云数据迁移

用户指南

文档版本 01

发布日期 2023-06-14

华为云计算技术有限公司

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云计算技术有限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，华为云计算技术有限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。

目 录

1 IAM 权限管理

1.1 创建 IAM 用户并授权使用 CDM

如果您需要对您所拥有的数据集成服务（CDM）进行精细的权限管理，您可以使用**统一身份认证服务**（Identity and Access Management，简称IAM），通过IAM，您可以：

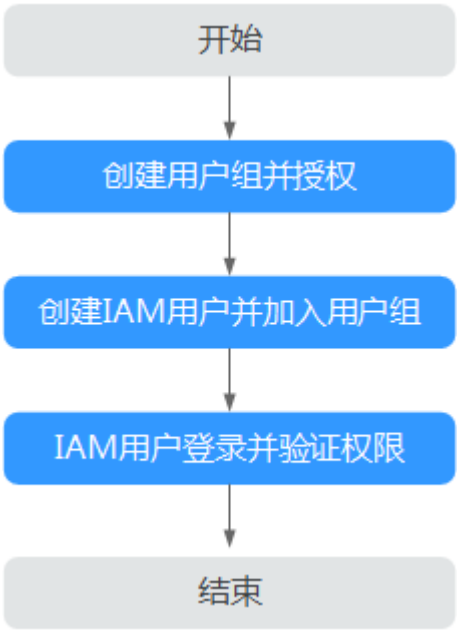
- 根据企业的业务组织，在您的华为云账号中，给企业中不同职能部门的员工创建IAM用户，让员工拥有唯一安全凭证，并使用CDM资源。
- 根据企业用户的职能，设置不同的访问权限，以达到用户之间的权限隔离。
- 将CDM资源委托给更专业、高效的其他华为云账号或者云服务，这些账号或者云服务可以根据权限进行代运维。

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

本章节为您介绍对用户授权的方法，操作流程如**图1-1**所示。

示例流程

图 1-1 IAM 用户授权流程



1. **创建用户组并授权**

   在IAM控制台创建用户组，并授予CDM集群只读权限“CDM ReadOnlyAccess”。

2. **创建用户并加入用户组**

   在IAM控制台创建用户，并将其加入**1**中创建的用户组。

3. **用户登录**并验证权限

   新创建的用户登录控制台，切换至授权区域，验证权限：

   - 在“服务列表”中选择“云数据迁移服务”，进入CDM主界面查看集群，若未提示权限不足，表示“CDM ReadOnlyAccess”已生效。
   - 在“服务列表”中选择除CDM服务外的任一服务，若提示权限不足，表示“CDM ReadOnlyAccess”已生效。

1.2 创建 CDM 自定义策略

如果系统预置的CDM权限策略，不满足您的授权要求，可以创建自定义策略。自定义策略中可以添加的授权项（Action）请参考**策略和授权项**。

目前华为云支持以下两种方式创建自定义策略：

- 可视化视图创建自定义策略：无需了解策略语法，按可视化视图导航栏选择云服务、操作、资源、条件等策略内容，可自动生成策略。
- JSON视图创建自定义策略：可以在选择策略模板后，根据具体需求编辑策略内容；也可以直接在编辑框内编写JSON格式的策略内容。

具体创建步骤请参见：**创建自定义策略**。本章为您介绍常用的CDM自定义策略样例。

CDM 自定义策略样例

- 示例1：授权用户创建CDM集群

```
{
   "Version": "1.1",
   "Statement": [
      {
         "Effect": "Allow",
         "Action": [
            "cdm:cluster:create"
         ]
      }
   ]
}
```

- 示例2：拒绝用户删除CDM集群

  拒绝策略需要同时配合其他策略使用，否则没有实际作用。用户被授予的策略中，一个授权项的作用如果同时存在Allow和Deny，则遵循**Deny优先原则**。

  如果您给用户授予CDM FullAccess的系统策略，但不希望用户拥有CDM FullAccess中定义的删除CDM集群权限，您可以创建一条拒绝删除CDM集群的自定义策略，然后同时将CDM FullAccess和拒绝策略授予用户，根据Deny优先原则，则用户可以对CDM执行除了删除CDM集群外的所有操作。拒绝策略示例如下：

```
{
   "Version": "1.1",
   "Statement": [
      {
         "Effect": "Deny",
         "Action": [
            "cdm:cluster:delete"
         ]
      }
   ]
}
```

- 示例3：多个授权项策略

  一个自定义策略中可以包含多个授权项，且除了可以包含本服务的授权项外，还可以包含其他服务的授权项，可以包含的其他服务必须跟本服务同属性，即都是项目级服务或都是全局级服务。多个授权语句策略描述如下：

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Action": [
                "cdm:cluster:list",
                "cdm:cluster:get",
                "ecs:*:get*",
                "ecs:*:list*",
                "vpc:*:get*",
                "vpc:*:list*",
                "evs:*:get*",
                "evs:*:list*",
                "bss:*:view*"
            ],
            "Effect": "Allow"
        }
    ]
}
```

2 支持的数据源

2.1 支持的数据源（2.9.3.300）

数据集成有两种迁移方式，支持的数据源有所不同：

- 表/文件迁移：适用于数据入湖和数据上云场景下，表或文件级别的数据迁移，请参见**表/文件迁移支持的数据源类型**。
- 整库迁移：适用于数据入湖和数据上云场景下，离线或自建数据库整体迁移场景，请参见**整库迁移支持的数据源类型**。

说明

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

表/文件迁移支持的数据源类型

表/文件迁移可以实现表或文件级别的数据迁移。

表/文件迁移时支持的数据源如**表2-1**所示。

表 2-1 表/文件迁移支持的数据源

数据源分类	源端数据源	对应的目的端数据源	说明
数据仓库	数据仓库服务（DWS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS ClickHouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	不支持DWS物理机纳管模式。
	数据湖探索（DLI）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS ClickHouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable），MongoDB • 搜索：Elasticsearch，云搜索服务（CSS）	MongoDB建议使用的版本：4.2。
	MRS ClickHouse	数据仓库：MRS ClickHouse，数据湖探索（DLI）	MRS ClickHouse建议使用的版本：21.3.4.X。

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

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

数据源分类	源端数据源	对应的目的端数据源	说明
	Apache HBase Apache Hive Apache HDFS	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• Apache数据源不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。 • Apache HBase建议使用的版本： – 2.1.X – 1.3.X • Apache Hive暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X • Apache HDFS建议使用的版本： – 2.8.X – 3.1.X
对象存储	对象存储服务（OBS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 对象存储服务之间的迁移，推荐使用对象存储迁移服务OMS。 • 不支持二进制文件导入到数据库或NoSQL。

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

数据源分类	源端数据源	对应的目的端数据源	说明
	MySQL PostgreSQL Oracle	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	Microsoft SQL Server	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	

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

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

数据源分类	源端数据源	对应的目的端数据源	说明
	MRS Kafka	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• MRS Kafka不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。
搜索	Elasticsearch	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	Elasticsearch仅支持非安全模式。
	云搜索服务（CSS）		导入数据到CSS推荐使用Logstash，请参见**使用Logstash导入数据到Elasticsearch**。

说明

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

整库迁移支持的数据源类型

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

数据集成支持整库迁移的数据源如**表2-2**所示。

表 2-2 整库迁移支持的数据源

数据源分类	数据源	读取	写入	说明
数据仓库	数据仓库服务（DWS）	支持	支持	-
Hadoop （仅支持本地存储，不支持存算分离场景，不支持Ranger场景，不支持ZK开启SSL场景）	MRS HBase	支持	支持	整库迁移仅支持导出到MRS HBase。 建议使用的版本： ● 2.1.X ● 1.3.X
	MRS Hive	支持	支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	FusionInsight HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	FusionInsight Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	Apache HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	Apache Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X

数据源分类	数据源	读取	写入	说明
	MRS Hudi	支持	支持	支持本地存储、存算分离场景。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
关系数据库	云数据库 MySQL	支持	支持	不支持OLTP到OLTP迁移，此场景推荐通过数据复制服务DRS进行迁移。
	云数据库 PostgreSQL	支持	支持	
	云数据库 SQL Server	支持	支持	
	MySQL	支持	不支持	
	PostgreSQL	支持	不支持	
	Microsoft SQL Server	支持	不支持	
	Oracle	支持	不支持	
	SAP HANA	支持	不支持	● 仅支持2.00.050.00.1592305219版本。 ● 仅支持Generic Edition。 ● 不支持BW/4 FOR HANA。 ● 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 ● 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 ● 迁移时不支持目的端自动建表。

数据源分类	数据源	读取	写入	说明
	达梦数据库 DM	支持	不支持	仅支持导出到DWS、Hive
NoSQL	分布式缓存服务（DCS）	不支持	支持	仅支持MRS到DCS迁移。
	文档数据库服务（DDS）	支持	支持	仅支持DDS和MRS之间迁移。
	表格存储服务（CloudTable）	支持	支持	-

2.2 支持的数据源（2.9.2.200）

数据集成有两种迁移方式，支持的数据源有所不同：

- 表/文件迁移：适用于数据入湖和数据上云场景下，表或文件级别的数据迁移，请参见**表/文件迁移支持的数据源类型**。
- 整库迁移：适用于数据入湖和数据上云场景下，离线或自建数据库整体迁移场景，请参见**整库迁移支持的数据源类型**。

说明

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

表/文件迁移支持的数据源类型

表/文件迁移可以实现表或文件级别的数据迁移。

表/文件迁移时支持的数据源如**表2-3**所示。

表 2-3 表/文件迁移支持的数据源

数据源分类	源端数据源	对应的目的端数据源	说明
数据仓库	数据仓库服务（DWS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS ClickHouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	不支持DWS物理机纳管模式。
	数据湖探索（DLI）		-
	MRS ClickHouse	数据仓库：MRS ClickHouse，数据湖探索（DLI）	MRS ClickHouse建议使用的版本：21.3.4.X。
Hadoop	MRS HDFS	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 支持本地存储，仅MRS Hive、MRS Hudi支持存算分离场景。 • 仅MRS Hive支持Ranger场景。 • 不支持ZK开启SSL场景。 • MRS HDFS建议使用的版本： – 2.8.X – 3.1.X • MRS HBase建议使用的版本： – 2.1.X – 1.3.X • MRS Hive、MRS Hudi暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X
	MRS HBase		

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

数据源分类	源端数据源	对应的目的端数据源	说明
	Apache HBase Apache Hive Apache HDFS	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• Apache数据源不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。 • Apache HBase建议使用的版本： – 2.1.X – 1.3.X • Apache Hive暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X • Apache HDFS建议使用的版本： – 2.8.X – 3.1.X
对象存储	对象存储服务（OBS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 对象存储服务之间的迁移，推荐使用对象存储迁移服务OMS。 • 不支持二进制文件导入到数据库或NoSQL。

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

数据源分类	源端数据源	对应的目的端数据源	说明
	MySQL PostgreSQL Oracle	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	Microsoft SQL Server	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	

数据源分类	源端数据源	对应的目的端数据源	说明
	SAP HANA	• 数据仓库：数据湖探索（DLI） • Hadoop：MRS Hive	SAP HANA数据源存在如下约束： • SAP HANA不支持作为目的端。 • 仅支持2.00.050.00.1592305219版本。 • 仅支持Generic Edition。 • 不支持BW/4 FOR HANA。 • 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 • 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 • 迁移时不支持目的端自动建表。
	分库	• 数据仓库：数据湖探索（DLI） • Hadoop：MRS HBase，MRS Hive • 搜索：Elasticsearch，云搜索服务（CSS） • 对象存储：对象存储服务（OBS）	分库数据源不支持作为目的端。 分库指的是同时连接多个后端数据源，该连接可作为作业源端，将多个数据源的数据合一迁移到其他数据源上。
NoSQL	Redis	Hadoop：MRS HDFS，MRS HBase，MRS Hive	除了表格存储服务（CloudTable）外，其他NoSQL数据源不支持作为目的端。
	文档数据库服务（DDS）		
	MongoDB		

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

数据源分类	源端数据源	对应的目的端数据源	说明
	MRS Kafka	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• MRS Kafka不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。
搜索	Elasticsearch	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	Elasticsearch仅支持非安全模式。
	云搜索服务（CSS）		导入数据到CSS推荐使用Logstash，请参见**使用Logstash导入数据到Elasticsearch**。

说明

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

整库迁移支持的数据源类型

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

数据集成支持整库迁移的数据源如**表2-4**所示。

表 2-4 整库迁移支持的数据源

数据源分类	数据源	读取	写入	说明
数据仓库	数据仓库服务（DWS）	支持	支持	-
Hadoop （仅支持本地存储，不支持存算分离场景，不支持Ranger场景，不支持ZK开启SSL场景）	MRS HBase	支持	支持	整库迁移仅支持导出到MRS HBase。 建议使用的版本： ● 2.1.X ● 1.3.X
	MRS Hive	支持	支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	FusionInsight HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	FusionInsight Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	Apache HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	Apache Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X

数据源分类	数据源	读取	写入	说明
关系数据库	云数据库 MySQL	支持	支持	不支持OLTP到OLTP迁移，此场景推荐通过数据复制服务DRS进行迁移。
	云数据库 PostgreSQL	支持	支持	
	云数据库 SQL Server	支持	支持	
	MySQL	支持	不支持	
	PostgreSQL	支持	不支持	
	Microsoft SQL Server	支持	不支持	
	Oracle	支持	不支持	
	SAP HANA	支持	不支持	• 仅支持2.00.050.00.1592305219版本。 • 仅支持Generic Edition。 • 不支持BW/4 FOR HANA。 • 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 • 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 • 迁移时不支持目的端自动建表。
	达梦数据库 DM	支持	不支持	仅支持导出到DWS、Hive
NoSQL	Redis	支持	支持	-
	文档数据库服务（DDS）	支持	支持	仅支持DDS和MRS之间迁移。
	表格存储服务（CloudTable）	支持	支持	-

2.3 支持的数据类型

配置字段映射时，数据源支持的数据类型请参见表2-5，以确保数据完整导入到目的端。

表 2-5 支持的数据类型

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

MySQL 数据库迁移时支持的数据类型

源端为MySQL数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 2-6 开源 MySQL 数据库作为源端时支持的数据类型

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

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

类别	类型	简要释义	存储格式示例	Hive	DWS
	BIGINT	一个大的整数，可以带符号。如果有符号，允许范围为-9223372036854775808到92233720368547758 07。如果无符号，允许的范围是从0到184467440737095516 15. 可以指定最多20位的宽度。	5236	BIGINT	BIGINT
	BIGINT UNSIGNED	BIGINT的无符号形式。	-	BIGINT	BIGINT
	MEDIUMINT	一个中等大小的整数，可以带符号。如果有符号，允许范围为-8388608至8388607。 如果无符号，允许的范围是从0到16777215，可以指定最多9位的宽度。	-128，127	INT	INTEGER
	MEDIUMINT UNSIGNED	MEDIUMINT的无符号形式。	-	BIGINT	INTEGER
	TINYINT	一个非常小的整数，可以带符号。如果是有符号，它允许的范围是从-128到127。 如果是无符号，允许的范围是从0到255，可以指定多达4位数的宽度。	100	TINYINT	SMALLINT
	TINYINT UNSIGNED	TINYINT的无符号形式。	-	TINYINT	SMALLINT
	BOOL	MySQL的bool实际上就是tinyint（1）。	-128、127	SMALLINT	BYTEA

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

类别	类型	简要释义	存储格式示例	Hive	DWS
日期时间	DATE	以YYYY-MM-DD格式的日期，在1000-01-01和9999-12-31之间。 例如，1973年12月30日将被存储为1973-12-30。	1999-10-01	DATE	TIMESTAMP
	TIME	用于存储时、分、秒信息。	'09:10:21'或'9:10:21'	不支持（String）	TIME
	DATETIME	日期和时间组合以YYYY-MM-DD HH:MM:SS格式，在1000-01-01 00:00:00到9999-12-31 23:59:59之间。例如，1973年12月30日下午3:30，会被存储为1973-12-30 15:30:00。	'1973-12-30 15:30:00'	TIMESTAMP	TIMESTAMP
	TIMESTAMP	1970年1月1日午夜之间的时间戳，到2037的某个时候。这看起来像前面的DATETIME格式，无需只是数字之间的连字符; 1973年12月30日下午3点30分将被存储为19731230153000（YYYYMMDDHHMMSS）。	19731230153000	TIMESTAMP	TIMESTAMP
	YEAR（M）	以2位或4位数字格式来存储年份。如果长度指定为2（例如YEAR（2）），年份就可以为1970至2069（70～69）。如果长度指定为4，年份范围是1901-2155，默认长度为4。	2000	不支持（String）	不支持
多媒体（二进制）	BINARY（M）	字节数为M，允许长度为0-M的变长二进制字符串，字节数为值得长度加1。	0x2A3B4058（二进制数据）	不支持	BYTEA
	VARBINARY（M）	字节数为M，允许长度为0-M的定长二进制字符串。	0x2A3B4059（二进制数据）	不支持	BYTEA

类别	类型	简要释义	存储格式示例	Hive	DWS
	TEXT	字段的最大长度是65535个字符。 TEXT是“二进制大对象”，并用来存储大的二进制数据，如图像或其他类型的文件。	0x5236（二进制数据）	不支持	不支持
	TINYTEXT	0-255字节短文本二进制字符串。	-	-	不支持
	MEDIUMTEXT	0-167772154字节中等长度文本二进制字符串。	-	-	不支持
	LONGTEXT	0-4294967295字节极大长度文本二进制字符串。	-	-	不支持
	BLOB	字段的最大长度是65535个字符。 BLOB是“二进制大对象”，并用来存储大的二进制数据，如图像或其他类型的文件。BLOB大小写敏感。	0x5236（二进制数据）	不支持	不支持
	TINYBLOB	0-255字节短文本二进制字符串。	-	不支持	不支持
	MEDIUMBLOB	0-167772154字节中等长度文本二进制字符串。	-	不支持	不支持
	LONGBLOB	0-4294967295字节极大长度文本二进制字符串。	0x5236（二进制数据）	不支持	不支持
特殊类型	SET	SET是一个字符串对象，可以有零或多个值，其值来自表创建时规定的允许的一列值。指定包括多个SET成员的SET列值时各成员之间用逗号（ ‘，’ ）间隔开。这样SET成员值本身不能包含逗号。	-	-	不支持
	JSON	-	-	不支持	不支持（TEXT）

类别	类型	简要释义	存储格式示例	Hive	DWS
	ENUM	当定义一个ENUM，要创建它的值的列表，这些是必须用于选择的项（也可以是NULL）。例如，如果想要字段包含“A”或“B”或“C”，那么可以定义为ENUM为 ENUM（“A”，“B”，“C”）也只有这些值（或NULL）才能用来填充这个字段。	-	不支持	不支持

Oracle 数据库迁移时支持的数据类型

源端为Oracle数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 2-7 Oracle 数据库作为源端时支持的数据类型

类别	类型	简要释义	Hive	DWS
字符串	char	定长字符串，会用空格填充来达到最大长度。	CHAR	CHAR
	nchar	包含unicode格式数据的定长字符串。	CHAR	CHAR
	varchar2	是VARCHAR的同义词。这是一个变长字符串，与CHAR类型不同，它不会用空格将字段或变量填充至最大长度。	VARCHAR	VARCHAR
	nvarchar2	包含unicode格式数据的变长字符串。	VARCHAR	VARCHAR
数值	number	能存储精度最多高达38位的数字。	DECIMAL	NUMERIC
	binary_float	2位单精度浮点数。	FLOAT	FLOAT8
	binary_double	64位双精度浮点数。	DOUBLE	FLOAT8
	long	能存储最多2GB的字符数据。	不支持	不支持
日期时间	date	7字节的定宽日期/时间数据类型，其中包含7个属性：世纪、世纪中的哪一年、月份、月中的哪一天、小时、分钟、秒。	DATE	TIMESTAMP

类别	类型	简要释义	Hive	DWS
	timestamp	7字节或11字节的定宽日期/时间数据类型，它包含小数秒。	TIMESTAMP	TIMESTAMP
	timestamp with time zone	3字节的timestamp，提供了时区支持。	TIMESTAMP	TIME WITH TIME ZONE
	timestamp with local time zone	7字节或11字节的定宽日期/时间数据类型，在数据的插入和读取时会发生时区转换。	TIMESTAMP	不支持（TEXT）
	interval year to month	5字节的定宽数据类型，用于存储一个时段。	不支持	不支持（TEXT）
	interval day to second	11字节的定宽数据类型，用于存储一个时段。将时段存储为天/小时/分钟/秒数，还可以有9位小数秒。	不支持	不支持（TEXT）
多媒体（二进制）	raw	一种变长二进制数据类型，采用这种数据类型存储的数据不会发生字符集转换。	不支持	不支持
	long raw	能存储多达2GB的二进制信息。	不支持	不支持
	blob	能够存储最多4GB的数据。	不支持	不支持
	clob	在Oracle 10g及以后的版本中允许存储最多（4GB）×（数据库块大小）字节的数据。CLOB包含要进行字符集转换的信息。这种数据类型很适合存储纯文本信息。	String	不支持
	nclob	这种类型能够存储最多4GB的数据。当字符集发生转换时，这种类型会受到影响。	不支持	不支持
	bfile	可以在数据库列中存储一个oracle目录对象和一个文件名，用户可以通过它来读取这个文件。	不支持	不支持
其他类型	rowid	实际上是数据库表中行的地址，它有10字节长。	不支持	不支持
	urowid	是一个通用的rowid，没有固定的rowid的表。	不支持	不支持

SQL Server 数据库迁移时支持的数据类型

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

表 2-8 SQL Server 数据库作为源端时支持的数据类型

类别	类型	简要释义	Hive	DWS	Oracle
字符串数据类型	char	定长字符串，会用空格填充来达到最大长度。	CHAR	CHAR	CHAR
	nchar	包含unicode格式数据的定长字符串。	CHAR	CHAR	CHAR
	varchar	可变长度的字符串是以长度为1到255之间字符数（高版本的MySQL超过255）；例如： VARCHAR（25）； 创建VARCHAR类型字段时，必须定义长度。	VARCHAR	VARCHAR	VARCHAR
	nvarchar	与varchar类似，存储可变长度Unicode字符数据。	VARCHAR	VARCHAR	VARCHAR
数值数据类型	int	int存储在4个字节中，其中一个二进制位表示符号位，其它31个二进制位表示长度和大小，可以表示-2的31次方~2的31次方-1范围内的所有整数。	INT	INTEGER	INT
	bigint	bigint存储在8个字节中，其中一个二进制位表示符号位，其它63个二进制位表示长度和大小，可以表示-2的63次方~2的63次方-1范围内的所有整数。	BIGINT	BIGINT	NUMBER
	smallint	smallint类型的数据占用了两个字节的存储空间，其中一个二进制位表示整数值的正负号，其它15个二进制位表示长度和大小，可以表示-2的15次方~2的15次方-1范围内的所有整数。	SMALLINT	SMALLINT	NUMBER
	tinyint	tinyint类型的数据占用了一个字节的存储空间，可以表示0~255范围内的所有整数。	TINYINT	TINYINT	NUMBER
	real	可以存储正的或者负的十进制数值。	DOUBLE	FLOAT4	NUMBER
	float	其中为用于存储float数值尾数的位数（以科学计数法表示），因此可以确定精度和存储大小。	FLOAT	FLOAT8	binary_float

类别	类型	简要释义	Hive	DWS	Oracle
	decimal	带固定精度和小数位数的数值数据类型。	DECIMAL	NUMERIC	NUMBER
	numeric	用于存储零、正负定点数。	DECIMAL	NUMERIC	NUMBER
日期时间数据类型	date	存储用字符串表示的日期数据。	DATE	TIMESTAMP	DATE
	time	以字符串形式记录一天的某个时间。	不支持（String）	TIME	不支持
	datetime	用于存储时间和日期数据。	TIMESTAMP	TIMESTAMP	不支持
	datetime2	datetime的扩展类型，其数据范围更大，默认的最小精度最高，并具有可选的用户定义的精度。	TIMESTAMP	TIMESTAMP	不支持
	smalldatetime	smalldatetime类型与datetime类型相似，只是其存储范围是从1900年1月1日到2079年6月6日，当日期时间精度较小时，可以使用smalldatetime，该类型数据占用4个字节的存储空间。	TIMESTAMP	TIMESTAMP	不支持
	datetimeoffset	用于定义一个采用24小时制与日期相组合并可识别时区的时间。	不支持（String）	TIMESTAMP	不支持
多媒体数据类型（二进制）	text	用于存储文本数据。	不支持（String）	不支持（String）	不支持
	netxt	与text类型作用相同，为长度可变的非Unicode数据。	不支持（String）	不支持（String）	不支持
	image	长度可变的二进制数据，用于存储照片、目录图片或者图画。	不支持（String）	不支持（String）	不支持
	binary	长度为n个字节的固定长度二进制数据，其中n是从1~8000的值。	不支持（String）	不支持（String）	不支持
	varbinary	可变长度二进制数据。	不支持（String）	不支持（String）	不支持
货币数据类型	money	用于存储货币值。	不支持（String）	不支持（String）	不支持

类别	类型	简要释义	Hive	DWS	Oracle
	smallmoney	与money类型相似，输入数据时在前面加上一个货币符号，如人民币为￥或其它定义的货币符号。	不支持（String）	不支持（String）	不支持
位数据类型	bit	位数据类型，只取0或1为值，长度1字节。bit值经常当作逻辑值用于判断true（1）或false（0），输入非0值时系统将其替换为1。	不支持	不支持	不支持
其他数据类型	rowversion	每个数据都有一个计数器，当对数据库中包含rowversion列的表执行插入或者更新操作时，该计数器数值就会增加。	不支持	不支持	不支持
	uniqueidentifier	16字节的GUID（Globally Unique Identifier，全球唯一标识符），是Sql Server根据网络适配器地址和主机CPU时钟产生的唯一号码，其中，每个为都是0~9或a~f范围内的十六进制数字。	不支持	不支持	不支持
	cursor	游标数据类型。	不支持	不支持	不支持
	sql_variant	用于存储除文本，图形数据和timestamp数据外的其它任何合法的Sql Server数据，可以方便Sql Server的开发工作。	不支持	不支持	不支持
	table	用于存储对表或视图处理后的结果集。	不支持	不支持	不支持
	xml	存储xml数据的数据类型。可以在列中或者xml类型的变量中存储xml实例。存储的xml数据类型表示实例大小不能超过2GB。	不支持	不支持	不支持

PostgreSQL 数据库迁移时支持的数据类型

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

表 2-9 PostgreSQL 数据库作为源端时支持的数据类型

类别	类型	简要释义	Hive	DWS	DLI
字符	char	定长字符串，存储右空格填充到指定的长度。	CHAR	CHAR	不支持（String）
	varchar	变长字符串，不会用空格将字段或变量填充至最大长度。	CARCHAR	CARCHAR	不支持（String）

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

类别	类型	简要释义	Hive	DWS	DLI
	time	只用于一日内时间，无时区。	不支持（String）	TIME	不支持（String）
	timez	只用于一日内时间，有时区。	不支持（String）	TIMEZ	不支持（String）
	interval	时间间隔。	不支持（String）	不支持（String）	不支持（String）
位串类型	bit	定长位串，例如：b'000101'。	不支持（String）	不支持（String）	不支持（String）
	varbit	可变长位串，例如：b'101'。	不支持（String）	不支持（String）	不支持（String）
货币类型	money	存储在8个字节中，它允许的范围是从-92233720368547.5808到922337203685477.5807。	DOUBLE	MONEY	DECIMAL（P，S）
布尔类型	boolean	存储在1个字节中，可以取值为 1、0 或 NULL。	BOOLEAN	BOOLEAN	BOOLEAN
文本类型	text	变长文本，无长度限制。	不支持（String）	不支持（String）	不支持（String）

DWS 数据库迁移时支持的数据类型

源端为DWS数据库时，支持的数据类型如下：

表 2-10 DWS 数据库作为源端时支持的数据类型

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

类别	类型	简要释义
	numeric	用于存储零、正负定点数。
	real	与double相同。
	int	int存储在4个字节中，其中一个二进制位表示符号位，其它31个二进制位表示长度和大小，可以表示-2的31次方~2的31次方-1范围内的所有整数。
	bigint	bigint存储在8个字节中，其中一个二进制位表示符号位，其它63个二进制位表示长度和大小，可以表示-2的63次方~2的63次方-1范围内的所有整数。
	smallint	smallint类型的数据占用了两个字节的存储空间，其中一个二进制位表示整数值的正负号，其它15个二进制位表示长度和大小，可以表示-2的15次方~2的15次方-1范围内的所有整数。
	tinyint	tinyint类型的数据占用了一个字节的存储空间，可以表示0~255范围内的所有整数。
日期时间	date	存储日期数据。
	timestamp	存储日期和时间数据，无时区。
	time	只用于一日内时间，无时区。
位串类型	bit	定长位串，例如： b'000101'。
布尔类型	boolean	存储在1个字节中，可以取值为 1、0 或 NULL。
文本类型	text	变长文本，无长度限制。

神通（ST）数据库迁移时支持的数据类型

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

表 2-11 神通（ST）数据库作为源端时支持的数据类型

类别	类型	简要释义	存储格式示例	MRS Hive	MRS Hudi
字符	VARCHAR	用于存储指定定长字符串。	‘a’或‘aaaaa’	VARCHAR（765）	STRING
	BPCHAR	用于存储指定变长字符串。	‘a’或‘aaaaa’	VARCHAR（765）	STRING

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

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

SAP HANA 数据库迁移时支持的数据类型

源端为SAP HANA数据库时，支持的数据类型如下：

表 2-12 SAP HANA 数据库作为源端时支持的数据类型

类别	类型	简要释义
字符	VARCHAR	用于存储指定定长字符串。
	NVARCHAR	包含unicode格式数据的变长字符串。
	TEXT	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。
数值	BIGINT	用于存储有符号整数，精度为19，标度为0。
	TINYINT	用于存储有符号整数，精度为3，标度为0。
	SMALLINT	SMALLINT类型的数据占用了两个字节的存储空间，其中一个二进制位表示整数值的正负号，其它15个二进制位表示长度和大小，可以表示-2的15次方~2的15次方-1范围内的所有整数。
	REAL	可以存储正的或者负的十进制数值。
	DECIMAL	带固定精度和小数位数的数值数据类型。
	FLOAT	用于存储带二进制精度的浮点数。
	DOUBLE	用于存储指明双精度的浮点数。
日期时间	DATE	用于存储年、月、日信息。
	TIME	用于存储时、分、秒信息。
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。
多媒体	CLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。
	NCLOB	这种类型能够存储最多4GB的数据。当字符集发生转换时，这种类型会受到影响。
布尔类型	BOOLEAN	存储在1个字节中，可以取值为 1、0 或 NULL。

DLI 数据库迁移时支持的数据类型

源端为DLI数据库时，支持的数据类型如下：

表 2-13 DLI 数据库作为源端时支持的数据类型

类别	类型	简要释义
字符	CHAR	用于存储指定定长字符串。
	VARCHAR	与CHAR相同。
	STRING	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。
数值	BIGINT	用于存储有符号整数，精度为19，标度为0。
	TINYINT	用于存储有符号整数，精度为3，标度为0。
	SMALLINT	SMALLINT类型的数据占用了两个字节的存储空间，其中一个二进制位表示整数值的正负号，其它15个二进制位表示长度和大小，可以表示-2的15次方~2的15次方-1范围内的所有整数。
	INT	用于存储有符号整数，精度为10，标度为0。
	DECIMAL	带固定精度和小数位数的数值数据类型。
	FLOAT	用于存储带二进制精度的浮点数。
	DOUBLE	用于存储指明双精度的浮点数。
日期时间	DATE	用于存储年、月、日信息。
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。
布尔类型	BOOLEAN	存储在1个字节中，可以取值为 1、0 或 NULL。

Elasticsearch/云搜索服务（CSS）数据库迁移时支持的数据类型

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

表 2-14 Elasticsearch/云搜索服务（CSS）数据库作为源端时支持的数据类型

类别	类型	简要释义	存储格式示例	MySQL
字符	keyword	用于存储字符串。	“keyword”	String
	text	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。	“long string”	TEXT

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

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

达梦数据库迁移时支持的数据类型

源端为达梦数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 2-15 达梦数据库作为源端时支持的数据类型

类别	类型	简要释义	存储格式示例	Hive	DWS
字符	CHAR	用于存储指定定长字符串。	‘a’或‘aaaaa’	CHAR	CHAR
	CHARACTER	与 CHAR 相同。	‘a’或‘aaaaa’	CHAR	CHAR
	VARCHAR	用于存储指定变长字符串.	‘a’或‘aaaaa’	VARCHAR	VARCHAR
	VARCHAR2	与 VARCHAR 相同。	‘a’或‘aaaaa’	VARCHAR	VARCHAR

类别	类型	简要释义	存储格式示例	Hive	DWS
数值	NUMERIC	用于存储零、正负定点数。	52.36	DECIMAL	NUMERIC
	DECIMAL	与 NUMERIC 相似。	52.36	DECIMAL	NUMERIC
	DEC	与 DECIMAL 相同。	52.36	DECIMAL	NUMERIC
	NUMBER	与 NUMERIC 相同。	52.36	DECIMAL	NUMERIC
	INTEGER	用于存储有符号整数，精度为10，标度为0。	5236	INT	INTEGER
	INT	与 INTEGER 相同。	5236	INT	INTEGER
	BIGINT	用于存储有符号整数，精度为19，标度为0。	5236	BIGINT	BIGINT
	TINYINT	用于存储有符号整数，精度为3，标度为0。	100	TINYINT	SMALLINT
	SMALLINT	用于存储有符号整数，精度为5，标度为0。	9999	SMALLINT	SMALLINT
	BYTE	与 TINYINT 相似，精度为3，标度为0。	100	TINYINT	SMALLINT
	BINARY	用于存储定长二进制数据。	0x2A3B4058	BINARY（NULL）	BYTEA（NULL）
	VARBINARY	用于存储可变长二进制数据。	0x2A3B4058	BINARY（NULL）	BYTEA（NULL）
	FLOAT	用于存储带二进制精度的浮点数。	52.36	FLOAT	FLOAT8
	DOUBLE	与FLOAT类似。	52.36	DOUBLE	FLOAT8
	REAL	用于存储带二进制精度的浮点数，但它不能由用户指定使用的精度。	52.3	FLOAT	FLOAT4
	DOUBLE PRECISION	用于存储指明双精度的浮点数。	52.3	DOUBLE	FLOAT8

类别	类型	简要释义	存储格式示例	Hive	DWS
位串	BIT	用于存储整数数据1、0 或 NULL。	1、0 或 NULL	TINYINT（1 0 NULL）	BOOLEAN（true false NULL）
日期时间	DATE	用于存储年、月、日信息。	1999-10-01'、'1999/10/01'或'1999.10.01'	DATE	TIMESTAMP
	TIME	用于存储时、分、秒信息。	'09:10:21'或'9:10:21'	不支持（String）	TIME
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。	2002-12-12 09:10:21'，'2002-12-12 9:10:21' '2002/12/12 09:10:21' 或'2002.12.12 09:10:21'	TIMESTAMP	TIMESTAMP
	TIME WITH TIME ZONE	用于存储一个带时区的 TIME 值，其定义是在 TIME 类型的后面加上时区信息。	'09:10:21 +8:00'，'09:10:21+8:00'或'9:10:21+8:00'	不支持（String）	TIME WITH TIME ZONE
	TIMESTAMP WITH TIME ZONE	用于存储一个带时区的 TIMESTAMP 值，其定义是TIMESTAMP类型的后面加上时区信息。	2002-12-12 09:10:21 +8:00'，'2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00'或'2002.12.12 09:10:21 +8:00'	TIMESTAMP	TIMESTAMP WITH TIME ZONE

类别	类型	简要释义	存储格式示例	Hive	DWS
	TIMESTAMP WITH LOCAL TIME ZONE	用于存储一个本地时区的 TIMESTAMP 值，能够将标准时区类型 TIMESTAMP WITH TIME ZONE 类型转化为本地时区类型。	2002-12-12 09:10:21 +8:00'，'2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00'或'2002.12.12 09:10:21 +8:00'	不支持（String）	不支持（TEXT）
	DATETIME WITH TIME ZONE	同TIMESTAMP WITH TIME ZONE。	2002-12-12 09:10:21 +8:00'，'2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00'或'2002.12.12 09:10:21 +8:00'	TIMESTAMP	TIMESTAMP WITH TIME ZONE
	INTERVAL YEAR	描述一个若干年的间隔，引导精度规定了年的取值范围。	INTERVAL '0015' YEAR	不支持（String）	不支持（VARCHAR）
	INTERVAL YEAR TO MONTH	描述一个若干年若干月的间隔，引导精度规定了年的取值范围。	INTERVAL '0015-08' YEAR TO MONTH	不支持（String）	不支持（VARCHAR）
	INTERVAL MONTH	描述一个若干月的间隔，引导精度规定了月的取值范围。	INTERVAL '0015' MONTH	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY	描述一个若干日的间隔，引导精度规定了日的取值范围。	INTERVAL '150' DAY	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY TO HOUR	描述一个若干日若干小时的间隔，引导精度规定了日的取值范围。	INTERVAL '923' DAY TO HOUR	不支持（String）	不支持（VARCHAR）

类别	类型	简要释义	存储格式示例	Hive	DWS
	INTERVAL DAY TO MINUTE	描述一个若干日若干小时若干分钟的间隔，引导精度规定了日的取值范围。	INTERVAL '09 23:12' DAY TO MINUTE	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY TO SECOND	描述一个若干日若干小时若干分钟若干秒的间隔，引导精度规定了日的取值范围。	INTERVAL '09 23:12:01.1' DAY TO SECOND	不支持（String）	不支持（VARCHAR）
	INTERVAL HOUR	描述一个若干小时的间隔，引导精度规定了小时的取值范围。	INTERVAL '150' HOUR	不支持（String）	不支持（VARCHAR）
	INTERVAL HOUR TO MINUTE	描述一个若干小时若干分钟的间隔，引导精度规定了小时的取值范围。	INTERVAL '23:12' HOUR TO MINUTE	不支持（String）	不支持（VARCHAR）
	INTERVAL HOUR TO SECOND	描述一个若干小时若干分钟若干秒的间隔，引导精度规定了小时的取值范围。	INTERVAL '23:12:01.1' HOUR TO SECOND	不支持（String）	不支持（VARCHAR）
	INTERVAL MINUTE	描述一个若干分钟的间隔，引导精度规定了分钟的取值范围。	INTERVAL '150' MINUTE	不支持（String）	不支持（VARCHAR）
	INTERVAL MINUTE TO SECOND	描述一个若干分钟若干秒的间隔，引导精度规定了分钟的取值范围。	INTERVAL '12:01.1' MINUTE TO SECOND	不支持（String）	不支持（VARCHAR）
	INTERVAL SECOND	描述一个若干秒的间隔，引导精度规定了秒整数部分的取值范围。	INTERVAL '51.1' SECOND	不支持（String）	不支持（VARCHAR）

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

3 管理集群

3.1 创建集群

操作场景

目前CDM采用独立集群的方式为用户提供安全可靠的数据迁移服务，各集群之间相互隔离，不可相互访问。目前一个集群只支持一个服务器。

前提条件

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

说明

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

  注：如果配置了VPC对等连接，可能会出现对端VPC子网与CDM管理网重叠，从而无法访问对端VPC中数据源的情况。推荐使用公网做跨VPC数据迁移，或联系管理员在CDM后台为VPC对等连接添加特定路由。
- 当CDM集群与其他云服务所在的区域不一致时，则需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP、CDM云上安全组出方向放通云下数据源所在的主机、数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
- 另外，如果创建了企业项目，则企业项目也会影响CDM集群与其他云服务的网络互通，只有企业项目一致的云服务才能打通网络。

操作步骤

步骤1 进入**购买云数据迁移服务**界面。

步骤2 配置CDM集群参数，各参数说明如**表3-1**所示。

表 3-1 CDM 集群参数

参数名称	样例	说明
当前区域	中国-香港	选择CDM集群的区域，不同区域的资源之间内网不互通。
可用区	可用区2	请参见**可用区**。
集群名称	cdm-aff1	自定义CDM集群名称。 **说明** CDM集群创建后，不支持修改集群名称。
实例类型	cdm.large	目前CDM支持以下规格供用户选择： • cdm.large：8核CPU、16G内存的虚拟机，最大带宽/基准带宽为3/0.8 Gbps，集群作业并发数上限为16。 • cdm.xlarge：16核CPU、32G内存的虚拟机，最大带宽/基准带宽为10/4 Gbps，集群作业并发数上限为32，适合使用10GE高速带宽进行TB级以上的数据量迁移。 • cdm.4xlarge：64核CPU、128G内存的虚拟机，最大带宽/基准带宽为40/36 Gbps，集群作业并发数上限为128。 • 购买DataArts Studio赠送的4核CPU、8G内存的虚拟机，仅支持作业单并发运行。
虚拟私有云	vpc1	CDM集群所属VPC、子网、安全组，需确保CDM集群与待连接的数据源能正常通信。用户可以根据CDM迁移的数据源端、目的端所处网络进行选择： • 如果CDM集群与待连接的数据源所属不同的VPC，或者待连接的为本地数据源时，CDM集群需要绑定EIP，通过公网通信。 • 如果待连接的数据源为云上服务，则推荐CDM集群的网络配置与该云服务一致，此时CDM集群不用绑定EIP，通过内网通信。 • 如果待连接的数据源为云上服务，CDM与它在同一个VPC但所属不同子网，则可以通过配置安全组规则来使CDM集群与云服务间的网络互通。 VPC、子网、安全组的详细操作，请参见《**虚拟私有云用户指南**》。 **说明** • 目前CDM实例创建完成后不支持切换VPC、子网、安全组，请谨慎选择。 • 此处支持选择共享VPC子网，即由VPC的所有者将VPC内的子网共享给当前账号，由当前账号在购买CDM集群时选择共享VPC子网。通过共享VPC子网功能，可以简化网络配置，帮助您统一配置和运维多个账号下的资源，有助于提升资源的管控效率，降低运维成本。如何共享VPC子网，请参考《**共享VPC**》。
子网	subnet-1	
安全组	sg-1	

参数名称	样例	说明
企业项目	default	在管理控制台，单击右上角的“企业”，可进入企业项目管理界面创建企业项目。 企业项目管理服务是一种云资源管理方式，具体请参见《企业管理用户指南》。
标签	cluster_owner:cdm	高级配置参数选择自定义时可配置标签参数。 如果您需要使用同一标签标识多种云资源，可以自定义填写标签键及对应的标签值，后续可在TMS标签系统中可筛选出同一标签的云资源。 **说明** • 一个集群最多可添加10个标签。 • 标签键（key）的最大长度为36个字符，标签值（value）的最大长度为43个字符。
消息通知	否	开启后，支持配置20个手机号码或邮箱，作业（目前仅支持表/文件迁移的作业）失败时、EIP异常时会发送短信或邮件通知用户。

步骤3 查看当前配置，确认无误后单击“立即购买”进入规格确认界面。

说明

集群创建好以后不支持修改规格，如果需要使用更高规格，需要重新创建。

步骤4 单击“提交”，系统开始自动创建CDM集群，在“集群管理”界面可查看创建进度。

----结束

3.2 解绑/绑定集群的 EIP

操作场景

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

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

说明

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

前提条件

- 已创建CDM集群。
- 已拥有EIP配额，才能绑定EIP。

操作步骤

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

图 3-1 集群列表

说明

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

步骤2 对相应需要操作的集群可以进行绑定EIP或解绑EIP的操作。

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

步骤3 单击“确定”绑定或解绑EIP。

----结束

3.3 重启集群

操作场景

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

前提条件

已创建CDM集群。

重启集群

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

图 3-2 集群列表

说明

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

步骤2 选择集群操作列中的“更多 > 重启”，进入重启集群确认界面。

图 3-3 重启集群

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

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

----结束

3.4 删除集群

操作场景

当您确认不再使用当前集群后，可以删除当前CDM集群。

⚠ 注意

删除CDM集群后集群以及数据都销毁且无法恢复，请您谨慎操作！

删除集群前，请您确认如下注意事项：

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

前提条件

已创建CDM集群。

删除集群

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

图 3-4 集群列表

说明

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

步骤2 通过以下两种方式进入删除集群确认界面。

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

步骤3 输入“DELETE”后单击“确定”，即开始删除CDM集群。

图 3-5 删除集群 1

----结束

3.5 下载集群日志

操作场景

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

前提条件

已创建CDM集群。

操作步骤

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

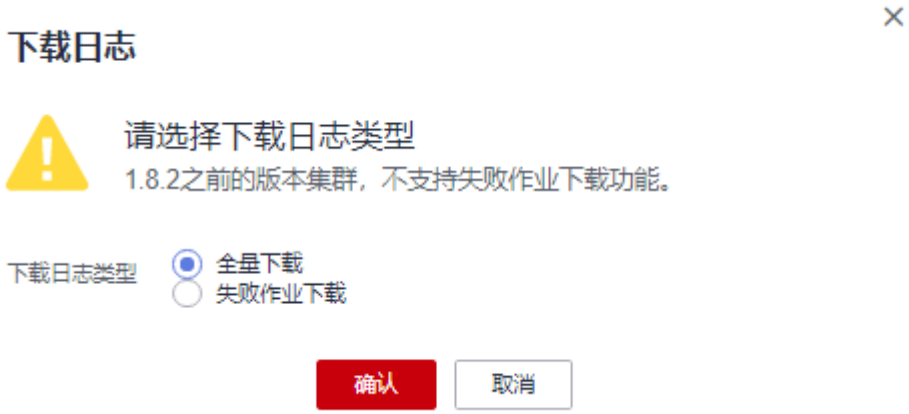
图 3-6 集群列表

说明

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

步骤2 选择集群操作列中的“更多 > 下载日志”，选择下载日志类型。

图 3-7 下载日志类型



步骤3 确认后，即可下载日志到本地。

----结束

3.6 查看集群基本信息/修改集群配置

操作场景

CDM集群已经创建成功后，您可以查看集群基本信息，并修改集群的配置。

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

      注意，用户隔离关闭后需要重启集群VM才能生效。
  - 最大抽取并发数：限制作业运行的总抽取并发数，如果当前所有作业总并发数超出限制，超出部分将排队等待。

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

    说明

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

前提条件

已创建CDM集群。

查看集群基本信息

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

图 3-8 集群列表

说明

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

步骤2 单击集群名称，可查看集群的基本信息。

----结束

修改集群配置

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

图 3-9 集群列表

说明

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

步骤2 单击集群名称后，选择“集群配置”页签，可修改消息通知、用户是否隔离以及最大抽取并发数的配置。

步骤3 修改完成后单击“保存”，返回集群管理界面。

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

图 3-10 重启集群

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

步骤5 选择“重启集群VM”后单击“确定”。

----结束

3.7 管理集群标签

操作场景

CDM集群已经创建成功后，支持新增、修改及删除CDM集群的标签。使用标签可以标识多种云资源，后续在TMS标签系统或者CDM集群管理列表中可筛选出同一标签的云资源。

说明

一个CDM集群最多可新增10个标签。

前提条件

已创建CDM集群。

操作步骤

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

图 3-11 集群列表

说明

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

步骤2 单击集群名称后，选择“标签”页签。

图 3-12 修改集群配置

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

图 3-13 添加标签

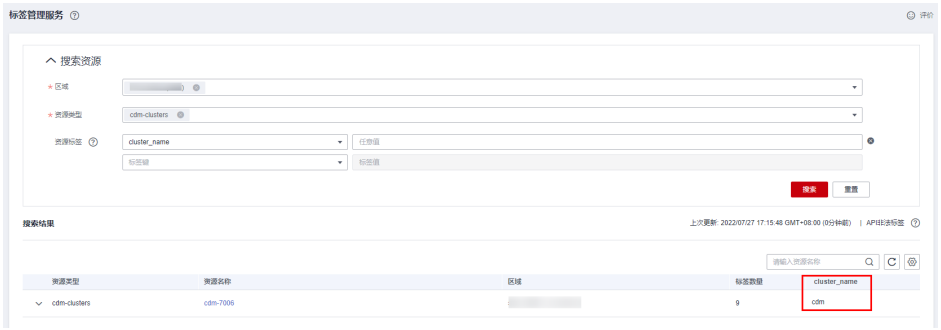
说明

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

步骤4 （可选）在标签列表中，单击标签操作列“删除”，删除CDM集群标签。

步骤5 通过以下两种方式筛选出所配置标签的资源。

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



- 在集群列表中，单击标签搜索，筛选出所配置标签的资源。

----结束

3.8 查看监控指标

3.8.1 支持的监控指标

功能说明

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

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

- 本章节描述了CDM上报云监控的监控指标的命名空间、监控指标列表和维度定义。
- 如果您需要查看CDM相关的监控指标，请参见**查看监控指标**。
- 如果您需要在监控数据满足指定条件时发送报警通知，可参见**设置告警规则**。

前提条件

使用CDM监控功能，需获取CES相关权限。

命名空间

SYS.CDM

监控指标

CDM集群支持的监控指标如**表3-2**所示。

表 3-2 CDM 支持的监控指标

指标ID	指标名称	指标含义	取值范围	测量对象	监控周期（原始指标）
bytes_in	网络流入速率	该指标用于统计每秒流入测量对象的网络流量。 单位：字节/秒。	≥ 0 bytes/s	CDM集群实例	1分钟
bytes_out	网络流出速率	该指标用于统计每秒流出测量对象的网络流量。 单位：字节/秒。	≥ 0 bytes/s	CDM集群实例	1分钟
cpu_usage	CPU使用率	该指标用于统计测量对象的CPU使用率。 单位：%。	0%～100%	CDM集群实例	1分钟
mem_usage	内存使用率	该指标用于统计测量对象的内存使用率。 单位：%。	0%～100%	CDM集群实例	1分钟

指标ID	指标名称	指标含义	取值范围	测量对象	监控周期（原始指标）
pg_pending_job	排队作业数	该指标用于统计该CDM实例中处于PENDING状态的作业数。 单位：Count/个。	>=0	CDM集群实例	1分钟
pending_threads	排队抽取并发数	该指标用于统计该CDM实例中处于Waiting状态的抽取并发线程数。 单位：Count/个。	>=0	CDM集群实例	1分钟
disk_usage	磁盘利用率	该指标为从物理机层面采集的磁盘使用率，数据准确性低于从弹性云服务器内部采集的数据。 单位：%。	0.001%~90%	CDM集群实例	1分钟
disk_io	磁盘io	该指标为从物理机层面采集的磁盘每秒读取和写入的字节数，数据准确性低于从弹性云服务器内部采集的数据。 单位：Byte/sec	0~10GB	CDM集群实例	1分钟
tomcat_heap_usage	堆内存使用率	该指标为从物理机层面采集的堆内存使用率，数据准确性低于从弹性云服务器内部采集的数据。 单位：%。	0.001%~90%	CDM集群实例	1分钟
tomcat_connect	tomcat并发连接数	该指标为从物理机层面采集的tomcat并发连接数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟
tomcat_thread_count	tomcat线程数	该指标为从物理机层面采集的tomcat所占线程数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟
pg_connect	数据库连接数	该指标为从物理机层面采集的postgres数据库连接数。 单位：Count/个。	0~2147483647	CDM集群实例	1分钟

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

维度

Key	Value
instance_id	云数据迁移服务实例

3.8.2 设置告警规则

操作场景

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

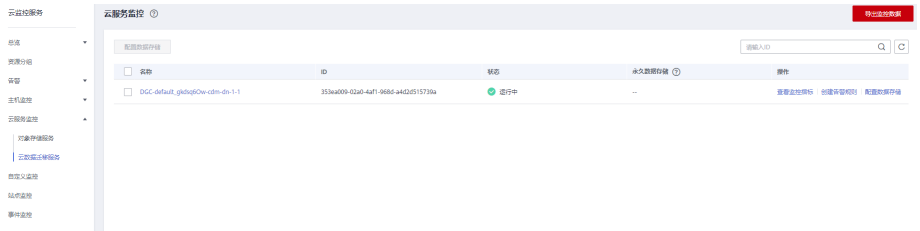
设置CDM集群的告警规则包括设置告警规则名称、监控对象、监控指标、告警阈值、监控周期和是否发送通知等参数。本节介绍了设置CDM集群告警规则的具体方法。

操作步骤

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

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

图 3-14 “云数据迁移服务”服务监控项



步骤3 根据界面提示设置CDM集群的告警规则。

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

说明

更多关于监控告警的信息，请参见**云监控用户指南**。

----结束

3.8.3 查看监控指标

操作场景

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

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

前提条件

- CDM集群正常运行。

  重启失败、不可用状态的集群，无法查看其监控指标。当集群再次启动或恢复后，即可正常查看。
- CDM集群已正常运行一段时间（约10分钟）。

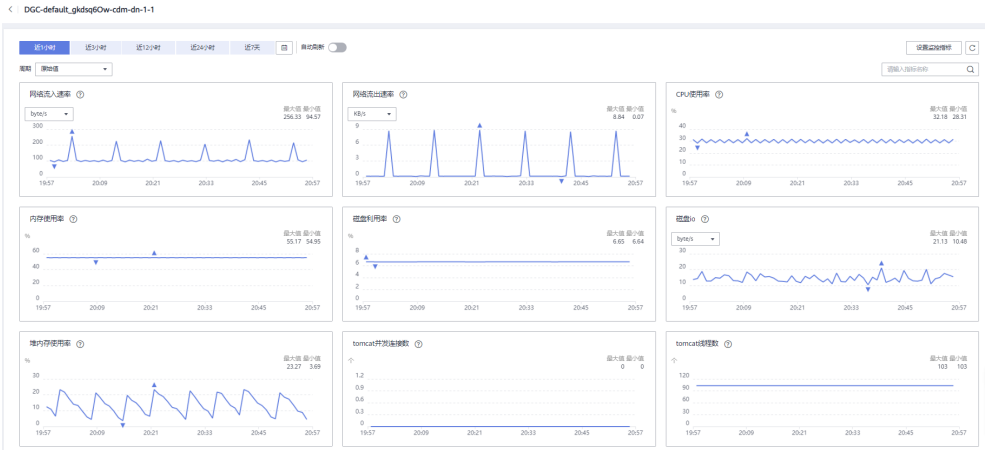
  对于新创建的集群，需要等待一段时间，才能查看上报的监控数据和监控视图。

操作步骤

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

步骤2 在CDM监控页面，可查看所有监控指标的小图。

图 3-15 查看监控指标



步骤3 单击小图右上角的 ，可进入大图模式查看。

步骤4 您可以在左上角选择时长作为监控周期，查看一段时间的指标变化情况。

----结束

4 管理连接

4.1 新建连接

操作场景

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

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

约束限制

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

前提条件

- 已具备CDM集群。
- CDM集群与目标数据源可以正常通信。
  - 如果目标数据源为云下的数据库，则需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP、CDM云上安全组出方向放通云下数据源所在的主机、数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
  - 如果目标数据源为云上服务（如DWS、MRS及ECS等），则网络互通需满足如下条件：
    - CDM集群与云上服务处于不同区域的情况下，需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP，数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
    - CDM集群与云上服务同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，

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

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

- 已获取待连接数据源的地址、用户名和密码，且该用户拥有数据导入、导出的操作权限。
- 使用Agent时需用主账户给子账户赋予CDM操作权限。

新建连接

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

图 4-1 集群列表

说明

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

步骤2 在CDM主界面，单击左侧导航上的“集群管理”，选择CDM集群后的“作业管理 > 连接管理 > 新建连接”。选择连接器类型，如**图4-2**所示。

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

图 4-2 选择连接器类型

步骤3 选择数据源类型后，单击“下一步”配置连接参数，这里以创建MySQL连接为例。

每种数据源的连接参数不同，您可以根据所选择的连接器类型在**表4-1**中查找对应参数。

表 4-1 连接参数分类

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

连接器类型	参数说明
MongoDB	连接本地MongoDB数据库时，具体参数请参见**配置MongoDB连接**。
文档数据库服务（DDS）	连接DDS时，具体参数请参见**配置DDS连接**。
• Redis • 分布式缓存服务（DCS）	连接Redis或DCS时，具体参数请参见**配置Redis连接**。
• MRS Kafka • Apache Kafka	连接MRS Kafka或Apache Kafka数据源时，具体参数请参见**配置Kafka连接**。
数据接入服务（DIS）	连接DIS时，具体参数请参见**配置DIS连接**。
云搜索服务 Elasticsearch	连接云搜索服务或Elasticsearch时，具体参数请参见**配置云搜索服务（CSS）连接**。
数据湖探索（DLI）	连接数据湖探索服务时，具体参数请参见**配置DLI连接**。
DMS Kafka	连接DMS的Kafka队列时，具体参数请参见**配置DMS Kafka连接**。
Cassandra	连接Cassandra时，具体参数请参见**配置Cassandra连接**。
MRS Hudi	连接MRS Hudi时，具体参数请参见**配置MRS Hudi连接**。
MRS ClickHouse	连接MRS ClickHouse时，具体参数请参见**配置MRS ClickHouse连接**。
神通数据库（ST）	连接神通数据库（ST）时，具体参数请参见**配置神通（ST）连接**。

说明

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

步骤4 连接的参数配置完成后单击“测试”，可测试连接是否可用。或者直接单击“保存”，保存时也会先检查连接是否可用。

受网络和数据源的影响，部分连接测试的时间可能需要30～60秒。

----结束

管理连接

CDM支持对已创建的连接进行以下操作：

- 删除：支持删除未被任何作业使用的连接，也支持批量删除连接。

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

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

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

步骤2 在连接管理界面找到需要修改的连接：

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

----结束

4.2 管理驱动

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

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

前提条件

- 已创建集群。
- 已参见**表4-2**下载对应的驱动。
- 已参见**配置FTP/SFTP连接**创建SFTP连接并将对应的驱动上传至线下文件服务器（可选）。

如何获取驱动

不同类型的关系数据库，需要适配不同类型的驱动。注意，上传的驱动版本不必与待连接的数据库版本相匹配，直接参考**表4-2**获取建议版本的JDK8 .jar格式驱动即可。

表 4-2 获取驱动

关系数据库类型	驱动名称	获取地址	建议版本
• 云数据库 MySQL • MySQL	MYSQL	**https://downloads.mysql.com/archives/c-j/**	5.1.48版本，获取mysql-connector-java-5.1.48.jar
Oracle	ORACLE_6 ORACLE_7 ORACLE_8	驱动包下载地址：**https://www.oracle.com/database/technologies/appdev/jdbc-downloads.html** 历史版本驱动包下载地址：**https://repo1.maven.org/maven2/com/oracle/database/jdbc/**	ojdbc8的12.2.0.1版本，获取ojdbc8.jar **说明** 不支持使用新版本（如Oracle Database 21c (21.3) drivers），会导致创建作业时无法获取模式名。
• 云数据库 PostgreSQL • PostgreSQL	POSTGRESQL	**https://mvnrepository.com/artifact/org.postgresql/postgresql**	PostgreSQL推荐使用42.3.4版本，获取postgresql-42.3.4.jar
金仓数据库	POSTGRESQL	**https://mvnrepository.com/artifact/org.postgresql/postgresql**	金仓数据库推荐使用42.2.9版本PostgreSQL驱动，获取postgresql-42.2.9.jar
GaussDB数据库	POSTGRESQL	GaussDB JDBC驱动请在**GaussDB官方文档**中搜索“JDBC包、驱动类和环境类”，然后选择实例对应版本的文档，参考文档获取gsjdbc4.jar。	请从对应版本的发布包中获取gsjdbc4.jar
• 云数据库 SQL Server • Microsoft SQL Server	SQLServer	**https://docs.microsoft.com/en-us/sql/connect/jdbc/release-notes-for-the-jdbc-driver?view=sql-server-ver15#previous-releases**	4.2版本，获取sqljdbc42.jar

操作步骤

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

图 4-3 上传驱动

更新驱动需要重启cdm集群才能生效。

驱动名称	驱动库名	建议版本 ⓘ	备注	操作
MYSQL	mysql-connector-java-5.1.48.jar	建议版本5.1.48，获取mysql-connector-java-5.1.48.jar，请参考 管理驱动 获取。		上传 | 从sftp复制
ORACLE_6	ojdbc6.jar	建议版本12.1.0.2，获取ojdbc6.jar，请参考 管理驱动 获取。	oracle < 12.1	上传 | 从sftp复制
ORACLE_8	ojdbc8.jar	建议版本12.2.0.1，获取ojdbc8.jar，请参考 管理驱动 获取。	oracle > 12.1	上传 | 从sftp复制
ORACLE_7	ojdbc6-11.2.0.4.jar	建议版本12.1.0.2，获取ojdbc7.jar，请参考 管理驱动 获取。	oracle = 12.1	上传 | 从sftp复制
POSTGRESQL	postgresql-42.1.4.jar	建议版本42.3.4，获取postgresql-42.3.4.jar，请参考 管理驱动 获取。		上传 | 从sftp复制
SQLSERVER	sqljdbc42.jar	建议版本4.2，获取sqljdbc42.jar，请参考 管理驱动 获取。		上传 | 从sftp复制
POSTGRESQL_KINGBASE	kingbase8-8.6.0.jar	建议版本与KINGBASE数据库一致，请参考 管理驱动 获取。	KINGBASE database	上传 | 从sftp复制
DORIS	mysql-connector-java-5.1.48.jar	请参考 管理驱动 获取。		上传 | 从sftp复制
DM	DmJdbcDriver18.jar	DM JDBC驱动jar包请从DM安装目录/dmdbms/drivers/jdbc中获取DmJdbcDriver18.jar。		上传 | 从sftp复制

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

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

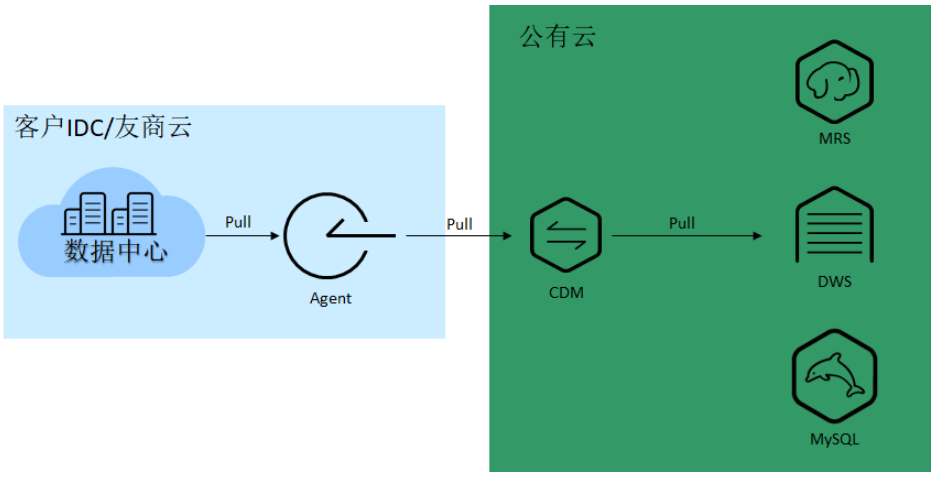
步骤3 （可选）在驱动更新场景下，上传驱动后必须在CDM集群列表中重启集群才能更新生效。

----结束

4.3 管理 Agent

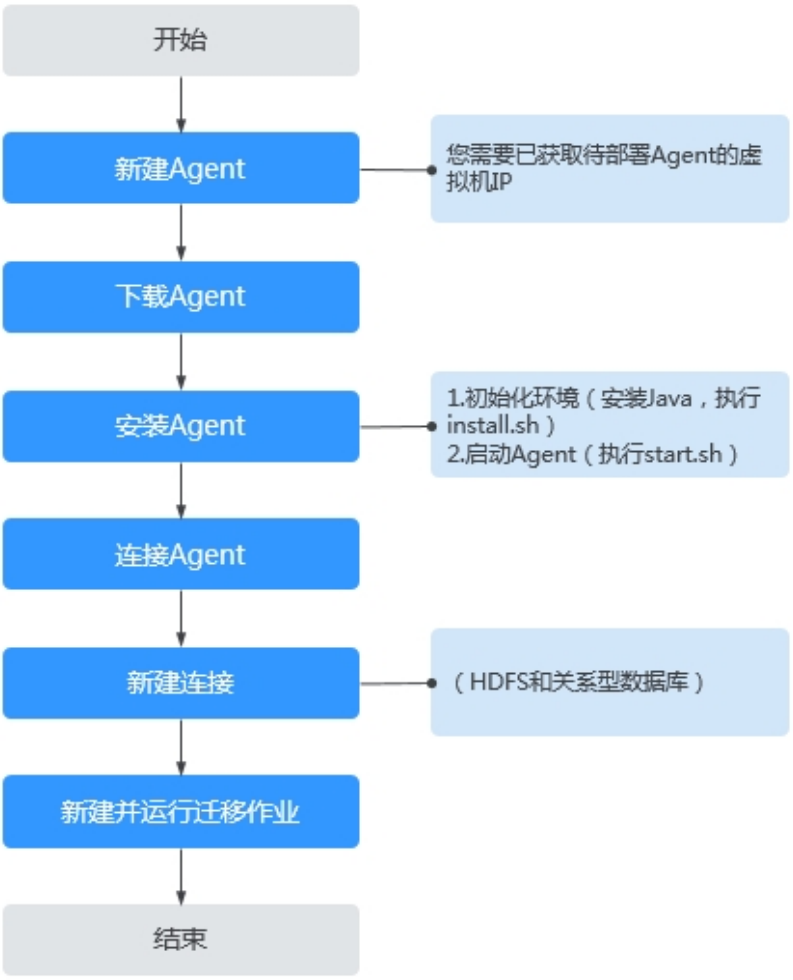
对于HDFS和关系型数据库类型的数据源，不方便暴露节点的场景，可选择在源端网络中部署Agent。CDM通过Agent拉取客户内部数据源的数据，但不支持写入数据。

图 4-4 场景图



Agent的使用流程如图4-5所示。

图 4-5 Agent 使用流程



前提条件

- 已具备CDM集群。
- 已具备Linux主机（例如Linux版本ECS云服务器）。该Linux主机对vCPUs、内存、磁盘等规格无特殊要求，但须满足以下条件：
  - 需要已安装64位版本java 8并配置java环境变量。
  - 授予Ruby用户（若无Ruby用户则需手动创建）在/tmp目录下的写权限。

新建 Agent

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

图 4-6 配置 Agent

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

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

----结束

安装并启动 Agent

步骤1 在Agent管理页面，找到已成功创建的Agent。如图4-7所示，下载Agent。

图 4-7 下载 Agent

步骤2 将下载的Agent压缩包，上传至待部署Agent的Linux主机上。

说明

该Linux主机对vCPUs、内存、磁盘等规格无特殊要求，但须满足以下条件：

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

步骤3 解压安装包后执行如下命令安装Agent。

sh sbin/install.sh

步骤4 如果需要通过Agent连接关系数据库，则需要将对应的驱动（参考**管理驱动**获取）上传至Agent安装目录下的/server/jdbc，并修改在同目录下properties文件里对应数据库驱动的版本号。

步骤5 以root用户执行如下命令，在/server/jdbc路径将新上传驱动的所有者和所属组修改为Ruby。

chown Ruby.Ruby * -R

步骤6 安装完成后，执行如下命令启动Agent。

su Ruby

sh sbin/start.sh

步骤7 执行如下命令检查Agent进程是否启动。

ps -ef | grep cdm

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

----结束

连接 Agent

步骤1 在Agent管理页面，找到已成功创建的Agent。如图4-8所示，连接Agent。

图 4-8 连接 Agent

步骤2 Agent连接成功后，即可在创建连接中选择Agent。

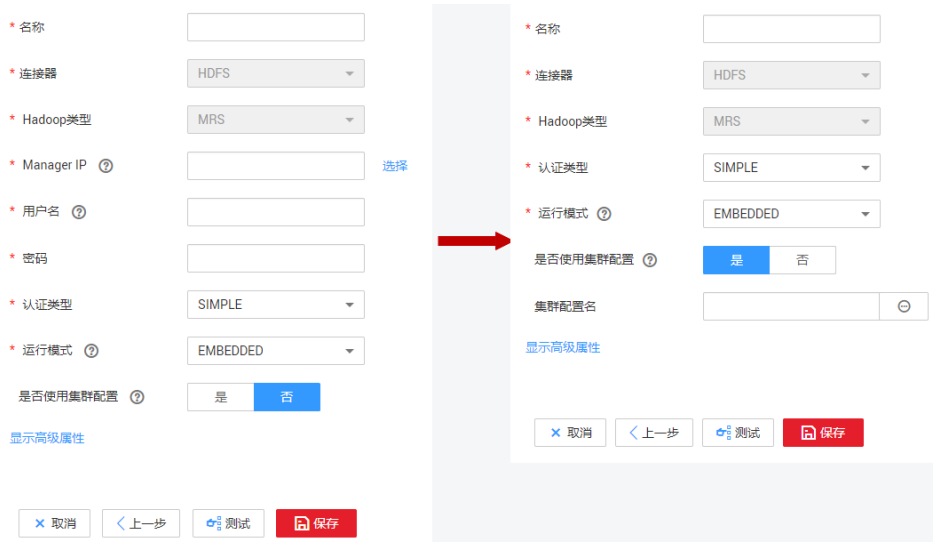
----结束

4.4 管理集群配置

集群配置管理支持新建、编辑或删除Hadoop集群配置。

Hadoop集群配置主要用于新建Hadoop类型连接时，能够简化复杂的连接参数配置，如图4-9所示。

图 4-9 使用集群配置前后对比



CDM支持的Hadoop类型连接主要包括以下几类：

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

操作场景

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

前提条件

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

获取集群配置文件和 Keytab 文件

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

表 4-3 集群配置文件和 Keytab 文件获取方式

Hadoop类型连接	集群配置文件获取方式	Keytab文件获取方式
MRS集群 ● MRS HDFS ● MRS HBase ● MRS Hive ● MRS Hudi ● MRS ClickHouse	针对MRS 3.x版本集群： 1. 登录FusionInsight Manager。 2. 选择“集群 > > 待操作的集群名称 > 概览 > 更多 > 下载客户端”，界面显示“下载集群客户端”对话框。 3. 对话框中选择“仅配置文件”，平台类型和服务端保持一致，其他保持默认即可，单击确认后进行本地下载。 4. 获取下载的tar包，此即为FusionInsight集群配置文件。 针对MRS 2.x及之前版本集群： 1. 登录MRS管理控制台。 2. 选择“集群列表 > 现有集群”，单击集群名称进入集群详情页面，单击“组件管理”。 3. 单击“下载客户端”。“客户端类型”选择“仅配置文件”，“下载路径”选择“服务器端”或“远端主机”，自定义文件保存路径后，单击“确定”开始生成客户端配置文件。 4. 将生成的配置文件，保存到本地路径。 具体可参见MapReduce服务文档。	针对MRS 3.x版本集群： 1. 登录FusionInsight Manager。 2. 通过“系统 > 权限 > 用户”，选择所需用户所在行，单击“更多 > 下载认证凭据”下载认证凭据文件。 3. 获取下载的tar包，此即为FusionInsight集群Keytab文件。 针对MRS 2.x及之前版本集群： 1. 登录MRS服务的Manager，单击“系统设置”。在“权限配置”区域，单击“用户管理”。 2. 在需导出keytab文件用户所在的行，选择“更多 > 下载认证凭据”下载认证文件，待文件自动生成后指定保存位置，并妥善保管该文件。 具体可参见MapReduce服务文档。

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

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

说明

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

操作步骤

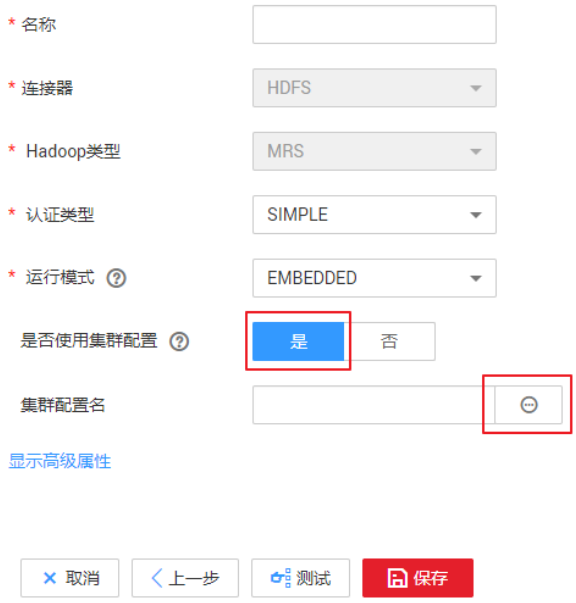
1. 进入CDM主界面，进入集群管理界面。选择CDM集群后的“作业管理 > 连接管理 > 集群配置管理”。
2. 在集群配置管理界面，选择“新建集群配置”，配置参数填写如下：

   图 4-10 新建集群配置

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

图 4-11 使用集群配置



4.5 配置 OBS 连接

OBS连接目的端OBS桶需添加读写权限，并在连接时不需要认证文件。

说明

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

连接OBS时，相关连接参数如表4-4所示。

表 4-4 OBS 连接的参数

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

参数名	说明	取值样例
端口	数据传输协议端口，https是443，http是80。	443
OBS桶类型	用户下拉选择即可，一般选择为“对象存储”。	对象存储
访问标识(AK)	AK和SK分别为登录OBS服务器的访问标识与密钥。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。 您可以通过如下方式获取访问密钥。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图4-12所示。 **图 4-12** 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)		-

4.6 配置 PostgreSQL/SQLServer 连接

连接PostgreSQL/SQLServer时，相关参数如表4-5所示，金仓和GaussDB数据源可通过PostgreSQL连接器进行连接，支持的迁移作业的源端、目的端情况与PostgreSQL数据源一致。

说明

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

表 4-5 PostgreSQL/SQLServer 连接参数

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

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”，可获取用户的实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	不同的数据库端口不同，请根据具体情况配置。例如： SQLServer默认端口：1433 PostgreSQL默认端口：5432
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**管理Agent**中已创建的Agent。	-
驱动版本	不同类型的关系数据库，需要适配不同的驱动，更多详情请参见**如何获取驱动**。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 常见配置举例如下： • **connectTimeout=60与socketTimeout=300**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位s），避免超时导致失败。 • **useCursorFetch=false**：CDM作业默认打开了JDBC连接器与关系型数据库通信使用二进制协议开关，即useCursorFetch=true。部分第三方可能存在兼容问题导致迁移时间转换出错，可以关闭此开关。 • **trustServerCertificate=true**：在创建安全连接的时候可能会报PKIX错误，建议设置为true。	sslmode=require

参数名	说明	取值样例
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。	"

4.7 配置数据仓库服务（DWS）连接

连接数据仓库服务（DWS）时，相关参数如表4-6所示。

说明

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

表 4-6 数据仓库服务（DWS）连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	dws_link
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	不同的数据库端口不同，请根据具体情况配置。
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择管理Agent中已创建的Agent。	-
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	"
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000

参数名	说明	取值样例
SSL加密	可选参数，支持通过SSL加密方式连接数据库，暂不支持自建的数据库。	是 **说明** 启用SSL加密需确保DWS本身已启用SSL加密。
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 常见配置举例如下： • **connectTimeout=60与socketTimeout=300**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位s），避免超时导致失败。 • **useCursorFetch=false**：CDM作业默认打开了JDBC连接器与关系型数据库通信使用二进制协议开关，即useCursorFetch=true。部分第三方可能存在兼容问题导致迁移时间转换出错，可以关闭此开关；开源MySQL数据库支持useCursorFetch参数，无需对此参数进行设置。	sslmode=require **说明** 启用SSL加密后sslmode值不设置可能会导致连接失败。

4.8 配置云数据库 MySQL/MySQL 数据库连接

连接MySQL数据库连接时，相关参数如表4-7所示。

说明

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

表 4-7 MySQL 数据库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mysql_link
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的MySQL数据库实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	3306
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm

参数名	说明	取值样例
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**管理Agent**中已创建的Agent。	-
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
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。	-
SSL加密	可选参数，支持通过SSL加密方式连接数据库，暂不支持自建的数据库。	是

参数名	说明	取值样例
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 常见配置举例如下： • **connectTimeout=600000**与**socketTimeout=300000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。 • **tinyInt1isBit=false**或**mysql.bool.type.transform=false**：MySQL默认开启配置tinyInt1isBit=true，将TINYINT(1)当作BIT也就是Types.BOOLEAN来处理，会将1或0读取为true或false从而导致迁移失败，此时可关闭配置避免迁移报错。 • **useCursorFetch=false**：CDM作业默认打开了JDBC连接器与关系型数据库通信使用二进制协议开关，即useCursorFetch=true。部分第三方可能存在兼容问题导致迁移时间转换出错，可以关闭此开关；开源MySQL数据库支持useCursorFetch参数，无需对此参数进行设置。 • **allowPublicKeyRetrieval=true**：MySQL默认关闭允许公钥检索机制，因此连接MySQL数据源时，如果TLS不可用、使用RSA公钥加密时，可能导致连接报错。此时可打开公钥检索机制，避免连接报错。	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

4.9 配置 Oracle 数据库连接

连接Oracle数据库时，连接参数如表4-8所示。

说明

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

表 4-8 Oracle 数据库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	oracle_link
数据库服务器	配置为要连接的数据库的IP地址或域名。	192.168.0.1
端口	配置为要连接的数据库的端口。	默认端口：1521
数据库连接类型	选择Oracle数据库连接类型： ● Service Name：通过SERVICE_NAME连接Oracle数据库。 ● SID：通过SID连接Oracle数据库。	SID
实例名称	配置Oracle实例ID，用于实例区分各个数据库。“数据库连接类型”选择“SID”时才有该参数。	dbname
数据库名称	配置为要连接的数据库名称。“数据库连接类型”选择“Service Name”时才有该参数。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**管理Agent**中已创建的Agent。	-
Oracle版本	创建Oracle连接时才有该参数，根据您Oracle数据库的版本来选择。当出现“java.sql.SQLException: Protocol violation异常”时，可以尝试更换版本号。	高于12.1
一次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。 Oracle到DWS迁移时，可能出现目的端写太久导致迁移超时的情况。此时请减少Oracle源端“一次请求行数”参数值的设置。	1000
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 常见配置举例如下： ● **oracle.net.CONNECT_TIMEOUT=60000**与**oracle.jdbc.ReadTimeout=300000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与读取超时时间（单位ms），避免超时导致失败。	sslmode=require

参数名	说明	取值样例
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	"
驱动版本	不同类型的关系数据库，需要适配不同的驱动，更多详情请参见**如何获取驱动**。	-

4.10 配置 DLI 连接

连接数据湖探索（DLI）服务时，相关参数如**表4-9**所示。

说明

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

表 4-9 DLI 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	dli_link
访问标识(AK)	访问DLI数据库时鉴权所需的AK和SK。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如**图4-13**所示。 **图 4-13** 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)		-

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

4.11 配置 Hive 连接

目前CDM支持连接的Hive数据源有以下几种：

- MRS Hive
- FusionInsight Hive
- Apache Hive

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

MRS Hive

用户具有MRS Hive连接的表的访问权限时，才能在字段映射时看到表。

MRS Hive连接适用于华为云上的MapReduce服务。MRS Hive的连接参数如表4-10所示。

说明

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

表 4-10 MRS Hive 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hivelink
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。根据服务端Hive版本设置。	HIVE_3_X

参数名	说明	取值样例
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。 **说明** ● 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对MRS组件的库、表、列进行操作，还需要参考MRS文档添加对应组件的库、表、列操作权限。 ● 如果CDM集群为2.9.0之前的版本，或MRS集群为3.1.0之前的版本，则所创建的用户需要具备Manager_administrator或System_administrator权限，才能在CDM创建连接。 ● 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。	cdm
密码	访问MRS Manager的用户密码。	-
开启LDAP认证	通过代理连接的时候，此项可配置。 当MRS Hive对接外部LDAP开启了LDAP认证时，连接Hive时需要使用LDAP账号与密码进行认证，此时必须开启此参数，否则会连接失败。	否
LDAP用户名	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的用户名。	-
LDAP密码	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否

参数名	说明	取值样例
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图4-14所示。 **图 4-14** 单击新增访问密钥  3. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id和Secret Access Key）。 **说明** • 每个用户仅允许新增两个访问密钥。 • 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台界面获取。请在生成后妥善保管。	- -
运行模式	“HIVE_3_X”版本支持该参数。支持以下模式： • EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 • STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式。 **说明** STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	EMBEDDED
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

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

常见配置举例如下：

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

FusionInsight Hive

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

FusionInsight Hive的连接参数如**表4-11**所示。

表 4-11 FusionInsight Hive 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hivelink
Manager IP	FusionInsight Manager平台的地址。	127.0.0.1
Manager端口	FusionInsight Manager平台的端口。	28443
CAS Server端口	与FusionInsight对接的CAS Server的端口。	20009
认证类型	访问集群的认证类型： • SIMPLE：非安全模式选择Simple鉴权。 • KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。	HIVE_3_X

参数名	说明	取值样例
用户名	登录FusionInsight Manager平台的用户名。	cdm
密码	FusionInsight Manager平台的密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图4-15所示。 **图 4-15** 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

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

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

常见配置举例如下：

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

Apache Hive

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

Apache Hive的连接参数如**表4-12**所示。

表 4-12 Apache Hive 连接参数

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

参数名	说明	取值样例
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图4-16所示。 **图 4-16** 单击新增访问密钥  3. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id和Secret Access Key）。 **说明** • 每个用户仅允许新增两个访问密钥。 • 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台界面获取。请在生成后妥善保管。	-
密钥(SK)		-
运行模式	“HIVE_3_X”版本支持该参数。支持以下模式： • EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 • STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式。 **说明** STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	EMBEDDED
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否

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

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

常见配置举例如下：

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

4.12 配置 HBase 连接

目前CDM支持连接的HBase数据源有以下几种：

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

说明

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

MRS HBase

连接MRS上的HBase数据源时，相关参数如**表4-13**所示。

说明

- 新建MRS连接前，需在MRS中添加一个kerberos认证用户并登录MRS管理页面更新其初始密码，然后使用该新建用户创建MRS连接。
- 如需连接MRS 2.x版本的集群，请先创建2.x版本的CDM集群。CDM 1.8.x版本的集群无法连接MRS 2.x版本的集群。
- 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。网络互通需满足如下条件：
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP，MRS集群可以访问公网且防火墙规则已开放连接端口。
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**章节。
  - 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。

表 4-13 MRS 上的 HBase 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mrs_hbase_link
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，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。	-

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

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

FusionInsight HBase

连接FusionInsight HD上的HBase数据源时，相关参数如**表4-14**所示。

表 4-14 FusionInsight HBase 连接参数

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

参数名	说明	取值样例
CAS Server端口	与FusionInsight对接的CAS Server的端口。	20009
用户名	登录FusionInsight Manager平台的用户名。	cdm
密码	FusionInsight Manager平台的密码。	-
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	KERBEROS
HBase版本	HBase版本。	HBASE_2_X
运行模式	“HBASE_2_X”版本支持该参数。选择HBase连接的运行模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式。 **说明** STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	STANDALONE
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hbase_01

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

Apache HBase

连接Apache Hadoop上的HBase数据源时，相关参数如**表4-15**所示。

表 4-15 Apache HBase 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hadoop_hbase_link
ZK链接地址	HBase的Zookeeper链接地址。 格式：<host1>:<port>,<host2>:<port>,<host3>:<port>	zk1.example.com:2181,zk2.example.com:2181,zk3.example.com:2181
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	KERBEROS
IP与主机名映射	输入IP和主机名。 如果配置文件使用主机名，需要配置所有IP与主机的映射，多个主机之间使用空格进行分隔。	IP：10.3.6.9 主机名：hostname01
HBase版本	HBase版本。	HBASE_2_X
运行模式	“HBASE_2_X”版本支持该参数。选择HBase连接的运行模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式。 **说明** STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。	STANDALONE
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hbase_01

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

4.13 配置 HDFS 连接

目前CDM支持连接的HDFS数据源有以下几种：

- MRS HDFS
- FusionInsight HDFS
- Apache HDFS

说明

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

MRS HDFS

连接MRS上的HDFS数据源时，相关参数如表4-16所示。

说明

- 新建MRS连接前，需在MRS中添加一个kerberos认证用户并登录MRS管理页面更新其初始密码，然后使用该新建用户创建MRS连接。
- 如需连接MRS 2.x版本的集群，请先创建2.x版本的CDM集群。CDM 1.8.x版本的集群无法连接MRS 2.x版本的集群。
- 如果选择集群后连接失败，请检查MRS集群与作为Agent的CDM实例是否网络互通。网络互通需满足如下条件：
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群处于不同区域的情况下，需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP，MRS集群可以访问公网且防火墙规则已开放连接端口。
  - DataArts Studio实例（指DataArts Studio实例中的CDM集群）与MRS集群同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**章节。
  - 此外，还需确保该MRS集群与DataArts Studio工作空间所属的企业项目相同，如果不同，您需要修改工作空间的企业项目。

表 4-16 MRS 上的 HDFS 连接参数

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

参数名	说明	取值样例
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。 **说明** ● 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对MRS组件的库、表、列进行操作，还需要参考MRS文档添加对应组件的库、表、列操作权限。 ● 如果CDM集群为2.9.0之前的版本，或MRS集群为3.1.0之前的版本，则所创建的用户需要具备Manager_administrator或System_administrator权限，才能在CDM创建连接。 ● 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。	cdm
密码	访问MRS Manager的用户密码。	-
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE

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

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

FusionInsight HDFS

连接FusionInsight HD上的HDFS数据源时，相关参数如**表4-17**所示。

表 4-17 FusionInsight HDFS 连接参数

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

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

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

Apache HDFS

连接Apache Hadoop上的HDFS数据源时，相关参数如表4-18所示。

表 4-18 Apache HDFS 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	hadoop_hdfs_link
URI	表示NameNode URI地址。可以填写为：hdfs://***namenode实例的ip***:8020。	hdfs://***IP***:8020
认证类型	访问集群的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	KERBEROS
运行模式	选择HDFS连接的运行模式： ● EMBEDDED：连接实例与CDM运行在一起，该模式性能较好。 ● STANDALONE：连接实例运行在独立进程。如果CDM需要对接多个Hadoop数据源（MRS、Hadoop或CloudTable），并且既有KERBEROS认证模式又有SIMPLE认证模式，只能使用STANDALONE模式或者配置不同的Agent。 **说明**：STANDALONE模式主要是用来解决版本冲突问题的运行模式。当同一种数据连接的源端或者目的端连接器的版本不一致时，存在jar包冲突的情况，这时需要将源端或目的端放在STANDALONE进程里，防止冲突导致迁移失败。 ● Agent：连接实例运行在Agent上。对于Apache HDFS，仅当“认证类型”为“SIMPLE”时，才可以选择Agent运行模式。	STANDALONE
IP与主机名映射	运行模式选择“EMBEDDED”、“STANDALONE”时，该参数有效。 如果HDFS配置文件使用主机名，需要配置IP与主机的映射。格式：IP与主机名之间使用空格分隔，多对映射使用分号或回车换行分隔。	10.1.6.9 hostname01 10.2.7.9 hostname02
Agent	认证类型选择“SIMPLE”，并且运行模式选择“Agent”时配置，选择连接Agent中已创建的Agent。	-
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否

参数名	说明	取值样例
集群配置名	当“是否使用集群配置”为“是”或“认证类型”为“KERBEROS”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hdfs_01

4.14 配置 FTP/SFTP 连接

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

说明

- 当前仅支持Linux操作系统的FTP 服务器。
- 作业运行中禁止修改密码或者更换用户。在作业运行过程中修改密码或者更换用户，密码不会立即生效且作业会运行失败。

连接FTP或SFTP服务器时，连接参数相同，如**表4-19**所示。

表 4-19 FTP/SFTP 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	ftp_link
主机名或IP	FTP或SFTP服务器的IP地址或者主机名。	ftp.apache.org
端口	FTP或SFTP服务器的端口，FTP默认值为21；SFTP默认值为22。	21
用户名	登录FTP或SFTP服务器的用户名。	cdm
密码	登录FTP或SFTP服务器的密码。	-

4.15 配置 Redis 连接

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

Redis连接不支持SSL加密的Redis数据源。

连接本地Redis数据库时，相关参数如**表4-20**所示。

说明

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

表 4-20 Redis 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	redis_link
Redis部署方式	Redis部署方式： ● Single：表示单机部署。 ● Cluster：表示集群部署。 ● Proxy：表示通过代理部署。	Single
Redis服务器列表	Redis服务器地址列表，输入格式为“数据库服务器域名或IP地址：端口”。多个服务器列表间以“;”分隔。	192.168.0.1:7300;192.168.0.2:7301
密码	连接Redis的密码。	-
Redis数据库索引	Redis分库的索引标识。 Redis的分库，相当于关系型数据库中的database。分库总数可以在Redis配置文件中设置，默认是16个，分库名称是一个整数（0～15），不是一个字符串。	0
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。 **说明** ● 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对MRS组件的库、表、列进行操作，还需要参考MRS文档添加对应组件的库、表、列操作权限。 ● 如果CDM集群为2.9.0之前的版本，或MRS集群为3.1.0之前的版本，则所创建的用户需要具备Manager_administrator或System_administrator权限，才能在CDM创建连接。 ● 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。	cdm

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

4.16 配置 DDS 连接

DDS连接适用于华为云上的文档数据库服务，常用于从DDS同步数据到大数据平台。

连接云服务DDS时，相关参数如**表4-21**所示。

说明

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

表 4-21 DDS 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	dds_link
服务器列表	服务器地址列表，输入格式为“数据库服务器域名或IP地址：端口”。多个服务器列表间以“;”分隔。	192.168.0.1:7300;192.168.0.2:7301
数据库名称	要连接的DDS数据库名称。	DB_dds
用户名	连接DDS的用户名。	cdm
密码	连接DDS的密码。	-
直连模式	适用于主节点网络通，副本节点网络不通场景。 **说明** • 直连模式服务器列表只能配一个ip。 • 直连适用于主节点网络通，副本节点网络不通场景。	否

4.17 配置 CloudTable 连接

连接CloudTable时，相关参数如**表4-22**所示。

说明

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

表 4-22 CloudTable 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	cloudtable_link
ZK链接地址	可通过CloudTable服务的集群管理界面获取该参数值。	cloudtable-cdm-zk1.cloudtable.com:2181,cloudtable-cdm-zk2.cloudtable.com:2181
IAM统一身份认证	如果所需连接的CloudTable集群在创建时开启了“IAM统一身份认证”，该参数需设置为“是”，否则设置为“否”。 当选择IAM统一身份认证时，需要输入用户名、AK和SK。	否
用户名	登录CloudTable集群的用户名。	admin
AK	登录CloudTable集群的访问标识。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。	-
SK	登录CloudTable集群的密钥。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。	-
是否使用集群配置	您可以通过使用集群配置，简化Hadoop连接参数配置。	否
集群配置名	仅当“是否使用集群配置”为“是”时，此参数有效。此参数用于选择用户已经创建好的集群配置。 集群配置的创建方法请参见**管理集群配置**。	hadoop_01

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

4.18 配置 MongoDB 连接

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

连接本地MongoDB数据库时，相关参数如**表4-23**所示。

说明

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

表 4-23 MongoDB 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mongodb_link
服务器列表	MongoDB服务器地址列表，输入格式为“数据库服务器域名或IP地址：端口”。多个服务器列表间以“;”分隔。	192.168.0.1:7300;192.168.0.2:7301
数据库名称	要连接的MongoDB数据库名称。	DB_mongodb
用户名	连接MongoDB的用户名。	cdm
密码	连接MongoDB的密码。	-
直连模式	适用于主节点网络通，副本节点网络不通场景。 **说明** • 直连模式服务器列表只能配一个ip。 • 直连适用于主节点网络通，副本节点网络不通场景。	否
连接属性	自定义连接属性，支持MongoDB属性，单位为ms。连接属性如下： • socketTimeout，默认socketTimeout=60000 • maxWaitTime，默认maxWaitTime=10000 • connectTimeout，默认connectTimeout=10000 • serverSelectionTimeout，默认serverSelectionTimeout=5000	socketTimeout=60000

4.19 配置 Cassandra 连接

说明

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

表 4-24 Cassandra 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	mongodb_link
服务节点	一个或者多个节点的地址，以“;”分隔。建议同时配置多个节点。	192.168.0.1;192.168.0.2
端口	连接的Cassandra节点的端口号。	9042
用户名	连接Cassandra的用户名。	cdm

参数名	说明	取值样例
密码	连接Cassandra的密码。	-
连接超时时长	可选参数，单击“显示高级属性”后显示。 连接超时时长，单位秒。	5
读取超时时长	可选参数，单击“显示高级属性”后显示。 读取超时时长，单位秒。小于或等于0表示不超时。	12

4.20 配置 DIS 连接

连接DIS时，相关参数如表4-25所示。

说明

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

表 4-25 DIS 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	dis_link
区域	DIS所在的区域。	-
终端节点	待连接DIS的URL，URL一般格式为：https://Endpoint。 终端节点（Endpoint）即调用API的**请求地址**，不同服务不同区域的终端节点不同。本服务的Endpoint可从**终端节点Endpoint**获取。	-
访问标识(AK)	登录DIS服务器的访问标识。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。	-
密钥(SK)	登录DIS服务器的密钥。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。	-
项目ID	DIS的项目ID。	-

4.21 配置 Kafka 连接

MRS Kafka

连接MRS上的Kafka数据源时，相关参数如表4-26所示。

说明

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

表 4-26 MRS Kafka 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	kafka_link
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
用户名	需要配置MRS Manager的用户名和密码。 如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。 **说明** • 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对MRS组件的库、表、列进行操作，还需要参考MRS文档添加对应组件的库、表、列操作权限。 • 如果CDM集群为2.9.0之前的版本，或MRS集群为3.1.0之前的版本，则所创建的用户需要具备Manager_administrator或System_administrator权限，才能在CDM创建连接。 • 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。	-
密码	访问MRS Manager的用户密码。	-
认证类型	访问MRS的认证类型： • SIMPLE：非安全模式选择Simple鉴权。 • KERBEROS：安全模式选择Kerberos鉴权。	是

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

Apache Kafka

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

连接Apache Hadoop上的Kafka数据源时，相关参数如表4-27所示。

表 4-27 Apache Kafka 连接参数

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

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

4.22 配置 DMS Kafka 连接

连接DMS的Kafka队列时，相关参数如表4-28所示。

说明

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

表 4-28 DMS Kafka 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	dms_link
服务类型	选择DMS Kafka版本，目前只有专享版。	专享版
Kafka Broker	Kafka专享版实例的地址，格式为 host:port。	-
Kafka SASL_SSL	选择是否打开客户端连接Kafka专享版实例时SSL认证的开关。当DMS Kafka实例的连接信息中启用的安全协议为“SASL_SSL”时需要开启。 开启Kafka SASL_SSL，则数据加密传输，安全性更高，但性能会下降。 **说明** 启用SSL认证后，Kafka会将Kafka Broker连接地址视做域名不断进行解析，导致性能消耗。建议修改CDM集群对应的ECS主机（通过集群IP查找对应的ECS主机）中的“/etc/hosts”文件，为其添加Broker连接地址的自映射，以便客户端能够快速解析实例的Broker。例如Kafka Broker地址配置为10.154.48.120时，hosts文件中的自映射配置为： `10.154.48.120 10.154.48.120`	是
用户名	开启Kafka SASL_SSL时显示该参数，表示连接DMS Kafka的用户名。	-

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

4.23 配置云搜索服务（CSS）连接

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

说明

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

连接云搜索服务(CSS)时，相关参数如**表4-29**所示。

表 4-29 云搜索服务(CSS)连接参数

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

参数名	说明	取值样例
https访问	CSS集群开启安全认证模式时显示此参数。该参数表示开启https访问，https访问相较于http访问更安全。	是

4.24 配置 Elasticsearch 连接

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

说明

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

连接Elasticsearch时，相关参数如表4-30所示。

表 4-30 Elasticsearch 连接参数

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

4.25 配置达梦数据库 DM 连接

连接达梦数据库 DM时，相关参数如表4-31所示。

说明

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

表 4-31 达梦数据库 DM 连接参数

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

参数名	说明	取值样例
端口	配置为要连接的数据库的端口。	不同的数据库端口不同，请根据具体情况配置。
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'

4.26 配置 SAP HANA 连接

连接SAP HANA时，相关参数如表4-32所示。

说明

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

表 4-32 SAP HANA 连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	sap_link
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	不同的数据库端口不同，请根据具体情况配置。
数据库名称	配置为要连接的数据库名称。	dbname

参数名	说明	取值样例
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**管理Agent**中已创建的Agent。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 常见配置举例如下： • **connectTimeout=360000**与**socketTimeout=360000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。 • **useCursorFetch=false**：CDM作业默认打开了JDBC连接器与关系型数据库通信使用二进制协议开关，即useCursorFetch=true。部分第三方可能存在兼容问题导致迁移时间转换出错，可以关闭此开关；开源MySQL数据库支持useCursorFetch参数，无需对此参数进行设置。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'

4.27 配置分库连接

分库指的是同时连接多个后端数据源，该连接可作为作业源端，将多个数据源的数据合一迁移到其他数据源上。连接参数如**表4-33**所示。

说明

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

表 4-33 分库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	my_link
用户名	待连接数据库的用户。 仅当“数据源列表”中某个后端数据库A未配置用户名密码时，该配置对A生效。如果后端数据库B已配置用户名密码，此处配置不对B生效。	cdm
密码	待连接数据库的用户密码。 仅当“数据源列表”中某个后端数据库A未配置用户名密码时，该配置对A生效。如果后端数据库B已配置用户名密码，此处配置不对B生效。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**管理Agent**中已创建的Agent。	-
后端数据源	输入后端数据库的类型，当前仅支持MYSQL。	MYSQL
数据源列表	输入后端数据库的IP、端口、数据库名称、账户名、密码，以“:”隔开。即ip:port:dbs:username:password，其中username:password可以不填，此时以“用户名”、“密码”配置为准。 如果此处有多个后端数据库，需要确保表结构一致，并使用“|”分隔数据源。如果密码包含“|”或者“:”，可使用“\”转义。 例如“192.168.3.0:3306:cdm|192.168.2.2:3306:cdm:user:password”表示，第一个后端数据库IP为192.168.3.0，端口为3306，数据库名称为cdm，账户名密码以“用户名”、“密码”处配置为准；第二个后端数据库IP为192.168.2.2，端口为3306，数据库名称为cdm，账户名为“user”、密码为“password”。	192.168.3.0:3306:cdm|192.168.2.2:3306:cdm:user:password
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。	sslmode=require
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'

4.28 配置 MRS Hudi 连接

连接MRS Hudi时，相关参数如表4-34所示。

说明

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

表 4-34 Hudi 连接参数

参数名	说明	取值样例
名称	连接名称。	Hudilink
Manager IP	MRS Manager的浮动IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。	127.0.0.1
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	KERBEROS
账号	登录MRS Manager的账号。	cdm
密码	登录MRS Manager的密码。	-
OBS支持	是否支持OBS存储，如果hudi表数据存储在OBS，需要打开此开关。	是

参数名	说明	取值样例
访问标识（AK）	“OBS支持”设置为“是”时，呈现此参数。 AK和SK分别为登录OBS服务器的访问标识与密钥。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。 您可以通过如下方式获取访问密钥。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图4-17所示。 **图 4-17** 单击新增访问密钥  3. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id和Secret Access Key）。 **说明** • 每个用户仅允许新增两个访问密钥。 • 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台界面获取。请在生成后妥善保管。	-
密钥（SK）		-
OBS测试路径	“OBS支持”设置为“是”时，呈现此参数。 请填写完整的文件路径，将调用元数据查询接口来校验路径的访问权限。 **说明** • 如果是对象存储，路径需要填写到对象级别，否则会报错404，例如：“obs://bucket/dir/test.txt”。 • 如果是并行文件系统，则可以只填写到目录级别。例如：“obs://bucket/dir”。	obs://bucket/dir/test.txt
属性配置	需要集成的表名，多个表名使用英文逗号“,”分开，请务必配置，不要有空格，默认无需配置。	-

4.29 配置 MRS ClickHouse 连接

连接MRS ClickHouse时，相关参数如表4-35所示。

说明

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

表 4-35 ClickHouse 连接参数

参数名	说明	取值样例
名称	连接名称。	cklink
数据库服务器	配置为要连接的数据库的IP地址或域名。 登录MRS ClickHouse数据源所在集群的Manager页面，选择“集群 > 服务 > ClickHouse > 实例”，查看ClickHouseServer所在的“业务IP”。	192.168.0.1
端口	配置为要连接的数据库的端口。 **说明** • 如果使用Server节点，开启“SSL加密”，配置默认端口。登录MRS ClickHouse数据源所在集群的Manager页面，选择“集群 > 服务 > ClickHouse > 实例”，配置ClickHouseServer的默认端口，非安全模式MRS集群配置“http_port”参数对应的端口，安全模式MRS集群配置“https_port”参数对应的端口。 • 如果使用Balancer节点，开启“SSL加密”，配置默认端口。登录MRS ClickHouse数据源所在集群的Manager页面，选择“集群 > 服务 > ClickHouse > 实例”，配置ClickHouseBalancer的默认端口，非安全模式MRS集群配置“lb_http_port”参数对应的端口，安全模式MRS集群配置“lb_https_port”参数对应的端口。 • 如果MRS ClickHouse是安全集群，则需配置为https默认端口。	8123
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
SSL加密	可选参数，支持通过SSL加密方式连接数据库，暂不支持自建的数据库。	否
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'

4.30 配置神通（ST）连接

连接神通（ST）数据库连接时，相关参数如表4-36所示。

说明

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

表 4-36 神通（ST）数据库连接参数

参数名	说明	取值样例
名称	连接的名称，根据连接的数据源类型，用户可自定义便于记忆、区分的连接名。	st_link
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的数据库实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	3306
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择**管理Agent**中已创建的Agent。	-
引用符号	可选参数，连接引用表名或列名时的分隔符号，参考对应数据库的产品文档进行配置。	'
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
单次请求行数	可选参数，单击“显示高级属性”后显示。 指定每次请求获取的行数，根据数据源端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
连接属性	可选参数，单击“添加”可增加多个指定数据源的JDBC连接器的属性，参考对应数据库的JDBC连接器说明文档进行配置。 常见配置举例如下： • **connectTimeout=360000**与**socketTimeout=360000**：迁移数据量较大、或通过查询语句检索全表时，会由于连接超时导致迁移失败。此时可自定义连接超时时间与socket超时时间（单位ms），避免超时导致失败。	sslmode=require

5 管理作业

5.1 新建表/文件迁移作业

操作场景

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

约束限制

- 记录脏数据功能依赖于OBS服务。
- 作业导入时，JSON文件大小不超过1MB。
- 单文件传输大小不超过1TB。
- 配置源端和目的端参数时，字段名不可包含&和%。

前提条件

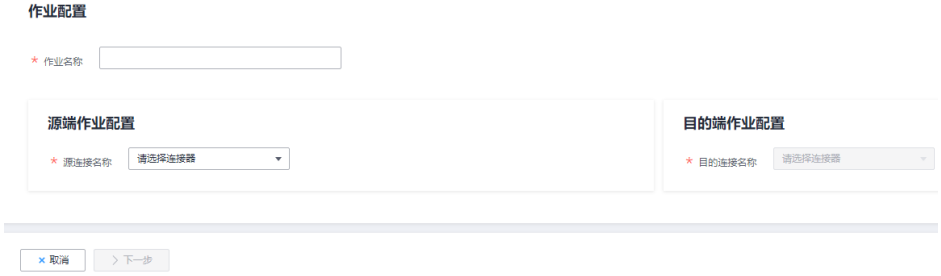
- 已新建连接，详情请参见**新建连接**。
- CDM集群与待迁移数据源可以正常通信。

操作步骤

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

步骤2 选择“表/文件迁移 > 新建作业”，进入作业配置界面。

图 5-1 新建表/文件迁移的作业

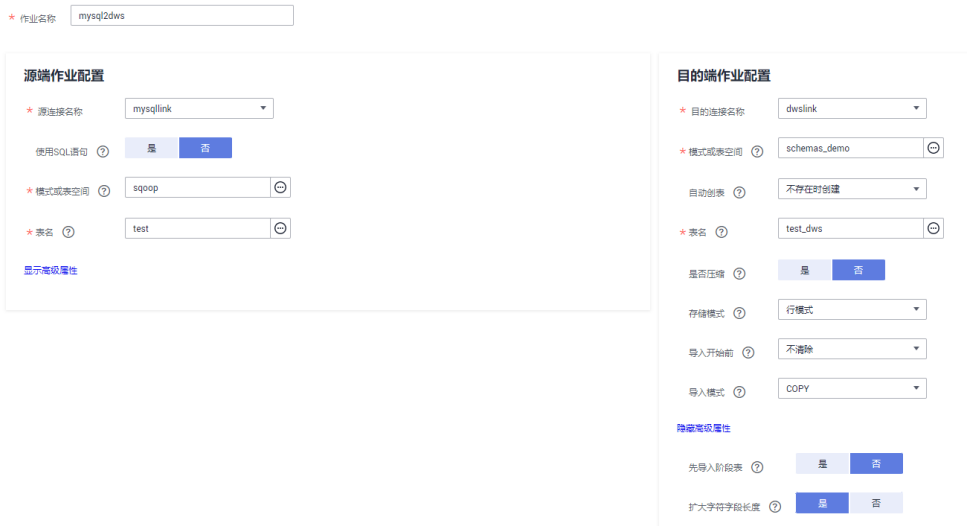


步骤3 选择源连接、目的连接：

- 作业名称：用户自定义任务名称，名称由中文、数字、字母、中划线、下划线、点号，且首字符不能是中划线或点号组成，长度必须在1到240个字符之间，例如“oracle2rds_t”。
- 源连接名称：选择待迁移数据的数据源，作业运行时将从此端复制导出数据。
- 目的连接名称：选择将数据迁移到哪个数据源，作业运行时会将数据导入此端。

步骤4 选择源连接后，配置作业参数，例如迁移MySQL到DWS时，如图5-2所示。

图 5-2 新建作业



每种数据源对应的作业参数不一样，其它类型数据源的作业参数请根据表5-1和表5-2选择。

表 5-1 源端作业参数说明

源端类型	说明	参数配置
OBS	支持以CSV、JSON或二进制格式抽取数据，其中二进制方式不解析文件内容，性能快，适合文件迁移。	参见配置OBS源端参数。
• MRS HDFS • FusionInsight HDFS • Apache HDFS	支持以CSV、Parquet或二进制格式抽取HDFS数据，支持多种压缩格式。	参见配置HDFS源端参数。
• MRS HBase • FusionInsight HBase • Apache HBase • CloudTable	支持从MRS、FusionInsight HD、开源Apache Hadoop的HBase，或CloudTable服务导出数据，用户需要知道HBase表的所有列族和字段名。	参见配置HBase/CloudTable源端参数。

源端类型	说明	参数配置
• MRS Hive • FusionInsight Hive • Apache Hive	支持从Hive导出数据，使用JDBC接口抽取数据。 Hive作为数据源，CDM自动使用Hive数据分片文件进行数据分区。	参见**配置Hive源端参数**。
DLI	支持从DLI导出数据。	参见**配置DLI源端参数**。
• FTP • SFTP	支持以CSV、JSON或二进制格式抽取FTP/SFTP的数据。	参见**配置FTP/SFTP源端参数**。
• HTTP	用于读取一个公网HTTP/HTTPS URL的文件，包括第三方对象存储的公共读取场景和网盘场景。 当前只支持从HTTP URL导出数据，不支持导入。	参见**配置HTTP源端参数**。
数据仓库 DWS	支持从数据仓库 DWS导出数据。	参见**配置DWS源端参数**。
SAP HANA	支持从SAP HANA导出数据。	参见**配置SAP HANA源端参数**。
• 云数据库 PostgreSQL • 云数据库 SQL Server • Microsoft SQL Server • PostgreSQL	支持从云端的数据库服务导出数据。 这些非云服务的数据库，既可以是用户在本地数据中心自建的数据库，也可以是用户在ECS上部署的，还可以是第三方云上的数据库服务。	从这些数据源导出数据时，CDM使用JDBC接口抽取数据，源端作业参数相同，详细请参见**配置PostgreSQL/SQL Server源端参数**。
MySQL	支持从MySQL导出数据。	参见**配置MySQL源端参数**。
Oracle	支持从Oracle导出数据。	参见**配置Oracle源端参数**。
分库	支持从分库导出数据。	参见**配置分库源端参数**。
• MongoDB • 文档数据库服务（DDS）	支持从MongoDB或DDS导出数据。	参见**配置MongoDB/DDS源端参数**。
Redis	支持从开源Redis导出数据。	参见**配置Redis源端参数**。
数据接入服务（DIS）	仅支持导出数据到云搜索服务。	参见**配置DIS源端参数**。

源端类型	说明	参数配置
• Apache Kafka • DMS Kafka • MRS Kafka	仅支持导出数据到云搜索服务。	参见**配置Kafka/DMS Kafka源端参数**。
• 云搜索服务 • Elasticsearch	支持从云搜索服务或Elasticsearch导出数据。	参见**配置Elasticsearch/云搜索服务源端参数**。
MRS Hudi	支持从MRS Hudi导出数据。	参见**配置MRS Hudi源端参数**。
MRS ClickHouse	支持从MRS ClickHouse导出数据。	参见**配置MRS ClickHouse源端参数**。
神通（ST）	支持从神通（ST）导出数据。	参见**配置神通（ST）源端参数**。
达梦数据库 DM	支持从达梦数据库 DM导出数据。	参见**配置达梦数据库DM源端参数**。

步骤5 配置目的端作业参数，根据目的端数据类型配置对应的参数，具体如**表5-2**所示。

表 5-2 目的端作业参数说明

目的端类型	说明	参数配置
OBS	支持使用CSV或二进制格式批量传输大量文件到OBS。	参见**配置OBS目的端参数**。
MRS HDFS	导入数据到HDFS时，支持设置压缩格式。	参见**配置HDFS目的端参数**。
MRS HBase CloudTable	支持导入数据到HBase，创建新HBase表时支持设置压缩算法。	参见**配置HBase/CloudTable目的端参数**。
MRS Hive	支持快速导入数据到MRS的Hive。	参见**配置Hive目的端参数**。
• MySQL • SQL Server • PostgreSQL	支持导入数据到云端的数据库服务。	使用JDBC接口导入数据，参见**配置MySQL/SQL Server/PostgreSQL目的端参数**。
DWS	支持导入数据到数据仓库DWS。	参见**配置DWS目的端参数**。
Oracle	支持导入数据到Oracle。	参见**配置Oracle目的端参数**。

目的端类型	说明	参数配置
数据湖探索（DLI）	支持导入数据到DLI服务。	参见**配置DLI目的端参数**。
Elasticsearch或云搜索服务	支持导入数据到云搜索服务。	参见**配置Elasticsearch/云搜索服务（CSS）目的端参数**。
MRS Hudi	支持快速导入数据到MRS的Hudi。	参见**配置MRS Hudi目的端参数**。
MRS Clickhouse	支持快速导入数据到MRS的Clickhouse。	参见**配置MRS ClickHouse目的端参数**。
MongoDB	支持快速导入数据到MongoDB。	参见**配置MongoDB目的端参数**。

步骤6 作业参数配置完成后，单击“下一步”进入字段映射的操作页面。

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

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

图 5-3 字段映射

说明

- 如果字段映射关系不正确，用户可以通过拖拽字段来调整映射关系。
- 如果在字段映射界面，CDM通过获取样值的方式无法获得所有列（例如从HBase/CloudTable/MongoDB导出数据时，CDM有较大概率无法获得所有列，以及SFTP/FTP迁移数据到DLI的链路场景），则可以单击后选择“添加新字段”来手动增加，确保导入到目的端的数据完整。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- 支持通过字段映射界面的，可自定义添加常量、变量及表达式。
- 当作业源端为OBS、迁移CSV文件时，并且配置“解析首行为列名”参数的场景下显示列名。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datatime）。
- Hive作为源端数据源时，支持array、map类型的数据读取。
- 当使用二进制格式进行文件到文件的迁移时，没有字段映射这一步。
- 如果是导入到数据仓库服务（DWS），则还需在目的字段中选择分布列，建议按如下顺序选取分布列：
  1. 有主键可以使用主键作为分布列。
  2. 多个数据段联合做主键的场景，建议设置所有主键作为分布列。
  3. 在没有主键的场景下，如果没有选择分布列，DWS会默认第一列作为分布列，可能会有数据倾斜风险。

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

图 5-4 新建转换器

CDM支持以下转换器：

- 脱敏：隐藏字符串中的关键数据。
  例如要将“12345678910”转换为“123****8910”，则参数配置如下：
  - “起始保留长度”为“3”。
  - “结尾保留长度”为“4”。
  - “替换字符”为“*”。
- 去前后空格：自动删除字符串前后的空值。

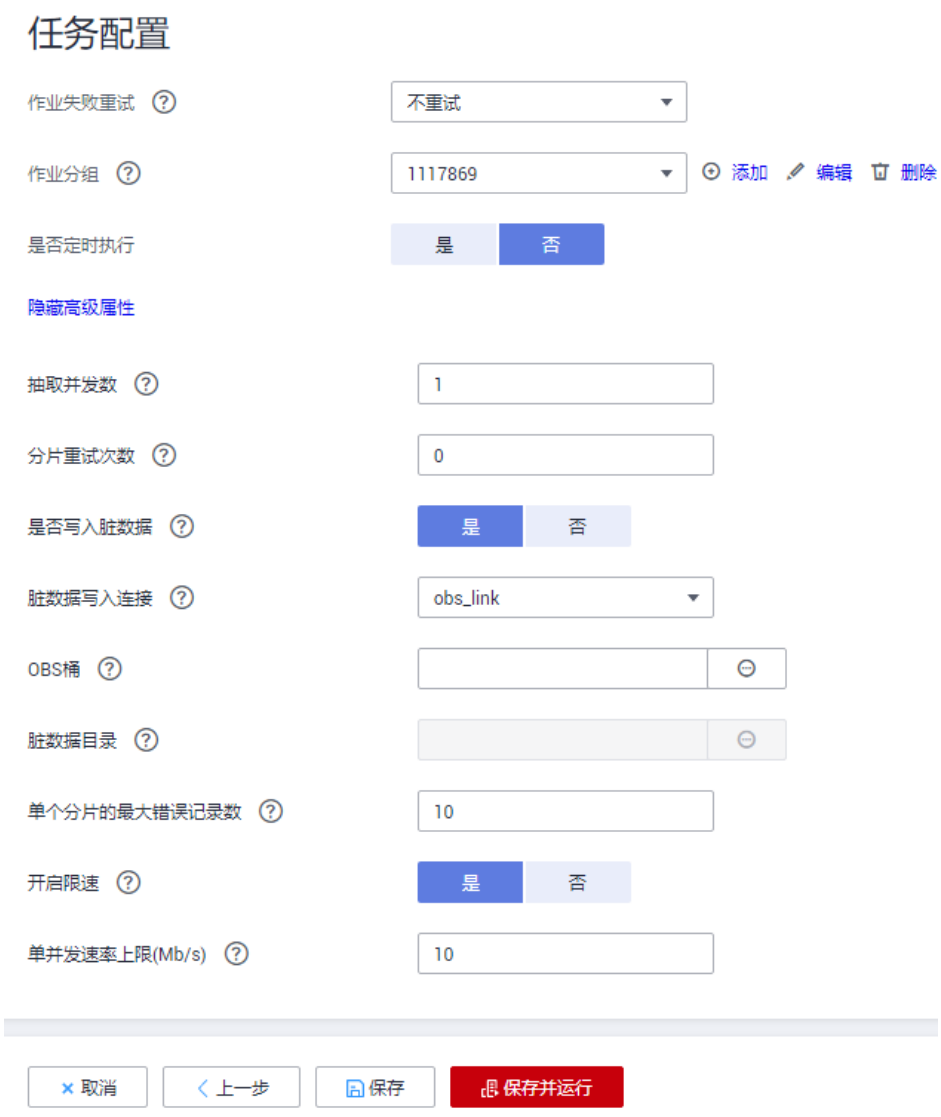
- 字符串反转：自动反转字符串，例如将“ABC”转换为“CBA”。
- 字符串替换：将选定的字符串替换。
- 表达式转换：使用JSP表达式语言（Expression Language）对当前字段或整行数据进行转换，详细请参见**字段转换**。
- 去换行：将字段中的换行符（\n、\r、\r\n）删除。

说明

作业源端开启“使用SQL语句”参数时不支持配置转换器。

步骤8 单击“下一步”配置任务参数，单击“显示高级属性”展开可选参数。

图 5-5 任务参数



各参数说明如**表5-3**所示。

表 5-3 任务配置参数

参数	说明	取值样例
作业失败重试	如果作业执行失败，可选择自动重试三次或者不重试。 建议仅对文件类作业或启用了导入阶段表的数据库作业配置自动重试，避免自动重试重复写入数据导致数据不一致。 **说明** 如果通过DataArts Studio数据开发使用参数传递并调度CDM迁移作业时，不能在CDM迁移作业中配置“作业失败重试”参数，如有需要请在数据开发中的CDM节点配置“失败重试”参数。	不重试
作业分组	选择作业的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。	DEFAULT
是否定时执行	如果选择“是”，可以配置作业自动启动的时间、重复周期和有效期，具体请参见**配置定时任务**。 **说明** 如果通过DataArts Studio数据开发调度CDM迁移作业，此处也配置了定时任务，则两种调度均会生效。为了业务运行逻辑统一和避免调度冲突，推荐您启用数据开发调度即可，无需配置CDM定时任务。	否

参数	说明	取值样例
抽取并发数	配置作业抽取并发数，控制将CDM作业拆分为多少Task。 CDM通过数据迁移作业，将源端数据迁移到目的端数据源中。其中，主要运行逻辑如下： 1. 数据迁移作业提交运行后，CDM会根据作业配置中的“抽取并发数”参数，将每个作业拆分为多个Task，即作业分片。 **说明** 不同源端数据源的作业分片维度有所不同，因此某些作业可能出现未严格按作业“抽取并发数”参数分片的情况。 2. CDM依次将Task提交给运行池运行。根据集群配置管理中的“最大抽取并发数”参数，超出规格的Task排队等待运行。 因此作业抽取并发数和集群最大抽取并发数参数设置为适当的值可以有效提升迁移速度。 作业抽取并发数的配置原则如下： 1. 迁移的目的端为文件时，CDM不支持多并发，此时应配置为单进程抽取数据。 2. 表中每行数据大小为1MB以下的可以设置多并发抽取，超过1MB的建议单线程抽取数据。 3. 作业抽取并发数可参考集群最大抽取并发数配置，但不建议超过集群最大抽取并发数上限。 4. 目的端为DLI数据源时，抽取并发数建议配置为1，否则可能会导致写入失败。 其中，集群最大抽取并发数的设置与CDM集群规格有关，并发数上限建议配置为vCPU核数*2。例如8核16GB规格集群的最大抽取并发数上限为16。	1
加载（写入）并发数	加载（写入）时并发执行的Loader数量。 仅当HBase或Hive作为目的数据源时该参数才显示。	3
分片重试次数	每个分片执行失败时的重试次数，为0表示不重试。	0
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据。 CDM中脏数据指的是数据格式非法的数据。当源数据中存在脏数据时，建议您打开此配置。否则可能导致迁移作业失败。 **说明** 脏数据当前仅支持写入到OBS桶路径中。因此仅当已具备OBS连接时，此参数才可以配置。	是

参数	说明	取值样例
脏数据写入连接	当“是否写入脏数据”为“是”才显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	obs_link
OBS桶	当“脏数据写入连接”为OBS类型的连接时，才显示该参数。 写入脏数据的OBS桶的名称。	dirtydata
脏数据目录	“是否写入脏数据”选择为“是”时，该参数才显示。 OBS上存储脏数据的目录，只有在配置了脏数据目录的情况下才会记录脏数据。 用户可以进入脏数据目录，查看作业执行过程中处理失败的数据或者被清洗过滤掉的数据，针对该数据可以查看源数据中哪些数据不符合转换、清洗规则。	/user/dirtydir
单个分片的最大错误记录数	当“是否写入脏数据”为“是”才显示该参数。 单个map的错误记录超过设置的最大错误记录数则任务自动结束，已经导入的数据不支持回退。推荐使用临时表作为导入的目标表，待导入成功后再改名或合并到最终数据表。	0
开启限速	设置限速可以保护源端读取压力，速率代表CDM传输速率，而非网卡流量。 **说明** • 支持对非二进制文件迁移的作业进行单并发限速。 • 如果作业配置多并发则实际限制速率需要乘以并发数。 • 文件到文件的二进制传输不支持限速功能。	是
单并发速率上限(Mb/s)	开启限速情况下设置的单并发速率上限值，如果配置多并发则实际速率限制需要乘以并发数。 **说明** 限制速率为大于1的整数。	20
中间队列缓存大小(MB)	数据写入时中间队列缓存大小，取值范围为1-500，默认值为64。 如果单行数据超过该值，可能会导致迁移失败。如果该值设置过大时，可能会影响集群正常运行。请酌情设置，无特殊场景请使用默认值。例如：64	64

步骤9 单击“保存”，或者“保存并运行”回到作业管理界面，可查看作业状态。

说明

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

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

----结束

5.2 新建整库迁移作业

操作场景

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

说明

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

支持整库迁移的数据源请参见**支持的数据源**。

约束限制

配置源端和目的端参数时，字段名不可包含&和%。

前提条件

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

操作步骤

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

步骤2 选择“整库迁移 > 新建作业”，进入作业参数配置界面。

图 5-6 创建整库迁移作业

步骤3 配置源端作业参数，根据待迁移的数据库类型配置对应参数，如表5-4所示。

表 5-4 源端作业参数

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

源端数据库类型	源端参数	参数说明	取值样例
DDS	数据库名称	待迁移的数据库名称，源连接中配置的用户需要拥有读取该数据库的权限。	ddsdb
	查询筛选	创建用于匹配文档的筛选器。 例如：{HTTPStatusCode:{$gt:"400",$lt:"500"},HTTPMethod:"GET"}。	-

步骤4 配置目的端作业参数，根据待导入数据的云服务配置对应参数，如表5-5所示。

表 5-5 目的端作业参数

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

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

步骤6 单击“下一步”配置任务参数。

图 5-7 任务参数

各参数说明如表5-6所示。

表 5-6 任务配置参数

参数	说明	取值样例
同时执行的表个数	抽取时并发执行的表的数量。	3
抽取并发数	设置同时执行的抽取任务数，一般保持默认即可。	1
是否写入脏数据	选择是否记录脏数据，默认不记录脏数据。	是
脏数据写入连接	当“是否写入脏数据”为“是”才显示该参数。 脏数据要写入的连接，目前只支持写入到OBS连接。	obs_link
OBS桶	当“脏数据写入连接”为OBS类型的连接时，才显示该参数。 写入脏数据的OBS桶的名称。	dirtydata
脏数据目录	“是否写入脏数据”选择为“是”时，该参数才显示。 OBS上存储脏数据的目录，只有在配置了脏数据目录的情况下才会记录脏数据。 用户可以进入脏数据目录，查看作业执行过程中处理失败的数据或者被清洗过滤掉的数据，针对该数据可以查看源数据中哪些数据不符合转换、清洗规则。	/user/dirtydir

参数	说明	取值样例
单个分片的最大错误记录数	当“是否写入脏数据”为“是”才显示该参数。 单个map的错误记录超过设置的最大错误记录数则任务自动结束，已经导入的数据不支持回退。推荐使用临时表作为导入的目标表，待导入成功后再改名或合并到最终数据表。	0

步骤7 单击“保存”，或者“保存并运行”。

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

----结束

说明

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

5.3 配置作业源端参数

5.3.1 配置 OBS 源端参数

作业中源连接为**OBS连接**时，源端作业参数如**表5-7**所示。

高级属性里的参数为可选参数，默认隐藏，单击界面上的“显示高级属性”后显示。

表 5-7 源端为 OBS 时的作业参数

参数类型	参数名	说明	取值样例
基本参数	桶名	待迁移数据所在的桶名。	BUCKET_2

参数类型	参数名	说明	取值样例
	源目录或文件	“列表文件”选择为“否”时，才有该参数。 待迁移数据的目录或单个文件路径。文件路径支持输入多个文件（最多50个），默认以“|”分隔，也可以自定义文件分隔符，具体请参见**文件列表迁移**。 待迁移数据的目录，将迁移目录下的所有文件（包括所有嵌套子目录及其子文件）。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	FROM/example.csv
	文件格式	指CDM以哪种格式解析数据，可选择以下格式： ● CSV格式：以CSV格式解析源文件，用于迁移文件到数据表的场景。 ● 二进制格式：选择“二进制格式”时不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。 ● JSON格式：以JSON格式解析源文件，一般都是用于迁移文件到数据表的场景。	CSV格式
	列表文件	当“文件格式”选择为“二进制格式”时，才有该参数。 打开列表文件功能时，支持读取OBS桶中文件（如txt文件）的内容作为待迁移文件的列表。该文件中的内容应为待迁移文件的绝对路径（不支持目录），例如直接写为如下内容： /052101/DAY20211110.data /052101/DAY20211111.data	是
	列表文件源连接	当“列表文件”选择为“是”时，才有该参数。可选择列表文件所在的OBS连接。	OBS_test_link

参数类型	参数名	说明	取值样例
	列表文件OBS桶	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶名。	01
	列表文件或目录	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶中的绝对路径或目录。 此处建议选择为文件的绝对路径。当选择为目录时，也支持迁移子目录中的文件，但如果目录下文件量过大，可能会导致集群内存不足。	/0521/Lists.txt
	JSON类型	当“文件格式”选择为“JSON格式”时，才有该参数。JSON文件中存储的JSON对象的类型，可以选择“JSON对象”或“JSON数组”。	JSON对象
	记录节点	当“文件格式”选择为“JSON格式”并且“JSON类型”为“JSON对象”时，才有该参数。对该JSON节点下的数据进行解析，如果该节点对应的数据为JSON数组，那么系统会以同一模式从该数组中提取数据。多层嵌套的JSON节点以字符“.”分割。	data.list
高级属性	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n
	字段分隔符	文件中的字段分隔符，使用Tab键作为分隔符请输入“\t”。当“文件格式”选择为“CSV格式”时，才有该参数。	,
	使用包围符	选择“是”时，包围符内的字段分隔符会被视为字符串值的一部分，目前CDM默认的包围符为："。	否
	使用转义符	选择“是”时，CSV数据行中的\作为转义符使用。 选择“否”时，CSV中的\作为数据不会进行转义。CSV只支持\作为转义符。	是
	使用正则表达式分隔字段	选择是否使用正则表达式分隔字段，当选择“是”时，“字段分隔符”参数无效。当“文件格式”选择为“CSV格式”时，才有该参数。	是
	正则表达式	分隔字段的正则表达式，正则表达式写法请参考**正则表达式分隔半结构化文本**。	^(\d.*\d) (\w*) \[(.*)\] ([\w\.]*) (\w.*).*

参数类型	参数名	说明	取值样例
	前N行为标题行	“文件格式”选择“CSV格式”时才有该参数。在迁移CSV文件到表时，CDM默认是全部写入，如果该参数选择“是”，CDM会将CSV文件的前N行数据作为标题行，不写入目的端的表。	否
	标题行数	“前N行为标题行”选择“是”时才有该参数。抽取数据时将被跳过的标题行数。 **说明** 标题行数不为空，取值为1-99之间的整数。	1
	解析首行为列名	“前N行为标题行”选择“是”时才有该参数。选择是否将标题的首行解析为列名，在配置字段映射时会在原字段中显示该列名。 **说明** ● 标题行数大于1时，当前仅支持解析标题的首行作为列名。 ● 列名不支持“&”字符，否则会导致作业迁移失败，需修改CSV文件“&”字符即可正常迁移。	是
	编码类型	文件编码类型，例如：“UTF-8”或“GBK”。只有文本文件可以设置编码类型，当“文件格式”选择为“二进制格式”时，该参数值无效。	GBK
	压缩格式	选择对应压缩格式的源文件： ● 无：表示传输所有格式的文件。 ● GZIP：表示只传输GZIP格式的文件。 ● ZIP：表示只传输ZIP格式的文件。 ● TAR.GZ：表示只传输TAR.GZ格式的文件。	无
	压缩文件后缀	压缩格式非无时，显示该参数。 该参数需要解压缩的文件后缀名。当一批文件中以该值为后缀时，才会执行解压缩操作，否则保持原样传输。当输入*或为空时，所有文件都会被解压。	*
	启动作业标识文件	选择是否开启作业标识文件的功能。当源端路径下存在启动作业的标识文件时才启动作业，否则会挂起等待一段时间，等待时长在下方“等待时间”中配置。	否

参数类型	参数名	说明	取值样例
	标识文件名	选择开启作业标识文件的功能时，需要指定启动作业的标识文件名。指定文件后，只有在源端路径下存在该文件的情况下才会运行任务。该文件本身不会被迁移。	ok.txt
	等待时间	选择开启作业标识文件的功能时，如果源路径下不存在启动作业的标识文件，作业挂机等待的时长，当超时后任务会失败。 等待时间设置为0时，当源端路径下不存在标识文件，任务会立即失败。 单位：秒。	10
	文件分隔符	“源目录或文件”参数中如果输入的是多个文件路径，CDM使用这里配置的文件分隔符来区分各个文件，默认为|。	|
	过滤类型	满足过滤条件的路径或文件会被传输，该参数有“无”、“通配符”和“正则表达式”三种选择。具体使用方法可参见**文件增量迁移**。	通配符
	目录过滤器	“过滤类型”选择“通配符”、“正则表达式”时，用通配符过滤目录，符合过滤器规则的目录，允许进行迁移。支持配置多个路径，中间使用“,”分隔。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*input
	文件过滤器	“过滤类型”选择“通配符”、“正则表达式”时，用通配符过滤目录下的文件，符合过滤器规则的文件，允许进行迁移。支持配置多个文件，中间使用“,”分隔。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*.csv,*.txt
	时间过滤	选择“是”时，可以根据文件的修改时间，选择性的传输文件。	是

参数类型	参数名	说明	取值样例
	起始时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间大于等于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-06-01 00:00:00
	终止时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}**表示：只迁移修改时间为当前时间以前的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-01 00:00:00
	忽略不存在原路径/文件	如果将其设为是，那么作业在源路径不存在的情况下也能成功执行。	否
	MD5文件名后缀	“文件格式”选择“二进制格式”时，该参数才显示。 校验CDM抽取的文件，是否与源文件一致，详细请参见**MD5校验文件一致性**。	.md5

说明

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

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

5.3.2 配置 HDFS 源端参数

作业中源连接为**HDFS连接**时，即从MRS HDFS、FusionInsight HDFS、Apache HDFS导出数据时，源端作业参数如**表5-8**所示。

表 5-8 HDFS 作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	源连接名称	由用户下拉选择即可。	hdfs_to_cdm
	源目录或文件	“列表文件”选择为“否”时，才有该参数。 待迁移数据的目录或单个文件路径。 待迁移数据的目录，将迁移目录下的所有文件（包括所有嵌套子目录及其子文件）。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	/user/cdm/
	文件格式	传输数据时所用的文件格式，可选择以下文件格式： • CSV格式：以CSV格式解析源文件，用于迁移文件到数据表的场景。 • 二进制格式：选择“二进制格式”时不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。 • Parquet格式：以Parquet格式解析源文件，用于HDFS数据导到表的场景。	CSV格式

参数类型	参数名	说明	取值样例
	列表文件	当“文件格式”选择为“二进制格式”时，才有该参数。 打开列表文件功能时，支持读取OBS桶中文件（如txt文件）的内容作为待迁移文件的列表。该文件中的内容应为待迁移文件的绝对路径（不支持目录），文件内容示例如下： /mrs/job-properties/application_1634891604621_0014/job.properties /mrs/job-properties/application_1634891604621_0029/job.properties	是
	列表文件源连接	当“列表文件”选择为“是”时，才有该参数。可选择列表文件所在的OBS连接。	OBS_test_link
	列表文件OBS桶	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶名。	01
	列表文件或目录	当“列表文件”选择为“是”时，才有该参数。该参数表示列表文件所在的OBS桶中的绝对路径或目录。	/0521/Lists.txt
高级属性	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n
	字段分隔符	文件中的字段分隔符，使用Tab键作为分隔符请输入“\t”。当“文件格式”选择为“CSV格式”时，才有该参数。	,
	首行为标题行	“文件格式”选择“CSV格式”时才有该参数。在迁移CSV文件到表时，CDM默认是全部写入，如果该参数选择“是”，CDM会将CSV文件的前N行数据作为标题行，不写入目的端的表。	否
	编码类型	文件编码类型，例如：“UTF-8”或“GBK”。只有文本文件可以设置编码类型，当“文件格式”选择为“二进制格式”时，该参数值无效。	GBK
	启动作业标识文件	选择是否开启作业标识文件的功能。当源端路径下存在启动作业的标识文件时才启动作业，否则会挂起等待一段时间，等待时长在下方“等待时间”中配置。	ok.txt

参数类型	参数名	说明	取值样例
	过滤类型	满足过滤条件的路径或文件会被传输，该参数有“无”、“通配符”和“正则表达式”三种选择。具体使用方法可参见**文件增量迁移**。	-
	目录过滤器	“过滤类型”选择“通配符”、“正则表达式”时，用通配符过滤目录，符合过滤器规则的目录，允许进行迁移。支持配置多个路径，中间使用“,”分隔。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*input
	文件过滤器	“过滤类型”选择“通配符”、“正则表达式”时，用通配符过滤目录下的文件，符合过滤器规则的文件，允许进行迁移。支持配置多个文件，中间使用“,”分隔。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*.csv
	时间过滤	选择“是”时，可以根据文件的修改时间，选择性的传输文件。	是
	起始时间	“过滤类型”选择“时间过滤器”时，可以指定一个时间值，当文件的修改时间大于等于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-01 00:00:00

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

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

5.3.3 配置 HBase/CloudTable 源端参数

作业中源连接为**HBase连接**或**CloudTable连接**时，即从MRS HBase、FusionInsight HBase、Apache HBase或者CloudTable导出数据时，源端作业参数如**表5-9**所示。

说明

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

表 5-9 HBase/CloudTable 作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	表名	导出数据的HBase表名。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_2
	列族	可选参数，导出数据所属的列族。	CF1&CF2

参数类型	参数名	说明	取值样例
高级属性	切分Rowkey	可选参数，选择是否拆分Rowkey，默认为“否”。	是
	Rowkey分隔符	可选参数，用于拆分Rowkey的分隔符，若不设置则不切分。	|
	起始时间	可选参数，起始时间（包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间及以后的数据。 该参数支持配置为时间宏变量，使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-01-01 20:00:00
	终止时间	可选参数，终止时间（不包含该值），格式为“yyyy-MM-dd HH:mm:ss”，表示只抽取该时间以前的数据。 该参数支持配置为时间宏变量，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-02-01 20:00:00

5.3.4 配置 Hive 源端参数

作业中源连接为**Hive连接**时，源端作业参数如**表5-10**所示。

表 5-10 Hive 作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	数据库名称	输入或选择数据库名称。单击输入框后面的按钮可进入数据库选择界面。	default

参数类型	参数名	说明	取值样例
	表名	输入或选择Hive表名。单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_E
	读取方式	包括HDFS和JDBC两种读取方式。默认为HDFS方式，如果没有使用WHERE条件进行数据过滤及在字段映射页面添加新字段的需求，选择HDFS方式即可。 • HDFS文件方式读取数据时，性能较好，但不支持使用WHERE条件进行数据过滤及在字段映射页面添加新字段。 • JDBC方式读取数据时，支持使用WHERE条件进行数据过滤及在字段映射页面添加新字段。	HDFS
	使用SQL语句	导出关系型数据库的数据时，您可以选择使用自定义SQL语句导出。	否

参数类型	参数名	说明	取值样例

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

高级属性	分区过滤条件	读取方式为HDFS时，单击“显示高级属性”后显示此参数。 该参数表示抽取指定值的partition，属性名称为分区名称，属性值可以配置多个值（空格分隔），也可以配置为字段取值范围，接受时间宏函数。详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	• 单/多值过滤场景属性值：${dateformat(yyyyMMdd, -1, DAY)} ${dateformat(yyyyMMdd)} • 范围过滤场景属性值：${value} >= ${dateformat(yyyyMMdd, -7, DAY)} && ${value} < ${dateformat(yyyyMMdd)}

参数类型	参数名	说明	取值样例
	Where子句	读取方式为JDBC时，单击“显示高级属性”后显示此参数。 填写该参数表示指定抽取的WHERE子句，不指定则抽取整表。如果要迁移的表中没有WHERE子句的字段，则会迁移失败。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	age > 18 and age <= 60

说明

Hive作为数据源，CDM自动使用Hive数据分片文件进行数据分区。

5.3.5 配置 DLI 源端参数

作业中源连接为**DLI连接**时，源端作业参数如**表5-11**所示。

表 5-11 DLI 作为源端时的作业参数

参数名	说明	取值样例
资源队列	选择目的表所属的资源队列。 DLI的default队列无法在迁移作业中使用，您需要在DLI中新建SQL队列。	cdm
数据库名称	写入数据的数据库名称。	dli
表名	写入数据的表名。	car_detail
分区	用于抽取分区的信息。	year=2020,location=sun

5.3.6 配置 FTP/SFTP 源端参数

作业中源连接为**FTP/SFTP连接**时，源端作业参数如**表5-12**所示。

高级属性里的参数为可选参数，默认隐藏，单击界面上的“显示高级属性”后显示。

表 5-12 FTP/SFTP 作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	源目录或文件	待迁移数据的目录或单个文件路径。文件路径支持输入多个文件（最多50个），默认以“|”分隔，也可以自定义文件分隔符，具体请参见**文件列表迁移**。 待迁移数据的目录，将迁移目录下的所有文件（包括所有嵌套子目录及其子文件）。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	/ftp/a.csv|/ftp/b.txt
	文件格式	指CDM以哪种格式解析数据，可选择以下格式： • CSV格式：以CSV格式解析源文件，用于迁移文件到数据表的场景。 • 二进制格式：选择“二进制格式”时不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。 • JSON格式：以JSON格式解析源文件，一般都是用于迁移文件到数据表的场景。 **说明** 当目的端为OBS数据源时，仅支持配置二进制格式。	CSV格式
	JSON类型	当“文件格式”选择为“JSON格式”时，才有该参数。JSON文件中存储的JSON对象的类型，可以选择“JSON对象”或“JSON数组”。	JSON对象
	记录节点	当“文件格式”选择为“JSON格式”并且“JSON类型”为“JSON对象”时，才有该参数。对该JSON节点下的数据进行解析，如果该节点对应的数据为JSON数组，那么系统会以同一模式从该数组中提取数据。多层嵌套的JSON节点以字符“.”分割。	data.list
高级属性	使用rfc4180解析器	当“文件格式”选择为“CSV格式”时，才有该参数。是否使用rfc4180解析器解析CSV文件。	否

参数类型	参数名	说明	取值样例
	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。当“文件格式”选择为“CSV格式”时，才有该参数。	\n
	字段分隔符	文件中的字段分隔符，使用Tab键作为分隔符请输入“\t”。当“文件格式”选择为“CSV格式”时，才有该参数。	,
	使用包围符	选择“是”时，包围符内的字段分隔符会被视为字符串值的一部分，目前CDM默认的包围符为："。	否
	使用转义符	选择“是”时，CSV数据行中的\作为转义符使用。 选择“否”时，CSV中的\作为数据不会进行转义。CSV只支持\作为转义符。	是
	使用正则表达式分隔字段	选择是否使用正则表达式分隔字段，当选择“是”时，“字段分隔符”参数无效。当“文件格式”选择为“CSV格式”时，才有该参数。	是
	正则表达式	当“使用正则表达式分隔字段”选择为“是”时，才有该参数。 分隔字段的正则表达式，正则表达式写法请参考**正则表达式分隔半结构化文本**。	^(\d.*\d) (\w*) \[(.*) \] ([\w\.]*) (\w.*).*
	首行为标题行	“文件格式”选择“CSV格式”时才有该参数。在迁移CSV文件到表时，CDM默认是全部写入，如果该参数选择“是”，CDM会将CSV文件的前N行数据作为标题行，不写入目的端的表。	是
	编码类型	文件编码类型，例如：“UTF-8”或“GBK”。只有文本文件可以设置编码类型，当“文件格式”选择为“二进制格式”时，该参数值无效。	UTF-8
	压缩格式	选择对应压缩格式的源文件： ● 无：表示传输所有格式的文件。 ● GZIP：表示只传输GZIP格式的文件。 ● ZIP：表示只传输ZIP格式的文件。 ● TAR.GZ：表示只传输TAR.GZ格式的文件。	无
	压缩文件后缀	压缩格式非无时，显示该参数。 该参数需要解压缩的文件后缀名。当一批文件中以该值为后缀时，才会执行解压缩操作，否则保持原样传输。当输入*或为空时，所有文件都会被解压。	*

参数类型	参数名	说明	取值样例
	启动作业标识文件	选择是否开启作业标识文件的功能。当源端路径下存在启动作业的标识文件时才启动作业，否则会挂起等待一段时间，等待时长在下方“等待时间”中配置。	是
	文件分隔符	“源目录或文件”参数中如果输入的是多个文件路径，CDM使用这里配置的文件分隔符来区分各个文件，默认为|。	|
	标识文件名	选择开启作业标识文件的功能时，需要指定启动作业的标识文件名。指定文件后，只有在源端路径下存在该文件的情况下才会运行任务。该文件本身不会被迁移。	ok.txt
	等待时间	选择开启作业标识文件的功能时，如果源路径下不存在启动作业的标识文件，作业挂机等待的时长，当超时后任务会失败。 等待时间设置为0时，当源端路径下不存在标识文件，任务会立即失败。 单位：秒。	10
	过滤类型	满足过滤条件的路径或文件会被传输，该参数有“无”、“通配符”和“正则表达式”三种选择。具体使用方法可参见**文件增量迁移**。	无
	目录过滤器	“过滤类型”选择“通配符”和“正则表达式”时，符合过滤器规则的目录，允许进行迁移。支持配置多个路径，中间使用“,”分隔。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*input,*out
	文件过滤器	“过滤类型”选择“通配符”和“正则表达式”时，符合过滤器规则的文件，允许进行迁移。支持配置多个文件，中间使用“,”分隔。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	*.csv
	时间过滤	选择“是”时，可以根据文件的修改时间，选择性的传输文件。	是

参数类型	参数名	说明	取值样例
	起始时间	“时间过滤”选择“是”时，可以指定一个时间值，当文件的修改时间大于等于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}**表示：只迁移最近90天内的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-01 00:00:00
	终止时间	“时间过滤”选择“是”时，可以指定一个时间值，当文件的修改时间小于该时间才会被传输，输入的时间格式需为“yyyy-MM-dd HH:mm:ss”。 该参数支持配置为时间宏变量，例如**${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}**表示：只迁移修改时间为当前时间以前的文件。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	2019-07-30 00:00:00
	忽略不存在原路径/文件	如果将其设为“是”，那么作业在源路径不存在的情况下也能成功执行。	否
	标识文件类型	选择开启作业标识文件的功能时，该参数才显示。 ● MARK_DONE：只有在源端路径下存在标识文件的情况下才会执行迁移任务。 ● MARK_DOING：只有在源端路径下不存在标识文件的情况下才会执行迁移任务。	MARK_DOING
	是否跳过空行	“文件格式”选择“CSV格式”时，该参数才显示。 如果某行数据为空，则跳过此行。	否
	null值	“文件格式”选择“二进制格式”时，该参数才显示。 由于文本文件中无法用字符串定义null值，此配置项定义将何种字符串标识为null。	否

参数类型	参数名	说明	取值样例
	MD5文件名后缀	“文件格式”选择“二进制格式”时，该参数才显示。 校验CDM抽取的文件，是否与源文件一致，详细请参见**MD5校验文件一致性**。	.md5

5.3.7 配置 HTTP 源端参数

作业中源连接为HTTP连接时，源端作业参数如**表5-13**所示。当前只支持从HTTP URL导出数据，不支持导入。

表 5-13 HTTP/HTTPS 作为源端时的作业参数

参数名	说明	取值样例
文件URL	通过使用GET方法，从HTTP/HTTPS协议的URL中获取数据。 用于读取一个公网HTTP/HTTPS URL的文件，包括第三方对象存储的公共读取场景和网盘场景。	https://bucket.obs.myhuaweicloud.com/object-key
列表文件	选择“是”，将待上传的文本文件中所有URL对应的文件拉取到OBS，文本文件记录的是HDFS上的文件路径。	是
列表文件源连接	文本文件存储在OBS桶中，这里需要选择已建立的OBS连接。	obs_link
列表文件OBS桶	存储文本文件的OBS桶名称。	obs-cdm
列表文件或目录	在OBS中存储文本文件的文件自定义目录，多级目录可用“/”进行分隔。	test1
文件格式	当前CDM只支持选择“二进制格式”，不解析文件内容直接传输，不要求原文件格式必须为二进制。	二进制格式
压缩格式	选择对应压缩格式的源文件进行迁移： ● 无：表示传输所有格式的文件。 ● GZIP：表示只传输GZIP格式的文件。 ● ZIP：表示只传输ZIP格式的文件。 ● TAR.GZ：表示只传输TAR.GZ格式的文件。	无
压缩文件后缀	压缩格式非无时，显示该参数。 该参数需要解压缩的文件后缀名。当一批文件中以该值为后缀时，才会执行解压缩操作，否则保持原样传输。当输入*或为空时，所有文件都会被解压。	*

参数名	说明	取值样例
文件分隔符	传输多个文件时，CDM使用这里配置的文件分隔符来区分各个文件，默认为|。列表文件选择“是”时，不显示该参数。	|
QUERY参数	• 该参数设置为“是”时，上传到OBS的对象使用的对象名，为去掉query参数后的字符。 • 该参数设置为“否”时，上传到OBS的对象使用的对象名，包含query参数。	否
忽略不存在原路径/文件	如果将其设为是，那么作业在源路径不存在的情况下也能成功执行。	否
MD5文件名后缀	校验CDM抽取的文件，是否与源文件一致，详细请参见**MD5校验文件一致性**。	.md5

5.3.8 配置 PostgreSQL/SQL Server 源端参数

作业中源连接为从云数据库 PostgreSQL、云数据库 SQL Server、PostgreSQL、Microsoft SQL Server导出的数据时，源端作业参数如**表5-14**所示。

表 5-14 PostgreSQL/SQL Server 作为源端时的作业参数

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

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

参数类型	参数名	说明	取值样例
	模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： • **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 • ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 • ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **说明** 表名支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有表（要求表中的字段个数和类型都一样）。例如： • **table***表示导出所有以“table”开头的表。 • ***table**表示导出所有以“table”结尾的表。 • ***table***表示表名中只要有“table”字符串，就全部导出。	table

参数类型	参数名	说明	取值样例
高级属性	抽取分区字段	“使用SQL语句”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** 抽取分区字段支持TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。	id
	Where子句	“使用SQL语句”选择“否”时，显示该参数，表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	分区字段是否允许空值	是否允许分区字段包含空值。	是
	按表分区抽取	支持从分区表的各个分区并行抽取数据。启用该功能时，可以通过下面的“表分区”参数指定具体的表分区。 ● 该功能不支持非分区表。 ● 仅支持源端数据源为PostgreSQL时配置该参数。 ● 数据库用户需要具有系统视图**dba_tab_partitions**和**dba_tab_subpartitions**的**SELECT**权限。	否
	拆分作业	选择“是”，会根据“作业拆分字段”值，将作业拆分为多个子作业并发执行。 **说明** 仅支持目的端为DLI和Hive时配置该参数及*作业拆分字段*、*拆分字段最小值*、*拆分字段最大值*、*子作业个数*参数。	是
	作业拆分字段	“拆分作业”选择“是”时，显示该参数，使用该字段将作业拆分为多个子作业并发执行。	-
	拆分字段最小值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最小值。	-

参数类型	参数名	说明	取值样例
	拆分字段最大值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最大值。	-
	子作业个数	“拆分作业”选择“是”时，显示该参数，根据“作业拆分字段”的最小值和最大值限定的数据范围，将作业拆分为多少个子作业执行。	-

5.3.9 配置 DWS 源端参数

作业中源连接为**DWS连接**时，源端作业参数如**表5-15**所示。

表 5-15 DWS 作为源端时的作业参数

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

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

参数类型	参数名	说明	取值样例
	模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： • **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 • ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 • ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **说明** 表名支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有表（要求表中的字段个数和类型都一样）。例如： • **table***表示导出所有以“table”开头的表。 • ***table**表示导出所有以“table”结尾的表。 • ***table***表示表名中只要有“table”字符串，就全部导出。	table

参数类型	参数名	说明	取值样例
高级属性	Where子句	“使用SQL语句”选择“否”时，显示该参数，表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	抽取分区字段	“使用SQL语句”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** 抽取分区字段支持TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。	id
	分区字段含有空值	是否允许分区字段包含空值。	是
	拆分作业	选择“是”，会根据“作业拆分字段”值，将作业拆分为多个子作业并发执行。 **说明** 仅支持目的端为DLI和Hive时配置该参数及*作业拆分字段*、*拆分字段最小值*、*拆分字段最大值*、*子作业个数*参数。	是
	作业拆分字段	“拆分作业”选择“是”时，显示该参数，使用该字段将作业拆分为多个子作业并发执行。	-
	拆分字段最小值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最小值。	-
	拆分字段最大值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最大值。	-
	子作业个数	“拆分作业”选择“是”时，显示该参数，根据“作业拆分字段”的最小值和最大值限定的数据范围，将作业拆分为多少个子作业执行。	-

5.3.10 配置 SAP HANA 源端参数

SAP HANA作为源端作业参数如**表5-16**所示。

表 5-16 SAP HANA 作源端时的作业参数

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

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

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

参数类型	参数名	说明	取值样例
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **说明** 表名支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有表（要求表中的字段个数和类型都一样）。例如： • **table***表示导出所有以“table”开头的表。 • ***table**表示导出所有以“table”结尾的表。 • ***table***表示表名中只要有“table”字符串，就全部导出。	table
高级属性	Where子句	“使用SQL语句”选择“否”时，显示该参数，表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	抽取区分字段	“使用SQL语句”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** 抽取分区字段支持TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。	id

5.3.11 配置 MySQL 源端参数

作业中源连接为**云数据库MySQL/MySQL数据库连接**时，源端作业参数如**表5-17**所示。

表 5-17 MySQL 作为源端时的作业参数

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

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

模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的账号是否有元数据查询的权限。 该参数支持配置正则表达式，实现导出满足规则的所有数据库。	SCHEMA_E

参数名	说明	取值样例
表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 该参数支持配置正则表达式，实现导出满足规则的所有数据库。例如：表名配置为*user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
抽取分区字段	“使用SQL语句”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** 抽取分区字段支持TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。	id
Where子句	“使用SQL语句”选择“否”时，显示该参数，表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
分区字段含有空值	是否允许分区字段包含空值。	是
拆分作业	选择“是”，会根据“作业拆分字段”值，将作业拆分为多个子作业并发执行。 **说明** 仅支持目的端为DLI和Hive时配置该参数及*作业拆分字段*、*拆分字段最小值*、*拆分字段最大值*、*子作业个数*参数。	是

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

5.3.12 配置 Oracle 源端参数

作业中源连接为**Oracle数据库连接**，源端作业参数如**表5-18**所示。

表 5-18 Oracle 作为源端时的作业参数

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

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

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

参数名	说明	取值样例
抽取分区字段	“按表分区抽取”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** 抽取分区字段支持TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。	id
Where子句	“使用SQL语句”选择“否”时，显示该参数，表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
分区字段含有空值	“按表分区抽取”选择“否”时，显示该参数，表示是否允许分区字段包含空值。	是
按表分区抽取	从Oracle导出数据时，支持从分区表的各个分区并行抽取数据。启用该功能时，可以通过下面的“表分区”参数指定具体的Oracle表分区。 • 该功能不支持非分区表。 • 数据库用户需要具有系统视图**dba_tab_partitions**和**dba_tab_subpartitions**的**SELECT**权限。	否
表分区	输入需要迁移数据的Oracle表分区，多个分区以&分隔，不填则迁移所有分区。 如果有子分区，以“分区.子分区”的格式填写，例如“P2.SUBP1”。	P0&P1&P2.SUBP1&P2.SUBP3
拆分作业	选择“是”，会根据“作业拆分字段”值，将作业拆分为多个子作业并发执行。 **说明** 仅支持目的端为DLI和Hive时配置该参数及*作业拆分字段*、*拆分字段最小值*、*拆分字段最大值*、*子作业个数*参数。	是
作业拆分字段	“拆分作业”选择“是”时，显示该参数，使用该字段将作业拆分为多个子作业并发执行。	-
拆分字段最小值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最小值。	-

参数名	说明	取值样例
拆分字段最大值	“拆分作业”选择“是”时，显示该参数，表示抽取数据时“作业拆分字段”的最大值。	-
子作业个数	“拆分作业”选择“是”时，显示该参数，根据“作业拆分字段”的最小值和最大值限定的数据范围，将作业拆分为多少个子作业执行。	-

说明

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

5.3.13 配置分库源端参数

作业中源连接为**分库连接**，源端作业参数如**表5-19**所示。

表 5-19 分库作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	模式或表空间	表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，分库连接时此处默认展示对应第一个后端连接的表空间。用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的账号是否有元数据查询的权限。 该参数支持配置正则表达式，实现导出满足规则的所有数据库。例如：表名配置为 *user_[0-9]{1,2}*，会匹配 user_0 到 user_9，user_00 到 user_99 的表。	SCHEMA_E

参数类型	参数名	说明	取值样例
	表名	表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 该参数支持配置正则表达式，实现导出满足规则的所有数据库。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
高级属性	Where子句	表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'

说明

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

5.3.14 配置 MongoDB/DDS 源端参数

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

作业中源连接为**MongoDB连接**时，即从本地MongoDB或DDS导出数据时，源端作业参数如**表5-20**所示。

表 5-20 MongoDB/DDS 作为源端时的作业参数

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

5.3.15 配置 Redis 源端参数

第三方云的Redis服务无法支持作为源端。如果是用户在本地数据中心或ECS上自行搭建的Redis支持作为源端或目的端。

作业中源连接为从本地Redis导出的数据时，源端作业参数如**表5-21**所示。

表 5-21 Redis 作为源端时的作业参数

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

参数类型	参数名	说明	取值样例
	值存储类型	仅支持以下数据格式： • STRING：不带列名，如“值1，值2”形式。 • HASH：带列名，如“列名1=值1，列名2=值2”的形式。	STRING
高级属性	键分隔符	用来分隔关系型数据库的表和列名。	_
	值分隔符	以STRING方式存储时，列之间的分隔符。	;
	字段相同	“值存储类型”参数值为“HASH”显示该参数。 哈希键内有相同的字段。	是

5.3.16 配置 DIS 源端参数

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

作业中源连接为**DIS连接**时，源端作业参数如所**表5-22**示。

表 5-22 DIS 作为源端时的作业参数

参数类型	参数	说明	取值样例
基本参数	DIS通道	DIS的通道名。	dis
	是否持久运行	用户自定义是否永久运行。设置为长久运行的任务，如果DIS系统发生中断，任务也会失败结束。	是
	DIS分区ID	DIS分区ID，该参数支持输入多个分区ID，使用英文逗号（,）分隔。	0,1,2
	偏移量参数	设置从DIS拉取数据时的初始偏移量： • 最新：最大偏移量，即拉取最新的数据。 • 上次停止处：从上次停止处继续读取。 • 最早：最小偏移量，即拉取最早的数据。	最新
	APP名字	配置用户数据消费程序的唯一标识符，不存在时会自动创建。	cdm

参数类型	参数	说明	取值样例
	数据格式	解析数据时使用的格式： ● 二进制格式：适用于文件迁移场景，不解析数据内容原样传输。 ● CSV格式：以CSV格式解析源数据。 ● JSON格式：以JSON格式解析源数据。	二进制格式
	字段分隔符	数据格式为“CSV格式”时呈现此参数。默认为逗号，使用Tab键作为分隔符请输入“\t”。	,
	记录分隔符	数据格式为“CSV格式”或“JSON格式”时呈现此参数。用于配置每条记录之间的分割符。	,
高级属性	最大消息数/poll	可选参数，每次向DIS请求数据限制最大请求记录数。	100

5.3.17 配置 Kafka/DMS Kafka 源端参数

作业中源连接为**Kafka连接**或**DMS Kafka连接**时，源端作业参数如**表5-23**所示。

表 5-23 Kafka 作为源端时的作业参数

参数类型	参数	说明	取值样例
基本参数	Topics	支持单个或多个topic。	est1,est2
	数据格式	解析数据时使用的格式： ● 二进制格式：适用于文件迁移场景，不解析数据内容原样传输。 ● CSV格式：以CSV格式解析源数据。 ● JSON：以JSON格式解析源数据。 ● CDC（DRS）：以DRS格式解析源数据。 ● CDC（JSON）：以JSON格式解析源数据。 ● CDC（DRS_AVRO）：以DRS_AVRO格式解析源数据。 ● CDC（DRS_JSON）：以DRS_JSON格式解析源数据。	二进制格式
	偏移量参数	从Kafka拉取数据时的初始偏移量： ● 最新：最大偏移量，即拉取最新的数据。 ● 最早：最小偏移量，即拉取最早的数据。 ● 已提交：拉取已提交的数据。 ● 时间范围：拉取时间范围内的数据。	最新

参数类型	参数	说明	取值样例
	抽取数据最大运行时间	持续拉取数据时间。如天调度作业，根据每天topic产生的数据量，配置足够的拉取时间。单位：分钟。	60
	等待时间	当配置为60时，如果消费者60s内从Kafka拉取数据返回一直为空（一般是已经读完主题中的全部数据，也可能是网络或者Kafka集群可用性原因），则立即停止任务，否则持续重试读取数据。单位：秒。	60
	消费组ID	用户指定消费组ID。 如果是从DMS Kafka导出数据，专享版请任意输入，标准版请输入有效的消费组ID。	sumer-group
	开始时间(>=)	“偏移量参数”选择为“时间范围”时配置。拉取数据的开始时间，包含设置时间点的数据。	2020-12-20 12:00:00
	结束时间(<)	“偏移量参数”选择为“时间范围”时配置。拉取数据的结束时间，不包含设置时间点的数据。	2020-12-20 20:00:00
	字段分隔符	“数据格式”选择为“CSV格式”时配置。默认为空格，使用Tab键作为分隔符请输入“\t”。	,
	记录分隔符	“数据格式”选择为“CSV格式”、“JSON”时配置。默认为空格，使用Tab键作为分隔符请输入“\t”。	,
高级参数	使用配置文件	“数据格式”选择为“CDC场景”时配置，用于配置OBS文件。	否
	OBS链接	选择OBS连接器信息。	obs_link
	OBS桶	选择OBS桶。	obs_test
	配置文件	选择OBS的配置文件。	/obs/config.csv
	最大消息数/poll	可选参数，每次向Kafka请求数据限制最大请求记录数。	100
	最大时间间隔/poll	可选参数，向Kafka请求数据的最大时间间隔。	100
	通知Topic	发送通知数据到通知Topic中。在CDC场景中，通知的内容是记录生成文件列表的文件名。	notice

5.3.18 配置 Elasticsearch/云搜索服务源端参数

作业中源连接为**配置Elasticsearch连接**或**配置云搜索服务（CSS）连接**时，源端作业参数如**表5-24**所示。

表 5-24 Elasticsearch/云搜索服务作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	索引	Elasticsearch的索引，类似关系数据库中的数据库名称。索引名称只能全部小写，不能有大写。	index
	类型	Elasticsearch的类型，类似关系数据库中的表名称。类型名称只能全部小写，不能有大写。 **说明** Elasticsearch搜索引擎7.x及以上版本不支持自定义类型，只能使用_doc类型。此处即使自定义也不会生效。	_doc
高级属性	拆分nested类型字段	可选参数，选择是否将nested字段的json内容拆分，例如：将“a:{ b:{ c:1, d:{ e:2, f:3 } } }”拆成三个字段“a.b.c”、“a.b.d.e”、“a.b.d.f”。	否

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

参数类型	参数名	说明	取值样例
		• 暂不支持通过Elasticsearch的query DSL（即DSL语法，Domain Sepcified Language）查询方式对源数据进行过滤。	
	抽取元字段	表示是否抽取索引的元字段，目前只支持（_index、_type、_id、_score）例如：_index、_type、_id、_score	是
	分页大小	Elasticsearch分页查询，用来设置分页size的大小。	1000
	ScrollId超时时间配置	Elasticsearch scroll查询时会记录一个scroll_id，超时或者scroll查询结束后会清除请求的srcoll_id，通过设置这个超时时间配置，来指定scroll_id超时时间。	5

5.3.19 配置 MRS Hudi 源端参数

作业中源连接为**MRS Hudi连接**时，源端作业参数如**表5-25**所示。

表 5-25 MRS Hudi 作为源端时的作业参数

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

参数类型	参数名	说明	取值样例
高级属性	Where子句	填写该参数表示指定抽取的Where子句，不指定则抽取整表。如果要迁移的表中没有Where子句的字段，则会迁移失败。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	age > 18 and age <= 60

5.3.20 配置 MRS ClickHouse 源端参数

作业中源连接为**MRS ClickHouse连接**时，源端作业参数如**表5-26**所示。

表 5-26 MRS ClickHouse 作为源端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	源连接名称	选择已配置的MRS ClickHouse连接。	ck_from_cdm
	模式或表空间	单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的账号是否有元数据查询的权限。 **说明** 该参数支持配置正则表达式，实现导出满足规则的所有数据库。	default
	表名	单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账号是否有元数据查询的权限。 **说明** 该参数支持配置正则表达式，实现导出满足规则的所有数据库。	TBL_E

参数类型	参数名	说明	取值样例
高级属性	Where子句	填写该参数表示指定抽取的WHERE子句，不指定则抽取整表。如果要迁移的表中没有WHERE子句的字段，则会迁移失败。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	age > 18 and age <= 60

5.3.21 配置达梦数据库 DM 源端参数

从达梦数据库 DM导出数据时，源端作业参数如**表5-27**所示。

表 5-27 达梦数据库 DM 作为源端时的作业参数

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

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

参数类型	参数名	说明	取值样例
	模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： • **SCHEMA***表示导出所有以“SCHEMA”开头的数据库。 • ***SCHEMA**表示导出所有以“SCHEMA”结尾的数据库。 • ***SCHEMA***表示数据库名称中只要有“SCHEMA”字符串，就全部导出。	SCHEMA_E
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **说明** 表名支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有表（要求表中的字段个数和类型都一样）。例如： • **table***表示导出所有以“table”开头的表。 • ***table**表示导出所有以“table”结尾的表。 • ***table***表示表名中只要有“table”字符串，就全部导出。	table

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

5.3.22 配置神通（ST）源端参数

从神通（ST）导出数据时，源端作业参数如**表5-28**所示。

表 5-28 神通（ST）作为源端时的作业参数

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

参数类型	参数名	说明	取值样例

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

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

参数类型	参数名	说明	取值样例
	表名	“使用SQL语句”选择“否”时，显示该参数，表示要抽取的表名。单击输入框后面的按钮可进入表的选择界面，用户也可以直接输入表名称。 如果选择界面没有待选择的表，请确认表是否已经创建，或者对应连接里的账号是否有元数据查询的权限。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。 **说明** 表名支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有表（要求表中的字段个数和类型都一样）。例如： • **table***表示导出所有以“table”开头的表。 • ***table**表示导出所有以“table”结尾的表。 • ***table***表示表名中只要有“table”字符串，就全部导出。	table
高级属性	抽取分区字段	“使用SQL语句”选择“否”时，显示该参数，表示抽取数据时使用该字段进行数据切分，CDM依据此字段将作业分割为多个任务并发执行。一般使用数据均匀分布的字段，例如以自然增长的序号字段作为分区字段。 单击输入框后面的按钮可进入字段选择界面，用户也可以直接输入抽取分区字段名。 **说明** 抽取分区字段支持TINYINT、SMALLINT、INTEGER、BIGINT、REAL、FLOAT、DOUBLE、NUMERIC、DECIMAL、BIT、BOOLEAN、DATE、TIME、TIMESTAMP类型，建议该字段带有索引。	id
	Where子句	“使用SQL语句”选择“否”时，显示该参数，表示配置抽取范围的Where子句，不配置时抽取整表。 该参数支持配置为时间宏变量，实现抽取指定日期的数据，详细说明请参见**关系数据库增量迁移**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'

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

5.4 配置作业目的端参数

5.4.1 配置 OBS 目的端参数

作业中目的连接为**OBS连接**时，即导入数据到云服务OBS时，目的端作业参数如**表5-29**所示。

高级属性里的参数为可选参数，默认隐藏，单击界面上的“显示高级属性”后显示。

表 5-29 OBS 作为目的端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	桶名	写入数据的OBS桶名。	bucket_2
	写入目录	写入数据到OBS服务器的目录，目录前面不加“/”。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	directory/

参数类型	参数名	说明	取值样例
	文件格式	写入后的文件格式，可选择以下文件格式： ● CSV格式：按CSV格式写入，适用于数据表到文件的迁移。 ● 二进制格式：选择“二进制格式”时不解析文件内容直接传输，CDM会原样写入文件，不改变原始文件格式，适用于文件到文件的迁移。 如果是文件类数据源（FTP/SFTP/HDFS/OBS）之间相互迁移数据，此处的“文件格式”只能选择与源端的文件格式一致。 **说明** ● 当源端为MRS Hive数据源时，仅支持配置CSV格式。 ● 当源端为FTP/SFTP数据源时，仅支持配置二进制格式。	CSV格式
	重复文件处理方式	当源端为HDFS数据源时配置。 只有文件名和文件大小都相同才会判定为重复文件。写入时如果出现文件重复，可选择如下处理方式： ● 替换重复文件 ● 跳过重复文件 ● 停止任务 具体使用方法可参见**文件增量迁移**。	跳过重复文件
高级属性	加密方式	选择是否对上传的数据进行加密，以及加密方式： ● 无：不加密，直接写入数据。 ● KMS：使用数据加密服务中的KMS进行加密。如果启用KMS加密则无法进行数据的MD5校验。 详细使用方法请参见**迁移文件时加解密**。	KMS
	KMS ID	写入文件时加密使用的密钥，“加密方式”选择“KMS”时显示该参数。单击输入框后面的，可以直接选择在数据加密服务中已创建好的KMS密钥。 ● 当使用与CDM集群相同项目下的KMS密钥时，不需要修改下面的“项目ID”参数。 ● 当用户使用其它项目下的KMS密钥时，需要修改下面的“项目ID”参数。	53440ccb-3e73-4700-98b5-71ff5476e621

参数类型	参数名	说明	取值样例
	项目ID	KMS ID所属的项目ID，该参数默认值为当前CDM集群所属的项目ID。 • 当“KMS ID”与CDM集群在同一个项目下时，这里的“项目ID”保持默认即可。 • 当“KMS ID”使用的是其它项目下的KMS ID时，这里需要修改为KMS所属的项目ID。	9bd7c4bd54e5417198f9591bef07ae67
	复制Content-Type属性	“文件格式”为“二进制”，且源端、目的端都为对象存储时，才有该参数。 选择“是”后，迁移对象文件时会复制源文件的Content-Type属性，主要用于静态网站的迁移场景。 归档存储的桶不支持设置Content-Type属性，所以如果开启了该参数，目的端选择写入的桶时，必须选择非归档存储的桶。	否
	换行符	文件中的换行符，默认自动识别“\n”、“\r”或“\r\n”。“文件格式”为“二进制格式”时该参数值无效。	\n
	字段分隔符	文件中的字段分隔符。“文件格式”为“二进制格式”时该参数值无效。	,
	写入文件大小	源端为数据库时该参数才显示，支持按大小分成多个文件存储，避免导出的文件过大，单位为MB。	1024
	校验MD5值	使用“二进制格式”传输文件时，才能校验MD5值。选择校验MD5值时，无法使用KMS加密。 计算源文件的MD5值，并与OBS返回的MD5值进行校验。如果源端已经存在MD5文件，则直接读取源端的MD5文件与OBS返回的MD5值进行校验，具体请参见**MD5校验文件一致性**。	是
	记录校验结果	当选择校验MD5值时，可以选择是否记录校验结果。	是
	校验结果写入连接	可以指定任意一个OBS连接，将MD5校验结果写入该连接的桶下。	obslink
	OBS桶	写入MD5校验结果的OBS桶。	cdm05
	写入目录	写入MD5校验结果的目录。	/md5/
	编码类型	文件编码类型，例如：“UTF-8”或“GBK”。“文件格式”为“二进制格式”时该参数值无效。	GBK

参数类型	参数名	说明	取值样例
	使用包围符	“文件格式”为“CSV格式”，才有该参数，用于将数据库的表迁移到文件系统的场景。 选择“是”时，如果源端数据表中的某一个字段内容包含字段分隔符或换行符，写入目的端时CDM会使用双引号（"）作为包围符将该字段内容括起来，作为一个整体存储，避免其中的字段分隔符误将一个字段分隔成两个，或者换行符误将字段换行。例如：数据库中某字段为hello,world，使用包围符后，导出到CSV文件的时候数据为"hello,world"。	否
	首行为标题行	从关系型数据库导出数据到OBS，“文件格式”为“CSV格式”时，才有该参数。 在迁移表到CSV文件时，CDM默认是不迁移表的标题行，如果该参数选择“是”，CDM在才会将表的标题行数据写入文件。	否
	作业成功标识文件	当作业执行成功时，会在写入目录下生成一个标识文件，文件名由用户指定。不指定时默认关闭该功能。	finish.txt
	文件夹模式	从关系型数据库导出数据到OBS，才有该参数。 启用后将会以根目录-表名-数据类型-数据的文件夹模型生成文件。例如：raw_schema/tbl_student/datas/tbl_student_1.csv	是
	Blog/Clog文件扩展名	“文件夹模式”为“是”时，才有该参数。文件夹模式下自定义Blob/Clog数据的文件扩展名。	.dat/.jpg/.png
	自定义目录层次	选择“是”时，支持迁移后的文件按照自定义的目录存储。即只迁移文件，不迁移文件所归属的目录。	是
	目录层次	自定义迁移后文件的存储路径，支持时间宏变量。 **说明** 源端为关系型数据库数据源时，目录层次为源端表名+自定义目录，其他场景下为自定义目录。	${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}

参数类型	参数名	说明	取值样例
	自定义文件名	从关系型数据库导出数据到OBS，且“文件格式”为“CSV格式”时，才有该参数。 用户可以通过该参数自定义OBS端生成的文件名，支持以下自定义方式： • 字符串，支持特殊字符。例如“cdm#”，则生成的文件名为“cdm#.csv”。 • 时间宏，例如“${timestamp()}”，则生成的文件名为“1554108737.csv”。 • 表名宏，例如“${tableName}”，则生成的文件名为源表名“sqltabname.csv”。 • 版本宏，例如“${version}”，则生成的文件名为集群版本号“2.9.2.200.csv”。 • 字符串和宏（时间宏/表名宏/版本宏）任意组合，例如“cdm#${timestamp()}_${version}”，则生成的文件名为“cdm#1554108737_2.9.2.200.csv”。	cdm

5.4.2 配置 HDFS 目的端参数

作业中目的连接为**HDFS连接**时，目的端作业参数如**表5-30**所示。

表 5-30 HDFS 作为目的端时的作业参数

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

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

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

说明

HDFS文件编码只能为“UTF-8”，故HDFS不支持设置文件编码类型。

5.4.3 配置 HBase/CloudTable 目的端参数

作业中目的连接为**HBase连接**或**CloudTable连接**时，即导入数据到以下数据源时，目的端作业参数如**表5-31**所示。

表 5-31 HBase/CloudTable 作为目的端时的作业参数

参数名	说明	取值样例
表名	写入数据的HBase表名。如果是创建新HBase表，支持从源端复制字段名。单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	TBL_2
导入前清空数据	选择目的端表中数据的处理方式： ● 是：任务启动前会清除目标表中数据。 ● 否：导入前不清空目标表中的数据，如果选“否”且表中有数据，则数据会追加到已有的表中。	是
自动创表	只有当源端为关系数据库时，才有该参数。表示写入表数据时，用户选择的操作： ● 不自动创建：不自动建表。 ● 不存在时创建：当目的端的数据库没有“表名”参数中指定的表时，CDM会自动创建该表。如果“表名”参数配置的表已存在，则不创建，数据写入到已存在的表中。 **说明** Hbase自动建表包含列族与协处理器Coprocessor信息。其他属性按默认值设置，不跟随源端。	不自动创建
Row key拼接分隔符	可选参数，用于多列合并作为rowkey，默认为空格。	,
Rowkey冗余	可选参数，是否将选做Rowkey的数据同时写入HBase的列，默认值“否”。	否
压缩算法	可选参数，创建新HBase表时采用的压缩算法，默认为值“NONE”。 ● NONE：不压缩。 ● SNAPPY：压缩为Snappy格式。 ● GZ：压缩为GZ格式。	NONE

参数名	说明	取值样例
WAL开关	选择是否开启HBase的预写日志机制（WAL，Write Ahead Log）。 ● 是：开启后如果出现HBase服务器宕机，则可以从WAL中回放执行之前没有完成的操作。 ● 否：关闭时能提升写入性能，但如果HBase服务器宕机可能会造成数据丢失。	否
匹配数据类型	● 是：源端数据库中的Short、Int、Long、Float、Double、Decimal类型列的数据，会转换为Byte[]数组（二进制）写入HBase，其他类型的按字符串写入。如果这几种类型中，有合并做rowkey的，则依然当字符串写入。 该功能作用是：降低存储占用空间，存储更高效；特定场景下rowkey分布更均匀。 ● 否：源端数据库中所有类型的数据，都会按照字符串写入HBase。	否

5.4.4 配置 Hive 目的端参数

作业中目的连接为**Hive连接**时，目的端作业参数如**表5-32**所示。

表 5-32 Hive 作为目的端时的作业参数

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

参数名	说明	取值样例
自动创表	只有当源端为关系数据库时，才有该参数。表示写入表数据时，用户选择的操作： ● 不自动创建：不自动建表。 ● 不存在时创建：当目的端的数据库没有“表名”参数中指定的表时，CDM会自动创建该表。如果“表名”参数配置的表已存在，则不创建，数据写入到已存在的表中。 ● 先删除后创建：CDM先删除“表名”参数中指定的表，然后再重新创建该表。 **说明** ● 自动建表只同步列注释，表注释不会被同步。 ● 自动建表不支持同步主键。	不自动创建
导入前清空数据	选择目的端表中数据的处理方式： ● 是：任务启动前会清除目标表中数据。 ● 否：导入前不清空目标表中的数据，如果选“否”且表中有数据，则数据会追加到已有的表中。	是
待清空分区	“导入前清空数据”设置为“是”时，呈现此参数。 填写待清空分区信息后，表示清空该分区的数据。	单分区：year=2020,location=sun; 多分区：['year=2020,location=sun', 'year=2021,location=earth'].
执行Analyze语句	数据全部写入完成后会异步执行ANALYZE TABLE语句，用于优化Hive表查询速度，执行的SQL如下： ● 非分区表：ANALYZE TABLE tablename COMPUTE STATISTICS ● 分区表：ANALYZE TABLE tablename PARTITION(partcol1[=val1], partcol2[=val2], ...) COMPUTE STATISTICS **说明** “执行Analyze语句”参数配置仅用于单表迁移场景。	是

说明

- Hive作为目的端时，会自动创建存储格式为ORC的表。
- 由于文件格式限制，当前仅支持ORC与Parquet格式写入复杂类型。
- 源端Hive包含array和map类型时，目的端表格式只支持ORC和parquet复杂类型。若目的端表格式为RC和TEXT时，会对源数据进行处理，支持成功写入。
- 因map类型为无序的数据结构，迁移到目的端的数据类型可能跟源端顺序不一致。
- Hive作为迁移的目的时，如果存储格式为Textfile，在Hive创建表的语句中需要显式指定分隔符。例如：

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

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

当作业将数据导入到MySQL/SQL Server/PostgreSQL时，目的端作业参数如表5-33所示。

表 5-33 MySQL、SQL Server、PostgreSQL 作为目的端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	模式或表空间	待写入数据的数据库名称，支持自动创建Schema。单击输入框后面的按钮可选择模式或表空间。	schema

参数类型	参数名	说明	取值样例
	自动创表	只有当源端为关系数据库时，才有该参数。表示写入表数据时，用户选择的操作： ● 不自动创建：不自动建表。 ● 不存在时创建：当目的端的数据库没有“表名”参数中指定的表时，CDM会自动创建该表。如果“表名”参数配置的表已存在，则不创建，数据写入到已存在的表中。 ● 先删除后创建：CDM先删除“表名”参数中指定的表，然后再重新创建该表。	不自动创建
	表名	写入数据的目标表名，单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
	导入开始前	导入数据前，选择是否清除目的表的数据： ● 不清除：写入数据前不清除目标表中数据，数据追加写入。 ● 清除全部数据：写入数据前会清除目标表中数据。 ● 清除部分数据：需要配置“where条件”参数，CDM根据条件选择性删除目标表的数据。	清除部分数据
	where条件	**“导入开始前”**参数选择为“清除部分数据”时配置，配置后导入前根据where条件删除目的表的数据。	age > 18 and age <= 60
	约束冲突处理	导入数据到云数据库 MySQL且当迁移数据出现冲突时的处理方式。 ● insert into：当存在主键、唯一性索引冲突时，数据无法写入并将以脏数据的形式存在。 ● replace into：当存在主键、唯一性索引冲突时，会先删除原有行、再插入新行，替换原有行的所有字段。 ● on duplicate key update，当存在主键、唯一性索引冲突时，目的表中约束冲突的行除开唯一约束列的其他数据列将被更新。	insert into

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

5.4.6 配置 Oracle 目的端参数

作业中目的连接为**Oracle数据库连接**时，目的端作业参数如**表5-34**所示。

表 5-34 Oracle 作为目的端时的作业参数

参数类型	参数名	说明	取值样例
基本参数	模式或表空间	待写入数据的数据库名称，支持自动创建Schema。单击输入框后面的按钮可选择模式或表空间。	schema
	表名	写入数据的目标表名，单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
	导入开始前	导入数据前，选择是否清除目的表的数据： ● 不清除：写入数据前不清除目标表中数据，数据追加写入。 ● 清除全部数据：写入数据前会清除目标表中数据。 ● 清除部分数据：需要配置“where条件”参数，CDM根据条件选择性删除目标表的数据。	清除部分数据
	where条件	**“导入开始前”**参数选择为“清除部分数据”时配置，配置后导入前根据where条件删除目的表的数据。	age > 18 and age <= 60
高级参数	先导入阶段表	如果选择“是”，则启用事务模式迁移，CDM会自动创建临时表，先将数据导入到该临时表，导入成功后再通过数据库的事务模式将数据迁移到目标表中，导入失败则将目的表回滚到作业开始之前的状态，具体请参见**事务模式迁移**。 默认为“否”，CDM直接将数据导入到目的表。这种情况下如果作业执行失败，已经导入到目标表中的数据不会自动回滚。 **说明** 如果“导入开始前”选择“清除部分数据”或“清除全部数据”，CDM的事务模式不会回滚已经删除的数据。	否
	导入前准备语句	执行任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句。	create temp table
	导入后完成语句	执行任务之后执行的SQL语句，目前仅允许执行一条SQL语句。	merge into

参数类型	参数名	说明	取值样例
	loader线程数	每个loader内部启动的线程数，可以提升写入并发数。 **说明** 不支持“约束冲突处理”策略为“replace into”或“on duplicate key update”的并发场景。	1

5.4.7 配置 DWS 目的端参数

作业中目的连接为**DWS连接**时，目的端作业参数如**表5-35**所示。

表 5-35 目的端为 DWS 时的作业参数

参数名	说明	取值样例
模式或表空间	待写入数据的数据库名称，支持自动创建Schema。单击输入框后面的按钮可选择模式或表空间。	schema
自动创表	只有当源端为关系数据库时，才有该参数。表示写入表数据时，用户选择的操作： • 不自动创建：不自动建表。 • 不存在时创建：当目的端的数据库没有“表名”参数中指定的表时，CDM会自动创建该表。如果“表名”参数配置的表已存在，则不创建，数据写入到已存在的表中。 • 先删除后创建：CDM先删除“表名”参数中指定的表，然后再重新创建该表。 当选择在DWS端自动创表时，DWS的表与源表的字段类型映射关系见**在DWS端自动建表时的字段类型映射**。 **说明** 自动建表只同步列注释，表注释不会被同步。	不自动创建
表名	写入数据的目标表名，单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
是否压缩	导入数据到DWS且选择自动创表时，用户可以指定是否压缩存储。	否

参数名	说明	取值样例
存储模式	导入数据到DWS且选择自动创表时，用户可以指定存储模式： ● 行模式：表的数据将以行式存储，适用于点查询（返回记录少，基于索引的简单查询），或者增删改比较多的场景。 ● 列模式：表的数据将以列式存储，适用于统计分析类查询（group、join多的场景），或者即席查询（查询条件不确定，行模式表扫描难以使用索引）的场景。	行模式
导入模式	导入数据到DWS时，用户可以指定导入模式： ● COPY模式，源数据经过管理节点后，复制到DWS的DataNode节点。 ● UPSERT模式，数据发生主键或唯一约束冲突时，更新除了主键和唯一约束列的其他列数据。	COPY
导入开始前	导入数据前，选择是否清除目的表的数据： ● 不清除：写入数据前不清除目标表中数据，数据追加写入。 ● 清除全部数据：写入数据前会清除目标表中数据。 ● 清除部分数据：需要配置“where条件”参数，CDM根据条件选择性删除目标表的数据。	清除部分数据
where条件	**“导入开始前”**参数选择为“清除部分数据”时配置，配置后导入前根据where条件删除目的表的数据。	age > 18 and age <= 60
先导入阶段表	如果选择“是”，则启用事务模式迁移，CDM会自动创建临时表，先将数据导入到该临时表，导入成功后再通过数据库的事务模式将数据迁移到目标表中，导入失败则将目的表回滚到作业开始之前的状态。 默认为“否”，CDM直接将数据导入到目的表。这种情况下如果作业执行失败，已经导入到目标表中的数据不会自动回滚。 **说明** 如果“导入开始前”选择“清除部分数据”或“清除全部数据”，CDM的事务模式不会回滚已经删除的数据。	否

参数名	说明	取值样例
扩大字符字段长度	当选择自动创表时，迁移过程中可将字符类型的字段长度扩大为原来的3倍，再写入到目的表中。如果源端数据库与目的端数据库字符编码不一样，但目的表字符类型字段与源表一样，在迁移数据时，可能会有出现长度不足的错误。 应用场景主要是将有中文内容的字符字段导入到DWS时，需要自动将字符长度放大3倍。 在导入中文内容的字符到DWS时，如果作业执行失败，且日志中出现类似“value too long for type character varying”的错误，则可以通过启用该功能解决。 **说明** 当启动该功能时，也会导致部分字段消耗用户相应的3倍存储空间。	否
使用非空约束	当选择自动创建目的表时，如果选择使用非空约束，则目的表字段的是否非空约束，与原表具有相应非空约束的字段保持一致。	是
导入前准备语句	执行任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句。	create temp table
导入后完成语句	执行任务之后执行的SQL语句，目前仅允许执行一条SQL语句。	merge into
loader线程数	每个loader内部启动的线程数，可以提升写入并发数。	1

在 DWS 端自动建表时的字段类型映射

CDM在数据仓库服务（Data Warehouse Service，简称DWS）中自动建表时，DWS的表与源表的字段类型映射关系如图5-8所示。例如使用CDM将Oracle整库迁移到DWS，CDM在DWS上自动建表，会将Oracle的**NUMBER(3,0)**字段映射到DWS的**SMALLINT**。

图 5-8 自动建表的字段映射

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

说明

自动建表场景不支持创建索引。

5.4.8 配置 DDS 目的端参数

作业中目的连接为**DDS连接**时，即导入数据到文档数据库服务（DDS）时，目的端作业参数如**表5-36**所示。

表 5-36 DDS 作为目的端时的作业参数

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

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

作业中目的连接为**配置Elasticsearch连接**或**配置云搜索服务（CSS）连接**时，即将数据导入到Elasticsearch/云搜索服务（CSS）时，目的端作业参数如**表5-37**所示。

须知

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

表 5-37 Elasticsearch/云搜索服务（CSS）作为目的端时的作业参数

参数名	说明	取值样例
索引	待写入数据的Elasticsearch的索引，类似关系数据库中的数据库名称。CDM支持自动创建索引和类型，索引和类型名称只能全部小写，不能有大写。	index
类型	待写入数据的Elasticsearch的类型，类似关系数据库中的表名称。类型名称只能全部小写，不能有大写。 **说明** Elasticsearch搜索引擎7.x及以上版本不支持自定义类型，只能使用_doc类型。此处即使自定义也不会生效。	type
管道ID	该参数用于数据传到Elasticsearch后，通过Elasticsearch的数据转换pipeline进行数据格式变换。 目的端为Elasticsearch时需要先在kibana中创建管道ID。 目的端为CSS时不需要创建管道ID，此参数填写配置文件名称，默认为name。	目的端为Elasticsearch时：pipeline_id 目的端为CSS时：name（name为配置文件名称）
开启路由	开启路由后，支持指定某一列的值作为路由写入Elasticsearch。 **说明** 开启路由前建议先建好目的端索引，可提高查询效率。	否
路由字段	**“开启路由”**参数选择为“是”时配置，用于配置目的端路由字段。目的端索引存在但是获取不到字段信息时，支持手动填写字段。路由字段允许为空，为空时写入Elasticsearch不指定routing值。	value1

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

5.4.10 配置 DLI 目的端参数

作业中目的连接为**DLI连接**时，即将数据导入到数据湖探索服务（DLI）时，目的端作业参数如**表5-38**所示。

说明

使用CDM服务迁移数据到DLI时，DLI要在OBS的*dli-trans**内部临时桶生成数据文件，因此在需要赋予使用AK/SK对应的账号对*dli-trans**桶的读、写、创建目录对象等权限，OBS权限策略添加请参见**新增OBS桶授权策略**。

表 5-38 DLI 作为目的端时的作业参数

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

参数名	说明	取值样例
空字符串作为null	如果设置为true，空字符串将作为null。	否
清空数据方式	导入前清空数据，如果设置为true时，呈现此参数。 TRUNCATE：删除标准数据。 INSERT_OVERWRITE：新增数据插入，同主键数据覆盖。 **说明** 当源端为Kafka时，如果DLI导入前清空数据，则不支持INSERT_OVERWRITE。	TRUNCATE
分区	“导入前清空数据”设置为“是”时，呈现此参数。 填写分区信息后，表示清空该分区的数据。	year=2020,location=sun

新增 OBS 桶授权策略

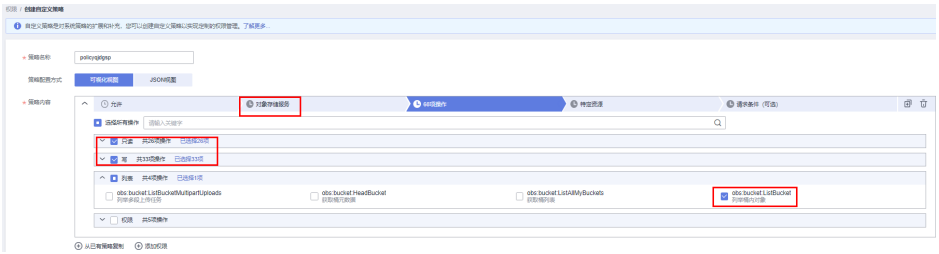
步骤1 登录统一身份认证服务控制台。

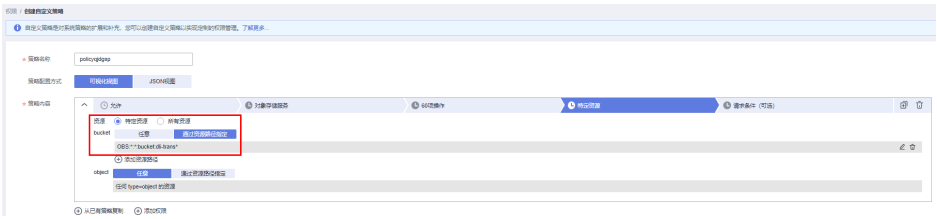
步骤2 在左侧导航窗格中，选择“权限管理>权限”页签，单击右上方的“创建自定义策略”。

图 5-9 创建自定义策略

步骤3 输入策略名称并选择对象存储服务后，配置策略内容，如图5-10所示。

图 5-10 配置策略内容





步骤4 填写策略描述后单击“确定”，完成对象存储服务自定义策略创建。

----结束

5.4.11 配置 MRS Hudi 目的端参数

作业中目的连接为**MRS Hudi连接**时，目的端作业参数如**表5-39**所示。

表 5-39 MRS Hudi 作为目的端时的作业参数

通用配置		
配置项	配置说明	推荐配置
目的连接名称	选择已配置的MRS Hudi连接。	hudi_to_cdm
数据库名称	输入或选择写入数据的数据库名称。单击输入框后面的按钮可进入数据库选择界面。	dbadmin
表名	单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	cdm
自动创表	是否自动创建Hudi表。 ● 不自动创建：不自动建表。 ● 不存在时创建：当目的端的数据库没有“表名”参数中指定的表时，CDM会自动创建该表。如果“表名”参数配置的表已存在，则不创建，数据写入到已存在的表中。	不自动创表
导入前清空数据	选择目的端表中数据的处理方式： ● 是：任务启动前会清除目标表中数据。 ● 否：导入前不清空目标表中的数据，如果选“否”且表中有数据，则数据会追加到已有的表中。	否

通用配置		
全量模式写Hoodie	选择写Hoodie模式，默认选“是”表示全量模式，“否”表示微批模式。 ● 全量模式为异步分片写入Hoodie，适用于一次全量写入场景。 ● 微批模式为异步分批写入Hoodie，适用于对入库时间SLA要求较为严格的场景，以及对资源消耗较小，对MOR表存储类型在线进行压缩的场景。 **说明** 运行-失败重试期间不允许修改此模式。	是
批次数据大小	“全量模式写Hoodie”设置为“否”时，使用微批模式呈现此参数。 用于设置单个批次写Hoodie的数据行数，默认100000行。	100000
使用入库时间字段	将一个字段标记为入库时间字段，自动建表时将此字段自动加到建表语句中，写入Hudi时将把此字段的值替换为当前时间，不自动建表时选择已经存在的入库时间字段。	是
入库时间字段名称	“使用入库时间字段”设置为“是”时，呈现此参数。 用于记录写入Hudi的时间。 **说明** ● 对于已存在目的端表中带有入库时间字段的，可以直接使用已有的timestamp类型字段。 ● 对于自动建表的场景，该字段会被拼接到建表语句中，类型为timestamp，该字段名称不能与源端的字段有重复（包括自定义字段）。	cdc_last_update_date
Hudi建表配置		
Location	存储在OBS或HDFS上数据库表的文件路径。	-
Hudi表类型	Hudi表存储类型。 ● MOR表：数据先写入avro格式的日志文件，读取时合并到parquet文件。 ● COW表：数据直接写入parquet文件。	MOR
Hudi表主键	对Hudi建表设置主键，多个值以逗号隔开。	-
Hudi表生成器类	主键生成类型，实现org.apache.hudi.keygen.KeyGenerator从传入记录中提取键值。	-

通用配置		
Hudi表预聚合键	对Hudi建表设置预聚合键，当两个记录拥有相同的主键时，保留precombine字段值较大的记录。 **说明** 如果没有时间字段，可以设置和主键一样的字段，当遇到主键冲突时，保留最新的记录。	ts
Hudi表分区字段	对Hudi建表设置分区字段，多个值以逗号隔开。	-
Hudi表压缩策略（是否开启写入压缩）	在线进行压缩，仅对MOR表生效。	是
Hudi表清除策略（保留提交数）	清除时保留的提交数。	1
Hudi表归档策略（最小保留提交数）	归档时保留的最小提交数。	1
Hudi表归档策略（最大保留提交数）	归档时保留的最大提交数。	100
Hudi表配置	对Hudi建表设置自定义参数属性，此处填入的参数将会在options中生效。例如：主键、combineKey、索引。	-

5.4.12 配置 MRS ClickHouse 目的端参数

作业中目的连接为**MRS ClickHouse连接**时，目的端作业参数如**表5-40**所示。

说明

当作业源端为MRS ClickHouse、DWS及Hive时：

- 若int及float类型字段为null时，创建MRS ClickHouse表格时字段类型需设置为nullable()，否则写入到MRS ClickHouse的值会为0。
- 请确认目的端表引擎是否为ReplicatedMergeTree引擎，该引擎自带去重机制，且去重数据不能准确预测，选用该引擎应保证数据唯一性，否则会造成不唯一数据被忽略写入，或尝试替换其他表引擎，例如MergeTree。

表 5-40 MRS ClickHouse 作为目的端时的作业参数

参数名	说明	取值样例
模式或表空间	单击输入框后面的按钮可选择模式或表空间。	schema

参数名	说明	取值样例
表名	输入或选择写入数据的目标表名。 单击输入框后面的按钮可进入表的选择界面。 该参数支持配置为时间宏变量，且一个路径名中可以有多个宏定义变量。使用时间宏变量和定时任务配合，可以实现定期同步新增数据，详细说明请参见**使用时间宏变量完成增量同步**。 **说明** 如果配置了时间宏变量，通过DataArts Studio数据开发调度CDM迁移作业时，系统会将时间宏变量替换为“数据开发作业计划启动时间-偏移量”，而不是“CDM作业实际启动时间-偏移量”。	table
导入开始前	导入数据前，选择是否清除目的表的数据： • 不清除：写入数据前不清除目标表中数据，数据追加写入。 • 清除全部数据：写入数据前会清除目标表中数据。 • 清除部分数据：需要配置“where条件”参数，CDM根据条件选择性删除目标表的数据。	清除部分数据
是否在集群操作	**“导入开始前”**参数选择为“清除部分数据”或“清除全部数据”时，显示该参数。如果设置为是，将对集群中的所有节点进行全部/部分数据清除操作。	是
where条件	**“导入开始前”**参数选择为“清除部分数据”时配置，配置后导入前根据where条件删除目的表的数据。	age > 18 and age <= 60

5.4.13 配置 MongoDB 目的端参数

作业中目的连接为**MongoDB连接**时，目的端作业参数如**表5-41**所示。

表 5-41 MongoDB 作为目的端时的作业参数

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

参数名	说明	取值样例
迁移行为	将记录迁移到MongoDB目的端时，选择需要进行的插入行为操作。 ● 新增：将文件记录直接插入指定的集合。 ● 有则新增，无则替换：以指定的过滤键作为查询条件。如果在集合中找到匹配的记录，则替换该记录（找到多条匹配记录时，只会替换找到的第一条记录）。如果不存在，则添加新记录。 ● 替换：使用指定的过滤键作为查询条件。如果在集合中找到匹配的记录，则替换该记录（找到多条匹配记录时，只会替换找到的第一条记录）。如果没有，则不会添加新记录。	新增
导入前准备语句	执行任务前需要先执行的MongoDB查询语句。 **说明** ● “导入前准备语句”格式是json，只有两个键值对，第一个键值对是配置操作类别，key是"type"，value只支持"remove"和"drop"。第二个键值对是针对不同操作类别，需要配置的数据条件或者集合名称。 ● 导入前准备语句的执行不会影响即将写入的数据内容。	{"type":"remove","json":"{$or:[{Pid:{$gt:'0',$lt:'2'}},{X:{$gt:'50',$lt:'80'}}]}"}

5.5 配置字段映射

操作场景

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

约束限制

- 作业源端开启“使用SQL语句”参数时不支持配置转换器。
- 如果在字段映射界面，CDM通过获取样值的方式无法获得所有列（例如从HBase/CloudTable/MongoDB导出数据时，CDM有较大概率无法获得所有列），则可以单击后选择“添加新字段”来手动增加，确保导入到目的端的数据完整。
- 关系数据库、Hive、MRS Hudi及DLI做源端时，不支持获取样值功能。
- SQLServer作为目的端数据源时，不支持timestamp类型字段的写入，需修改为其他时间类型字段写入（如datetime）。

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

新增字段

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

图 5-11 字段映射

目前支持以下类型自定义字段：

- **常量**

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

- **变量**

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

- **表达式**

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

新建转换器

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

图 5-12 新建转换器

CDM可以在迁移过程中对字段进行转换，目前支持以下字段转换器：

- **脱敏**
  隐藏字符串中的关键信息，例如要将“12345678910”转换为“123****8910”，则配置如下：
  - “起始保留长度”为“3”。
  - “结尾保留长度”为“4”。
  - “替换字符”为“*”。
- **去前后空格**
  自动去字符串前后的空值，不需要配置参数。
- **字符串反转**
  自动反转字符串，例如将“ABC”转换为“CBA”，不需要配置参数。
- **字符串替换**
  替换字符串，需要用户配置被替换的对象，以及替换后的值。
- **去换行**
  将字段中的换行符（\n、\r、\r\n）删除。
- **表达式转换**
  数据进行转换过程中，替换内容包含特殊字符时，需要先使用\将该字符转义成普通字符。
  - 表达式支持以下两个环境变量：
    - value：当前字段值。
    - row：当前行，数组类型。
  - 表达式支持的工具类用法罗列如下，未列出即表示不支持：
    i. 如果当前字段为字符串类型，将字符串全部转换为小写，例如将“aBC”转换为“abc”。

表达式：StringUtils.lowerCase(value)

ii. 将当前字段的字符串全部转为大写。

表达式：StringUtils.upperCase(value)

iii. 如果想将第1个日期字段格式从“2018-01-05 15:15:05”转换为“20180105”。

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

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

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

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

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

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

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

vii. 如果当前字段值为数值类型，转换后值为当前值的两倍。

表达式：value*2

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

表达式：value=="true"?"Y":"N"

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

表达式：empty value? "Default":value

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

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

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

表达式：CommonUtils.randomUUID()

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

表达式：StringUtils.capitalize(value)

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

表达式：StringUtils.uncapitalize(value)

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

表达式：StringUtils.center(value,*4*)

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

表达式：StringUtils.chomp(value)

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

表达式：StringUtils.contains(value,"*a*")

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

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

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

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

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

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

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

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

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

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

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

表达式：StringUtils.equals(value,"*ABC*")

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

表达式：StringUtils.indexOf(value,"*ab*")

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

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

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

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

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

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

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

表达式：StringUtils.isAlpha(value)

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

表达式：StringUtils.isAlphanumeric(value)

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

表达式：StringUtils.isAlphanumericSpace(value)

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

表达式：StringUtils.isAlphaSpace(value)

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

表达式：StringUtils.isAsciiPrintable(value)

xxxii. 如果字符串为空或null，返回布尔值true，否则返回false。

表达式：StringUtils.isEmpty(value)

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

表达式：StringUtils.isNumeric(value)

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

表达式：StringUtils.left(value,*2*)

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

表达式：StringUtils.right(value,*2*)

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

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

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

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

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

表达式：StringUtils.length(value)

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

表达式：StringUtils.remove(value,"*ue*")

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

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

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

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

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

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

替换内容包含特殊字符时，需要先把该字符转义成普通字符，例如，客户想通过该表达式把字符串中 \t 去掉时，需要配置为：StringUtils.replace(value,"\\t","")（即把 \ 再次转义）。

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

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

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

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

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

表达式：StringUtils.strip(value,"*xyz*b")

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

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

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

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

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

表达式：StringUtils.substring(value,*2*)

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

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

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

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

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

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

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

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

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

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

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

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

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

表达式：StringUtils.trim(value)

lvi. 将当前字符串转换为字节，如果转换失败，则返回0。

表达式：NumberUtils.toByte(value)

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

表达式：NumberUtils.toByte(value,*1*)

lviii. 将当前字符串转换为Double数值，如果转换失败，则返回0.0d。

表达式：NumberUtils.toDouble(value)

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

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

lx. 将当前字符串转换为Float数值，如果转换失败，则返回0.0f。

表达式：NumberUtils.toFloat(value)

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

表达式：NumberUtils.toFloat(value,*1.1f*)

lxii. 将当前字符串转换为Int数值，如果转换失败，则返回0。

表达式：NumberUtils.toInt(value)

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

表达式：NumberUtils.toInt(value,*1*)

lxiv. 将字符串转换为Long数值，如果转换失败，则返回0。

表达式：NumberUtils.toLong(value)

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

表达式：NumberUtils.toLong(value,*1L*)

lxvi. 将字符串转换为Short数值，如果转换失败，则返回0。

表达式：NumberUtils.toShort(value)

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

表达式：NumberUtils.toShort(value,*1*)

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

表达式：CommonUtils.ipToLong(value)

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

表达式：HttpsUtils.downloadMap("*url*")

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

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

lxxi. 取出缓存的IP与地址映射对象。

表达式：CommonUtils.getCache("*ipList*")

lxxii.判断是否有IP与地址映射缓存。

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

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

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

lxxiv如果value值为空或者null时，则返回字符串“aaa”，否则返回value。

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

特殊链路说明

- 当源端为DLI，目的端为DWS时，DLI的tinyint类型字段映射为DWS的smallint类型字段。
- 当源端为Hudi，目的端为DWS时，Hudi的Double类型字段映射为DWS的Float类型字段。

5.6 配置定时任务

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

说明

- CDM在配置定时作业时，不要为大量任务设定相同的定时时间，应该错峰调度，避免出现异常。
- 如果通过DataArts Studio数据开发调度CDM迁移作业，此处也配置了定时任务，则两种调度均会生效。为了业务运行逻辑统一和避免调度冲突，推荐您启用数据开发调度即可，无需配置CDM定时任务。
- 定时任务功能原理：采用Java Quartz定时器，类似Cron表达式配置。对起始时间解析出分，小时，天，月。构造出cronb表达式。

  以配置天调度为例：重复周期选择1天：若当前时间2022/10/14 12:00，配置起始时间为2022/10/14 00:00。任务在2022/10/15 00:00执行；若当前时间2022/10/14 12:00，配置起始时间为2022/10/15 00:00。任务在2022/10/15 00:00执行。

  重复周期选择2天：若当前时间2022/10/14 12:00，配置起始时间为2022/10/14 00:00。任务在2022/10/16 00:00执行；若当前时间2022/10/14 12:00，配置起始时间为2022/10/15 00:00。任务在2022/10/16 00:00执行。

分钟

CDM支持配置每几分钟执行一次作业，定时任务周期不建议小于5分钟。

- 开始时间：表示定时配置生效的时间，也是第一次自动执行作业的时间。
- 重复周期（分）：从开始时间起，每多少分钟执行一次作业。

- 结束时间：该参数为可选参数，如果不配置则表示一直自动执行。如果配置了结束时间，则会在该时间停止自动执行作业。

图 5-13 重复周期为分钟

例如上图表示：从2023年1月1日0时0分开始第一次自动执行作业，每30分钟自动执行一次，2023年12月31日23时59分之后不再自动执行。

小时

CDM支持配置每几小时执行一次作业。

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

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

图 5-14 重复周期为小时

例如上图表示：定时配置从2023年1月1日0时0分生效，0:10时开始第一次自动执行作业，0:30第二次，0:50第三次，以后每2小时重复三次，2023年12月31日23时59分之后不再自动执行。

天

CDM支持配置每几天执行一次作业。

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

图 5-15 重复周期为天

例如上图表示：从2023年1月1日0时0分开始第一次自动执行，每3天自动执行一次，配置一直有效。

周

CDM支持配置每几周执行一次作业。

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

图 5-16 重复周期为周

例如上图表示：在2023年1月1日0时0分以后，每2周的周二、周六、周日的0时0分，便自动执行作业，直到2023年12月31日23时59分不再自动执行。

月

CDM支持配置每几月执行一次作业。

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

图 5-17 重复周期为月

例如上图表示：从2023年1月1日0点开始，每月5日、25日的0点自动执行作业，直到2023年12月31日23时59分不再自动执行。

5.7 作业配置管理

CDM作业管理界面的“配置管理”页签，主要操作如下：

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

最大抽取并发数

最大抽取并发数即集群最大抽取并发数。

说明

此处的“最大抽取并发数”参数与集群配置处的“最大抽取并发数”参数同步，在任意一处修改即可生效。

CDM通过数据迁移作业，将源端数据迁移到目的端数据源中。其中，主要运行逻辑如下：

1. 数据迁移作业提交运行后，CDM会根据作业配置中的“抽取并发数”参数，将每个作业拆分为多个Task，即作业分片。

   说明

   不同源端数据源的作业分片维度有所不同，因此某些作业可能出现未严格按作业“抽取并发数”参数分片的情况。

2. CDM依次将Task提交给运行池运行。根据集群配置管理中的“最大抽取并发数”参数，超出规格的Task排队等待运行。

因此作业抽取并发数和集群最大抽取并发数参数设置为适当的值可以有效提升迁移速度，您可参考下文有效配置抽取并发数。

1. 集群最大抽取并发数的上限建议为vCPU核数*2，如表5-42所示。

   表 5-42 集群最大抽取并发数配置建议

规格名称	vCPUs/内存	集群并发数上限参考
cdm.large	8核 16GB	16
cdm.xlarge	16核 32GB	32
cdm.4xlarge	64核 128GB	128

2. 作业抽取并发数的配置原则如下：
   a. 迁移的目的端为文件时，CDM不支持多并发，此时应配置为单进程抽取数据。
   b. 表中每行数据大小为1MB以下的可以设置多并发抽取，超过1MB的建议单线程抽取数据。
   c. 作业抽取并发数可参考集群最大抽取并发数配置，但不建议超过集群最大抽取并发数上限。
   d. 目的端为DLI数据源时，抽取并发数建议配置为1，否则可能会导致写入失败。

定时备份/恢复

该功能依赖于OBS服务。

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

  表 5-43 定时备份参数

参数	说明	配置样例
定时备份	自动备份功能的开关，该功能只备份作业，不会备份连接。	开
备份策略	• 所有作业：不管作业处于什么状态，CDM会备份所有表/文件迁移作业、整库迁移的作业。不备份历史作业。 • 分组作业：选择备份某一个或多个分组下的作业。	所有作业
备份周期	选择备份周期： • 日：每天零点执行一次。 • 周：每周一零点执行一次。 • 月：每月1号零点执行一次。	日

参数	说明	配置样例
备份写入OBS连接	CDM通过该连接，将作业备份到OBS，需要用户提前在“连接管理”界面创建好OBS连接。	obslink
OBS桶	存储备份文件的OBS桶。	cdm
备份数据目录	存储备份文件的目录。	/cdm-bk/

- 恢复作业

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

  您可以单击备份列表操作列的“恢复备份”来恢复CDM作业。

作业参数的环境变量

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

这里以批量替换作业中OBS桶名为例进行介绍。

1. 在CDM作业管理界面，单击“配置管理”页签，配置环境变量。

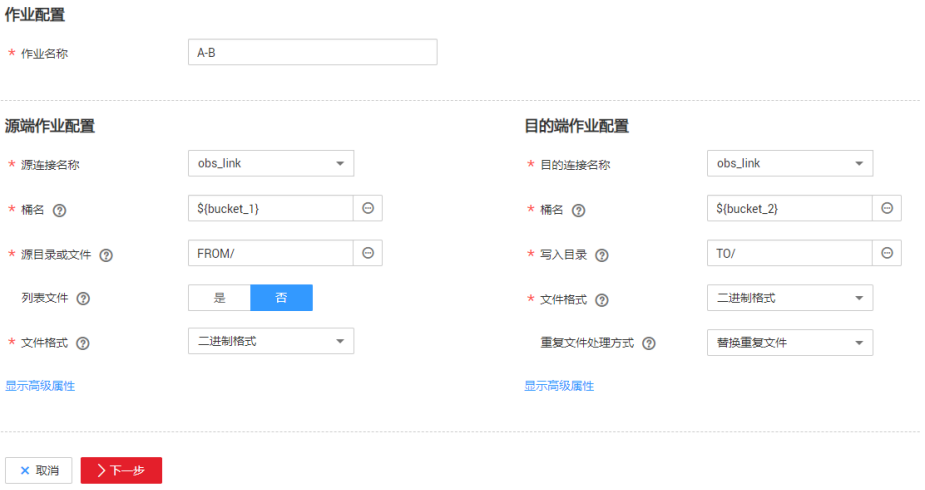
   ```
   bucket_1=A
   bucket_2=B
   ```

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

2. 在创建CDM迁移作业的界面，迁移桶A的数据到桶B。

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

   图 5-18 桶名配置为环境变量

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

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

5.8 管理单个作业

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

查看

- **查看作业状态**

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

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

- **查看历史记录**

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

- **查看作业日志**

  在历史记录界面可查看作业所有的日志。

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

- **查看作业JSON**

  直接编辑作业的JSON文件，作用等同于修改作业的参数配置。

- **源目的统计查询**

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

修改

- **修改作业参数**

  可重新配置作业参数，但是不能重新选择源连接和目的连接。

- **编辑作业JSON**

  直接编辑作业的JSON文件，作用等同于修改作业的参数配置。

操作步骤

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

步骤2 单击“历史作业”可以查看最近1个月所有执行过的历史作业。

步骤3 单击“表/文件迁移”显示作业列表，可对单个作业执行如下操作：

- 修改作业参数：单击作业操作列的“编辑”可修改作业参数。
- 运行作业：单击作业操作列的“运行”可手动启动作业。
- 查看历史记录：单击作业操作列的“历史记录”进入历史记录界面，可查看该作业的历史执行记录、读取和写入的统计数据。在历史记录界面单击“日志”，可查看作业执行的日志信息。
- 删除作业：选择作业操作列的“更多 > 删除”可删除作业。

- 停止作业：选择作业操作列的“更多 > 停止”可停止作业。
- 查看作业JSON：选择作业操作列的“更多 > 查看作业JSON”，可查看该作业的JSON定义。
- 编辑作业JSON：选择作业操作列的“更多 > 编辑作业JSON”，可直接编辑该作业的JSON文件，作用等同于修改作业的参数配置。
- 配置定时任务：选择作业操作列的“更多 > 配置定时任务”，可选择在有效期内周期性启动作业，具体请参考**配置定时任务**。
- 日志：选择作业操作列的“更多 > 日志”，可查看该作业最近的一次日志。

  也可以在历史记录界面可查看作业所有的日志。
- 失败重试：选择作业操作列的“更多 > 失败重试”，可以对执行失败的作业，选择自动重试三次或者不重试。

步骤4 修改完成后单击“保存”或“保存并运行”。

----结束

5.9 批量管理作业

操作场景

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

- 作业分组管理
- 批量运行作业
- 批量删除作业
- 批量导出作业
- 批量导入作业

批量导出、导入作业的功能，适用以下场景：

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

操作步骤

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

步骤2 单击“表/文件迁移”显示作业列表，提供以下批量操作：

- **作业分组**

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

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

- **批量运行作业**
  勾选一个或多个作业后，单击“运行”可批量启动作业。
- **批量删除作业**
  勾选一个或多个作业后，单击“删除”可批量删除作业。
- **批量导出作业**
  单击“导出”，弹出批量导出页面，如图5-19。

  图 5-19 批量导出页面

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

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

  说明

  由于安全原因，CDM导出作业时没有导出连接密码，连接密码全部使用“Add password here”替换。
- **批量导入作业**
  单击“导入”，选择JSON格式的文件导入或文本导入。
  - 文件导入：待导入的作业文件必须为JSON格式（大小不超过1M）。如果待导入的作业文件是之前从CDM中导出的，则导入前必须先编辑JSON文件，将“Add password here”替换为对应连接的正确密码，再执行导入操作。
  - 文本导入：无法正确上传本地JSON文件时可选择该方式。将作业的JSON文本直接粘贴到输入框即可。

说明

当前导入时不支持覆盖已有作业。

----结束

6 查看审计日志

6.1 如何查看审计日志

概述

云审计服务（Cloud Trace Service，CTS）可以记录CDM相关的操作事件，用于支撑安全分析、合规审计、资源管理和问题定位等常见应用场景。

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

前提条件

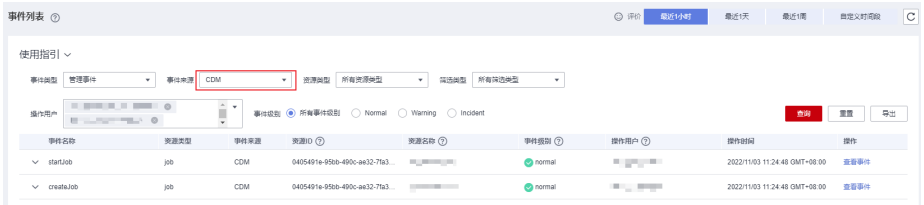
已开通云审计服务。开通方式请参见**开通云审计服务**。

操作步骤

1. 登录管理控制台，在服务列表中选择“云审计服务 CTS”，进入云审计服务控制台。
2. 在云审计服务控制台，默认展示事件列表，您可以通过筛选来查询对应的操作事件。

   其中，CDM的操作事件您可以在“事件来源”中筛选“CDM”进行查看。

   图 6-1 CDM 操作事件

3. 在需要查看的事件左侧，单击事件名称左侧箭头 ，展开该记录的详细信息。
4. 在需要查看的记录右侧，单击“查看事件”，弹窗中显示了该操作事件结构的详细信息。

   更多关于云审计的信息，请参见**云审计服务用户指南**。

6.2 支持云审计的关键操作

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

表 6-1 支持云审计的关键操作列表

操作名称	资源类型	事件名称
创建集群	cluster	createCluster
删除集群	cluster	deleteCluster
修改集群配置	cluster	modifyCluster
开机	cluster	startCluster
重启	cluster	restartCluster
导入作业	cluster	clusterImportJob
绑定弹性IP	cluster	bindEip
解绑弹性IP	cluster	unbindEip
创建连接	link	createLink
修改连接	link	modifyLink
测试连接	link	verifyLink
删除连接	link	deleteLink
创建任务	job	createJob
修改任务	job	modifyJob
删除任务	job	deleteJob
启动任务	job	startJob
停止任务	job	stopJob

7 关键操作指导

7.1 增量迁移原理介绍

7.1.1 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 7-1 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 7-2 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。

7.1.2 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7-3**所示，参数配置如下：

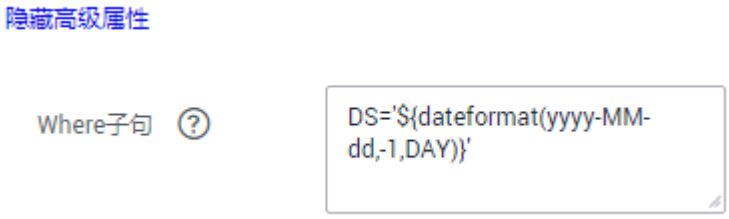
图 7-3 表数据

SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	s	2017-05-01
2	5	s	2017-05-01
3	1	g	2017-05-02
4	4	o	2017-05-02
5	6	a	2017-05-02
6	7	n	2017-05-02
7	1	g	2017-05-02
8	4	o	2017-05-02
9	6	a	2017-05-02
10	7	n	2017-05-02
11	2	fa	2017-10-15
12	3	te	2017-10-15
13	2	fa	2017-10-15
14	3	te	2017-10-15

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

图 7-4 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。

7.1.3 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 7-5 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。

7.1.4 MongoDB/DDS 增量迁移

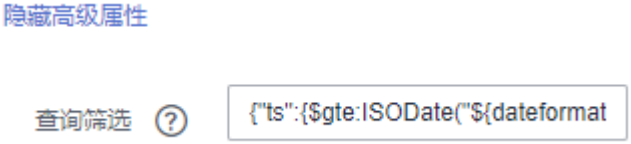
使用CDM导出MongoDB或者DDS的数据时，支持导出指定时间段内的数据，配合CDM的定时任务，可以实现MongoDB/DDS的增量迁移。

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 7-6 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。

7.2 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7-1**所示。

表 7-1 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7-7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。

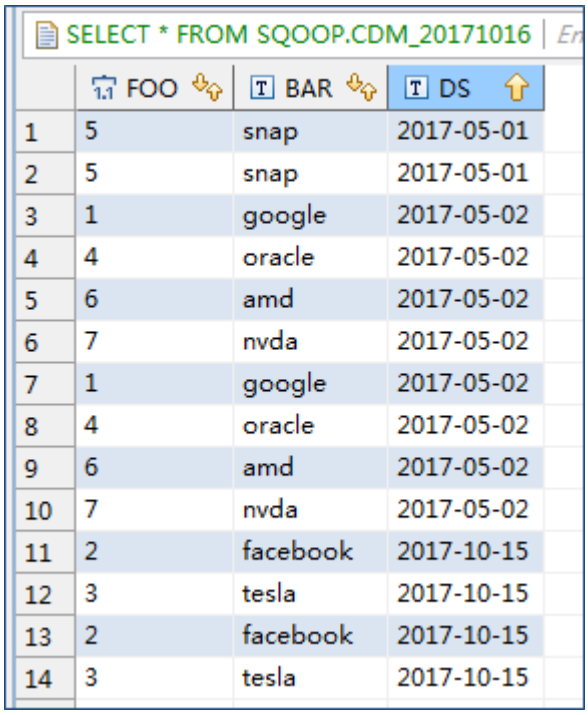
图 7-7 源表名和写入目录配置为时间宏变量

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7-8所示。

图 7-8 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。

7.3 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 7-9 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。

7.4 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。

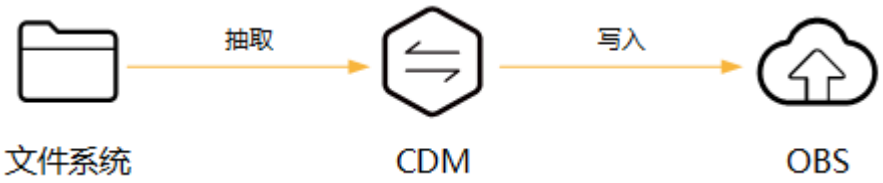
说明

- 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。

7.5 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7-10**所示。

图 7-10 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**。

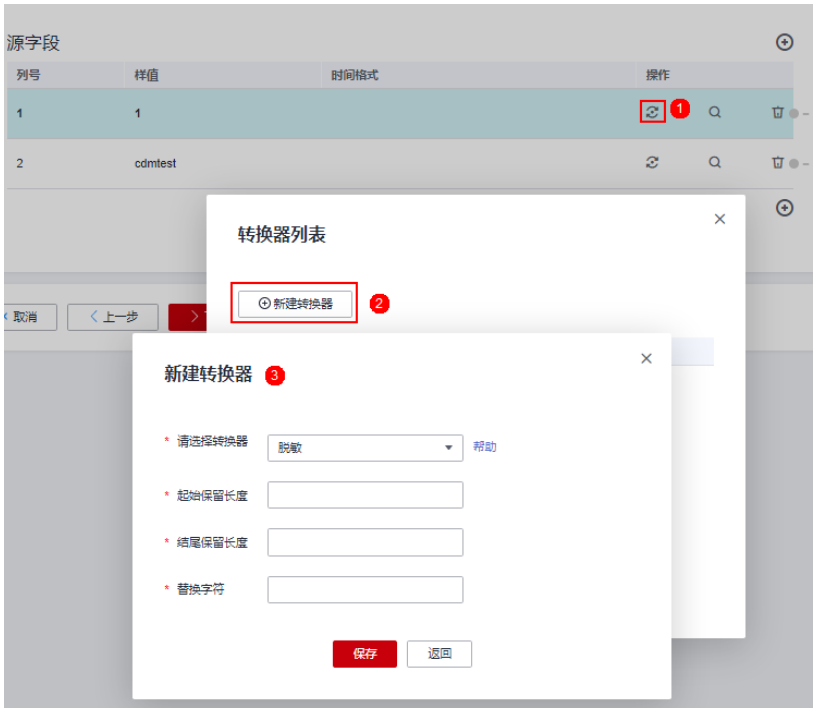
7.6 字段转换器配置指导

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 7-11 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 7-12 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 7-13 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。

数据进行转换过程中，替换内容包含特殊字符时，需要先使用\将该字符转义成普通字符。

- 表达式支持以下两个环境变量：
  - value：当前字段值。
  - row：当前行，数组类型。
- 表达式支持的工具类用法罗列如下，未列出即表示不支持：
  a. 如果当前字段为字符串类型，将字符串全部转换为小写，例如将“aBC”转换为“abc”。
     表达式：StringUtils.lowerCase(value)
  b. 将当前字段的字符串全部转为大写。
     表达式：StringUtils.upperCase(value)
  c. 如果想将第1个日期字段格式从“2018-01-05 15:15:05”转换为“20180105”。

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

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

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

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

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

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

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

g. 如果当前字段值为数值类型，转换后值为当前值的两倍。

表达式：value*2

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

表达式：value=="true"?"Y":"N"

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

表达式：empty value? "Default":value

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

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

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

表达式：CommonUtils.randomUUID()

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

表达式：StringUtils.capitalize(value)

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

表达式：StringUtils.uncapitalize(value)

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

表达式：StringUtils.center(value,*4*)

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

表达式：StringUtils.chomp(value)

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

表达式：StringUtils.contains(value,"*a*")

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

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

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

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

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

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

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

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

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

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

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

表达式：StringUtils.equals(value,"*ABC*")

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

表达式：StringUtils.indexOf(value,"*ab*")

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

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

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

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

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

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

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

表达式：StringUtils.isAlpha(value)

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

表达式：StringUtils.isAlphanumeric(value)

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

表达式：StringUtils.isAlphanumericSpace(value)

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

表达式：StringUtils.isAlphaSpace(value)

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

表达式：StringUtils.isAsciiPrintable(value)

af. 如果字符串为空或null，返回布尔值true，否则返回false。

表达式：StringUtils.isEmpty(value)

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

表达式：StringUtils.isNumeric(value)

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

表达式：StringUtils.left(value,*2*)

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

表达式：StringUtils.right(value,*2*)

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

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

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

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

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

表达式：StringUtils.length(value)

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

表达式：StringUtils.remove(value,"*ue*")

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

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

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

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

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

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

替换内容包含特殊字符时，需要先把该字符转义成普通字符，例如，客户想通过该表达式把字符串中 \t 去掉时，需要配置为：StringUtils.replace(value,"\\t","")（即把 \ 再次转义）。

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

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

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

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

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

表达式：StringUtils.strip(value,"*xyz*b")

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

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

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

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

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

表达式：StringUtils.substring(value,*2*)

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

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

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

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

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

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

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

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

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

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

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

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

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

表达式：StringUtils.trim(value)

bd. 将当前字符串转换为字节，如果转换失败，则返回0。

表达式：NumberUtils.toByte(value)

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

表达式：NumberUtils.toByte(value,*1*)

bf. 将当前字符串转换为Double数值，如果转换失败，则返回0.0d。

表达式：NumberUtils.toDouble(value)

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

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

bh. 将当前字符串转换为Float数值，如果转换失败，则返回0.0f。

表达式：NumberUtils.toFloat(value)

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

表达式：NumberUtils.toFloat(value,*1.1f*)

bj. 将当前字符串转换为Int数值，如果转换失败，则返回0。

表达式：NumberUtils.toInt(value)

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

表达式：NumberUtils.toInt(value,*1*)

bl. 将字符串转换为Long数值，如果转换失败，则返回0。

表达式：NumberUtils.toLong(value)

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

表达式：NumberUtils.toLong(value,*1L*)

bn. 将字符串转换为Short数值，如果转换失败，则返回0。

表达式：NumberUtils.toShort(value)

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

表达式：NumberUtils.toShort(value,*1*)

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

表达式：CommonUtils.ipToLong(value)

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

表达式：HttpsUtils.downloadMap("*url*")

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

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

bs. 取出缓存的IP与地址映射对象。

表达式：CommonUtils.getCache("*ipList*")

bt. 判断是否有IP与地址映射缓存。

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

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

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

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

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

7.7 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 7-14 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。

7.8 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。

7.9 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 7-15 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 7-2 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 7-3 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 7-4 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 7-5 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 7-6 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

7.10 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。

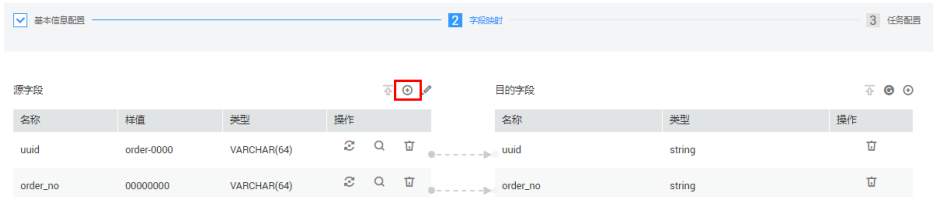
创建表/文件迁移作业

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 7-16 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 7-17 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7-7**。

图 7-18 添加字段

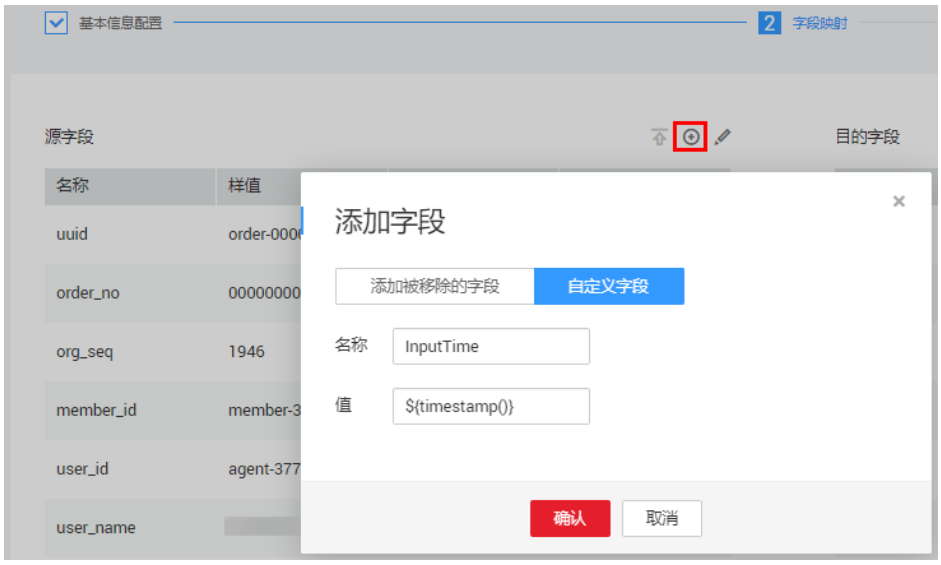


表 7-7 时间变量宏定义具体展示

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

7.11 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 7-8 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7-8**。

3. **编码类型**

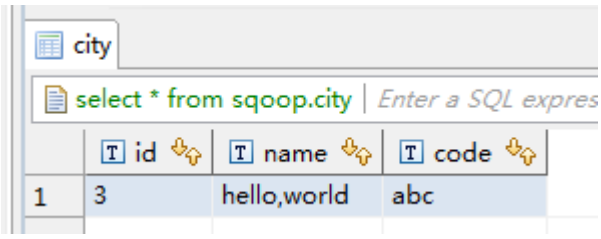
   文件的编码类型，默认是UTF-8，中文的编码有时会采用GBK。

   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7-19**所示，数据库的name字段的值中包含了字段分隔符逗号：

     图 7-19 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文件为源端，并且其中数据是被包围符括起来的时候，如果想把数据正确的导入到数据库，就需要在源端开启“使用包围符”，这样包围符内的值的，会写入一个字段内。

5. **使用正则表达式分隔字段**

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

6. **首行为标题行**

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

7. **写入文件大小**

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

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

JSON 格式

这里主要介绍JSON文件格式的以下内容：

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

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

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

表 7-9 示例

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

b. 示例二

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

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


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

表 7-10 示例

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

c. 示例三

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

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

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

表 7-11 示例

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

d. 示例四

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

表 7-12 示例

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

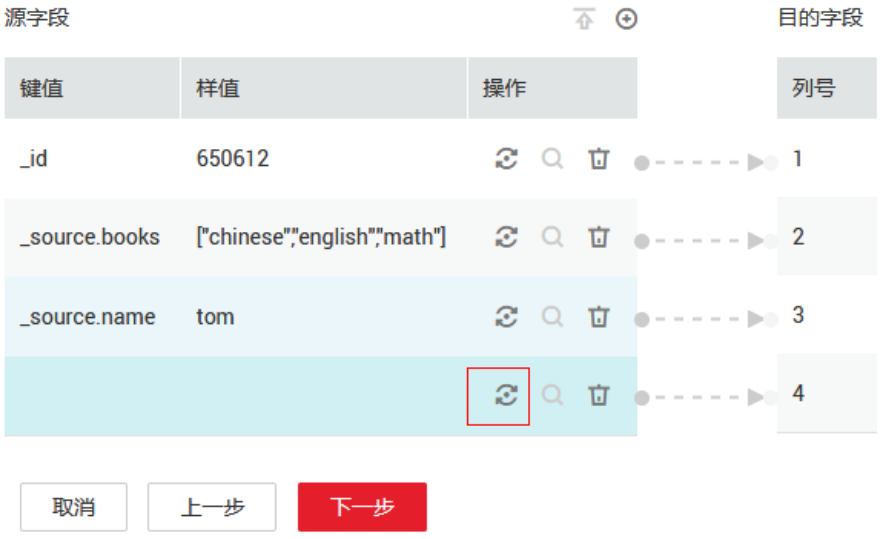
则需要在作业第一步指定文件格式为“JSON格式”，指定JSON类型为“JSON对象”，并且指定记录节点为“hits.hits”，然后在作业第二步添加转换器，操作步骤如下：

i. 单击添加字段，新增一个字段。

图 7-20 添加字段

ii. 在添加的新字段后面，单击添加字段转换器。

图 7-21 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 7-22 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：

   - 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7-8**。

   - 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

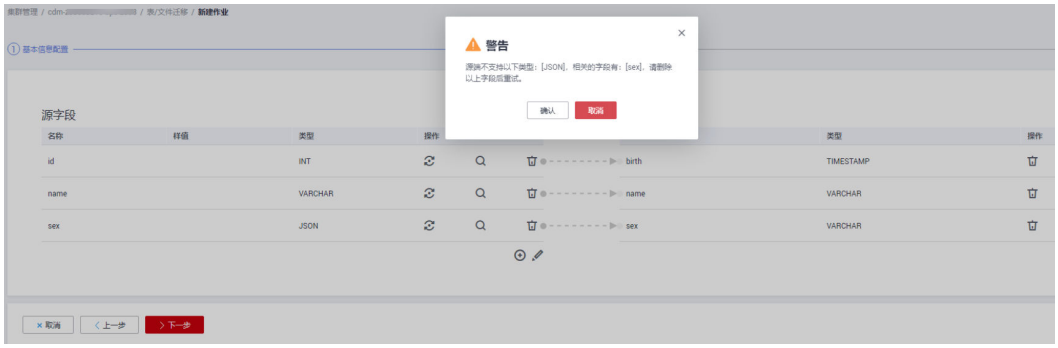
   - 场景：使用CDM先将MySQL中的某张表（表的某个字段值中包含了换行符\n）导出到CSV格式的文件中，然后再使用CDM将导出的CSV文件导入到MRS HBase，发现导出的CSV文件中出现了数据被截断的情况。
   - 解决方法：指定换行符。

     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。

7.12 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

8 使用教程

8.1 创建 MRS Hive 连接器

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

前提条件

- 已创建CDM集群。
- 已获取MRS集群的Manager IP、管理员账号和密码，且该账号拥有数据导入、导出的操作权限。
- MRS集群和CDM集群之间网络互通，网络互通需满足如下条件：
  - CDM集群与云上服务处于不同区域的情况下，需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP，数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
  - CDM集群与云上服务同区域情况下，同虚拟私有云、同子网、同安全组的不同实例默认网络互通；如果同虚拟私有云但是子网或安全组不同，还需配置路由规则及安全组规则，配置路由规则请参见**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**章节。
  - 此外，您还必须确保该云服务的实例与CDM集群所属的企业项目必须相同，如果不同，需要修改工作空间的企业项目。

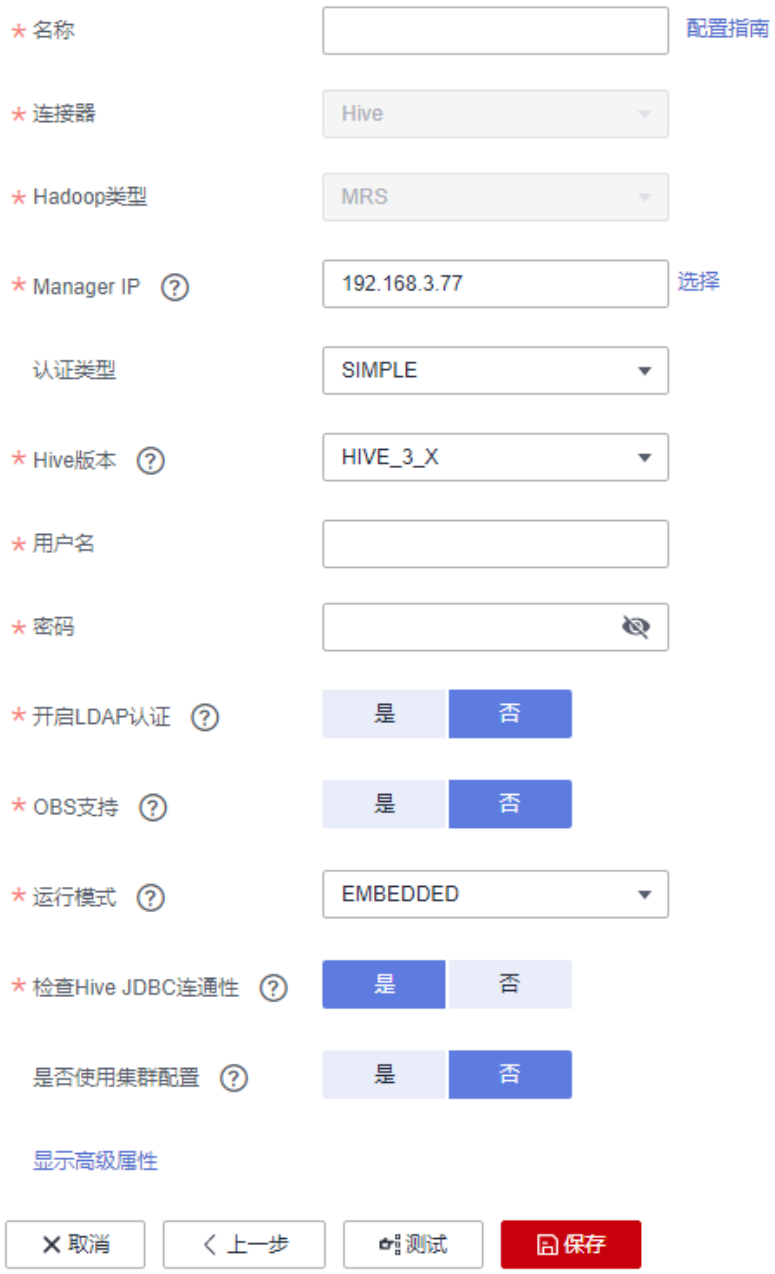
新建 MRS hive 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图8-1**所示。

图 8-1 选择连接器类型

步骤2 连接器类型选择“MRS Hive”后单击“下一步”，配置MRS Hive连接的参数，如图8-2所示。

图 8-2 创建 MRS Hive 连接



步骤3 单击“显示高级属性”可查看更多可选参数，这里保持默认，必填参数如下表所示。

表 8-1 MRS Hive 连接参数

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

参数名	说明	取值样例
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。根据服务端Hive版本设置。	HIVE_3_X
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。 **说明** ● 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对MRS组件的库、表、列进行操作，还需要参考MRS文档添加对应组件的库、表、列操作权限。 ● 如果CDM集群为2.9.0之前的版本，或MRS集群为3.1.0之前的版本，则所创建的用户需要具备Manager_administrator或System_administrator权限，才能在CDM创建连接。 ● 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。	cdm
密码	访问MRS Manager的用户密码。	-
开启LDAP认证	通过代理连接的时候，此项可配置。 当MRS Hive对接外部LDAP开启了LDAP认证时，连接Hive时需要使用LDAP账号与密码进行认证，此时必须开启此参数，否则会连接失败。	否
LDAP用户名	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的用户名。	-
LDAP密码	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否

参数名	说明	取值样例
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图8-3所示。 **图 8-3** 单击新增访问密钥  3. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id和Secret Access Key）。 **说明** – 每个用户仅允许新增两个访问密钥。 – 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台界面获取。请在生成后妥善保管。	- -
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

说明

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

步骤4 单击“保存”回到连接管理界面，完成MRS Hive连接器的配置。

----结束

8.2 创建 MySQL 连接器

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

前提条件

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

新建 MySQL 连接器

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

步骤2 在“驱动管理”页面，单击MySQL驱动“建议版本”列中的资料链接，按照相应指导获取驱动文件。

步骤3 在“驱动管理”页面中，选择以下方式上传MySQL驱动。

方式一：单击对应驱动名称右侧操作列的“上传”，选择本地已下载的驱动。

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

步骤4 在“集群管理”界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图8-4**所示。

图 8-4 选择连接器类型

步骤5 连接器类型选择“MySQL”后单击“下一步”，配置MySQL连接的参数，参数如表8-2所示。

表 8-2 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。	192.168.1.110
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	是
local_infile字符集	mysql通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	适配mysql的驱动。	-
Agent	单击“选择”，选择已创建的Agent。	-
单次请求行数	指定每次请求获取的行数。	1000

参数名	说明	取值样例
单次提交行数	支持通过Agent从源端提取数据	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'
单次写入行数	指定单次批量写入的行数，当写入行数累计到单次批量提交行数时提交一次，该值应当小于单次提交行数。	100

步骤6 单击“保存”回到连接管理界面，完成MySQL连接器的配置。

说明

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

----结束

8.3 MySQL 数据迁移到 MRS Hive 分区表

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

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

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

下面使用CDM将MySQL数据导入到MRS Hive分区表为例进行介绍。

操作场景

假设MySQL上有一张表trip_data，保存了自行车骑行记录，里面有起始时间、结束时间，起始站点、结束站点、骑手ID等信息，trip_data表字段定义如图8-5所示。

图 8-5 MySQL 表字段

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

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

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

前提条件

- 已经购买MRS。
- 已获取连接MySQL数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有MySQL数据库的读写权限。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了MySQL数据库驱动。

在 MRS Hive 上创建 Hive 分区表

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

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

说明

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

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

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

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

图 8-6 集群列表

说明

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

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图8-7**所示。

图 8-7 选择连接器类型

步骤2 选择“云数据库 MySQL”后单击“下一步”，配置云数据库 MySQL连接的参数。

图 8-8 创建 MySQL 连接

单击“显示高级属性”可查看更多可选参数，具体请参见**配置云数据库MySQL/MySQL数据库连接**。这里保持默认，必填参数如**表8-3**所示。

表 8-3 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。	-

参数名	说明	取值样例
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	否
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

步骤3 单击“保存”回到连接管理界面。

说明

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

----结束

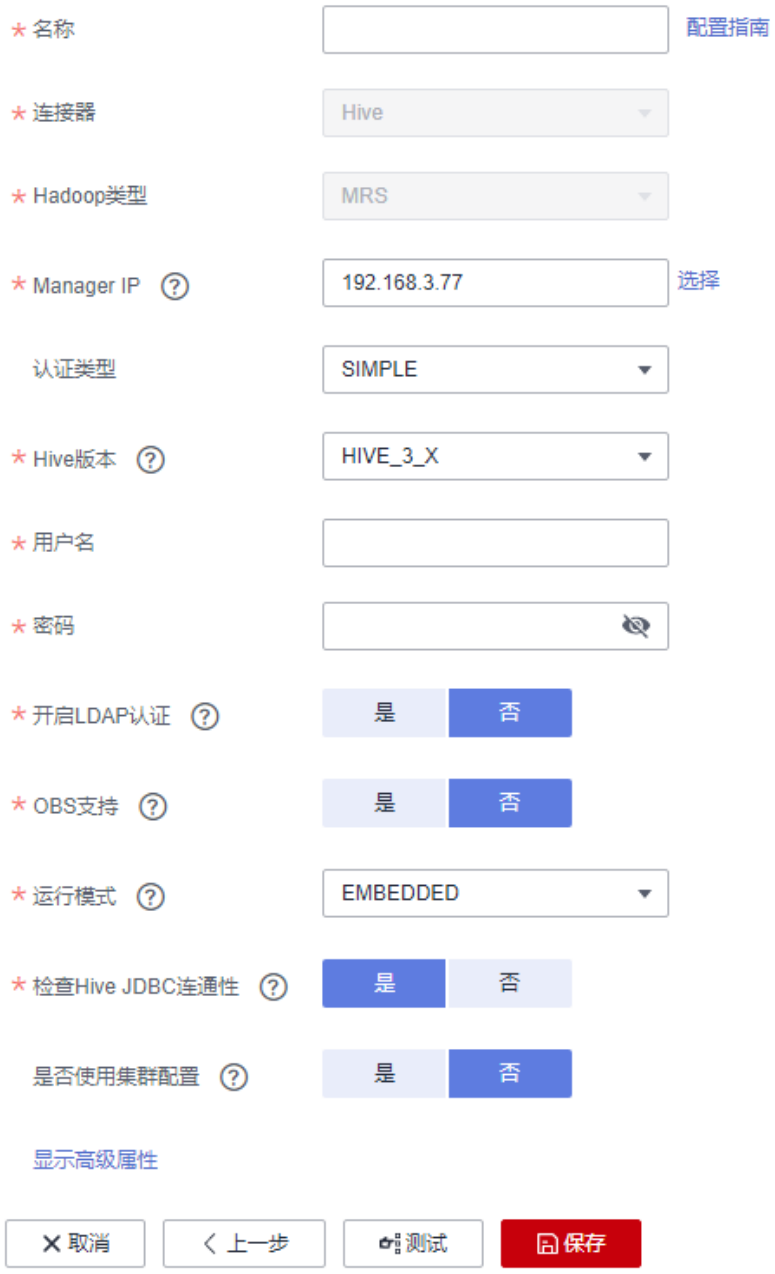
创建 Hive 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图8-9**所示。

图 8-9 选择连接器类型

步骤2 连接器类型选择“MRS Hive”后单击“下一步”配置Hive连接参数，如图8-10所示。

图 8-10 创建 MRS Hive 连接



各参数说明如表8-4所示，需要您根据实际情况配置。

表 8-4 MRS Hive 连接参数

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

参数名	说明	取值样例
认证类型	访问MRS的认证类型： ● SIMPLE：非安全模式选择Simple鉴权。 ● KERBEROS：安全模式选择Kerberos鉴权。	SIMPLE
Hive版本	Hive的版本。根据服务端Hive版本设置。	HIVE_3_X
用户名	选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。 如果要创建MRS安全集群的数据连接，不能使用admin用户。因为admin用户是默认的管理页面用户，这个用户无法作为安全集群的认证用户来使用。您可以创建一个新的MRS用户，然后在创建MRS数据连接时，“用户名”和“密码”填写为新建的MRS用户及其密码。 **说明** ● 如果CDM集群为2.9.0版本及之后版本，且MRS集群为3.1.0及之后版本，则所创建的用户至少需具备Manager_viewer的角色权限才能在CDM创建连接；如果需要对MRS组件的库、表、列进行操作，还需要参考MRS文档添加对应组件的库、表、列操作权限。 ● 如果CDM集群为2.9.0之前的版本，或MRS集群为3.1.0之前的版本，则所创建的用户需要具备Manager_administrator或System_administrator权限，才能在CDM创建连接。 ● 仅具备Manager_tenant或Manager_auditor权限，无法创建连接。	cdm
密码	访问MRS Manager的用户密码。	-
开启LDAP认证	通过代理连接的时候，此项可配置。 当MRS Hive对接外部LDAP开启了LDAP认证时，连接Hive时需要使用LDAP账号与密码进行认证，此时必须开启此参数，否则会连接失败。	否
LDAP用户名	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的用户名。	-
LDAP密码	当“开启LDAP认证”参数选择为“是”时，此参数是必选项。 填写为MRS Hive开启LDAP认证时配置的密码。	-
OBS支持	需服务端支持OBS存储。在创建Hive表时，您可以指定将表存储在OBS中。	否

参数名	说明	取值样例
访问标识(AK)	当“OBS支持”参数选择为“是”时，此参数是必选项。请注意，此处AK/SK对应的账号应具备OBS Buckets Viewer系统权限，否则会无法访问OBS并报“403 AccessDenied”错误。 您需要先创建当前账号的访问密钥，并获得对应的AK和SK。 1. 登录控制台，在用户名下拉列表中选择“我的凭证”。 2. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图8-11所示。 **图 8-11** 单击新增访问密钥  3. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id和Secret Access Key）。 **说明** – 每个用户仅允许新增两个访问密钥。 – 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台界面获取。请在生成后妥善保管。	- -
密钥(SK)		
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

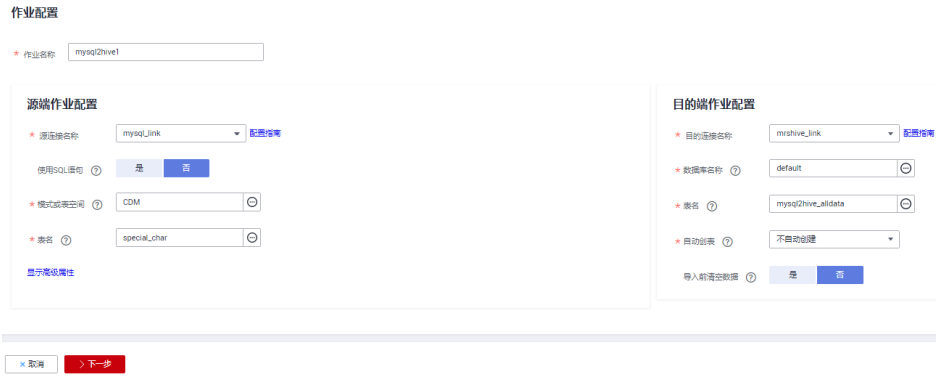
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建数据迁移任务，如**图8-12**所示。

图 8-12 创建 MySQL 到 Hive 的迁移任务



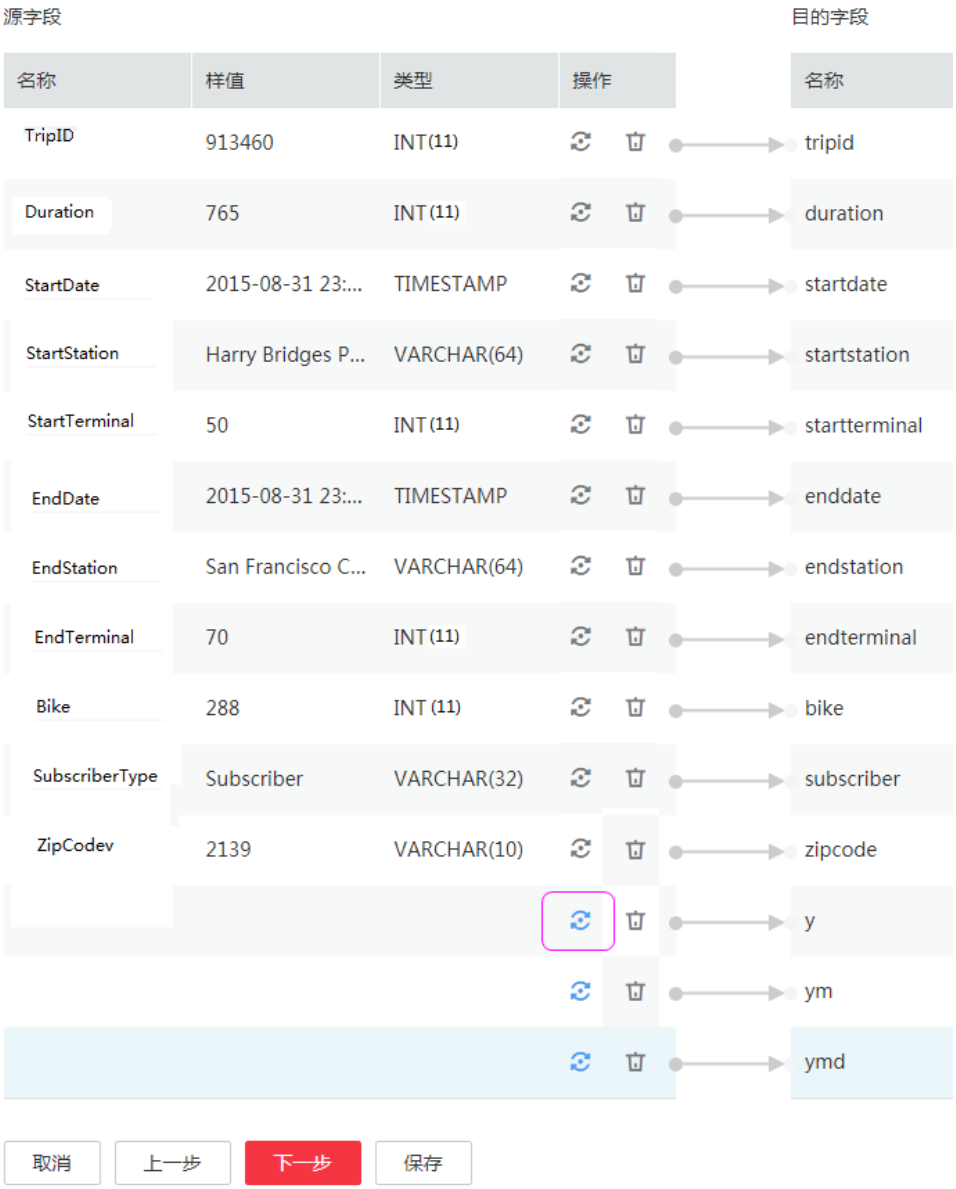
说明

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

步骤2 作业参数配置完成后，单击“下一步”，进入字段映射界面，如**图8-13**所示。

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

图 8-13 Hive 字段映射



步骤3 单击 进入转换器列表界面，再选择“新建转换器 > 表达式转换”，如图8-14所示。

y、ym、ymd字段的表达式分别配置如下：

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

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

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

图 8-14 配置表达式

说明

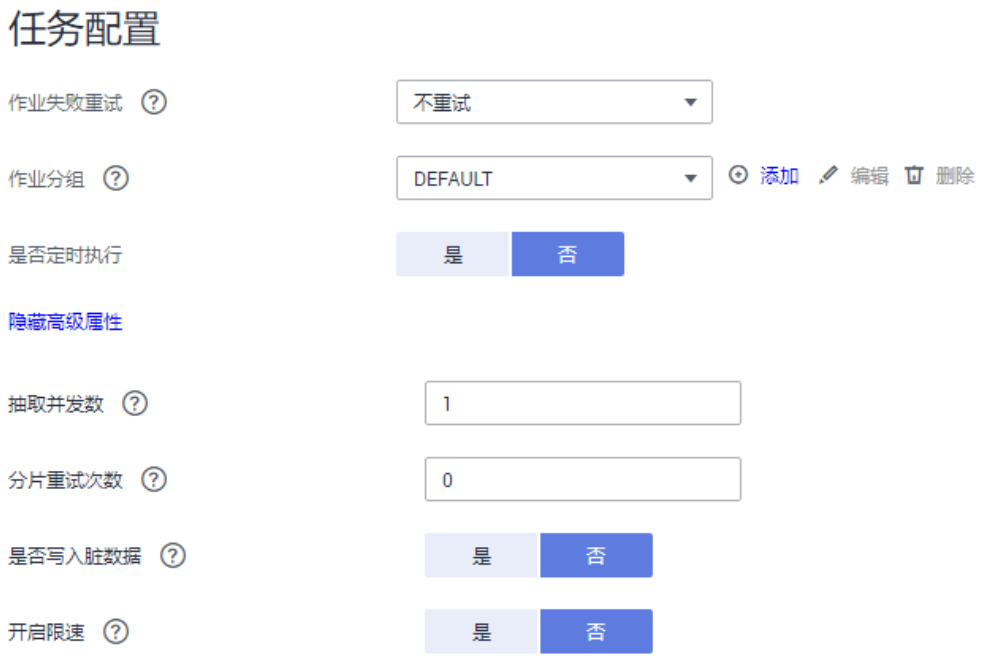
CDM的表达式已经预置常用字符串、日期、数值等类型的字段内容转换，详细请参见**字段转换**。

步骤4 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

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

图 8-15 任务配置



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

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

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

8.4 MySQL 数据迁移到 OBS

操作场景

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

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

前提条件

- 已获取OBS的访问域名、端口，以及AK、SK。
- 已获取连接MySQL数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有MySQL数据库的读写权限。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了MySQL数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

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

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

说明

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

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图8-16**所示。

图 8-16 选择连接器类型

步骤2 选择“云数据库 MySQL”后单击“下一步”，配置云数据库 MySQL连接的参数。

图 8-17 创建 MySQL 连接

单击“显示高级属性”可查看更多可选参数，具体请参见**配置云数据库MySQL/MySQL数据库连接**。这里保持默认，必填参数如**表8-5**所示。

表 8-5 MySQL 连接参数

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

参数名	说明	取值样例
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	否
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

步骤3 单击“保存”回到连接管理界面。

说明

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

----结束

创建 OBS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图8-18**所示。

图 8-18 选择连接器类型

步骤2 连接器类型选择“对象存储服务（OBS）”后，单击“下一步”配置OBS连接参数，如图8-20所示。

- 名称：用户自定义连接名称，例如“obslink”。
- OBS服务器、端口：配置为OBS实际的地址信息。
- 访问标识（AK）、密钥（SK）：登录OBS的AK、SK。
  您可以通过如下方式获取访问密钥。
  a. 登录控制台，在用户名下拉列表中选择“我的凭证”。
  b. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图8-19所示。

图 8-19 单击新增访问密钥

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

说明

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

图 8-20 创建 OBS 连接

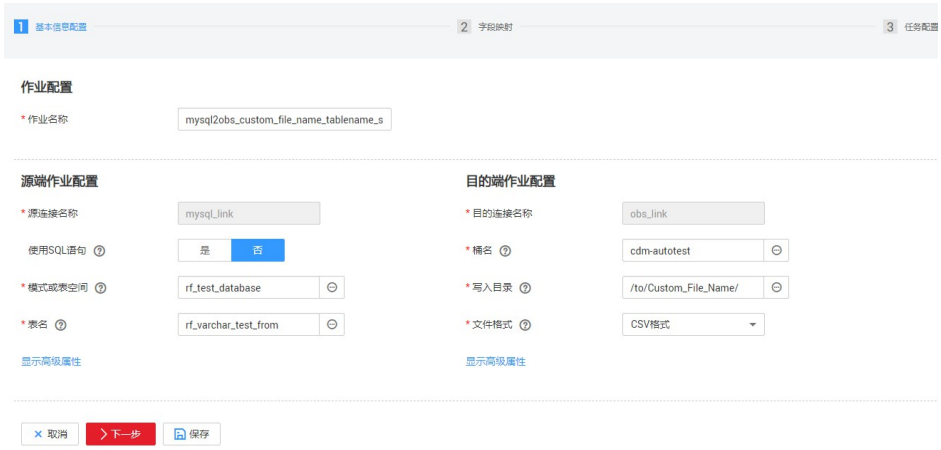
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

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

图 8-21 创建 MySQL 到 OBS 的迁移任务



- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建MySQL连接**中的“mysqllink”。
  - 使用SQL语句：否。
  - 模式或表空间：待抽取数据的模式或表空间名称。
  - 表名：要抽取的表名。
  - 其他可选参数一般情况下保持默认即可。
- 目的端作业配置
  - 目的连接名称：选择**创建OBS连接**中的“obslink”。
  - 桶名：待迁移数据的桶。
  - 写入目录：写入数据到OBS服务器的目录。
  - 文件格式：迁移数据表到文件时，文件格式选择“CSV格式”。
  - 高级属性里的可选参数一般情况下保持默认即可。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如**图8-22**所示。

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

图 8-22 表到文件的字段映射

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

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

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

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

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

8.5 MySQL 数据迁移到 DWS

操作场景

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

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

前提条件

- 已获取DWS数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有DWS数据库的读、写和删除权限。
- 已获取连接MySQL数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有MySQL数据库的读写权限。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了MySQL数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

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

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

说明

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

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图8-23**所示。

图 8-23 选择连接器类型

步骤2 选择“云数据库 MySQL”后单击“下一步”，配置云数据库 MySQL连接的参数。

图 8-24 创建 MySQL 连接

单击“显示高级属性”可查看更多可选参数，具体请参见**配置云数据库MySQL/MySQL数据库连接**。这里保持默认，必填参数如**表8-6**所示。

表 8-6 MySQL 连接参数

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

参数名	说明	取值样例
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	否
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

步骤3 单击“保存”回到连接管理界面。

说明

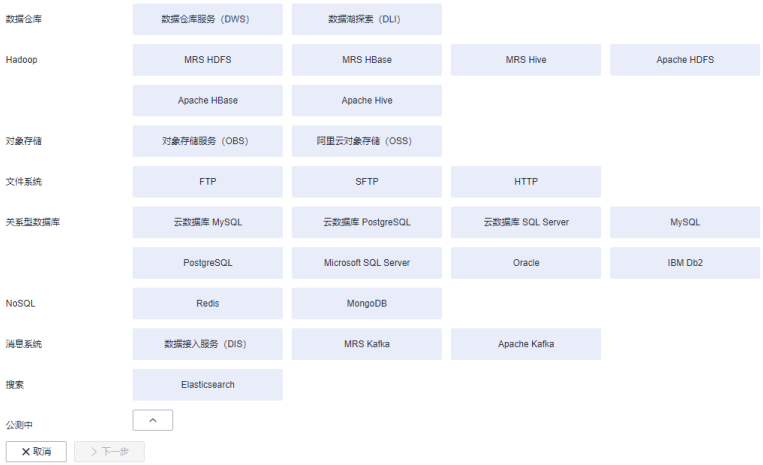
如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

创建 DWS 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图8-25**所示。

图 8-25 选择连接器类型



步骤2 连接器类型选择“数据仓库服务（DWS）”后单击“下一步”配置DWS连接参数，必填参数如表8-7所示，可选参数保持默认即可。

表 8-7 DWS 连接参数

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

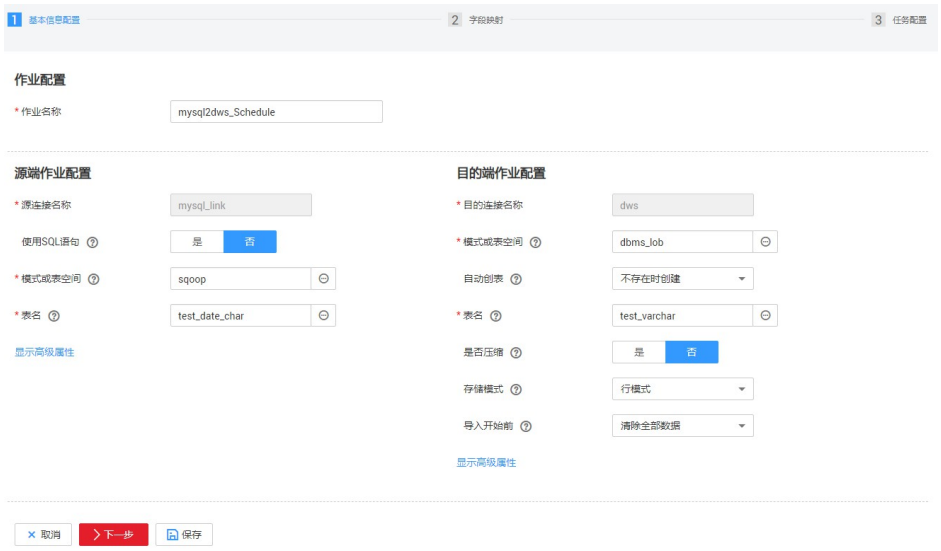
步骤3 单击“保存”完成创建连接。

----结束

创建迁移作业

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

图 8-26 创建 MySQL 到 DWS 的迁移任务



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

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如**图8-27**所示。

- 如果字段映射顺序不匹配，可通过拖拽字段调整。

- CDM的表达式已经预置常用字符串、日期、数值等类型的字段内容转换，详细请参见**字段转换**。

图 8-27 表到表的字段映射

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

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

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

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

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

8.6 MySQL 整库迁移到 RDS 服务

操作场景

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

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

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

4. 创建整库迁移作业

前提条件

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

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

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

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

图 8-28 集群列表

说明

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

----结束

创建 MySQL 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图8-29**所示。

图 8-29 选择连接器类型

步骤2 选择“云数据库 MySQL”后单击“下一步”，配置云数据库 MySQL连接的参数。

图 8-30 创建 MySQL 连接

单击“显示高级属性”可查看更多可选参数，具体请参见**配置云数据库MySQL/MySQL数据库连接**。这里保持默认，必填参数如**表8-8**所示。

表 8-8 MySQL 连接参数

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

参数名	说明	取值样例
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	否
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

步骤3 单击“保存”回到连接管理界面。

说明

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

----结束

创建 RDS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图8-31**所示。

图 8-31 选择连接器类型

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

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

说明

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

步骤3 单击“保存”回到连接管理界面。

----结束

创建整库迁移作业

步骤1 两个连接创建完成后，选择“整库迁移 > 新建作业”，开始创建迁移任务，如图8-32所示。

图 8-32 创建整库迁移作业

- 作业名称：用户自定义整库迁移的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建MySQL连接**中的“mysqllink”。
  - 模式或表空间：选择从本地MySQL的哪个数据库导出数据。
- 目的端作业配置
  - 目的连接名称：选择**创建RDS连接**中的“rds_link”。
  - 模式或表空间：选择将数据导入到RDS的哪个数据库。
  - 自动创表：选择“不存在时创建”，当RDS数据库中没有本地MySQL数据库里的表时，CDM会自动在RDS数据库中创建那些表。
  - 导入开始前：选择“是”，当RDS数据库中存在与本地MySQL数据库重名的表时，CDM会清除RDS中重名表里的数据。
  - 约束冲突处理：选择“insert into”，当迁移数据出现唯一约束冲突时的处理方式。
  - 高级属性里的可选参数保持默认即可。

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

步骤3 单击“保存并运行”，CDM会立即开始执行整库迁移任务。

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

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

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

----结束

8.7 Oracle 数据迁移到云搜索服务

操作场景

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

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

前提条件

- 已经开通了云搜索服务，且获取云搜索服务集群的IP地址和端口。
- 已获取Oracle数据库的IP、数据库名、用户名和密码。
- 如果Oracle数据库是在本地数据中心或第三方云上，需要确保Oracle可通过公网IP访问，或者已经建立好了企业内部数据中心到华为云的VPN通道或专线。
- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了Oracle数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群的VPC必须和云搜索服务集群所在VPC一致，且推荐子网、安全组也与云搜索服务一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

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

说明

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

----结束

创建云搜索服务连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图8-33**所示。

图 8-33 选择连接器类型

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

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

图 8-34 创建云搜索服务连接

步骤3 单击“保存”回到连接管理界面。

----结束

创建 Oracle 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如图8-35所示。

图 8-35 选择连接器类型

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

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

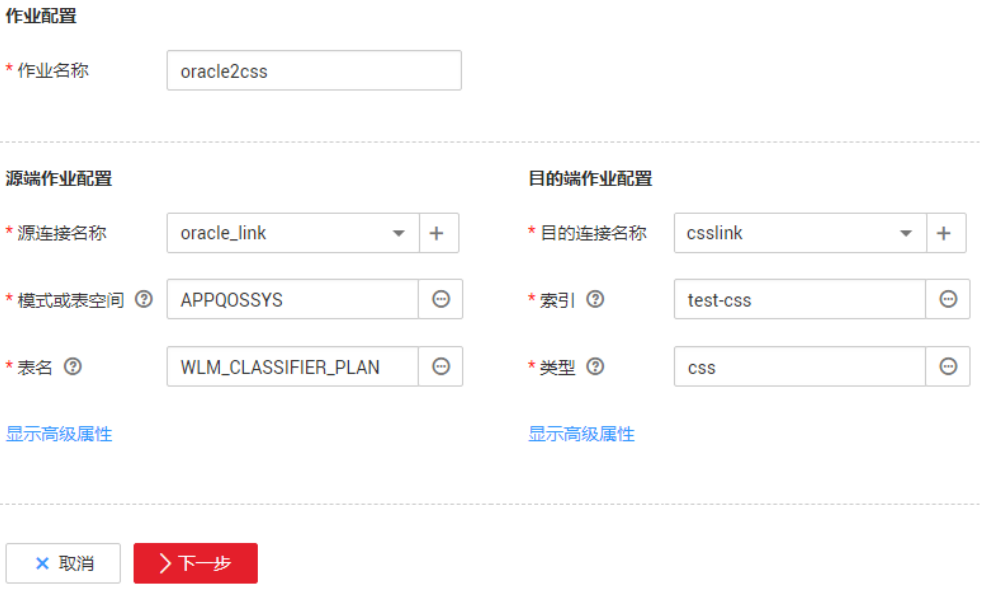
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

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

图 8-36 创建 Oracle 到云搜索服务的迁移任务



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

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如图8-37所示。

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

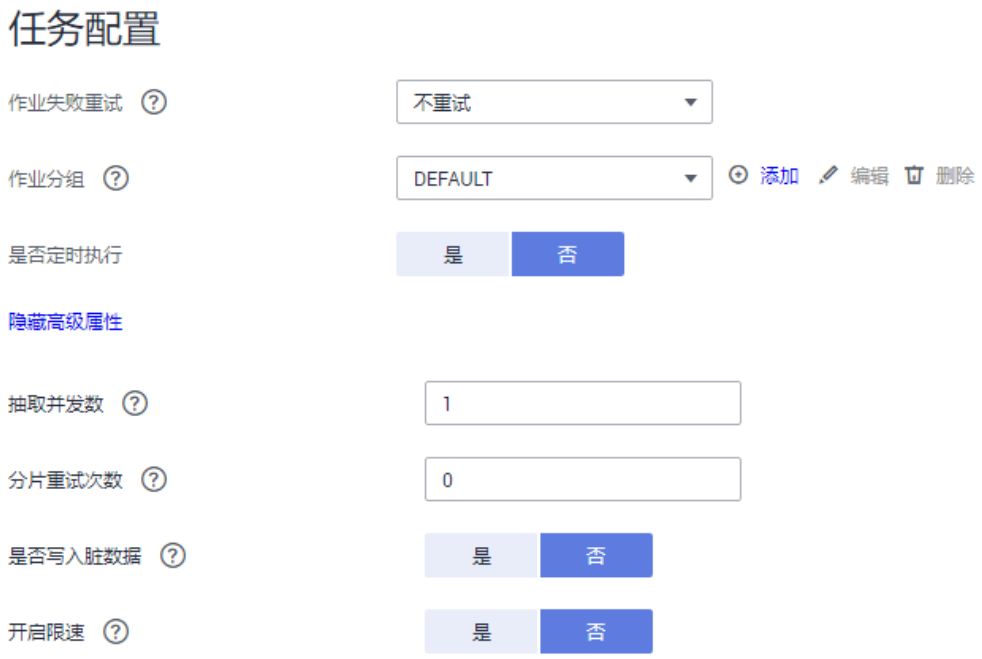
图 8-37 云搜索服务的字段映射

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

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

图 8-38 任务配置



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

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

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

8.8 Oracle 数据迁移到 DWS

操作场景

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

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

前提条件

- 已购买DWS集群，并且已获取DWS数据库的IP地址、端口、数据库名称、用户名、密码，且该用户拥有DWS数据库的读、写和删除权限。
- 已获取Oracle数据库的IP、数据库名、用户名和密码。
- 如果Oracle数据库是在本地数据中心或第三方云上，需要确保Oracle可通过公网IP访问，或者已经建立好了企业内部数据中心到华为云的VPN通道或专线。

- 已在CDM集群的“作业管理 > 连接管理 > 驱动管理”页面，上传了Oracle数据库驱动。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群所在VPC、子网、安全组，选择与DWS集群所在的网络一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

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

说明

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

----结束

创建 Oracle 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图8-39**所示。

图 8-39 选择连接器类型

步骤2 连接器类型选择“Oracle”后单击“下一步”，配置Oracle连接参数，参数说明如**表8-9**所示。

图 8-40 创建 Oracle 连接

表 8-9 Oracle 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	oracle_link
数据库服务器	数据库服务器域名或IP地址。	192.168.0.1
端口	Oracle数据库的端口。	3306
数据库连接类型	Oracle数据库连接类型。	Service Name
数据库名称	要连接的数据库。	db_user
用户名	拥有Oracle数据库的读取权限的用户。	admin
密码	Oracle数据库的登录密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	是
Agent	单击“选择”，选择已创建的Agent。	-
ORACLE版本	默认使用最新版本驱动，若不兼容请尝试其他版本。	高于12.1
驱动版本	需要适配的驱动。	-
一次请求行数	指定每次请求获取的行数。	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'

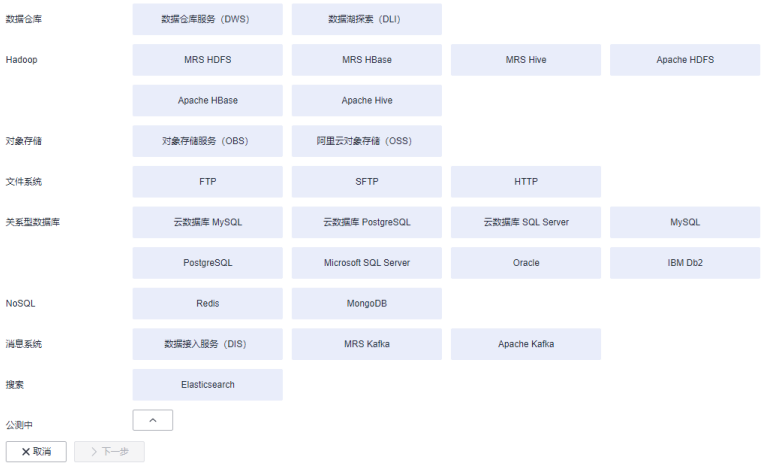
步骤3 单击“保存”回到连接管理界面。

----结束

创建 DWS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如图8-41所示。

图 8-41 选择连接器类型



步骤2 连接器类型选择“数据仓库服务（DWS）”后单击“下一步”配置DWS连接参数，必填参数如**表8-10**所示，可选参数保持默认即可。

表 8-10 DWS 连接参数

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

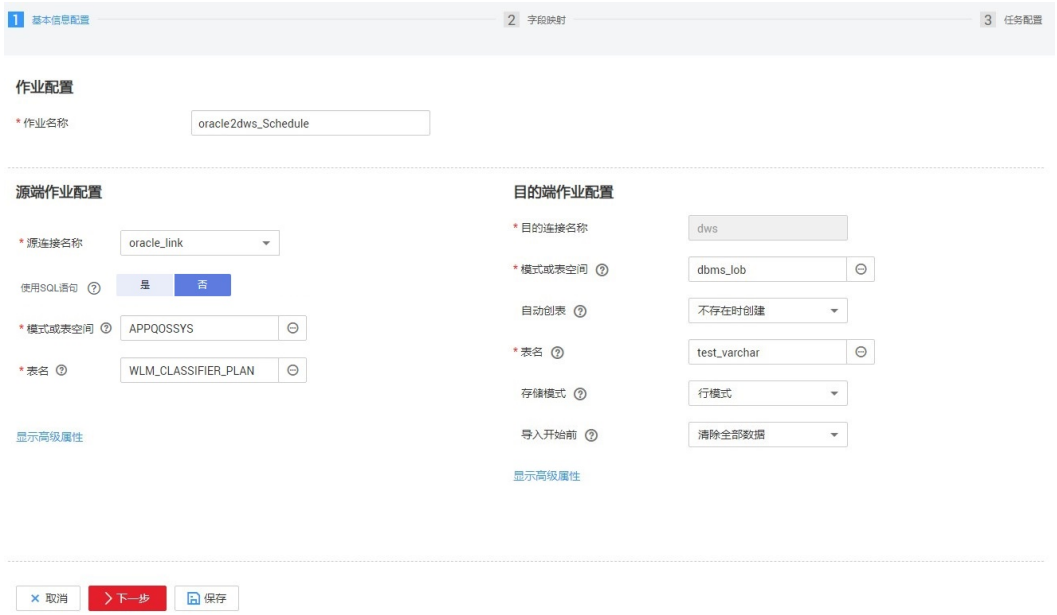
步骤3 单击“保存”完成创建连接。

----结束

创建迁移作业

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

图 8-42 创建 Oracle 到 DWS 的迁移任务



- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建Oracle连接**中的“oracle_link”。
  - 模式或表空间：待迁移数据的数据库名称。
  - 表名：待迁移数据的表名。
  - 高级属性里的可选参数一般情况下保持默认即可。
- 目的端作业配置
  - 目的连接名称：选择**创建DWS连接**中的连接“dwslink”。
  - 模式或表空间：选择待写入数据的DWS数据库。
  - 自动创表：只有当源端和目的端都为关系数据库时，才有该参数。
  - 表名：待写入数据的表名，可以手动输入一个不存在表名，CDM会在DWS中自动创建该表。
  - 存储模式：可以根据具体应用场景，建表的时候选择行存储还是列存储表。一般情况下，如果表的字段比较多（大宽表），查询中涉及到的列不多的情况下，适合列存储。如果表的字段个数比较少，查询大部分字段，那么选择行存储比较好。
  - 扩大字符字段长度：当目的端和源端数据编码格式不一样时，自动建表的字符字段长度可能不够用，配置此选项后CDM自动建表时会将字符字段扩大3倍。
  - 导入前清空数据：任务启动前，是否清除目的表中数据，用户可根据实际需要选择。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如**图8-43**所示。

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

图 8-43 表到表的字段映射

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

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

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

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

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

说明

如遇目的端写太久导致迁移超时，请减少Oracle连接器中“一次请求行数”参数值的设置。

8.9 OBS 数据迁移到云搜索服务

操作场景

CDM支持在云上各服务之间相互迁移数据，本章节介绍如何通过CDM将数据从OBS迁移到云搜索服务中，流程如下：

1. **创建CDM集群**
2. **创建云搜索服务连接**
3. **创建OBS连接**
4. **创建迁移作业**

前提条件

- 已获取OBS的访问域名、端口，以及AK、SK。
- 已经开通了云搜索服务，且获取云搜索服务集群的IP地址和端口。

创建 CDM 集群

如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群的VPC必须和云搜索服务集群所在VPC一致，且推荐子网、安全组也与云搜索服务一致。
- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

创建云搜索服务连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图8-44**所示。

图 8-44 选择连接器类型

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

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

图 8-45 创建云搜索服务连接

步骤3 单击“保存”回到连接管理界面。

----结束

创建 OBS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如图8-46所示。

图 8-46 选择连接器类型

步骤2 连接器类型选择“对象存储服务（OBS）”后，单击“下一步”配置OBS连接参数，如图8-48所示。

- 名称：用户自定义连接名称，例如“obslink”。
- OBS服务器、端口：配置为OBS实际的地址信息。
- 访问标识（AK）、密钥（SK）：登录OBS的AK、SK。
  您可以通过如下方式获取访问密钥。
  a. 登录控制台，在用户名下拉列表中选择“我的凭证”。
  b. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图8-47所示。

图 8-47 单击新增访问密钥

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

说明

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

图 8-48 创建 OBS 连接

步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

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

图 8-49 创建 OBS 到云搜索服务的迁移任务

- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建OBS连接**中的“obslink”。
  - 桶名：待迁移数据的桶。
  - 源目录或文件：待迁移数据的路径，也可以迁移桶下的所有目录、文件。
  - 文件格式：迁移文件到数据表时，文件格式选择“CSV格式”。
  - 高级属性里的可选参数一般情况下保持默认即可。
- 目的端作业配置
  - 目的连接名称：选择**创建云搜索服务连接**中的“csslink”。
  - 索引：待写入数据的Elasticsearch索引，也可以输入一个新的索引，CDM会自动在云上搜索服务中创建。
  - 类型：待写入数据的Elasticsearch类型，可输入新的类型，CDM支持在目的端自动创建类型。
  - 高级属性里的可选参数一般情况下保持默认即可。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段，如**图8-50**所示。

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

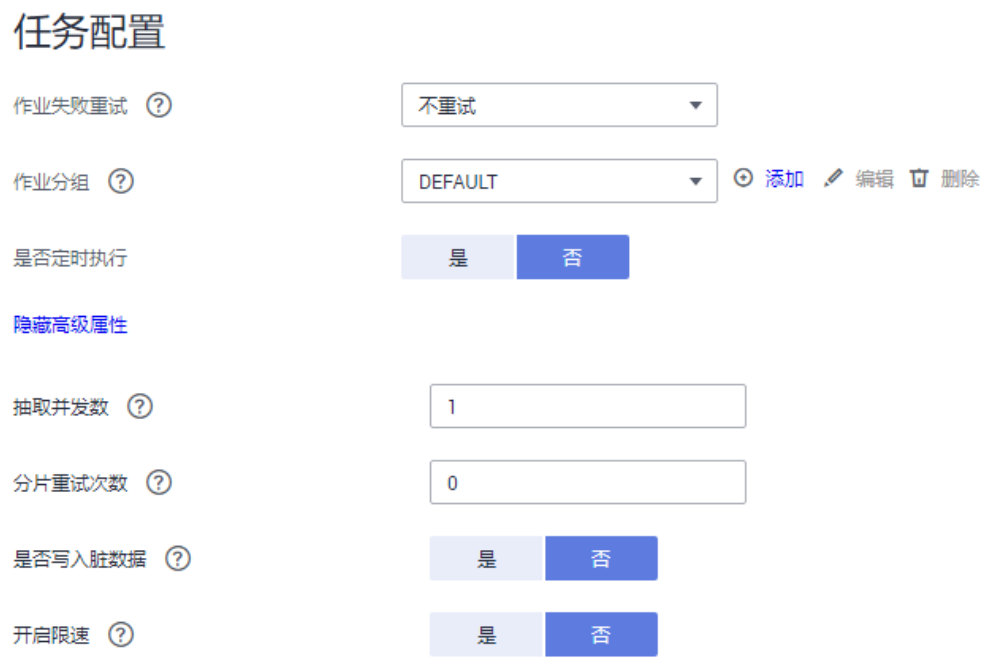
图 8-50 云搜索服务的字段映射

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

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

图 8-51 任务配置



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

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

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

8.10 OBS 数据迁移到 DLI 服务

操作场景

数据湖探索（Data Lake Insight，简称DLI）提供大数据查询服务，本章节介绍使用CDM将OBS的数据迁移到DLI，使用流程如下：

1. **创建CDM集群**
2. **创建DLI连接**
3. **创建OBS连接**
4. **创建迁移作业**

前提条件

- 已经开通了OBS和DLI，并且当前用户拥有OBS的读取权限。
- 已经在DLI服务中创建好资源队列、数据库和表。

创建 CDM 集群

如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

该场景下，如果CDM集群只是用于迁移OBS数据到DLI，不需要迁移其他数据源，则CDM集群所在的VPC、子网、安全组选择任一个即可，没有要求，CDM通过内网访问DLI和OBS。主要是选择CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。

创建 DLI 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图8-52**所示。

图 8-52 选择连接器类型

步骤2 连接器类型选择“数据湖探索（DLI）”后单击“下一步”，配置DLI连接参数，如图8-53所示。

- 名称：用户自定义连接名称，例如“dlilink”。
- 访问标识（AK）、密钥（SK）：访问DLI数据库的AK、SK。
- 项目ID：DLI所属区域的项目ID。

图 8-53 创建 DLI 连接

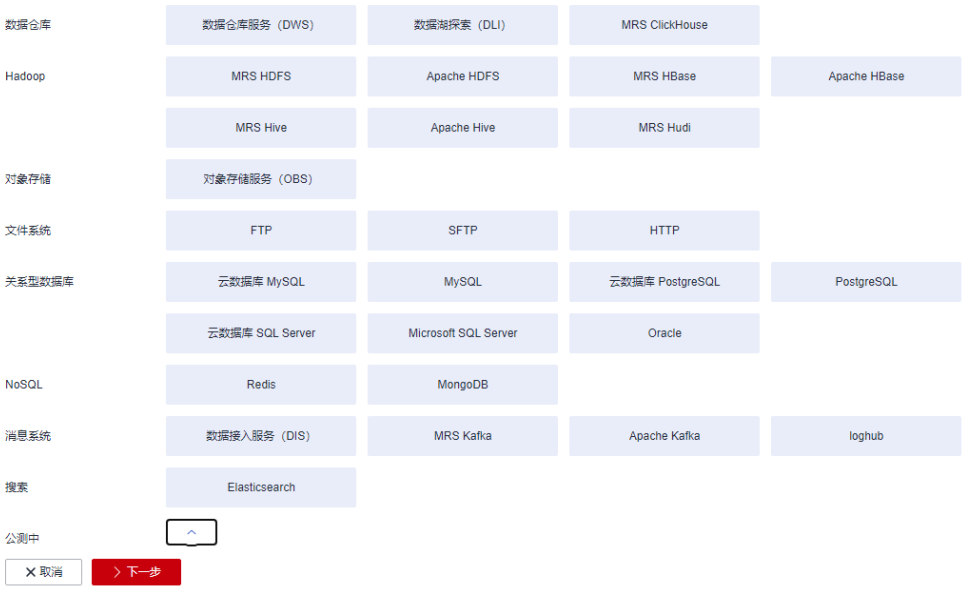
步骤3 单击“保存”回到连接管理界面。

----结束

创建 OBS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如图8-54所示。

图 8-54 选择连接器类型



步骤2 连接器类型选择“对象存储服务（OBS）”后，单击“下一步”配置OBS连接参数，如图8-56所示。

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

  您可以通过如下方式获取访问密钥。

  a. 登录控制台，在用户名下拉列表中选择“我的凭证”。

  b. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图8-55所示。

  图 8-55 单击新增访问密钥

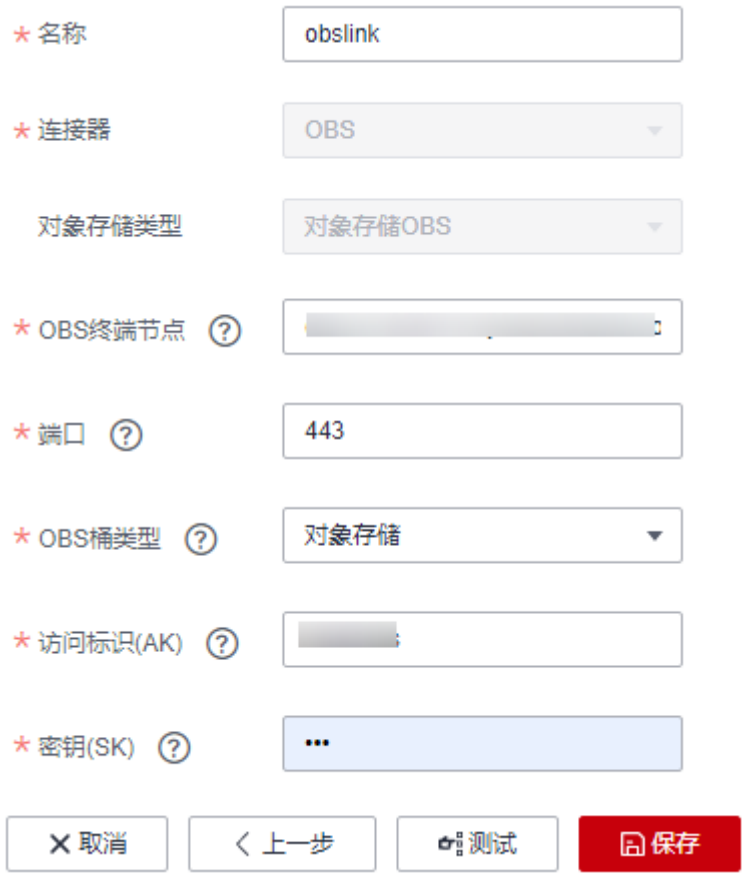
  

  c. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id和Secret Access Key）。

说明

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

图 8-56 创建 OBS 连接



步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从OBS迁移数据到DLI的任务，如图8-57所示。

图 8-57 创建 OBS 到 DLI 的迁移任务

- 作业名称：用户自定义作业名称。
- 源连接名称：选择**创建OBS连接**中的“obslink”。
  - 桶名：待迁移数据所属的桶。
  - 源目录或文件：待迁移数据的具体路径。
  - 文件格式：传输文件到数据表时，这里选择“CSV格式”或“JSON格式”。
  - 高级属性里的可选参数保持默认。
- 目的连接名称：选择**创建DLI连接**中的“dlilink”。
  - 资源队列：选择目的表所属的资源队列。
  - 数据库名称：写入数据的数据库名称。
  - 表名：写入数据的目的表。CDM暂不支持在DLI中自动创表，这里的表需要先在DLI中创建好，且该表的字段类型和格式，建议与待迁移数据的字段类型、格式保持一致。
  - 导入前清空数据：导入数据前，选择是否清空目的表中的数据，这里保持默认“否”。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段。

- 如果字段映射顺序不匹配，可通过拖拽字段调整。
- CDM支持迁移过程中转换字段内容，详细请参见**字段转换**。

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行可开启。这里保持默认值“否”。

- 抽取并发数：设置同时执行的抽取任务数，适当的抽取并发数可以提升迁移效率，配置原则请参见**性能调优**。这里保持默认值“1”。
- 是否写入脏数据：如果需要将作业执行过程中处理失败的数据、或者被清洗过滤掉的数据写入OBS中，以便后面查看，可通过该参数配置，写入脏数据前需要在CDM先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。

图 8-58 任务配置

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

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

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

8.11 MRS HDFS 数据迁移到 OBS

操作场景

CDM支持文件到文件类数据的迁移，本章节以MRS HDFS-->OBS为例，介绍如何通过CDM将文件类数据迁移到文件中。流程如下：

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

前提条件

- 已获取OBS的访问域名、端口，以及AK、SK。
- 已经购买了MRS。
- 拥有EIP配额。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

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

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

说明

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

----结束

创建 MRS HDFS 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图8-59**所示。

图 8-59 选择连接器类型

步骤2 连接器类型选择“MRS HDFS”后单击“下一步”，配置MRS HDFS链接参数。

- 名称：用户自定义连接名称，例如“mrs_hdfs_link”。
- Manage IP：MRS Manager的IP地址，可以单击输入框后的“选择”来选定已创建的MRS集群，CDM会自动填充下面的鉴权参数。
- 用户名：选择KERBEROS鉴权时，需要配置MRS Manager的用户名和密码。

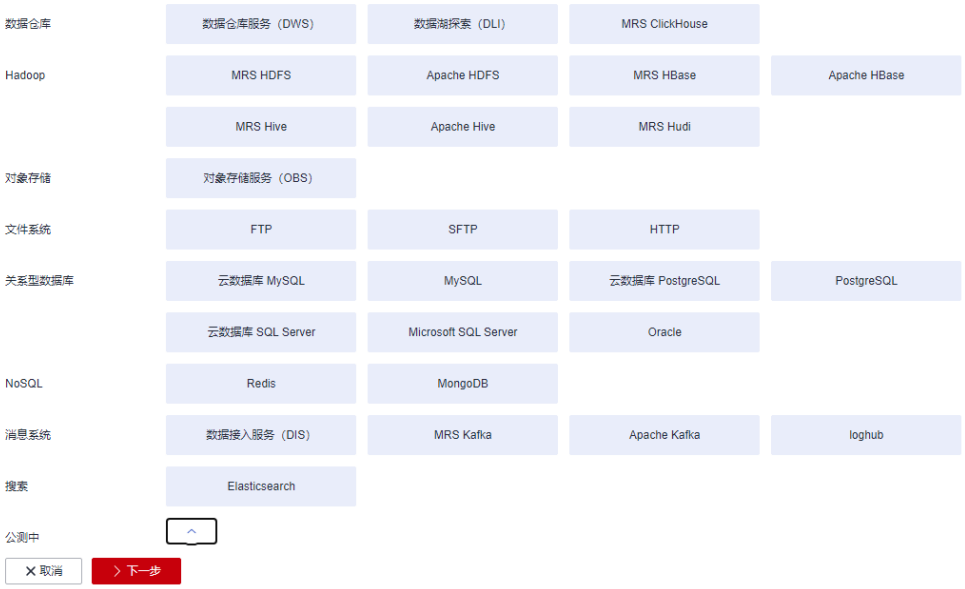
  从HDFS导出目录时，如果需要创建快照，这里配置的用户需要HDFS系统的管理员权限。
- 密码：访问MRS Manager的用户密码。
- 认证类型：访问MRS的认证类型。
- 运行模式：选择HDFS连接的运行模式。

----结束

创建 OBS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如图8-60所示。

图 8-60 选择连接器类型



步骤2 连接器类型选择“对象存储服务（OBS）”后，单击“下一步”配置OBS连接参数，如图8-62所示。

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

  您可以通过如下方式获取访问密钥。

  a. 登录控制台，在用户名下拉列表中选择“我的凭证”。
  b. 进入“我的凭证”页面，选择“访问密钥 > 新增访问密钥”，如图8-61所示。

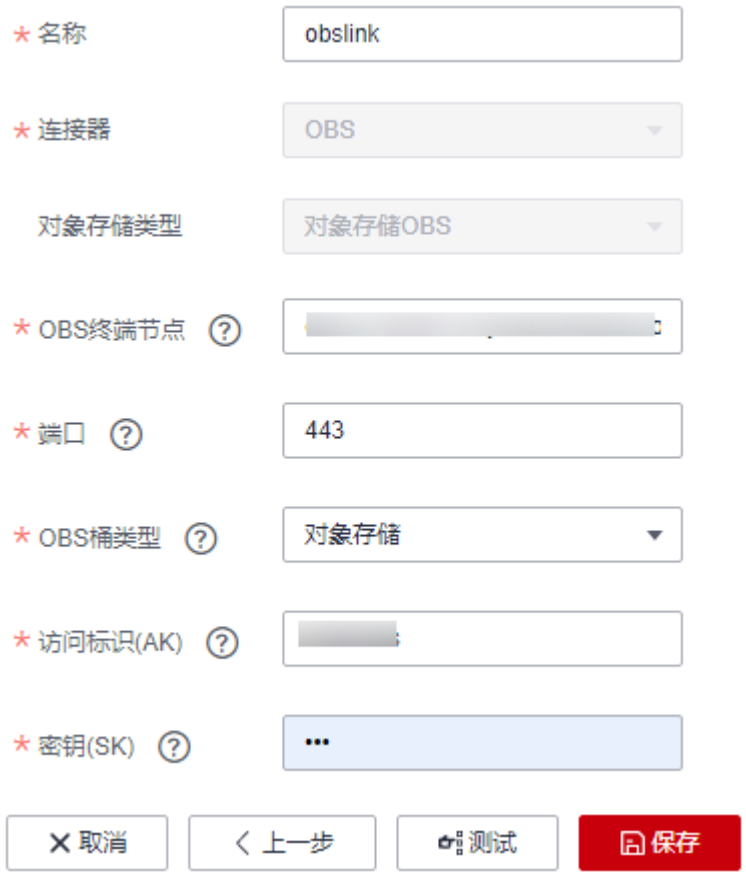
图 8-61 单击新增访问密钥

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

说明

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

图 8-62 创建 OBS 连接



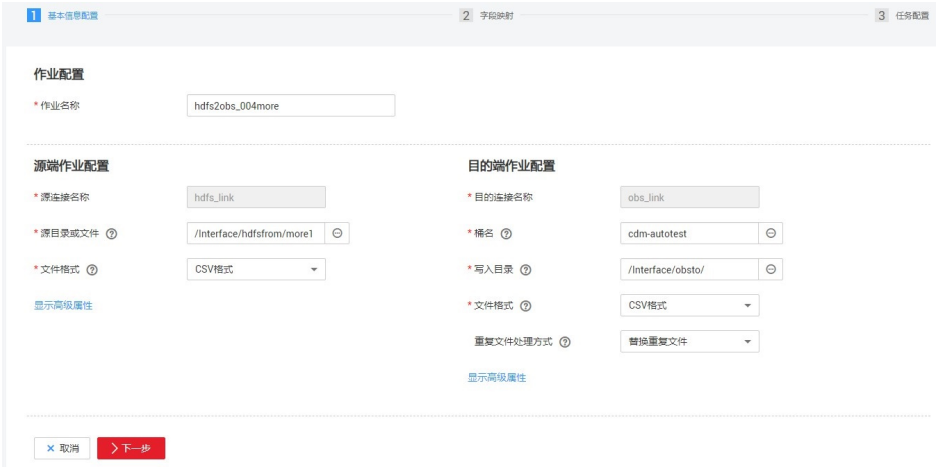
步骤3 单击“保存”回到连接管理界面。

----结束

创建迁移作业

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

图 8-63 创建 MRS HDFS 到 OBS 的迁移任务



- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建MRS HDFS连接**中的“hdfs_llink”。
  - 源目录或文件：待迁移数据的目录或单个文件路径。
  - 文件格式：传输数据时所用的文件格式，这里选择“二进制格式”。不解析文件内容直接传输，不要求文件格式必须为二进制。适用于文件到文件的原样复制。
  - 其他可选参数一般情况下保持默认即可。
- 目的端作业配置
  - 目的连接名称：选择**创建OBS连接**中的“obs_link”。
  - 桶名：待迁移数据的桶。
  - 写入目录：写入数据到OBS服务器的目录。
  - 文件格式：迁移文件类数据到文件时，文件格式选择“二进制格式”。
  - 高级属性里的可选参数一般情况下保持默认即可。

步骤2 单击“下一步”进入字段映射界面，CDM会自动匹配源和目的字段。

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

步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。

- 是否定时执行：如果需要配置作业定时自动执行，可打开此配置。这里保持默认值“否”。
- 抽取并发数：设置同时执行的抽取任务数。CDM支持多个文件的并发抽取，调大参数有利于提高迁移效率
- 是否写入脏数据：否，文件到文件属于二进制迁移，不存在脏数据。
- 作业运行完是否删除：这里保持默认值“不删除”。根据使用场景，也可配置为“删除”，防止迁移作业堆积。

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

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

在历史记录界面单击“日志”，可查看作业的日志信息。

----结束

8.12 Elasticsearch 整库迁移到云搜索服务

操作场景

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

1. **创建CDM集群并绑定EIP**
2. **创建云搜索服务连接**
3. **创建Elasticsearch连接**
4. **创建整库迁移作业**

前提条件

- 拥有EIP配额。
- 已经开通了云搜索服务，且获取云搜索服务集群的IP地址和端口。
- 已获取本地Elasticsearch数据库的服务器IP、端口、用户名和密码。

  如果Elasticsearch服务器是在本地数据中心或第三方云上，需要确保Elasticsearch可通过公网IP访问，或者是已经建立好了企业内部数据中心到华为云的VPN通道或专线。

创建 CDM 集群并绑定 EIP

步骤1 如果是独立CDM服务，参考**创建集群**创建CDM集群；如果是作为DataArts Studio服务CDM组件使用，参考**创建集群**创建CDM集群。

关键配置如下：

- CDM集群的规格，按待迁移的数据量选择，一般选择cdm.medium即可，满足大部分迁移场景。
- CDM集群的VPC必须和云搜索服务集群所在VPC一致，且推荐子网、安全组也与云搜索服务一致。

- 如果安全控制原因不能使用相同子网和安全组，那么需要确保安全组规则能允许CDM访问云搜索服务集群。

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

说明

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

----结束

创建云搜索服务连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如图8-64所示。

图 8-64 选择连接器类型

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

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

图 8-65 创建云搜索服务连接

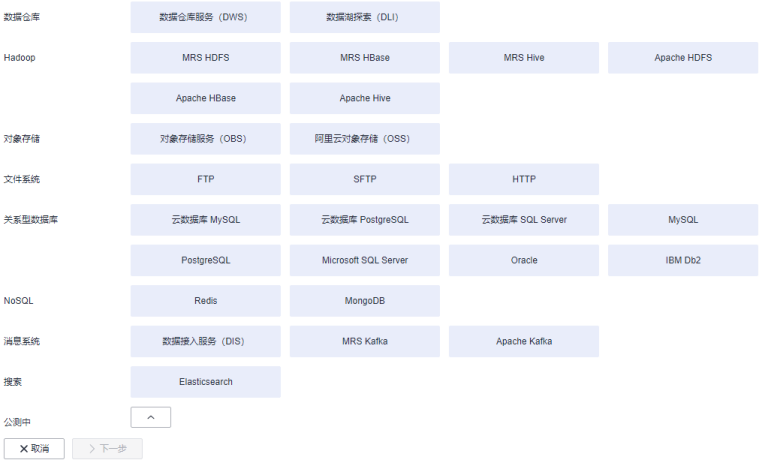
步骤3 单击“保存”回到连接管理界面。

----结束

创建 Elasticsearch 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如图8-66所示。

图 8-66 选择连接器类型



步骤2 连接器类型选择“Elasticsearch”后单击“下一步”，配置Elasticsearch连接参数，Elasticsearch连接参数与云搜索服务的连接参数一样：

- 名称：用户自定义连接名称，例如“es_link”。
- Elasticsearch服务器列表：配置为本地Elasticsearch数据库的IP地址、端口，多个地址之间使用分号（；）分隔。

步骤3 单击“保存”回到连接管理界面。

----结束

创建整库迁移作业

步骤1 选择“整库迁移 > 新建作业”，开始创建Elasticsearch整库迁移到云搜索服务的任务。

图 8-67 创建 Elasticsearch 整库迁移作业

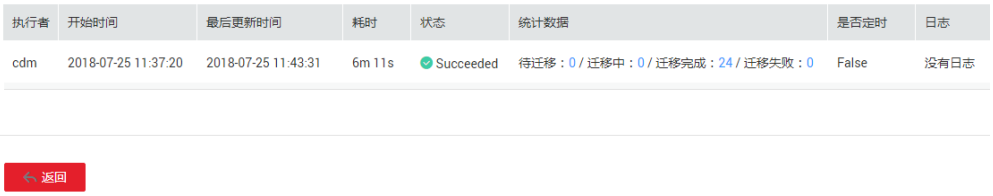
- 作业名称：用户自定义便于记忆、区分的任务名称。
- 源端作业配置
  - 源连接名称：选择**创建Elasticsearch连接**中的“es_link”。
  - 索引：单击输入框后面的按钮，可选择本地Elasticsearch数据库中的一个索引，也可以手动输入索引名称，名称只能全部小写。需要一次迁移多个索引时，这里可配置为通配符，CDM会迁移所有符合通配符条件的索引。例如这里配置为cdm*时，CDM将迁移所有名称为cdm开头的索引：cdm01、cdmB3、cdm_45……
- 目的端作业配置
  - 目的连接名称：选择**创建云搜索服务连接**中的“csslink”。
  - 索引：待写入数据的索引，这里可以选择一个云搜索服务中已存在的索引，也可以手动输入一个不存在的索引名称，名称只能全部小写，CDM会自动在云搜索服务中创建该索引。一次迁移多个索引时，该参数将被禁止配置，CDM自动在目的端创建索引。
  - 导入前清空数据：如果上面选择的索引，在云搜索服务中已存在，这里可以选择导入数据前是否清空该索引中的数据。如果选择不清空，则数据追加写入该索引。

步骤2 作业配置完成后，单击“保存并运行”，回到作业管理界面，在整库迁移的作业管理界面可查看执行进度和结果。

本地Elasticsearch索引中的每个类型都会生成一个子作业并发执行，可以单击作业名查看子作业进度。

步骤3 作业执行完成后，单击作业操作列的“历史记录”，可查看该作业的历史执行记录、读取和写入的统计数据，以及日志信息（子作业才有日志）。

图 8-68 作业执行记录



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
cdm	2018-07-25 11:37:20	2018-07-25 11:43:31	6m 11s	Succeeded	待迁移：0 / 迁移中：0 / 迁移完成：24 / 迁移失败：0	False	没有日志

返回

----结束

