EECS 4415 Project Proposal

Taxi vs. Uber in NYC

Mahmoud Alsaeed
York University

Ashkan Moatamed
York University

Koko Nanah Ji
York University

Dong Hoon Lee
York University

1 Domain Description and Motivation

1.1 What is the data domain?

The data domain for this project is transportation and more specifically Uber and Taxis. The data sets we are using contain information about the pickup time, trip time, fares, pickup address, drop off the address and other miscellaneous information regarding New York City customers.

1.2 What is the goal of your project?

In this project, we want to do a compound analysis to examine Taxi and Uber datasets. This analysis aims to extract patterns between the customers of Taxi and Uber services and find the correlations between them as well as the differences. The results of this analysis can be then used to further improve the services that Taxis and Uber provide and help them better understand their customers.

1.3 What is the motivation for rigorous data analytics?

The need arises from the demand for a better analysis platform which can extract as much insight as possible about the behavior of customers in New York City. Furthermore, rigorous data analysis allows us to make predictions that can help the service providers improve and personalize their services.

1.4 What are the questions that you want to answer?

- Economics Analysis: Is Uber driving the Taxi industry out of business in New York City?
- Residential Analysis: What area (i.e., rural vs. urban) seems to be more interested in using Uber/Taxi?
- Time and Space Analysis: In a given location, does the ratio of Taxi/Uber usage change during the day?
- Business Improvement: What is the ratio of Taxi to non-Taxi in a given location? Is this in any way related to how the weaker business is advertising/promoting their services?
- Ambitious Question: How does the income of a household correlate to the usage of a service (i.e., Taxi vs. Uber)?

1.5 Why is the analysis important?

It is necessary to gain insights from the collected data and to answer the questions that can help service providers improve their business and gain a better understanding of their customers.

1.6 What are a few potential applications?

- Provide the findings of this analysis to Taxi companies and Uber to help them improve their revenue and customer service based on data instead of their intuition.
- Provide real-time analysis to the customers (for example through an app) to help them make better choices regarding which service to use at what time.

2 The architecture of the Proposed Solution

2.1 Data analytics architecture

Process the data in Spark's batch mode:

- Use HDFS with a Spark distributed cluster to retain scalability, high fault tolerance, and efficient execution.
- Store the data in a columnar structure to improve the performance of analytics.
- Perform aggregation using the Spark API.
- Display the results using MatPlotLib or Folium Python module.

2.2 Description of the data collection/ingestion process, data storage, data processing, data serving and data visualization.

Data Ingestion:

- Retrieve the data from Kaggle as a batch and preprocess it as needed to enforce schema (i.e., schema on write).
- Perform data cleaning on Spark (through flatMap and reduce functions).

Data Storage:

- Columnar distributed storage.

Data Processing:

- Discover patterns and make predictions using the Spark API and specifically MLlib.

Data Serving:

- Text file containing all positions and times.

Data visualization:

- Graphs created with MatPlotLib and/or Folium.

2.3 Overall architecture and data flow in the system

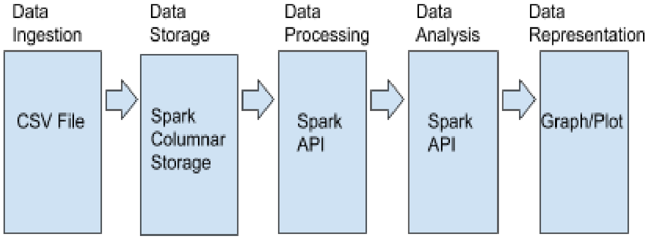

Figure 1**: The Architecture and the flow of data in the proposed solution.**

2.4 Limitations and difficulties with the chosen approach

- Incomplete data (i.e., lacking some vital information that can be used to improve the analysis further) due to the challenge of obtaining data from companies like Uber because of privacy laws. Furthermore, some countries may have stricter privacy policies than others, which entirely prohibit sharing any information about the trips taken by customers. This in turn negatively impacts the scalability of this application into those regions.
- Suffering from no significant limitations regarding scalability except the financial one!
- Difficulty in combining the schemata of two different datasets and inferring knowledge from the given data (Latitudes and Longitudes vs. Addresses).
- Difficulty in parallelizing the process between different stages since each stage needs the previous one to fully finish before it can start (i.e., serialized between stages).

3 System Evaluation and Data Analysis

3.1 How will you evaluate your system and architecture?

Evaluation will be done based on the following criterion.

1. Data cleaning evaluation based on data quality to ascertain whether the system can detect incorrect and wrong data by removing it in the preprocessing.

2. Testing the system's ability to deal with stragglers and machine failures which may even happen mid-processing.

3. Overall performance evaluation when running on more massive datasets (with potentially high correlation).

4. Stress testing the system by overloading with many processes to determine the capabilities of the scheduler.

3.2 What results do you plan to obtain?

- Detailed (potentially interactive) map(s) of "hotspots" across time and space for Taxis and Uber.
- A detailed map of customer service preferences.

3.3 What type of data analysis will you perform?

Preprocessing and Batch analysis to answer all of the raised questions.

3.4 How this type of analysis is adequate for the data, problem and the issues posed?

Due to the nature of the data, it should be easily scalable to other cities and even other countries. However, with traditional methods (i.e., storing all of the data on a single machine with an RDBMS), the solution will not be very feasible. Whereas with Spark, we can perform many rigorous analyses and even build prediction algorithms while maintaining scalability for more massive datasets.

3.5 What other datasets can be used?

New York zoning map can give us a clear view of the significance of each area regarding its district type (i.e., business/commercial/residential). We can further analyze the locations by finding the household income of each residential area and try to relate that to the customers using a service (Taxi or Uber). Through this, we can give more meaningful analysis about the customers to the business owners to help them understand their clients better.

3.6 What are the steps you need to take to scale your solution?

We can do either of the following and Spark will take care of the rest.

1. Adding computational power through better CPUs and GPUs or even adding new computers.
2. Adding storage through more RAM and disk or even adding new hardware.