

解决方案实践

轻量离线大数据解决方案

文档版本 1.0
发布日期 2023-12-08



版权所有 © 华为技术有限公司 2023。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



HUAWEI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

漏洞声明

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该政策可参考华为公司官方网站的网址：<https://www.huawei.com/cn/psirt/vul-response-process>。

如企业客户须获取漏洞信息，请访问：<https://securitybulletin.huawei.com/enterprise/cn/security-advisory>。

目录

1 方案概述	1
2 资源和成本规划	3
3 实施步骤	4
3.1 准备工作.....	4
3.2 快速部署.....	8
3.3 开始使用.....	12
3.4 快速卸载.....	27
4 附录	28
5 修订记录	29

1 方案概述

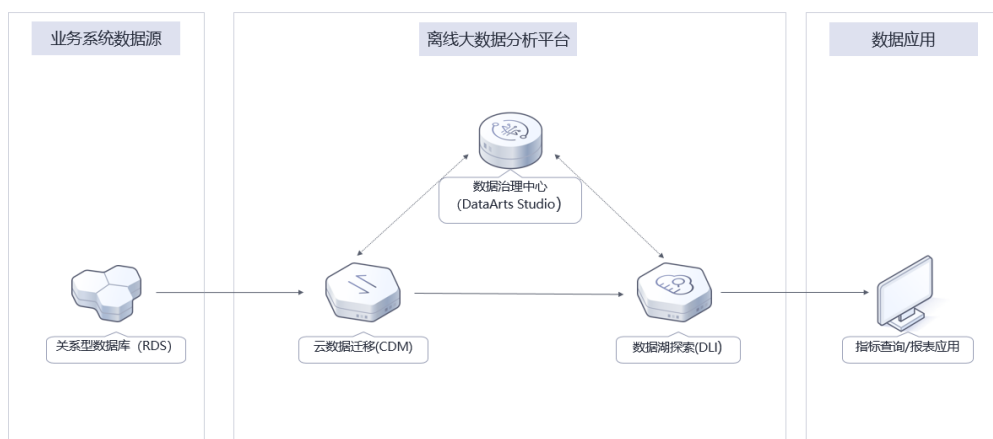
应用场景

此解决方案通过华为云数据治理中心 DataArts Studio、数据湖探索服务 DLI、云数据迁移 CDM、云数据库 RDS等服务的组合，可以快速构建一个轻量化离线数据分析平台，可一站式完成业务数据采集、分析，指标查询、做展示等工作，帮助企业实现业务数据洞察。

方案架构

该解决方案部署架构如下图所示：

图 1-1 方案架构



该解决方案会部署如下资源：

- 数据迁移服务 CDM：高效、易用的批量/增量数据迁移和同步服务，提供了40+数据源到数据湖仓的集成能力；
- 数据治理中心 DataArts Studio：提供数据标准和数据模型的规范设计，以及数据分析流程的在线开发、任务编排调度等的一站式数据治理能力；

- 数据湖探索服务 DLI：是一个完全兼容Spark、FLink生态，支持流处理、批处理、交互式分析的Serverless大数据分析服务；

方案优势

- 免部署
本方案提供的都是云原生服务，客户无需关注服务器，无需自行部署软件，一键式、分钟级完成整个方案的资源发放。
- 轻开发
本方案全流程的高度可视化，没有复杂的代码编写过程，对技术人员的技术要求不高，可以大大降低大数据分析的门槛，只要懂数据库会SQL就能够进行大数据分析。
- 易运维
本方案为客户完全屏蔽了具体软件引擎的复杂度，让客户技术团队从传统的软件问题定位、处理、调优等运维工作中彻底解放，完全聚焦在业务开发、数据价值创造上。

约束与限制

- 部署该解决方案之前，您需注册华为云账户，完成实名认证，且帐号不能处于欠费或冻结状态，请根据[表2-1](#)中预估价格，确保余额充足。

2 资源和成本规划

该解决方案主要部署如下资源，不同产品的花费仅供参考，实际以收费账单为准，具体请参考华为云[官网价格](#)：

表 2-1 资源和成本规划

华为云服务	配置示例	每月预估花费
数据治理中心 DataArts Studio	<ul style="list-style-type: none">区域：华北-北京四计费模式：包年包月版本：初级版规格：cdm.medium 4核 8GB购买量：1	2000元
数据迁移服务 CDM	<ul style="list-style-type: none">按需计费：2.25/小时区域：华北-北京四计费模式：按需计费规格：cdm.large 8核 16GB购买数量：1	$2.25 * 24 * 30 = 1620$ 元
数据湖探索服务 DLI	<ul style="list-style-type: none">按需计费：5.6元/小时区域：华北-北京四计费模式：按需计费类型：SQL队列CPU架构：X86规格：16 CUs购买数量：1	$5.6 * 24 * 30 = 4032$ 元
合计	-	7652元

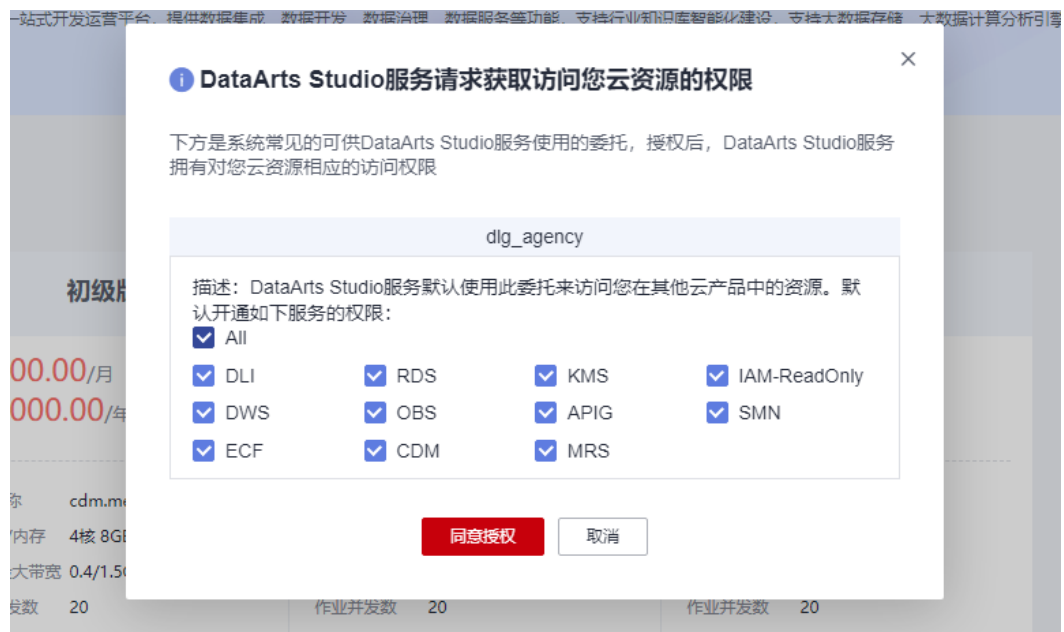
3 实施步骤

- 3.1 准备工作
- 3.2 快速部署
- 3.3 开始使用
- 3.4 快速卸载

3.1 准备工作

步骤1 登录[数据治理中心](#)，系统会自动弹出“云资源访问授权”的对话框，提示您对所列出的服务进行委托授权。勾选所有服务并单击“同意授权”，系统会自动创建委托。更多详情介绍请参考[DataArts Studio基础包](#)。

图 3-1 委托授权



步骤2 登录[华为云官网控制台](#)，打开[虚拟私有云子网](#)查看子网，打开源端业务数据库 RDS所属的子网，获取网络ID。

图 3-2 子网网络 ID



步骤3 在**网络控制台**，查看安全组，打开源端业务数据库 RDS所配置的安全组，获取安全组 ID。

图 3-3 安全组 ID



----结束

创建 rf_amdin_trust 委托

步骤1 进入华为云官网，打开**控制台管理**界面，鼠标移动至个人帐号处，打开“统一身份认证”菜单。

图 3-4 控制台管理界面



图 3-5 统一身份认证菜单



步骤2 进入“委托”菜单，搜索“rf_admin_trust”委托。

图 3-6 委托列表



- 如果委托存在，则不用执行接下来的创建委托的步骤
- 如果委托不存在时执行接下来的步骤创建委托

步骤3 单击步骤2界面中的“创建委托”按钮，在委托名称中输入“rf_admin_trust”，选择“普通帐号”，委托的帐号，输入“op_svc_IAC”，单击“下一步”。

图 3-7 创建委托



步骤4 在搜索框中输入” Tenant Administrator” 权限，并勾选搜索结果。

图 3-8 选择策略



步骤5 选择“所有资源”，并单击下一步完成配置。

图 3-9 设置授权范围



步骤6 “委托”列表中出现“rf_admin_trust”委托则创建成功。

图 3-10 委托列表



----结束

3.2 快速部署

本章节主要帮助用户快速部署“轻量离线大数据解决方案”。

表 3-1 参数填写说明

参数名称	类型	是否必填	参数解释	默认值
subnet_id	String	必填	子网ID, 该模板使用已有子网, 请选择后端业务服务器所在虚拟私有云VPC内子网, 查询子网ID请参考部署指南。	空
secgroup_id	String	必填	安全组ID, 该模板使用已有安全组, 建议选择后端业务服务器相同安全组, 查询安全组ID请参考部署指南。	空
cdm_cluster_name	String	必填	云数据迁移集群CDM名称。集群名称在4位到64位之间, 必须以字母开头, 可以包含字母、数字、中划线或者下划线, 不能包含其他的特殊字符。	lightweight_offline_big_data_demo
dli_queue_name	String	必填	数据湖探索服务dli新建队列名称。名称只能包含数字、英文字母和下划线, 但不能是纯数字, 且不能以下划线开头。长度限制: 1~128个字符。	lightweight_offline_big_data_demo
dataarts_studio_name	String	必填	数据治理中心DataArts Studio名称。名称在4位到64位之间, 必须以字母开头, 可以包含字母、数字、中划线或者下划线, 不能包含其他的特殊字符。	lightweight_offline_big_data_demo
dli_cu_count	Number	必填	与队列绑定的最小计算单元个数。设置值当前只支持16, 64, 256。默认16。	16

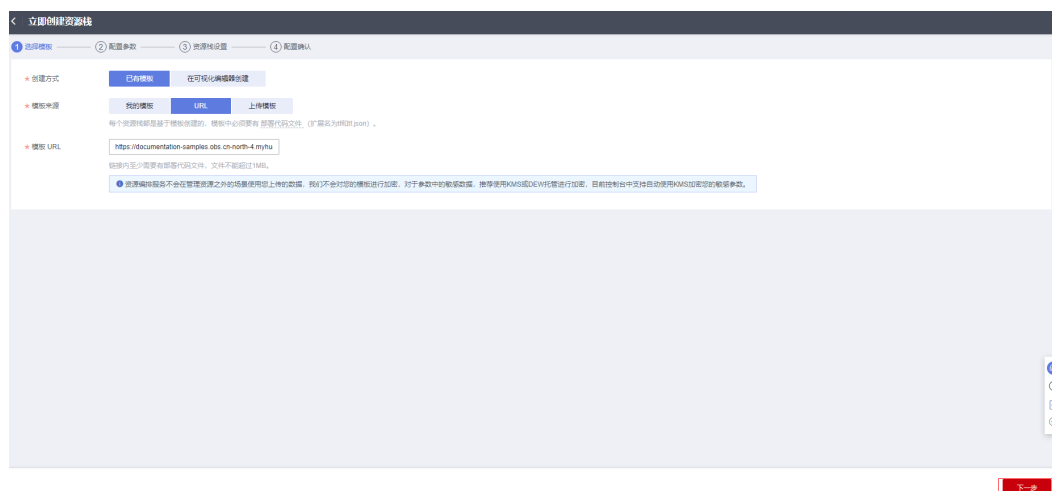
步骤1 登录[华为云解决方案实践](#), 选择“轻量化离线大数据解决方案”, 单击“一键部署”, 跳转至解决方案创建堆栈界面。

图 3-11 解决方案实施库



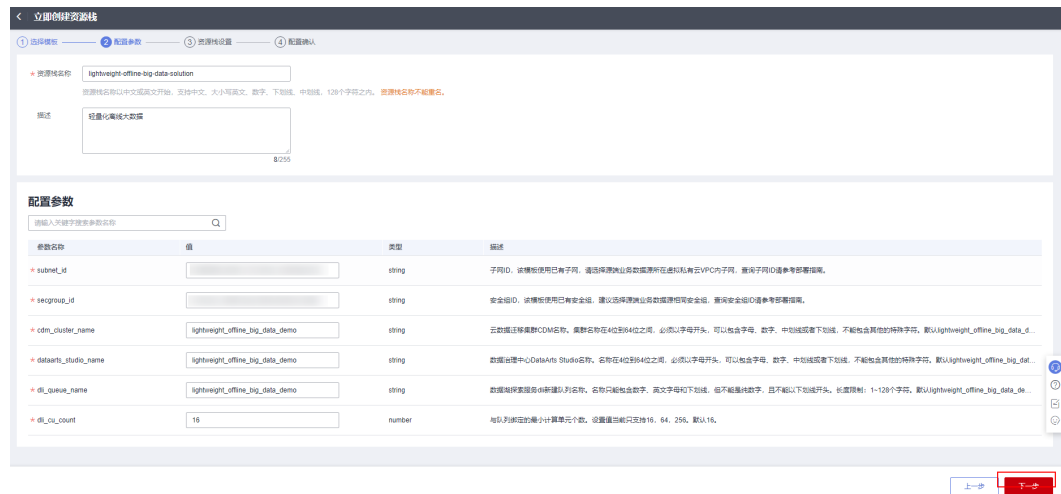
步骤2 在选择模板界面中，单击“下一步”。

图 3-12 选择模板



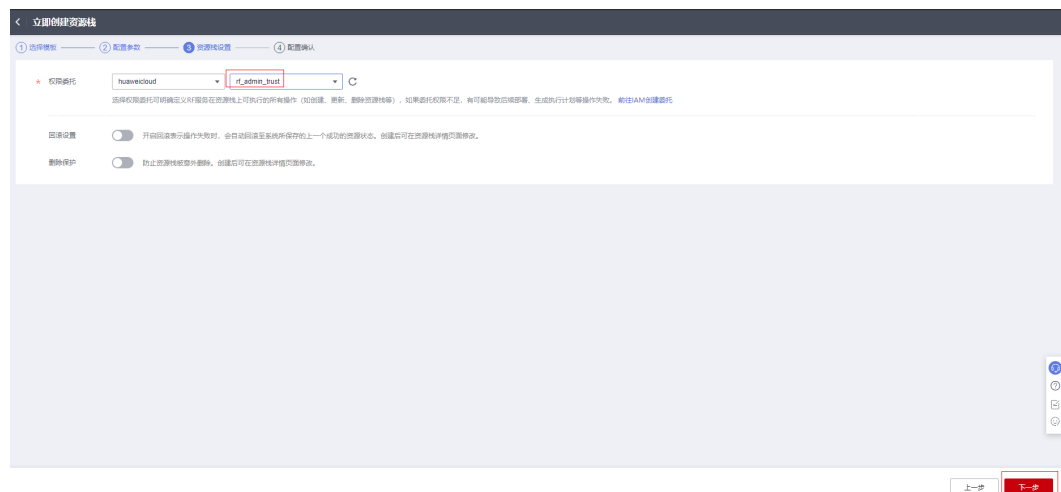
步骤3 在配置参数界面中，自定义填写堆栈名称，根据表3-1配置参数信息，单击“下一步”。

图 3-13 配置参数



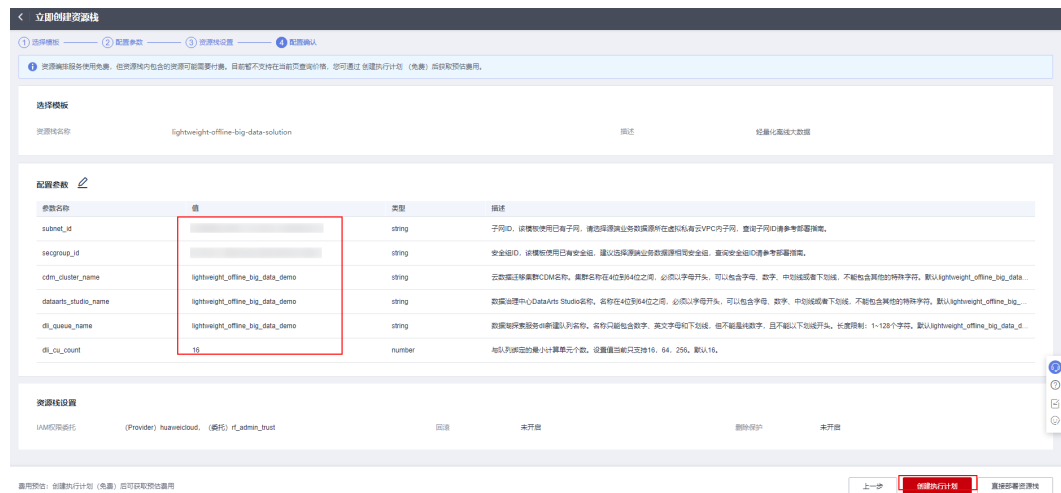
步骤4 在资源栈设置页面中，权限委托选择“rf_admin_trust”，单击“下一步”。

图 3-14 高级配置



步骤5 在配置确认页面中，单击“创建执行计划”。

图 3-15 配置确认



步骤6 在弹出的创建执行计划框中，自定义填写执行计划名称，单击“确定”。

图 3-16 创建执行计划



步骤7 等待执行计划状态为“创建成功，待部署”后，单击“部署”，并且在弹出的执行计划确认框中单击“执行”。

图 3-17 执行计划创建成功

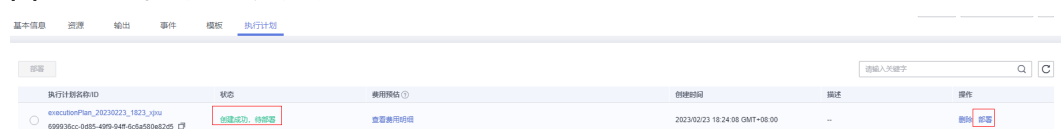


图 3-18 执行计划确认

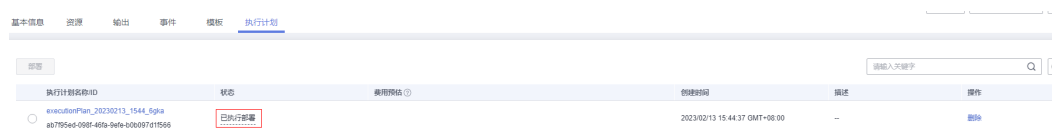


步骤8 等待解决方案自动部署。部署成功后，单击“事件”，回显结果如下：

图 3-19 资源创建成功



图 3-20 执行计划完成



----结束

3.3 开始使用

说明

本解决方案以真实数据（脱敏后）为基础，为用户提供配置样例，供用户借鉴参考

数据迁移示例

步骤1 登录华为云[云数据库RDS](#)控制台，进入RDS列表，选择业务数据源登录。

图 3-21 登录 RDS



步骤2 查看业务数据源"cyz_test"数据库, "t_trade_order"和"t_user_store_info"两张表的基本信息(数据已提前导入), 执行SQL获取对应需要在DLI中创建表的语句。

图 3-22 数据库的基本信息



图 3-23 建表结果



获取DLI建表语句:

```
CREATE TABLE ods_ddl_convert (
  source VARCHAR(100) NOT NULL,
  data_type1 VARCHAR(100) NOT NULL,
  target VARCHAR(100) NOT NULL,
  data_type2 VARCHAR(100),
  update_time VARCHAR(26),
  PRIMARY KEY (source, data_type1, target)
) ENGINE = InnoDB DEFAULT CHARSET = utf8mb4 COMMENT = '数据库表结构转换';

INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql', 'bigint', 'hive', 'BIGINT', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql', 'binary', 'hive', 'BINARY', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql', 'char',
```

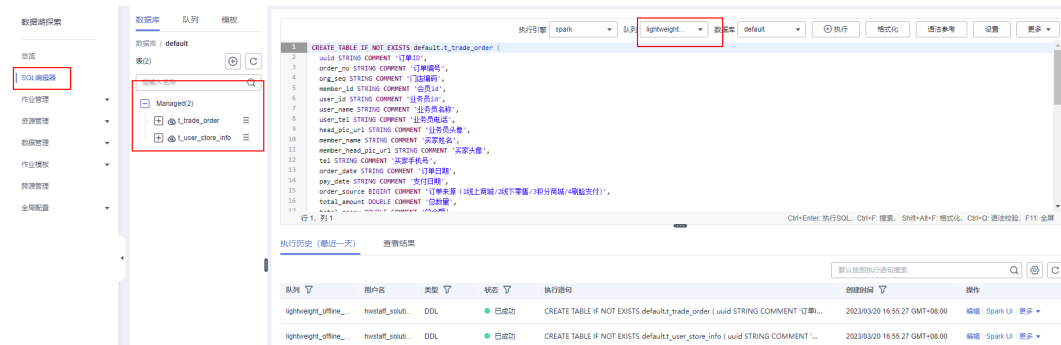
```
'hive', 'STRING', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'datetime', 'hive', 'STRING', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'decimal', 'hive', 'DOUBLE', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'double', 'hive', 'DOUBLE', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'float', 'hive', 'DOUBLE', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql', 'int',
'hive', 'BIGINT', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql', 'json',
'hive', 'MAP<STRING,STRING>', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'mediumtext', 'hive', 'STRING', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'smallint', 'hive', 'BIGINT', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql', 'text',
'hive', 'STRING', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'time', 'hive', 'STRING', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'timestamp', 'hive', 'STRING', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'tinyint', 'hive', 'BIGINT', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'varbinary', 'hive', 'BINARY', '2019-07-31 00:00:00');
INSERT INTO ods_ddl_convert (source, data_type1, target, data_type2, update_time) VALUES ('mysql',
'varchar', 'hive', 'STRING', '2019-07-31 00:00:00');

SET SESSION group_concat_max_len = 102400;
SELECT
  a.TABLE_NAME ,
  concat('DROP TABLE IF EXISTS ',a.TABLE_NAME,';CREATE TABLE IF NOT EXISTS ',a.TABLE_NAME ,
('group_concat(concat(a.COLUMN_NAME,' ',
  c.data_type2," COMMENT '",COLUMN_COMMENT,'"") order by a.TABLE_NAME,a.ORDINAL_POSITION) ,
  ") COMMENT '",b.TABLE_COMMENT ,"' STORED AS parquet;") AS col_name
FROM
  (
    SELECT
      TABLE_SCHEMA,
      TABLE_NAME,
      COLUMN_NAME,
      ORDINAL_POSITION,
      DATA_TYPE,
      COLUMN_COMMENT
    FROM
      information_schema.COLUMNS
    WHERE
      TABLE_SCHEMA='cyz_test'
  ) AS a
LEFT JOIN
  information_schema.TABLES AS b
ON
  a.TABLE_NAME=b.TABLE_NAME
AND a.TABLE_SCHEMA=b.TABLE_SCHEMA
#选择源为mysql，目标为hive
LEFT JOIN
  (
    select
      *
    from ods_ddl_convert
    where source='mysql' and target='hive'
  ) AS c
ON
  a.DATA_TYPE=c.data_type1
where b.TABLE_TYPE='BASE TABLE'
and a.TABLE_NAME not like 'ods_%'
GROUP BY
```

a.TABLE_NAME,
b.TABLE_COMMENT

步骤3 登录数据湖探索 DLI服务控制台（首次使用您可以设置临时日志存储，根据弹窗设置 OBS即可），选择SQL编辑->复制下方语句->选择模板自动创建的 lightweight_offline_big_data_demo队列->default数据库->单击执行，确保左侧SQL编辑器数据库中表创建成功。

图 3-24 在 default 数据库中建表



在数据湖探索 DLI的default数据库中执行的建表语句：

```
CREATE TABLE IF NOT EXISTS default.t_user_store_info (  
  uuid STRING COMMENT '主键ID',  
  org_seq STRING COMMENT '组织机构编码',  
  auth_code STRING COMMENT '授权码',  
  store_name STRING COMMENT '店铺姓名',  
  business_license STRING COMMENT '营业执照',  
  store_type BIGINT COMMENT '店铺类型(1一代店、2二代店、3三代店、4四代店、5五代店、6六代店)',  
  manager_area DOUBLE COMMENT '经营面积',  
  building_area DOUBLE COMMENT '建筑面积',  
  practical_area DOUBLE COMMENT '实用面积',  
  leasehold_area DOUBLE COMMENT '租赁面积',  
  warehouse_area DOUBLE COMMENT '仓库面积',  
  open_status BIGINT COMMENT '营业状态(1-营业/0-停业)',  
  open_time STRING COMMENT '营业时间',  
  open_date STRING COMMENT '开业日期',  
  province STRING COMMENT '所属省',  
  city STRING COMMENT '所属城市',  
  business_district_type BIGINT COMMENT '商圈类型',  
  company_name STRING COMMENT '经销商公司名',  
  legal_person STRING COMMENT '法人',  
  legal_person_phone STRING COMMENT '法人手机号',  
  store_owner_name STRING COMMENT '店主姓名',  
  store_owner_phone STRING COMMENT '店主手机号',  
  store_phone STRING COMMENT '店铺电话',  
  store_owner_grade STRING COMMENT '店主评级',  
  store_address STRING COMMENT '店铺地址',  
  store_address_detail STRING COMMENT '店铺详细地址',  
  store_lng STRING COMMENT '店铺经度',  
  store_lat STRING COMMENT '店铺纬度',  
  doorway_pic STRING COMMENT '门头照片',  
  store_owner_pic STRING COMMENT '店主照片',  
  store_manager_pic STRING COMMENT '店长照片',  
  status BIGINT COMMENT '状态',  
  create_time STRING COMMENT '创建时间',  
  create_user STRING COMMENT '创建人',  
  update_time STRING COMMENT '更新时间',  
  update_user STRING COMMENT '更新人',  
  gradelevel_id STRING COMMENT '等级id',  
  store_manager_phone STRING COMMENT '店长电话',  
  store_manager_name STRING COMMENT '店长姓名',
```

```
merid STRING COMMENT '工行Merid(12位识别码)',
agreement STRING COMMENT '工行协议号',
wechat_id STRING COMMENT '微信识别码',
up_org_zone STRING COMMENT '所属片区',
base_code STRING COMMENT '所属CBD',
base_person_code STRING COMMENT 'CBD所属业务员',
up_org_seq STRING COMMENT '所属片区组织编码',
open_begin_time STRING COMMENT '营业开始时间',
open_end_time STRING COMMENT '营业结束时间',
base_name STRING COMMENT '所属CBD名称',
base_person_name STRING COMMENT 'CBD所属业务员姓名',
store_code STRING COMMENT '店铺编码',
shut_down_time STRING COMMENT '店铺停业时间',
is_icbc BIGINT COMMENT '是否是工行对公账户(1-是 2-否)',
front_card STRING COMMENT '法人身份证正面照片',
back_card STRING COMMENT '法人身份证背面照片',
settlement_pic STRING COMMENT '结算帐号照片',
created_date STRING COMMENT '建档日期',
icbc_switch BIGINT COMMENT '银行解锁(1-已解锁, 2-已锁定)',
icbc_update_time STRING COMMENT '档案更新时间',
bank_account STRING COMMENT '银行帐号',
bank_name STRING COMMENT '银行名称'
) COMMENT '门店信息' STORED AS parquet;
CREATE TABLE IF NOT EXISTS default.t_trade_order (
  uuid STRING COMMENT '订单ID',
  order_no STRING COMMENT '订单编号',
  org_seq STRING COMMENT '门店编码',
  member_id STRING COMMENT '会员id',
  user_id STRING COMMENT '业务员id',
  user_name STRING COMMENT '业务员名称',
  user_tel STRING COMMENT '业务员电话',
  head_pic_url STRING COMMENT '业务员头像',
  member_name STRING COMMENT '买家姓名',
  member_head_pic_url STRING COMMENT '买家头像',
  tel STRING COMMENT '买家手机号',
  order_date STRING COMMENT '订单日期',
  pay_date STRING COMMENT '支付日期',
  order_source BIGINT COMMENT '订单来源(1线上商城/2线下零售/3积分商城/4刷脸支付)',
  total_amount DOUBLE COMMENT '总数量',
  total_money DOUBLE COMMENT '总金额',
  salesevent_deduct_money DOUBLE COMMENT '活动抵扣金额',
  discount_money DOUBLE COMMENT '定向券扣减金额',
  received_money DOUBLE COMMENT '实收金额',
  coupon_deduct_money DOUBLE COMMENT '通用券抵扣金额',
  pay_money DOUBLE COMMENT '应付款金额',
  pay_method BIGINT COMMENT '付款方式(0线下支付/1线上支付/2微信/3支付宝/4银行卡)',
  delivery_method BIGINT COMMENT '配送方式(1门店自提/2配送服务/3快递服务)',
  order_status BIGINT COMMENT '订单状态(0待支付/1待发货/2待收货/3已完成/4已取消)',
  pay_status BIGINT COMMENT '支付状态(0待支付/1已支付)',
  buyer_note STRING COMMENT '买家留言',
  tdcode STRING COMMENT '提货码',
  tdqrcode STRING COMMENT '提货二维码',
  order_bonuspoint DOUBLE COMMENT '本次积分',
  status BIGINT COMMENT '0 删除 1成功',
  create_user STRING COMMENT '创建人',
  create_time STRING COMMENT '创建时间',
  update_user STRING COMMENT '更新人',
  update_time STRING COMMENT '更新时间',
  pay_time STRING COMMENT '支付时间',
  delivery_time STRING COMMENT '发货时间',
  completed_time STRING COMMENT '完成时间(收货时间)',
  canceled_time STRING COMMENT '取消时间',
  member_coupon_id STRING COMMENT '会员卡卷id',
  return_status BIGINT COMMENT '退货状态(0正常单/1被退货单/2退货单)',
  return_order_id STRING COMMENT '退货单原单id/被退货单退货单id',
  auto_completed_status BIGINT COMMENT '自动收货延期状态(0未延期/1已延期)',
  auto_completed_date STRING COMMENT '自动收货日期',
  is_customize_order BIGINT COMMENT '是否定制订单(1是0否)',
  customize_type BIGINT COMMENT '定制类型(1、2)',
```

```
customize_status BIGINT COMMENT '定制状态(1审核中/2已驳回/3审核成功待生产/4生产中/5生产完成待配送/6已配送)',
customize_price DOUBLE COMMENT '定制单价',
customize_amount DOUBLE COMMENT '定制总金额'
) COMMENT '订单信息' STORED AS parquet;
```

步骤4 登录[云数据迁移服务 CDM](#)控制台，选择模板自动创建的集群，单击右侧的作业管理。选择连接管理，单击新建连接，选择数据湖探索（DLI）单击 下一步。

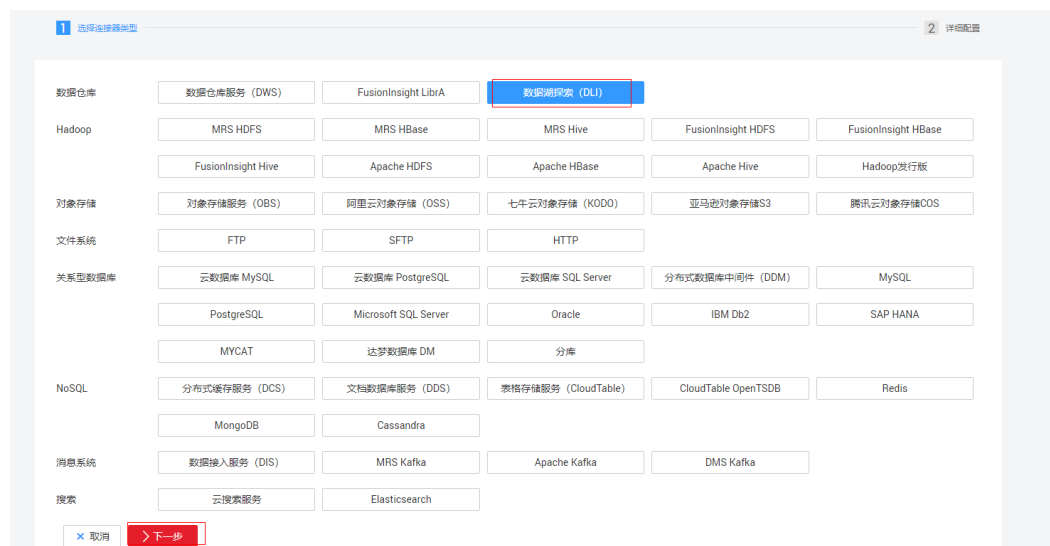
图 3-25 作业管理

集群名称	集群状态	内网地址	公网地址	企业项目	操作
lightweight_offline_big_data_demo_AGENT_wkQzNVM	运行中	192.168.0.117	--	default	作业管理 弹性伸缩 更多
lightweight_offline_big_data_demo	运行中	192.168.0.208	--	default	作业管理 弹性伸缩 更多

图 3-26 连接管理



图 3-27 数据湖探索



步骤5 填写名称“dli-link”，输入华为云ak、sk（[如何获取AK/SK](#)），单击保存。

图 3-28 新建连接

The screenshot shows a 'New Connection' configuration form. It includes the following fields and values:

- * 名称: dli-link
- * 连接器: DLI
- * 访问标识(AK) (?): IA
- * 密钥(SK) (?):
- * 项目ID: 0f4db96fe080f5fe2f8ac012c

At the bottom of the form, there are four buttons: '取消' (Cancel), '上一步' (Previous Step), '测试' (Test), and '保存' (Save).

步骤6 单击新建连接，选择“云数据库 MySQL”，填写名称“rds-link”，选择“rds-dgc”源端业务数据库实例，端口“3306”，数据库名称“cyz_test”，用户名/密码，单击保存。

图 3-29 数据库连接

第一次创建数据库连接时，需到“[驱动管理](#)”上传对应驱动。

* 名称	<input type="text" value="rds-link"/>	
* 连接器	<input type="text" value="关系数据库"/>	
数据库类型	<input type="text" value="MySQL"/>	
* 数据库服务器 ?	<input type="text" value="192.168.1.34"/>	选择
* 端口 ?	<input type="text" value="3306"/>	
* 数据库名称 ?	<input type="text" value="cyz_test"/>	
* 用户名 ?	<input type="text" value="root"/>	
* 密码 ?	<input type="password" value="....."/>	
使用本地API ?	<input type="radio" value="是"/> <input checked="" type="radio" value="否"/>	
使用Agent ?	<input type="radio" value="是"/> <input checked="" type="radio" value="否"/>	

[显示高级属性](#)

注意

- 首次连接RDS需要上传驱动，单击图中上方的蓝色字体“[驱动管理](#)”进行驱动上传，驱动版本参考[驱动下载地址](#)。
- RDS和CDM在同一个VPC时可以直接连接，否则RDS和CDM均需绑定EIP后使用公网ip，同时RDS安全组添加CDM的入方向规则，放开22号端口，安全组规则添加参考[添加安全组规则](#)。

步骤7 (a)选择表/文件迁移，单击新建作业，输入作业名称“t_user_store_info”，作业信息按照下图进行填写。配置完成后单击下一步，字段会自动映射，无需修改，单击下一步，单击保存。

(b)选择表/文件迁移，单击新建作业，输入作业名称“t_trade_order”，作业信息按照下图进行填写。配置完成后单击下一步，字段会自动映射，无需修改，单击下一步，单击保存。

图 3-30 表 1 作业配置

作业配置

* 作业名称

源端作业配置

* 源连接名称 +

使用SQL语句 是 否

* 模式或表空间 ⊖

* 表名 ⊖

隐藏高级属性

Where子句

抽取分区字段 ⊖

分区字段是否允许空值 是 否

拆分作业 是 否

目的端作业配置

* 目的连接名称 +

* 资源队列 ⊖

* 数据库名称 ⊖

* 表名 ⊖

导入前清空数据 是 否

清空数据方式 ⊖

分区

图 3-31 表 2 作业配置

作业配置

* 作业名称

源端作业配置

* 源连接名称 +

使用SQL语句 是 否

* 模式或表空间 ⊖

* 表名 ⊖

隐藏高级属性

Where子句

抽取分区字段 ⊖

分区字段是否允许空值 是 否

拆分作业 是 否

目的端作业配置

* 目的连接名称 +

* 资源队列 ⊖

* 数据库名称 ⊖

* 表名 ⊖

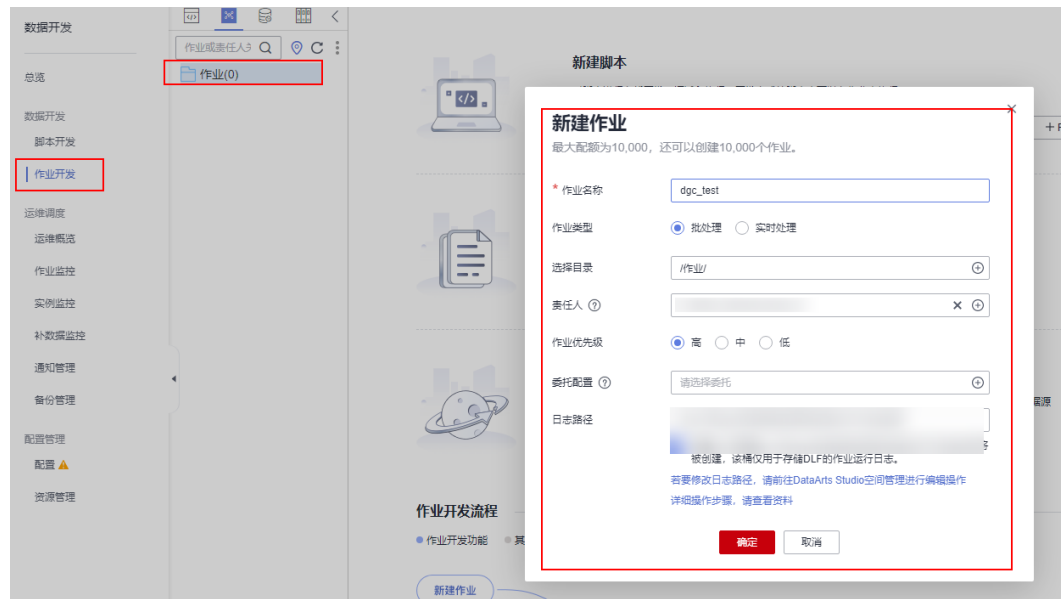
导入前清空数据 是 否

清空数据方式 ⊖

分区

步骤8 登录**数据治理中心 DataArts Studio**控制台，选择模板自动创建的实例，选择default工作空间，单击数据开发，在“作业”目录右键，单击新建作业，输入作业名称“dgc_test”作业信息默认即可，单击确定。

图 3-32 新建作业

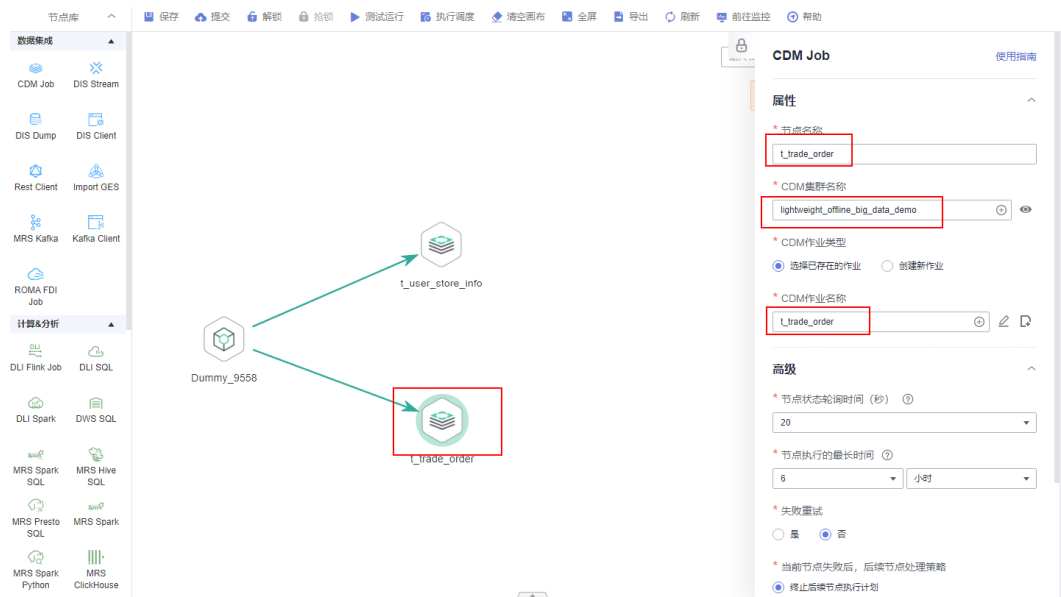


步骤9 在弹出的界面中选择“节点库”，拖动1个Dummy节点、2个CDM Job节点，按下图填写配置。

图 3-33 CDM job 节点 1 配置



图 3-34 CDM job 节点 2 配置

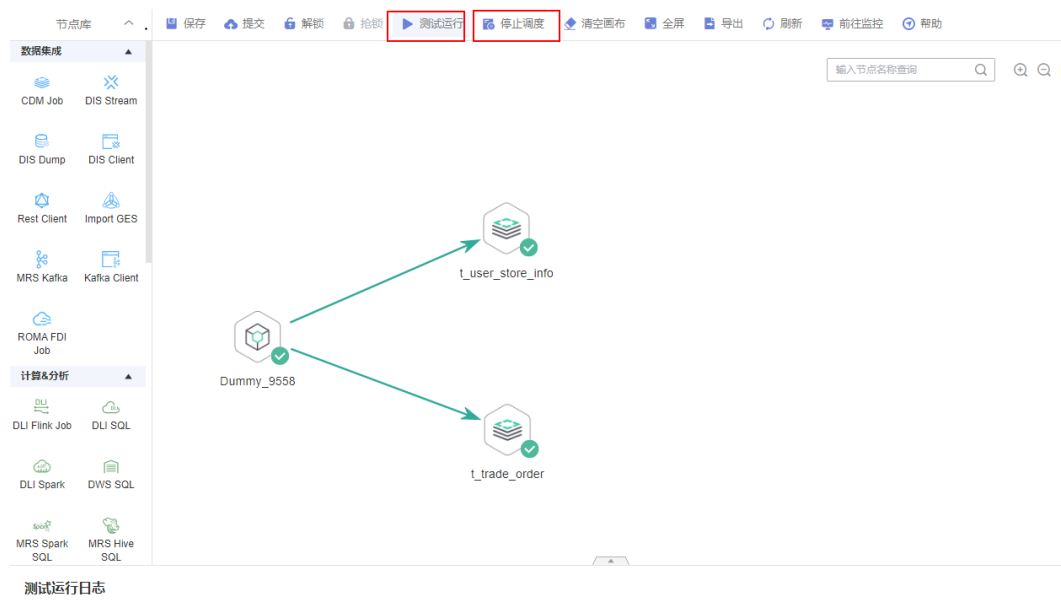


步骤10 按照截图填好信息后，修改调度信息。单击“保存并提交版本”，单击“执行调度”，单击“测试运行”。

图 3-35 调度配置



图 3-36 执行调度后测试运行



步骤11 执行CDM作业，将RDS的两张表“t_user_store_info”、“t_trade_order”的数据同步到DLI的表中。

图 3-37 数据迁移成功

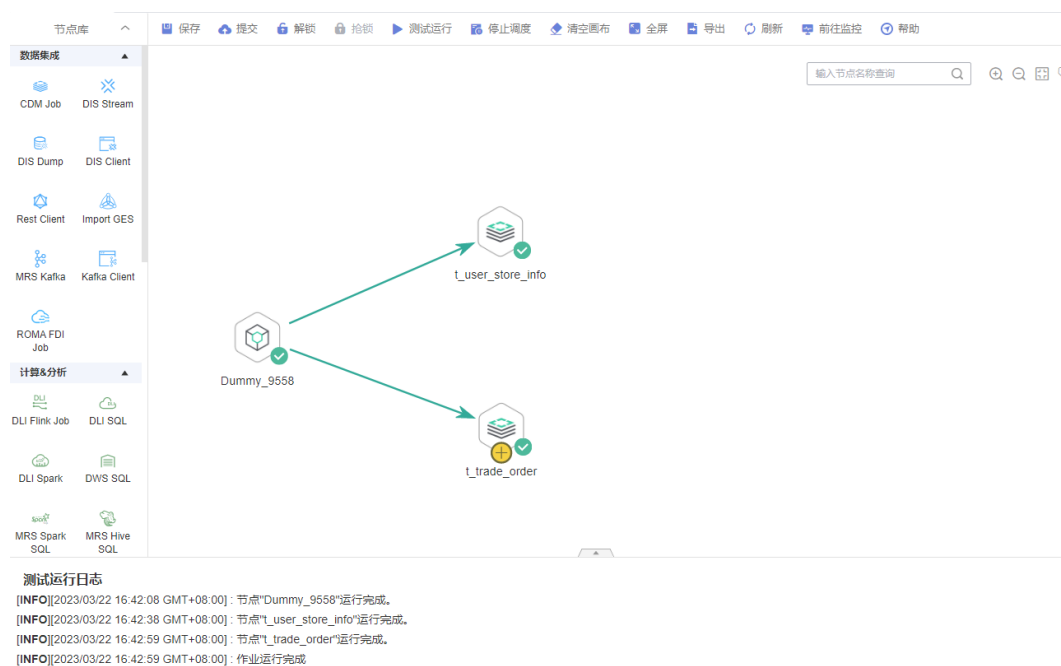
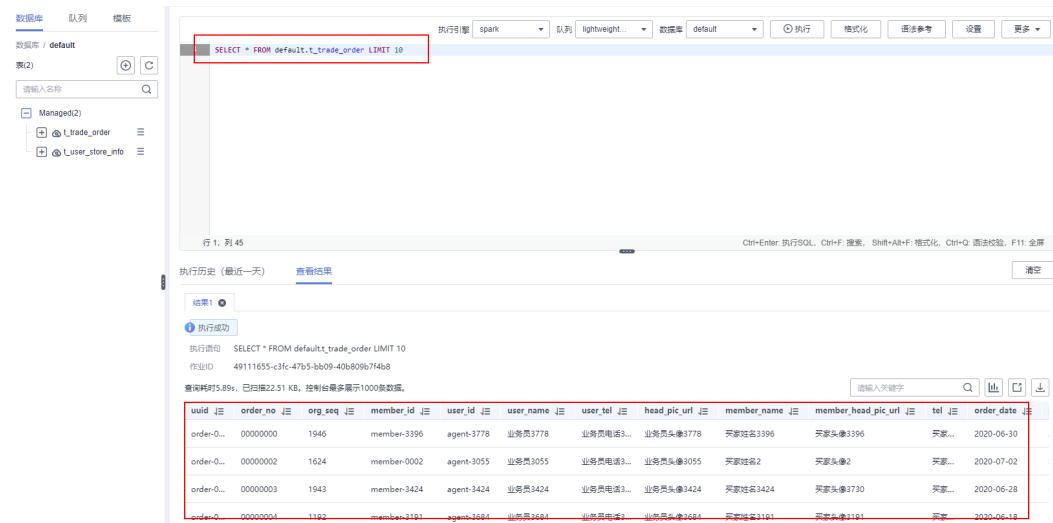


图 3-38 在 DLI 查看数据迁移结果



----结束

数据分析处理示例

步骤1 在完成数据迁移基础上，拖动1个DLI SQL节点到右侧的画布中，按下图填写配置。

图 3-39 DLI SQL 节点配置



对应的DLI SQL语句为（在这里实现业务定制化需求）：

```

DROP TABLE IF EXISTS dwd_sell_info;
CREATE TABLE IF NOT EXISTS dwd_sell_info AS
SELECT
  user_name,
  sum(total_money) as sell_total_money
FROM
  default.t_trade_order
group by
  user_name
    
```

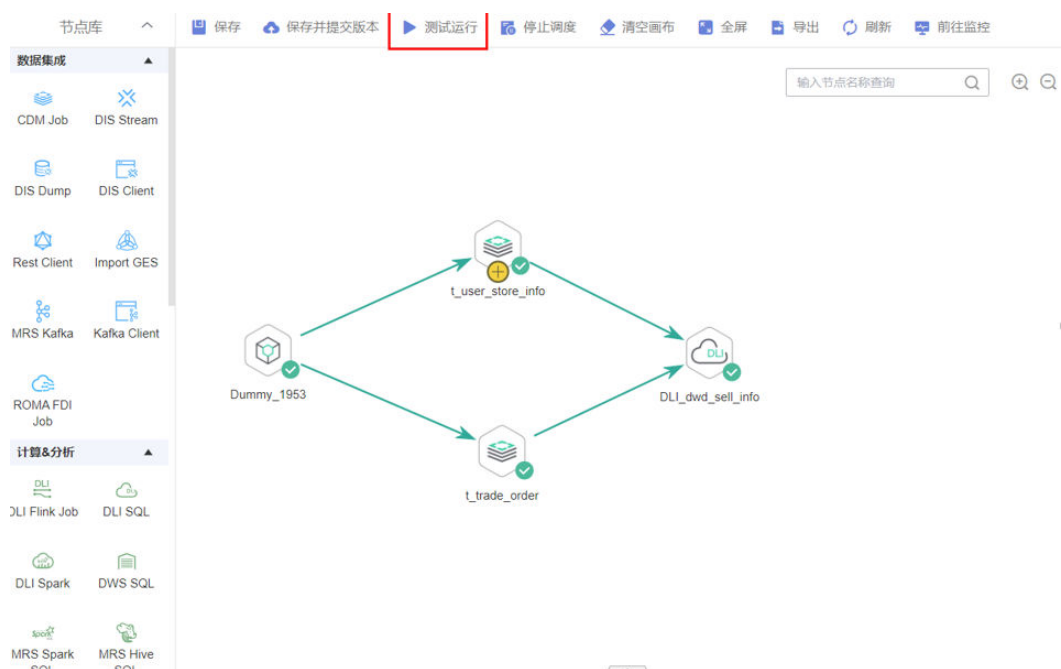
```
order by  
sell_total_money desc  
limit  
10;
```

步骤2 按照截图填好信息后，修改调度信息。单击“保存并提交版本”，单击“执行调度”，单击“测试运行”。

图 3-40 调度配置

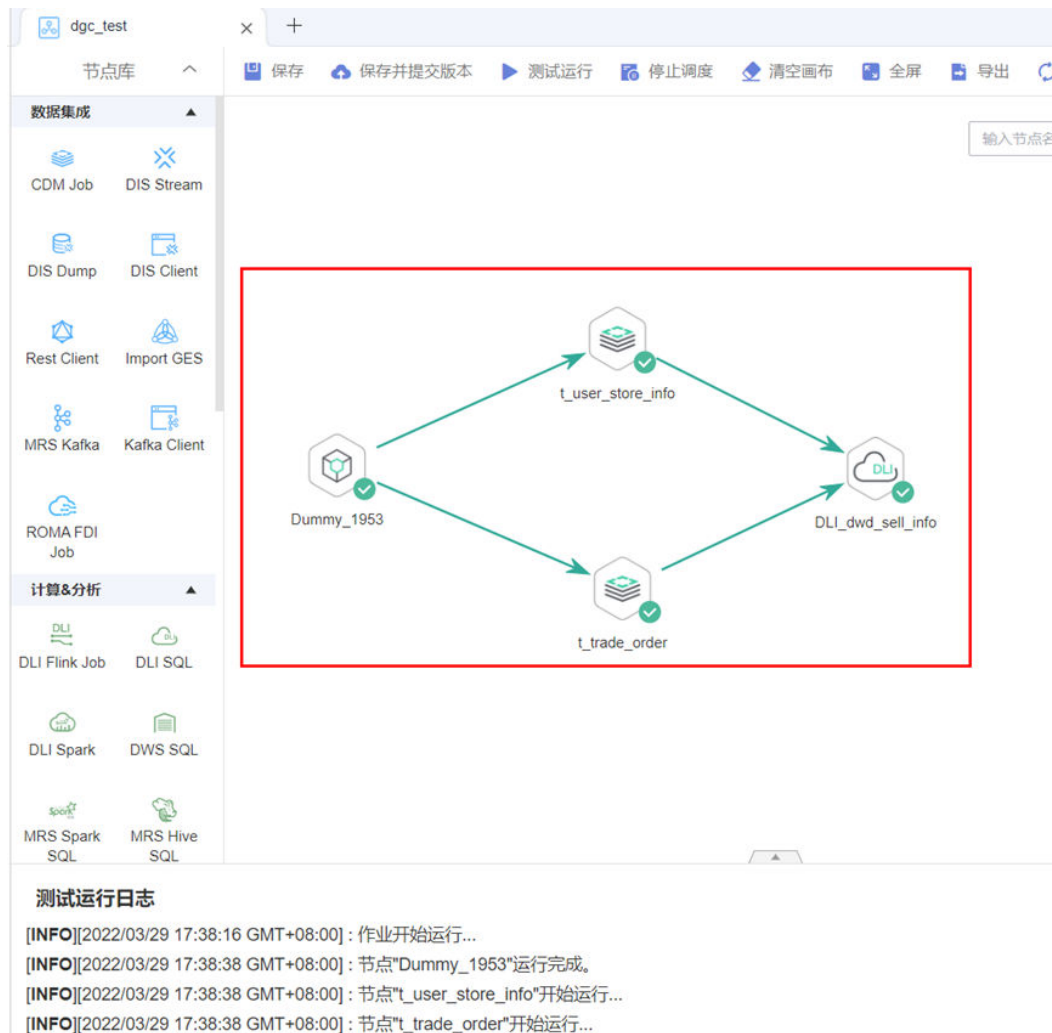


图 3-41 执行调度后测试运行



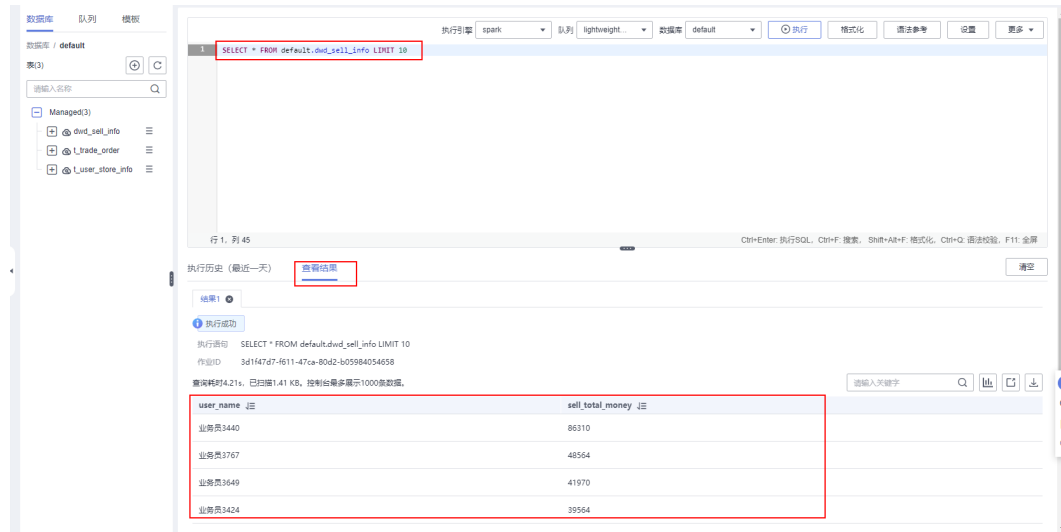
步骤3 执行CDM作业，将RDS的两张表“t_user_store_info”、“t_trade_order”的数据更新到DLI的表中，并做分析处理，作业执行成功。

图 3-42 执行成功



执行成功

图 3-43 在 DLI 查看数据分析处理结果



---结束

3.4 快速卸载

步骤1 登录应用编排服务AOS堆栈，找到该解决方案创建的堆栈，单击堆栈名称最右侧“删除”按钮，在弹出的“删除堆栈”提示框输入Delete，单击“确定”进行解决方案卸载。

图 3-44 一键删除



---结束

4 附录

名词解释

基本概念、云服务简介、专有名词解释

- 云数据迁移 CDM: **云数据迁移** (Cloud Data Migration, 简称CDM), 是一个高效、易用的批量数据迁移服务。CDM围绕大数据迁移上云和智能数据湖解决方案, 提供了简单易用的迁移能力和多种数据源到数据湖的集成能力, 降低了客户数据源迁移和集成的复杂性, 有效地提高您数据迁移和集成的效率。
数据湖探索 DLI: **数据湖探索** (Data Lake Insight, 简称DLI) 是完全兼容Apache Spark、Apache Flink、openLooKeng (基于Presto) 生态, 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, 支持数据入湖、数据仓库、BI、AI融合等能力。
- 数据治理中心 DataArts Studio: **数据治理中心**(DataArts Studio)是数据全生命周期一站式开发运营平台, 提供数据集成、数据开发、数据治理、数据服务、数据可视化等功能, 支持行业知识库智能化建设, 支持大数据存储、大数据计算分析引擎等数据底座, 帮助企业客户快速构建数据运营能力。
- 云数据库 RDS for MySQL: **云数据库** (简称RDS) 是一种基于云计算平台的即开即用、稳定可靠、弹性伸缩、便捷管理的在线云数据库服务。

5 修订记录

发布日期	修订记录
2023-03-30	第一次正式发布。