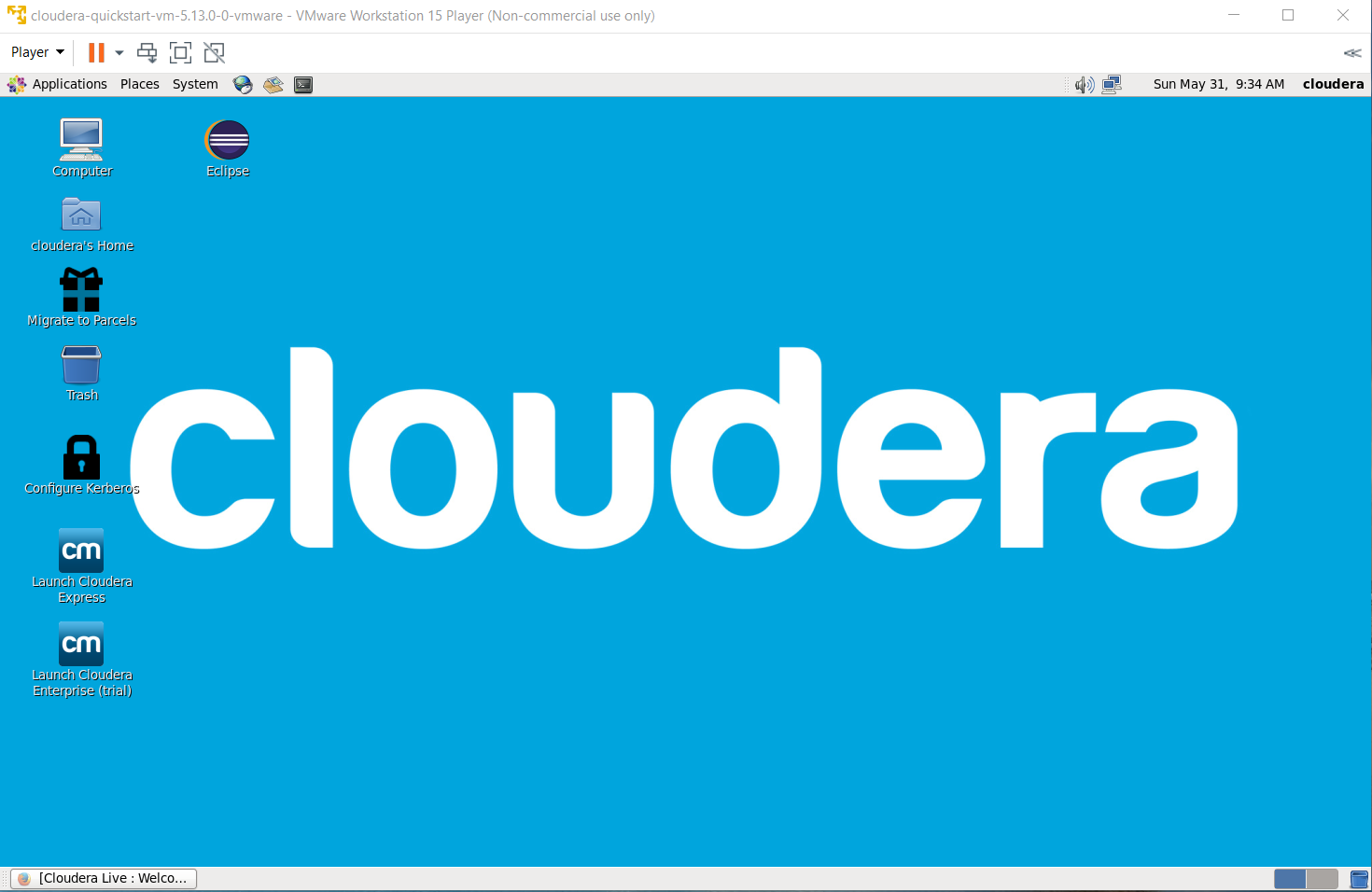
**درس : Big Data استاد : خانم دکتر صفری**

**دانشجو : رقیه یزدانی تمرین : Hive**

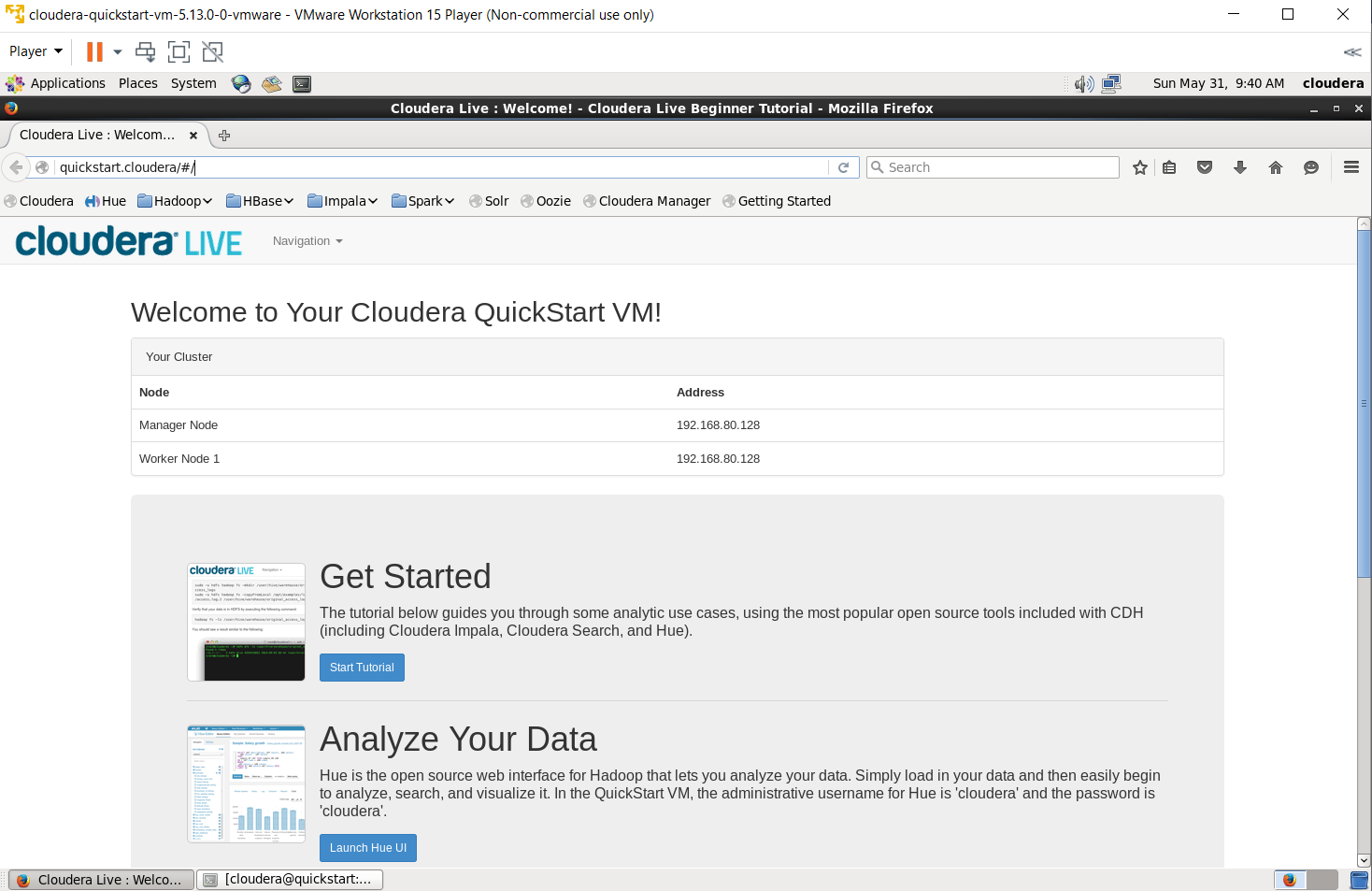
# Hive

# ایجاد بستر لازم برای امکان کار با Hive

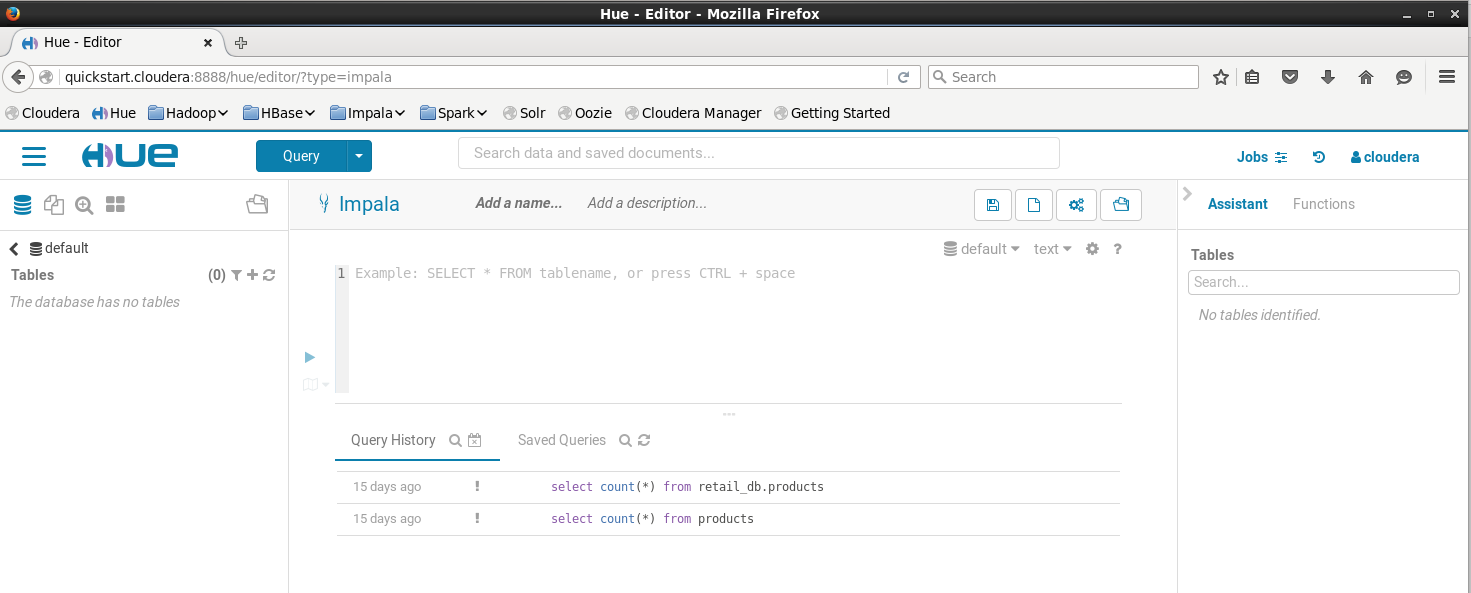
برای انجام این تمرین و تمرینهای بعدی، فایل cloudera-quickstart-vm-5.13.0-0-vmware که درواقع ماشین مجازی cloudera می‌باشد، با حجم 5.6 GB دانلود و برای امکان استفاده از آن، نرم افزار VMWare Player 15.5 با حجم 141 MB دانلود و نصب گردید. پس از بالا آوردن VM با استفاده از VMWare Player، محیط آن بصورت زیر می‌باشد. سیستم عامل آن لینوکس CentOS 6.7 می‌باشد و مجموعه ای از نرم افزارهای حوزه Big Data از جمله Hadoop و Hive و Hbase و ... روی آن نصب و پیکربندی شده و آماده استفاده می‌باشد :



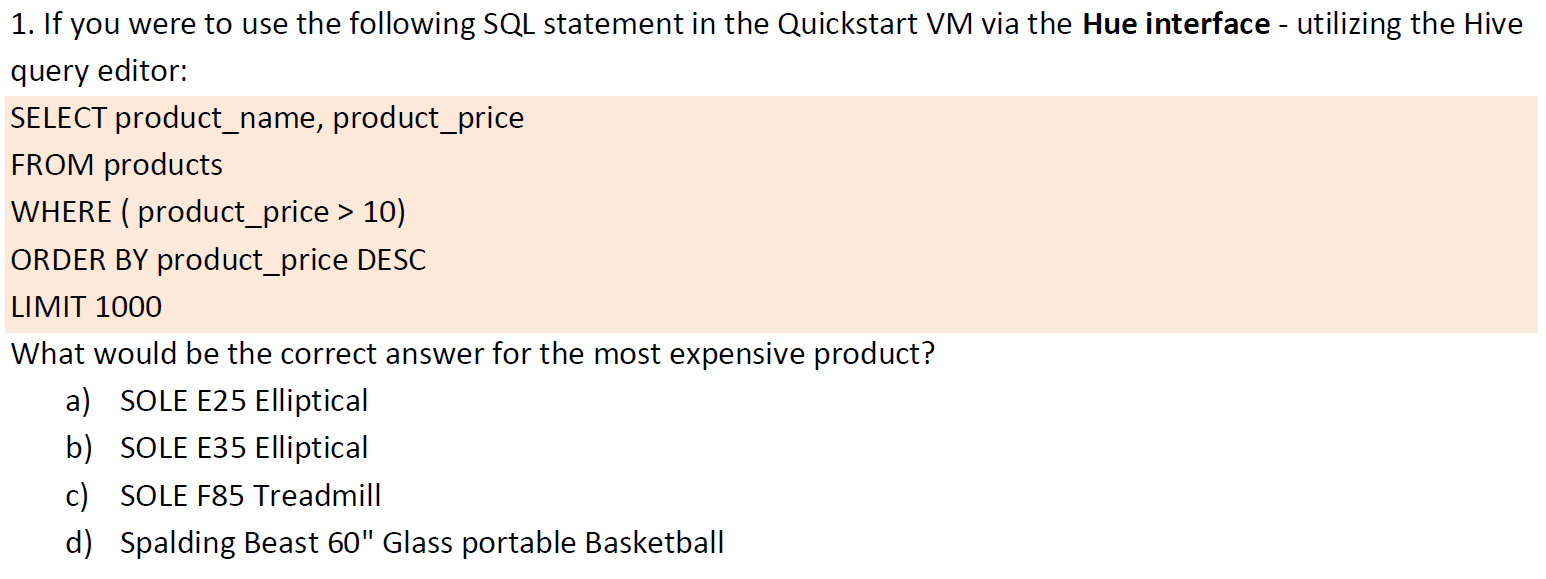
پس از اتصال به کنسول web کلودرا که بصورت زیر می‌باشد، صفحه ای مشابه زیر نمایش داده می‌شود :



با کلیک بر روی آیکن Hue و وارد کردن کلمه cloudera به عنوان user و password وارد محیط Hue interface می‌شویم :

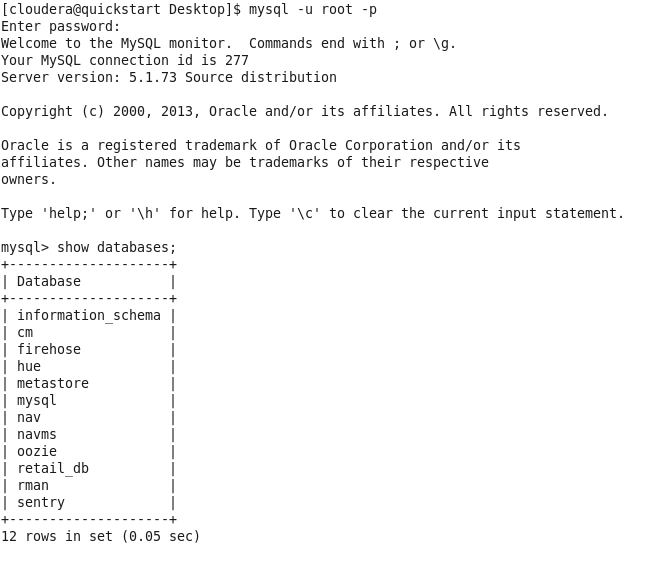


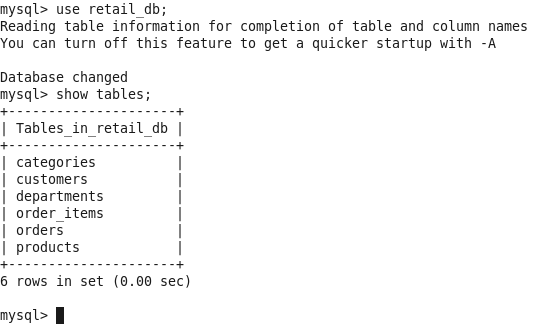
# تمرین اول :



برای انجام این تمرین، در واقع نیاز به دیتابیس reatil\_db می‌باشد که یکی از جداول آن products است و در query بالا استفاده شده است.

روی ماشین مجازی کلودرا، دیتابیس رابطه ای mysql از قبل نصب شده و تعدادی دیتابیس sample در آن وجود دارد که یکی از آنها reatil\_db است. در شکلهای زیر، مراحل اتصال به دیتابیس mysql و اطمینان از وجود دیتابیس retail\_db در آن نشان داده شده است : (پسورد دیتابیس cloudera می باشد)





برای امکان کار با این دیتابیس در Hive، لازم است که ابتدا این اطلاعات به Hive منتقل شود. پس ابتدا از طریق ترمینال به Hive وصل شده و یک دیتابیس به این نام می‌سازیم:

$ hive

hive> show databases;

hive> create database retail\_db;

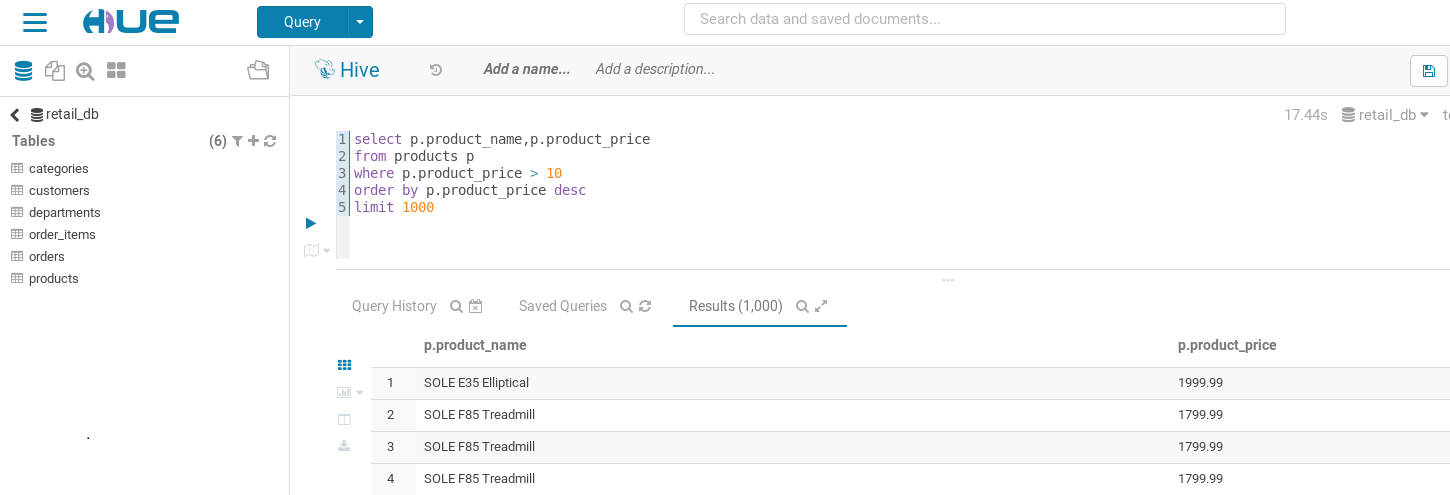
سپس باید اطلاعات را از mysql به این دیتابیس جدید، import ‌کنیم. برای اینکار از ابزار sqoop استفاده می‌کنیم. این ابزار که بصورت command-line می‌باشد، برای انتقال اطلاعات از دیتابیس رابطه‌ای به hadoop کاربرد دارد. در واقع نام آن هم برگرفته از sql+hadoop می‌باشد. دستور زیر را اجرا می‌کنیم:

$ sqoop import-all-tables –connect jdbc:mysql://localhost/retail\_db –username root

--password cloudera –hive-import –hive-database retail\_db

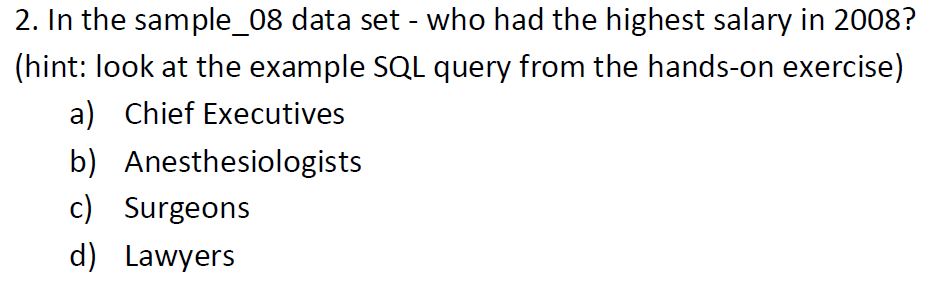
حال در اینترفیس Hue از مسیر زیر، وارد بخش hive شده و پس از انتخاب دیتابیس reatil\_db، query مربوط به این تمرین را نوشته و اجرا می‌کنیم. شکل زیر نتیجه اجرا را نشان می‌دهد :

Hue Interface 🡪 Query 🡪 Editor 🡪 Hive



بنابراین گزینه دوم (SOLE E35 Elliptical) صحیح است که گرانترین محصول دارای قیمت بالای 10 می‌باشد.

# تمرین دوم :



برای حل این تمرین، نیاز به دیتاست sample\_08 می‌باشد. برای اینکار با دستورات زیر، این دیتاستها را دانلود می‌کنیم :

$ cd /tmp

$ wget <https://raw>.githubusercontent.com/HortonworksUniversity/Security\_Labs/master/labdata/sample\_07.csv

$ wget <https://raw>.githubusercontent.com/HortonworksUniversity/Security\_Labs/master/labdata/sample\_08.csv

سپس مشابه تمرین قبل، به CLI Hive متصل شده و دستورات زیر را اجرا می‌کنیم. در این دستورات، ابتدا جداول ساخته شده و سپس دیتا از فایلهای csv دانلود شده در مرحله قبل، به آنها کپی می‌شود :

-- create table

CREATE TABLE sample\_07 (

code string,

description string,

total\_emp int,

salary int )

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘\t’

STORED AS TextFile;

--load data into table

load data local inpath ‘/tmp/sample\_07.csv’ into table sample\_07;

-- create another table

CREATE TABLE sample\_08 (

code string ,

description string ,

total\_emp int ,

salary int )

ROW FORMAT DELIMITED

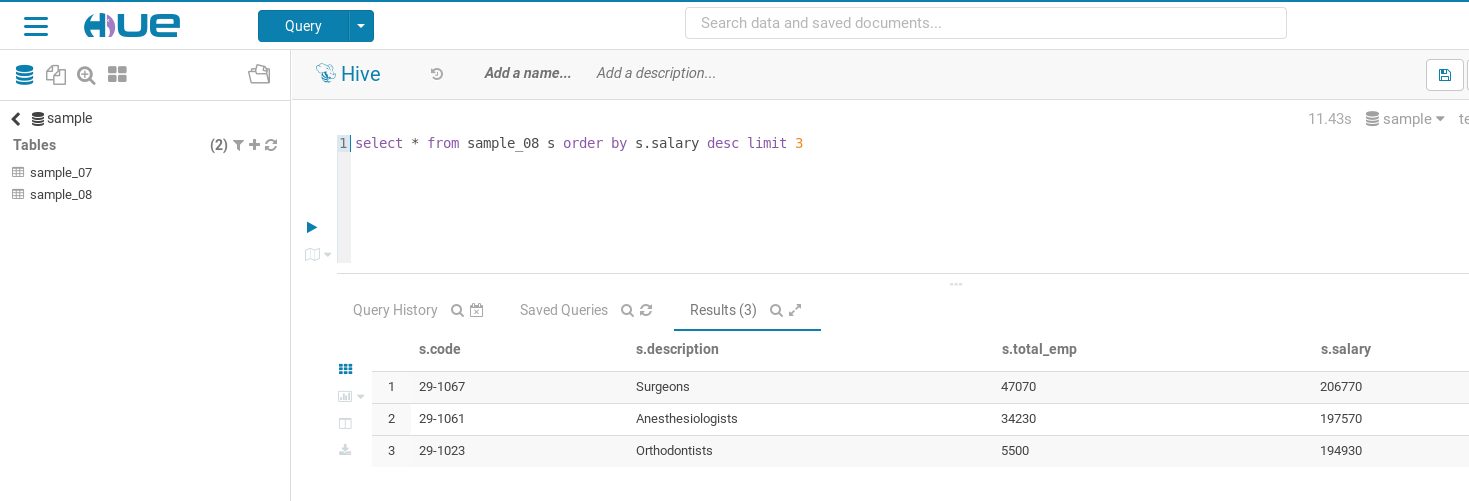
FIELDS TERMINATED BY ‘\t’

STORED AS TextFile;

--load data into a table

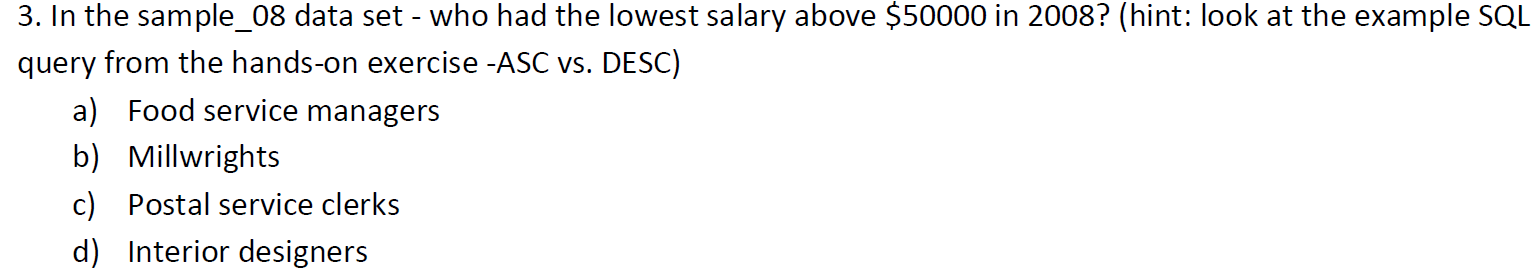
load data local inpath ‘/tmp/sample\_08.csv’ into table sample\_08;

حال برای پاسخ به این تمرین که بالاترین حقوق در سال 2008 را در دیتاست sample\_08 خواسته است، query زیر را اجرا می‌کنیم:

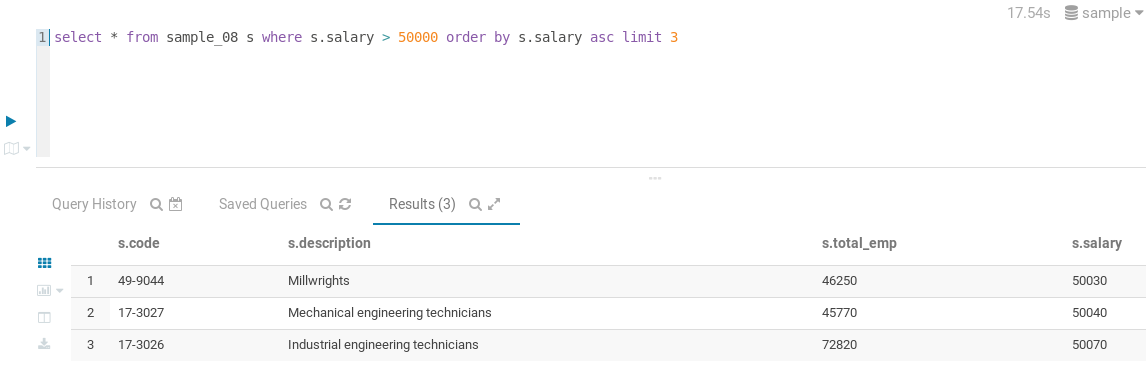


بنابراین با توجه به نتیجه query بالا، پاسخ تمرین، گزینه سوم یعنی Surgeons می‌باشد.

# تمرین سوم :

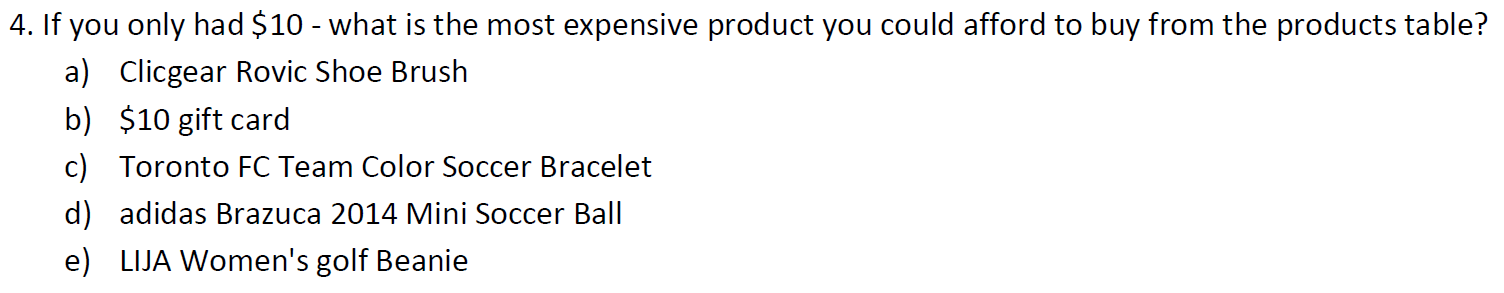


برای پاسخ به این سوال که چه کسی پایین ترین حقوق بالای 50000 دلار را در سال 2008 داشته است، query زیر را اجرا می‌کنیم :

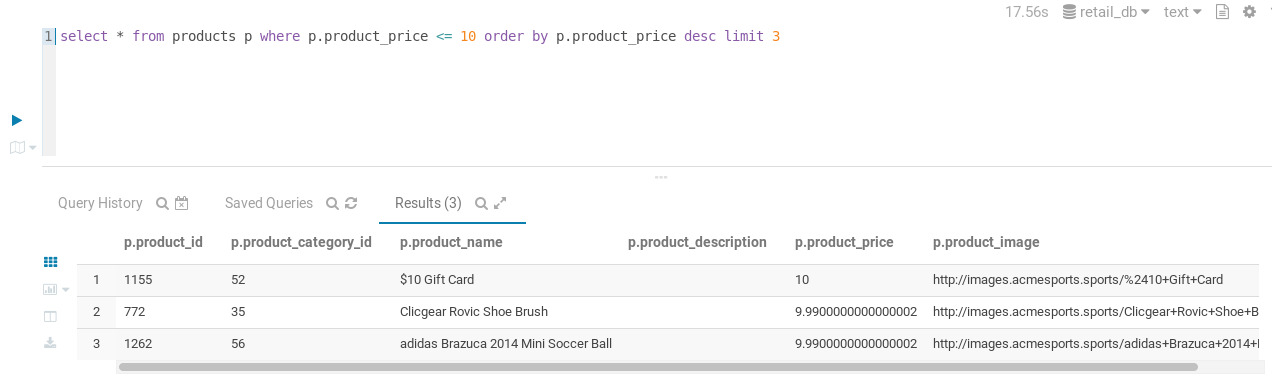


بر اساس نتیجه query، پاسخ سوال، گزینه دوم صحیح است.

# تمرین چهارم :



برای پاسخ به این سوال، query زیر نوشته و اجرا شد، که در نتیجه آن، گزینه دوم (10$ gift card) صحیح است.



برای ادامه تمرینها نیاز به دیتاست bayareabikeshare می‌باشد که مشابه قبل، ابتدا فایلهای csv مربوطه، دانلود شد، سپس به ازای هر فایل، یک جدول در hive ایجاد گردید و بعد دیتای فایل مربوطه در آن import شد. یکی از تفاوتهای این اسکریپتها نسبت به قبل، این است که در این دیتاست، از کاراکتر ‘,’ بجای ‘\t’ به عنوان جداکننده استفاده شده و همچنین پس از ایجاد جدول، با دستور alter table، تنظیم شد تا اولین سطر از فایل csv که header می‌باشد هنگام import در نظر گرفته نشود. در ادامه، اسکریپتها و همچنین تصویر اجرای آن آورده شده است :

$ wget <https://raw.githubusercontent.com/udacity/data-analyst/master/projects/bike_sharing/201402_trip_data.csv>

$ wget <https://raw.githubusercontent.com/udacity/data-analyst/master/projects/bike_sharing/201408_trip_data.csv>

$ wget <https://raw.githubusercontent.com/udacity/data-analyst/master/projects/bike_sharing/201502_trip_data.csv>

برای هر یک از فایلهای csv دانلود شده، باید مراحل زیر انجام شود. اسکریپتها، برای اولین فایل است. برای دوفایل بعدی هم باید عملیات بصورت مشابه انجام شود :

-- create table

CREATE TABLE 201402\_trip\_data (

trip\_id string,

duration int,

start\_date string,

start\_station string,

start\_terminal string,

end\_date string,

end\_station string,

end\_terminal string,

bike string,

subscription\_type string,

zip\_code string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TextFile;

ALTER TABLE 201402\_trip\_data SET TBLPROPERTIES ("skip.header.line.count"="1");

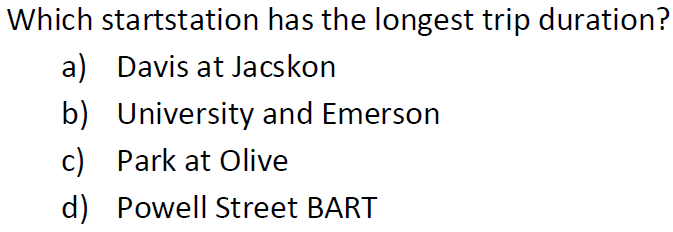
--load data into table

load data local inpath '/tmp/201402\_trip\_data.csv' into table 201402\_trip\_data;

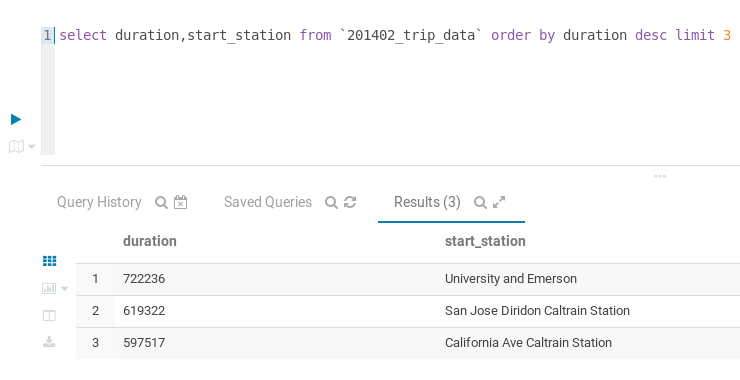
تصویر اجرای اسکریپت در hive



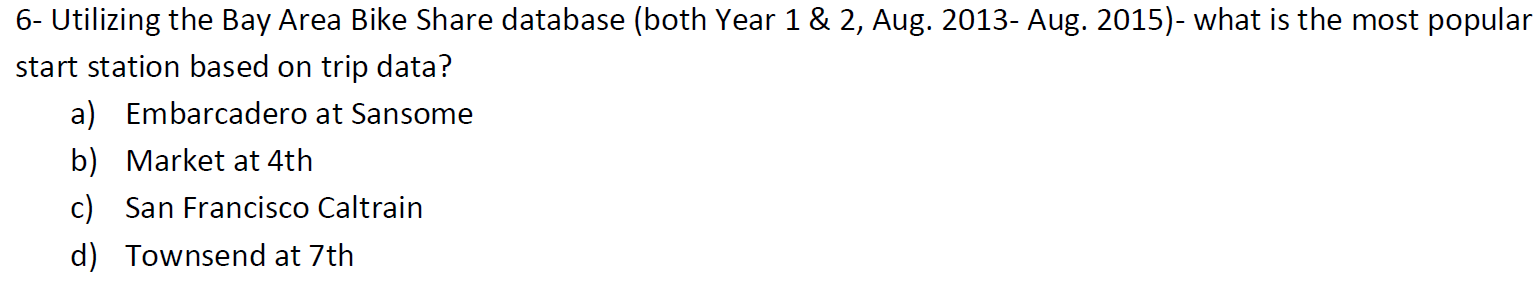
# تمرین پنجم :



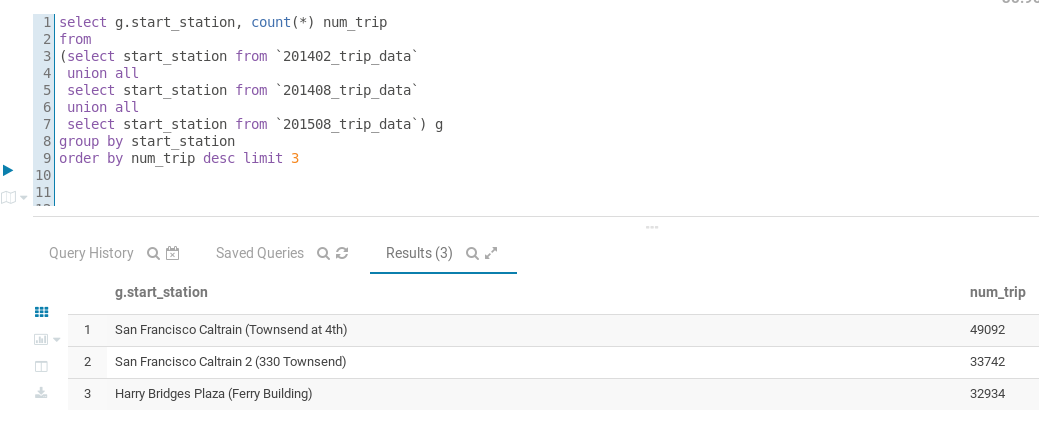
برای پاسخ به این سوال، اسکریپت زیر نوشته و اجرا شد، که بر اساس نتیجه اجرا، پاسخ صحیح گزینه دوم(University and Emerson) می باشد.



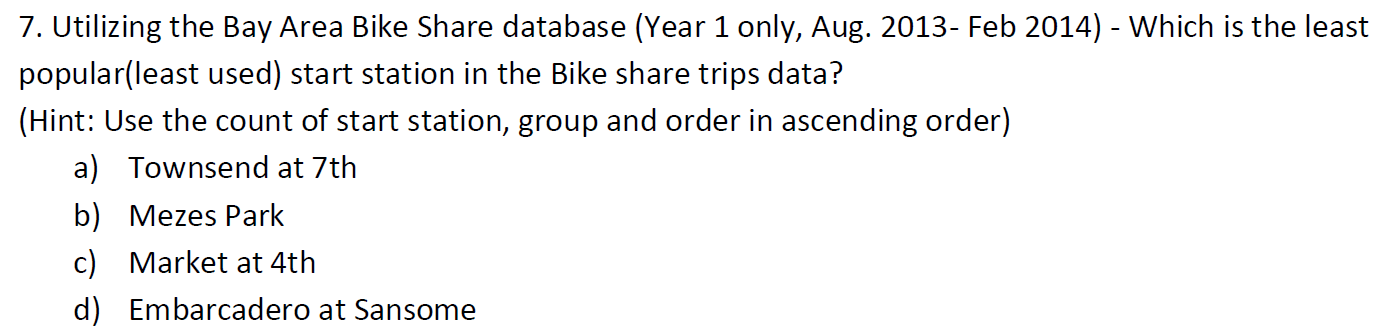
# تمرین ششم :



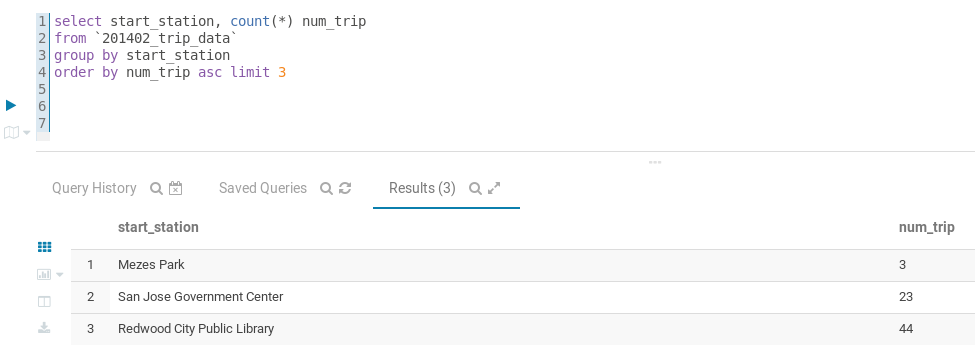
برای انجام این تمرین؛ لازم است که با داده های هر سه جدول کار کنیم؛ ابتدا ستون start\_station از هر سه جدول، union شده است و سپس با استفاده از group by و count تعداد سفر شروع از هر ایستگاه بدست آمده است. با توجه به نتیجه query؛ گزینه سوم (San Francisco Caltrain) صحیح است.



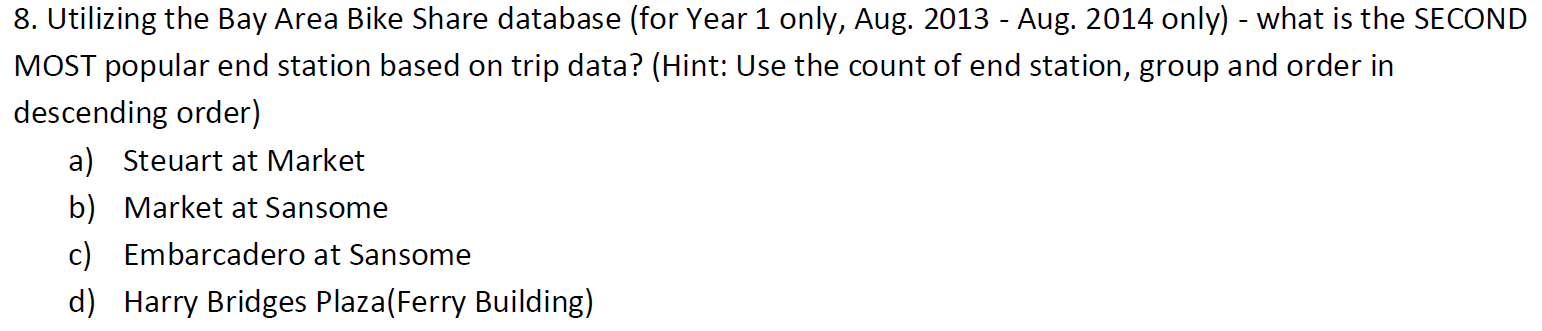
# تمرین هفتم :



بر اساس نتیجه query مربوطه، گزینه دوم (Mezes Park) صحیح است.



# تمرین هشتم :



برای این تمرین، ابتدا ستون end\_station از دو جدول مربوط به محدوده زمانی تعیین شده، با هم union شده و سپس با استفاده از group by و count و همچینین مرتب کردن نزولی، پاسخ سوال بدست امده است که بر اساس خروجی query، گزینه سوم (Embarcadero at Sansome) صحیح است.

