Lab #6: Sqoop, Imapala, and Hive

Part I. Sqoop Questions

Answer the questions below, based on the tables from the *retail_db* database. ***Note***: you do not need to save a docker image after completing this section, but it is recommended that you test the commands using Hadoop. Because Sqoop uses Map Reduce, which is resource intensive, you may need to restart Hadoop (which is best done by creating a new container) in order to test the commands. You may also want to use a single mapper (by including *-m 1*), as this does not require as much memory and will be faster than using the default setting of 4 mappers.

1. What is the Sqoop command to import the *customers* table from the *retail_db* database into the HDFS directory *hdfs://user/cloudera/customers*.

2. What is the Sqoop command to import the product id, product name, and price of all products with a *product_category_id* of 38 directly into the folder *hdfs://user/cloudera/product_38/*. Note that to import the table contents directly into a folder, the *--target-dir* argument should be used.

3. What is the Sqoop command to import a table containing the number of customers from each state. The table should contain one column for the state and one column for the count, and should be imported into the directory */user/cloudera/state_counts/*

Part II. Hive/Impala Queries and Data Directories

For these questions, use the example databases provided by Hue.

1. Write a query to display the number of web page views by country from the *web_logs* table, sorted in decreasing order by number of views.

2. Write a query to display the code, description, and salary for the top 10 salaries from the *sample_08* table (Note that salaries with NULL values should not be displayed).

3. The file *more_salaries.txt* contains the code, description, and salary information for additional jobs. Assuming this file is in the */home/cloudera* directory of your container, what *hdfs* command can you use to copy this file to the appropriate HDFS location, so that the additional jobs are added to the *sample_08* table? Note: you can confirm that the new salaries have been added by running the following:

```
/* refresh tables */
invalidate metadata;

/* query new salaries */
select * from sample_08 where code like "99-%";
```

4. Write a query to summarize customer e-mail preferences by outputting the e-mail preference frequency (for each user, this is and the number of customers with that preference.

5. Write a query to display the names of all customers who have elected to receive e-mail surveys (i.e., their e-mail preferences for receiving surveys is to *true*).