**Assignment 3**

**Big Data Analytics**

MET CS777

Faculty - Farshid Alizadeh-Shabdiz, PhD, MBA

## **1 Description**

The goal is to analyze a data set consisting of New York City Taxi trip reports in the Year 2013. The dataset was released under the FOIL (The Freedom of Information Law) and made public by Chris Whong

(https://chriswhong.com/open-data/foil\_nyc\_taxi/).

# 2 Taxi Data set

The data set itself is a simple text file. Each taxi trip report is a different line in the file. Among other things, each trip report includes the starting point, the drop-off point, corresponding timestamps, and information related to the payment. The data are reported by the time that the trip ended, i.e., upon arrive in the order of the drop-off timestamps. The attributes present on each line of the file are in order as it was shown in table 1. The data files are in comma separated values (CSV) format.

Table

Description automatically generated

Table 1: Taxi data set fields.

The data files are in comma separated values (CSV) format. Example lines from the file are:

07290D3599E7A0D62097A346EFCC1FB5,E7750A37CAB07D0DFF0AF7E3573AC141, 2013-01-01,00:00:00,2013-01-01 00:02:00,120,0.44,-73.956528,40.716976,-73.962440, 40.715008,CSH,3.50,0.50,0.50,0.00,0.00,4.50

22D70BF00EEB0ADC83BA8177BB861991,3FF2709163DE7036FCAA4E5A3324E4BF, 2013-01-01,00:02:00,2013-01-01 00:02:00,0,0.00,0.000000,0.000000,0.000000,0.000000, CSH,27.00,0.00,0.50,0.00,0.00,27.50

0EC22AAF491A8BD91F279350C2B010FD,778C92B26AE78A9EBDF96B49C67E4007, 2013-01-01,00:01:00,2013-01-01 00:03:00,120,0.71,-73.973145,40.752827,-73.965897 73.965897,40.760445,CSH,4.00,0.50,0.50,0.00,0.00,5.00

Follow PySpark Code below to cleanup the data and get the required field.

Text

Description automatically generated

You can also pre-process the data and store it in your own cluster storage.

# Obtaining the Dataset

The data set (93 MB compressed, uncompressed 384 MB) is available at the following url on Google Cloud : gs://meetcs777/taxi-data-sorted-small.csv.bz2

You can download or access the data sets using the following internal URLs:

|  |  |
| --- | --- |
|  | **Google Cloud** |
| Small data set | gs://metcs777-fa/taxi-data-sorted-small.csv.bz2 |
| Large data set | gs://metcs777-fa/taxi-data-sorted-large.csv.bz2 |

**Assignment Tasks**

Task 1 - Top-10 Active Taxis (25 points)

Many different taxis have had multiple drivers. Write and execute a Python program that computes the top ten taxis that have had the largest number of drivers. Your output should be a set of (medallion, number of drivers) pairs.

Note: You should consider that this is a real world data set that might include wrongly formatted data lines. You should clean up the data before the main processing, a line might not include all of the fields. If a data line is not correctly formatted, you should drop that line and do not consider it.

Task 2 - Top-10 Best Drivers (35 Points)

We would like to figure out who the top 10 best drivers are in terms of their average earned money per minute spent carrying a customer. The total amount field is the total money earned on a trip. In the end, we are interested in computing a set of (driver, money per minute) pairs.

Task 3 - Best time of the day to Work on Taxi (40 Points)

We would like to know which hour of the day is the best time for drivers that has the highest profit per miles. Consider the surcharge amount in dollar for each taxi ride (without tip amount) and the distance in miles, and sum up the rides for each hour of the day (24 hours) – consider the pickup time for your calculation. The profit ratio is the ration surcharge in dollar divided by the travel distance in miles for each specific time of the day.

Profit Ratio = (Surcharge Amount in US Dollar) / (Travel Distance in miles) We are interested to know the time of the day that has the highest profit ratio.

Task 4 – Advanced question (10 points)

Here are two further tasks for advanced groups.

* What percentage of taxi customers pay with cash and what percentage use electronic cards? Analyze these payment methods for different time of the day and provide a list of percentages for each hour of the day? As a result provide two numbers for total percentages and a list like (hour of the day, percent paid card)
* We would like to measure the efficiency of taxis drivers by finding out their average earned money per mile. (Consider the total amount which includes tips, as their earned money) Implement a Spark job that can find out the top-10 efficient taxi divers.

Task 4 – More Advanced questions for interested students – needs research (No Grade)

* What are mean, median, first and third quantiles of tip amount? How do find the median?
* Using the IQR outlier detection method find out the top-10 outliers.

# **Academic Misconduct Regarding Programming**

In a programming class like our class, there is sometimes a very fine line between ”cheating” and acceptable and beneficial interaction between peers. Thus, it is very important that you fully understand what is and what is not allowed in terms of collaboration with your classmates. We want to be 100% precise, so that there can be no confusion.

The rule on collaboration and communication with your classmates is very simple: you cannot transmit or receive code from or to anyone in the class in any way—visually (by showing someone your code), electronically (by emailing, posting, or otherwise sending someone your code), verbally (by reading code to someone) or in any other way we have not yet imagined. Any other collaboration is acceptable.

The rule on collaboration and communication with people who are not your classmates (or your TAs or instructor) is also very simple: it is not allowed in any way, period. This disallows (for example) posting any questions of any nature to programming forums such as StackOverflow. As far as going to the web and using Google, we will apply the ”two line rule”. Go to any web page you like and do any search that you like. But you cannot take more than two lines of code from an external resource and actually include it in your assignment in any form. Note that changing variable names or otherwise transforming or obfuscating code you found on the web does not render the ”two line rule” inapplicable. It is still a violation to obtain more than two lines of code from an external resource and turn it in, whatever you do to those two lines after you first obtain them.

Furthermore, you should cite your sources. Add a comment to your code that includes the URL(s) that you consulted when constructing your solution. This turns out to be very helpful when you’re looking at something you wrote a while ago and you need to remind yourself what you were thinking.