数据治理中心

最佳实践

文档版本 01

发布日期 2024-04-29

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 数据迁移进阶实践

1.1 增量迁移原理介绍

1.1.1 文件增量迁移

CDM支持对文件类数据源进行增量迁移，全量迁移完成之后，第二次运行作业时可以导出全部新增的文件，或者只导出特定的目录/文件。

目前CDM支持以下文件增量迁移方式：

1. **增量导出指定目录的文件**
   - 适用场景：源端数据源为文件类型（OBS/HDFS/FTP/SFTP）。这种增量迁移方式，只追加写入文件，不会更新或删除已存在的记录。
   - 关键配置：**文件/路径过滤器**+定时执行作业。
   - 前提条件：源端目录或文件名带有时间字段。
2. **增量导出指定时间以后的文件**
   - 适用场景：源端数据源为文件类型（OBS/HDFS/FTP/SFTP）。这里的指定时间，是指文件的修改时间，当文件的修改时间大于等于指定的起始时间，CDM才迁移该文件。
   - 关键配置：**时间过滤**+定时执行作业。
   - 前提条件：无。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为（数据开发作业计划启动时间-偏移量），而不是（CDM作业实际启动时间-偏移量）。

文件/路径过滤器

- 参数位置：在创建表/文件迁移作业时，如果源端数据源为文件类型，那么源端作业参数的高级属性中可以看到“过滤类型”参数，该参数可选择：通配符或正则表达式。
- 参数原理：“过滤类型”选择“通配符”时，CDM就可以通过用户配置的通配符过滤文件或路径，CDM只迁移满足指定条件的文件或路径。

- 配置样例：

  例如源端文件名带有时间字段“2017-10-15 20:25:26”，这个时刻生成的文件为“/opt/data/file_20171015202526.data”，则在创建作业时，参数配置如下：

  a. 过滤类型：选择“通配符”。

  b. 文件过滤器：配置为**“*${dateformat(yyyyMMdd,-1,DAY)}*”**（这是CDM支持的日期宏变量格式，详见**时间宏变量使用解析**）。

  图 1-1 文件过滤

  c. 配置作业定时自动执行，“重复周期”为1天。

这样每天就可以把昨天生成的文件都导入到目的端目录，实现增量同步。

文件增量迁移场景下，“路径过滤器”的使用方法同“文件过滤器”一样，需要路径名称里带有时间字段，这样可以定期增量同步指定目录下的所有文件。

时间过滤

- 参数位置：在创建表/文件迁移作业时，如果源端数据源为文件类型，那么源端作业配置下的高级属性中，“时间过滤”参数选择“是”。
- 参数原理：“起始时间”和“终止时间”参数中输入时间值后，只有修改时间介于起始时间和终止时间之间（时间区间为左闭右开，即等于起始时间也在区间之内）的文件才会被CDM迁移。
- 配置样例：

  例如需要CDM只同步2021年1月1日~2022年1月1日生成的文件到目的端，则参数配置如下：

  a. 时间过滤器：选择为“是”。

  b. 起始时间：配置为**2021-01-01 00:00:00**（格式要求为yyyy-MM-dd HH:mm:ss）。

  c. 终止时间：配置为**2022-01-01 00:00:00**（格式要求为yyyy-MM-dd HH:mm:ss）

  图 1-2 时间过滤

这样CDM作业就只迁移2021年1月1日~2022年1月1日时间段内生成的文件，下次作业再启动时就可以实现增量同步。

1.1.2 关系数据库增量迁移

CDM支持对关系型数据库进行增量迁移，全量迁移完成之后，可以增量迁移指定时间段内的数据（例如每天晚上0点导出前一天新增的数据）。

- **增量迁移指定时间段内的数据**
  - 适用场景：源端为关系型数据库，目的端没有要求。
  - 关键配置：**Where子句**+定时执行作业。
  - 前提条件：数据表中有时间日期字段或时间戳字段。

关系数据库增量迁移方式，只对数据表追加写入，不会更新或删除已存在的记录。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为（数据开发作业计划启动时间-偏移量），而不是（CDM作业实际启动时间-偏移量）。

Where 子句

- 参数位置：在创建表/文件迁移作业时，如果源端为关系型数据库，那么在源端作业参数的高级属性下面可以看到“Where子句”参数。
- 参数原理：通过“Where子句”参数可以配置一个SQL语句（例如：age > 18 and age <= 60），CDM只导出该SQL语句指定的数据；不配置时导出整表。

  Where子句支持配置为**时间宏变量**，当数据表中有时间日期字段或时间戳字段时，配合定时执行作业，能够实现抽取指定日期的数据。
- 配置样例：

  假设数据库表中存在表示时间的列DS，类型为“varchar(30)”，插入的时间格式类似于“2017-xx-xx”，如**图1-3**所示，参数配置如下：

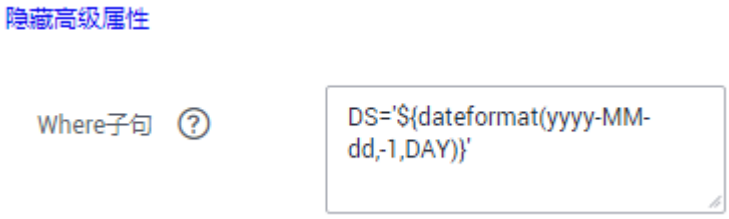
图 1-3 表数据

SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	snap	2017-05-01
2	5	snap	2017-05-01
3	1	google	2017-05-02
4	4	oracle	2017-05-02
5	6	amd	2017-05-02
6	7	nvda	2017-05-02
7	1	google	2017-05-02
8	4	oracle	2017-05-02
9	6	amd	2017-05-02
10	7	nvda	2017-05-02
11	2	facebook	2017-10-15
12	3	tesla	2017-10-15
13	2	facebook	2017-10-15
14	3	tesla	2017-10-15

a. Where子句：配置为**DS='${dateformat(yyyy-MM-dd,-1,DAY)}'**。

图 1-4 Where 子句



b. 配置定时任务：重复周期为1天，每天的凌晨0点自动执行作业。

这样就可以每天0点导出前一天产生的所有数据。Where子句支持配置多种**时间宏变量**，结合CDM定时任务的重复周期：分钟、小时、天、周、月，可以实现自动导出任意指定日期内的数据。

1.1.3 HBase/CloudTable 增量迁移

使用CDM导出HBase（包括MRS HBase、FusionInsight HBase、Apache HBase）或者表格存储服务（CloudTable）的数据时，支持导出指定时间段内的数据，配合CDM的定时任务，可以实现HBase/CloudTable的增量迁移。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为（数据开发作业计划启动时间-偏移量），而不是（CDM作业实际启动时间-偏移量）。

在创建CDM表/文件迁移的作业，源连接选择为HBase连接或CloudTable连接时，高级属性的可选参数中可以配置时间区间。

图 1-5 HBase 时间区间

- 起始时间（包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间及以后的数据。
- 终止时间（不包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间以前的数据。

这2个参数支持配置为**时间宏变量**，例如：

- 起始时间配置为**${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}**时，表示只导出昨天以后的数据。
- 终止时间配置为**${dateformat(yyyy-MM-dd HH:mm:ss)}**时，表示只导出当前时间以前的数据。

这2个参数同时配置后，CDM就只导出前一天内的数据，再将该作业配置为每天0点执行一次，就可以增量同步每天新生成的数据。

1.1.4 MongoDB/DDS 增量迁移

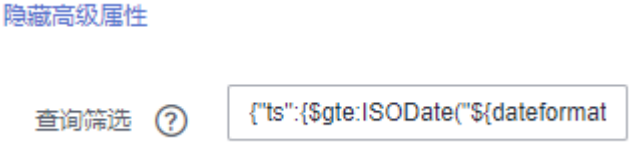
使用CDM导出MongoDB或者DDS的数据时，支持导出指定时间段内的数据，配合CDM的定时任务，可以实现MongoDB/DDS的增量迁移。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为（数据开发作业计划启动时间-偏移量），而不是（CDM作业实际启动时间-偏移量）。

在创建CDM表/文件迁移的作业，源连接选择为MongoDB连接或者DDS连接时，高级属性的可选参数中可以配置查询筛选。

图 1-6 MongoDB 查询筛选



此参数支持配置为**时间宏变量**，例如起始时间配置为**{"ts":{$gte:ISODate("${dateformat(yyyy-MM-dd'T'HH:mm:ss.SSS'Z',-1,DAY)}")}}**，表示查找ts字段中大于时间宏转换后的值，即只导出昨天以后的数据。

参数配置后，CDM就只导出前一天内的数据，再将该作业配置为每天0点执行一次，就可以增量同步每天新生成的数据。

1.2 时间宏变量使用解析

在创建表/文件迁移作业时，CDM支持在源端和目的端的以下参数中配置时间宏变量：

- 源端的源目录或文件
- 源端的表名
- “通配符”过滤类型中的目录过滤器和文件过滤器
- “时间过滤”中的起始时间和终止时间
- 分区过滤条件和Where子句
- 目的端的写入目录
- 目的端的表名

支持通过宏定义变量表示符“${}”来完成时间类型的宏定义，当前支持两种类型：dateformat和timestamp。

通过时间宏变量+定时执行作业，可以实现数据库增量同步和文件增量同步。

说明

如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为（数据开发作业计划启动时间-偏移量），而不是（CDM作业实际启动时间-偏移量）。

dateformat

dateformat支持两种形式的参数：

- dateformat(format)

  format表示返回日期的格式，格式定义参考“java.text.SimpleDateFormat.java”中的定义。

  例如当前日期为“2017-10-16 09:00:00”，则“yyyy-MM-dd HH:mm:ss”表示“2017-10-16 09:00:00”。

- dateformat(format, dateOffset, dateType)
  - format表示返回日期的格式。
  - dateOffset表示日期的偏移量。
  - dateType表示日期的偏移量的类型。

    目前dateType支持以下几种类型：SECOND（秒），MINUTE（分钟），HOUR（小时），DAY（天），MONTH（月），YEAR（年）。

说明

其中MONTH（月），YEAR（年）的偏移量类型存在特殊场景：

- 对于年、月来说，若进行偏移后实际没有该日期，则按照日历取该月最大的日期。
- 不支持在源端和目的端的“时间过滤”参数中的起始时间、终止时间使用年、月的偏移。

例如当前日期为“2023-03-01 09:00:00”，则：

- “dateformat(yyyy-MM-dd HH:mm:ss, -1, YEAR)”表示当前时间的前一年，也就是“2022-03-01 09:00:00”。
- “dateformat(yyyy-MM-dd HH:mm:ss, -3, MONTH)”表示当前时间的前三月，也就是“2022-12-01 09:00:00”。
- “dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)”表示当前时间的前一天，也就是“2023-02-28 09:00:00”。
- “dateformat(yyyy-MM-dd HH:mm:ss, -1, HOUR)”表示当前时间的前一小时，也就是“2023-03-01 08:00:00”。
- “dateformat(yyyy-MM-dd HH:mm:ss, -1, MINUTE)”表示当前时间的前一分钟，也就是“2023-03-01 08:59:00”。
- “dateformat(yyyy-MM-dd HH:mm:ss, -1, SECOND)”表示当前时间的前一秒，也就是“2023-03-01 08:59:59”。

timestamp

timestamp支持两种形式的参数：

- timestamp()

  返回当前时间的时间戳，即从1970年到现在的毫秒数，如1508078516286。
- timestamp(dateOffset, dateType)

  返回经过时间偏移后的时间戳，“dateOffset”和“dateType”表示日期的偏移量以及偏移量的类型。

  例如当前日期为“2017-10-16 09:00:00”，则“timestamp(-10, MINUTE)”返回当前时间点10分钟前的时间戳，即“1508115000000”。

时间变量宏定义具体展示

假设当前时间为“2017-10-16 09:00:00”，时间变量宏定义具体如**表1-1**所示。

表 1-1 时间变量宏定义具体展示

宏变量	含义	实际显示效果
${dateformat(yyyy-MM-dd)}	以yyyy-MM-dd格式返回当前时间。	2017-10-16
${dateformat(yyyy/MM/dd)}	以yyyy/MM/dd格式返回当前时间。	2017/10/16
${dateformat(yyyy_MM_dd HH:mm:ss)}	以yyyy_MM_dd HH:mm:ss格式返回当前时间。	2017_10_16 09:00:00

宏变量	含义	实际显示效果
${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前一天。	2017-10-15 09:00:00
${timestamp()}	返回当前时间的时间戳，即1970年1月1日（00:00:00 GMT）到当前时间的毫秒数。	1508115600000
${timestamp(-10, MINUTE)}	返回当前时间点10分钟前的时间戳。	1508115000000
${timestamp(dateformat(yyyyMMdd))}	返回今天0点的时间戳。	1508083200000
${timestamp(dateformat(yyyyMMdd,-1,DAY))}	返回昨天0点的时间戳。	1507996800000
${timestamp(dateformat(yyyyMMddHH))}	返回当前整小时的时间戳。	1508115600000

路径和表名的时间宏变量

如图1-7所示，如果将：

- 源端的“表名”配置为“CDM_/${dateformat(yyyy-MM-dd)}”。
- 目的端的“写入目录”配置为“/opt/ttxx/${timestamp()}”。

经过宏定义转换，这个作业表示：将Oracle数据库的“SQOOP.CDM_20171016”表中数据，迁移到HDFS的“/opt/ttxx/1508115701746”目录中。

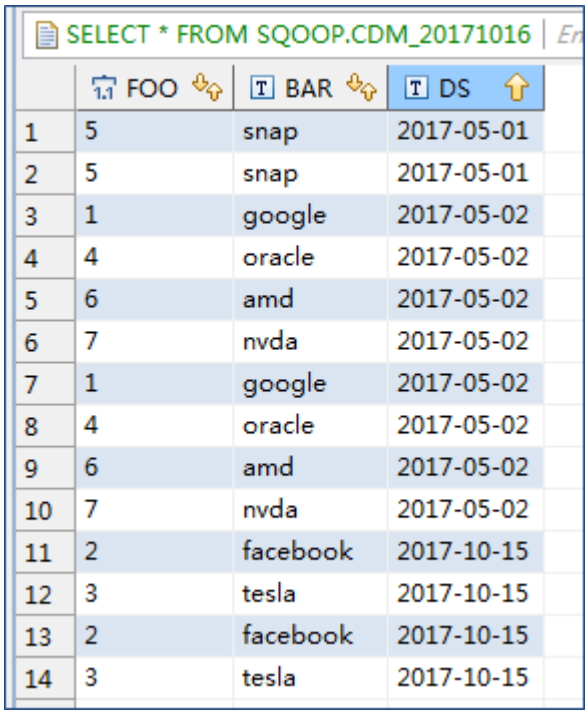
图 1-7 源表名和写入目录配置为时间宏变量

目前也支持一个表名或路径名中有多个宏定义变量，例如“/opt/ttxx/${dateformat(yyyy-MM-dd)}/${timestamp()}”，经过转换后为“/opt/ttxx/2017-10-16/1508115701746”。

Where 子句中的时间宏变量

以SQOOP.CDM_20171016表为例，该表中存在表示时间的列DS，如图1-8所示。

图 1-8 表数据



SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	snap	2017-05-01
2	5	snap	2017-05-01
3	1	google	2017-05-02
4	4	oracle	2017-05-02
5	6	amd	2017-05-02
6	7	nvda	2017-05-02
7	1	google	2017-05-02
8	4	oracle	2017-05-02
9	6	amd	2017-05-02
10	7	nvda	2017-05-02
11	2	facebook	2017-10-15
12	3	tesla	2017-10-15
13	2	facebook	2017-10-15
14	3	tesla	2017-10-15

假设当前时间为“2017-10-16”，要导出前一天的数据（即DS=‘2017-10-15’），则可以在创建作业时配置“Where子句”为**DS='${dateformat(yyyy-MM-dd,-1,DAY)}'**，即可将符合DS=‘2017-10-15’条件的数据导出。

时间宏变量和定时任务配合完成增量同步

这里列举两个简单的使用场景：

- 数据库表中存在表示时间的列DS，类型为“varchar(30)”，插入的时间格式类似于“2017-xx-xx”。

  定时任务中，重复周期为1天，每天的凌晨0点执行定时任务。配置“Where子句”为**DS='${dateformat(yyyy-MM-dd,-1,DAY)}'**，这样就可以在每天的凌晨0点导出前一天产生的所有数据。

- 数据库表中存在表示时间的列time，类型为“Number”，插入的时间格式为时间戳。

  定时任务中，重复周期为1天，每天的凌晨0点执行定时任务。配置“Where子句”为**time between ${timestamp(-1,DAY)} and ${timestamp()}**，这样就可以在每天的凌晨0点导出前一天产生的所有数据。

其它的配置方式原理相同。

1.3 事务模式迁移

CDM的事务模式迁移，是指当CDM作业执行失败时，将数据回滚到作业开始之前的状态，自动清理目的表中的数据。

- 参数位置：创建表/文件迁移的作业时，如果目的端为关系型数据库，在目的端作业配置的高级属性中，可以通过“先导入阶段表”参数选择是否启用事务模式。

- 参数原理：如果启用，在作业执行时CDM会自动创建临时表，先将数据导入到该临时表，导入成功后再通过数据库的事务模式将数据迁移到目标表中；导入失败则将目的表回滚到作业开始之前的状态。

图 1-9 事务模式迁移

说明

如果“导入开始前”选择“清除部分数据”或“清除全部数据”，CDM的事务模式不会回滚已经删除的数据。

1.4 迁移文件时加解密

在迁移文件到文件系统时，CDM支持对文件加解密，目前支持以下加密方式：

- AES-256-GCM加密
- KMS加密

AES-256-GCM 加密

目前只支持AES-256-GCM（NoPadding）。该加密算法在目的端为加密，在源端为解密，支持的源端与目的端数据源如下。

- 源端支持的数据源：HDFS（使用二进制格式传输时支持）。
- 目的端支持的数据源：HDFS（使用二进制格式传输时支持）。

下面分别以HDFS导出加密文件时解密、导入文件到HDFS时加密为例，介绍AES-256-GCM加解密的使用方法。

- **源端配置解密**

  创建从HDFS导出文件的CDM作业时，源端数据源选择HDFS、文件格式选择二进制格式后，在“源端作业配置”的“高级属性”中，配置如下参数。

  a. 加密方式：选择“AES-256-GCM”。
  b. 数据加密密钥：这里的密钥必须与加密时配置的密钥一致，否则解密出来的数据会错误，且系统不会提示异常。
  c. 初始化向量：这里的初始化向量必须与加密时配置的初始化向量一致，否则解密出来的数据会错误，且系统不会提示异常。

  这样CDM从HDFS导出加密过的文件时，写入目的端的文件便是解密后的明文文件。

- **目的端配置加密**

  创建CDM导入文件到HDFS的作业时，目的端数据源选择HDFS、文件格式选择二进制格式后，在“目的端作业配置”的“高级属性”中，配置如下参数。

  a. 加密方式：选择“AES-256-GCM”。
  b. 数据加密密钥：用户自定义密钥，密钥由长度64的十六进制数组成，不区分大小写但必须64位，例如“DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B”。
  c. 初始化向量：用户自定义初始化向量，初始化向量由长度32的十六进制数组成，不区分大小写但必须32位，例如“5C91687BA886EDCD12ACBC3FF19A3C3F”。

  这样在CDM导入文件到HDFS时，目的端HDFS上的文件便是经过AES-256-GCM算法加密后的文件。

KMS 加密

说明

源端解密不支持KMS。

CDM目前只支持导入文件到OBS时，目的端使用KMS加密，表/文件迁移和整库迁移都支持。在“目的端作业配置”的“高级属性”中配置。

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

当启用KMS加密功能后，用户上传对象时，数据会加密成密文存储在OBS。用户从OBS下载加密对象时，存储的密文会先在OBS服务端解密为明文，再提供给用户。

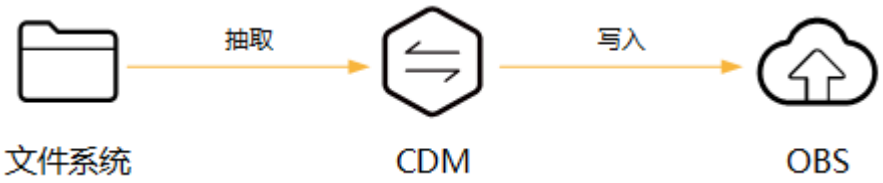
说明

- 如果选择使用KMS加密，则无法**使用MD5校验一致性**。
- 如果这里使用其它项目的KMS ID，则需要修改“项目ID”参数为KMS ID所属的项目ID；如果KMS ID与CDM在同一个项目下，“项目ID”参数保持默认即可。
- 使用KMS加密后，OBS上对象的加密状态不可以修改。
- 使用中的KMS密钥不可以删除，如果删除将导致加密对象不能下载。

1.5 MD5 校验文件一致性

CDM数据迁移以抽取-写入模式进行，CDM首先从源端抽取数据，然后将数据写入到目的端。在迁移文件到OBS时，迁移模式如**图1-10**所示。

图 1-10 迁移文件到 OBS



在这个过程中，CDM支持使用MD5检验文件一致性。

- **抽取时**
  - 该功能支持源端为OBS、HDFS、FTP、SFTP、HTTP。可校验CDM抽取的文件，是否与源文件一致。
  - 该功能由源端作业参数“MD5文件名后缀”控制（“文件格式”为“二进制格式”时生效），配置为源端文件系统中的MD5文件名后缀。
  - 当源端数据文件同一目录下有对应后缀的保存md5值的文件，例如build.sh和build.sh.md5在同一目录下。若配置了“MD5文件名后缀”，则只迁移有MD5值的文件至目的端，没有MD5值或者MD5不匹配的数据文件将迁移失败，MD5文件自身不被迁移。
  - 若未配置“MD5文件名后缀”，则迁移所有文件。
- **写入时**
  - 该功能目前只支持目的端为OBS。可校验写入OBS的文件，是否与CDM抽取的文件一致。
  - 该功能由目的端作业参数“校验MD5值”控制，读取文件后写入OBS时，通过HTTP Header将MD5值提供给OBS做写入校验，并将校验结果写入OBS桶（该桶可以不是存储迁移文件的桶）。如果源端没有MD5文件则不校验。

说明

- 迁移文件到文件系统时，目前只支持校验CDM抽取的文件是否与源文件一致（即只校验抽取的数据）。
- 迁移文件到OBS时，支持抽取和写入文件时都校验。
- 如果选择使用MD5校验，则无法**使用KMS加密**。

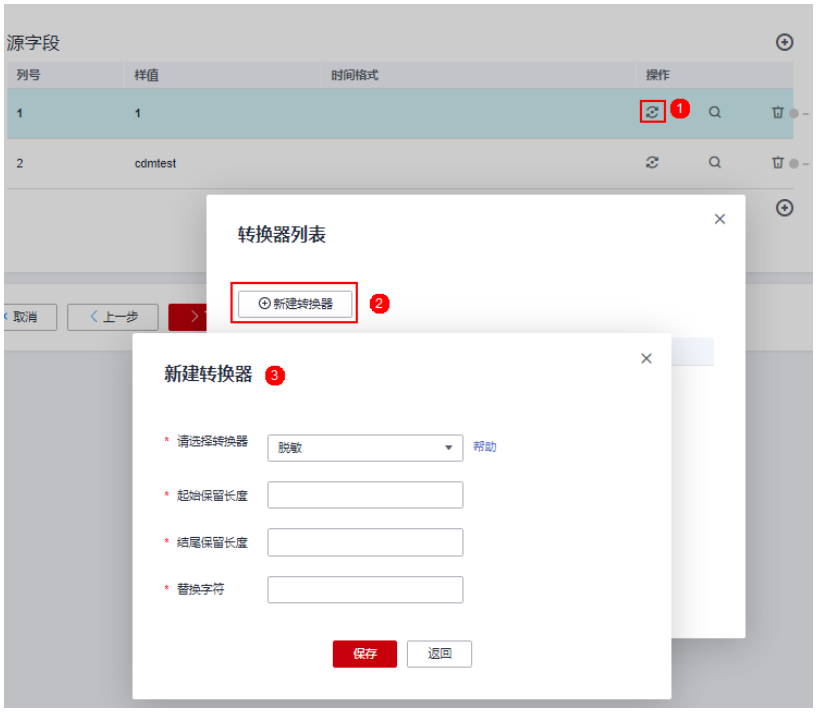
1.6 字段转换器配置指导

操作场景

- 作业参数配置完成后，将进行字段映射的配置，您可以单击操作列下 创建字段转换器。
- 如果是文件类数据源（FTP/SFTP/HDFS/OBS）之间相互迁移数据，且源端“文件格式”配置为“二进制格式”（即不解析文件内容直接传输），则没有字段映射这一步骤。

在创建表/文件迁移作业的字段映射界面，可新建字段转换器，如下图所示。

图 1-11 新建字段转换器



CDM可以在迁移过程中对字段进行转换，目前支持以下字段转换器：

- 脱敏
- 去前后空格
- 字符串反转
- 字符串替换
- 去换行
- 表达式转换

约束限制

- 作业源端开启“使用SQL语句”参数时不支持配置转换器。

- 如果在字段映射界面，CDM通过获取样值的方式无法获得所有列（例如从HBase/CloudTable/MongoDB导出数据时，CDM有较大概率无法获得所有列），则可以单击后选择“添加新字段”来手动增加，确保导入到目的端的数据完整。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datetime）。
- 当作业源端为OBS、迁移CSV文件时，并且配置“解析首行为列名”参数的场景下显示列名。
- 当使用二进制格式进行文件到文件的迁移时，没有配置字段转换器这一步。
- 自动创表场景下，需在目的端表中提前手动新增字段，再在字段映射里新增字段。
- 添加完字段后，新增的字段在界面不显示样值，不会影响字段值的传输，CDM会将字段值直接写入目的端。
- 如果字段映射关系不正确，您可以通过拖拽字段、单击对字段批量映射两种方式来调整字段映射关系。
- 创建表达式转换器时，表达式的功能是对该字段的数据进行处理，故不建议使用时间宏，如需使用，请根据以下场景处理（源端是文件类的配置时仅支持**方式一**）：
  - 方式一：新建表达式转换器时，表达式需要用''包围。

    ${dateformat(yyyy-MM-dd)}不加引号使用时，解析成2017-10-16之后还会进行运算，将'-'识别为减号，导致结果为1991，**须使用'${dateformat(yyyy-MM-dd)}'**，即'2017-10-16'。

    图 1-12 使用''包围表达式

  - 方式二：源字段中新增自定义字段，在样值中填写时间宏变量，重新进行字段映射处理。

    图 1-13 源字段新增自定义字段

- 如果是导入到数据仓库服务（DWS），则还需在目的字段中选择分布列，建议按如下顺序选取分布列：

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

脱敏

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

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

去前后空格

自动去字符串前后的空值，不需要配置参数。

字符串反转

自动反转字符串，例如将“ABC”转换为“CBA”，不需要配置参数。

字符串替换

替换字符串，需要用户配置被替换的对象，以及替换后的值。

去换行

将字段中的换行符（\n、\r、\r\n）删除。

表达式转换

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

- 表达式支持以下两个环境变量：
  - value：当前字段值。
  - row：当前行，数组类型。
- 表达式支持的工具类用法罗列如下，未列出即表示不支持：
  a. 如果当前字段为字符串类型，将字符串全部转换为小写，例如将“aBC”转换为“abc”。

     表达式：StringUtils.lowerCase(value)
  b. 将当前字段的字符串全部转为大写。

     表达式：StringUtils.upperCase(value)
  c. 如果想将第1个日期字段格式从“2018-01-05 15:15:05”转换为“20180105”。

     表达式：DateUtils.format(DateUtils.parseDate(row[0],"yyyy-MM-dd HH:mm:ss"),"yyyyMMdd")

d. 如果想将时间戳转换成“yyyy-MM-dd hh:mm:ss”格式的日期字符串的类型，例如字段值为“1701312046588”，转化后为“2023-11-30 10:40:46”。

表达式：DateUtils.format(NumberUtils.toLong(value),"yyyy-MM-dd HH:mm:ss")

e. 如果想将“yyyy-MM-dd hh:mm:ss”格式的日期字符串转换成时间戳的类型。

表达式：DateUtils.getTime(DateUtils.parseDate(value,"yyyy-MM-dd hh:mm:ss"))

f. 如果当前字段值为“yyyy-MM-dd”格式的日期字符串，需要截取年，例如字段值为“2017-12-01”，转换后为“2017”。

表达式：StringUtils.substringBefore(value,"-")

g. 如果当前字段值为数值类型，转换后值为当前值的两倍。

表达式：value*2

h. 如果当前字段值为“true”，转换后为“Y”，其它值则转换后为“N”。

表达式：value=="true"?"Y":"N"

i. 如果当前字段值为字符串类型，当为空时，转换为“Default”，否则不转换。

表达式：empty value? "Default":value

j. 如果想将日期字段格式从“2018/01/05 15:15:05”转换为“2018-01-05 15:15:05”。

表达式：DateUtils.format(DateUtils.parseDate(value,"yyyy/MM/dd HH:mm:ss"),"yyyy-MM-dd HH:mm:ss")

k. 获取一个36位的UUID（Universally Unique Identifier，通用唯一识别码）。

表达式：CommonUtils.randomUUID()

l. 如果当前字段值为字符串类型，将首字母转换为大写，例如将“cat”转换为“Cat”。

表达式：StringUtils.capitalize(value)

m. 如果当前字段值为字符串类型，将首字母转换为小写，例如将“Cat”转换为“cat”。

表达式：StringUtils.uncapitalize(value)

n. 如果当前字段值为字符串类型，使用空格填充为指定长度，并且将字符串居中，当字符串长度不小于指定长度时不转换，例如将“ab”转换为长度为4的“ab”。

表达式：StringUtils.center(value,*4*)

o. 删除字符串末尾的一个换行符（包括“\n”、“\r”或者“\r\n”），例如将“abc\r\n\r\n”转换为“abc\r\n”。

表达式：StringUtils.chomp(value)

p. 如果字符串中包含指定的字符串，则返回布尔值true，否则返回false。例如“abc”中包含“a”，则返回true。

表达式：StringUtils.contains(value,"*a*")

q. 如果字符串中包含指定字符串的任一字符，则返回布尔值true，否则返回false。例如“zzabyycdxx”中包含“z”或“a”任意一个，则返回true。

表达式：StringUtils.containsAny(value,"*za*")

r. 如果字符串中不包含指定的所有字符，则返回布尔值true，包含任意一个字符则返回false。例如“abz”中包含“xyz”里的任意一个字符，则返回false。

表达式：StringUtils.containsNone(value,"*xyz*")

s. 如果当前字符串只包含指定字符串中的字符，则返回布尔值true，包含任意一个其它字符则返回false。例如“abab”只包含“abc”中的字符，则返回true。

表达式：StringUtils.containsOnly(value,"*abc*")

t. 如果字符串为空或null，则转换为指定的字符串，否则不转换。例如将空字符转换为null。

表达式：StringUtils.defaultIfEmpty(value,*null*)

u. 如果字符串以指定的后缀结尾（包括大小写），则返回布尔值true，否则返回false。例如“abcdef”后缀不为null，则返回false。

表达式：StringUtils.endsWith(value,*null*)

v. 如果字符串和指定的字符串完全一样（包括大小写），则返回布尔值true，否则返回false。例如比较字符串“abc”和“ABC”，则返回false。

表达式：StringUtils.equals(value,"*ABC*")

w. 从字符串中获取指定字符串的第一个索引，没有则返回整数-1。例如从“aabaabaa”中获取“ab”的第一个索引1。

表达式：StringUtils.indexOf(value,"*ab*")

x. 从字符串中获取指定字符串的最后一个索引，没有则返回整数-1。例如从“aFkyk”中获取“k”的最后一个索引4。

表达式：StringUtils.lastIndexOf(value,"*k*")

y. 从字符串中指定的位置往后查找，获取指定字符串的第一个索引，没有则转换为“-1”。例如“aabaabaa”中索引3的后面，第一个“b”的索引是5。

表达式：StringUtils.indexOf(value,"*b*",*3*)

z. 从字符串获取指定字符串中任一字符的第一个索引，没有则返回整数-1。例如从“zzabyycdxx”中获取“z”或“a”的第一个索引0。

表达式：StringUtils.indexOfAny(value,"*za*")

aa. 如果字符串仅包含Unicode字符，返回布尔值true，否则返回false。例如“ab2c”中包含非Unicode字符，返回false。

表达式：StringUtils.isAlpha(value)

ab. 如果字符串仅包含Unicode字符或数字，返回布尔值true，否则返回false。例如“ab2c”中仅包含Unicode字符和数字，返回true。

表达式：StringUtils.isAlphanumeric(value)

ac. 如果字符串仅包含Unicode字符、数字或空格，返回布尔值true，否则返回false。例如“ab2c”中仅包含Unicode字符和数字，返回true。

表达式：StringUtils.isAlphanumericSpace(value)

ad. 如果字符串仅包含Unicode字符或空格，返回布尔值true，否则返回false。例如“ab2c”中包含Unicode字符和数字，返回false。

表达式：StringUtils.isAlphaSpace(value)

ae. 如果字符串仅包含ASCII可打印字符，返回布尔值true，否则返回false。例如“!ab-c~”返回true。

表达式：StringUtils.isAsciiPrintable(value)

af. 如果字符串为空或null，返回布尔值true，否则返回false。

表达式：StringUtils.isEmpty(value)

ag. 如果字符串中仅包含Unicode数字，返回布尔值true，否则返回false。

表达式：StringUtils.isNumeric(value)

ah. 获取字符串最左端的指定长度的字符，例如获取“abc”最左端的2位字符“ab”。

表达式：StringUtils.left(value,*2*)

ai. 获取字符串最右端的指定长度的字符，例如获取“abc”最右端的2位字符“bc”。

表达式：StringUtils.right(value,*2*)

aj. 将指定字符串拼接至当前字符串的左侧，需同时指定拼接后的字符串长度，如果当前字符串长度不小于指定长度，则不转换。例如将“yz”拼接到“bat”左侧，拼接后长度为8，则转换后为“yzyzybat”。

表达式：StringUtils.leftPad(value,*8*,"*yz*")

ak. 将指定字符串拼接至当前字符串的右侧，需同时指定拼接后的字符串长度，如果当前字符串长度不小于指定长度，则不转换。例如将“yz”拼接到“bat”右侧，拼接后长度为8，则转换后为“batyzyzy”。

表达式：StringUtils.rightPad(value,*8*,"*yz*")

al. 如果当前字段为字符串类型，获取当前字符串的长度，如果该字符串为null，则返回0。

表达式：StringUtils.length(value)

am. 如果当前字段为字符串类型，删除其中所有的指定字符串，例如从“queued”中删除“ue”，转换后为“qd”。

表达式：StringUtils.remove(value,"*ue*")

an. 如果当前字段为字符串类型，移除当前字段末尾指定的子字符串。指定的子字符串若不在当前字段的末尾，则不转换，例如移除当前字段“www.domain.com”后的“.com”。

表达式：StringUtils.removeEnd(value,"*.com*")

ao. 如果当前字段为字符串类型，移除当前字段开头指定的子字符串。指定的子字符串若不在当前字段的开头，则不转换，例如移除当前字段“www.domain.com”前的“www.”。

表达式：StringUtils.removeStart(value,"*www.*")

ap. 如果当前字段为字符串类型，替换当前字段中所有的指定字符串，例如将“aba”中的“a”用“z”替换，转换后为“zbz”。

表达式：StringUtils.replace(value,"*a*","*z*")

aq. 如果当前字段为字符串类型，一次替换字符串中的多个字符，例如将字符串“hello”中的“h”用“j”替换，“o”用“y”替换，转换后为“jelly”。

表达式：StringUtils.replaceChars(value,"*ho*","*jy*")

ar. 如果字符串以指定的前缀开头（区分大小写），则返回布尔值true，否则返回false，例如当前字符串“abcdef”以“abc”开头，则返回true。

表达式：StringUtils.startsWith(value,"*abc*")

as. 如果当前字段为字符串类型，去除字段中首、尾处所有指定的字符，例如去除“abcyx”中首尾所有的“x”、“y”、“z”和“b”，转换后为“abc”。

表达式：StringUtils.strip(value,"*xyz*b")

at. 如果当前字段为字符串类型，去除字段末尾所有指定的字符，例如去除当前字段末尾的"abc"字符串。

表达式：StringUtils.stripEnd(value,*"abc"*)

au. 如果当前字段为字符串类型，去除字段开头所有指定的字符，例如去除当前字段开头的所有空格。

表达式：StringUtils.stripStart(value,*null*)

av. 如果当前字段为字符串类型，获取字符串指定位置后（索引从0开始，包括指定位置的字符）的子字符串，指定位置如果为负数，则从末尾往前计算位置，末尾第一位为-1。例如获取“abcde”第2个字符（即c）及之后的字符串，则转换后为“cde”。

表达式：StringUtils.substring(value,*2*)

aw. 如果当前字段为字符串类型，获取字符串指定区间（索引从0开始，区间起点包括指定位置的字符，区间终点不包含指定位置的字符）的子字符串，区间位置如果为负数，则从末尾往前计算位置，末尾第一位为-1。例如获取“abcde”第2个字符（即c）及之后、第4个字符（即e）之前的字符串，则转换后为“cd”。

表达式：StringUtils.substring(value,*2*,4)

ax. 如果当前字段为字符串类型，获取当前字段里第一个指定字符后的子字符串。例如获取“abcba”中第一个“b”之后的子字符串，转换后为“cba”。

表达式：StringUtils.substringAfter(value,"*b*")

ay. 如果当前字段为字符串类型，获取当前字段里最后一个指定字符后的子字符串。例如获取“abcba”中最后一个“b”之后的子字符串，转换后为“a”。

表达式：StringUtils.substringAfterLast(value,"*b*")

az. 如果当前字段为字符串类型，获取当前字段里第一个指定字符前的子字符串。例如获取“abcba”中第一个“b”之前的子字符串，转换后为“a”。

表达式：StringUtils.substringBefore(value,"*b*")

ba. 如果当前字段为字符串类型，获取当前字段里最后一个指定字符前的子字符串。例如获取“abcba”中最后一个“b”之前的子字符串，转换后为“abc”。

表达式：StringUtils.substringBeforeLast(value,"*b*")

bb. 如果当前字段为字符串类型，获取嵌套在指定字符串之间的子字符串，没有匹配的则返回null。例如获取“tagabctag”中“tag”之间的子字符串，转换后为“abc”。

表达式：StringUtils.substringBetween(value,"*tag*")

bc. 如果当前字段为字符串类型，删除当前字符串两端的控制字符（char≤32），例如删除字符串前后的空格。

表达式：StringUtils.trim(value)

bd. 将当前字符串转换为字节，如果转换失败，则返回0。

表达式：NumberUtils.toByte(value)

be. 将当前字符串转换为字节，如果转换失败，则返回指定值，例如指定值配置为1。

表达式：NumberUtils.toByte(value,*1*)

bf. 将当前字符串转换为Double数值，如果转换失败，则返回0.0d。

表达式：NumberUtils.toDouble(value)

bg. 将当前字符串转换为Double数值，如果转换失败，则返回指定值，例如指定值配置为1.1d。

表达式：NumberUtils.toDouble(value,*1.1d*)

bh. 将当前字符串转换为Float数值，如果转换失败，则返回0.0f。

表达式：NumberUtils.toFloat(value)

bi. 将当前字符串转换为Float数值，如果转换失败，则返回指定值，例如配置指定值为1.1f。

表达式：NumberUtils.toFloat(value,*1.1f*)

bj. 将当前字符串转换为Int数值，如果转换失败，则返回0。

表达式：NumberUtils.toInt(value)

bk. 将当前字符串转换为Int数值，如果转换失败，则返回指定值，例如配置指定值为1。

表达式：NumberUtils.toInt(value,*1*)

bl. 将字符串转换为Long数值，如果转换失败，则返回0。

表达式：NumberUtils.toLong(value)

bm. 将当前字符串转换为Long数值，如果转换失败，则返回指定值，例如配置指定值为1L。

表达式：NumberUtils.toLong(value,*1L*)

bn. 将字符串转换为Short数值，如果转换失败，则返回0。

表达式：NumberUtils.toShort(value)

bo. 将当前字符串转换为Short数值，如果转换失败，则返回指定值，例如配置指定值为1。

表达式：NumberUtils.toShort(value,*1*)

bp. 将当前IP字符串转换为Long数值，例如将“10.78.124.0”转换为LONG数值是“172915712”。

表达式：CommonUtils.ipToLong(value)

bq. 从网络读取一个IP与物理地址映射文件，并存放到Map集合，这里的URL是IP与地址映射文件存放地址，例如“http://10.114.205.45:21203/sqoop/IpList.csv”。

表达式：HttpsUtils.downloadMap("*url*")

br. 将IP与地址映射对象缓存起来并指定一个key值用于检索，例如“ipList”。

表达式：CommonUtils.setCache("*ipList*",HttpsUtils.downloadMap("*url*"))

bs. 取出缓存的IP与地址映射对象。

表达式：CommonUtils.getCache("*ipList*")

bt. 判断是否有IP与地址映射缓存。

表达式：CommonUtils.cacheExists("*ipList*")

bu. 根据指定的偏移类型（month/day/hour/minute/second）及偏移量（正数表示增加，负数表示减少），将指定格式的时间转换为一个新时间，例如将“2019-05-21 12:00:00”增加8个小时。

表达式：DateUtils.getCurrentTimeByZone("*yyyy-MM-dd HH:mm:ss*",value, "*hour*", *8*)

bv. 如果value值为空或者null时，则返回字符串"aaa"，否则返回value。

表达式：StringUtils.defaultIfEmpty(value,"aaa")

1.7 新增字段操作指导

操作场景

- 作业参数配置完成后，将进行字段映射的配置，您可以通过字段映射界面的可自定义新增字段。
- 如果是文件类数据源（FTP/SFTP/HDFS/OBS）之间相互迁移数据，且源端“文件格式”配置为“二进制格式”（即不解析文件内容直接传输），则没有字段映射这一步骤。
- 其他场景下，CDM会自动匹配源端和目的端数据表字段，需用户检查字段映射关系和时间格式是否正确，例如：源字段类型是否可以转换为目的字段类型。

您可以单击字段映射界面的选择“添加新字段”自定义新增字段，通常用于标记数据库来源，以确保导入到目的端数据的完整性。

图 1-14 字段映射

目前支持以下类型自定义字段：

- **常量**

  常量参数即参数值是固定的参数，不需要重新配置值。例如“lable”=“friends”用来标识常量值。

- **变量**

  您可以使用时间宏、表名宏、版本宏等变量来标记数据库来源信息。变量的语法：${variable}，其中“variable”指的是变量。例如“input_time”=“${timestamp()}”用来标识当前时间的时间戳。

- **表达式**

  您可以使用表达式语言根据运行环境动态生成参数值。表达式的语法：#{expr}，其中“expr”指的是表达式。例如“time”=“#{DateUtil.now()}”用来标识当前日期字符串。

约束限制

- 如果在字段映射界面，CDM通过获取样值的方式无法获得所有列（例如从HBase/CloudTable/MongoDB导出数据时，CDM有较大概率无法获得所有列），则可以单击后选择“添加新字段”来手动增加，确保导入到目的端的数据完整。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datetime）。
- 当作业源端为OBS、迁移CSV文件时，并且配置“解析首行为列名”参数的场景下显示列名。

- 当使用二进制格式进行文件到文件的迁移时，没有字段映射这一步。
- 自动创表场景下，需在目的端表中提前手动新增字段，再在字段映射里新增字段。
- 添加完字段后，新增的字段在界面不显示样值，不会影响字段值的传输，CDM会将字段值直接写入目的端。
- 如果字段映射关系不正确，您可以通过拖拽字段、单击  对字段批量映射两种方式来调整字段映射关系。
- 如果是导入到数据仓库服务（DWS），则还需在目的字段中选择分布列，建议按如下顺序选取分布列：
  a. 有主键可以使用主键作为分布列。
  b. 多个数据段联合做主键的场景，建议设置所有主键作为分布列。
  c. 在没有主键的场景下，如果没有选择分布列，DWS会默认第一列作为分布列，可能会有数据倾斜风险。
- 如CDM不支持源端迁移字段类型，请参见**不支持数据类型转换规避指导**将字段类型转换为CDM支持的类型。

1.8 指定文件名迁移

从FTP/SFTP/OBS导出文件时，CDM支持指定文件名迁移，用户可以单次迁移多个指定的文件（最多50个），导出的多个文件只能写到目的端的同一个目录。

在创建表/文件迁移作业时，如果源端数据源为FTP/SFTP/OBS，CDM源端的作业参数“源目录或文件”支持输入多个文件名（最多50个），文件名之间默认使用“|”分隔，您也可以自定义文件分隔符，从而实现文件列表迁移。

说明

1. 迁移文件或对象时支持文件级增量迁移（通过配置跳过重复文件实现），但不支持断点续传。
   例如要迁移3个文件，第2个文件迁移到一半时由于网络原因失败，再次启动迁移任务时，会跳过第1个文件，从第2个文件开始重新传，但不能从第2个文件失败的位置重新传。
2. 文件迁移时，单个任务支持千万数量的文件，如果待迁移目录下文件过多，建议拆分到不同目录并创建多个任务。

1.9 正则表达式分隔半结构化文本

在创建表/文件迁移作业时，对简单CSV格式的文件，CDM可以使用字段分隔符进行字段分隔。但是对于一些复杂的半结构化文本，由于字段值也包含了分隔符，所以无法使用分隔符进行字段分隔，此时可以使用正则表达式分隔。

正则表达式参数在源端作业参数中配置，要求源连接为对象存储或者文件系统，且“文件格式”必须选择“CSV格式”。

图 1-15 正则表达式参数

在迁移CSV格式的文件时，CDM支持使用正则表达式分隔字段，并按照解析后的结果写入目的端。正则表达式语法请参考对应的相关资料，这里举例下面几种日志文件的正则表达式的写法：

- **Log4J日志**
- **Log4J审计日志**
- **Tomcat日志**
- **Django日志**
- **Apache server日志**

Log4J 日志

- 日志样例：
  ```
  2018-01-11 08:50:59,001 INFO
  [org.apache.sqoop.core.SqoopConfiguration.configureClassLoader(SqoopConfiguration.java:251)]
  Adding jars to current classloader from property: org.apache.sqoop.classpath.extra
  ```
- 正则表达式为：
  ```
  ^(\d.*\d) (\w*)  \[(.*)\] (\w.*).*
  ```
- 解析出的结果如下：

表 1-2 Log4J 日志解析结果

列号	样值
1	2018-01-11 08:50:59,001
2	INFO
3	org.apache.sqoop.core.SqoopConfiguration.configureClassLoader(SqoopConfiguration.java:251)
4	Adding jars to current classloader from property: org.apache.sqoop.classpath.extra

Log4J 审计日志

- 日志样例：
```
2018-01-11 08:51:06,156 INFO
[org.apache.sqoop.audit.FileAuditLogger.logAuditEvent(FileAuditLogger.java:61)]
user=sqoop.anonymous.user    ip=189.xxx.xxx.75    op=show    obj=version    objId=x
```
- 正则表达式为：
```
^(\d.*\d) (\w*) \[(.*)\] user=(\w.*)    ip=(\w.*)    op=(\w.*)    obj=(\w.*)    objId=(.*).*
```
- 解析结果如下：

表 1-3 Log4J 审计日志解析结果

列号	样值
1	2018-01-11 08:51:06,156
2	INFO
3	org.apache.sqoop.audit.FileAuditLogger.logAuditEvent(FileAuditLogger.java:61)
4	sqoop.anonymous.user
5	189.xxx.xxx.75
6	show
7	version
8	x

Tomcat 日志

- 日志样例：
```
11-Jan-2018 09:00:06.907 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log OS
Name:               Linux
```
- 正则表达式为：
```
^(\d.*\d) (\w*) \[(.*)\] ([\w\.]*) (\w.*).*
```
- 解析结果如下：

表 1-4 Tomcat 日志解析结果

列号	样值
1	11-Jan-2018 09:00:06.907
2	INFO
3	main
4	org.apache.catalina.startup.VersionLoggerListener.log
5	OS Name:Linux

Django 日志

- 日志样例：
  ```
  [08/Jan/2018 20:59:07 ] settings     INFO     Welcome to Hue 3.9.0
  ```
- 正则表达式为：
  ```
  ^\[(.*)\] (\w*)     (\w*)     (.*).*
  ```
- 解析结果如下：

  表 1-5 Django 日志解析结果

列号	样值
1	08/Jan/2018 20:59:07
2	settings
3	INFO
4	Welcome to Hue 3.9.0

Apache server 日志

- 日志样例：
  ```
  [Mon Jan 08 20:43:51.854334 2018] [mpm_event:notice] [pid 36465:tid 140557517657856] AH00489: Apache/2.4.12 (Unix) OpenSSL/1.0.1t configured -- resuming normal operations
  ```
- 正则表达式为：
  ```
  ^\[(.*)\] \[(.*)\] \[(.*)\] (.*).*
  ```
- 解析结果如下：

  表 1-6 Apache server 日志解析结果

列号	样值
1	Mon Jan 08 20:43:51.854334 2018
2	mpm_event:notice
3	pid 36465:tid 140557517657856
4	AH00489: Apache/2.4.12 (Unix) OpenSSL/1.0.1t configured -- resuming normal operations

1.10 记录数据迁移入库时间

CDM在创建表/文件迁移的作业，支持连接器源端为关系型数据库时，在表字段映射中使用时间宏变量增加入库时间字段，用以记录关系型数据库的入库时间等用途。

前提条件

- 已创建连接器源端为关系型数据库，以及目的端数据连接。
- 目的端数据表中已有时间日期字段或时间戳字段。如自动创表场景下，需提前在目的端表中手动创建时间日期字段或时间戳字段。

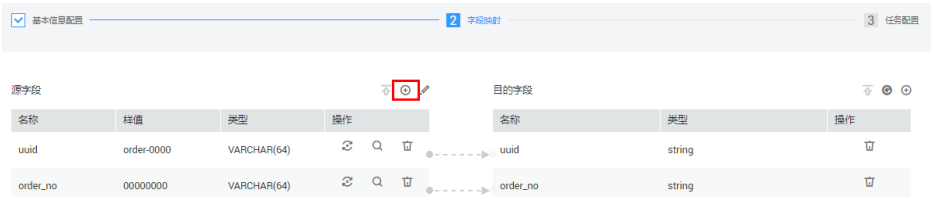
创建表/文件迁移作业

步骤1 在创建表/文件迁移作业时，选择已创建的源端连接器、目的端连接器。

图 1-16 配置作业

步骤2 单击“下一步”，进入“字段映射”配置页面后，单击源字段图标。

图 1-17 配置字段映射



步骤3 选择“自定义字段”页签，填写字段名称及字段值后单击“确认”按钮，例如：

名称：InputTime。

值：${timestamp()}，更多时间宏变量请参见**表1-7**。

图 1-18 添加字段

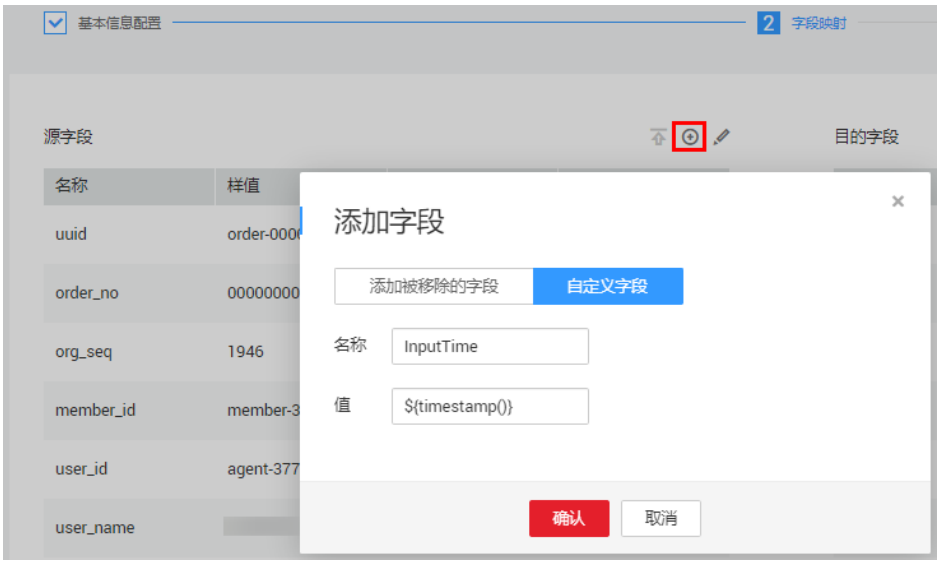


表 1-7 时间变量宏定义具体展示

宏变量	含义	实际显示效果
${dateformat(yyyy-MM-dd)}	以yyyy-MM-dd格式返回当前时间。	2017-10-16
${dateformat(yyyy/MM/dd)}	以yyyy/MM/dd格式返回当前时间。	2017/10/16
${dateformat(yyyy_MM_dd HH:mm:ss)}	以yyyy_MM_dd HH:mm:ss格式返回当前时间。	2017_10_16 09:00:00
${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	以yyyy-MM-dd HH:mm:ss格式返回时间，时间为当前时间的前一天。	2017-10-15 09:00:00
${timestamp()}	返回当前时间的时间戳，即1970年1月1日（00:00:00 GMT）到当前时间的毫秒数。	1508115600000
${timestamp(-10, MINUTE)}	返回当前时间点10分钟前的时间戳。	1508115000000
${timestamp(dateformat(yyyyMMdd))}	返回今天0点的时间戳。	1508083200000
${timestamp(dateformat(yyyyMMdd,-1,DAY))}	返回昨天0点的时间戳。	1507996800000
${timestamp(dateformat(yyyyMMddHH))}	返回当前整小时的时间戳。	1508115600000

说明

- 添加完字段后，新增的字段在界面不显示样值，不会影响字段值的传输，CDM会将字段值直接写入目的端。
- 这里“添加字段”中“自定义字段”的功能，要求源端连接器为JDBC连接器、HBase连接器、MongoDB连接器、ElasticSearch连接器、Kafka连接器，或者目的端为HBase连接器。
- 添加完字段后，请确保自定义入库时间字段与目的端表字段类型相匹配。

步骤4 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

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

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

在历史记录界面单击“日志”，可查看作业的日志信息。

步骤7 前往目的端数据源查看数据迁移的入库时间。

----结束

1.11 文件格式介绍

在创建CDM作业时，有些场景下源端、目的端的作业参数中需要选择“文件格式”，这里分别介绍这几种文件格式的使用场景、子参数、公共参数、使用示例等。

- **CSV格式**
- **JSON格式**
- **二进制格式**
- **文件格式的公共参数**
- **文件格式问题解决方法**

CSV 格式

如果想要读取或写入某个CSV文件，请在选择“文件格式”的时候选择“CSV格式”。CSV格式的主要有以下使用场景：

- 文件导入到数据库、NoSQL。
- 数据库、NoSQL导出到文件。

选择了CSV格式后，通常还可以配置以下可选子参数：

1.换行符

2.字段分隔符

3.编码类型

4.使用包围符

5.使用正则表达式分隔字段

6.首行为标题行

7.写入文件大小

1. **换行符**

   用于分隔文件中的行的字符，支持单字符和多字符，也支持特殊字符。特殊字符可以使用URL编码输入，例如：

   表 1-8 特殊字符对应的 URL 编码

特殊字符	URL编码
空格	%20
Tab	%09
%	%25
回车	%0d
换行	%0a
标题开头\u0001 (SOH)	%01

2. **字段分隔符**

   用于分隔CSV文件中的列的字符，支持单字符和多字符，也支持特殊字符，详见**表1-8**。

3. **编码类型**

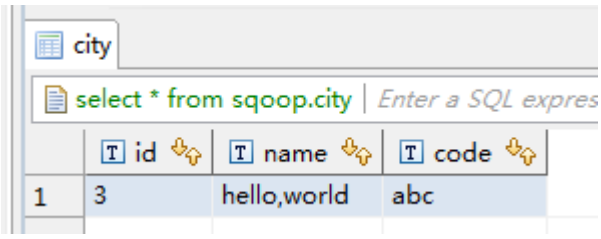
   文件的编码类型，默认是UTF-8，中文的编码有时会采用GBK。

   如果源端指定该参数，则使用指定的编码类型去解析文件；目的端指定该参数，则写入文件的时候，以指定的编码类型写入。

4. **使用包围符**

   - 数据库、NoSQL导出到CSV文件（“使用包围符”在目的端）：当源端某列数据的字符串中出现字段分隔符时，目的端可以通过开启“使用包围符”，将该字符串括起来，作为一个整体写入CSV文件。CDM目前只使用双引号（""）作为包围符。如**图1-19**所示，数据库的name字段的值中包含了字段分隔符逗号：

     图 1-19 包含字段分隔符的字段值

     不使用包围符的时候，导出的CSV文件，数据会显示为：

     ```
     3.hello,world,abc
     ```

     如果使用包围符，导出的数据则为：

     ```
     3,"hello,world",abc
     ```

     如果数据库中的数据已经包含了双引号（""），那么使用包围符后，导出的CSV文件的包围符会是三个双引号（"""）。例如字段的值为：a"hello,world"c，使用包围符后导出的数据为：

     ```
     """a"hello,world"c"""
     ```

   - CSV文件导出到数据库、NoSQL（“使用包围符”在源端）：CSV文件为源，并且其中数据是被包围符括起来的时候，如果想把数据正确的导入到数据库，就需要在源端开启“使用包围符”，这样包围符内的值的，会写入一个字段内。

5. **使用正则表达式分隔字段**

   这个功能是针对一些复杂的半结构化文本，例如日志文件的解析，详见：**使用正则表达式分隔半结构化文本**。

6. **首行为标题行**

   这个参数是针对CSV文件导出到其它地方的场景，如果源端指定了该参数，CDM在抽取数据时将第一行作为标题行。在传输CSV文件的时候会跳过标题行，这时源端抽取的行数，会比目的端写入的行数多一行，并在日志文件中进行说明跳过了标题行。

7. **写入文件大小**

   这个参数是针对数据库导出到CSV文件的场景，如果一张表的数据量比较大，那么导出到CSV文件的时候，会生成一个很大的文件，有时会不方便下载或查看。

这时可以在目的端指定该参数，这样会生成多个指定大小的CSV文件，避免导出的文件过大。该参数的数据类型为整型，单位为MB。

JSON 格式

这里主要介绍JSON文件格式的以下内容：

- **CDM支持解析的JSON类型**
- **记录节点**
- **从JSON文件复制数据**

1. **CDM支持解析的JSON类型：JSON对象、JSON数组。**
   - JSON对象：JSON文件包含单个对象，或者以行分隔/串连的多个对象。
     i. 单一对象JSON：
     ```
     {
        "took" : 190,
        "timed_out" : false,
        "total" : 1000001,
        "max_score" : 1.0
     }
     ```
     ii. 行分隔的JSON对象：
     ```
     {"took" : 188, "timed_out" : false, "total" : 1000003, "max_score" : 1.0 }
     {"took" : 189, "timed_out" : false, "total" : 1000004, "max_score" : 1.0 }
     ```
     iii. 串连的JSON对象：
     ```
     {
       "took": 190,
       "timed_out": false,
       "total": 1000001,
       "max_score": 1.0
     }
     {
       "took": 191,
       "timed_out": false,
       "total": 1000002,
       "max_score": 1.0
     }
     ```
   - JSON数组：JSON文件是包含多个JSON对象的数组。
     ```
     [{
        "took" : 190,
        "timed_out" : false,
        "total" : 1000001,
        "max_score" : 1.0
     },
     {
        "took" : 191,
        "timed_out" : false,
        "total" : 1000001,
        "max_score" : 1.0
     }]
     ```
2. **记录节点**

   记录数据的根节点。该节点对应的数据为JSON数组，CDM会以同一模式从该数组中提取数据。多层嵌套的JSON节点以字符“.”分割。
3. **从JSON文件复制数据**

   a. 示例一：从行分隔/串连的多个对象中提取数据。JSON文件包含了多个JSON对象，例如：
   ```
   {
     "took": 190,
     "timed_out": false,
     "total": 1000001,
   ```



```
  "max_score": 1.0
}
{
  "took": 191,
  "timed_out": false,
  "total": 1000002,
  "max_score": 1.0
}
{
  "took": 192,
  "timed_out": false,
  "total": 1000003,
  "max_score": 1.0
}
```

如果您想要从该JSON对象中提取数据，使用以下格式写入到数据库，只需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON对象”，然后在作业第二步进行字段匹配即可。

took	timedOut	total	maxScore
190	false	1000001	1.0
191	false	1000002	1.0
192	false	1000003	1.0

b. 示例二：从记录节点中提取数据。JSON文件包含了单个的JSON对象，但是其中有效的数据在一个数据节点下，例如：

```
{
  "took": 190,
  "timed_out": false,
  "hits": {
     "total": 1000001,
     "max_score": 1.0,
     "hits":
      [{
        "_id": "650612",
        "_source": {
           "name": "tom",
           "books": ["book1","book2","book3"]
        }
      },
      {
         "_id": "650616",
         "_source": {
            "name": "tom",
            "books": ["book1","book2","book3"]
         }
      },
    {
         "_id": "650618",
         "_source": {
            "name": "tom",
            "books": ["book1","book2","book3"]
         }
      }]
  }
}
```

如果想以如下格式写入到数据库，则需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON对象”，并且指定记录节点为“hits.hits”，然后在作业第二步进行字段匹配。

ID	SourceName	SourceBooks
650612	tom	["book1","book2","book3"]
650616	tom	["book1","book2","book3"]
650618	tom	["book1","book2","book3"]

c. 示例三：从JSON数组中提取数据。JSON文件是包含了多个JSON对象的JSON数组，例如：

```
[{
    "took" : 190,
    "timed_out" : false,
    "total" : 1000001,
    "max_score" : 1.0
 },
 {
    "took" : 191,
    "timed_out" : false,
    "total" : 1000002,
    "max_score" : 1.0
 }]
```

如果想以如下格式写入到数据库，需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON数组”，然后在作业第二步进行字段匹配。

took	timedOut	total	maxScore
190	false	1000001	1.0
191	false	1000002	1.0

d. 示例四：在解析JSON文件的时候搭配转换器。在示例二前提下，想要把hits.max_score字段附加到所有记录中，即以如下格式写入到数据库中：

ID	SourceName	SourceBooks	MaxScore
650612	tom	["book1","book2","book3"]	1.0
650616	tom	["book1","book2","book3"]	1.0
650618	tom	["book1","book2","book3"]	1.0

则需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON对象”，并且指定记录节点为“hits.hits”，然后在作业第二步添加转换器，操作步骤如下：

i. 单击添加字段，新增一个字段。

图 1-20 添加字段

ii. 在添加的新字段后面，单击 添加字段转换器。

图 1-21 添加字段转换器

iii. 创建“表达式转换”的转换器，表达式输入“1.0”，然后保存。

图 1-22 配置字段转换器

二进制格式

如果想要在文件系统间按原样复制文件，则可以选择二进制格式。二进制格式传输文件到文件的速率高、性能稳定，且不需要在作业第二步进行字段匹配。

- **文件传输的目录结构**

  CDM的文件传输，支持单文件，也支持一次传输目录下所有的文件。传输到目的端后，目录结构会保持原样。

- **增量迁移文件**

  使用CDM进行二进制传输文件时，目的端有一个参数“重复文件处理方式”，可以用作文件的增量迁移，具体请参见**文件增量迁移**。

  增量迁移文件的时候，选择“重复文件处理方式”为“跳过重复文件”，这样如果源端有新增的文件，或者是迁移过程中出现了失败，只需要再次运行任务，已经迁移过的文件就不会再次迁移。

- **写入到临时文件**

  二进制迁移文件时候，可以在目的端指定是否写入到临时文件。如果指定了该参数，在文件复制过程中，会将文件先写入到一个临时文件中，迁移成功后，再进行rename或move操作，在目的端恢复文件。

- **生成文件MD5值**

  对每个传输的文件都生成一个MD5值，并将该值记录在一个新文件中，新文件以“.md5”作为后缀，并且可以指定MD5值生成的目录。

文件格式的公共参数

- **启动作业标识文件**

  这个主要用于自动化场景中，CDM配置了定时任务，周期去读取源端文件，但此时源端的文件正在生成中，CDM此时读取会造成重复写入或者是读取失败。所以，可以在源端作业参数中指定启动作业标识文件为“ok.txt”，在源端生成文件成功后，再在文件目录下生成“ok.txt”，这样CDM就能读取到完整的文件。

  另外，可以设置超时时间，在超时时间内，CDM会周期去查询标识文件是否存在，超时后标识文件还不存在的话，则作业任务失败。

  启动作业标识文件本身不会被迁移。

- **作业成功标识文件**

  文件系统为目的端的时候，当任务成功时，在目的端的目录下，生成一个空的文件，标识文件名由用户来指定。一般和“启动作业标识文件”搭配使用。

  这里需要注意的是，不要和传输的文件混淆，例如传输文件为finish.txt，但如果作业成功标识文件也设置为finish.txt，这样会造成这两个文件相互覆盖。

- **过滤器**

  使用CDM迁移文件的时候，可以使用过滤器来过滤文件。支持通过通配符或时间过滤器来过滤文件。

  - 选择通配符时，CDM只迁移满足过滤条件的目录或文件。
  - 选择时间过滤器时，只有文件的修改时间晚于输入的时间才会被传输。

  例如：用户的“/table/”目录下存储了很多数据表的目录，并且按天进行了划分：DRIVING_BEHAVIOR_20180101～DRIVING_BEHAVIOR_20180630，保存了DRIVING_BEHAVIOR从1月到6月的所有数据。如果只想迁移DRIVING_BEHAVIOR的3月份的表数据。那么需要在作业第一步指定源目录为“/table”，过滤类型选择“通配符”，然后指定“路径过滤器”为“DRIVING_BEHAVIOR_201803*”。

文件格式问题解决方法

1. 数据库的数据导出到CSV文件，由于数据中含有分隔符逗号，造成导出的CSV文件中数据混乱。

   CDM提供了以下几种解决方法：

   a. 指定字段分隔符

      使用数据库中不存在的字符，或者是极少见的不可打印字符来作为字段分隔符。例如：可以在目的端指定“字段分隔符”为“%01”，这样导出的字段分隔符就是“\u0001”，详情可见**表1-8**。

   b. 使用包围符

      在目的端作业参数中开启“使用包围符”，这样数据库中如果字段包含了字段分隔符，在导出到CSV文件的时候，CDM会使用包围符将该字段括起来，使之作为一个字段的值写入CSV文件。

2. 数据库的数据包含换行符

   场景：使用CDM先将MySQL中的某张表（表的某个字段值中包含了换行符\n）导出到CSV格式的文件中，然后再使用CDM将导出的CSV文件导入到MRS HBase，发现导出的CSV文件中出现了数据被截断的情况。

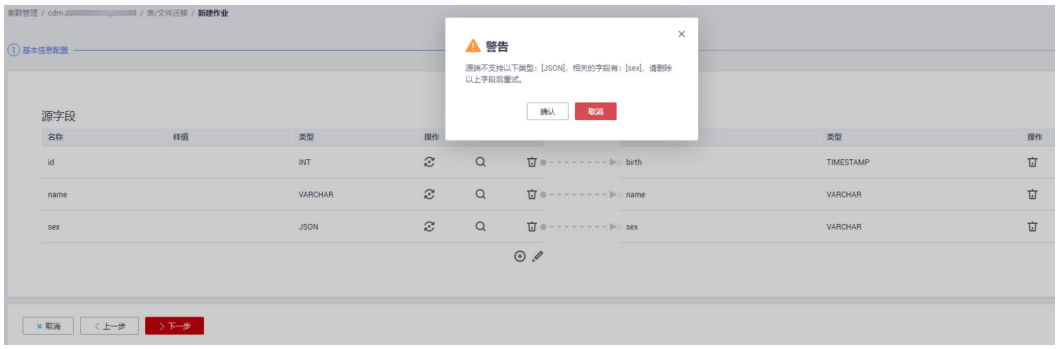
   解决方法：指定换行符。

   在使用CDM将MySQL的表数据导出到CSV文件时，指定目的端的换行符为“%01”（确保这个值不会出现在字段值中），这样导出的CSV文件中换行符就是“%01”。然后再使用CDM将CSV文件导入到MRS HBase时，指定源端的换行符为“%01”，这样就避免了数据被截断的问题。

1.12 不支持数据类型转换规避指导

操作场景

CDM在配置字段映射时提示字段的数据类型不支持，要求删除该字段。如果需要使用该字段，可在源端作业配置中使用SQL语句对字段类型进行转换，转换成CDM支持的类型，达到迁移数据的目的。



操作步骤

步骤1 修改CDM迁移作业，通过使用SQL语句的方式迁移。

说明

SQL语句格式为：“select id,cast(原字段名 as INT) as 新字段名可以和原字段名一样 from schemaName.tableName；”

例如：select `id`, `name`, cast(`sex` AS char(255)) AS `sex` from `test_1117869`.`test_no_support_type`;

步骤2 转换后的字段就转换为CDM支持的数据类型。

----结束

2 数据开发进阶实践

2.1 周期调度依赖策略

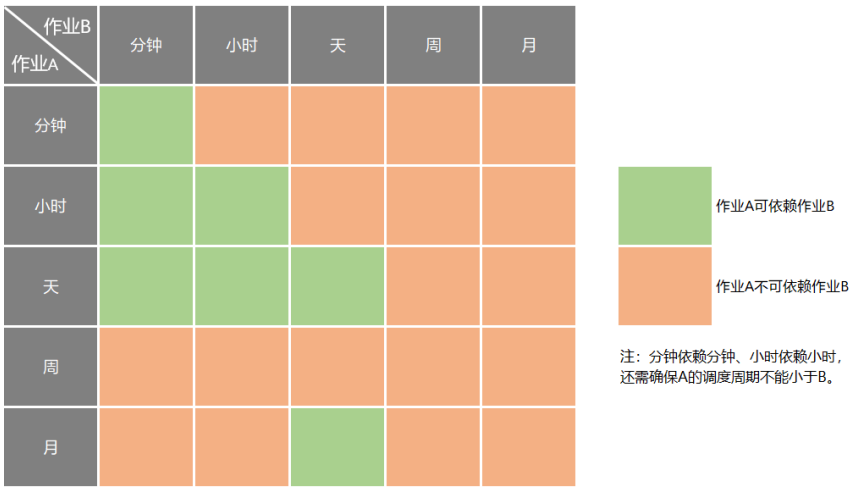
2.1.1 传统周期调度依赖和自然周期调度依赖对比

数据开发当前支持两种调度依赖策略：传统周期调度依赖和自然周期调度依赖。

传统周期调度依赖，只支持同周期或者大周期依赖于小周期，不支持小周期依赖于大周期。详细说明如下：

- 同周期依赖，依赖时间段范围为从当前批次时间往前推一个周期。
- 跨周期依赖，依赖时间段范围为上一个周期时间段内。

图 2-1 传统周期作业依赖关系全景图

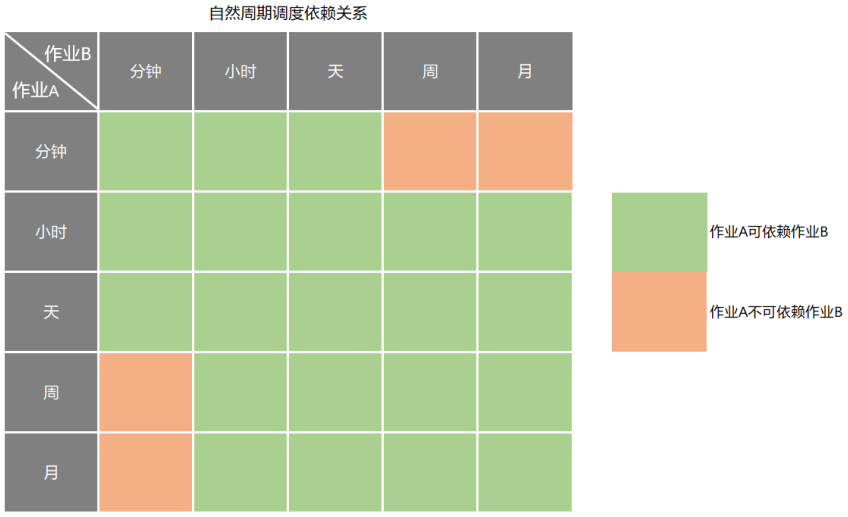


自然周期调度依赖，支持同周期、跨周期（大周期依赖于小周期、小周期依赖于大周期等）调度周期依赖，对于作业依赖来说，比较灵活，能够满足用户的复杂业务场景。详细依赖推断规则说明如下：

- 规则一：天、小时任务按自然天、自然小时，推断出任务的依赖关系。
- 规则二：周、月任务按当日自然天，推断出任务的依赖关系。

- 规则三：大周期依赖小周期，如，天任务依赖小时任务，只依赖小周期最后一个任务成功与否。
  自然天：[00:00:00-23:59:59]
  自然小时： [00:00-59:59]

图 2-2 自然周期调度作业依赖关系全景图



如何确认当前的周期调度依赖是传统周期调度依赖还是自然周期调度依赖?

创建了一个作业A，以小时进行调度，同时创建一个作业B，以天进行调度。

- 作业A在关联依赖作业时，如果可以选到依赖作业B，则是自然周期调度。（支持小周期依赖于大周期）
- 作业A在关联依赖作业时，如果不能选到依赖作业B，则是传统周期调度。（不支持小周期依赖于大周期）

图 2-3 作业依赖

2.1.2 传统周期调度

解释说明

周期调度作业支持设置调度周期符合条件的作业为依赖作业。设置依赖作业的操作详情请参考**配置作业调度任务（批处理作业）**章节。

例如周期调度作业A，可设置其依赖作业为作业B，如**图2-4**所示进行配置。则仅当其依赖的作业B在某段时间内所有实例运行完成、且不存在失败实例时，才开始执行作业A。

说明

- 依赖的作业B的“某段时间”，计算方法如下，详见后文**设置依赖作业后的作业运行原理**。
  - 同周期依赖，如分钟依赖分钟、小时依赖小时或天依赖天时，“某段时间”为 **(作业A执行时间-作业A周期时间, 作业A执行时间]** 。
  - 跨周期依赖：如小时依赖分钟、天依赖分钟、天依赖小时或月依赖天时，“某段时间”为 **[上一作业A调度周期的自然起点, 当前作业A调度周期的自然起点)** 。
- 作业A是否判断其依赖的作业B的实例状态，与“依赖的作业失败后，当前作业处理策略”参数有关，具体如下：
  - “依赖的作业失败后，当前作业处理策略”参数配置为“挂起”或“取消执行”后，当其依赖的作业B在某段时间内存在运行失败实例，则作业A“挂起”或“取消执行”。
  - “依赖的作业失败后，当前作业处理策略”参数配置为“继续执行”，只要其依赖的作业B在某段时间内所有实例跑完（不判断其状态），则作业A就继续执行。

图 2-4 作业依赖属性

本章节主要介绍**设置依赖作业的条件**，以及**设置依赖作业后的作业运行原理**。

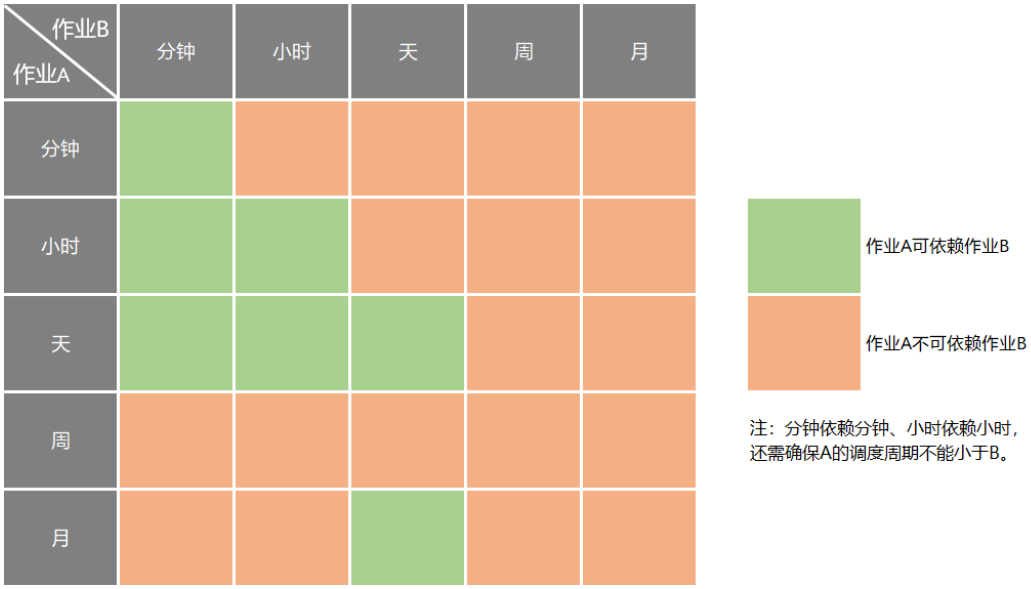
设置依赖作业的条件

当前周期调度作业的调度周期包括分钟、小时、天、周、月这五种周期，周期调度作业A如果要配置依赖作业为周期调度作业B，则调度周期必须符合以下要求：

- 作业A的调度周期不能比依赖作业B小。例如，作业A和作业B同为分钟/小时调度，A的间隔时间小于B的间隔时间，则作业A不能设置作业B为依赖作业；作业A为分钟调度，作业B为小时调度，则作业A不能设置作业B为依赖作业。
- 作业A和依赖作业B中不能有任一调度周期为周。例如，作业A的调度周期为周或作业B的调度周期为周，则作业A不能设置作业B为依赖作业。
- 调度周期为月的作业只能依赖调度周期为天的作业。例如，作业A的调度周期为月，则作业A只能设置调度周期为天的作业为依赖作业。

不同调度周期的作业，其允许配置的依赖作业调度周期总结如图2-5所示。

图 2-5 作业依赖关系全景图



设置依赖作业后的作业运行原理

同周期依赖和跨周期依赖的作业运行原理有所差异。为方便说明，本例中假设“依赖的作业失败后，当前作业处理策略”参数设置为“继续执行”，作业A不判断作业B的实例运行状态；如果该参数设置为“挂起”或“取消执行”，则作业A还会额外判断作业B的实例中是否存在失败实例。

- **同周期依赖**：即作业A与其依赖作业B为相同调度周期，如分钟依赖分钟、小时依赖小时或天依赖天。

  同周期依赖的情况下，当作业A的依赖作业配置为作业B后，作业A会在 **(作业A执行时间-作业A周期时间, 作业A执行时间]** 时间区间内检查是否有作业B的实例运行，只有在此期间作业B的实例运行完成才会运行作业A。

  示例1：作业A依赖作业B，均为分钟调度。作业A的开始时间10:00，周期时间20分钟；作业B的开始时间10:00，周期时间10分钟。则会出现如下情况：

表 2-1 示例 1：同周期作业依赖情况

时间点	作业B（分钟调度，开始时间10:00，周期时间10分钟）	作业A（分钟调度，开始时间10:00，周期时间20分钟）
10:00	执行	检查 **(09:40, 10:00]** 区间，有作业B实例运行，待作业B执行完成后，执行作业A
10:10	执行	-
10:20	执行	检查 **(10:00, 10:20]** 区间，有作业B实例运行，待作业B执行完成后，执行作业A
10:30	执行	-
...	...	...

示例2：作业A依赖作业B，均为天调度。作业A的开始时间为8月1日09:00；作业B的开始时间8月1日10:00。则会出现如下情况：

表 2-2 示例 2：同周期作业依赖情况

时间点	作业B（天调度，开始时间为8月1日10:00）	作业A（天调度，开始时间8月1日09:00）
8月1日 09:00	-	检查 **(7月31日09:00, 8月1日09:00]** 区间，无作业B实例运行，不执行作业A
8月1日 10:00	执行	-
8月2日 09:00	-	检查 **(8月1日09:00, 8月2日09:00]** 区间，有作业B实例运行，待作业B执行完成后，执行作业A
8月2日 10:00	执行	-
...	...	...

说明

天作业依赖天作业，上游作业调度时间早于下游作业，下游作业才能依赖到上游当天的作业。

- **跨周期依赖**：即作业A与其依赖作业B为不同调度周期，如小时依赖分钟、天依赖分钟、天依赖小时或月依赖天。

跨周期依赖的情况下，当作业A的依赖作业配置为作业B后，作业A会在 **[上一作业A调度周期的自然起点, 当前作业A调度周期的自然起点)** 时间区间内检查是否有作业B的实例运行，只有在此期间作业B的实例运行完成才会运行作业A。

说明

调度周期的自然起点定义如下：

- 调度周期为小时：**上一调度周期的自然起点**为上一小时的零分零秒，**当前调度周期的自然起点**为当前小时的零分零秒。
- 调度周期为天：**上一调度周期的自然起点**为昨天的零点零分零秒，**当前调度周期的自然起点**为今天的零点零分零秒。
- 调度周期为月：**上一调度周期的自然起点**为上个月1号的零点零分零秒，**当前调度周期的自然起点**为当月1号的零点零分零秒。

示例3：作业A依赖作业B，作业A为天调度，作业B为小时调度。作业A的每天02:00执行；作业B的开始时间00:00，间隔时间10小时。则会出现如下情况：

表 2-3 示例 3：跨周期作业依赖情况

时间点	作业B（小时调度，开始时间00:00，间隔时间10小时）	作业A（天调度，每天02:00执行）
第1天 00:00	执行	-
第1天 02:00	-	检查 **[第0天00:00:00, 第1天00:00:00)** 区间，无作业B实例运行，不执行
第1天 10:00	执行	-
第1天 20:00	执行	-
第2天 00:00	执行	-
第2天 02:00	-	检查 **[第1天00:00:00, 第2天00:00:00)** 区间，有作业B实例运行完成，执行作业A
第2天 10:00	执行	-
第2天 20:00	执行	-
...	...	...

示例4：作业A依赖作业B，作业A为月调度，作业B为天调度。作业A的每月1号、2号的02:00执行；作业B在8月1日00:00开始执行。则会出现如下情况：

表 2-4 示例 4：跨周期作业依赖情况

时间点	作业B（天调度，8月1日00:00执行）	作业A（月调度，每月1号、2号的02:00执行）
8月1日00:00	执行	-
8月1日02:00	-	检查 **[7月1日00:00:00, 8月1日00:00:00)** 区间，无作业B实例运行，不执行
8月2日00:00	执行	-
8月2日02:00	-	检查 **[7月1日00:00:00, 8月1日00:00:00)** 区间，无作业B实例运行，不执行
...	-	...
9月1日00:00	执行	-
9月1日02:00	-	检查 **[8月1日00:00:00, 9月1日00:00:00)** 区间，有作业B实例运行完成，执行作业A
9月2日00:00	执行	-
9月2日02:00	-	检查 **[8月1日00:00:00, 9月1日00:00:00)** 区间，有作业B实例运行完成，执行作业A
...	...	...

2.1.3 自然周期调度

解释说明

DataArts Studio支持自然周期的调度方式。通过各个节点的调度依赖配置结果，有序的运行业务流程中各个节点，保障业务数据有效、适时地产出。

调度依赖就是节点间的上下游依赖关系，在DataArts Studio中，上游任务节点运行完成且运行成功，下游任务节点才会开始运行。

配置调度依赖后，可以保障调度任务在运行时能取到正确的数据（当前节点依赖的上游节点成功运行后，DataArts Studio通过节点运行的状态识别到上游表的最新数据已产生，下游节点再去取数）。避免下游节点取数据时，上游表数据还未正常产出，导致下游节点取数出现问题。

在配置依赖关系时，支持配置同周期的依赖和上一周期的依赖。

同周期依赖的原理，详情参考**自然周期调度之同周期依赖原理**。

上一周期依赖的原理，详情参考**自然周期调度之上一周期依赖原理**。

说明

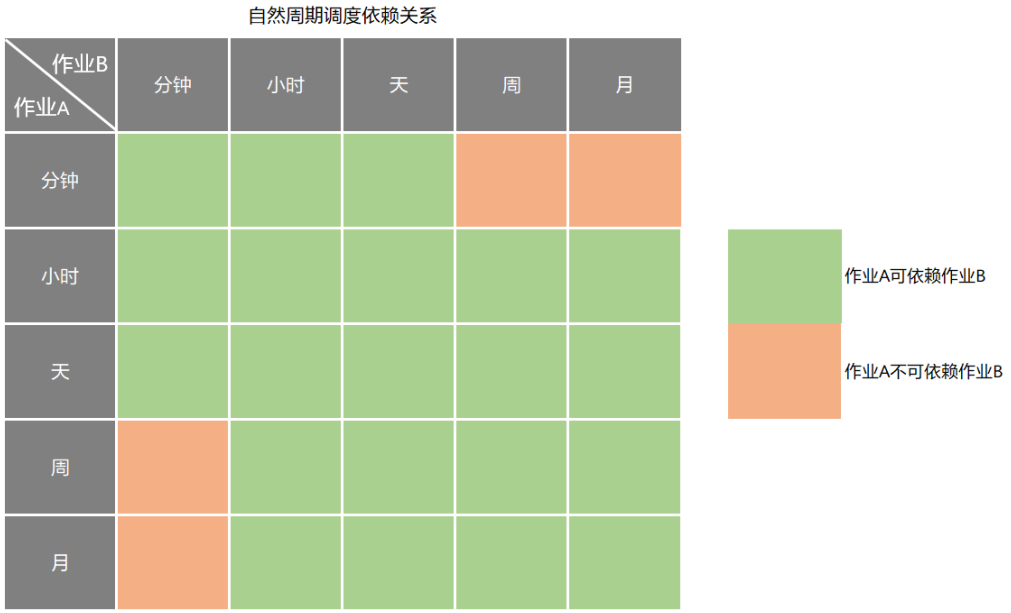
当分钟的调度周期不能被小时整除时，周期调度就不是严格按照间隔周期去跑，而是按照cron表达式的规则，每个小时的零点触发去跑，再往后推间隔。

2.1.4 自然周期调度之同周期依赖原理

解释说明

即作业A依赖于作业B的相同调度周期的运行实例。周期单位包括分钟、小时、天、周、月这五种，不同调度周期的作业，其允许配置的依赖作业调度周期总结如图2-6所示。

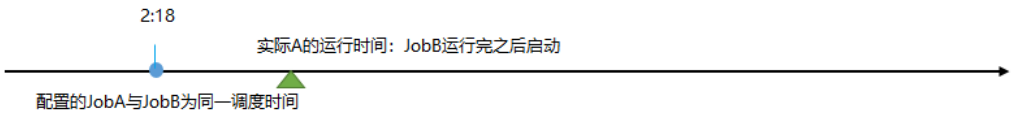
图 2-6 同周期作业依赖关系全景图



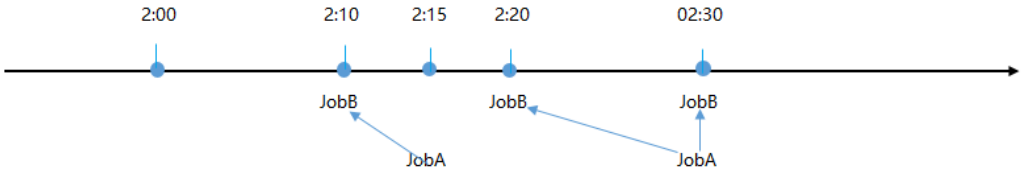
分钟依赖分钟

规则：分钟是最小调度粒度，没有自然分钟周期的概念，依赖策略是往前推一个调度周期找依赖实例。

举例1：A依赖B，为同周期分钟作业，在同一时间点，B执行完后开始执行A。



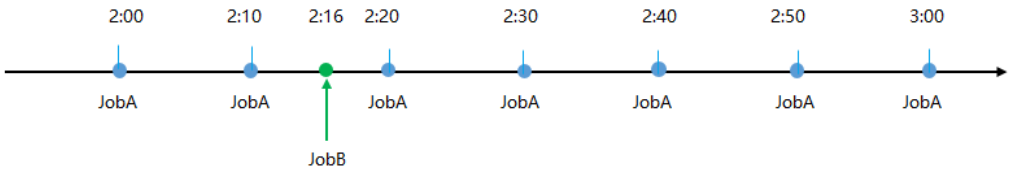
举例2：A依赖B，A为15分钟周期，B为10分钟周期，A往前推15分钟（包括当前启动整点），依赖范围内的B实例，在2:15分执行A任务依赖1个B实例（2:10分），2:30执行的A任务依赖两个B实例（2:20和2:30）。它的边界范围为(0分:15分]，前开后闭区间。



分钟依赖小时

规则：分钟级作业依赖自然小时的上一周期作业执行完成后，再执行。

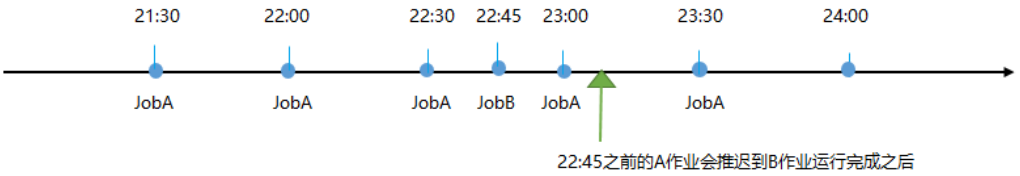
举例：A依赖B，A分钟级作业依赖B小时级作业，A每10分钟触发，B是每小时第16分钟执行，那么作业A实例会在B作业上一周期执行完成后再执行。



分钟依赖天

规则：分钟作业依赖自然天的作业，需等天作业执行完成后再执行。

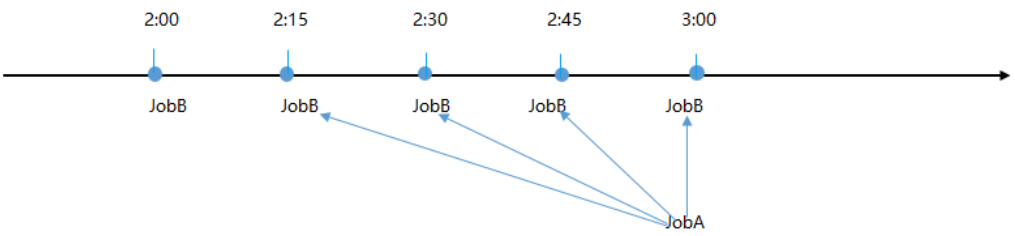
例如：A依赖B，A分钟作业依赖B天作业，A每30分钟执行，B是22:45执行，那么22:45之前的A作业实例都会推迟到B作业执行完成后再执行。



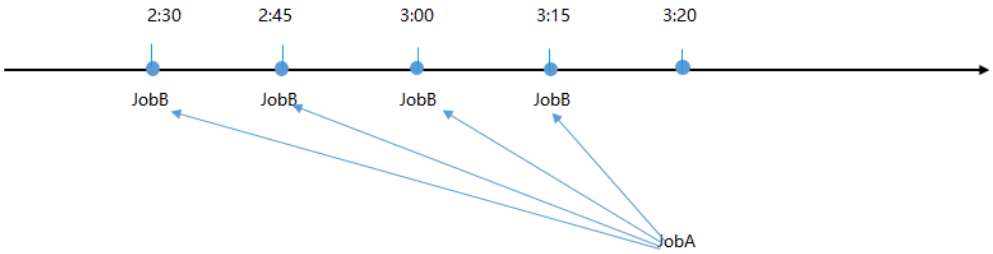
小时依赖分钟

规则：小时作业依赖分钟作业，往前推到上一个自然小时范围内的所有分钟级实例。区间是前开后闭。

举例1：A依赖B，A为小时作业，每个小时0分执行，B为15m分钟作业；B执行完后执行A。



举例2：A依赖B，A为小时作业，启动时间3:20，B为15m作业，会依赖往前推一个小时内的所有B实例。



如果勾选“最近”的按钮，小时作业只依赖所选作业最近的一个运行实例，比如A在3:20开始调度，A依赖B最近的3:00调度的一个运行实例。

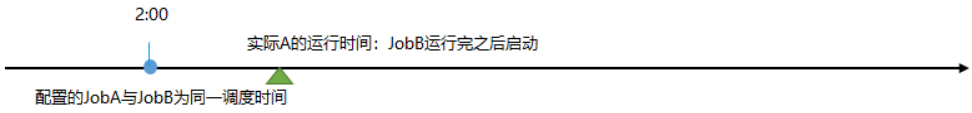
说明

如果作业A在零点进行调度，所依赖作业B可以是昨天的分钟任务。

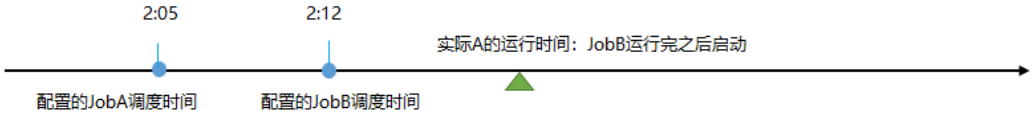
小时依赖小时

规则：每个自然小时周期内的实例产生依赖，区间边界是自然小时[00:00, 00:59]。

举例1：A依赖B，在同自然小时内，无论A、B设置在什么时间点执行，A永远在B之后执行。



举例2：A依赖B，A在每小时5分0秒执行，B在12分执行，A会等B执行完成后执行。



说明

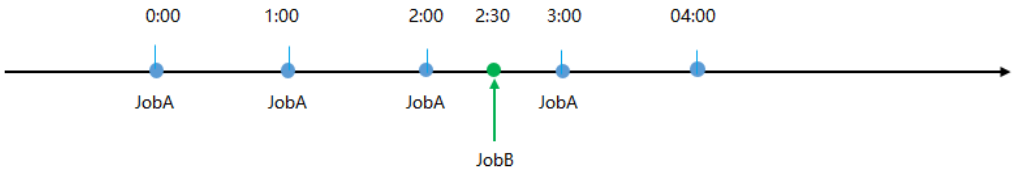
离散小时依赖离散小时：

- 自然天内，依赖关系中的上游、下游任务数量一致，上下游周期数一致。
- 自然天内，上下游任务数量不一致，下游任务运行当天生成的周期实例，将根据就近原则挂载依赖，依赖距离自己定时运行时间最近的上游实例。从index向前找上游依赖实例，依赖上游一整个区间内的实例；向前未找到依赖的实例时，需要向后找，向后查找时，只依赖最近的一个实例。

小时依赖天

规则：小时作业依赖自然天的作业，需等天作业执行完成后再执行。

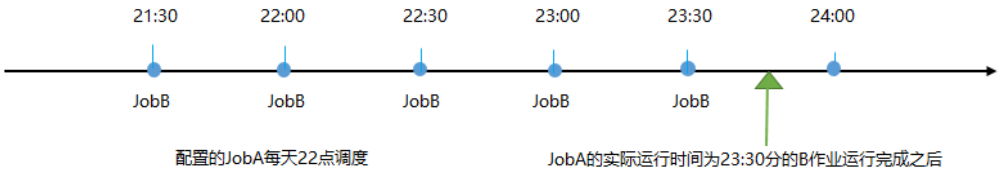
举例：A依赖B，A小时作业依赖B天作业，A每小时整点指定，B是2:30指定，那么2:30之前的A作业实例都会推迟到B作业执行完成后再执行。



天依赖分钟

规则：按自然天，天周期作业实例依赖一天内所有分钟级作业的实例。

举例：A依赖B，A为天作业，每天22点调度，依赖B分钟作业，每30分钟调度一次，A依赖所有B在自然天内的实例， A会在最后一个B作业实例执行完成后执行。

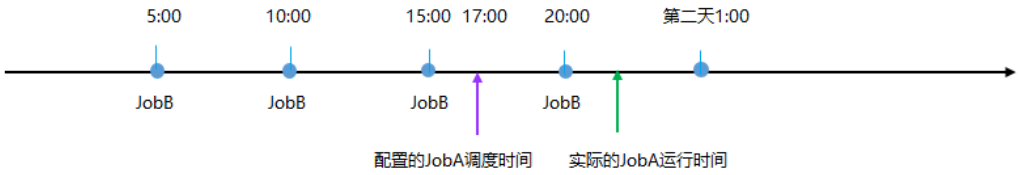


如果勾选“最近”的按钮，天作业只依赖所选作业最近的一个运行实例，比如A在每天22点开始调度，A依赖B最近的21:30调度的一个运行实例。

天依赖小时

规则：按自然天，天周期作业实例依赖一天内所有小时作业的实例。A为天作业，依赖B小时作业，A依赖所有B在自然天内的实例，A会在最后一个B小时作业实例执行完成后执行。

举例：A依赖B，A配置的调度时间为每天17点执行一次，B从0点开始，每5个小时执行一次，那么A实际执行时间为JobB在20点的实例运行完之后开始运行。

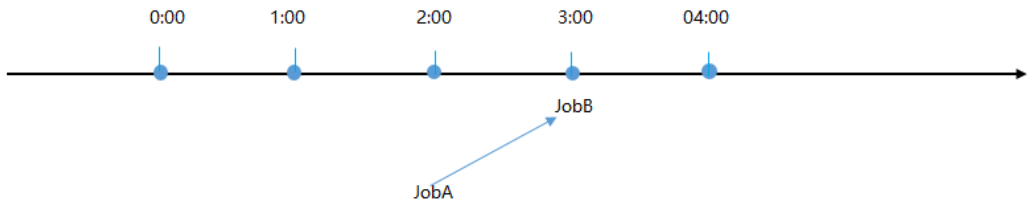


如果勾选“最近”的按钮，天作业只依赖所选作业最近的一个运行实例，比如A在每天17点开始调度，A依赖B最近的15:00调度的一个运行实例。

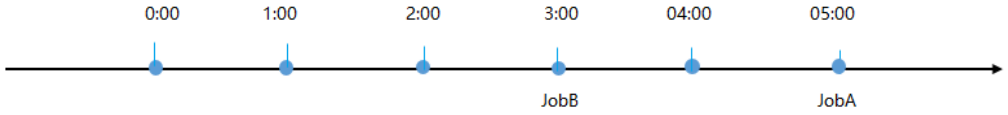
天依赖天

规则：按自然天内的实例进行依赖，不会跨天向前推找依赖实例。在同自然天内A依赖B ，无论A、B设置在什么时间点执行，A永远在B之后执行。 天区间为[00:00:00, 23:59:59]

举例1：A依赖B，A在2:00执行，B在3:00执行，A会等B在3:00执行完成后执行。



举例2：A依赖B，A在5:00执行，B在3:00执行，A在B执行完成后，在5:00执行。



天依赖周

规则：依赖自然天。

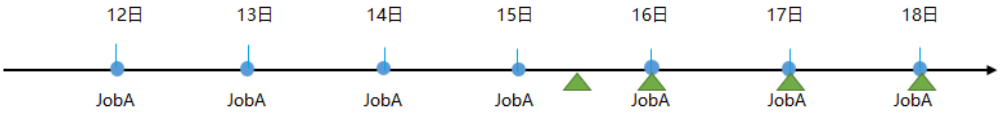
举例1：A依赖B，A作业每天执行，B作业每周三执行。在A作业执行时，B作业当天非周三，未执行，A作业则直接执行。

举例2：A依赖B，A作业每天执行，B作业每周三执行。在A作业执行时，B作业正好当天是周三，会执行，则A作业等待B作业执行完成后，开始执行。

天依赖月

规则：天作业依赖自然月的作业，需等月作业执行完成后再执行。

举例：A依赖B，A为天作业每天执行一次，B为月作业每月15号执行一次。A实际会在每月15号B作业执行完成后执行。



周依赖小时

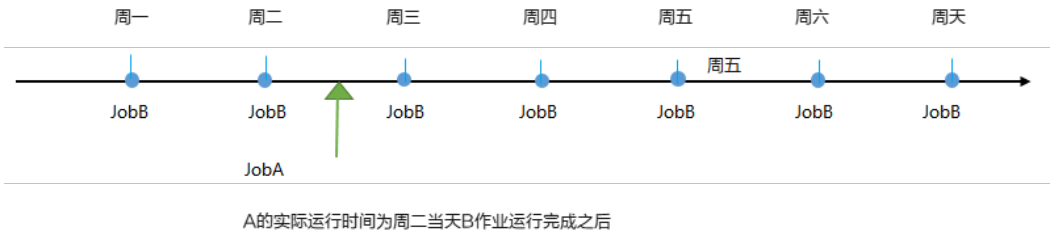
规则：依赖自然天，范围是[前一天的零点，当天的零点)，查找当天的小时任务作业B是否全部执行完成，然后执行周任务作业A。

举例：A依赖B，A作业每周一调度，B作业每小时第50分钟执行。则A作业会一直等待B作业执行，一直到B作业周一最后一个任务23:50分的任务执行完成后，开始执行A作业。

周依赖天

规则：周作业只依赖同一天调度执行的作业。

举例：A依赖B，A作业计划周二执行，B作业每天运行，A周二的作业会在周二的B作业执行完后再执行。

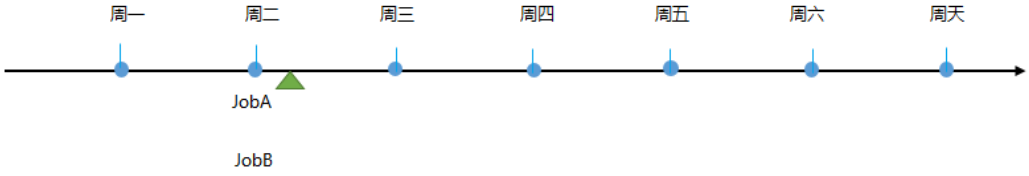


周依赖周

规则：周作业只依赖同一天调度执行的作业实例。

举例1：A依赖B，A作业计划周二执行；B作业计划周三执行；作业A依赖作业B，实际上A会在周二执行，不会等到周三B执行完。

举例2：A、B作业都是周二执行，A依赖B，A会等B执行完执行。



周依赖月

规则：依赖自然天

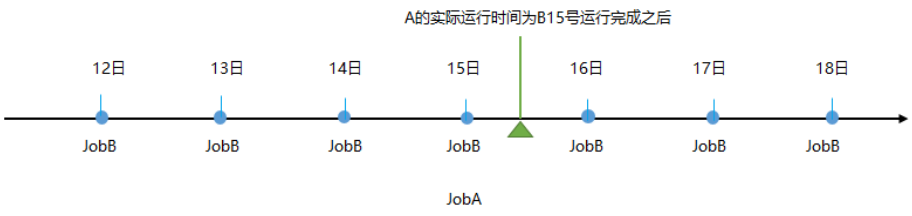
举例1：A依赖B，A作业每周三执行，B作业每月10号执行。在A作业执行时，如果正好是10号，A作业会等待B作业执行完成后执行。

举例2：A依赖B，A作业每周三执行，B作业每月10号执行。在A作业执行时，如果不是10号，则A作业直接执行。

月依赖天

规则：A依赖B，月调度任务，只依赖于当前的天任务完成后，即可运行。

举例：A依赖B，A为月作业，依赖B天作业， A依赖所有B在月任务当前的那个实例，A就可运行。



月依赖周

规则：依赖自然天

举例1：A依赖B，A作业每月10号执行，B作业每周三执行。在A作业执行时，B作业当天非周三，未执行，A作业则直接执行。

举例2：A依赖B，A作业每月10号执行，B作业每周三执行。在A作业执行时，B作业当天正好是周三，则A作业等待B作业执行完成后开始执行。

月依赖月

规则：依赖自然天

举例1：A依赖B，A作业每月1号执行，B作业每月2号执行，A作业1号正常执行，B作业不阻塞A作业执行。

举例2：A依赖B，A作业和B作业都是2号执行，A作业会等待B作业执行完成后开始执行。

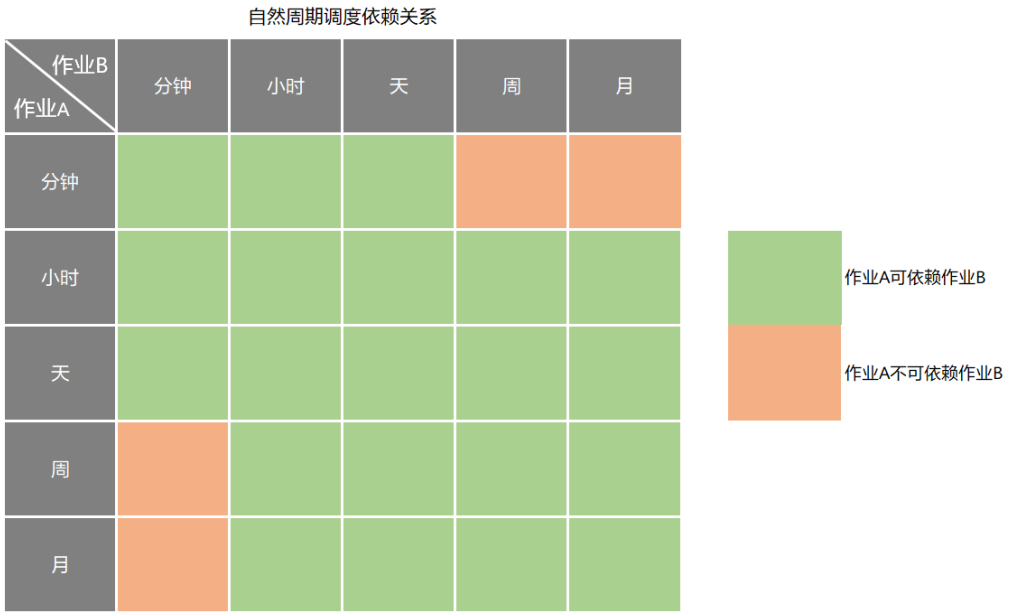
举例3：A依赖B，A作业每月3号执行，B作业每月2号执行，3号A作业依赖2号B作业。

2.1.5 自然周期调度之上一周期依赖原理

自然周期调度的概念

自然周期调度作业的调度周期包括分钟、小时、天、周、月这五种周期，不同调度周期的作业，其允许配置的依赖作业调度周期总结如图2-7所示。

图 2-7 上一周期作业依赖关系全景图



即作业A的调度依赖于作业B的上一调度周期，包含以下场景：

分钟依赖分钟

规则：分钟是最小调度粒度，没有自然分钟周期的概念，依赖策略是根据调度周期长的作业，往前推一个调度周期找依赖实例。

例如：A依赖B，A和B都是从每小时的0分开始，A每隔10分钟运行一次，B每隔15分钟运行一次，A的实际运行实际是上一小时的45分B作业运行完成之后。

分钟依赖小时

规则：分钟级作业依赖上一个自然小时周期的作业执行完成后，再执行。

例如：A依赖B，A分钟级作业依赖B小时级作业，A每10分钟触发，B是每小时第18分钟执行，那么作业A实例会在该周期的第10分钟运行。

分钟依赖天

规则：分钟作业依赖自然天作业的前一个周期，需等上一个天周期作业执行完成后再执行。

例如：A依赖B，A分钟作业依赖B天作业，A每5分钟执行，B是2：30执行，那么A需要依赖前一天2:30之后B作业的实例。

小时依赖分钟

规则： 小时作业依赖分钟作业，往前推到上一个自然小时范围内的所有分钟级实例。区间是前开后闭。

举例1：A依赖B，A为小时作业，每个小时0分执行，B为15m作业；A会依赖上一小时B在45分钟时的作业实例。

举例2：A依赖B，A为小时作业，启动时间3:20， B为15m作业，A会依赖B在3点15时生成的作业实例。

小时依赖小时

规则：每个自然小时周期内的实例产生依赖，区间边界是自然小时[00:00, 00:59]，依赖策略是调度周期长的作业，往前推一个调度周期找依赖实例。

A依赖B，在同自然小时内，无论A、B设置在什么时间点执行，A永远在B的前一周期完成后执行。

举例：A在每小时5分0秒执行，B在12分执行，A会在每小时5分时依赖B上一小时生成的实例。

说明

离散小时依赖离散小时：

- 自然天内，依赖关系中的上游、下游任务数量一致，上下游周期数一致。
- 自然天内，上下游任务数量不一致，下游任务运行当天生成的周期实例，将根据就近原则挂载依赖，依赖距离自己定时运行时间最近的上游实例。从index向前找上游依赖实例，依赖上游一整个区间内的实例；向前未找到依赖的实例时，需要向后找，向后查找时，只依赖最近的一个实例。

小时依赖天

规则：小时作业依赖自然天的作业，需等天作业的前一周期执行完成后再执行。

举例：A依赖B，A小时作业依赖B天作业，A每小时整点指定，B是2:30指定，那么A作业执行会依赖前一天2点30分B作业的运行实例。

天依赖分钟

规则：按自然天，天周期作业实例依赖前一天内所有分钟级作业的实例。

举例：A依赖B，A为天作业，依赖B分钟作业， A依赖B在当天内的最后一个实例， A会在最后一个B作业实例执行完成后执行。

天依赖小时

规则：按自然天，天周期作业实例依赖前一天内最后一周期小时作业的实例。

举例：A依赖B，A为天作业，依赖B小时作业，A依赖B在前一天最后一个周期的小时作业实例执。

天依赖天

规则：按自然天的上一个周期实例进行依赖。 在同自然天内A依赖B ，无论A、B设置在什么时间点执行，A永远依赖B的前一周期实例执行。 天区间为[00:00:00, 23:59:59]

举例：A在2:00执行，B在3:00执行，A会依赖B在前一个周期3:00执行的实例，在当前周期2点执行。

天依赖周

规则：依赖自然天。

举例1：A依赖B，A作业每天执行，B作业每周三执行。在A作业执行时，前一天非周三，B作业未执行，A作业则直接执行。

举例2：A依赖B，A作业每天执行，B作业每周三执行。在A作业执行时，前一天是周三，B作业会执行，则A作业等待B作业执行完成后，开始执行。

天依赖月

规则：天作业依赖自然月的作业，需等月作业执行完成上一周期后再执行。

例如：A依赖B，A为天作业每天执行一次，B为月作业每月15号执行一次。A作业的执行依赖B作业上个月15号的运行实例。

周依赖小时

规则：依赖自然天，范围是[前一天的零点，当天的零点)，查找当天的小时任务作业B是否全部执行完成，然后执行周任务作业A。

举例：A依赖B，A作业每周一调度，B作业每小时第50分钟执行。则A作业会一直等待B作业执行，一直到B作业上周日最后一个任务23:50分的任务执行完成后，开始执行A作业。

周依赖天

规则：周作业依赖前一天的天作业，需等前一天作业完成后再执行。

例如：A依赖B，B为天作业每天执行一次，A为周作业每周一执行一次。A作业的执行依赖B作业前一天的运行实例。

周依赖周

规则：周作业依赖前一天的周作业，需等前一天周作业完成后再执行，如果前一天没有实例，则不需要依赖。

例如：A依赖B，B为周作业每周一执行一次，A为周作业每周二执行一次。A作业的执行依赖B作业周一的运行实例。

周依赖月

规则：依赖自然天

举例1：A依赖B，A作业每周三执行，B作业每月10号执行。在A作业执行时，如果前一天正好是10号，A作业会等待B作业执行完成后执行。

举例2：A依赖B，A作业每周三执行，B作业每月10号执行。在A作业执行时，如果前一天不是10号，则A作业直接执行。

月依赖天

规则：月作业依赖前一天的天作业，需等前一天作业完成后再执行。

例如：A依赖B，B为天作业每天执行一次，A为月作业每月一执行一次。A作业的执行依赖B作业前一天的运行实例。

月依赖周

规则：依赖自然天

举例1：A依赖B，A作业每月10号执行，B作业每周三执行。在A作业执行时，前一天非周三，B作业未执行，A作业则直接执行。

举例2：A依赖B，A作业每月10号执行，B作业每周三执行。在A作业执行时，前一天正好是周三，B作业会执行，则A作业等待B作业执行完成后开始执行。

月依赖月

规则：依赖自然天

举例1：A依赖B，A作业每月1号执行，B作业每月2号执行，A作业1号正常执行，B作业不阻塞A作业执行。

举例2：A依赖B，A作业每月3号执行，B作业每月2号执行，A作业需要等待2号的B作业运行完成后开始执行。

举例3：A依赖B，A作业每月3号执行，B作业每月2号执行，3号A作业依赖2号B作业。

2.2 补数据场景使用介绍

适用场景

在某项目搬迁场景下，当您需要补充以前时间段内的历史业务数据，需要查看历史数据的详细信息时，可以使用补数据特性。

补数据是指作业执行一个调度任务，在过去某一段时间里生成一系列的实例。用户可以通过补数据，修正历史中出现数据错误的作业实例，或者构建更多的作业记录以便调试程序等。

说明

- 补数据作业除了支持SQL脚本，其他节点也支持。
- 如果SQL脚本的内容有变化，补数据作业运行的是最新版本的脚本。
- 使用补数据功能时，如SQL中变量是DATE，脚本中就写${DATE}，在作业参数中会自动增加脚本参数DATE，脚本参数DATE的值支持使用EL表达式。如果是变量时间的话，需要使用DateUtil内嵌对象的表达式，平台会自动转换成历史日期。EL表达式用法可参考**EL表达式**。
- 补数据作业除了支持作业参数，脚本参数或者全局环境变量也支持。

约束条件

- 只有数据开发作业配置了周期调度，才支持使用补数据功能。

使用案例

案例场景

在某企业的产品数据表中，有一个记录产品销售额的源数据表A，现在需要把产品销售额的历史数据导入的目的表B里面，需要您配置补数据作业的相关操作。

需要导入的列表情况如**表1**所示。

表 2-5 需要导入的列表情况

源数据表名	目的表名
A	B

配置方法

1. 准备源表和目的表。为了便于后续作业运行验证，需要先创建DWS源数据表和目的表，并给源数据表插入数据。
   a. 创建DWS表。您可以在DataArts Studio数据开发中，新建DWS SQL脚本执行以下SQL命令：
   ```
   /* 创建数据表 */
   CREATE TABLE A (PRODUCT_ID INT, SALES INT, DATE DATE);
   CREATE TABLE B (PRODUCT_ID INT, SALES INT, DATE DATE);
   ```
   b. 给源数据表插入示例数据。您可以在DataArts Studio数据开发模块中，新建DWS SQL脚本执行以下SQL命令：

```
/* 源数据表插入示例历史数据 */
INSERT INTO A VALUES ('1','60', '2022-03-01');
INSERT INTO A VALUES ('2','80', '2022-03-01');
INSERT INTO A VALUES ('1','50', '2022-02-28');
INSERT INTO A VALUES ('2','55', '2022-02-28');
INSERT INTO A VALUES ('1','60', '2022-02-27');
INSERT INTO A VALUES ('2','45', '2022-02-27');
```

2. 开发一个补数据的脚本。开发脚本时，脚本表达式里面必须包含时间变量（例如，SQL中变量是DATE，脚本中就写${DATE}）。在作业参数配置里面，您可以在3中编写脚本参数DATE的语句表达式。

   在“脚本开发”界面，在编辑器中输入开发语句。

```
INSERT INTO B (SELECT * FROM A WHERE DATE = ${DATE})
```

图 2-8 开发脚本

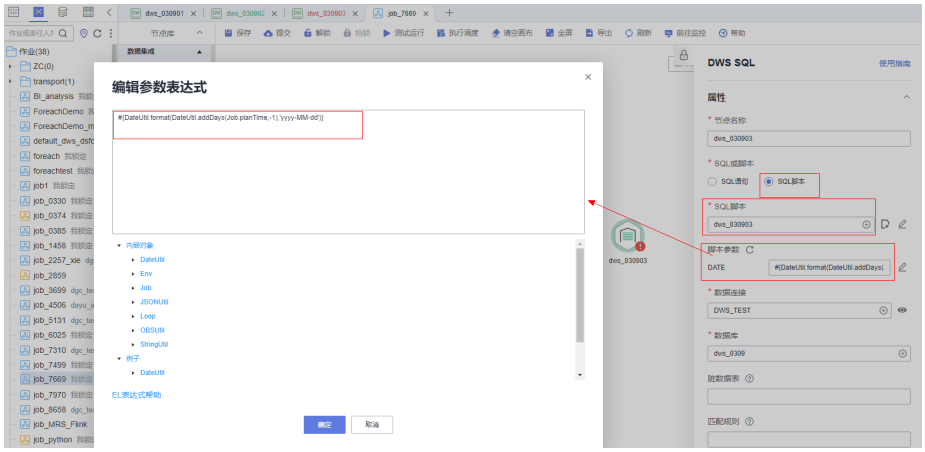
```
-- DWS sql
-- ********************************************************************** --
-- author: [redacted]
-- create time: 2023/05/23 17:03:02 GMT+08:00
-- ********************************************************************** --
INSERT INTO B (SELECT * FROM A WHERE DATE = ${DATE})
```

   脚本编写完成后，保存并提交此脚本的最新版本。

3. 开发一个补数据的批处理作业。开发作业时，您需要配置节点属性参数和调度周期。

   在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。

图 2-9 节点参数



说明

- 如果作业所关联的SQL脚本使用了参数，此处显示脚本参数名称（例如DATE），请在参数名称后的输入框配置参数值。参数值支持使用EL表达式，EL表达式用法可参考**EL表达式**。

  如果参数是时间的话，请您查看下DateUtil内嵌对象的表达式例子，平台会自动替换成补数据的历史日期（由补数据的业务日期所决定）。

  您也可以直接编写SQL语句，编写SQL表达式。

- 若关联的SQL脚本，脚本参数发生变化，可单击刷新按钮同步，也可以单击进行编辑。
- 脚本参数的举例如下所示。

  例如：#{DateUtil.format(DateUtil.addDays(Job.planTime,-1),'yyyy-MM-dd')}

  - Job.planTime表示作业计划时间，yyyy-MM-dd表示时间格式。
  - 如果作业计划时间三月二号，减去一天就是三月一号。补数据时，配置的补数据的业务日期就会替换作业计划时间。
  - Job.planTime会把作业计划时间通过表达式转化为yyyy-MM-dd格式的时间。

配置补数据作业的调度周期。单击界面右侧的调度配置，配置补数据作业的调度周期，该使用指导配置周期设置为天。

图 2-10 配置调度周期

说明

- 作业调度周期设置为天，每天会进行作业调度，并生成一个调度实例。您可以在“实例监控”页面中，查看补数据实例的运行状态。用户可以在该页面中查看作业的实例信息，并根据需要对实例进行更多操作。
- 该作业调度时间从2023/03/09开始生效，每天2点调度一次作业。
- 执行以下SQL命令，查询目的表B里面是否存在源表A的数据。

  SELECT * FROM B

参数配置完成后，保存并提交此作业的最新版本，测试运行该作业。

单击“执行调度”，让该作业运行起来。

4. 创建补数据。

   您在创建了一个周期调度作业后，用户需要为该任务进行补数据的操作。

a. 在数据开发主界面的左侧导航栏，选择“运维调度 > 作业监控”。

b. 单击“批作业监控”页签，进入批作业的监控页面。在该作业的“操作”列，选择“更多 > 补数据”。进入“补数据”页面。

如果您需要补充**2023-02-27至2023-03-01**之间的历史数据，补数据的业务日期需要设置为**2023-02-28至2023-03-02**，该业务日期系统会自动传给作业计划时间，脚本时间变量DATE的表达式中，定义的时间为作业计划时间减去一天，即作业计划时间的前一天时间为补数据的时间范围（**2023-02-27至2023-03-01**）。

图 2-11 补数据

表 2-6 参数说明

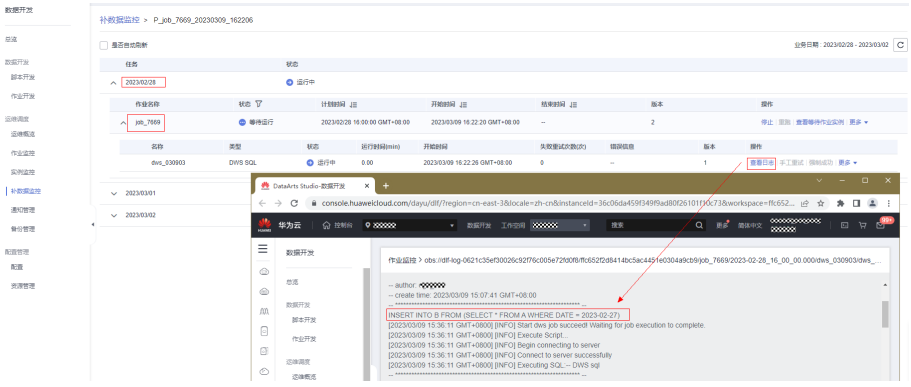
参数	说明
补数据名称	系统自动生成一个补数据的任务名称，允许修改。
作业名称	系统自动显示需要补数据的作业名称。

参数	说明
业务日期	选择需要补数据的时间段。这个业务日期会传递给作业的计划时间。作业运行时，作业计划时间就会被补数据里面的业务时间替换掉。 **说明** 一个作业可进行多次补数据。但多次补数据的业务日期需要避免交叉重叠，否则可能导致数据重复或混乱，用户请谨慎操作。 如果勾选了“按日期倒序补数据”，则系统按照日期倒序补跑，每日内的补数顺序仍是正序。 **说明** • 该功能适合在各日数据不耦合的条件下使用。 • 为保证补数可以倒序进行，补数据作业对更早日期作业实例的依赖关系将被忽略。
并行周期数	设置同时执行的实例数量，最多可同时执行5个实例。 **说明** 请根据实际情况配置并行周期数，例如CDM作业实例，不可同时执行补数据操作，并行周期数只可设置为1。
需要补数据的上下游作业	可选。选择需要补数据的下游作业（指依赖于当前作业的作业），支持多选。

c. 单击“确定”，系统会根据作业的调度周期开始补数据。

d. 在“补数据监控”页面中，查看补数据的任务状态、业务日期、并行周期数、补数据作业名称，以及停止运行中的任务，同时您可以查看补数据的详细信息。

图 2-12 补数据详细信息



e. 执行以下SQL命令，查询目的表B里面是否存在源表A的历史数据。

```
SELECT * FROM B
```

2.3 作业调度支持每月最后一天

场景描述

在配置作业调度时，可以选择每个月的最后一天执行。如果您需要配置作业的调度时间为每月最后一天，请参考下面两种方法。

表 2-7 配置每月最后一天进行调度

配置方法	优势	如何配置
调度周期配置为天，通过条件表达式进行判断是否为每月最后一天	可以灵活适用多种场景。只需要编写条件表达式就可以灵活调度作业去运行。例如，每月最后一天，每月七号等。	方法1
调度周期配置为月，勾选每月最后一天	通过配置调度周期来执行任务调度。不用编写开发语句，通过勾选需要调度的时间去执行任务。	方法2

方法 1

在DataArts Studio中配置一个每天调度执行的作业，然后在作业里面新增一个Dummy节点（空节点，不处理实际的业务数据），在Dummy节点与后续执行任务的节点的连线上，您可以配置条件表达式，判断当前是否为每个月的最后一天。如果是最后一天，则执行后续节点，否则跳过后续节点。

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
2. 任务配置为天调度，如下图：

图 2-13 调度周期配置为天

3. 在节点的连线上，单击右键，选择设置条件，配置条件表达式。通过表达式来判断，是否执行后续的业务节点。

图 2-14 设置条件表达式

4. 表达式配置方法如下所示。

```
#{DateUtil.getDay(DateUtil.addDays(Job.planTime,1)) == 1 ? "true" : "false"}
```

表达式的含义是：获取当前的时间点，往后推一天，判断是不是1号，如果是，则表明当前是每个月的最后一天，执行后续节点。如果不是，则跳过后续的业务节点。

图 2-15 条件表达式

如果用户的作业是每个月的最后一天执行，可以按照上面的方法进行配置。

如果用户的作业是每月7号执行，可以按照下面的方法进行配置。

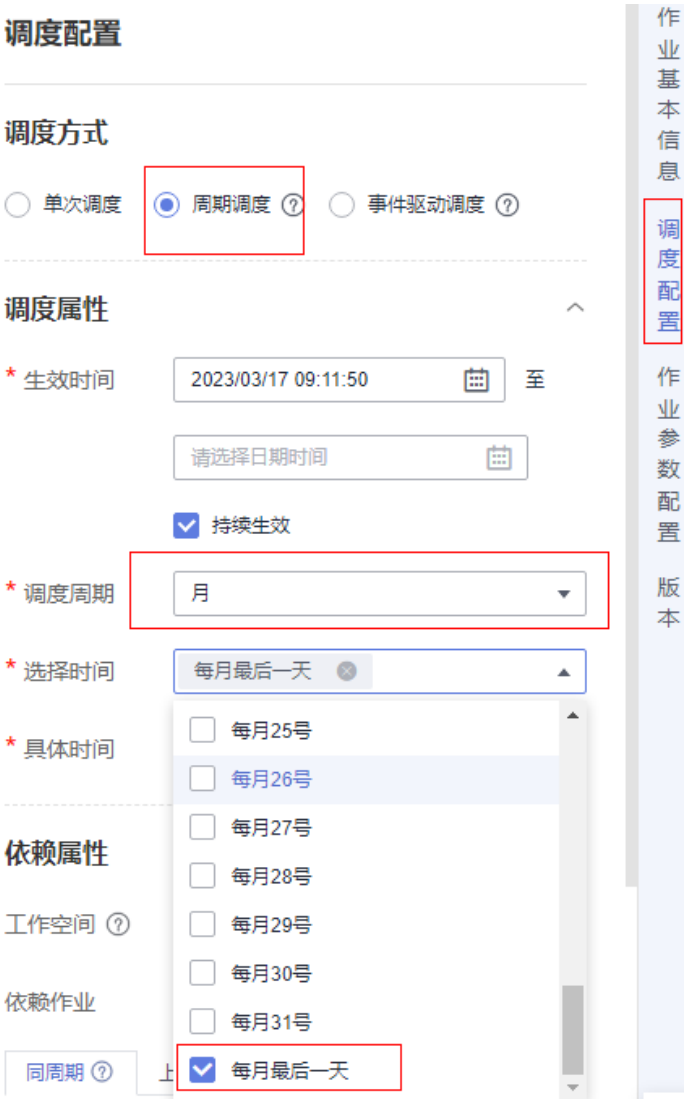
判断是否为7号，表达式配置方法如下所示。

```
#{DateUtil.getDay(Job.planTime) == 7 ? "true" : "false"}
```

方法 2

1. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
2. 单击作业画布右侧“调度配置”页签，进入调度配置页面。
3. 调度方式选择“周期调度”，调度周期选择“月”，选择时间为“每月最后一天”，如下图所示。

图 2-16 调度时间为每月最后一天



调度时间配置好之后，在每个月的最后一天，所配置的作业会按照调度时间去自动运行。

2.4 获取 SQL 节点的输出结果值

当您在数据开发模块进行作业开发，需要获取SQL节点的输出结果值，并将结果应用于后续作业节点或判断时，可参考本教程获取SQL节点的输出结果。

场景说明

使用EL表达式**#{Job.getNodeOutput("前一节点名")}**获取的前一节点的输出结果时，输出结果为二维数组形式，形如[["Dean",...,"08"],...,["Smith",...,"53"]]所示。为获取其中的值，本案例提供了如**表2-8**所示的两个常见方法示例。

表 2-8 获取结果值常见方法

方法	关键配置	适用场景要求
通过StringUtil提取输出结果值	当SQL节点的输出结果只有一个字段，形如[["11"]]所示时，可以通过StringUtil内嵌对象EL表达式分割二维数组，获取前一节点输出的字段值： #{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("*前一节点名*"),"]")[0],"[")[0],"\\"")[0]}	通过StringUtil提取输出结果值配置简单，但对适用场景有如下要求： • 前一SQL节点的输出结果只有一个字段，形如[["11"]]所示。 • 输出结果值数据类型为String，需要应用场景支持String数据类型。例如当需要使用IF条件判断输出结果值的数值大小时，不支持String类型，则不能使用本方法。
通过For Each节点提取输出结果值	通过For Each节点，循环获取数据集中二维数组的值： • For Each节点数据集：#{Job.getNodeOutput('*前一节点名*')} • For Each节点子作业参数：#{Loop.current[*索引*]}	通过For Each节点输出结果值适用场景更广泛，但需将作业拆分为主作业和子作业。

通过 StringUtil 提取输出结果值

场景说明

通过StringUtil内嵌对象EL表达式分割二维数组结果，获取前一节点输出的字段值，输出结果类型为String。

本例中，MRS Hive SQL节点返回单字段二维数组，Kafka Client节点发送的数据定义为StringUtil内嵌对象EL表达式，通过此表达式即可分割二维数组，获取MRS Hive SQL节点输出的字段值。

说明

为便于查看最终获得的结果值，本例选择Kafka Client节点进行演示。在实际使用中，您可以根据您的业务需求选择后续节点类型，在节点任务中应用StringUtil内嵌对象EL表达式，即可获取前一节点返回的数据值。

图 2-17 作业样例

其中，Kafka Client节点的关键配置为“发送数据”参数，取值如下：

```
#{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("count95"),"]")[0],"[")[0],"\\"")[0]}
```

配置方法

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 选择“空间管理”页签，在工作空间列表中，找到所需要的工作空间，单击工作空间的“数据开发”，系统跳转至数据开发页面。

步骤3 构造原始表格student_score。新建临时Hive SQL脚本，选择Hive连接和数据库后，粘贴如下SQL语句并运行，运行成功后即可删除此脚本。

```
CREATE TABLE `student_score` (`name` String COMMENT '', `score` INT COMMENT '');
INSERT INTO
    student_score
VALUES
    ('ZHAO', '90'),
    ('QIAN', '88'),
    ('SUN', '93'),
    ('LI', '94'),
    ('ZHOU', '85'),
    ('WU', '79'),
    ('ZHENG', '87'),
    ('WANG', '97'),
    ('FENG', '83'),
    ('CEHN', '99');
```

步骤4 新建MRS Hive SQL节点调用的Hive SQL脚本。新建Hive SQL脚本，选择Hive连接和数据库后，粘贴如下SQL语句并提交版本，脚本命名为count95。

```
--从student_score表中统计成绩在95分以上的人数--
SELECT count(*) FROM student_score WHERE score> "95" ;
```

步骤5 在“作业开发”页面，新建数据开发作业。选择一个MRS Hive SQL节点和一个Kafka Client节点，选中连线图标并拖动，编排如图2-17所示的作业。

步骤6 配置MRS Hive SQL节点参数。SQL脚本选择步骤4中提交的脚本count95，选择Hive连接和数据库。

图 2-18 配置 MRS Hive SQL 节点参数

步骤7 配置Kafka Client节点参数。发送数据定义为：#{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("count95"),"]")[0],"[")[0],"\\"")[0]}，选择Kafka连接和Topic名称。

图 2-19 配置 Kafka Client 节点参数

步骤8 作业节点配置完成后，选择测试运行。待作业测试运行成功后，在Kafka Client节点上右键查看日志，可以发现MRS Hive SQL节点返回的二维数组**[["2"]]**已被清洗为**2**。

说明

您可以将Kafka Client节点中的发送数据定义为**#{Job.getNodeOutput("count95")}**，然后作业运行后查看Kafka Client节点日志，则可以验证MRS Hive SQL节点返回的结果为二维数组**[["2"]]**。

图 2-20 查看 Kafka Client 节点日志

----结束

通过 For Each 节点提取输出结果值

场景说明

结合For Each节点及其支持的Loop内嵌对象EL表达式#{Loop.current[0]}，循环获取前一节点输出的结果值。

本例中，MRS Hive SQL节点返回多字段的二维数组，选择For Each节点和EL表达式**#{Loop.current[]}**，再通过For Each循环调用Kafka Client节点子作业，Kafka Client

节点发送的数据也定义为#{Loop.current[]}，通过此配置即可获取MRS Hive SQL节点输出的结果值。

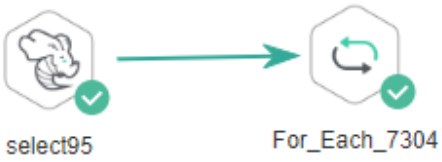
说明

为便于查看最终获得的结果值，本例中For Each节点子作业选择Kafka Client节点进行演示。在实际使用中，您可以根据您的业务需求选择子作业节点类型，在节点任务中应用Loop内嵌对象EL表达式，即可获取For Each前一节点返回的结果值。

For Each节点主作业编排如图2-21所示。其中，For Each节点的关键配置如下：

- 数据集：数据集就是HIVE SQL节点的Select语句的执行结果。使用EL表达式**#{Job.getNodeOutput("select95")}**，其中**select95**为前一个节点的名称。
- 子作业参数：子作业参数是子作业中定义的参数名，然后在主作业中定义的参数值，传递到子作业以供使用。此处子作业参数名定义为**name**和**score**，其值为分别为数据集中的第一列和第二列数值，使用EL表达式**#{Loop.current[0]}**和**#{Loop.current[1]}**。

图 2-21 主作业样例



而For Each节点中所选的子作业，则需要定义For Each节点中的子作业参数名，以便让主作业识别参数定义，作业如图2-22所示。

图 2-22 子作业样例

配置方法

开发子作业

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 选择“空间管理”页签，在工作空间列表中，找到所需要的工作空间，单击工作空间的“数据开发”，系统跳转至数据开发页面。

步骤3 在“作业开发”页面，新建数据开发子作业**EL_test_slave**。选择一个Kafka Client节点，并配置作业参数，编排图2-22所示的作业。

此处需将参数名填写为**name**和**score**，仅用于主作业的For Each节点识别子作业参数；参数值无需填写。

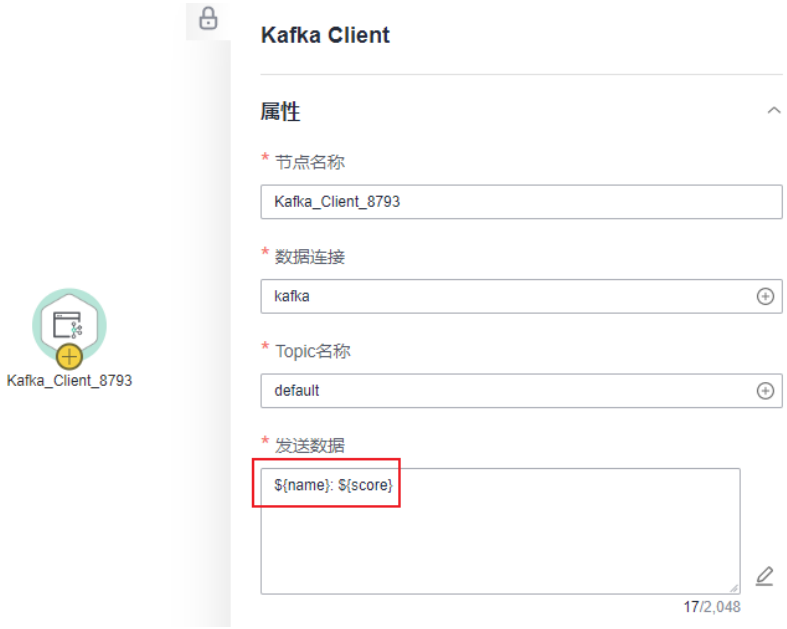
步骤4 配置Kafka Client节点参数。发送数据定义为：**${name}: ${score}**，选择Kafka连接和Topic名称。

说明

此处不能使用EL表达式#{Job.getParam("job_param_name")}，因为此表达式只能直接获取当前作业里配置的参数的value，并不能获取到父作业传递过来的参数值，也不能获取到工作空间里面配置的全局变量，作用域仅为本作业。

而表达式${job_param_name}，既可以获取到父作业传递过来的参数值，也可以获取到全局配置的变量。

图 2-23 配置 Kafka Client 节点参数



步骤5 配置完成后提交子作业。

----结束

开发主作业

步骤1 在“作业开发”主页面，进入脚本开发。

步骤2 构造原始表格student_score。新建临时Hive SQL脚本，选择Hive连接和数据库后，粘贴如下SQL语句并运行，运行成功后即可删除此脚本。

```
CREATE TABLE `student_score` (`name` String COMMENT '', `score` INT COMMENT '');
INSERT INTO
    student_score
VALUES
    ('ZHAO', '90'),
```

```
('QIAN', '88'),
('SUN', '93'),
('LI', '94'),
('ZHOU', '85'),
('WU', '79'),
('ZHENG', '87'),
('WANG', '97'),
('FENG', '83'),
('CEHN', '99');
```

步骤3 新建MRS Hive SQL节点调用的Hive SQL脚本。新建Hive SQL脚本，选择Hive连接和数据库后，粘贴如下SQL语句并提交版本，脚本命名为select95。

```
--从student_score表中展示成绩在95分以上的姓名和成绩--
SELECT * FROM student_score WHERE score> "95" ;
```

步骤4 在“作业开发”页面，新建数据开发主作业**EL_test_master**。选择一个HIVE SQL节点和一个For Each节点，选中连线图标并拖动，编排图2-21所示的作业。

步骤5 配置MRS Hive SQL节点参数。SQL脚本选择步骤3中提交的脚本select95，选择Hive连接和数据库。

图 2-24 配置 MRS Hive SQL 节点参数

步骤6 配置For Each节点属性，如图2-25所示。

- 子作业：子作业选择已经开发完成的子作业**EL_test_slave**。
- 数据集：数据集就是HIVE SQL节点的Select语句的执行结果。使用EL表达式**#{Job.getNodeOutput("select95")}**，其中**select95**为前一个节点的名称。
- 子作业参数：子作业参数是子作业中定义的参数名，然后在主作业中定义的参数值，传递到子作业以供使用。此处子作业参数名定义为**name**和**score**，其值为分别为数据集中的第一列和第二列数值，使用EL表达式**#{Loop.current[0]}**和**#{Loop.current[1]}**。

图 2-25 配置 For Each 节点参数

步骤7 配置完成后保存作业。

----结束

测试运行主作业

步骤1 单击主作业**EL_test_master**画布上方的“测试运行”按钮，测试作业运行情况。主作业运行后，会通过For Each节点循环调用运行子作业**EL_test_slave**。

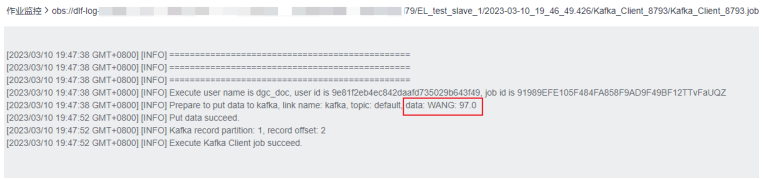
步骤2 单击左侧导航栏中的“实例监控”，进入实例监控中查看作业运行结果。

步骤3 待作业运行完成后，从实例监控中找到子作业**EL_test_slave**的循环运行结果，如图2-26所示。

图 2-26 子作业运行结果

步骤4 查看子作业**EL_test_slave**在循环运行中的结果日志，从日志中可以看到，结合For Each节点及其支持的Loop内嵌对象EL表达式，成功获取For Each前一节点输出的结果值。

图 2-27 查看日志



----结束

2.5 IF 条件判断教程

当您在数据开发模块进行作业开发编排时，想要实现通过设置条件，选择不同的执行路径，可使用IF条件判断。

本教程包含以下三个常见场景举例。

- **根据前一个节点的执行状态进行IF条件判断**
- **根据前一个节点的输出结果进行IF条件判断**
- **多IF条件下当前节点的执行策略**

IF条件的数据来源于EL表达式，通过EL表达式，根据具体的场景选择不同的EL表达式来达到目的。您可以参考本教程，根据您的实际业务需要，开发您自己的作业。

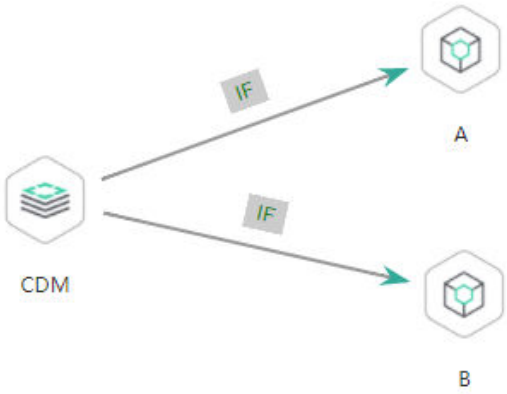
EL表达式用法可参考**EL表达式**。

根据前一个节点的执行状态进行 IF 条件判断

场景说明

根据前一个CDM节点是否执行成功，决定执行哪一个IF条件分支。基于**图2-28**的样例，说明如何设置IF条件。

图 2-28 作业样例



配置方法

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 选择“空间管理”页签，在工作空间列表中，找到所需要的工作空间，单击工作空间的“数据开发”，系统跳转至数据开发页面。

步骤3 在“作业开发”页面，新建数据开发作业，然后分别选择CDM节点和两个Dummy节点，选中连线图标并拖动，编排图2-28所示的作业。

其中CDM节点的失败策略需要设置为“继续执行下一节点”。

图 2-29 配置 CDM 节点的失败策略

步骤4 右键单击连线，选择“设置条件”，在弹出的“编辑EL表达式”文本框中输入IF条件。

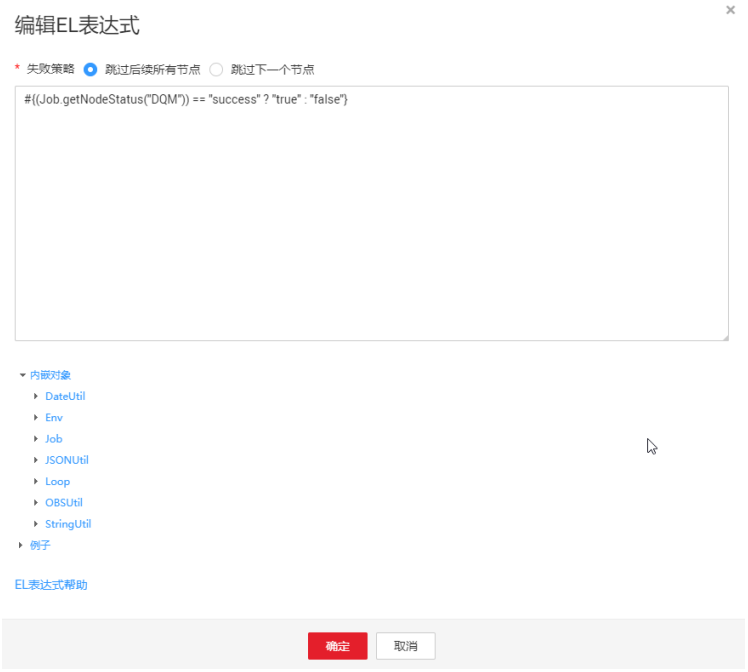
每一个条件分支都需要填写IF条件，IF条件为通过EL表达式语法填写三元表达式。当三元表达式结果为true的时候，才会执行连线后面的节点，否则后续节点将被跳过。

此Demo中使用的EL表达式为“#{Job.getNodeStatus("node_name")}”，这个表达式的作用为获取指定节点的执行状态，成功状态返回success，失败状态返回fail。本例使用中，IF条件表达式分别为：

- 上面的A分支IF条件表达式为： #{(Job.getNodeStatus("CDM")) == "success" ? "true" : "false"}
- 下面的B分支IF条件表达式为：#{(Job.getNodeStatus("CDM")) == "fail" ? "true" : "false"}

输入IF条件表达式后，配置IF条件匹配失败策略，可选择仅跳过相邻的下一个节点，或者跳过该IF分支后续所有节点。配置完成后单击确定，保存作业。

图 2-30 配置失败策略



步骤5 测试运行作业，并前往实例监控中查看执行结果。

步骤6 待作业运行完成后，从实例监控中查看作业实例的运行结果，如图2-31所示。可以看到运行结果是符合预期的，当前CDM执行的结果为fail的时候，跳过A分支，执行B分支。

图 2-31 作业运行结果

----结束

根据前一个节点的输出结果进行 IF 条件判断

场景说明

目标场景：通过HIVE SQL统计成绩在85分以上的人数，并将执行结果作为参数传递到下一个节点，通过与人数通过标准进行数值比较，然后决定执行哪一个IF条件分支。

场景分析：由于HIVE SQL节点的Select语句执行结果为单字段的二维数组，因此为获取二维数组中的值，EL表达式**#{Loop.dataArray[][]}**或**#{Loop.current[]}**均可以实现，且当前只有For Each节点支持Loop表达式，所以HIVE SQL节点后面需要连接一个For Each节点。

说明

此场景下不能使用StringUtil表达式
#{StringUtil.split(StringUtil.split(StringUtil.split(Job.getNodeOutput("前一节点名"),"]")[0],"[")[0],"\\"")[0]}替代Loop表达式，因为StringUtil表达式最终获取的数据类型为String，无法与标准数据Int比较大小。

作业编排如图2-32所示：

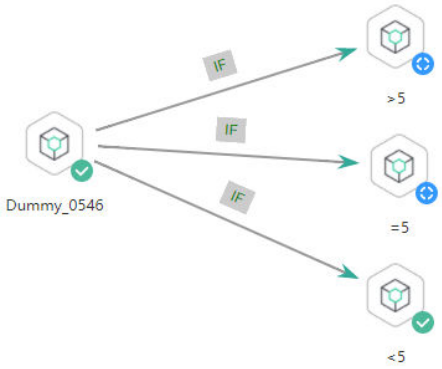
图 2-32 主作业样例

其中，For Each节点的关键配置如下：

- 数据集：数据集就是HIVE SQL节点的Select语句的执行结果。使用EL表达式**#{Job.getNodeOutput('HIVE')}**，其中HIVE为前一个节点的名称。
- 子作业参数：子作业参数是子作业中定义的参数，可以将主作业前一个节点的输出，传递到子作业以供使用。此处变量名为**result**，其值为数据集中的某一列，使用EL表达式**#{Loop.dataArray[0][0]}**或**#{Loop.current[]}**，本例以**{Loop.dataArray[0][0]}**为例进行说明。

而For Each节点中所选的子作业，需要根据For Each节点传过来的子作业参数，决定执行For Each中子作业的哪一个IF条件分支，作业编排如图2-33所示。

图 2-33 子作业样例



其中，子作业的关键配置为IF条件设置，本例使用表达式${result}获取作业参数的值。

说明

此处不能使用EL表达式#{Job.getParam("job_param_name")}，因为此表达式只能直接获取当前作业里配置的参数的value，并不能获取到父作业传递过来的参数值，也不能获取到工作空间里面配置的全局变量，作用域仅为本作业。

而表达式${job_param_name}，既可以获取到父作业传递过来的参数值，也可以获取到全局配置的变量。

配置方法

开发子作业

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 选择“空间管理”页签，在工作空间列表中，找到所需要的工作空间，单击工作空间的“数据开发”，系统跳转至数据开发页面。

步骤3 在“作业开发”页面，新建数据开发子作业For Each。选择四个Dummy节点，选中连线图标并拖动，编排图2-33所示的作业。

步骤4 右键单击节点间的连线，选择“设置条件”，在弹出的“编辑EL表达式”文本框中输入IF条件。

每一个条件分支都需要填写IF条件，IF条件为通过EL表达式语法填写三元表达式。当三元表达式结果为true的时候，才会执行连线后面的节点，否则后续节点将被跳过。

- 上面的>5分支，IF条件表达式为：#{${result} > 5 ? "true" : "false"}
- 中间的=5分支，IF条件表达式为：#{${result} == 5 ? "true" : "false"}
- 下面的<5分支，IF条件表达式为：#{${result} < 5 ? "true" : "false"}

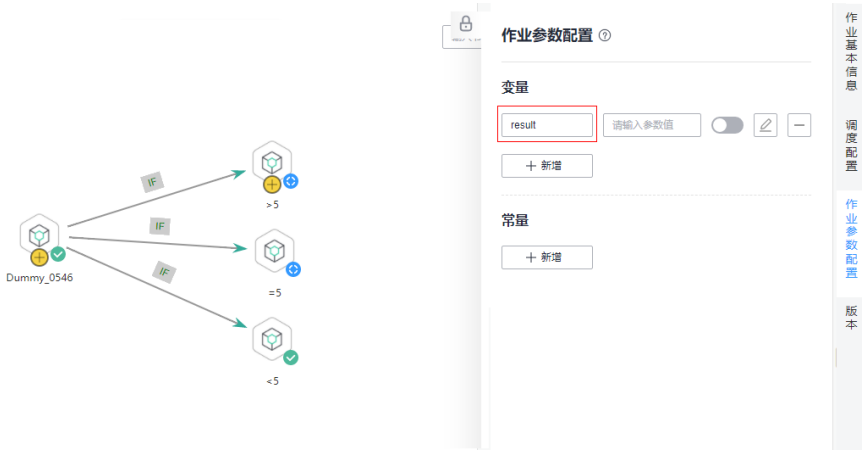
输入IF条件表达式后，配置IF条件匹配失败策略，可选择仅跳过相邻的下一个节点，或者跳过该IF分支后续所有节点。

说明

表达式中包含多条件的场景下，可以通过“||”联合多个条件。例如：
#{(${result} >= 19 || ${result} <=9) ? "true" : "false"}

步骤5 配置作业参数。此处需将参数名填写为**result**，仅用于主作业testif中的For Each节点识别子作业参数；参数值无需填写。

图 2-34 配置作业参数



步骤6 配置完成后保存作业。

----结束

开发主作业

步骤1 在“作业开发”页面，新建数据开发主作业testif。选择HIVE SQL节点和For Each节点，选中连线图标并拖动，编排图2-32所示的作业。

步骤2 配置HIVE SQL节点属性。此处配置为引用SQL脚本，SQL脚本的语句如下所示。其他节点属性参数无特殊要求。


```
--从student_score表中统计成绩在85分以上的人数--
SELECT count(*) FROM student_score WHERE score> "85" ;
```

图 2-35 HIVE SQL 脚本执行结果

步骤3 配置For Each节点属性，如图2-36所示。

- 子作业：子作业选择已经开发完成的子作业“foreach”。
- 数据集：数据集就是HIVE SQL节点的Select语句的执行结果。使用EL表达式**#{Job.getNodeOutput('HIVE')}**，其中HIVE为前一个节点的名称。
- 子作业参数：子作业参数是子作业中定义的参数，可以将主作业前一个节点的输出，传递到子作业以供使用。此处变量名为子作业参数名**result**，其值为数据集中的某一列，使用EL表达式**#{Loop.dataArray[0][0]}**。

图 2-36 For Each 节点属性

步骤4 配置完成后保存作业。

----结束

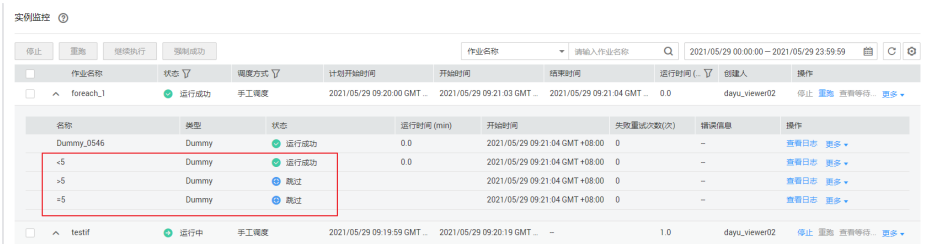
测试运行主作业

步骤1 单击主作业画布上方的“测试运行”按钮，测试作业运行情况。主作业运行后，会通过For Each节点自动调用运行子作业。

步骤2 单击左侧导航栏中的“实例监控”，进入实例监控中查看作业运行结果。

步骤3 待作业运行完成后，从实例监控中查看子作业foreach的运行结果，如图2-37所示。可以看到运行结果是符合预期的，当前HIVE SQL执行的结果是4，所以>5和=5的分支被跳过，执行<5这个分支成功。

图 2-37 子作业运行结果



----结束

多 IF 条件下当前节点的执行策略

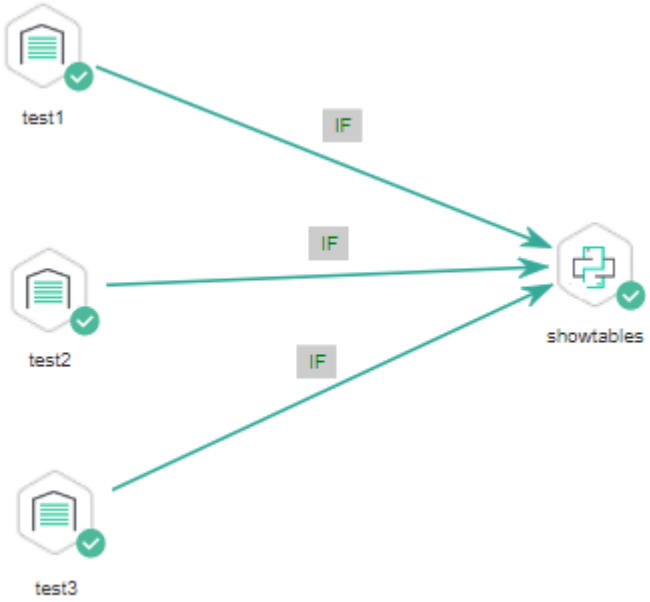
如果当前节点的执行依赖多个IF条件的节点，执行的策略包含逻辑或和逻辑与两种。

当执行策略配置为逻辑或，则表示多个IF判断条件只要任意一个满足条件，则执行当前节点。

当执行策略配置为逻辑与，则表示多个IF判断条件需要所有条件满足时，才执行当前节点。

如果没有配置执行策略，系统默认为逻辑或处理。

图 2-38 多 IF 条件作业样例



配置方法

配置执行策略

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 选择“空间管理”页签，在工作空间列表中，找到所需要的工作空间，单击工作空间的“数据开发”，系统跳转至数据开发页面。

步骤3 在数据开发模块，单击“配置管理 > 配置”，单击“默认项配置”。

步骤4 “多IF策略”可设置为“逻辑与”或者“逻辑或”。

步骤5 单击“保存”。

----结束

开发作业

步骤1 在“作业开发”页面，新建一个数据开发作业。

步骤2 拖动三个DWS SQL算子作为父节点，一个Python算子作为子节点，选中连线图标并拖动，编排图2-38所示的作业。

步骤3 右键单击节点间的连线，选择“设置条件”，在弹出的“编辑EL表达式”文本框中输入IF条件。

每一个条件分支都需要填写IF条件，IF条件为通过EL表达式语法填写三元表达式。

- test1节点IF条件表达式为：#{(Job.getNodeStatus("test1")) == "success" ? "true" : "false"},

- test2节点IF条件表达式为：#{(Job.getNodeStatus("test2")) == "success" ? "true" : "false"},
- test3节点IF条件表达式为：#{(Job.getNodeStatus("test3")) == "success" ? "true" : "false"},

此处表达式均采用前一个节点的执行状态进行IF条件判断。

输入IF条件表达式后，配置IF条件匹配失败策略，可选择仅跳过相邻的下一个节点，或者跳过该IF分支后续所有节点。

----结束

测试运行作业

步骤1 单击作业画布上方的“保存”按钮，保存完成编排的作业。

步骤2 单击作业画布上方的“测试运行”按钮，测试作业运行情况。

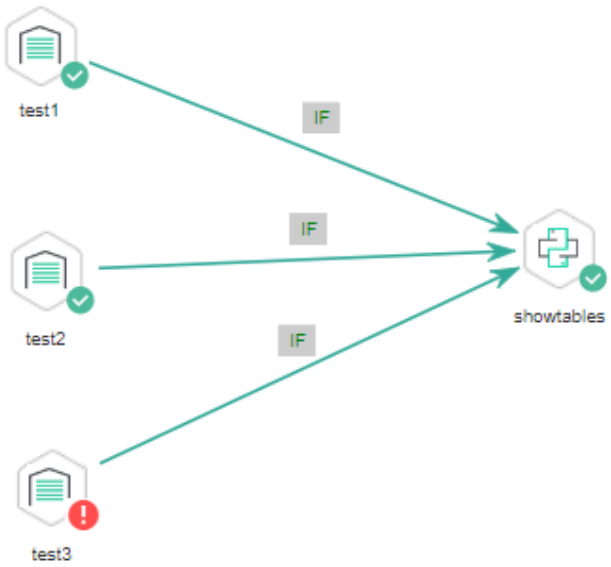
test1运行成功，则对应的IF条件为true；

test2运行成功，则对应的IF条件为true；

test3运行失败，则对应的IF条件为false。

当多IF策略配置为“逻辑或”时，showtables节点运行完成，作业运行完成。详细情况如下所示。

图 2-39 配置为“逻辑或”的作业运行情况

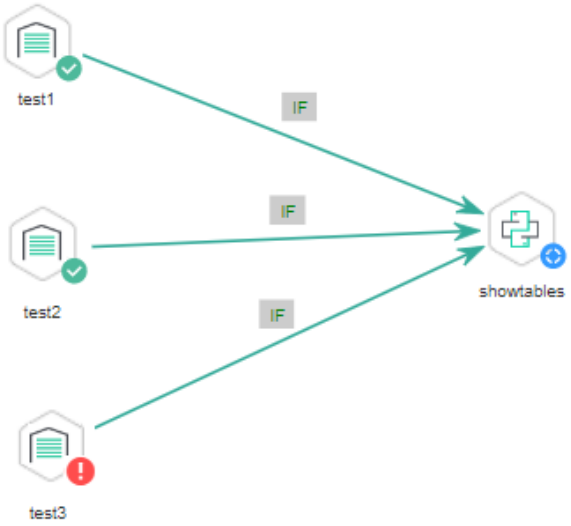


测试运行日志

[INFO][2022/03/31 15:53:35 GMT+08:00] : 作业开始运行...
[INFO][2022/03/31 15:54:12 GMT+08:00] : 节点"test1"开始运行...
[INFO][2022/03/31 15:54:12 GMT+08:00] : 节点"test2"开始运行...
[INFO][2022/03/31 15:54:12 GMT+08:00] : 节点"test3"开始运行...
[INFO][2022/03/31 15:54:22 GMT+08:00] : 节点"test1"运行完成。
[INFO][2022/03/31 15:54:22 GMT+08:00] : 节点"test2"运行完成。
[ERROR][2022/03/31 15:54:53 GMT+08:00] : 节点"test3"运行失败。
[INFO][2022/03/31 15:55:03 GMT+08:00] : 节点"showtables"开始运行...
[INFO][2022/03/31 15:55:13 GMT+08:00] : 节点"showtables"运行完成。
[INFO][2022/03/31 15:55:13 GMT+08:00] : 作业运行完成

当多IF策略配置为“逻辑与”时，showtables节点跳过，作业运行完成。详细情况如下所示。

图 2-40 配置为“逻辑与”的作业运行情况



测试运行日志

[INFO][2022/03/31 15:51:38 GMT+08:00] : 作业开始运行...
[INFO][2022/03/31 15:52:16 GMT+08:00] : 节点"test2"运行完成。
[INFO][2022/03/31 15:52:16 GMT+08:00] : 节点"test1"运行完成。
[INFO][2022/03/31 15:52:16 GMT+08:00] : 节点"test3"开始运行...
[ERROR][2022/03/31 15:52:56 GMT+08:00] : 节点"test3"运行失败。
[INFO][2022/03/31 15:53:06 GMT+08:00] : 节点"showtables"已跳过
[INFO][2022/03/31 15:53:17 GMT+08:00] : 作业运行完成

----结束

2.6 获取 Rest Client 节点返回值教程

Rest Client节点可以执行华为云内的RESTful请求。

本教程主要介绍如何获取Rest Client的返回值，包含以下两个使用场景举例。

- **通过“响应消息体解析为传递参数定义”获取返回值**
- **通过EL表达式获取返回值**

通过“响应消息体解析为传递参数定义”获取返回值

如图2-41所示，第一个Rest Client调用了MRS服务查询集群列表的API，图2-42为API返回值的JSON消息体。

- 使用场景：需要获取集群列表中第一个集群的cluster Id，然后作为参数传递给后面的节点使用。

- 关键配置：在第一个Rest Client的“响应消息体解析为传递参数定义”配置中，配置clusterId=clusters[0].clusterId，后续的Rest Client节点就可以用${clusterId}的方式引用到集群列表中的第一个集群的cluster Id。

  说明

  响应消息体解析为参数传递定义时，传递的参数名（例如clusterId）在该作业的所有节点参数中需要保持唯一性，避免和其他参数同名。

图 2-41 Rest Clien 作业样例 1

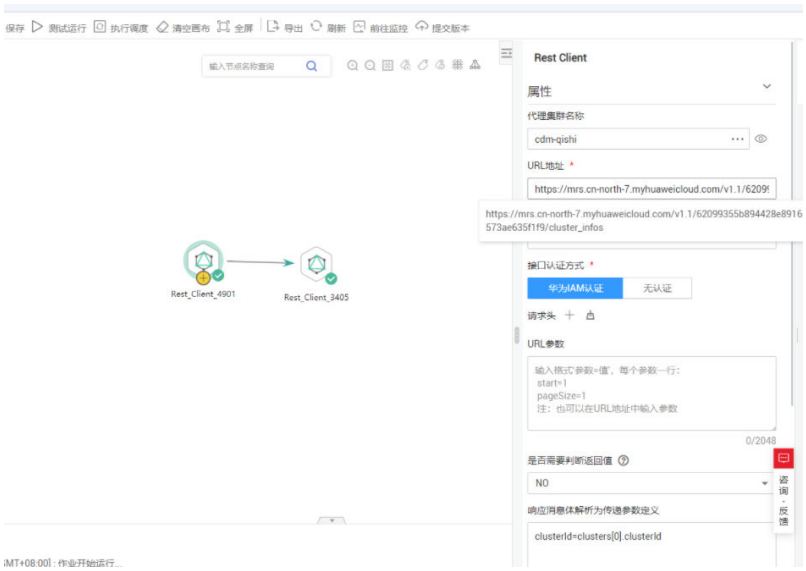


图 2-42 JSON 消息体


```
{
    "clusterTotal":31,
    "clusters":[
        {
            "clusterId":"6ea1b5c2-6526-4ef8-9c8f-4105b63fa893",
            "clusterName":"mrs_hbase22",
            "totalNodeNum":"2",
            "clusterState":"running",
            "stageDesc":null,
            "createAt":"1620378935",
            "updateAt":"1620611307",
            "chargingStartTime":"1620380067",
            "billingType":"Metered",
            "dataCenter":"cerxth-7",
            "vpc":"vpc-dlf",
            "vpcId":"f35aee01-c4a3-47c1-8d92-9df430537de4",
            "duration":"0",
            "fee":"0.0",
            "hadoopVersion":"",
            "componentList":[
                {
                    "id":"218051",
                    "componentId":"MRS_2.1.0_001",
                    "componentName":"Hadoop",
                    "componentVersion":"3.1.1",
                    "external_datasources":null,
                    "componentDesc":"A distributed data storage and processing framework for large data sets, including core components such as HDFS, YARN, and MapReduce.",
                    "componentDescEn":null,
                    "multi_service_name":null
                },
```

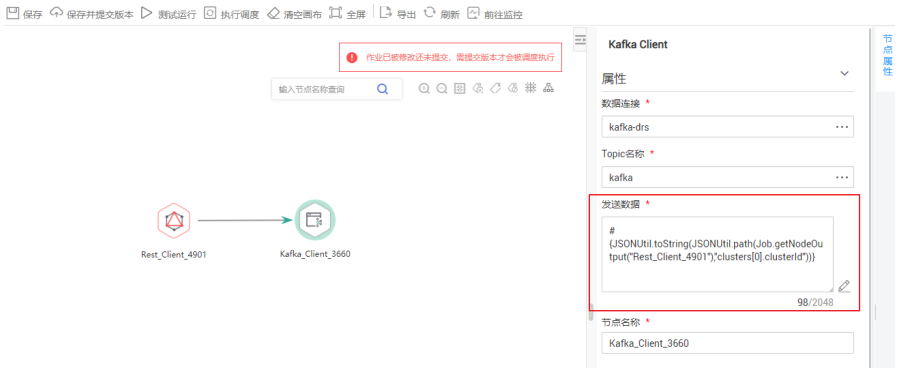

通过 EL 表达式获取返回值

Rest Client算子可与EL表达式相配合，根据具体的场景选择不同的EL表达式来实现更丰富的用法。您可以参考本教程，根据您的实际业务需要，开发您自己的作业。EL表达式用法可参考**EL表达式**。

如图2-43所示，Rest Client调用了MRS服务查询集群列表的API，然后执行Kafka Client发送消息。

- 使用场景：Kafka Client发送字符串消息，消息内容为集群列表中第一个集群的cluster Id。
- 关键配置：在Kafka Client中使用如下EL表达式获取Rest API返回消息体中的特定字段：
  #{JSONUtil.toString(JSONUtil.path(Job.getNodeOutput("Rest_Client_4901"),"clusters[0].clusterId"))}

图 2-43 Rest Client 作业样例 2



2.7 For Each 节点使用介绍

适用场景

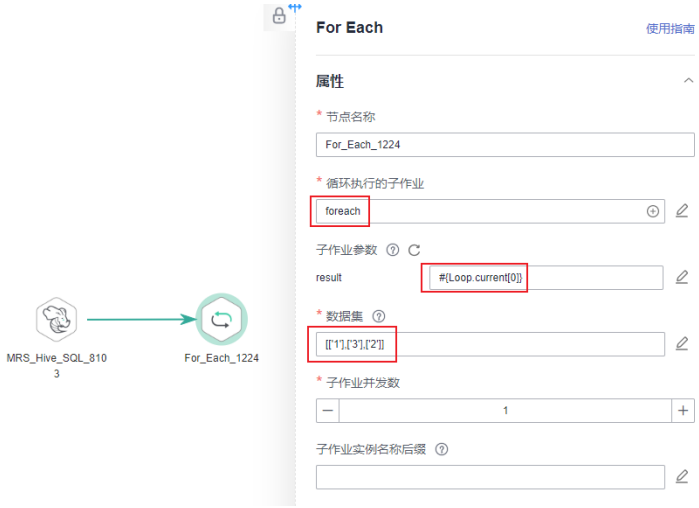
当您进行作业开发时，如果某些任务的参数有差异、但处理逻辑全部一致，在这种情况下您可以通过For Each节点避免重复开发作业。

For Each节点可指定一个子作业循环执行，并通过数据集对子作业中的参数进行循环替换。关键参数如下：

- 子作业：选择需要循环执行的作业。
- 数据集：即不同子任务的参数值的集合。可以是给定的数据集，如“[['1'],['3'],['2']]”；也可以是EL表达式如“#{Job.getNodeOutput('preNodeName')}”，即前一个节点的输出值。
- 子作业参数：参数名即子作业中定义的变量；参数值一般配置为数据集中的某组数据，每次运行中会将参数值传递到子作业以供使用。例如参数值填写为：#{Loop.current[0]}，即将数据集中每行数据的第一个数值遍历传递给子作业。

For Each节点举例如图2-44所示 。从图中可以看出，子作业“foreach”中的参数名为“result”，参数值为一维数组数据集“[['1'],['3'],['2']] ”的遍历（即第一次循环为1，第二次循环为3，第三次循环为2）。

图 2-44 for each 节点



For Each 节点与 EL 表达式

要想使用好For Each节点，您必须对EL表达式有所了解。EL表达式用法请参考**EL表达式**。

下面为您展示For Each节点常用的一些EL表达式。

- #{Loop.dataArray}：For循环节点输入的数据集，是一个二维数组。
- #{Loop.current}：由于For循环节点在处理数据集的时候，是一行一行进行处理的，那Loop.current就表示当前处理到的某行数据，Loop.current是一个一维数组，一般定义格式为#{Loop.current[0]}、#{Loop.current[1]}或其他，0表示遍历到当前行的第一个值。
- #{Loop.offset}：For循环节点在处理数据集时当前的偏移量，从0开始。
- #{Job.getNodeOutput('preNodeName')}：获取前面节点的输出。

使用案例

案例场景

因数据规整要求，需要周期性地将多组DLI源数据表数据导入到对应的DLI目的表，如**表1**所示。

表 2-9 需要导入的列表情况

源数据表名	目的表名
a_new	a
b_2	b
c_3	c
d_1	d
c_5	e
b_1	f

如果通过SQL节点分别执行导入脚本，需要开发大量脚本和节点，导致重复性工作。在这种情况下，我们可以使用For Each节点进行循环作业，节省开发工作量。

配置方法

步骤1 准备源表和目的表。为了便于后续作业运行验证，需要先创建DLI源数据表和目的表，并给源数据表插入数据。

1. 创建DLI表。您可以在DataArts Studio数据开发中，新建DLI SQL脚本执行以下SQL命令，也可以在数据湖探索（DLI）服务控制台中的SQL编辑器中执行以下SQL命令：

```
/* 创建数据表 */
CREATE TABLE a_new (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE b_2 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE c_3 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE d_1 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE c_5 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE b_1 (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE a (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE b (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE c (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE d (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE e (name STRING, score INT) STORED AS PARQUET;
CREATE TABLE f (name STRING, score INT) STORED AS PARQUET;
```

2. 给源数据表插入数据。您可以在DataArts Studio数据开发模块中，新建DLI SQL脚本执行以下SQL命令，也可以在数据湖探索（DLI）服务控制台中的SQL编辑器中执行以下SQL命令：

```
/* 源数据表插入数据 */
INSERT INTO a_new VALUES ('ZHAO','90'),('QIAN','88'),('SUN','93');
INSERT INTO b_2 VALUES ('LI','94'),('ZHOU','85');
INSERT INTO c_3 VALUES ('WU','79');
INSERT INTO d_1 VALUES ('ZHENG','87'),('WANG','97');
INSERT INTO c_5 VALUES ('FENG','83');
INSERT INTO b_1 VALUES ('CEHN','99');
```

步骤2 准备数据集数据。您可以通过以下方式之一获取数据集：

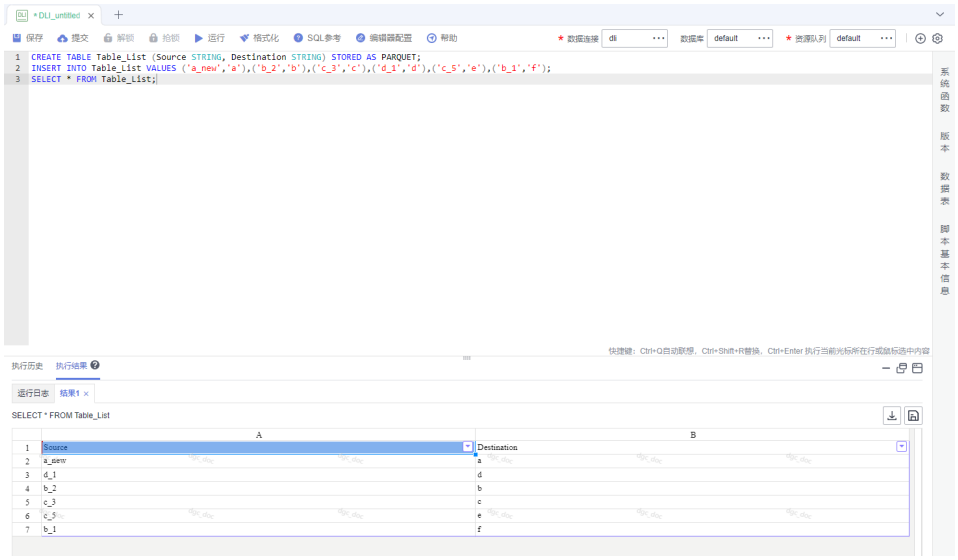
1. 您可以将**表1**数据导入到DLI表中，然后将SQL脚本读取的结果作为数据集。
2. 您可以将**表1**数据保存在OBS的CSV文件中，然后通过DLI SQL或DWS SQL创建OBS外表关联这个CSV文件，然后将OBS外表查询的结果作为数据集。DLI创建外表请参见**OBS输入流**，DWS创建外表请参见**创建外表**。
3. 您可以将**表1**数据保存在HDFS的CSV文件中，然后通过HIVE SQL创建Hive外表关联这个CSV文件，然后将HIVE外表查询的结果作为数据集。MRS创建外表请参见**创建表**。

本例以方式1进行说明，将**表1**中的数据导入到DLI表（Table_List）中。您可以在DataArts Studio数据开发模块中，新建DLI SQL脚本执行以下SQL命令导入数据，也可以在数据湖探索（DLI）服务控制台中的SQL编辑器中执行以下SQL命令：

```
/* 创建数据表TABLE_LIST，然后插入表1数据，最后查看生成的表数据 */
CREATE TABLE Table_List (Source STRING, Destination STRING) STORED AS PARQUET;
INSERT INTO Table_List VALUES ('a_new','a'),('b_2','b'),('c_3','c'),('d_1','d'),('c_5','e'),('b_1','f');
SELECT * FROM Table_List;
```

生成的Table_List表数据如下：

图 2-45 Table_List 表数据



步骤3 创建要循环运行的子作业ForeachDemo。在本次操作中，定义循环执行的是一个包含了DLI SQL节点的任务。

1. 进入DataArts Studio数据开发模块选择“作业开发”页面，新建作业ForeachDemo，然后选择DLI SQL节点，编排图2-46所示的作业。

   DLI SQL的语句中把要替换的变量配成${}这种参数的形式。在下面的SQL语句中，所做的操作是把${Source}表中的数据全部导入${Destination}中，${fromTable}、${toTable} 就是要替换的变量参数。SQL语句为：

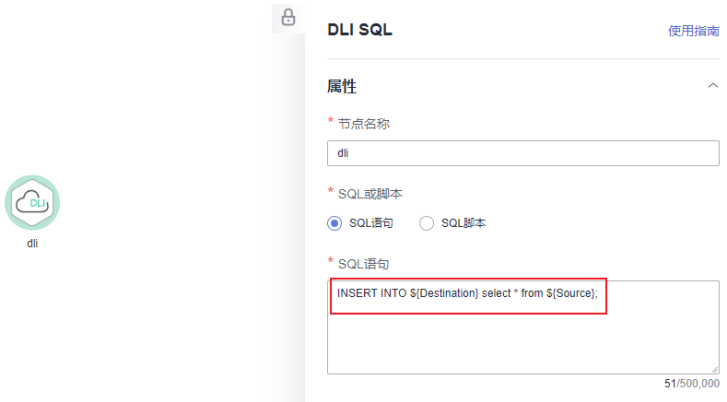
   ```
   INSERT INTO ${Destination} select * from ${Source};
   ```

   说明

   此处不能使用EL表达式#{Job.getParam("job_param_name")}，因为此表达式只能直接获取当前作业里配置的参数的value，并不能获取到父作业传递过来的参数值，也不能获取到工作空间里面配置的全局变量，作用域仅为本作业。

   而表达式${job_param_name}，既可以获取到父作业传递过来的参数值，也可以获取到全局配置的变量。

图 2-46 循环执行子作业



2. 配置完成SQL语句后，在子作业中配置作业参数。此处仅需要配置参数名，用于主作业ForeachDemo_master中的For Each节点识别子作业参数；参数值无需填写。

图 2-47 配置子作业参数

3. 配置完成后保存作业。

步骤4 创建For Each节点所在的主作业ForeachDemo_master。

1. 进入DataArts Studio数据开发模块选择“作业开发”页面，新建数据开发主作业ForeachDemo_master。选择DLI SQL节点和For Each节点，选中连线图标并拖动，编排图2-48所示的作业。

图 2-48 编排作业

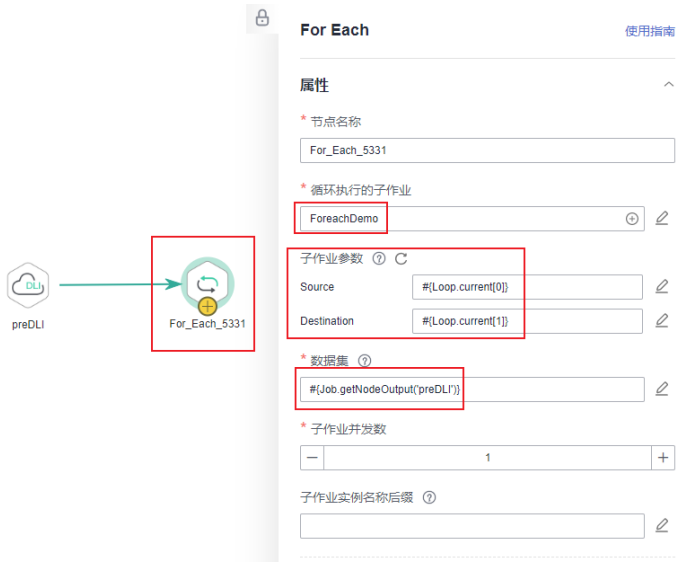
2. 配置DLI SQL节点属性，此处配置为SQL语句，语句内容如下所示。DLI SQL节点负责读取DLI表Table_List中的内容作为数据集。

```
SELECT * FROM Table_List;
```

图 2-49 DLI SQL 节点配置

3. 配置For Each节点属性。
   - 子作业：子作业选择**步骤2**已经开发完成的子作业“ForeachDemo”。
   - 数据集：数据集就是DLI SQL节点的Select语句的执行结果。使用EL表达式**#{Job.getNodeOutput('preDLI')}**，其中preDLI为前一个节点的名称。
   - 子作业参数：用于将数据集中的数据传递到子作业以供使用。Source对应的是数据集Table_List表的第一列，Destination是第二列，所以配置的EL表达式分别为**#{Loop.current[0]}**、**#{Loop.current[1]}**。

图 2-50 配置 For Each 节点

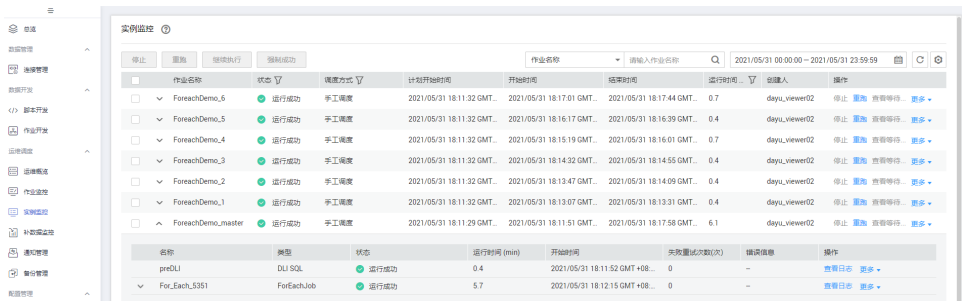


4. 配置完成后保存作业。

步骤5 测试运行主作业。

1. 单击主作业画布上方的“测试运行”按钮，测试作业运行情况。主作业运行后，会通过For Each节点自动调用运行子作业。
2. 单击左侧导航栏中的“实例监控”，进入实例监控中查看作业运行情况。等待作业运行成功后，就能查看For Each节点生成的子作业实例，由于数据集中有6行数据，所以这里就对应产生了6个子作业实例。

图 2-51 查看作业实例

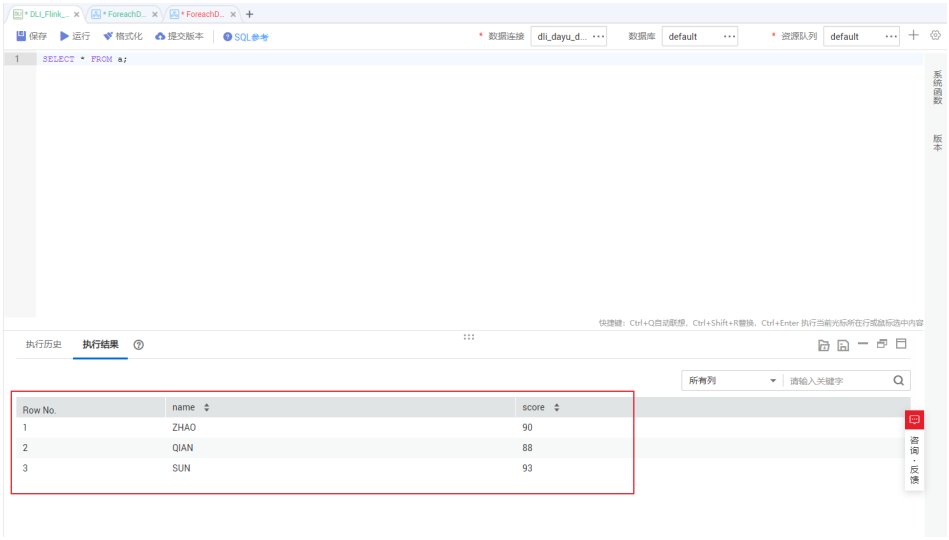


3. 查看对应的6个DLI目的表中是否已被插入预期的数据。您可以在DataArts Studio数据开发模块中，新建DLI SQL脚本执行以下SQL命令导入数据，也可以在数据湖探索（DLI）服务控制台中的SQL编辑器中执行以下SQL命令：

```
/* 查看表a数据，其他表数据请修改命令后运行 */
SELECT * FROM a;
```

将查询到的表数据与**给源数据表插入数据**步骤中的数据进行对比，可以发现数据插入符合预期。

图 2-52 目的表数据



----结束

更多案例参考

For Each节点可与其他节点配合，实现更丰富的功能。您可以参考以下案例，了解For Each节点的更多用法。

- **通过CDM节点批量创建分表迁移作业**
- **根据前一个节点的输出结果进行IF条件判断**

2.8 数据开发调用数据质量算子并且作业运行的时候需要传入质量参数

由于数据质量作业在执行SQL语句时不支持传参，通过数据开发调用数据质量算子，运行的时候可以把数据质量作业的参数传递给数据质量算子作业，实现数据质量的参数传递。

使用场景

数据质量需要传递参数到数据质量算子作业里面并且能够正常运行。

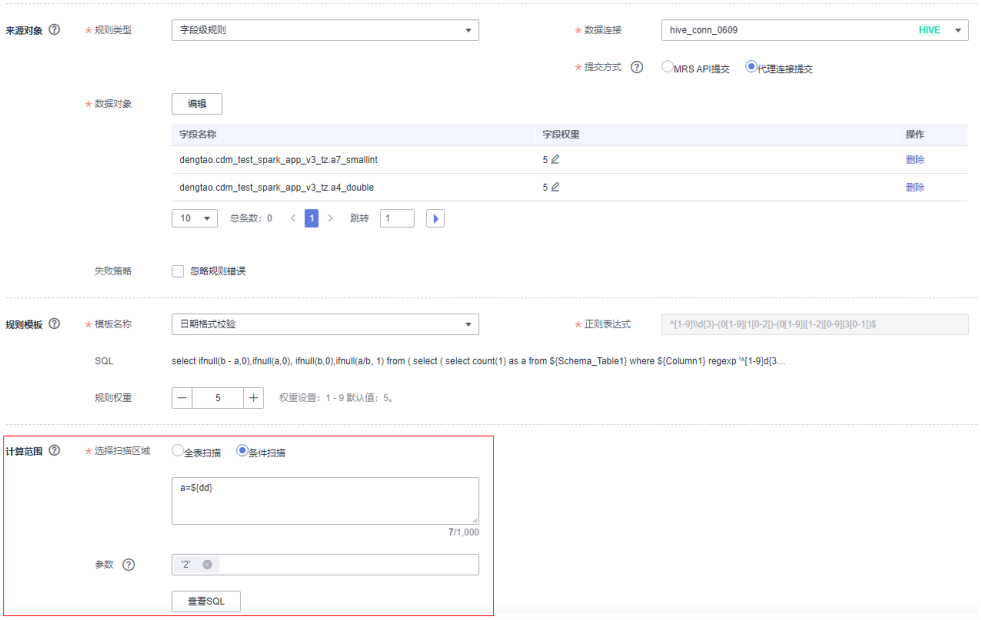
配置方法

创建质量作业

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。
2. （可选）选择“数据质量监控 > 质量作业”，新建目录。如果已存在可用的目录，可以不用新建目录。

3. 在“质量作业”页面单击“新建”，进入“基本配置”页面，配置质量作业的基本信息。
4. 单击“下一步”进入“规则配置”页面，配置质量作业的相关规则。在“计算范围”的“条件扫描”里面配置数据质量作业参数，如下图所示。

图 2-53 设置数据质量参数



5. 单击“下一步”，依次配置告警、订阅、调度等信息。配置质量作业的详细操作请参见**新建质量作业**。
6. 单击“提交”。数据质量作业配置完成。

配置数据开发作业

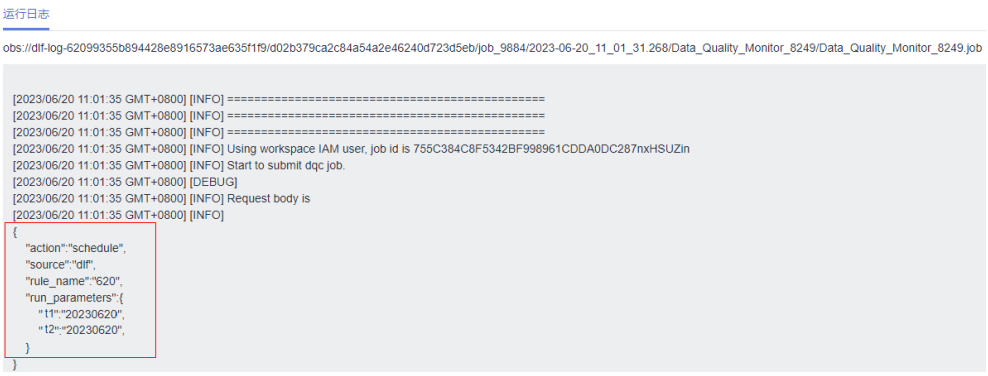
1. 登录DataArts Studio控制台。选择实例，单击“进入控制台”，选择对应工作空间的“数据开发”模块，进入数据开发页面。
2. 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发”。
3. 创建一个批处理的Pipeline作业并进入作业配置页面。
4. 选择Data_Quality_Monitor数据质量监控算子，将该节点拖入空白页面。并配置节点属性参数。

图 2-54 配置 Data_Quality_Monitor 节点属性

5. 配置调度周期。

图 2-55 配置调度周期

6. 提交版本并执行调度。
7. 在“作业监控”查看作业运行日志。



2.9 跨空间进行作业调度

适用场景

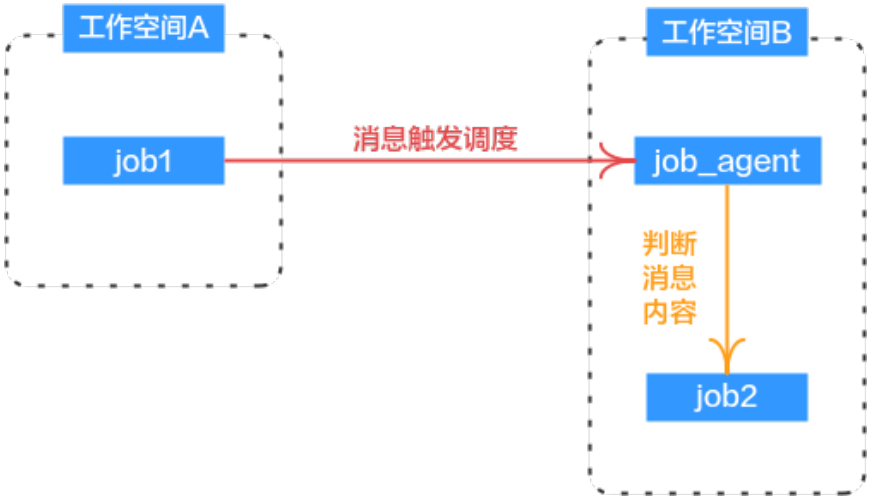
当您已按照工作空间划分权限，不同空间用户只能操作本空间的作业。但是不同的工作空间之间的作业如果存在依赖关系，可参见本教程操作实现跨空间作业调度。

方案说明

DataArts Studio数据开发模块支持以事件触发的方式运行作业，因此通过DIS或者MRS Kafka作为作业依赖纽带，可以跨空间实现作业调度。

如下图，工作空间A中的job1运行完成后，可以使用DIS Client或Kafka Client发送消息触发中继作业job_agent；job_agent配置事件触发调度，根据DIS Client或Kafka Client发送的消息触发运行后，判断消息是否符合预期，符合则触发job2作业运行，否则不再触发job2运行。

图 2-56 调度方案



前提条件

以下条件满足其一即可：

- 已具备DIS通道。
- 已具备MRS服务Kafka组件，并已分别在工作空间A和B的管理中心组件内，创建MRS Kafka连接。

配置方法（DIS Client）

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 单击第一个工作空间A的“数据开发”，系统跳转至数据开发页面，新建数据开发作业job1。分别选择Dummy节点和DIS Client节点，选中连线图标并拖动，编排如图2-57所示的作业。

- Dummy节点不执行任何操作，本例选择Dummy节点仅为演示操作，实际使用中您可以用其他作业节点替代。
- DIS Client节点用于发送消息。您需要选择DIS所属Region和通道，并将发送数据配置为EL表达式**job1,#{DateUtil.getDay(Job.startTime)}**。则当本作业执行完成后，将使用DIS Client发送一条字符串消息：job1,***作业执行日期***。例如2月15日作业job1执行，实际的消息则为：**job1,15**。
- 作业调度等其他作业参数无需配置，保持默认即可。

图 2-57 job1 作业 DIS Client 节点配置

步骤3 在另一个工作空间B，新建数据开发作业job_agent。分别选择Dummy节点和Subjob节点，选中连线图标并拖动，编排图2-58所示的作业。

图 2-58 job_agent 作业调度配置

- Dummy节点不执行任何操作，本例选择Dummy节点用于设置Dummy节点到Subjob节点之间连线的IF条件。
- Subjob节点用于将需要后续执行的作业job2作为子作业引用执行。实际使用中您可以引用已有作业，也可以使用其他作业节点替代Subjob节点。
- 作业的调度方式设置为“事件驱动调度”，DIS通道名称选择为工作空间A中job1作业中DIS Client节点所选择的通道，用于通过DIS消息触发作业运行。
- IF判断条件设置，用于校验DIS Client节点发送的消息是否符合预期，符合才会继续执行Subjob节点，否则跳过。

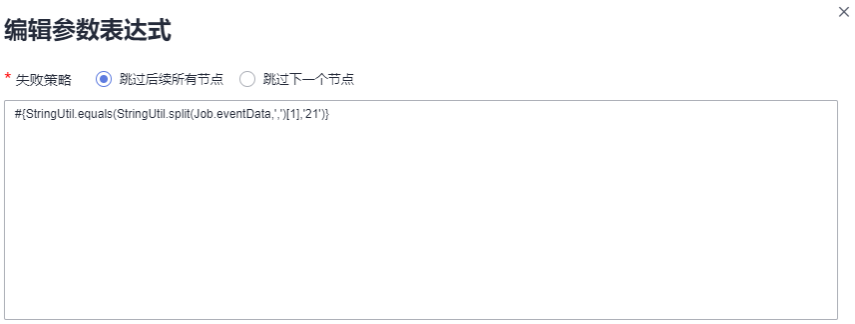
  右键单击连线，选择“设置条件”，在弹出的“编辑参数表达式”文本框中输入IF判断条件，失败策略保持默认即可。IF判断条件为通过EL表达式语法填写三元表达式，当三元表达式结果为true的时候，才会执行连线后面的节点，否则后续节点将被跳过。


```
#{StringUtil.equals(StringUtil.split(Job.eventData,',')[1],'21')}
```

该IF判断条件表示，仅当从DIS通道获取的消息逗号后的部分为“21”时，即每月21日时，才执行后续的作业节点。

说明

如果您需要匹配多条消息记录，可以添加多个Dummy节点并分别添加到Subjob节点的IF条件，然后将数据开发组件配置项中的“多IF策略”设置为“逻辑或”即可。



步骤4 测试运行作业job_agent，在工作空间A的作业job1未运行的情况下，前往实例监控中查看执行结果是否符合预期。

由于作业job1未运行即未发送消息，则job_agent作业中的Subjob节点被跳过，证明IF条件判断生效。

图 2-59 Subjob 节点被跳过

步骤5 启动调度job_agent。然后测试运行工作空间A作业job1，待job1实例运行成功后，前往工作空间B实例监控中查看作业运行结果是否符合预期。

- job_agent被触发运行。
- 如果当天日期和IF条件中的日期匹配，则job_agent作业中的Subjob节点成功运行、子作业job2也执行完成。否则Subjob节点被跳过。

图 2-60 Subjob 节点成功运行

----结束

配置方法（Kafka Client）

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 单击第一个工作空间A的“数据开发”，系统跳转至数据开发页面，新建数据开发作业job1。分别选择Dummy节点和Kafka Client节点，选中连线图标并拖动，编排如图2-61所示的作业。

- Dummy节点不执行任何操作，本例选择Dummy节点仅为演示操作，实际使用中您可以用其他作业节点替代。
- Kafka Client节点用于发送消息。您需要选择Kafka连接和Topic名称，并将发送数据配置为EL表达式**job1,#{DateUtil.getDay(Job.startTime)}**。则当本作业执行完成后，将使用Kafka Client发送一条字符串消息：job1,***作业执行日期***。例如2月15日作业job1执行，实际的消息则为：**job1,15**。
- 作业调度等其他作业参数无需配置，保持默认即可。

图 2-61 job1 作业 Kafka Client 节点配置

步骤3 在另一个工作空间B，新建数据开发作业job_agent。分别选择Dummy节点和Subjob节点，选中连线图标并拖动，编排图2-62所示的作业。

图 2-62 job_agent 作业调度配置

- Dummy节点不执行任何操作，本例选择Dummy节点用于设置Dummy节点到Subjob节点之间连线的IF条件。
- Subjob节点用于将需要后续执行的作业job2作为子作业引用执行。实际使用中您可以引用已有作业，也可以使用其他作业节点替代Subjob节点。
- 作业的调度方式设置为“事件驱动调度”，连接名称和Topic选择为工作空间B中的Kafka连接和Topic，需要与工作空间A中job1作业中Kafka Client节点所选择的Kafka连接和Topic相对应，用于通过Kafka消息触发作业运行。
- IF判断条件设置，用于校验Kafka Client节点发送的消息是否符合预期，符合才会继续执行Subjob节点，否则跳过。

  右键单击连线，选择“设置条件”，在弹出的“编辑参数表达式”文本框中输入IF判断条件，失败策略保持默认即可。IF判断条件为通过EL表达式语法填写三元表

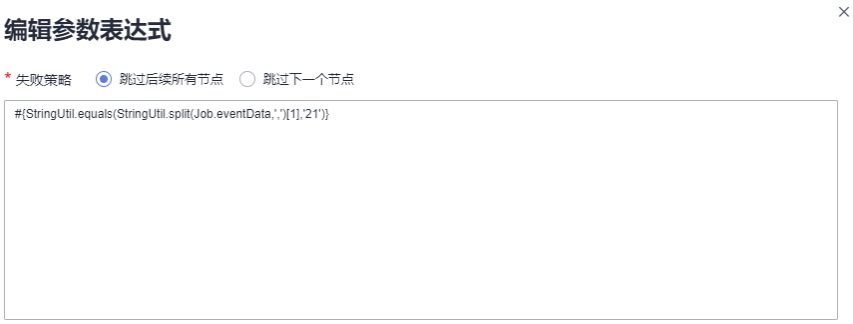
达式，当三元表达式结果为true的时候，才会执行连线后面的节点，否则后续节点将被跳过。

```
#{StringUtil.equals(StringUtil.split(Job.eventData,',')[1],'21')}
```

该IF判断条件表示，仅当从Kafka通道获取的消息逗号后的部分为“21”时，即每月21日时，才执行后续的作业节点。

说明

如果您需要匹配多条消息记录，可以添加多个Dummy节点并分别添加到Subjob节点的IF条件，然后将数据开发组件配置项中的“多IF策略”设置为“逻辑或”即可。



步骤4 测试运行作业job_agent，在工作空间A的作业job1未运行的情况下，前往实例监控中查看执行结果是否符合预期。

由于作业job1未运行即未发送消息，则job_agent作业中的Subjob节点被跳过，证明IF条件判断生效。

图 2-63 Subjob 节点被跳过

步骤5 启动调度job_agent。然后测试运行工作空间A作业job1，待job1实例运行成功后，前往工作空间B实例监控中查看作业运行结果是否符合预期。

- job_agent被触发运行。
- 如果当天日期和IF条件中的日期匹配，则job_agent作业中的Subjob节点成功运行、子作业job2也执行完成。否则Subjob节点被跳过。

图 2-64 Subjob 节点成功运行

----结束

3 跨工作空间的 DataArts Studio 数据搬迁

3.1 概述

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。

操作前准备

- 已创建新的工作空间，新建工作空间的用户需要具备 Administrator或Tenant Administrator权限。
- 执行数据搬迁的用户，至少应具备新旧两个工作空间的开发者权限。
- CDM集群和数据服务专享版集群在工作空间之间相互隔离，建议您在新空间提前准备好对应旧空间的集群。
- 搬迁依赖于OBS功能，请您提前规划OBS桶和文件夹目录。
- DataArts Studio数据搬迁时，依赖各组件的备份或导入导出能力。您可以根据自己的数据需求，自由选择搬迁哪个组件的数据。
  - **管理中心数据搬迁**
  - **数据集成数据搬迁**
  - **数据架构数据搬迁**
  - **数据开发数据搬迁**
  - **数据质量数据搬迁**
  - **数据目录数据搬迁**
  - **数据安全数据搬迁**
  - **数据服务数据搬迁**

3.2 管理中心数据搬迁

管理中心数据搬迁依赖于管理中心的资源迁移功能。

资源迁移支持迁移的资源包含数据服务、数据目录和管理中心数据连接。

约束与限制

- 资源导入可以基于OBS服务，也支持从本地导入。
- 名称相同的采集任务不支持被重复迁移。
- 名称相同的分类和标签不支持被重复迁移。
- 待导入的资源应为通过导出获取的zip文件，导入时系统会进行资源校验。
- 由于安全原因，导出连接时没有导出连接密码，需要在导入时自行输入。
- 仅企业版支持数据目录（分类、标签、采集任务）导出，专家版暂不支持。
- 导入文件时，OBS和本地方式均限制文件大小不超过10M。

旧空间导出资源

请您登录控制台首页，选择并进入旧工作空间的“管理中心”模块，然后执行如下操作进行资源导出。

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

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

步骤3 在管理中心页面，单击“资源迁移”，进入资源迁移页面。

图 3-1 资源迁移

步骤4 单击“新建导出”，配置文件的OBS存储位置和文件名称。

图 3-2 选择导出文件

步骤5 单击“下一步”，勾选导出的模块。

图 3-3 勾选导出的模块

步骤6 单击“下一步”，等待导出完成，资源包导出到所设置的OBS存储位置。

图 3-4 导出完成

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

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

图 3-5 下载导出结果

----结束

新空间导入资源

请您登录控制台首页，选择并进入新工作空间的“管理中心”模块，然后执行如下操作进行资源导入。

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

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

步骤3 在管理中心页面，单击“资源迁移”，进入资源迁移页面。

图 3-6 资源迁移

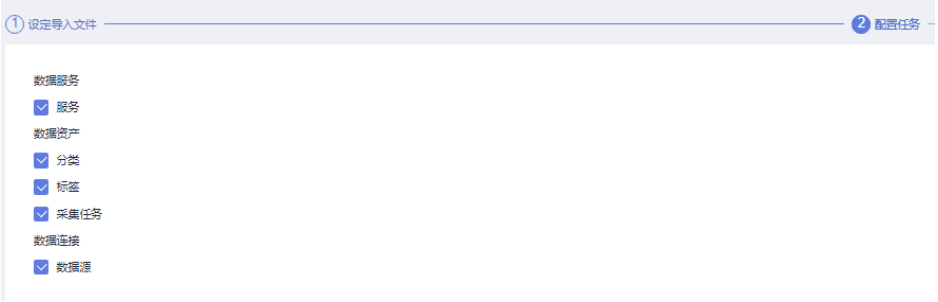
步骤4 单击“新建导入”，选择导入方式后，配置待导入资源的OBS或本地路径。待导入的资源应为通过导出获取的zip文件。

图 3-7 配置待导入的资源存储路径

步骤5 单击“新建导入”，上传待导入资源。待导入的资源应为通过导出获取的zip文件

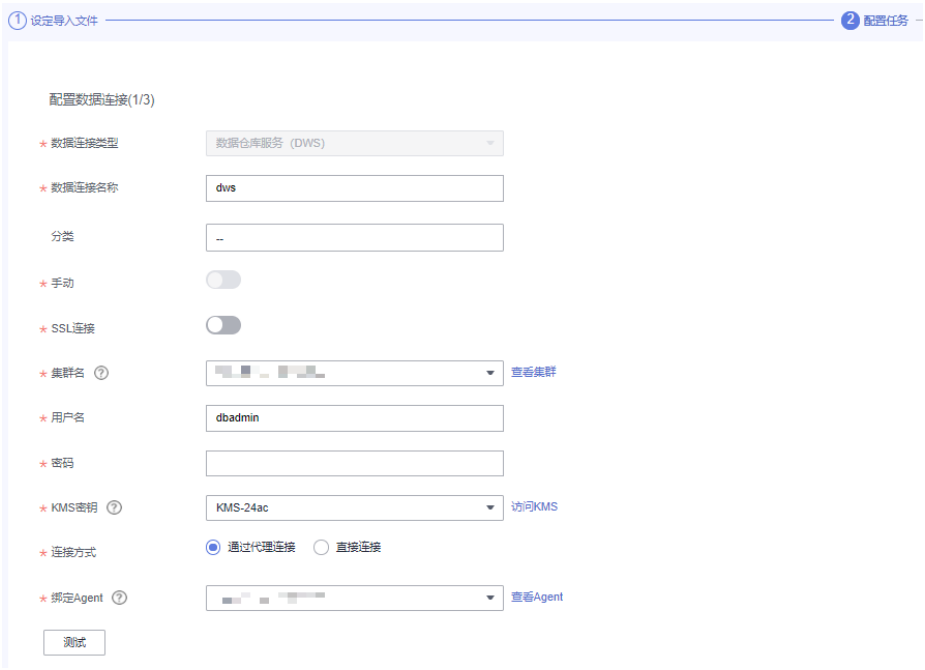
步骤6 单击“下一步”，勾选导入的资源类型。

图 3-8 勾选导入的资源类型



步骤7 如果选择导入数据源，则单击“下一步”需要配置数据连接。

图 3-9 配置数据连接



步骤8 单击“下一步”，等待导入任务下发，导入任务成功下发后系统提示“导入开始”。

图 3-10 导入开始

步骤9 系统提示“导入开始”后，单击“确定”，可在资源迁移任务列表中查看导入结果。

其中存在子任务失败时，可单击红色子任务名，查看失败原因。

图 3-11 查看导入结果

----结束

搬迁后验证

在新空间的资源导入完成后，您可以在新空间查看并验证如下导入资源是否与旧空间一致：

- 管理中心的数据连接。
- 数据目录的元数据采集任务，元数据的分类和标签。
- 数据服务中发布的API。

3.3 数据集成数据搬迁

数据集成数据搬迁依赖于CDM的批量导入导出作业功能。

CDM上支持导出的数据包括配置的连接和作业，可以导出到本地保存。

约束与限制

- 数据集成中的集群配置、环境变量等数据不支持导入导出，如有需要，请您进行手动配置同步。
- 由于安全原因，CDM不会将对应数据源的连接密码导出。因此在重新导入前，需要通过手工编辑导出的JSON文件补充密码或在导入窗口配置密码。
- 从本地导入数JSON文件时，导入文件大小不能超过1M。

旧空间导出作业和连接

请您登录控制台首页，选择并进入旧工作空间的“数据集成”模块，然后执行如下操作进行批量导出。

步骤1 在CDM主界面，单击左侧导航上的“集群管理”，单击集群“操作”列的“作业管理”，进入到“表/文件迁移”界面。

步骤2 单击作业列表上方的“导出”按钮，准备导出连接和作业。

图 3-12 批量导出

步骤3 在弹出的窗口中，选择“全部作业和连接”，单击“确认”，导出所有作业和连接。

图 3-13 全部导出

步骤4 导出成功后，通过浏览器下载地址，获取到导出的JSON文件。

----结束

新空间导入作业和连接

请您登录控制台首页，选择并进入新工作空间的“数据集成”模块，然后执行如下操作进行批量导入。

步骤1 在CDM主界面，单击左侧导航上的“集群管理”，单击集群“操作”列的“作业管理”，进入到“表/文件迁移”界面。

步骤2 单击作业列表上方的“导入”按钮，准备导入JSON文件。

图 3-14 批量导入

步骤3 在弹出的窗口中，选择导出作业获取的JSON文件，上传JSON文件。

图 3-15 选择 JSON 文件

步骤4 JSON文件上传成功后，单击“设置密码”，配置数据连接的密码或SK。

图 3-16 进入设置密码

步骤5 在设置密码弹窗中，依次输入各数据连接的密码或SK，完成后单击确认，回到导入作业界面。

图 3-17 设置密码

步骤6 在导入作业界面，单击确认，开始导入。

图 3-18 开始导入

步骤7 导入完成后，界面会显示导入情况。如果存在导入失败的情况，请您根据系统报错原因提示，调整后重新导入。

----结束

搬迁后验证

在新空间的作业和连接导入完成后，您可以在新空间查看并验证作业和连接是否与旧空间一致，以确保导入成功。

3.4 数据架构数据搬迁

数据架构数据搬迁依赖于数据架构的导入导出功能。

约束与限制

- 导入关系建模表/实体、维度建模维度/事实表、维度建模汇总表前请确保已创建管理中心连接，确保数据连接可用。
- 数据架构中的时间限定、审核中心和配置中心数据不支持导入导出。如有涉及，请您在其他数据迁移前，先进行手动配置同步。
- 数据架构支持最大导入文件大小为4Mb；支持最大导入指标个数为3000个；支持一次最大导出500张表。

旧空间导出表数据

请您登录控制台首页，选择并进入旧工作空间的“数据架构”模块，然后执行如下操作依次**导出流程**、**导出主题**、**导出码表**、**导出数据标准**、**导出关系建模表/实体**、**导出维度建模维度/事实表**、**导出业务指标**、**导出技术指标**、**导出维度建模汇总表**。

导出流程

步骤1 在数据架构主界面，单击左侧导航栏的“流程设计”，进入流程设计页面。

步骤2 单击列表上方的“导出”按钮，直接导出所有流程。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-19 导出流程

----结束

导出主题

步骤1 在数据架构主界面，单击左侧导航栏的“主题设计”，进入主题设计页面。

步骤2 单击列表上方的“更多 > 导出”，直接导出所有主题。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-20 导出主题

----结束

导出码表

步骤1 在数据架构主界面，单击左侧导航栏的“码表管理”，进入码表管理页面。

步骤2 选择需要导出的码表，然后单击列表上方的“更多 > 导出”按钮，导出所选码表。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-21 导出码表

----结束

导出数据标准

步骤1 在数据架构主界面，单击左侧导航栏的“数据标准”，进入数据标准页面。

步骤2 选择需要导出的数据标准，然后单击列表上方的“更多 > 导出”按钮，导出所选数据标准。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-22 导出数据标准

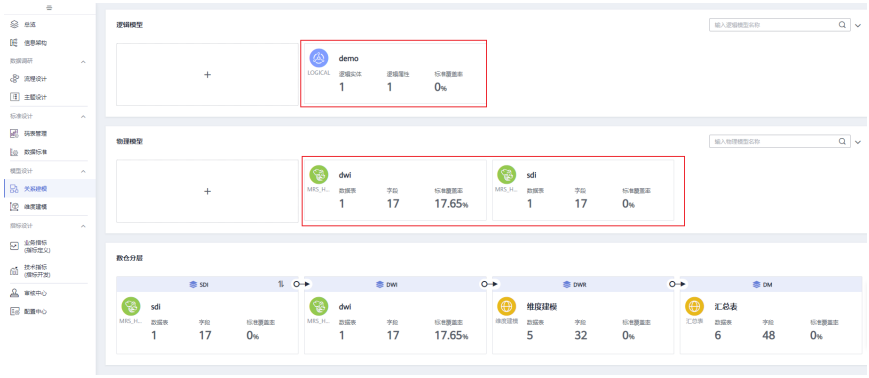
----结束

导出关系建模表/实体

步骤1 在数据架构主界面，单击左侧导航栏的“关系建模”，进入模型设计页面。

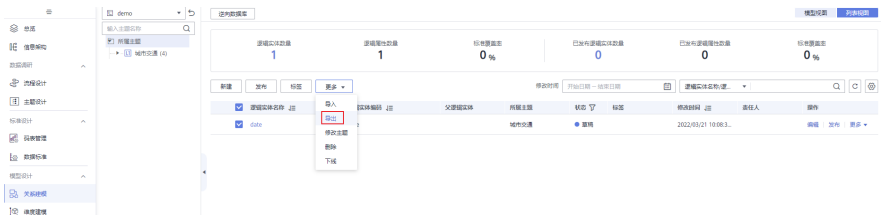
步骤2 进入任一需要导出的逻辑模型或物理模型，然后在模型内部导出表/实体。本例以进入逻辑模型demo为例进行说明。

图 3-23 进入模型内部



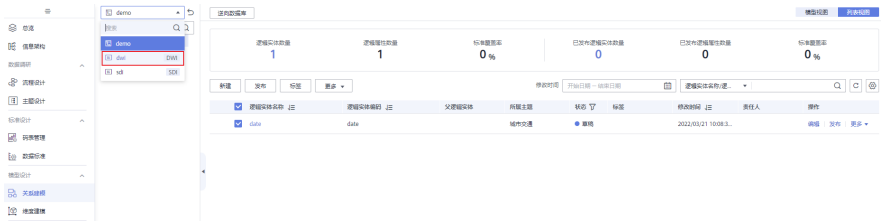
步骤3 在模型内部，然后选择所需导出的表/实体，单击列表上方的“更多 > 导出”按钮，导出所选关系建模表/实体，建议导出对象选择为“表”。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-24 导出关系建模表/实体



步骤4 在主题树上方，依次选择其他模型，进入模型后重复**步骤3**，依次下载其他模型的表/实体。

图 3-25 选择其他模型依导出



----结束

导出维度建模维度/事实表

步骤1 在数据架构主界面，单击左侧导航栏的“维度建模”，进入维度建模页面。

步骤2 在“维度”页签选择所需导出的维度，单击列表上方的“更多 > 导出”按钮，导出所选维度。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-26 导出维度

步骤3 选择“事实表”，然后选择所需导出的事实表，单击列表上方的“更多 > 导出”按钮，导出所选事实表。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-27 导出事实表

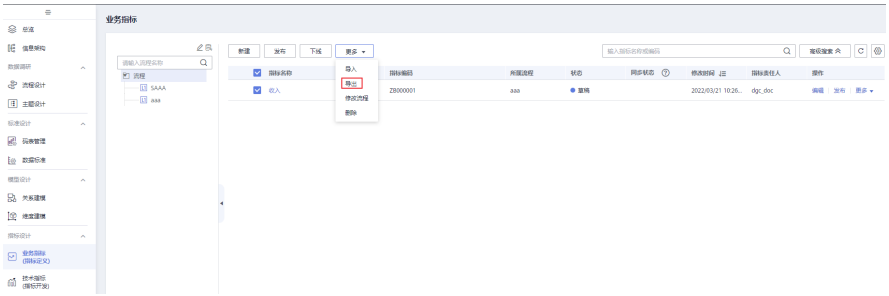
----结束

导出业务指标

步骤1 在数据架构主界面，单击左侧导航栏的“业务指标”，进入业务指标页面。

步骤2 选择所需导出的业务指标，单击列表上方的“更多 > 导出”按钮，导出所选业务指标。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-28 导出业务指标



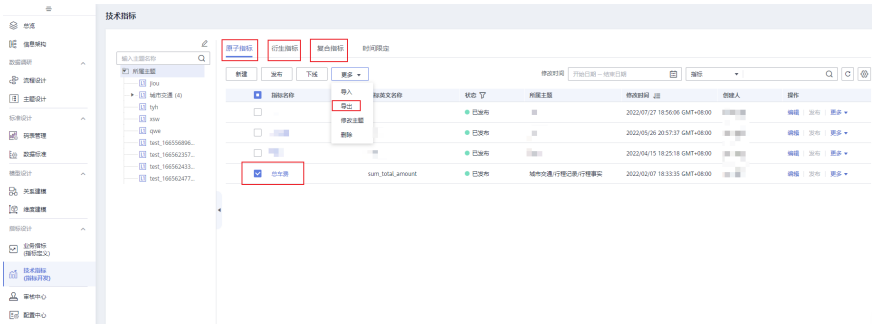
----结束

导出技术指标

步骤1 在数据架构主界面，单击左侧导航栏的“技术指标”，进入技术指标页面。

步骤2 分别进入“原子指标”、“衍生指标”和“复合指标”，选择所需导出的技术指标，单击列表上方的“更多 > 导出”按钮，导出所选技术指标。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-29 导出技术指标



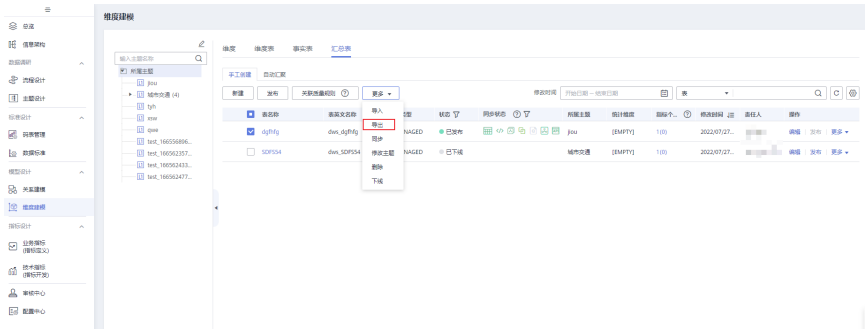
----结束

导出维度建模汇总表

步骤1 在数据架构主界面，单击左侧导航栏的“维度建模”，进入维度建模页面。

步骤2 选择“汇总表”，然后选择所需导出的汇总表，单击列表上方的“更多 > 导出”按钮，导出所选汇总表。导出完成后，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-30 导出汇总表



----结束

新空间导入表数据

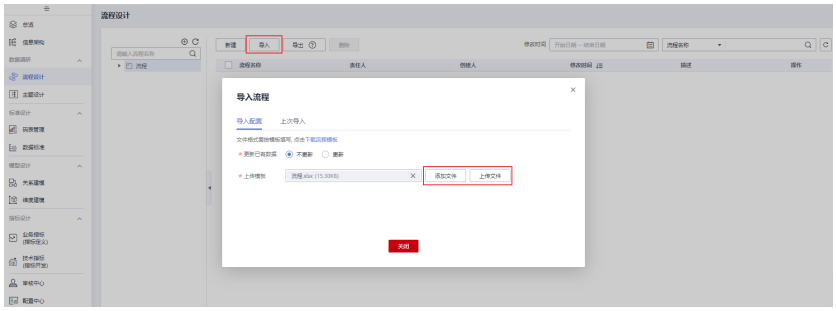
请您登录控制台首页，选择并进入新工作空间的“数据架构”模块，然后执行如下操作依次**导入流程**、**导入主题**、**导入码表**、**导入数据标准**、**导入关系建模表/实体**、**导入维度建模维度/事实表**、**导入业务指标**、**导入技术指标**、**导入维度建模汇总表**。

导入流程

步骤1 在数据架构主界面，单击左侧导航栏的“流程设计”，进入流程设计页面。

步骤2 单击列表上方的“导入”按钮，在弹出的导入窗口中，选择并上传需要导入的流程文件。

图 3-31 导入流程



步骤3 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-32 导入流程成功

----结束

导入主题

步骤1 在数据架构主界面，单击左侧导航栏的“主题设计”，进入主题设计页面。

步骤2 单击列表上方的“更多 > 导入”按钮，在弹出的导入窗口中，选择并上传需要导入的主题文件。

图 3-33 导入主题

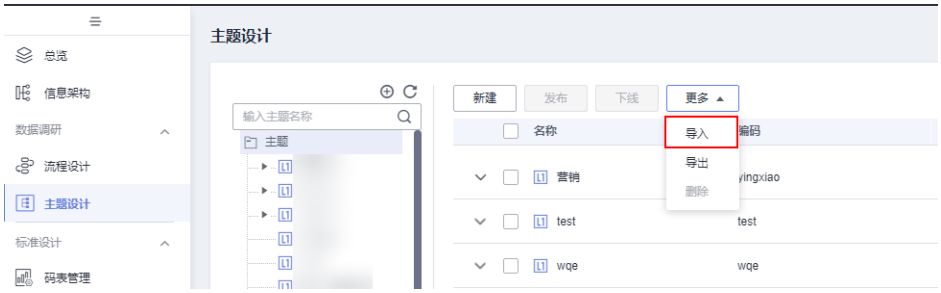
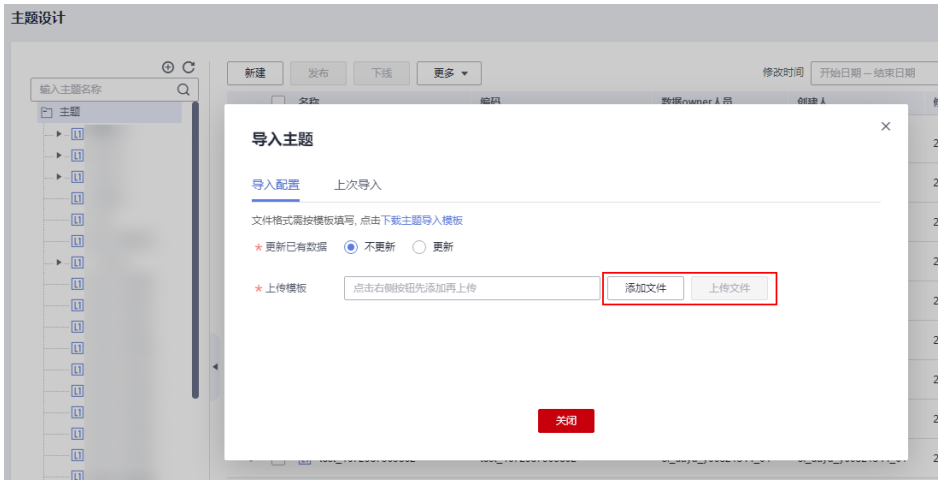


图 3-34 添加文件



步骤3 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-35 导入主题成功

步骤4 导入成功后，请单击“发布”，使其处于“已发布”状态。

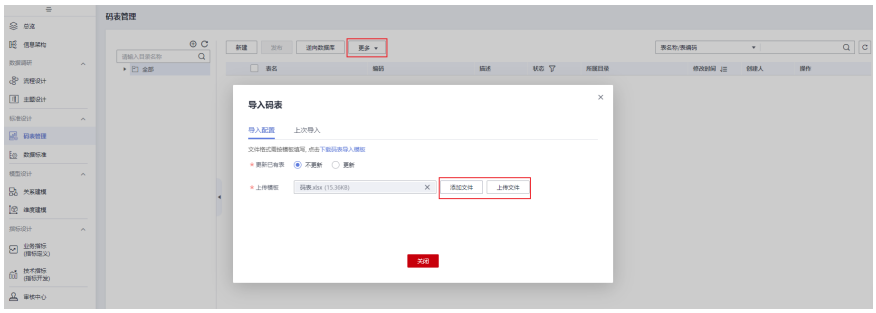
----结束

导入码表

步骤1 在数据架构主界面，单击左侧导航栏的“码表管理”，进入码表管理页面。

步骤2 单击列表上方的“更多 > 导入”按钮，在弹出的导入窗口中，选择并上传需要导入的码表文件。

图 3-36 导入码表



步骤3 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-37 导入码表成功

步骤4 导入成功后，请单击“发布”，使其处于“已发布”状态。

----结束

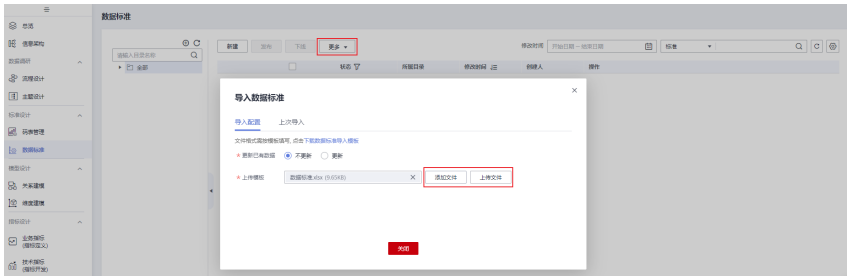
导入数据标准

步骤1 在数据架构主界面，单击左侧导航栏的“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，会显示制定数据标准模板的页面，请参考旧空间的“配置中心 > 标准模板管理”页面，修改新空间数据标准模板，完成后单击“确定”。

步骤3 单击列表上方的“更多 > 导入”按钮，在弹出的导入窗口中，选择并上传需要导入的数据标准文件。

图 3-38 导入数据标准



步骤4 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-39 导入数据标准成功

步骤5 导入成功后，请单击“发布”，使其处于“已发布”状态。

----结束

导入关系建模表/实体

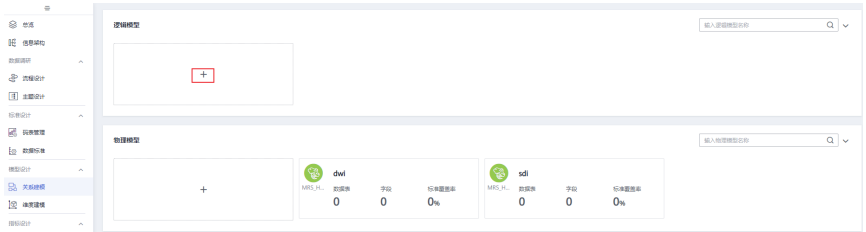
步骤1 在数据架构主界面，单击左侧导航栏的“关系建模”，进入模型设计页面。

步骤2 在“关系建模”页面，如果当前未创建过关系模型，系统会弹出“新建分层治理模型”提示框。请参考旧空间的“关系建模”页面，建立新空间的SDI和DWI层模型，完成后单击“确定”。

图 3-40 新建分层治理模型

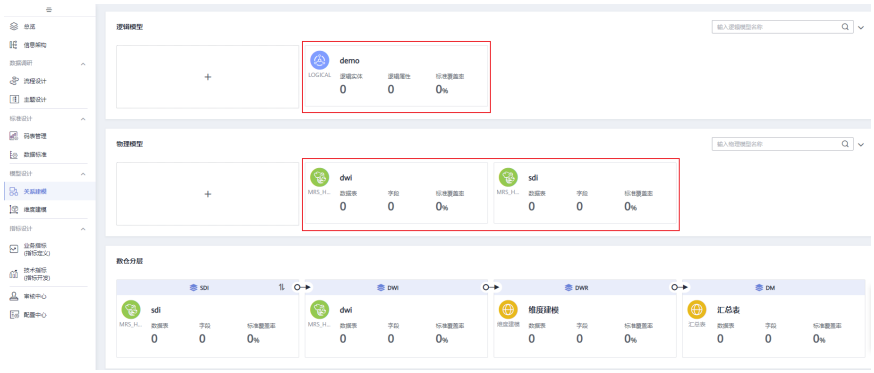
步骤3 如果旧空间还新建有“逻辑模型”，请单击+按钮新建模型。

图 3-41 新建逻辑模型



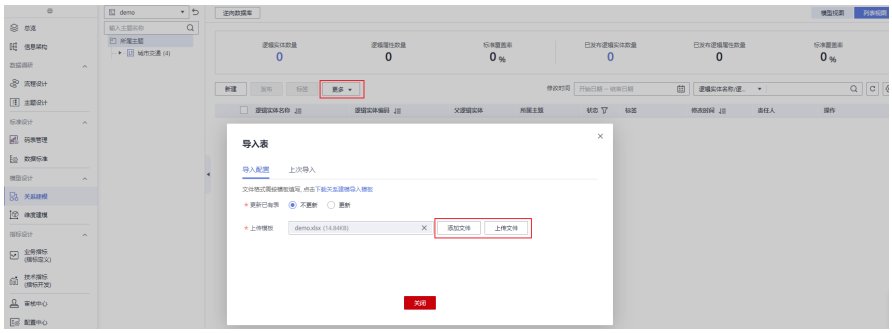
步骤4 进入任一需要导入的逻辑模型或物理模型，然后在模型内部导入表/实体。本例以进入逻辑模型demo为例进行说明。

图 3-42 进入模型内部



步骤5 在模型内部，单击列表上方的“更多 > 导入”按钮，在弹出的导入窗口中，选择并上传需要导入的表/实体文件。

图 3-43 导入关系建模表/实体



步骤6 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-44 导入关系建模表/实体成功

步骤7 在主题树上方，依次选择其他模型，进入模型后重复**步骤5**~**步骤6**，依次导入其他模型的表/实体。

图 3-45 选择其他模型依次导入

步骤8 导入成功后，请单击“发布”，使其处于“已发布”状态。

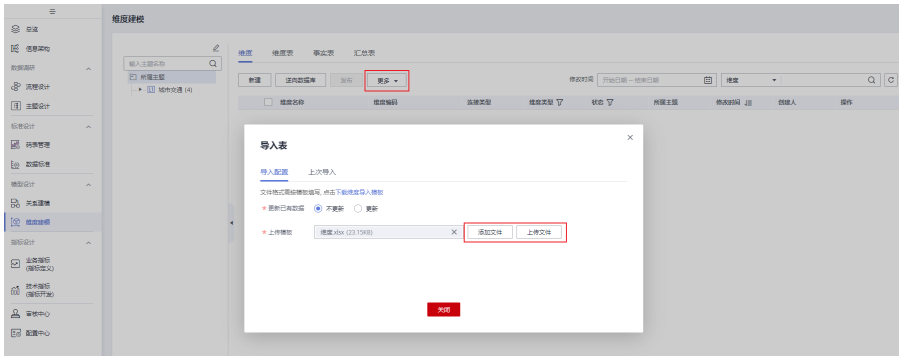
----结束

导入维度建模维度/事实表

步骤1 在数据架构主界面，单击左侧导航栏的“维度建模”，进入维度建模页面。

步骤2 选择“维度”，单击列表上方的“更多 > 导入”按钮，在弹出的导入窗口中，选择并上传需要导入的维度文件。

图 3-46 导入维度

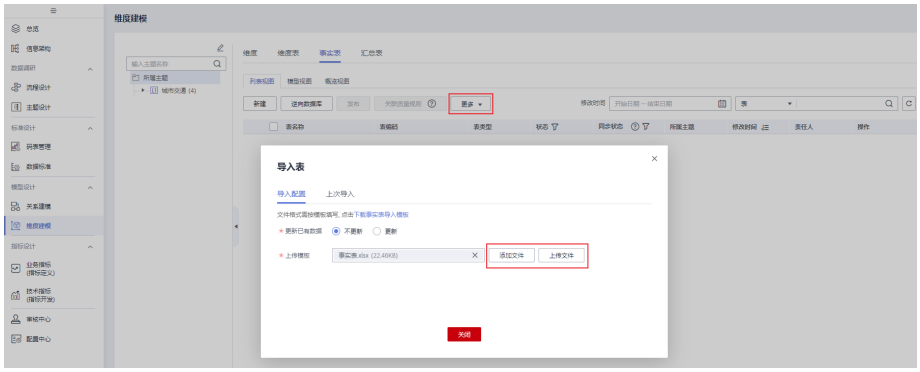


步骤3 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-47 导入维度成功

步骤4 选择“事实表”，单击列表上方的“更多 > 导入”按钮，在弹出的导入窗口中，选择并上传需要导入的事实表文件。

图 3-48 导入事实表



步骤5 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-49 导入事实表成功

步骤6 导入成功后，请单击“发布”，使其处于“已发布”状态。

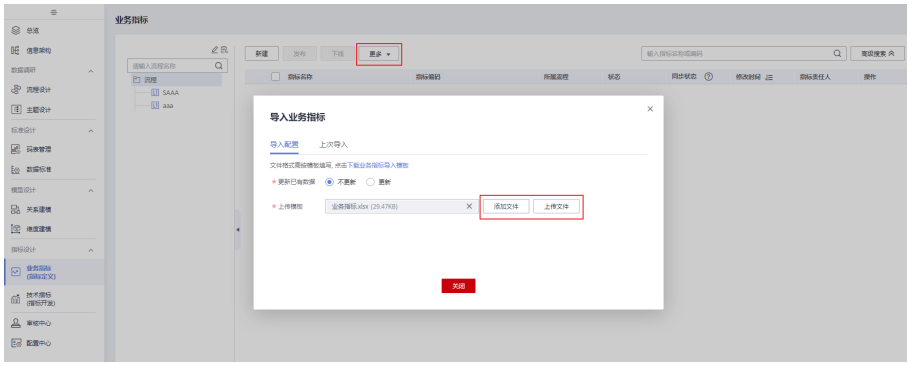
----结束

导入业务指标

步骤1 在数据架构主界面，单击左侧导航栏的“业务指标”，进入业务指标页面。

步骤2 单击列表上方的“更多 > 导入”按钮，在弹出的导入窗口中，选择并上传需要导入的业务指标文件。

图 3-50 导入业务指标



步骤3 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-51 导入业务指标成功

步骤4 导入成功后，请单击“发布”，使其处于“已发布”状态。

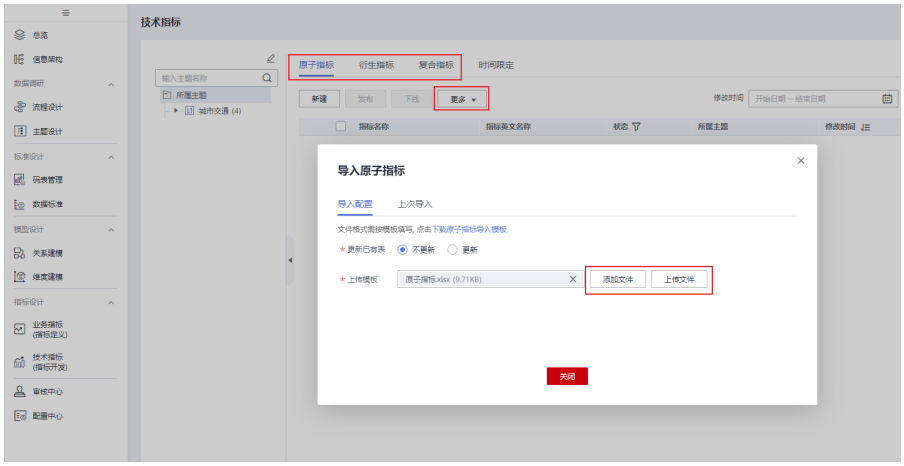
----结束

导入技术指标

步骤1 在数据架构主界面，单击左侧导航栏的“技术指标”，进入业务指标页面。

步骤2 分别进入“原子指标”、“衍生指标”和“复合指标”，单击列表上方的“更多 > 导入”按钮，在弹出的导入窗口中，选择并上传需要导入的技术指标文件。

图 3-52 导入技术指标



步骤3 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-53 导入技术指标成功

步骤4 导入成功后，请单击“发布”，使其处于“已发布”状态。

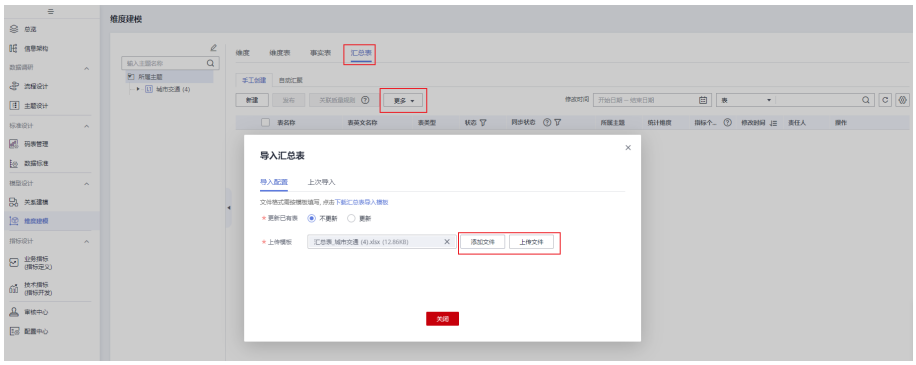
----结束

导入维度建模汇总表

步骤1 在数据架构主界面，单击左侧导航栏的“维度建模”，进入维度建模页面。

步骤2 选择“汇总表”，单击列表上方的“更多 > 导入”按钮，在弹出的导入窗口中，选择并上传需要导入的汇总表文件。

图 3-54 导入汇总表



步骤3 上传文件后系统开始自动导入，导入成功后系统会显示导入的情况。

图 3-55 导入汇总表成功

步骤4 导入成功后，请单击“发布”，使其处于“已发布”状态。

----结束

搬迁后验证

在新空间的表数据导入完成后，您可以在新空间查看并验证模型和表数据等是否与旧空间一致，以确保导入成功。

3.5 数据开发数据搬迁

数据开发数据搬迁依赖于数据开发的脚本、作业、环境变量、资源导入导出功能。

约束与限制

- 已完成**管理中心数据搬迁**。
- 数据开发中的通知配置、备份管理、作业标签、委托配置、默认项等数据不支持导入导出，如有涉及，请您进行手动配置同步。

- 导入脚本、作业、环境变量、资源功能部分依赖于OBS服务。

旧空间导出数据

请您登录控制台首页，选择并进入旧工作空间的“数据开发”模块，然后执行如下操作依次**导出脚本**、**导出作业**、**导出环境变量**、**导出资源**。

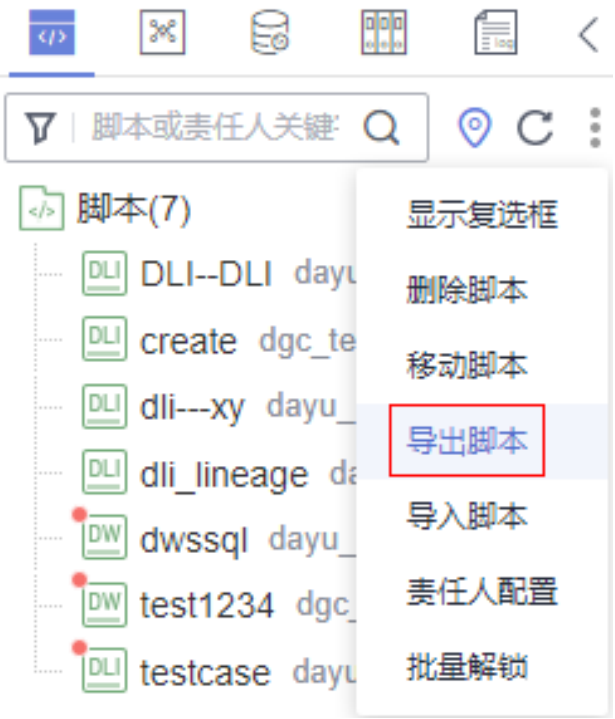
导出脚本

步骤1 在数据开发主界面，单击左侧导航上的“脚本开发”，进入脚本目录。

步骤2 单击脚本目录中的，选择“显示复选框”。

步骤3 勾选需要导出的脚本，单击 > 导出脚本。导出完成后，即可通过浏览器下载地址，获取到导出的zip文件。

图 3-56 选择并导出脚本



步骤4 在弹出的“导出脚本”界面，选择需要导出的脚本的状态，单击“确定”。

图 3-57 导出脚本

----结束

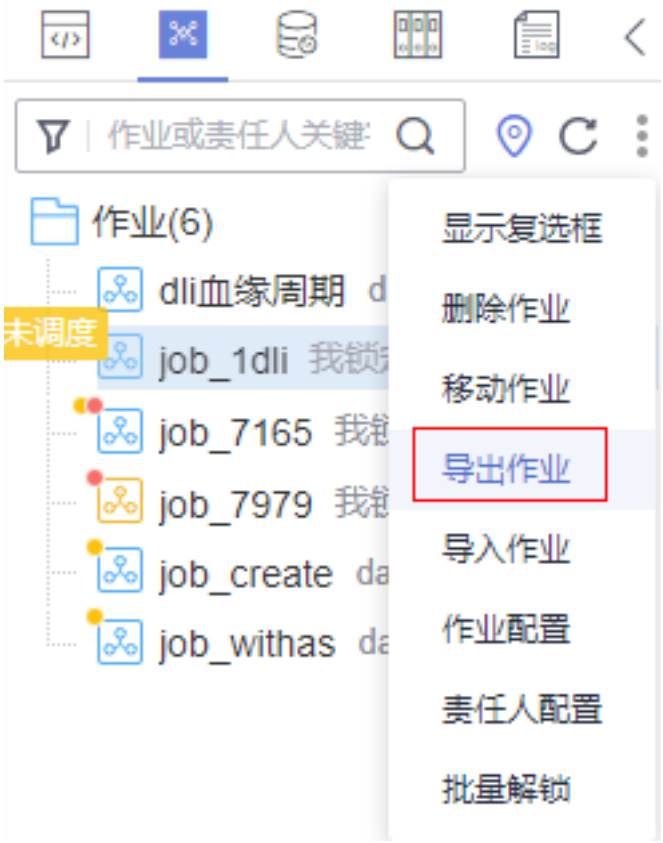
导出作业

步骤1 单击脚本目录树上方的，切换到作业界面。

步骤2 单击作业目录中的，选择“显示复选框”。

步骤3 勾选需要导出的作业，单击 > 导出作业，可选择“只导出作业”或“导出作业及其依赖脚本和资源定义”。导出完成后，即可通过浏览器下载地址，获取到导出的zip文件。

图 3-58 选择并导出作业



步骤4 在弹出的“导出作业”界面，选择需要导出的作业范围和状态，单击“确定”，可以在下载中心查看导入结果。

图 3-59 导出作业

----结束

导出环境变量

步骤1 单击左侧导航上的“配置”，进入环境变量页面。

步骤2 单击环境变量配置下的“导出”，导出环境变量。

图 3-60 导出环境变量

----结束

导出资源

步骤1 单击左侧导航上的“资源管理”，进入资源管理页面。

步骤2 单击资源目录中的 ，选择“显示复选框”。

图 3-61 显示资源复选框

步骤3 勾选需要导出的资源，单击 > 导出资源。导出完成后，即可通过浏览器下载地址，获取到导出的zip文件。

图 3-62 选择并导出资源

----结束

新空间导入数据

请您登录控制台首页，选择并进入新工作空间的“数据开发”模块，然后执行如下操作依次**导入资源**、**导入环境变量**、**导入脚本**、**导入作业**。

导入资源

步骤1 在数据开发主界面，单击左侧导航上的“资源管理”，进入资源管理页面。

步骤2 单击资源目录中的，选择“导入资源”。

图 3-63 选择导入资源

步骤3 在弹出的导入资源窗口中，“文件位置”选择为“本地”，选择从旧空间导出的资源文件，“重名处理策略”默认选择“覆盖”，单击下一步。

图 3-64 导入资源

步骤4 资源开始导入，导入成功后系统会显示导入的资源名。

图 3-65 导入资源成功

----结束

导入环境变量

步骤1 单击左侧导航上的“配置”，进入环境变量页面。

步骤2 单击环境变量配置下的“导入”，导入环境变量。

图 3-66 选择导入环境变量

步骤3 在弹出的导入环境变量窗口中，“文件位置”选择为“本地”，选择从旧空间导出的环境变量文件，“重名处理策略”默认选择“覆盖”，单击下一步。

图 3-67 导入环境变量

步骤4 环境变量开始导入，导入前系统会提示是否要修改变量值，确定后环境变量即可导入成功。

图 3-68 导入结果确认

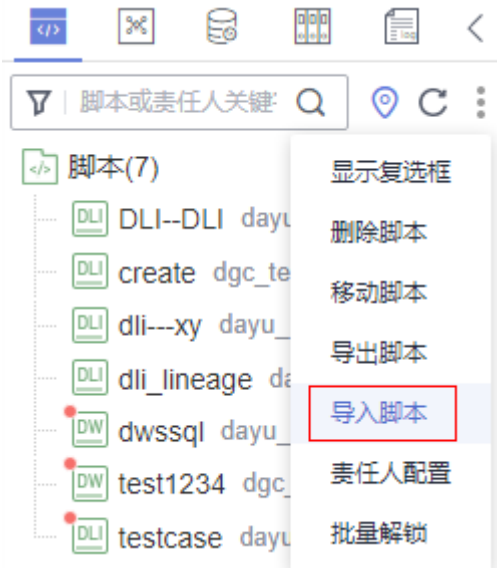
----结束

导入脚本

步骤1 单击左侧导航上的“脚本开发”，进入脚本目录。

步骤2 单击脚本目录中的，选择“导入脚本”。

图 3-69 选择导入脚本



步骤3 在弹出的导入脚本窗口中，“文件位置”选择为“本地”，选择从旧空间导出的脚本文件，“重名处理策略”默认选择“覆盖”，单击下一步。

图 3-70 导入脚本

步骤4 脚本开始导入，导入成功后系统会显示导入的脚本名。

图 3-71 导入脚本成功

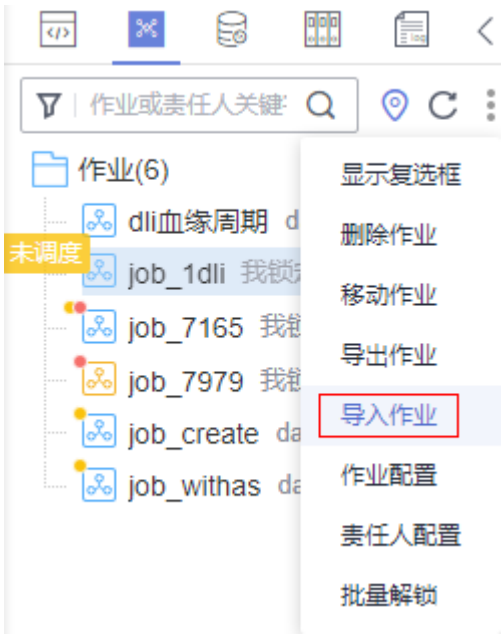
----结束

导入作业

步骤1 单击脚本目录树上方的，切换到作业界面。

步骤2 单击作业目录中的，选择“导入作业”。

图 3-72 选择导入作业



步骤3 在弹出的导入作业窗口中，“文件位置”选择为“本地”，选择从旧空间导出的作业文件，单击下一步。

图 3-73 导入作业

步骤4 作业开始导入，导入成功后系统会显示导入的作业名。

图 3-74 导入作业成功

----结束

搬迁后验证

在新空间的脚本、作业、环境变量、资源数据导入完成后，您可以在新空间查看并验证这些数据是否与旧空间一致，以确保导入成功。

3.6 数据质量数据搬迁

数据质量数据搬迁依赖于数据质量监控的规则模板、质量作业、对账作业导入导出功能。

约束与限制

- 已完成**管理中心数据搬迁**。
- 业务指标监控中的指标、规则、业务场景等数据均不支持导入导出，如有涉及，请您进行手动配置同步。
- 系统支持将自定义的规则模板批量导出，一次最多可导出200个规则模板。
- 系统支持将自定义的规则模板批量导入，一次最大可导入4M数据的文件。
- 系统支持批量导出质量作业，一次最多可导出200个质量作业。导出作业时，导出的单元格内容最大长度支持65534个字符。
- 系统支持批量导入质量作业，一次最大可导入4M数据的文件。导入作业时，导入的单元格内容最大长度支持65534个字符。

- 系统支持批量导出对账作业，一次最多可导出200个对账作业。导出作业时，导出的单元格内容最大长度支持65534个字符。
- 系统支持批量导入对账作业，一次最大可导入4M数据的文件。导入作业时，导出的单元格内容最大长度支持65534个字符。

旧空间导出数据

请您登录控制台首页，选择并进入旧工作空间的“数据质量”模块，然后执行如下操作依次**导出规则模板**、**导出质量作业**、**导出对账作业**。

导出规则模板

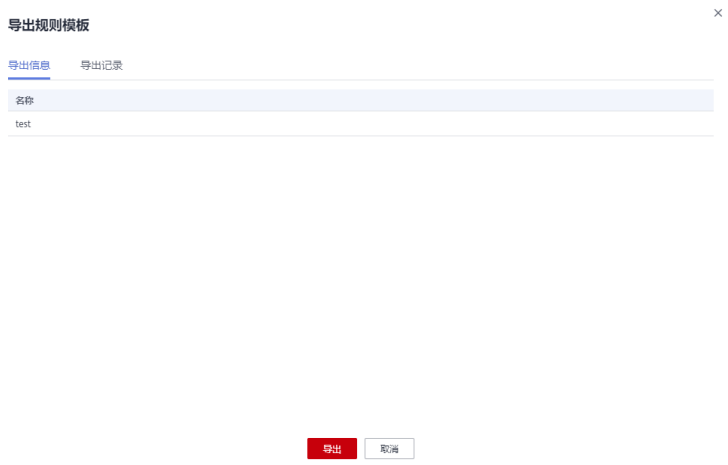
步骤1 在数据质量主界面，单击左侧导航上的“规则模板”，进入规则模板列表。

步骤2 在规则模板列表，选择自定义的规则模板，然后单击“导出”。

图 3-75 批量导出规则模板

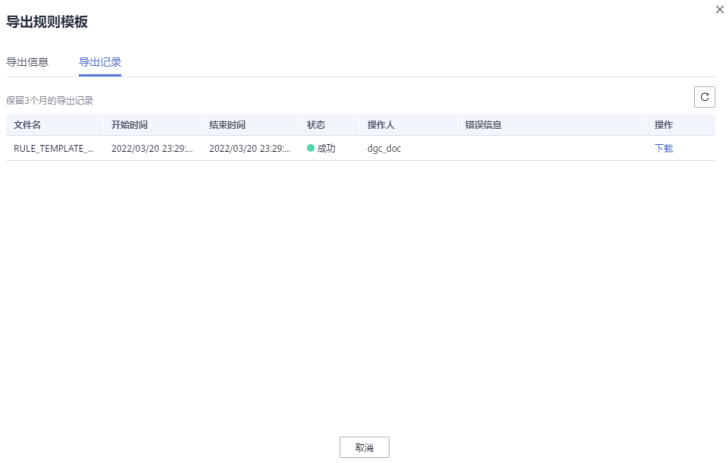
步骤3 在弹出的导出窗口中，确认选择无误后单击“导出”，导出规则模板。

图 3-76 规则模板导出确认



步骤4 导出成功后，在导出记录中单击“下载”，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-77 获取规则模板导出结果



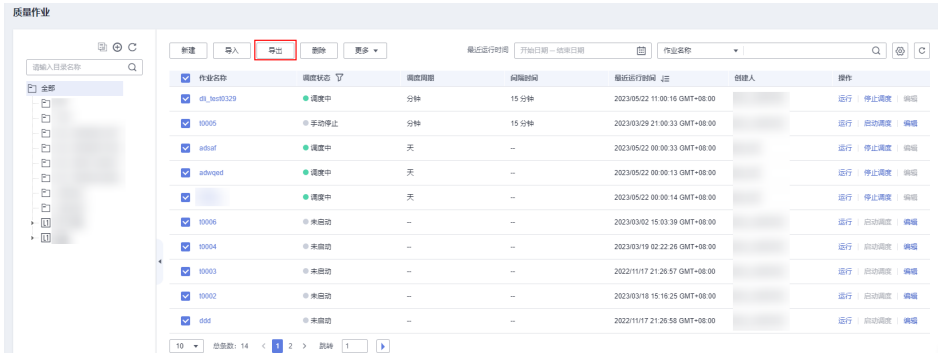
----结束

导出质量作业

步骤1 单击左侧导航上的“质量作业”，进入质量作业列表。

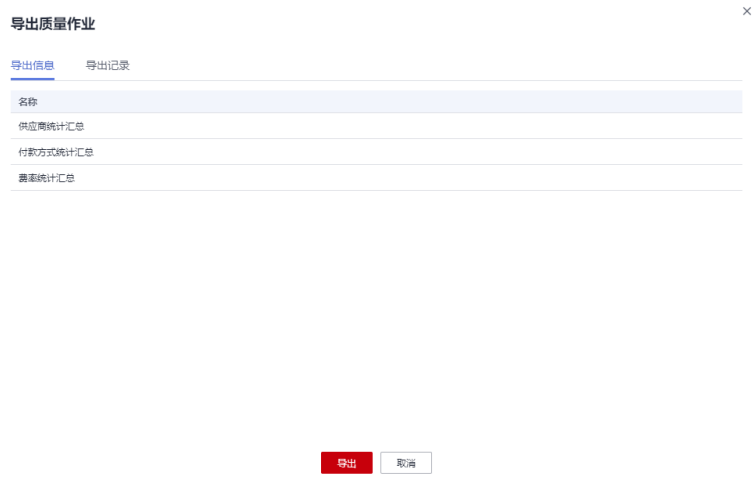
步骤2 在质量作业列表，选择需要迁移的质量作业，然后单击“导出”。

图 3-78 批量导出质量作业



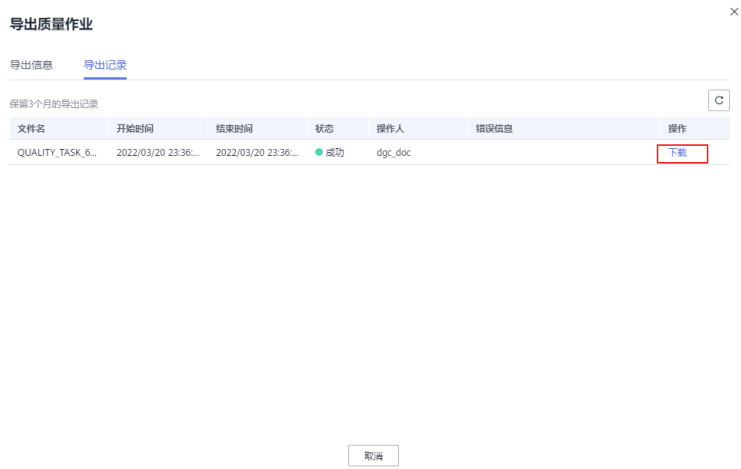
步骤3 在弹出的导出窗口中，确认选择无误后单击“导出”，导出质量作业。

图 3-79 质量作业导出确认



步骤4 导出成功后，在导出记录中单击“下载”，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-80 获取质量作业导出结果



----结束

导出对账作业

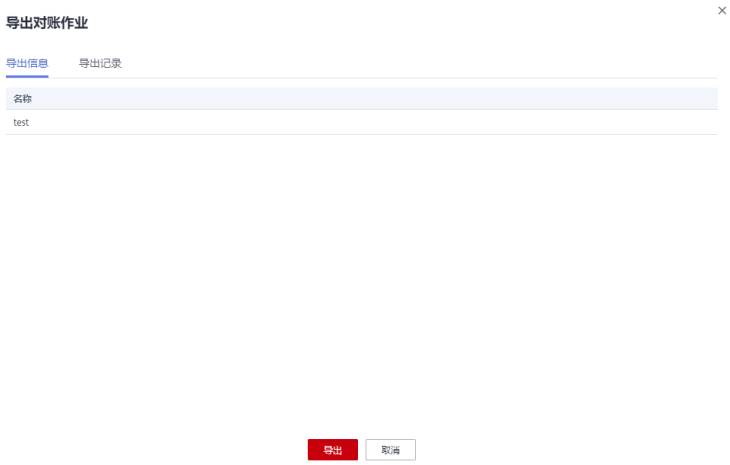
步骤1 单击左侧导航上的“对账作业”，进入对账作业列表。

步骤2 在对账作业列表，选择需要迁移的对账作业，然后单击“导出”。

图 3-81 批量导出对账作业

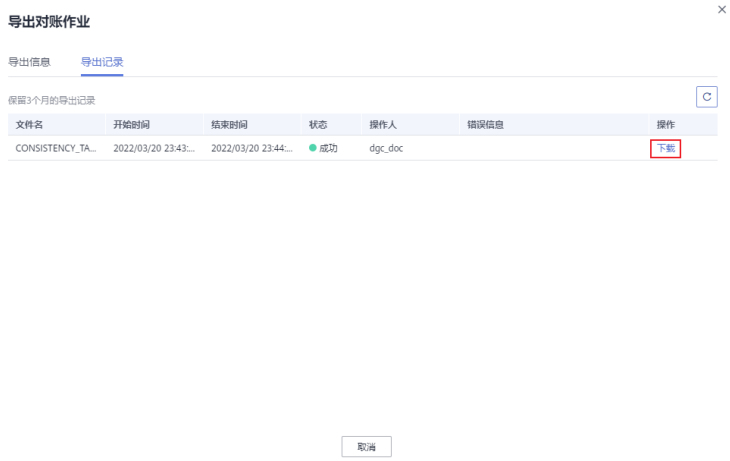
步骤3 在弹出的导出窗口中，确认选择无误后单击“导出”，导出对账作业。

图 3-82 对账作业导出确认



步骤4 导出成功后，在导出记录中单击“下载”，即可通过浏览器下载地址，获取到导出的xlsx文件。

图 3-83 获取对账作业导出结果



----结束

新空间导入数据

请您登录控制台首页，选择并进入新工作空间的“数据质量”模块，然后执行如下操作依次**导入规则模板**、**导入质量作业**、**导入对账作业**。

导入规则模板

步骤1 在数据质量主界面，单击左侧导航上的“规则模板”，进入规则模板列表。

步骤2 在规则模板列表，单击“导入”。

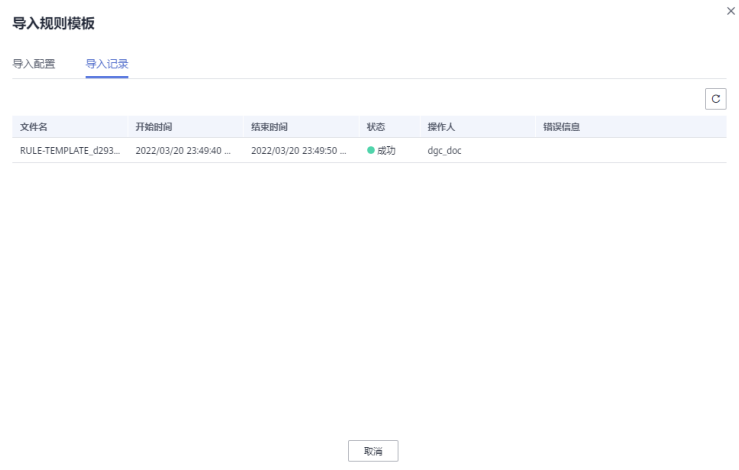
图 3-84 批量导入规则模板

步骤3 在弹出的导入窗口中，选择从旧空间导出的规则模板文件，然后选择目录的映射路径，“重名处理策略”默认选择“终止”，最后单击“导入”。

图 3-85 规则模板导入

步骤4 在导入记录中，可查看导入状态，显示为成功后即成功导入。

图 3-86 查看规则模板导入结果



----结束

导入质量作业

步骤1 单击左侧导航上的“质量作业”，进入质量作业列表。

步骤2 在质量作业列表，单击“导入”。

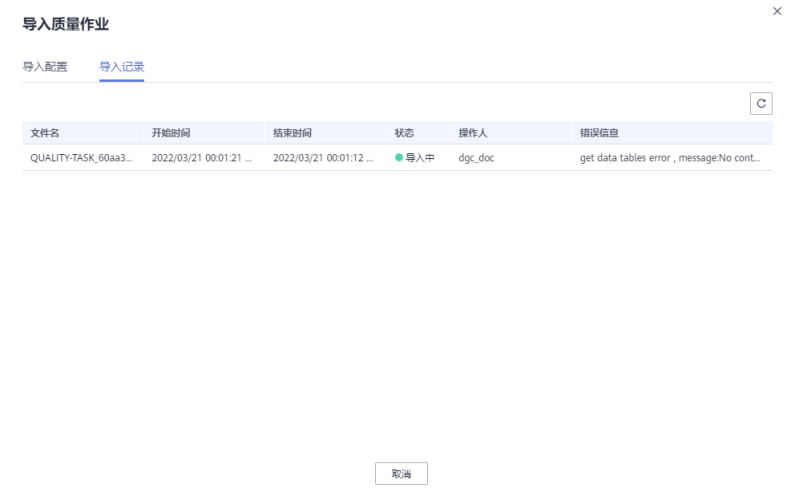
图 3-87 批量导入质量作业

步骤3 在弹出的导入窗口中，选择从旧空间导出的质量作业文件，然后选择数据连接、集群和目录等的映射路径，“重名处理策略”默认选择“终止”，最后单击“导入”。

图 3-88 质量作业导入

步骤4 在导入记录中，可查看导入状态，显示为成功后即成功导入。

图 3-89 查看质量作业导入结果



----结束

导入对账作业

步骤1 单击左侧导航上的“对账作业”，进入对账作业列表。

步骤2 在对账作业列表，单击“导入”。

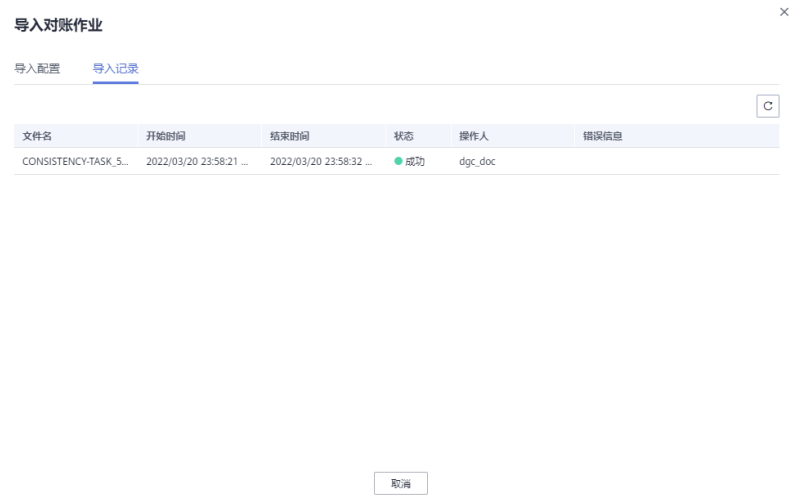
图 3-90 批量导入对账作业

步骤3 在弹出的导入窗口中，选择从旧空间导出的对账作业文件，然后选择数据连接、集群和目录等的映射路径，“重名处理策略”默认选择“终止”，最后单击“导入”。

图 3-91 对账作业导入

步骤4 在导入记录中，可查看导入状态，显示为成功后即成功导入。

图 3-92 查看对账作业导入结果



----结束

搬迁后验证

在新空间的规则模板、质量作业、对账作业导入完成后，您可以在新空间查看并验证规则模板、质量作业、对账作业是否与旧空间一致，以确保导入成功。

3.7 数据目录数据搬迁

数据目录数据搬迁依赖于管理中心的资源迁移功能，详见**管理中心数据搬迁**。

说明

当前管理中心支持搬迁的数据目录数据包含分类、标签、采集任务，数据目录中的业务资产、技术资产、指标资产均不支持直接导入导出。

您可以通过导入管理中心和数据架构数据，并运行新导入的采集任务重新生成业务资产、技术资产、指标资产。

3.8 数据安全数据搬迁

当前暂不支持数据安全数据的导入导出，需要您手动同步各项配置数据和任务。

3.9 数据服务数据搬迁

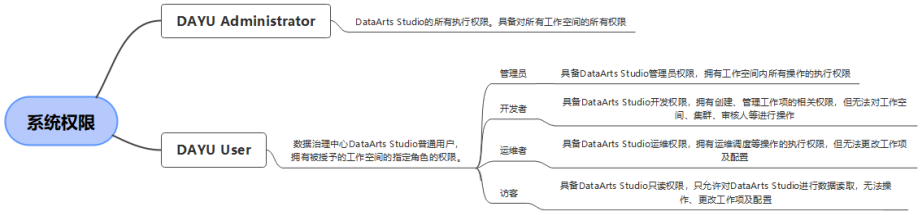
数据服务数据搬迁依赖于管理中心的资源迁移功能，详见**管理中心数据搬迁**。

4 如何最小化授权用户使用 DataArts Studio

实践场景及目标

某数据运营工程师专职负责数据质量监控相关工作，仅需要服务数据质量组件的操作权限。

图 4-1 权限体系



服务的权限体系如图4-1所示。如果项目管理员直接赋予该数据运营工程师IAM账号“DAYU User系统角色+工作空间开发者角色”权限，则会出现如下非必需权限过大的风险：

1. 依赖服务权限过大：服务作为平台型服务，DAYU User系统角色预置了依赖服务（如MRS、DWS等相关服务）的管理员权限。当为数据运营工程师IAM账号授予DAYU User系统角色后，会导致其拥有依赖服务的管理员权限。

为了解决此问题，项目管理员可以按照如下解决方案进行权限最小化配置，这样既能满足实际业务使用，也避免了权限过大的风险。

1. 为数据运营工程师IAM账号授予DAYU User系统角色权限，然后删除IAM账号中的依赖服务权限，再赋予依赖服务的最小权限合集。

操作流程

1. **创建用户组并授予系统角色DAYU User**：创建数据运营工程师IAM账号所在的用户组，并授予**DAYU User**权限。
2. **去除用户组依赖服务权限并配置最小权限合集**：为用户组去除默认的依赖服务管理员权限，然后配置最小权限。
3. **创建IAM用户并加入用户组**：为数据运营工程师创建IAM账号，并加入到用户组中。

4. **添加工作空间成员并配置角色**：将新创建的IAM用户加入到工作空间并配置角色。
5. **用户登录并验证权限**：使用新创建的用户登录控制台，验证权限配置是否符合预期。

创建用户组并授予系统角色 DAYU User

步骤1 使用华为账号登录统一身份认证服务IAM控制台。

步骤2 在IAM服务控制台中，单击“用户组”，在用户组页面单击右上方的“创建用户组”。

图 4-2 创建用户组

步骤3 在“创建用户组”界面，输入“用户组名称”DQC。

图 4-3 用户组名称

步骤4 单击“确定”，用户组创建完成，用户组列表中显示新创建的用户组。

说明

您最多可以创建20个用户组，如果当前资源配额无法满足业务需要，您可以申请扩大配额，具体方法请参见：**如何申请扩大配额?**

步骤5 在用户组列表中，单击新建用户组右侧的“授权”。

图 4-4 进入用户组权限设置页面

步骤6 在搜索框中输入DAYU User，勾选该系统角色，单击“下一步”。

图 4-5 角色授权

说明

请勿勾选“**DAYU Administrator**”权限，“**DAYU Administrator**”权限具有DataArts Studio服务的所有执行权限，不受工作空间权限管控。

步骤7 授权范围方案选择需要授予的区域项目，单击“确定”，完成授权。

说明

DataArts Studio部署时通过物理区域划分，为项目级服务。授权时，“授权范围方案”如果选择“所有资源”，则该权限在所有区域项目中都生效；如果选择“指定区域项目资源”，则该权限仅对此项目生效。IAM用户授权完成后，访问DataArts Studio时，需要先切换至授权区域。

图 4-6 设置最小授权范围

----结束

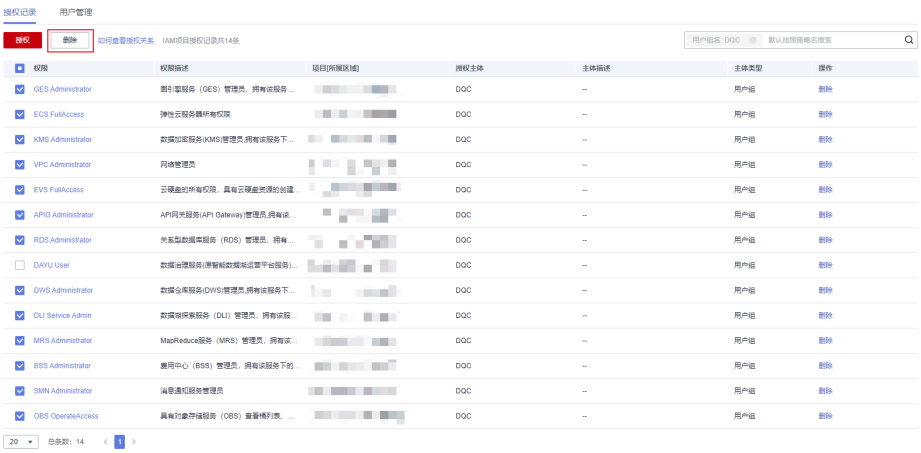
去除用户组依赖服务权限并配置最小权限合集

步骤1 在IAM服务控制台中，单击“用户组”，在用户组页面单击创建的DQC用户组名，进入用户组详情页面。

图 4-7 进入用户组详情

步骤2 在用户组详情页面下方的授权记录区域，条数切换到20条，展开所有14条授权记录。勾选除DAYU User外的所有依赖服务权限，并单击列表上方的删除。

图 4-8 删除依赖服务权限



步骤3 依赖服务权限删除成功后，返回IAM服务控制台首页，单击“权限管理 > 权限”，在权限页面单击右上方的“创建自定义策略”。

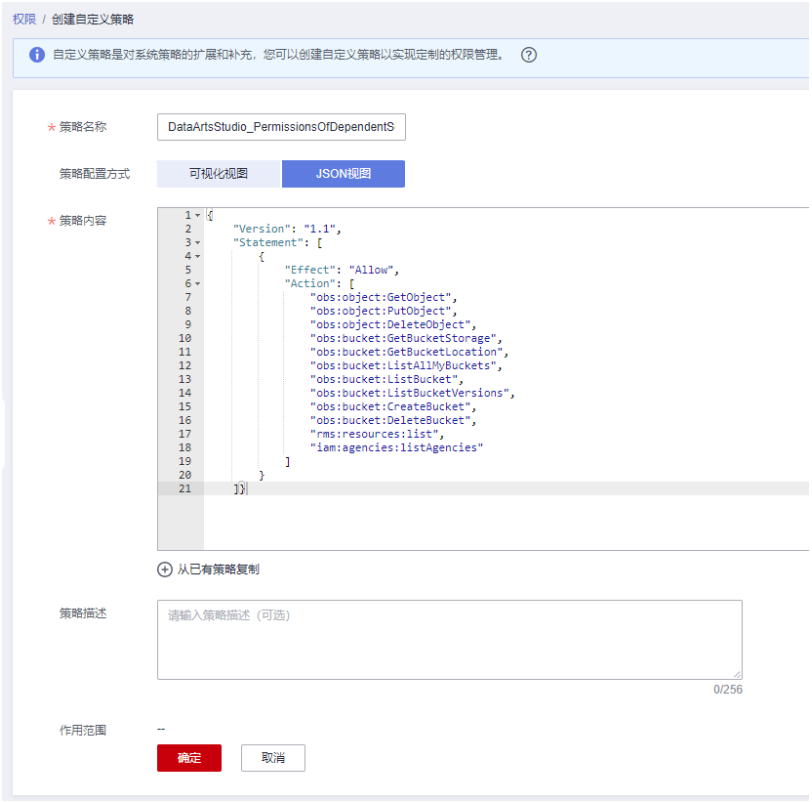
图 4-9 创建自定义策略

步骤4 在自定义策略配置页面，策略配置方式切换至JSON视图，然后按照如下策略内容，分别创建**DataArtsStudio_PermissionsOfDependentServices_global**和**DataArtsStudio_PermissionsOfDependentServices_region**自定义策略。

说明

- 创建自定义策略时，暂不支持同时选全局级云服务和项目级云服务，因此需要将依赖服务自定义策略拆分为两条分别创建。
- 策略内容来自于DataArts Studio服务各组件功能所需依赖服务的最小权限，详情请参见**权限管理**。

图 4-10 创建自定义策略示例



- 依赖的全局级（global级）云服务的自定义策略 **DataArtsStudio_PermissionsOfDependentServices_global**：

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "obs:object:GetObject",
                "obs:object:PutObject",
                "obs:object:DeleteObject",
                "obs:bucket:GetBucketStorage",
                "obs:bucket:GetBucketLocation",
                "obs:bucket:ListAllMyBuckets",
                "obs:bucket:ListBucket",
                "obs:bucket:ListBucketVersions",
                "obs:bucket:CreateBucket",
                "obs:bucket:DeleteBucket",
                "rms:resources:list",
                "iam:agencies:listAgencies"
            ]
        }
    ]
}
```

- 依赖的项目级（region级）云服务的自定义策略 **DataArtsStudio_PermissionsOfDependentServices_region**：

```
{
   "Version": "1.1",
   "Statement": [
      {
         "Effect": "Allow",
         "Action": [
            "cdm:cluster:get",
            "cdm:cluster:list",
            "cdm:cluster:create",
            "cdm:link:operate",
            "cdm:job:operate",
            "ces:*:get",
            "ces:*:list",
            "cloudtable:*:get",
            "cloudtable:*:list",
      "css:*:get",
            "css:*:list",
            "dis:streams:list",
            "dis:transferTasks:list",
            "dli:queue:submitJob",
            "dli:queue:cancelJob",
            "dli:table:insertOverwriteTable",
            "dli:table:insertIntoTable",
            "dli:table:alterView",
            "dli:table:alterTableRename",
            "dli:table:compaction",
            "dli:table:truncateTable",
            "dli:table:alterTableDropColumns",
            "dli:table:alterTableSetProperties",
            "dli:table:alterTableChangeColumn",
            "dli:table:showSegments",
            "dli:table:alterTableRecoverPartition",
            "dli:table:dropTable",
            "dli:table:update",
            "dli:table:alterTableDropPartition",
            "dli:table:alterTableAddPartition",
            "dli:table:alterTableAddColumns",
            "dli:table:alterTableRenamePartition",
            "dli:table:delete",
            "dli:table:alterTableSetLocation",
            "dli:table:describeTable",
            "dli:table:showPartitions",
            "dli:table:showCreateTable",
            "dli:table:showTableProperties",
            "dli:table:select",
            "dli:resource:updateResource",
            "dli:resource:useResource",
            "dli:resource:getResource",
            "dli:resource:listAllResource",
            "dli:resource:deleteResource",
            "dli:database:explain",
            "dli:database:createDatabase",
            "dli:database:dropFunction",
            "dli:database:createFunction",
            "dli:database:displayAllDatabases",
      "dli:database:displayAllTables",
            "dli:database:displayDatabase",
            "dli:database:describeFunction",
            "dli:database:createView",
            "dli:database:createTable",
            "dli:database:showFunctions",
            "dli:database:dropDatabase",
            "dli:group:useGroup",
            "dli:group:updateGroup",
            "dli:group:listAllGroup",
            "dli:group:getGroup",
            "dli:group:deleteGroup",
```

```
            "dli:column:select",
            "dli:jobs:start",
            "dli:jobs:export",
            "dli:jobs:update",
            "dli:jobs:list",
            "dli:jobs:listAll",
            "dli:jobs:get",
            "dli:jobs:delete",
            "dli:jobs:create",
            "dli:jobs:stop",
        "dli:variable:update",
        "dli:variable:delete",
            "dws:cluster:list",
            "dws:cluster:getDetail",
            "dws:openAPICluster:getDetail",
            "ecs:servers:get",
        "ecs:servers:list",
            "ecs:servers:stop",
            "ecs:servers:start",
        "ecs:flavors:get",
            "ecs:cloudServerFlavors:get",
            "ecs:cloudServers:list",
            "ecs:availabilityZones:list",
            "ges:graph:access",
            "ges:metadata:create",
            "ges:jobs:list",
            "ges:graph:operate",
            "ges:jobs:getDetail",
            "ges:graph:getDetail",
            "ges:graph:list",
            "ges:metadata:list",
            "ges:metadata:getDetail",
            "ges:metadata:delete",
            "ges:metadata:operate",
            "kms:cmk:get",
            "kms:cmk:list",
            "kms:cmk:create",
            "kms:cmk:decrypt",
            "kms:cmk:encrypt",
            "kms:dek:create",
            "kms:dek:encrypt",
            "kms:dek:decrypt",
            "mrs:cluster:get",
            "mrs:cluster:list",
            "mrs:job:get",
            "mrs:job:list",
            "mrs:job:submit",
            "mrs:job:stop",
            "mrs:job:delete",
            "mrs:sql:execute",
            "mrs:sql:cancel",
            "rds:*:get",
            "rds:*:list",
            "smn:topic:publish",
            "smn:topic:list",
        "vpc:publicIps:list",
            "vpc:publicIps:get",
            "vpc:vpcs:get",
            "vpc:vpcs:list",
            "vpc:subnets:get",
            "vpc:securityGroups:get",
            "vpc:firewalls:list",
            "vpc:routeTables:list",
            "vpc:subNetworkInterfaces:list"
          ]
        }
    ]
}
```

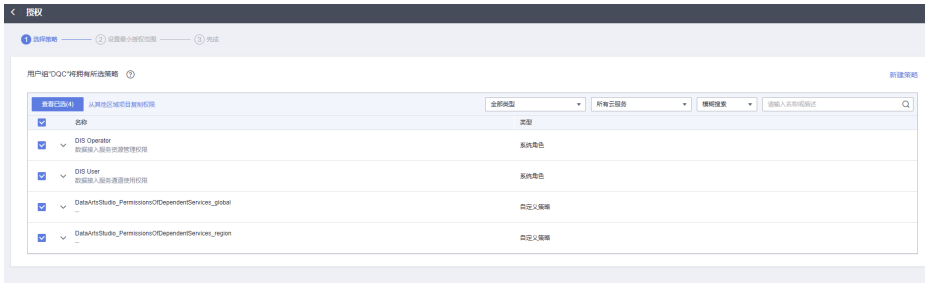
步骤5 自定义策略创建完成后，再次进入“用户组”，单击DQC用户组后的“授权”，进入授权操作。首先选择“角色与策略授权”，勾选如下系统角色和自定义策略为对象授权。

- 系统角色DIS Operator和DIS User
- 自定义策略DataArtsStudio_PermissionsOfDependentServices_global
- 自定义策略DataArtsStudio_PermissionsOfDependentServices_region

说明

仅当在数据开发组件作业中通过**DLI Spark**节点选择自定义镜像时，需要容器镜像服务中的镜像读取权限，推荐账号管理员通过镜像授权的方式为用户授予权限（SWR管理员权限账号登录容器镜像服务SWR控制台，在左侧导航栏选择“我的镜像”，进入所需自定义镜像的镜像详情页面，为用户授予该镜像的读取权限）。否则，则需要为用户授予SWR Administrator权限。

图 4-11 为用户组配置依赖服务最小权限合集



步骤6 授权成功后，依赖服务最小权限配置完成。

----结束

创建 IAM 用户并加入用户组

步骤1 在IAM服务控制台中，左侧导航窗格中，选择“用户”，单击右上方的“创建用户”。

图 4-12 创建用户

步骤2 在“创建用户”页面按照下图配置“用户信息”，完成配置后单击页面右下角的“下一步”。

- 用户信息：用户名填写为DataArts Studio-DQC。
- 访问方式：选择“管理控制台访问”和“编程访问”。

说明

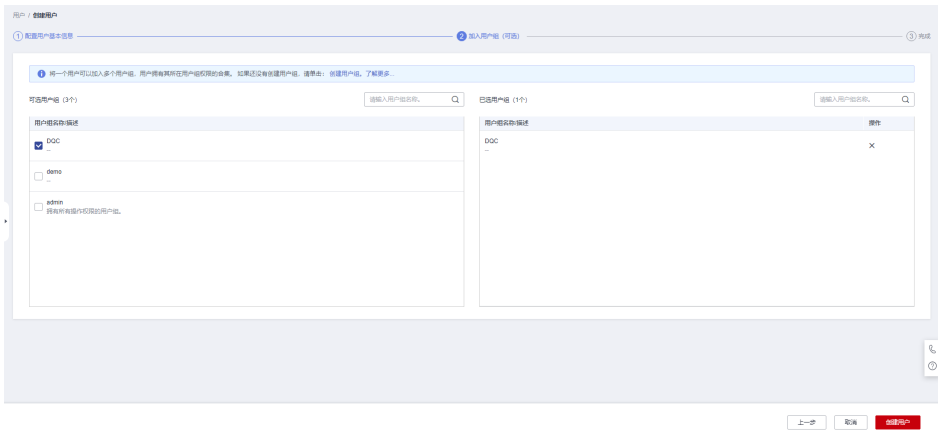
仅当创建IAM用户时的访问方式勾选“编程访问”后，此IAM用户才能通过认证鉴权，从而使用API、SDK等方式访问DataArts Studio。

- 凭证类型：勾选访问密钥和密码，推荐为用户自定义初始密码。
- 登录保护：根据需求选择，一般无需开启。

图 4-13 配置用户信息

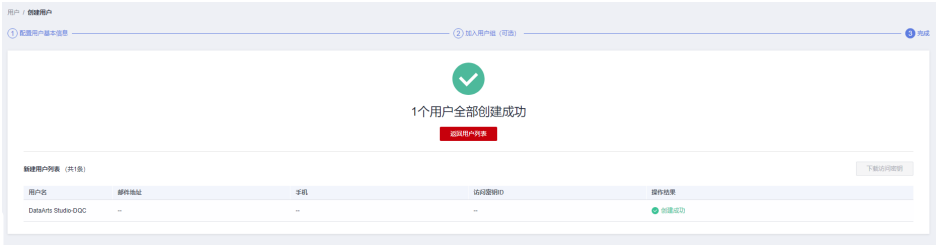
步骤3 选择将用户加入到DQC用户组，单击页面右下角的创建用户。

图 4-14 创建用户按钮



步骤4 创建完成可返回用户列表查看。

图 4-15 创建成功



----结束

添加工作空间成员并配置角色

步骤1 使用华为账号登录DataArts Studio管理控制台的首页，单击“空间管理”。

图 4-16 空间管理

步骤2 选择需要加入的工作空间，单击“编辑”。

图 4-17 编辑工作空间

步骤3 在空间信息界面，单击“添加”。

图 4-18 添加成员

步骤4 将新建的IAM用户加入工作空间，单击“确定”。

- 用户类型：选择“按用户添加”。
- 成员账号：选择**创建IAM用户并加入用户组**章节中新建的IAM用户。
- 设置角色：选择角色。

图 4-19 添加成员

步骤5 加入到工作空间后，该用户即可拥有DataArts Studio数据质量组件的操作权限，其余组件仅有查看权限但无法编辑。

----结束

用户登录并验证权限

步骤1 以**创建IAM用户并加入用户组**章节中新建的IAM账号登录华为云控制台，切换至授权区域。

步骤2 在“服务列表”中选择数据治理中心，进入DataArts Studio实例卡片。从实例卡片进入控制台首页后，确认能否正常查看工作空间列表情况。

步骤3 进入已添加当前用户的工作空间，进入各功能组件（例如管理中心和数据质量），查看能否正常进行数据质量业务操作。

----结束

5 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，详情请参考如下操作指导：

- **统计数据表行数**
- **统计数据库大小**

统计数据表行数

对于不同类型的数据源，DataArts Studio提供了多种方式来查看表的行数。

- 对于DWS、DLI、RDS、MRS Presto、MRS Hive、MRS Spark、等数据源，您可以在数据开发组件执行对应类型的统计表行数的SQL脚本，来查看表行数。
  ```
  select count(*) from tablename
  ```
- 对于DWS、DLI、RDS、MRS Hive、MRS Spark、Oracle等数据源，您可以在数据质量组件执行质量作业，来查看表行数。

对于非上述数据源，建议您参考数据源侧的操作说明，在数据源侧直接查看表行数。

本例以通过DataArts Studio数据质量作业获取表行数的操作为例进行说明，这种方式可以同时统计同一数据库下多个表的行数。

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 5-1 选择数据质量

步骤2 单击“质量作业”，进入质量作业列表。

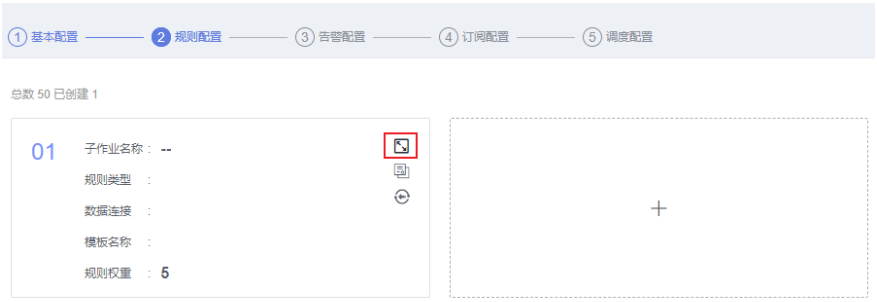
步骤3 单击“创建 ”，进入质量作业基本配置页面，如下图所示。

- 作业名称：CountingRows。
- 所属目录：选择作业存放目录。
- 作业级别：保持默认即可。

图 5-2 基本配置

步骤4 单击“下一步”，进入“规则配置”页面。单击子作业的打开图标，进入子作业配置页面。

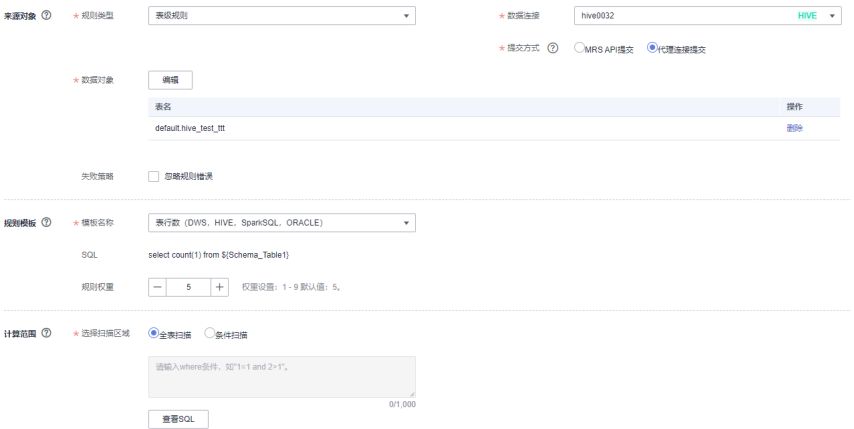
图 5-3 进入子作业配置



步骤5 单击子作业的打开图标，进入子作业的配置页面，配置规则信息。

- 基本信息：非必填项，保持默认即可。
- 来源对象：
  - 规则类型：选择“表级规则”。
  - 数据连接：选择在管理中心组件中创建的数据源连接。
  - 数据对象：选择待统计的数据表。
  - 其他参数保持默认即可。
- 规则模板：
  - 模板名称：选择“表行数（DWS，HIVE，SparkSQL，ORACLE）”。
  - 其他参数保持默认即可。
- 计算范围：选择“全表扫描”。
- 告警条件：非必填，保持默认即可。

图 5-4 子作业规则配置



步骤6 单击“下一步”，进入“告警配置”页面。

告警条件选择“子规则告警条件”，表达式可以自定义，此处可配置为“${1}<=0”，表示总行数小于等于0时触发告警。

图 5-5 告警配置

步骤7 单击“下一步”，进入“订阅配置”页面。

如果开启通知状态，需选择通知类型，并选择主题。通知类型有“触发告警”和“运行成功”两类，可根据实际业务场景选择。

步骤8 单击“下一步”，进入“调度配置”页面。

调度方式分为“单次调度”和“周期调度”。单次统计选择“单次调度”即可。

步骤9 单击“提交”，进入质量作业列表页面。

图 5-6 质量作业列表

步骤10 在CountingRows作业操作列，单击“运行”，生成作业对应的实例。

步骤11 单击“运维管理”，进入作业实例列表界面，找到对应的作业实例。待实例运行完成后，单击“结果&日志”，在“运行结果”页签，可查看该质量作业的运行结果，即待统计表的总行数。

图 5-7 查看表的总行数

----结束

统计数据库大小

您可以直接在数据目录组件中查看数据库大小。

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 5-8 选择数据目录

步骤2 在“总览”页面的“资产总览”页签，单击技术资产下数据库的统计数量，即可查看每个库对应的表数量及大小。

图 5-9 查看技术资产

图 5-10 查看数据量

----结束

6 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。

本章以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。

前提条件

- 已在数据仓库服务创建DWS集群，确保与DataArts Studio实例网络互通，并且具有KMS密钥的查看权限。
- 已在Map Reduce服务创建MRS集群，确保与DataArts Studio实例网络互通。
- 已创建CDM集群，详情请参见购买**批量数据迁移增量包**。

创建数据迁移连接

步骤1 登录DataArts Studio控制台，单击相应工作空间后的“数据集成”。

步骤2 在集群管理页面，单击所创建集群操作列“作业管理”，进入“作业管理”页面。

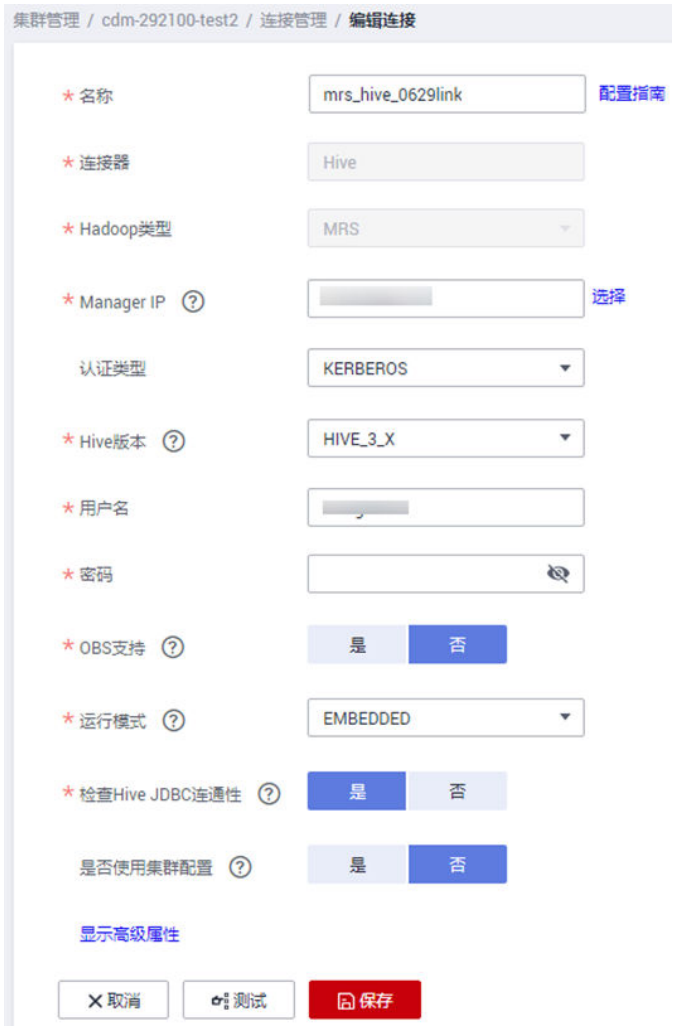
图 6-1 作业管理页面

步骤3 在连接管理页签中，单击“新建连接”，创建DWS数据连接，参数说明请参见**配置DWS连接**。

图 6-2 配置 DWS 连接

步骤4 同上述步骤，创建MRS Hive数据连接，参数说明请参见**配置MRS Hive连接**。

图 6-3 配置 MRS Hive 连接



----结束

创建并执行数据迁移作业

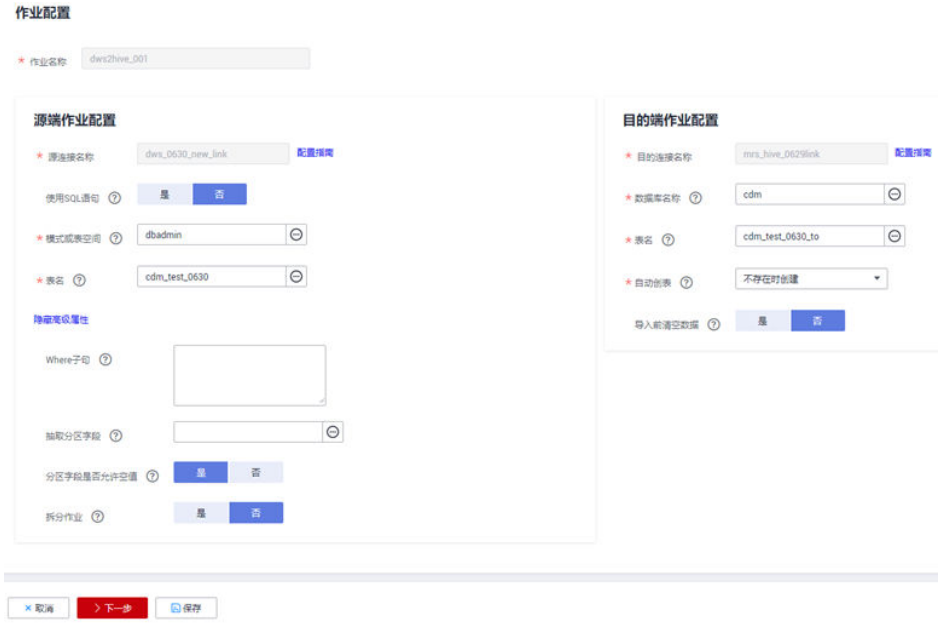
步骤1 登录DataArts Studio控制台，单击相应工作空间后的“数据集成”。

步骤2 在“集群管理”页面，单击所创建集群操作列“作业管理”，进入“作业管理”页面。

步骤3 在表/文件迁移页签中，单击新建作业，创建数据迁移作业。

步骤4 配置DWS源端作业参数、MRS Hive目的端作业参数，参数说明请参见**配置DWS源端参数**、**配置MRS Hive目的端作业参数**。

图 6-4 作业配置



步骤5 配置作业字段映射及任务配置，单击“保存并运行”，执行CDM作业。

步骤6 在“表/文件迁移”作业列表中，查看作业执行情况。

图 6-5 查看作业运行情况

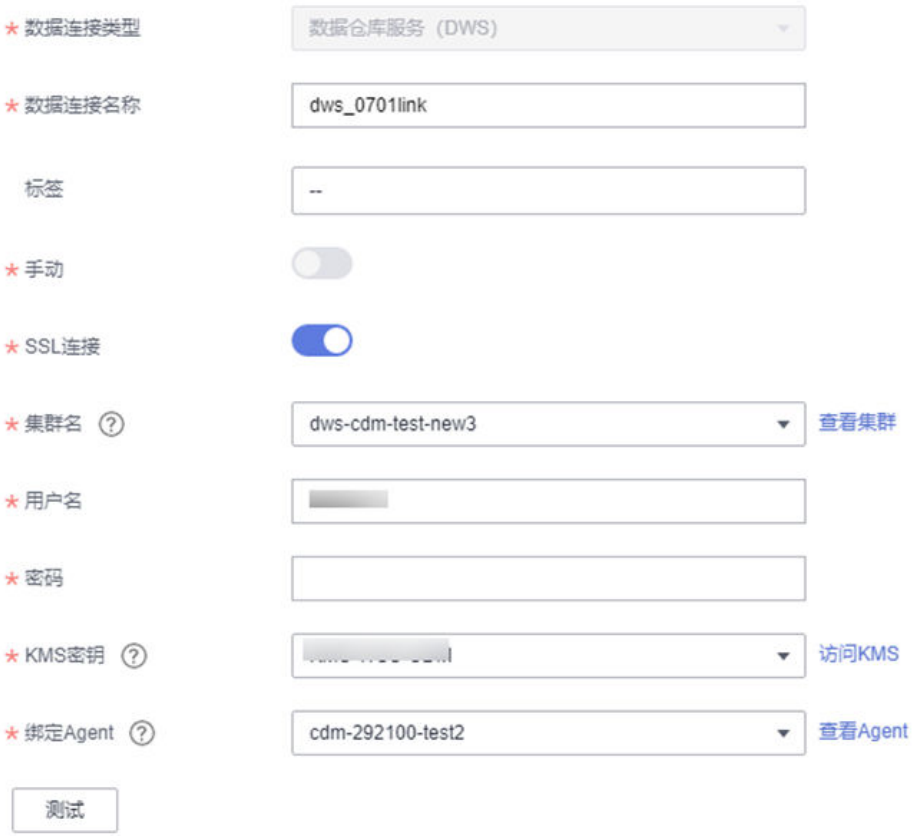
----结束

创建数据连接

步骤1 登录DataArts Studio控制台，单击相应工作空间后的“管理中心”。

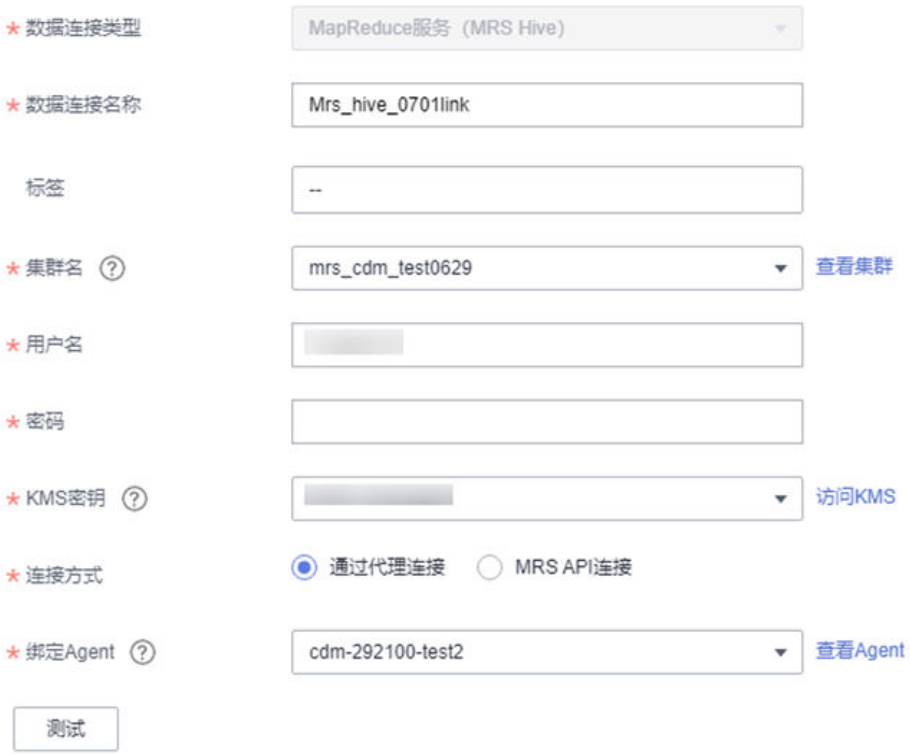
步骤2 在DataArts Studio管理中心模块中，单击“创建数据连接”，创建DWS数据连接，参数说明请参见**DWS数据连接**。

图 6-6 创建 DWS 数据连接



步骤3 同上述步骤创建MRS Hive数据连接，参数说明请参见**MRS Hive数据连接**。

图 6-7 创建 MRS Hive 数据连接



----结束

创建对账作业

步骤1 登录DataArts Studio控制台，单击相应工作空间后的“数据质量”。

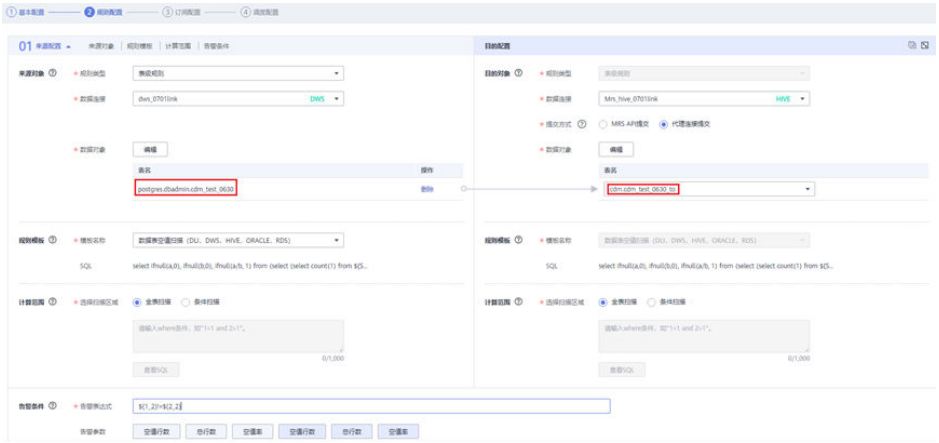
步骤2 在DataArts Studio数据质量模块，选择左侧导航菜单“数据质量监控->对账作业。”

步骤3 单击“新建”，配置对账作业的基本信息，如图6-8所示。

图 6-8 配置对账作业基本信息

步骤4 单击“下一步”，进入规则配置页面。您需要单击规则卡片中的，然后配置对账规则，选择数据迁移前后两张数据表，并配置告警规则，如图6-9所示。

图 6-9 配置对账规则



说明

- 源端和目的端的信息需要分别配置。
- 配置告警条件，其中单击左侧的表行数（${1_1}）表示左侧源端选中表的行数，单击右侧表行数（${2_1}）表示目的端表行数。此处配置告警条件为${1_1}!=${2_1}，表示当左侧表行数与右侧表行数不一致时，触发报警并显示报警状态。

步骤5 单击“下一步”，配置订阅信息，如图6-10所示。

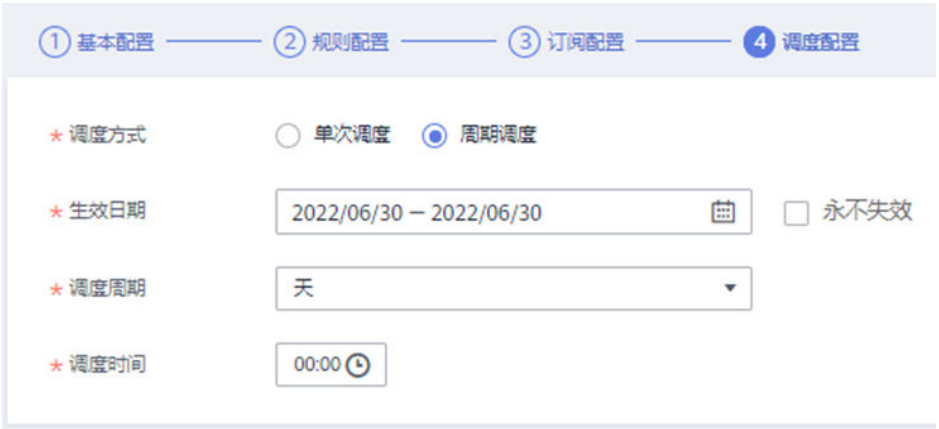
图 6-10 配置订阅信息

说明

勾选触发告警表示作业报警时发送通知到对应的SMN主题，勾选运行成功表示不报警时发送通知到SMN主题。

步骤6 单击“下一步”，配置调度方式，如图6-11所示。

图 6-11 调度配置



说明

单次调度表示需要手动触发运行，周期性调度表示会按照配置定期触发作业运行。此处以当天配置为例，设置每15分钟触发运行一次对账作业为例的配置。

步骤7 单击“提交”， 完成对账作业的创建。

----结束

执行对账作业并查看结果分析

步骤1 在数据质量模块左侧导航栏中，选择“数据质量监控 > 对账作业”。

步骤2 单击对账作业操作列中的“运行”，运行对账作业。

图 6-12 运行对账作业

步骤3 在数据质量模块左侧导航栏中，选择“数据质量监控 > 运维管理”，进入运维管理页面。

图 6-13 进入运维管理页面

步骤4 作业执行完成后，单击“结果&日志”，查看对账作业运行结果，如果源端和目的端表行数一致，则迁移成功。

图 6-14 查看运行结果

说明

- 运行结果中，左侧表示源端表行数规则运行结果，右侧表示目的端表行数规则运行结果。
- 误差率表示两端数据行数的差异比率，此处误差率为0表示两端一致。

----结束

7 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。

说明

- 本示例介绍的参数传递功能仅支持CDM 2.8.6版本及以上集群。
- 本示例以执行迁移Oracle数据到MRS Hive的CDM作业为例，介绍通过数据开发使用参数传递功能灵活调度CDM作业。

前提条件

已购买数据集成增量包。

创建 CDM 迁移作业

步骤1 登录控制台，选择实例，单击“进入控制台”，单击相应工作空间后的“数据集成”。

步骤2 在集群管理页面，单击集群操作列“作业管理”，进入“作业管理”页面，如图7-1所示。

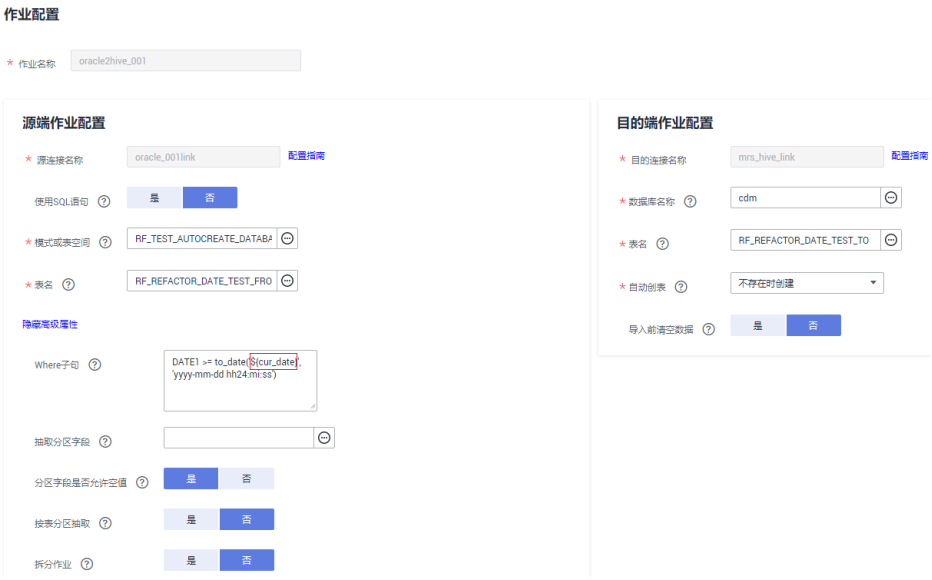
图 7-1 集群管理

步骤3 在“连接管理”页签中，单击“新建连接”，分别创建Oracle数据连接和MRS Hive数据连接，详情请参见**新建Oracle数据连接**和**新建MRS Hive数据连接**。

步骤4 在“表/文件迁移”页签中，单击“新建作业”，创建数据迁移作业。

步骤5 配置Oracle源端参数、MRS hive目的端参数，并配置传递参数，参数形式为 ${varName}，本示例参数为${cur_date}，如图7-2所示。

图 7-2 配置作业



说明

不能在CDM迁移作业中配置“作业失败重试”参数，如有需要请在数据开发中的CDM节点配置“失败重试”参数。

----结束

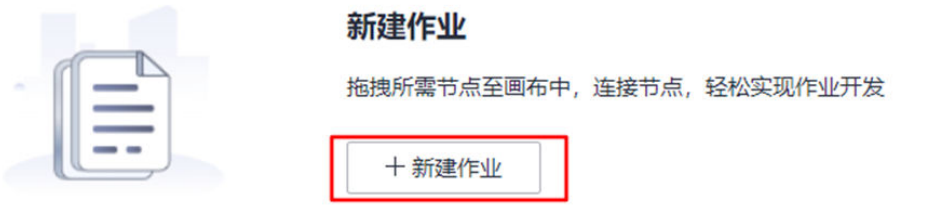
创建并执行数据开发作业

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

步骤2 在数据开发主界面的左侧导航栏，选择“数据开发> 作业开发” 。

步骤3 在“作业开发”界面中，单击“新建作业”，如图7-3所示。

图 7-3 新建作业



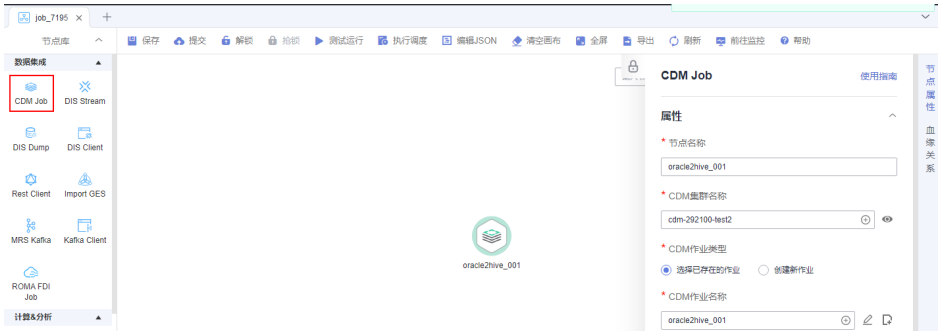
步骤4 在弹出的“新建作业”页面，配置如所示的参数。单击“确定”，创建作业。

表 7-1 作业参数

参数	说明
作业名称	自定义作业的名称，只能包含英文字母、数字、中文、“-”、“_”、“.”，且长度为1～128个字符。
作业类型	选择作业的类型。 ● 批处理作业：按调度计划定期处理批量数据，主要用于实时性要求低的场景。批作业是由一个或多个节点组成的流水线，以流水线作为一个整体被调度。被调度触发后，任务执行一段时间必须结束，即任务不能无限时间持续运行。 批处理作业可以配置作业级别的调度任务，即以作业为一整体进行调度，具体请参见**配置作业调度任务（批处理作业）**。 ● 实时处理作业：处理实时的连续数据，主要用于实时性要求高的场景。实时作业是由一个或多个节点组成的业务关系，每个节点可单独被配置调度策略，而且节点启动的任务可以永不下线。在实时作业里，带箭头的连线仅代表业务上的关系，而非任务执行流程，更不是数据流。 实时处理作业可以配置节点级别的调度任务，即每一个节点可以独立调度，具体请参见**配置作业调度任务（实时作业）**。
创建方式	选择作业的创建方式。 ● 创建空作业：创建一个空的作业。 ● 基于模板创建：使用数据开发模块提供的模板来创建。
选择目录	选择作业所属的目录，默认为根目录。
责任人	填写该作业的责任人。
作业优先级	选择作业的优先级，提供高、中、低三个等级。
委托配置	配置委托后，作业执行过程中，以委托的身份与其他服务交互。 **说明** 作业级委托优先于工作空间级委托。
日志路径	选择作业日志的OBS存储路径。日志默认存储在以dlf-log-{Projectid}命名的桶中。 **说明** ● 若您想自定义存储路径，请参见**（可选）修改作业日志存储路径**选择您已在OBS服务侧创建的桶。 ● 请确保您已具备该参数所指定的OBS路径的读、写权限，否则系统将无法正常写日志或显示日志。

步骤5 在数据开发作业中添加CDM Job节点，并关联已创建的CDM作业，如**图7-4**所示。

图 7-4 关联 CDM 作业



步骤6 在作业参数中配置业务需要的参数，如图7-5所示。

图 7-5 配置作业参数

说明

作业调度执行的过程中，会将该参数值传递给CDM作业，传递的参数“cur_date”可以配置为本示例“2021-11-10 00:00:00”固定参数值，也可以配置为EL表达式，例如：计划运行日期的前一天：#{DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyy-MM-dd")}，更多EL表达式请参见**EL表达式**。

步骤7 保存并提交作业版本，单击“测试运行”，执行数据开发作业。

步骤8 数据开发作业执行成功后，单击右上角的“前往监控”，进入“作业监控”页面，查看生成的任务或实例是否符合需求，如图7-6所示。

图 7-6 查看运行结果

----结束

8 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。

这里以DWS导入到OBS为例，介绍DLF配合CDM实现增量迁移的流程：

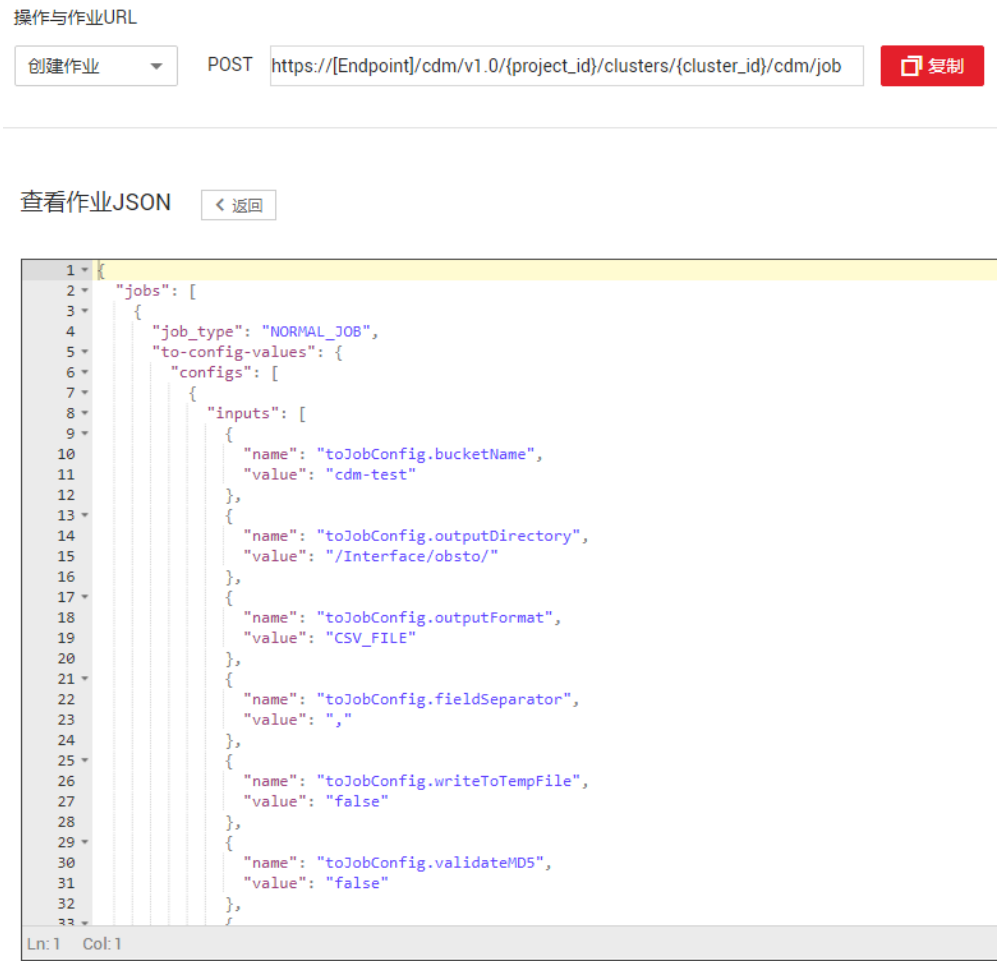
1. **获取CDM作业的JSON**
2. **修改JSON**
3. **创建DLF作业**

获取 CDM 作业的 JSON

1. 进入CDM主界面，创建一个DWS到OBS的表/文件迁移作业。
2. 在CDM“作业管理”界面的“表/文件迁移”页签下，找到已创建的作业，单击作业操作列的“更多 > 查看作业JSON”，如**图8-1**所示。

   您也可以使用其它已创建好的CDM作业JSON。

图 8-1 查看作业 JSON



3. 作业JSON就是创建CDM作业的请求消息体模板，URL地址中[Endpoint]、{project_id}、{cluster_id}需要替换为您实际的信息：
   - [Endpoint]：终端节点。

     终端节点（Endpoint）即调用API的**请求地址**，不同服务不同区域的终端节点不同。本服务的Endpoint可从**终端节点Endpoint**获取。
   - {project_id}：项目ID。
   - {cluster_id}：集群ID，可在CDM集管理界面，单击集群名称查看。

修改 JSON

根据您的业务需要，可以修改JSON Body。这里以1天为周期，where子句作为抽取数据时的判断条件（一般使用时间字段来作为增量迁移时的判断条件），每天迁移昨天新增的数据。

1. 修改where子句，增量某个时间段的数据：

```
{
  "name": "fromJobConfig.whereClause",
  "value": "_timestamp >= '${startTime}' and _timestamp < '${currentTime}'"
}
```

说明

- 源端数据库是数据仓库服务DWS或者MySQL时，对于时间的判断可以写成以下两种：
  ```
  _timestamp >= '2018-10-10 00:00:00' and _timestamp < '2018-10-11 00:00:00'
  或者
  _timestamp between '2018-10-10 00:00:00' and '2018-10-11 00:00:00'
  ```
- 如果源端数据库是Oracle，where子句应该写成：
  ```
  _timestamp >= to_date (2018-10-10 00:00:00' , 'yyyy-mm-dd hh24:mi:ss' ) and _timestamp < to_date (2018-10-10 00:00:00' , 'yyyy-mm-dd hh24:mi:ss' )
  ```

2. 每个周期的增量数据导入到不同的目录：
   ```
   {
     "name": "toJobConfig.outputDirectory",
     "value": "dws2obs/${currentTime}"
   }
   ```
3. 作业名改成动态的，否则会因为作业重名而无法创建：
   ```
   "to-connector-name": "obs-connector",
   "from-link-name": "dws_link",
   "name": "dws2obs-${currentTime}"
   ```

如果需要修改更多参数，请参见《**云数据迁移API参考**》，这里修改后的JSON样例如下：

```
{
  "jobs": [
    {
      "job_type": "NORMAL_JOB",
      "to-config-values": {
        "configs": [
          {
            "inputs": [
              {
                "name": "toJobConfig.bucketName",
                "value": "cdm-test"
              },
              {
                "name": "toJobConfig.outputDirectory",
                "value": "dws2obs/${currentTime}"
              },
              {
                "name": "toJobConfig.outputFormat",
                "value": "CSV_FILE"
              },
              {
                "name": "toJobConfig.fieldSeparator",
                "value": ","
              },
              {
                "name": "toJobConfig.writeToTempFile",
                "value": "false"
              },
              {
                "name": "toJobConfig.validateMD5",
                "value": "false"
              },
              {
                "name": "toJobConfig.encodeType",
                "value": "UTF-8"
              },
              {
                "name": "toJobConfig.duplicateFileOpType",
                "value": "REPLACE"
              },
              {
                "name": "toJobConfig.kmsEncryption",
                "value": "false"
              }
            ],
            "name": "toJobConfig"
```



```
        }
       ]
      },
      "from-config-values": {
       "configs": [
        {
         "inputs": [
          {
           "name": "fromJobConfig.schemaName",
           "value": "dws_database"
          },
          {
           "name": "fromJobConfig.tableName",
           "value": "dws_from"
          },
          {
           "name": "fromJobConfig.whereClause",
           "value": "_timestamp >= '${startTime}' and _timestamp < '${currentTime}'"
          },
          {
           "name": "fromJobConfig.columnList",
           "value":
"_tiny&_small&_int&_integer&_bigint&_float&_double&_date&_timestamp&_char&_varchar&_text"
          }
         ],
         "name": "fromJobConfig"
        }
       ]
      },
      "from-connector-name": "generic-jdbc-connector",
      "to-link-name": "obs_link",
      "driver-config-values": {
       "configs": [
        {
         "inputs": [
          {
           "name": "throttlingConfig.numExtractors",
           "value": "1"
          },
          {
           "name": "throttlingConfig.submitToCluster",
           "value": "false"
          },
          {
           "name": "throttlingConfig.numLoaders",
           "value": "1"
          },
          {
           "name": "throttlingConfig.recordDirtyData",
           "value": "false"
          },
          {
           "name": "throttlingConfig.writeToLink",
           "value": "obs_link"
          }
         ],
         "name": "throttlingConfig"
        },
        {
         "inputs": [],
         "name": "jarConfig"
        },
        {
         "inputs": [],
         "name": "schedulerConfig"
        },
        {
         "inputs": [],
         "name": "transformConfig"
```



```
        },
        {
          "inputs": [],
          "name": "smnConfig"
        },
        {
          "inputs": [],
          "name": "retryJobConfig"
        }
      ]
    },
    "to-connector-name": "obs-connector",
    "from-link-name": "dws_link",
    "name": "dws2obs-${currentTime}"
  }
 ]
}
```

创建 DLF 作业

1. 在DLF创建如图8-2所示的Rest Client节点数据开发作业，详细操作请参见《数据治理中心DataArts Studio 用户指南》的新建作业章节。

   各节点与作业的配置详情请参见后续步骤。

   图 8-2 DLF 作业

2. 配置“创建作业”节点。

   DLF通过Rest Client节点调用REST接口创建CDM迁移作业。配置Rest Client节点的属性如下：

   a. 节点名称：您自定义名称，例如“创建CDM作业”。注意区分：在DLF作业中，CDM的迁移作业只是作为节点运行。

   b. URL地址：配置为获取CDM作业的JSON中获取的URL，格式为https://{*Endpoint*}/cdm/v1.0/{*project_id*}/clusters/{*cluster_id*}/cdm/job。

   c. HTTP方法：创建CDM作业的HTTP请求方法为“POST”。

   d. 添加如下两个请求头：

      - Content-Type = application/json
      - X-Language = en-us

   e. 请求消息体：输入修改JSON里面修改完成后的CDM作业JSON。

图 8-3 创建 CDM 作业的节点属性

3. 配置“运行作业”节点。

   创建CDM作业的节点配置完后，还需要在后面添加运行CDM作业的REST节点，具体请参见《**云数据迁移API参考**》中的“启动作业”章节。配置RestAPI节点的属性如下：

   a. 节点名称：运行作业。

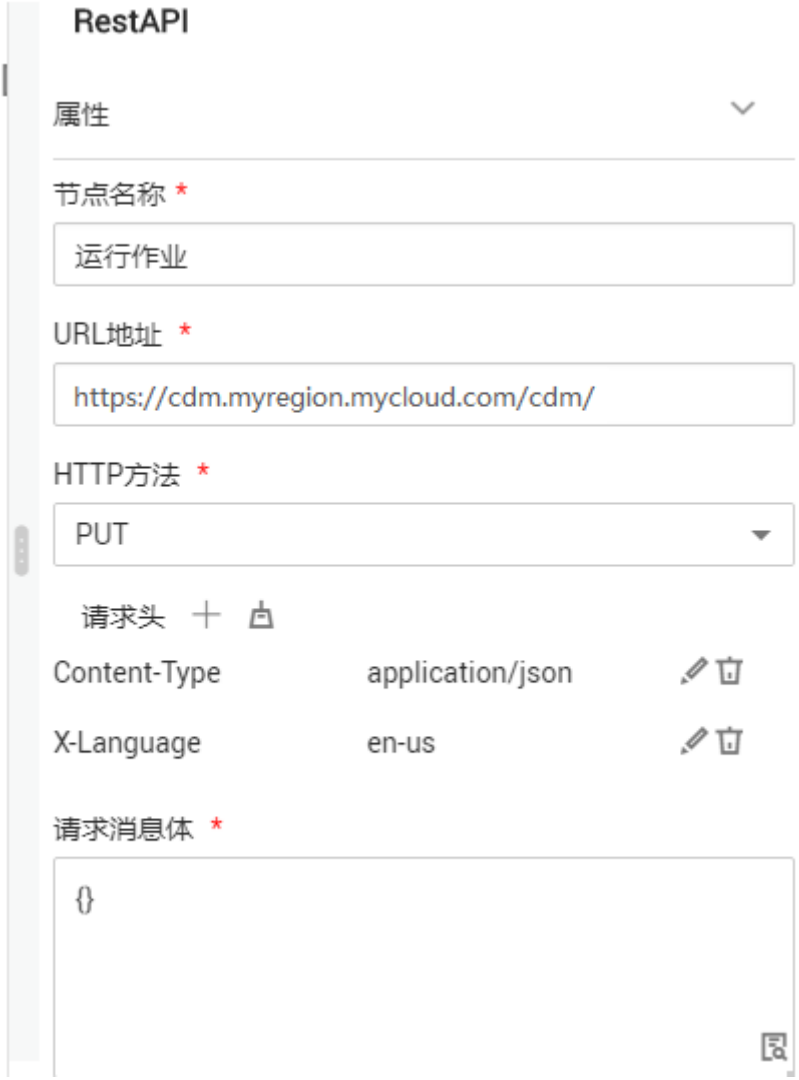
   b. URL地址：其中project_id、cluster_id和**2. 配置“创建作业”节点**中的保持一致，作业名需要配置为“dws2obs-${currentTime}”。格式为https://{*Endpoint*}/cdm/v1.0/{*project_id*}/clusters/{*cluster_id*}/cdm/job/{*job_name*}/start。

   c. HTTP方法：运行CDM作业的HTTP请求方法为“PUT”。

   d. 请求头：

      - Content-Type = application/json
      - X-Language = en-us

图 8-4 运行 CDM 作业的节点属性



4. 配置“等待作业运行完成”节点。

   由于CDM作业是异步运行的，运行作业的REST请求返回200，不代表数据已经迁移成功。后续有计算作业依赖CDM的迁移作业时，需要一个RestAPI节点去周期判断迁移是否成功，如果CDM迁移成功，再去做计算操作。查询CDM迁移是否成功的API，具体请参见《**云数据迁移API参考**》中“查询作业状态”章节。

   运行CDM作业的REST节点配置完成后，添加等待CDM作业完成节点，节点属性为：

   a. 节点名称：等待作业运行完成。

   b. URL地址：格式为https://{*Endpoint*}/cdm/v1.0/{*project_id*}/clusters/{*cluster_id*}/cdm/job/{*job_name*}/status。其中project_id、cluster_id和**2. 配置“创建作业”节点**中的保持一致，作业名需要配置为“dws2obs-${currentTime}”。

   c. HTTP方法：查询CDM作业状态的HTTP请求方法为“GET”。

   d. 请求头：

      - Content-Type = application/json

- X-Language = en-us

e. 是否需要判断返回值：选择“YES”。
f. 返回值字段路径：配置为submissions[0].status。
g. 请求成功标志位：配置为SUCCEEDED。
h. 其他参数保持默认即可。

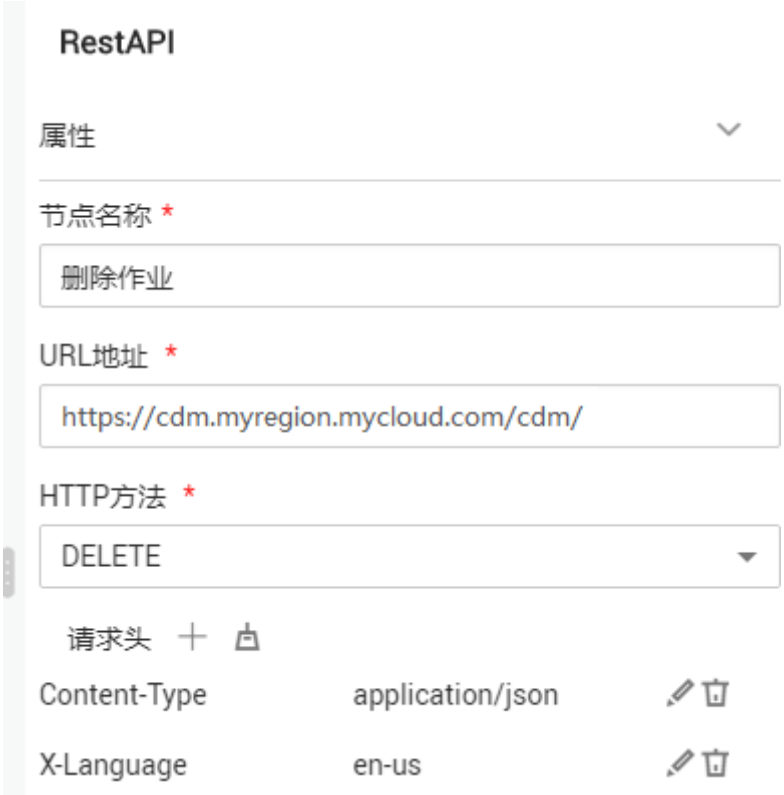
5. （可选）配置“删除作业运行完成”节点。

这里的删除作业可根据实际需要选择。由于DLF是通过周期创建CDM作业来实现增量迁移，因此会累积大量的作业在CDM集群上，所以可在迁移成功后，删除已经运行成功的作业。如果您需要删除，在查询CDM作业状态的节点后面，添加删除CDM作业的RestAPI节点即可，DLF会调用《**云数据迁移API参考**》中的“删除作业”接口。

删除CDM作业的节点属性为：

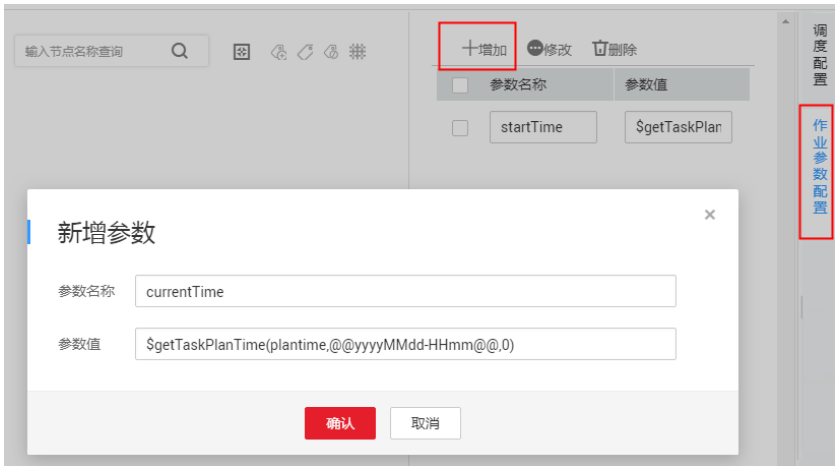
a. 节点名称：删除作业。
b. URL地址：格式为https://{*Endpoint*}/cdm/v1.0/{*project_id*}/clusters/{*cluster_id*}/cdm/job/{*job_name*}。其中project_id、cluster_id和**2. 配置“创建作业”节点**中的保持一致，作业名需要配置为“dws2obs-${currentTime}”。
c. HTTP方法：删除CDM作业的HTTP请求方法为“DELETE”。
d. 请求头：
   - Content-Type = application/json
   - X-Language = en-us
e. 其他参数保持默认即可。

图 8-5 删除 CDM 作业节点配置



6. 如果需要在迁移完后进行计算操作，可在后续添加各种计算节点，完成数据计算。
7. 配置DLF作业参数。
   a. 配置DLF作业参数，如图8-6所示。
      - startTime = $getTaskPlanTime(plantime,@@yyyyMMddHHmmss@@,-24*60*60)
      - currentTime = $getTaskPlanTime(plantime,@@yyyyMMdd-HHmm@@,0)

      图 8-6 DLF 作业参数配置

   b. 保存DLF作业后，选择“调度配置 > 周期调度”，调度周期配置为1天。
      这样，DLF配合CDM就实现了每天迁移昨天新增的数据。

9 通过 CDM 节点批量创建分表迁移作业

适用场景

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。

本教程中，源端MySQL数据库中存在三张分表，分别是mail01、mail02和mail03，且表结构一致，数据内容不同。目的端为MRS Hive服务。

操作前提

- 已创建CDM集群。
- 已经开通了MRS Hive服务。
- 已经在MRS Hive服务中创建了数据库和表。

创建连接

步骤1 登录DataArts Studio控制台，找到所需要的DataArts Studio实例，单击实例卡片上的“进入控制台”，进入概览页面。

步骤2 找到所需要的工作空间，单击工作空间的“数据集成”，系统跳转至数据集成页面。

步骤3 单击CDM集群“操作”列的“作业管理”，进入作业管理界面。

步骤4 单击“连接管理->驱动管理”，参考**管理驱动**，上传MySQL数据库驱动。

步骤5 选择“连接管理 > 新建连接”，新建MySQL连接。连接器类型选择“MySQL”，然后单击“下一步”配置连接参数，参数说明如**表9-1**所示。配置完成后，单击“保存”回到连接管理界面。

表 9-1 MySQL 数据库连接参数

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

参数名	说明	取值样例
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的MySQL数据库实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	3306
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
使用本地API	可选参数，选择是否使用数据库本地API加速。 创建MySQL连接时，CDM会自动尝试启用MySQL数据库的local_infile系统变量，开启MySQL的LOAD DATA功能加快数据导入，提高导入数据到MySQL数据库的性能。注意，开启本参数后，日期类型将不符合格式的会存储为0000-00-00，更多详细信息可在MySQL官网文档查看。 如果CDM自动启用失败，请联系数据库管理员启用local_infile参数或选择不使用本地API加速。 如果是导入到RDS上的MySQL数据库，由于RDS上的MySQL默认没有开启LOAD DATA功能，所以同时需要修改MySQL实例的参数组，将“local_infile”设置为“ON”，开启该功能。 **说明** 如果RDS上的“local_infile”参数组不可编辑，则说明是默认参数组，需要先创建一个新的参数组，再修改该参数值，并应用到RDS的MySQL实例上，具体操作请参见《关系型数据库用户指南》。	是
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择已创建的Agent。	-
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
单次提交行数	可选参数，单击“显示高级属性”后显示。 指定每次批量提交的行数，根据数据目的端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	-

参数名	说明	取值样例
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 常见配置举例如下： • **connectTimeout=360000**与**socketTimeout=360000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。 • **tinyInt1isBit=false**或**mysql.bool.type.transform=false**：MySQL默认开启配置tinyInt1isBit=true，将TINYINT(1)当作BIT也就是Types.BOOLEAN来处理，会将1或0读取为true或false从而导致迁移失败，此时可关闭配置避免迁移报错。 • **useCursorFetch=false**：CDM作业默认打开了JDBC连接器与关系型数据库通信使用二进制协议开关，即useCursorFetch=true。部分第三方可能存在兼容问题导致迁移时间转换出错，可以关闭此开关；开源MySQL数据库支持useCursorFetch参数，无需对此参数进行设置。 • **allowPublicKeyRetrieval=true**：MySQL默认关闭允许公钥检索机制，因此连接MySQL数据源时，如果TLS不可用、使用RSA公钥加密时，可能导致连接报错。此时可打开公钥检索机制，避免连接报错。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'
单次写入行数	指定单次批量写入的行数，当写入行数累计到单次批量提交行数时提交一次，该值应当小于单次提交行数。	100

步骤6 再次选择“连接管理 > 新建连接”，新建MRS Hive连接。连接器类型选择“MRS Hive”，然后单击“下一步”配置连接参数，参数说明如表9-2所示。配置完成后，单击“保存”回到连接管理界面。

表 9-2 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。	hive

参数名	说明	取值样例
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	KERBEROS
Hive版本	Hive的版本。根据服务端Hive版本设置。	HIVE_3_X
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。 **说明** ● 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对应组件的进行库、表、数据的操作，还需要添加对应组件的用户组权限。 ● 如果CDM集群为2.9.0之前的版本，或MRS集群为3.1.0之前的版本，则所创建的用户需要具备Manager_administrator、Manager_tenant或System_administrator权限，才能在CDM创建连接。	cdm
密码	访问MRS Manager的用户密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否
运行模式	“HIVE_3_X”版本支持该参数。支持以下模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端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

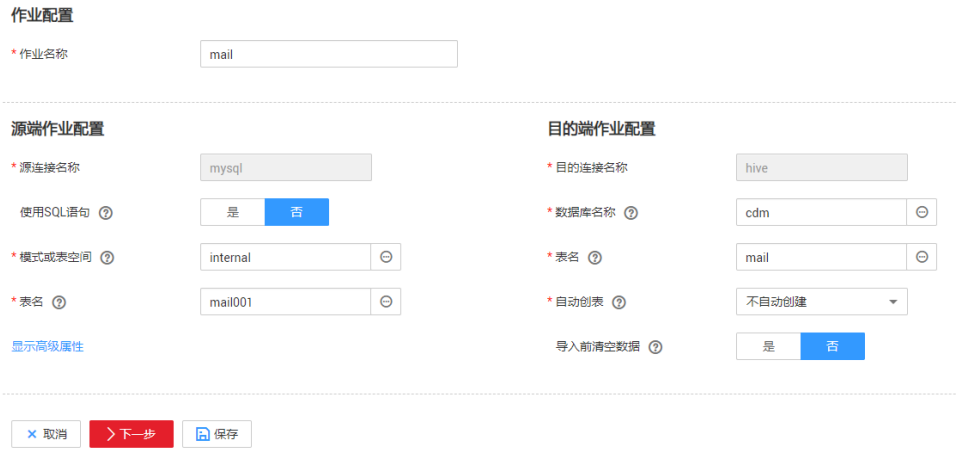
参数名	说明	取值样例
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否

----结束

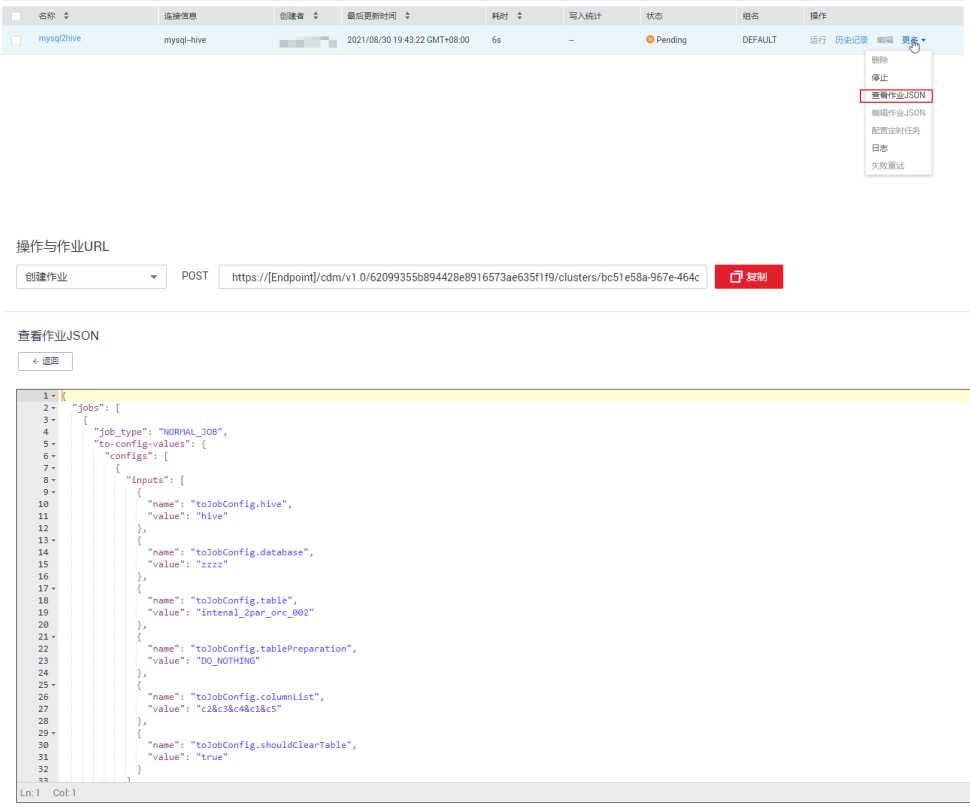
创建样例作业

步骤1 单击CDM集群“操作”列的“作业管理”，进入作业管理界面。

步骤2 进入“表/文件迁移”页签，单击“新建作业”创建MySQL第一个分表mail001到MRS Hive目标表mail的数据集成作业，具体如下图所示。



步骤3 样例作业创建完毕后，如下图查看作业JSON，并复制作业JSON，用于后续数据开发作业配置。



----结束

创建数据开发作业

步骤1 单击工作空间的“数据开发”，进入DataArts Studio数据开发模块。

步骤2 创建子作业“分表作业”，选择CDM节点，节点属性中作业类型配置为“创建新作业”，并将**步骤2**中复制的作业JSON粘贴到“CDM作业消息体”中。

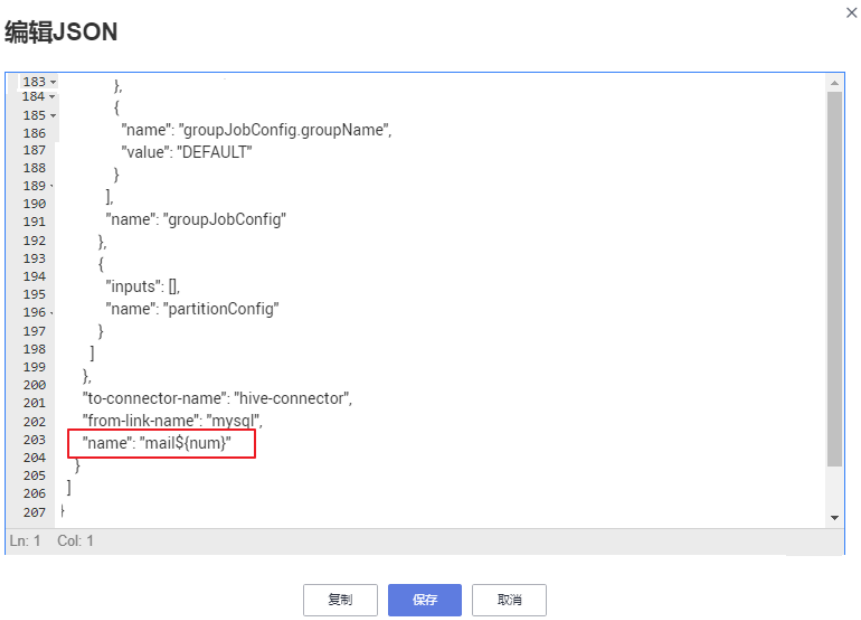
步骤3 编辑“CDM作业消息体”。

1. 由于源表有三个，分别为mail001、mail002、mail003，因此需要将作业JSON中的“fromJobConfig.tableName”属性值配置为“mail${num}”，即源表名是通过参数配置。如下图所示：

2. 由于数据迁移作业名不能重复，因此修改JSON中作业名称“name”属性值配置为“mail${num}”，目的是创建多个CDM集成作业，避免作业名称重复。如下图所示：

说明

如果需要创建分库的作业，也可将作业JSON中的源连接修改为变量，方便替换。



步骤4 添加作业参数num，用于作业JSON中调用。如下图所示：

添加完成后单击“保存并提交版本”，以保存子作业。

步骤5 创建主作业“集成管理”，选择For Each节点，每次循环调用分表作业，分别将参数001、002、003传递给子作业，生成不同的分表抽取任务。

关键配置如下：

- 子作业：选择“分表作业”
- 数据集：[['001'],['002'],['003']]
- 子作业参数：@@#{Loop.current[0]}@@

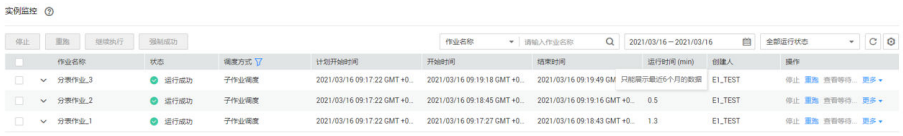
说明

此处子作业参数的EL表达式需要添加@@。如果不加@@包围，数据集001会被识别为1，导致源表名不存在的问题。

如下图所示：

配置完成后点击“保存并提交版本”，以保存主作业。

步骤6 创建主作业和子作业完成后，通过测试运行主作业“集成管理”，检查数据集成作业创建情况。运行成功后，创建并运行CDM子作业成功。



----结束

注意事项

- 由于CDM版本不同，某些属性可能不支持，比如fromJobConfig.BatchJob。当创建任务报错时，需要在请求体中删除该属性。如下图所示：

- CDM节点配置为创建作业时，节点运行会检测是否有同名CDM作业。
  - 如果CDM作业未运行，则按照请求体内容更新同名作业。
  - 如果同名CDM作业正在运行中，则等待作业运行完成。此时该CDM作业可能被其他任务启动，可能会导致数据抽取不符合预期（如作业配置未更新、运行时间宏未替换正确等），因此请注意不要启动或者创建多个同名作业。

10 基于 MRS Hive 表构建图数据并自动导入 GES

10.1 场景说明

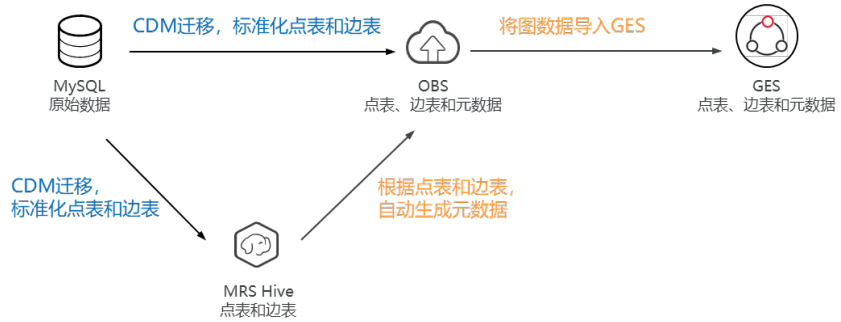
图引擎服务（Graph Engine Service，简称GES）是针对以“关系”为基础的“图”结构数据，进行查询、分析的服务，并广泛应用于社交关系分析、营销推荐、舆情及社会化聆听、信息传播、防欺诈等具有丰富关系数据的场景。

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。

场景说明

本案例基于某电影网站的用户和评分数据，使用DataArts Studio将MySQL原始数据处理为标准点数据集和边数据集，并同步到OBS和MRS Hive中，然后通过Import GES节点自动生成元数据后，将图数据导入到GES服务中。

图 10-1 业务场景说明



需要额外说明的是，GES图数据格式包含三部分：点数据集、边数据集以及元数据，如果原始数据不符合GES指定的格式，则需要将数据整理为GES支持的格式。

- 点数据集用于存放点数据。
- 边数据集用于存放边数据。

- 元数据用于描述点数据集和边数据集中的数据格式。

GES相关概念和图数据介绍请参见**一般图数据格式**。

约束限制

通过Import GES节点自动生成元数据时，有如下约束限制：

1. 生成元数据时，目前仅支持选择单标签（Label）场景的点表和边表。如果点表或边表中存在多个标签，则生成的元数据会存在缺失。
2. 生成元数据xml文件是手动单击“生成元数据”触发的，如果在该节点在后续的作业调度运行中，点表和边表结构发生变化，元数据xml文件并不会随之更新，需要手动进入新建元数据窗口，再次单击“生成元数据”重新生成新的元数据xml文件。
3. 生成的元数据xml文件，属性（Property）中的数据复合类型（Cardinality），目前仅支持填写为“single”类型，不支持自定义。
4. 生成元数据功能本身，支持一次生成多对点表和边表的元数据xml文件。但考虑到Import GES节点的“边数据集”和“点数据集”参数，分别只能选择一张表，建议您在有多对点表和边表的情况下，分拆多个Import GES节点分别导入，以确保导入图数据时，元数据与每对点表和边表能够一一对应。

10.2 准备工作

操作环境准备

- 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、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。
- 您需要在MRS服务控制台，创建一个包含Hive组件的MRS集群，用于通过存储其中的点数据集和边数据集生成元数据。建议创建MRS集群时，相关网络参数与DataArts Studio实例中的CDM集群的所在区域、虚拟私有云、子网、安全组保持一致，默认内网互通，否则还需手动打通MRS集群与CDM集群的网络。二者的企业项目也应保持一致。

  说明

  由于创建MRS集群时仅支持自动创建安全组，建议您可以先创建MRS安全集群，然后在购买DataArts Studio实例时选择同虚拟私有云、同子网、以及MRS集群自动创建的安全组（以“mrs_*集群名_随机字符*”命名），以确保DataArts Studio实例和MRS集群网络默认互通。

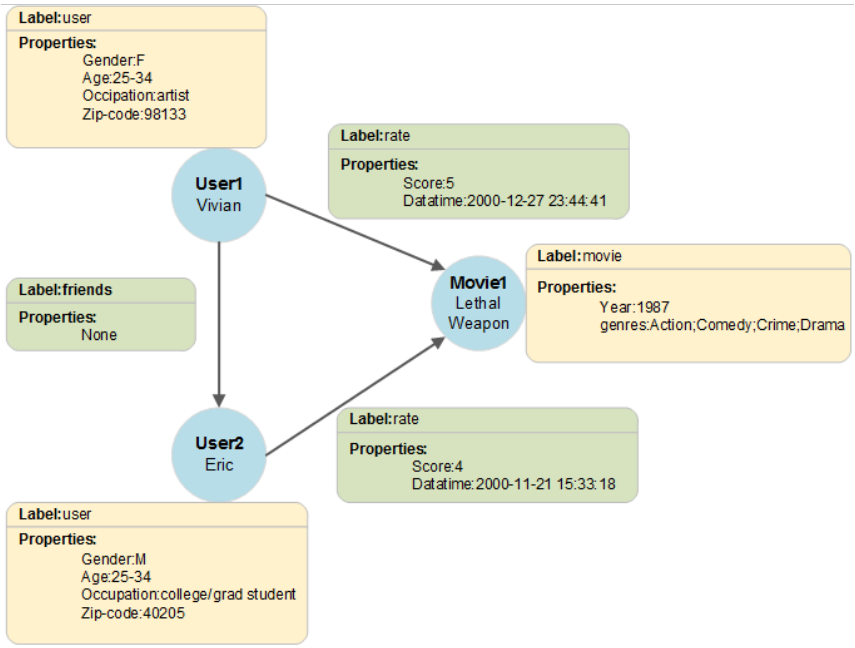
  如果您已购买DataArts Studio实例，然后才开始创建MRS集群，则您需要在“虚拟私有云VPC”服务控制台的“访问控制 > 安全组”界面对MRS集群创建的安全组（以“mrs_*集群名_随机字符*”命名）添加规则，放通入方向的DataArts Studio实例安全组，详情请参见**如何配置安全组规则**章节。
- 您需要在云数据库RDS服务控制台，创建一个MySQL数据库实例，用于模拟原始数据源。建议创建MySQL数据库时，相关网络参数与DataArts Studio实例中的CDM集群的所在区域、虚拟私有云、子网、安全组保持一致，默认内网互通，否则还需手动打通MySQL数据库与CDM集群的网络。二者的企业项目也应保持一致。
- 您需要准备OBS桶，用于保存生成的元数据。OBS桶与DataArts Studio实例中的CDM集群的所在区域保持一致，企业项目也应相同。

- 您需要在图引擎GES服务控制台，创建一个图，用于将图数据导入其中，并进行可视化图形分析。GES与DataArts Studio实例中的CDM集群的所在区域保持一致，企业项目也应相同。

数据源准备

本示例原始数据包含**用户表vertex_user**，**电影表vertex_movie**，**朋友关系表edge_friends**和**电影评分表edge_rate**。关系说明如**图10-2**所示。

图 10-2 图数据说明



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4张样例数据表，分别为**用户表vertex_user**，**电影表vertex_movie**，**朋友关系表edge_friends**和**电影评分表edge_rate**。具体数据如下：

- 用户表vertex_user.csv：

```
Vivian,F,25-34,artist,98133
Mercedes,F,Under 18,K-12 student,10562
Katherine,F,35-44,lawyer,79101
Stuart,M,25-34,programmer,30316
Jacob,M,25-34,artist,55408
Editha,F,56+,homemaker,46911
Cassandra,F,56+,artist,55113
Sarah,F,18-24,other or not specified,55105
Hayden,M,56+,academic/educator,30030
Jeffery,M,25-34,self-employed,45242
Bonnie,F,50-55,technician/engineer,19716
Serena,F,35-44,programmer,44106
Sidney,M,18-24,writer,85296
Leander,M,50-55,doctor/health care,98237
Fred,M,35-44,other or not specified,30906
Roger,M,45-49,technician/engineer,73069
Ella,F,25-34,other or not specified,94402
Ray,M,18-24,college/grad student,90241
Eric,M,18-24,college/grad student,40205
Frances,F,56+,retired,1234
Allison,F,18-24,sales/marketing,49505
Willy,M,25-34,technician/engineer,38104
Lance,M,18-24,college/grad student,6459
June,F,25-34,other or not specified,13326
Marshal,M,50-55,scientist,7746
Max,M,35-44,executive/managerial,91107
Hardy,M,35-44,academic/educator,22181
Jordan,M,25-34,artist,8817
Reed,M,18-24,college/grad student,89146
Glendon,M,35-44,self-employed,46214
Kevin,M,56+,retired,2356
Evan,M,45-49,programmer,53718
Clark,M,56+,academic/educator,85718
Johnny,M,56+,retired,52003
Caleb,M,50-55,retired,41076
Janet,F,35-44,homemaker,61270
Sue,F,50-55,self-employed,13207
Margaret,F,45-49,academic/educator,1609
Luke,M,35-44,executive/managerial,44306
William,M,45-49,programmer,37914
Lena,F,35-44,other or not specified,42420
Solomon,M,45-49,scientist,64081-8102
Cary,M,35-44,executive/managerial,55124
Colin,M,25-34,executive/managerial,44115
Kenny,M,25-34,college/grad student,74074
Gavin,M,25-34,programmer,24060
Donald,M,35-44,programmer,95864
Wayne,M,18-24,scientist,94606
Frank,M,18-24,college/grad student,2906
Alexander,M,18-24,college/grad student,61801
```



```
Isaiah,M,25-34,other or not specified,33142
Josephine,F,25-34,college/grad student,78728
Joshua,M,35-44,executive/managerial,54016
August,M,35-44,customer service,64801
Jessie,F,18-24,clerical/admin,60640
Yvette,F,35-44,artist,94109
Albert,M,25-34,other or not specified,40515
Eugene,M,35-44,other or not specified,40504
Rachel,F,35-44,doctor/health care,33314
Constance,F,50-55,executive/managerial,10022
Larry,M,45-49,technician/engineer,2067
Mike,M,25-34,other or not specified,30606
Hank,M,50-55,programmer,44286
Daniel,M,45-49,technician/engineer,37923
Wesley,M,25-34,executive/managerial,35244
Gina,F,35-44,sales/marketing,60202
Teresa,F,45-49,academic/educator,43202
Terry,M,35-44,writer,80222
Leo,M,50-55,academic/educator,93105
Bruce,M,50-55,academic/educator,19087-3622
Terence,M,25-34,writer,14450
Alice,F,25-34,academic/educator,79928
Benjamin,M,25-34,technician/engineer,48092
Sharon,F,18-24,college/grad student,55406
Ryan,M,18-24,college/grad student,26241
Mason,M,25-34,technician/engineer,92584
Gloria,F,56+,retired,60506
Tom,M,25-34,writer,10010
Melissa,F,35-44,doctor/health care,23507
David,M,25-34,clerical/admin,19147
Alex,M,18-24,college/grad student,10013
Florence,F,35-44,academic/educator,23508
Darwin,M,45-49,customer service,98502
Michael,M,18-24,other or not specified,31211
Brown,M,25-34,executive/managerial,90210
Jimmy,M,25-34,writer,94122
Jay,M,18-24,programmer,43650
Gladys,F,18-24,programmer,5055
Denny,M,45-49,tradesman/craftsman,2557
Jack,M,50-55,other or not specified,94025
Edison,M,45-49,executive/managerial,85287-2702
Neil,M,35-44,scientist,48187
Jennifer,F,35-44,writer,75093
Caspar,M,25-34,other or not specified,3766
Mickey,M,18-24,programmer,97205
Arthur,M,25-34,executive/managerial,2139
Christine,F,25-34,academic/educator,32303
Adeline,F,Under 18,other or not specified,1036
Cody,M,18-24,college/grad student,78705
Hillary,F,35-44,executive/managerial,21117
```

- 电影表vertex_movie.csv：

```
American Beauty,1999,Comedy;Drama
Airplane!,1980,Comedy
Rushmore,1998,Comedy
Predator,1987,Action;Sci-Fi;Thriller
There's Something About Mary,1998,Comedy
The Shawshank Redemption,1994,Drama
Election,1999,Comedy
Clueless,1995,Comedy;Romance
The Crying Game,1992,Drama;Romance;War
Back to the Future,1985,Comedy;Sci-Fi
The Talented Mr. Ripley,1999,Drama;Mystery;Thriller
Life Is Beautiful (La vita ии bella),1997,Comedy;Drama
2001: A Space Odyssey,1968,Drama;Mystery;Sci-Fi;Thriller
Jaws,1975,Action;Horror
Jerry Maguire,1996,Drama;Romance
The Hunt for Red October,1990,Action;Thriller
Close Encounters of the Third Kind,1977,Drama;Sci-Fi
Star Wars: Episode IV - A New Hope,1977,Action;Adventure;Fantasy;Sci-Fi
```



```
Rocky,1976,Action;Drama
The Usual Suspects,1995,Crime;Thriller
A Clockwork Orange,1971,Sci-Fi
Psycho,1960,Horror;Thriller
The Godfather: Part II,1974,Action;Crime;Drama
Annie Hall,1977,Comedy;Romance
Terminator 2: Judgment Day,1991,Action;Sci-Fi;Thriller
Pleasantville,1998,Comedy
Chinatown,1974,Film-Noir;Mystery;Thriller
Independence Day (ID4),1996,Action;Sci-Fi;War
Star Wars: Episode V - The Empire Strikes Back,1980,Action;Adventure;Drama;Sci-Fi;War
Face/Off,1997,Action;Sci-Fi;Thriller
Total Recall,1990,Action;Adventure;Sci-Fi;Thriller
Blade Runner,1982,Film-Noir;Sci-Fi
The Terminator,1984,Action;Sci-Fi;Thriller
Robocop,1987,Action;Crime;Sci-Fi
The Rock,1996,Action;Adventure;Thriller
Superman,1978,Action;Adventure;Sci-Fi
The Full Monty,1997,Comedy
Raising Arizona,1987,Comedy
Lethal Weapon,1987,Action;Comedy;Crime;Drama
Platoon,1986,Drama;War
The Fifth Element,1997,Action;Sci-Fi
The Patriot,2000,Action;Drama;War
Clerks,1994,Comedy
Being John Malkovich,1999,Comedy
The Mask,1994,Comedy;Crime;Fantasy
Grosse Pointe Blank,1997,Comedy;Crime
```

- 朋友关系表edge_friends.csv

```
Gloria,David
Brown,Mason
Terence,Kenny
Clark,Brown
Mickey,Janet
Mickey,Margaret
Hayden,Constance
Frank,Janet
Lena,Darwin
Leo,Jimmy
Mercedes,Gavin
Hillary,Bruce
Leo,Neil
Terence,August
Sue,Wayne
Max,Denny
Max,Josephine
Hillary,Michael
Constance,Janet
Florence,Donald
Alice,Jacob
Roger,Sidney
Margaret,Frances
Roger,Fred
Fred,Donald
Margaret,Gavin
Fred,Gavin
Rachel,Janet
Alexander,Clark
Darwin,Cassandra
Jordan,Vivian
Terry,Larry
Hardy,Kevin
Terry,Rachel
Mercedes,Marshal
Marshal,Sharon
Jeffery,Tom
Terence,Max
Katherine,Stuart
Luke,Cassandra
```

```
Michael,Arthur
Luke,Editha
Neil,Mason
Darwin,Jessie
Marshal,Alex
Hardy,Margaret
Alexander,Eric
Mercedes,Caspar
Brown,Clark
Roger,Kevin
Benjamin,Max
Jessie,Adeline
Michael,Luke
Jimmy,Gloria
Isaiah,Frances
June,Darwin
Editha,Vivian
Caspar,Cassandra
Bruce,Denny
Caspar,Jacob
Isaiah,Ella
Mason,Ryan
Mercedes,Eugene
Roger,Josephine
Wayne,Alice
Hayden,Denny
Alexander,Colin
Larry,August
Jimmy,Brown
Jacob,William
Hardy,Gladys
Jessie,Caspar
Mason,Terence
June,Jennifer
Hardy,Arthur
Alexander,Solomon
Larry,Wayne
Larry,Gavin
Ella,Ray
Ella,Eric
Alice,Janet
Larry,Willy
Isaiah,Solomon
Benjamin,Leander
Isaiah,Sue
Caspar,Jordan
Ella,Jordan
Vivian,Eric
Max,Jay
Ryan,Hank
Ella,Colin
Luke,Alexander
Luke,Joshua
Wayne,Caspar
Wayne,Denny
Editha,Marshal
Ryan,Jessie
Michael,Cassandra
Solomon,Hillary
Jordan,Josephine
```

- 电影评分表edge_rate.csv：

```
Vivian,Lethal Weapon,5,2000/12/27 23:44
Mercedes,Raising Arizona,4,2000/12/27 23:51
Katherine,The Rock,3,2000/12/27 20:12
Stuart,The Mask,2,2000/12/27 20:00
Jacob,Face/Off,4,2000/12/27 20:12
Editha,There's Something About Mary,5,2000/12/27 20:06
Cassandra,Superman,4,2000/12/27 20:11
Sarah,American Beauty,4,2000/12/27 20:13
```



```
Hayden,Lethal Weapon,3,2000/12/27 20:09
Jeffery,2001: A Space Odyssey,4,2000/12/23 1:48
Bonnie,A Clockwork Orange,3,2000/12/22 23:23
Serena,Lethal Weapon,4,2000/12/22 23:24
Sidney,Raising Arizona,4,2000/12/22 23:24
Leander,Clerks,5,2000/12/12 16:58
Fred,Superman,5,2000/12/18 1:17
Roger,A Clockwork Orange,5,2000/12/13 23:54
Ella,Robocop,5,2000/12/13 23:44
Ray,The Talented Mr. Ripley,3,2000/12/14 0:24
Eric,Psycho,5,2002/1/3 20:29
Frances,The Godfather: Part II,2,2000/12/10 18:45
Allison,Independence Day (ID4),3,2000/12/13 23:58
Willy,Clerks,4,2002/1/3 20:46
Lance,There's Something About Mary,5,2000/12/13 23:43
June,Superman,4,2002/1/3 20:41
Marshal,Being John Malkovich,5,2000/12/10 18:40
Max,Predator,4,2000/12/10 18:32
Hardy,Total Recall,3,2000/12/10 18:39
Jordan,American Beauty,4,2000/12/13 23:57
Reed,Lethal Weapon,1,2000/12/10 18:37
Glendon,Airplane!,4,2000/12/13 23:46
Kevin,Raising Arizona,4,2000/12/13 23:51
Evan,Jerry Maguire,1,2000/12/13 23:58
Clark,The Hunt for Red October,5,2000/12/13 23:46
Johnny,2001: A Space Odyssey,3,2000/12/14 0:16
Caleb,Clerks,4,2000/12/9 16:45
Janet,Lethal Weapon,2,2000/12/9 16:16
Sue,Close Encounters of the Third Kind,4,2000/12/9 16:14
Margaret,Star Wars: Episode IV - A New Hope,2,2000/12/9 16:04
Luke,Clueless,2,2000/12/8 19:02
William,The Terminator,2,2000/12/8 19:03
Lena,Robocop,5,2000/12/8 18:59
Solomon,Lethal Weapon,5,2000/12/8 18:59
Cary,Airplane!,5,2000/12/8 19:00
Colin,The Usual Suspects,4,2000/12/5 20:59
Kenny,Clueless,5,2000/12/5 20:52
Gavin,A Clockwork Orange,4,2000/12/5 20:52
Donald,The Talented Mr. Ripley,3,2000/12/5 20:52
Wayne,Back to the Future,3,2000/12/5 20:56
Frank,Being John Malkovich,4,2000/12/5 20:53
Alexander,Predator,5,2000/12/5 20:52
Isaiah,Jaws,4,2000/12/5 20:48
Josephine,Chinatown,3,2000/12/5 20:55
Joshua,The Mask,4,2000/12/5 20:54
August,Platoon,4,2000/12/5 20:53
Jessie,Election,4,2000/12/5 20:52
Yvette,Rocky,5,2000/12/5 20:52
Albert,The Fifth Element,4,2000/12/5 20:55
Eugene,Clueless,4,2000/12/5 17:59
Rachel,Lethal Weapon,5,2000/12/5 17:58
Constance,Raising Arizona,4,2000/12/5 17:59
Larry,The Usual Suspects,4,2000/12/5 15:07
Mike,The Crying Game,5,2000/12/5 15:21
Hank,Independence Day (ID4),4,2000/12/5 15:21
Daniel,There's Something About Mary,4,2000/12/5 15:10
Wesley,Lethal Weapon,5,2000/12/2 19:51
Gina,The Godfather: Part II,3,2000/12/2 19:55
Teresa,Total Recall,4,2000/12/2 19:44
Terry,2001: A Space Odyssey,4,2000/12/2 19:53
Leo,A Clockwork Orange,5,2000/11/28 23:22
Bruce,The Full Monty,2,2000/11/28 23:12
Terence,Predator,5,2000/11/28 23:07
Alice,Jaws,5,2000/11/28 23:20
Benjamin,Psycho,3,2000/11/28 23:08
Sharon,Total Recall,5,2000/11/28 23:13
Ryan,Election,5,2000/11/28 23:18
Mason,The Fifth Element,2,2000/11/28 23:26
Gloria,The Usual Suspects,5,2000/11/28 12:57
```



```
Tom,Clueless,3,2000/11/28 13:09
Melissa,A Clockwork Orange,3,2000/12/8 15:10
David,The Talented Mr. Ripley,5,2000/12/25 13:24
Alex,Independence Day (ID4),4,2000/11/28 13:14
Florence,Star Wars: Episode V - The Empire Strikes Back,2,2000/12/8 15:23
Darwin,The Full Monty,2,2000/11/28 13:16
Michael,Being John Malkovich,4,2000/12/25 14:44
Brown,Predator,5,2000/11/28 13:01
Jimmy,Lethal Weapon,4,2000/12/8 15:07
Jay,Jaws,4,2000/11/28 13:07
Gladys,Psycho,4,2000/11/28 13:08
Denny,The Godfather: Part II,3,2000/12/25 13:25
Jack,Annie Hall,4,2000/12/8 15:05
Edison,The Mask,3,2000/11/28 13:11
Neil,Face/Off,4,2000/12/8 15:22
Jennifer,There's Something About Mary,3,2000/12/25 6:17
Caspar,Superman,3,2000/12/8 15:09
Mickey,Total Recall,1,2000/11/28 13:14
Arthur,American Beauty,3,2000/12/8 15:18
Christine,Platoon,3,2000/12/2 13:21
Adeline,Raising Arizona,4,2000/12/8 15:15
Cody,Blade Runner,1,2000/12/8 15:22
Hillary,Election,3,2000/11/28 12:57
```

在管理中心创建数据连接

在本示例中，我们需要将MySQL原始数据同步到MRS Hive中并按照GES图导入要求标准化，然后基于MRS Hive生成元数据。

因此在准备工作中，需要先在管理中心创建MRS连接。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 10-3 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10-4**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 10-4 创建 MRS Hive 数据连接

----结束

创建数据表

本例中为了方便演示，我们需要通过数据集成将CSV格式的样例数据导入到MySQL数据库中，之后MySQL数据库即作为案例场景中的原始数据源端。因此在数据导入中，需要在MySQL数据库中预先创建原始数据表。

正式业务流程中，MySQL数据库源端数据需要导入OBS数据库作为点数据集和边数据集，这种到OBS的数据集成场景无需提前创建表。但MySQL数据库源端数据导入到MRS Hive时，需要在MRS Hive数据库中预先创建标准数据表。

因此，本例共涉及MySQL数据库创建原始数据表和在MRS Hive数据库中创建标准数据表。本例以执行SQL方式建表为例进行说明。

步骤1 创建MySQL原始数据表。在MySQL中选择原始表所在的数据库后，执行如下SQL语句，按照**数据源准备**中的原始数据结构创建4个原始数据表。

```
DROP TABLE IF EXISTS `edge_friends`;
CREATE TABLE `edge_friends` (
  `user1` varchar(32) DEFAULT NULL,
```



```
    `user2` varchar(32) DEFAULT NULL
);

DROP TABLE IF EXISTS `edge_rate`;
CREATE TABLE `edge_rate` (
    `user` varchar(32) DEFAULT NULL,
    `movie` varchar(64) DEFAULT NULL,
    `score` int(11) unsigned DEFAULT NULL,
    `datatime` varchar(32) DEFAULT NULL
);

DROP TABLE IF EXISTS `vertex_movie`;
CREATE TABLE `vertex_movie` (
    `movie` varchar(64) DEFAULT NULL,
    `year` varchar(32) DEFAULT NULL,
    `genres` varchar(64) DEFAULT NULL
);

DROP TABLE IF EXISTS `vertex_user`;
CREATE TABLE `vertex_user` (
    `user` varchar(32) DEFAULT NULL,
    `gender` varchar(32) DEFAULT NULL,
    `age` varchar(32) DEFAULT NULL,
    `occupation` varchar(32) DEFAULT NULL,
    `zip-code` varchar(32) DEFAULT NULL
);
```

步骤2 创建MRS Hive标准数据表。

将原始数据结构根据GES图导入的要求标准化。则点表vertex_user和vertex_movie需要在第二列补充标签label，边表edge_rate和edge_friends需要在第三列补充标签label。

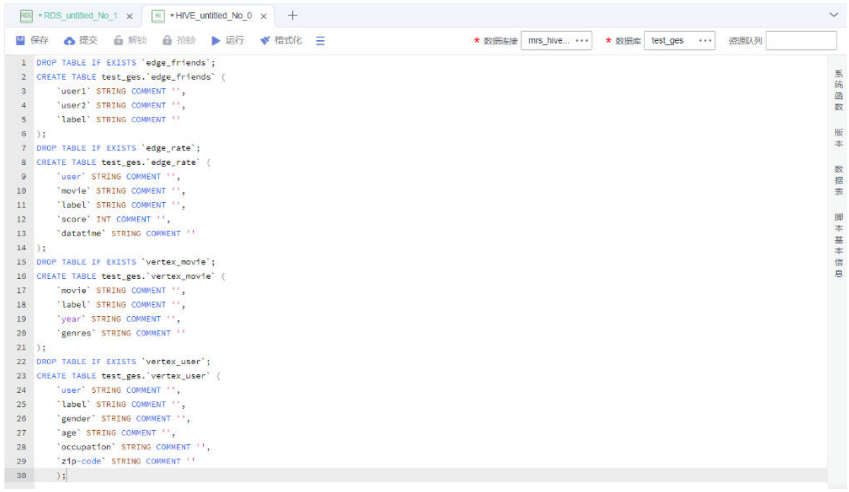
须知

点数据集和边数据集应符合GES图数据格式要求。图数据格式要求简要介绍如下，详情可参见**一般图数据格式**。

- 点数据集罗列了各个点的数据信息。一行为一个点的数据。格式如下所示，id是点数据的唯一标识。
  ```
  id,label,property 1,property 2,property 3,…
  ```
- 边数据集罗列了各个边的数据信息，一行为一条边的数据。GES中图规格是以边的数量进行定义的，如一百万边。格式如下所示，id 1、id 2是一条边的两个端点的id。
  ```
  id 1, id 2, label, property 1, property 2, …
  ```

您可以在DataArts Studio数据开发模块，选择**在管理中心创建数据连接**中创建的MRS Hive数据连接，并选择数据库后，执行如下SQL语句，在MRS Hive数据库中创建一个标准数据表。

图 10-5 创建 MRS Hive 标准数据表



```
DROP TABLE IF EXISTS `edge_friends`;
CREATE TABLE test_ges.`edge_friends` (
    `user1` STRING COMMENT '',
    `user2` STRING COMMENT '',
    `label` STRING COMMENT ''
);

DROP TABLE IF EXISTS `edge_rate`;
CREATE TABLE test_ges.`edge_rate` (
    `user` STRING COMMENT '',
    `movie` STRING COMMENT '',
    `label` STRING COMMENT '',
    `score` INT COMMENT '',
    `datatime` STRING COMMENT ''
);

DROP TABLE IF EXISTS `vertex_movie`;
CREATE TABLE test_ges.`vertex_movie` (
    `movie` STRING COMMENT '',
    `label` STRING COMMENT '',
    `year` STRING COMMENT '',
    `genres` STRING COMMENT ''
);

DROP TABLE IF EXISTS `vertex_user`;
CREATE TABLE test_ges.`vertex_user` (
    `user` STRING COMMENT '',
    `label` STRING COMMENT '',
    `gender` STRING COMMENT '',
    `age` STRING COMMENT '',
    `occupation` STRING COMMENT '',
    `zip-code` STRING COMMENT ''
);
```

----结束

10.3 创建数据集成作业

本章节将介绍如何创建DataArts Studio数据集成作业。

本例中，需要创建如下三类集成作业：

1. **OBS到MySQL迁移作业**：为方便演示，需要将OBS中的CSV格式的样例数据导入到MySQL数据库中。

2. **MySQL到OBS迁移作业**：正式业务流程中，需要将MySQL中的原始样例数据需要导入OBS中，并标准化为点数据集和边数据集。
3. **MySQL到MRS Hive迁移作业**：正式业务流程中，需要将MySQL中的原始样例数据需要导入MRS Hive中，并标准化为点数据集和边数据集。

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。

新建数据集成连接

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 10-6 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 10-7 集群管理

步骤3 进入作业管理后，选择“连接管理”。

图 10-8 连接管理

步骤4 创建集成任务所需的OBS连接、云数据库MySQL连接和MRS Hive连接。

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 10-9 创建 OBS 连接

表 10-1 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10-10所示。 **图 10-10** 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)		-

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“云数据库 MySQL”，单击“下一步”，然后如下图所示配置连接参数，单击“保存”。

图 10-11 创建 MySQL 连接

表 10-2 MySQL 连接参数

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

参数名	说明	取值样例
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	否
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

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 10-12 创建 MRS Hive 连接

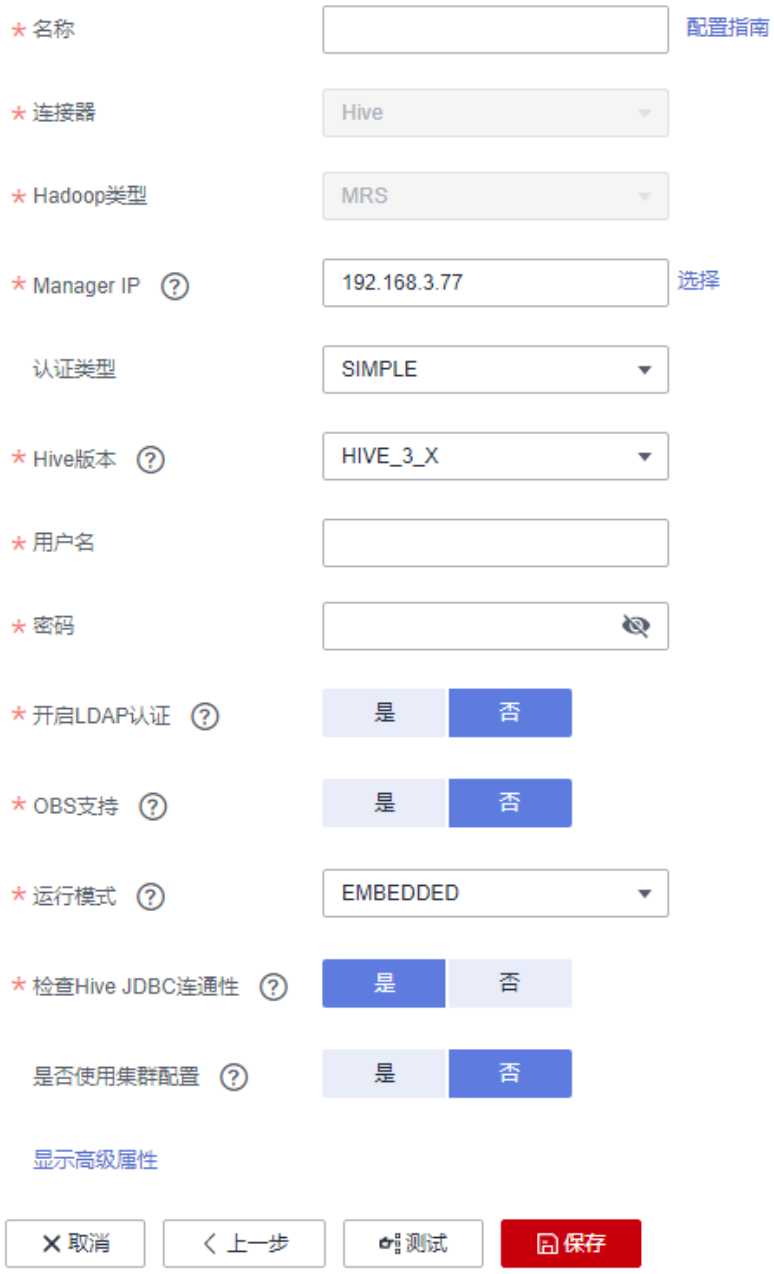


表 10-3 MRS Hive 连接参数

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”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”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10-13所示。 **图 10-13** 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 OBS 到 MySQL 迁移作业

为方便演示，需要将OBS中的CSV格式的样例数据导入到MySQL数据库中。

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 10-14 表/文件迁移

步骤3 按照如下步骤将**数据源准备**中的4张原始数据表，依次从OBS迁移到MySQL数据库中。

1. 配置作业vertex_user_obs2rds。

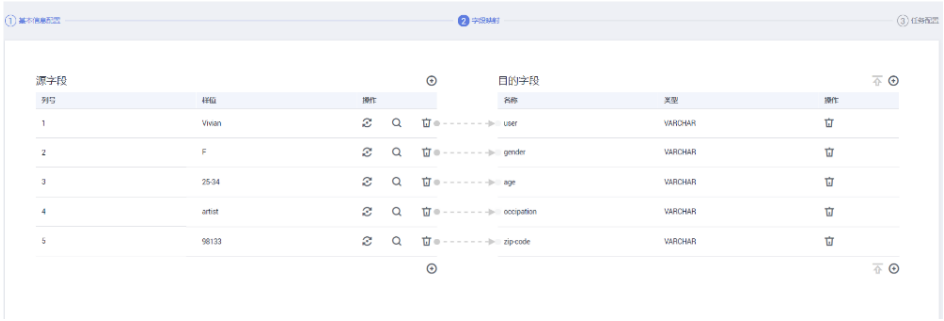
   源端的“源目录或文件”选择在**数据源准备**中上传到OBS的vertex_user.csv，由于表中有中文字符还需额外配置高级属性“编码类型”为“GBK”。目的端的“表名”选择在**创建MySQL原始数据表**中创建的vertex_user表。然后单击“下一步”。

   图 10-15 vertex_user_obs2rds 作业配置

2. 在字段映射中，检查字段映射顺序是否正确。如果字段映射顺序正确，单击下一步即可。

图 10-16 vertex_user_obs2rds 字段映射

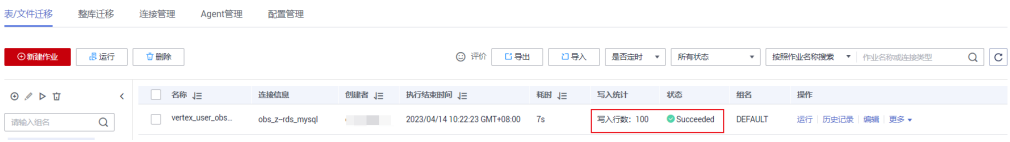


3. 任务配置无需修改，直接保存并运行即可。

图 10-17 任务配置

步骤4 等待作业运行完成后，如果作业成功，则vertex_user表已成功迁移到MySQL数据库中。

图 10-18 vertex_user_obs2rds 作业运行成功



步骤5 参考**步骤2**到**步骤4**，完成vertex_movie_obs2rds、edge_friends_obs2rds和edge_rate_obs2rds作业的创建，将4张原始表从OBS迁移到MySQL中。

----结束

新建 MySQL 到 OBS 迁移作业

正式业务流程中，需要将MySQL中的原始样例数据需要导入OBS中，并标准化为点数据集和边数据集。

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 10-19 表/文件迁移

步骤3 按照如下步骤将MySQL中的4张原始数据表，依次迁移到OBS桶中。

1. 配置作业vertex_user_rds2obs。

   源端的“表名”选择在**新建OBS到MySQL迁移作业**中迁移到MySQL的vertex_user。目的端的“写入目录”注意选择非原始数据所在目录以避免文件覆盖，“文件格式”按照GES图导入格式要求设置为“CSV格式”，由于表中有中文字符还需额外配置高级属性“编码类型”为“GBK”。

   注意：目的端高级属性需要额外配置“自定义文件名”，取值为“${tableName}”。如果不配置，则迁移到OBS的CSV文件名会带上时间戳等额外字段，导致每次运行迁移作业获取的文件名不一致，无法每次迁移后自动导入GES图数据。

   其他高级属性无需配置，单击“下一步”。

   图 10-20 vertex_user_rds2obs 作业基础配置

图 10-21 vertex_user_rds2obs 作业高级配置

2. 在字段映射中，根据GES图数据的要求，此处需要新增字段label，作为图文件的标签。
   - vertex_user：label取值为user，并将此字段调整至第2列。
   - vertex_movie：label取值为movie，并将此字段调整至第2列。
   - edge_friends：label取值为friends，并将此字段调整至第3列。
   - edge_rate：label取值为rate，并将此字段调整至第3列。

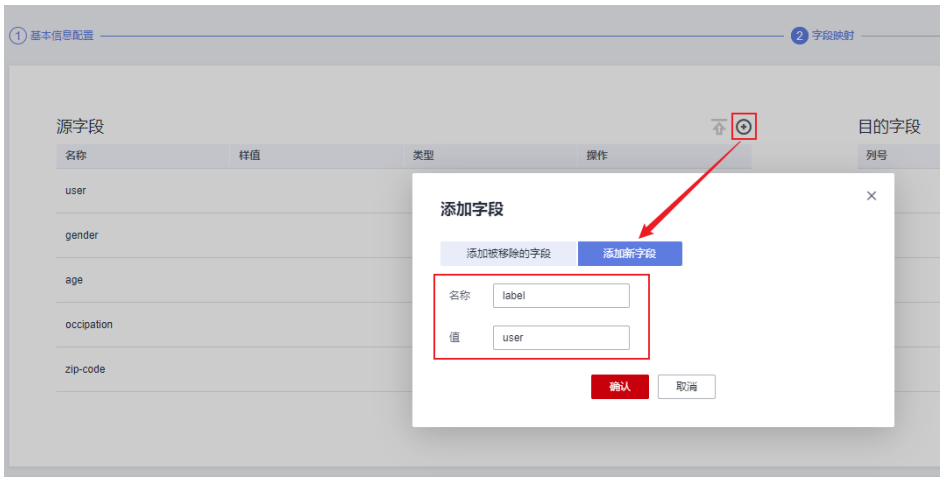
   将原始数据结构根据GES图导入的要求标准化。则点表vertex_user和vertex_movie需要在第二列补充标签label，边表edge_rate和edge_friends需要在第三列补充标签label。

须知

点数据集和边数据集应符合GES图数据格式要求。图数据格式要求简要介绍如下，详情可参见**一般图数据格式**。

- 点数据集罗列了各个点的数据信息。一行为一个点的数据。格式如下所示，id是点数据的唯一标识。
  ```
  id,label,property 1,property 2,property 3,…
  ```
- 边数据集罗列了各个边的数据信息，一行为一条边的数据。GES中图规格是以边的数量进行定义的，如一百万边。格式如下所示，id 1、id 2是一条边的两个端点的id。
  ```
  id 1, id 2, label, property 1, property 2, …
  ```

图 10-22 vertex_user_rds2obs 新增字段映射



3. 调整字段顺序，点数据集将label调整至第2列，边数据集将label调整至第3列。调整完成后如**图10-24**所示，然后单击下一步。

图 10-23 vertex_user_rds2obs 调整字段顺序

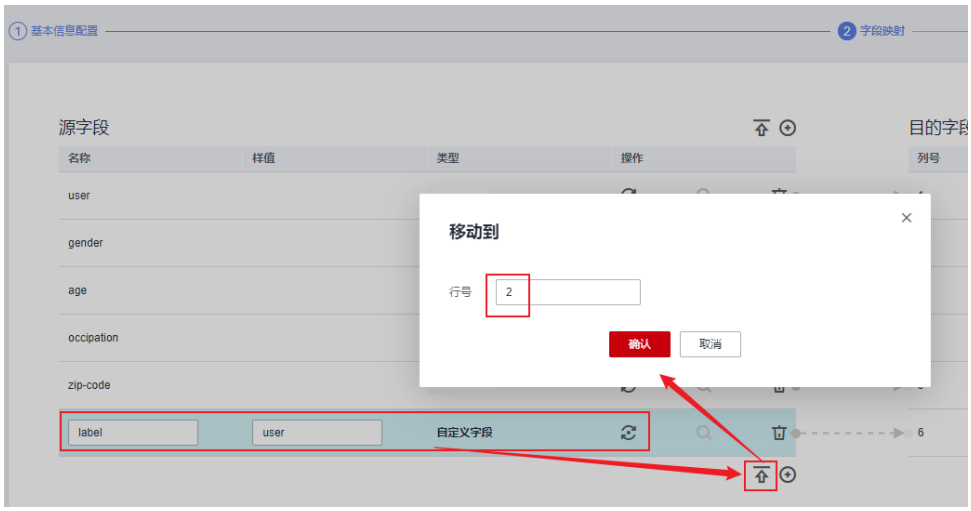
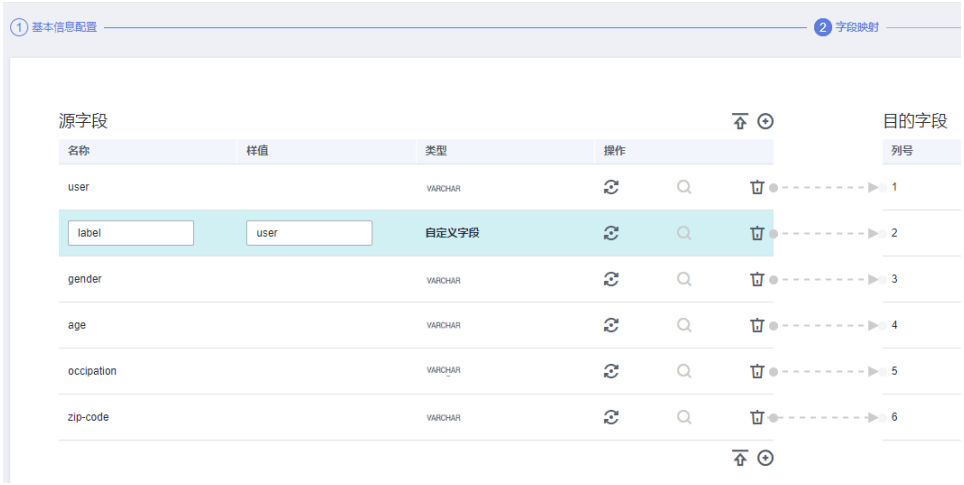


图 10-24 vertex_user_rds2obs 字段映射

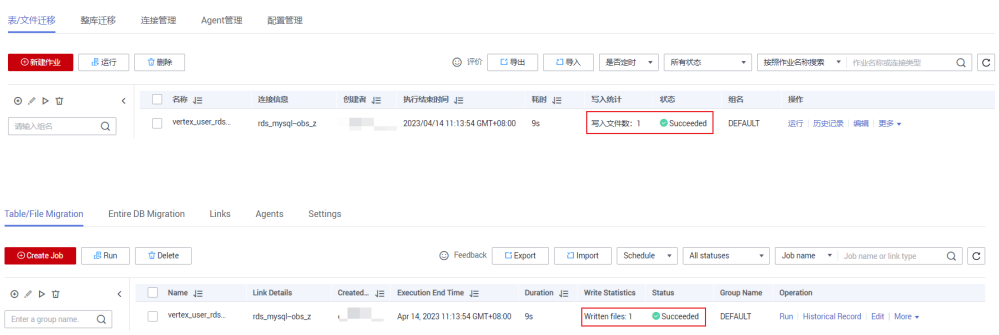


4. 任务配置无需修改，直接保存并运行即可。

图 10-25 任务配置

步骤4 等待作业运行完成后，如果作业成功，则vertex_user.csv表已成功写入到OBS桶中。

图 10-26 vertex_user_rds2obs 作业运行成功



步骤5 参考**步骤2**到**步骤4**，完成vertex_movie_rds2obs、edge_friends_rds2obs和edge_rate_rds2obs作业的创建，将4张原始表从MySQL标准化到OBS桶中。

----结束

新建 MySQL 到 MRS Hive 迁移作业

正式业务流程中，需要将MySQL中的原始样例数据需要导入MRS Hive中，并标准化为点数据集和边数据集。

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”。

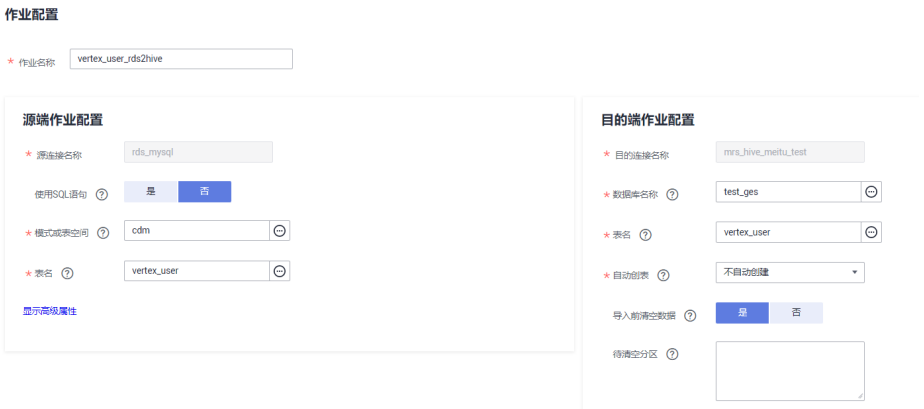
图 10-27 表/文件迁移

步骤3 按照如下步骤将MySQL中的4张原始数据表，依次迁移到MRS Hive中。

1. 配置作业vertex_user_rds2hive。

   源端的“表名”选择在**新建OBS到MySQL迁移作业**中迁移到MySQL的vertex_user，目的端的“表名”选择在**创建MRS Hive标准数据表**中创建的vertex_user表。其他参数配置如图所示，无需配置高级属性，然后单击“下一步”。

   图 10-28 vertex_user_rds2hive 作业基础配置



2. 在字段映射中，根据GES图数据的要求，此处需要新增字段label，作为图文件的标签。
   - vertex_user：label取值为user，并将此字段调整至第2列。
   - vertex_movie：label取值为movie，并将此字段调整至第2列。
   - edge_friends：label取值为friends，并将此字段调整至第3列。

- edge_rate：label取值为rate，并将此字段调整至第3列。

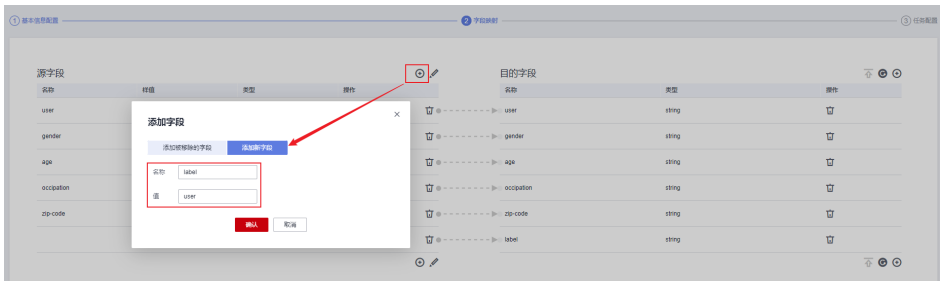
将原始数据结构根据GES图导入的要求标准化。则点表vertex_user和vertex_movie需要在第二列补充标签label，边表edge_rate和edge_friends需要在第三列补充标签label。

须知

点数据集和边数据集应符合GES图数据格式要求。图数据格式要求简要介绍如下，详情可参见**一般图数据格式**。

- 点数据集罗列了各个点的数据信息。一行为一个点的数据。格式如下所示，id是点数据的唯一标识。
  ```
  id,label,property 1,property 2,property 3,…
  ```
- 边数据集罗列了各个边的数据信息，一行为一条边的数据。GES中图规格是以边的数量进行定义的，如一百万边。格式如下所示，id 1、id 2是一条边的两个端点的id。
  ```
  id 1, id 2, label, property 1, property 2, …
  ```

图 10-29 vertex_user_rds2hive 新增字段映射



3. 调整字段顺序，点文件中将label调整至第2列，边文件将label调整至第3列。调整完成后如**图10-31**所示，然后单击下一步。

图 10-30 vertex_user_rds2hive 调整字段顺序

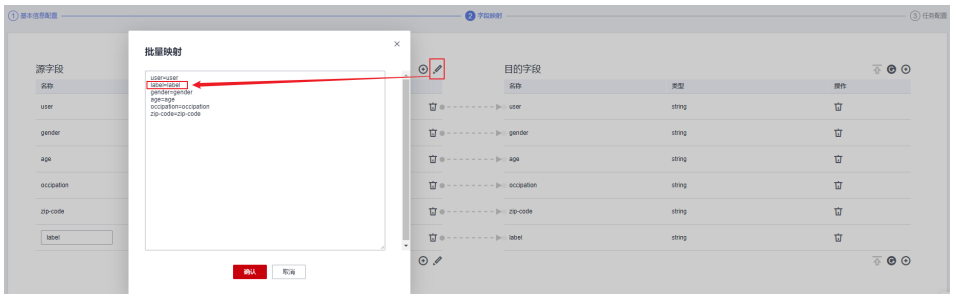
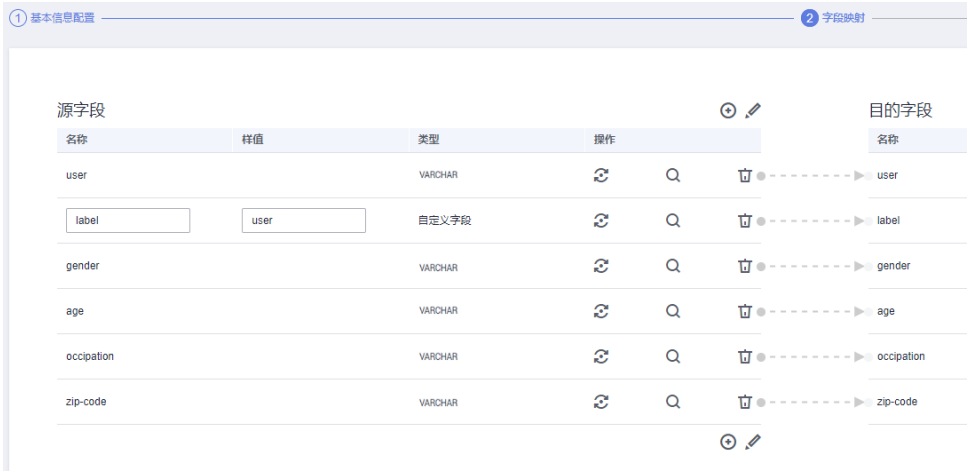


图 10-31 vertex_user_rds2hive 字段映射



4. 任务配置无需修改，直接保存并运行即可。

图 10-32 任务配置

步骤4 等待作业运行完成后，如果作业成功，则vertex_user表已成功迁移到MRS Hive中。

图 10-33 vertex_user_rds2hive 作业运行成功

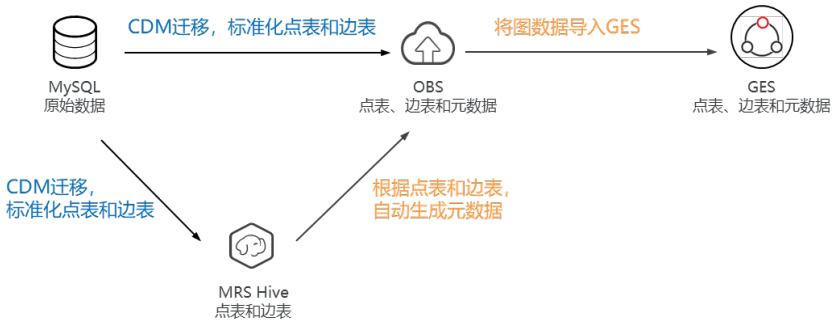
步骤5 参考**步骤2**到**步骤4**，完成vertex_movie_rds2hive、edge_friends_rds2hive和edge_rate_rds2hive作业的创建，将4张原始表从MySQL标准化到MRS Hive中。

----结束

10.4 开发并调度 Import GES 作业

本章节介绍通过数据开发调用数据集成作业，将MySQL原始数据定期同步到OBS和MRS Hive中，并标准化为GES点/边数据集。然后基于标准化点/边数据集，自动生成图的元数据，实现最终将图数据（点数据集、边数据集和元数据）定期导入到GES服务中。

图 10-34 业务场景说明



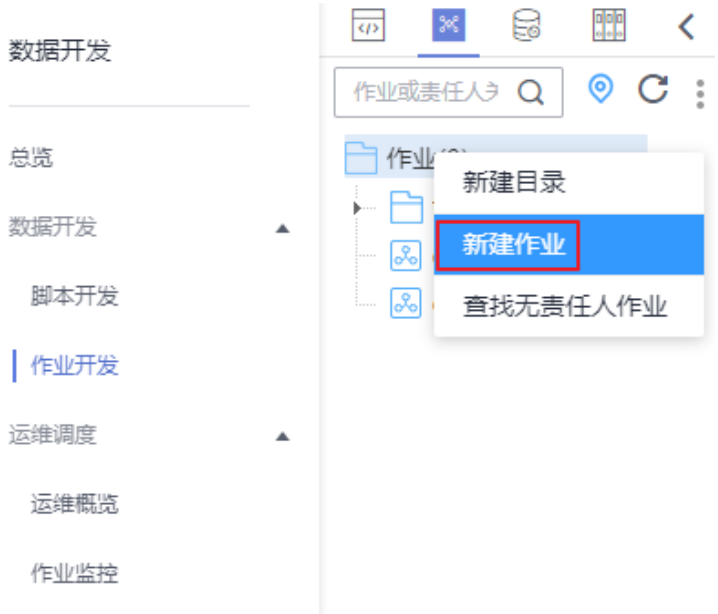
开发并调度 Import GES 作业

假设MySQL中的原始数据表是每日更新的，我们希望每天都能将基于原始数据的最新图数据更新到GES中，则需要使用数据开发按如下步骤编排作业，并定期调度。

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，作业名称可以命名为“import_ges”。

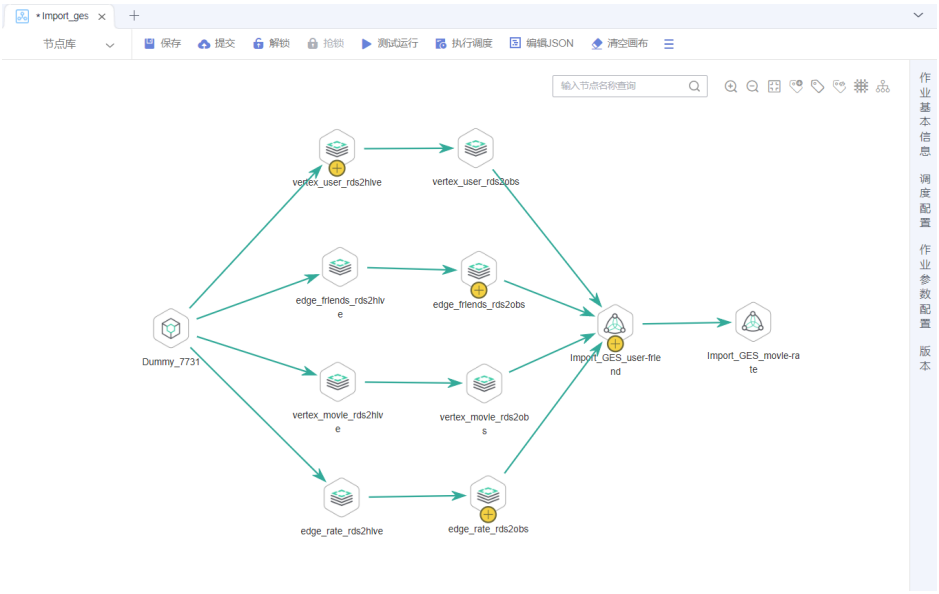
图 10-35 新建作业



步骤3 在作业开发页面，拖动1个Dummy节点、8个CDM Job节点、和2个Import GES节点到画布中，选中连线图标并拖动，编排图10-36所示的作业。

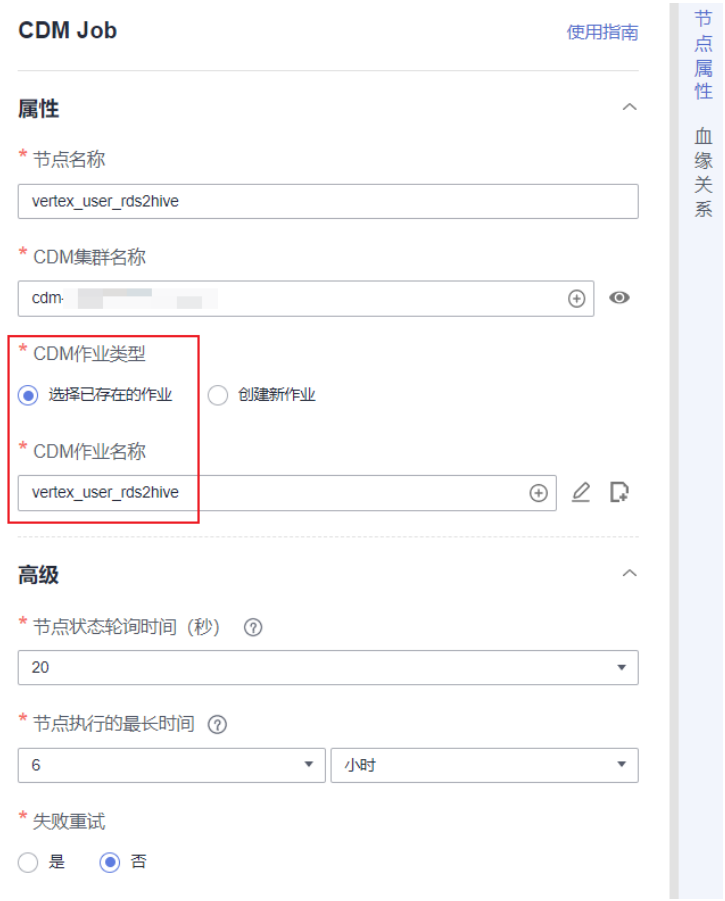
其中Dummy节点不执行任何操作，只作为起始点的标识。CDM Job节点用于调用在**创建数据集成作业**中创建的数据集成作业。Import GES节点用于将图数据导入GES。

图 10-36 编排作业



步骤4 分别配置作业中的8个CDM Job节点。调用已创建的数据集成作业，将原始数据标准化为GES点/边数据集，并同步到OBS和MRS Hive中。

图 10-37 配置 CDM 节点



CDM节点说明：

- vertex_user_rds2hive（CDM Job节点）：在节点属性中，选择**创建数据集成作业**中的CDM集群，并关联CDM作业“vertex_user_rds2hive”。
- vertex_user_rds2obs（CDM Job节点）：在节点属性中，选择**创建数据集成作业**中的CDM集群，并关联CDM作业“vertex_user_rds2obs”。
- edge_friends_rds2hive（CDM Job节点）：在节点属性中，选择**创建数据集成作业**中的CDM集群，并关联CDM作业“edge_friends_rds2hive”。
- edge_friends_rds2obs（CDM Job节点）：在节点属性中，选择**创建数据集成作业**中的CDM集群，并关联CDM作业“edge_friends_rds2obs”。
- vertex_movie_rds2hive（CDM Job节点）：在节点属性中，选择**创建数据集成作业**中的CDM集群，并关联CDM作业“vertex_movie_rds2hive”。
- vertex_movie_rds2obs（CDM Job节点）：在节点属性中，选择**创建数据集成作业**中的CDM集群，并关联CDM作业“vertex_movie_rds2obs”。
- edge_rate_rds2hive（CDM Job节点）：在节点属性中，选择**创建数据集成作业**中的CDM集群，并关联CDM作业“edge_rate_rds2hive”。
- edge_rate_rds2obs（CDM Job节点）：在节点属性中，选择**创建数据集成作业**中的CDM集群，并关联CDM作业“edge_rate_rds2obs”。

步骤5 分别配置作业中的2个Import GES节点。由于1个Import GES节点只能选择一张点表和一张边表，并生成对应的元数据，因此本示例中使用2个Import GES节点依次进行导入。

Import GES节点说明：

- Import_GES_user-friend：在节点属性中，选择图名称后，边数据集和点数据集分别填写为“edge_friends”边表和“vertex_user”点表。另外，应配置为不允许重复边，否则定期调度后将产生大量重复边。

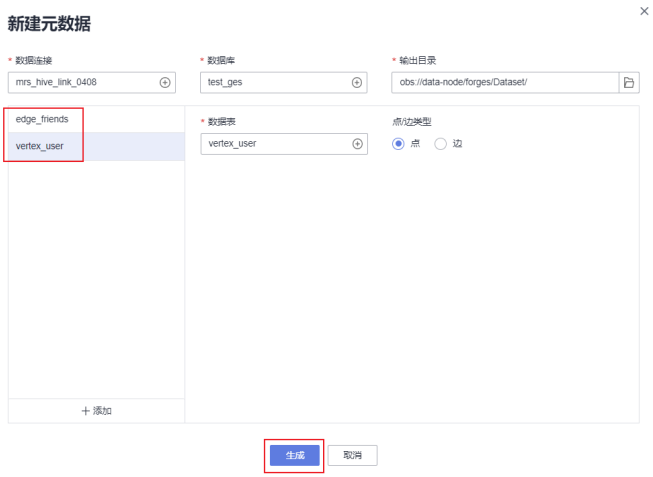
  注意，“元数据来源”需要选择为“新建元数据”，然后单击“元数据”参数后的生成按钮，弹出新建元数据窗口，如**图10-39**所示。在新建元数据窗口内，分别选择MRS中的“edge_friends”边表和“vertex_user”点表，输出目录可以设置为OBS点表和边表所在目录，然后单击生成，系统会自动在“元数据”参数处回填已生成的元数据Schema所在的OBS目录。

- Import_GES_movie-rate：在节点属性中，选择图名称后，边数据集和点数据集分别填写为“edge_rate”边表和“vertex_movie”点表。另外，应配置为不允许重复边，否则定期调度后将产生大量重复边。

  注意，“元数据来源”需要选择为“新建元数据”，然后单击“元数据”参数后的生成按钮，弹出新建元数据窗口，如**图10-39**所示。在新建元数据窗口内，分别选择MRS中的“edge_rate”边表和“vertex_movie”点表，输出目录可以设置为OBS点表和边表所在目录，然后单击生成，系统会自动在“元数据”参数处回填已生成的元数据Schema所在的OBS目录。

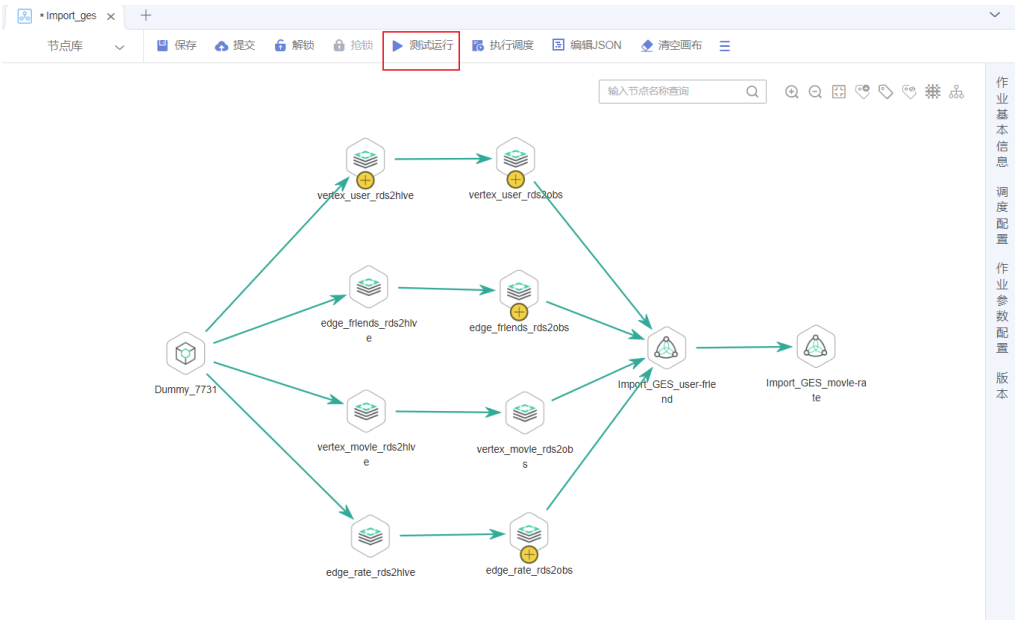
图 10-38 配置 Import GES 节点

图 10-39 新建元数据



步骤6 作业编排完成后，单击，测试运行作业。

图 10-40 测试运行作业



步骤7 如果作业运行正常，单击“调度配置”，配置作业的调度策略。

图 10-41 调度配置

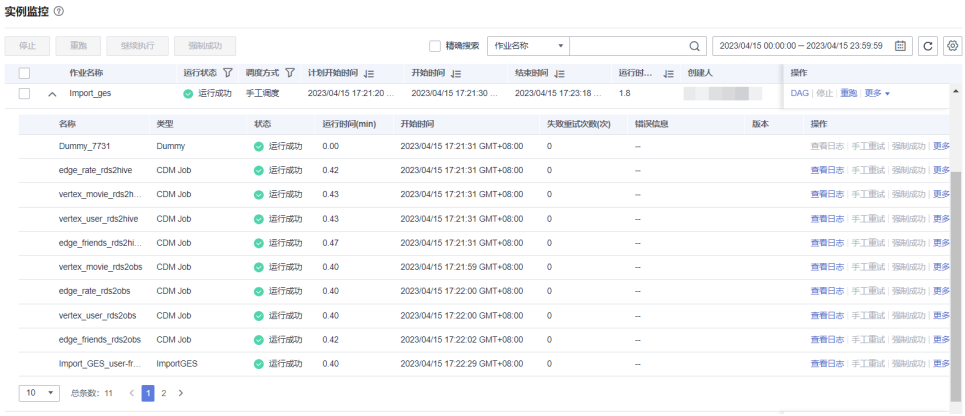
说明：

- 2023/04/01 00:00开始，每天00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8 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，每日数据将自动导入到GES图中。

步骤9 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10-42所示。

图 10-42 查看作业执行情况



----结束

10.5 分析图数据

通过GES直接对图数据进行可视化分析。

前提条件

已完成**开发并调度Import GES作业**，且作业运行成功。

通过 GES 分析数据

1. 进入图引擎服务GES控制台，在“图管理”页面中单击对应图后的“访问”按钮。

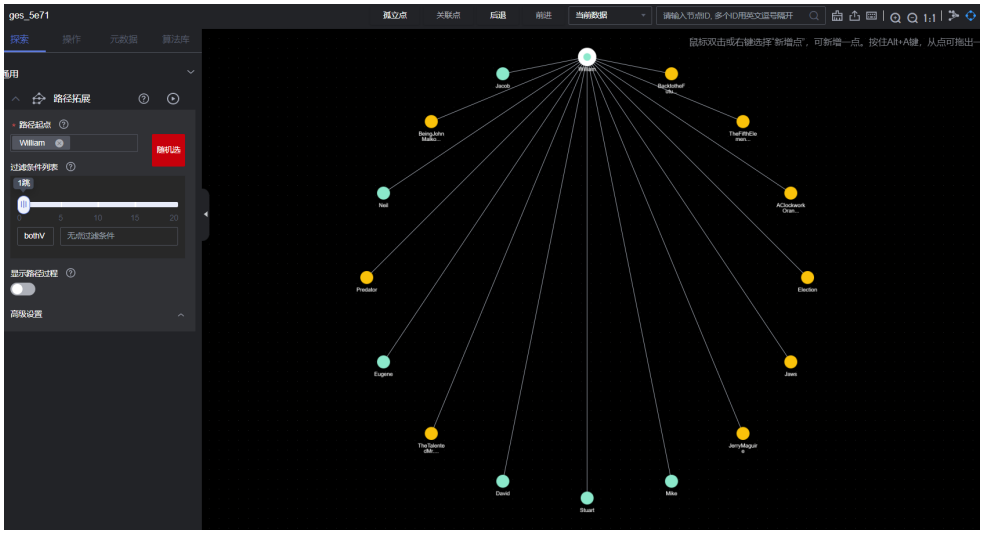
   图 10-43 访问图

2. 参考**访问图和分析图**，对导入的图数据进行可视化分析。

   本例以图探索功能为例，查看用户William相关的用户与电影情况，如**图10-44**所示。

图 10-44 分析图数据



11 案例：贸易数据统计与分析

11.1 场景介绍

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。

场景描述

H公司是国内一家收集主要贸易国贸易统计及买家数据的商业机构，拥有大量的贸易统计数据库，其数据广泛应用于产业研究、行业研究、国际贸易促进等方面。

在这之前，H公司采用其自建的大数据集群，并安排专人维护，每年固定购买电信联通双线专用带宽，在机房、电力、专网、服务器、运维方面进行高额投入，但其在面对客户不断变化的业务诉求时，因为人员投入不足，大数据集群能力不匹配，而无法聚焦业务创新，使得存量100T的数据只有4%的利用率。

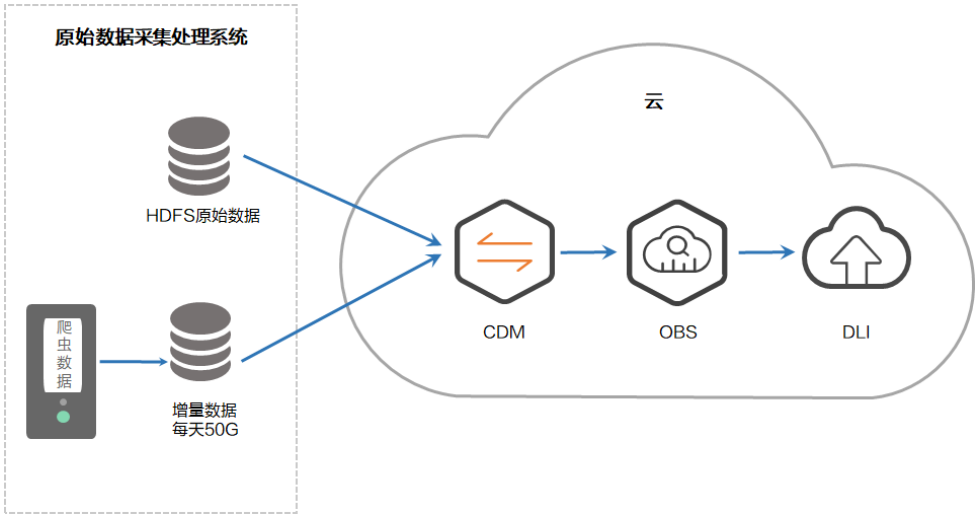
在将本地的贸易统计数据迁移到华为云之后，基于华为公有云的大数据分析能力，可帮助H公司屏蔽大数据基础设施复杂的构建、维护过程，使其客户人员可以全身心聚焦业务创新，盘活100T的存量数据，使资产最大化变现。

CDM和DLI服务按需收费，帮助H公司客户释放了维护人员并降低了专用带宽成本，使得维护成本相比线下数据中心降低了70%，且使用门槛低，可实现已有数据的平滑迁移，使新业务上线周期相比之前缩短了50%。

场景任务

根据客户原始数据采集处理系统中已有的H公司的数据（例如：贸易详单数据和基础信息数据），基于CDM+OBS+DLI完成贸易统计分析。

图 11-1 场景方案



说明

DLI创建OBS外表，对OBS表数据存储格式有所要求：

- 使用DataSource语法创建OBS表时，支持orc，parquet，json，csv，carbon，avro类型。
- 使用Hive语法创建OBS表时，支持TEXTFILE, AVRO, ORC, SEQUENCEFILE, RCFILE, PARQUET, CARBON类型。

如果原始数据表存储格式不满足要求，您可以通过CDM将原始数据直接导入到DLI中进行分析，无需上传OBS。

数据说明

- 贸易详单数据

  包括主要贸易国货物贸易统计数据。

表 11-1 贸易详单数据

字段名称	字段类型	字段说明
hs_code	string	进出口商品编码列表
country	smallint	国家基础信息
dollar_value	double	交易金额
quantity	double	交易量
unit	smallint	计量单位
b_country	smallint	目标国家基础信息
imex	smallint	进出口类型
y_year	smallint	年
m_month	smallint	月

- 基础信息数据
  贸易详单数据中维度字段对应的相关字典数据信息。

表 11-2 国家基础信息表（country）

字段名称	字段类型	字段说明
countryid	smallint	国家编码
country_en	string	国家英文名称
country_cn	string	国家中文名称

表 11-3 更新时间信息表（updatetime）

字段名称	字段类型	字段说明
countryid	smallint	国家编码
imex	smallint	进出口类型
hs_len	smallint	商品编码长度
minstartdate	string	最小开始时间
startdate	string	开始时间
newdate	string	更新时间
minnewdate	string	最小更新时间

表 11-4 进出口商品编码信息表（hs246）

字段名称	字段类型	字段说明
id	bigint	编号
hs	string	商品编码
hs_cn	string	商品中文名称
hs_en	string	商品英文名称

表 11-5 单位信息表（unit_general）

字段名称	字段类型	字段说明
id	smallint	计量单位编码
unit_en	string	计量单位英文名称
unit_cn	string	计量单位中文名称

11.2 操作流程概述

流程介绍

使用CDM+OBS+DLI进行贸易统计分析的操作过程主要包括2个步骤：

1. **使用CDM上传数据到OBS**
   a. 通过CDM将H公司存量数据上传到对象存储服务OBS。
   b. 通过CDM作业的定时任务，每天自动上传增量数据到OBS。
2. **使用DLI分析数据**
   通过DLI直接分析OBS中的业务数据，支撑H公司客户进行贸易统计分析。

11.3 使用 CDM 上传数据到 OBS

11.3.1 上传存量数据

1. 使用**华为云专线**，搭建用户本地数据中心与华为云VPC之间的专属连接通道。
2. 创建OBS桶，并记录OBS的访问域名、端口和AK/SK。
3. 创建CDM集群。

   说明

   DataArts Studio实例中已经包含一个CDM集群（试用版除外），如果该集群已经满足需求，您无需再购买数据集成增量包，可以跳过这部分内容。

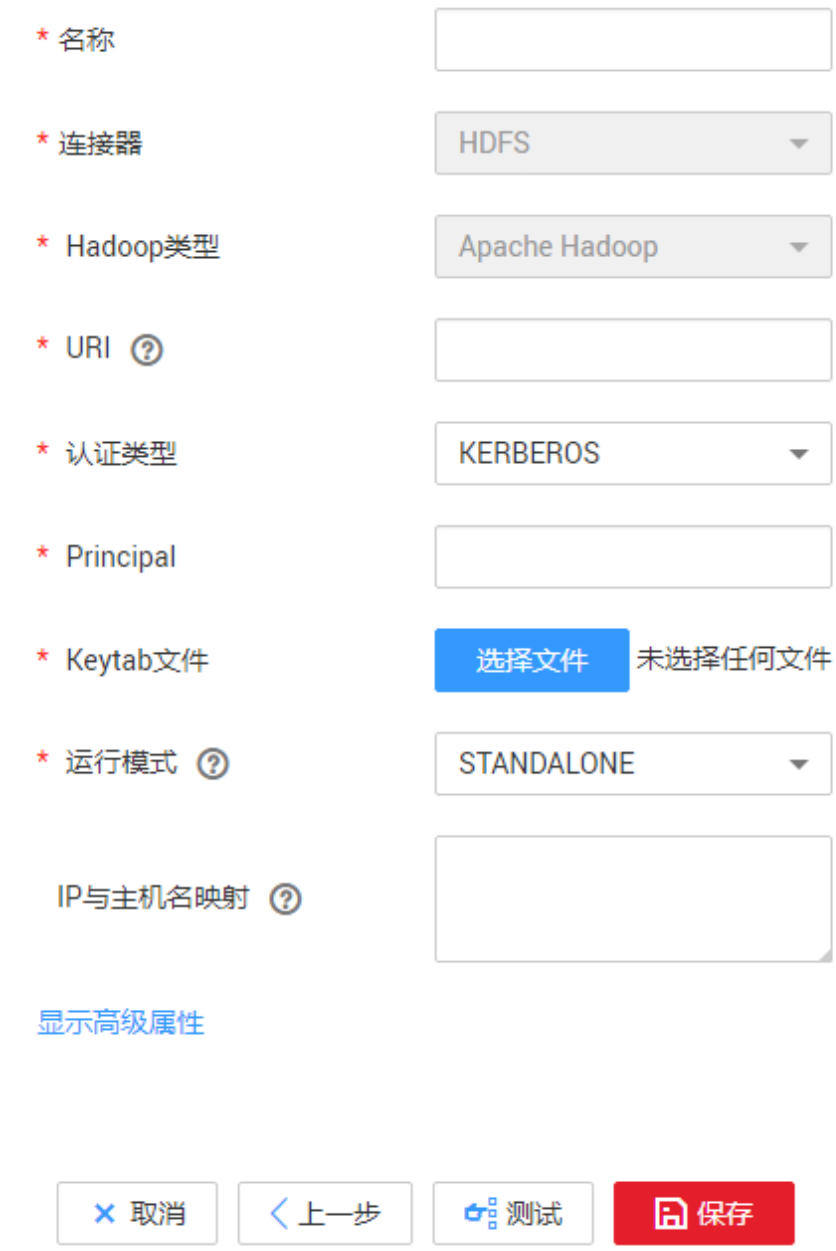
   如果您需要再创建新的CDM集群，请参考**购买DataArts Studio增量包**，完成购买数据集成增量包的操作。

   - 实例类型：选择“cdm.xlarge”，该实例类型适用大部分迁移场景。
   - 虚拟私有云：CDM集群的VPC，选择用户本地数据中心与云专线连通的VPC。
   - 子网、安全组：这里没有要求，任选一个即可。
4. 集群创建完成后，选择集群后面的“作业管理 > 连接管理 > 新建连接”，进入选择连接类型的界面，如**图11-2**所示。

图 11-2 选择连接器类型

5. 连接H公司本地的Apache Hadoop HDFS文件系统时，连接类型选择“Apache HDFS”，然后单击“下一步”。

图 11-3 创建 HDFS 连接



说明

- 名称：用户自定义连接名称，例如“hdfs_link”。
- URI：配置为H公司HDFS文件系统的Namenode URI地址。
- 认证类型：安全模式Hadoop选择KERBEROS鉴权，通过获取客户端的principal和keytab文件进行认证。
- Principal、Keytab文件：用于认证的账号Principal和keytab文件，可以联系Hadoop管理员获取。

6. 单击“保存”，CDM会自动测试连接是否可用。
   - 如果可用则提示保存成功，系统自动跳转到连接管理界面。

- 如果测试不可用，需要重新检查连接参数是否配置正确，或者H公司防火墙是否允许CDM集群的EIP访问数据源。

7. 单击“新建连接”来创建OBS连接，连接类型选择“对象存储服务（OBS）”后单击“下一步”，配置OBS连接参数，如图11-4所示。

图 11-4 创建 OBS 连接

说明

- 名称：用户自定义连接名称，例如“obslink”。
- OBS终端节点：配置为OBS的域名或IP地址，例如“obs.myhuaweicloud.com”。
- 端口：OBS服务器的端口，例如“443”。
- OBS桶类型：根据实际情况下拉选择即可。
- 访问标识（AK）、密钥（SK）：访问OBS数据库的AK、SK。可在管理控制台单击用户名，选择“我的凭证 > 访问密钥”后获取。

8. 单击“保存”，系统回到连接管理界面。
9. 选择“表/文件迁移 > 新建作业”，创建迁移H公司贸易数据到OBS的作业， 如图11-5所示。

图 11-5 创建作业

说明

- 作业名称：用户自定义作业名称。
- 源端作业配置：
  - 源连接名称：选择5创建的HDFS连接“hdfs_link”。
  - 源目录或文件：配置为H公司贸易数据在本地的存储路径，可以是一个目录，也可以是单独一个文件。这里配置为目录，CDM会迁移整个目录下的文件到OBS。
  - 文件格式：选择“二进制格式”。这里的文件格式是指CDM传输数据时所用的格式，不会改变原始文件自身的格式。迁移文件到文件时，推荐使用“二进制格式”，传输的效率和性能都最优。
- 目的端作业配置：
  - 目的连接名称：选择7创建的OBS连接“obslink”。
  - 桶名、写入目录：在OBS中储存贸易数据的路径，CDM会将文件写入到该路径下。
  - 文件格式：与源端一样，选择“二进制格式”，原始文件自身的格式不会改变。
  - 重复文件处理方式：这里选择“跳过重复文件”。只有当源端和目的端存在文件名、文件大小都相同的文件时，CDM才会判定该文件为重复文件，这时CDM将跳过该文件，不迁移到OBS。

10. 单击“下一步”配置任务参数，迁移存量数据时，参数配置保持默认即可。
11. 单击“保存并运行”，进入作业管理界面，查看作业执行进度和结果。
12. 作业执行成功之后，单击作业后面的“历史记录”查看作业的写入行数、读取行数、写入字节、写入文件数和执行日志。

11.3.2 上传增量数据

1. 使用CDM将存量数据上传完后，单击该作业后的“编辑”，直接修改该作业。
2. 保持作业基本参数不变，单击“下一步”修改任务参数，如图11-6所示。

图 11-6 定时任务配置

3. 勾选“是否定时执行”，配置定时任务：
   - “重复周期”配置为1天。
   - “开始时间”配置为每天凌晨0点1分执行。

   这样CDM每天凌晨自动执行全量迁移，但因为“重复文件处理方式”选择了“跳过重复文件”，相同名称且相同大小的文件不迁移，所以只会上传每天新增的文件。
4. 单击“保存”，完成CDM的增量同步配置。

11.4 分析数据

通过DLI直接对OBS数据进行贸易统计分析。

前提条件

DLI创建OBS外表，对OBS表数据存储格式有所要求：

- 使用DataSource语法创建OBS表时，支持orc，parquet，json，csv，carbon，avro类型。
- 使用Hive语法创建OBS表时，支持TEXTFILE, AVRO, ORC, SEQUENCEFILE, RCFILE, PARQUET, CARBON类型。

如果原始数据表存储格式不满足要求，您可以通过CDM将原始数据直接导入到DLI中进行分析，无需上传OBS。

通过 DLI 分析数据

1. 进入数据湖探索DLI控制台，参考DLI用户指南中的**创建数据库**创建数据库。
2. 参考**创建OBS表**创建OBS外表，包括贸易统计数据库、贸易详单信息表和基础信息表。
3. 基于业务需求，在DLI控制台中开发相应的SQL脚本进行贸易统计分析。

12 案例：车联网大数据业务上云

12.1 场景介绍

场景描述

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。

约束限制

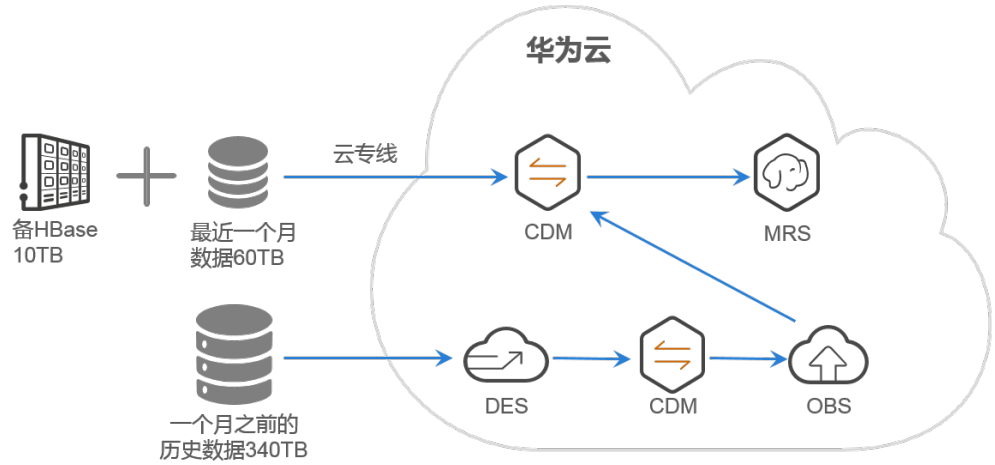
本方案仅支持迁移数据到MRS 1.x版本集群，2.x及之后版本无法通过HBase修复命令重建HBase表。

须知

迁移的目标集群为2.x及之后版本时，HBase修复命令已经不再支持，通过HBase数据目录迁移的方式无法使用。

迁移方案

图 12-1 迁移方案



H公司的车联网大数据业务平台当前CDH（Cloudera Hadoop） HBase集群中共有854张表约400TB，备HBase集群中共有149张表，约10TB数据。最近一个月新增的数据量是60TB。

使用CDM将CDH集群中的HBase HFile抽取出来存入到MRS（MapReduce） HDFS中，再通过HBase修复命令重建HBase表。基于这种迁移方案，可以使用以下2种迁移方式同时进行：

1. CDM通过专线直接迁移近一个月的数据以及备HBase集群的数据：
   CDH → CDM（华为云） → MRS

   说明

   使用云专线直接迁移时的优缺点：
   - 优点：数据无需做多次的搬迁，缩短整体搬迁周期。
   - 缺点：在数据大量传输过程中会占用专线带宽，对客户并行进行的业务存在影响，跨越多个交换机设备。
2. CDM通过**DES（数据快递服务）**迁移1个月前的历史数据，迁移路径如下：
   CDH → DES → CDM（华为云） → OBS → CDM（华为云） → MRS

   说明

   DES适用场景：数据量大，用户私有云与华为云无专线打通，用户私有云网络到公网带宽有限。
   - 优点：传输可靠性高，受专线以及网络质量影响较小。
   - 缺点：迁移方式耗时较长。

12.2 迁移准备

前提条件

- CDH HBase的版本号小于或等于MRS HBase的版本号。

- 待迁移的表在迁移过程中不能有写入，Split，Merge等操作。
- 使用华为云专线搭建CDH集群与华为云VPC之间的专属连接通道。

迁移流程

1. 预估迁移数据量、迁移时间。
2. 输出详细待迁移数据表、文件个数、大小，用于后续校验。
3. 分批配置迁移任务，保证迁移进度与速度。
4. 校验文件个数以及文件大小。
5. 在MRS中恢复HBase表并验证。

准备数据

项目	数据项	说明	取值示例
DES盒子	挂载地址	DES盒子在客户的虚拟机挂载的地址。	//*虚拟机IP*/huawei
	存储管理系统	DES盒子的存储管理系统，与管理IP相关。	https://*管理IP*:8088/deviceManager/devicemanager/login/login.html
	用户名	登录存储管理系统的用户名。	admin
	密码	登录密码。	-
CDH集群	NameNode IP	客户CDH集群的主NameNode IP。	192.168.2.3
	HDFS的端口	一般默认为9000。	9000
	HDFS URI	客户CDH集群中HDFS的NameNode URI地址。	hdfs://192.168.2.3:9000
OBS	OBS终端节点	OBS的Endpoint。	obs.ap-southeast-1.myhuaweicloud.com
	OBS桶	存放CDH一个月前历史数据的OBS桶。	cdm
	AK/SK	连接OBS的AK/SK。	-
MRS	Manager IP	MRS Manager的IP地址。	192.168.3.11

12.3 CDM 迁移近一个月的数据

备HBase集群中约10TB数据，最近一个月新增的数据量约60TB，总共约70TB。H公司安装的云专线为20GE端口，支持CDM超大规格的集群（cdm.xlarge），综合考虑迁移

时间、成本、性能等，这里使用2个CDM超大规格集群并行迁移。CDM集群规格如**表12-1**所示。

表 12-1 CDM 集群规格

实例类型	核数/内存	最大带宽/基准带宽	并发作业数	适用场景
cdm.large	8核/16G	3/0.8 Gbps	16	单表规模≥1000万条。
cdm.xlarge	16核/32G	10/4 Gbps	32	适合10GE高速带宽进行TB以上的数据量迁移。
cdm.4xlarge	64核/128G	40/36 Gbit/s	64	-

说明

其他场景中，可根据情况选择多个CDM集群同时迁移，加快迁移效率。MRS HDFS多副本策略会占用网络带宽，影响迁移速率。

华为云 CDM 创建连接

1. 创建2个CDM集群：

   说明

   DataArts Studio实例中已经包含一个CDM集群（试用版除外），如果该集群已经满足需求，您无需再购买数据集成增量包，可以跳过这部分内容。

   如果您需要再创建新的CDM集群，请参考**购买DataArts Studio增量包**，完成购买数据集成增量包的操作。

   - 集群规格选择“cdm.xlarge”。
   - 集群所属的VPC与MRS所属的VPC一致，同时也要与云专线连通的VPC的一致。
   - 其它参数可以自定义，或者保持默认。

2. 创建CDH HDFS连接：
   a. 单击CDM集群操作列的“作业管理”，进入作业管理界面。
   b. 选择“连接管理 > 新建连接”，进入连接器类型的选择界面，选择“Apache HDFS”。

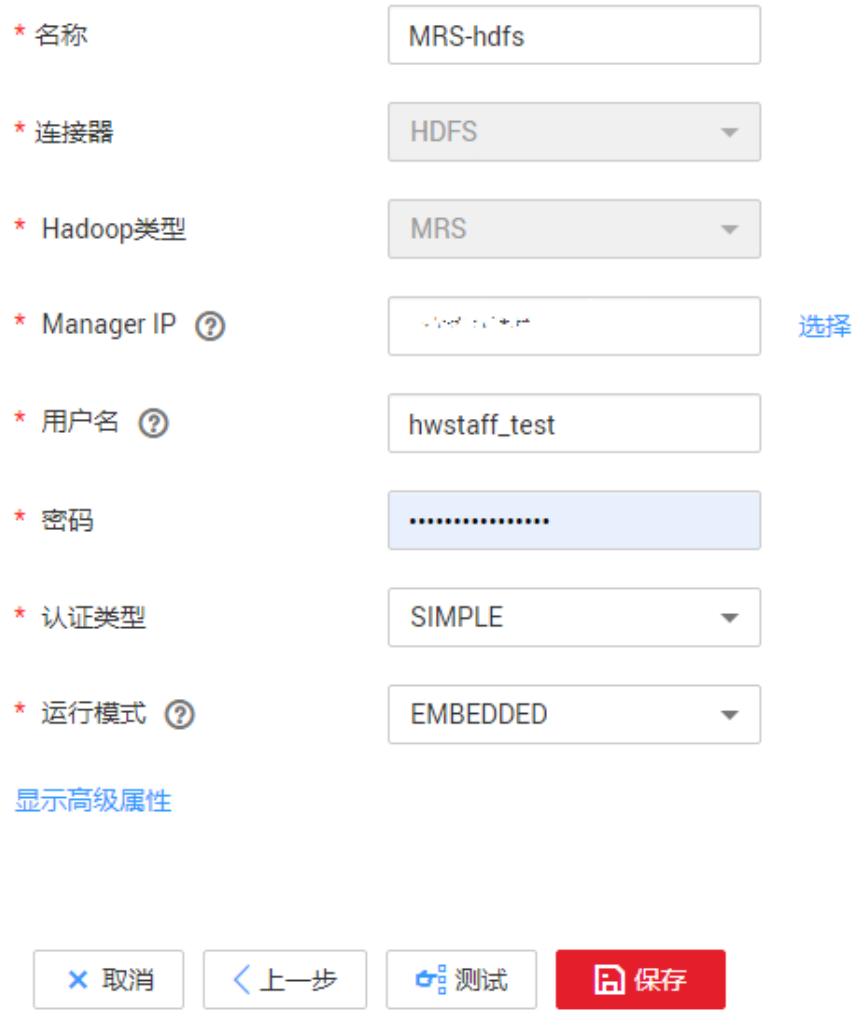
c. 单击“下一步”，配置连接参数，依次填写相关信息。URI格式为“hdfs://NameNode IP:端口”，若CDH没有启动Kerberos认证则“认证类型”选择“SIMPLE”。

d. 单击“测试”，如果右上角显示“测试成功”，表示连接成功，单击“保存”。

3. 创建MRS HDFS连接：

a. 在作业管理界面，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，选择“MRS HDFS”。

b. 单击“下一步”，配置连接参数，依次填写相关信息。“认证类型”选择“SIMPLE”，运行模式保持默认即可。



c. 单击“测试”，如果右上角显示“测试成功”，表示连接成功，单击“保存”。

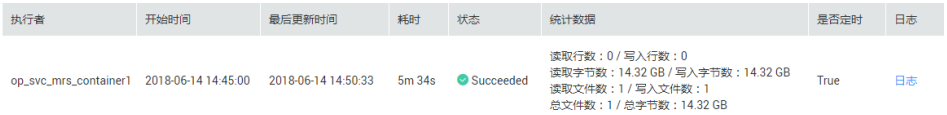
华为云 CDM 创建迁移作业

1. 在CDM集群的作业管理界面，选择“表/文件迁移 > 新建作业”，每个表文件的目录作为一个迁移作业。

- 源端作业配置
  - 源连接名称：选择上面创建的**CDH HDFS连接**。
  - 源目录或文件：选择CDH中HBase的HBase表所在目录。例如“/hbase/data/default/table_20180815”，表示迁移“table_20180815”这个目录下所有文件。
  - 文件格式：文件的复制要选择“二进制格式”。
- 目的端作业配置
  - 目的连接名称：选择上面创建的**MRS HDFS连接**。
  - 写入目录：选择MRS HBase的目录，例如“/hbase/data/default/table_20180815/”。这个目录必须带有表名（例如这里的表名是table_20180815），如果该目录不存在，CDM会自动创建该目录。
  - 文件格式：同源端相同，选择“二进制格式”。
- 其它可选参数保持默认即可。

2. 单击“下一步”进行任务配置，其中抽取并发数默认为3，适当增加可以增加迁移速率，本例中设置为8，其它参数保持默认即可。

3. 重复上述步骤创建其它迁移目录的作业，参数配置都相同。2个CDM集群的作业个数平均分配，并发执行。
4. 作业执行完成后，可在作业的“历史记录”中查看详细的数据统计。



执行者	开始时间	最后更新时间	耗时	状态	统计数据	是否定时	日志
op_svc_mrs_container1	2018-06-14 14:45:00	2018-06-14 14:50:33	5m 34s	Succeeded	读取行数：0 / 写入行数：0 读取字节数：14.32 GB / 写入字节数：14.32 GB 读取文件数：1 / 写入文件数：1 总文件数：1 / 总字节数：14.32 GB	True	日志

12.4 DES 迁移一个月前的历史数据

迁移流程

1. 通过脚本将一个月前的历史数据导入到DES盒子。DES盒子的相关操作请参见**数据快递服务 DES**。
2. DES将数据快递到华为云数据中心。
3. 使用华为云CDM将DES中的数据迁移到华为云OBS。
4. 使用华为云CDM将OBS数据迁移到MRS。

其中CDM相关操作，与**CDM迁移近一个月的数据**相同，都是使用二进制直接传输文件目录，2个集群并发执行作业。

注意事项

- 当迁移动作影响到客户的HDFS集群时，需要手动停止作业。

- 如果作业出现大批量的失败：
  a. 先检查DES盒子是否被写满。如果写满，需要清除最近写入的目录，保证后面写入的数据都是完整的。
  b. 再检查网络是否连通。
  c. 检查客户的HDFS集群。检查是否有指标异常的现象，如果有，则需要暂停迁移任务。

12.5 MRS 中恢复 HBase 表

CDH HBase表目录已经迁移到MRS HBase后，可以使用命令恢复。对于那些会变化的数据，需要使用快照保证数据不变，然后再迁移并恢复。

约束限制

本方案仅支持迁移数据到MRS 1.x版本集群，2.x及之后版本无法通过HBase修复命令重建HBase表。

须知

迁移的目标集群为2.x及之后版本时，HBase修复命令已经不再支持，通过HBase数据目录迁移的方式无法使用。

使用命令恢复历史不变的数据

这里以恢复“/hbase/data/default/table_20180811”表为例，恢复步骤如下：

1. 进入MRS Client所在的节点，例如master1节点。
2. 切换为omm用户。
   su – omm
3. 加载环境变量。
   source /opt/client/bigdata_env
4. 执行修改目录权限命令。
   hdfs dfs –chown *omm:hadoop* **–R** */hbase/data/default/table_20180811*
   - omm:hadoop：表示用户名，实际场景中请替换。
   - /hbase/data/default/table_20180811：表示表所在路径。
5. 执行恢复元数据命令。
   hbase hbck –fixMeta *table_20180811*
6. 执行Region上线命令。
   hbase hbck –fixAssignments *table_20180811*
7. 出现“Status: OK”则说明恢复表成功。

使用快照迁移并恢复会变的数据

1. 在源端CDH集群HBase shell中执行：
   flush <*table name*>

2. 在源端CDH集群HBase shell执行：

   compact *<table name>*

3. 如果表没有打开Snap功能，则执行：

   hadoop dfsadmin -allowSnapshot $path

4. 创建HDFS Snapshot，例如命名s0：

   hdfs dfs -createSnapshot *<snapshotDir>* [*s0*]

   hdfs dfs -createSnapshot test

5. CDM通过HDFS Snapshot复制文件至MRS。CDM的作业配置：
   - “源目录或文件”输入：/hbase/data/default/src_test/.snapshot/s0
   - 目的端“写入目录”输入：/hbase/data/default/*表名*

6. 执行fixMeta和fixAssignments等命令恢复表，参考**使用命令恢复历史不变的数据**。

7. 在CDH集群中删除快照：

   hdfs dfs -deleteSnapshot *<snapshotDir> s0*

恢复表时的问题处理

1. 执行完fixMeta命令后，报错显示“xx inconsistent”：

   fixMeta命令是校验HDFS和HBase元数据一致性，出现这个提示是正常情况，继续执行fixAssignments命令即可。

2. 执行完fixAssignments命令后，报错显示“xx inconsistent”：

   fixAssignments是让所有Region上线，偶尔会出现部分Region上线较慢，可以再执行一次以下命令检查一下：

   hbase hbck *表名*

   如果出现“Status : OK”则表示HBase表恢复成功。

3. 执行完fixAssignments命令后，错误提示多个region有相同的startkey，部分region存在overlap情况：

   可以再执行：

   hbase hbck –fixHdfsOverlaps *表名*

   执行完毕后再执行fixMeta和fixAssignments命令。

13 案例：搭建实时报警平台

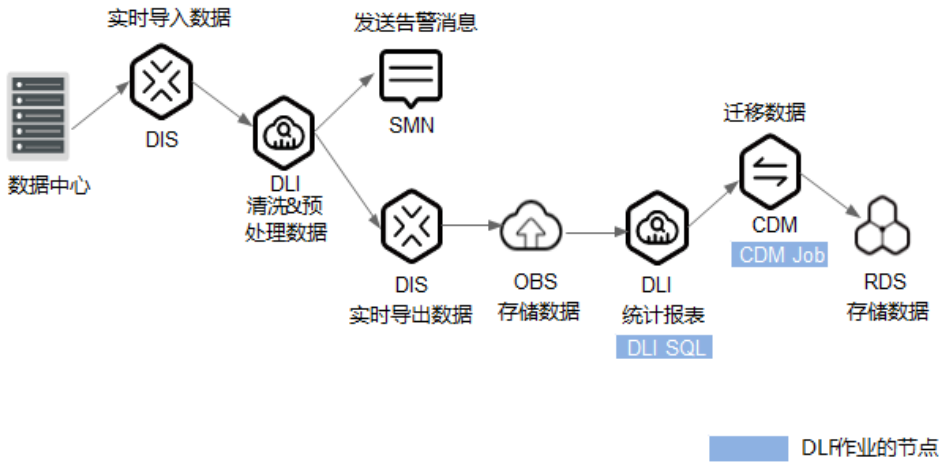
在本实践用户可以了解到如何搭建一个简单的实时报警平台，该平台将应用多个云服务，结合数据开发模块的作业编辑和作业调度功能来实现。

假设客户有一个数据中心部署了很多应用，需要建立统一的运维系统，实时接收应用的告警信息。

- 当告警级别达到严重及以上级别时，向用户发送一条消息。
- 每天提供一个运维报表，统计各应用的告警级别数据。

为解决以上场景的需求，我们设计了如下方案：

图 13-1 方案设计



操作流程如下：

1. 实时数据导入：通过数据接入服务（DIS）将数据中心的告警数据实时导入到数据湖探索（DLI）。
2. 数据清洗和预处理：DLI对告警数据进行数据清洗和预处理。
3. 发送告警消息：当告警级别超过指定值时向用户发送短信。
4. 数据导出和存储：清洗过的数据进入DIS通道，DIS根据导入时间将告警数据按日期存放到OBS。
5. 输出告警统计报表：通过DLI SQL脚本建立外部分区数据表，以及按照告警分区时间和告警类别进行统计。

6. 迁移数据：告警统计表计算完成后，将数据通过云数据迁移服务（CDM）统一导出到RDS MySQL数据库。

环境准备

- 已开通对象存储服务（OBS），并创建桶，例如“obs://dlfexample/alarm_info”、“obs://dlfexample/alarm_count_info”，分别用于存放原始告警表和告警统计报表。
- 已开通数据治理中心DataArts Studio，并具备CDM集群“cdm-alarm”，用于**创建CDM作业**。
- 已开通数据湖探索服务（DLI）。
- 已开通消息通知服务（SMN）。

数据准备

原始告警表为数据中心的实时数据，包含告警ID、告警级别。示例数据如**表13-1**所示。

表 13-1 原始数据示例

alarm_id	alarm_type
00440114	3
00440121	5
00440122	6
00440123	7
00440124	8
00440126	0

创建 DIS 通道

我们需要在DIS服务控制台创建两个DIS通道，分别用于实时数据导入到DLI、实时数据导出到OBS。

步骤1 创建实时数据导入到DLI的通道，通道名称为“dis-alarm-input”。

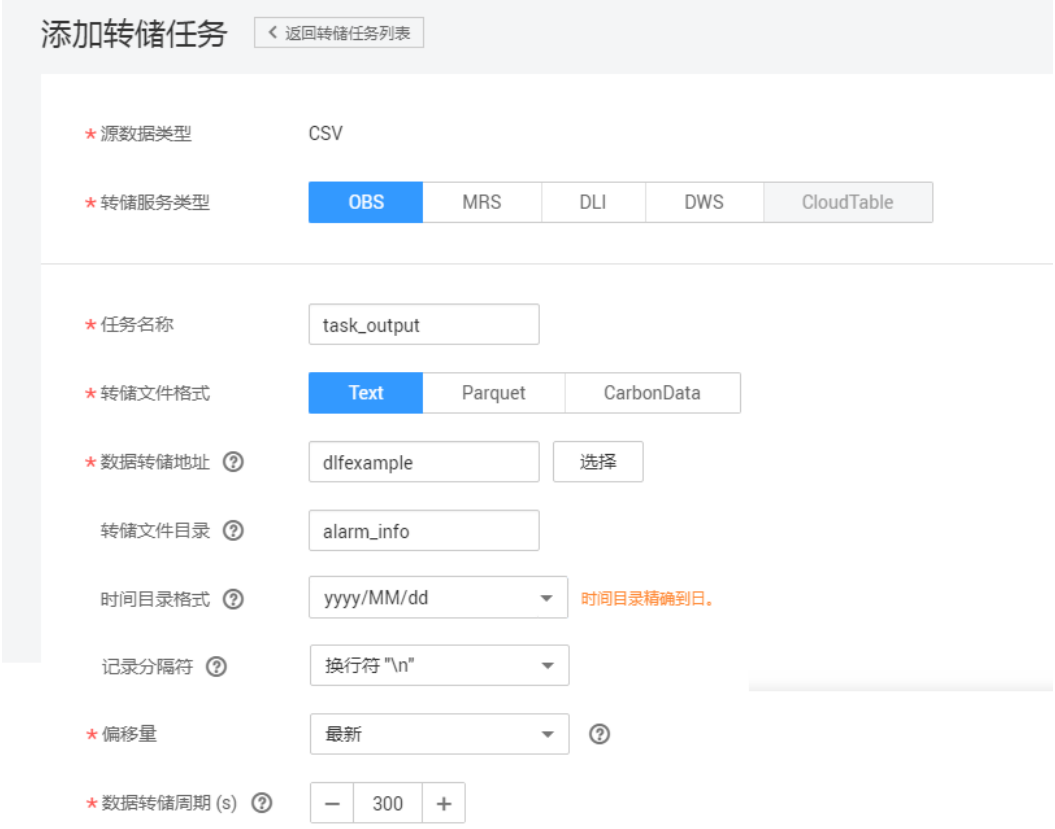
图 13-2 创建 input 通道

步骤2 创建实时数据导出到OBS的通道，通道名称为“dis-alarm-output”。

图 13-3 创建 output 通道

为dis-alarm-output通道配置转储任务，将通道中的数据按照导出时间转储到OBS的“obs://dlfexample/alarm_info”目录下。

图 13-4 output 通道配置转储任务



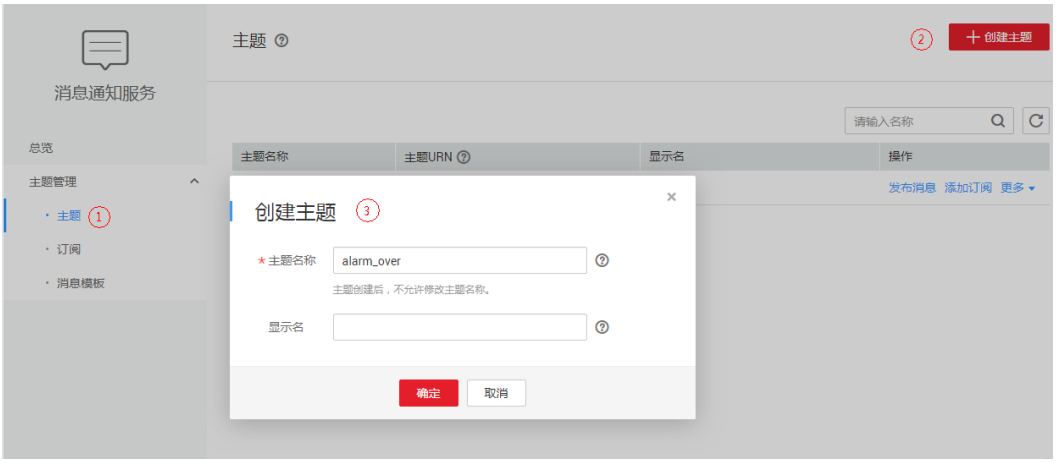
----结束

创建 SMN 主题

我们需要创建一个SMN主题并添加订阅，将需要收到告警通知的用户添加到订阅终端中。

步骤1 创建一个SMN主题，主题名称为“alarm_over”。

图 13-5 创建 SMN 主题



步骤2 为**步骤1**中的主题添加订阅，指定告警消息类型和需要接收告警通知的用户。

图 13-6 添加订阅

关键参数说明：

- 协议：选择“短信”，当告警级别达到指定值时向用户发送短信通知。
- 订阅终端：填写需要接收告警通知的用户手机号码。

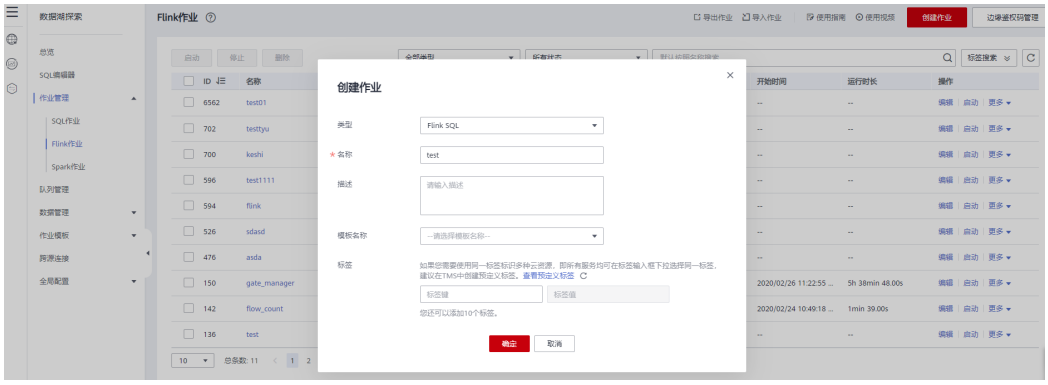
----结束

使用 DLI 作业管理构建告警通知工程

DIS通道（**创建DIS通道**）和SMN主题（**创建SMN主题**）创建完成后，我们就可以在DLI中构建告警通知工程。

步骤1 在DLI中创建一个Flink作业，作业名称为“test”。

图 13-7 创建 Flink SQL 作业



步骤2 编辑**步骤1**中创建的Flink SQL作业，在SQL编辑器中输入语句。

图 13-8 编辑 Flink SQL 作业

SQL语句实现的功能：

1. DIS通过工具上传实时数据至DLI，使用**步骤1**中创建的“dis-alarm-input”通道。
2. 判断告警级别，当告警级别达到指定值时向用户发送短信通知。
3. DLI处理过的数据再通过DIS导出到OBS中，使用**步骤2**中创建的“dis-alarm-output”通道。

```
CREATE SOURCE STREAM alarm_info (
  alarm_id STRING,
  alarm_type INT
)
WITH (
  type = "dis",
  region = "cn-south-1",
  channel = "dis-alarm-input",
  partition_count = "1",
  encode = "csv",
  field_delimiter = ","
);
CREATE SINK STREAM over_alarm (
  alarm_over STRING /* over speed message */
)
WITH (
  type = "smn",
  region = "cn-south-1",
  topic_urn = "urn:smn:cn-south-1:6f2bf33af5104f45ab85de31d7841f5a:alarm_over",
  message_subject = "alarm",
  message_column = "alarm_over"
);
INSERT INTO over_alarm
SELECT "your alarm over (" || CAST(alarm_type as CHAR(20)) || ") ."
FROM alarm_info
WHERE alarm_type > 8;
CREATE SINK STREAM alarm_info_output (
  alarm_id STRING,
  alarm_type INT
)WITH (
  type ="dis",
  region = "cn-south-1",
  channel = "dis-alarm-output",
  PARTITION_KEY = "alarm_type",
  encode = "csv",
  field_delimiter = ","
```



```
);
INSERT INTO alarm_info_output
SELECT *
FROM alarm_info
WHERE alarm_type > 0;
```

步骤3 Flink SQL作业开发完成后，保存并启动作业。

----结束

使用 DLI SQL 脚本开发构建告警报表脚本

我们需要通过SQL脚本在DLI中新建OBS表来存放数据表，然后再构建一个SQL脚本来统计告警信息。

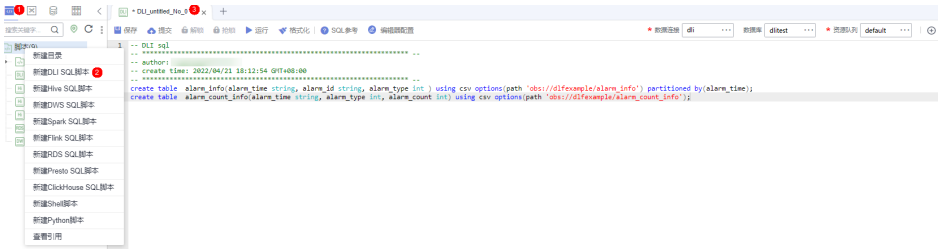
步骤1 在DataArts Studio管理中心模块创建一个到DLI的连接，数据连接名称为“dli”。

步骤2 进入数据开发模块，在DLI中创建一个数据库，用于存放数据表，数据库名称为“dlitest”。

步骤3 创建一个DLI SQL脚本，通过SQL语句来创建数据表alarm_info，alarm_count_info。

其中，alarm_info、alarm_count_info都为OBS表，数据存储在OBS中，分别用于存放原始告警表、告警统计报表。

图 13-9 创建数据表



关键操作说明：

- 图13-9中的脚本开发区为临时调试区，关闭脚本页签后，开发区的内容将丢失。如需保留该SQL脚本，请单击，将脚本保存至指定的目录中。

关键参数说明：

- 数据连接：步骤1中创建的DLI数据连接。
- 数据库：步骤2中创建的数据库。
- 资源队列：使用DLI提供的默认资源队列“default”。
- SQL语句：如下所示。

```
create table  alarm_info(alarm_time string, alarm_id string, alarm_type int ) using csv options(path 'obs://dlfexample/alarm_info') partitioned by(alarm_time);
create table  alarm_count_info(alarm_time string, alarm_type int, alarm_count int) using csv options(path 'obs://dlfexample/alarm_count_info');
```

步骤4 单击运行脚本，创建alarm_info、alarm_count_info数据表。

步骤5 清空编辑器中步骤4的SQL语句，重新输入SQL语句。

```
ALTER TABLE alarm_info  ADD PARTITION (alarm_time = ${dayParam})
LOCATION 'obs://dlfexample/alarm_info/${obsPathYear}';
insert into alarm_count_info
```

```
select alarm_time,alarm_type,count(alarm_type) from alarm_info where alarm_time = ${dayParam} group by alarm_time,alarm_type;
```

SQL语句实现的功能：

1. 在OBS的“obs://dlfexample/alarm_info”目录下，根据日期新建DLI分区。假设当前日期为2018/10/10，那么在“obs://dlfexample/alarm_info”目录下新建“2018/10/09”的DLI分区，用于存放前一天的数据表。
2. 按照告警分区时间和告警类别进行统计，将统计结果插入alarm_count_info数据表。

关键参数说明：

- ${dayParam}：dayParam是指alarm_info表分区值，在脚本编辑器下方输入具体的参数值“$getCurrentTime(@@yyyyMMdd@@,-24*60*60)”。
- ${obsPathYear}：obsPathYear是指OBS分区目录路径，在脚本编辑器下方输入具体的参数值“$getCurrentTime(@@yyyy/MM/dd@@,-24*60*60)”。

步骤6 脚本调试无误后，我们需要保存该脚本，脚本名称为“dli_partition_count”。在后续的作业中设置为定期执行该脚本（**使用DLF作业开发和作业调度每天定时输出告警统计报表**），实现定期输出告警统计报表。

----结束

创建 CDM 作业

方案的最后一步需要将OBS中的告警统计报表迁移到RDS MySQL中，我们选择使用CDM来实现该功能。

关键参数说明：

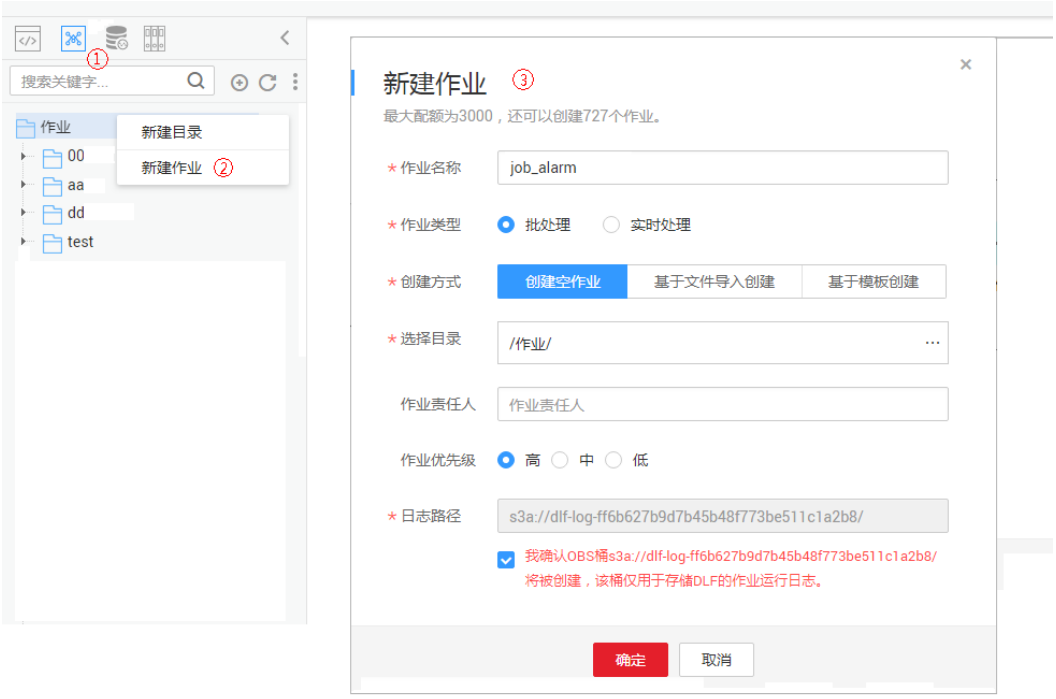
- 作业名称：obs_rds，在后续的作业中设置为定期执行该作业（**使用DLF作业开发和作业调度每天定时输出告警统计报表**），实现定期迁移数据。
- 源端：存储告警统计报表的OBS目录，源连接“obs_link”需要提前在CDM中创建好。
- 目的端：即将存储告警统计报表的RDS MySQL空间，目的连接“mysql_link”需要提前在CDM中创建好。

使用 DLF 作业开发和作业调度每天定时输出告警统计报表

告警统计报表的脚本（**使用DLI SQL脚本开发构建告警报表脚本**）和数据迁移的CDM作业（**创建CDM作业**）创建完成后，我们在数据开发模块中构建一个作业每天自动执行，那么就可以每天输出告警统计报表、每天自动迁移数据。

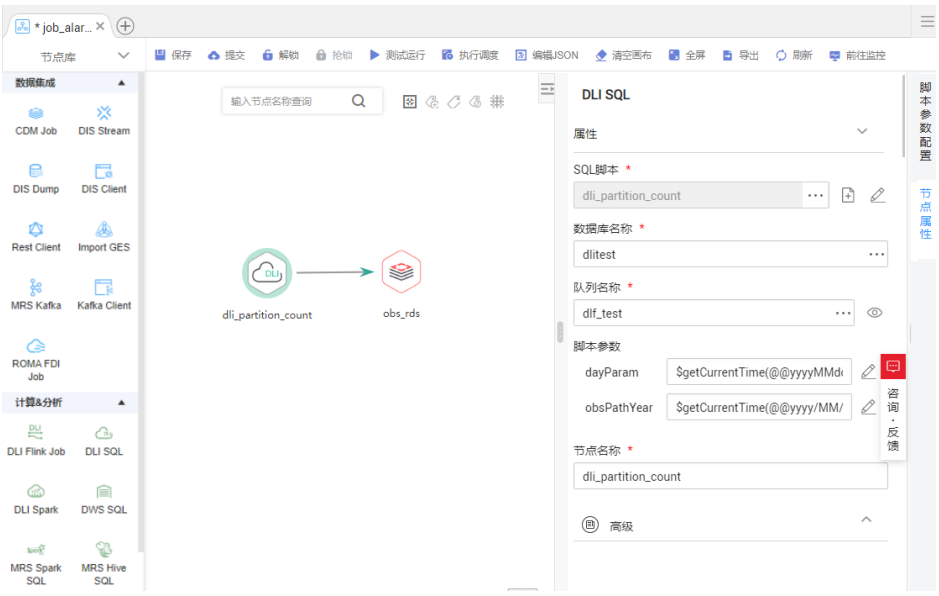
步骤1 创建一个批处理作业，作业名称为“job_alarm”。

图 13-10 创建 DLF 作业



步骤2 然后进入到作业开发页面，拖动DLI SQL和CDM Job节点到画布中，连接并配置节点的属性。

图 13-11 连接和配置节点属性



关键说明：

- dli_partition_count（DLI SQL节点）：在节点属性中，关联**使用DLI SQL脚本开发构建告警报表脚本**中开发完成的DLI SQL脚本“dli_partition_count”。
- obs_rds（CDM Job节点）：在节点属性中，关联**创建CDM作业**中创建的CDM作业“obs_rds”。

步骤3 作业编排完成后，单击 ，测试运行作业。

步骤4 如果日志运行正常，单击右侧的“调度配置”，配置作业的调度策略。

图 13-12 调度配置

说明：

- 2018/10/10至2018/11/09，每天2点执行一次作业。

步骤5 最后我们需要保存作业并提交版本，执行调度作业（单击 ），实现作业每天自动运行。

----结束

