核人

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

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

   图 4-26 选择数据架构

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

您也可以添加自己当前账号为审核人，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 4-27 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 4-28 功能配置

3. 单击“确定”完成配置。

主题设计

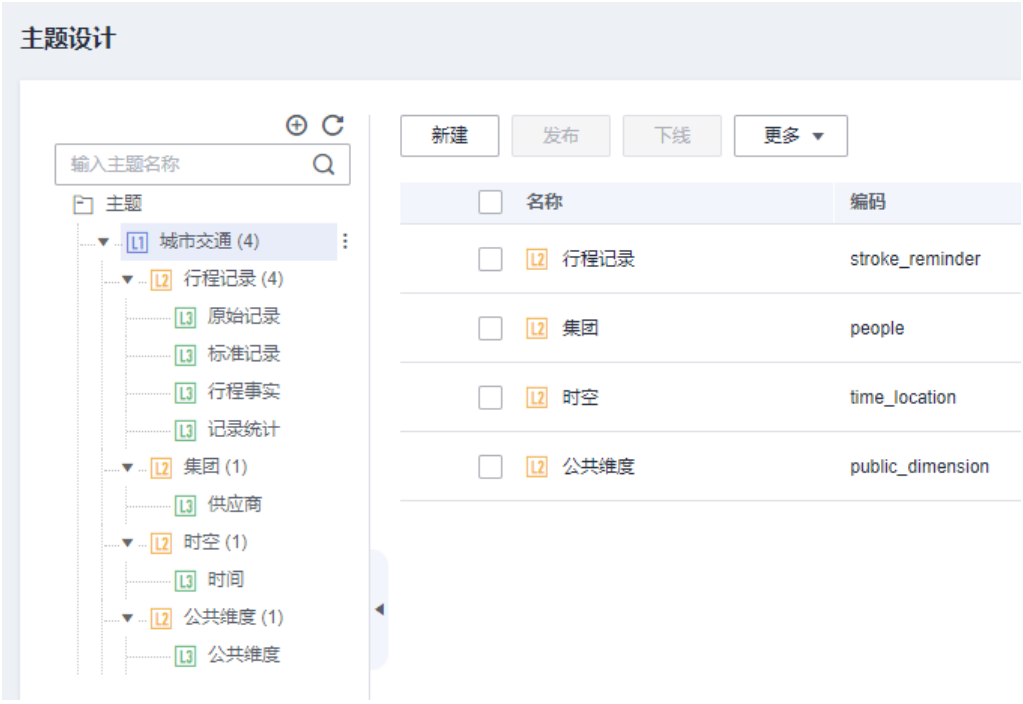
在本示例中，主题设计如表4-6所示，说明如下：

- 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 4-6 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
		时空	time_location	时间	date
		公共维度	public_dimension	公共维度	public_dimension

图 4-29 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。

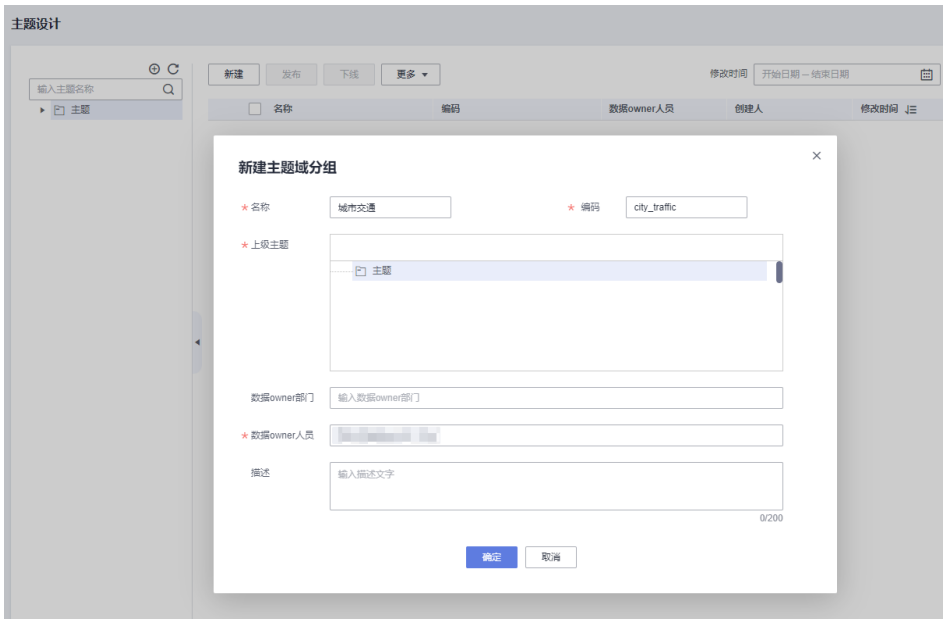
步骤3 在数据架构控制台，单击左侧菜单栏的“配置中心”。选择“ 主题流程配置”，使用默认的3层层级。

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 4-30 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 4-31 新建 L1 层主题



在弹出窗口中，按图4-31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 4-32 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。

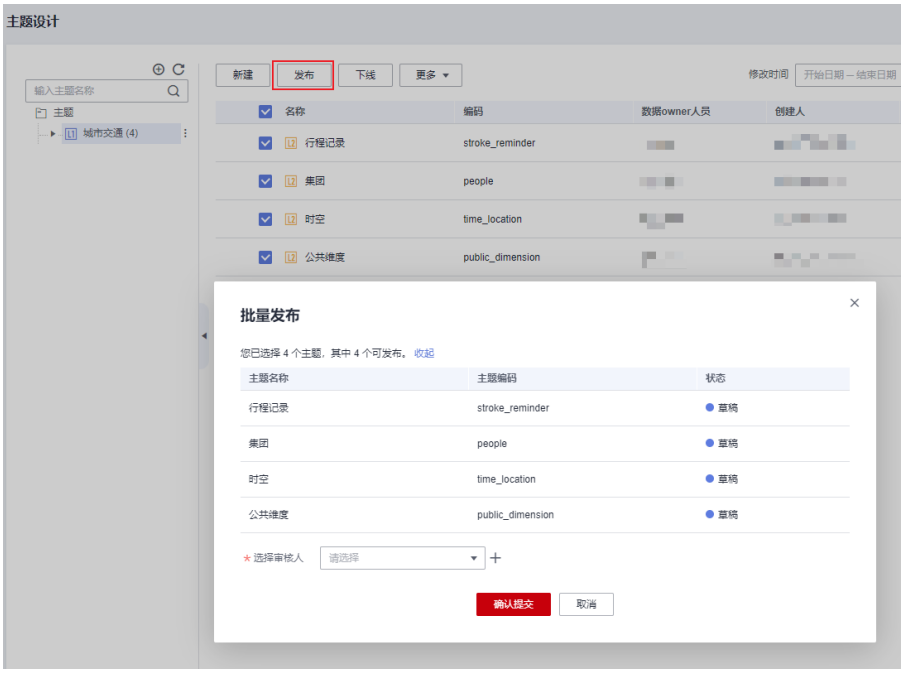
以主题域“行程记录”为例，新建主题域的步骤如下，其他主题域也请参照以下步骤进行添加：

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 4-33 创建 L2 层主题

2. 在弹出窗口中，“名称”和“编码”请参照表4-6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 4-34 发布主题域



步骤7 新建业务对象。

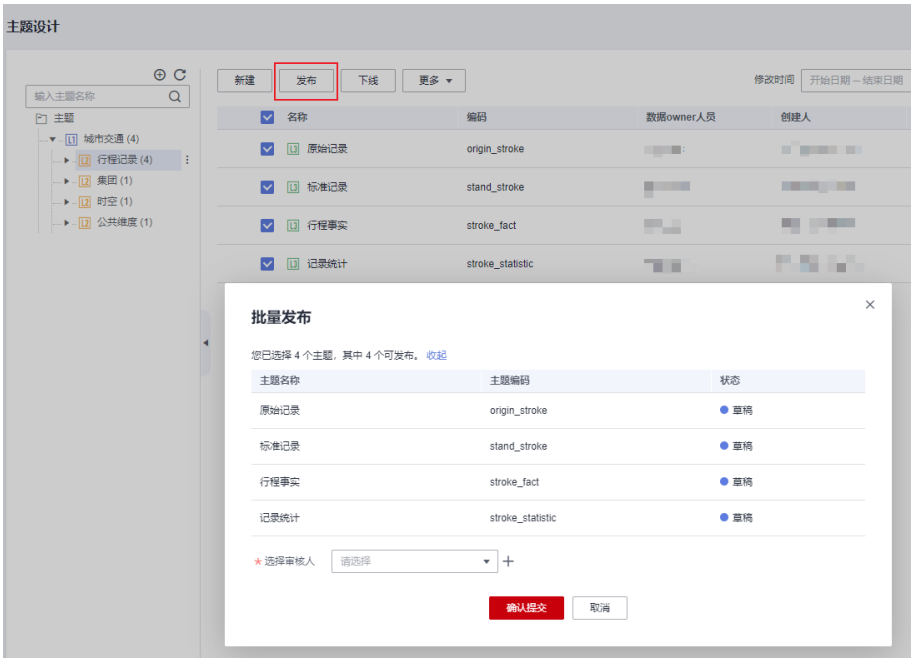
- 在主题域“行程记录”下，新建4个业务对象：原始记录、标准记录、行程事实、记录统计。
- 在主题域“集团”下，新建1个业务对象：供应商。

- 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4-6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 4-35 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4-7所示的3个码表：

表 4-7 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 4-36 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 4-37 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。

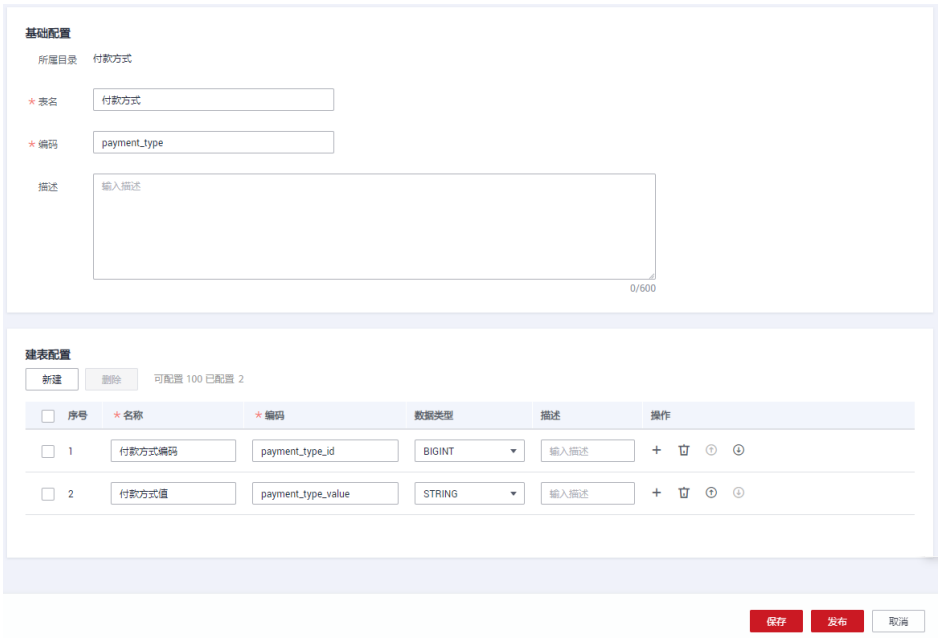
以新建“付款方式”码表为例，新建码表步骤如下，其他码表也请参照以下步骤完成新建：

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 4-38 码表管理

2. 在新建码表页面中，请参考表4-7配置参数，然后单击“保存”。

图 4-39 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 4-40 供应商码表

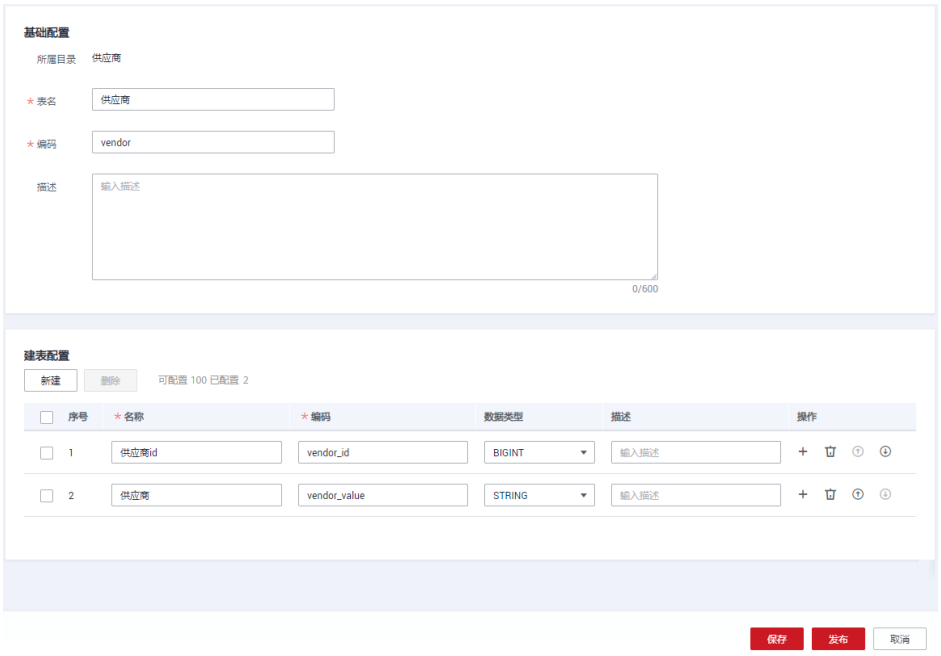
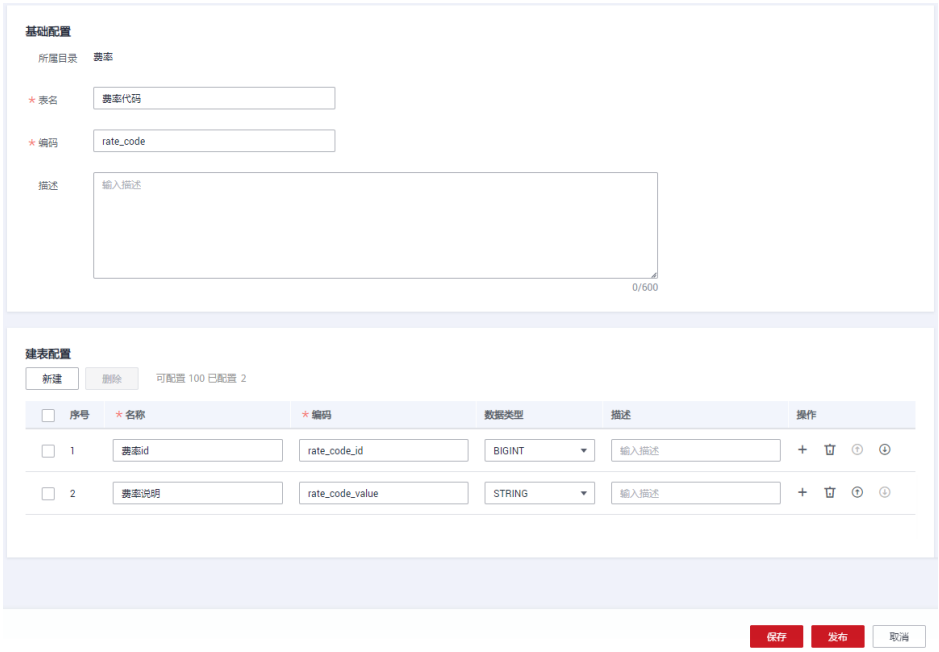


图 4-41 费率码表



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4-8所示的数值。

表 4-8 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4-9所示的数值。

表 4-9 供应商码表的数值

供应商id vendor_id	供应商 vendor_value
1	A Company

供应商id vendor_id	供应商 vendor_value
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4-10所示的数值。

表 4-10 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4-11所示的3个数据标准：

表 4-11 数据标准

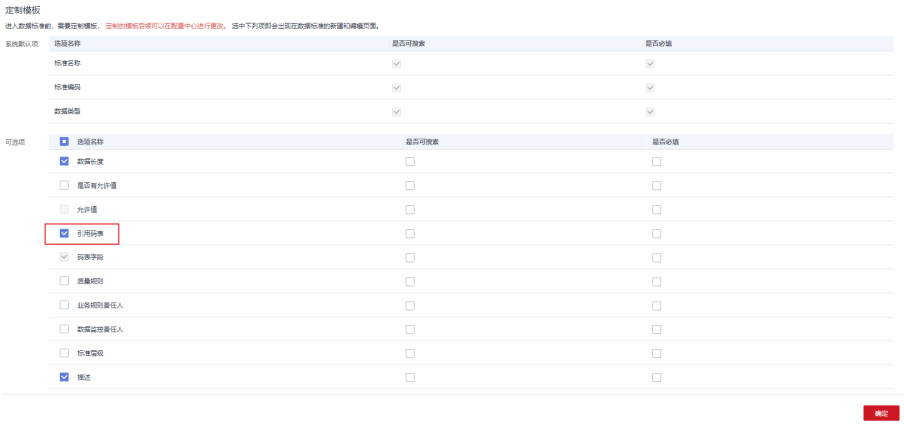
目录	*标准名称	*标准编码（自定义）	*数据类型	数据长度	引用码表	*码表字段	描述
付款方式	付款方式	payment_type	长整型(BIGINT)	无	付款方式	付款方式编码	无
供应商	供应商	vendor	长整型(BIGINT)	无	供应商	供应商id	无

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
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 4-42 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 4-43 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 4-44 数据标准-付款方式

图 4-45 数据标准-供应商

图 4-46 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

关系建模：新建 SDI 层和 DWI 层两个模型

在关系建模中，分别新建SDI层和DWI层两个关系模型，并通过逆向数据库导入原始数据表到SDI层的关系模型中，在DWI层模型中新建一个“标准出行数据”的标准化的业务表。

步骤1 在数据架构控制台，单击左侧导航树中的“关系建模”。

- 如果当前未创建过关系模型，系统会弹出“新建分层治理模型”提示框。您可以新建一个SDI层关系模型，命名为“sdi”，再新建一个DWI层关系模型，命名为“dwi”。单击“确定”即可。

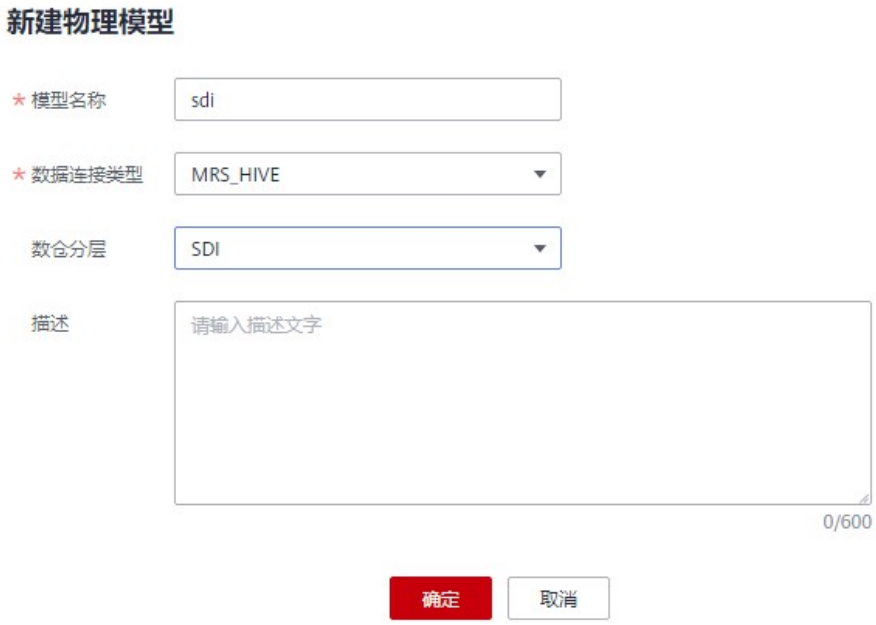
图 4-47 “新建分层治理模型”提示框

- 如果不是首次创建，单击 + 新建物理模型，如下图所示。

图 4-48 新建逻辑模型

a. 先新建一个SDI层关系模型，命名为“sdi”。在物理模型页签中，单击+，新建模型，配置如下参数，单击“确定”。

图 4-49 新建 SDI 物理模型



b. 再新建一个DWI层关系模型，命名为“dwi”。在物理模型页签中，单击+，新建模型，配置如下参数，单击“确定”。

图 4-50 新建 DWI 模型

步骤2 在“数仓分层”页签中，单击新建的SDI关系模型，展开，选中业务对象“城市交通 > 行程记录 > 原始记录”，单击“逆向数据库”，通过逆向数据库，导入原始表。

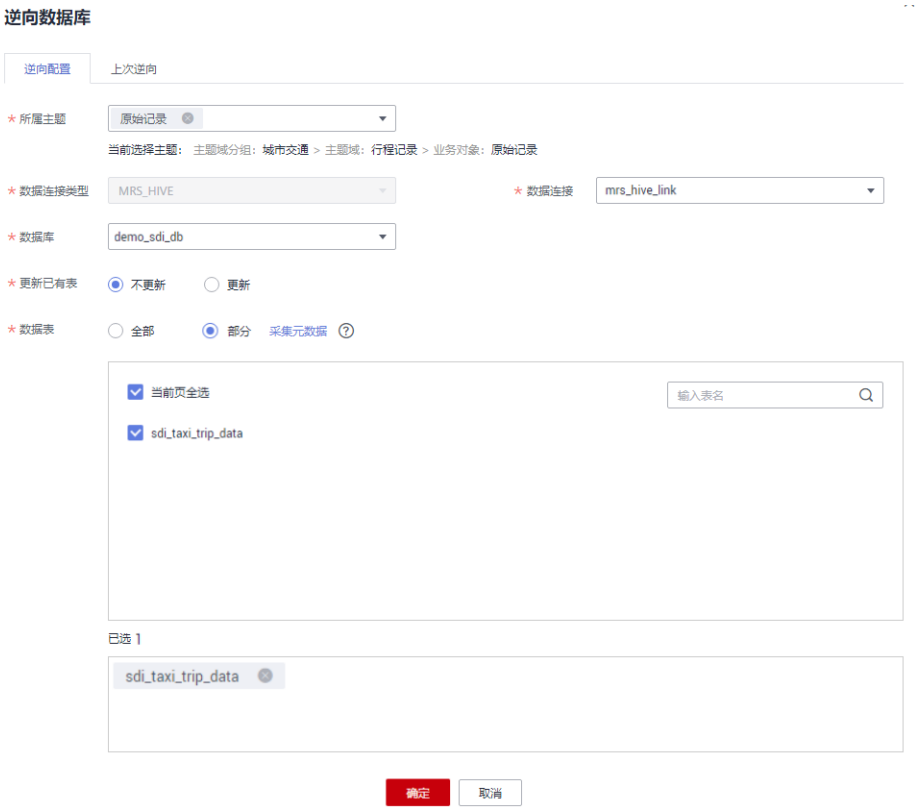
说明

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。数据资产采集的具体操作，请参见**步骤4：元数据采集**。

图 4-51 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 4-52 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 4-53 查看表

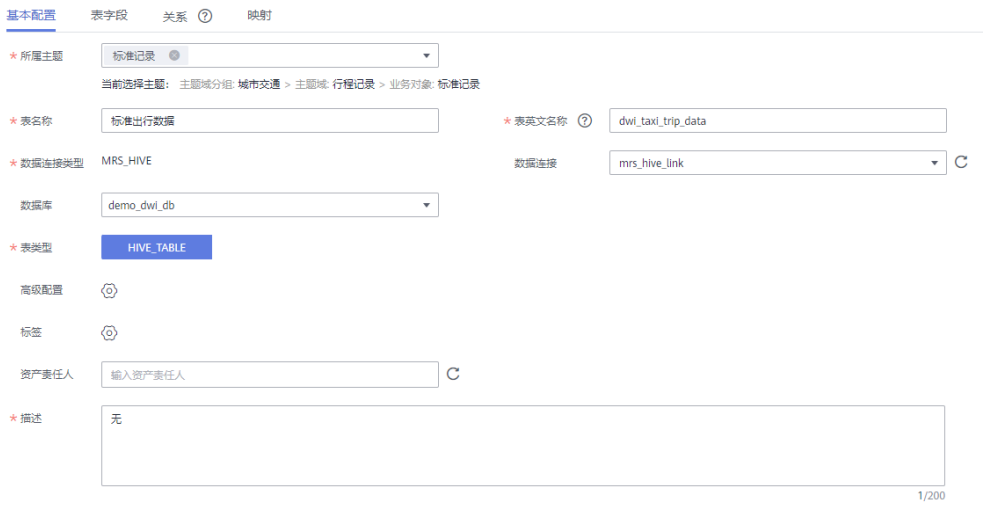
步骤3 请参照以下步骤，新建一个“标准出行数据”的标准化的业务表。

1. 在“数仓分层”页签中，单击新建的DWI关系模型，展开，选中DWI模型中的业务对象“城市交通 > 行程记录 > 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 4-12 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 4-54 行程数据表基本配置



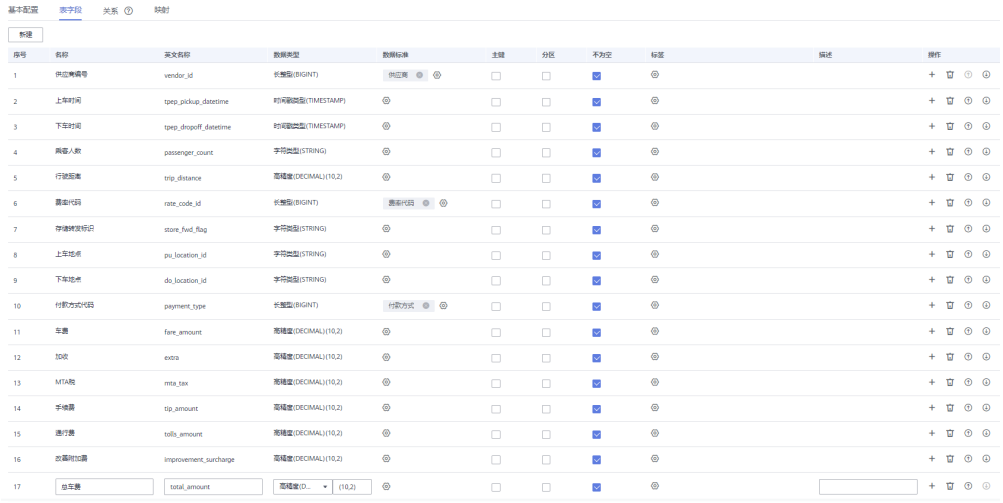
3. 单击“下一步”，进入“表字段”标签页。单击“新建”，在标准出行数据表中，依次添加如**表4-13**所示的字段，并单击字段供应商编号、费率代码、付款方式的“数据标准”列中的 按钮，分别关联数据标准“供应商”、“费率代码”和“付款方式”。添加完成后如**图4-55**所示。

表 4-13 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	-

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
6	费率代码	rate_code_id	长整型(BIGINT)	费率代码	不勾选	不勾选	勾选	_
7	存储转发标识	store_fwd_flag	字符类型(STRING)	-	不勾选	不勾选	勾选	_
8	上车地点	pu_location_id	字符类型(STRING)	-	不勾选	不勾选	勾选	_
9	下车地点	do_location_id	字符类型(STRING)	-	不勾选	不勾选	勾选	_
10	付款方式代码	payment_type	长整型(BIGINT)	付款方式	不勾选	不勾选	勾选	_
11	车费	fare_amount	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	_
12	加收	extra	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	_
13	MTA税	mta_tax	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	_
14	手续费	tip_amount	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	_
15	通行费	tolls_amount	高精度(DECIMAL)(10,2)	-	不勾选	不勾选	勾选	_
16	改善附加费	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 4-55 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**

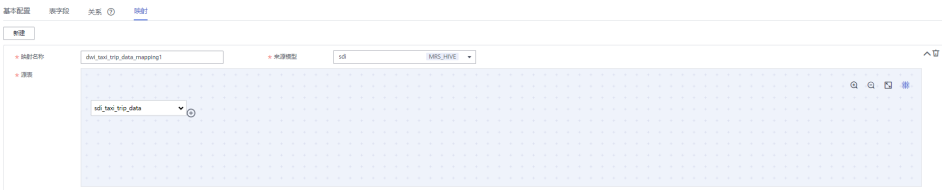
  完成表的新建后，您可以在表中为字段关联质量规则，完成关联后，当表发布成功后，就会在DataArts Studio数据质量中自动创建质量作业，如果当前表已经发布，则系统会自动更新质量作业。有关关联质量规则的更多信息，请参见**关联质量规则**。

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4-56**。
   - **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 4-56 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4-57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 4-57 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 4-58 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 4-59 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 4-60 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。

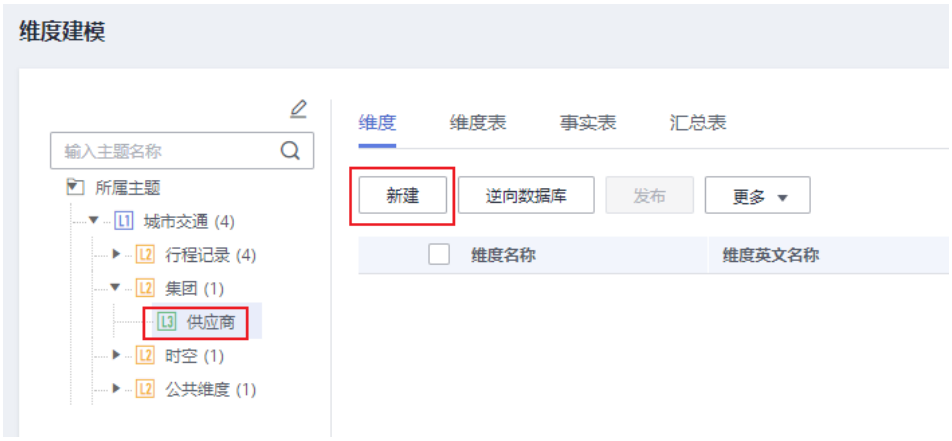
步骤2 新建如表4-14所示的3个码表维度。

表 4-14 码表维度

*所属主题	*维度名称	*维度英文名称	*维度类型	*资产责任人	描述	*数据连接类型	*数据连接	*数据库	选择码表
供应商	供应商	dim_vendor	码表维度	-	无	MRS_HIVE	mrs_hive_link	demo_dwr_db	供应商
公共维度	费率代码	dim_rate_code	码表维度	-	无	MRS_HIVE	mrs_hive_link	demo_dwr_db	费率
公共维度	付款方式	dim_payment_type	码表维度	-	无	MRS_HIVE	mrs_hive_link	demo_dwr_db	付款方式

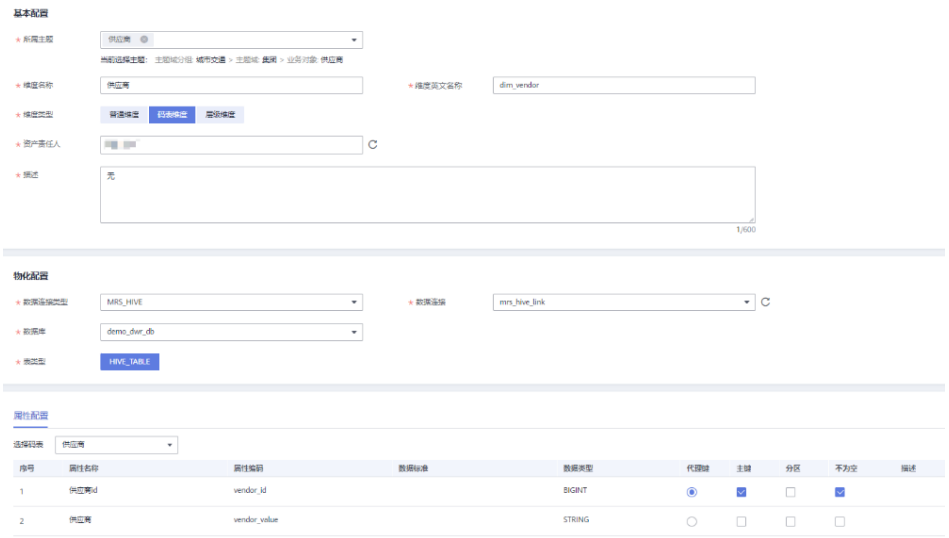
1. 在“维度建模”页面进入“维度”标签页，在主题树中选中“城市交通 > 集团 > 供应商”，然后单击“新建”新建供应商维度。

图 4-61 维度建模



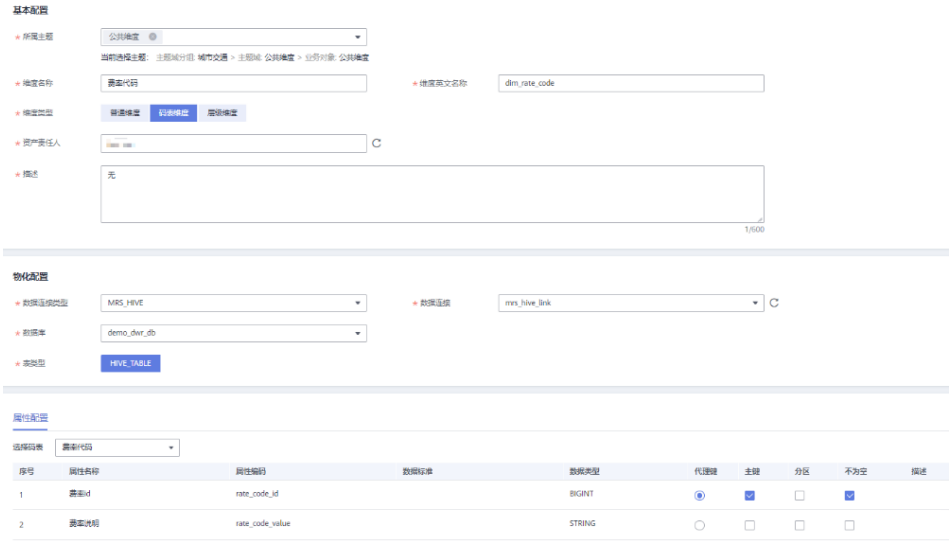
2. 在新建维度页面，如下图所示配置参数，然后单击“保存”完成维度的新建。

图 4-62 新建维度



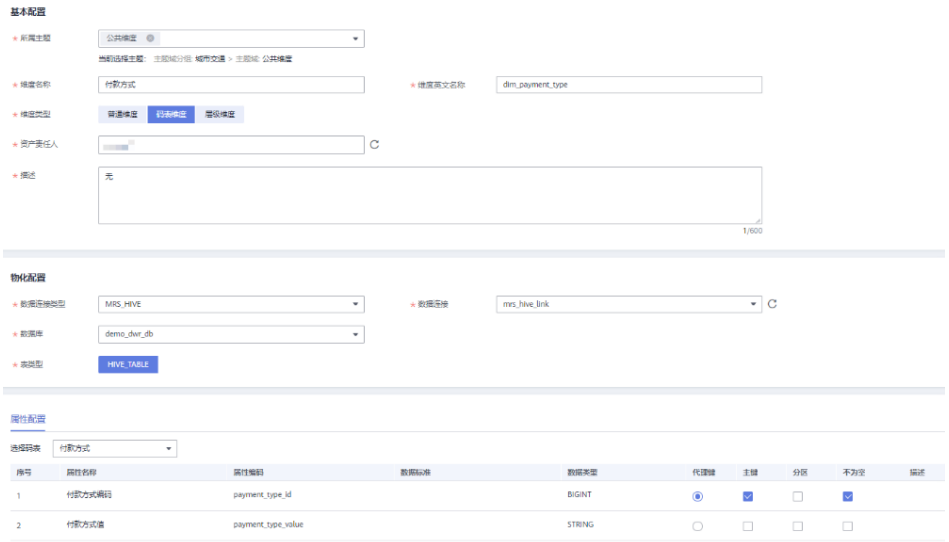
3. 在“维度建模”页面进入“维度”标签页，在主题树中选中“城市交通 > 公共维度 > 公共维度”，然后单击“新建”新建费率代码维度。在新建维度页面，配置如下，配置完成后单击“保存”。

图 4-63 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 4-64 付款方式维度



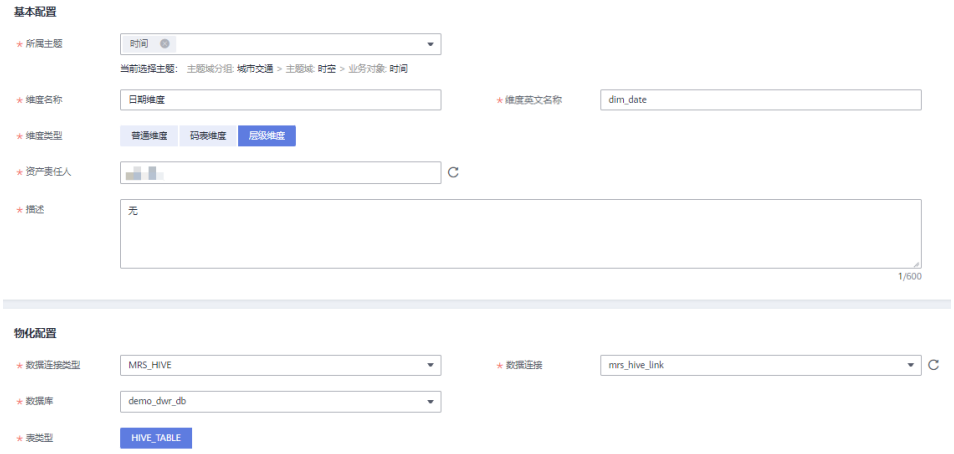
步骤3 新建一个层级维度“日期维度”。

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 4-15 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 4-65 日期维度



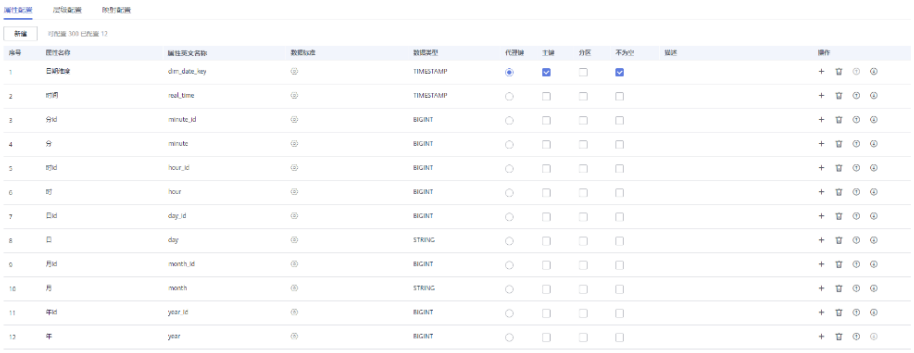
3. 属性配置如下：

表 4-16 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 4-66 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 4-67 层级 1

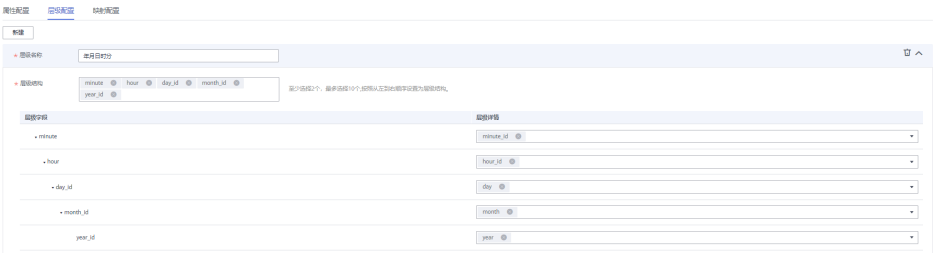


图 4-68 层级 2

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 4-69 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4-17所示。

表 4-17 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

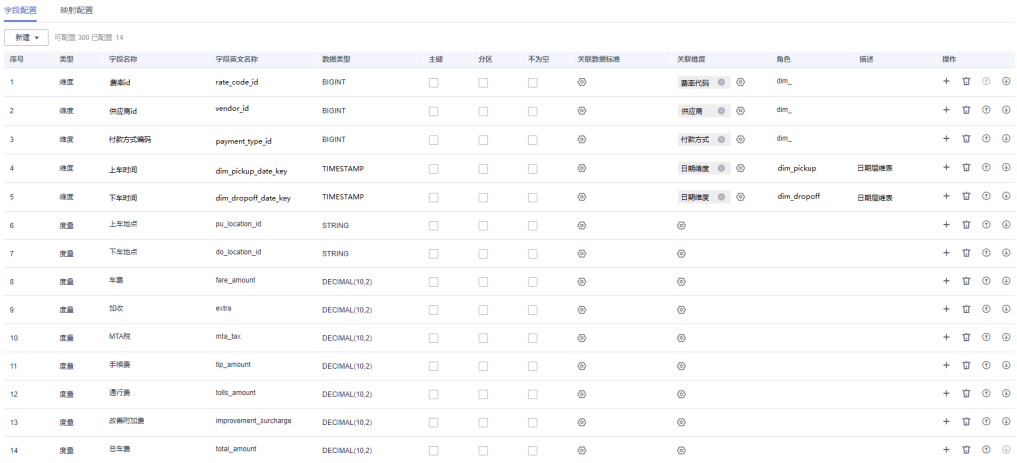
在“字段配置”区域，选择“新建 > 度量”，依次新建如表4-18所示的字段。

表 4-18 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 4-70 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4-19和表4-20所示的技术指标：

表 4-19 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 4-20 衍生指标

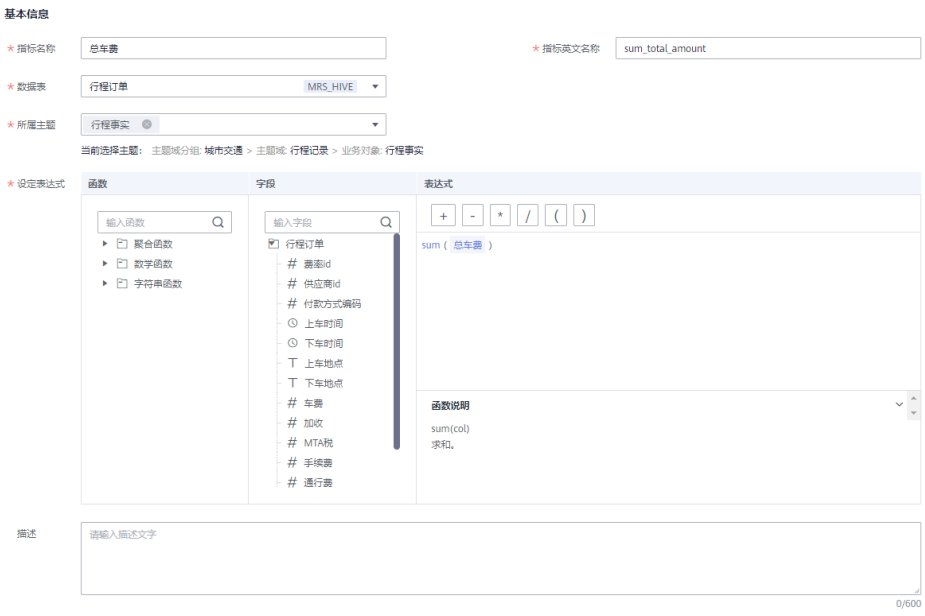
指标	*数据表	*所属主题	*原子指标	统计维度	时间限定	通用限定
基于付款方式维度统计总车费	行程订单	记录统计	总车费	付款方式	无	无
基于费率代码维度统计总车费	行程订单	记录统计	总车费	费率代码	无	无
基于供应商和下车时间维度统计总车费	行程订单	记录统计	总车费	供应商，行程订单.下车时间	无	无

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 4-71 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 4-72 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 4-73 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 4-74 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

维度建模：在 DM 层新建并发布汇总表

在DM数据集市层，您需要新建如表4-21所示的汇总表。

表 4-21 汇总表

*所属主题	*表名称	*表英文名称	统计维度	数据连接类型	*数据连接	*数据库	资产责任人	描述
记录统计	付款方式统计汇总	dws_payment_type	付款方式	MRS_HIVE	mrs_hive_link	demo_dm_db	-	无
记录统计	费率统计汇总	dws_rate_code	费率代码	MRS_HIVE	mrs_hive_link	demo_dm_db	-	无
记录统计	供应商统计汇总	dws_vendor	供应商,行程订单.下车时间	MRS_HIVE	mrs_hive_link	demo_dm_db	-	无

步骤1 在数据架构控制台，单击左侧导航树中的“维度建模”，进入维度建模页面。

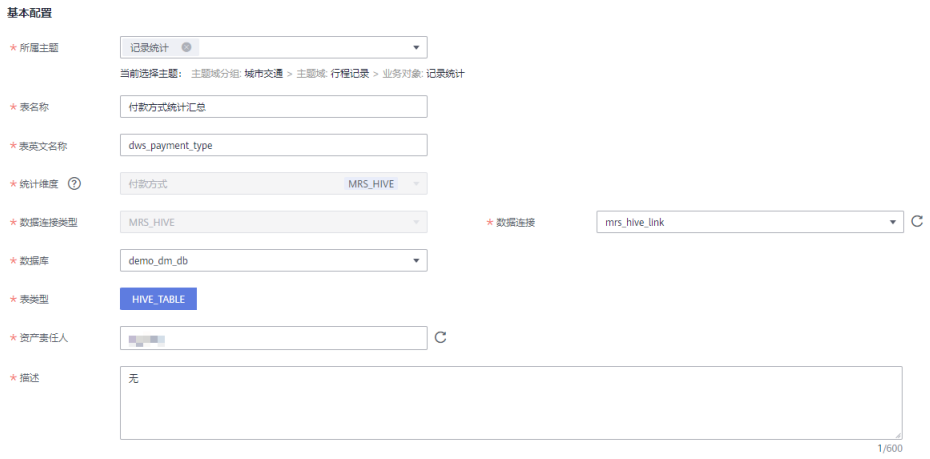
步骤2 单击“汇总表”页签，进入汇总表页面。

步骤3 新建3个汇总表：付款方式统计汇总表、费率统计汇总表、供应商统计汇总表。

1. 在“汇总表”页面，在主题树中选中“城市交通 > 行程记录 > 记录统计”，然后单击“新建”新建付款方式统计汇总表。在新建汇总表页面，配置如下，配置完成后单击“保存”。

在新建汇总表页面，基本配置如下：

图 4-75 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 4-76 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)”。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 4-77 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 4-78 费率统计汇总-基本配置

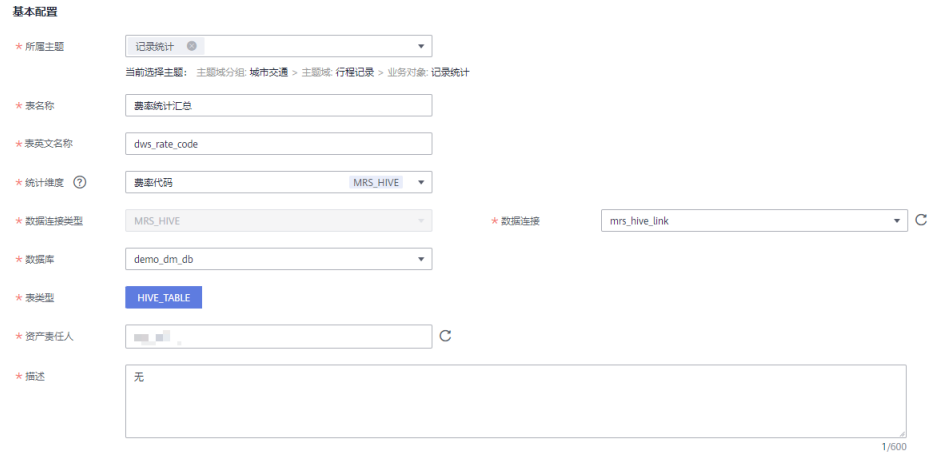
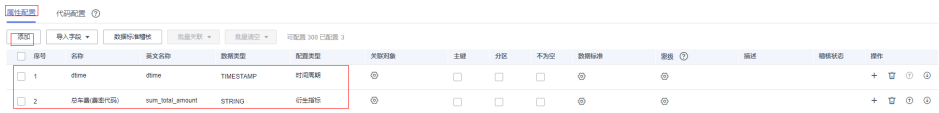


图 4-79 费率统计汇总-属性配置



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 4-80 供应商统计汇总-基本配置

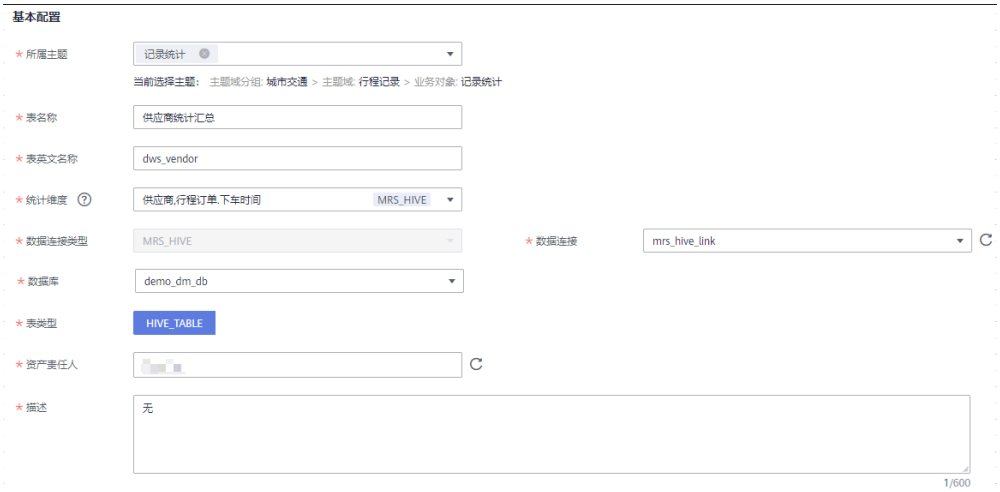


图 4-81 供应商统计汇总-属性配置

步骤4 返回维度建模页面的“汇总表”标签页后，勾选建好的3个汇总表，单击“发布”。

步骤5 在弹出框中选择审核人，单击“确认提交”，等待审核人员审核通过后，汇总表会自动创建。如果当前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“维度建模 > 汇总表”页面，在列表中找到刚发布的汇总表，在“同步状态”一列中可以查看汇总表的同步状态。

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

----结束

审核人员审核对象

步骤1 使用审核人员账号，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”。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2 在左侧导航树中，单击“审核中心”，在“待我审核”页签的列表中选中需要审核的对象，然后单击“批量审核”。

步骤3 输入审核意见后，单击“批量通过”完成审核。

----结束

4.7 步骤 6：数据开发

DataArts Studio数据开发模块可管理多种大数据服务，提供一站式的大数据开发环境、全托管的大数据调度能力，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。

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、版本管理、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。

在DataArts Studio数据开发模块中，您将完成以下步骤：

1. **数据管理**
2. **脚本开发**
3. **作业开发**
   a. 历史数据到源数据表，使用数据集成将历史数据从OBS导入到SDI贴源层的原始数据表。
   b. 历史数据清洗，使用数据开发的MRS Hive SQL脚本将源数据表清洗之后导入DWI层的标准出行数据表。
   c. 将基础数据插入维度表中。
   d. 将DWI层的标准出行数据导入DWR层的事实表中。
   e. 数据汇总，通过Hive SQL将出租车行程订单事实表中的数据进行汇总统计并写入汇总表。

4. **运维调度**

数据管理

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。主要包含建立数据连接、新建数据库、新建数据表等操作。

在本例中，相关数据管理操作已经在**步骤2：准备工作**中完成，本步骤可跳过。

脚本开发

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 在左侧导航栏中，单击“脚本开发”，再右键单击“脚本”选择“新建目录”，在弹出框中输入目录名称例如“transport”，然后单击“确定”。

步骤3 在脚本目录树中，右键单击目录名称transport，选择菜单“新建Hive SQL脚本”。

步骤4 在新建的HIVE_untitled脚本中，选择数据连接mrs_hive_link，选择数据库demo_dwr_db，然后输入脚本内容。

图 4-82 编辑脚本

该脚本用于将付款方式、费率代码、供应商的基本信息写入到相应的维度表中。脚本内容如下：

```
truncate table dim_payment_type;
truncate table dim_rate_code;
truncate table dim_vendor;

INSERT INTO dim_payment_type VALUES ("1","Credit card");
INSERT INTO dim_payment_type VALUES ("2","Cash");
INSERT INTO dim_payment_type VALUES ("3","No charge");
INSERT INTO dim_payment_type VALUES ("4","Dispute");
INSERT INTO dim_payment_type VALUES ("5","Unknown");
INSERT INTO dim_payment_type VALUES ("6","Voided trip");

INSERT INTO dim_rate_code VALUES ("1","Standard rate");
INSERT INTO dim_rate_code VALUES ("2","JFK");
INSERT INTO dim_rate_code VALUES ("3","Newark");
INSERT INTO dim_rate_code VALUES ("4","Nassau or Westchester");
INSERT INTO dim_rate_code VALUES ("5","Negotiated fare");
INSERT INTO dim_rate_code VALUES ("6","Group ride");

INSERT INTO dim_vendor VALUES ("1","A Company");
INSERT INTO dim_vendor VALUES ("2","B Company");
```


步骤5 单击“运行”按钮，测试脚本是否正确。

图 4-83 运行脚本

步骤6 测试通过后，单击“保存”按钮，在弹出框中输入脚本名称如：demo_taxi_dim_data，选择保存的脚本路径并单击“提交”按钮提交版本。

图 4-84 保存脚本

图 4-85 提交脚本版本

步骤7 重复**步骤4**~**步骤6**的步骤，完成如下脚本的创建。

1. 脚本名称：demo_etl_sdi_dwi，该脚本用于将SDI贴源层的原始数据写入到DWI层的标准出行数据表中。脚本内容如下：

```
INSERT INTO
  demo_dwi_db.dwi_taxi_trip_data
SELECT
  `vendorid`,
  cast(
    regexp_replace(
      `tpep_pickup_datetime`,
      '(\\d{2})/(\\d{2})/(\\d{4}) (\\d{2}):(\\d{2}):(\\d{2}) (.*)',
      '$3-$1-$2 $4:$5:$6'
    ) as TIMESTAMP
  ),
  cast(
    regexp_replace(
      `tpep_dropoff_datetime`,
      '(\\d{2})/(\\d{2})/(\\d{4}) (\\d{2}):(\\d{2}):(\\d{2}) (.*)',
      '$3-$1-$2 $4:$5:$6'
    ) as TIMESTAMP
  ),
  `passenger_count`,
  `trip_distance`,
  `ratecodeid`,
  `store_fwd_flag`,
  `pulocationid`,
  `dolocationid`,
  `payment_type`,
  `fare_amount`,
  `extra`,
  `mta_tax`,
  `tip_amount`,
  `tolls_amount`,
  `improvement_surcharge`,
  `total_amount`
FROM
  demo_sdi_db.sdi_taxi_trip_data
WHERE
  trip_distance > 0
  and total_amount > 0
  and payment_type in (1, 2, 3, 4, 5, 6)
  and vendorid in (1, 2)
  and ratecodeid in (1, 2, 3, 4, 5, 6)
  and tpep_pickup_datetime < tpep_dropoff_datetime
  and tip_amount >= 0
  and fare_amount >= 0
  and extra >= 0
  and mta_tax >= 0
  and tolls_amount >= 0
  and improvement_surcharge >= 0
  and total_amount >= 0
  and (fare_amount + extra + mta_tax + tip_amount + tolls_amount + improvement_surcharge) =
total_amount;
```

2. 脚本名称：demo_etl_dwi_dwr_fact，该脚本用于将DWI层的标准出行数据写入到DWR层的事实表中。脚本内容如下：

```
INSERT INTO
  demo_dwr_db.fact_stroke_order
SELECT
  rate_code_id,
  vendor_id,
  payment_type,
  tpep_dropoff_datetime,
  tpep_pickup_datetime,
  pu_location_id,
  do_location_id,
  fare_amount,
  extra,
  mta_tax,
  tip_amount,
```

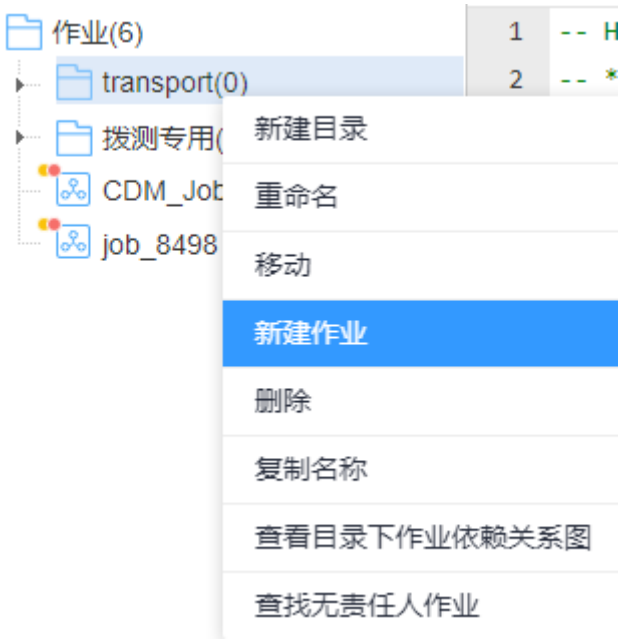
```
  tolls_amount,
  improvement_surcharge,
  total_amount
FROM
  demo_dwi_db.dwi_taxi_trip_data;
```

----结束

作业开发

1. 在DataArts Studio数据开发控制台的左侧导航栏中，单击“作业开发”，然后右键单击“作业”选择菜单“新建目录”，在目录树下根据需要创建作业目录，例如“transport”。
2. 右键单击作业目录，在弹出菜单中单击“新建作业”。

图 4-86 作业

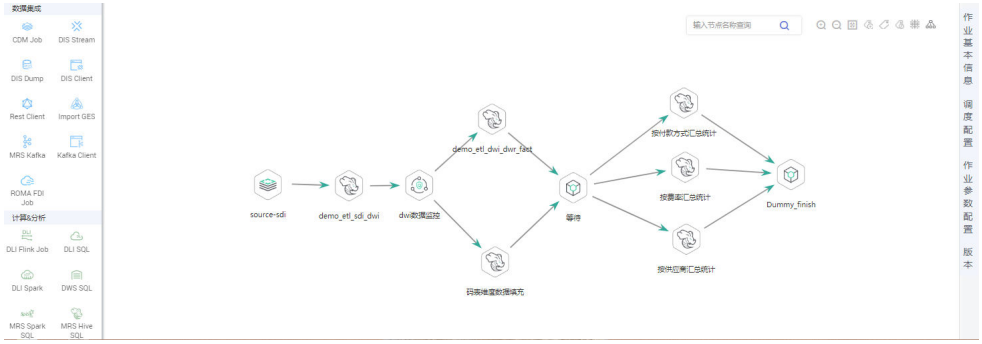


3. 在弹出弹框中输入“作业名称”如demo_taxi_trip_data，“作业类型”选择“批处理”，其他参数保留默认值，单击“确定”完成批作业创建。

图 4-87 新建批处理作业

4. 如下图所示，编排批作业。

图 4-88 编排作业



每个节点配置如下：

- **source_sdi节点**：为CDM Job节点，通过CDM节点将OBS上的数据导入到MRS Hive的原始表中。其中CDM集群名称和作业名称分别选择在**步骤3：数**

据集成中的集群和迁移作业（图中仅为示例，以实际集群名和迁移作业名为准）。

图 4-89 source_sdi 节点属性

- **demo_etl_sdi_dwi节点**：为MRS Hive SQL节点，用于清洗过滤SDI贴源层上原始表中的数据，将合法数据写入数据架构中DWI层标准出行数据表dwi_taxi_trip_data中。其中，“SQL脚本”请选择在**脚本开发**中创建的脚本demo_etl_sdi_dwi。

图 4-90 demo_etl_sdi_dwi 节点属性

- **dwi数据监控节点**：为Data Quality Monitor节点，用于监控DWI层的标准出行数据的质量。其中，“数据质量规则名称”请选择发布DWI层标准出行数据表时自动生成的质量规则“标准出行数据”。

  图 4-91 dwi 数据监控节点

- **demo_etl_dwi_dwr_fact节点**：为MRS Hive SQL节点，用于将DWI上的原始数据写入DWR层的事实表fact_stroke_order中。其中，“SQL脚本”请选择在**脚本开发**中创建的脚本demo_etl_dwi_dwr_fact。

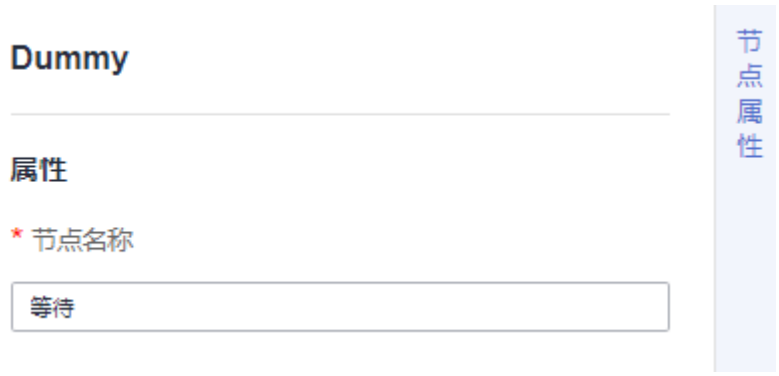
图 4-92 demo_etl_dwi_dwr_fact 节点属性

- **码表维度数据填充节点**：为MRS Hive SQL节点，用于将付款方式、费率代码和供应商的集成数据写入DWR层相应的维度表中。其中，“SQL脚本”请选择在**脚本开发**中创建的脚本demo_taxi_dim_data。

图 4-93 码表维度数据填充节点属性

- **等待节点**：不做任何事情，等待前面的节点运行结束。

  图 4-94 等待节点

- **按付款方式汇总统计节点**：为MRS Hive SQL节点，按付款方式维度统计汇总截止到当前日期的收入。该节点是从发布汇总表“付款方式统计汇总”时自动生成的数据开发作业（作业名称以demo_dm_db_dws_payment_type_开头，命名规则为“数据库名称_汇总表编码”）中复制的，复制节点后需手动配置该节点的“数据连接”和“数据库”参数，“数据库”需设置为事实表所在的数据库。

  说明

  数据开发作业自动生成功能需在**管理配置中心**中勾选“创建数据开发作业”实现。

图 4-95 按付款方式汇总统计节点属性

- **按费率汇总统计节点**：为MRS Hive SQL节点，按费率代码维度统计汇总截止到当前日期的收入。该节点是从发布汇总表“费率统计汇总”时自动生成的数据开发作业（作业名称以demo_dm_db_dws_rate_code_开头，命名规则为“数据库名称_汇总表编码”）中复制的，复制节点后需手动配置该节点的“数据连接”和“数据库”参数，“数据库”需设置为事实表所在的数据库。

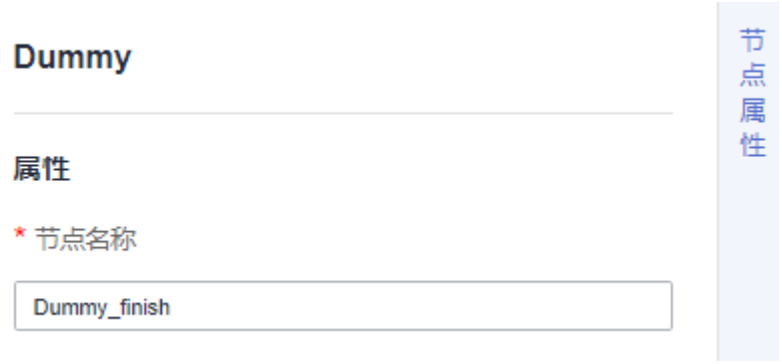
图 4-96 按费率汇总统计节点属性

- **按供应商汇总统计节点**：为MRS Hive SQL节点，按供应商维度统计汇总截止到当前日期各时间维度的收入。该节点是从发布汇总表“供应商统计汇总”时自动生成的数据开发作业（作业名称以demo_dm_db_dws_vendor_开头，命名规则为“数据库名称_汇总表编码”）中复制的，复制节点后需手动配置该节点的“数据连接”和“数据库”参数，“数据库”需设置为事实表所在的数据库。

图 4-97 按供应商汇总统计节点属性

- **Dummy_finish节点**：不做任何事情，作为作业结束的标记。

  图 4-98 Dummy_finish 节点



5. 作业编排好之后，您可以通过测试运行来测试作业编排是否正确。
6. 您可以根据需要，配置作业的调度方式。单击右侧“调度配置”页签，展开配置页面。当前支持单次调度、周期调度和事件驱动调度作业。

图 4-99 配置作业的调度方式

7. 调度配置完成后，您需要单击“保存”按钮保存作业并单击“提交”按钮提交作业版本。然后单击“执行调度”来启动作业的调度。

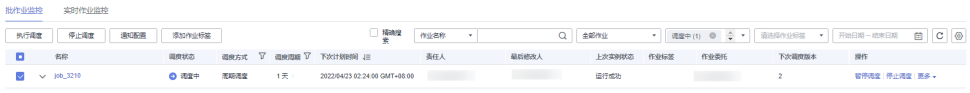
图 4-100 保存并提交作业与执行调度

运维调度

您可以通过运维调度功能，查看作业以及作业实例的运行状态。

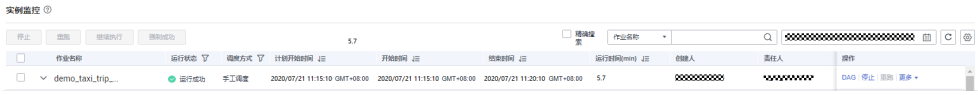
1. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。
2. 单击“批作业监控”页签，进入批作业监控界面。
3. 批作业监控提供了对批处理作业的状态进行监控的能力。您可以查看批作业的调度状态、调度频率、调度开始时间等信息，勾选作业名称前的复选框，并进行“执行调度”/“停止调度”/“通知配置”，相应操作。

图 4-101 批量处理作业



4. 单击左侧导航栏，选择“运维调度 > 实例监控”，进入实例监控界面。
   在运维调度的“实例监控”页面，可以查看作业实例的运行详情以及运行日志等。

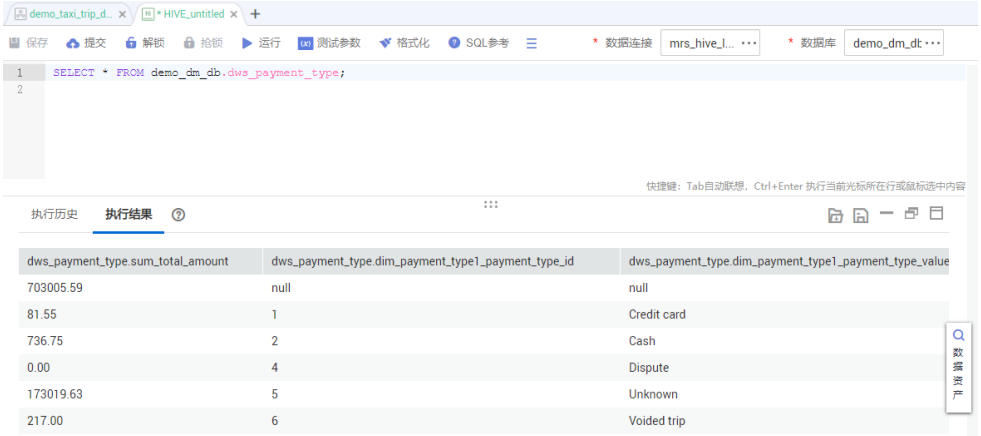
图 4-102 实例监控



5. 作业运行成功后，您可以在DataArts Studio数据目录中查看汇总表的数据预览，具体操作请参见**步骤8：数据目录管理**。您也可以在数据开发的“脚本开发”页面新建一个Hive SQL脚本，执行以下命令查询结果，执行成功后返回类似如下的结果：

```
SELECT * FROM demo_dm_db.dws_payment_type;
```

图 4-103 查询结果



4.8 步骤 7：数据质量监控

数据质量监控DQC（Data Quality Control）模块是对数据库里的数据质量进行质量管理的工具。您可从完整性、有效性、及时性、一致性、准确性、唯一性六个维度进行单列、跨列、跨行和跨表的分析。

在DataArts Studio数据质量模块中，可以对业务指标和数据质量进行监控。

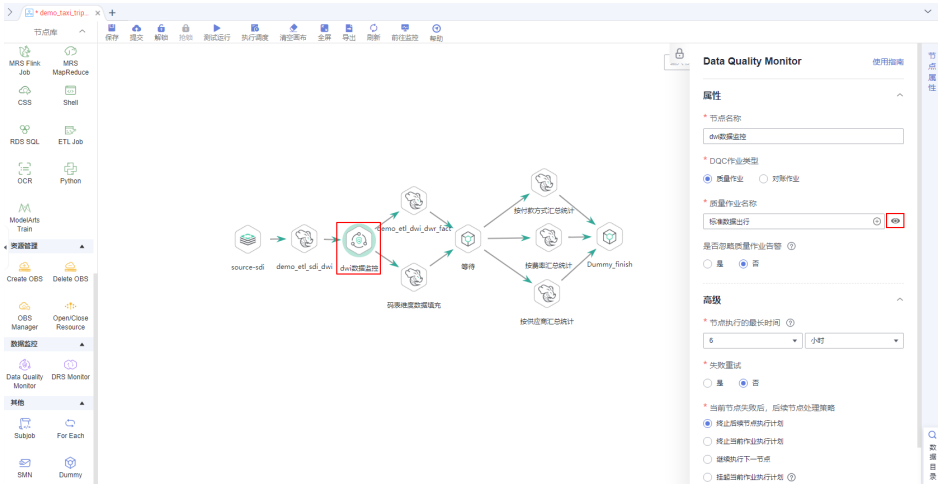
查看质量作业

在DataArts Studio数据开发中，作业运行成功后，您可以登录DataArts Studio数据质量控制台查看质量作业运行结果。

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DataArts Studio作业开发控制台，选择**步骤6：数据开发**中新建的作业，单击数据质量监控节点，然后在该节点的节点属性中，单击“数据质量规则名称”后的 按钮，可以跳转到DataArts Studio数据质量控制台的“质量作业”页面。

图 4-104 质量作业节点

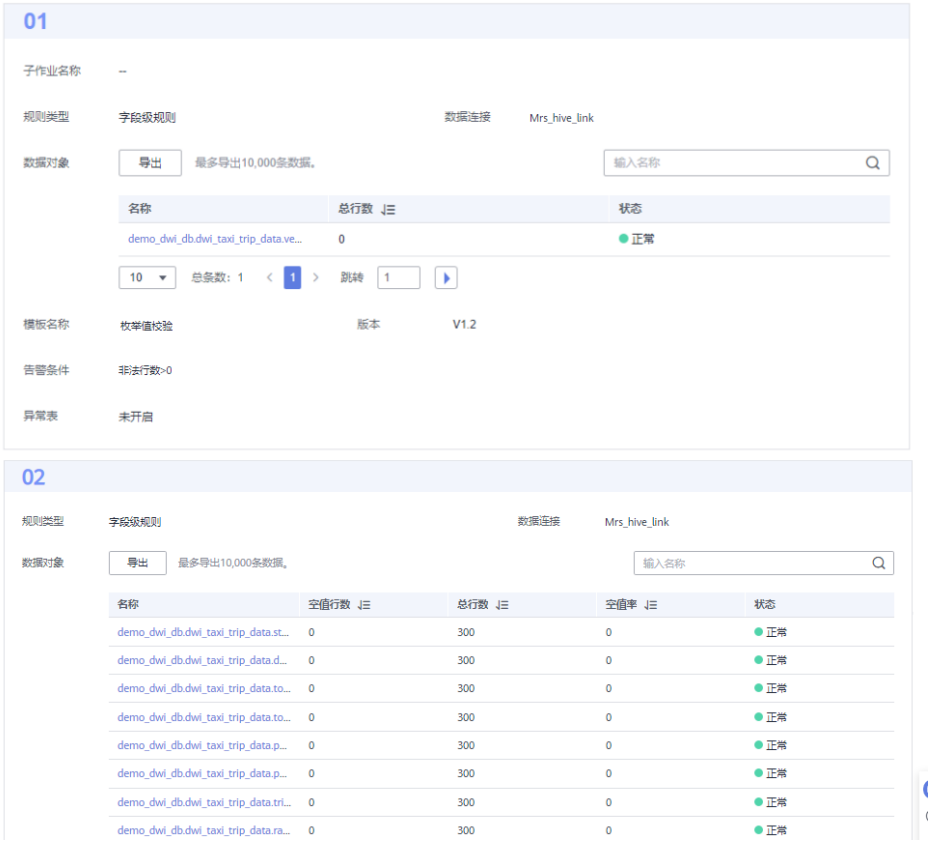


步骤3 在数据质量页面，单击质量作业名称，可以查看质量作业的基础配置。

图 4-105 质量作业列表

步骤4 单击左侧导航栏中的“运维管理”，单击操作列的“结果&日志”按钮，可查看质量作业的运行结果。

图 4-106 质量作业运行结果



----结束

监控业务指标

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，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，可以判断业务指标是否满足质量规则。本例通过监控出租车一天的运营收入，对于当天收入低于500进行预警。具体请参考如下步骤：

步骤1 在DataArts Studio控制台首页，选择实例，单击“进入控制台”，选择对应工作空间的“数据质量”模块，进入数据质量页面。

图 4-107 选择数据质量

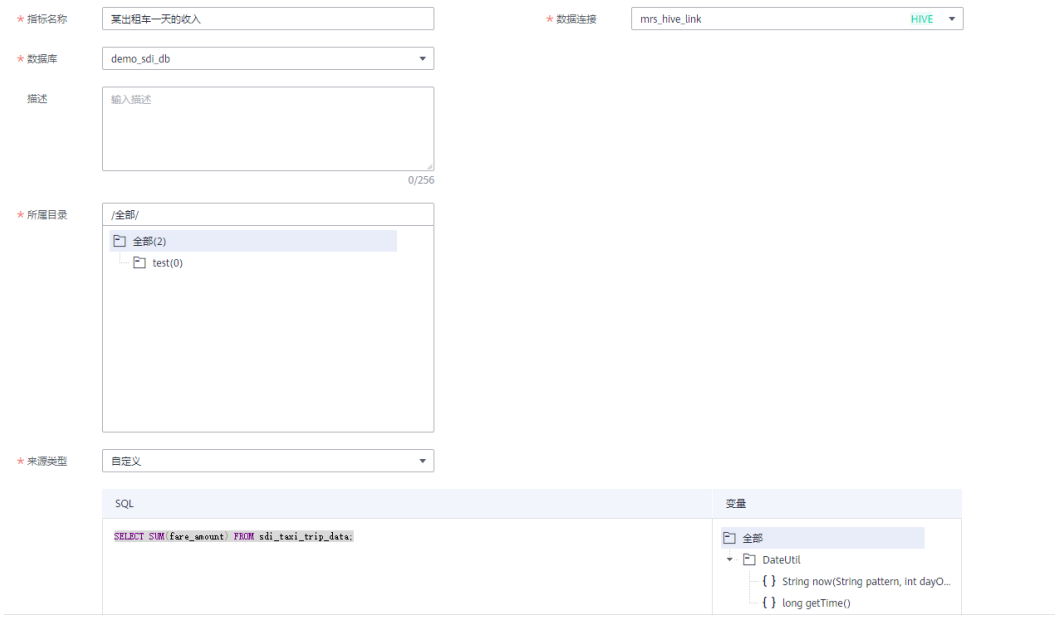
步骤2 选择“业务指标监控 > 指标管理”。

步骤3 单击“新建”，在弹出的对话框中，配置相关参数，新建指标。

SQL语句如下：

```
SELECT SUM(fare_amount) FROM sdi_taxi_trip_data;
```

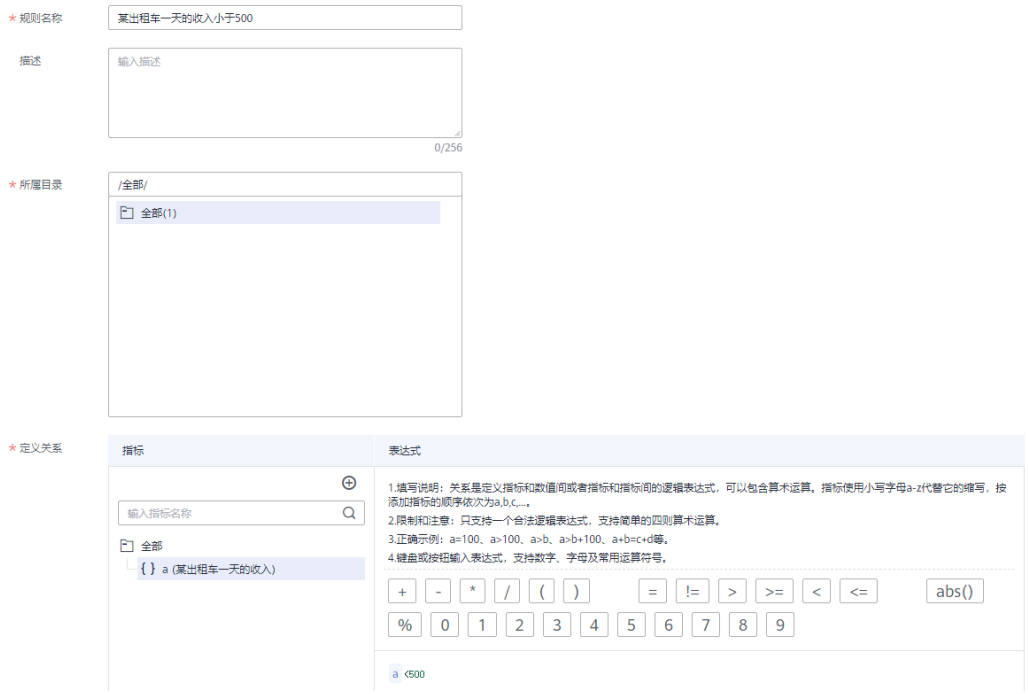
图 4-108 新建指标



步骤4 选择“业务指标监控 > 规则管理”。

步骤5 单击“新建”，在弹出的对话框中，配置相关参数，新建规则。

图 4-109 新建规则

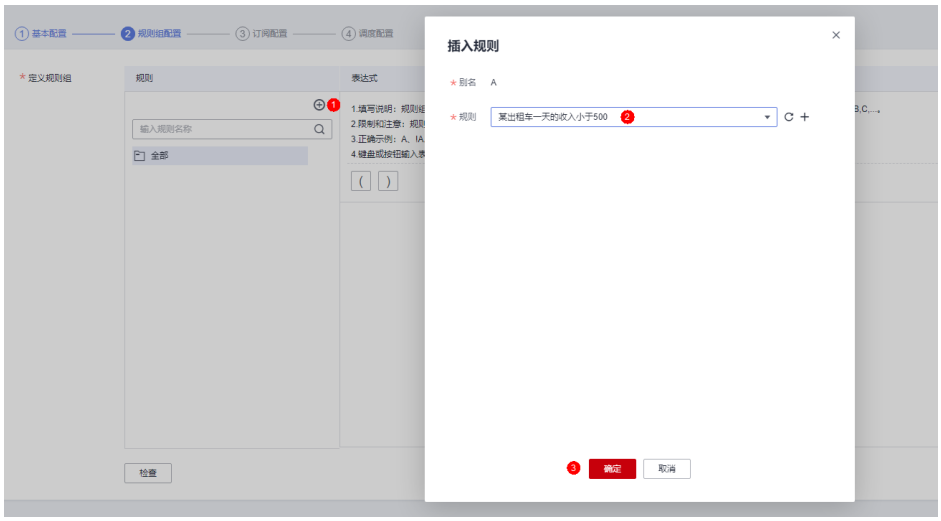


步骤6 选择“业务指标监控 > 业务场景管理”。

步骤7 单击“新建”，在弹出的对话框中，配置相关参数，新建场景。

图 4-110 基本配置

图 4-111 规则组配置

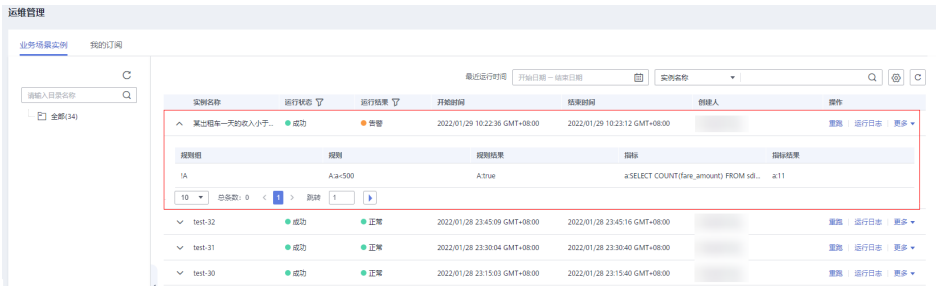


单击“下一步”，选择调度方式，支持单次调度和周期调度两种方式。

步骤8 在业务场景管理列表中，单击操作列的“运行”，再跳转到运维管理模块。

步骤9 单击“运行结果”，查看具体的指标监控情况。

图 4-112 运行结果



说明

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。

----结束

4.9 步骤 8：数据目录管理

在DataArts Studio数据目录模块中，您可以查看数据地图，详情请参见**数据目录**章节。数据地图包含业务资产和技术资产，业务资产就是指逻辑实体和业务对象，技术资产就是指数据连接、数据库对象等。

本章节介绍如何在DataArts Studio数据目录中查看业务资产和技术资产。例如，在技术资产的事实表中，您可以查看数据血缘等详细信息，在技术资产的汇总表中，您可以查看预览结果等详细信息。

查看业务资产和技术资产

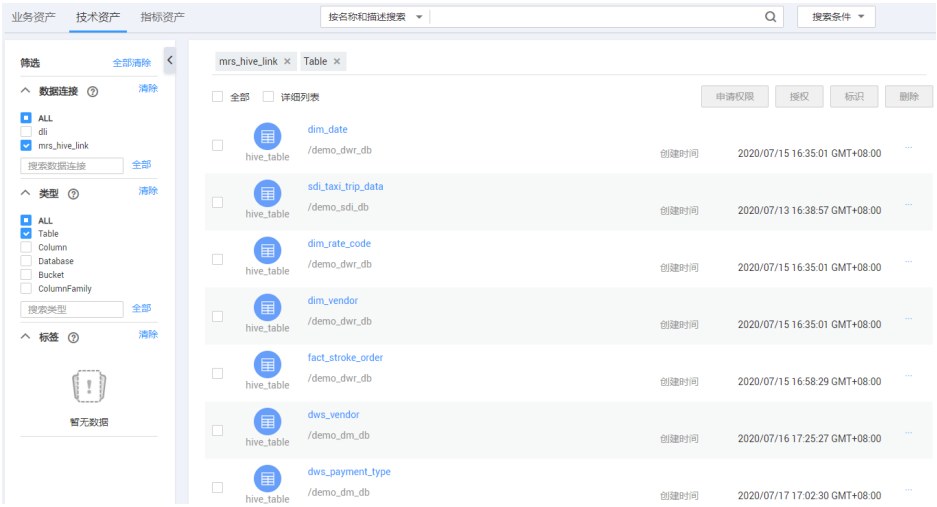
步骤1 在DataArts Studio控制台首页，选择实例，单击“进入控制台”，选择对应工作空间的“数据目录”模块，进入数据目录页面。

图 4-113 选择数据目录

步骤2 在左侧导航栏单击“数据目录”，选择“业务资产”页签，然后在筛选条件中选择业务对象，将显示符合条件的业务资产。

步骤3 选择“技术资产”页签，然后在筛选条件中“数据连接”选择所需查看的连接，“类型”选择“Table”，右侧页面将显示符合条件的所有的元数据。

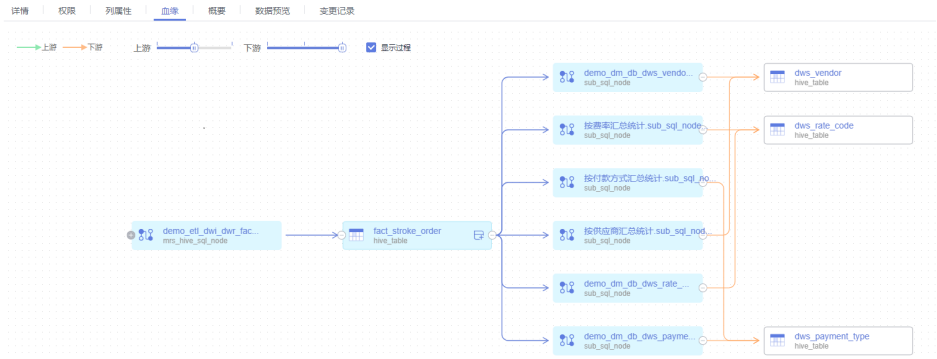
图 4-114 技术资产



步骤4 在资产列表中，单击所需查看的元数据名称，即可查看详情信息。

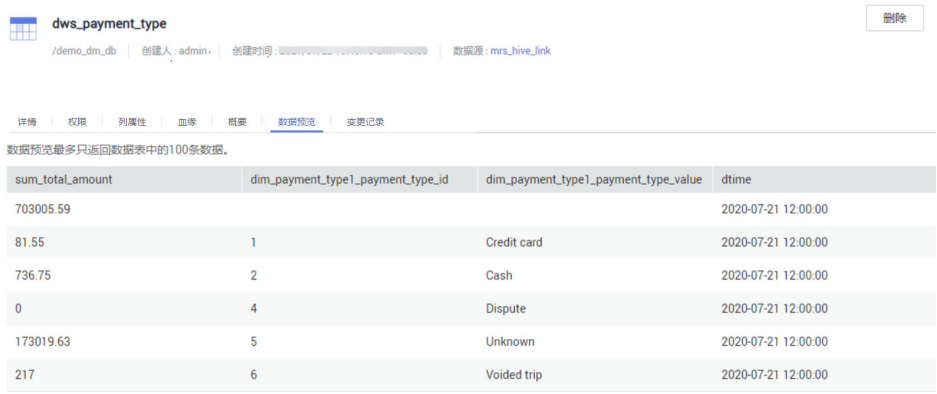
例如，在资产列表中，找到事实表fact_stroke_order，单击事实表名称，即可查看事实表的详情信息。在详情页面，进入“血缘”页签，可查看事实表的输入输出血缘信息。

图 4-115 血缘



在资产列表中，找到汇总表，例如dws_payment_type，单击汇总表名称，即可查看汇总表的详情信息。在详情页面，进入“数据预览”页签，可查看汇总表的预览数据。

图 4-116 数据预览



sum_total_amount	dim_payment_type1_payment_type_id	dim_payment_type1_payment_type_value	dtime
703005.59			2020-07-21 12:00:00
81.55	1	Credit card	2020-07-21 12:00:00
736.75	2	Cash	2020-07-21 12:00:00
0	4	Dispute	2020-07-21 12:00:00
173019.63	5	Unknown	2020-07-21 12:00:00
217	6	Voided trip	2020-07-21 12:00:00

----结束

4.10 步骤 9：服务退订（可选）

本开发场景中，DataArts Studio、OBS、MRS和DWS服务均会产生相关费用。在使用过程中，如果您额外进行了通知配置，可能还会产生以下相关服务的费用：

- SMN服务：如果您在使用DataArts Studio各组件过程中开启了消息通知功能，则会产生消息通知服务费用，收费标准请参见**SMN价格详情**。
- EIP服务：如果您为数据集成集群或数据服务专享版集群开通了公网IP，则会产生弹性公网IP服务费用，收费标准请参见**EIP价格详情**。
- DEW服务：在数据集成或创建管理中心连接时，如果启用了KMS，则会产生密钥管理费用，收费标准请参见**DEW价格详情**。
- APIG服务：在使用数据服务共享版发布API到API网关共享版后，如果调用API，则会产生API网关的调用API费用和流量费用，收费标准请参见**APIG价格详情**。

在场景开发完成后，如果您不再使用DataArts Studio及相关服务，请及时进行退订和资源删除，避免持续产生费用。

表 4-22 相关服务退订方式

服务	计费说明	退订方式
DataArts Studio	**DataArts Studio计费说明**	DataArts Studio实例仅支持包周期计费。您可以根据需要参考**云服务退订**退订DataArts Studio包年包月套餐。
OBS	**OBS计费说明**	OBS服务支持按需和包周期计费，套餐包暂不支持退订。本例中使用按需计费，完成后删除新建的存储桶即可；另外，DataArts Studio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，在退订DataArts Studio后可以一并删除。
MRS	**MRS计费说明**	MRS服务支持按需和包周期计费。本例中使用按需计费，完成后删除MRS集群即可。如果使用包周期计费，您需要参考**云服务退订**退订包年包月套餐，并删除MRS集群。
DWS	**DWS计费说明**	DWS服务支持按需和包周期计费。本例中使用按需计费，完成后删除DWS集群即可。如果使用包周期计费，您需要参考**云服务退订**退订包年包月套餐，并删除DWS集群。
SMN	**SMN计费说明**	SMN服务按实际用量付费，退订DataArts Studio服务后不会再产生通知，您也可以直接删除SMN服务已产生的主题和订阅。
EIP	**EIP计费说明**	EIP服务支持按需和包周期计费，本例中使用按需计费，完成后删除EIP即可。如果使用包周期计费，您需要参考**云服务退订**退订包年包月套餐，并删除EIP。
DEW	**DEW计费说明**	KMS密钥管理按密钥实例进行按需计费，您可以直接删除DEW服务已产生的KMS密钥。
APIG	**APIG计费说明**	如果您使用的是数据服务专享版，则不涉及此项费用。 使用数据服务专享版时，共享版API网关按实际使用量计费，包含API调用量（次数）和流量费用（下行流量）两个维度。退订DataArts Studio服务后不会再产生API调用，您也可以直接删除发布到APIG网关上的API。

5 入门实践

当您参考**准备工作**章节完成注册华为账号、购买DataArts Studio实例（DataArts Studio企业版）、创建工作空间等一系列操作后，可以根据自身的业务需求使用DataArts Studio提供的一系列常用实践。

表 5-1 常用最佳实践

实践		描述
数据迁移	**数据迁移进阶实践**	本最佳实践提供了数据集成CDM组件的高阶使用技巧，例如如何实现增量迁移、时间宏变量表达式写法等。
数据开发	**数据开发进阶实践**	本最佳实践提供了数据开发DLF组件的高阶使用技巧，例如如何使用IF条件判断、For Each节点使用等。
DataArts Studio+X	**跨工作空间的DataArts Studio数据搬迁**	实例内的工作空间包含了完整的功能，工作空间的划分通常按照分子公司（集团、子公司、部门等）、业务领域（采购、生产、销售等）或者实施环境（开发、测试、生产等），没有特定的划分要求。 随着业务的不断发展，您可能进行了更细致的工作空间划分。这种情况下，您可以参考本文档，将原有工作空间的数据（包含管理中心数据连接、数据集成连接和作业、数据架构表、数据开发脚本、数据开发作业、数据质量作业等），搬迁到新建的工作空间中。

实践		描述
	如何授权其他用户使用DataArts Studio	某数据运营工程师负责本公司的数据质量监控，仅需要数据质量组件的权限。管理员如果直接赋予该数据运营工程师“开发者”的预置角色，则会出现其他组件权限过大的风险。 为了解决此问题，项目管理员可以创建一个基于“开发者”预置角色的自定义角色“Developer_Test”，在“开发者”角色权限的基础上为其去除其他组件的增删改及操作权限，并赋予该数据运营工程师此角色，既能满足实际业务使用，也避免了权限过大的风险。
	如何查看表行数和库大小	在数据治理流程中，我们常常需要统计数据表行数或数据库的大小。其中，数据表的行数可以通过SQL命令或数据质量作业获取；数据库大小可以直接在数据目录组件中查看，详情可参考本实践。
	通过数据质量对比数据迁移前后结果	数据对账对数据迁移流程中的数据一致性至关重要，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。本文以DWS数据迁移到MRS Hive分区表为例，介绍如何通过DataArts Studio中的数据质量模块实现数据迁移前后的一致性校验。
	通过数据开发使用参数传递灵活调度CDM作业	如果CDM作业接收来自数据开发作业配置的参数，则在数据开发模块可以使用诸如EL表达式传递动态参数来调度CDM作业。
	通过数据开发实现数据增量迁移	DataArts Studio服务的DLF组件提供了一站式的大数据协同开发平台，借助DLF的在线脚本编辑、周期调度CDM的迁移作业，也可以实现增量数据迁移。本文以DWS导入到OBS为例，介绍DLF配合CDM实现增量迁移的流程
	通过CDM节点批量创建分表迁移作业	业务系统中，数据源往往会采用分表的形式，以减少单表大小，支持复杂的业务应用场景。在这种情况下，通过CDM进行数据集成时，需要针对每张表创建一个数据迁移作业。您可以参考本教程，通过数据开发模块的For Each节点和CDM节点，配合作业参数，实现批量创建分表迁移作业。

实践		描述
	基于MRS Hive表构建图数据并自动导入GES	在DataArts Studio中，您可以将原始数据表按照GES数据导入要求处理为标准点数据集和边数据集，并通过自动生成元数据功能，将图数据（点数据集、边数据集和元数据）定期导入到GES服务中，在GES中对最新数据进行可视化图形分析。
案例	案例：贸易数据统计与分析	使用云数据迁移（Cloud Data Migration，简称CDM）将本地贸易统计数据导入到OBS，再使用数据湖探索（Data Lake Insight，简称DLI）进行贸易统计分析，帮助H咨询公司以极简、极低成本构建其大数据分析平台，使得该公司更好地聚焦业务，持续创新。
	案例：车联网大数据业务上云	为搭建H公司车联网业务集团级的云管理平台，统一管理、部署硬件资源和通用类软件资源，实现IT应用全面服务化、云化，CDM（Cloud Data Migration，简称CDM）助力H公司做到代码“0”改动、数据“0”丢失迁移上云。
	案例：搭建实时报警平台	在本实践用户可以了解到如何搭建一个简单的实时报警平台，该平台将应用多个云服务，结合数据开发模块的作业编辑和作业调度功能来实现。