

## 企业数据经营分析实战操作指导

### 1 任务介绍

沃尔玛公司是一家美国的跨国零售企业，在保持高速发展的同时，沉淀了海量的销售数据。如何利用华为云数据湖探索服务从海量数据中发现数据的价值，是大数据应用在精准营销中的关键问题。

本实践利用的DLI Serverless Spark先对沃尔玛公开的销售数据进行简单的预处理，然后利用DLI SQL进行查询分析，最后利用永洪BI对数据进行可视化呈现。我们将这些未处理前的销售数据已经准备好并存储在华为云对象存储服务OBS中了。任务过程为：

- 1、获取原始数据；
- 2、上传数据至华为云OBS进行存储；
- 3、使用DLI Spark对原始数据去空处理；
- 4、使用DLI SQL对数据进行查询分析；
- 5、DLI对接BI进行分析结果的可视化呈现。

### 2 任务执行

#### 2.1 数据准备

##### 2.1.1 原始数据下载地址。

访问：<http://obs-salepredict.obs.cn-north-1.myhwclouds.com/index.html>



点击下载[OriginData.csv](#).

### 2.1.2 数据格式介绍

year_month_day_timestamp	时间戳
TOPyear	年
month_of_year	月
day_of_month	日
product_id	商品id
week	星期几
all_shop_number	涉及门店个数
sale_quantity	当日销售量

### 2.1.3 上传数据

1. 在华为云官网页面上方的导航栏，选择“产品”。在“存储”列表中，单击“对象存储服务 OBS”进入OBS产品页面。

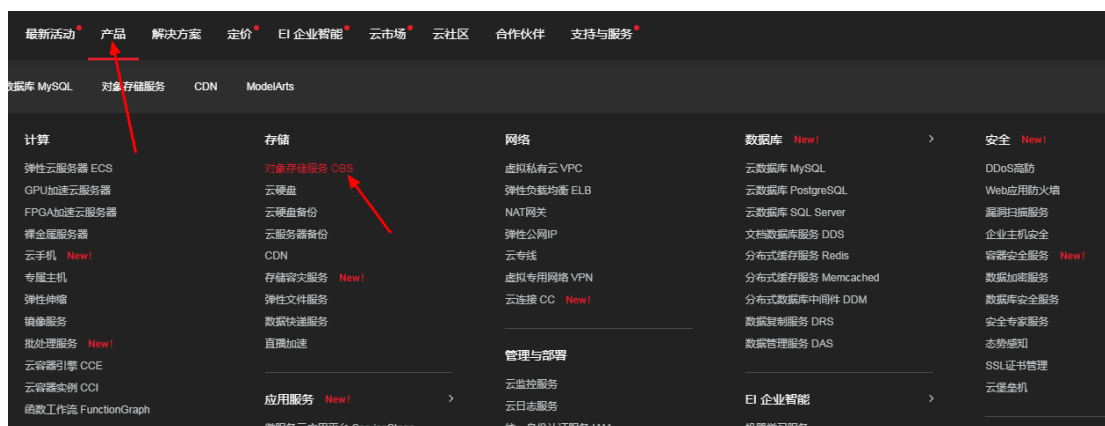


图 1 OBS产品入口

2. 在OBS产品页面，单击“进入控制台”进入华为云对象存储服务（OBS）控制台。在OBS控制台页面，单击右上角“创建桶”，进入创建桶页面，如下图所示，建立OBS桶，桶名称自定义即可，区域选择华北-北京一区。

创建桶

< 返回桶列表

\* 区域

华北-北京一

\* 桶名称

obs-salepredict

\* 存储类别

☒ 标准存储 ☐ 低频访问存储 ☐ 归档存储

适用于有大量热点文件或小文件，且需要频繁访问（平均一个月多次）并快速获取数据的业务场景。

\* 桶策略

☒ 私有 ☐ 公共读 ☐ 公共读写

桶的所有者拥有完全控制权限，其他用户在未经授权的情况下均无访问权限。

标签

所选区域不支持标签功能。

高级设置

不配置

立即配置

3. 将OriginData.csv上传到这个桶里。注意文件名按照本教程的填写，不能改成其他的名字，避免后续步骤出错。

## 2.2 数据处理

在本地编写Spark程序。因为原始数据中有些特征的空，我们写个Spark程序过滤掉，以此来熟悉DLI Spark的功能。程序Jar包已准备好并存储在OBS上，直接下载使用即可。

### 2.2.1 上传程序 Jar 包到 OBS

访问链接：<http://obs-salepredict.obs.cn-north-1.myhwclouds.com/index.html>

下载SalesDemo.jar，进入对象存储服务OBS上传jar包。



桶列表 > obs-salepredict



对象 已删除对象 碎片

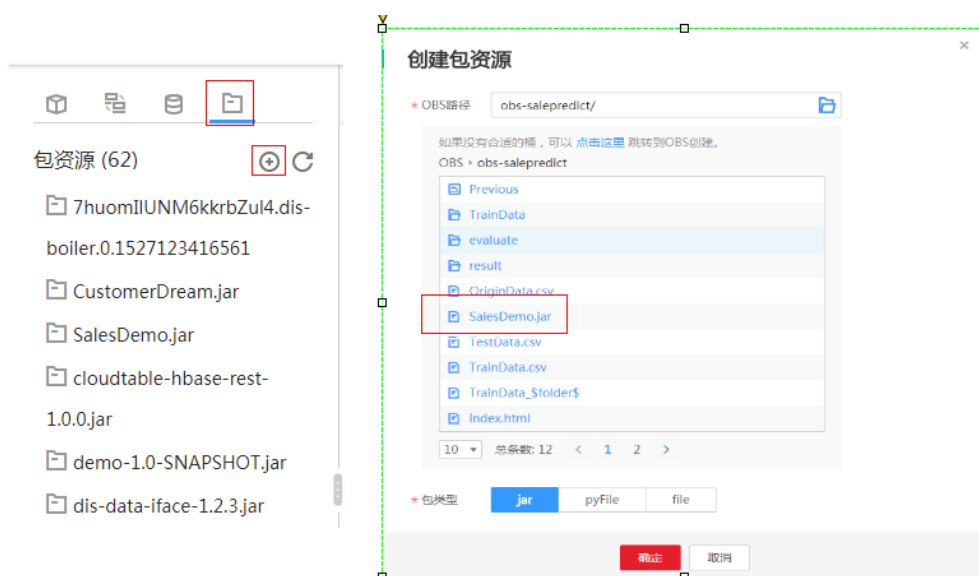
对象是数据存储的基本单位，在OBS中文件和文件夹都是对象。您可以上传任何类型（文本、图片、视频等）的文件，并在桶中对这些文件进行管理。 [了解更多](#)

[上传文件](#) [新建文件夹](#) [删除](#) [恢复](#)



<input type="checkbox"/>	名称	存储类别	大小	加密状态	恢复状态	最后修改时间
<input type="checkbox"/>	result	-	-	-	-	-
<input type="checkbox"/>	evaluate	-	-	-	-	-
<input type="checkbox"/>	OriginData.csv	标准存储	993.900 KB	未加密	-	2018/07/19 15:06:40 GMT
<input type="checkbox"/>	index.html	标准存储	3.083 KB	未加密	-	2018/07/19 15:00:37 GMT
<input type="checkbox"/>	TestData.csv	标准存储	30.327 KB	未加密	-	2018/07/19 10:17:28 GMT
<input type="checkbox"/>	TrainData.csv	标准存储	993.900 KB	未加密	-	2018/07/19 10:17:11 GMT

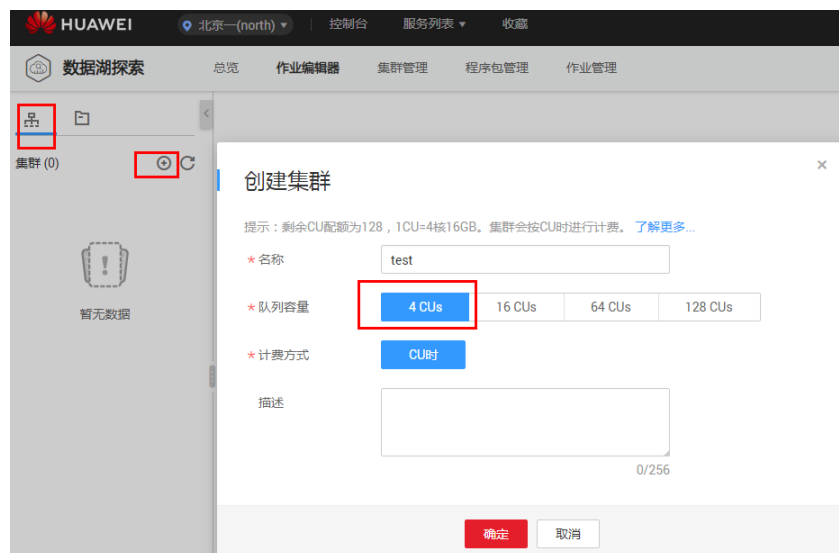
### 2.2.2 导入包到 DLI

访问<https://console.huaweicloud.com/dli/?region=cn-north-1#/dashboard>，进入DLI服务控制台，单击“Spark 作业”进入Spark作业页面，单击左侧导航栏中的图标，单击创建包资源，在创建包资源页面，选择前一步骤上传到OBS的jar包。



## 2.2.3 新建集群

单击左侧导航栏中的 ，单击  进入创建集群页面。参考下图创建一个4CU的集群。





## 2.2.4 新建 Spark 作业

在作业编辑器页面，选择程序包，主类填写：

com.huawei.www.SalesDemo，参数1：填写自己的AK，参数2：填写自己的SK，参数3：填写2.1.3中创建的OBS桶名。

AK和SK的获取，请访问


<https://console.huaweicloud.com/iam/?region=cn-north-1&locale=zh-cn#/myCredential> 页面，管理访问密钥，新增访问密钥创建AK,SK。请保存好生成的AK,SK，后续无法在华为云上查看AK,SK。



← → ↻ 🔒 https://console.huaweicloud.com/iam/?region=cn-north-1&locale=zh-cn#/myCredential

HUAWEI 控制台 服务列表 ▾ 收藏

**我的凭证**

  
更换

用户名: [redacted]

用户ID: af8[redacted]3

账号名: Y[redacted]

账号ID: ce[redacted]8850

已验证邮箱: [redacted] [绑定](#)

已验证手机: [redacted] [修改](#)

密码: 安全程度强 [强] [中] [弱] [修改](#)

登录验证方式: 关闭 [修改](#)

虚拟MFA: ● 未绑定 | [绑定](#)

项目列表

**管理访问密钥**

访问密钥对账号具有完全的访问权限，如果访问密钥泄露，会带来数据泄露风险，为了账号安全性，建议您定期更换并妥善保管访问密钥。

[新增访问密钥](#) 您还可以添加0个访问密钥。

最后单击 **执行** 按钮提交执行。

总览 作业编辑器 集群管理 程序包管理 作业管理 [购买CUI套餐包](#) [执行历史](#)

Spark gatetest **执行** [设置](#)

作业名称: 请输入作业名称

\* 程序包: SalesDemo.jar

\* 主类: com.huawei.www.SalesDemo

参数:

- URI: [redacted] [AK](#)
- pYuYeWtoJs: [redacted] [SK](#)
- obs-salepredict [OBS桶名](#)

## 2.2.5 查看作业状态

单击作业管理，进入作业管理页面，选择**Spark**就可以看到自己的作业的运行

状态和运行日志

数据湖探索 查询编辑器 数据管理 资源管理 模型管理 **作业管理** [立即下载](#) [常用链接](#)

SQL SPARK 所属集群: ALL 请输入作业ID 🔍

作业ID	作业名称	作业状态	所属集群	创建时间	运行时长	操作
25d1f362-b13f-49fd-9b1f-968a302b7e85	--	starting	gat_test	--	--	<a href="#">删除</a> <a href="#">提交日志</a> <b>运行日志</b>

提交日志:

```
tracking URL: https://lqy-cluster-30-001226001/proxy/application_1529384265897_0401/
user: mls | org.apache.spark.internal.Logging$class.logInfo(Logging.scala:54)
2018-07-19 21:53:29.512 | INFO | pool-1-thread-1 | Shutdown hook called | org.apache.spark.internal.Logging$class.logInfo(Logging.scala:54)
2018-07-19 21:53:29.513 | INFO | pool-1-thread-1 | Deleting directory /tmp/spark-a241aa2-6d9d-42b5-a4d8-11de5d1d37c2 |
org.apache.spark.internal.Logging$class.logInfo(Logging.scala:54)


stderr:
YARN Diagnostics:
```

## 2.2.6 查看 OBS 上是否生成了数据



## 2.3 数据查询与分析

### 2.3.1 新建队列

访问<https://console.huaweicloud.com/dli/?region=cn-north-1#/dashboard>，进入DLI服务控制台，单击“SQL作业”进入SQL作业页面，选择左侧导航栏的，单击创建一个4CU的队列。





## 创建队列

提示：剩余CU配额为96，1CU=4核16GB。[了解更多...](#)

\* 队列名称

default

\* 队列容量

4 CUs

16 CUs

64 CUs

128 CUs

\* 计费方式

CU时



描述

0/256

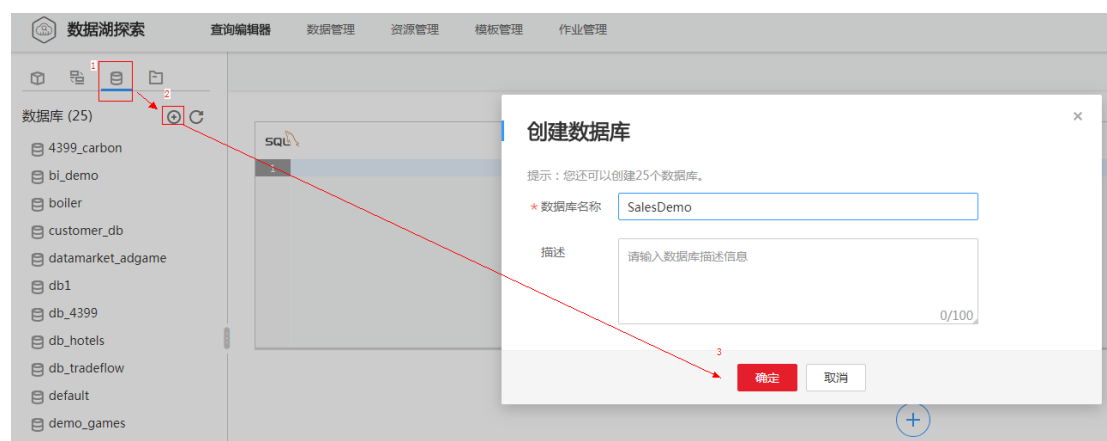
确定

取消

### 2.3.2 新建数据库

在左侧导航栏中，单击 切换到数据库，单击 新建数据库，输入数据库名称

SalesDemo

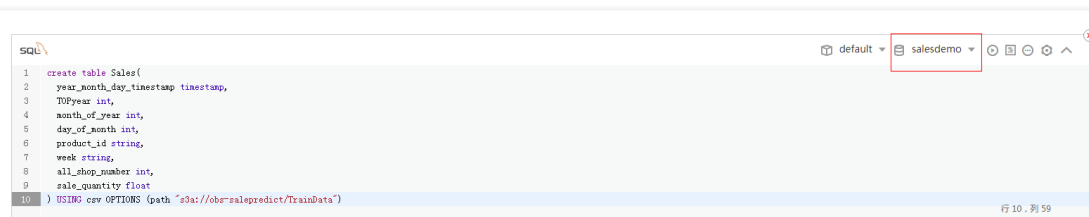


### 2.3.3 新建表


建表提供图形化的方式和SQL语句的方式。首先介绍通过SQL语句创建表的方式

式。单击“作业编辑器”，在编辑器栏选择创建的数据库salesdemo，在编辑器中输入以下SQL，注意SQL中的OBS桶名修改为自己创建的桶名：

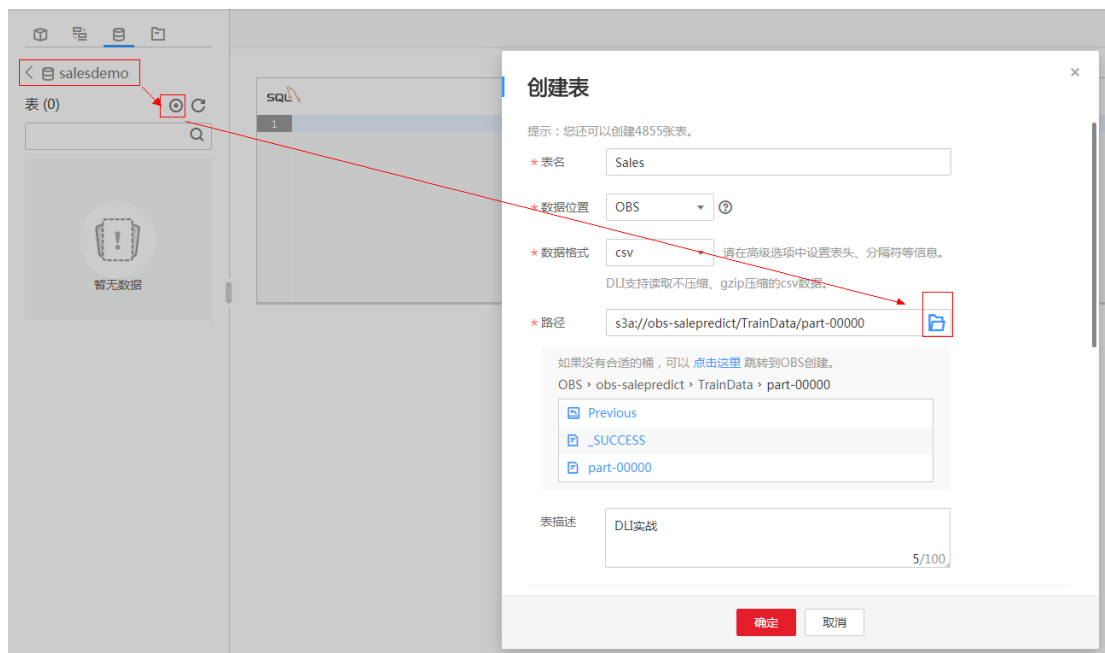
```
create table Sales(
    year_month_day_timestamp long,
    TOPyear int,
    month_of_year int,
    day_of_month int,
    product_id string,
    week string,
    all_shop_number int,
    sale_quantity float
) USING csv OPTIONS (path "s3a://obs-salepredict/TrainData",header
'true' )
```



执行即可。

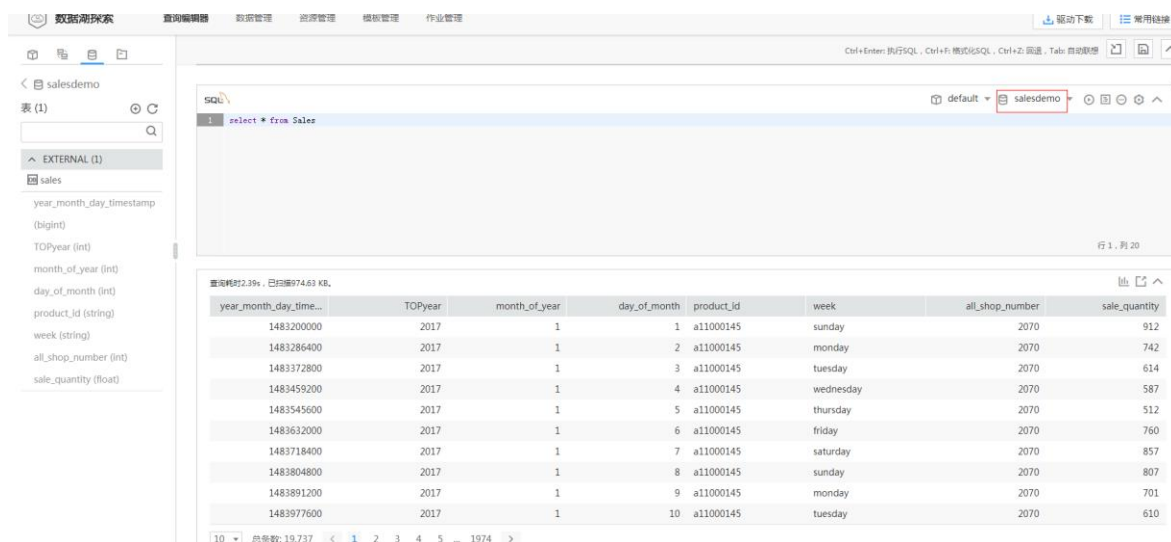
也可以使用图形化界面创建表，双击左侧导航栏中创建的数据库，然后单击创建表。

如下图创建表。



### 2.3.4 数据查询

单击“**作业编辑器**”，在编辑器栏选择创建的数据库salesdemo，在编辑器中输入SQL查询表中的数据。



## 2.4 数据展示

DLI对接到BI进行数据的展示，该步骤只给大家**做个演示**，大家不用做实际操作。访问<http://49.4.0.83:9090/bi/>

### 2.4.1 添加数据源

打开BI页面后，单击“添加数据源”，添加一个GENERIC类型的数据源。参考下图填写数据源内容。驱动填写com.huawei.dli.jdbc.DliDriver，URL为jdbc:luxor://dli.cn-north-

1.myhuaweicloud.com/cbb9c2d6c9d1474f9a50f39baff62337?queueName=test;databaseName=SalesDemo;authenticationMode=aksk;accessKey=**ak**;secretKey=**sk**;regionName=cn-north-1;serviceName=dli

将URL中accesskey, secretkey的值替换成自己的ak,sk。 QueueName ,  
databasename的值替换成之前创建的队列和数据库。点击测试连接, 能看到  
测试成功, 然后点击上方保存。注意ak,sk,queueName,databasename替换时  
URL中不要有多余的空格, 否则测试连接时会报错。

### 数据库

选择数据源: GENERIC \*

☐ 仅对有写权限的用户可见

### 连接属性

驱动: com.huawei.dli.jdbc.DliDriver \*

URL: 自定义协议 jdbc:luxor://dli.cn-north-1.myhuaweicloud.com/cbb9c2d6 \*

服务器登录: 无身份验证 \*

用户名:

密码:

最大连接数: 10

别名类型:

默认数据库: SalesDemo

### 编码转换

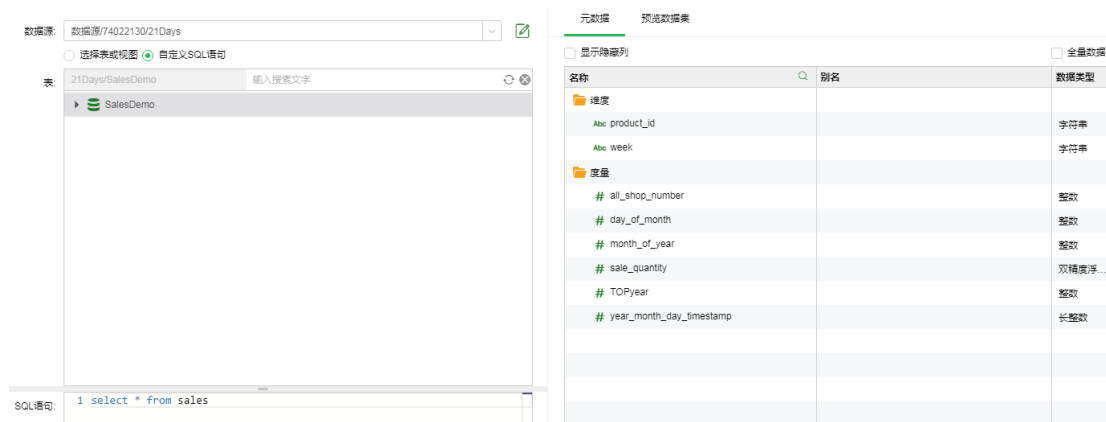
数据库编码:

转换编码:

测试连接

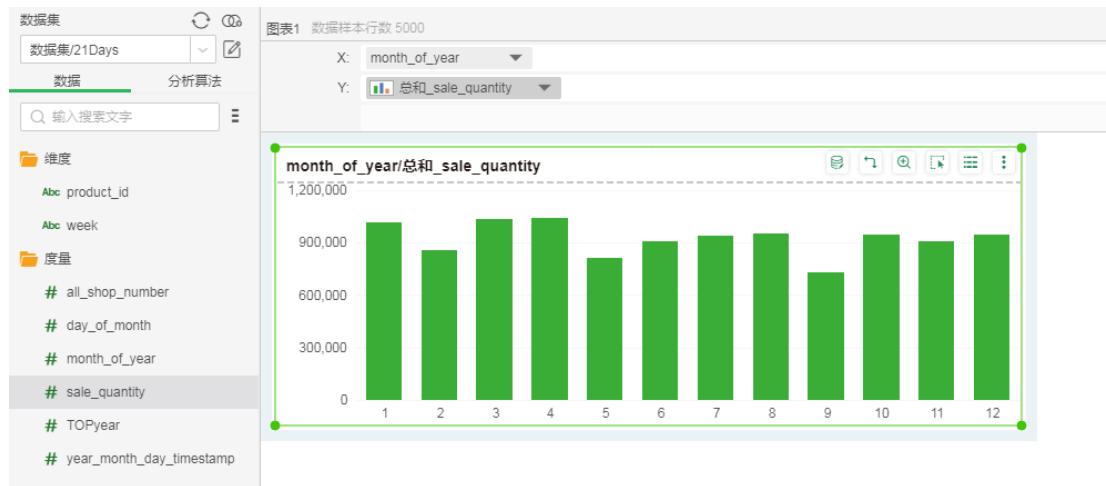
表结构模式:

## 2.4.2 添加数据集



## 2.4.3 制作报告

选择数据集，选择柱状图组件，将month\_of\_year拖放到X轴并转换成维度组，sale\_quantity托放入Y轴。



## 2.4.4 查看报告

选择查看报告，选择我的仪表板，点击21Days.

或者点击该链接

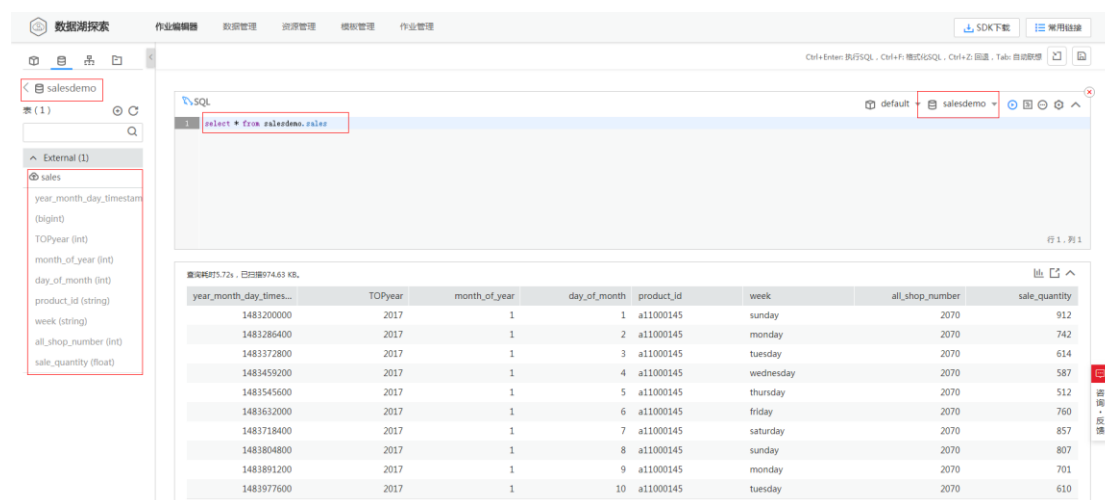
[http://49.4.0.83:9090/bi/?proc=1&action=viewer&hback=true&db=\\_\\_M](http://49.4.0.83:9090/bi/?proc=1&action=viewer&hback=true&db=__M)

[Y\\_DB\\_!2f!21Days.db&isAir=false&browserType=chrome](#)

### 3 打卡任务

#### 3.1 任务描述

在数据湖探索服务能看到自己建立的表，并能查询到结果，参考2.3.4中结果，如下图。



#### 3.2 打卡要求

对包含查询结果的页面进行截图，截图中需包含学员的华为云用户名信息，如下图所示。



HUAWEI

北京—(north)

控制台

服务列表

收藏

数据湖探索

作业编辑器

数据管理

资源管理

模板管理

作业管理

SDK下载

常用链接

数据湖探索

表 (1)

External (1)

sales

year\_month\_day\_timestamp (bigint)

TOPYear (int)

month\_of\_year (int)

day\_of\_month (int)

product\_id (string)

week (string)

all\_shop\_number (int)

sale\_quantity (float)

SQL

select \* from salesdemo.sales

行 1, 列 1

查询耗时: 5.72s, 已扫描 974.63 KB

year_month_day_times...	TOPYear	month_of_year	day_of_month	product_id	week	all_shop_number	sale_quantity
1483200000	2017	1	1	a11000145	sunday	2070	912
1483286400	2017	1	2	a11000145	monday	2070	742
1483372800	2017	1	3	a11000145	tuesday	2070	614
1483459200	2017	1	4	a11000145	wednesday	2070	587
1483545600	2017	1	5	a11000145	thursday	2070	512
1483632000	2017	1	6	a11000145	friday	2070	760
1483718400	2017	1	7	a11000145	saturday	2070	857
1483804800	2017	1	8	a11000145	sunday	2070	807
1483891200	2017	1	9	a11000145	monday	2070	701
1483977600	2017	1	10	a11000145	tuesday	2070	610