**BUAD 5722: Big Data**

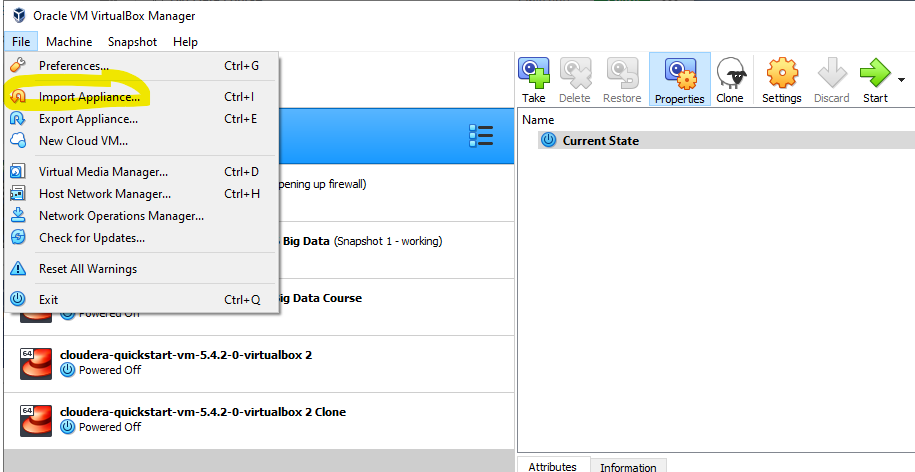
**Prof. Arturo Castellanos**

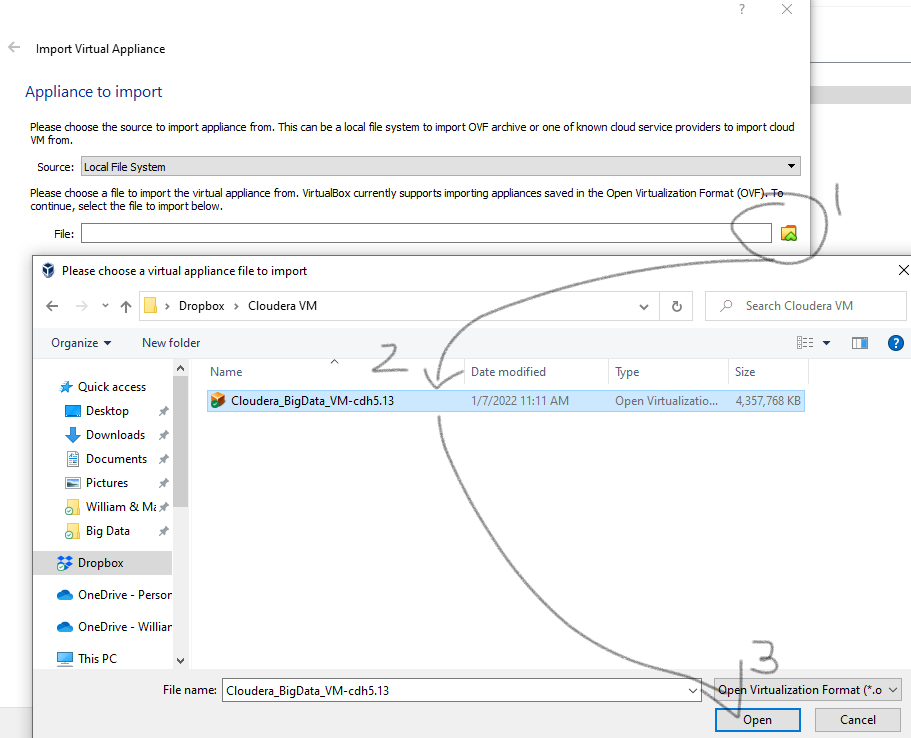
**Due: Feb 14th, 2022.**

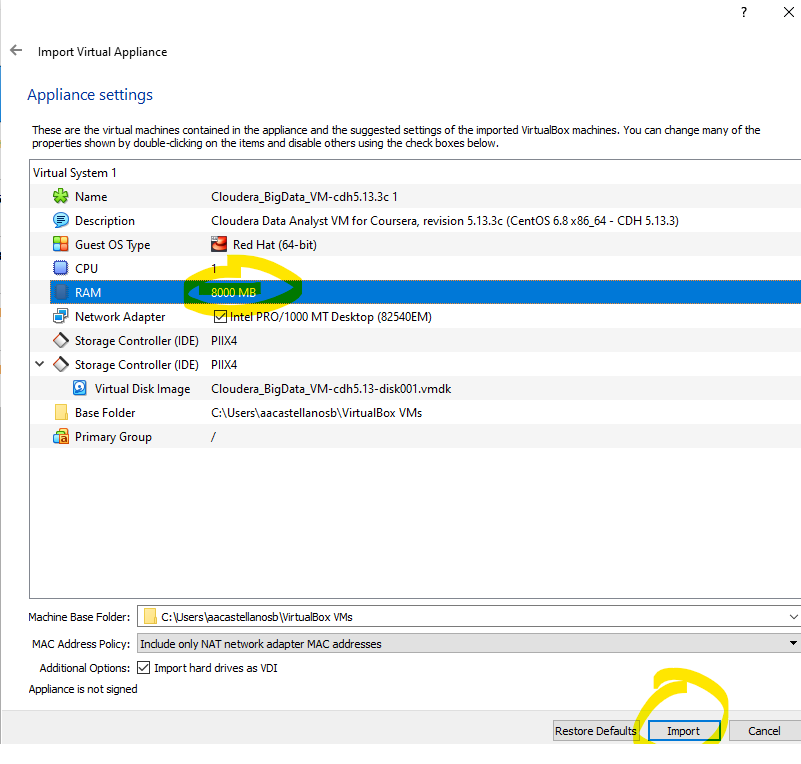
**Assignment – Import database from MySQL using Sqoop and writing HiveQL queries in HUE**

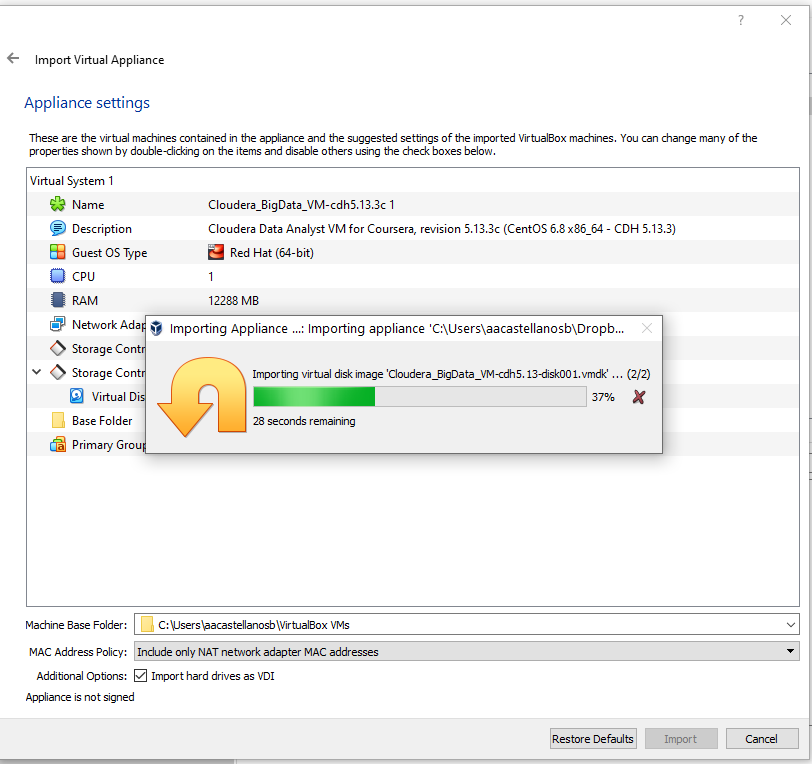
# Lab steps:

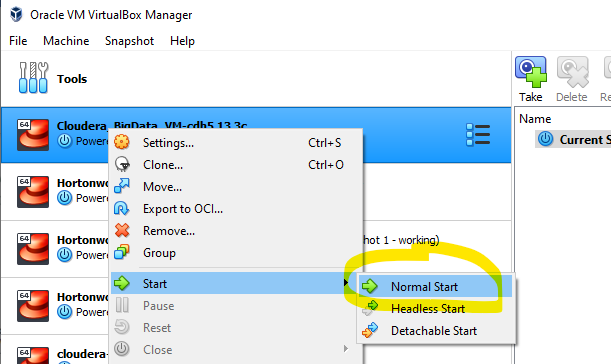
1. Downloading Virtualbox and Loading Cloudera’s quickstart VM
2. Check Olist\_db Dump file (this is a dump from the MySQL database)
3. Import into a MySQL database inside the Cloudera Quickstart VM
4. Check the database was created and it has data in it
5. Import all tables into HIVE using Sqoop import
6. Check if the tables are in HIVE and write your first HIVE QL query:
7. Access the HIVE database via HUE (web browser)
8. Answer questions by writing HIVEQL queries
9. **Loading Cloudera’s quickstart VM**
10. Install Virtualbox: Download from <https://www.virtualbox.org/>
11. Download [Quickstart VM image](https://www.dropbox.com/s/0a1zdfm3xkh33wu/Cloudera_BigData_VM-cdh5.13.ova?raw=1) (downloadable link)
12. Load the quickstart VM:











1. Create a shared folder between the VM and your local computer. [Watch this video](https://www.youtube.com/watch?v=5Ijhj2IcdFQ) on how to do it.
2. **Check Olist\_db Dump file (this is a dump from the MySQL database)**

Download the [olist\_dump\_tables.sql](https://www.dropbox.com/s/futtemd3zlh37xu/olist_db_dump_tables.sql?raw=1) (downloadable link) and [olist\_dump\_values.sql](https://www.dropbox.com/s/0uo0pe3ds3blido/olist_db_dump_values.sql?raw=1) (downloadable link) and take a look at the files.

1. **Next, we will import the olist\_db into MySQL inside the Quickstart VM.**

To do so, open a terminal window. Make sure you navigate to the shared directory where you can access the sql\_dump\_tables.sql and sql\_dump\_values.sql files before (a):

1. mysql -u your\_username -p your\_passw

**enter:** mysql -utraining -ptraining

1. Once you are in the mysql terminal:

First, we create the database with its tables:

**enter in terminal:**

mysql>source olist\_db\_dump\_tables.sql

Second, we insert the data into the database we just created:

**enter in terminal:**

mysql >use olist\_db;

mysql >SET FOREIGN\_KEY\_CHECKS = 0;

mysql >source olist\_db\_dump\_values.sql;

1. **Check the database was created and it has data in it**

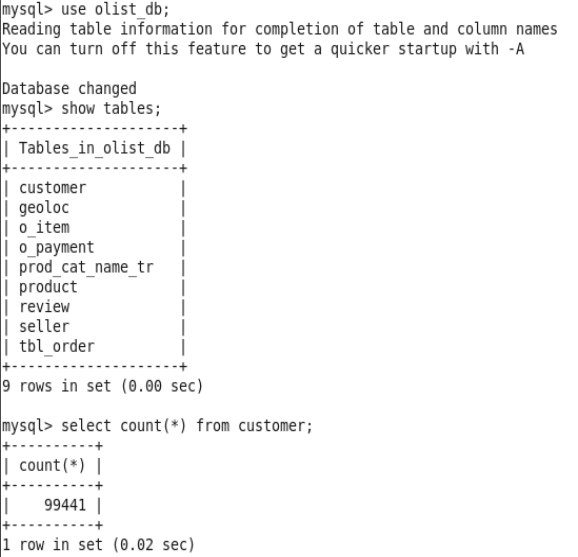
In terminal enter: mysql -utraining -ptraining

mysql > *use olist\_db;*

mysql > *show tables;*

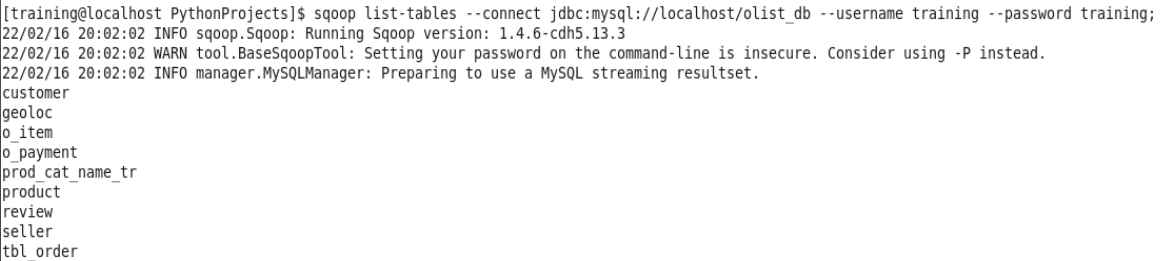
mysql > *select count(\*) from customer;*

**Output:**



|  |  |
| --- | --- |
| **Make sure all the data was imported into the MySQL database in Quickstart VM** | |
|  |  |

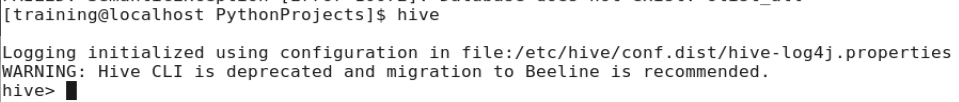
* List the tables in the Olist\_db database.  
  $ sqoop list-tables --connect jdbc:mysql://localhost/olist\_db --username training --password training;



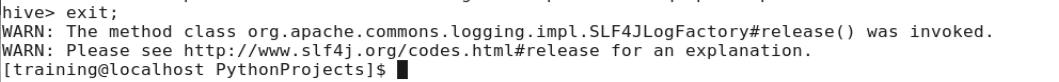
You can use sqoop to move a file to HDFS (see appendix) or directly to HIVE:

1. **Move all tables from MySQL database to HIVE:**
2. **create a database in Hive:**

* In terminal, enter *hive* to access Hive:

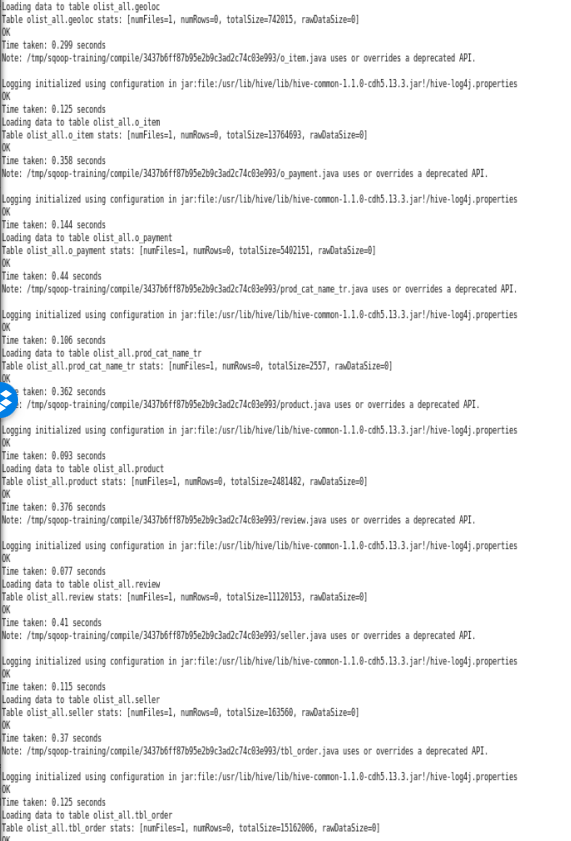


* create a database named: olist\_db => hive > CREATE DATABASE olist\_db;
* Exit Hive to terminal by entering exit;



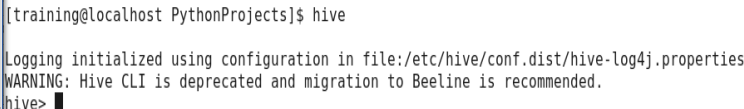
1. **Pull data from MySQL into the newly created database in Hive olist\_db using sqoop:**

sqoop import-all-tables --connect jdbc:mysql://localhost/olist\_db --username training --password training -m 1 --hive-import --hive-overwrite --hive-database olist\_db --hive-home /user/hive/warehouse --warehouse-dir=/user/hive/warehouse/olist\_staging.db

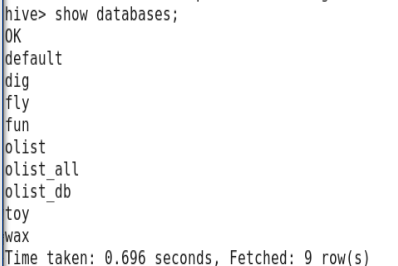


1. **Check if the table is in HIVE and write your first HIVE QL query**

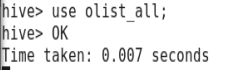
Enter in terminal: *hive*

**

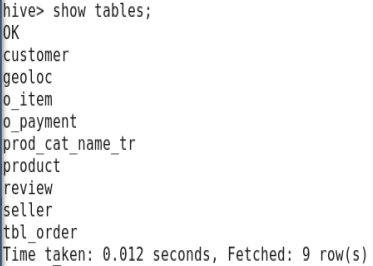
hive > *show databases;*

**

hive > use olist\_db;

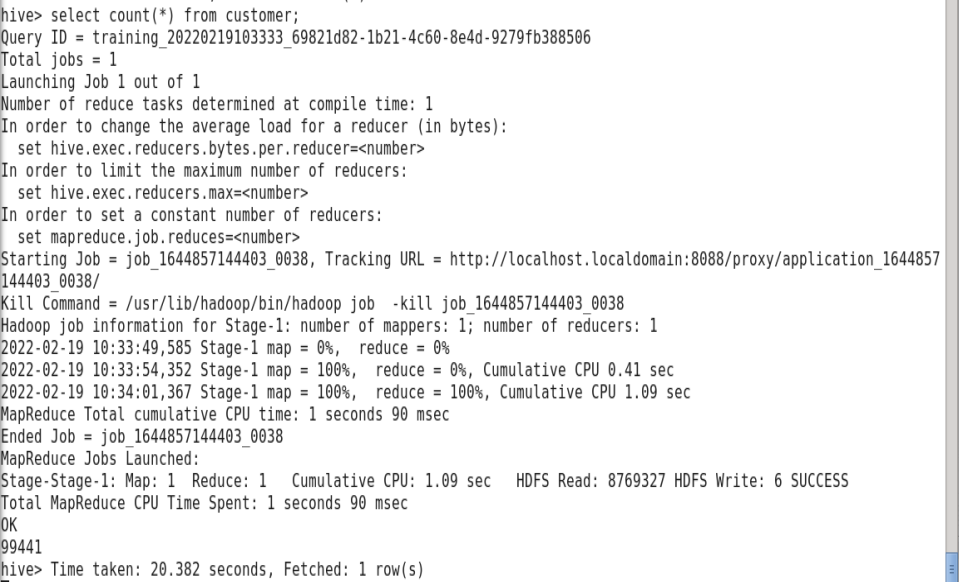


hive > show tables;

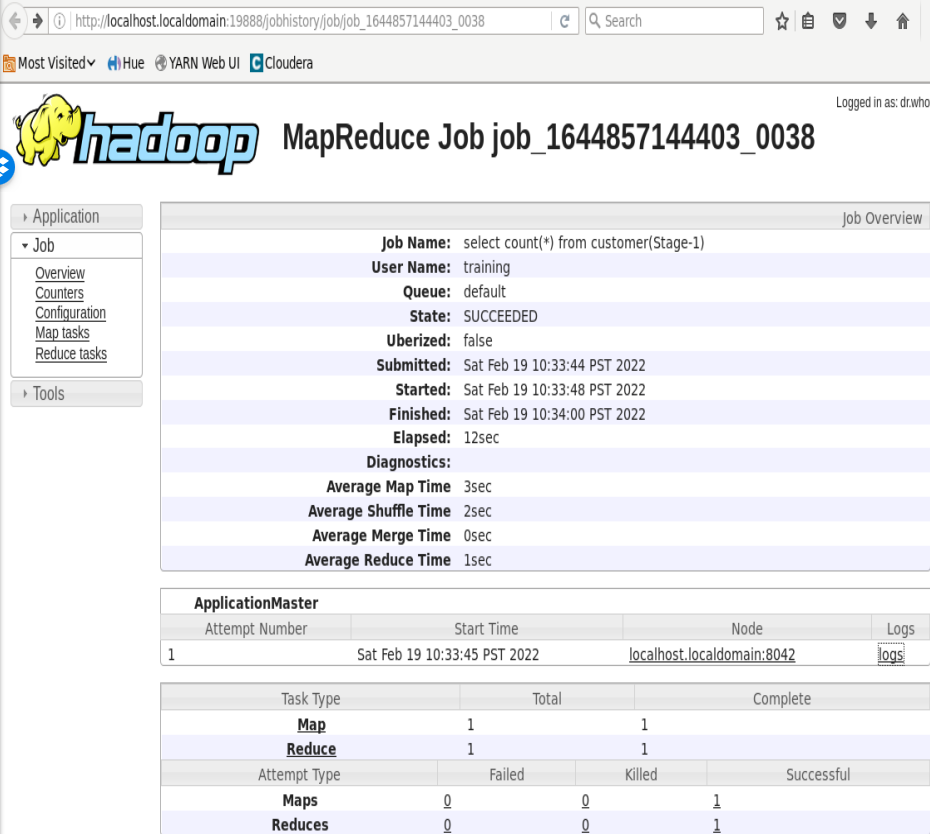


Next, we will write a query to retrieve the number of rows in the customer table. HIVE will write a map-reduce job that will execute such query.

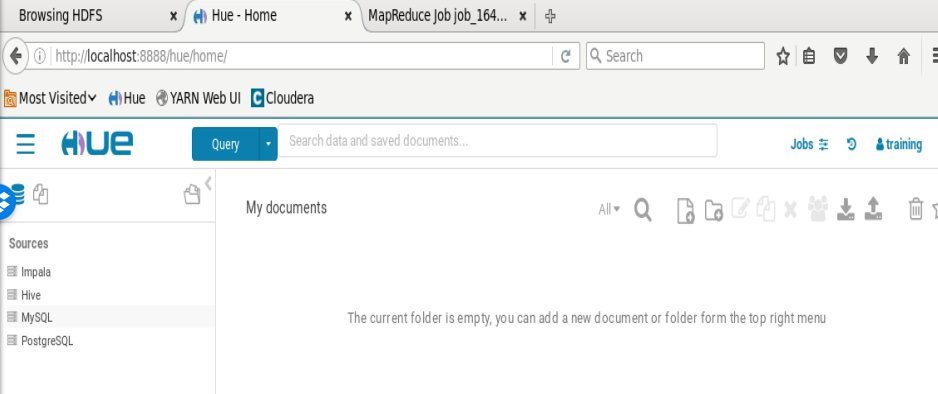
hive > select count(\*) from customer;



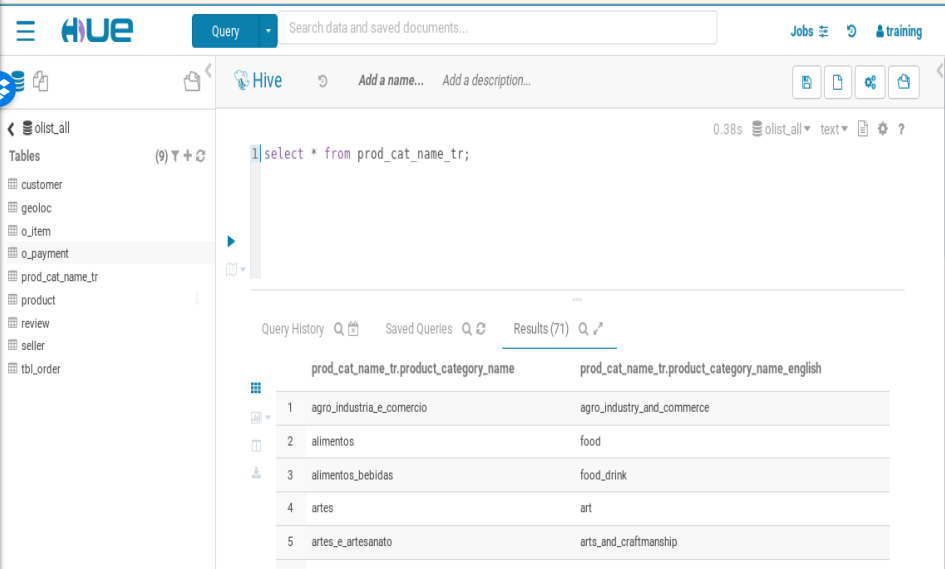
We can access the job tracking URL via the browser:



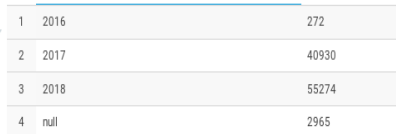
1. **Access the HIVE tables via HUE**



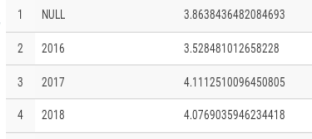
Click on the drop-down arrow next to Query, select Editor -> Hive and select all rows from prod\_cat\_name\_tr table:

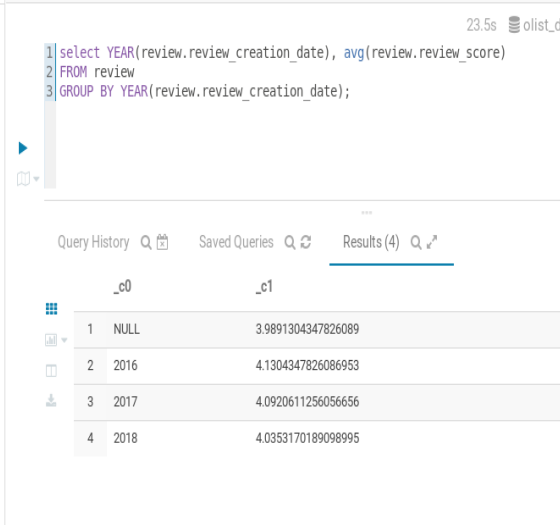


1. Answer the following questions by writing HIVE QL queries in Hue:
2. What is the total number of orders placed by year?

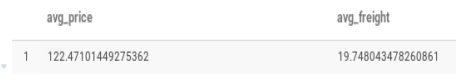


1. What is the average review score over the years?





1. What is the average item price and average freight value for items sold by seller with id 00ee68308b45bc5e2660cd833c3f81cc





# APPENDIX - other Sqoop capabilities

**Move an individual table to HDFS (do not run these commands):**

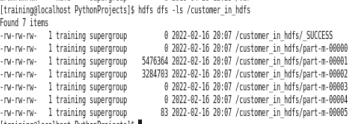
* Use Sqoop to import the customer table in the olist\_db database and save it in HDFS under /customer\_in\_HDFS:

$ sqoop import --connect jdbc:mysql://localhost/olist\_db --username training --password training --table customer --target-dir /customer\_in\_hdfs --null-non-string '\\N';

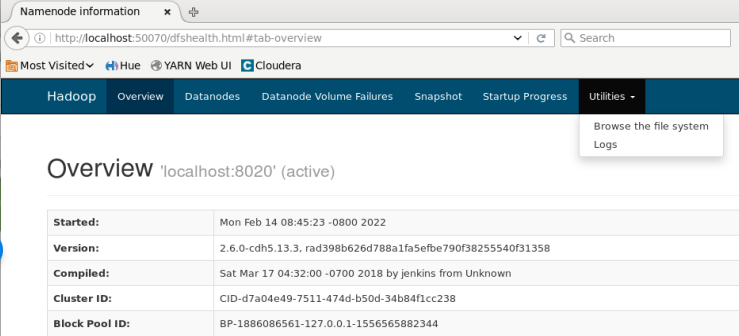
$ hdfs dfs -ls /

This shows you the contents of the root directory in HDFS.

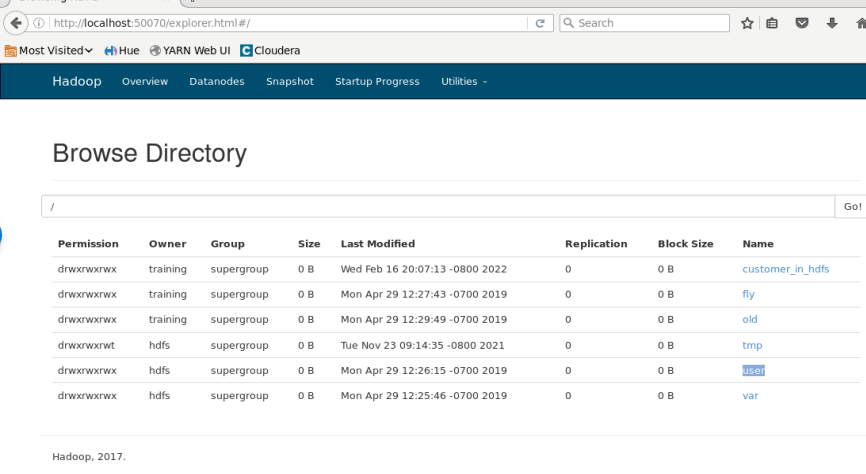
Navigate to your customer\_in\_HDFS directory:

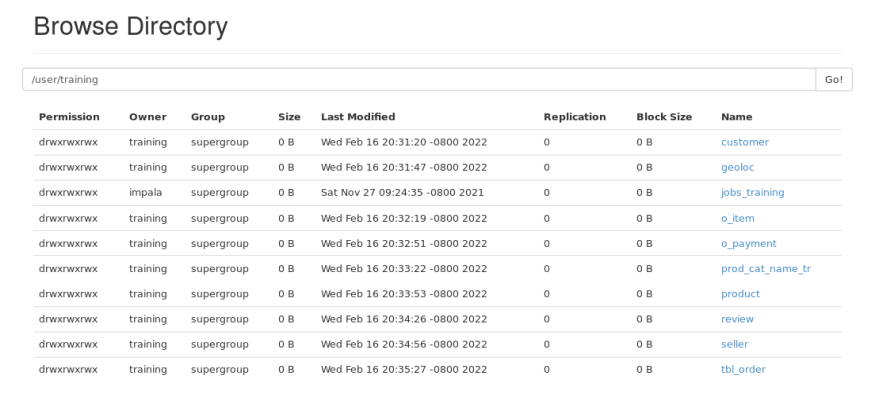


Open the browser (Firefox in Quickstart VM) and go type the following url: localhost:50070, click on Utilities -> Browse the file system

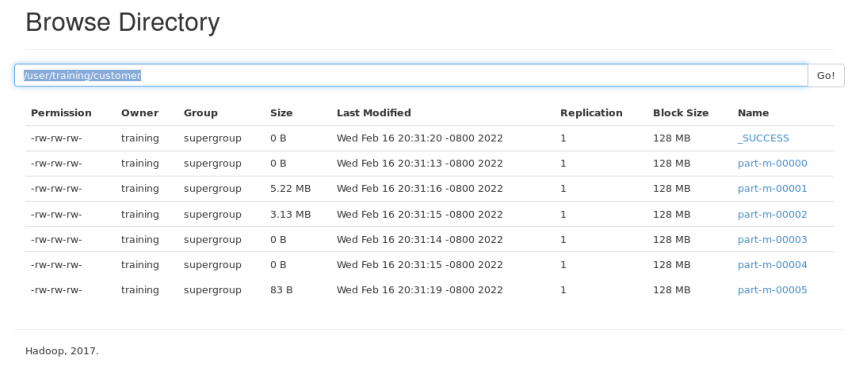


Note that in the case where we did not specified a target directory, those tables are stored in the **“user”** directory. Click on the *user 🡪 training* directory.

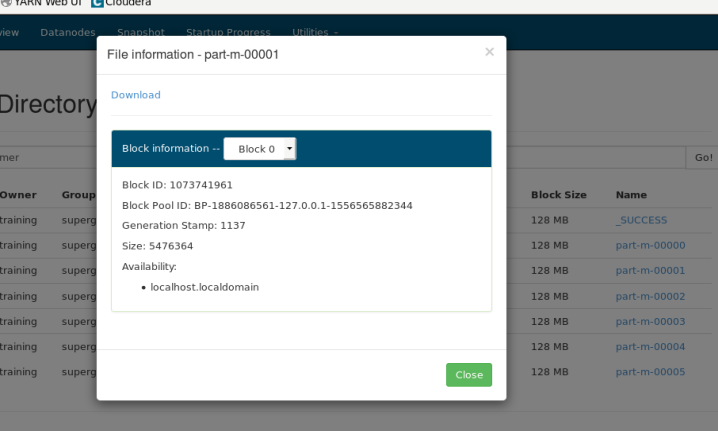




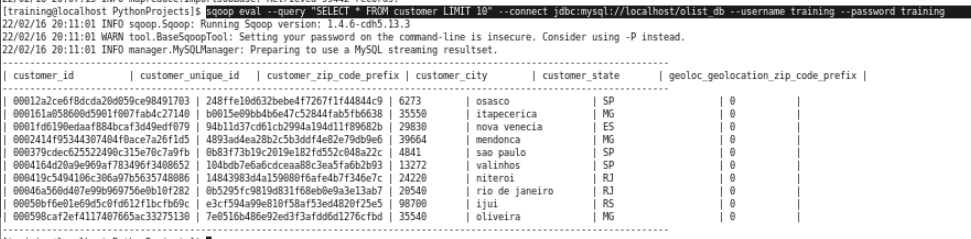
Note that for each of the tables imported using Sqoop, you will have the number of files as many mappers you had. In this case 6.



You can click on any of the parts to see the data:



* Evaluate the data:  
  $ sqoop eval --query "SELECT \* FROM customer LIMIT 10" --connect jdbc:mysql://localhost/olist\_db --username training --password training



**Move an invidual table to HIVE (do not run this command):**

sqoop import --connect jdbc:mysql://localhost/olist\_db --username training --password training --table customer -m 1 --hive-import --create-hive-table --hive-table customer--target-dir /user/hive/warehouse/ --enclosed-by '\"' --fields-terminated-by , --escaped-by \\

**Importing Incremental Updates (do not run this command):**$ sqoop import \ --connect jdbc:mysql://localhost/olist\_db --username training --password training --incremental append --null-non-string ' \\N' --table customer --target-dir /olist\_db/customer --check-column customer\_id --last-value <largest\_customer\_id>

# References:

[*https://sqoop.apache.org/docs/1.4.2/SqoopUserGuide.html*](https://sqoop.apache.org/docs/1.4.2/SqoopUserGuide.html)