NYC Taxi Trip and Fare Data Analytics using BigData

Umang Patel [#1], Anil Chandan [#2]

[#] *Department of Computer Science and Engineering*

University of Bridgeport, USA

[1] umapatel@my.bridgeport.edu

[2] achindam@my.bridgeport.edu

***Abstract*—As there is an amassed evolution in the metropolitan zones, urban data are apprehended and have become certainly manageable for first-hand prospects for data – driven analysis which can be recycled for an improvement of folks who lives in urban zone. This particular project highlights, the prevailing focus on the dataset of NYC taxi trips and fare. Traditionally the data captured from the NYC Taxi & Limousine commission was physically analysed by various analyst to find the superlative practice to follow and derives the output from it which would eventually aids the people who commute via taxis. Later during early 2000 the taxi services where exponentially developed and the data capture by NYC was in GB's, which was very difficult to analyse manually. To overcome these hitches BigData was under the limelight to analyse such a colossal dataset. There were around 180 million taxi ride in city of New York in 2014. BigData can effortlessly analyse the thousands of GB within a fractions seconds and expedite the process. This data can be analysed for several purposes like avoiding traffics, lower rate where services are not functioning more frequency than a cab on crown location and many more. This information can be used by numerous authorities and industries for their own purpose. Government official can use this data to deliver supplementary public transport service. The company like Uber can use this data for their own taxi service.**

I. INTRODUCTION

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 [1] 50,000 vehicles and 100,000 drivers are existing in NYC Taxi and Limousine Commission.

There are many misperceptions in [1] TLC (Taxi and Limousine Commission) of the New York city, that how the taxi services should be disseminated in the city that too based on certain assumptions, like most pickups, time, distance, airport timings. In order to provide very good taxi service and plan for effective integration in city transportation system, it's very important to analyse the demand for taxi transportation. The dataset provides the relating information such as where taxis are used, when taxis are used and factors which tend public to use taxi as divergent to other modes of transportation.

In present day transportation dataset contains large quantity of information than the preceding data. For specimen, from the year 2010 TLC using the Global Positioning System(GPS) data type for every taxi trip, including the time and location (latitude and longitude) of the pickup and drop-off. In this a complete traffic data which contains nearly 180 million rows of data in the year 2014. Due to huge amount of data, this data is example of "BigData." Using BigData it's easy to develop procedures to clean and process the data so it used for analyse the raw data into useful way in transportation service.

The core objective of this is to analyse the factors for demand for taxis, to find the most pickups, drop-offs of public based on their location, time of most traffic and how to overcome the needs of the public. Explicitly, the key contributions of this paper are as follows:

- Primitively to the best of our knowledge, we conduct the analysis that recommends the top driver based on the most distance travelled, most fare collected, most time travelled and most efficient driver. It will help commission to award such drivers and encourage such drivers to get most out of them. Analysis was done with the help of MapReduce programming.
- Furthermore, to achieve our goal, we proposed our subsequent analysis i.e. Analysis on Region. Here we accumulate information which was allied with location like PickUp Latitude, PickUp Longitude, DropOff Latitude and DropOff Longitude. Expending PickUp Latitude and PickUp Longitude we appraise the most PickUp locations and same for the DropOff locations. This will help to provide more taxis on the most PickUp location and so on. Analysis was thru with the help of MapReduce programming
- To quantify the Total Pick-ups and Drop-offs by Time of Day based on location we used Hive to analyse it. The third analysis was based on the total PickUp and DropOff for a day per hour and location. For executing such intricate query, we used Hive which is a part of Hadoop ecosystem. The final output consists of the total PickUp or DropOff count for every hour of a day based on location.
- Ultimately, we analysed the fare to get the Drivers revenue. This analysis consists of both gross and net revenue to get the Driver Fare Revenue (Gross and Net). This analysis was prepared with the help of another Hadoop ecosystem technology called as Pig. The PigLatin language is used to write the query.

Our evaluation effort is encyclopaedic. We test our all analysis on a real world dataset consisting of GPS records from more than 14, 000 [2] taxicabs in a big metropolitan space with a population of more than 10 million. The rest of the paper is organized as follows. Section II introduces the related work. Section III proposes our main problem statements. Section IV depicts our Performance Evaluation. Section, followed by the conclusion in Section V

II. Related Works

Now we are going to fetch the section focusing Technologies which are used in analysing huge dataset. The complete analysis is analyzed using BigData with Hadoop and Hadoop Ecosystem. Following are the brief definition of the BigData, Hadoop, MapReduce and Hive and Pig a part of BigData Hadoop Ecosystem

A. BigData

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data [8].

B. Hadoop

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework [9].

C. MapReduce

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations [10].

D. Hive

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. While initially developed by Facebook, Apache Hive is now used and developed by other companies such as Netflix. Amazon maintains a software fork of Apache Hive that is included in Amazon Elastic MapReduce on Amazon Web Services [11].

E. Pig

Pig is a high-level platform for creating MapReduce programs used with Hadoop. The language for this platform is called Pig Latin. Pig Latin abstracts the programming from the Java MapReduce idiom into a notation which makes MapReduce programming high level, similar to that of SQL for RDBMS systems. Pig Latin can be extended using UDF (User Defined Functions) which the user can write in Java, Python, JavaScript, Ruby or Groovy and then call directly from the language. Pig was originally developed at Yahoo Research around 2006 for researchers to have an ad-hoc way of creating and executing map-reduce jobs on very large data sets. In 2007, it was moved into the Apache Software Foundation [12].

Also, this section focus on earlier big data projects on NYC taxi dataset, namely to optimize taxi usage, and on big data infrastructures and applications for transport data events.

- Transdec (Demiryurek et al. 2010) is a project of the University of California to create a big data infrastructure adapted to transport. It's built on three tiers comparable to the MVC (Model, View, Controller) model for transport data [5]
- (Jagadish et al. 2014) propose a big data infrastructure based on five steps: data acquisition, data cleaning and information extraction, data integration and aggregation, big data analysis and data interpretation [7].
- (Yuan et al. 2013), (Ge et al. 2010), (Lee et al. 2004) worked a transport project to help taxi companies optimize their taxi usage. They work on optimising the odds of a client needing a taxi to meet an empty taxi, optimizing travel time from taxi to clients, based on historical data collected from running taxis [6].

III. Problem Definition

During this course of study about NYC taxi data sets consequence are entirely centred on certain analysis. These problem statements are elucidated clearly one after the other in the sequence

A. Problem Definition I – Analysis on Individual

The problem definition I consists of analysis on individual driver. This analysis will be done using MapReduce programming. Following are the analysis which will be done using MapReduce programming:

1) Driver with most distance travelled: This analysis will be on individual basis where we will analyze and determine the driver travelled with most distance in miles.

MapReduce Function:

Mapper

map(Hack_license, Trip_distance) =
Emit(inter_Hack_license, inter_Trip_distance)

Reducer:

reduce(inter_ Hack_license, inter_Trip_distance) =
Emit(Hack_license, Max_Trip_distance).

2) Driver with most fare collected: This analysis will be on individual basis where we will analyze and determine the drive collected most fare in dollars including fare amount, surcharges, mta tax, and tip amount i.e. total amount

MapReduce Function:

Mapper:

map(Hack_license, Total_amount) = Emit(inter_Hack_license, inter_Total_amount)

Reducer:

reduce(inter_ Hack_license, inter_Total_amount) = Emit(Hack_license, Max_Total_amount).

3) Driver with most time travelled: This analysis will be on individual basis where we will analyze and determine the driver spend most travel time in seconds.

MapReduce Function:

Mapper:

map(Hack_license, Trip_time) = Emit(inter_Hack_license, inter_Trip_time)

Reducer:

reduce(inter_ Hack_license, inter_Trip_time) = Emit(Hack_license, Max_Trip_time).

4) Driver with most efficiency based on distance and time: This analysis will be on individual basis where we will analyze and determine the most efficient driver. It can be determine with distance / time with criteria of minimum distance travelled.

MapReduce Function:

Mapper:

map(Hack_license, Trip_distance / Trip_time) = Emit(inter_Hack_license, inter_Trip_distance / inter_ Trip_time)

Reducer:

reduce(inter_ Hack_license, inter_Trip_distance / inter_ Trip_time) = Emit(Hack_license, Min_effeciency)

B. Problem Definition II – Analysis on Region

The problem definition II consists of analysis on a region. This analysis will be done using MapReduce programming. Following are the analysis which will be done using MapReduce programming:

1) Most pick up location: This analysis will be based on region i.e. the latitude and longitude of the pickup. So pickup location is combination of Pickup_latitude and Pickup_longitude.

MapReduce Function:

Mapper:

map(Pickup_location) = Emit(inter_Pickup_location,1)

Reducer:

reduce(inter_Pickup_location, 1) = Emit(Pickup_location, Sum).

2) Most drop off location: This analysis will be based on region i.e. the latitude and longitude of the dropoff. So dropoff location is combination of dropoff_latitude and dropoff_longitude.

MapReduce Function:

Mapper:

map(Dropoff_location) = Emit(inter_ Dropoff_location,1)

Reducer:

reduce(inter_ Dropoff_location, 1) = Emit(Dropoff_location, Sum).

C. Problem Definition III – Analysis based on time and location

This section will through some light on how we determine some complex analysis by using Hive. The hive is a part of Hadoop ecosystem which is a software that facilitates querying and managing large dataset using simple SQL like commands. It is built on top of the Hadoop. In this analysis we will determine average of total pickup and drop-offs by time in a day based on location. Fig 1 shows the Average Total Pick-ups and Drop-offs by Time of Day based on location.

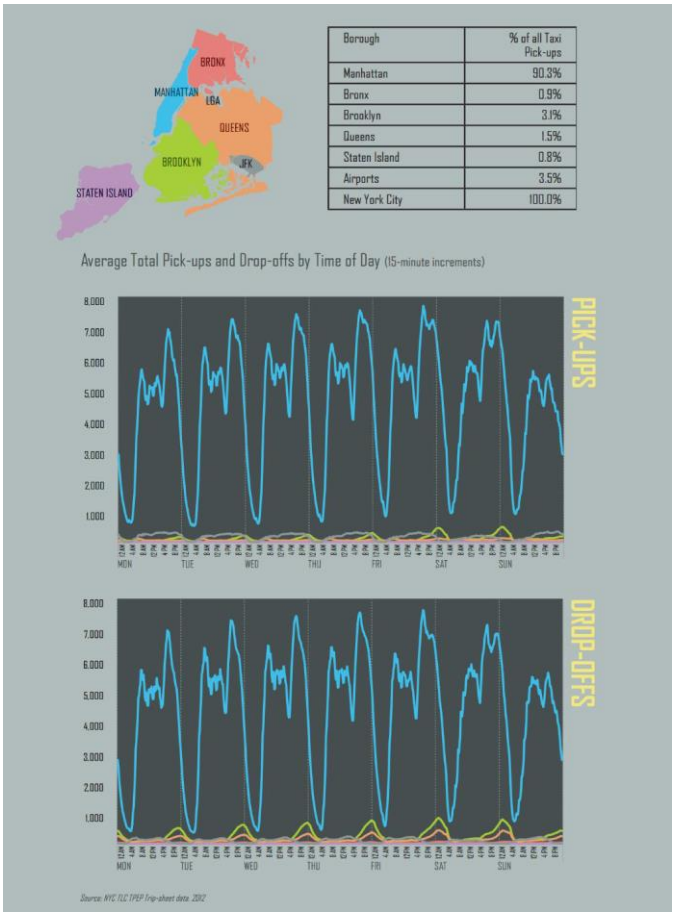

Fig. 1. Average Total Pick-ups and Drop-offs by Time of Day based on location [2]

Fig. 2 and Fig. 3 shows an example of actual outputs generated from the output of the Hive query

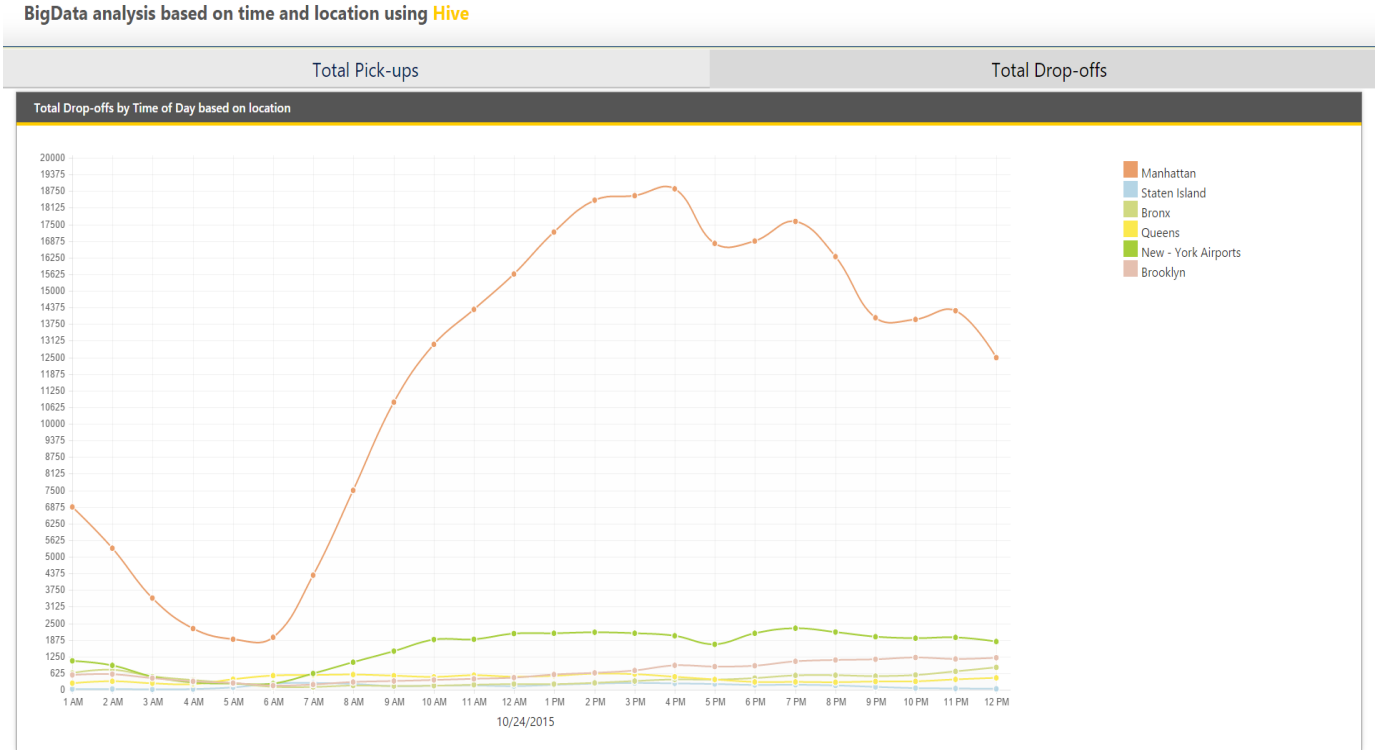

Fig. 2. Average Total Drop-offs by Time of Day based on location

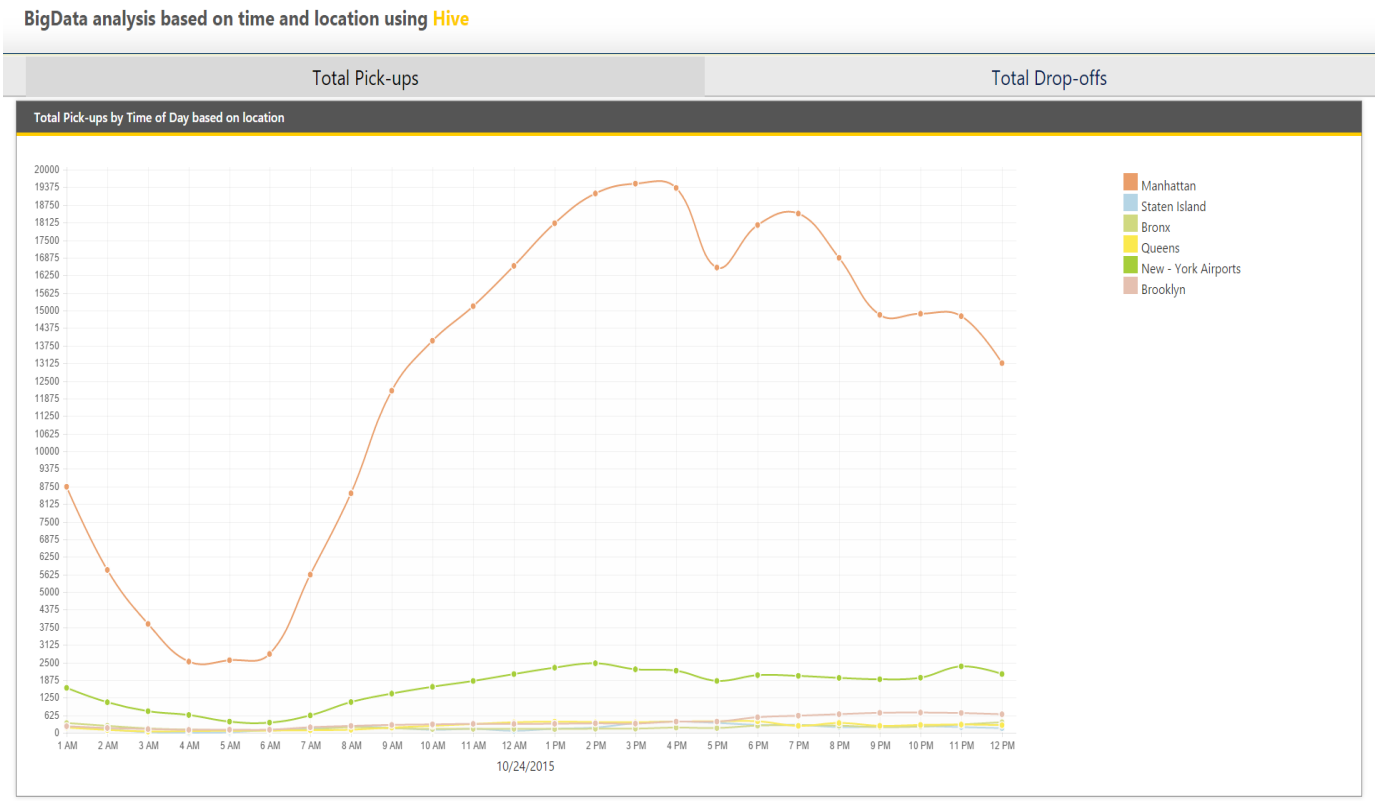

Fig. 3. Average Total Pick-ups by Time of Day based on location

D. Problem Definition III – Analysis based on Fare

In this we will determine some complex analysis using Pig. The pig is a part of Hadoop ecosystem which is a platform that facilitates MapReduce program used with hadoop and the language used by this platform is called as Pig Lation. It is built on top of the Hadoop. Basically Pig Latin abstracts the programming from Java MapReduce expression into a notation which makes MapReduce programming high level, similar to that of SQL in RDBMS. The 10 lines of pig is similar to 200 lines of java code.

In this analysis we will determine the average revenue per hour which includes both Gross revenue and Net revenue. Fig.4 shows Average Driver Fare Revenue per hour (Gross and Net).

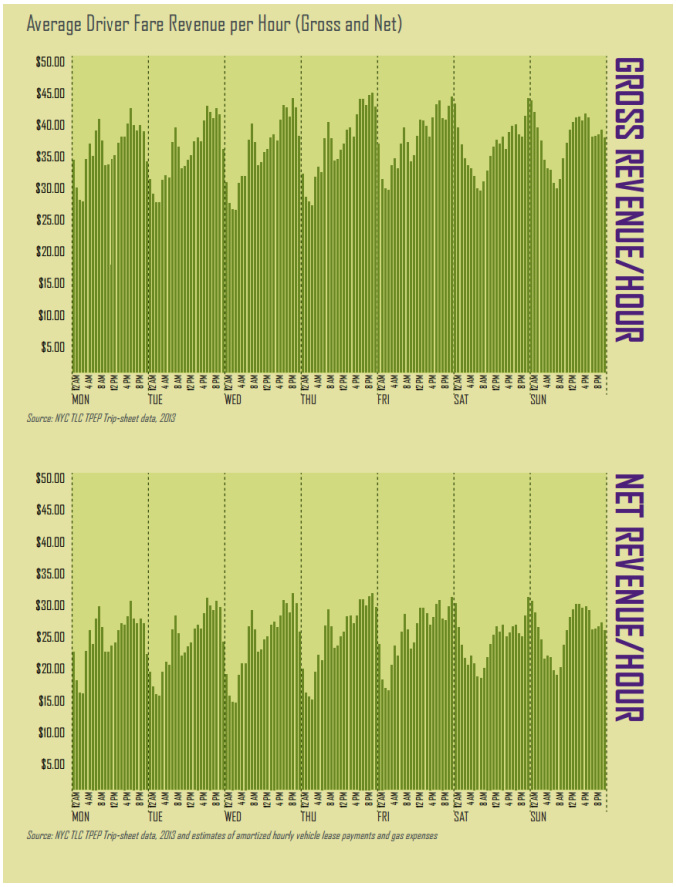

Fig. 4. Average Driver Fare Revenue per hour (Gross and Net) [2]

IV. Performance Evaluation

In this section, we are evaluated performance between MapReduce and Hive of all the Problem definition of defined in Problem Definition 1 and Problem Definition 2. The result was as expected MapReduce runs faster than hive. But when it comes to complex query like in Problem Definition 3 and Problem Definition 4 than hive is preferred as it is easy to write and understand.

TABLE I
Performance Evaluation Using MapReduce and Hive

Problem Definition	**Time In Seconds**	
	Hive	**MapReduce**
Problem Definition 1.1	20	15
Problem Definition 1.2	15	12
Problem Definition 1.3	22	19
Problem Definition 1.4	45	35
Problem Definition 2.1	10	8
Problem Definition 2.2	9	8

V. Conclusion

In this assignment, we generated new system that ropes in the visual exploration of big origin-destination and spatio-temporal data. The most vital section of this project is a visual query model that sanctions the users to quickly select data slices and use them. This project will be helpful for many industries in the near future because of its virtuous balance between simplicity and expressiveness. By associating this data with other sources of information about neighborhood needs, employment, and model to explain the chronological difference of travel demand for taxis. The Fig. 5 shows the location with most pickup and dropoff.

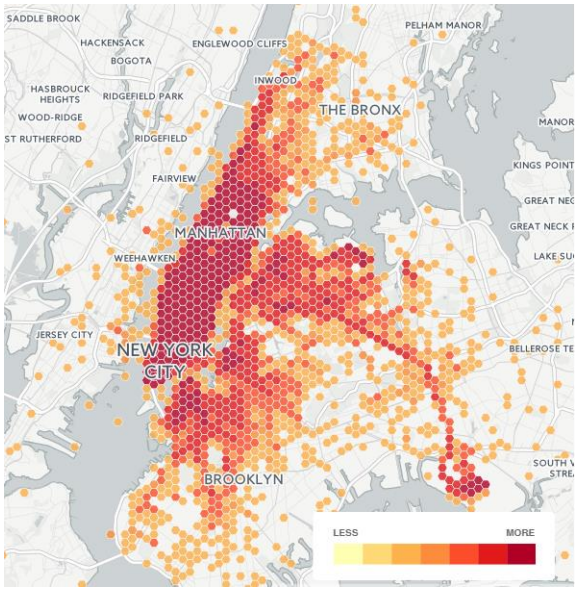

Fig. 4. Location with most frequent locations in NYC

The first analysis i.e. Analysis on Individual can helpful for determining the ability of the individual driver. We can determine the ability like efficient, quick, accuracy etc. which can be helpful in evaluating the individual and reward the individual who is doing great work or train the individual who is struggling to do the good work. In second analysis we determine which region has highest pickup and drop-off

location, it helps vendor to provide more taxis where there is more pickup and lessen the number of taxi where there is more drop-off. In this way system will work more efficiently. In the third analysis, we determine the Average Total Pick-ups and Drop-offs by Time of Day based on location. Using this analysis, we will determine the number of pickup and drop –off location in a particular region for a particular time. This will be helpful for providing more number of taxis in a region. In the last analysis Average Driver Fare Revenue per hour (Gross and Net) will help us in deriving the business is in profit or loss. Currently only few queries can be analyze but in future this data can be more analyze to get more benefits from this data. We can also analyze on trips between Penn Station and the three nearest airports in the NYC region to show how mode choice is affected by the size of the traveling group, the travelers' valuation of time, and the time of day. Ultimately, we conclude the usefulness of these project of trip provides planners, engineers, and decision makers with information about how people use the transportation system. In this case, by identifying the factors that drive taxi demand, forecasts can be made about how this demand can be expected to grow and change as neighborhoods evolve. As decisions are made regarding the regulation of the taxi industry, the provision of transit service, and urban development, these models are useful for forming a complete and holistic vision of how travel patterns and use of modes can be expected to respond.

ACKNOWLEDGMENT

The special vote of Thanks to the Taxi & Commission of New York City for offering the data capitalised in this particular paper. Our second most Important Thanks to Prof. Jeongkyu Lee and the Department of Computer Science and Engineering for their incessant support.

REFERENCES

[1] NYC Taxi & Limousine Comission. http://www.nyc.gov/html/tlc/html/about/about.shtml.
[2] Taxicab fact book http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf
[3] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica. Blinkdb: Queries with bounded errors and bounded response times on very large data. In Proc. EuroSys, pages 29–42, 2013.
[4] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, pages 161–170. IEEE, 2011.
[5] Demiryurek, U., Banaei-Kashani, F. & Shahabi, C., 2010. TransDec:A Spatiotemporal Query Processing Framework for Transportation Systems. IEEE, pp.1197–1200.
[6] Ge, Y. et al., 2010. An energy-efficient mobile recommender system. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10. N
[7] Jagadish, H.V. et al., 2014. Big Data and Its Technical Challenges
[8] "BigData," https://en.wikipedia.org/wiki/Big_data [Accessed November 28, 2015]
[9] "Hadoop," https://en.wikipedia.org/wiki/Apache_Hadoop [Accessed November 28, 2015]
[10] "MapReduce," https://en.wikipedia.org/wiki/MapReduce [Accessed November 28, 2015]
[11] "Hive," https://en.wikipedia.org/wiki/Apache_Hive [Accessed November 28, 2015]
[12] "Pig," https://en.wikipedia.org/wiki/Pig_(programming_tool) [Accessed November 28, 2015]