Spatio- temporal Analytics: Correlation between taxi trips and events in NYC

Anirudhan Rajagopalan
Computer Science,
New York University
ajr619@nyu.edu

Narasimman Sairam
Computer Science,
New York University
ns3184@nyu.edu

Shrenik Shah
Information Systems,
New York University
sbs539@nyu.edu

Abstract-

In modern cities like New York City, more and more vehicles, such as taxis, have been equipped with GPS devices for localization and navigation. Gathering and analyzing these large-scale real-world digital traces have provided us an unprecedented opportunity to understand the city dynamics and reveal the hidden social and economic patterns. The objective of this paper is to analyze temporal and location based characteristics of taxi trips along with the events and event performers data we can analyze the most happening place in New York City on a given date. In this paper, we aim to discover correlation between taxi trips and NYC events from a large-scale yellow taxi dataset and NYC eventful data. By the results of this study we can predicting the most happening place in New York City for any given date in the past.

Keywords—analytics, map reduce, social networks, check-ins, big data, Location Based Social Networks

I.INTRODUCTION

In this paper, we will be finding the most popular place in NYC i.e. place with most number of people in a time range. As a New Yorker the naive answer would be some pub in the middle of Manhattan or some live event that took place on that day. However there are many pubs and lot of events that are happening every day in New York, we will look to find places that attracts unusually large population from New York and compare empirically, given a date and duration of time. In order to answer this analytically, we will be using historic events data along with traffic data from yellow cabs.

Driven by travel demand, the distribution and density of taxi passenger pick-up and drop-off points reflect the attractiveness of an area and thus, can be used to find out hot spots and the movement of human flow, to form an analytic for understanding the most happening place. Retrieving data from the traditional public transportation (e.g. bus, train, metro) can provide a relevant database of samples and general passengers' movement. However, does not always provide the exact origin and destination for each passenger, since these transportation modes rely on pre-designated stops and paths, and usually the ticket validation is only performed on the pick-up. The taxi service can be a way to retrieve large dataset of information with a higher precision when we focus the origin and destination of each trip. It can pick-up the passengers right where they are standing, and then drop-off them precisely in the desirable destination, without being bounded to a pre-determined path.

In this study, we are clustering the events from eventful dataset using the latitudes, longitudes of events into different neighborhoods. Then, we segregate taxi trip records into groups of 30 minutes interval each along the date range of the data set using spark. These minor groups of taxi records are partitioned by neighborhoods by means of latitude and longitude of taxi trips using Euclidian distance since the area under consideration is relatively flat. After that, the outliers [peaks] of the groups are detected using moving averages. Finally, we determine the correlation between the peaks and an occurrence of an event.

II.MOTIVATION

As far as a using this data for a business venture, any hotel chain owner or event planning agency would find this information valuable. There are a plethora of vacation planning sites looking to set themselves apart from their competition. This data could be neatly packaged into a small, independent app users can view before going for entertainment activities.

Understanding the distribution of hot spots, and people's interests to these areas are valuable for location-based service (LBS), Location-based Social Network (LBSN), transport management and urban planning, etc.

III.RELATED WORK

The data we are using for this paper is static and historic in nature with periodic refreshes. Keeping in mind the possibilities of ever increasing volume of data, we decided to future-proof our implementation by preparing for big data volume right from the offset.

The data we are using for this paper is static and historic in nature with periodic refreshes. Keeping in mind the possibilities of ever increasing volume of data, we decided to future-proof our implementation by preparing for big data volume right from the offset. In the paper “Urban mobility and taxi flow”, Marco Veloso and Santi Phithakkitnukoon perform analysis of taxi flow can help better understand the urban mobility. In this work, they analyze 177,169 taxi trips collected in Lisbon, Portugal, to explore the relationships between pick-up and drop-off locations; the behavior between the previous drop-off to the following pick-up; and the impact of area type in taxi services. Their predictability analysis shows that individual taxi trips are relatively random. With Bayesian approach given time of the day, day of the week, weather condition, area type, and the current pick-up location, only 5% of all trips are predictable. R.W. Scholaz and Y.Lu in their Detection of dynamic activity patterns at a collective level from large-volume trajectory data paper defined the typology for the different dynamic patterns of activity hot spots and proposed a method to reveal and predict the space-time dynamic patterns of activity hot spots. GPS trajectories of 536 taxi cabs over 22 days in San Francisco were used for empirical analyses. The life cycle of a selected activity hot spot was described in detail. Poisson distribution was used to estimate the probability of a certain number of activity instances occurring at a certain tract during a certain hour.

IV. DESIGN

A] System Architecture

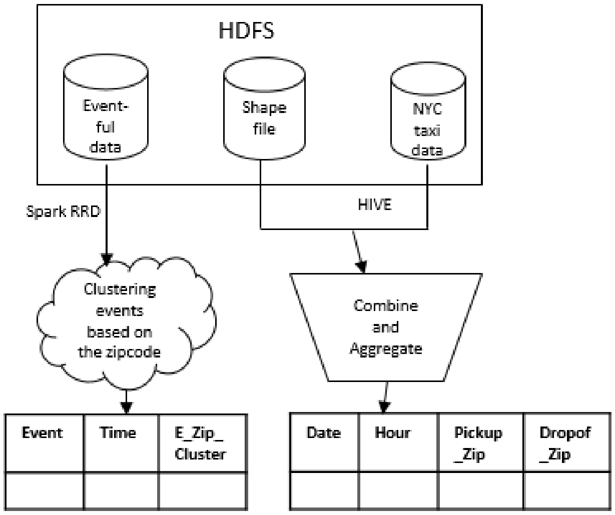

Fig.1a: Phase 1-Data Cleansing and Data Integration

HDFS: It is a fault tolerant filesystem designed for storing very large files with streaming data access patterns, running on clusters of commodity hardware.

Hive: It is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. It provides an SQL-like language called HiveQL with schema on read and transparently converts queries to map/reduce, Apache Tez and Spark jobs.

Apache Spark : It is an extremely fast general purpose engine for large-scale data processing cluster computing system. It provides high-level APIs for writing applications quickly in Java, Scala, Python and R, and supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing. Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3.

Eventful: Eventful dataset contains information regarding entertainment events for concerts, sports, family fun and nightlife that are taking place in local markets throughout the world. Using Eventful API’s we can leverage features and functionalities of Eventful’s data to filter data by geographic area, category, or other event parameters.

NYC Taxi Dataset: This dataset includes trip records from all trips completed in yellow taxis in NYC in 2014. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations in latitude-longitude, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided by NYC open data under Taxicab Passenger Enhancement Program (TPEP).

Zipcode is a group of five or nine numbers that are added to a postal address for classifying a neighborhood which was primarily used to assist in the sorting of mail.

Shape File: It contains non topological geometrical information regarding attributes of geographic features of a location. The geometry for a feature is stored as a shape comprising a set of vector coordinates in a geospatial vector storage format known as ESRI vector. It is commonly used by GIS software to latitude, longitude to identify a neighborhood by means of ESRI API’s.

Moving Average: It is a calculation to analyze data points by creating series of averages of different subsets of the full data set. It is called 'moving' because it is continually recomputed as new data becomes available, it progresses by dropping the earliest value and adding the latest value. Moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles.

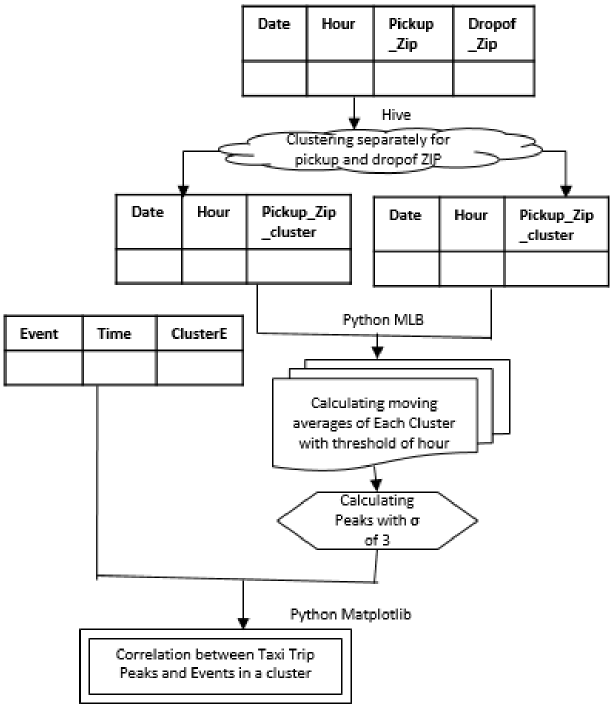

Fig. 1b: Data transformation and Correlation

B] Implementation

Parse Events dataset that consists of around 2,00,000 event details of the past three years.

One of the challenges in solving this problem is to identify the clustering mechanism. It is not straight forward to cluster the events as we have latitudes and longitudes.

So, to cluster the events based on the zip code, we get it from the NYC shape file.

A shape file is an Esri vector data storage format for storing the location, shape, and attributes of geographic features. It is stored as a set of related files and contains one feature class. We use the shape file and determine the zip code of a given (latitude, longitude).

Once we identified the zip code, we can now cluster the events based on the neighborhood. So, we get clusters of events (event, date, time, cluster)

Parse NYC Yellow taxi dataset for the past three years. As in events dataset, this also has latitude and longitude of the pickup and drop off locations of the trip. So, we parse the dataset and identify

Aggregate the number of taxi trips taken to all neighborhoods for every hour of a day.

For each neighborhood, calculate the moving average of the aggregate count of trips for every hour of a day.

Find outliers with a standard deviation of 3. We get the peaks which means the set (date, hour, cluster) will give us the cluster where there were more number of taxi trips than usual.

Now, we find correlation between the taxi trip peaks and the list of events happening on the same hour of the day grouped under the same cluster.

If there is a strong positive correlation, then we can mark that event as one of the popular ones.

Taxi trips are just one among the many different factors that we can take into consideration before predicting the most popular event in the city.

The more factors we consider, the more accurate our prediction will be.

When calculating successive values, a new value comes into the sum and an old value drops out, meaning a full summation each time is unnecessary for this simple case,

$$SMA_{\text{today}} = SMA_{\text{yesterday}} + \frac{p_M}{n} - \frac{p_{M-n}}{n}$$

V.RESULTS

(Future… In this section, you can describe: Your experimental setup/issues with data/ performance/ etc. Describe your experiments, describe what you learned. Did you prove or disprove your hypothesis? Were some results unexpected? Why?)

VI.FUTURE WORK

We have taken NYC Taxi dataset as one of the factors to determine the most happening place of the city. But, the correlation between the events and the taxi trips is not sufficient to solve this problem. There are many other factors that will increase the accuracy of the prediction model.

We should also include social media analytics, popularity of the performers of the event, prior popularity of the event, real time check-in data and so on. We should weigh these factors based on their importance to the model.

VII.CONCLUSION

This analytics will help in accurately identifying popular places which can help people to figure out where to go and out on a particular day. This data can also help location based services to optimize their profits and transport management and urban planning, etc.

ACKNOWLEDGEMENT

We would like to Thank Professor McIntosh for helping out with paper research. We would also like to thank publications like IEEE, Elsevier and NYU for providing free access to these useful resources.

We would also like to thank NYU HPC services team without which we could not have ran the project efficiently and economically.

REFERENCES

[1] *Temporal Data Classification and Rule Extraction using a Probabilistic Decision Tree* by Mojtaba Malek Akhlagh, Shing Chiang Tan, Faiiaz Khak,

[2] Advanced Analytics with SparkO'Reilly Media Inc.,Sebastopol, CA, April 2015.

[3] *Finding Highly Correlated Pairs Efficiently with Powerful Pruning* by Jian Zhang and Joan Feigenbaum

[4*] Spatial Framework for Hadoop from ESRI-http://esri.github.io/gis-tools-for-hadoop/*

[5] Investigating Socio-cultural Behavior of Users Reflected in Different Social Channels on K-pop by Yonghwan Kim, Dahee Lee, Jung Eun Hahm, Namgi Han, Min Song.

[6] Detection of dynamic activity patterns at a collective level from large-volume trajectory data by R.W. Scholaz and Y.Lu .

[7] Urban mobility and taxi flow", Marco Veloso and Santi Phithakkitnukoon.