CPSC 4330 Big Data Analytics

In-class Exercise 13
Create, Populate Hive-Managed Tables and Run Hive Queries

In this exercise, you will create and populate Hive-managed tables. You will also write HiveQL queries to analyze data in Hive tables that have been populated with data.

Prepare the Data

1. On Docker terminal, type “hive” to start hive, create a database named “dualcore”. You should be able to see that there is a directory “dualcore.db” created in /user/hive/warehouse/. In the subsequent steps, you will create four tables (“customers”, “products”, “orders”, “order_details”) in the “dualcore” database and populate each table.

2. On Docker terminal, create the following directories on HDFS.

- /user/hive/warehouse/dualcore.db/customers
- /user/hive/warehouse/dualcore.db/products
- /user/hive/warehouse/dualcore.db/orders
- /user/hive/warehouse/dualcore.db/order_details

3. Download four files (“customers”, “products”, “orders”, and “order_details”) from Canvas and copy them to Docker. Then copy each file to their corresponding directory on HDFS. For example, copy the file “customers” to the HDFS directory /user/hive/warehouse/dualcore.db/customers/; copy the file “products” to the HDFS directory /user/hive/warehouse/dualcore.db/products/, etc.

4. Write Hive scripts to create tables for customers, products, orders and order_details. See below for the data types in each table. When you create tables, remember to specify the delimiter and directory of their data. Note that the four files are all tab-delimited.

(1) Customers

cust_id int
fname string
lname string
address string
city string
state string
zipcode string

(2) Products

prod_id int
brand string
name string
price int
cost int
shipping_wt int

(3) Orders

order_id int
cust_id int
order_date timestamp

(4) Order_details

order_id int
prod_id int

Running a Query from the Hive Shell

Dualcore ran a contest in which customers posted videos of interesting ways to use their new tablets. A $5,000 prize will be awarded to the customer whose video received the highest rating.

However, the registration data was lost due to an RDBMS crash, and the only information they have is from the videos. The winning customer introduced herself only as “Bridget from Kansas City” in her video.

1. You will need to write a Hive query that identifies the winner’s record in the customer table so that Dualcore can send her the $5,000 prize.

Question: Which customer did your query identify as the winner of the $5,000 prize?

Running a HiveQL Script

The rules for the contest described earlier require that the winner bought the advertised tablet from Dualcore between May 1, 2013 and May 31, 2013. Before Dualcore can authorize the accounting department to pay the $5,000 prize, you must ensure that Bridget is eligible.

1. Write a HiveQL script to find out the date when Bridget bought the advertised tablet (product id = 1274348).

Question: Did Bridget order the advertised tablet in May?

2. Write a query to count the number of records in the *customers* table.

Question: How many customers does Dualcore serve?

3. Write a query to find the ten states with the most customers.

Question: Which state has the most customers?