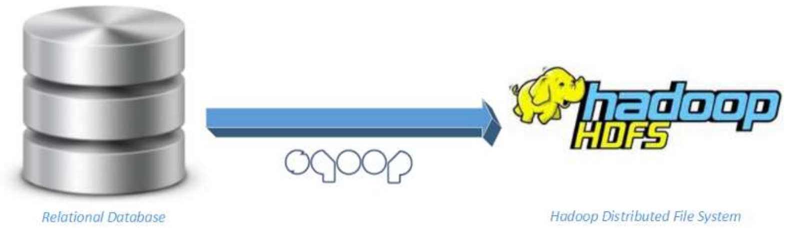
Tutorial 3: Apache Sqoop with Cloudera

CN7031 - Big Data Analytics

Dr Fahimeh Jafari (f.jafari@uel.ac.uk)

LEARNING OUTCOMES: After completing this tutorial, you should:

- Have gotten a hands-on experience in deploying Apache Sqoop with Cloudera
- Understand Sqoop concepts and commands
- Be able to import data into HDFS
- Using Hue to Verify/Query Tables



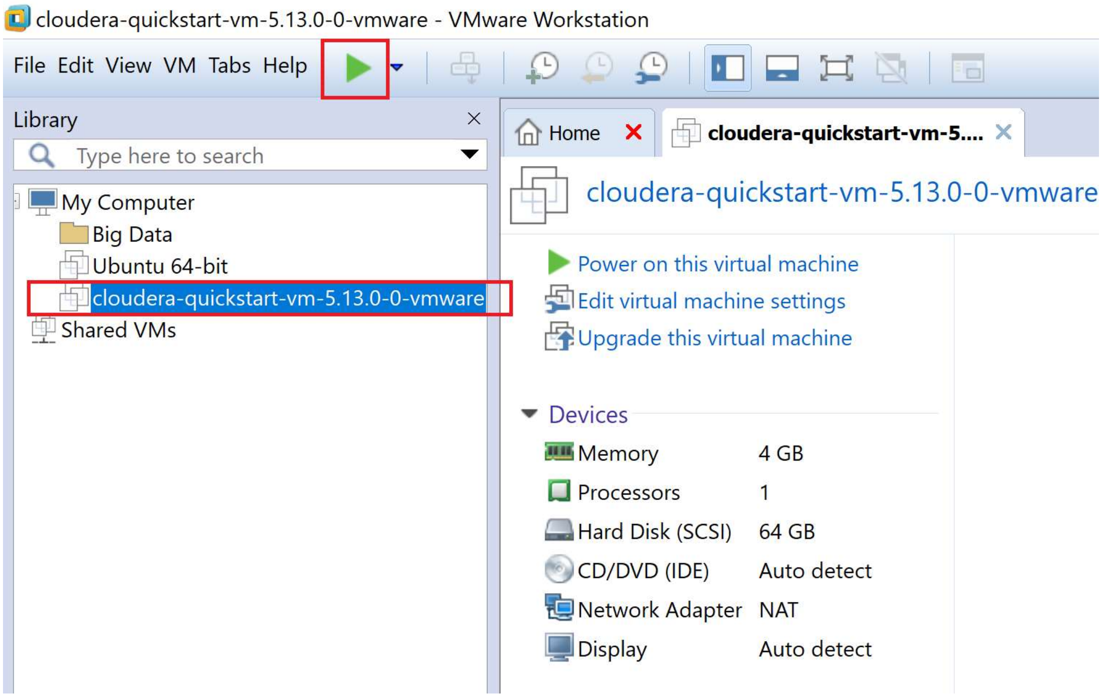
NOTE: Tutorial Submission (OPTIONAL)

You can submit the results of your work by taking screenshots (wherever pointed), pasting them into a word file and sending the word/pdf file through the submission link provided in Moodle.

Exercise 1: Launch Cloudera

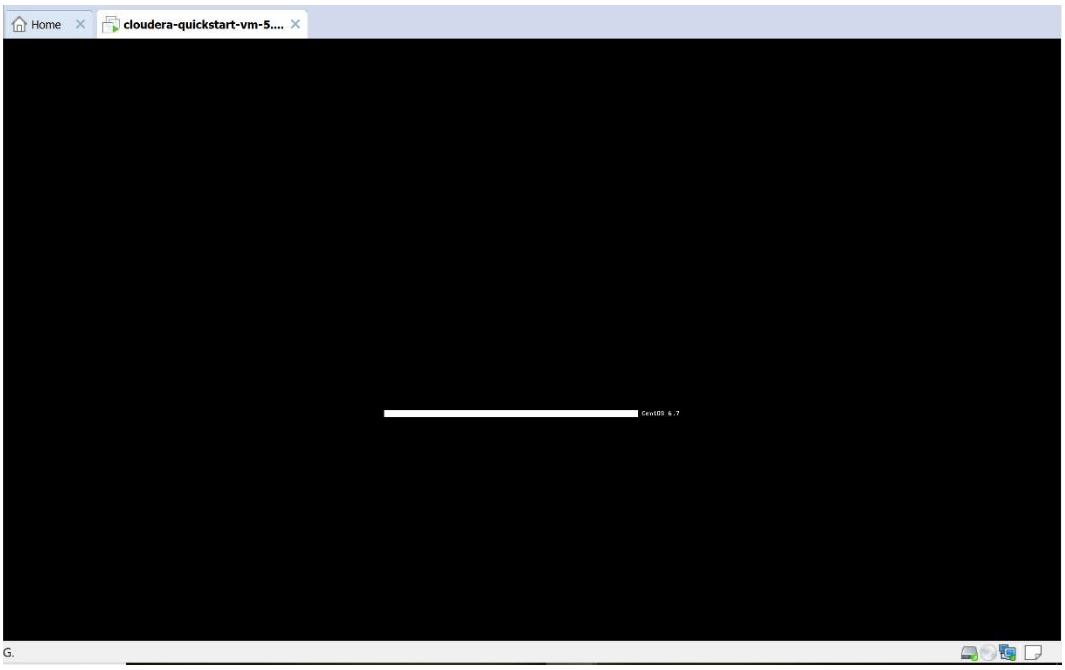
Please use the following instructions to launch the `Cloudera Quickstart VM` with `VMware`.

1. **Launch Cloudera VM.** When you open it, the `cloudera-quickstart-vm-5.13.0-0-vmware` will appear on the left in the VMware window. Select it and click the `Start` button to launch the VM.

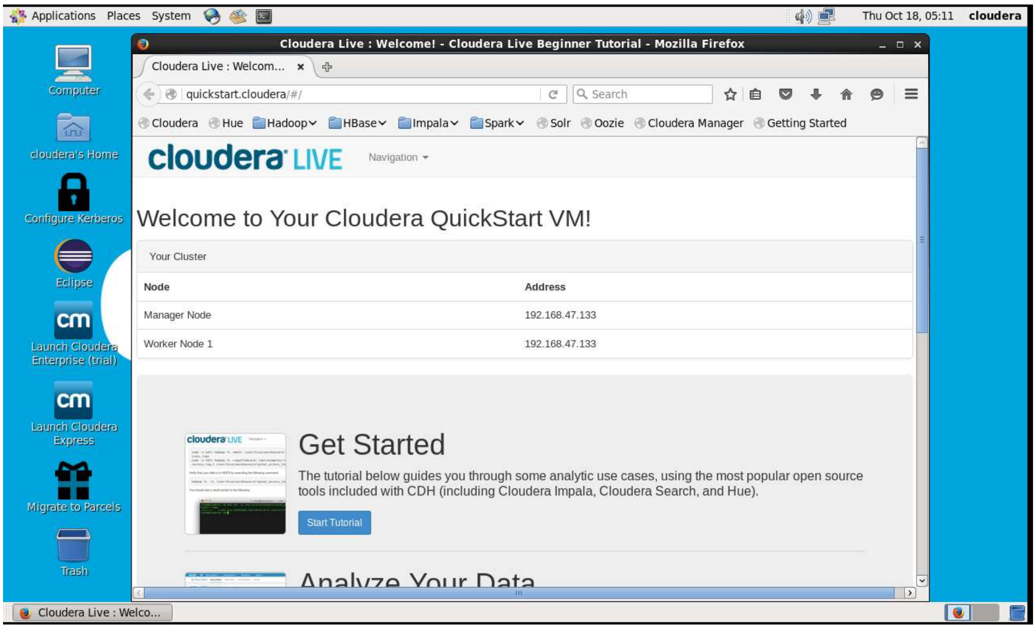

cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation
File Edit View VM Tabs Help
Library
Type here to search
My Computer
Big Data
Ubuntu 64-bit
cloudera-quickstart-vm-5.13.0-0-vmware
Shared VMs
Home
cloudera-quickstart-vm-5....
cloudera-quickstart-vm-5.13.0-0-vmware
Power on this virtual machine
Edit virtual machine settings
Upgrade this virtual machine
Devices

Memory	4 GB
Processors	1
Hard Disk (SCSI)	64 GB
CD/DVD (IDE)	Auto detect
Network Adapter	NAT
Display	Auto detect

2. **Cloudera VM booting.** It will take several minutes for the Virtual Machine to start. The booting process takes a long time since many Hadoop tools are started.



3. **The Cloudera VM desktop.** Once the booting process is complete, the desktop will appear with a browser.


Applications Places System
Thu Oct 18, 05:11 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
Computer
cloudera's Home
Configure Kerberos
Eclipse
Launch Cloudera Enterprise (trial)
Launch Cloudera Express
Migrate to Parcels
Trash
Start Tutorial
Cloudera Live : Welco...

Welcome to Your Cloudera QuickStart VM!

Your Cluster

Node	Address
Manager Node	192.168.47.133
Worker Node 1	192.168.47.133

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).

Analyze Your Data

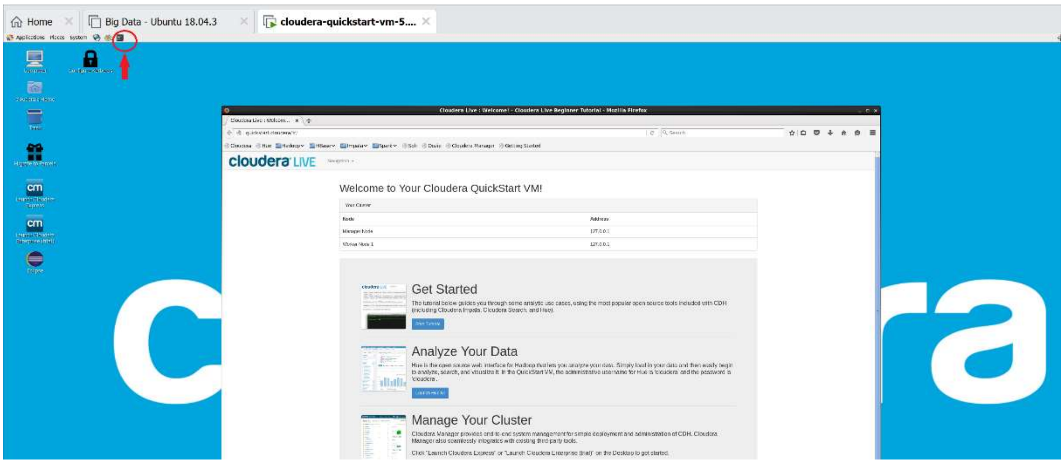
Exercise 2: Work with MySQL

MySQL is an open-source database management system, commonly installed as part of the popular LAMP (Linux, Apache, MySQL, PHP/Python/Perl) stack. It uses a relational database and SQL (Structured Query Language) to manage its data.

We don’t need to install MySQL package as it’s installed in Cloudera image.

Step 1: How to Access the MySQL shell

1. You should first open a terminal, which you can do by clicking on the black "Terminal" icon at the top of your screen. Once it is open, you can launch the MySQL.

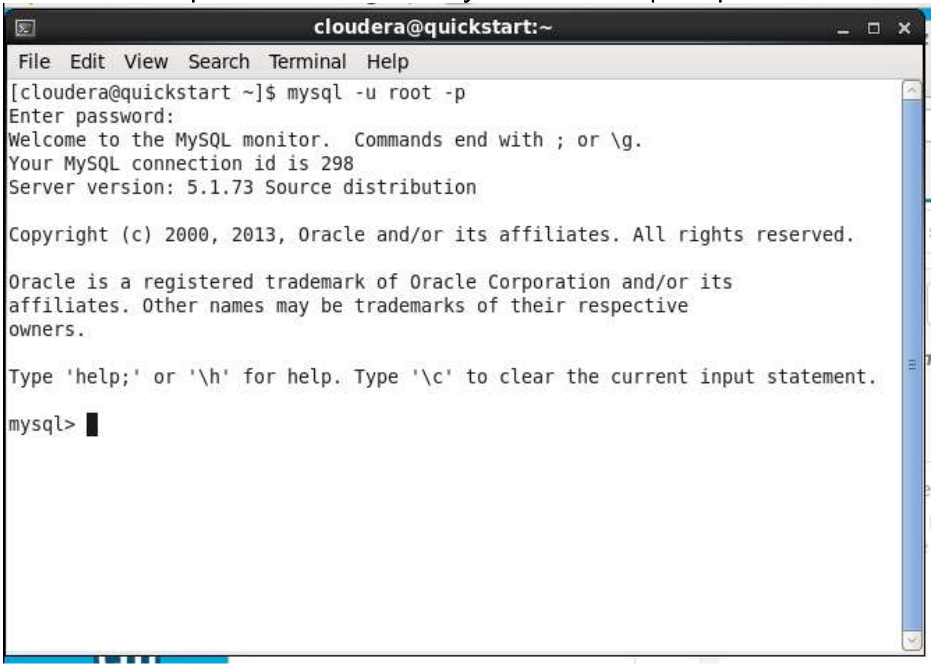

Home
Big Data - Ubuntu 18.04.3
cloudera-quickstart-vm-5....
cloudera LIVE
Welcome to Your Cloudera QuickStart VM!
Get Started
Analyze Your Data
Manage Your Cluster

2. Run MySQL by typing the following command:

```
mysql -u root -p
```

3. Then you’ll be asked to enter MySQL password. Type **cloudera**

   You’ll then be presented with the MySQL monitor prompt:



```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 298
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> 
```

Two points to keep in mind:

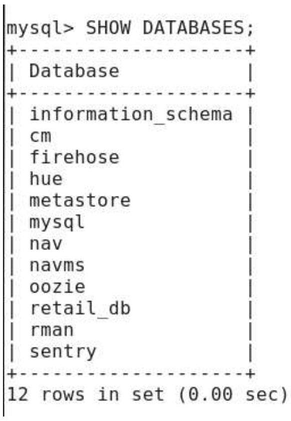
- All MySQL commands end with a semicolon; if the phrase does not end with a semicolon, the command will not execute.
- Also, although it is **not** required, MySQL commands are usually written in uppercase and databases, tables, usernames, or text are in lowercase to make them easier to distinguish. However, the MySQL command line is not case sensitive.

Step 2: Show Database and Tables

MySQL organizes its information into databases; each one can hold tables with specific data. Show existing databases with

```
SHOW DATABASES;
```

Output



```
mysql> SHOW DATABASES;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| cm                 |
| firehose           |
| hue                |
| metastore          |
| mysql              |
| nav                |
| navms              |
| oozie              |
| retail_db          |
| rman               |
| sentry             |
+--------------------+
12 rows in set (0.00 sec)
```

(Take Screenshot #1)

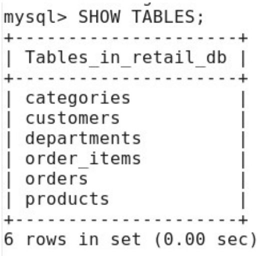
In this tutorial we works with `retail_db` database So, access to this databases with

```
USE retail_db;
```

You can quickly check what tables are available by typing:

```
SHOW TABLES;
```

Output



```
mysql> SHOW TABLES;
+---------------------+
| Tables_in_retail_db |
+---------------------+
| categories          |
| customers           |
| departments         |
| order_items         |
| orders              |
| products            |
+---------------------+
6 rows in set (0.00 sec)
```

(Take Screenshot #2)

Step 3: Exit from MySQL

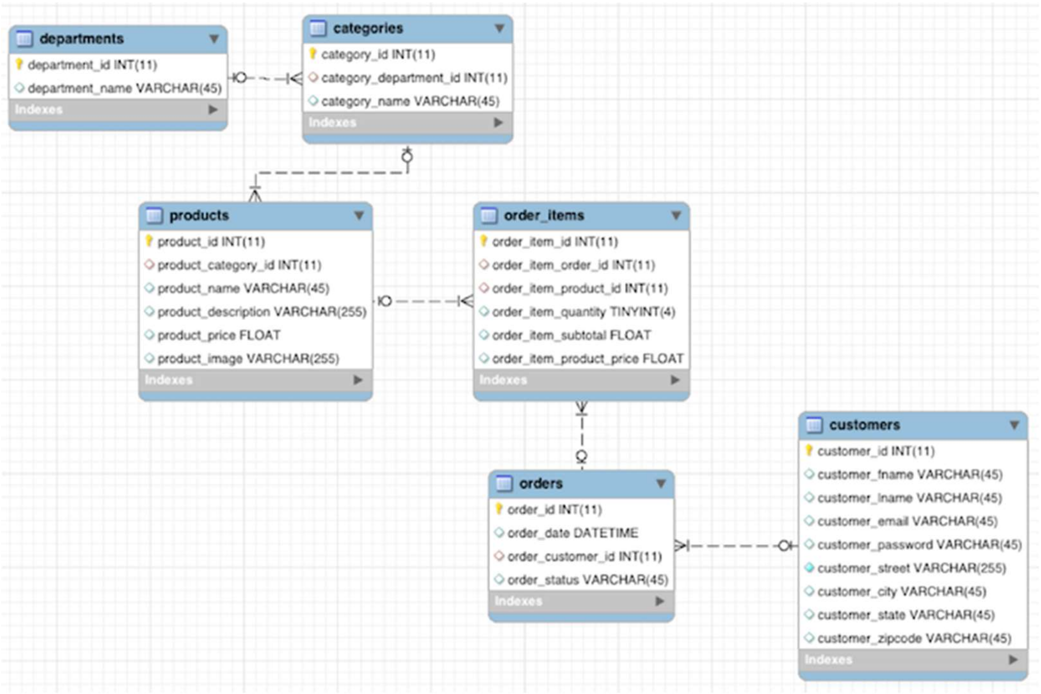
Type following command to exit from MySQL

```
exit
```

Exercise 3: Importing data from MySQL database to Hadoop HDFS

To analyse the transaction data in the new platform, we need to ingest it into the Hadoop Distributed File System (HDFS). We need to find a tool that easily transfers structured data from a RDBMS to HDFS, while preserving structure. That enables us to query the data, but not interfere with or break any regular workload on it.

This exercise describes how to import data from MySQL database to Hadoop HDFS. We are going to import `retail_db` into the Hadoop with Sqoop. The flowchart of `retail_db` is in the following picture [Source: https://www.cloudera.com/developers/get-started-with-hadoop-tutorial/exercise-1.html]

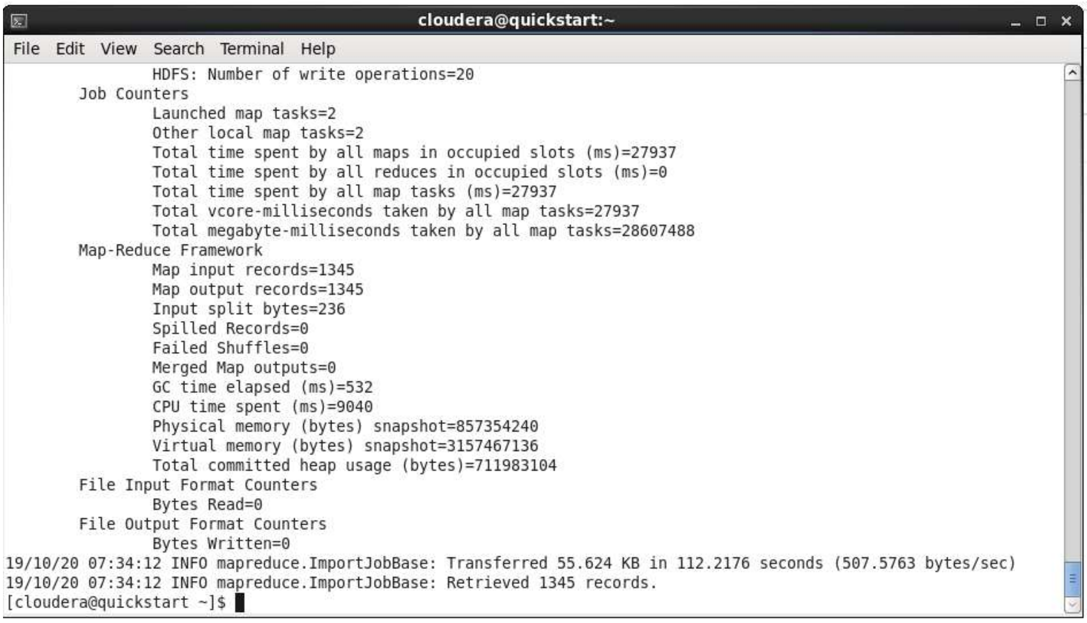

departments
department_id INT(11)
department_name VARCHAR(45)
Indexes
categories
category_id INT(11)
category_department_id INT(11)
category_name VARCHAR(45)
Indexes
products
product_id INT(11)
product_category_id INT(11)
product_name VARCHAR(45)
product_description VARCHAR(255)
product_price FLOAT
product_image VARCHAR(255)
Indexes
order_items
order_item_id INT(11)
order_item_order_id INT(11)
order_item_product_id INT(11)
order_item_quantity TINYINT(4)
order_item_subtotal FLOAT
order_item_product_price FLOAT
Indexes
customers
customer_id INT(11)
customer_fname VARCHAR(45)
customer_lname VARCHAR(45)
customer_email VARCHAR(45)
customer_password VARCHAR(45)
customer_street VARCHAR(255)
customer_city VARCHAR(45)
customer_state VARCHAR(45)
customer_zipcode VARCHAR(45)
Indexes
orders
order_id INT(11)
order_date DATETIME
order_customer_id INT(11)
order_status VARCHAR(45)
Indexes

Open a new terminal and run the following Sqoop command line by line. **It would take a while to be completed, because Sqoop launches MapReduce job.**

NOTE: You must type the commands yourself. Don’t copy & past them as it wouldn’t work.

```
[cloudera@quickstart ~]$ sqoop import-all-tables \
-m 2 \
--connect jdbc:mysql://localhost:3306/retail_db \
--username=root \
--password=cloudera \
--compression-codec=snappy \
--as-parquetfile \
--warehouse-dir /user/hive/warehouse
```

After about 1 minute, you have to see the following screenshot as the confirmation:



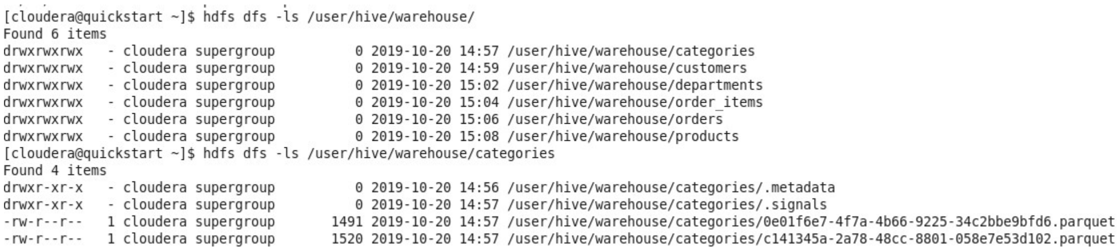
```
		HDFS: Number of write operations=20
	Job Counters
		Launched map tasks=2
		Other local map tasks=2
		Total time spent by all maps in occupied slots (ms)=27937
		Total time spent by all reduces in occupied slots (ms)=0
		Total time spent by all map tasks (ms)=27937
		Total vcore-milliseconds taken by all map tasks=27937
		Total megabyte-milliseconds taken by all map tasks=28607488
	Map-Reduce Framework
		Map input records=1345
		Map output records=1345
		Input split bytes=236
		Spilled Records=0
		Failed Shuffles=0
		Merged Map outputs=0
		GC time elapsed (ms)=532
		CPU time spent (ms)=9040
		Physical memory (bytes) snapshot=857354240
		Virtual memory (bytes) snapshot=3157467136
		Total committed heap usage (bytes)=711983104
	File Input Format Counters
		Bytes Read=0
	File Output Format Counters
		Bytes Written=0
19/10/20 07:34:12 INFO mapreduce.ImportJobBase: Transferred 55.624 KB in 112.2176 seconds (507.5763 bytes/sec)
19/10/20 07:34:12 INFO mapreduce.ImportJobBase: Retrieved 1345 records.
[cloudera@quickstart ~]$ 
```

(Take Screenshot #3)

Verification:

When this command is complete, confirm that your data files exist in HDFS. These commands will show the directories and the files inside them that make up your tables:

```
hdfs dfs -ls /user/hive/warehouse/
hdfs dfs -ls /user/hive/warehouse/categories/
```



```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/hive/warehouse/
Found 6 items
drwxrwxrwx   - cloudera supergroup          0 2019-10-20 14:57 /user/hive/warehouse/categories
drwxrwxrwx   - cloudera supergroup          0 2019-10-20 14:59 /user/hive/warehouse/customers
drwxrwxrwx   - cloudera supergroup          0 2019-10-20 15:02 /user/hive/warehouse/departments
drwxrwxrwx   - cloudera supergroup          0 2019-10-20 15:04 /user/hive/warehouse/order_items
drwxrwxrwx   - cloudera supergroup          0 2019-10-20 15:06 /user/hive/warehouse/orders
drwxrwxrwx   - cloudera supergroup          0 2019-10-20 15:08 /user/hive/warehouse/products
[cloudera@quickstart ~]$ hdfs dfs -ls /user/hive/warehouse/categories
Found 4 items
drwxr-xr-x   - cloudera supergroup          0 2019-10-20 14:56 /user/hive/warehouse/categories/.metadata
drwxr-xr-x   - cloudera supergroup          0 2019-10-20 14:57 /user/hive/warehouse/categories/.signals
-rw-r--r--   1 cloudera supergroup       1491 2019-10-20 14:57 /user/hive/warehouse/categories/0e01f6e7-4f7a-4b66-9225-34c2bbe9bfd6.parquet
-rw-r--r--   1 cloudera supergroup       1520 2019-10-20 14:57 /user/hive/warehouse/categories/c141345a-2a78-48cc-8801-058e7e53d102.parquet
```

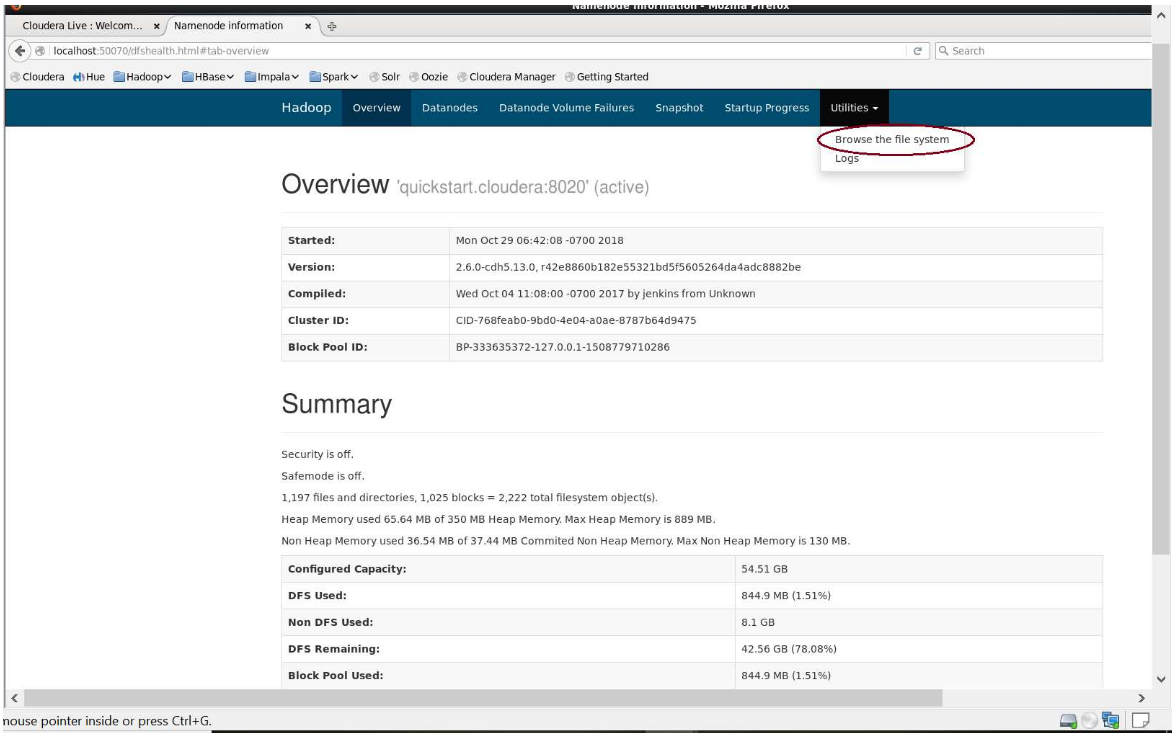
(Take Screenshot #4)

Note: The number of `.parquet` files shown will be equal to what was passed to Sqoop with the `–m` parameter. This is the number of 'mappers' that Sqoop will use in its MapReduce jobs. It could also be thought of as the number of simultaneous connections to your database, or the number of disks/Data Nodes you want to spread the data across. So, on a single-node you will just see one, but larger clusters will have a greater number of files.

OR, complete the verification step visually (localhost:50070) through the following steps:

- Open a new tab in Firefox and go to http://localhost:50070/

- Then click on `Utilities->Brows the file systems`



Overview 'quickstart.cloudera:8020' (active)

Started:	Mon Oct 29 06:42:08 -0700 2018
Version:	2.6.0-cdh5.13.0, r42e8860b182e55321bd5f5605264da4adc8882be
Compiled:	Wed Oct 04 11:08:00 -0700 2017 by jenkins from Unknown
Cluster ID:	CID-768feab0-9bd0-4e04-a0ae-8787b64d9475
Block Pool ID:	BP-333635372-127.0.0.1-1508779710286

Summary

Security is off.

Safemode is off.

1,197 files and directories, 1,025 blocks = 2,222 total filesystem object(s).

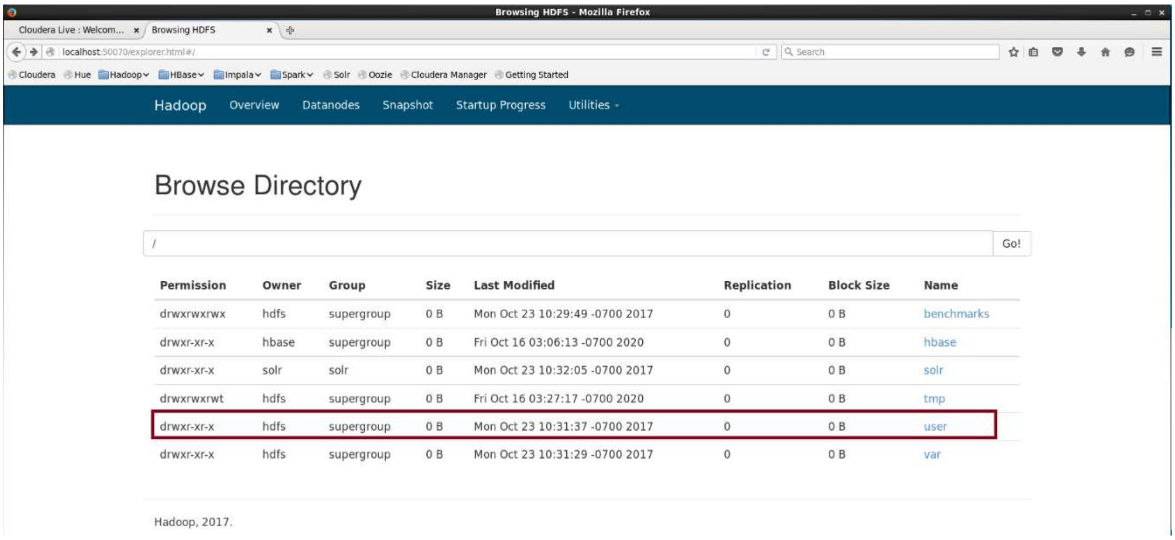
Heap Memory used 65.64 MB of 350 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 36.54 MB of 37.44 MB Commited Non Heap Memory. Max Non Heap Memory is 130 MB.

Configured Capacity:	54.51 GB
DFS Used:	844.9 MB (1.51%)
Non DFS Used:	8.1 GB
DFS Remaining:	42.56 GB (78.08%)
Block Pool Used:	844.9 MB (1.51%)

nouse pointer inside or press Ctrl+G.

You should see `user` Directory.



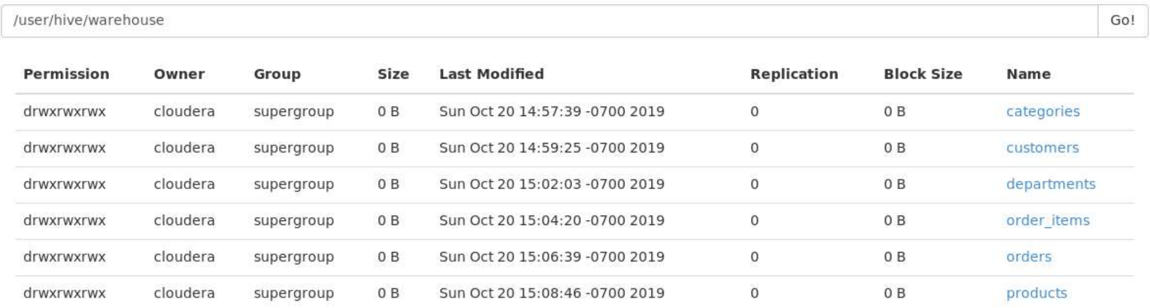
Browse Directory

/

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxrwx	hdfs	supergroup	0 B	Mon Oct 23 10:29:49 -0700 2017	0	0 B	benchmarks
drwxr-xr-x	hbase	supergroup	0 B	Fri Oct 16 03:06:13 -0700 2020	0	0 B	hbase
drwxr-xr-x	solr	solr	0 B	Mon Oct 23 10:32:05 -0700 2017	0	0 B	solr
drwxrwxrwt	hdfs	supergroup	0 B	Fri Oct 16 03:27:17 -0700 2020	0	0 B	tmp
drwxr-xr-x	hdfs	supergroup	0 B	Mon Oct 23 10:31:37 -0700 2017	0	0 B	user
drwxr-xr-x	hdfs	supergroup	0 B	Mon Oct 23 10:31:29 -0700 2017	0	0 B	var

Hadoop, 2017.

- Click on it and then go to `hive-> warehouse.` Now you should see the directories and the files that make up your tables:

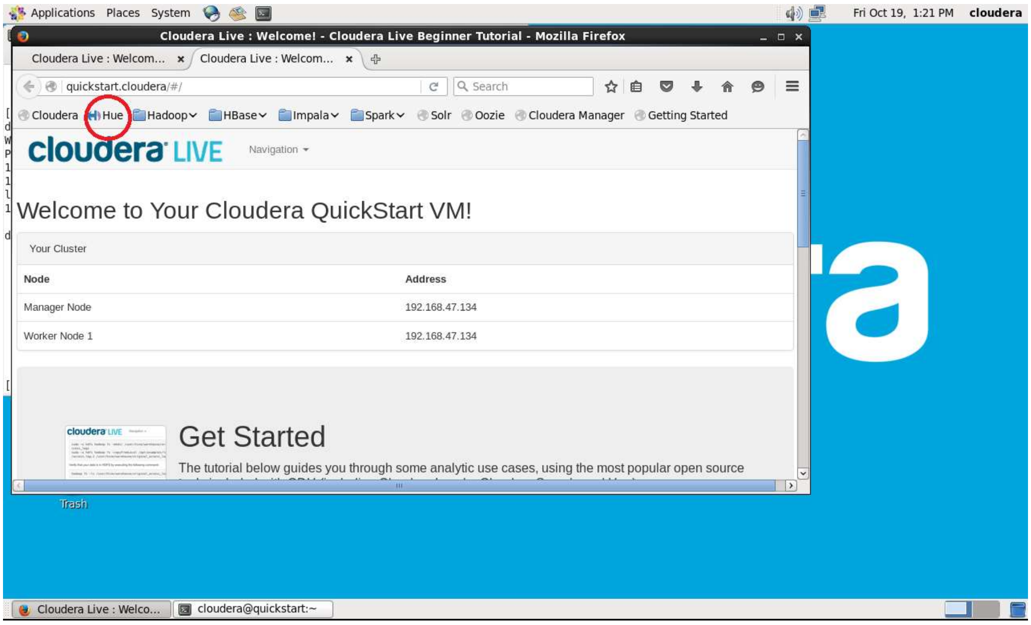


Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxrwx	cloudera	supergroup	0 B	Sun Oct 20 14:57:39 -0700 2019	0	0 B	categories
drwxrwxrwx	cloudera	supergroup	0 B	Sun Oct 20 14:59:25 -0700 2019	0	0 B	customers
drwxrwxrwx	cloudera	supergroup	0 B	Sun Oct 20 15:02:03 -0700 2019	0	0 B	departments
drwxrwxrwx	cloudera	supergroup	0 B	Sun Oct 20 15:04:20 -0700 2019	0	0 B	order_items
drwxrwxrwx	cloudera	supergroup	0 B	Sun Oct 20 15:06:39 -0700 2019	0	0 B	orders
drwxrwxrwx	cloudera	supergroup	0 B	Sun Oct 20 15:08:46 -0700 2019	0	0 B	products

(Take Screenshot #5)

Exercise 4: Using Hue to Verify/Query Tables

We're going to use Hue's Impala app to query our tables. Hue provides a web-based interface for many of the tools in CDH and can be found on port 8888 of your Manager Node. In the QuickStart VM, click on Hue as shown in the picture.


Applications Places System
Fri Oct 19, 1:21 PM cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera
Hue
Hadoop
HBase
Impala
Spark
Solr
Oozie
Cloudera Manager
Getting Started
cloudera LIVE
Navigation
Trash
Cloudera Live : Welco...
cloudera@quickstart:~

Welcome to Your Cloudera QuickStart VM!

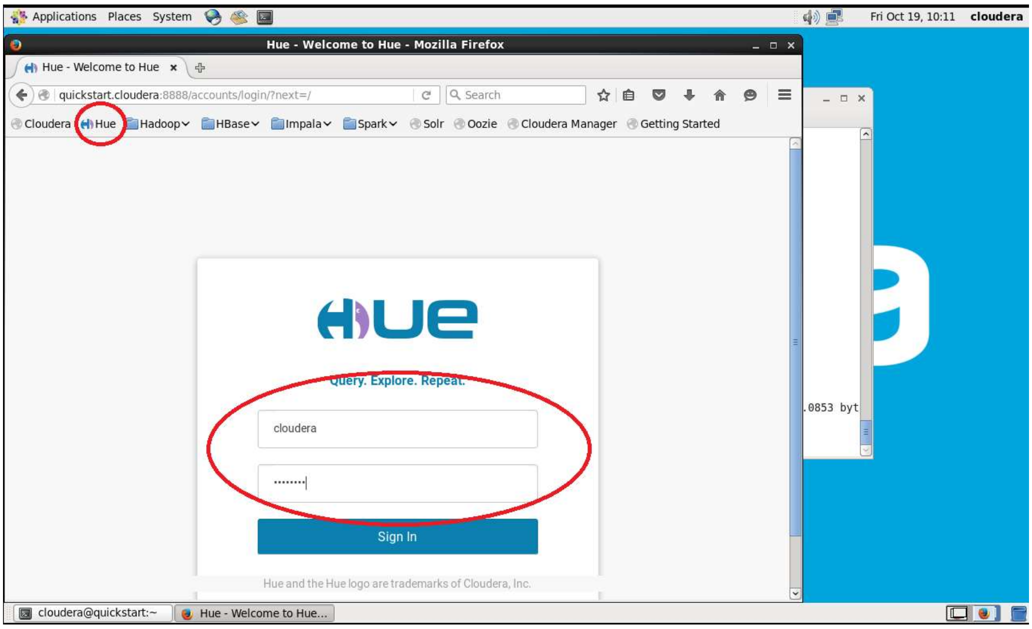
Your Cluster

Node	Address
Manager Node	192.168.47.134
Worker Node 1	192.168.47.134

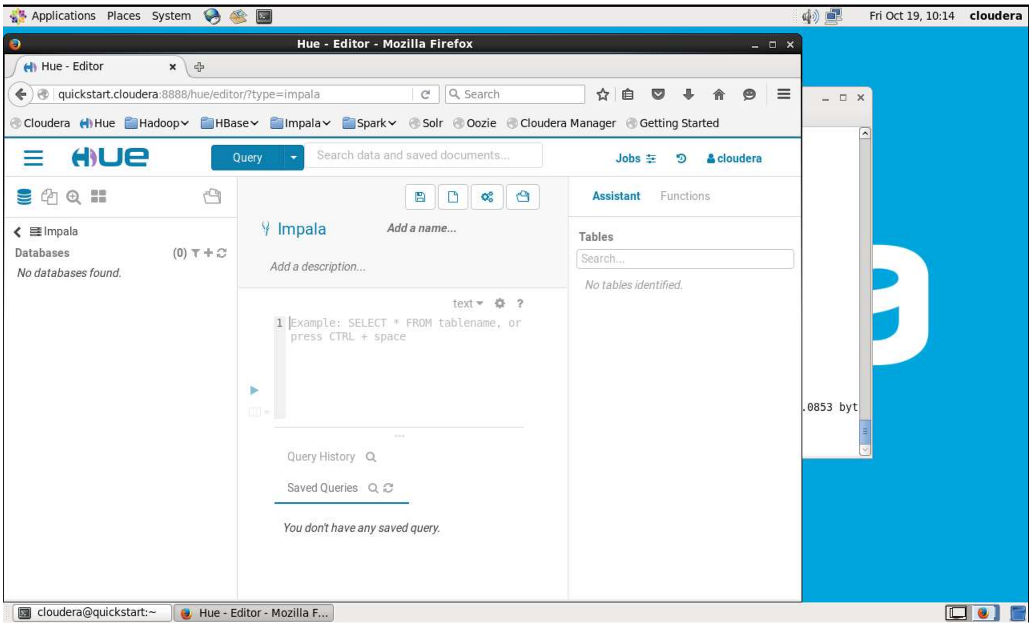
Get Started

The tutorial below guides you through some analytic use cases, using the most popular open source

Then enter the administrator username for Hue which is 'cloudera'. For the password, please first try 'BigDataadmin', if it doesn’t work, try ‘cloudera'.

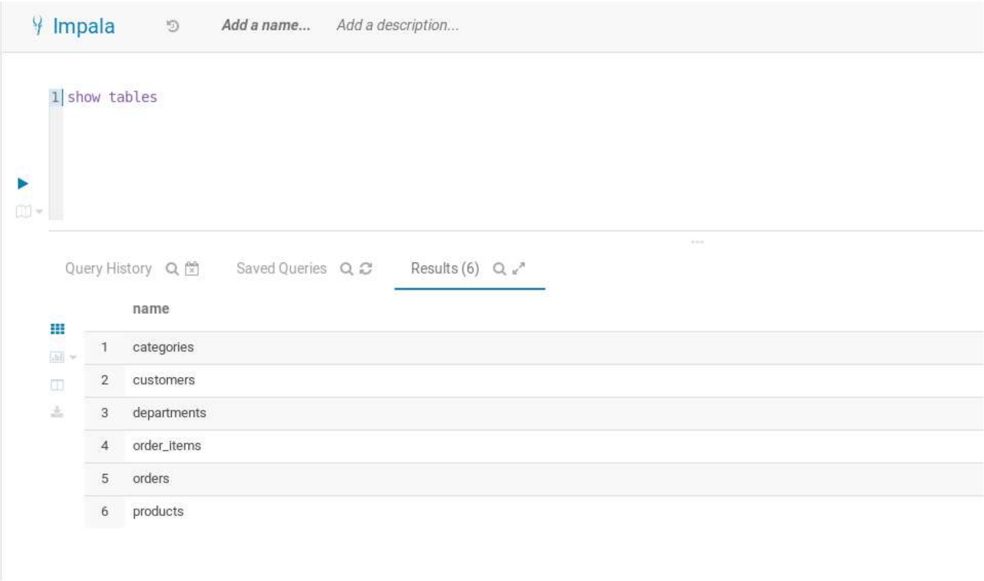

Applications Places System
Fri Oct 19, 10:11 cloudera
Hue - Welcome to Hue - Mozilla Firefox
Hue - Welcome to Hue
quickstart.cloudera:8888/accounts/login/?next=/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
HUE
Query. Explore. Repeat.
cloudera
••••••••
Sign In
Hue and the Hue logo are trademarks of Cloudera, Inc.
cloudera@quickstart:~
Hue - Welcome to Hue...

Once you are inside of Hue, click on **Query → Editor → Impala** to launch the Impala Editor.


Applications Places System
Fri Oct 19, 10:14 cloudera
Hue - Editor - Mozilla Firefox
Hue - Editor
quickstart.cloudera:8888/hue/editor/?type=impala
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
HUE
Query
Search data and saved documents...
Jobs
cloudera
Impala
Databases (0)
No databases found.
Impala
Add a name...
Add a description...
text
Example: SELECT * FROM tablename, or press CTRL + space
Query History
Saved Queries
You don't have any saved query.
Assistant Functions
Tables
Search...
No tables identified.
cloudera@quickstart:~
Hue - Editor - Mozilla F...

To save time during queries, Impala does not poll constantly for metadata changes. So, the first thing we must do is tell Impala that its metadata is out of date. Then we should see our tables show up, ready to be queried. To this, type the following commands in the editor and then click on the play icon to run the commands.

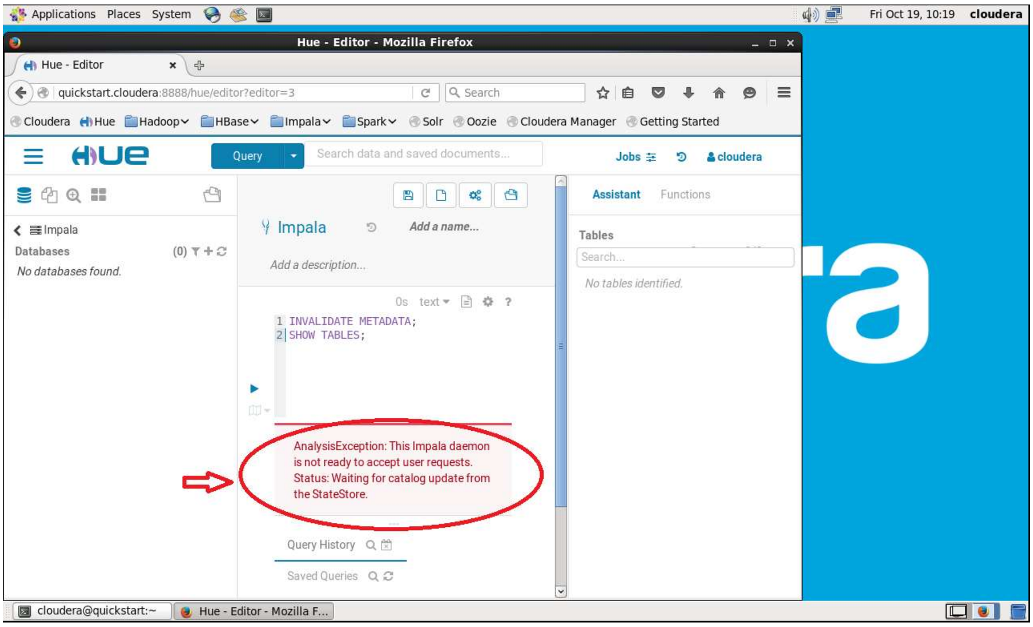
```
invalidate metadata;

show tables;
```



```
show tables
```

	name
1	categories
2	customers
3	departments
4	order_items
5	orders
6	products

[Exceptional Error:] You may get the following error when you run the commands. This is because both services `impala-state-store` and `impala-catalog` are down and need to be restarted.

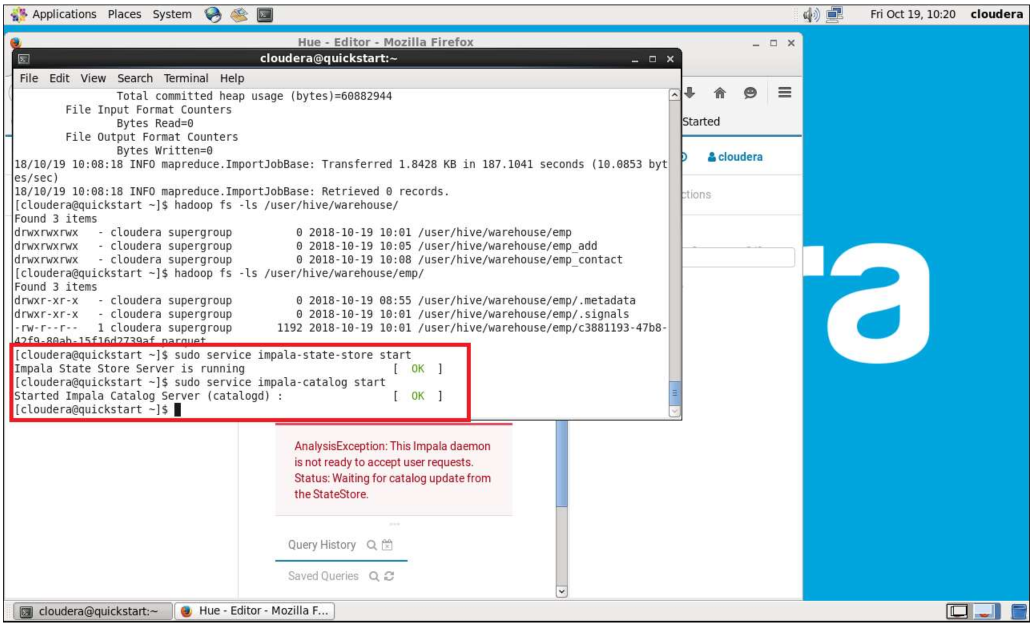

Applications Places System
Fri Oct 19, 10:19 cloudera
Hue - Editor - Mozilla Firefox
Hue - Editor
quickstart.cloudera:8888/hue/editor?editor=3
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
HUE
Query
Search data and saved documents...
Jobs
cloudera
Impala
Databases (0)
No databases found.
Impala
Add a name...
Add a description...
0s text
Assistant Functions
Tables
Search...
No tables identified.
Query History
Saved Queries
cloudera@quickstart:~
Hue - Editor - Mozilla F...

```
INVALIDATE METADATA;
SHOW TABLES;
```

AnalysisException: This Impala daemon is not ready to accept user requests. Status: Waiting for catalog update from the StateStore.

To fix the error, **back to the terminal and type the following commands:**

- `sudo service impala-state-store start`
- `sudo service impala-catalog start`



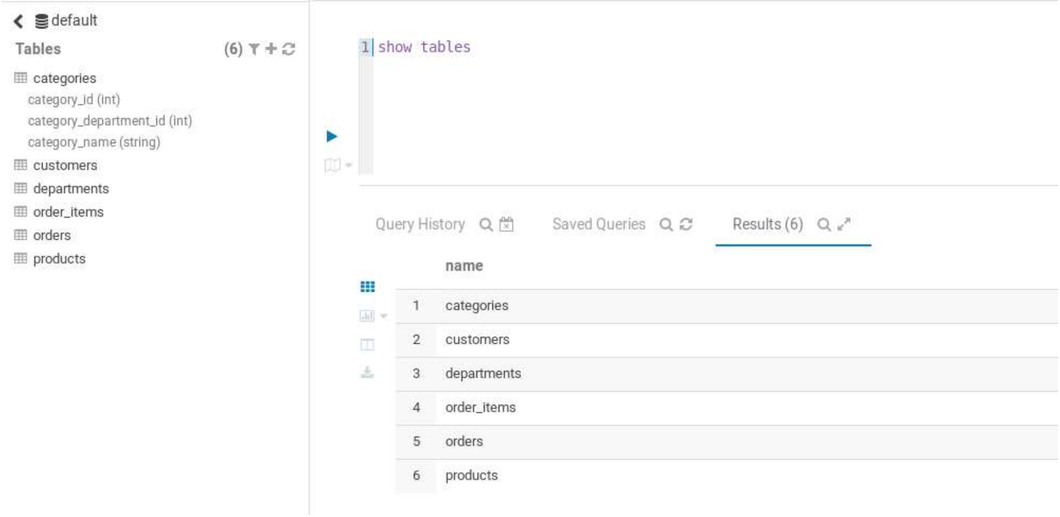
Applications Places System
Fri Oct 19, 10:20 cloudera
Hue - Editor - Mozilla Firefox
cloudera@quickstart:~
File Edit View Search Terminal Help
Started
cloudera
Query History
Saved Queries
cloudera@quickstart:~
Hue - Editor - Mozilla F...

```
        Total committed heap usage (bytes)=60882944
    File Input Format Counters
        Bytes Read=0
    File Output Format Counters
        Bytes Written=0
18/10/19 10:08:18 INFO mapreduce.ImportJobBase: Transferred 1.8428 KB in 187.1041 seconds (10.0853 byt
es/sec)
18/10/19 10:08:18 INFO mapreduce.ImportJobBase: Retrieved 0 records.
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/
Found 3 items
drwxrwxrwx   - cloudera supergroup          0 2018-10-19 10:01 /user/hive/warehouse/emp
drwxrwxrwx   - cloudera supergroup          0 2018-10-19 10:05 /user/hive/warehouse/emp_add
drwxrwxrwx   - cloudera supergroup          0 2018-10-19 10:08 /user/hive/warehouse/emp_contact
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/emp/
Found 3 items
drwxr-xr-x   - cloudera supergroup          0 2018-10-19 08:55 /user/hive/warehouse/emp/.metadata
drwxr-xr-x   - cloudera supergroup          0 2018-10-19 10:01 /user/hive/warehouse/emp/.signals
-rw-r--r--   1 cloudera supergroup       1192 2018-10-19 10:01 /user/hive/warehouse/emp/c3881193-47b8-
42f9-80ab-15f16d2739af.parquet
[cloudera@quickstart ~]$ sudo service impala-state-store start
Impala State Store Server is running                       [  OK  ]
[cloudera@quickstart ~]$ sudo service impala-catalog start
Started Impala Catalog Server (catalogd) :                 [  OK  ]
[cloudera@quickstart ~]$ 
```

AnalysisException: This Impala daemon is not ready to accept user requests. Status: Waiting for catalog update from the StateStore.

Now click on Hue icon again to refresh the page and then run the query. You can now see the tables in the result menu.

Now, your transaction data is readily available for structured queries in CDH, it's time to address a few questions. You can make a few basic/advanced SQL queries to get familiar with the retail_db database. Also, you can see the tables' properties visually in the Hue platform as shown in the following picture.



- categories
  - category_id (int)
  - category_department_id (int)
  - category_name (string)
- customers
- departments
- order_items
- orders
- products

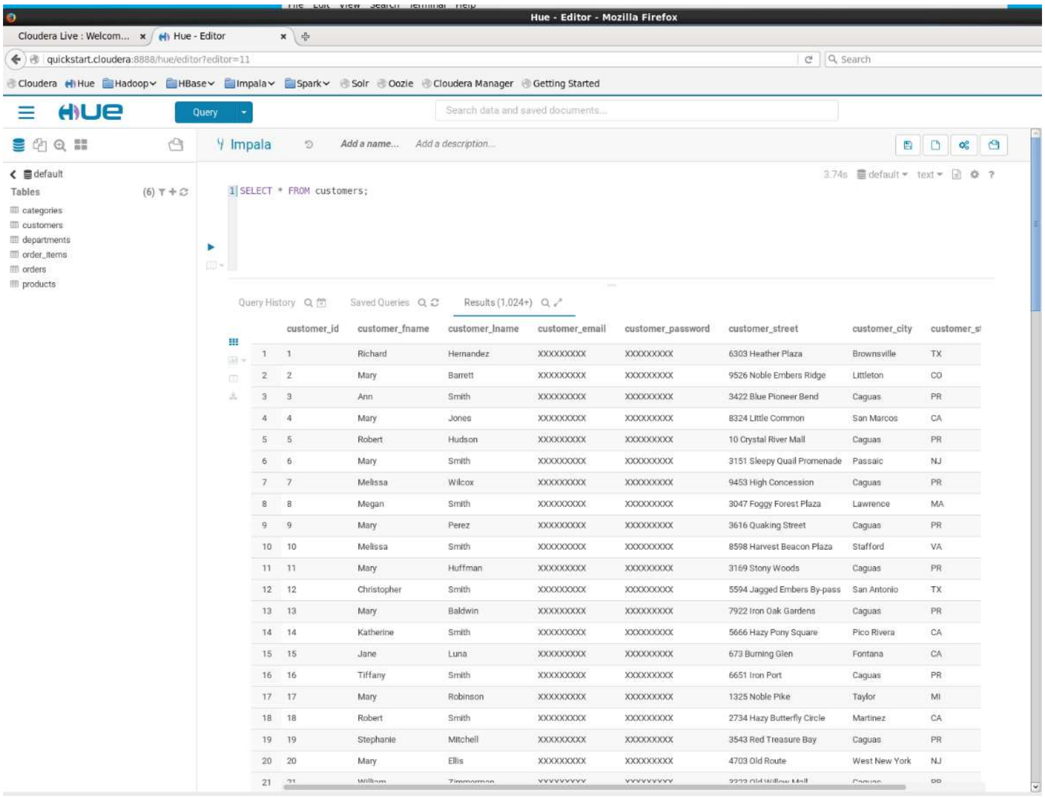
```
show tables
```

	name
1	categories
2	customers
3	departments
4	order_items
5	orders
6	products

Type and run the following SQL query in the editor to show the content of the `customer` table:

```
SELECT * FROM customers;
```

It shows you the `customer` table data in the result window.



```
SELECT * FROM customers;
```

	customer_id	customer_fname	customer_lname	customer_email	customer_password	customer_street	customer_city	customer_s
1	1	Richard	Hernandez	XXXXXXXXX	XXXXXXXXX	6303 Heather Plaza	Brownsville	TX
2	2	Mary	Barrett	XXXXXXXXX	XXXXXXXXX	9526 Noble Embers Ridge	Littleton	CO
3	3	Ann	Smith	XXXXXXXXX	XXXXXXXXX	3422 Blue Pioneer Bend	Caguas	PR
4	4	Mary	Jones	XXXXXXXXX	XXXXXXXXX	8324 Little Common	San Marcos	CA
5	5	Robert	Hudson	XXXXXXXXX	XXXXXXXXX	10 Crystal River Mall	Caguas	PR
6	6	Mary	Smith	XXXXXXXXX	XXXXXXXXX	3151 Sleepy Quail Promenade	Passaic	NJ
7	7	Melissa	Wilcox	XXXXXXXXX	XXXXXXXXX	9453 High Concession	Caguas	PR
8	8	Megan	Smith	XXXXXXXXX	XXXXXXXXX	3047 Foggy Forest Plaza	Lawrence	MA
9	9	Mary	Perez	XXXXXXXXX	XXXXXXXXX	3616 Quaking Street	Caguas	PR
10	10	Melissa	Smith	XXXXXXXXX	XXXXXXXXX	8598 Harvest Beacon Plaza	Stafford	VA
11	11	Mary	Huffman	XXXXXXXXX	XXXXXXXXX	3169 Stony Woods	Caguas	PR
12	12	Christopher	Smith	XXXXXXXXX	XXXXXXXXX	5594 Jagged Embers By-pass	San Antonio	TX
13	13	Mary	Baldwin	XXXXXXXXX	XXXXXXXXX	7922 Iron Oak Gardens	Caguas	PR
14	14	Katherine	Smith	XXXXXXXXX	XXXXXXXXX	5666 Hazy Pony Square	Pico Rivera	CA
15	15	Jane	Luna	XXXXXXXXX	XXXXXXXXX	673 Burning Glen	Fontana	CA
16	16	Tiffany	Smith	XXXXXXXXX	XXXXXXXXX	6651 Iron Port	Caguas	PR
17	17	Mary	Robinson	XXXXXXXXX	XXXXXXXXX	1325 Noble Pike	Taylor	MI
18	18	Robert	Smith	XXXXXXXXX	XXXXXXXXX	2734 Hazy Butterfly Circle	Martinez	CA
19	19	Stephanie	Mitchell	XXXXXXXXX	XXXXXXXXX	3543 Red Treasure Bay	Caguas	PR
20	20	Mary	Ellis	XXXXXXXXX	XXXXXXXXX	4703 Old Route	West New York	NJ
21	21	William	Zimmerman	XXXXXXXXX	XXXXXXXXX	3323 Old Willow Mall	Caguas	PR

(Take Screenshot #6)

Extra Task: Exporting data from Hadoop to MySQL

You can use the export command to export data back from the HDFS to MySQL.

```
$ sqoop export (generic-args) (export-args)

$ sqoop-export (generic-args) (export-args)
```

The target table must exist in the target database. The files that are given as input to the Sqoop contain records, which are called rows in the table. Those are read and parsed into a set of records and delimited with a user-specified delimiter. The default operation is to insert all the records from the input files into the database table using the INSERT statement. In update mode, Sqoop generates the UPDATE statement that replaces the existing record in the database. It only understands HDFS directories, not Hive or Hcatalog.

NOTE: It is mandatory that the table to be exported is created manually and is present in the MySQL database.

Let us take the customer's data in the file, in HDFS. The customer's data is available in the retail file in the retail_db directory in HDFS. It is mandatory that the table to be exported is created manually and is present in the database from where it has to be exported. So, now we are creating a table named customers_export in MySQL and SELECT all the data from the customers table.

NOTE: You must type the commands yourself. Don't copy and paste them as it wouldn't work.

1. Either keep the mysql terminal open, or issue a new terminal and invoke the following;

```
mysql> use retail_db;
```

2. Create a new table "customers_export" to start the importing / exporting from MySQL to HDFS

```
CREATE TABLE customers_export AS
SELECT
   customer_id,
   customer_fname,
   customer_lname,
   customer_email,
   customer_password,
   customer_street,
   customer_city,
   customer_state,
   customer_zipcode
FROM customers
WHERE 1=0;
```

3. Verify the new table "customers_export" exists;

```
mysql> show tables;
```

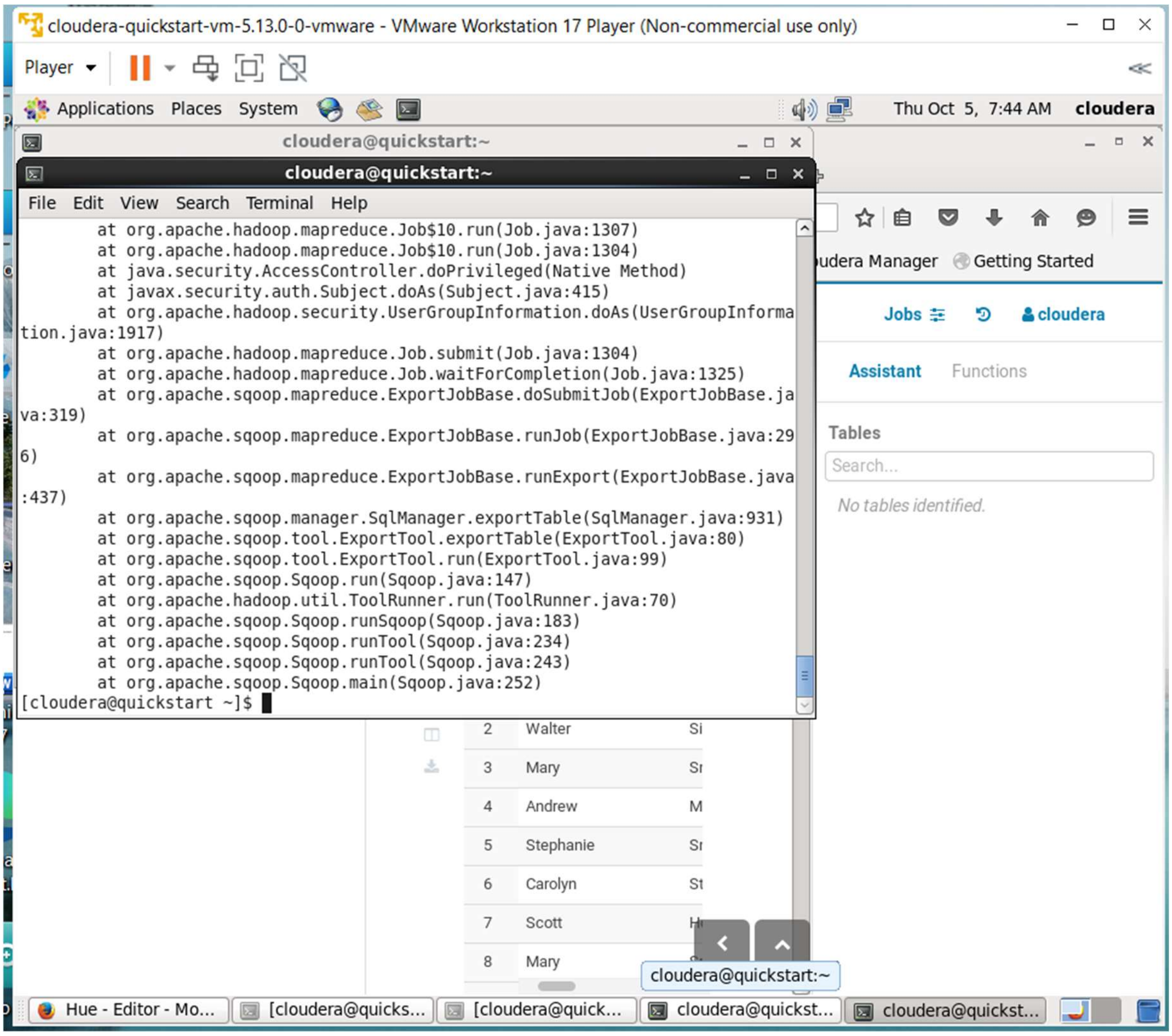
4. Verify it's currently empty;

```
mysql> select * from customers_export;
```

5. Now we import from MySQL to HDFS

```
[cloudera@quickstart ~]$  sqoop import \

  --connect jdbc:mysql://quickstart.cloudera:3306/retail_db \

  --username root \

  --password cloudera \

  --table customers \

  --target-dir /retail_db/customers
```

After about 3 minutes, you must see the following screenshot as the confirmation:



```
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1307)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1304)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:415)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInforma
tion.java:1917)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1304)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1325)
        at org.apache.sqoop.mapreduce.ExportJobBase.doSubmitJob(ExportJobBase.ja
va:319)
        at org.apache.sqoop.mapreduce.ExportJobBase.runJob(ExportJobBase.java:29
6)
        at org.apache.sqoop.mapreduce.ExportJobBase.runExport(ExportJobBase.java
:437)
        at org.apache.sqoop.manager.SqlManager.exportTable(SqlManager.java:931)
        at org.apache.sqoop.tool.ExportTool.exportTable(ExportTool.java:80)
        at org.apache.sqoop.tool.ExportTool.run(ExportTool.java:99)
        at org.apache.sqoop.Sqoop.run(Sqoop.java:147)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
        at org.apache.sqoop.Sqoop.runSqoop(Sqoop.java:183)
        at org.apache.sqoop.Sqoop.runTool(Sqoop.java:234)
        at org.apache.sqoop.Sqoop.runTool(Sqoop.java:243)
        at org.apache.sqoop.Sqoop.main(Sqoop.java:252)
[cloudera@quickstart ~]$ 
```

2	Walter	Si
3	Mary	Sı
4	Andrew	M
5	Stephanie	Sı
6	Carolyn	St
7	Scott	H
8	Mary	

(Take Screenshot #7)

6. Now we export from HDFS into MySQL to fill in the data of the new table "customers_export"

```
[cloudera@quickstart ~]$  sqoop export \

 --table customers_export \

 --connect jdbc:mysql://quickstart.cloudera:3306/retail_db \

 --username root \

 --password cloudera \

 --export-dir /retail_db/customers
```

7. Verify export transaction by issuing the following in the same open "MySQL" terminal.

```
mysql> select * FROM customers_export;
```

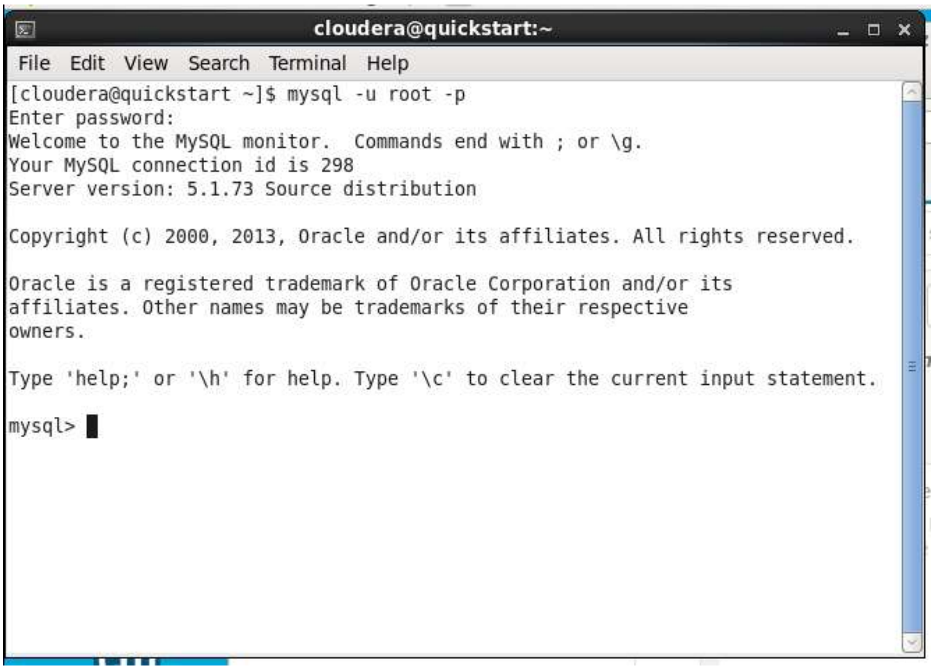
Or, open a new mysql terminal to verify exportation, and then follow the following instruction to verify.

Run MySQL by typing the following command:

```
mysql -u root -p
```

Then you'll be asked to enter MySQL password. Type **cloudera**

You'll then be presented with the MySQL monitor prompt:



```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 298
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> 
```

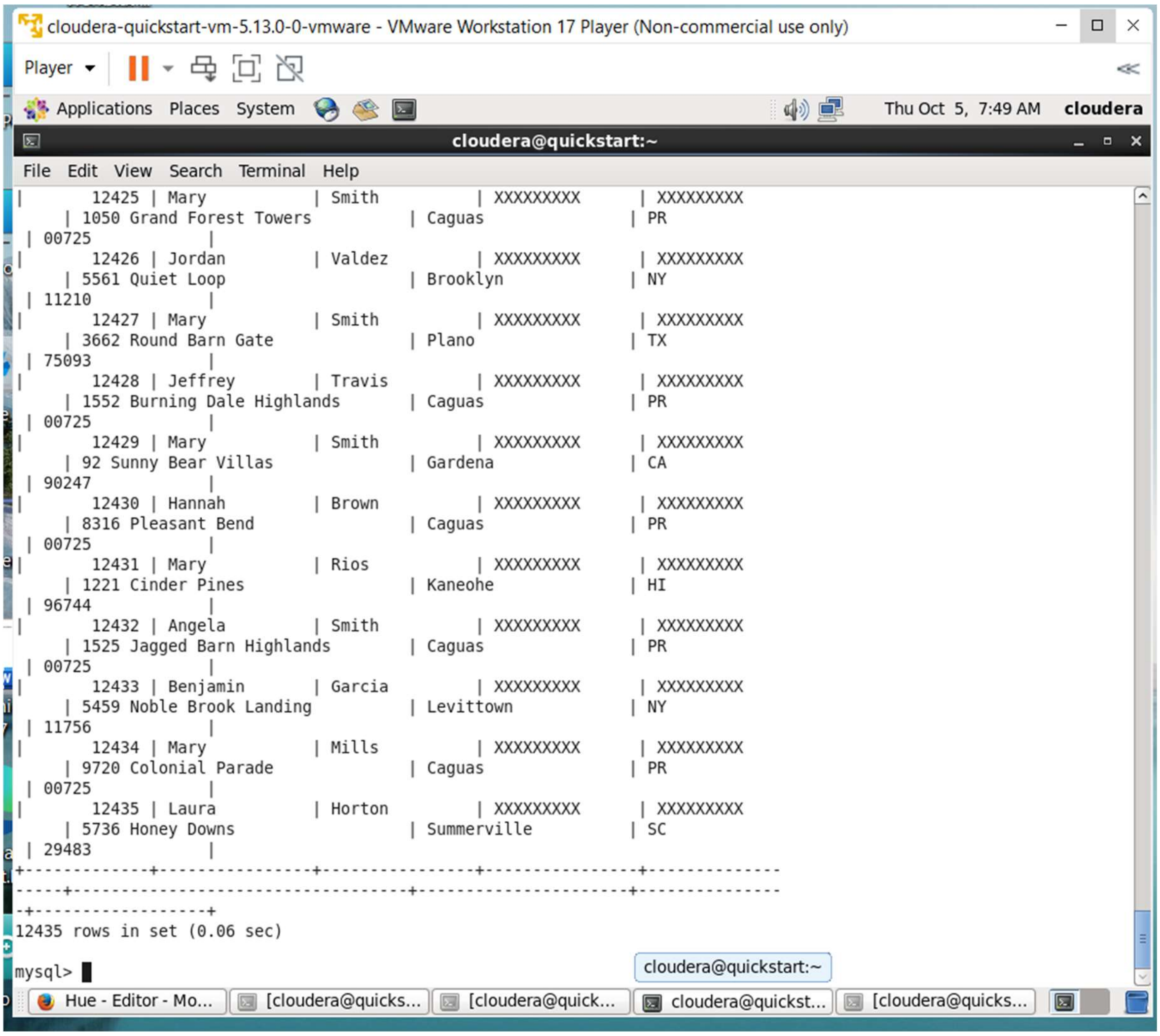
To get access to these databases invoke the following:

```
USE retail_db;
```

The following command is used to verify the table in the MySQL command line:

```
mysql> select * from customers_export;
```

After running this command, you will see the customer's data in MySQL


cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
Thu Oct 5, 7:49 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
cloudera@quickstart:~
Hue - Editor - Mo...
[cloudera@quicks...
[cloudera@quick...
cloudera@quickst...
[cloudera@quicks...

```
|       12425 | Mary           | Smith          | XXXXXXXXX      | XXXXXXXXX
     | 1050 Grand Forest Towers           | Caguas               | PR
 | 00725            |
|       12426 | Jordan         | Valdez         | XXXXXXXXX      | XXXXXXXXX
     | 5561 Quiet Loop                    | Brooklyn             | NY
 | 11210            |
|       12427 | Mary           | Smith          | XXXXXXXXX      | XXXXXXXXX
     | 3662 Round Barn Gate               | Plano                | TX
 | 75093            |
|       12428 | Jeffrey        | Travis         | XXXXXXXXX      | XXXXXXXXX
     | 1552 Burning Dale Highlands        | Caguas               | PR
 | 00725            |
|       12429 | Mary           | Smith          | XXXXXXXXX      | XXXXXXXXX
     | 92 Sunny Bear Villas               | Gardena              | CA
 | 90247            |
|       12430 | Hannah         | Brown          | XXXXXXXXX      | XXXXXXXXX
     | 8316 Pleasant Bend                 | Caguas               | PR
 | 00725            |
|       12431 | Mary           | Rios           | XXXXXXXXX      | XXXXXXXXX
     | 1221 Cinder Pines                  | Kaneohe              | HI
 | 96744            |
|       12432 | Angela         | Smith          | XXXXXXXXX      | XXXXXXXXX
     | 1525 Jagged Barn Highlands         | Caguas               | PR
 | 00725            |
|       12433 | Benjamin       | Garcia         | XXXXXXXXX      | XXXXXXXXX
     | 5459 Noble Brook Landing           | Levittown            | NY
 | 11756            |
|       12434 | Mary           | Mills          | XXXXXXXXX      | XXXXXXXXX
     | 9720 Colonial Parade               | Caguas               | PR
 | 00725            |
|       12435 | Laura          | Horton         | XXXXXXXXX      | XXXXXXXXX
     | 5736 Honey Downs                   | Summerville          | SC
 | 29483            |
+-------------+----------------+----------------+----------------+--------------
-----+-----------------------------------+----------------------+---------------
-+------------------+
12435 rows in set (0.06 sec)

mysql> 
```