MSCI: 6110 Fall 2019 Big Data Management and Analytics Homework 3

Due 10/17/2019 6PM. Submit on ICON Dropbox

Total points: 100

Instructions:

1. For HW3 please submit **a zip file with an R script with your code + all figure image files generated in the questions**. Mark questions and your explanations using R comments in your R script.

2. Your code should be able to run correctly in SparkR. Load all the libraries needed. No need to include the install.packages() functions. You can assume all these packages are installed.

Q1 (20 pts). Start sparkR session correctly. Use the nyc_taxi_Jan table in Hive you previously created for HW1. Create a Spark DataFrame in R and load all the taxi trips started on **Jan 1**. Show the top 25 rows of this Spark DataFrame.

Q2 (20 pts). Write R code in SparkR to find trips on Jan 1 paid for by credit card (CRD) and trips paid for by cash (CSH) from the Spark DataFrame created in Q1. Save these two groups of trips into two new Spark DataFrames, namely, "Trips_CRD" and "Trips_CSH". Use the Spark DataFrame functions (e.g., SELECT, WHERE, groupBy) instead of using the sql() function in this question. Show the first 20 rows of each DataFrame to validate the results.

Q3 (20 pts). Use the two DataFrames created in Q2 to perform the following analysis in SparkR. First, remove trips with no tip information (null). Then calculate the percentage of tip with respect to all the other charges (i.e., tip_percent = tip_amount/(total_payment – tip_amount)). Generate two new Spark DataFrames with the extra tip_percent column added to the "Trips_CRD" and the "'Trips_CSH" DataFrames, respectively. You will get two new Spark DataFrames, "Trips_CRD_Pct" and "Trips_CSH_Pct". Again, you must use Spark DataFrame functions to solve this problem, rather than sending queries to Hive using sql().

Q4 (15 pts). Use the spark DataFrames you obtained from Q3 to calculate on average how many percent passengers tip (tip_percent) when paying by credit card and by cash, respectively. The results should be two numbers. Again you must not use the sql() function.

Q5. (25 pts) For each day in January, calculate the total fare revenue of all the taxi drivers. The total revenue of a day is the total amount of money paid in all the trips that **end** on that day. You need to load the whole month's data again into a Spark DataFrame (using sql()). Then use the SparkR DataFrame functions (e.g., SELECT, WHERE, groupBy) to aggregate the results into another Spark DataFrame with 31 rows. Convert the second spark DataFrame to an R data frame. Finally, generate a line chart to visualize the data, where the X-axis is the day of month and the Y-axis is the total revenue of all the taxi drivers on that day. Name this image file "plot.png". Save the image file and include it in your submission.