Analyzing the correlation between taxi trips and events

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 this paper, we perform an exploratory analysis to visualize the relationships between yellow taxi trips and the events that happen for a given neighborhood. 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. Overall, our work shows how the combination of open datasets and data generated by mobile applications can allow researchers and practitioners alike to understand complex phenomena in the urban domain.**

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

I.INTRODUCTION

In this paper, we will be finding the correlation between taxi trips and events in the New York City. In order to answer this analytically, we have the data of all the events that happened in the past four years and the taxi trips data from yellow cabs.

Driven by travel demand, the distribution and density of taxi passenger’s 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, it 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.

II.MOTIVATION

The problem of finding the most happening place in a city given a date and time is a difficult and interesting one. The questions like ‘Where can I go to have a good time right now?’, ‘Which will be the most happening place in the city?’ arise often to people who are planning to spend some time out there while not having a destination in mind. The answers to these questions open up many avenues for a variety of people. Understanding the distribution of hot spots, and people’s interests to these areas are valuable for tourists who wants to explore the city, business owners looking to open up a new outlet in a place where more number of people visit, location-based service (LBS), Location-based Social Network (LBSN), transport management and urban planning.

As far as using this data for a business venture, any hotel chain owner or event planning agency would find this information valuable. There is 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 that users can view before going for entertainment activities.

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.

In the paper, Urban mobility and taxi flow[7] the authors perform analysis of taxi flow that helps us to better understand the urban mobility. They analyze 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.

In the paper Detection of dynamic activity patterns at a collective level from large-volume trajectory [6], they 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. SYSTEM DESIGN

[A] System Architecture

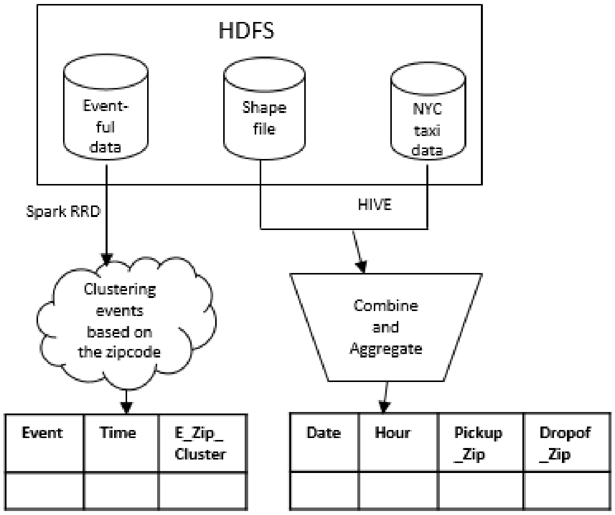


Figure 1: Phase 1-Data Cleansing and Data Integration

Events dataset: Eventful is the world's largest collection of events, taking place in local markets throughout the world, from concerts and sports to singles events and political rallies. Using Eventful API's, we can leverage the features and functionalities of events data to filter by geographic area, category, or other event parameters.

The New York City Taxi Dataset: The Freedom of Information Law in the United States encourages public authorities to release their data where appropriate to the benefit of the citizens. The dataset describes taxi journeys in New York City during the full course of 2013, and informs us on the origin and destination points of taxi trips in terms of geographic latitude and longitude coordinates. This mobility dataset counts 30GB of mobility data representing almost 170 million trips.

Big Data tools:

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.

Apache Spark: It is an extremely fast general purpose engine for large-scale data processing cluster computing system. It supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing.

Moving Average: It is a calculation to analyze data points by creating series of averages of different subsets of the full data set. 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,

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

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.

[B] Implementation

1. We begin with processing of the events dataset that consists of almost 200K event details of the past three years.
2. We cluster the events based on the zip code, calculated using a shape file.
3. 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 to determine the zip code of a given geo-coordinate (latitude, longitude). Now, the clusters of events (event, date, time, cluster) has been calculated.
4. Next, we process New York Yellow taxi dataset that counts 30GB describing the taxi trips taken in the past one year.
5. Similar to the events data, taxi trips also has geo-coordinates (latitude and longitude) of the pickup and drop off locations of every trip.

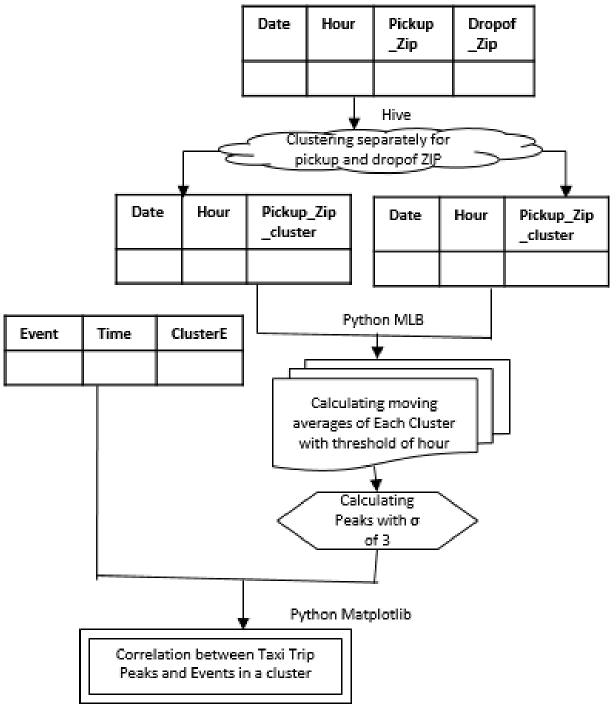

Date
Hour
Pickup_Zip
Dropof_Zip
Hive
Clustering separately for pickup and dropof ZIP
Date
Hour
Pickup_Zip_cluster
Date
Hour
Pickup_Zip_cluster
Event
Time
ClusterE
Python MLB
Calculating moving averages of Each Cluster with threshold of hour
Calculating Peaks with σ of 3
Python Matplotlib
Correlation between Taxi Trip Peaks and Events in a cluster

Fig. 1b: Data transformation and Correlation

6. We determine the zipcode of every trip and aggregate the number of trips taken for every hour of all days of every zipcode. So, we get a tuple of format (date, hour, zipcode, count).
7. For each neighborhood, calculate the moving average of the aggregated trip count for every hour of a day. A moving average is a calculation to analyze data points by creating series of averages of different subsets of the full data set.
8. Once we have the moving average and the aggregates of all the data points, we find the outliers (called as peaks) with a standard deviation of 3. We get the peaks which are tuples of format (date, hour, cluster).
9. On a graph with time (2013-14) along X-axis and aggregate counts along Y-axis, we plot the data points of moving average, aggregates of the trips and the events happened for a peak hour of the busiest neighborhood in New York.
10. On observing the above graph, we can identify the data points that confirm the positive correlation between the trips and the events.
11. Taxi trips are just one among the many different factors that we can take into consideration to predict the most happening place in the city.
12. Adding more factors and applying weights to these factors will provide us the results with much higher level of accuracy.

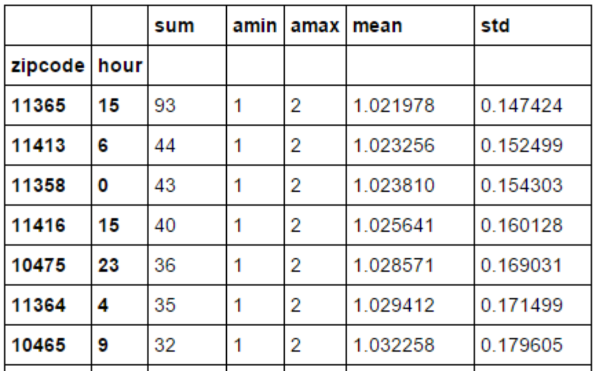
V.RESULTS

Before we observe the results, here are the experimental issues faced during the course of this project.

- There are not many public datasets available to get a list of all events that happened in the past. This is one of the important dataset necessary to solve the problem. We found a resource (www.api.eventful.com) that maintains the list of all events. Although the dataset is not very accurate, this was one single resource we had with respect to events.
- A major issue was grouping latitude-longitude of taxi trips into zip code in order to identify neighborhood of trip and make the data usable.
- Difficulty in grouping clusters by density of events as distribution of events across the city was not uniform, i.e. at some places the density of events was extremely large while at most of other places the density of events was very sparse.
- Difficulty in parsing huge taxi dataset in Spark because Spark RDD's have in memory computation.

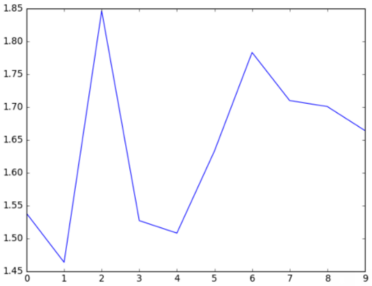
Following are the set of observations made:

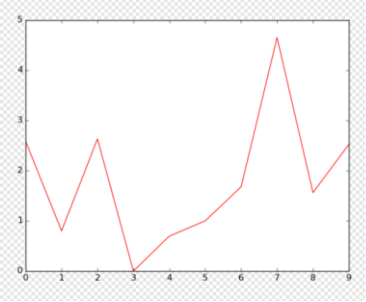
1. *Table containing the aggregated values for taxi trips for the tuple(zipcode, hour)*



		sum	amin	amax	mean	std
zipcode	hour					
11365	15	93	1	2	1.021978	0.147424
11413	6	44	1	2	1.023256	0.152499
11358	0	43	1	2	1.023810	0.154303
11416	15	40	1	2	1.025641	0.160128
10475	23	36	1	2	1.028571	0.169031
11364	4	35	1	2	1.029412	0.171499
10465	9	32	1	2	1.032258	0.179605

2. *Scatter plot of the dataset:*





Hence, based on the above observations, for any given tuple(zipcode, date, hour), if there is an increase in the number of events and a corresponding increase in the number of taxi trips, then there is a positive correlation between them.

Although, this is not sufficient information to determine the most happening place in the city, taxi trip does play a major contribution.

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 to improve 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.

VIII.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 run the project efficiently and economically.

IX.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] *Sensing Urban mobility and taxi flow*, Marco Veloso and Santi Phithakkitnukoon.