

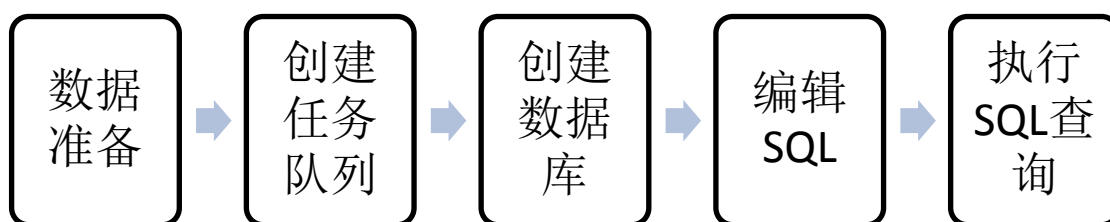
轻松探索数据背后的价值 – 数据湖探索的使用

1 任务介绍

华为云数据湖探索服务 (**Data Lake Insight** , 简称DLI) 是完全托管的数据分析服务, 用户无需管理任何服务器, 即开即用; 服务支持标准SQL, 完全兼容Spark生态接口, 提供云上多态数据的统一分析能力。

本实践主要学习DLI服务的**基本数据分析功能**, 实现如何利用DLI服务**对芝加哥地区的酒店消费水平进行分析**。

2 任务过程



3 任务执行

3.1 数据准备：上传原始数据至华为云 OBS 存储

1. 测试数据下载。

访问：[http://obs-salepredict.obs.cn-north-](http://obs-salepredict.obs.cn-north-1.myhwclouds.com/index.html)

[1.myhwclouds.com/index.html](http://obs-salepredict.obs.cn-north-1.myhwclouds.com/index.html)

点击chicago.csv下载原始数据。

2. 进入华为云对象存储服务（OBS）控制台

进入 OBS 控制台之前，请先注册并登录华为云账号。免费注册链接：

<https://reg.huaweicloud.com/registerui/public/custom/register.html?>

[locale=zh-cn#/register](https://reg.huaweicloud.com/registerui/public/custom/register.html?locale=zh-cn#/register)

2.1 在华为云官网页面上方的导航栏，选择“产品”。

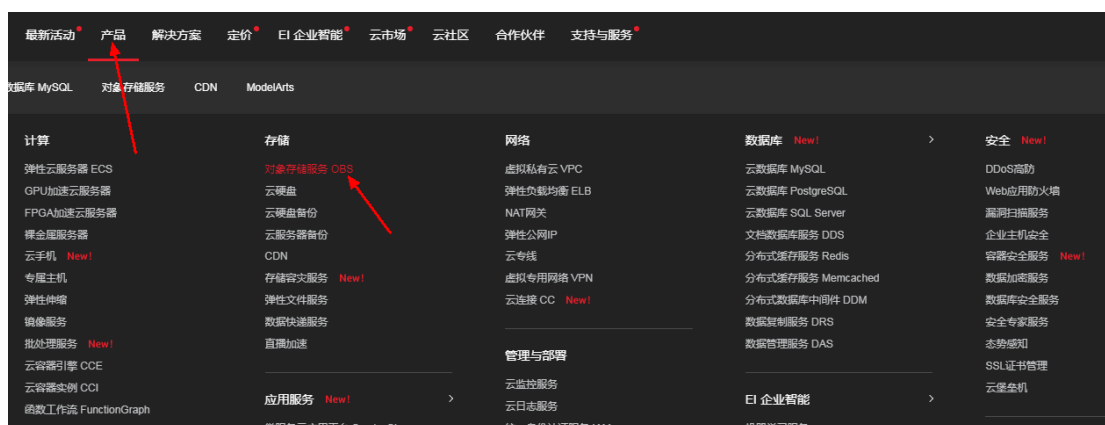


图 1 OBS产品入口

2.2 在“存储”列表中，单击“对象存储服务 OBS”进入 OBS 产品页面。

或直接点击链接进入 OBS 产品页面：

<https://www.huaweicloud.com/product/obs.html>

2.3 在“对象存储服务 OBS”页面，单击“进入控制台”。

3. 创建桶

进入对象存储服务控制台页面，单击右上角“创建桶”，进入如下图所示的创建桶页面。首次使用的用户，需**先进行实名认证**。单击页面右下角“去认证”链接（如下图）按照提示完成认证。

创建桶 < 返回桶列表

* 区域 华北-北京一

* 桶名称 obs-da64

* 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。

高级设置

按用量收费
创建免费，使用阶段按照用量收费。

您尚未实名认证，开通云服务需要先进行实名认证。去认证>>

图 2 创建桶

参考图 2，“区域”选择“华北-北京一”，“存储类别”为“标准存储”，“桶策略”为“私有”，输入“桶名称”或使用默认桶名称。单击“立即创建”即可完成创建桶操作。

4. 上传数据

OBS 桶创建完成之后，新创建的桶名会出现在桶列表中，单击所创建的桶名。进入桶“概览”页面。

单击左侧列表中的“对象”。单击新建文件夹，创建一个存放文件的目录。单击创建的目录名，选择“上传文件”，将要处理的文件 chicago.csv 上传到该目录下。

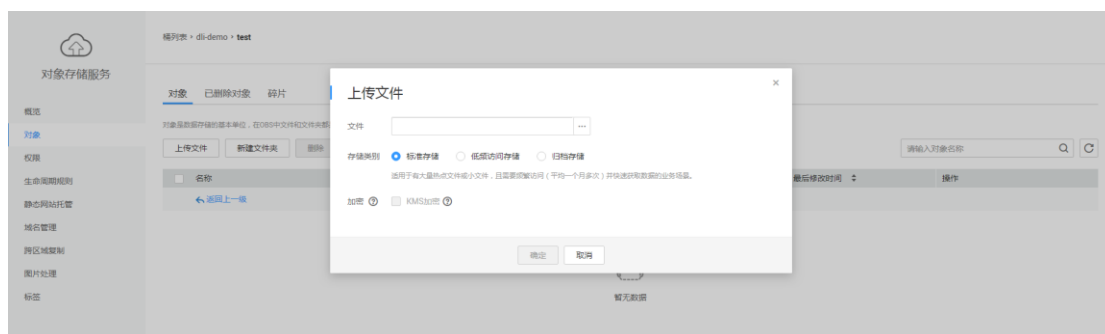


图 3 上传文件到OBS桶

以上操作也可通过 OBS Browser 完成，OBS Browser 下载链接：

<https://static.huaweicloud.com/upload/files/tools/OBSBrowser.zip>

OBS Browser 使用指导参考：

https://support.huaweicloud.com/clientogw-obs/obs_03_0064.html

3.2 创建 DLI OBS 表

1. 进入数据湖探索 (DLI) 控制台

1.1 在华为云官网页面上方的导航栏，选择“EI 企业智能”。

1.2 在“EI 大数据”列表中，单击“数据湖探索”。

或直接点击链接：

<https://www.huaweicloud.com/product/dli.html>



1.3 在“数据湖探索 DLI”页面，单击“进入控制台”。

2. 创建队列

2.1 进入 DLI 控制台，单击 SQL 作业的“**创建作业**”超链接（如图 4），进入 SQL 作业的“作业编辑器”界面。



图 4 DLI控制台

2.2 在左侧导航栏中，单击  选择队列，单击 （如图 5），参考图 6 创建容量为 4CU 的队列 test。

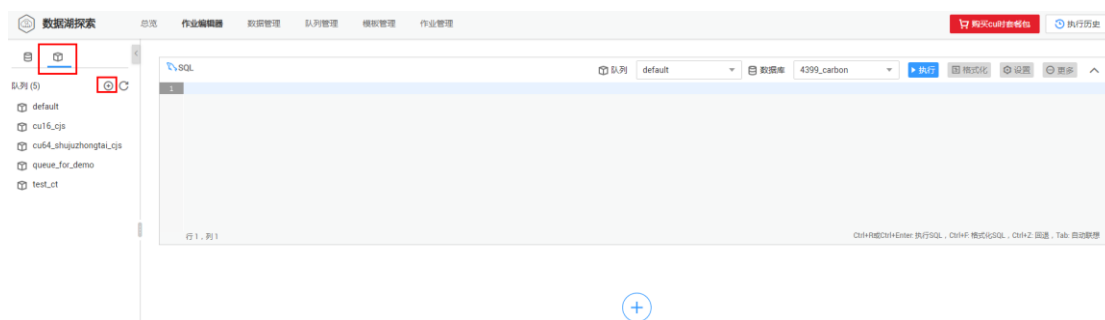


图 5 创建队列操作

×

创建队列

提示：剩余CU配额为95，1CU=4核16GB。[了解更多...](#)

★ 队列名称

test

★ 队列容量

4CU

16CU

64CU

★ 计费方式

按CU时

按SQL计算量

描述



0/256

确定

取消

图 6 创建队列界面

3. 创建数据库

在“作业编辑器”左侧导航栏中，单击  选择数据库，单击 （如图 7）创建数据库 DB1。

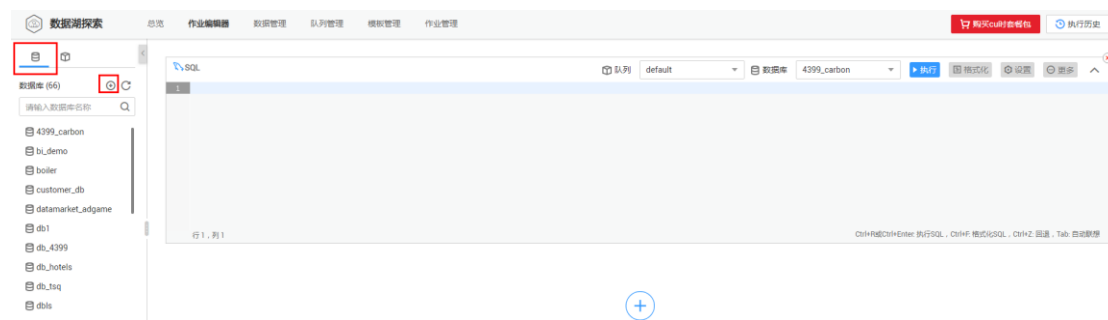


图 7 创建数据库操作

创建数据库

提示：您还可以创建27个数据库。

* 数据库名称

DB1

描述

test

4/100


确定

取消

图 8 创建数据库

数据库创建成功后，创建的数据库 DB1 会显示在左侧导航栏“数据库”列表中。

4. 创建 OBS 表

4.1 单击数据库 DB1，单击“表”右侧，参考图 9 创建表 chicago。

创建表

提示：您还可以创建4873张表。

* 表名

chicago

* 数据位置

OBS

?

* 数据格式

csv

请在高级选项中设置表头、分隔符等信息。

DLI支持读取不压缩、gzip压缩的csv数据。

* 路径

s3a://obs-712/part-00000

当前不支持中国东北区OBS的数据。

表描述

test

4/100

* 列名

room_id

* 类型

long

列描述

请输入列描述

操作

* 列名

survey_id

* 类型

int

列描述

请输入列描述

操作

确定

取消

图 9 创建表

其中，

- ◆ 数据位置：OBS
- ◆ 数据格式：csv
- ◆ “路径”为 [3.1 章节 4](#) 中在 OBS 所上传的数据所在路径，格式为：s3a://bucket_name/dir_name（s3a://桶名称/目录名称）
- ◆ 输入所建表中的“列名”和“类型”。具体如下：

列名	类型
room_id	long

survey_id	int
host_id	long
room_type	string
country	string
city	string
borough	string
neighborhood	string
reviews	int
overall_satisfaction	float
accommodates	int
bedrooms	float
bathrooms	string
price	float
minstay	string
last_modified	string
latitude	double
longitude	double

location	string
----------	--------

单击“确定”后，在左侧“表”下，将显示所建表“chicago”。

也可直接在 SQL 作业编辑窗口中，输入 SQL 建表语句建表，语句内容如下（注意 path 路径改为自己的 csv 文件路径）：

```
CREATE TABLE los_angeles(
    room_id long, survey_id int, host_id long, room_type string,
    country string, city string, borough string, neighborhood string,
    reviews int, overall_satisfaction float, accommodates int,
    bedrooms float, bathrooms string, price float, minstay string,
    last_modified string, latitude double, longitude double, location
    string
) USING csv OPTIONS (path "s3a://bucket_name/dir_name" )
```

选择创建的队列和数据库，单击编辑窗口右上的执行图标，执行建表语句创建表，如下图。

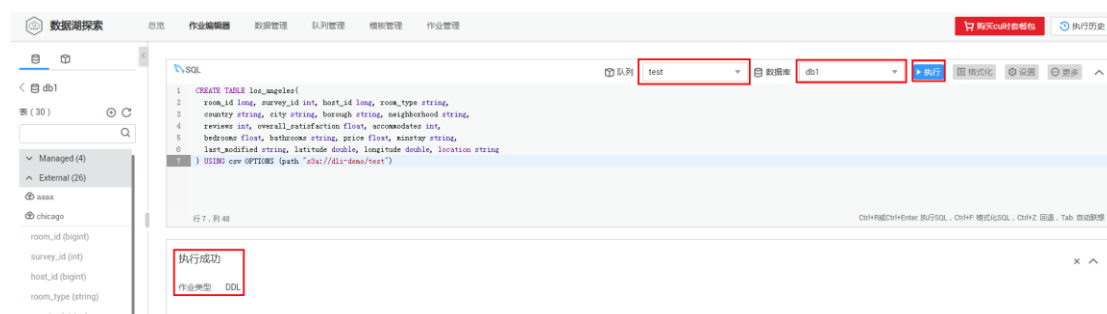


图 10 通过SQL创建表

3.3 SQL 编辑与查询

单击上节创建的表名 “chicago” ，会在表下显示其列名，双击表名，会在右侧的SQL编辑窗口中显示输入的查询语句，根据需要对查询语句进行修改。例如输入：

```
select
    case when price <= 50 then '1 (<50)'
        when price >50 and price <=100 then '2 (50-100)'
        when price >100 and price <=150 then '3 (100-150)'
        when price >150 and price <=200 then '4 (150-200)'
        when price >200 and price <=250 then '5 (200-250)'
        when price >250 and price <=300 then '6 (250-300)'
        when price >300 and price <=500 then '7 (300-500)'
        else '8 (>500)' end as price_level,
    count(*) / (select count(*) from chicago) as percentage
from
    db1.chicago
group by price_level
order by price_level
```

输入查询语句后，在编辑器上方选择所使用的队列和数据库，单击执行，执行查询语句，等待一段时间，DLI 将会返回查询结果，如图 11 所示。

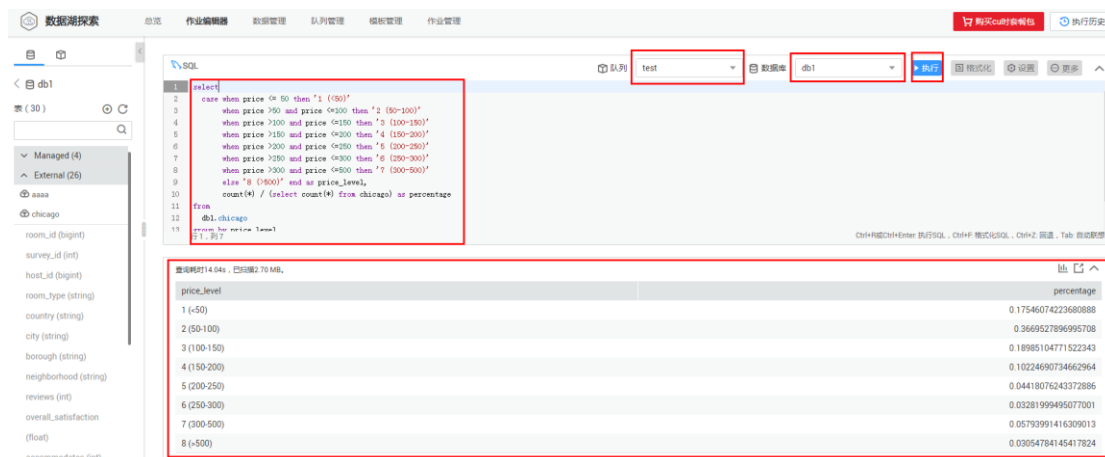


图 11 SQL查询界面

单击执行结果窗口右上的  图标，以图表方式展示查询结果，如图 12 所示。

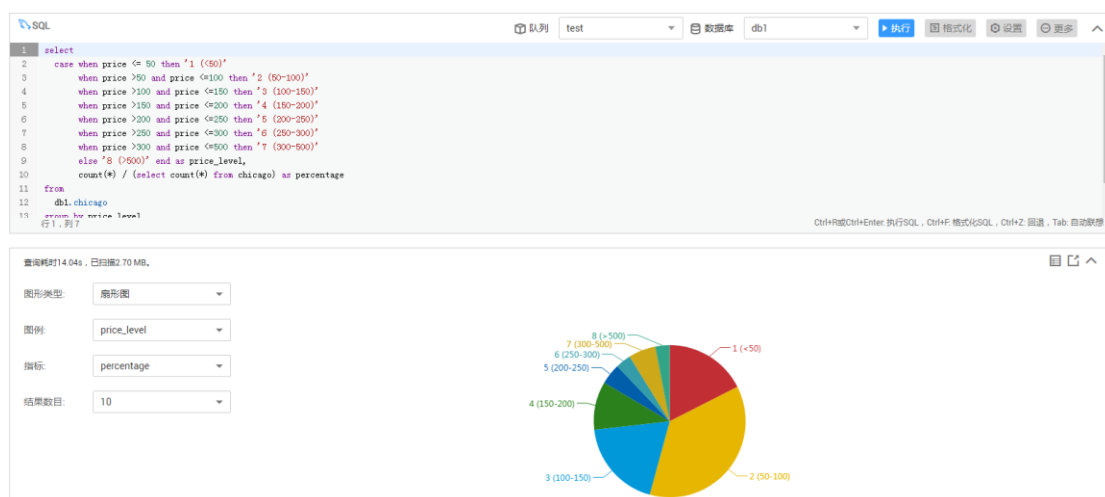


图 12 以图表方式展示查询结果

从图中可见，50~100 美元价格的酒店房间占比最大。低于 50 美元和 100~150 美元价格的酒店房间占比相当。

4 打卡任务

任务描述：熟悉 DLI 的操作界面，参考任务执行过程，使用 DLI 服务计算各个价格区间的酒店房间数量的百分比，以扇形图展示，并截图反馈结果。

打卡要求：对包含扇形图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。

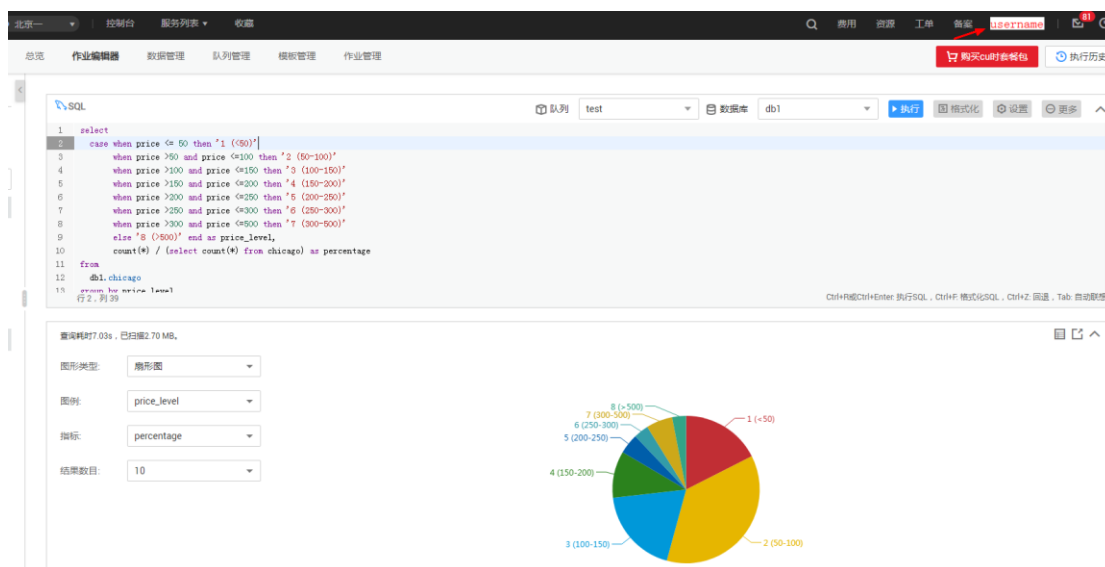


图 13 打卡截图参考

附录：数据湖探索服务用户指南

https://support.huaweicloud.com/usermanual-dli/zh-cn_topic_0067628621.html