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

## Due: 9/25/2019 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 Linux/Hadoop/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:**

(1) Read the data descriptions and samples here to understand the fields. This is the Yellow Taxi Trip Data from NYC. https://data.cityofnewyork.us/Transportation/2014-Yellow-Taxi-Trip-Data/gn7m-em8n. (Update 9/15: The data fields in this link are no longer consistent with the actual data we have. We have less columns in our dataset. Please check the following link for correct columns: https://www.biz.uiowa.edu/faculty/xzhou/Teaching/nyc_data_columns.txt)

(2) Log into the server through the shell access terminal. Start an interactive session with the PSC server with 3 nodes requested (this may take a few minutes). You can decide your time for reservation. Load the Hadoop and Hive packages using the command discussed in class.

No need to submit the commands for the above steps.

1. (10pts) First, create a Linux folder in your home directory called "Homework" using Linux commands. Then enter this folder. Visit the HDFS path /user/ever930/hw1 using Hadoop commands and check what file is in there. There should be a file, which stores the taxi trip data for NYC in August 2014. Obtain this file from HSDF and save it (use the same file name) into the Linux folder "Homework" you just created. Submit the Linux/HDFS commands in every step. One in each line. Do not take screenshots.

2. (20 pts) Start Hive. Create a new database named after your hawkid. Create a Hive table "nyc_taxi_Aug" for the NYC Yellow Taxi Data (August 2014). Load all the yellow taxi August 2014 data into the nyc_taxi_Aug table. Make sure to check the headers. There are two rows in the file before the data, where the first row is the column names and the second row is a blank row.

3. (20 pts) Write a HiveQL query to show the top 10 longest trips in the nyc_taxi_Aug table. Show all the columns. Sort the results by distance in descending order.

4. (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.

5. (30 pts) Calculate the average speed of taxis in NYC during each hour of day (0-23) on August 15. The average speed within an hour is calculated as the average of the speed of all the trips that started and ended both in this hourly slot. Ignore trips that run across different hourly slots. The speed of a single trip is calculated as total distance divided by total time duration (miles/hour). Sort the results by hour in ascending order.