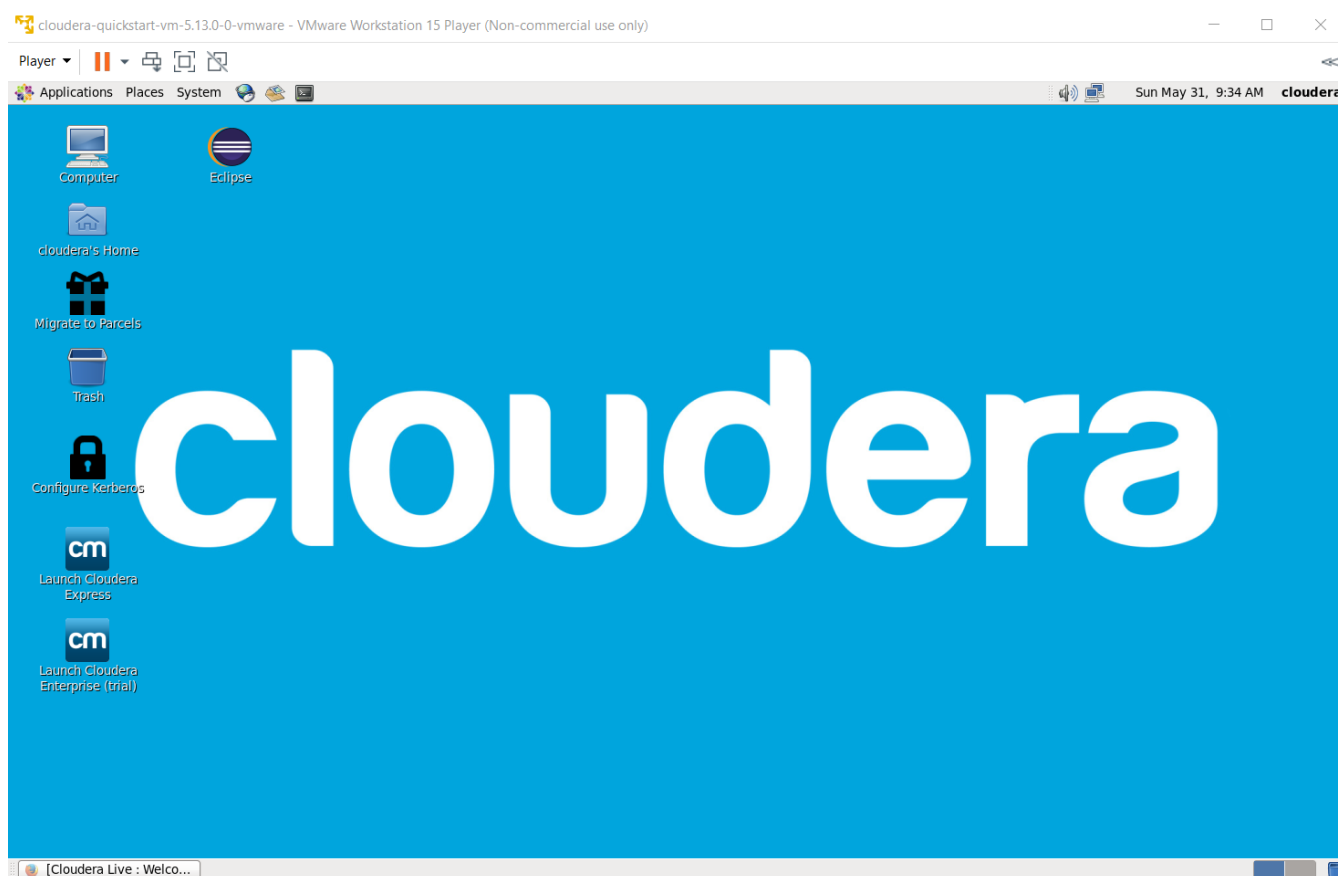


Hive

ایجاد بستر لازم برای امکان کار با Hive

برای انجام این تمرین و تمرینهای بعدی، فایل cloudera-quickstart-vm-5.13.0-0-vmware مجازی cloudera می باشد، 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 می باشد. سیستم عامل آن لینوکس CentOS 6.7 می باشد 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 می باشد :



پس از اتصال به کنسول web کلودرا که بصورت زیر می باشد، صفحه ای مشابه زیر نمایش داده می شود :

cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)

Player ▾ | Applications | Places | System | Sun May 31, 9:40 AM | cloudera

Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox

quickstart.cloudera/#/

Cloudera | Hue | Hadoop ▾ | HBase ▾ | Impala ▾ | Spark ▾ | Solr | Oozie | Cloudera Manager | Getting Started

cloudera LIVE

Navigation ▾

Welcome to Your Cloudera QuickStart VM!

Your Cluster	
Node	Address
Manager Node	192.168.80.128
Worker Node 1	192.168.80.128

Get Started

The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).

[Start Tutorial](#)

Analyze Your Data

Hue is the open source web interface for Hadoop that lets you analyze your data. Simply load in your data and then easily begin to analyze, search, and visualize it. In the QuickStart VM, the administrative username for Hue is 'cloudera' and the password is 'cloudera'.

[Launch Hue UI](#)

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

Hue - Editor - Mozilla Firefox

Hue - Editor | quickstart.cloudera:8888/hue/editor?type=impala

Cloudera | Hue | Hadoop ▾ | HBase ▾ | Impala ▾ | Spark ▾ | Solr | Oozie | Cloudera Manager | Getting Started

HUE

Query ▾ | Search data and saved documents...

Jobs | cloudera

Assistant | Functions

default ▾ | text ▾ | ?

Tables

Search...

No tables identified.

Query History | Saved Queries

Time	Query
15 days ago	select count(*) from retail_db.products
15 days ago	select count(*) from products

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 `retail_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 `retail_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 retail_db، query مربوط به این تمرین را نوشته و اجرا می‌کنیم. شکل زیر نتیجه اجرا را نشان می‌دهد:

Hue Interface → Query → Editor → Hive

The screenshot shows the Hue web interface. On the left, a sidebar lists the tables in the 'retail_db' database: categories, customers, departments, order_items, orders, and products. The main area is divided into two sections. The top section is the 'Query' editor, where a SQL query is entered:

```
1 select p.product_name, p.product_price
2 from products p
3 where p.product_price > 10
4 order by p.product_price desc
5 limit 1000
```

The bottom section shows the 'Results (1,000)' table. It has two columns: 'p.product_name' and 'p.product_price'. The first four rows are visible:

	p.product_name	p.product_price
1	SOLE E35 Elliptical	1999.99
2	SOLE F85 Treadmill	1799.99
3	SOLE F85 Treadmill	1799.99
4	SOLE F85 Treadmill	1799.99

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

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:

The screenshot shows the Hue web interface. On the left, a sidebar lists 'sample' and its tables 'sample_07' and 'sample_08'. The main area displays a Hive query: `select * from sample_08 s order by s.salary desc limit 3`. Below the query editor, the 'Results (3)' tab is active, showing a table with 4 columns: s.code, s.description, s.total_emp, and s.salary. The results are as follows:

	s.code	s.description	s.total_emp	s.salary
1	29-1067	Surgeons	47070	206770
2	29-1061	Anesthesiologists	34230	197570
3	29-1023	Orthodontists	5500	194930

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 :

The screenshot shows the Hue web interface with a new Hive query: `select * from sample_08 s where s.salary > 50000 order by s.salary asc limit 3`. The 'Results (3)' tab is active, showing a table with 4 columns: s.code, s.description, s.total_emp, and s.salary. The results are as follows:

	s.code	s.description	s.total_emp	s.salary
1	49-9044	Millwrights	46250	50030
2	17-3027	Mechanical engineering technicians	45770	50040
3	17-3026	Industrial engineering technicians	72820	50070

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.

17.56s retail_db text

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

برای ادامه تمرینها نیاز به دیتاست bayareabikeshare می باشد که مشابه قبل، ابتدا فایل های 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 تفاوت های 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 می باشد 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 فایل های 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 Jacson
- 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.


```
1 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?

- Embarcadero at Sansome
- Market at 4th
- San Francisco Caltrain
- 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.

```
1 select g.start_station, count(*) num_trip
2 from
3 (select start_station from `201402_trip_data`
4 union all
5 select start_station from `201408_trip_data`
6 union all
7 select start_station from `201508_trip_data`) g
8 group by start_station
9 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.

```
1 select start_station, count(*) num_trip
2 from `201402_trip_data`
3 group by start_station
4 order by num_trip asc limit 3
5
6
7
```

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 آمده است که بر اساس خروجی query، گزینه سوم (Embarcadero at Sansome) صحیح است.

```

1 select g.end_station, count(*) num_trip
2 from
3 (select end_station from `201402_trip_data`
4  union all
5  select end_station from `201408_trip_data`) g
6 group by end_station
7 order by num_trip desc limit 3
8
9

```

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