Analyzing New York City Taxi Data with Spark Structured Streaming

Please upload the files in one ZIP on Moodle by X

This project aims to use Spark Structured Streaming to analyse the value of different stocks over time.

The Dataset

You can download the NYC taxi dataset from the link http://www.andresmh.com/nyctaxitrips/. However, it is huge. So, we have already prepared a sample that you can find in this link.

Each row of the file after the header represents a single taxi ride in CSV format. For each ride, we have some attributes of the cab (a hashed version of the medallion number) as well as the driver (a hashed version of the hack license, which is what licenses to drive taxis are called), some temporal information about when the trip started and ended, and the longitude/latitude coordinates for where the passenger(s) were picked up and dropped off.

We are mainly interested in each Trip's:

- Some Unique ID for the car (license)
- Pick-up location
- Pick-up time
- Drop-off location
- Drop-off time

Additionally, in the dataset archive, you will find a .geojson file that contains the geographical boundaries of the different boroughs of New York City. This information is needed to compute the answers to the queries needed on the data.

Development environment

For this project, you should reuse the same environment you used for the practice related to Spark and Spark Structured Streaming.
Use the notebooks you find on Moodle, in particular

- The notebook **Kafka_Producer_for_Project** reads the file and ingests the data into Kafka with the schema needed.
    - Consider that for the timestamp, it uses the current one.
    - Put the csv with the data in the same folder.
- The notebook **Project Template** registers Spark to the stream and puts it in a manageable form. Be careful not to edit the cell already there unless you know what you are doing ;).

What do we need to compute?

Imagine you want to create a dashboard that shows real-time statistics about what's at the taxi drivers

1. [Query 1] Utilization over a window of 5, 10, and 15 minutes per taxi/driver. This can be computed by computing the idle time per taxi. How does it change? Is there an optimal window?
2. [Query 2] The average time it takes for a taxi to find its next fare(trip) per destination borough. This can be computed by finding the time difference, e.g. in seconds, between the trip's drop off and the next trip's pick up within a given unit of time
3. [Query 3] The number of trips that started and ended within the same borough in the last hour
4. [Query 4] The number of trips that started in one borough and ended in another one in the last hour

Grading Rubric

This project contributes 15% of the total grade. The breakdown of the grade is as follows:

Item	Points
Query1 – utilization	4
Query 2	2
Query 3	2
Query 4	2
Final presentation (better with data viz)	5 (+1)
Total	15

Please upload the files in one ZIP on Moodle by X