Study of relationship between NYC taxi trips and crime complaints data

Aishwarya Budhkar
New York University
New York, United States
asb862@nyu.edu

Ambika Singhania
New York University
New York, United States
as12578@nyu.edu

***Abstract*—**

In this paper we studied the relationship between NYC taxi trips and crime complaints for different zones and boroughs in NYC. Taxis are a proxy of a city's movement. Thus, they can give a clear picture that which areas in a city are more travelled. We used the Yellow taxi trips dataset and NYPD historic crime dataset for our analysis. The data is first cleaned using Map/Reduce. Then the cleaned data is analyzed using Hive. We studied the number of crimes versus taxi in different zones and also the relationship borough wise.

Keywords—Hadoop MapReduce, Hive, Big Data Analytics, Big Data, NYC Taxi, NYC Crime, Crime, Taxi, NYPD open data

I. Introduction

These days crimes are increasing at a high rate which propose a challenge to the law enforcement agencies. A huge amount of data about the different types of crimes is collected and stored annually. The data can be analyzed using big data technologies to find potential solution for the increasing crime rate. As taxis in New York are equipped with GPS sensors a lot of data about the taxi pickup zone, drop off zone, fare amount, etc. is stored.

Due the large volume of data, the traditional systems find it difficult to process the data. The big data framework can help to discover patterns in data efficiently with great speed. The term big data is referred to data with large volume, velocity and veracity. In big data analytics we look at big data and find patterns, incomprehensible relations and other insights which can be used to prove or disprove assumptions. It is used on a large scale due to its fast development and many frameworks are provided for big data analytics. We deal with big data resources, tools and techniques, big data analytics and its applications.

The remainder of paper is organized as follows: Section II describes why the analytic is important; Section III describes the related research work in the field; Section IV includes the design and implementation details along with the description of datasets; Section V describes the experiment details and observations from the analytics; Section VI states some general future work directions. Section VII states the conclusion derived.

II. Motivation

Safety is crucial in every city. Safety perception might influence people's behavior and their travel preferences. People might get an idea regarding the safety of their destination and can choose the travelling option according to their convenience. The taxi companies can use this data to improve their service by providing more taxis in the area of greater requirement. Thus, people can check if hiring a taxi is the general trend in that area due to any criminal activity and thus can stay safe by preferring to walk less. Taxi companies can get monetary benefits due to more usage of taxi.

Understanding crime and its pattern can also be beneficial for people and companies who want to buy a new house or want to start a new business establishment. This also has an impact on walkability score of a region, thus it can have a relationship with taxi pickups and drop-offs. This analytic might be a good stepping stone for prediction of crime counts in a given zone or borough of NY. This might also be useful for organizations like NYPD to curb the crimes in zones with prediction of high crime rate.

III. Related Work

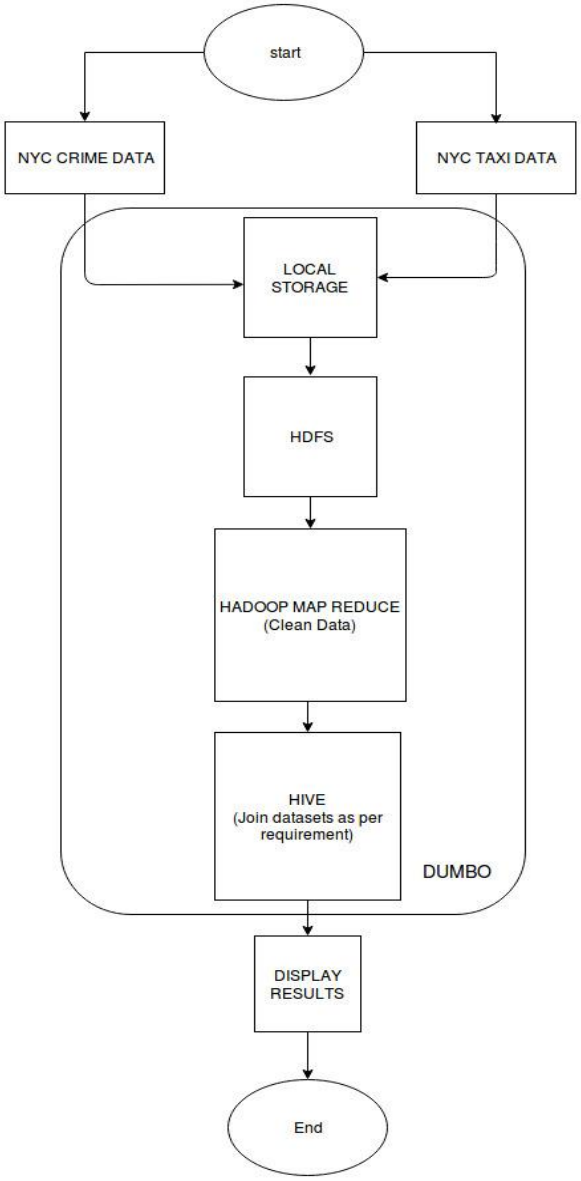
Crime data is analyzed to maintain law and order and provide safety for citizens. Big data is voluminous, complex data from different sources. Processing and storing huge amount of data is a challenge which cannot be handled using the traditional methods and tools. Big data doesn't just mean large size but also high speed with which new data is generated and huge variety of big data and number of uncertainties in big data like missing, duplicate and incomplete data and large value contained in big data. [3] Hadoop is an analytics framework engine providing scalability and faster processing of huge data. The paper proposes a framework for analyzing crime trends. Understanding what factors caused higher crime is important to make policies for better life of citizens [4]. Twitter, taxi and Foursquare data help to improve the prediction accuracy of NYC crime [6].

Transportation has been proved as the most vital service in large cities. Diverse modes of transportation are accessible. In large cities in the United States and cities around the world, taxi mode of conveyance plays a foremost role and used as the best substitute for the general public use of transportation to get their

necessities. For instance, by today in New York there are nearly 50,000 vehicles and 100,000 drivers are existing in NYC Taxi and Limousine Commission [2]. We have used this paper to understand the dynamics of taxi commute in NY. Also, this paper helped us to better understand how to use Big Data technologies to analyze Taxi data.

IV. Design and Implementation

A. The following design diagram describes the process flow. First, we collect the taxi and crime datasets and put them in HDFS. To clean the data, we use Hadoop Map/Reduce. The cleaned data is then joined using Hive and analytic is performed on Hive. We can use Map/Reduce for analytics. Finally, the results are visualized.



B. Description of Datasets

We have used two datasets: NYC Yellow Taxi Trips Data and NYPD Complaints Historic Data.

The NYPD Complaints Historic Data had many columns. It is 2 GB in size. Of those, we have used following for our analytic:

Column Name	DataType (length)	Limits
CMPLNT_FR_DT	STRING (9)	2006-2019 03/31/2019
CMPLNT_FR_TM	STRING (8)	00:00:00-23:59:59 23:50:00
CMPLNT_TO_DT	STRING (9)	2006-2019 04/01/2019
CMPLNT_TO_TM	STRING (8)	00:00:00-23:59:59 00:10:00
OFNS_DESC	STRING (Variable length)	No max limit only what compiler can store ASSAULT 3 & RELATED OFFENSES
BORO_NM	STRING (Variable Name)	No max limit only what compiler can store MANHATTAN
Latitude	FLOAT 64	Compiler specific float limit 40.85358740100002
Longitude	FLOAT 64	Compiler specific float limit 73.90059135599995

The columns used from NYC Yellow Taxi Trips Data are given below. It is 10 GB in size

Column name	Description	Data Type
VendorID	A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.	Integer

Pickup_date	Date of pickup for particular taxi trip. We filtered it for year 2018	datetime
Drop date	Date of dropping for particular taxi trip. We filtered it for year 2018.	datetime
Pickup_zone	Pickup Zone signifies the location of pickup	Float Range – (1-265)
Drop_zone	Drop Zone signifies the location of pickup	Float Range – (1-265)
borough	This is NYC Borough for the trip	String (max size – 20) Values (BROOKLYN, MANHATTAN, BRONX, QUEENS, STATEN ISLAND)

V. Results

We started with cleaning our datasets. We mainly wanted to identify the relationship between taxi pickups and crime counts in a zone. For this purpose, we removed some columns which were not of use for our current analytic like fare-type in Taxi Data. Since it was a huge dataset, we encountered lot of challenges in its cleaning phase. Certain columns had vague values like year was 1026 in some rows.

In both Crime as well as Taxi datasets, there were different number of columns in some rows. We used Hadoop MapReduce for cleaning the datasets. We started by developing the MapReduce cleaning code and cleaning a subset of data using the code, after this once we determined that code was working as expected, we ingested data in larger batches. It is then we started understanding that there were irregularities in data, when we subjected the same code to larger dataset, it failed or didn't give correct output. We then went back to improving our code and thus handle each new scenario that was encountered. It was with such multiple iterations that we could finally clean all our data.

Taxi data had pickup and drop-off zoneid in place of Latitude and Longitude for location, where as crime data had latitude and longitude. These zoneids are assigned to regions in NY and are defined by Taxi and Limousine Commission.

We then used a Taxi-zone dataset that mapped latitude and longitude to these TLC zones. We then joined this file with crime dataset to obtain zones corresponding to latitude and longitude. This was also done using MapReduce in cleaning phase of project.

After cleaning the data, we uploaded all the clean data into Hive tables. After this we performed various aggregations on our tables. We also grouped the taxi pickups, drops and crimes by zones to obtain their respective counts for each zone. We also performed sorting and various other operations on our cleaned data. We expected that zones with highest pickups would have highest crime count as well, but after all these operations, our data did not conform to this. Then we referred to some other experiments in the domain and understood that it's too ambitious to expect direct relationship in such real-life datasets, we then started finding patterns at borough level.

We observed that Taxi pickups are also a function of population density in a borough, but number of pickups were affected by other factors also and crime was clearly one of those. We could firmly say this by patterns like,

1. Bronx and Brooklyn have nearly same population density but Brooklyn has higher crime count and also higher number of pickups (17 times more pickup)
2. Staten island has least number of pickups as well as least crime count.
3. Queens has second highest number of pickups after Manhattan has less than 1/3rd population density compared to Manhattan but it's crime count is similar to Manhattan which explains why Queens has second highest number of pickups

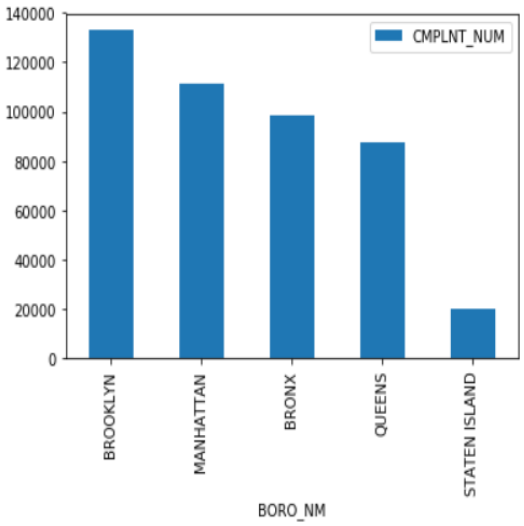
Number of Taxi pickups per borough

Borough	Number of taxi pickups
Staten Island	2290
Bronx	87495
Brooklyn	1492647
Queens	6260821
Manhattan	81471339

Population density per borough

Borough	Population Density per mile
Bronx	34,653
Bronx	37,137
Brooklyn	72,033
Queens	21,460
Manhattan	8,112

Complaint count per borough in NYC



VI. Future Work

To get a deeper understanding of the data, in future we want to do spatial analysis of zones and create some heatmaps which can highlight some features of the data that is difficult to capture otherwise. As a number of socio-economic factors affect the crime rate in a region, we would like to add datasets like US Census data, Subway usage data, Foursquare Venues data, and other Point of Interest datasets for better understanding of the crime rate. Machine Learning models like Regression and Random Forest models can be used for predicting crime in the zones or boroughs.

VII. Conclusion

We conclude that the total number of crime occurrences is positively correlated with the number of pickups by taxi. The reason is people might prefer to use a taxi instead of walking or using public transportation in regions of high crime.

To check the goodness of our analytic we compared our results with researches/ findings in the domain and our results were aligning with theirs [9] [10]. We realized that we were not able to fully capture the whole picture due to many different socio-economic factors affecting the taxi pickups in a region.

However, our analytic will provide a guideline for other researchers. Addition of different Point of Interest datasets will lead to better understanding of the relationship between crime occurrences and taxi pickups and drop-offs.

Acknowledgment

We are thankful to the NYC HPC support for quickly responding to all our questions related to Dumbo. Thanks to Cloudera for providing us with access to Hadoop Cluster. Thanks Professor Suzanne Mcintosh and our grader Shreya Pandey for guiding us, helping us with all the blockers and constantly encouraging us.

References

1. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
2. Patel, Umang. (2015). NYC Taxi Trip and Fare Data Analytics using BigData. 10.13140/RG.2.1.3511.0485.
3. Arushi Jain. Vishal Bhatnagar. Crime Data Analysis Using Pig with Hadoop. Procedia Computer Science. Volume 78. 2016. Pages 571-578.
4. Hongjian Wang. Daniel Kifer. Corina Graif. Zhenhui Li. Crime Rate Inference with Big Data. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 13-17. 2016. San Francisco. California. USA.
5. Sathyadevan, Shiju. S Devan. S Gangadharan Surya. (2014). Crime Analysis and Prediction Using Data Mining. 10.1109/CNSC.2014.6906719.
6. Vomfell Lara. Härdle Wolfgang Karl. Lessmann Stefan. (2018). Improving Crime Count Forecasts Using Twitter and Taxi Data. Decision Support Systems. 113. 10.1016/j.dss.2018.07.003.
7. A.M.S. Osman. A novel big data analytics framework for smart cities. Future Generation Computer Systems (2018).
8. Deri, Joya & Moura, Jose. (2015). Taxi data in New York city: A network perspective. 1829-1833. 10.1109/ACSSC.2015.7421468.
9. Li, Kebing, "Investigating the effect crime has on Uber and Yellow Taxi pickups in NYC" (2019). Honors Theses.Paper 924.
10. Kadar, C. & Pletikosa, I. Mining large-scale human mobility data for longterm crime prediction. EPJ Data Science 7, 26 (2018).