云数据迁移

最佳实践

文档版本 02

发布日期 2024-08-30

华为云计算技术有限公司

版权所有 © 华为云计算技术有限公司 2024。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

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

 版权所有 © 华为云计算技术有限公司

 版权所有 © 华为云计算技术有限公司

1 使用教程

1.1 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1-1**所示。

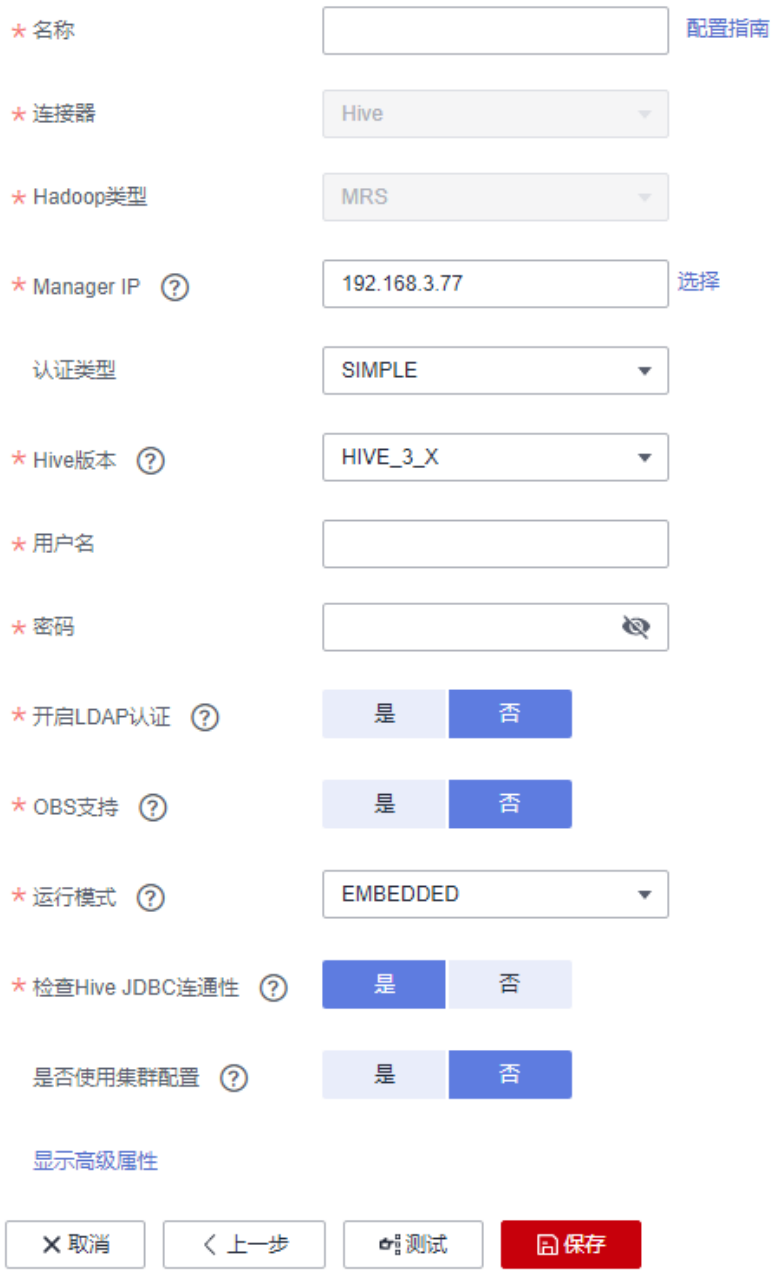
版权所有 © 华为云计算技术有限公司

图 1-1 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1-2所示。

版权所有 © 华为云计算技术有限公司

图 1-2 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 1-1 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

版权所有 © 华为云计算技术有限公司

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

版权所有 © 华为云计算技术有限公司

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1-3所示。 **图 1-3** 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

版权所有 © 华为云计算技术有限公司

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

1.2 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1-4**所示。

版权所有 © 华为云计算技术有限公司

图 1-4 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。

表 1-2 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	Agent功能待下线，无需配置。	-
local_infile字符集	mysql通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	适配mysql的驱动。	-
Agent	Agent功能待下线，无需配置。	-
单次请求行数	指定每次请求获取的行数。	1000

版权所有 © 华为云计算技术有限公司

参数名	说明	取值样例
单次提交行数	可选参数，单击“显示高级属性”后显示。 指定每次批量提交的行数，根据数据目的端和作业数据规模的大小配置该参数。如果配置过大或过小，可能影响作业的时长。	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'
单次写入行数	指定单次批量写入的行数，当写入行数累计到单次批量提交行数时提交一次，该值应当小于单次提交行数。	100

步骤6 单击“保存”回到连接管理界面，完成MySQL连接器的配置。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

1.3 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1-5所示。

版权所有 © 华为云计算技术有限公司

图 1-5 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。

版权所有 © 华为云计算技术有限公司

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 1-6 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1-7**所示。

图 1-7 选择连接器类型

版权所有 © 华为云计算技术有限公司

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。

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1-3**所示。

表 1-3 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。	-
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	Agent功能待下线，无需配置。	-
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

步骤3 单击“保存”回到连接管理界面。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

创建 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1-8**所示。

版权所有 © 华为云计算技术有限公司

图 1-8 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。

各参数说明如**表1-4**所示，需要您根据实际情况配置。

表 1-4 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
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

版权所有 © 华为云计算技术有限公司

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

版权所有 © 华为云计算技术有限公司

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1-9所示。 **图 1-9** 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

版权所有 © 华为云计算技术有限公司

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

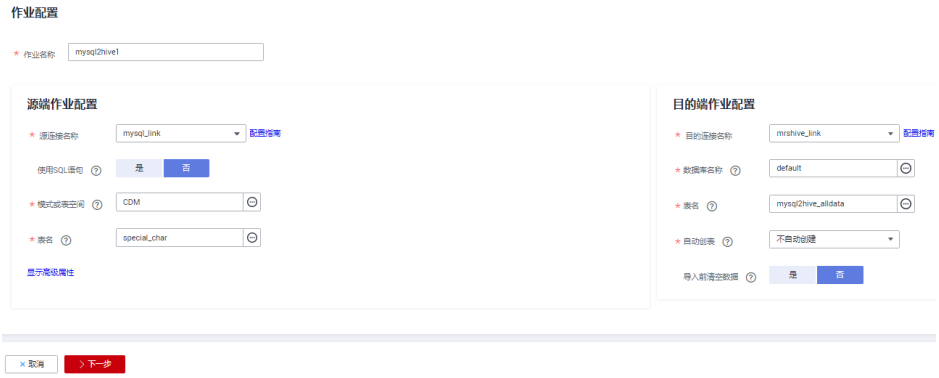
步骤3 单击“保存”回到连接管理界面。

----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1-10**所示。

图 1-10 创建 MySQL 到 Hive 的迁移任务



说明

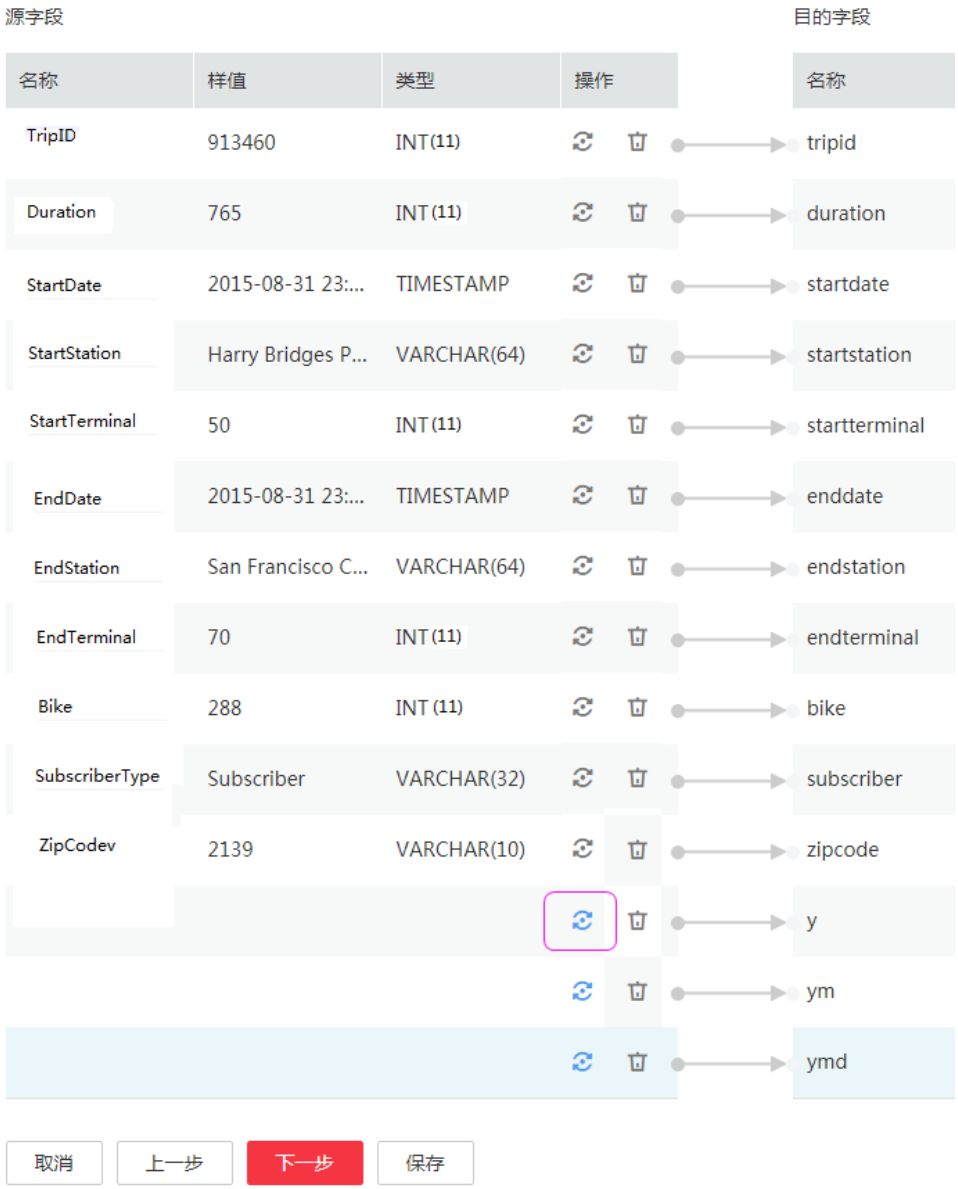
“导入前清空数据”选“是”，这样每次导入前，会将之前已经导入到Hive表的数据清空。

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1-11**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。

版权所有 © 华为云计算技术有限公司

图 1-11 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1-12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")

版权所有 © 华为云计算技术有限公司

图 1-12 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**。

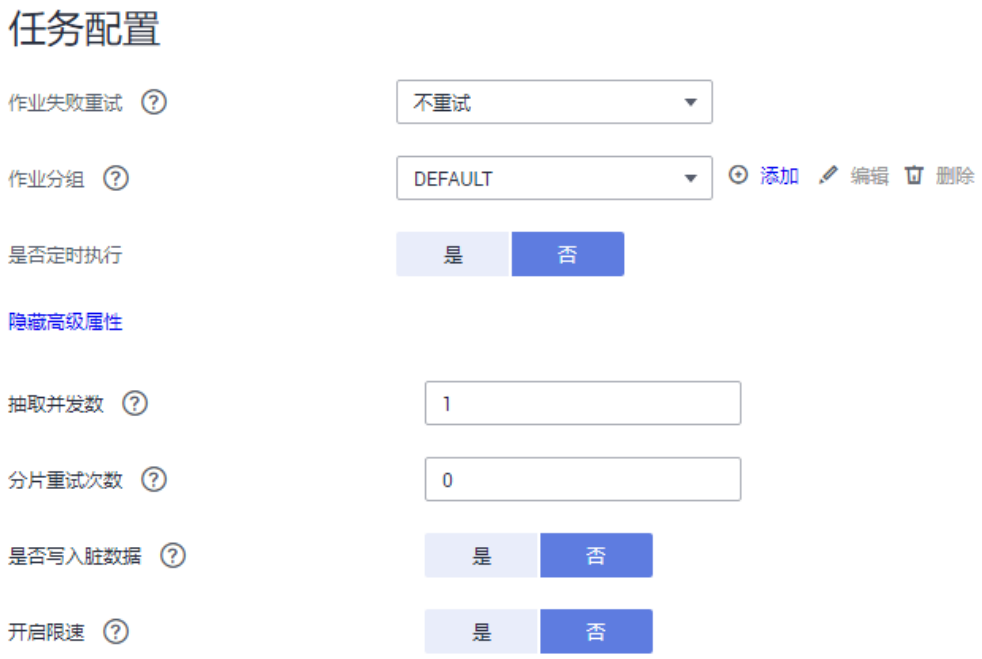
步骤4 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

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。

版权所有 © 华为云计算技术有限公司

图 1-13 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

1.4 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。

版权所有 © 华为云计算技术有限公司

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1-14**所示。

图 1-14 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。

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1-5**所示。

表 1-5 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

版权所有 © 华为云计算技术有限公司

参数名	说明	取值样例
数据库服务器	MySQL数据库的IP地址或域名。	-
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	Agent功能待下线，无需配置。	-
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

步骤3 单击“保存”回到连接管理界面。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

创建 OBS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图1-15**所示。

版权所有 © 华为云计算技术有限公司

图 1-15 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1-17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1-16所示。

图 1-16 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。

版权所有 © 华为云计算技术有限公司

图 1-17 创建 OBS 连接

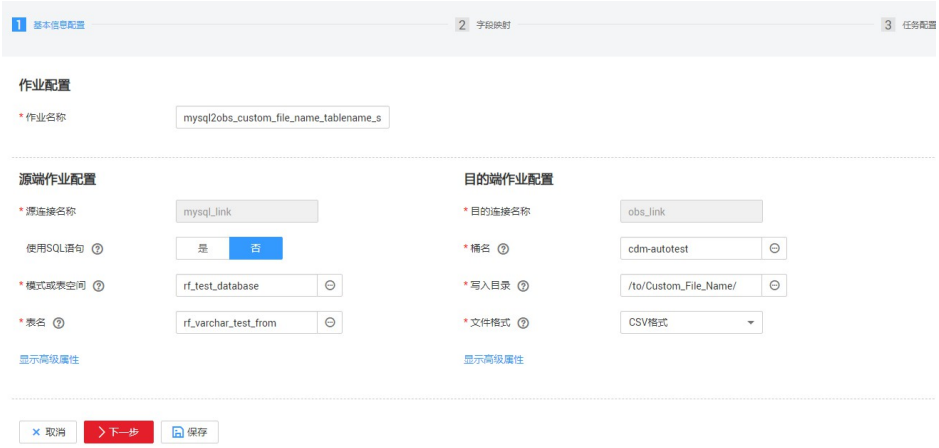
步骤3 单击“保存”回到连接管理界面。

----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 1-18 创建 MySQL 到 OBS 的迁移任务



 版权所有 © 华为云计算技术有限公司

- 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1-19**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 1-19 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。

版权所有 © 华为云计算技术有限公司

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

1.5 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1-20**所示。

版权所有 © 华为云计算技术有限公司

图 1-20 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。

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1-6**所示。

表 1-6 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
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	Agent功能待下线，无需配置。	-
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8

版权所有 © 华为云计算技术有限公司

参数名	说明	取值样例
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

步骤3 单击“保存”回到连接管理界面。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

创建 DWS 连接

步骤1 在CDM集群管理界面，单击集群后的“作业管理”，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，如**图1-21**所示。

图 1-21 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1-7**所示，可选参数保持默认即可。

表 1-7 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

版权所有 © 华为云计算技术有限公司

参数名	说明	取值样例
数据库名称	DWS数据库的名称。	db_demo
用户名	拥有DWS数据库的读、写和删除权限的用户。	dbadmin
密码	用户的密码。	-
使用Agent	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
导入模式	COPY模式：将源数据经过DWS管理节点后复制到数据节点。如果需要通过Internet访问DWS，只能使用COPY模式。	COPY

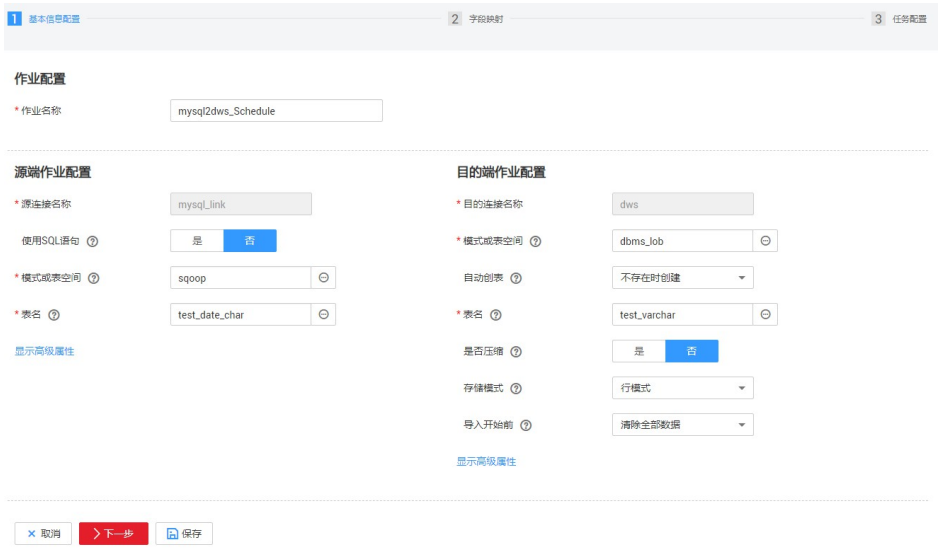
步骤3 单击“保存”完成创建连接。

----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从MySQL导出数据到DWS的任务。

图 1-22 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。

 版权所有 © 华为云计算技术有限公司

  - 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1-23**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 1-23 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。

 版权所有 © 华为云计算技术有限公司

- 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

1.6 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。

 版权所有 © 华为云计算技术有限公司

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 1-24 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1-25所示。

图 1-25 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。

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1-8**所示。

表 1-8 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

版权所有 © 华为云计算技术有限公司

参数名	说明	取值样例
数据库服务器	MySQL数据库的IP地址或域名。	-
端口	MySQL数据库的端口。	3306
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	Agent功能待下线，无需配置。	-
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

步骤3 单击“保存”回到连接管理界面。

说明

如果保存时出错，一般是由于MySQL数据库的安全设置问题，需要设置允许CDM集群的EIP访问MySQL数据库。

----结束

创建 RDS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图1-26**所示。

版权所有 © 华为云计算技术有限公司

图 1-26 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1-27所示。

版权所有 © 华为云计算技术有限公司

图 1-27 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

 版权所有 © 华为云计算技术有限公司

1.7 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1-28**所示。

版权所有 © 华为云计算技术有限公司

图 1-28 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。

版权所有 © 华为云计算技术有限公司

图 1-29 创建云搜索服务连接

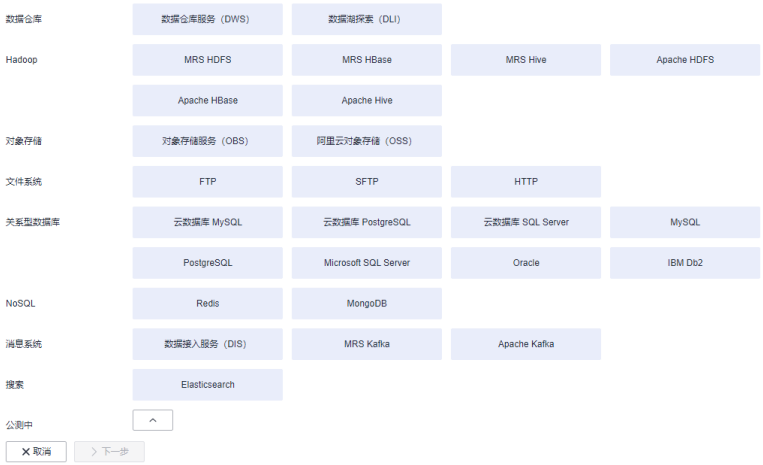
步骤3 单击“保存”回到连接管理界面。

----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1-30所示。

图 1-30 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：

版权所有 © 华为云计算技术有限公司

- 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。

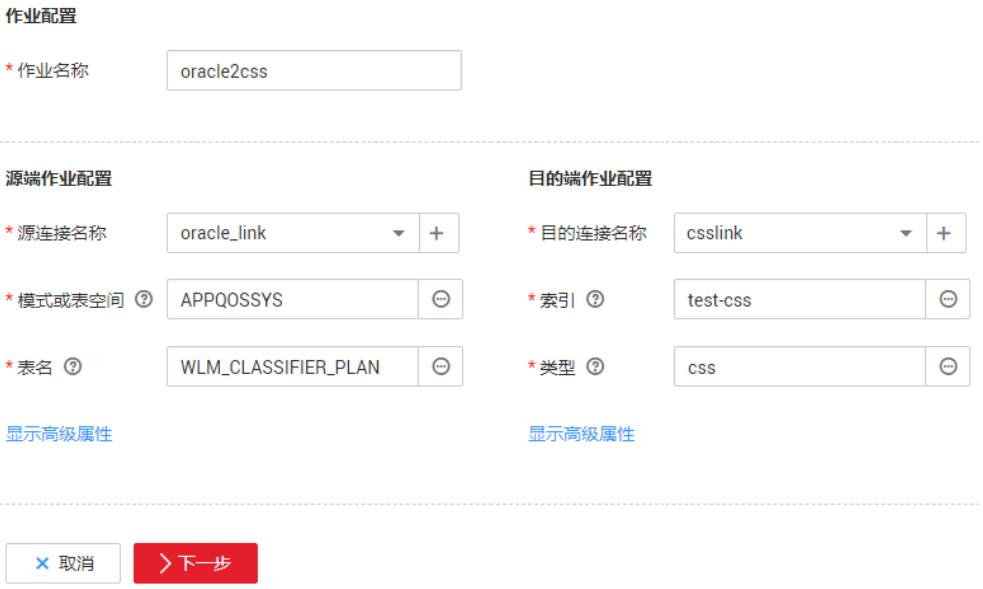
步骤3 单击“保存”回到连接管理界面。

----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 1-31 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。

版权所有 © 华为云计算技术有限公司

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1-32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 1-32 云搜索服务的字段映射

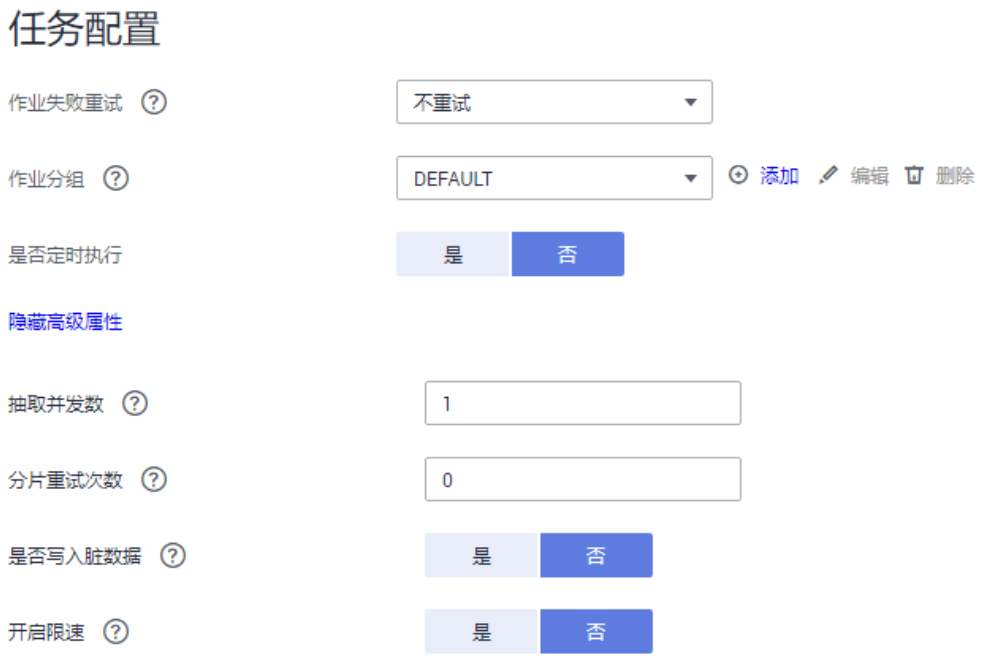
步骤3 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

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。

版权所有 © 华为云计算技术有限公司

图 1-33 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

1.8 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。

版权所有 © 华为云计算技术有限公司

- 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1-34**所示。

图 1-34 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1-9**所示。

版权所有 © 华为云计算技术有限公司

图 1-35 创建 Oracle 连接

版权所有 © 华为云计算技术有限公司

表 1-9 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	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
ORACLE版本	默认使用最新版本驱动，若不兼容请尝试其他版本。	高于12.1
驱动版本	需要适配的驱动。	-
一次请求行数	指定每次请求获取的行数。	1000
连接属性	自定义连接属性。	useCompression=true
引用符号	连接引用表名或列名时的分隔符号。默认为空。	'

步骤3 单击“保存”回到连接管理界面。

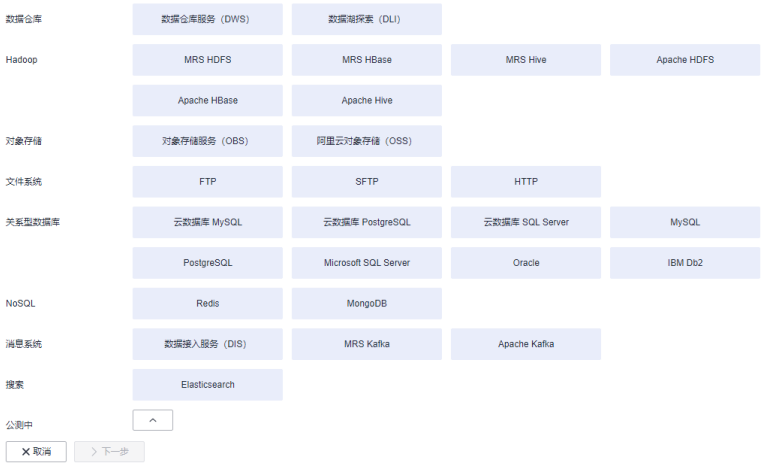
----结束

创建 DWS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如图1-36所示。

版权所有 © 华为云计算技术有限公司

图 1-36 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1-10**所示，可选参数保持默认即可。

表 1-10 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	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
导入模式	COPY模式：将源数据经过DWS管理节点后复制到数据节点。如果需要通过Internet访问DWS，只能使用COPY模式。	COPY

步骤3 单击“保存”完成创建连接。

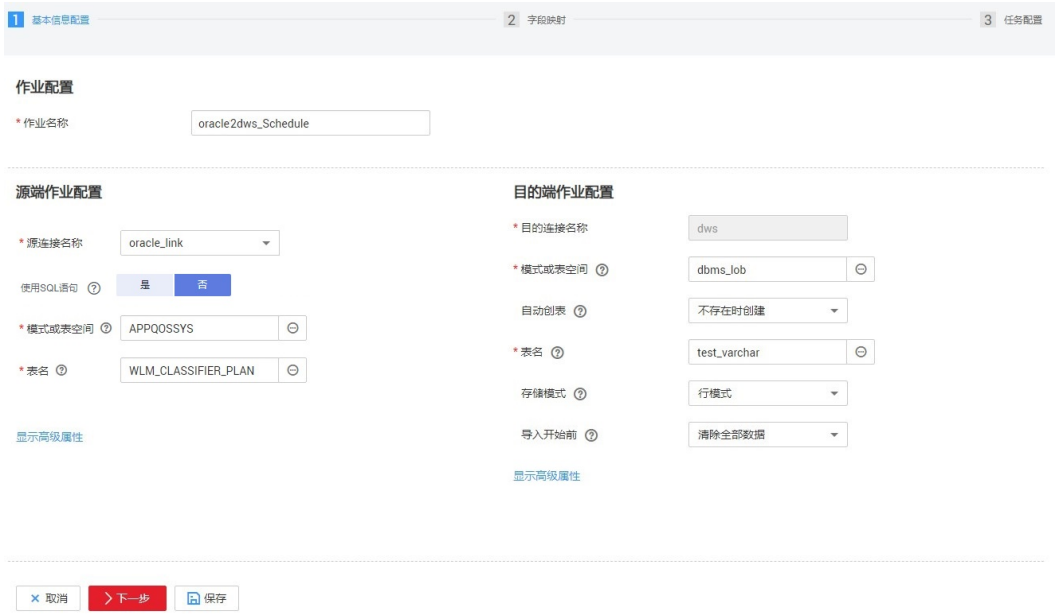
----结束

创建迁移作业

步骤1 选择“表/文件迁移 > 新建作业”，开始创建从Oracle导出数据到DWS的任务。

版权所有 © 华为云计算技术有限公司

图 1-37 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1-38**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**。

 版权所有 © 华为云计算技术有限公司

图 1-38 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。

1.9 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**

版权所有 © 华为云计算技术有限公司

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1-39**所示。

图 1-39 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。

版权所有 © 华为云计算技术有限公司

图 1-40 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1-41所示。

版权所有 © 华为云计算技术有限公司

图 1-41 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1-43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1-42所示。

图 1-42 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。

版权所有 © 华为云计算技术有限公司

图 1-43 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。

 版权所有 © 华为云计算技术有限公司

图 1-44 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1-45**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**。

 版权所有 © 华为云计算技术有限公司

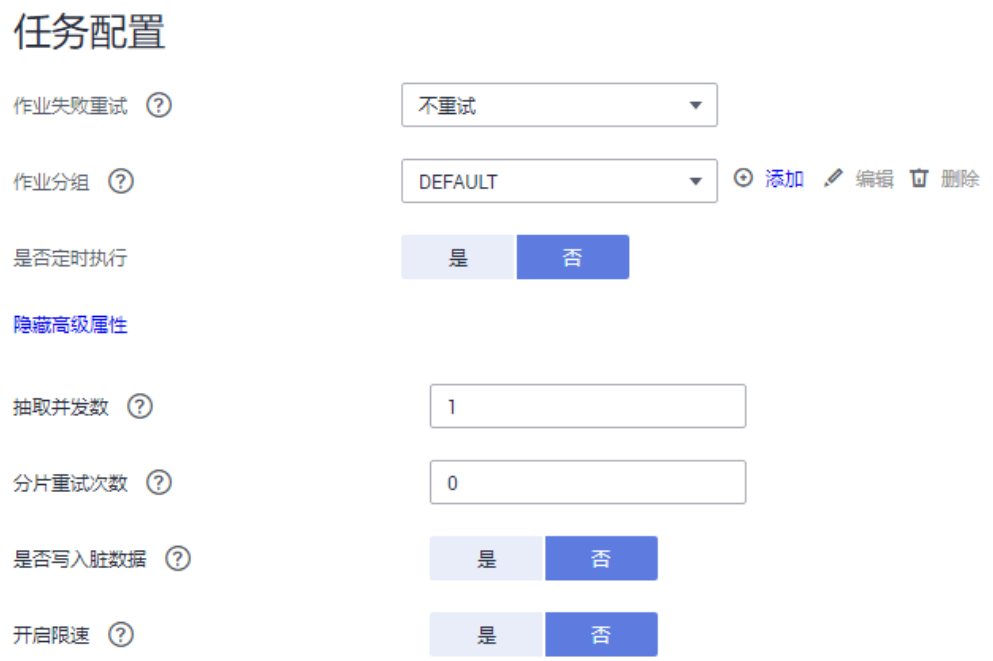
图 1-45 云搜索服务的字段映射

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 1-46 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。

版权所有 © 华为云计算技术有限公司

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

1.10 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1-47**所示。

 版权所有 © 华为云计算技术有限公司

图 1-47 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1-48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 1-48 创建 DLI 连接

版权所有 © 华为云计算技术有限公司

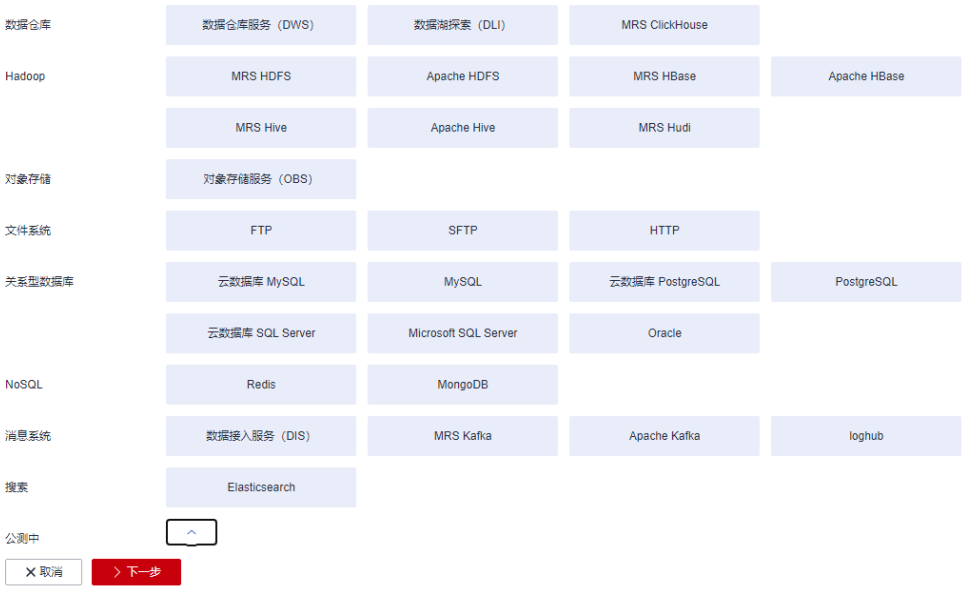
步骤3 单击“保存”回到连接管理界面。

----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1-49所示。

图 1-49 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1-51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1-50所示。

图 1-50 单击新增访问密钥

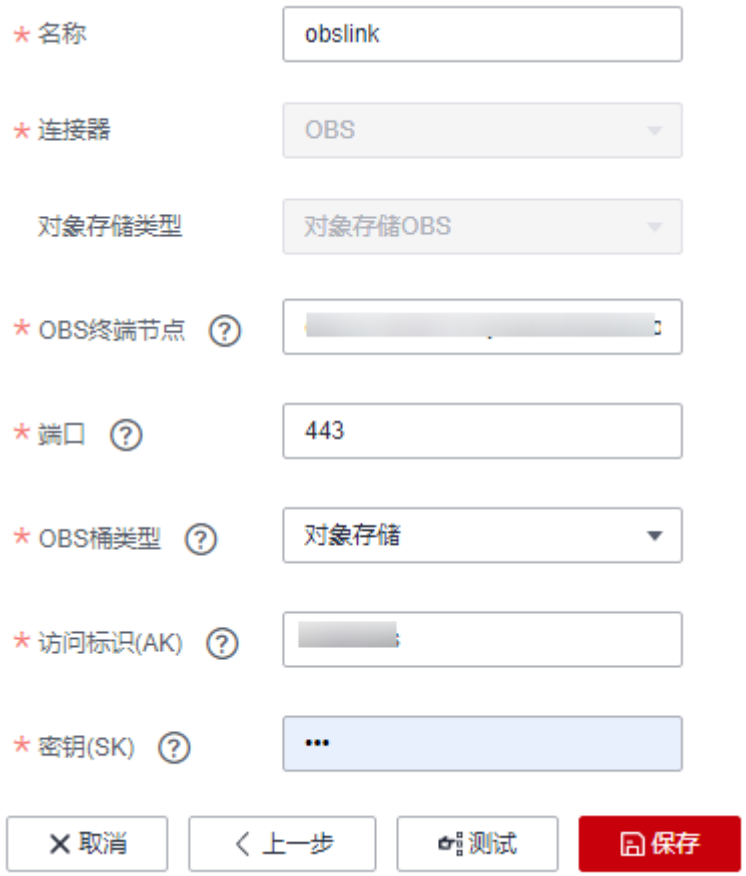
  c. 单击“确定”，根据浏览器提示，保存密钥文件。密钥文件会直接保存到浏览器默认的下载文件夹中。打开名称为“credentials.csv”的文件，即可查看访问密钥（Access Key Id和Secret Access Key）。

版权所有 © 华为云计算技术有限公司

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 1-51 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1-52所示。

版权所有 © 华为云计算技术有限公司

图 1-52 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”。

版权所有 © 华为云计算技术有限公司

- 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 1-53 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

1.11 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**

版权所有 © 华为云计算技术有限公司

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1-54**所示。

图 1-54 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。

版权所有 © 华为云计算技术有限公司

- 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1-55所示。

图 1-55 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1-57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1-56所示。

版权所有 © 华为云计算技术有限公司

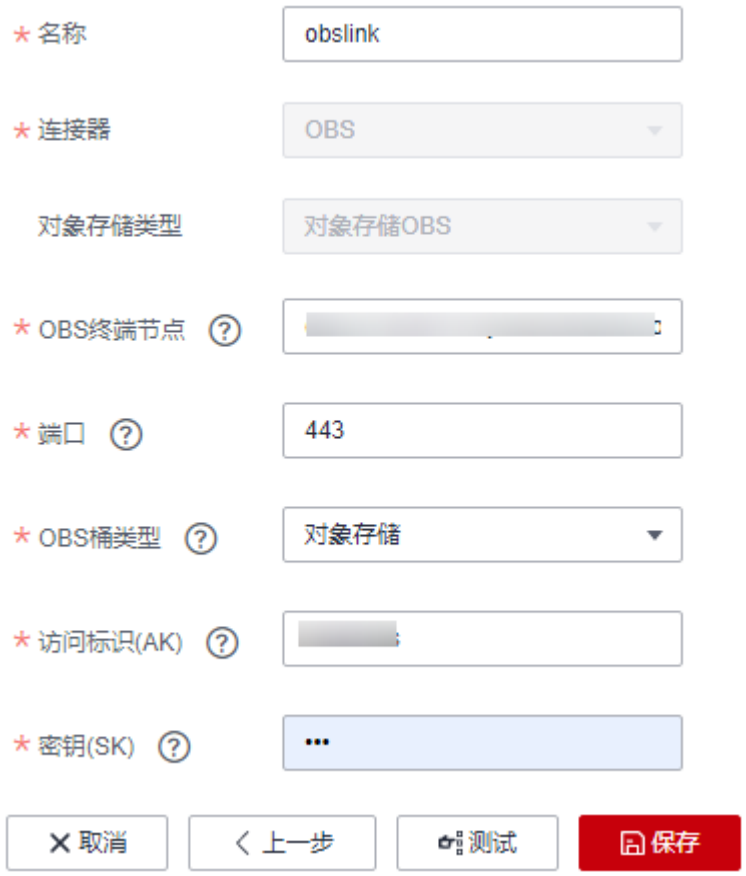
图 1-56 单击新增访问密钥

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 1-57 创建 OBS 连接



步骤3 单击“保存”回到连接管理界面。

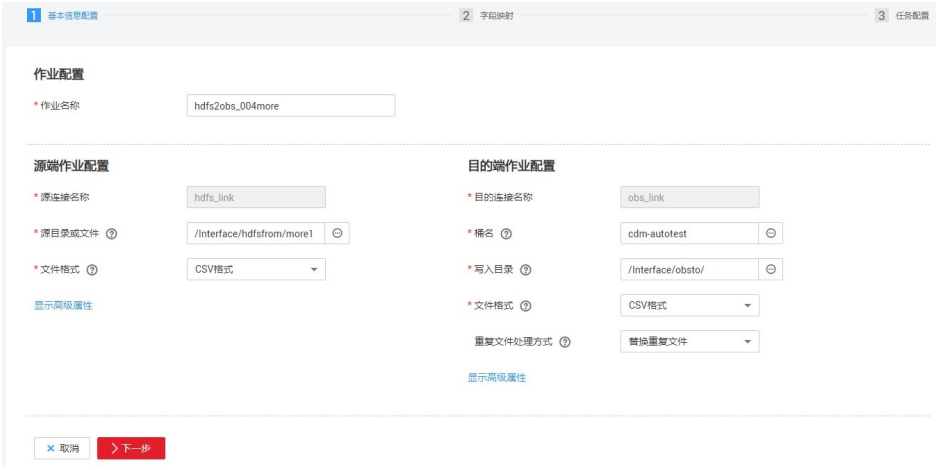
----结束

版权所有 © 华为云计算技术有限公司

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 1-58 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。

版权所有 © 华为云计算技术有限公司

- 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

1.12 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。

版权所有 © 华为云计算技术有限公司

- 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1-59所示。

图 1-59 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。

版权所有 © 华为云计算技术有限公司

图 1-60 创建云搜索服务连接

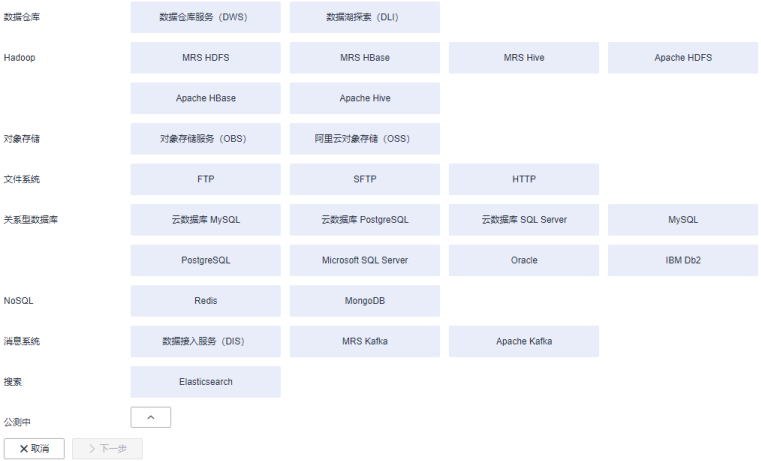
步骤3 单击“保存”回到连接管理界面。

----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1-61所示。

图 1-61 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：

 版权所有 © 华为云计算技术有限公司

- 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 1-62 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。

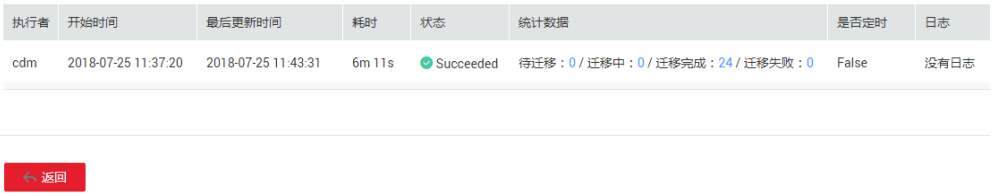
 版权所有 © 华为云计算技术有限公司

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 1-63 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

版权所有 © 华为云计算技术有限公司

2 进阶实践

2.1 增量迁移原理介绍

2.1.1 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。

版权所有 © 华为云计算技术有限公司

- 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 2-1 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：

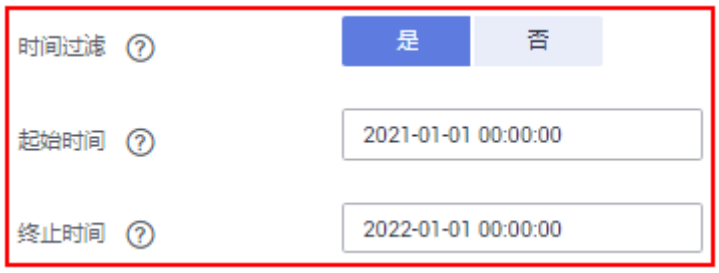
  例如需要CDM只同步2021年1月1日~2022年1月1日生成的文件到目的端，则参数配置如下：

  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 2-2 时间过滤

版权所有 © 华为云计算技术有限公司

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。

2.1.2 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2-3**所示，参数配置如下：

版权所有 © 华为云计算技术有限公司

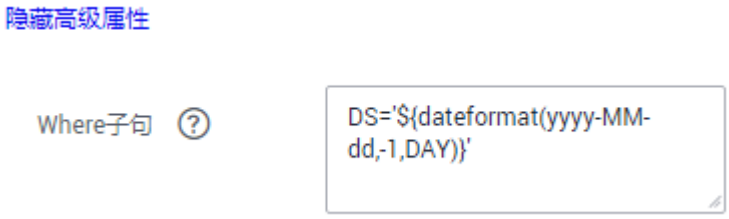
图 2-3 表数据

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 2-4 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。

2.1.3 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。

版权所有 © 华为云计算技术有限公司

图 2-5 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。

2.1.4 MongoDB/DDS 增量迁移

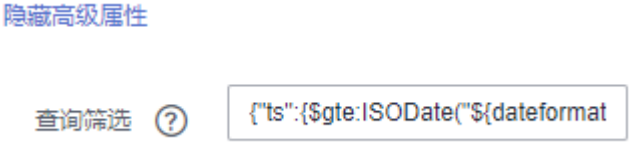
使用CDM导出MongoDB或者DDS的数据时，支持导出指定时间段内的数据，配合CDM的定时任务，可以实现MongoDB/DDS的增量迁移。

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 2-6 MongoDB 查询筛选



版权所有 © 华为云计算技术有限公司

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。

2.2 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）。

版权所有 © 华为云计算技术有限公司

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2-1**所示。

表 2-1 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

版权所有 © 华为云计算技术有限公司

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2-7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 2-7 源表名和写入目录配置为时间宏变量

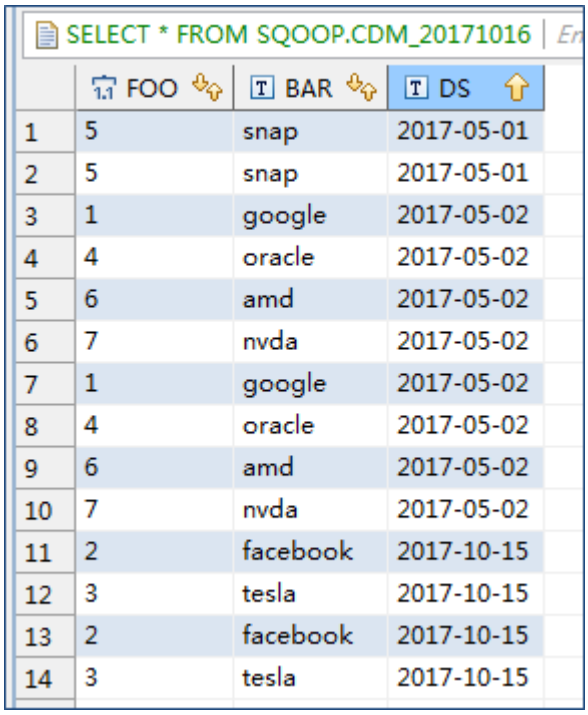
目前也支持一个表名或路径名中有多个宏定义变量，例如“/opt/ttxx/${dateformat(yyyy-MM-dd)}/${timestamp()}”，经过转换后为“/opt/ttxx/2017-10-16/1508115701746”。

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2-8所示。

版权所有 © 华为云计算技术有限公司

图 2-8 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。

2.3 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。

版权所有 © 华为云计算技术有限公司

- 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 2-9 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。

2.4 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。

版权所有 © 华为云计算技术有限公司

- 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。

版权所有 © 华为云计算技术有限公司

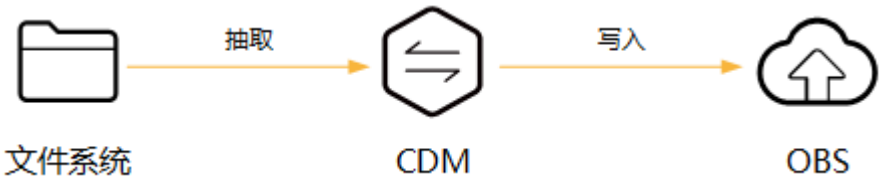
说明

- 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。

2.5 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2-10**所示。

图 2-10 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**。

版权所有 © 华为云计算技术有限公司

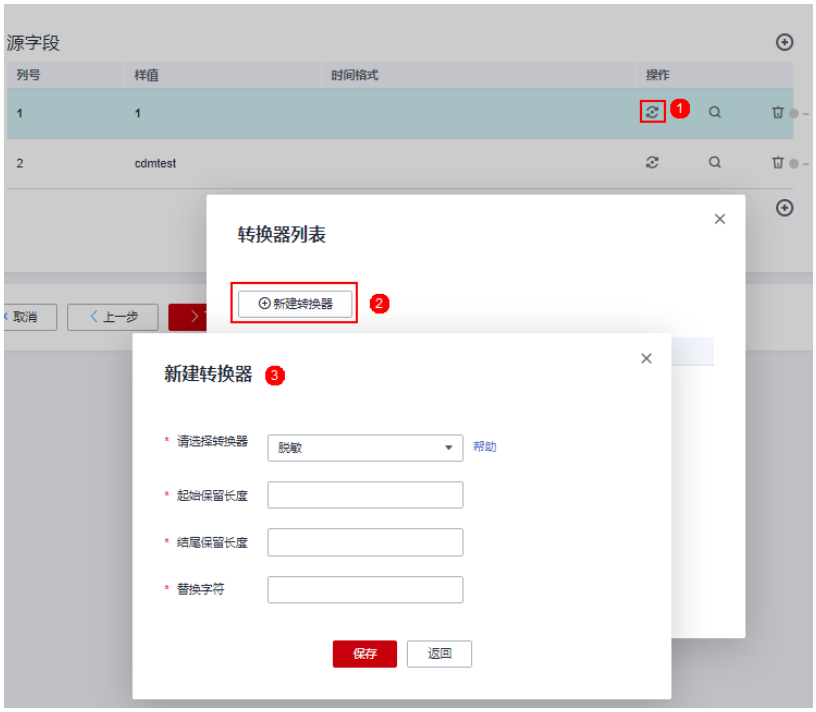
2.6 字段转换器配置指导

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 2-11 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。

版权所有 © 华为云计算技术有限公司

- 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 2-12 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 2-13 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：

版权所有 © 华为云计算技术有限公司

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”。

 版权所有 © 华为云计算技术有限公司

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。

版权所有 © 华为云计算技术有限公司

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)

版权所有 © 华为云计算技术有限公司

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*")

版权所有 © 华为云计算技术有限公司

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)

版权所有 © 华为云计算技术有限公司

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。

版权所有 © 华为云计算技术有限公司

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")

2.7 指定文件名迁移

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

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

说明

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

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

2.8 正则表达式分隔半结构化文本

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

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

图 2-14 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：

版权所有 © 华为云计算技术有限公司

- 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 2-2 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 2-3 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

版权所有 © 华为云计算技术有限公司

列号	样值
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 2-4 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 2-5 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

版权所有 © 华为云计算技术有限公司

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 2-6 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

2.9 记录数据迁移入库时间

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

前提条件

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

创建表/文件迁移作业

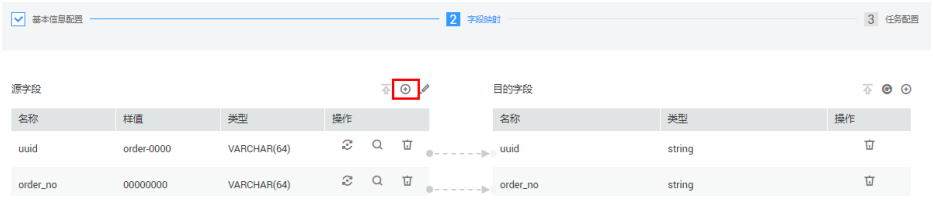
步骤1 在创建表/文件迁移作业时，选择已创建的源端连接器、目的端连接器。

图 2-15 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。

 版权所有 © 华为云计算技术有限公司

图 2-16 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2-7**。

图 2-17 添加字段

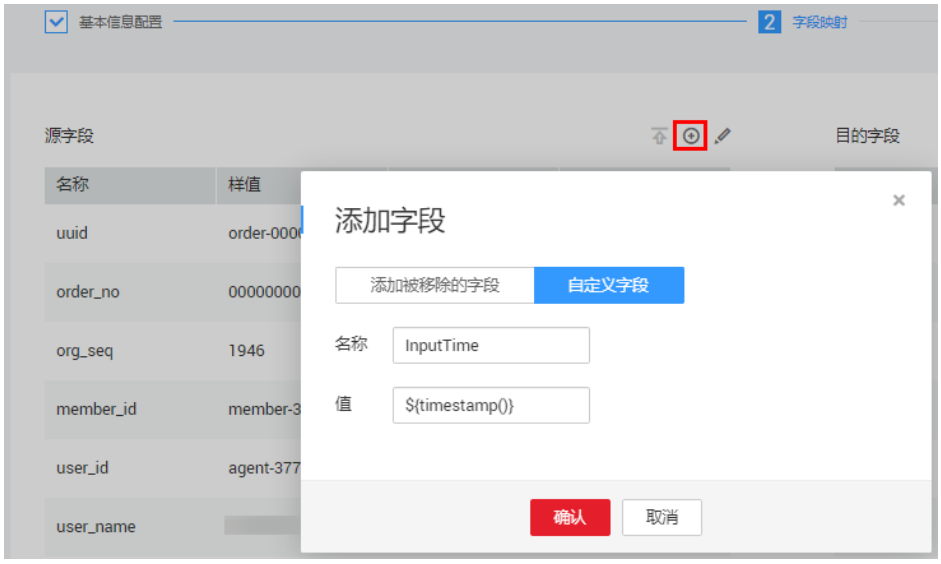


表 2-7 时间变量宏定义具体展示

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

版权所有 © 华为云计算技术有限公司

宏变量	含义	实际显示效果
${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

2.10 文件格式介绍

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

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

版权所有 © 华为云计算技术有限公司

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 2-8 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2-8**。

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。

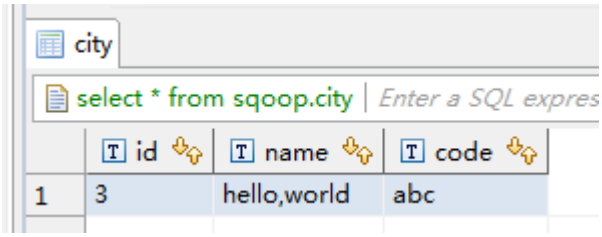
4. **使用包围符**

   - 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

版权所有 © 华为云计算技术有限公司

（""）作为包围符。如图2-18所示，数据库的name字段的值中包含了字段分隔符逗号：

图 2-18 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。

版权所有 © 华为云计算技术有限公司

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
}
```


版权所有 © 华为云计算技术有限公司

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 2-9 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 2-10 示例

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

版权所有 © 华为云计算技术有限公司

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 2-11 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 2-12 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。

版权所有 © 华为云计算技术有限公司

图 2-19 添加字段

ii. 在添加的新字段后面，单击 添加字段转换器。

图 2-20 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。

版权所有 © 华为云计算技术有限公司

图 2-21 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。

版权所有 © 华为云计算技术有限公司

- **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2-8**。

   - 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。

版权所有 © 华为云计算技术有限公司

3 通过数据开发使用参数传递灵活调度 CDM 作业

如果CDM作业接收来自数据开发作业配置的参数，则在数据开发模块可以使用诸如EL表达式传递动态参数来调度CDM作业。

说明

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

前提条件

已购买数据集成增量包。

创建 CDM 迁移作业

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

步骤2 在集群管理页面，单击集群操作列“作业管理”，进入“作业管理”页面，如图3-1所示。

图 3-1 集群管理

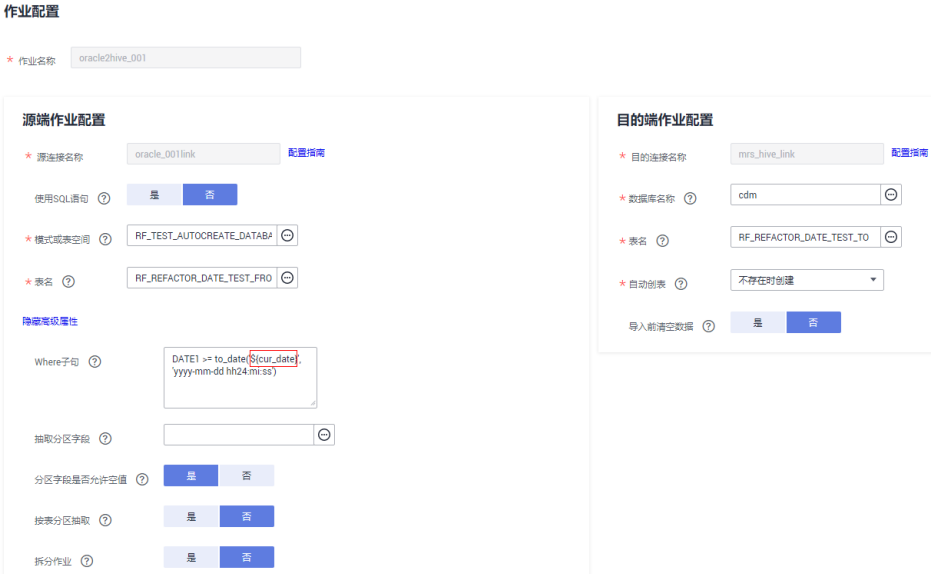
步骤3 在“连接管理”页签中，单击“新建连接”，分别创建Oracle数据连接和MRS Hive数据连接，详情请参见**新建Oracle数据连接**和**新建MRS Hive数据连接**。

步骤4 在“表/文件迁移”页签中，单击“新建作业”，创建数据迁移作业。

步骤5 配置Oracle源端参数、MRS hive目的端参数，并配置传递参数，参数形式为 ${varName}，本示例参数为${cur_date}，如图3-2所示。

 版权所有 © 华为云计算技术有限公司

图 3-2 配置作业



说明

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

----结束

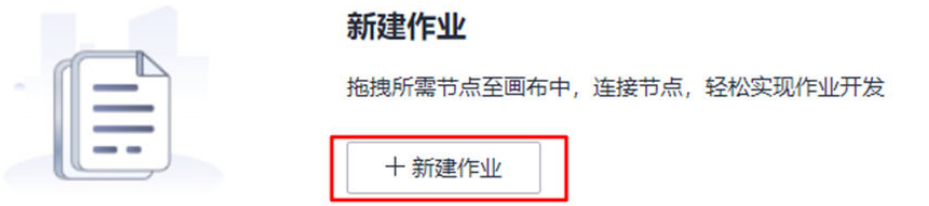
创建并执行数据开发作业

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

步骤2 在数据开发主界面的左侧导航栏，选择“数据开发 > 作业开发” 。

步骤3 在“作业开发”界面中，单击“新建作业”，如图3-3所示。

图 3-3 新建作业



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

 版权所有 © 华为云计算技术有限公司

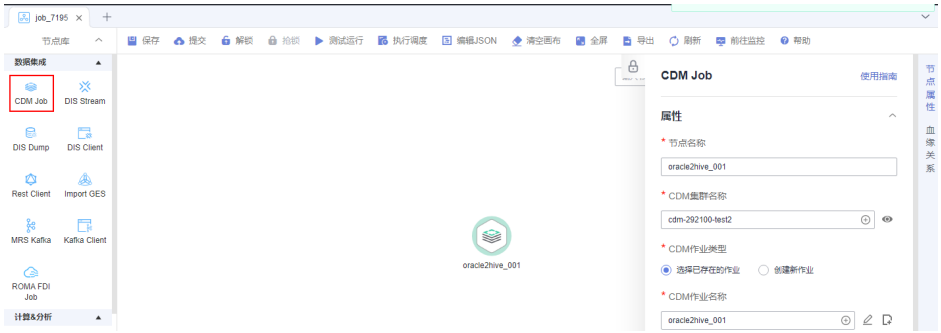
表 3-1 作业参数

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

步骤5 在数据开发作业中添加CDM Job节点，并关联已创建的CDM作业，如**图3-4**所示。

 版权所有 © 华为云计算技术有限公司

图 3-4 关联 CDM 作业



步骤6 在作业参数中配置业务需要的参数，如图3-5所示。

图 3-5 配置作业参数

说明

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

步骤7 保存并提交作业版本，单击“测试运行”，执行数据开发作业。

步骤8 数据开发作业执行成功后，单击右上角的“前往监控”，进入“作业监控”页面，查看生成的任务或实例是否符合需求，如图3-6所示。

版权所有 © 华为云计算技术有限公司

图 3-6 查看运行结果

----结束

 版权所有 © 华为云计算技术有限公司

4 通过数据开发实现数据增量迁移

DataArts Studio服务的DLF组件提供了一站式的大数据协同开发平台，借助DLF的在线脚本编辑、周期调度CDM的迁移作业，也可以实现增量数据迁移。

这里以DWS导入到OBS为例，介绍DLF配合CDM实现增量迁移的流程：

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

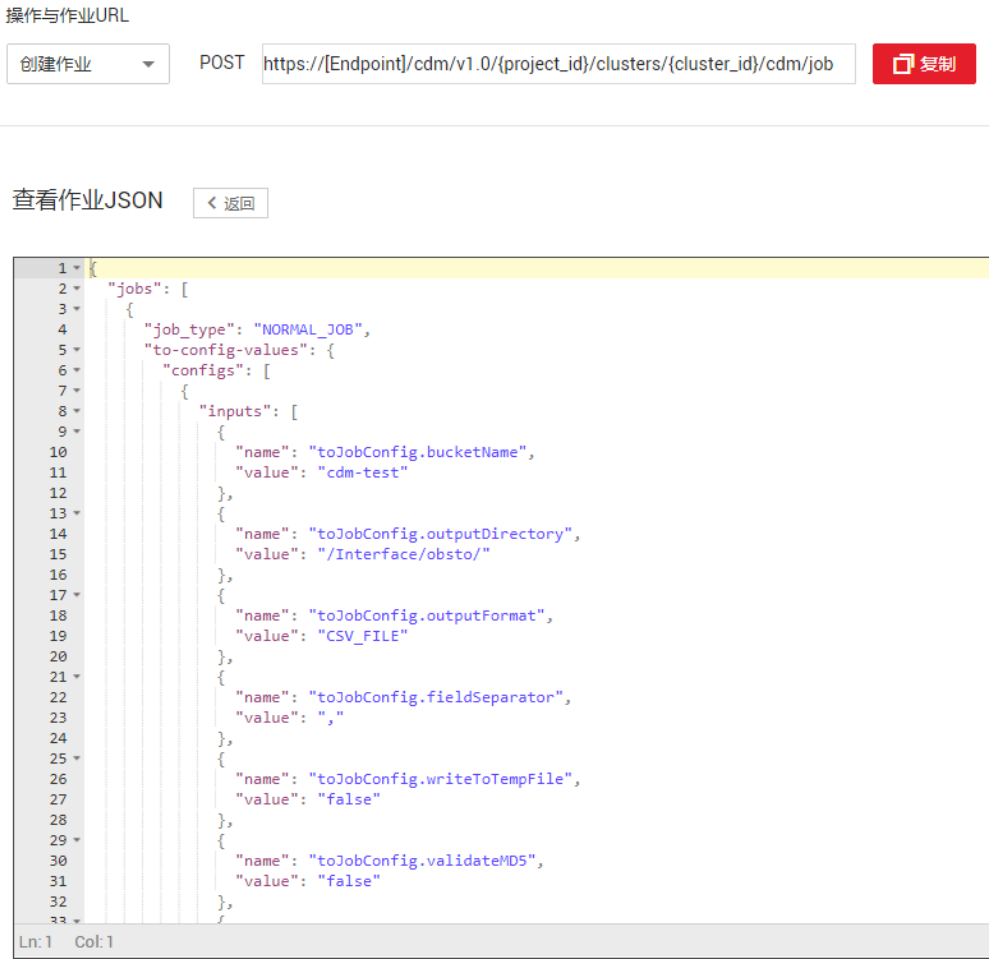
获取 CDM 作业的 JSON

1. 进入CDM主界面，创建一个DWS到OBS的表/文件迁移作业。
2. 在CDM“作业管理”界面的“表/文件迁移”页签下，找到已创建的作业，单击作业操作列的“更多 > 查看作业JSON”，如**图4-1**所示。

   您也可以使用其它已创建好的CDM作业JSON。

版权所有 © 华为云计算技术有限公司

图 4-1 查看作业 JSON



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

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

修改 JSON

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

1. 修改where子句，增量某个时间段的数据：

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


版权所有 © 华为云计算技术有限公司

说明

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

2. 每个周期的增量数据导入到不同的目录：

   ```
   {
     "name": "toJobConfig.outputDirectory",
     "value": "dws2obs/${currentTime}"
   }
   ```

3. 作业名改成动态的，否则会因为作业重名而无法创建：

   ```
   "to-connector-name": "obs-connector",
   "from-link-name": "dws_link",
   "name": "dws2obs-${currentTime}"
   ```

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

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

 版权所有 © 华为云计算技术有限公司

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

 版权所有 © 华为云计算技术有限公司

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

创建 DLF 作业

1. 在DLF创建如图4-2所示的Rest Client节点数据开发作业，详细操作请参见《数据治理中心DataArts Studio 用户指南》的新建作业章节。

   各节点与作业的配置详情请参见后续步骤。

   图 4-2 DLF 作业

2. 配置“创建作业”节点。

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

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

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

   c. HTTP方法：创建CDM作业的HTTP请求方法为“POST”。

   d. 添加如下两个请求头：

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

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

版权所有 © 华为云计算技术有限公司

图 4-3 创建 CDM 作业的节点属性

3. 配置“运行作业”节点。

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

   a. 节点名称：运行作业。

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

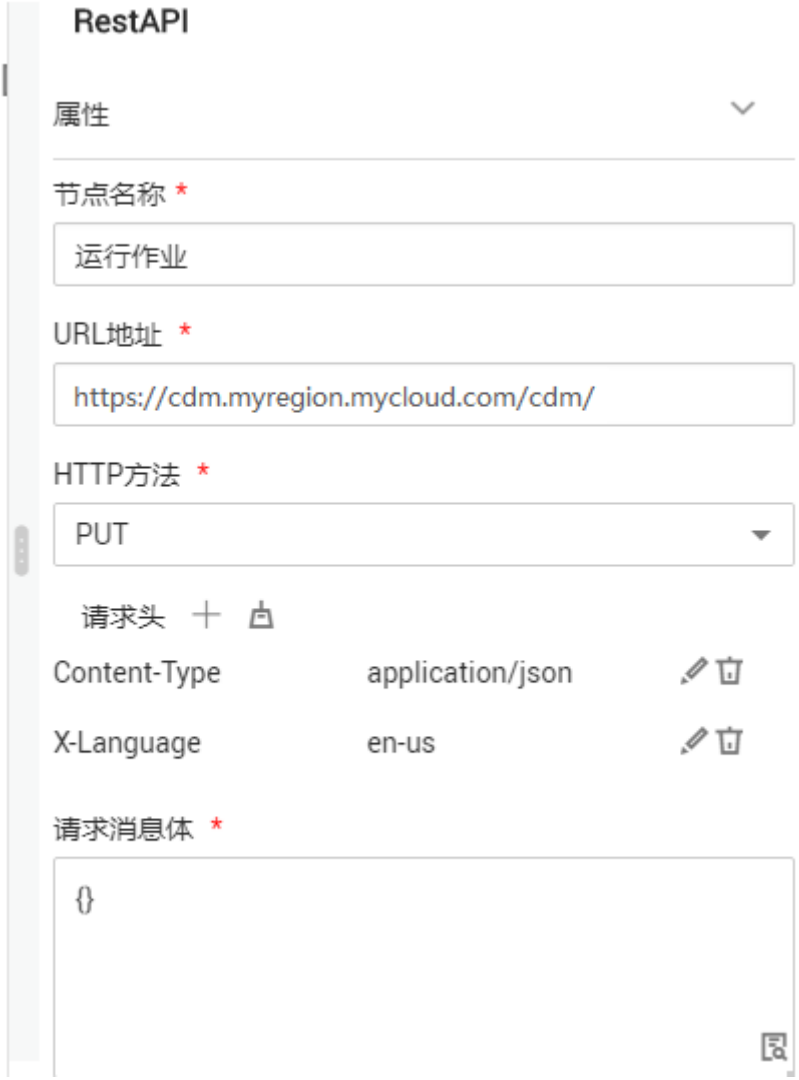
   c. HTTP方法：运行CDM作业的HTTP请求方法为“PUT”。

   d. 请求头：

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

 版权所有 © 华为云计算技术有限公司

图 4-4 运行 CDM 作业的节点属性



4. 配置“等待作业运行完成”节点。

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

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

   a. 节点名称：等待作业运行完成。

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

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

   d. 请求头：

      - Content-Type = application/json

 版权所有 © 华为云计算技术有限公司

- X-Language = en-us

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

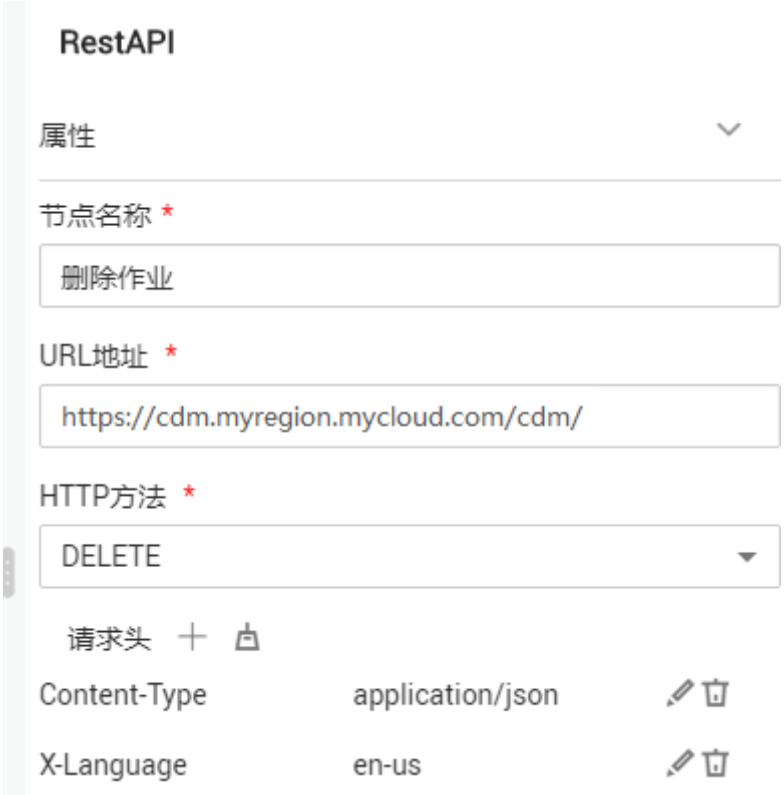
5. （可选）配置“删除作业运行完成”节点。

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

删除CDM作业的节点属性为：

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

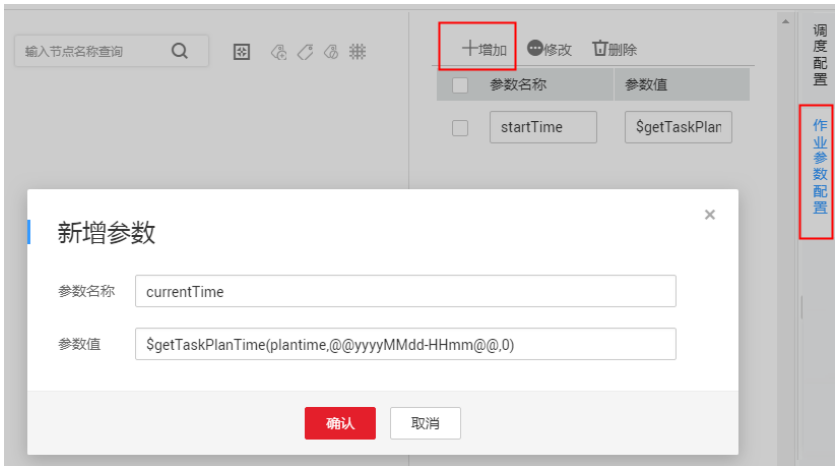
图 4-5 删除 CDM 作业节点配置



 版权所有 © 华为云计算技术有限公司

6. 如果需要在迁移完后进行计算操作，可在后续添加各种计算节点，完成数据计算。
7. 配置DLF作业参数。
   a. 配置DLF作业参数，如图4-6所示。
      - startTime = $getTaskPlanTime(plantime,@@yyyyMMddHHmmss@@,-24*60*60)
      - currentTime = $getTaskPlanTime(plantime,@@yyyyMMdd-HHmm@@,0)

      图 4-6 DLF 作业参数配置

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

版权所有 © 华为云计算技术有限公司

5 通过 CDM 节点批量创建分表迁移作业

适用场景

业务系统中，数据源往往会采用分表的形式，以减少单表大小，支持复杂的业务应用场景。

在这种情况下，通过CDM进行数据集成时，需要针对每张表创建一个数据迁移作业。您可以参考本教程，通过数据开发模块的For Each节点和CDM节点，配合作业参数，实现批量创建分表迁移作业。

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

操作前提

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

创建连接

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

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

步骤3 单击CDM集群“操作”列的“作业管理”，进入作业管理界面。

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

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

表 5-1 MySQL 数据库连接参数

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

版权所有 © 华为云计算技术有限公司

参数名	说明	取值样例
数据库服务器	配置为要连接的数据库的IP地址或域名。 单击输入框后的“选择”，可获取用户的MySQL数据库实例列表。	192.168.0.1
端口	配置为要连接的数据库的端口。	3306
数据库名称	配置为要连接的数据库名称。	dbname
用户名	待连接数据库的用户。该数据库用户需要有数据表的读写权限，以及对元数据的读取权限。	cdm
密码	用户名密码。	-
使用本地API	可选参数，选择是否使用数据库本地API加速。 创建MySQL连接时，CDM会自动尝试启用MySQL数据库的local_infile系统变量，开启MySQL的LOAD DATA功能加快数据导入，提高导入数据到MySQL数据库的性能。注意，开启本参数后，日期类型将不符合格式的会存储为0000-00-00，更多详细信息可在MySQL官网文档查看。 如果CDM自动启用失败，请联系数据库管理员启用local_infile参数或选择不使用本地API加速。 如果是导入到RDS上的MySQL数据库，由于RDS上的MySQL默认没有开启LOAD DATA功能，所以同时需要修改MySQL实例的参数组，将“local_infile”设置为“ON”，开启该功能。 **说明** 如果RDS上的“local_infile”参数组不可编辑，则说明是默认参数组，需要先创建一个新的参数组，再修改该参数值，并应用到RDS的MySQL实例上，具体操作请参见《关系型数据库用户指南》。	是
使用Agent	Agent功能待下线，无需配置。	-
Agent	Agent功能待下线，无需配置。	-
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	不同类型的关系数据库，需要适配不同的驱动。	-
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。	-

版权所有 © 华为云计算技术有限公司

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

步骤6 再次选择“连接管理 > 新建连接”，新建MRS Hive连接。连接器类型选择“MRS Hive”，然后单击“下一步”配置连接参数，参数说明如表5-2所示。配置完成后，单击“保存”回到连接管理界面。

表 5-2 MRS Hive 连接参数

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

版权所有 © 华为云计算技术有限公司

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

版权所有 © 华为云计算技术有限公司

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

----结束

创建样例作业

步骤1 单击CDM集群“操作”列的“作业管理”，进入作业管理界面。

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

图 5-1 新建作业

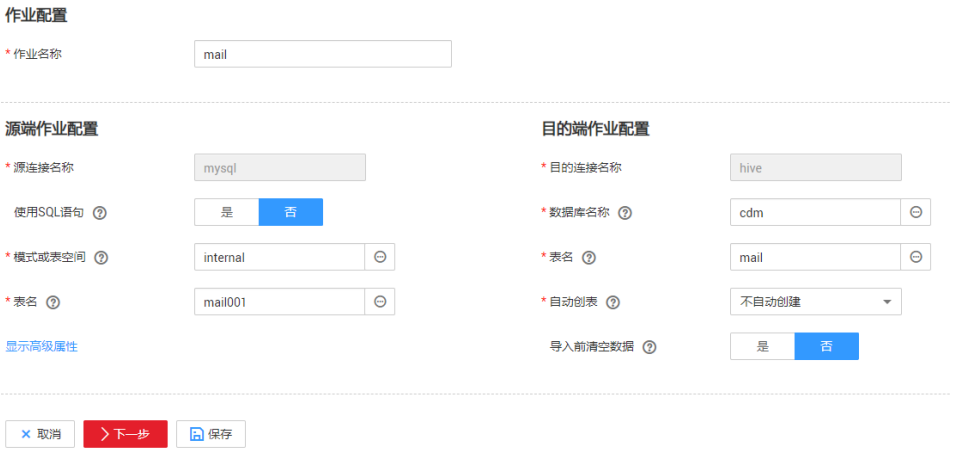


图 5-2 配置基本信息

 版权所有 © 华为云计算技术有限公司

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

图 5-3 查看作业 JSON

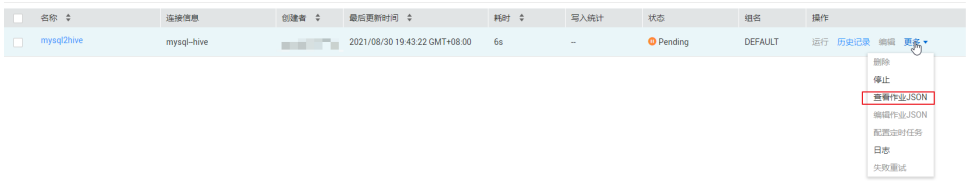
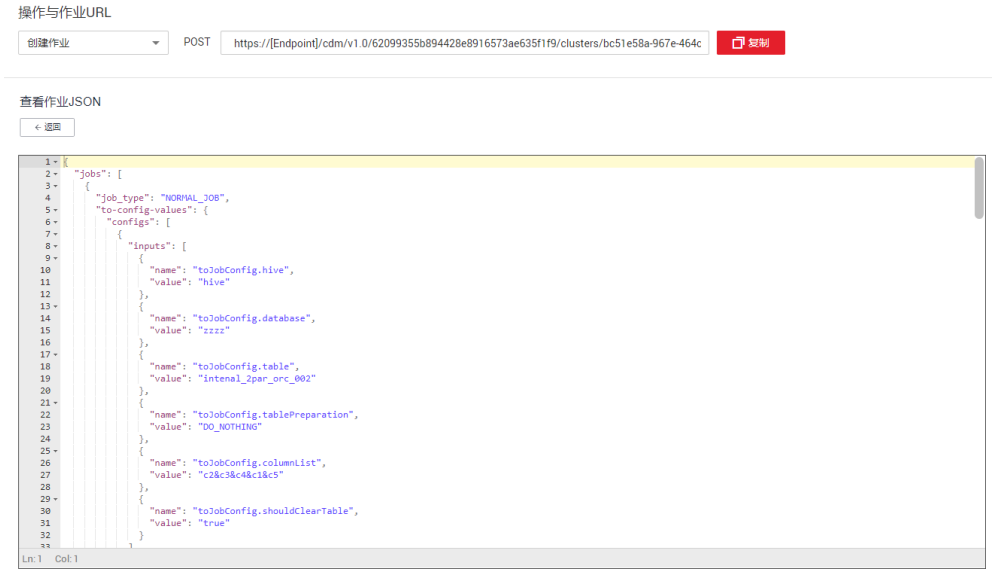


图 5-4 复制作业参数



----结束

创建数据开发作业

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

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

版权所有 © 华为云计算技术有限公司

图 5-5 配置 CDM 作业消息体

步骤3 编辑“CDM作业消息体”。

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

   图 5-6 编辑 JSON

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

   说明

   如果需要创建分库的作业，也可将作业JSON中的源连接修改为变量，方便替换。

版权所有 © 华为云计算技术有限公司

图 5-7 编辑 JSON

步骤4 添加作业参数num，用于作业JSON中调用。如下图所示：

图 5-8 添加作业参数 num

添加完成后单击“保存并提交版本”，以保存子作业。

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

关键配置如下：

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

说明

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

如下图所示：

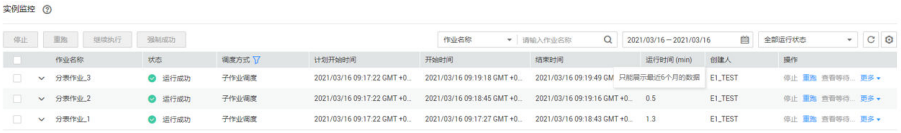
 版权所有 © 华为云计算技术有限公司

图 5-9 配置关键参数

配置完成后单击“保存并提交版本”，以保存主作业。

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

图 5-10 查看作业创建情况



----结束

注意事项

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

版权所有 © 华为云计算技术有限公司

图 5-11 修改属性

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

版权所有 © 华为云计算技术有限公司

6 贸易数据极简上云与统计分析

6.1 贸易数据上云场景介绍

使用云数据迁移（Cloud Data Migration，简称CDM）将本地贸易统计数据导入到OBS，再使用数据湖探索（Data Lake Insight，简称DLI）进行贸易统计分析，帮助H咨询公司以极简、极低成本构建其大数据分析平台，使得该公司更好地聚焦业务，持续创新。

场景描述

H公司是国内一家收集主要贸易国贸易统计及买家数据的商业机构，拥有大量的贸易统计数据库，其数据广泛应用于产业研究、行业研究、国际贸易促进等方面。

在这之前，H公司采用其自建的大数据集群，并安排专人维护，每年固定购买电信联通双线专用带宽，在机房、电力、专网、服务器、运维方面进行高额投入，但其在面对客户不断变化的业务诉求时，因为人员投入不足，大数据集群能力不匹配，而无法聚焦业务创新，使得存量100T的数据只有4%的利用率。

在将本地的贸易统计数据迁移到华为云之后，基于华为公有云的大数据分析能力，可帮助H公司屏蔽大数据基础设施复杂的构建、维护过程，使其客户人员可以全身心聚焦业务创新，盘活100T的存量数据，使资产最大化变现。

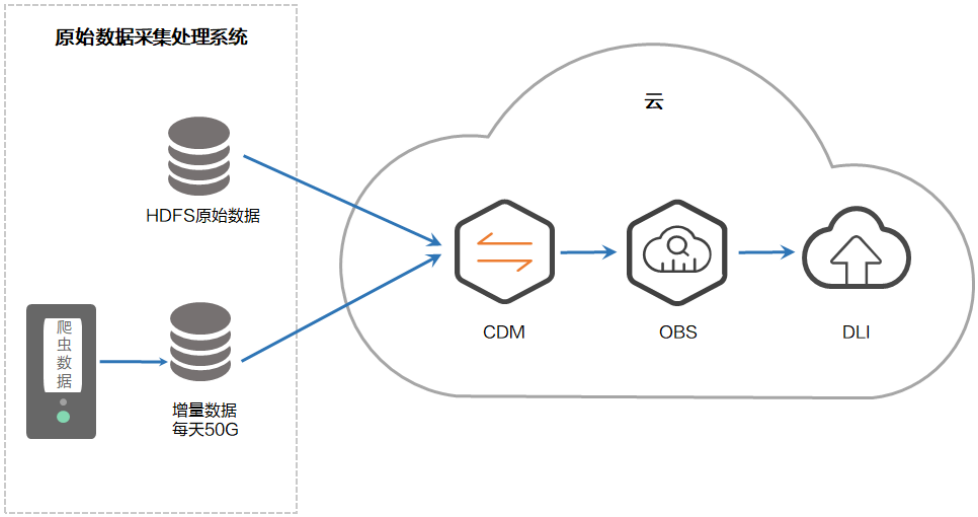
CDM和DLI服务按需收费，帮助H公司客户释放了维护人员并降低了专用带宽成本，使得维护成本相比线下数据中心降低了70%，且使用门槛低，可实现已有数据的平滑迁移，使新业务上线周期相比之前缩短了50%。

场景任务

根据客户原始数据采集处理系统中已有的H公司的数据（例如：贸易详单数据和基础信息数据），基于CDM+OBS+DLI完成贸易统计分析。

版权所有 © 华为云计算技术有限公司

图 6-1 场景方案



说明

DLI创建OBS外表，对OBS表数据存储格式有所要求：

- 使用DataSource语法创建OBS表时，支持orc，parquet，json，csv，carbon，avro类型。
- 使用Hive语法创建OBS表时，支持TEXTFILE, AVRO, ORC, SEQUENCEFILE, RCFILE, PARQUET, CARBON类型。

如果原始数据表存储格式不满足要求，您可以通过CDM将原始数据直接导入到DLI中进行分析，无需上传OBS。

数据说明

- 贸易详单数据

  包括主要贸易国货物贸易统计数据。

表 6-1 贸易详单数据

字段名称	字段类型	字段说明
hs_code	string	进出口商品编码列表
country	smallint	国家基础信息
dollar_value	double	交易金额
quantity	double	交易量
unit	smallint	计量单位
b_country	smallint	目标国家基础信息
imex	smallint	进出口类型
y_year	smallint	年
m_month	smallint	月

版权所有 © 华为云计算技术有限公司

- 基础信息数据
  贸易详单数据中维度字段对应的相关字典数据信息。

  表 6-2 国家基础信息表（country）

字段名称	字段类型	字段说明
countryid	smallint	国家编码
country_en	string	国家英文名称
country_cn	string	国家中文名称

  表 6-3 更新时间信息表（updatetime）

字段名称	字段类型	字段说明
countryid	smallint	国家编码
imex	smallint	进出口类型
hs_len	smallint	商品编码长度
minstartdate	string	最小开始时间
startdate	string	开始时间
newdate	string	更新时间
minnewdate	string	最小更新时间

  表 6-4 进出口商品编码信息表（hs246）

字段名称	字段类型	字段说明
id	bigint	编号
hs	string	商品编码
hs_cn	string	商品中文名称
hs_en	string	商品英文名称

  表 6-5 单位信息表（unit_general）

字段名称	字段类型	字段说明
id	smallint	计量单位编码
unit_en	string	计量单位英文名称
unit_cn	string	计量单位中文名称

版权所有 © 华为云计算技术有限公司

6.2 操作流程概述

流程介绍

使用CDM+OBS+DLI进行贸易统计分析的操作过程主要包括2个步骤：

1. **使用CDM上传数据到OBS**
   a. 通过CDM将H公司存量数据上传到对象存储服务OBS。
   b. 通过CDM作业的定时任务，每天自动上传增量数据到OBS。
2. **使用DLI分析数据**
   通过DLI直接分析OBS中的业务数据，支撑H公司客户进行贸易统计分析。

6.3 使用 CDM 上传数据到 OBS

6.3.1 上传存量数据

1. 使用**华为云专线**，搭建用户本地数据中心与华为云VPC之间的专属连接通道。
2. 创建OBS桶，并记录OBS的访问域名、端口和AK/SK。
3. 创建CDM集群。

   说明

   DataArts Studio实例中已经包含一个CDM集群（试用版除外），如果该集群已经满足需求，您无需再购买数据集成增量包，可以跳过这部分内容。

   如果您需要再创建新的CDM集群，请参考**购买批量数据迁移增量包**，完成购买数据集成增量包的操作。

   - 实例类型：选择“cdm.xlarge”，该实例类型适用大部分迁移场景。
   - 虚拟私有云：CDM集群的VPC，选择用户本地数据中心与云专线连通的VPC。
   - 子网、安全组：这里没有要求，分别任选一个即可。
4. 集群创建完成后，选择集群后面的“作业管理 > 连接管理 > 新建连接”，进入选择连接类型的界面，如**图6-2**所示。

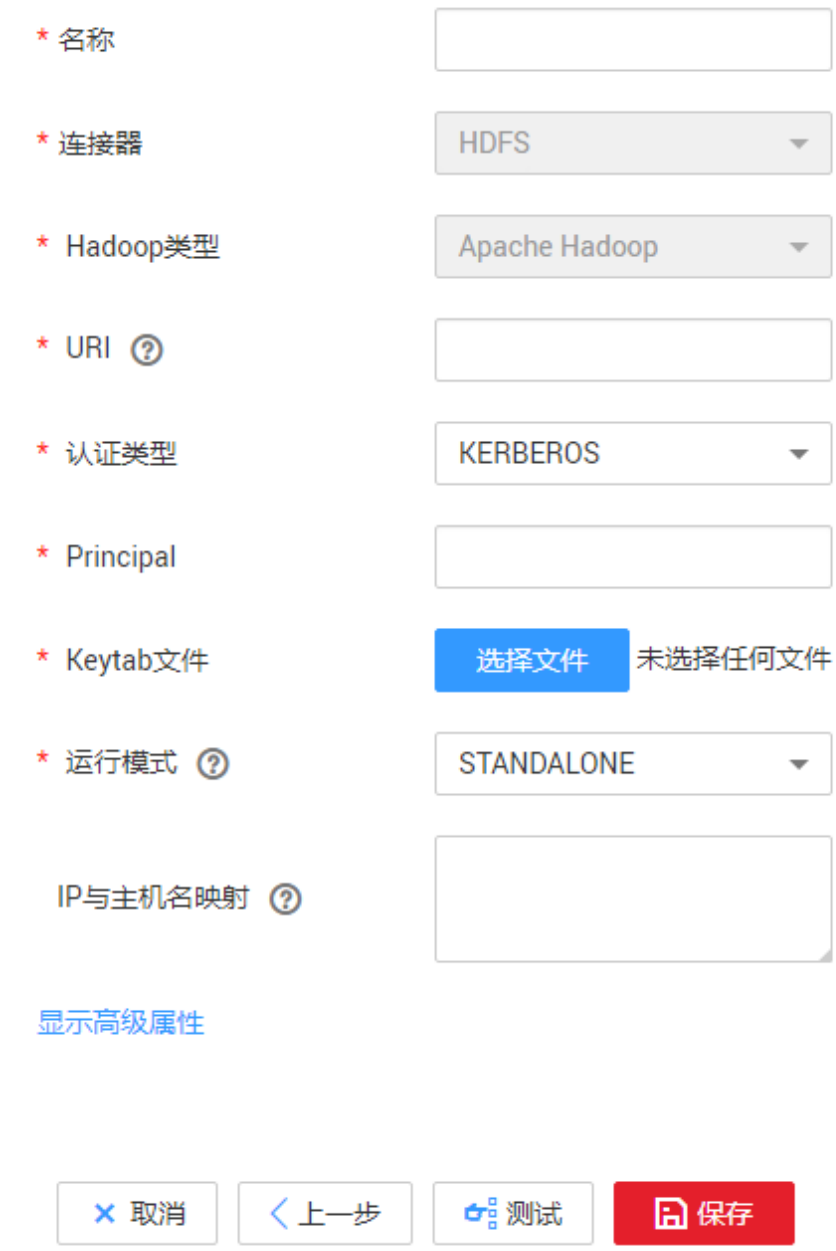
版权所有 © 华为云计算技术有限公司

图 6-2 选择连接器类型

5. 连接H公司本地的Apache Hadoop HDFS文件系统时，连接类型选择“Apache HDFS”，然后单击“下一步”。

 版权所有 © 华为云计算技术有限公司

图 6-3 创建 HDFS 连接



说明

- 名称：用户自定义连接名称，例如“hdfs_link”。
- URI：配置为H公司HDFS文件系统的Namenode URI地址。
- 认证类型：安全模式Hadoop选择KERBEROS鉴权，通过获取客户端的principal和keytab文件进行认证。
- Principal、Keytab文件：用于认证的账号Principal和keytab文件，可以联系Hadoop管理员获取。

6. 单击“保存”，CDM会自动测试连接是否可用。
   - 如果可用则提示保存成功，系统自动跳转到连接管理界面。

 版权所有 © 华为云计算技术有限公司

- 如果测试不可用，需要重新检查连接参数是否配置正确，或者H公司防火墙是否允许CDM集群的EIP访问数据源。

7. 单击“新建连接”来创建OBS连接，连接类型选择“对象存储服务（OBS）”后单击“下一步”，配置OBS连接参数，如图6-4所示。

图 6-4 创建 OBS 连接

说明

- 名称：用户自定义连接名称，例如“obslink”。
- OBS终端节点：配置为OBS的域名或IP地址，例如“obs.myhuaweicloud.com”。
- 端口：OBS服务器的端口，例如“443”。
- OBS桶类型：根据实际情况下拉选择即可。
- 访问标识（AK）、密钥（SK）：访问OBS数据库的AK、SK。可在管理控制台单击用户名，选择“我的凭证 > 访问密钥”后获取。

8. 单击“保存”，系统回到连接管理界面。
9. 选择“表/文件迁移 > 新建作业”，创建迁移H公司贸易数据到OBS的作业， 如图6-5所示。

 版权所有 © 华为云计算技术有限公司

图 6-5 创建作业

说明

- 作业名称：用户自定义作业名称。
- 源端作业配置：
  - 源连接名称：选择5创建的HDFS连接“hdfs_link”。
  - 源目录或文件：配置为H公司贸易数据在本地的存储路径，可以是一个目录，也可以是单独一个文件。这里配置为目录，CDM会迁移整个目录下的文件到OBS。
  - 文件格式：选择“二进制格式”。这里的文件格式是指CDM传输数据时所用的格式，不会改变原始文件自身的格式。迁移文件到文件时，推荐使用“二进制格式”，传输的效率和性能都最优。
- 目的端作业配置：
  - 目的连接名称：选择7创建的OBS连接“obslink”。
  - 桶名、写入目录：在OBS中储存贸易数据的路径，CDM会将文件写入到该路径下。
  - 文件格式：与源端一样，选择“二进制格式”，原始文件自身的格式不会改变。
  - 重复文件处理方式：这里选择“跳过重复文件”。只有当源端和目的端存在文件名、文件大小都相同的文件时，CDM才会判定该文件为重复文件，这时CDM将跳过该文件，不迁移到OBS。

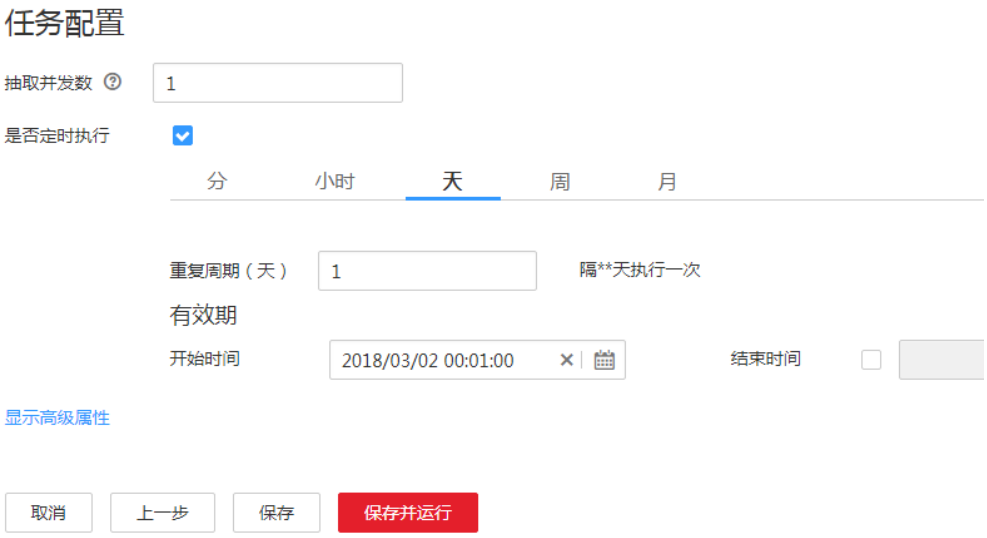
10. 单击“下一步”配置任务参数，迁移存量数据时，参数配置保持默认即可。
11. 单击“保存并运行”，进入作业管理界面，查看作业执行进度和结果。
12. 作业执行成功之后，单击作业后面的“历史记录”查看作业的写入行数、读取行数、写入字节、写入文件数和执行日志。

6.3.2 上传增量数据

1. 使用CDM将存量数据上传完后，单击该作业后的“编辑”，直接修改该作业。
2. 保持作业基本参数不变，单击“下一步”修改任务参数，如图6-6所示。

版权所有 © 华为云计算技术有限公司

图 6-6 定时任务配置



3. 勾选“是否定时执行”，配置定时任务：
   - “重复周期”配置为1天。
   - “开始时间”配置为每天凌晨0点1分执行。

   这样CDM每天凌晨自动执行全量迁移，但因为“重复文件处理方式”选择了“跳过重复文件”，相同名称且相同大小的文件不迁移，所以只会上传每天新增的文件。
4. 单击“保存”，完成CDM的增量同步配置。

6.4 分析数据

通过DLI直接对OBS数据进行贸易统计分析。

前提条件

DLI创建OBS外表，对OBS表数据存储格式有所要求：

- 使用DataSource语法创建OBS表时，支持orc，parquet，json，csv，carbon，avro类型。
- 使用Hive语法创建OBS表时，支持TEXTFILE, AVRO, ORC, SEQUENCEFILE, RCFILE, PARQUET, CARBON类型。

如果原始数据表存储格式不满足要求，您可以通过CDM将原始数据直接导入到DLI中进行分析，无需上传OBS。

通过 DLI 分析数据

1. 进入数据湖探索DLI控制台，参考DLI用户指南中的**创建数据库**创建数据库。
2. 参考**创建OBS表**创建OBS外表，包括贸易统计数据库、贸易详单信息表和基础信息表。
3. 基于业务需求，在DLI控制台中开发相应的SQL脚本进行贸易统计分析。

版权所有 © 华为云计算技术有限公司

7 车联网大数据零丢失搬迁入湖

7.1 车联网大数搬迁入湖简介场景介绍

场景描述

为搭建H公司车联网业务集团级的云管理平台，统一管理、部署硬件资源和通用类软件资源，实现IT应用全面服务化、云化，CDM（Cloud Data Migration，简称CDM）助力H公司做到代码“0”改动、数据“0”丢失迁移上云。

约束限制

本方案仅支持迁移数据到MRS 1.x版本集群，2.x及之后版本无法通过HBase修复命令重建HBase表。

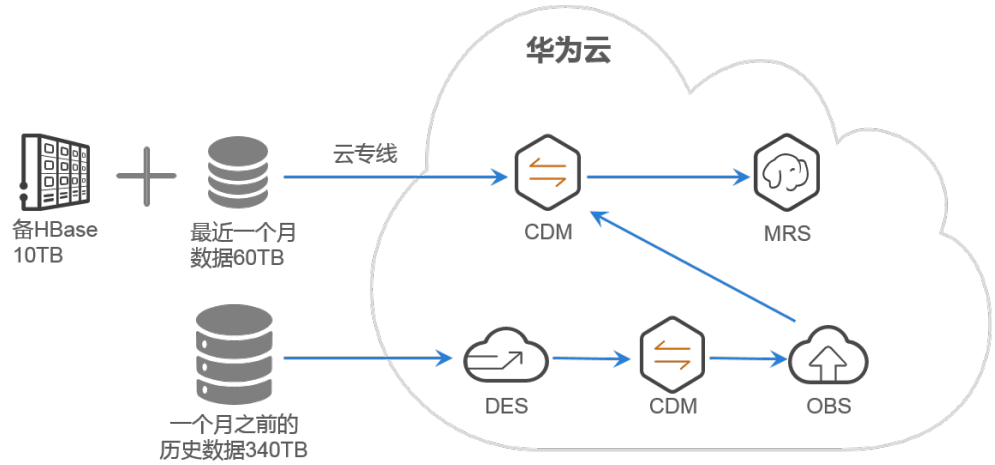
须知

迁移的目标集群为2.x及之后版本时，HBase修复命令已经不再支持，通过HBase数据目录迁移的方式无法使用。

版权所有 © 华为云计算技术有限公司

迁移方案

图 7-1 迁移方案



H公司的车联网大数据业务平台当前CDH（Cloudera Hadoop） HBase集群中共有854张表约400TB，备HBase集群中共有149张表，约10TB数据。最近一个月新增的数据量是60TB。

使用CDM将CDH集群中的HBase HFile抽取出来存入到MRS（MapReduce） HDFS中，再通过HBase修复命令重建HBase表。基于这种迁移方案，可以使用以下2种迁移方式同时进行：

1. CDM通过专线直接迁移近一个月的数据以及备HBase集群的数据：
   CDH → CDM（华为云） → MRS

   说明

   使用云专线直接迁移时的优缺点：
   - 优点：数据无需做多次的搬迁，缩短整体搬迁周期。
   - 缺点：在数据大量传输过程中会占用专线带宽，对客户并行进行的业务存在影响，跨越多个交换机设备。
2. CDM通过**DES（数据快递服务）**迁移1个月前的历史数据，迁移路径如下：
   CDH → DES → CDM（华为云） → OBS → CDM（华为云） → MRS

   说明

   DES适用场景：数据量大，用户私有云与华为云无专线打通，用户私有云网络到公网带宽有限。
   - 优点：传输可靠性高，受专线以及网络质量影响较小。
   - 缺点：迁移方式耗时较长。

7.2 迁移准备

前提条件

- CDH HBase的版本号小于或等于MRS HBase的版本号。

版权所有 © 华为云计算技术有限公司

- 待迁移的表在迁移过程中不能有写入，Split，Merge等操作。
- 使用**华为云专线**搭建CDH集群与华为云VPC之间的专属连接通道。

迁移流程

1. 预估迁移数据量、迁移时间。
2. 输出详细待迁移数据表、文件个数、大小，用于后续校验。
3. 分批配置迁移任务，保证迁移进度与速度。
4. 校验文件个数以及文件大小。
5. 在MRS中恢复HBase表并验证。

准备数据

项目	数据项	说明	取值示例
DES盒子	挂载地址	DES盒子在客户的虚拟机挂载的地址。	//*虚拟机IP*/huawei
	存储管理系统	DES盒子的存储管理系统，与管理IP相关。	https://*管理IP*:8088/deviceManager/devicemanager/login/login.html
	用户名	登录存储管理系统的用户名。	admin
	密码	登录密码。	-
CDH集群	NameNode IP	客户CDH集群的主NameNode IP。	192.168.2.3
	HDFS的端口	一般默认为9000。	9000
	HDFS URI	客户CDH集群中HDFS的NameNode URI地址。	hdfs://192.168.2.3:9000
OBS	OBS终端节点	OBS的Endpoint。	obs.ap-southeast-1.myhuaweicloud.com
	OBS桶	存放CDH一个月前历史数据的OBS桶。	cdm
	AK/SK	连接OBS的AK/SK。	-
MRS	Manager IP	MRS Manager的IP地址。	192.168.3.11

7.3 CDM 迁移近一个月的数据

备HBase集群中约10TB数据，最近一个月新增的数据量约60TB，总共约70TB。H公司安装的云专线为20GE端口，支持CDM超大规格的集群（cdm.xlarge），综合考虑迁移

版权所有 © 华为云计算技术有限公司

时间、成本、性能等，这里使用2个CDM超大规格集群并行迁移。CDM集群规格如表7-1所示。

表 7-1 CDM 集群规格

实例类型	核数/内存	最大带宽/基准带宽	并发作业数	适用场景
cdm.large	8核/16G	3/0.8 Gbps	16	单表规模≥1000万条。
cdm.xlarge	16核/32G	10/4 Gbps	32	适合10GE高速带宽进行TB以上的数据量迁移。
cdm.4xlarge	64核/128G	40/36 Gbit/s	64	-

说明

其他场景中，可根据情况选择多个CDM集群同时迁移，加快迁移效率。MRS HDFS多副本策略会占用网络带宽，影响迁移速率。

华为云 CDM 创建连接

1. 创建2个CDM集群：

   说明

   DataArts Studio实例中已经包含一个CDM集群（试用版除外），如果该集群已经满足需求，您无需再购买数据集成增量包，可以跳过这部分内容。

   如果您需要再创建新的CDM集群，请参考**购买批量数据迁移增量包**章节，完成购买数据集成增量包的操作。

   - 集群规格选择“cdm.xlarge”。
   - 集群所属的VPC与MRS所属的VPC一致，同时也要与云专线连通的VPC的一致。
   - 其它参数可以自定义，或者保持默认。

2. 创建CDH HDFS连接：
   a. 单击CDM集群操作列的“作业管理”，进入作业管理界面。
   b. 选择“连接管理 > 新建连接”，进入连接器类型的选择界面，选择“Apache HDFS”。

版权所有 © 华为云计算技术有限公司

图 7-2 选择连接器类型

c. 单击“下一步”，配置连接参数，依次填写相关信息。URI格式为“hdfs://NameNode IP:端口”，若CDH没有启动Kerberos认证则“认证类型”选择“SIMPLE”。

版权所有 © 华为云计算技术有限公司

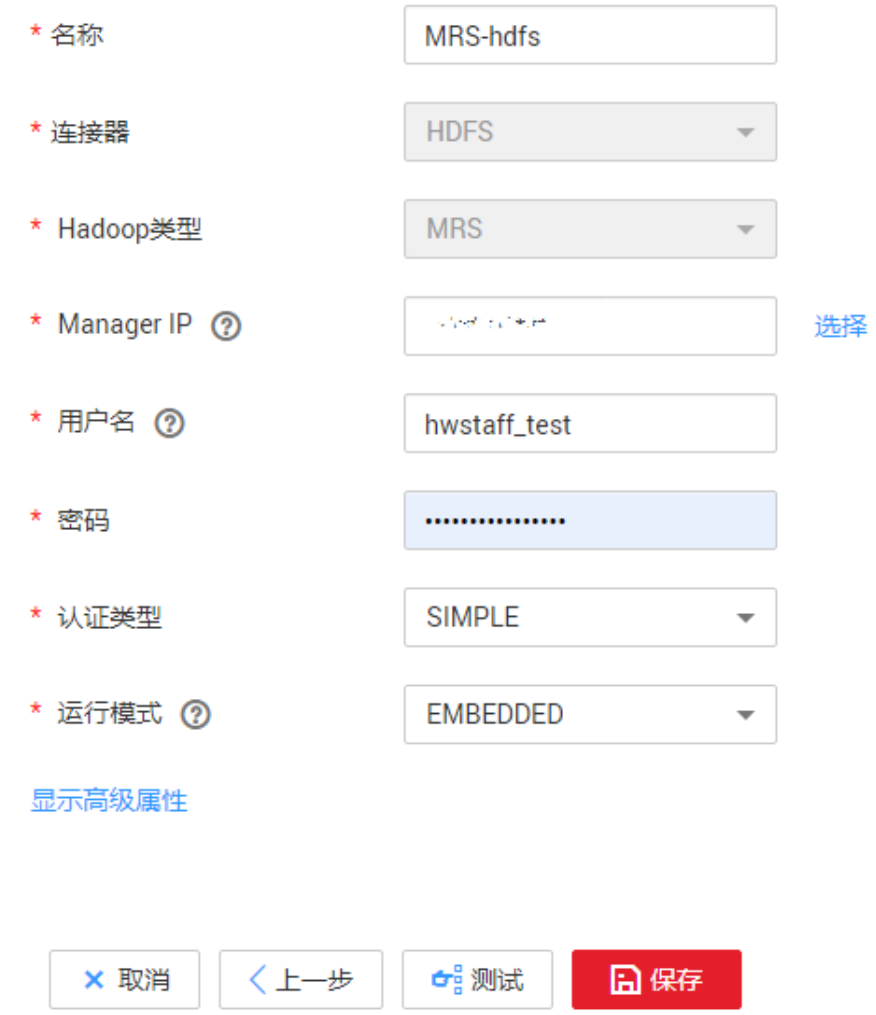
d. 单击“测试”，如果右上角显示“测试成功”，表示连接成功，单击“保存”。

3. 创建MRS HDFS连接：

a. 在作业管理界面，选择“连接管理 > 新建连接”，进入连接器类型的选择界面，选择“MRS HDFS”。

 版权所有 © 华为云计算技术有限公司

b. 单击“下一步”，配置连接参数，依次填写相关信息。“认证类型”选择“SIMPLE”，运行模式保持默认即可。



 版权所有 © 华为云计算技术有限公司

c. 单击“测试”，如果右上角显示“测试成功”，表示连接成功，单击“保存”。

华为云 CDM 创建迁移作业

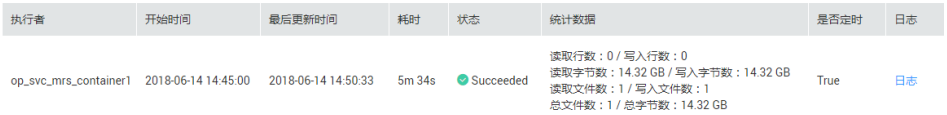
1. 在CDM集群的作业管理界面，选择“表/文件迁移 > 新建作业”，每个表文件的目录作为一个迁移作业。

- 源端作业配置
  - 源连接名称：选择上面创建的**CDH HDFS连接**。
  - 源目录或文件：选择CDH中HBase的HBase表所在目录。例如“/hbase/data/default/table_20180815”，表示迁移“table_20180815”这个目录下所有文件。
  - 文件格式：文件的复制要选择“二进制格式”。
- 目的端作业配置
  - 目的连接名称：选择上面创建的**MRS HDFS连接**。
  - 写入目录：选择MRS HBase的目录，例如“/hbase/data/default/table_20180815/”。这个目录必须带有表名（例如这里的表名是table_20180815），如果该目录不存在，CDM会自动创建该目录。
  - 文件格式：同源端相同，选择“二进制格式”。
- 其它可选参数保持默认即可。

2. 单击“下一步”进行任务配置，其中抽取并发数默认为3，适当增加可以增加迁移速率，本例中设置为8，其它参数保持默认即可。

 版权所有 © 华为云计算技术有限公司

3. 重复上述步骤创建其它迁移目录的作业，参数配置都相同。2个CDM集群的作业个数平均分配，并发执行。
4. 作业执行完成后，可在作业的“历史记录”中查看详细的数据统计。



执行者	开始时间	最后更新时间	耗时	状态	统计数据	是否定时	日志
op_svc_mrs_container1	2018-06-14 14:45:00	2018-06-14 14:50:33	5m 34s	Succeeded	读取行数：0 / 写入行数：0 读取字节数：14.32 GB / 写入字节数：14.32 GB 读取文件数：1 / 写入文件数：1 总文件数：1 / 总字节数：14.32 GB	True	日志

7.4 DES 迁移一个月前的历史数据

迁移流程

1. 通过脚本将一个月前的历史数据导入到DES盒子。DES盒子的相关操作请参见**数据快递服务 DES**。
2. DES将数据快递到华为云数据中心。
3. 使用华为云CDM将DES中的数据迁移到华为云OBS。
4. 使用华为云CDM将OBS数据迁移到MRS。

其中CDM相关操作，与**CDM迁移近一个月的数据**相同，都是使用二进制直接传输文件目录，2个集群并发执行作业。

注意事项

- 当迁移动作影响到客户的HDFS集群时，需要手动停止作业。

版权所有 © 华为云计算技术有限公司

- 如果作业出现大批量的失败：
  a. 先检查DES盒子是否被写满。如果写满，需要清除最近写入的目录，保证后面写入的数据都是完整的。
  b. 再检查网络是否连通。
  c. 检查客户的HDFS集群。检查是否有指标异常的现象，如果有，则需要暂停迁移任务。

7.5 MRS 中恢复 HBase 表

CDH HBase表目录已经迁移到MRS HBase后，可以使用命令恢复。对于那些会变化的数据，需要使用快照保证数据不变，然后再迁移并恢复。

约束限制

本方案仅支持迁移数据到MRS 1.x版本集群，2.x及之后版本无法通过HBase修复命令重建HBase表。

须知

迁移的目标集群为2.x及之后版本时，HBase修复命令已经不再支持，通过HBase数据目录迁移的方式无法使用。

使用命令恢复历史不变的数据

这里以恢复“/hbase/data/default/table_20180811”表为例，恢复步骤如下：

1. 进入MRS Client所在的节点，例如master1节点。
2. 切换为omm用户。
   su – omm
3. 加载环境变量。
   source /opt/client/bigdata_env
4. 执行修改目录权限命令。
   hdfs dfs –chown *omm:hadoop* **–R** */hbase/data/default/table_20180811*
   - omm:hadoop：表示用户名，实际场景中请替换。
   - /hbase/data/default/table_20180811：表示表所在路径。
5. 执行恢复元数据命令。
   hbase hbck –fixMeta *table_20180811*
6. 执行Region上线命令。
   hbase hbck –fixAssignments *table_20180811*
7. 出现“Status: OK”则说明恢复表成功。

使用快照迁移并恢复会变的数据

1. 在源端CDH集群HBase shell中执行：
   flush <*table name*>

 版权所有 © 华为云计算技术有限公司

2. 在源端CDH集群HBase shell执行：

   compact *<table name>*

3. 如果表没有打开Snap功能，则执行：

   hadoop dfsadmin -allowSnapshot $path

4. 创建HDFS Snapshot，例如命名s0：

   hdfs dfs -createSnapshot *<snapshotDir>* [*s0*]

   hdfs dfs -createSnapshot test

5. CDM通过HDFS Snapshot复制文件至MRS。CDM的作业配置：
   - “源目录或文件”输入：/hbase/data/default/src_test/.snapshot/s0
   - 目的端“写入目录”输入：/hbase/data/default/*表名*

6. 执行fixMeta和fixAssignments等命令恢复表，参考**使用命令恢复历史不变的数据**。

7. 在CDH集群中删除快照：

   hdfs dfs -deleteSnapshot *<snapshotDir> s0*

恢复表时的问题处理

1. 执行完fixMeta命令后，报错显示“xx inconsistent”：

   fixMeta命令是校验HDFS和HBase元数据一致性，出现这个提示是正常情况，继续执行fixAssignments命令即可。

2. 执行完fixAssignments命令后，报错显示“xx inconsistent”：

   fixAssignments是让所有Region上线，偶尔会出现部分Region上线较慢，可以再执行一次以下命令检查一下：

   hbase hbck *表名*

   如果出现“Status : OK”则表示HBase表恢复成功。

3. 执行完fixAssignments命令后，错误提示多个region有相同的startkey，部分region存在overlap情况：

   可以再执行：

   hbase hbck –fixHdfsOverlaps *表名*

   执行完毕后再执行fixMeta和fixAssignments命令。

 版权所有 © 华为云计算技术有限公司