数据湖探索

最佳实践

文档版本 01

发布日期 2024-09-19

华为技术有限公司

商标声明


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

注意

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

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

安全声明

漏洞处理流程

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

目 录

1 最佳实践内容概览

本指导从数据迁移、数据分析提供了完整的端到端最佳实践内容，帮助您更好的使用DLI进行大数据分析和处理。

数据迁移

您可以通过**云数据迁移服务**CDM轻松的将其他云服务或者业务平台的数据迁移至DLI。包括以下最佳实践内容：

- 迁移Hive数据至DLI，具体请参考**迁移Hive数据至DLI**。
- 迁移Kafka数据至DLI，具体请参考**迁移Kafka数据至DLI**。
- 迁移Elasticsearch数据至DLI，具体请参考**迁移Elasticsearch数据至DLI**。
- 迁移RDS数据至DLI，具体请参考**迁移RDS数据至DLI**。
- 迁移DWS数据至DLI，具体请参考**迁移DWS数据至DLI**。

数据分析

DLI应用于海量的日志数据分析和大数据ETL处理，助力各行业使能数据价值。当前数据分析最佳实践内容如下：

- 使用DLI进行车联网场景驾驶行为数据分析，具体请参考**使用DLI进行车联网场景驾驶行为数据分析**。
- 使用DLI将CSV数据转换为Parquet数据，具体请参考**使用DLI将CSV数据转换为Parquet数据**。
- 使用DLI进行电商BI报表分析，具体请参考**使用DLI进行电商BI报表分析**。
- 使用DLI进行账单分析与优化，具体请参考**使用DLI进行账单分析与优化**。

2 数据迁移

2.1 数据迁移概览

本文为您介绍数据迁移的最佳实践，您可以通过**云数据迁移服务**CDM轻松的将其他云服务或者业务平台的数据迁移至DLI。

DLI提供一站式的流处理、批处理、交互式分析的Serverless融合处理分析服务，采用批流融合高扩展性框架，为TB~EB级数据提供了更实时高效的多样性算力，可支撑更丰富的大数据处理需求。

数据迁移最佳实践

- 迁移Hive数据至DLI，具体请参考**迁移Hive数据至DLI**。
- 迁移Kafka数据至DLI，具体请参考**迁移Kafka数据至DLI**。
- 迁移Elasticsearch数据至DLI，具体请参考**迁移Elasticsearch数据至DLI**。
- 迁移RDS数据至DLI，具体请参考**迁移RDS数据至DLI**。
- 迁移DWS数据至DLI，具体请参考**迁移DWS数据至DLI**。

数据迁移数据类型映射

将其他云服务或业务平台数据迁移到DLI ，或者将DLI数据迁移到其他云服务或业务平台时，涉及到源和目的端数据类型的转换和映射，根据**表2-1**可以获取到源和目的端的数据类型映射关系。

表 2-1 数据类型映射表

MySQL	Hive	DWS	Oracle	PostgreSQL	Hologres	DLI Spark
CHAR	CHAR	CHAR	CHAR	CHAR	CHAR	CHAR
VARCHAR	VARCHAR	VARCHAR	VARCHAR	VARCHAR	VARCHAR	VARCHAR/ STRING

MySQL	Hive	DWS	Oracle	PostgreSQL	Hologres	DLI Spark
DECIMAL	DECIMAL	NUMERIC	NUMERIC	NUMERIC	DECIMAL	DECIMAL
INT	INT	INTEGER	NUMBER	INTEGER	INTEGER	INT
BIGINT	BIGINT	BIGINT	NUMBER	BIGINT	BIGINT	BIGINT/LONG
TINYINT	TINYINT	SMALLINT	NUMBER	SMALLINT	SMALLINT	TINYINT
SMALLINT	SMALLINT	SMALLINT	NUMBER	SMALLINT	SMALLINT	SMALLINT/SHORT
BINARY	BINARY	BYTEA	RAW	BYTEA	BYTEA	BINARY
VARBINARY	BINARY	BYTEA	RAW	BYTEA	BYTEA	BINARY
FLOAT	FLOAT	FLOAT4	FLOAT	DOUBLE	FLOAT4	FLOAT
DOUBLE	DOUBLE	FLOAT8	FLOAT	REAL/DOUBLE	FLOAT8	DOUBLE
DATE	DATE	TIMESTAMP	DATE	DATE	DATE	DATE
TIME	不支持（推荐使用：String）	TIME	DATE	TIME	TIME	不支持（推荐使用：String）
DATETIME	TIMESTAMP	TIMESTAMP	TIME	TIME	TIMESTAMP	TIMESTAMP
TINYINT	TINYINT	BOOLEAN	不支持	TINYINT	BOOLEAN	BOOLEAN
不支持（推荐使用：TEXT）	不支持（推荐使用：String）	不支持（推荐使用：TEXT）	不支持（推荐使用：VARCHAR）	不支持（推荐使用：TEXT）	不支持（推荐使用：TEXT）	ARRAY
不支持（推荐使用：TEXT）	不支持（推荐使用：String）	不支持（推荐使用：TEXT）	不支持（推荐使用：VARCHAR）	不支持（推荐使用：TEXT）	不支持（推荐使用：TEXT）	MAP

MySQL	Hive	DWS	Oracle	PostgreSQL	Hologres	DLI Spark
不支持（推荐使用：TEXT）	不支持（推荐使用：String）	不支持（推荐使用：TEXT）	不支持（推荐使用：VARCHAR）	不支持（推荐使用：TEXT）	不支持（推荐使用：TEXT）	STRUCT

说明

推荐使用：表示当前服务没有支持的标准数据类型，可以使用推荐的数据类型来替换使用。

2.2 迁移 Hive 数据至 DLI

本文为您介绍如何通过CDM数据同步功能，迁移MRS Hive数据至DLI。其他MRS Hadoop组件数据，均可以通过CDM与DLI进行双向同步。

前提条件

- 已创建DLI的SQL队列。

  注意

  创建DLI队列时**队列类型**需要选择为“**SQL队列**”。

- 已创建包含Hive组件的MRS安全集群。
  - 本示例创建的MRS集群和各组件版本如下：
    - MRS集群版本：MRS 3.1.0
    - Hive版本：3.1.0
    - Hadoop版本：3.1.1
  - 本示例创建MRS集群时开启了Kerberos认证。
- 已创建CDM迁移集群。创建CDM集群的操作可以参考**创建CDM集群**。

说明

- 如果目标数据源为云下的数据库，则需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP、CDM云上安全组出方向放通云下数据源所在的主机、数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
- 数据源为云上的MRS、DWS等服务时，网络互通需满足如下条件：

  i. 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。

  ii. 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。

  iii. 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。

本示例CDM集群的虚拟私有云、子网以及安全组和MRS集群保持一致。

步骤一：数据准备

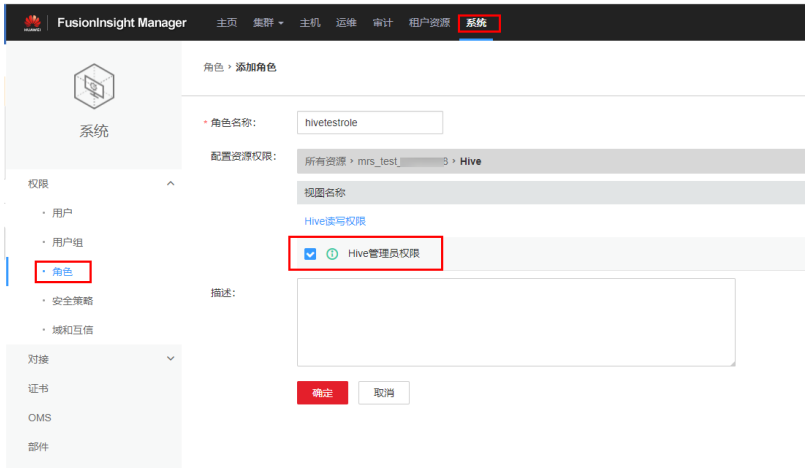
- MRS集群上创建Hive表和插入表数据。

  a. 参考**访问MRS Manager**登录MRS Manager。

  b. 在MRS Manager上，选择“系统 > 权限 > 角色”，单击“ 添加角色 ”，在添加角色页面分别配置参数。

  - 角色名称：输入自定义的“角色名称”，例如当前输入为：hivetestrole。
  - 配置资源权限：选择“*当前MRS集群的名称* > hive”，勾选“Hive管理员权限”。

  图 2-1 Manager 创建 Hive 的角色



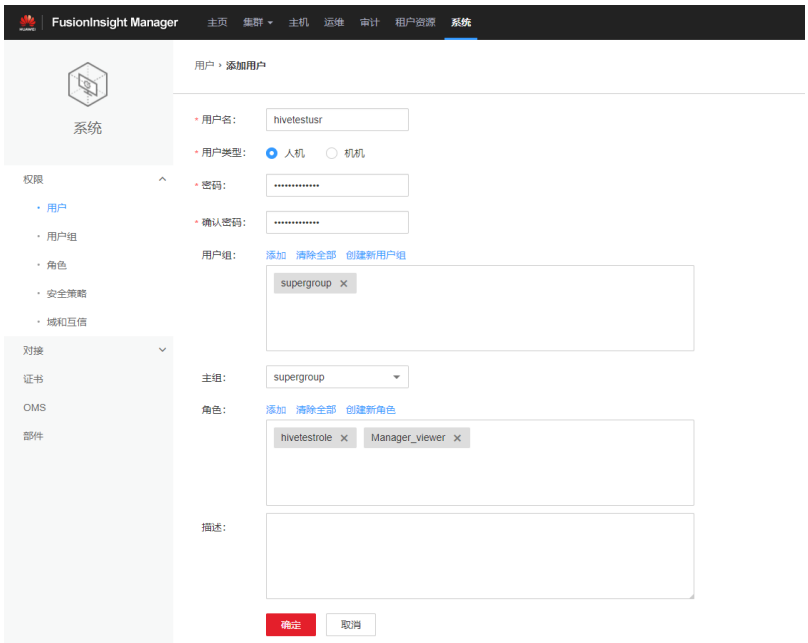
  更多MRS创建角色的操作说明可以参考：**创建Hive管理员角色**。

  c. 在MRS Manager上，选择“系统 > 权限 > 用户”，单击“添加用户”，在添加用户页面分别配置如下参数。

  i. 用户名：自定义的用户名。当前示例输入为：hivetestusr。

ii. 用户类型：当前选择为“人机”。
iii. 密码和确认密码：输入当前用户名对应的密码。
iv. 用户组和主组：选择supergroup
v. 角色：同时选择**b**中创建的角色和Manager_viewer角色。

图 2-2 MRS Manager 上创建 Hive 用户



d. 参考**安装MRS客户端**下载并安装Hive客户端。例如，当前Hive客户端安装在MRS主机节点的“/opt/hiveclient”目录上。
e. 以root用户进入客户端安装目录下。
例如：**cd /opt/hiveclient**
f. 执行以下命令配置环境变量。
source bigdata_env
g. 因为当前集群启用了Kerberos认证，则需要执行以下命令进行安全认证。认证用户为**c**中创建的用户。
kinit *c中创建的用户名*
例如，**kinit hivetestusr**
h. 执行以下命令连接Hive。
beeline
i. 创建表和插入表数据。
创建表：

```
create table user_info(id string,name string,gender string,age int,addr string);
```

插入表数据：

```
insert into table user_info(id,name,gender,age,addr) values("12005000201","A","男",19,"A城市");
insert into table user_info(id,name,gender,age,addr) values("12005000202","B","男",20,"B城市");
insert into table user_info(id,name,gender,age,addr) values("12005000202","B","男",20,"B城市");
```


说明

上述示例是通过创建表和插入表数据构造迁移示例数据。如果是迁移已有的Hive数据库和表数据，则可以通过以下命令获取Hive的数据库和表信息。

- 在Hive客户端执行如下命令获取数据库信息

  show databases
- 切换到需要迁移的Hive数据库

  use *Hive数据库名*
- 显示当前数据库下所有的表信息

  show tables
- 查询Hive表的建表语句

  show create table *Hive表名*

  查询出来的建表语句需要做一些处理，建表语句要符合DLI的建表语法，再到具体的DLI上执行。

- 在DLI上创建数据库和表。

  a. 登录DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列。

  在编辑器中输入以下语句创建数据库，例如当前创建迁移后的DLI数据库testdb。详细的DLI创建数据库的语法可以参考**创建DLI数据库**。

  ```
  create database testdb;
  ```

  b. 在数据库下创建表。

  说明

  如果是通过在MRS Hive中的“show create table *hive表名*”获取的建表语句，则需要修改该建表语句以符合DLI的建表语法。具体DLI的建表语法可以参考**创建DLI表**。

  ```
  create table user_info(id string,name string,gender string,age int,addr string);
  ```

步骤二：数据迁移

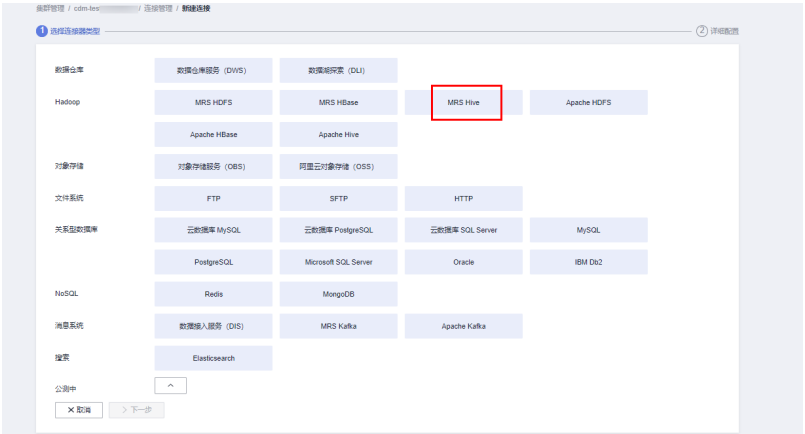
1. 配置CDM数据源连接。

   a. 配置源端MRS Hive的数据源连接。

      i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

      ii. 在作业管理界面，选择“连接管理”，单击“新建连接”，连接器类型选择“MRS Hive”，单击“下一步”。

图 2-3 创建 MRS Hive 数据源连接



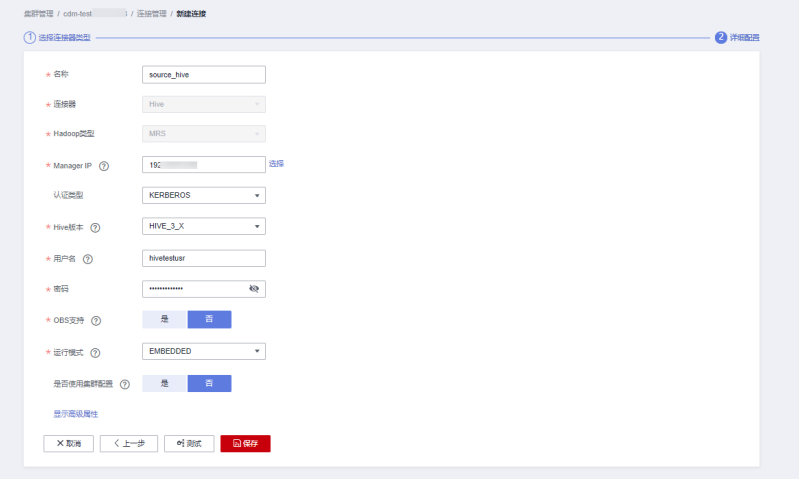
iii. 配置源端MRS Hive的数据源连接，具体参数配置如下。

表 2-2 MRS Hive 数据源配置

参数	值
名称	自定义MRS Hive数据源名称。例如当前配置为：source_hive
Manager IP	单击输入框旁边的“选择”按钮，选择当前MRS Hive集群即可自动关联出来Manager IP。
认证类型	如果当前MRS集群为普通集群则选择为SIMPLE，如果是MRS集群启用了Kerberos安全认证则选择为KERBEROS。 本示例选择为：KERBEROS。
Hive版本	根据当前创建MRS集群时候的Hive版本确定。当前Hive版本为3.1.0，则选择为：HIVE_3_X。
用户名	在c中创建的MRS Hive用户名。
密码	对应的MRS Hive用户名的密码。

其他参数保持默认即可。

图 2-4 CDM 配置 MRS Hive 数据源



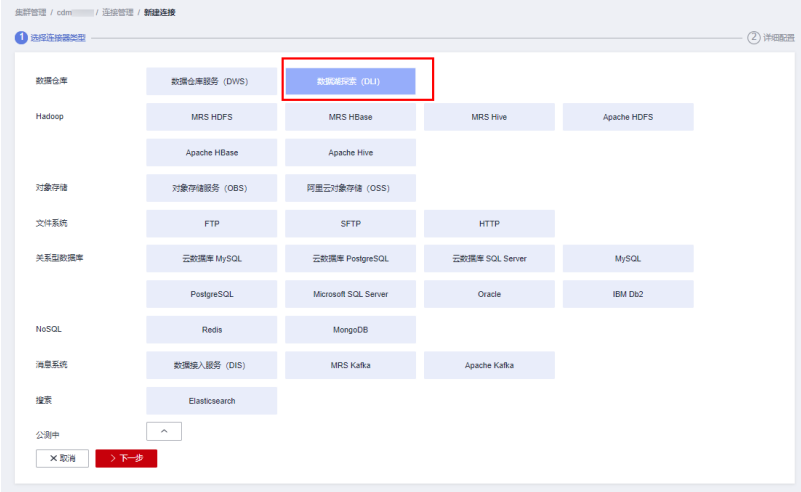
iv. 单击“保存”完成MRS Hive数据源配置。

b. 配置目的端DLI的数据源连接。

i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

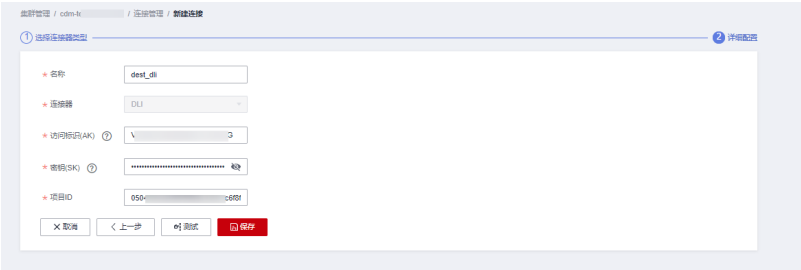
ii. 在作业管理界面，选择“连接管理”，单击“新建连接”，连接器类型选择“数据湖探索（DLI）”，单击“下一步”。

图 2-5 创建 DLI 数据源连接



iii. 配置目的端DLI数据源连接连接参数。

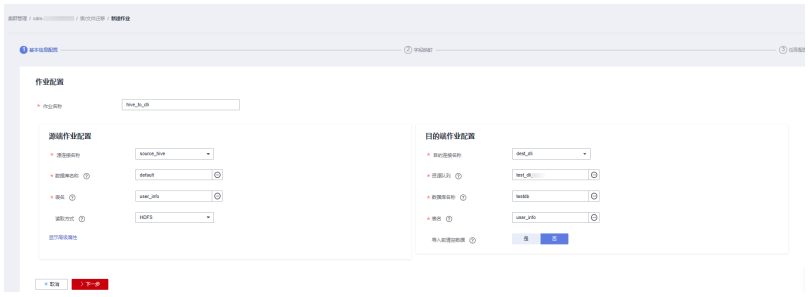
图 2-6 配置 DLI 数据源连接参数



配置完成后，单击“保存”完成DLI数据源配置。

2. 创建CDM迁移作业。
   a. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。
   b. 在“作业管理”界面，选择“表/文件迁移”，单击“新建作业”。
   c. 在新建作业界面，配置当前作业配置信息，具体参数参考如下：

图 2-7 新建 CDM 作业作业配置



i. 作业名称：自定义数据迁移的作业名称。例如，当前定义为：hive_to_dli。
ii. 源端作业配置，具体参考如下：

表 2-3 源端作业配置

参数名	参数值
源连接名称	选择**1.a**中已创建的数据源名称。
数据库名称	选择MRS Hive待迁移的数据库名称。例如当前待迁移的表数据数据库为“default”。
表名	待建议Hive数据表名。当前示例为在DLI上创建数据库和表中的“user_info”表。
读取方式	当前示例选择为：HDFS。具体参数含义如下： 包括HDFS和JDBC两种读取方式。默认为HDFS方式，如果没有使用WHERE条件做数据过滤及在字段映射页面添加新字段的需求，选择HDFS方式即可。 HDFS文件方式读取数据时，性能较好，但不支持使用WHERE条件做数据过滤及在字段映射页面添加新字段。 JDBC方式读取数据时，支持使用WHERE条件做数据过滤及在字段映射页面添加新字段。

更多参数的详细配置可以参考：**CDM配置Hive源端参数**。

iii. 目的端作业配置，具体参考如下：

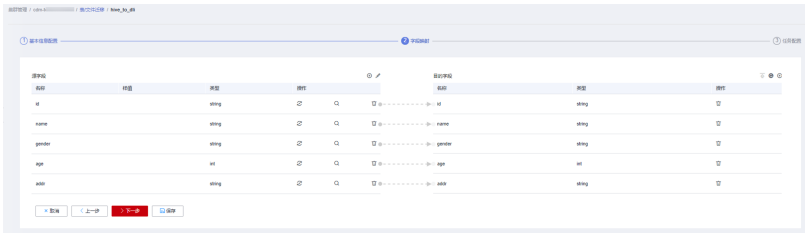
表 2-4 目的端作业配置

参数名	参数值
目的连接名称	选择**1.b**已创建的DLI数据源连接。
资源队列	选择已创建的DLI SQL类型的队列。
数据库名称	选择DLI下已创建的数据库。当前示例为**在DLI上创建数据库和表**中创建的数据库名，即为“testdb”。
表名	选择DLI下已创建的表名。当前示例为**在DLI上创建数据库和表**中创建的表名，即为“user_info”。
导入前清空数据	选择导入前是否清空目的表的数据。当前示例选择为“否”。 如果设置为是，任务启动前会清除目标表中数据。

更多参数的详细配置可以参考：**CDM配置DLI目的端参数**。

3. 单击“下一步”，进入到字段映射界面，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。
   - 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 2-8 字段映射



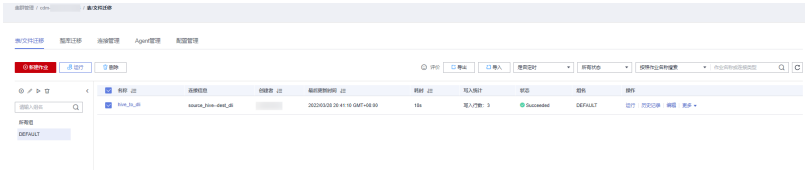
4. 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

   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，请参见**配置定时任务**。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。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，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。

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。

图 2-9 迁移作业进度和结果查询

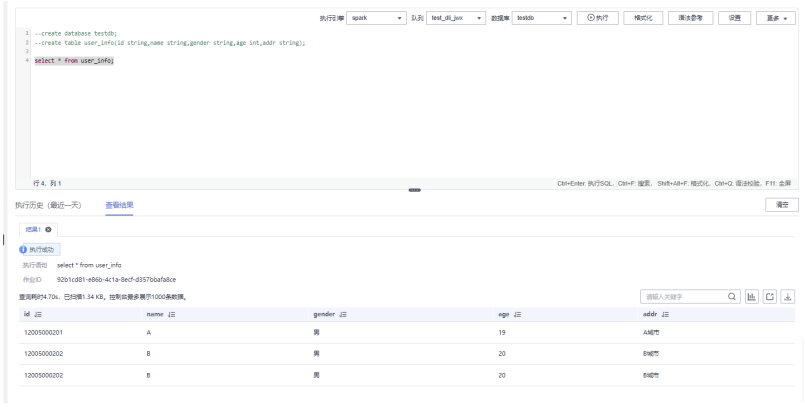


步骤三：结果查询

CDM迁移作业运行完成后，再登录到DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列，数据库选择已**a**已创建的数据库，执行DLI表查询语句，查询Hive表数据是否已成功迁移到DLI的“user_info”表中。

```
select * from user_info;
```

图 2-10 迁移后查询 DLI 的表数据



2.3 迁移 Kafka 数据至 DLI

本文为您介绍如何通过CDM数据同步功能，迁移MRS Kafka数据至DLI。

前提条件

- 已创建DLI的SQL队列。创建DLI队列的操作可以参考**创建DLI队列**。

  注意

  创建DLI队列时**队列类型**需要选择为“**SQL队列**”。

- 已创建包含Kafka组件的MRS安全集群。具体创建MRS集群的操作可以参考**创建MRS集群**。
  - 本示例创建的MRS集群版本为：MRS 3.1.0。
  - 本示例创建的MRS集群开启了Kerberos认证。
- 已创建CDM迁移集群。创建CDM集群的操作可以参考**创建CDM集群**。

  说明

  - 如果目标数据源为云下的数据库，则需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP、CDM云上安全组出方向放通云下数据源所在的主机、数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
  - 数据源为云上的MRS、DWS时，网络互通需满足如下条件：

    i. 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。

    ii. 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。

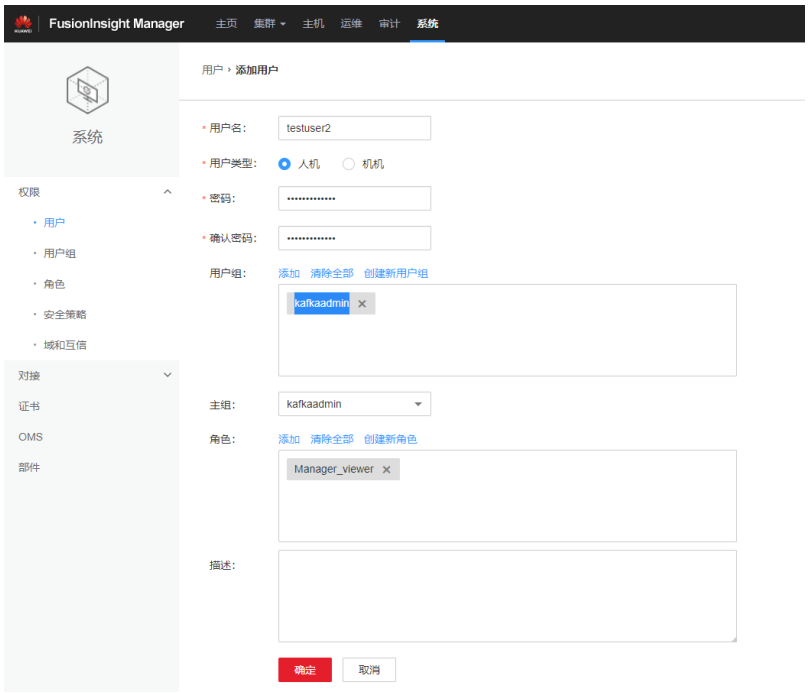
    iii. 此外，您还必须确保该云服务的实例与CDM集群所属的企业项目必须相同，如果不同，需要修改工作空间的企业项目。

  本示例CDM集群的虚拟私有云、子网以及安全组和创建的MRS集群保持一致。

步骤一：数据准备

- MRS集群上创建Kafka的Topic并且向Topic发送消息。
  a. 参考**访问MRS Manager**登录MRS Manager。
  b. 在MRS Manager上，选择“系统 > 权限 > 用户”，单击“添加用户”，在添加用户页面分别配置如下参数。
     i. 用户名：自定义的用户名。当前示例输入为：testuser2。
     ii. 用户类型：当前选择为“人机”。
     iii. 密码和确认密码：输入当前用户名对应的密码。
     iv. 用户组和主组：选择kafkaadmin。
     v. 角色：选择Manager_viewer角色。

     图 2-11 MRS Manager 上创建 Kafka 用户



  c. 在MRS Manager上，选择“集群 > 待操作的集群名称 > 服务 > ZooKeeper > 实例”，获取ZooKeeper角色实例的IP地址，为后续步骤做准备。
  d. 在MRS Manager上，选择“集群 > 待操作的集群名称 > 服务 > kafka > 实例”，获取kafka角色实例的IP地址，为后续步骤做准备。
  e. 参考**安装MRS客户端**下载并安装Kafka客户端。例如，当前Kafka客户端安装在MRS主机节点的“/opt/kafkaclient”目录上。
  f. 以root用户进入客户端安装目录下。
     例如：**cd /opt/kafkaclient**
  g. 执行以下命令配置环境变量。
     source bigdata_env
  h. 因为当前集群启用了Kerberos认证，则需要执行以下命令进行安全认证。认证用户为**b**中创建的用户。
     kinit *b中创建的用户名*
     例如，**kinit testuser2**

i. 执行以下命令创建名字为kafkatopic的Kafka Topic。

```
kafka-topics.sh --create --zookeeper ZooKeeper角色实例所在节点IP地址1:2181,ZooKeeper角色实例所在节点IP地址2:2181,ZooKeeper角色实例所在节点IP地址3:2181/kafka --replication-factor 1 --partitions 1 --topic kafkatopic
```

上述命令中的“ZooKeeper角色实例所在节点IP地址”即为**c**中获取的ZooKeeper实例IP。

j. 执行以下命令向kafkatopic发送消息。

```
kafka-console-producer.sh --broker-list Kafka角色实例所在节点的IP地址1:21007,Kafka角色实例所在节点的IP地址2:21007,Kafka角色实例所在节点的IP地址3:21007 --topic kafkatopic --producer.config /opt/kafkaclient/Kafka/kafka/config/producer.properties
```

上述命令中的“Kafka角色实例所在节点的IP地址”即为**d**中获取的Kafka实例IP。

发送测试消息内容如下：

```
{"PageViews":5, "UserID":"4324182021466249494", "Duration":146,"Sign":-1}
```

- 在DLI上创建数据库和表。

a. 登录DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列。

在编辑器中输入以下语句创建数据库，例如当前创建迁移后的DLI数据库testdb。详细的DLI创建数据库的语法可以参考**创建DLI数据库**。

```
create database testdb;
```

b. 创建数据库下的表。详细的DLI建表语法可以参考**创建DLI表**。

```
CREATE TABLE testdlitable(value STRING);
```

步骤二：数据迁移

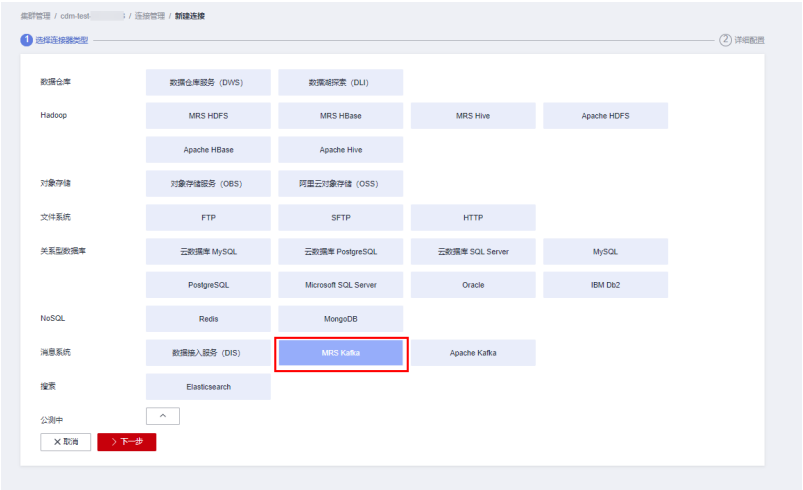
1. 配置CDM数据源连接。

a. 配置源端MRS Kafka的数据源连接。

i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

ii. 在作业管理界面，选择“连接管理”，单击“新建连接”，连接器类型选择“MRS Kafka”，单击“下一步”。

图 2-12 创建 MRS Kafka 数据源



iii. 配置源端MRS Kafka的数据源连接，具体参数配置如下。

表 2-5 MRS Kafka 数据源配置

参数	值
名称	自定义MRS Kafka数据源名称。例如当前配置为“source_kafka”。
Manager IP	单击输入框旁边的“选择”按钮，选择当前MRS Kafka集群即可自动关联出来Manager IP。
用户名	在**b**中创建的MRS Kafka用户名。
密码	对应MRS Kafka用户名的密码。
认证类型	如果当前MRS集群为普通集群则选择为SIMPLE，如果是MRS集群启用了Kerberos安全认证则选择为KERBEROS。 本示例选择为：KERBEROS。

更多参数的详细说明可以参考**CDM上配置Kafka连接**。

图 2-13 CDM 配置 MRS Kafka 数据源连接

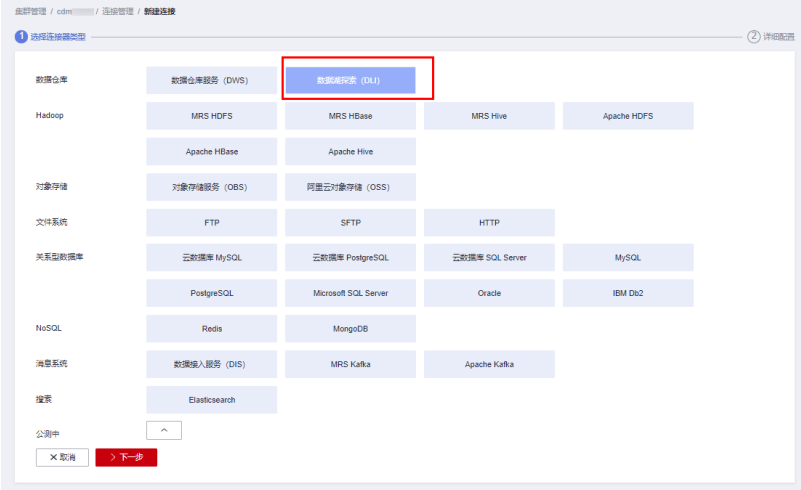
iv. 单击“保存”完成MRS Kafka数据源配置。

b. 配置目的端DLI的数据源连接。

i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

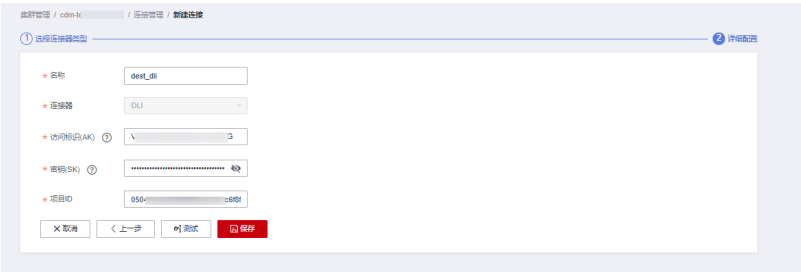
ii. 在作业管理界面，选择“连接管理”，单击“新建连接”，连接器类型选择“数据湖探索（DLI）”，单击“下一步”。

图 2-14 创建 DLI 数据源连接



iii. 配置目的端DLI数据源连接连接参数。具体参数配置可以参考**在CDM上配置DLI连接**。

图 2-15 配置 DLI 数据源连接参数



iv. 配置完成后，单击“保存”完成DLI数据源配置。

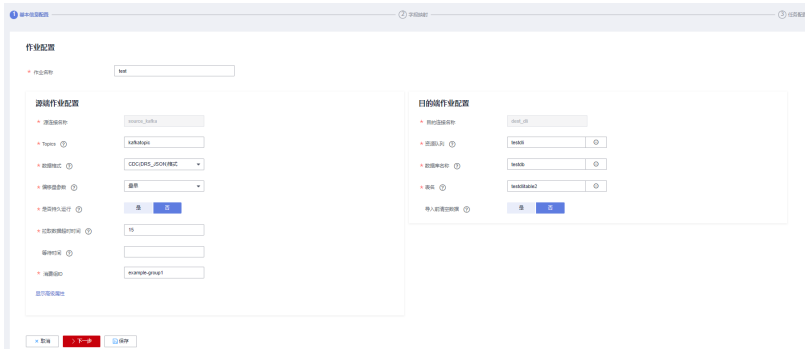
2. 创建CDM迁移作业。

a. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

b. 在“作业管理”界面，选择“表/文件迁移”，单击“新建作业”。

c. 在新建作业界面，配置当前作业配置信息，具体参数参考如下：

图 2-16 新建 CDM 作业作业配置



i. 作业名称：自定义数据迁移的作业名称。例如，当前定义为：test。

ii. 源端作业配置，具体参考如下：

表 2-6 源端作业配置

参数名	参数值
源连接名称	选择**1.a**中已创建的数据源名称。
Topics	选择MRS Kafka待迁移的Topic名称，支持单个或多个Topic。当前示例为：kafkatopic。
数据格式	根据实际情况选择当前消息格式。本示例选择为：CDC（DRS_JSON），以DRS_JSON格式解析源数据。
偏移量参数	从Kafka拉取数据时的初始偏移量。本示例当前选择为：最新。 • 最新：最大偏移量，即拉取最新的数据。 • 最早：最小偏移量，即拉取最早的数据。 • 已提交：拉取已提交的数据。 • 时间范围：拉取时间范围内的数据。
是否持久运行	用户自定义是否永久运行。当前示例选择为：否。
拉取数据超时时间	持续拉取数据多长时间超时，单位分钟。当前示例配置为：15。
等待时间	可选参数，超出等待时间还是无法读取到数据，则不再读取数据，单位秒。当前示例不配置该参数。
消费组ID	用户指定消费组ID。当前使用MRS Kafka默认的消息组ID：“example-group1”。

其他参数的详细配置说明可以参考：**CDM配置kafka源端参数**。

iii. 目的端作业配置，具体参考如下：

表 2-7 目的端作业配置

参数名	参数值
目的连接名称	选择**1.b**已创建的DLI数据源连接。
资源队列	选择已创建的DLI SQL类型的队列。
数据库名称	选择DLI下已创建的数据库。当前示例为**在DLI上创建数据库和表**中创建的数据库名，即为“testdb”。
表名	选择DLI下已创建的表名。当前示例为**在DLI上创建数据库和表**中创建的表名，即为“testdlitable”。

参数名	参数值
导入前清空数据	选择导入前是否清空目的表的数据。当前示例选择为“否”。 如果设置为是，任务启动前会清除目标表中数据。

详细的参数配置可以参考：**CDM配置DLI目的端参数**。

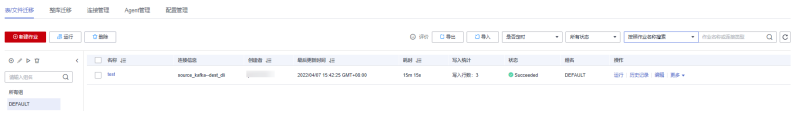
3. 单击“下一步”，进入到字段映射界面，CDM会自动匹配源和目的字段。
   - 如果字段映射顺序不匹配，可通过拖拽字段调整。
   - 如果选择在目的端自动创建类型，这里还需要配置每个类型的字段类型、字段名称。
   - CDM支持迁移过程中转换字段内容，详细请参见**字段转换**。

图 2-17 字段映射

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，请参见**配置定时任务**。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。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，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。
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。

图 2-18 迁移作业进度和结果查询



步骤三：结果查询

CDM迁移作业运行完成后，再登录到DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列，数据库选择已**a**已创建的数据库，执行DLI表查询语句，查询Kafka数据是否已成功迁移到DLI的“testdlitable”表中。

```
select * from testdlitable;
```

2.4 迁移 Elasticsearch 数据至 DLI

本文为您介绍如何通过CDM数据同步功能，迁移Elasticsearch类型的CSS集群数据至DLI。其他自建的Elasticsearch等服务数据，均可以通过CDM与DLI进行双向同步。

前提条件

- 已创建DLI的SQL队列。创建DLI队列的操作可以参考**创建DLI队列**。

  注意

  创建DLI队列时**队列类型**需要选择为“**SQL队列**”。

- 已创建Elasticsearch类型的CSS集群。具体创建CSS集群的操作可以参考**创建CSS集群**。

  本示例创建的CSS集群版本为：7.6.2，集群为非安全集群。

- 已创建CDM迁移集群。创建CDM集群的操作可以参考**创建CDM集群**。

  说明

  - 如果目标数据源为云下的数据库，则需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP、CDM云上安全组出方向放通云下数据源所在的主机、数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
  - 数据源为云上的CSS服务时，网络互通需满足如下条件：

    i. 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。

    ii. 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。

    iii. 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。

  本示例CDM集群的虚拟私有云、子网以及安全组和创建的CSS集群保持一致。

步骤一：数据准备

- CSS集群上创建索引并导入数据。

  a. 登录CSS管理控制台，选择“集群管理 > Elasticsearch”。

  b. 在集群管理界面，在已创建的CSS集群的“操作”列，单击“Kibana”访问集群。

  c. 在Kibana的左侧导航中选择“Dev Tools”，进入到Console界面。

  d. 在Console界面，执行如下命令创建索引“my_test”。

  ```
  PUT /my_test
  {
    "settings": {
      "number_of_shards": 1
    },
    "mappings": {
  ```



```
        "properties": {
       "productName": {
         "type": "text",
         "analyzer": "ik_smart"
       },
       "size": {
         "type": "keyword"
       }
     }
   }
 }
```

e. 在Console界面，执行如下命令，将数据导入到“my_test”索引中。

```
POST /my_test/_doc/_bulk
{"index":{}}
{"productName":"2017秋装新款文艺衬衫女装","size":"L"}
{"index":{}}
{"productName":"2017秋装新款文艺衬衫女装","size":"M"}
{"index":{}}
{"productName":"2017秋装新款文艺衬衫女装","size":"S"}
{"index":{}}
{"productName":"2018春装新款牛仔裤女装","size":"M"}
{"index":{}}
{"productName":"2018春装新款牛仔裤女装","size":"S"}
{"index":{}}
{"productName":"2017春装新款休闲裤女装","size":"L"}
{"index":{}}
{"productName":"2017春装新款休闲裤女装","size":"S"}
```

当返回结果信息中“**errors**”字段的值为“false”时，表示导入数据成功。

- 在DLI上创建数据库和表。

a. 登录DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列。

在编辑器中输入以下语句创建数据库，例如当前创建迁移后的DLI数据库testdb。详细的DLI创建数据库的语法可以参考**创建DLI数据库**。

```
create database testdb;
```

b. 创建数据库下的表。详细的DLI建表语法可以参考**创建DLI表**。

```
create table tablecss(size string, productname string);
```

步骤二：数据迁移

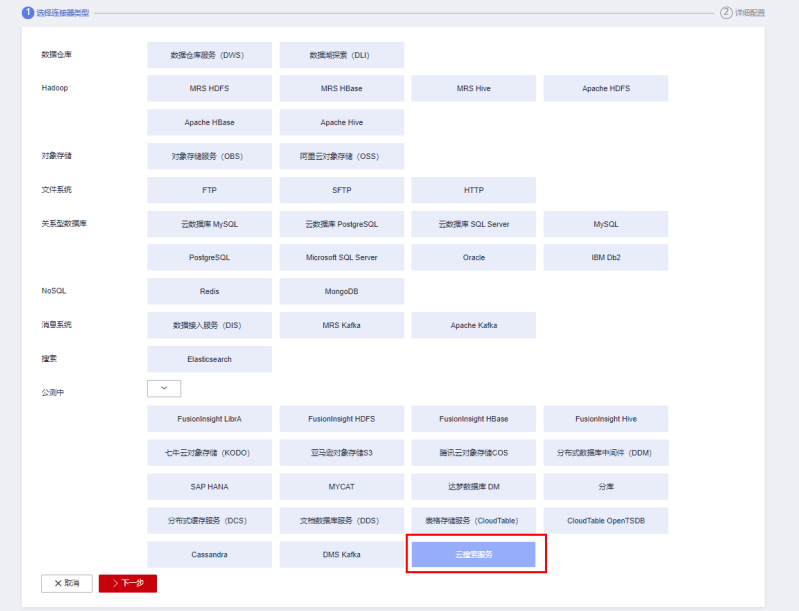
1. 配置CDM数据源连接。

a. 配置源端CSS的数据源连接。

i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

ii. 在作业管理界面，选择“连接管理”，单击“新建连接”，连接器类型选择“云搜索服务”，单击“下一步”。

图 2-19 创建 CSS 数据源



iii. 配置源端CSS的数据源连接，具体参数配置如下。详细参数配置可以参考**CDM上配置CSS连接**。

表 2-8 CSS 数据源配置

参数	值
名称	自定义CSS数据源名称。例如当前配置为“source_css”。
Elasticsearch服务器列表	单击输入框旁边的“选择”按钮，选择当前CSS集群即可自动关联出来Elasticsearch服务器列表。
安全模式认证	如果所需连接的CSS集群在创建时开启了“安全模式”，该参数需设置为“是”，否则设置为“否”。 本示例选择为“否”。

图 2-20 CDM 配置 CSS 数据源

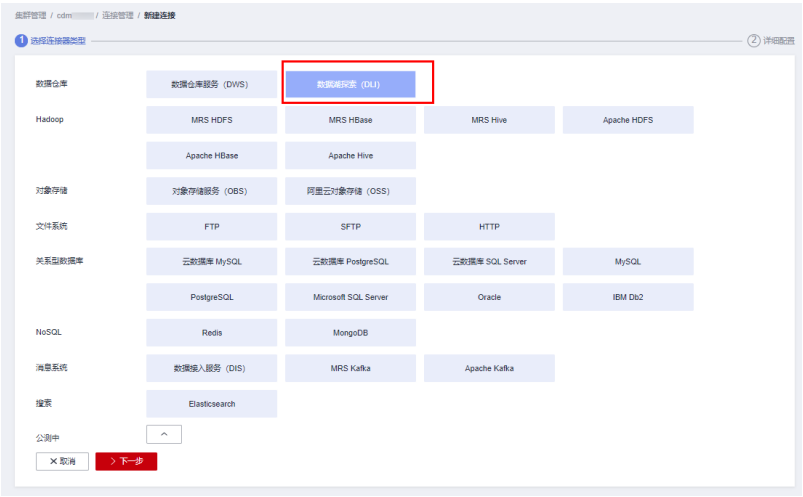
iv. 单击“保存”完成CSS数据源配置。

b. 配置目的端DLI的数据源连接。

i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

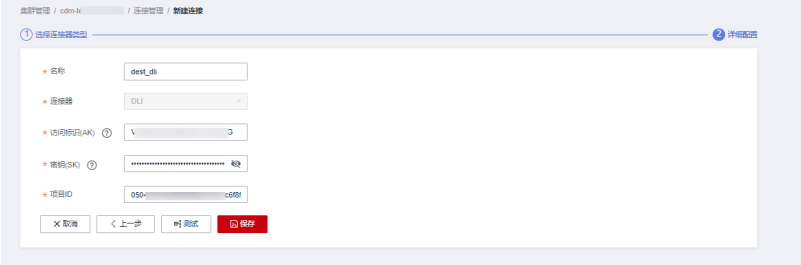
ii. 在作业管理界面，选择“连接管理”，单击“新建连接”，连接器类型选择“数据湖探索（DLI）”，单击“下一步”。

图 2-21 创建 DLI 数据源连接



iii. 配置目的端DLI数据源连接连接参数。具体参数配置可以参考**在CDM上配置DLI连接**。

图 2-22 配置 DLI 数据源连接参数



iv. 配置完成后，单击“保存”完成DLI数据源配置。

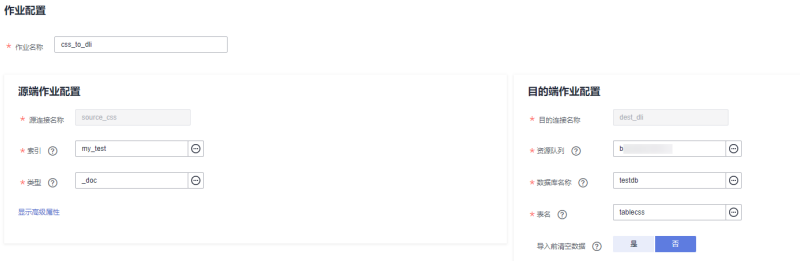
2. 创建CDM迁移作业。

a. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

b. 在“作业管理”界面，选择“表/文件迁移”，单击“新建作业”。

c. 在新建作业界面，配置当前作业配置信息，具体参数参考如下：

图 2-23 新建 CDM 作业作业配置



i. 作业名称：自定义数据迁移的作业名称。例如，当前定义为：css_to_dli。

ii. 源端作业配置，具体参考如下：

表 2-9 源端作业配置

参数名	参数值
源连接名称	选择**1.a**中已创建的数据源名称。
索引	选择CSS集群中创建的Elasticsearch索引名。当前示例为**CSS集群上创建索引并导入数据**中创建的索引“my_test”。 索引名称只能全部小写，不能有大写。
类型	Elasticsearch的类型，类似关系数据库中的表名称。类型名称只能全部小写，不能有大写。当前示例为：“_doc”。

更多其他参数说明可以参考：**CDM配置CSS源端参数**。

iii. 目的端作业配置，具体参考如下：

表 2-10 目的端作业配置

参数名	参数值
目的连接名称	选择**1.b**已创建的DLI数据源连接。
资源队列	选择已创建的DLI SQL类型的队列。
数据库名称	选择DLI下已创建的数据库。当前示例为**在DLI上创建数据库和表**中创建的数据库名，即为“testdb”。
表名	选择DLI下已创建的表名。当前示例为**在DLI上创建数据库和表**中创建的表名，即为“tablecss”。
导入前清空数据	选择导入前是否清空目的表的数据。当前示例选择为“否”。 如果设置为是，任务启动前会清除目标表中数据。

详细的参数配置可以参考：**CDM配置DLI目的端参数**。

3. 单击“下一步”，进入到字段映射界面，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。
   - 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**。

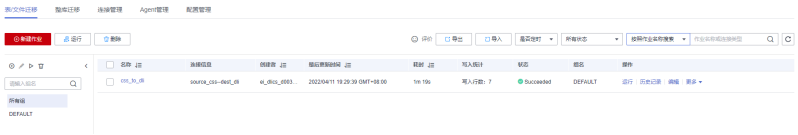
   图 2-24 字段映射

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，请参见**配置定时任务**。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。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，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。
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。

图 2-25 迁移作业进度和结果查询

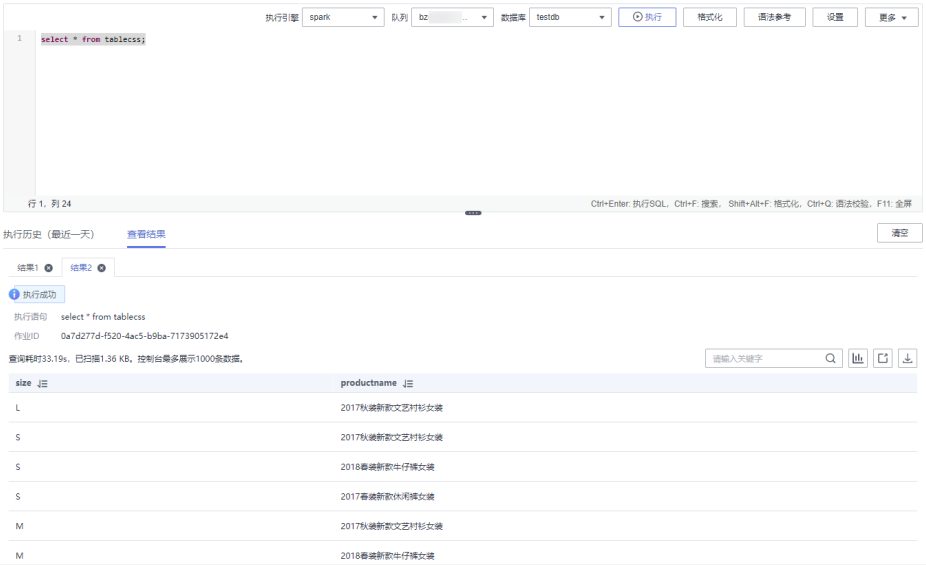


步骤三：结果查询

CDM迁移作业运行完成后，再登录到DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列，数据库选择已a中已创建的数据库，执行DLI表查询语句，查询CSS的数据是否已成功迁移到DLI的“tablecss”表中。

```
select * from tablecss;
```

图 2-26 迁移后查询 DLI 的表数据



2.5 迁移 RDS 数据至 DLI

本文为您介绍如何通过CDM数据同步功能，迁移关系型数据库RDS数据至DLI。其他关系型数据库数据都可以通过CDM与DLI进行双向同步。

前提条件

- 已创建DLI的SQL队列。创建DLI队列的操作可以参考**创建DLI队列**。

  注意

  创建DLI队列时**队列类型**需要选择为“**SQL队列**”。

- 已创建云数据库RDS的MySQL的数据库实例。具体创建RDS集群的操作可以参考**创建RDS MySQL数据库实例**。
  - 本示例RDS数据库引擎：MySQL

- 本示例RDS MySQL数据库版本：5.7。

- 已创建CDM迁移集群。创建CDM集群的操作可以参考**创建CDM集群**。

  说明

  - 如果目标数据源为云下的数据库，则需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP、CDM云上安全组出方向放通云下数据源所在的主机、数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
  - 数据源为云上服务RDS、MRS时，网络互通需满足如下条件：

    i. 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。

    ii. 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。

    iii. 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。

  本示例CDM集群的虚拟私有云、子网以及安全组和RDS MySQL实例保持一致。

步骤一：数据准备

- RDS的MySQL的数据库实例上创建数据库和表。

  a. 登录RDS管理控制台，在“实例管理”界面，选择已创建的MySQL实例，选择操作列的“更多 > 登录”，进入数据管理服务实例登录界面。

  b. 输入实例登录的用户名和密码。单击“登录”，即可进入MySQL数据库并进行管理。

  c. 在数据库实例界面，单击“新建数据库”，数据库名定义为：testrdsdb，字符集保持默认即可。

  d. 在已创建的数据库的操作列，单击“SQL查询”，输入以下创建表语句，创建RDS MySQL表。

```
CREATE TABLE tabletest (
    `id` VARCHAR(32) NOT NULL,
    `name` VARCHAR(32) NOT NULL,
    PRIMARY KEY (`id`)
)    ENGINE = InnoDB
    DEFAULT CHARACTER SET = utf8mb4;
```

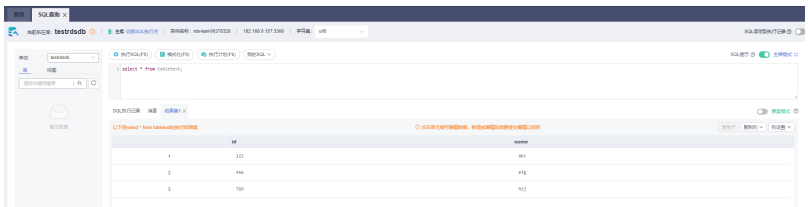
  e. 插入表数据。

```
insert into tabletest VALUES ('123','abc');
insert into tabletest VALUES ('456','efg');
insert into tabletest VALUES ('789','hij');
```

  f. 查询测试的表数据。

```
select * from tabletest;
```

图 2-27 查询 RDS 表数据



- 在DLI上创建数据库和表。
  a. 登录DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列。
     在编辑器中输入以下语句创建数据库，例如当前创建迁移后的DLI数据库testdb。详细的DLI创建数据库的语法可以参考**创建DLI数据库**。
     ```
     create database testdb;
     ```
  b. 在“SQL编辑器”中，数据库选择“testdb”，执行以下建表语句创建数据库下的表。详细的DLI建表语法可以参考**创建DLI表**。
     ```
     create table tabletest(id string,name string);
     ```

步骤二：数据迁移

1. 配置CDM数据源连接。
   a. 创建源端RDS数据库的连接。
      i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。
      ii. 首次创建RDS MySQL数据库连接时需要上传MySQL的驱动，单击“连接管理 > 驱动管理”，进入驱动管理界面。
      iii. 参考**CDM管理驱动**下载MySQL的驱动包到本地，将下载后驱动包本地解压，获取驱动的jar包文件。
      例如，当前下载MySQL驱动包压缩文件为“mysql-connector-java-5.1.48.zip”，解压后获取驱动文件“mysql-connector-java-5.1.48.jar”。
      iv. 返回到驱动管理界面，在驱动名称为MYSQL的操作列，单击“上传”，在“导入驱动文件”界面单击“添加文件”，将**1.a.iii**获取的驱动文件上传。
      v. 在驱动管理界面单击“返回”按钮回到连接管理界面，单击“新建连接”，连接器类型选择“云数据库 MySQL”，单击“下一步”。
      vi. 配置连接RDS的数据源连接参数，具体参数配置如下。

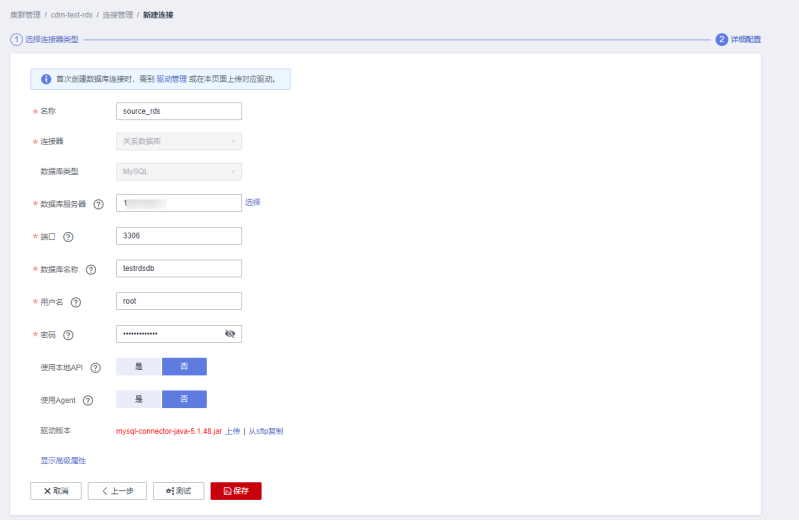
表 2-11 RDS MySQL 数据源配置

参数	值
名称	自定义RDS数据源名称。例如当前配置为：source_rds。
数据库服务	单击输入框旁边的“选择”按钮，选择当前已创建的RDS实例名即可自动关联出来数据库服务器地址。
端口	RDS实例的端口。选择数据库服务器后自动自动关联。
数据库名称	当前需要迁移的RDS MySQL数据库名称。当前示例为**c**中创建的数据库“testrdsdb”。
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。 本示例使用创建RDS MySQL数据库实例的默认用户“root”。

参数	值
密码	对应的RDS MySQL数据库用户的密码。

其他更多参数保持默认即可，如果需要了解详细参数说明，可以参考**配置关系数据库连接**。单击“保存”完成RDS MySQL数据源连接配置。

图 2-28 CDM 配置 RDS MySQL 数据源

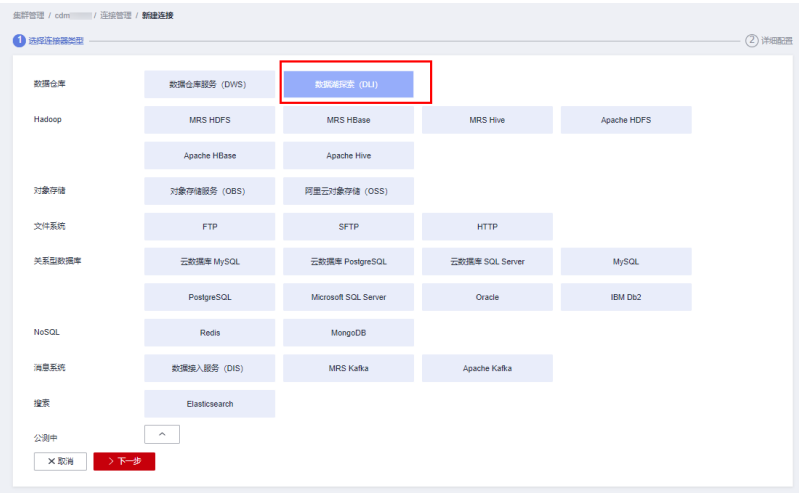


b. 创建目的端DLI数据源的连接。

i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

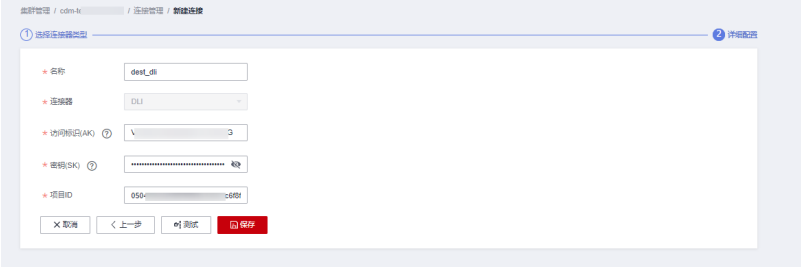
ii. 在作业管理界面，选择“连接管理”，单击“新建连接”，连接器类型选择“数据湖探索（DLI）”，单击“下一步”。

图 2-29 创建 DLI 数据源连接



i. 配置目的端DLI数据源连接。具体参数配置可以参考**在CDM上配置DLI连接**。

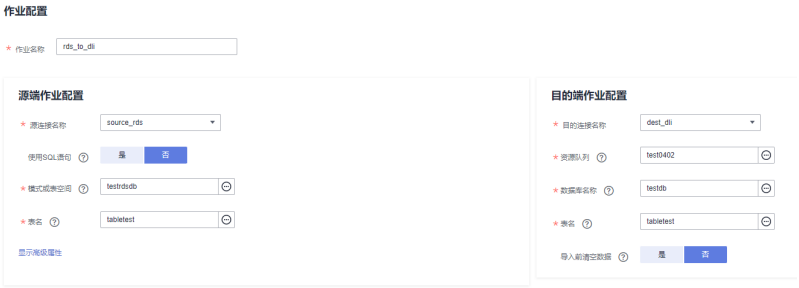
图 2-30 创建 DLI 数据源连接



配置完成后，单击“保存”完成DLI数据源配置。

2. 创建CDM迁移作业。
   a. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。
   b. 在“作业管理”界面，选择“表/文件迁移”，单击“新建作业”。
   c. 在新建作业界面，配置当前作业配置信息，具体参数参考如下：

图 2-31 CDM 数据迁移作业配置



   i. 作业名称：自定义数据迁移的作业名称。例如，当前定义为：rds_to_dli。
   ii. 源端作业配置，具体参考如下：

表 2-12 源端作业配置

参数名	参数值
源连接名称	选择1.a中已创建的数据源名称。
使用SQL语句	“使用SQL语句”选择“是”时，您可以在这里输入自定义的SQL语句，CDM将根据该语句导出数据。 本示例当前选择为“否”。
模式或表空间	选择RDS MySQL待迁移的数据库名称。例如当前待迁移的表数据数据库为“testrdsdb”。
表名	待迁移的RDS MySQL数据表名。当前为d中的“tabletest”表。

更多详细参数配置请参考**配置关系数据库源端参数**。

iii. 目的端参数配置，具体参考如下：

表 2-13 目的端作业配置

参数名	参数值
目的连接名称	选择已创建的DLI数据源连接。
资源队列	选择已创建的DLI SQL类型的队列。
数据库名称	选择DLI下已创建的数据库。当前示例为**在DLI上创建数据库和表**创建的数据库名，即为“testdb”。
表名	选择DLI下已创建的表名。当前示例为**在DLI上创建数据库和表**创建的表名，即为“tabletest”。
导入前清空数据	选择导入前是否清空目的表的数据。当前示例选择为“否”。 如果设置为是，任务启动前会清除目标表中数据。

详细的参数配置可以参考：**CDM配置DLI目的端参数**。

iv. 单击“下一步”，进入到字段映射界面，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。
- 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 2-32 字段映射

v. 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

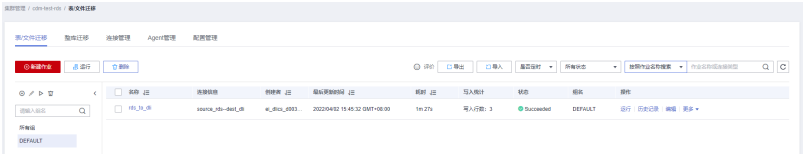
该步骤用户可以配置如下可选功能：

- 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，请参见**配置定时任务**。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。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，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。

vi. 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。

图 2-33 迁移作业进度和结果查询

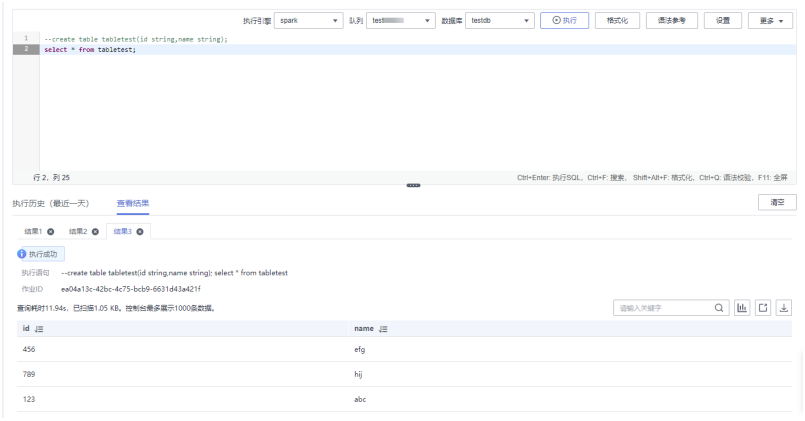


步骤三：结果查询

CDM迁移作业运行完成后，再登录到DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列，数据库选择**在DLI上创建数据库和表**已创建的数据库，执行DLI表查询语句，查询RDS MySQL表数据是否已成功迁移到DLI的“tabletest”表中。

```
select * from tabletest;
```

图 2-34 查询 DLI 表数据



2.6 迁移 DWS 数据至 DLI

本文为您介绍如何通过CDM数据同步功能，迁移数据仓库服务DWS数据至DLI。

前提条件

- 已创建DLI的SQL队列。创建DLI队列的操作可以参考**创建DLI队列**。

  注意

  创建DLI队列时**队列类型**需要选择为“**SQL队列**”。

- 已创建数据仓库服务DWS集群。具体创建DWS集群的操作可以参考**创建DWS集群**。
- 已创建CDM迁移集群。创建CDM集群的操作可以参考**创建CDM集群**。

说明

- 如果目标数据源为云下的数据库，则需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP、CDM云上安全组出方向放通云下数据源所在的主机、数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
- 数据源为云上的DWS、MRS等服务时，网络互通需满足如下条件：

  i. 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。

  ii. 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。

  iii. 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。

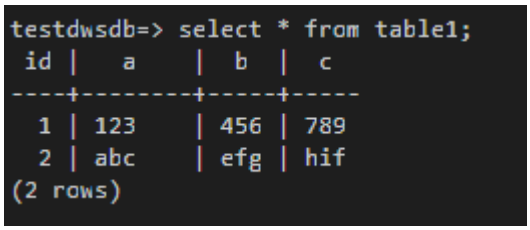
本示例CDM集群的虚拟私有云、子网以及安全组和DWS集群保持一致。

步骤一：数据准备

- DWS集群上创建数据库和表。
  a. 参考**使用gsql命令行客户端连接DWS集群**连接已创建的DWS集群。
  b. 执行以下命令连接DWS集群的默认数据库“gaussdb”：
  ```
  gsql -d gaussdb -h DWS集群连接地址 -U dbadmin -p 8000 -W password -r
  ```
     - gaussdb：DWS集群默认数据库。
     - DWS集群连接地址：请参见**获取集群连接地址**进行获取。如果通过公网地址连接，请指定为集群“公网访问地址”或“公网访问域名”，如果通过内网地址连接，请指定为集群“内网访问地址”或“内网访问域名”。如果通过弹性负载均衡连接，请指定为“弹性负载均衡地址”。
     - dbadmin：创建集群时设置的默认管理员用户名。
     - -W：默认管理员用户的密码。
  c. 在命令行窗口输入以下命令创建数据库“testdwsdb”。
  ```
  CREATE DATABASE testdwsdb;
  ```
  d. 执行以下命令，退出gaussdb数据库，连接新创建的数据库“testdwsdb”。
  ```
  \q
  gsql -d testdwsdb -h DWS集群连接地址 -U dbadmin -p 8000 -W password -r
  ```
  e. 执行以下命令创建表并插入数据。

  创建表：
  ```
  CREATE TABLE table1(id int, a char(6), b varchar(6),c varchar(6)) ;
  ```
  插入表数据：
  ```
  INSERT INTO table1 VALUES(1,'123','456','789');
  INSERT INTO table1 VALUES(2,'abc','efg','hif');
  ```
  f. 查询表数据确认数据插入成功。
  ```
  select * from table1;
  ```

图 2-35 查询表数据



- 在DLI上创建数据库和表。
  a. 登录DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列。

     在编辑器中输入以下语句创建数据库，例如当前创建迁移后的DLI数据库testdb。详细的DLI创建数据库的语法可以参考**创建DLI数据库**。

     ```
     create database testdb;
     ```

  b. 在“SQL编辑器”中，数据库选择“testdb”，执行以下建表语句创建数据库下的表。详细的DLI建表语法可以参考**创建DLI表**。

     ```
     create table tabletest(id INT, name1 string, name2 string, name3 string);
     ```

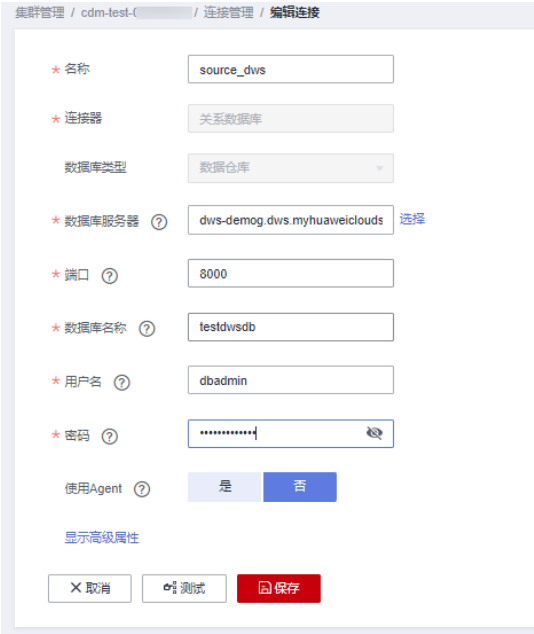
步骤二：数据迁移

1. 配置CDM数据源连接。
   a. 创建源端DWS数据库的连接。
      i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。
      ii. 在作业管理界面，选择“连接管理”，单击“新建连接”，连接器类型选择“数据仓库服务（DWS）”，单击“下一步”。
      iii. 配置连接DWS的数据源连接参数，具体参数配置如下。

表 2-14 DWS 数据源配置

参数	值
名称	自定义DWS数据源名称。例如当前配置为：source_dws。
数据库服务器	单击输入框旁边的“选择”按钮，选择当前已创建的DWS集群名称。
端口	DWS数据库的端口，默认为：8000。
数据库名称	当前需要迁移的DWS数据库名称。当前示例为**DWS集群上创建数据库和表**中创建的数据库“testdwsdb”。
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。 本示例使用创建DWS数据库实例的默认管理员用户“dbadmin”。
密码	对应的DWS数据库用户的密码。

图 2-36 CDM 配置 DWS 数据源



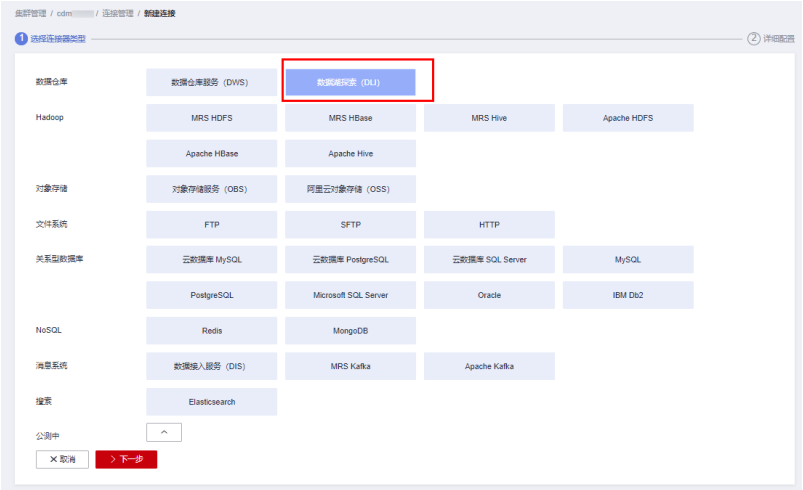
其他更多参数保持默认即可，如果需要了解更多参数说明，可以参考**配置关系数据库连接**。单击“保存”完成DWS数据源连接配置。

b. 创建目的端DLI数据源的连接。

i. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。

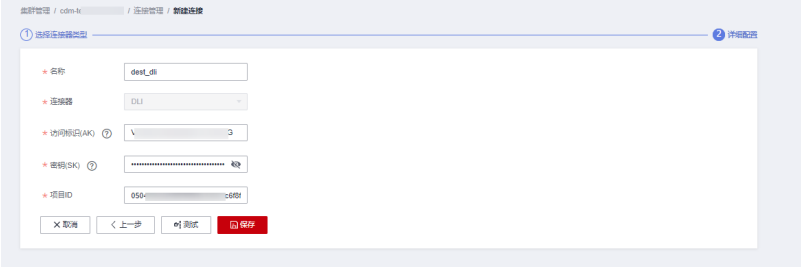
ii. 在作业管理界面，选择“连接管理”，单击“新建连接”，连接器类型选择“数据湖探索（DLI）”，单击“下一步”。

图 2-37 创建 DLI 数据源连接



i. 配置目的端DLI数据源连接。具体参数配置可以参考**在CDM上配置DLI连接**。

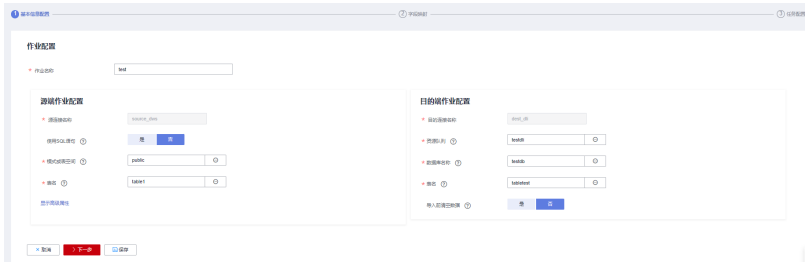
图 2-38 创建 DLI 数据源连接



配置完成后，单击“保存”完成DLI数据源配置。

2. 创建CDM迁移作业。
   a. 登录CDM控制台，选择“集群管理”，选择已创建的CDM集群，在操作列选择“作业管理”。
   b. 在“作业管理”界面，选择“表/文件迁移”，单击“新建作业”。
   c. 在新建作业界面，配置当前作业配置信息，具体参数参考如下：

图 2-39 CDM 数据迁移作业配置



   i. 作业名称：自定义数据迁移的作业名称。例如，当前定义为：test。
   ii. 源端作业配置，具体参考如下：

表 2-15 源端作业配置

参数名	参数值
源连接名称	选择1.a中已创建的数据源名称。
使用SQL语句	“使用SQL语句”选择“是”时，您可以在这里输入自定义的SQL语句，CDM将根据该语句导出数据。 本示例当前选择为“否”。

参数名	参数值
模式或表空间	“使用SQL语句”选择“否”时，显示该参数，表示待抽取数据的模式或表空间名称。单击输入框后面的按钮可进入模式选择界面，用户也可以直接输入模式或表空间名称。 本示例因为**DWS集群上创建数据库和表**中没有创建SCHEMA，则本参数为默认的“public”。 如果选择界面没有待选择的模式或表空间，请确认对应连接里的账号是否有元数据查询的权限。 **说明** 该参数支持配置通配符（*），实现导出以某一前缀开头或者以某一后缀结尾的所有数据库。例如： SCHEMA*表示导出所有以“SCHEMA”开头的数据库。 *SCHEMA表示导出所有以“SCHEMA”结尾的数据库。 *SCHEMA*表示数据库名称中只要有“SCHEMA”字符串，就全部导出。
表名	待迁移的DWS数据表名。当前为**DWS集群上创建数据库和表**中的“table1”表。

更多详细参数配置请参考**配置关系数据库源端参数**。

iii. 目的端作业参数配置，具体参考如下：

表 2-16 目的端作业配置

参数名	参数值
目的连接名称	选择已创建的DLI数据源连接。
资源队列	选择已创建的DLI SQL类型的队列。
数据库名称	选择DLI下已创建的数据库。当前示例为**在DLI上创建数据库和表**创建的数据库名，即为“testdb”。
表名	选择DLI下已创建的表名。当前示例为**在DLI上创建数据库和表**创建的表名，即为“tabletest”。
导入前清空数据	选择导入前是否清空目的表的数据。当前示例选择为“否”。 如果设置为是，任务启动前会清除目标表中数据。

详细的参数配置可以参考：**CDM配置DLI目的端参数**。

iv. 单击“下一步”，进入到字段映射界面，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。
- 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 2-40 字段映射

v. 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，请参见**配置定时任务**。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。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，写入脏数据前需要先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。

vi. 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。

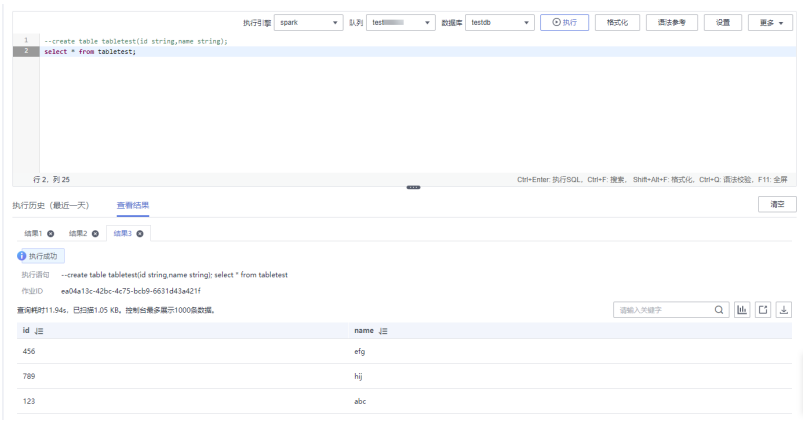
图 2-41 迁移作业进度和结果查询

步骤三：结果查询

CDM迁移作业运行完成后，再登录到DLI管理控制台，选择“SQL编辑器”，在SQL编辑器中“执行引擎”选择“spark”，“队列”选择已创建的SQL队列，数据库选择**在DLI上创建数据库和表**中已创建的数据库，执行DLI表查询语句，查询DWS表数据是否已成功迁移到DLI的“tabletest”表中。

```
select * from tabletest;
```

图 2-42 查询 DLI 表数据



3 数据分析

3.1 使用 DLI 进行车联网场景驾驶行为数据分析

应用场景

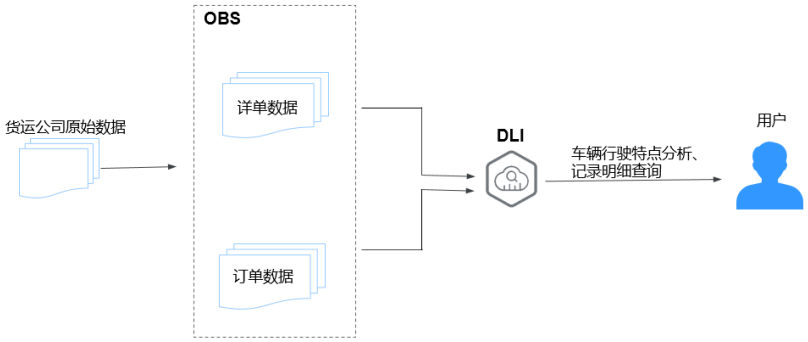
在车联网领域，云计算与大数据为企业提供了强大的分析挖掘能力，可以帮助企业和车队管理者更加科学、便捷地进行车辆数据管理与分析。

方案架构

根据已有的某货运公司车辆定时上报的详单数据和货运订单数据，DLI可以完成对该货运公司车辆行驶特点分析、记录明细的查询。

详细的数据说明请参考**数据说明**。

图 3-1 方案简介



流程指导

使用DLI进行驾驶行为数据分析的操作过程主要包括以下步骤：

步骤1：上传数据。将数据上传到对象存储服务OBS，为后面使用DLI完成数据分析做准备。

步骤2：分析数据。使用DLI对待分析的数据进行查询。

示例代码

具体样例数据及详细SQL语句可以通过**数据包**进行下载。

方案优势

- 数据免搬迁：DLI支持与多种数据源的对接，直接通过SQL建表就可以完成数据源的映射。
- 简单易用：直接使用标准SQL编写指标分析逻辑，无需关注背后复杂的分布式计算平台。
- 按需计费：日志分析按时效性要求按周期进行调度，每次调度之间存在大量空闲期。DLI按需计费只在使用期间收费，有效节约队列成本。

资源和成本规划

表 3-1 资源和成本规划

资源	资源说明	成本说明
OBS	需要创建一个OBS桶将数据上传到对象存储服务OBS，为后面使用DLI完成数据分析做准备。	OBS的使用涉及以下几项费用： • 存储费用：静态网站文件存储在OBS中产生的**存储费用**。 • 请求费用：用户访问OBS中存储的静态网站文件时产生的**请求费用**。 • 流量费用：用户使用自定义域名通过公网访问OBS时产生的**流量费用**。 实际产生的费用与存储的文件大小、用户访问所产生的请求次数和流量大小有关，请根据自己的业务进行预估。
DLI	在创建SQL作业前需购买队列，使用DLI的队列资源时，按照队列CU时进行计费。	如购买按需计费的队列，在使用队列资源时，按照队列CU时进行计费。 以小时为单位进行结算。不足一小时按一小时计费，小时数按整点计算。队列CU时按需计费的计算费用=单价*CU数*小时数。

数据说明

- 详单数据

  车辆上报的详单数据，包括定时上报的位置记录和异常的驾驶行为触发的告警事件数据。

  表 3-2 详单数据

字段名称	字段类型	字段说明
driverID	string	驾驶员ID
carNumber	string	车牌号

字段名称	字段类型	字段说明
latitude	double	纬度
longitude	double	经度
speed	int	速度
direction	int	方向
siteName	string	地点
time	timestamp	记录上报时间
isRapidlySpeedup	int	急加速标识，“1”表示急加速，“0”表示非急加速
isRapidlySlowdown	int	急减速
isNeutralSlide	int	空挡滑行
isNeutralSlideFinished	int	空挡滑行结束
neutralSlideTime	bigint	空挡滑行时长
isOverspeed	int	超速
isOverspeedFinished	int	超速结束
overspeedTime	bigint	超速时长
isFatigueDriving	int	疲劳驾驶
isHthrottleStop	int	停车轰油门
isOilLeak	int	用油异常

- 订单数据
  订单数据记录了货运订单相关的信息。

  表 3-3 订单数据

字段名称	字段类型	字段说明
orderNumber	string	订单号
driverID	string	驾驶员ID
carNumber	string	车牌号
customerID	string	客户ID
sourceCity	string	出发城市
targetCity	string	到达城市
expectArriveTime	timestamp	期望送达时间

字段名称	字段类型	字段说明
time	timestamp	记录产生时间
action	string	事件类型，包括创建订单、开始发货、货物送达、订单签收等事件

步骤 1：上传数据

将数据上传到对象存储服务OBS，为后面使用DLI完成数据分析做准备。

1. 下载OBS Browser+。下载地址请参考《**对象存储服务工具指南**》。
2. 安装OBS Browser+。安装步骤请参考《**对象存储服务工具指南**》。
3. 登录OBS Browser+。OBS Browser+支持AK方式登录，以及授权码登录两种登录方式。登录步骤请参考《**对象存储服务工具指南**》。
4. 通过OBS Browser+上传数据。

   在OBS Browser+页面单击“创建桶”，按照要求选择“区域”和填写“桶名”（例如：dli-demo），其他参数保持默认或根据需要选择，创建桶成功后，返回桶列表，单击桶dli-demo。OBS Browser+提供强大的拖拽上传功能，您可以将本地的一个或多个文件或者文件夹拖拽到对象存储的对象列表或者并行文件系统的对象列表中；同时您也可以将文件或文件夹拖拽到指定的目录上，这样可以上传到指定的目录中。

   单击**Best_Practice_01.zip**获取本示例的测试数据，将“Best_Practice_01.zip”压缩包解压。后续操作说明如下：

   - 详单数据：将解压后Data目录下的“detail-records”文件夹上传到OBS桶根目录下。
   - 订单数据：将解压后Data目录下的“order-records”文件夹上传到OBS桶根目录下。

步骤 2：分析数据

使用DLI对分析的数据进行查询。

1. 创建数据库、表。
   a. 在Console页面上方菜单栏中单击“产品”，单击“大数据”分类中的“数据湖探索 DLI”。
   b. 在DLI控制台总览页面左侧，单击“SQL编辑器”，进入SQL作业编辑器页面。
   c. 在SQL作业编辑器左侧，选择“数据库”页签，单击创建demo数据库，请参见**图3-2**。

图 3-2 创建数据库

说明

“default”为内置数据库，不能使用该数据库名。

d. 选择demo数据库，在编辑框中输入以下SQL语句：

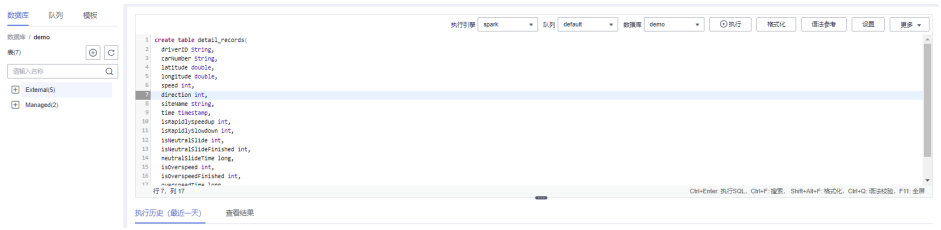
```
create table detail_records(
  driverID String,
  carNumber String,
  latitude double,
  longitude double,
  speed int,
  direction int,
  siteName String,
  time timestamp,
  isRapidlySpeedup int,
  isRapidlySlowdown int,
  isNeutralSlide int,
  isNeutralSlideFinished int,
  neutralSlideTime long,
  isOverspeed int,
  isOverspeedFinished int,
  overspeedTime long,
  isFatigueDriving int,
  isHthrottleStop int,
  isOilLeak int) USING CSV OPTIONS (PATH 'obs://dli-demo/detail-records/');
```

说明

使用该案例时，需将上述SQL语句中的文件路径修改为实际存放详单数据的OBS路径。

e. 单击“执行”，创建详单表detail_records，请参见图3-3。

图 3-3 创建详单表



f. 执行以下SQL语句，在demo数据库下创建告警事件表event_records，步骤同1.d和1.e。

```
create table event_records(
  driverID String,
  carNumber String,
  latitude double,
  longitude double,
  speed int,
  direction int,
  siteName String,
  time timestamp,
  isRapidlySpeedup int,
  isRapidlySlowdown int,
  isNeutralSlide int,
  isNeutralSlideFinished int,
  neutralSlideTime long,
  isOverspeed int,
  isOverspeedFinished int,
  overspeedTime long,
  isFatigueDriving int,
  isHthrottleStop int,
  isOilLeak int)
```

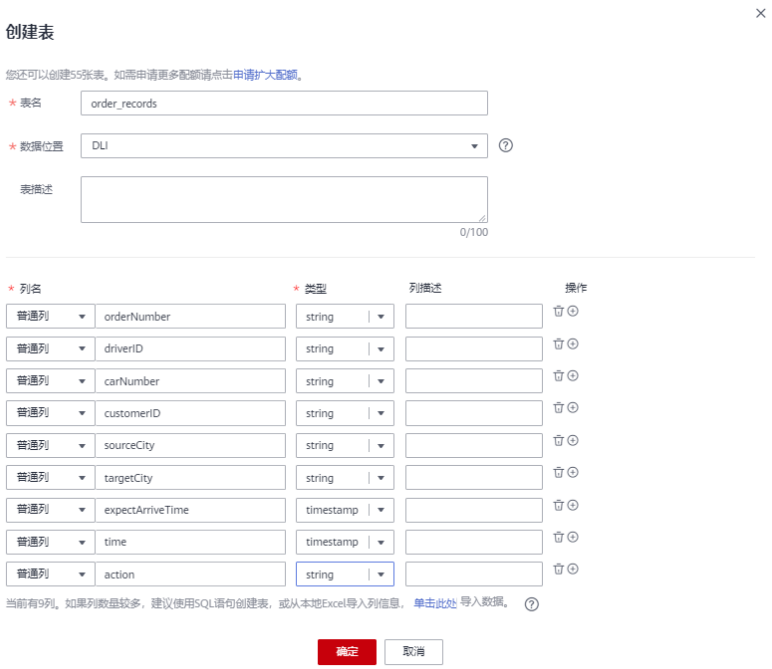
g. 执行以下SQL语句，将告警事件数据从详单中抽取出来插入到event_records表中。

```
insert into table event_records
(select *
from detail_records
where isRapidlySpeedup > 0
OR isRapidlySlowdown > 0
OR isNeutralSlide > 0
OR isNeutralSlideFinished > 0
OR isOverspeed > 0
OR isOverspeedFinished > 0
OR isFatigueDriving > 0
OR isHthrottleStop > 0
OR isOilLeak > 0)
```

h. 使用另一种方式创建订单表order_records。

在SQL作业编辑器左侧，选择“数据库”页签，单击数据库“demo”，单击表菜单右边的加号，创建表，数据位置选择DLI，请参见图3-4。字段类型请参见订单数据。

图 3-4 创建订单表

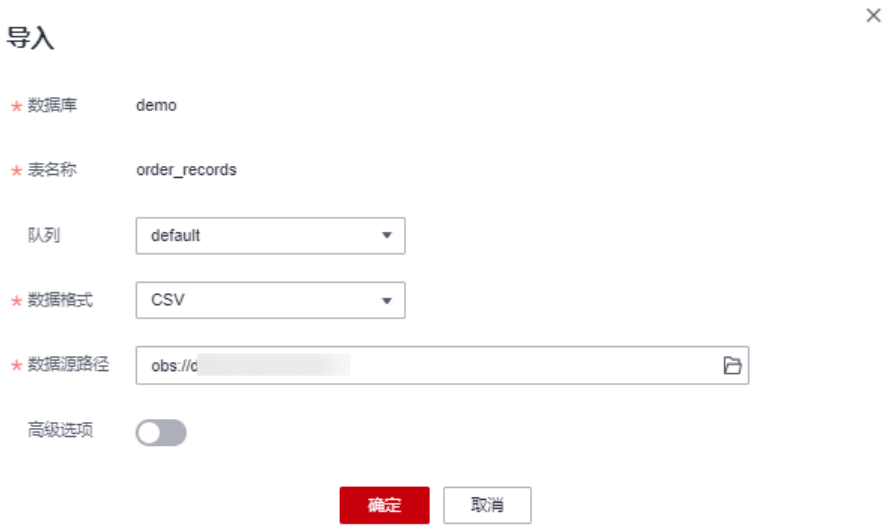


i. 将OBS数据导入到order_records表，单击“数据管理 > 库表管理”，单击demo数据库，进入“表管理”页面，单击order_records表对应“操作”列中的“更多”>“导入”，数据格式选择“CSV”，数据源路径为“obs://dli-demo/order-records/”，参数配置完成后单击“确定”。请参见图3-5。

说明

导入数据时，默认时间戳格式为“yyyy-MM-dd HH:mm:ss”，如果采用其他日期格式，可打开“高级选项”手动输入（本示例该选项不做修改）。

图 3-5 导入表数据



2. 执行查询

a. 执行以下SQL语句，对所有司机在某段时间的异常告警事件进行统计。

说明

常用查询语句可以在SQL编辑器中，选择“更多 > 设为模板”设置为模板。设为模板后，后续可以在模板管理页面找到对应模板进行SQL查询和修改。

具体操作为：选择“作业模板 > SQL模板 > 自定义模板”，在对应模板的操作列，单击“执行”会跳转到SQL语句编辑器，修改查询条件可以很方便地查找对应的数据。

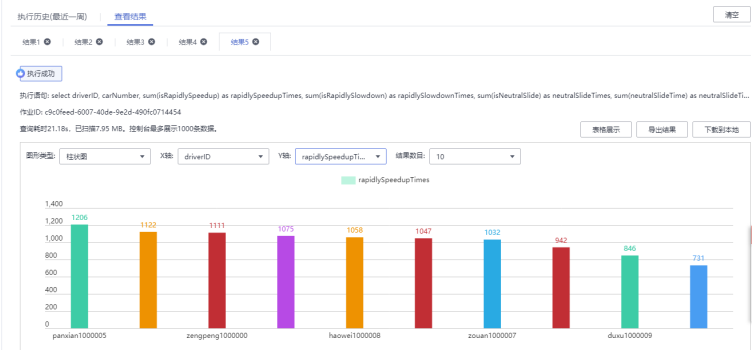
```
select
  driverID,
  carNumber,
  sum(isRapidlySpeedup) as rapidlySpeedupTimes,
  sum(isRapidlySlowdown) as rapidlySlowdownTimes,
  sum(isNeutralSlide) as neutralSlideTimes,
  sum(neutralSlideTime) as neutralSlideTimeTotal,
  sum(isOverspeed) as overspeedTimes,
  sum(overspeedTime) as overspeedTimeTotal,
  sum(isFatigueDriving) as fatigueDrivingTimes,
  sum(isHthrottleStop) as hthrottleStopTimes,
  sum(isOilLeak) as oilLeakTimes
from
  event_records
where
  time >= "2017-01-01 00:00:00"
  and time <= "2017-02-01 00:00:00"
group by
  driverID,
  carNumber
order by
  rapidlySpeedupTimes desc,
  rapidlySlowdownTimes desc,
  neutralSlideTimes desc,
  neutralSlideTimeTotal desc,
  overspeedTimes desc,
  overspeedTimeTotal desc,
  fatigueDrivingTimes desc,
  hthrottleStopTimes desc,
  oilLeakTimes desc
```

在查询结果中，单击“结果图形化”：

- “图形类型”选择“柱状图”
- “X轴”选择“driverID”
- “Y轴”选择“rapidlySpeedupTimes”
- “结果数目”选择“10”

展示结果如下：

图 3-6 急加速



b. 执行以下SQL语句，查询某个司机在某个时间段的详细记录。

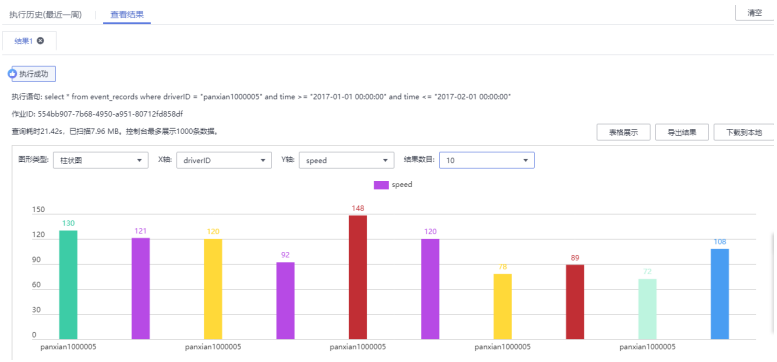
```
select
  *
from
  event_records
where
  driverID = "panxian1000005"
  and time >= "2017-01-01 00:00:00"
  and time <= "2017-02-01 00:00:00"
```

在查询结果中，单击 “结果图形化”：

- “图形类型”选择“柱状图”
- “X轴”选择“driverID”
- “Y轴”选择“speed”
- “结果数目”选择“10”

展示结果如下：

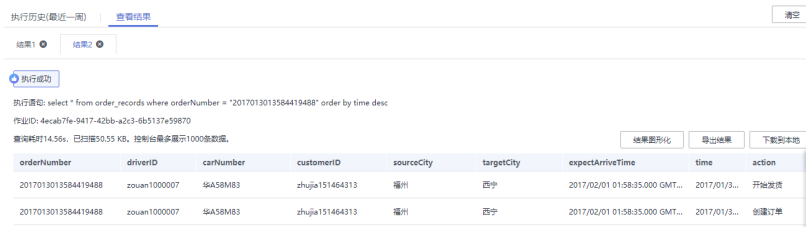
图 3-7 超速记录



c. 执行以下SQL语句，查询订单信息。

```
select
  *
from
  order_records
where
  orderNumber = "2017013013584419488"
order by
  time desc
```

图 3-8 订单信息



d. 执行以下SQL语句，根据司机和发车时间信息查询司机的详细行驶特点。

```
select
  driverID,
  carNumber,
  latitude,
  longitude,
```



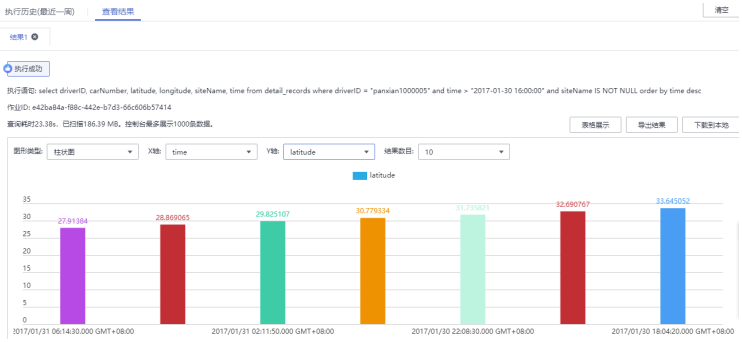
```
  siteName,
  time
from
  detail_records
where
  driverID = "panxian1000005"
  and time > "2017-01-30 16:00:00"
  and siteName IS NOT NULL
order by
  time desc
```

在查询结果中，单击“结果图形化”：

- “图形类型”选择“柱状图”
- “X轴”选择“time”
- “Y轴”选择“latitude”
- “结果数目”选择“10”

展示结果如下：

图 3-9 行驶信息



3.2 使用 DLI 将 CSV 数据转换为 Parquet 数据

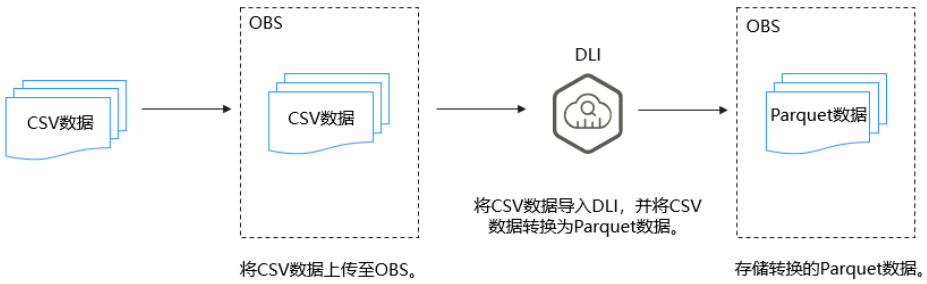
应用场景

Parquet是面向分析型业务的列式存储格式，这种格式可以加快查询速度，查询Parquet格式数据时，只检查所需要的列并对它们的值执行计算，也就是说，只读取一个数据文件或表的一小部分数据。Parquet还支持灵活的压缩选项，因此可以显著减少磁盘上的存储。使用DLI可轻松将CSV格式数据转换为Parquet格式数据。

方案架构

将CSV格式的数据上传到对象存储服务OBS，使用DLI将CSV数据转换为Parquet数据，并将转换后的Parquet数据存储到OBS中。

图 3-10 方案简介



流程指导

使用DLI将CSV数据转换为Parquet数据主要包括以下步骤：

步骤1：创建并上传数据。将数据上传到对象存储服务OBS。

步骤2：使用DLI将CSV数据转换为Parquet数据。将CSV数据导入DLI，并将CSV数据转换为Parquet数据。

方案优势

- **提升查询性能**

  如果您在HDFS上拥有基于文本的数据文件或者表，而且正在使用Spark SQL对数据执行查询操作，那么推荐将文本数据文件转换为Parquet数据文件，转换需要时间，但查询性能的提升在某些情况下可能达到约30倍或更高。
- **节省存储空间**

  Parquet还支持灵活的压缩选项，因此可以显著减少磁盘上的存储。存储的节省可高达约75%。

资源和成本规划

表 3-4 资源和成本规划

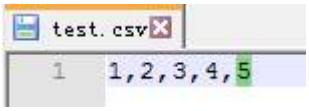
资源	资源说明	成本说明
OBS	需要创建一个OBS桶将数据上传到对象存储服务OBS，为后面使用DLI完成数据分析做准备。	OBS的使用涉及以下几项费用： • 存储费用：静态网站文件存储在OBS中产生的**存储费用**。 • 请求费用：用户访问OBS中存储的静态网站文件时产生的**请求费用**。 • 流量费用：用户使用自定义域名通过公网访问OBS时产生的**流量费用**。 实际产生的费用与存储的文件大小、用户访问所产生的请求次数和流量大小有关，请根据自己的业务进行预估。

资源	资源说明	成本说明
DLI	在创建SQL作业前需购买队列，使用DLI的队列资源时，按照队列CU时进行计费。	如购买按需计费的队列，在使用队列资源时，按照队列CU时进行计费。 以小时为单位进行结算。不足一小时按一小时计费，小时数按整点计算。队列CU时按需计费的计算费用=单价*CU数*小时数。

步骤 1：创建并上传数据

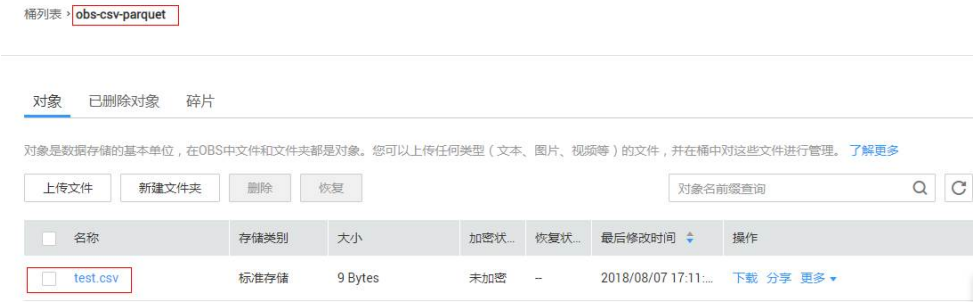
1. 创建CSV数据，例如，如图3-11所示test.csv：

   图 3-11 创建 test.csv 文件

2. 在OBS上建桶obs-csv-parquet，并将test.csv文件上传至OBS，如图3-12所示：

   图 3-12 上传 CSV 数据至 OBS

3. 在OBS上创建一个新的桶obs-parquet-data用于存储转换的Parquet数据。

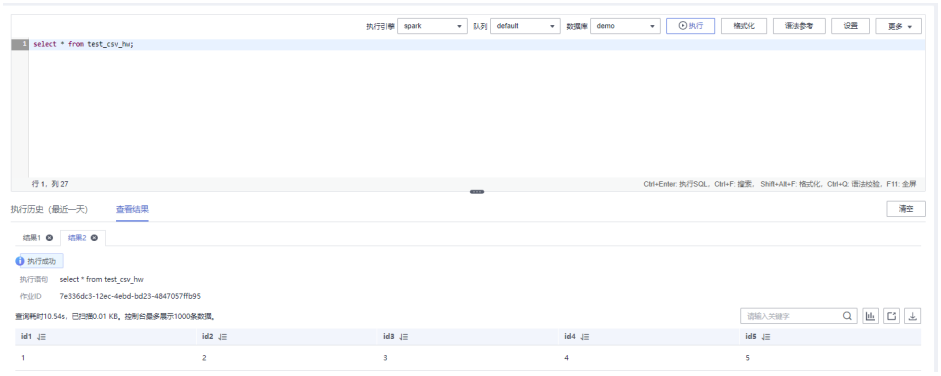
步骤 2：使用 DLI 将 CSV 数据转换为 Parquet 数据

1. 在DLI控制台总览页面左侧，单击“SQL编辑器”，进入SQL作业编辑器页面。
2. 在SQL作业编辑器左侧，选择“数据库”页签，单击创建名字为demo的数据库。
3. 在DLI的SQL编辑窗口，执行引擎选择“spark”，队列选择“default”，数据库选择为“demo”。输入以下建表语句，创建OBS表test_csv_hw并导入test.csv数据。

```
create table test_csv_hw(id1 int, id2 int, id3 int, id4 int, id5 int)
 using csv
 options(
 path 'obs://obs-csv-parquet/test.csv'
 )
```

4. 在DLI的SQL编辑窗口，执行以下语句可以查询表test_csv_hw中的数据。

图 3-13 查询表 test_csv_hw



5. 在DLI的SQL编辑窗口中创建OBS表test_parquet_hw。

```
create table `test_parquet_hw` (`id1` INT, `id2` INT, `id3` INT, `id4` INT, `id5` INT)
using parquet
options (
path 'obs://obs-parquet-data/'
)
```

说明

不需要指明具体的文件，因为在将数据从CSV格式转换为Parquet格式之前，不存在任何Parquet文件。

6. 在DLI的SQL编辑窗口中将CSV数据转换为Parquet数据并存储在OBS中。

```
insert into test_parquet_hw select * from test_csv_hw
```

7. 检查结果，如图3-14所示，系统自动创建了一个文件用于保存结果。

图 3-14 保存 Parquet 数据

3.3 使用 DLI 进行电商 BI 报表分析

应用场景

某商城作为中国一家自营式电商，在保持高速发展的同时，沉淀了数亿的忠实用户，积累了海量的真实数据。如何利用BI工具从历史数据中找出商机，是大数据应用在精准营销中的关键问题，也是所有电商平台在做智能化升级时所需要的核心技术。

本案例以某商城真实的用户、商品、评论数据（脱敏后）为基础，利用华为云数据湖探索、数据仓库服务以及永洪BI来分析用户和商品的各种数据特征，可为营销决策、广告推荐、信用评级、品牌监控、用户行为预测提供高质量的信息。

流程指导

使用DLI进行电商数据分析的操作过程主要包括以下步骤：

步骤1：上传数据。将数据上传到对象存储服务OBS，为后面使用DLI完成数据分析做准备。

步骤2：分析数据。使用DLI对待分析的数据进行查询。

数据说明

为保护用户的隐私和数据安全，所有数据均已进行了采样和脱敏。

- 用户数据

  表 3-5 用户数据

字段名称	字段类型	字段说明	取值范围
user_id	int	用户ID	脱敏
age	int	年龄段	-1表示未知
gender	int	性别	• 0表示男 • 1表示女 • 2表示保密
rank	Int	用户等级	有顺序的级别枚举，越高级别数字越大
register_time	string	用户注册日期	单位：天

- 商品数据

  表 3-6 商品数据

字段名称	字段类型	字段说明	取值范围
product_id	int	商品编号	脱敏
a1	int	属性1	枚举，-1表示未知
a2	int	属性2	枚举，-1表示未知
a3	int	属性3	枚举，-1表示未知
category	int	品类ID	脱敏
brand	int	品牌ID	脱敏

- 评价数据

表 3-7 评价数据

字段名称	字段类型	字段说明	取值范围
deadline	string	截止时间	单位：天
product_id	int	商品编号	脱敏
comment_num	int	累计评论数分段	• 0表示无评论 • 1表示有1条评论 • 2表示有2-10条评论 • 3表示有11-50条评论 • 4表示大于50条评论
has_bad_comment	int	是否有差评	0表示无，1表示有
bad_comment_rate	float	差评率	差评数占总评论数的比重

- 行为数据

表 3-8 行为数据

字段名称	字段类型	字段说明	取值范围
user_id	int	用户编号	脱敏
product_id	int	商品编号	脱敏
time	string	行为时间	-
model_id	string	模块编号	脱敏
type	string	• 浏览（指浏览商品详情页） • 加入购物车 • 购物车删除 • 下单 • 关注 • 点击	-

步骤 1：上传数据

将数据上传到对象存储服务OBS，为后面使用DLI完成数据分析做准备。

1. 下载OBS Browser+。下载地址请参考《**对象存储服务工具指南**》。
2. 安装OBS Browser+。安装步骤请参考《**对象存储服务工具指南**》。

3. 登录OBS Browser+。OBS Browser+支持AK方式登录，以及授权码登录两种登录方式。登录步骤请参考《**对象存储服务工具指南**》。
4. 通过OBS Browser+上传数据。

   在OBS Browser+页面单击“创建桶”，按照要求选择“区域”和填写“桶名”（例如：DLI-demo），创建桶成功后，返回桶列表，单击桶DLI-demo。OBS Browser+提供强大的拖拽上传功能，您可以将本地的一个或多个文件或者文件夹拖拽到对象存储的对象列表或者并行文件系统的对象列表中；同时您也可以将文件或文件夹拖拽到指定的目录上，这样可以上传到指定的目录中。

   单击**Best_Practice_04.zip**获取本示例的测试数据，解压“Best_Practice_04.zip”压缩包，解压后将data文件夹上传到OBS桶根目录下。测试数据目录说明如下：

   - user表数据：data/JData_User
   - product表数据：data/JData_Product
   - comment表数据：data/JData_Product/JData_Comment
   - action表数据：data/JData_Action

步骤 2：分析数据

1. 创建数据库、表

   a. 在portal页面上方菜单栏中单击“产品”，单击“大数据”分类中的“数据湖探索 DLI”。

   b. 创建demo数据库，在DLI控制台总览页面，选择“作业管理 > SQL作业”，单击“创建作业”，进入SQL作业编辑器。

   c. 在SQL作业编辑器左侧，选择“数据库”页签，单击创建demo数据库，请参见**图3-15**。

   图 3-15 创建数据库

> **说明**
>
> “default”为内置数据库，不能创建名为“default”的数据库。

d. 选择demo数据库，在编辑框中输入以下SQL语句：

```
create table user(
  user_id int,
  age int,
  gender int,
  rank int,
  register_time string
) USING csv OPTIONS (path "obs://DLI-demo/data/JData_User")
```

> **说明**
>
> 上述SQL语句中的文件路径为实际存放数据的OBS路径。

e. 单击“执行”，创建用户信息表user。

f. 用相同的方法创建商品表，评价表，行为表。

- 商品表

```
create table product(
  product_id int,
  a1 int,
  a2 int,
  a3 int,
  category int,
  brand int
) USING csv OPTIONS (path "obs://DLI-demo/data/JData_Product")
```

- 评价表

```
create table comment(
  deadline string,
  product_id int,
  comment_num int,
  has_bad_comment int,
  bad_comment_rate float
) USING csv OPTIONS (path "obs://DLI-demo/data/JData_Comment")
```

- 行为表

```
create table action(
  user_id int,
  product_id int,
  time string,
  model_id string,
  type string
) USING csv OPTIONS (path "obs://DLI-demo/data/JData_Action");
```

2. 执行查询

常用查询语句可以设置为模板，下次查询的时候在模板管理页面可以查看，具体操作可参见《数据湖探索用户指南》中的**《模板管理》**。

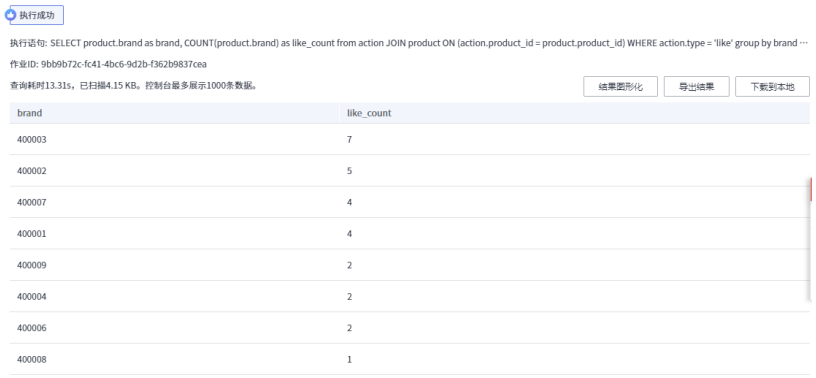
- 分析出10大用户点赞数最多的产品

i. 执行以下SQL语句，可以分析出10大用户点赞数最多的产品。

```
SELECT
  product.brand as brand,
  COUNT(product.brand) as like_count
from
  action
  JOIN product ON (action.product_id = product.product_id)
WHERE
  action.type = 'like'
group by
  brand
ORDER BY like_count desc
limit
  10
```

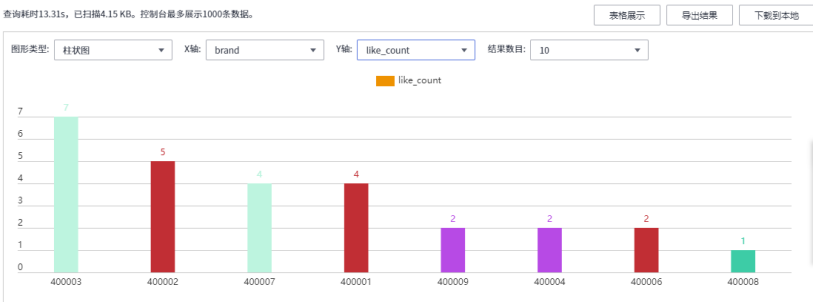

ii. 单击“执行”，运行结果如图3-16所示：

图 3-16 查询结果



iii. 单击  “结果图形化”，对结果进行图形展示：

图 3-17 结果图形化



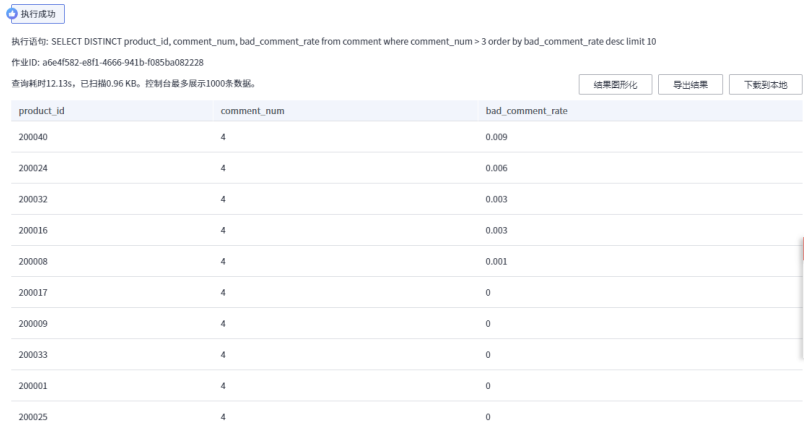
– 分析出10大评级最差的商品

i. 执行以下SQL语句，可以分析出10大评级最差的商品。

```
SELECT
  DISTINCT product_id,
  comment_num,
  bad_comment_rate
from
  comment
where
  comment_num > 3
order by
  bad_comment_rate desc
limit
  10
```

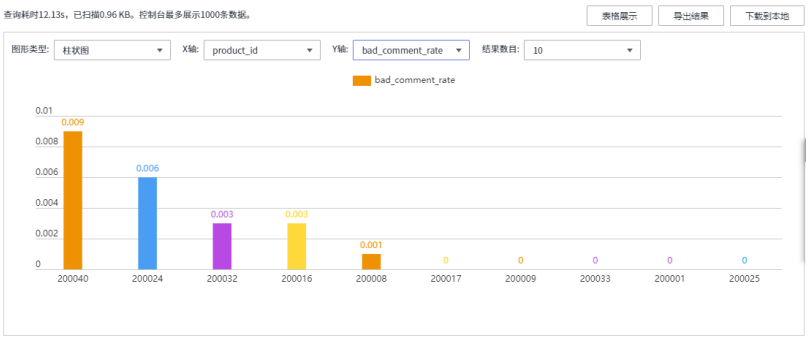
ii. 单击“执行”，运行结果如图3-18所示：

图 3-18 查询结果



iii. 单击  “结果图形化”，对结果进行图形展示：

图 3-19 结果图形化



此外，还可以分析用户的年龄分布、性别比例、商品评价情况、购买情况、浏览情况等。

3.4 使用 DLI 进行账单分析与优化

应用场景

本文主要介绍如何使用华为云DLI上的实际消费数据（文中涉及账户的信息已脱敏），在DLI的大数据分析平台上进行分析，找出费用优化的空间，并给出使用DLI过程中降低成本的一些优化措施。

流程介绍

使用DLI进行账单分析与优化的操作过程主要包括以下步骤：

步骤1：获取消费数据。获取账户的实际消费数据。

步骤2：分析账户消费结构并优化。在DLI上分析账户消费结构，找出开支较大的资源或用户，并给出降底成本的优化措施。

资源和成本规划

表 3-9 资源和成本规划

资源	资源说明	成本说明
DLI	数据湖探索（DLI）作为华为云大数据分析平台，其计费项包括存储费用与计算费用两项，计费类型包括包周期（包年包月），套餐包和按需计费三种。	DLI目前支持三种作业：SQL作业，Flink作业和Spark作业。 SQL作业的计费包括存储计费和计算计费，其中计算计费有包年包月计费和按需计费两种。 ● 包年包月计费根据购买周期进行扣费，推荐使用包年包月模式，价格优惠且在周期内独享计算资源。 ● 按需计费以小时为单位进行扣费。按需计费又分为按CU时计费和按扫描数据量计费，这两种计费方式是互斥的，可根据需要选择其中一种。建议优先选择按CU时计费，可资源独享，且成本核算清晰。同时，按CU时计费还提供套餐包的购买和使用。 – CU时资费=CU数*使用时长*单价。使用时长按自然小时计费，不足一个小时按一个小时计费。 – 扫描数据量资费=执行SQL时产生的扫描数据量*单价。如果计算任务超时或失败，则当次计算不收取费用。 ● Flink作业和Spark作业的计费只有计算计费，具体计费规则与SQL作业相同。 具体计费规则可以参考华为云官网**价格详情**。 。

步骤 1：获取消费数据

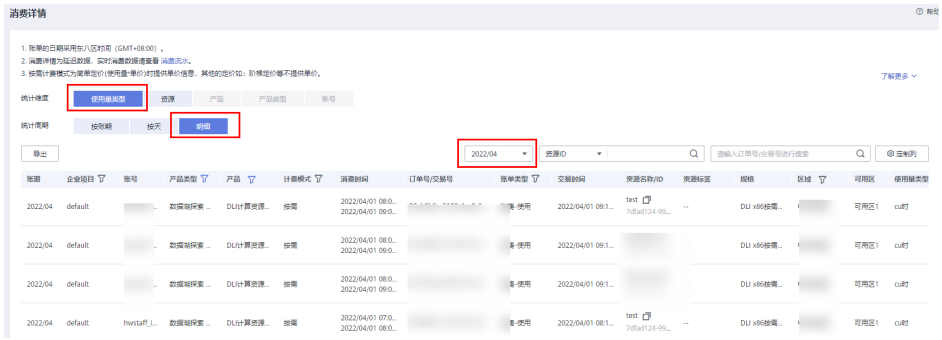
1. 获取消费明细数据。
   a. 使用华为云账户登录控制台。
   b. 通过“费用与成本”>“费用账单”进入费用中心。

图 3-20 费用账单

c. 在费用中心的“账单管理”界面，单击“消费详情”，统计维度选择“使用量类型”，统计周期选择“明细”，时间选择对应账期时间。

在显示数据的标题行，“产品类型”搜索并选择“数据湖探索 DLI”，“产品”搜索并选择“DLI计算资源使用量”，单击“导出”。在导出界面根据需要选择导出的时间和数据范围，单击“导出”跳转到导出记录界面。

图 3-21 费用汇总



d. 在导出记录界面，等待文件状态变为“文件生成完成”后，单击“下载”完成文件下载。

步骤 2：分析账户消费结构并优化

1. 在DLI上进行消费明细分析。

   a. 将**步骤1：获取消费数据**下载的消费明细数据上传到已建好的OBS桶中。

   b. 在数据湖探索服务中创建表。

      i. 登录DLI控制台，左侧导航栏单击“SQL编辑器”，执行引擎选择“spark”，选择执行的队列和数据库。本次演示队列和数据库选择“default”。

      ii. 下载的文件中包含时间用量等，按表头意义在DLI上创建表，具体可以参考如下示例，其中amount列为费用。

```
CREATE TABLE `spending` (
  account_period string,
  EnterpriseProject string,
  EnterpriseProjectID string,
  accountID string,
```

```
product_type_code string,
product_type string,
product_code string,
product_name string,
product_id string,
mode string,
time1 string,
use_start string,
use_end string,
orderid string,
ordertime string,
resource_type string,
resource_id string,
resouce_name string,
tag string,
skuid string,
`c22name` STRING,
`c23name` STRING,
`c24name` STRING,
`c25name` STRING,
`c26name` STRING,
`c27name` STRING,
`c28name` STRING,
`c29name` STRING,
size STRING,
`c31name` STRING,
`c32name` STRING,
`c33name` STRING,
`c34name` STRING,
`c35name` STRING,
`amount` STRING,
`c37name` STRING,
`c38name` STRING,
`c39name` STRING,
`c40name` STRING,
`c41name` STRING,
`c42name` STRING,
`c43name` STRING,
`c44name` STRING,
`c45name` STRING,
`c46name` STRING,
`c47name` STRING,
`c48name` STRING,
`c49name` STRING,
`c50name` STRING,
`c51name` STRING,
`c52name` STRING,
`c53name` STRING,
`c54name` STRING
) USING csv options (
path 'obs://xxx/Spendings(ByTransaction)_20200501_20200531.csv',
header true)
```

c. 查询该时间内消费最高的resource_id, resource_name。

通过以下语句，可以发现sql和flink队列使用的费用均为1842元，在总费用3754元中占比为98%。

```
select resource_id, resouce_name, sum(size)
  as usage, sum(amount)
  as sum_amount
  from spending
  group by resource_id, resouce_name
  order by sum_amount desc
```


图 3-22 查询结果

resource_id	resouce_name	usage	sum_amount
d91d4616-b10c-471a-820d-e676e6c5f4b4	sql	5264	1842.3999999999896
8163cc27-89ce-4bac-aa85-38cb753ee425	flink	5264	1842.3999999999896
9bd0736b-f8ca-4bfb-b3e7-0e391ef70d8b	null	48	14.399999999999999
dd3a12ff-c0af-4ad1-bbc1-8588f4d3661c	dlitest	32	11.2
f8265ef5-eb5f-4eff-b8d6-9ca91ed20009	test	16	5.6

d. 使用以下语句具体分析sql和flink这两个资源消费的时间段。

```
select * from spending where resource_id = 'd91d4616-b10c-471a-820d-e676e6c5f4b4' order by ordertime
```

可以发现sql队列从2020-05-14 17:00:00 GMT+08:00开始，每小时产生5.6元费用，持续到2020-05-28 10:00:00 GMT+08:00，说明这个sql队列在这段时间内持续使用。

同样，也可以发现flink队列在2020-05-14 17:00:00 GMT+08:00到2020-05-28 10:00:00 GMT+08:00这段 内持续使用。

2. 优化建议。

通过以上分析，了解到sql和flink这两个队列几乎是在持续使用的，建议通过购买包周期队列来降低使用成本。另外，对于明确需要使用多少CU时的作业，也可以提前购买对应的CU时套餐包，来降低使用成本。

企业中的业务模式较多且经常变化，成本管理员通常并不能全面及时了解花销较大的业务在哪里，哪些是合理的，哪些是不合理的，通过在DLI中对费用明细进行分析，可以及时发现企业花销不合理的地方，及时进行成本管理，进一步降低企业使用华为云的成本。

3.5 使用 DLI Flink SQL 进行电商实时业务数据分析

应用场景

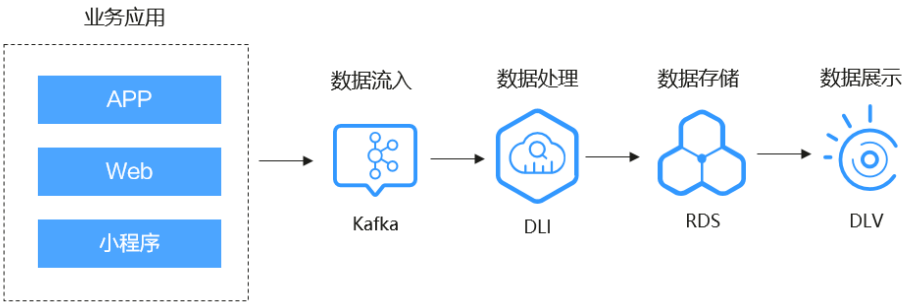
当前线上购物无疑是最火热的购物方式，而电商平台则又可以以多种方式接入，例如通过web方式访问、通过app的方式访问、通过微信小程序的方式访问等等。而电商平台则需要每天统计各平台的实时访问数据量、订单数、访问人数等等指标，从而能在显示大屏上实时展示相关数据，方便及时了解数据变化，有针对性地调整营销策略。而如何高效快捷地统计这些指标呢?

假设平台已经将每个商品的订单信息实时写入Kafka中，这些信息包括订单ID、订单生成的渠道(即web方式、app方式等)、订单时间、订单金额、折扣后实际支付金额、支付时间、用户ID、用户姓名、订单地区ID等信息。而我们需要做的，就是根据当前可以获取到的业务数据，实时统计每种渠道的相关指标，输出存储到数据库中，并进行大屏展示。

方案架构

使用DLI Flink完成电商业务实时数据的分析处理，获取各个渠道的销售汇总数据。

图 3-23 方案简介



流程指导

使用DLI Flink进行电商实时业务数据分析的操作过程主要包括以下步骤：

步骤1：创建资源。在您的账户下创建作业需要的相关资源，涉及VPC、DMS、DLI、RDS。

步骤2：获取DMS连接地址并创建Topic。获取DMS Kafka实例连接地址并创建DMS Topic。

步骤3：创建RDS数据库表。获取RDS实例内网地址，登录RDS实例创建RDS数据库及MySQL表。

步骤4：创建DLI增强型跨源。创建DLI增强型跨源，并测试队列与RDS、DMS实例连通性。

步骤5：创建并提交Flink作业。创建DLI Flink OpenSource SQL作业并运行。

步骤6：查询结果。查询Flink作业结果，使用DLV进行大屏展示。

方案优势

- 跨源分析：数据免搬迁，就可以关联分析存在OBS中的各个渠道的销售汇总数据。
- 纯SQL操作：DLI已对接多个数据源，直接通过SQL建表就可以完成数据源的映射。

资源和成本规划

表 3-10 资源和成本规划

资源	资源说明	成本说明
OBS	需要创建一个OBS桶将数据上传到对象存储服务OBS，为后面使用DLI完成数据分析做准备。	OBS的使用涉及以下几项费用： • 存储费用：静态网站文件存储在OBS中产生的**存储费用**。 • 请求费用：用户访问OBS中存储的静态网站文件时产生的**请求费用**。 • 流量费用：用户使用自定义域名通过公网访问OBS时产生的**流量费用**。 实际产生的费用与存储的文件大小、用户访问所产生的请求次数和流量大小有关，请根据自己的业务进行预估。
DLI	在创建SQL作业前需购买队列，使用DLI的队列资源时，按照队列CU时进行计费。	如购买按需计费的队列，在使用队列资源时，按照队列CU时进行计费。 以小时为单位进行结算。不足一小时按一小时计费，小时数按整点计算。队列CU时按需计费的计算费用=单价*CU数*小时数。
VPC	VPC丰富的功能帮助您灵活管理云上网络，包括创建子网、设置安全组和网络ACL、管理路由表、申请弹性公网IP和带宽等。	VPC本身不收取费用。 但如有互联网访问需求，您需要购买弹性公网IP。弹性公网IP提供“包年/包月”和“按需计费”两种计费模式。 了解**VPC计费说明**。
DMS Kafka	Kafka提供的消息队列服务，向用户提供计算、存储和带宽资源独占式的Kafka专享实例。	Kafka版支持按需和包周期两种付费模式。Kafka计费项包括Kafka实例和Kafka的磁盘存储空间。 了解**Kafka计费说明**。
RDS MySQL	数据库 RDS for MySQL提供在线云数据库服务。	RDS对您选择的数据库实例、数据库存储和备份存储（可选）收费。 了解**RDS计费说明**。
DLV	DLV适配云上云下多种数据源，提供丰富多样的可视化组件，快速定制数据大屏。	使用DLV服务的费用主要是DLV包年包月套餐的费用，您可以根据实际使用情况，选择合适的版本规格。

数据说明

- 数据源表：电商业务订单详情宽表

字段名	字段类型	说明
order_id	*string*	订单ID
order_channel	*string*	订单生成的渠道(即web方式、app方式等)
order_time	*string*	订单时间
pay_amount	*double*	订单金额
real_pay	*double*	实际支付金额
pay_time	*string*	支付时间
user_id	*string*	用户ID
user_name	*string*	用户姓名
area_id	*string*	订单地区ID

- 结果表：各渠道的销售总额实时统计表。

字段名	字段类型	说明
begin_time	*varchar(32)*	开始统计指标的时间
channel_code	*varchar(32)*	渠道编号
channel_name	*varchar(32)*	渠道名
cur_gmv	*double*	当天GMV
cur_order_user_count	*bigint*	当天付款人数
cur_order_count	*bigint*	当天付款订单数
last_pay_time	*varchar(32)*	最近结算时间
flink_current_time	*varchar(32)*	Flink数据处理时间

步骤 1：创建资源

如**表3-11**所示，完成VPC、DMS、RDS、DLI、DLV资源的创建。

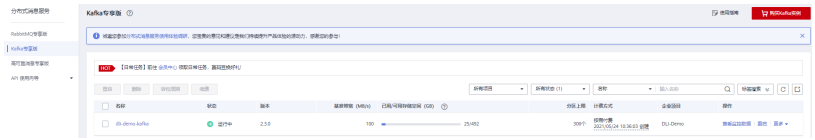
表 3-11 创建资源

资源类型	说明	操作指导
VPC	VPC为资源提供云上的网络管理服务。 **资源网络规划说明：** • Kafka与MySQL实例指定的VPC需为同一VPC。 • Kafka与MySQL实例所属VPC网段不得与创建的DLI队列网段冲突。	创建VPC和子网
DMS Kafka	本例中以DMS Kafka实例作为数据源。	DMS Kafka入门指引
RDS MySQL	本例中以使用RDS提供在线云数据库服务。	RDS MySQL快速入门
DLI	DLI提供实时业务数据分析。 创建DLI队列时请创建“包年包月”或者“按需-专属资源”模式的通用队列，否则无法创建增强型网络连接。	DLI 创建队列
DLV	DLV实时展现DLI队列处理后的结果数据。	DLV 创建大屏

步骤 2：获取 DMS 连接地址并创建 Topic

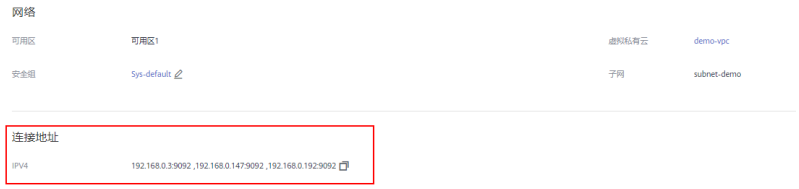
1. 在控制台单击“服务列表”，选择“分布式消息服务DMS”，单击进入DMS服务控制台页面。在“Kafka专享版”页面找到您所创建的Kafka实例。

图 3-24 Kafka 实例



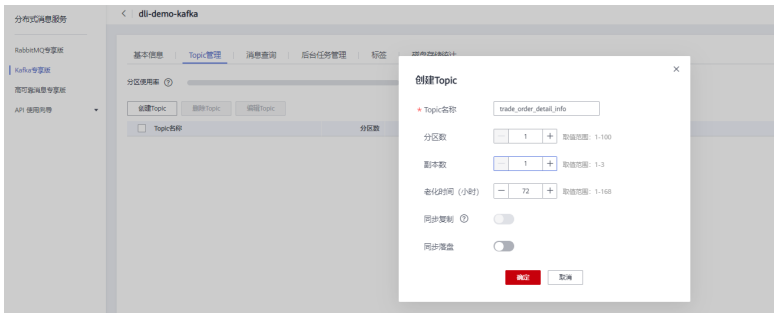
2. 进入实例详情页面。单击“基本信息”，获取“连接地址”。

图 3-25 获取连接地址



3. 单击“Topic管理”，创建一个Topic：trade_order_detail_info。

图 3-26 创建 Topic



Topic配置如下：

- 分区数：1
- 副本数：1
- 老化时间：72h
- 同步落盘：否

步骤 3：创建 RDS 数据库表

1. 在控制台单击“服务列表”，选择“云数据库RDS”，单击进入RDS页面。在“实例管理页面”，找到您已经创建的RDS实例，获取其内网地址。

图 3-27 内网地址

2. 单击所创建RDS实例的“登录”，跳转至“数据管理服务-DAS”。输入相关账户信息，单击“测试连接”。显示连接成功后，单击“登录”，进入“实例登录”页面。

图 3-28 实例登录

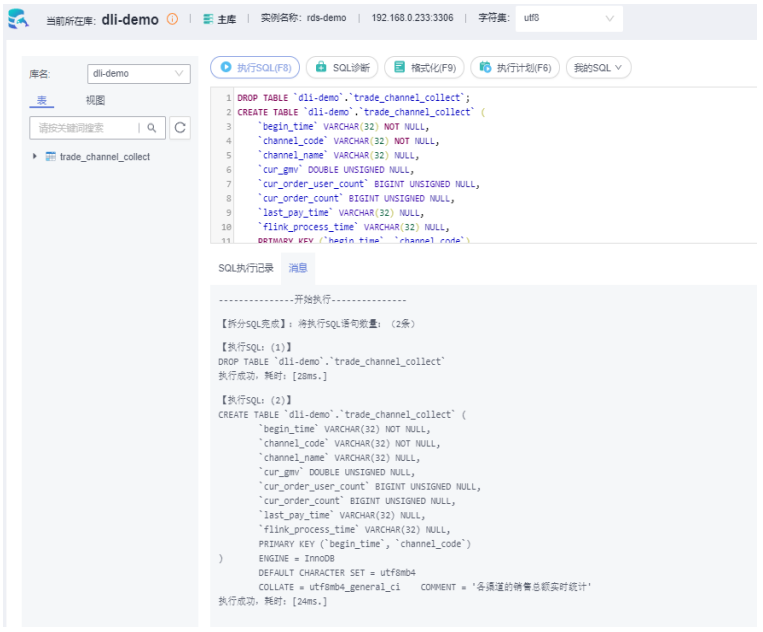
3. 登录RDS实例后，单击“新建数据库”，创建名称为“dli-demo”的数据库。

图 3-29 创建数据库

4. 单击“SQL操作”>“SQL查询”，执行如下SQL创建测试用MySQL表，表相关字段含义在**•数据说明**中有详细介绍。

```
DROP TABLE `dli-demo`.`trade_channel_collect`;
CREATE TABLE `dli-demo`.`trade_channel_collect` (
    `begin_time` VARCHAR(32) NOT NULL,
    `channel_code` VARCHAR(32) NOT NULL,
    `channel_name` VARCHAR(32) NULL,
    `cur_gmv` DOUBLE UNSIGNED NULL,
    `cur_order_user_count` BIGINT UNSIGNED NULL,
    `cur_order_count` BIGINT UNSIGNED NULL,
    `last_pay_time` VARCHAR(32) NULL,
    `flink_current_time` VARCHAR(32) NULL,
    PRIMARY KEY (`begin_time`, `channel_code`)
)   ENGINE = InnoDB
    DEFAULT CHARACTER SET = utf8mb4
    COLLATE = utf8mb4_general_ci
    COMMENT = '各渠道的销售总额实时统计';
```

图 3-30 创建表



步骤 4：创建 DLI 增强型跨源

1. 在控制台单击“服务列表”，选择“数据湖探索”，单击进入DLI服务页面。单击“资源管理 > 队列管理”，查询创建的DLI队列。

图 3-31 队列列表

2. 单击“全局配置 > 服务授权”，选中“VPC Administrator”，单击“更新委托权限”，赋予DLI操作用户VPC资源的权限，用于创建VPC的“对等连接”。

图 3-32 更新委托权限

3. 单击“跨源连接 > 增强型跨源 > 创建”，配置如下连接信息后单击“确定”。
   - 连接名称：增强型跨源名称。
   - 弹性资源池：选择您所创建的通用队列。
   - 虚拟私有云：选择 Kafka 与 MySQL 实例所在的VPC。
   - 子网：选择 Kafka 与 MySQL 实例所在的子网。

图 3-33 创建增强型跨源

增强型跨源创建完成后，在跨源列表中，对应的跨源连接状态会显示为“已激活”。

单击跨源连接的名称，详情页面显示连接状态为“ACTIVE”。

图 3-34 跨源连接状态

图 3-35 详情

4. 测试队列与RDS、DMS实例连通性。

   a. 单击“队列管理”，选择您所使用的队列，单击“操作”列中的“更多”>“测试地址连通性”。

图 3-36 检测地址连通性

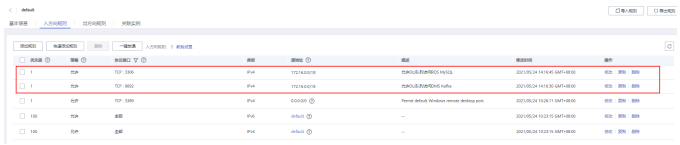
b. 输入DMS Kafka实例连接地址和步RDS MySQL实例内网地址，进行网络连通性测试。

测试结果显示可达，则DLI队列与Kafka、MySQL实例的网络已经联通。

图 3-37 测试结果

如果测试结果不可达，需要修改实例所在VPC的安全组规则，放开9092、3306端口对DLI队列的限制，DLI队列网段信息可以在队列的详情页中获取。

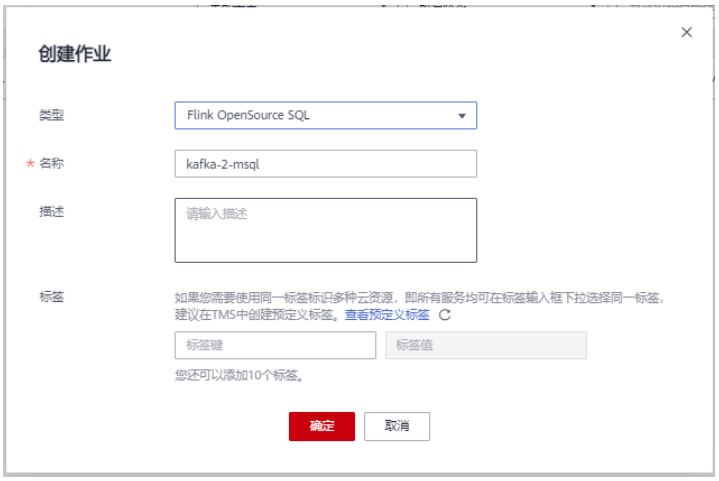
图 3-38 安全组规则



步骤 5：创建并提交 Flink 作业

1. 单击DLI控制台左侧“作业管理”，选择“Flink作业”。单击“创建作业”。
   - 类型：选择作业类型为：Flink OpenSource SQL。
   - 名称：自定义。

图 3-39 创建 Flink 作业



2. 单击“确定”，进入作业编辑作业页面，具体SQL示例如下，部分参数值需要根据RDS和DMS对应的信息进行修改。

```
--********************************************************************--
-- 数据源：trade_order_detail_info (订单详情宽表)
--********************************************************************--
create table trade_order_detail (
  order_id string,      -- 订单ID
  order_channel string,   -- 渠道
  order_time string,     -- 订单创建时间
  pay_amount double,     -- 订单金额
  real_pay double,      -- 实际付费金额
  pay_time string,      -- 付费时间
  user_id string,      -- 用户ID
  user_name string,     -- 用户名
  area_id string       -- 地区ID
) with (
  "connector.type" = "kafka",
  "connector.version" = "0.10",
  "connector.properties.bootstrap.servers" = "xxxx:9092,xxxx:9092,xxxx:9092", -- Kafka连接地址
  "connector.properties.group.id" = "trade_order",   -- Kafka groupID
  "connector.topic" = "trade_order_detail_info",     -- Kafka topic
  "format.type" = "json",
  "connector.startup-mode" = "latest-offset"
);

--********************************************************************--
-- 结果表：trade_channel_collect (各渠道的销售总额实时统计)
--********************************************************************--
create table trade_channel_collect(
  begin_time string,       --统计数据的开始时间
  channel_code string,      -- 渠道编号
  channel_name string,      -- 渠道名
  cur_gmv double,         -- 当天GMV
  cur_order_user_count bigint, -- 当天付款人数
  cur_order_count bigint,    -- 当天付款订单数
  last_pay_time string,     -- 最近结算时间
  flink_current_time string,
  primary key (begin_time, channel_code) not enforced
) with (
  "connector.type" = "jdbc",
  "connector.url" = "jdbc:mysql://xxxx:3306/xxxx",    -- mysql连接地址，jdbc格式
  "connector.table" = "xxxx",             -- mysql表名
  "connector.driver" = "com.mysql.jdbc.Driver",
  'pwd_auth_name'= '***xxxxx***', --DLI侧创建的Password类型的跨源认证名称。使用跨源认证则无需在作业中配置账号和密码。
  "connector.write.flush.max-rows" = "1000",
  "connector.write.flush.interval" = "1s"
);
```



```
--********************************************************************--
-- 临时中间表
--********************************************************************--
create view tmp_order_detail
as
select *
    , case when t.order_channel not in ("webShop", "appShop", "miniAppShop") then "other"
           else t.order_channel end as channel_code --重新定义统计渠道 只有四个枚举值[webShop、
appShop、miniAppShop、other]
    , case when t.order_channel = "webShop" then _UTF16"网页商城"
           when t.order_channel = "appShop" then _UTF16"app商城"
           when t.order_channel = "miniAppShop" then _UTF16"小程序商城"
           else _UTF16"其他" end as channel_name --渠道名称
from (
    select *
        , row_number() over(partition by order_id order by order_time desc ) as rn --去除重复订单数据
        , concat(substr("2021-03-25 12:03:00", 1, 10), " 00:00:00") as begin_time
        , concat(substr("2021-03-25 12:03:00", 1, 10), " 23:59:59") as end_time
    from trade_order_detail
    where pay_time >= concat(substr("2021-03-25 12:03:00", 1, 10), " 00:00:00") --取今天数据，为了方
便运行，这里使用"2021-03-25 12:03:00"替代cast(LOCALTIMESTAMP as string)
    and real_pay is not null
) t
where t.rn = 1;

-- 按渠道统计各个指标
insert into trade_channel_collect
select
      begin_time  --统计数据的开始时间
    , channel_code
    , channel_name
    , cast(COALESCE(sum(real_pay), 0) as double) as cur_gmv --当天GMV
    , count(distinct user_id) as cur_order_user_count --当天付款人数
    , count(1) as cur_order_count --当天付款订单数
    , max(pay_time) as last_pay_time --最近结算时间
    , cast(LOCALTIMESTAMP as string) as flink_current_time --flink任务中的当前时间
from tmp_order_detail
where pay_time >= concat(substr("2021-03-25 12:03:00", 1, 10), " 00:00:00")
group by begin_time, channel_code, channel_name;
```

说明

作业逻辑说明如下：

1. 创建一个Kafka源表，用来从Kafka指定Topic中读取消费数据；
2. 创建一个结果表，用来通过JDBC向MySQL中写入结果数据。
3. 实现相应的处理逻辑，以实现各个指标的统计。

   为了简化最终的处理逻辑，使用创建视图进行数据预处理。

   1. 利用over窗口条件和过滤条件结合以去除重复数据（该方式是利用了top N的方法），同时利用相应的内置函数concat和substr将当天的00:00:00作为统计的开始时间，当天的23:59:59作为统计结束时间，并筛选出支付时间在当天凌晨00:00:00后的订单数据进行统计（**为了方便模拟数据的构造，这里使用"2021-03-25 12:03:00"替代cast(LOCALTIMESTAMP as string)**）。
   2. 根据这些数据的订单渠道利用内置的条件函数设置channel_code和channel_name的值，从而获取了源表中的字段信息，以及begin_time、end_time和channel_code、channel_name的值。
4. 根据需要对相应指标进行统计和筛选，并将结果写入到结果表中。

3. 选择所创建的DLI通用队列提交作业。

图 3-40 Flink Opensource SQL 作业

4. 等待作业状态会变为“运行中”，单击作业名称，可以查看作业详细运行情况。

图 3-41 作业运行状态

5. 使用Kafka客户端向指定topic发送数据，模拟实时数据流。
   具体方法请参考**DMS-连接实例生产消费信息**。

图 3-42 模拟实时数据流

6. 发送命令如下：

```
sh kafka_2.11-2.3.0/bin/kafka-console-producer.sh --broker-list KafKa连接地址 --topic Topic名称
```

示例数据如下：

```
{"order_id":"202103241000000001", "order_channel":"webShop", "order_time":"2021-03-24 10:00:00", "pay_amount":"100.00", "real_pay":"100.00", "pay_time":"2021-03-24 10:02:03", "user_id":"0001", "user_name":"Alice", "area_id":"330106"}
{"order_id":"20210324160606001", "order_channel":"appShop", "order_time":"2021-03-24 16:06:06", "pay_amount":"200.00", "real_pay":"180.00", "pay_time":"2021-03-24 16:10:06", "user_id":"0001", "user_name":"Alice", "area_id":"330106"}
{"order_id":"20210325120202001", "order_channel":"miniAppShop", "order_time":"2021-03-25 12:02:02", "pay_amount":"60.00", "real_pay":"60.00", "pay_time":"2021-03-25 12:03:00", "user_id":"0002", "user_name":"Bob", "area_id":"330110"}
{"order_id":"20210325150505001", "order_channel":"qqShop", "order_time":"2021-03-25 15:05:05", "pay_amount":"500.00", "real_pay":"400.00", "pay_time":"2021-03-25 15:10:00", "user_id":"0003", "user_name":"Cindy", "area_id":"330108"}
{"order_id":"20210325202020001", "order_channel":"webShop", "order_time":"2021-03-24 20:20:20", "pay_amount":"600.00", "real_pay":"480.00", "pay_time":"2021-03-25 00:00:00", "user_id":"0004", "user_name":"Daisy", "area_id":"330102"}
{"order_id":"20210326080808001", "order_channel":"webShop", "order_time":"2021-03-25 08:08:08", "pay_amount":"300.00", "real_pay":"240.00", "pay_time":"2021-03-25 08:10:00", "user_id":"0004", "user_name":"Daisy", "area_id":"330102"}
{"order_id":"20210326131130001", "order_channel":"webShop", "order_time":"2021-03-25 13:13:13", "pay_amount":"100.00", "real_pay":"100.00", "pay_time":"2021-03-25 16:16:16", "user_id":"0004", "user_name":"Daisy", "area_id":"330102"}
{"order_id":"20210327060606001", "order_channel":"appShop", "order_time":"2021-03-25 06:06:06", "pay_amount":"50.50", "real_pay":"50.50", "pay_time":"2021-03-25 06:07:00", "user_id":"0001", "user_name":"Alice", "area_id":"330106"}
{"order_id":"20210327060606002", "order_channel":"webShop", "order_time":"2021-03-25 06:06:06", "pay_amount":"66.60", "real_pay":"66.60", "pay_time":"2021-03-25 06:07:00", "user_id":"0002", "user_name":"Bob", "area_id":"330110"}
{"order_id":"20210327060606003", "order_channel":"miniAppShop", "order_time":"2021-03-25 06:06:06", "pay_amount":"88.80", "real_pay":"88.80", "pay_time":"2021-03-25 06:07:00", "user_id":"0003", "user_name":"Cindy", "area_id":"330108"}
{"order_id":"20210327060606004", "order_channel":"webShop", "order_time":"2021-03-25 06:06:06", "pay_amount":"99.90", "real_pay":"99.90", "pay_time":"2021-03-25 06:07:00", "user_id":"0004", "user_name":"Daisy", "area_id":"330102"}
```

7. 单击DLI控制台左侧“作业管理”>“Flink作业”，单击**3**提交的Flink作业。在作业详情页面，可以看到处理的数据记录数。

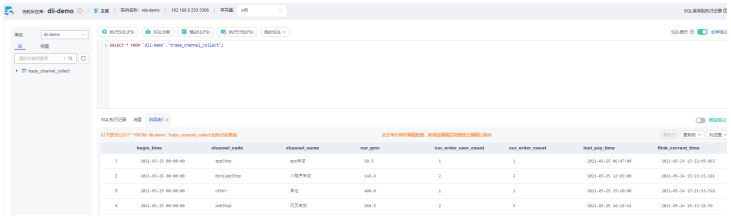
图 3-43 Flink 作业详情

步骤 6：查询结果

1. 参考2，登录MySQL实例，执行如下SQL语句，即可查询到经过Flink作业处理后的结果数据。

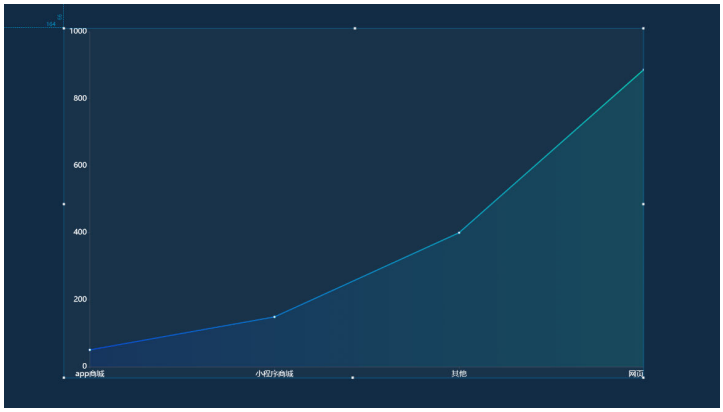
```
SELECT * FROM `dli-demo`.`trade_channel_collect`;
```

图 3-44 查询结果



2. 配置DLV大屏，执行SQL查询RDS MySQL，即可以实现大屏实时展示。具体配置方法可参考**DLV开发大屏**。

图 3-45 大屏展示



3.6 永洪 BI 对接 DLI 提交 Spark 作业

3.6.1 永洪 BI 对接准备工作

操作场景

永洪BI与DLI对接之前的准备工作。

操作步骤

步骤1 （可选）在公有云管理控制台上方的“服务列表”中选择“大数据”中的“数据湖探索”，单击右上角的“常用链接”下载DLI JDBC驱动（例如：dli-jdbc-1.1.0-jar-with-dependencies-jdk1.7.jar）。具体操作请参考**下载JDBC驱动包**。

步骤2 JDBC认证方式支持AK/SK方式和Token方式，建议采用AK/SK方式。

步骤3 询问永洪客服，获取永洪SaaS生产环境用户账号和密码。

步骤4 登入永洪SaaS生产环境，输入用户账号和密码。

----结束

3.6.2 永洪 BI 添加数据源

操作场景

在永洪SaaS生产环境中添加DLI的数据源。

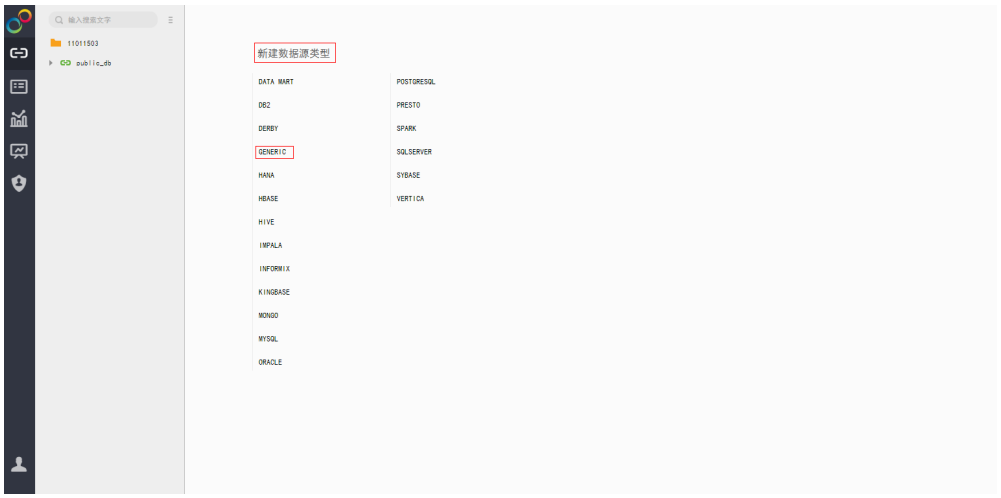
操作步骤

步骤1 在永洪SaaS生产环境主页，单击左侧导航栏中的“添加数据源”，请参见图3-46。

图 3-46 添加数据源

步骤2 “选择数据源类型”页面中，新建数据源类型选择“GENERIC”。请参见图3-47。

图 3-47 选择数据源类型



步骤3 添加数据源的相关配置，请参见图3-48。

“驱动”栏填写DLI JDBC的驱动：com.huawei.dli.jdbc.DliDriver。

“URL” 栏选择“自定义协议”，后面填写DLI jdbc的URL，URL的格式见表3-12，属性配置项说明见表3-13。

说明

- “表结构模式”可填写需访问的数据库名称，如果填写，后续创建数据集时，刷新表，页面上只可见该数据库下的表。如果不填写，后续创建数据集时，刷新表，页面上会显示所有数据库下的表。创建数据集请参考**永洪BI创建数据集**。
- 其他选项不需要填写，也无需勾选“需要登录”选项。

图 3-48 添加数据源配置

表 3-12 数据库连接参数

参数	描述
URL	URL的格式如下。 ***jdbc:dli://<endPoint>/<projectId>?<key1>=<val1>;<key2>=<val2>…*** **说明** • endpoint指DLI的终端节点，具体请参考**地区和终端节点**。 • projectId指项目编号，从华为云“基本信息>我的凭证”页面获取项目编号。 • “？”后面接其他配置项，每个配置项以“key=value”的形式列出，配置项之间以“;”隔开，详见**表3-13**

表 3-13 属性配置项

属性项（key）	必须配置	默认值（value）	描述
queuename	是	-	DLI服务的队列名称。
databasename	否	-	默认访问的数据库，URL中若不填此项，访问数据库的表时需采用db.table方式（如 select * from dbother.tabletest）。

属性项（key）	必须配置	默认值（value）	描述
authentication mode	是	token	身份认证方式，可以是token或aksk，永洪BI对接建议采用aksk认证方式。
accesskey	authenticationmode=aksk时必须配置	-	参考**永洪BI对接准备工作**。
secretkey	authenticationmode=aksk时必须配置	-	参考**永洪BI对接准备工作**。
regionname	authenticationmode=aksk时必须配置	-	具体请参考**地区和终端节点**。
servicename	authenticationmode=aksk时必须配置	-	由于是对接DLI，所以servicename=dli。
dli.sql.checkNoResultQuery	否	false	是否允许调用executeQuery接口执行没有返回结果的语句（如DDL）。 • “false”表示允许调用。 • “true”表示不允许调用。 **说明** 当dli.sql.checkNoResultQuery=false时，非查询语句会执行两次。

步骤4 在“添加数据源配置”页面工具栏中单击“测试连接”，测试通过后，单击“保存”，填写数据源名称，保存该数据源。

说明

目前没有根目录保存权限，需保存到已建文件夹目录下。

----结束

3.6.3 永洪 BI 创建数据集

操作场景

在永洪SaaS生产环境中创建DLI的数据集。

操作步骤

步骤1 在永洪SaaS生产环境主页，单击左侧导航栏中的“创建数据集”，请参见**图3-49**。

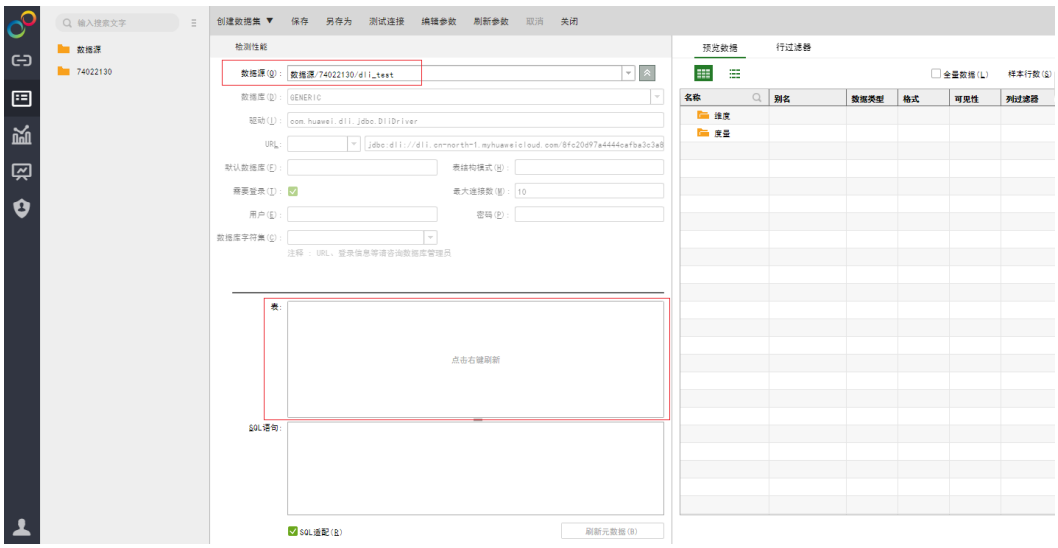
图 3-49 创建数据集

步骤2 在“数据集类型”页面中，选择创建“SQL数据集”，请参见图3-50。

图 3-50 创建 SQL 数据集

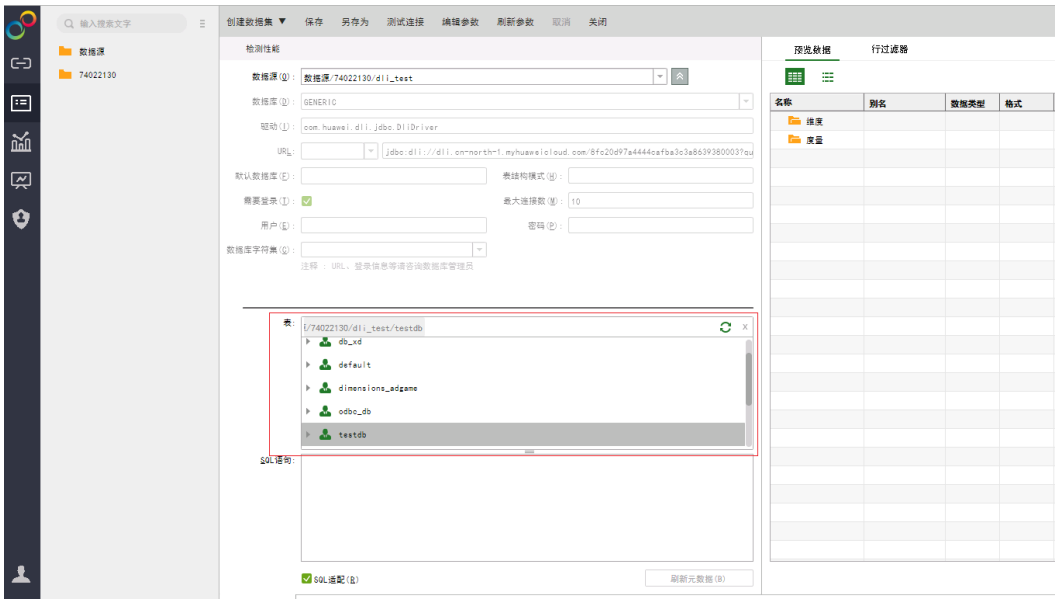
步骤3 在“创建数据集”页面中，左侧“数据源”栏选择已添加的DLI数据源，请参见图3-51。

图 3-51 选择数据源



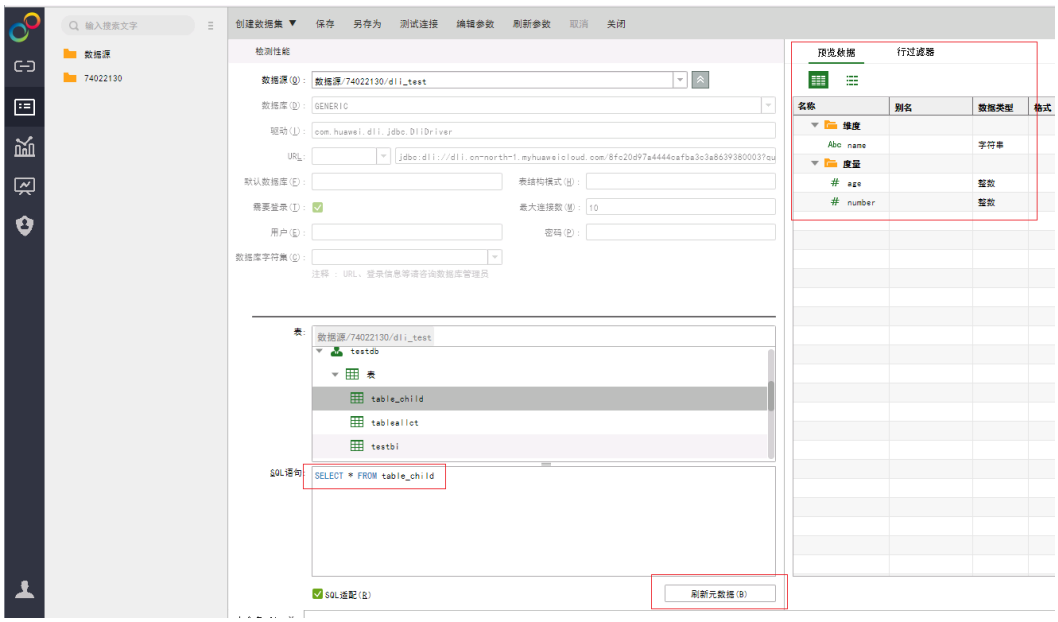
步骤4 左侧“表”栏中点击右键，刷新表，将列出所有数据库及数据库下面的数据表（这是添加数据源时，“表结构模式”没有配置时的情况），请参见图3-52。

图 3-52 刷新数据表



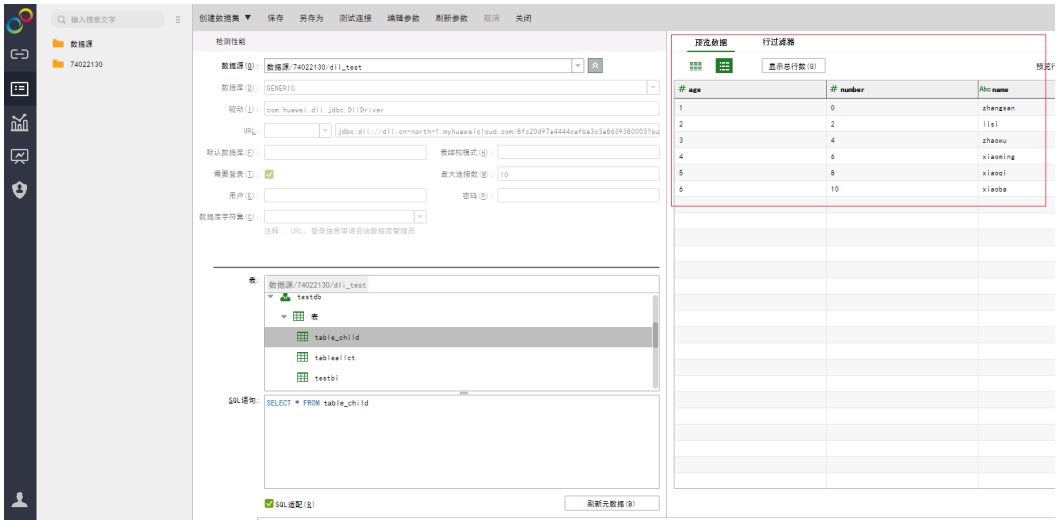
步骤5 在左侧“SQL语句”栏中执行表查询命令”select * from table_name“，点击“刷新元数据”，再单击右侧“预览数据”栏下左侧的“预览元数据”  ，可查询出该表的元数据（包括字段，字段类型等），请参见图3-53。

图 3-53 查询数据表



步骤6 单击右侧“预览数据”栏下右侧的“数据细节”  ，可查询出该表的数据，请参见图3-54。

图 3-54 查询数据表数据



步骤7 在“创建数据集”页面工具栏中单击“保存”，完成创建数据集。

----结束

3.6.4 永洪 BI 制作图表

操作场景

在永洪SaaS生产环境中制作图表。

操作步骤

步骤1 在永洪SaaS生产环境主页，单击左侧导航栏中的“制作报告”，请参见图3-55。

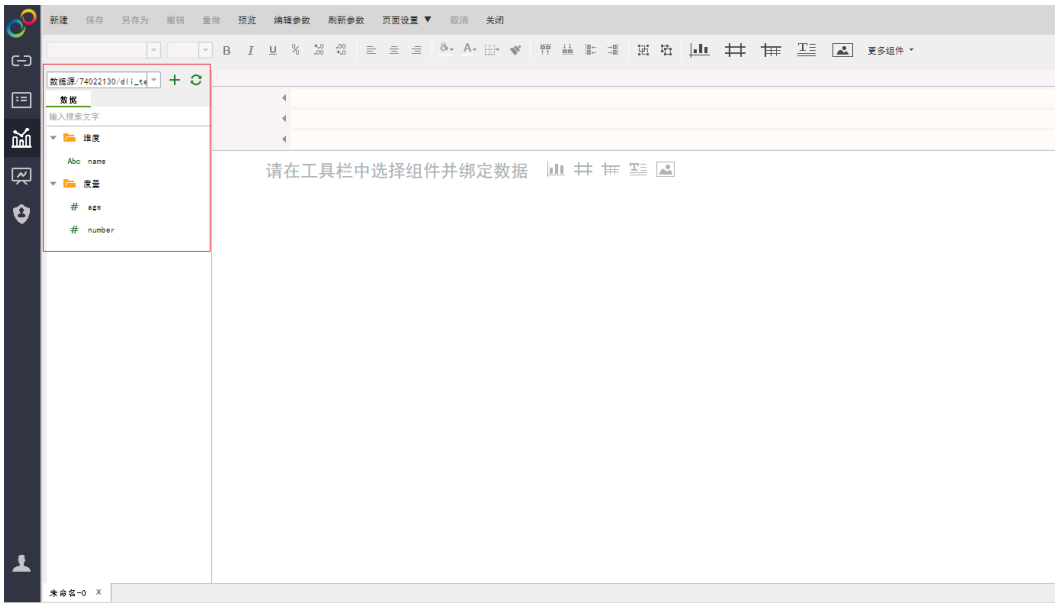
图 3-55 制作报告

步骤2 选择图表风格，请参见图3-56。

图 3-56 选择报告风格

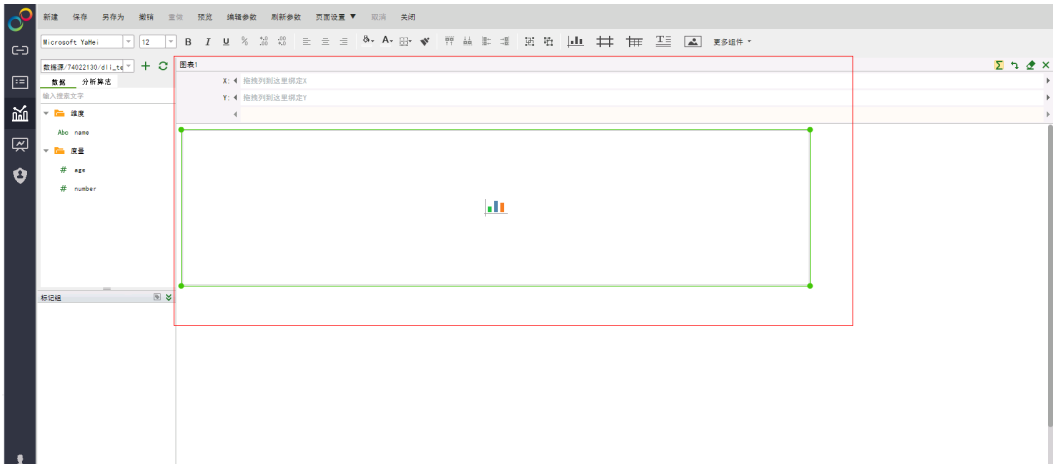
步骤3 选择“清爽绿主题”为例，在界面左侧下拉选择添加已创建的数据集，选择其中的一个表（例如table_child）作为数据源，会在下方的“数据”栏显示出该表的元数据（包括字段和字段类型），请参见图3-57。

图 3-57 选择表数据源



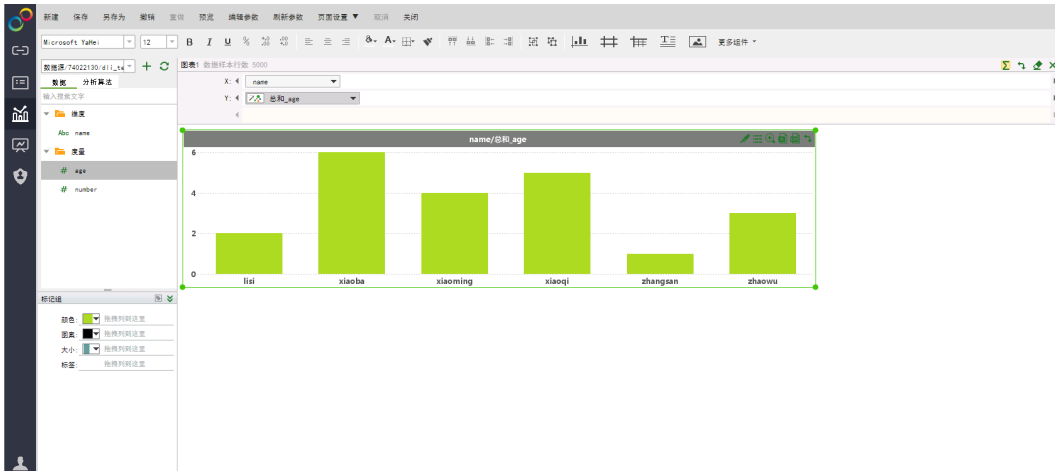
步骤4 在制作报告界面，制表组件主要包括图表、表、交叉表、列表过滤等，以新建图表为例，单击工具栏中的“新建图表 ”  ，将其拖入编辑区域，请参见图3-58。

图 3-58 新建图表



步骤5 选择“name”作为X变量，“age”作为Y变量，将其直接拖入对应的位置，系统将自动生成对应的柱状图，请参见图3-59。

图 3-59 生成图表



步骤6 在“制作图表”页面工具栏中单击“保存”，完成制作图表。

----结束

4 队列网络联通

4.1 配置 DLI 队列与内网数据源的网络联通

背景信息

DLI执行作业时如需访问外部数据源数据，如：DLI连接MRS、RDS、CSS、Kafka、DWS时，需要打通DLI和外部数据源之间的网络。DLI增强型跨源连接，底层采用对等连接的方式打通与目的数据源的vpc网络，通过点对点的方式实现数据互通。

创建增强型跨源连接网络不通的问题，可以根据本指导的整体流程和步骤进行排查验证。

整体流程

图 4-1 增强型跨源连接配置流程

前提条件

- 已创建DLI队列。创建队列详见**创建DLI队列**操作指导。

  注意

  队列的计费类型必须为“按需计费”且勾选“专属资源模式”。
  仅“专属资源模式”的“按需计费”资源才能创建增强型跨源链接。

- 已创建对应的外部数据源集群。具体对接的外部数据源根据业务自行选择。

表 4-1 创建各外部数据源参考

服务名	参考文档链接
RDS	**RDS MySQL快速入门**
DWS	**创建DWS集群**
DMS Kafka	**创建Kafka实例** **注意** 创建DMS Kafka实例时，**不能开启Kafka SASL_SSL**。
CSS	**创建CSS集群**
MRS	**创建MRS集群**

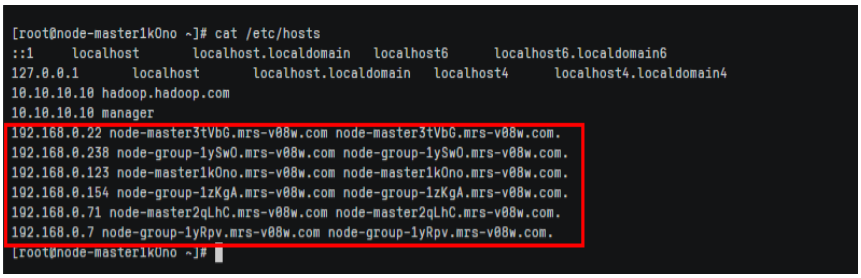
⚠ 注意

- 绑定跨源的**DLI队列网段和其他数据源子网网段不能重合**。
- 系统default队列不支持创建跨源连接。

步骤 1：获取外部数据源的内网 IP、端口和安全组

表 4-2 各数据源信息获取

数据源	参数获取
DMS Kafka	1. 在Kafka管理控制台，选择“Kafka专享版”，单击对应的Kafka名称，进入到Kafka的基本信息页面。 2. 在“连接信息”中获取该Kafka的“内网连接地址”，在“网络”中获取该实例的“虚拟私有云”和“子网”信息。 3. Kafka的基本信息页面，“网络 > 安全组”参数下获取Kafka的安全组。
RDS	在RDS控制台“实例管理”页面，单击对应实例名称，查看“连接信息”，获取“内网地址”、“虚拟私有云”、“子网”、“数据库端口”和“安全组”信息。
CSS	1. 在CSS管理控制台，选择“Elasticsearch > 集群管理”，单击已创建的CSS集群名称，进入到CSS的基本信息页面。 2. 在“基本信息”中获取CSS的“内网访问地址”、“虚拟私有云”、“子网”和“安全组”信息，方便后续操作步骤使用。
DWS	1. 在DWS管理控制台，选择“集群管理”，单击已创建的DWS集群名称，进入到DWS的基本信息页面。 2. 在“基本信息”的“数据库属性”中获取该实例的“内网IP”、“端口”，在“网络”中获取“虚拟私有云”、“子网”和“安全组”信息，方便后续操作步骤使用。

数据源	参数获取
MRS HBase	以MRS 3.x版本集群为例。 1. 登录MRS管理控制台，单击“集群列表 > 现有集群”，单击对应的集群名称，进入到集群概览页面。 2. 在集群概览页面“基本信息”中获取“虚拟私有云”、“子网”和“安全组”。 3. 因为在创建连接MRS HBase的作业时，需要用到MRS集群的ZooKeeper实例和端口，则还需要获取MRS集群主机节点信息。 a. 参考**访问MRS Manager**登录MRS Manager，在MRS Manager上，选择“集群 > *待操作的集群名称* > 服务 > ZooKeeper > 实例”，根据“主机名称”和“业务IP”获取ZooKeeper的主机信息。 b. 在MRS Manager上，选择“集群 > 待操作的集群名称 > 服务 > ZooKeeper > 配置 > 全部配置”，搜索参数“clientPort”，获取“clientPort”的参数值即为ZooKeeper的端口。 c. 使用root用户ssh登录任意一个MRS主机节点。具体请参考**登录MRS集群节点**。 d. 执行以下命令获取MRS对应主机节点的hosts信息，复制保存。 **cat /etc/hosts** 例如，查询结果参考如下，将内容复制保存，以备后续步骤使用。 

步骤 2：获取 DLI 队列网段

在DLI管理控制台，单击“资源管理 > 队列管理”，选择运行作业的队列，单击队列名称旁的 按钮，获取队列的网段信息。

步骤 3：外部数据源的安全组添加放通 DLI 队列网段的规则

1. 登录VPC控制台。
2. 在左侧导航树选择“访问控制 > 安全组”。
3. 单击外部数据源所属的安全组名称，进入安全组详情界面。

   您可以在对应数据源的管理控制台，参考**步骤1：获取外部数据源的内网IP、端口和安全组**获取对应数据源的安全组名称。
4. 在“入方向规则”页签中添加放通队列网段的规则。如**图4-2**所示。

   详细的入方向规则参数说明请参考**表4-3**。

图 4-2 添加入方向规则

表 4-3 入方向规则参数说明

参数	说明	取值样例
优先级	安全组规则优先级。 优先级可选范围为1-100，默认值为1，即最高优先级。优先级数字越小，规则优先级级别越高。	1
策略	安全组规则策略。	允许
协议端口	• 网络协议。目前支持“All”、“TCP”、“UDP”、“ICMP”和“GRE”等协议。 • 端口：允许远端地址访问指定端口，取值范围为：1～65535。	本例中选择TCP协议，端口值不填或者填写为**步骤1：获取外部数据源的内网IP、端口和安全组**获取的数据源的端口。
类型	IP地址类型。	IPv4
源地址	源地址用于放通来自IP地址或另一安全组内的实例的访问。	本例填写**步骤2：获取DLI队列网段**获取的队列网段。
描述	安全组规则的描述信息，非必填项。	_

步骤 4：创建增强型跨源连接

1. 登录DLI管理控制台，在左侧导航栏单击“跨源管理”，在跨源管理界面，单击“增强型跨源”，单击“创建”。
2. 在增强型跨源创建界面，配置具体的跨源连接参数。具体参考如下。
   - 连接名称：设置具体的增强型跨源名称。
   - 弹性资源池：选择DLI的队列。（未添加至资源池的队列，请直接选择队列名称。）

- 虚拟私有云：选择**步骤1：获取外部数据源的内网IP、端口和安全组**获取的外部数据源的虚拟私有云。
- 子网：选择**步骤1：获取外部数据源的内网IP、端口和安全组**获取的外部数据源的子网。
- 其他参数可以根据需要选择配置。

3. 参数配置完成后，单击“确定”完成增强型跨源配置。单击创建的跨源连接名称，查看跨源连接的连接状态，等待连接状态为：“已激活”后可以进行后续步骤。
4. 如果是连接MRS HBase，则还需要添加MRS的主机节点信息，具体步骤如下：
   a. 在“跨源管理 > 增强型跨源”中，在已创建的增强型跨源连接的“操作”列，单击“更多 > 修改主机信息”。
   b. 在“主机信息”参数中，将**步骤1：获取外部数据源的内网IP、端口和安全组**中获取到的MRS HBase主机节点信息拷贝追加进去。

   图 4-3 修改主机信息

   c. 单击“确定”完成主机信息添加。

步骤 5：测试网络连通性

1. 单击“队列管理”，选择操作的队列，在操作列，单击“更多 > 测试地址连通性”。
2. 在“测试连通性”界面，根据**步骤1：获取外部数据源的内网IP、端口和安全组**中获取的数据源的IP和端口，地址栏输入“数据源内网IP:数据源端口”，单击“测试”测试DLI到外部数据源网络是否可达。

   说明

   MRS HBase在测试网络连通性的时候，使用：**ZooKeeperIP地址:ZooKeeper端口，或者，ZooKeeper的主机信息:ZooKeeper端口**。

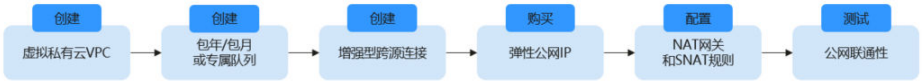
4.2 配置 DLI 队列与公网网络联通

操作场景

本节操作为您提供DLI队列在公网访问场景下网络打通的方法。通过配置SNAT规则，添加到公网的路由信息，可以实现队列到和公网的网络打通。

操作流程

图 4-4 配置 DLI 队列访问公网流程



步骤 1：创建 VPC

登录虚拟私有云控制台，创建虚拟私有云。创建的VPC供NAT访问公网使用。

创建VPC的具体操作请参考**创建虚拟私有云**。

图 4-5 创建 VPC

步骤 2：创建专属队列

本例以按需计费的专属资源队列为例。

注意

队列的计费类型必须为：“包年/包月”，“按需计费”（按需计费需勾选“专属资源模式”。）

仅“包年/包月”资源、“专属资源模式”的“按需计费”资源才能创建增强型跨源链接。

1. 登录DLI管理控制台。
2. 在“购买队列”页面，进行资源选型和参数配置。

   购买队列的详细参数请参考**创建队列**。

步骤 3：创建专属队列和 VPC 的增强型跨源连接

1. 在DLI管理控制台左侧导航栏中，选择“跨源管理”。
2. 选择“增强型跨源”页签，单击左上角的“创建”按钮。

   输入连接名称，选择创建的弹性资源池/队列，虚拟私有云，子网，输入主机信息（可选）。

图 4-6 创建增强型跨源连接

步骤 4：购买弹性公网 IP

1. 在“弹性公网IP”界面，单击“购买弹性公网IP”。
2. 根据界面提示配置参数。
   参数填写说明请参考**“购买弹性公网IP”**。

步骤 5：配置 NAT 网关

步骤1 创建NAT网关。

1. 登录控制台，在“服务列表”搜索“NAT网关”，进入网络控制台页面。
2. 单击“购买公网NAT网关”，配置NAT网关的相关信息。
   详细请参考《NAT网关用户指南》中**“购买公网NAT网关”**。

图 4-7 购买 NAT 网关

3. 配置完成后，单击“立即购买”。

   说明

   “虚拟私有云”为**步骤1：创建VPC**创建的VPC。

步骤2 添加路由。

进入VPC的路由表，配置路由规则。通常NAT创建成功会自动创建到NAT网关的路由。

目的地址为访问的公网IP地址，下一跳为NAT网关。

图 4-8 添加路由

步骤3 添加SNAT规则。

为新建的NAT网关添加SNAT规则，才能实现该子网下的主机与Internet互相访问。

1. NAT网关购买成功后，在NAT控制台，单击购买成功的NAT网关“名称”，进入NAT网关详情页面。
2. 选择“SNAT规则”页签，单击“添加SNAT规则”。
   详细请参考《NAT网关用户指南》中“**添加SNAT规则**”。
3. 使用场景选择云专线/云连接。
4. 添加专属队列所在的网段。
5. 绑定对应的弹性公网IP。

图 4-9 添加 SNAT 规则

6. 添加完成后，单击“确定”。

----结束

步骤 6：添加自定义路由

在增强型跨源连接页面添加自定义路由。此处添加的是访问的IP地址的路由信息。

详细操作请参考**自定义路由信息**。

图 4-10 增强型跨源链接添加测试路由信息

步骤 7：测试公网连通性

测试队列到公网的连通性。单击队列操作列下方的“更多 > 测试地址连通性”，输入访问的公网IP地址。

图 4-11 测试地址联通性