CPSC 4330 Big Data Analytics
Winter 2025

Homework 5

Due: 10:55 am, Monday, March 17

Problem 1:

Dualcore recently started a loyalty program to reward their best customers. Dualcore has a sample of the data (loyalty_data.txt) that contains information about customers who have signed up for the program, including their customer ID, first name, last name, email, loyalty level, phone numbers , a list of past order IDs, and a struct that summarizes the minimum, maximum, average, and total value of past orders. You will create the table, populate it with the provided data, and then run a few queries to reference some fields.

1. Write a HiveQL query 'hw5_1.hql' to create the table *loyalty_program*.

2. Write a query 'hw5_2.hql' to load the data in *loyalty_data.txt* (posted on Canvas) into Hive.

3. Write a query 'hw5_3.hql' to select the *HOME* phone number for customer ID 1200866.

4. Write a query 'hw5_4.hql' to select the third element from the *order_ids* for customer ID 1200866.

Problem 2:

For this problem, you will need customers, orders, order_details, and products tables that you have prepared in exercise 13.

1. Write a HiveQL query 'hw5_5.hql' to find how many products have been bought by the customer 1071189?

2. Write a HiveQL query 'hw5_6.hql' to find how many customers have spent more than 300000 on the total price of all products that s/he has bought?

3. Write a HiveQL query 'hw5_7.hql' to list the customers (cust_id only) who have not placed any order.

Problem 3:

In the database 'dualcore' (the one you created in exercise 13), create a table named *ratings* for storing tab-delimited records using this structure:

Field Name	Field Type
posted	TIMESTAMP
cust_id	INT
prod_id	INT
rating	TINYINT
message	STRING

Populate the 'ratings' table directly by copying product ratings data of 2012 ("ratings_2012.txt" posted on Canvas) to the table's directory in HDFS. Similarly, populate the 'ratings' table by copying product ratings data of 2013("ratings_2013.txt" posted on Canvas) to the table's directory in HDFS.

Customer ratings and feedback are great sources of information for both customers and retailers like Dualcore. However, customer comments are typically free-form text and must be handled differently.

Before delving into analyzing customer comments, you will begin by analyzing the numeric ratings customers have assigned to various products.

1. Write a HiveQL query 'hw5_8.hql' to find the product with the lowest average rating among products with at least 50 ratings.

2. We observed earlier that customers are very dissatisfied with one of the products that Dualcore sells. Although numeric ratings can help identify which product that is, they don't tell Dualcore why customers don't like the product. We could simply read through all the comments associated with that output to learn this information, but that approach doesn't scale. Now we want to analyze the comments to get more information on why customers don't like the product.

   Write a HiveQL query 'hw5_9.hql' to find the five most common trigrams (three-word combinations) in the comments for the product that you have identified in the previous question. Products' comments are in the "message" field of table *ratings*.

3. Among the patterns you see in the result of the previous question is the phrase "ten times more." This might be related to the complaints that the product is too expensive. Write a

query ‘hw5_10.hql’ to list the product’s (the product that you have identified in question 1 of this problem) comments that contain the phrase.

Submission

Zip all 10 files and use Canvas to submit.