CSC 343: Big Data Programming and Management
Lab #5: MySQL Assignment

Being familiar with *MySQL* will enable us to transfer data from MySQL into Hadoop, and to query and analyze data on HDFS using Hive (which provides an SQL-like interface) and Impala (which uses SQL).

1. Complete the Basic MySQL Tutorial at **http://www.mysqltutorial.org/**, from MySQL SELECT through MySQL Sub Query, but skipping the following: MySQL LEFT JOIN, MYSQL RIGHT JOIN, MYSQL CROSS JOIN, MYSQL Self Join MySQL ROLLUP)

2. Specify the following queries, using the mysqltutorial.org sample database:

    a. Select the first name, last name, and email address of all employees with the last name of ‘Patterson’.

    b. Select the first name and last name of all employees whose last name begins with a ‘P’.

    c. Select the customer name (which is a company name) and credit limit of all customers from CT, and sort them by credit limit from highest to lowest.

    d. Select the customer names and state for customers from CT, NY, and NJ.

    e. Find the total number of customers

    f. From the *payments* table, find the total amount spent by each customer. Display this information in two columns using the names *customerNumber* and *total*.

    g. Submit a query that shows the customer name, the payment amount, and the date.

    h. Find the customer with the highest single payment, and display only this customer (you can display the customer number or name), the payment date, and the payment amount.

3. Specify the following queries using the *retail_db* database available from the *gdancik/cloudera* image. To run mysql from this image, the *mysql* server must be running, which will be the case if you run *docker-quickstart*. Alternatively, you may create a container running *bash* (without any other cloudera options) and then execute `service mysqld start`. You can then access the mysql shell by running the following command: `mysql -u root -p`. You will be prompted for a password, which is *cloudera*.

a. Output the customer id, last name, and first name for all customers with the last name of ‘Jones’, using the column names *ID*, *firstName*, and *lastName*.

b. Output the customer id, last name, and first name for all customers whose last name is 3 characters long, using the column names *ID*, *firstName*, and *lastName*.

c. Output the 10 most common last names, using column names of *lastName* and *count*.

d. The total number of customers (12,435) is given by the query

```
select count(*) from customers;
```

and the total number of customers with orders (12,405) is given by

```
select count(distinct order_customer_id) from orders;
```

Therefore, there are 30 “customers” who have not placed orders. Output the customer ID, first name, and last name of these 30 customers. Note that this query is relatively slow (> 20 seconds for me) because we need to look up *customer_ids* in the *orders* table, but there is not an index on this column. To speed up this query, you can optionally add an index on this column using the command

```
create index customer_id_idx ON orders (order_customer_id);
```

With the index the query should take < 1 second.