درس : Big Data

استاد : خانم دکتر صفری

دانشجو : رقیه یزدانی

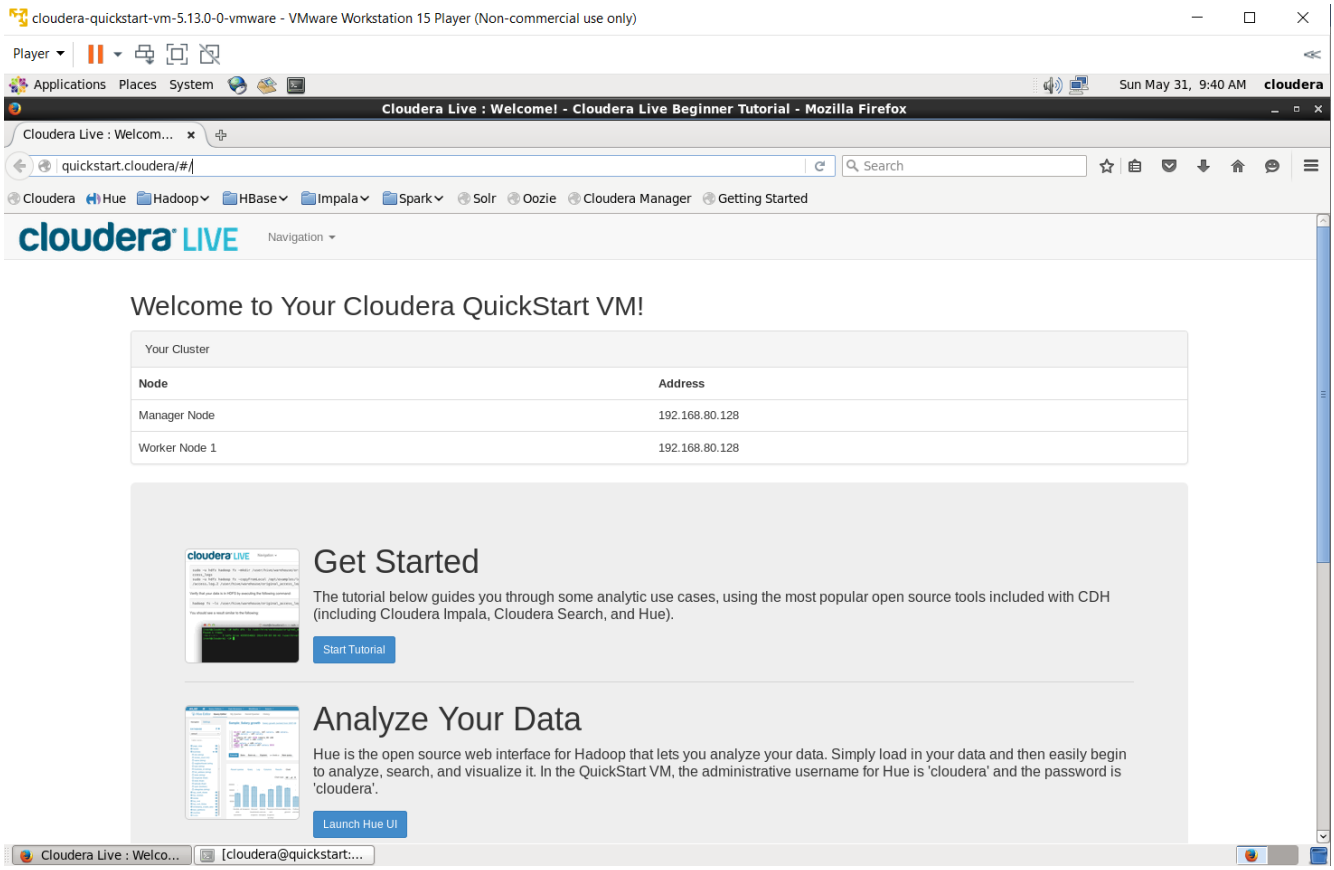
تمرین : Hive

Hive

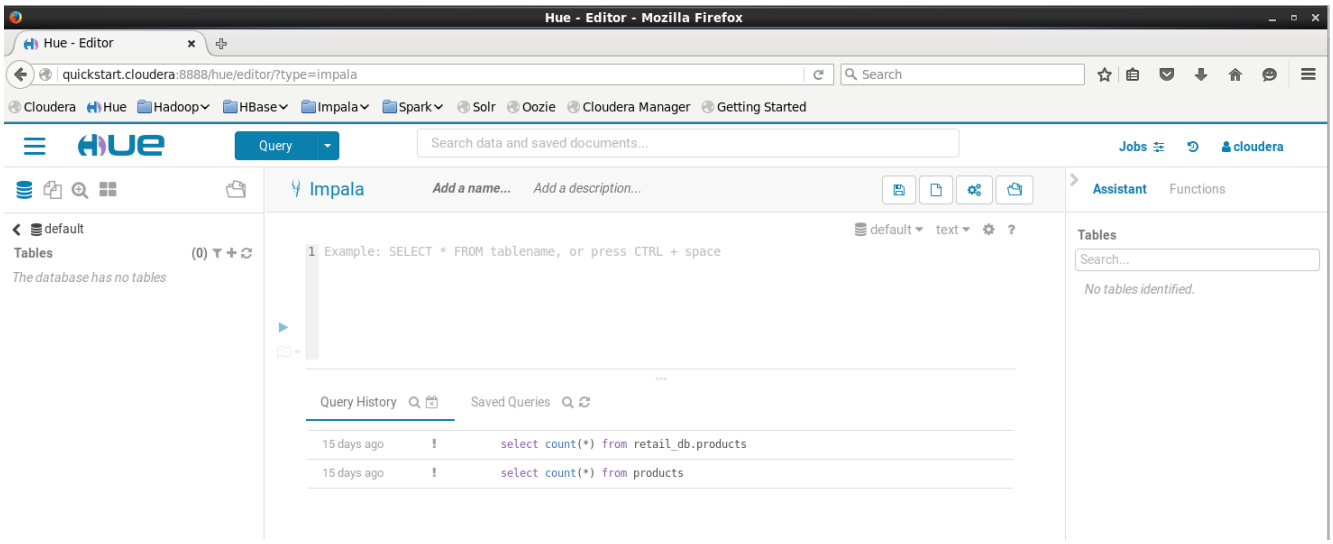
ایجاد بستر لازم برای امکان کار با Hive

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 :

1. If you were to use the following SQL statement in the Quickstart VM via the **Hue interface** - utilizing the Hive query editor:

```
SELECT product_name, product_price
FROM products
WHERE ( product_price > 10)
ORDER BY product_price DESC
LIMIT 1000
```

What would be the correct answer for the most expensive product?

a) SOLE E25 Elliptical
b) SOLE E35 Elliptical
c) SOLE F85 Treadmill
d) Spalding Beast 60" Glass portable Basketball

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)

```
[cloudera@quickstart Desktop]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 277
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| cm                 |
| firehose           |
| hue                |
| metastore          |
| mysql              |
| nav                |
| navms              |
| oozie              |
| retail_db          |
| rman               |
| sentry             |
+--------------------+
12 rows in set (0.05 sec)
```

```
mysql> use retail_db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+---------------------+
| Tables_in_retail_db |
+---------------------+
| categories          |
| customers           |
| departments         |
| order_items         |
| orders              |
| products            |
+---------------------+
6 rows in set (0.00 sec)

mysql>
```

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:

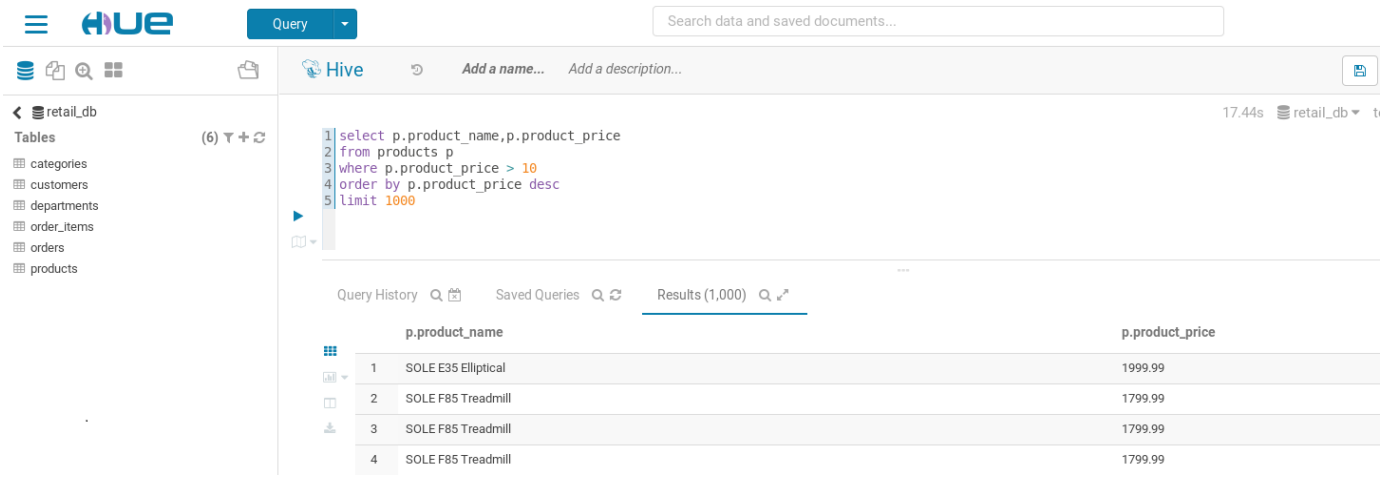
```
$ hive
hive> show databases;
hive> create database retail_db;
```

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 :

2. In the sample_08 data set - who had the highest salary in 2008?
(hint: look at the example SQL query from the hands-on exercise)

a) Chief Executives
b) Anesthesiologists
c) Surgeons
d) Lawyers

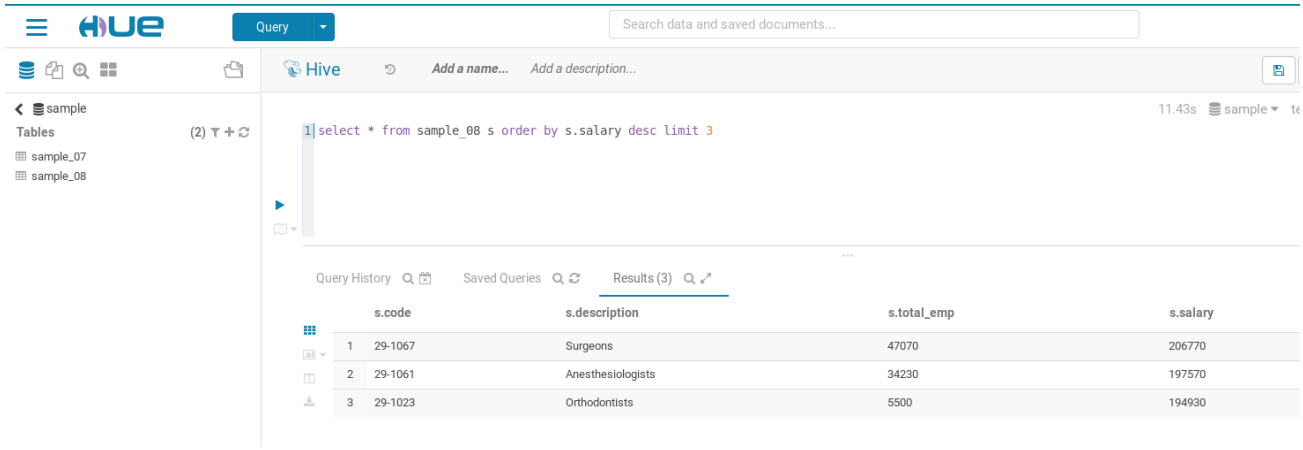
برای حل این تمرین، نیاز به دیتاست sample_08 می‌باشد. برای اینکار با دستورات زیر، این دیتاستها را دانلود می‌کنیم :

```
$ 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 Hive CLI 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:



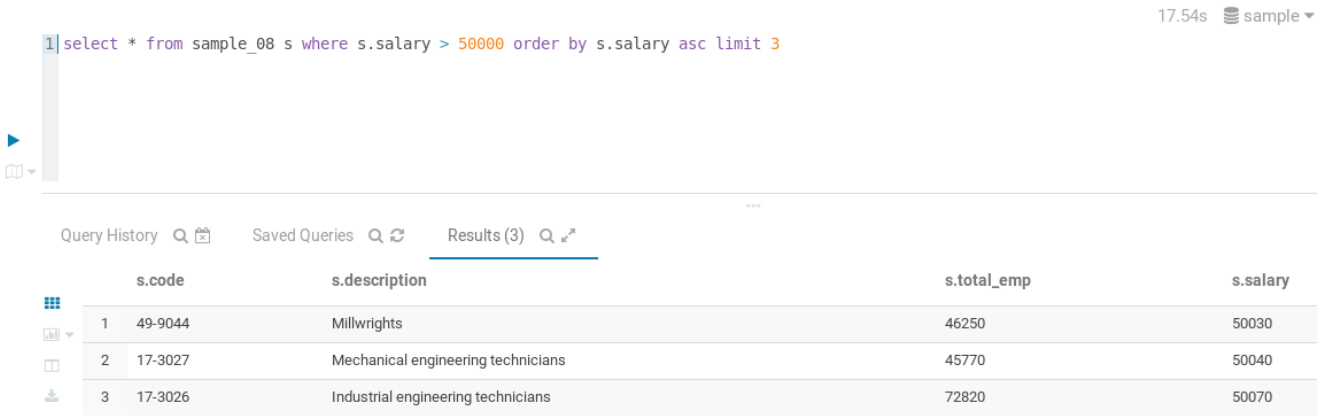
بنابراین با توجه به نتیجه query بالا، پاسخ تمرین، گزینه سوم یعنی Surgeons می‌باشد.

تمرین سوم :

3. In the sample_08 data set - who had the lowest salary above $50000 in 2008? (hint: look at the example SQL query from the hands-on exercise -ASC vs. DESC)

a) Food service managers
b) Millwrights
c) Postal service clerks
d) Interior designers

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 :

4. If you only had $10 - what is the most expensive product you could afford to buy from the products table?
    a) Clicgear Rovic Shoe Brush
    b) $10 gift card
    c) Toronto FC Team Color Soccer Bracelet
    d) adidas Brazuca 2014 Mini Soccer Ball
    e) LIJA Women's golf Beanie

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

```
1 select * from products p where p.product_price <= 10 order by p.product_price desc limit 3
```

17.56s retail_db text

Query History | Saved Queries | Results (3)

	p.product_id	p.product_category_id	p.product_name	p.product_description	p.product_price	p.product_image
1	1155	52	$10 Gift Card		10	http://images.acmesports.sports/%2410+Gift+Card
2	772	35	Clicgear Rovic Shoe Brush		9.9900000000000002	http://images.acmesports.sports/Clicgear+Rovic+Shoe+B
3	1262	56	adidas Brazuca 2014 Mini Soccer Ball		9.9900000000000002	http://images.acmesports.sports/adidas+Brazuca+2014+I

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

```
hive> 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
    > ;
OK
Time taken: 0.062 seconds
hive> ALTER TABLE 201402_trip_data SET TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.07 seconds
hive> load data local inpath '/tmp/201402_trip_data.csv' into table 201402_trip_data;
Loading data to table default.201402_trip_data
Table default.201402_trip_data stats: [numFiles=1, numRows=0, totalSize=17219022, rawDataSize=0]
OK
Time taken: 0.369 seconds
hive>
```

تمرین پنجم :

Which startstation has the longest trip duration?

a) Davis at Jacskon
b) University and Emerson
c) Park at Olive
d) Powell Street BART

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.

```
select duration,start_station from `201402_trip_data` order by duration desc limit 3
```

Query History | Saved Queries | Results (3)

	duration	start_station
1	722236	University and Emerson
2	619322	San Jose Diridon Caltrain Station
3	597517	California Ave Caltrain Station

تمرین ششم :

6- Utilizing the Bay Area Bike Share database (both Year 1 & 2, Aug. 2013- Aug. 2015)- what is the most popular start station based on trip data?

a) Embarcadero at Sansome
b) Market at 4th
c) San Francisco Caltrain
d) Townsend at 7th

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.

```
select g.start_station, count(*) num_trip
from
(select start_station from `201402_trip_data`
 union all
 select start_station from `201408_trip_data`
 union all
 select start_station from `201508_trip_data`) g
group by start_station
order by num_trip desc limit 3
```

Query History | Saved Queries | Results (3)

	g.start_station	num_trip
1	San Francisco Caltrain (Townsend at 4th)	49092
2	San Francisco Caltrain 2 (330 Townsend)	33742
3	Harry Bridges Plaza (Ferry Building)	32934

تمرین هفتم :

7. Utilizing the Bay Area Bike Share database (Year 1 only, Aug. 2013- Feb 2014) - Which is the least popular(least used) start station in the Bike share trips data?
(Hint: Use the count of start station, group and order in ascending order)

a) Townsend at 7th
b) Mezes Park
c) Market at 4th
d) Embarcadero at Sansome

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

```
select start_station, count(*) num_trip
from `201402_trip_data`
group by start_station
order by num_trip asc limit 3
```

Query History | Saved Queries | Results (3)

	start_station	num_trip
1	Mezes Park	3
2	San Jose Government Center	23
3	Redwood City Public Library	44

تمرین هشتم :

8. Utilizing the Bay Area Bike Share database (for Year 1 only, Aug. 2013 - Aug. 2014 only) - what is the SECOND MOST popular end station based on trip data? (Hint: Use the count of end station, group and order in descending order)

a) Steuart at Market
b) Market at Sansome
c) Embarcadero at Sansome
d) Harry Bridges Plaza(Ferry Building)

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 همچنین مرتب کردن نزولی، پاسخ سوال بدست امده است که بر اساس خروجی query، گزینه سوم (Embarcadero at Sansome) صحیح است.

```sql
select g.end_station, count(*) num_trip
from
(select end_station from `201402_trip_data`
 union all
 select end_station from `201408_trip_data`) g
group by end_station
order by num_trip desc limit 3
```

Query History | Saved Queries | Results (3)

	g.end_station	num_trip
1	San Francisco Caltrain (Townsend at 4th)	28369
2	Embarcadero at Sansome	15731
3	Harry Bridges Plaza (Ferry Building)	15383