MSCI: 6110 Fall 2019 Big Data Management and Analytics Homework 2

Due: 10/9/2019 3:30PM. Submit on ICON Dropbox

Total points: 100

Instructions:

1. Please submit a single .txt file. Name it as <your hawkid>.txt , for example, xunzhou.txt

2. Write your Hive commands to answer each question sequentially. Use comments, i.e. lines starting with double dash ("--") to list question numbers. Your code should be able to run correctly without any error on its own. Any explanatory text should be added as HiveQL comments, too.

Preparation: Request 3 nodes in the interactive mode. Before you run any query, run the following commands in Hive to enable dynamic partitioning. If you exit Hive, you must run them again the next time you start.

```
SET hive.exec.dynamic.partition = true;
SET hive.exec.dynamic.partition.mode = nonstrict;
SET hive.exec.max.dynamic.partitions.pernode = 1000;
```

1. (20 pts) Use the nyc_taxi_Aug table from HW1 for the following questions. Create another Hive table "nyc_taxi_Aug_part_day" for the NYC_Yellow Taxi Data and define partitions based on the day of the pickup time. Write a query to load all the data from nyc_taxi_Aug table into all the partitions of the new table. Make sure to load the correct part to each partition.

2. (20 pts) Calculate the total number of trips on day 31 (use pickup_datetime). Write two HiveQL queries using (a) the nyc_taxi_Aug table (without partitions) and (b) the nyc_taxi_ Aug_part_day table (with partitions), respectively. Report the "Total MapReduce CPU Time Spent" of these two queries (from the log) using HiveQL comments. Which one is faster?

3. (20 pts) Create another table called "nyc_taxi_Aug_buk_dist" for the same dataset. In this table, do not define any partitions. Instead, define 10 buckets based on the distance of the trips (trip_distance). Write a query to load the entire dataset from nyc_taxi_Aug to this table.

4. (20 pts) Write a Hive query to calculate the average fare amount of trips between 10 miles and 15 miles (inclusive). Do this query twice using (a) the nyc_taxi_Aug table, and (b) the table created in Q3, respectively. Report the "Total MapReduce CPU Time Spent" of these two queries (from the log) in using HiveQL comment. Which one is faster?

5. (20 pts) Compute statistics for all the columns of the table nyc_taxi_Aug. Show formatted statistics of the “passenger_count” column. Report the min and max values in a HiveQL comment line.