Assignment 1

MET CS 777 - Big Data Analytics
Spark Data Wrangling (20 points)

GitHub Classroom Invitation Link
https://classroom.github.com/a/uIbp6F5z

1 Description

The goal of this assignment is to implement a set of Spark programs in Python (using Apache Spark). Specifically, your Spark jobs will analyze a data set consisting of New York City Taxi trip reports in 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. 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 arriving in the order of the drop-off timestamps. The attributes present on each line of the file are, in order:

	Attribute	**Description**
0	medallion	an md5sum of the identifier of the taxi - vehicle bound (Taxi ID)
1	hack license	an md5sum of the identifier for the taxi license (Driver ID)
2	pickup datetime	time when the passenger(s) were picked up
3	dropoff datetime	time when the passenger(s) were dropped off
4	trip time in secs	duration of the trip
5	trip distance	trip distance in miles
6	pickup longitude	longitude coordinate of the pickup location
7	pickup latitude	latitude coordinate of the pickup location
8	dropoff longitude	longitude coordinate of the drop-off location
9	dropoff latitude	latitude coordinate of the drop-off location
10	payment type	the payment method -credit card or cash
11	fare amount	fare amount in dollars
12	surcharge	surcharge in dollars
13	mta tax	tax in dollars
14	tip amount	tip in dollars
15	tolls amount	bridge and tunnel tolls in dollars
16	total amount	total paid amount in dollars

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 00:02:00,0,0.00,0.00000,0.00000,0.0000,0.000000,
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
```

You can use the following notebook as a helper code to clean up the data, get the required field. Assignment 1-Helper Code.ipynb - Colaboratory (google.com) (https://colab.research.google.com/drive/1Uaa1MzSNqgCpAXzhezTim_GOUcmrpsA0)

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

NOTE:

- To submit the task to Cloud, please use the code_template.py.
- In your submission, only the code based on the code_template needs to be included.

3 Obtaining the Dataset

There are two versions of the dataset: the Small dataset (93 MB compressed, 384 MB uncompressed) for implementation and testing purposes (roughly 2 million taxi trips) and the large dataset (9GB compressed, 32GB uncompressed).

To download and use the dataset on your computer, use the following HTTPS links:

- Small dataset: https://storage.googleapis.com/met-cs-777-data/taxi-data-sorted-small.csv.bz2
- Large dataset: https://storage.googleapis.com/met-cs-777-data/taxi-data-sorted-large.csv.bz2

When running your code on the cloud, you can have direct access to the files using the following internal links:

- Small Data Set: gs://met-cs-777-data/taxi-data-sorted-small.csv.bz2
- Large Data Set: gs://met-cs-777-data/taxi-data-sorted-large.csv.bz2

4 Assignment Tasks

4.1 Task 1: Top-10 Active Taxis (5 points)

Based on the observation that numerous taxis have been operated by multiple drivers, you are required to develop and execute a Spark Python program to determine the top ten taxis that have had the highest count of distinct drivers. The program should output a set of pairs consisting of the taxi's medallion number and the corresponding count of drivers.

Note: The provided dataset is derived from real-world data, which may contain inaccurately formatted lines. Therefore, it is essential to cleanse the data prior to primary processing. If a line lacks certain fields or is improperly formatted, it should be dropped from consideration.

4.2 Task 2 - Top-10 Best Drivers (15 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.

4.3 Task 3 - The best time of the day to Work on Taxi (For Advanced Students - no points)

We would like to know which hour of the day is the best time for drivers that have the highest profit per mile. Consider the surcharge amount in dollars 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 dollars 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 in the time of day with the highest profit ratio.

4.4 Task 4 - (For Advanced Students - no points)

Here are some further tasks for advanced groups.

- How many percent of taxi customers pay with cash, and how many percent use electronic cards? Analyze these payment methods for different times of the day and provide a list of percentages for each day's time. 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 taxi drivers' efficiency 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 the top 10 efficient taxi divers.
- What are the mean, median, first, and third quantiles of tip amount? How do you find the median?
- Using the IQR outlier detection method, find out the top 10 outliers.

5 Important Considerations

5.1 Machines to Use

One thing to be aware of is that you can choose virtually any configuration for your Cloud Cluster - you can choose different numbers of machines and different configurations of those machines. And each is going to cost you differently! Since this is real money, it makes sense to develop your code and run your jobs locally, on your laptop, using the small data set. Once things are working, you'll then move to the Cloud.

As a proposal for this assignment, you can use the e2-standard-4 machines on the Google Cloud, one for the Master node and e2-highmem-8 (8 vCPU, 4 core, 64 GB memory) for two worker nodes. You will have a cluster with a total of 16 vCPU and 128GB RAM. 100 GB of disk space will be enough. The cost of the cluster for 5 hours, which is more than enough to finish the assignment, is estimated at $6.

Remember to delete your cluster after the calculation is finished!!!

More information regarding Google Cloud Pricing can be found here: https://cloud.google.com/products/calculator. As you can see, the average server costs around 50 cents per hour. That is not much, but **IT WILL ADD UP QUICKLY IF YOU FORGET TO SHUT OFF YOUR MACHINES**. Be very careful, and stop your machine as soon as you are done working. You can always come back and start your machine or create a new one easily when you begin your work again.

Another thing to be aware of is that Google and Amazon charge you when you move data around. To avoid such charges, do everything in the "Iowa (us-central1)" region. That's where data is, and that's where you should put your data and machines.

- You should document your code as well as possible.
- Your code should be compilable on a Unix-based operating system like Linux or macOS.

5.2 Academic Misconduct Regarding Programming

In a programming class like ours, there is sometimes a very fine line between "cheating" and acceptable and beneficial peer interaction. Thus, it is essential that you fully understand what is and what is not allowed in collaboration with your classmates. We want to be 100% precise so 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. However, you cannot take more than two lines of code from an external resource and actually include them in your assignment in any form. Note that changing variable names or otherwise transforming or obfuscating code you found online 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 can be very helpful when you're looking at something you wrote a while ago and need to remind yourself what you were thinking.

5.3 Turnin

- For each task and Spark job you ran, include a screenshot of the Spark History.
  It is important to include URLs in the screenshots to demonstrate that you executed your code on the cloud. Otherwise, we cannot verify whether the code was executed in your cloud account.

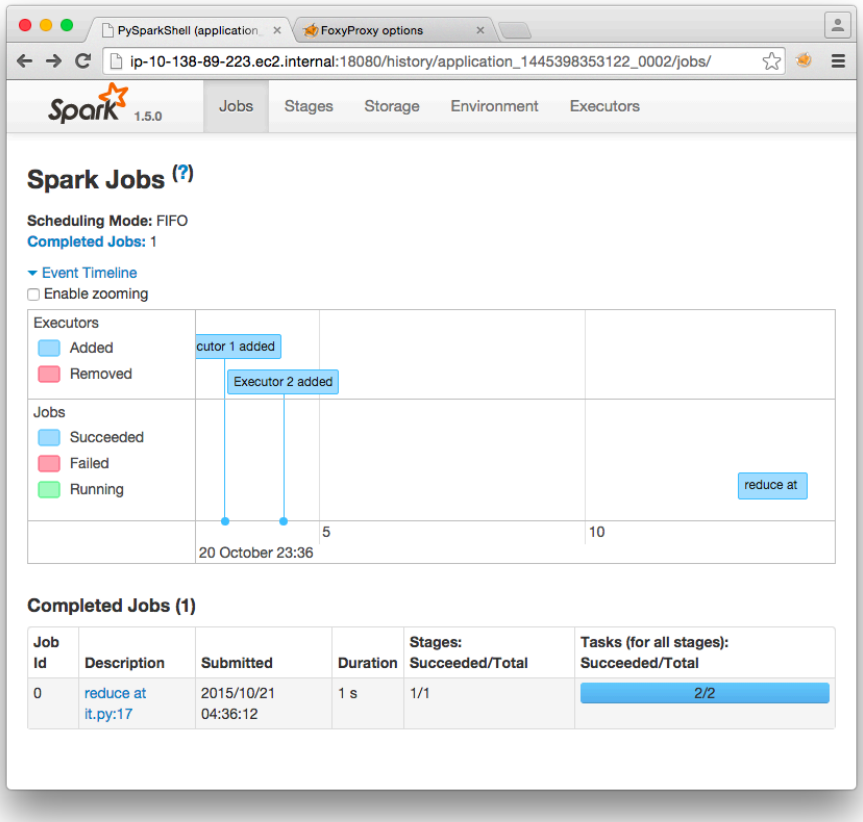

Figure 1: Screenshot of Spark History

- Fill in the results in the provided template.
- Please zip up all of your code and your document (use .zip only, please!), or attach each piece of code and your document to your submission individually.
- Please have the latest version of your code on GitHub. Zip the GitHub files and submit your latest version of the assignment work to Blackboard. We will consider the latest version on the Blackboard, but it should match your code on GitHub exactly.

Remember to delete your cluster after the calculation is finished!!!