Taxi or CitiBike - Which to choose?

Abstract—**This paper reports our findings of the NYC taxis and Citibikes using the NYC Taxi and Limousine Commissions (NYCTLC) Yellow and Green cabs dataset, the Citbike dataset and big data technologies. The citibikes have a predefined stations where people have to walk to get a ride and if the stations are too far, people don't prefer using it. So our paper targets at improving the usage of the citibikes, by individually analyzing each customer of the taxi's as well as guiding users decide which mode of transport to choose. Also we find out using K-means algorithm where should citibikes add new stations to increase their usage and benefit their customers.**

Index Terms—**Big Data, Green Taxi's, Yellow Taxi's, Citibikes, K-Means**

I. Introduction

In one of the largest cities around the world, NYC, the iconic yellow cabs have been facilitating thousands of people daily to commute within the city and it's boroughs daily. Along with the yellow cabs, the green cabs and CitiBikes are also commonly used by people to navigate in the city. The main goal of this paper is to help the common man which to choose - Cab(Yellow in Midtown and Downtown Manhattan and Green in the boroughs) or CitiBike by taking into consideration one of the most important aspect of time. For the analysis, we have used the NYC Taxi and Limousine Commission[1] dataset for 2016 and CitiBike Dataset[2] for 2016. It is important to note that the pickup and drop off locations of the CitiBikes is fixed as we have stations where the cycles are docked. However, the pickup and drop off locations for taxis can be anywhere. So we have mapped the pick up and drop off location of taxis to the nearest CitiBike station and have assumed that the person will walk until that station and take the bike after that. We are also planning to implement the k-means unsupervised machine learning algorithm to group together the pick up and drop off locations of Taxis by considering the latitudes and longitudes of a place. Then we can suggest if the new Citibike station can be set up at some location, so that people will prefer CitiBike more.

II. Related Work

We have analysed many papers on Taxi Data and on CitiBike Data, but we did not find anything that establishes the link between CitiBike and Taxi and this was the main motivation behind this paper. In paper[3], the authors have analysed the green and yellow cab data between 2009 and 2015, and have derived some useful statistics from it, like the distribution of pick up and drop off time at different times of the day,etc. They have also analysed the growth of the green cabs over time and hence, alongh with yellow cabs green cabs also form the important part of our analytics.

The authors of [4] have analysed around 500,000 taxi trips each day in NYC, providing useful insights into the enormous amount of data. They have developed a SQL like Query model with the help of which they have queried the data and plotted the density of taxis at any given place and time, which is another important aspect of our paper.

In one of the other websites that we have read for this project, [5], the author has explored the CitiBike data and has described the opinon of few people who use CitiBike frequently. He has also included some of the statistics of CitiBike; taking into consideration the hourly rides, popular places, etc. This has given us the better insights into the CitiBike Data

III. Design

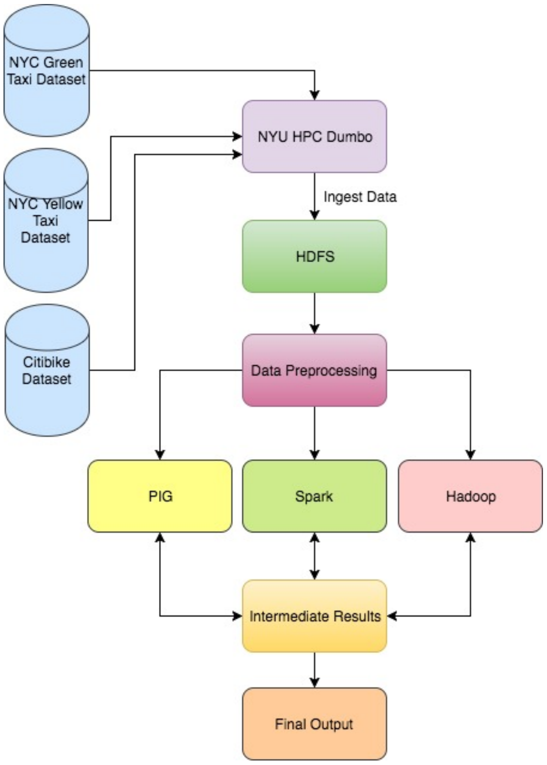

Fig. 1. Design Diagram

The fig[1] describes the workflow of our system. Initially we have taken 2 datasets, the Green, Yellow Cabs which is approximately around 30 GB combined for 2016. The other dataset is CitiBike Dataset which includes the information about CitiBike and is around 2 GB. We have loaded all this data into NYU HPC Dumbo and preprocessed the data using PIG and Hadoop MapReduce. We have also used kmeans in Spark's ML Lib library to do clustering.

References

[1] NYC Taxi and Limousine Comission. http://www.nyc.gov/html/tlc/html/about/about.shtml

[2] W. H. Cantrell, and W. A. Davis, “Amplitude modulator utilizing a high-Q class-E DC-DC converter”, *2003 IEEE MTT-S Int. Microwave Symp. Dig.*, vol. 3, pp. 1721-1724, June 2003.

[3] Sun, Huiyu, and Suzanne McIntosh. ”Big data mobile services for New York city taxi riders and drivers.” Mobile Services (MS), 2016 IEEE International Conference on. IEEE, 2016.

[4] Ferreira, Nivan, et al. ”Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips.” IEEE Transactions on Visualization and Computer Graphics 19.12 (2013): 2149-2158.

[5] What 22 Million Rides Tell Us About NYC Bike-Share. https://nextcity.org/daily/entry/citi-bike-new-york-city-bike-share-data