MSCI: 6110 Fall 2019 Big Data Management and Analytics Homework 1

Due: 9/19/2019 6:00PM. 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: Get the 2014 January NYC yellow taxi trip data and save it in your home directory. No need to submit code for this step (see Lecture 2 notes for details).

Read the data descriptions and samples here to understand the fields.
https://data.cityofnewyork.us/Transportation/2014-Yellow-Taxi-Trip-Data/gn7m-em8n

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;
```

1. (20 pts) Create a Hive table "nyc_taxi_Jan" for the NYC Yellow Taxi Data (January 2014). Load all the yellow taxi January 2014 data into the nyc_taxi_Jan table.

2. (20 pts) Write a HiveQL query to show the top 10 longest trips in the nyc_taxi_Jan table. Show all the columns. Sort the results by distance in descending order.

3. (20 pts) The drop-off/pick-up area of the LaGuardia Airport is between latitudes [40.766703, 40.774724] and longitudes [-73.877101, -73.859692]. Write a Hive query to find the total number of passengers picked up by yellow taxi at this airport during each hour of day (0 ~ 23) in the whole month. Sort the rows by hour in ascending order.

4. (20 pts) Create another Hive table "nyc_taxi_Jan_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_Jan table into all the partitions of the new table. Make sure to load the correct part of data to each partition.

5. (20 pts) Calculate the total number of trips on day 31 (use pickup_datetime). Write HiveQL using (a) the nyc_taxi_Jan table (without partitions) and (b) the nyc_taxi_Jan_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?