Arcenas, Carlos
De Guzman, Carlo
Seechung, Vito
Tiongson, Zen

CS129.1–A
Big Data Project

Analysis of New York Boroughs: In which places are passengers likely to be picked up?

1. Big data problem:

   Given the data of taxi and for-hire vehicle pick ups in New York City, how do we determine the top boroughs in NYC in which passengers are picked up?
   (Results will be visualized on a map diagram)

   Solving this problem can be helpful for drivers to know where he or she is most likely to get passengers, at different times of day. Solving this problem would also be helpful for passengers, as they would know which companies are more present in their area throughout the day, increasing the likelihood that they will find transport.

   Our solution would make use of the dataset to reflect the changing patterns of traffic for every hours of the day across the different boroughs of New York.

2. Description of source and data:

   According to the GitHub repository for the Uber response to the New York Taxi and Limousine Commission, as released through a Freedom of Information Law request:

   "The Uber TLC FOIL Response contains data on over 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. Trip-level data on 10 other for-hire vehicle (FHV) companies, as well as aggregated data for 329 FHV companies, is also included. All the files are as they were received on August 3, Sept. 15 and Sept. 22, 2015."

   The data is currently available on the Github repo from fivethirtyeight (https://github.com/fivethirtyeight/uber-tlc-foil-response), all provided in CSV form. Data will be obtained by simply downloading the provided files. After which, the CSV is converted into a JSON format using mongo import. Due to the large volume of the data that is in the Github repo, what will be used in the project is only the 2014 data (uber-raw-data-apr14.csv to uber-raw-data-sep14.csv).

   Our solution would consist of a time and location keying. In these datasets, the "Lat" and "Long" keys to form the actual coordinates of the pickups. The "Date/Time" key will be used to reflect the changing patterns coordinates across time.

Alongside this dataset, we will be making use of the borough boundaries geographic dataset created by the government of the City of New York (https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm/data) to accurately locate the pick-up points to their respective boroughs. Downloading the data first starts by navigating to New York (either by Navigate Location button or manual). New York Borough boundaries must be highlighted in this. The data is then exported and downloaded in GeoJSON format.

This data is used to group the coordinates from the Uber Data according to what borough of New York they actually took place in.

3. Description of Output

The output after MapReduce shows a mapping of coordinates based on hour. A single JSON object ID is the hour and borough number (two keys). Each object will also contain count of coordinates and an array of coordinates. The hour referrers to the time a passenger is picked up by a ride-sharing vehicle, the borough number indicates one of the following: 0 - Not in New York, 1 - Manhattan, 2 - Bronx, 3 - Brooklyn, 4 - Queens, 5 - Staten.

The count refers to the total number of pickups in that borough for the specific hour. Lastly, an object contains an array of points that consists of the lat and long coordinates of each pickup instance.

Based on the hour, borough, and count, a person may know the “peak time” or when the surge for ride-sharing vehicles occur. From this data, a driver may know which

borough to go and on what specific time will he/she go there. Likewise, passengers may also strategically plan out when and where he/she will book a ride-sharing vehicle so that the likelihood of getting picked-up is better.

4. Visualization of Data