Indian Institute of Information Technology, Allahabad

PROJECT REPORT

"Apache Flink Exploration and Analysis of NYC Taxi Ride Data"

Project Supervisor - Dr. Sonali Agarwal

Declaration by the Candidates

We, hereby declare that the project titled "Apache Flink Exploration and Analysis of NYC Taxi Ride Data " is a record of bonafide project work carried out by us under the guidance of *Dr. Sonali Agarwal* in partial fulfillment of the 7th semester Project work for the B.Tech (IT) Course in Indian Institute of Information Technology, Allahabad.

Saurabh Tanwar – IIT2014140

Sachin Agarwal – IIT2014501

Sacheendra Mohan Singh – IIT2014506

Certificate

This is to certify that the project report entitled "Apache Flink Exploration and Analysis of NYC Taxi Ride Data " submitted to Department of Information Technology, Indian Institute of Information Technology, Allahabad in partial fulfillment of the 7th semester Mini-project work, is a record of bonafide work carried out by:

Saurabh Tanwar – IIT2014140
Sachin Agarwal – IIT2014501
Sacheendra Mohan Singh – IIT2014506

under my supervision and guidance.

This report has not been submitted anywhere else for any other purpose.

Submission Date : 22/11/2017

Dr. Sonali Agarwal
Assistant Professor
Department of Information Technology
Indian Institute of Information Technology
Allahabad - 211012

Acknowledgement

We would like to express our special thanks of gratitude to Dr. Sonali Agarwal who gave us the opportunity to do this project titled “Apache Flink Exploration and Analysis of NYC Taxi Ride Data” . We appreciate her contribution, constant support and perseverance in this endeavor of ours. Her engagement through the process of this project has been precious and irreplaceable.

CONTENTS

ABSTRACT

With the amassing increase in development of urban cities, the urban data also needs to be analysed and studied for managing the city efficiently. Among the various variables in a city transport is a major one. Various modes of transportation are in use. In most of the large cities in US like New York, taxi is a prominent source of transportation. This data obtained using Global Positioning System(GPS) for every taxi ride has grown tremendously and needs to be managed and studied for providing various statistics.

Earlier the Taxi Ride data were analysed by analysts and various outputs were derived according to some criteria which would be helpful to people who commute. After early 2000, the data grew exponentially and it was in GB's, which became impossible to analyse by analysts. Thus to overcome this Big Data which had started prevailing came into picture. Using Big Data the analysis was done effortlessly in minutes. This analysis is helpful in many cases such as traffic management, finding popular locations, finding suitable market for various third party organisations and many more.

I. INTRODUCTION

I.I BACKGROUND

Transportation plays a very vital role in major cities. Various modes of transportation are in use. In most of the large cities in US like New York, taxi is a prominent source of transportation. Also compared to earlier times nowadays there is a rapid evolution in the urban cities, and metropolitan zones have become a rich source of data which can be a good source of first hand data to be used in various Big Data Projects.

I.II MOTIVATION

BigData analysis of cities is helpful in many purposes such as analysing traffic condition in city, providing good transport services. It can also be used by Government Officials to take necessary measures to manage traffic. Also third parties such as Uber can use this data to provide cab services in city.

Some projects based on analysis of Taxi Data data has been done using technologies such as Spark, Hadoop but not much has been done using Apache Flink. Thus, as there is not much work based on Apache Flink there is a need for the same.

II. PROBLEM STATEMENT AND OBJECTIVE

Given a stream of taxi ride events from the public data set of New York Taxi and Limousine Commision. The dataset consists of records about taxi trips in New York city from 2009 to 2015. We took some of its data, used Apache flink for data stream processing , elasticsearch as a backend to store major information, and tools such as Kibana for dashboard visualization to complete the following objectives.

Our aim is to use as much as features of Apache Flink as possible for analysing the data.

II.I OBJECTIVES:

1. To identify popular locations of New York city

It takes stream of taxi ride events and counts for each coordinate the no. of people that arrive there by taxi. Apache flink features used :- (map, filter, fold,window).

2. To identify popular pickup locations for any two particular locations.

It takes stream of taxi ride events and two destination locations as argument and counts for each location the no. of times it is used for travelling to either of the given destination. Apache flink features used :- (map, filter, fold,join,window) .

3. To split the datastream into day and night time and identify popular locations during those time.

It takes stream of taxi ride events and perform splits the stream into two new streams, computes for each stream the popular locations. Apache flink features used :- (map, filter, fold,split,window) .

4. To compare popular locations over time interval of about 6 months.

III . LITERATURE SURVEY

RELATED WORKS :

Now here we are going to discuss the Technologies which we focused on to analyse huge dataset of New York Taxi Ride. The complete analysis involve using BigData with Apache flink and Apache flink Ecosystem, Kibana for visualization and Elasticsearch for the data store. After that brief definition of BigData , Elasticsearch , kibana etc.

3.1 Big Data

"Big Data" tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. Big data “Size” is constantly moving target moving from few dozen terabytes to the petabytes of data. it usually includes data which is huge in amount and beyond the ability of any commonly used software tools.[1]

3.2 Apache Flink

Apache Flink is an open source stream processing framework developed by the Apache Software Foundation. The core of Apache Flink is a distributed streaming dataflow engine written in Java and Scala. Flink executes arbitrary dataflow programs in a data-parallel and pipelined manner.[2]

3.3 ElasticSearch

It is a data-collection tool .it is developed alongside a data-collection and log-parsing engine called Logstash. Elasticsearch is a search engine based on Lucene. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents. Elasticsearch is developed in Java and is released as open source under the terms of the Apache License.[3]

3.4 Kibana

KIBANA is an open source data visualization plugin for Elasticsearch. It provides visualization capabilities on top of the content indexed on an Elasticsearch cluster.The combination of Elasticsearch, Logstash, and Kibana (also known as ELK stack or Elastic stack) is available as products or service.[4]

Also, in this section we focus on earlier big data projects on NYC taxi dataset taxi dataset, namely to optimize taxi usage and to use as much as features of Apache Flink as possible for analysing the data.

- Transdec (Demiryurek et al. 2010) is a project which provide us with the end-to-end data driven system enabling spatiotemporal queries in transportation systems with,real-time, dynamic and historical data. University of California build this project to create a big data infrastructure adapted to transport and it's built on three tiers comparable to the MVC (Model, View,Controller) model for transport data. [5]
- (Jagadish et al. 2014) worked on this project for Exploring the inherent technical challenges in realizing the potential of Big Data .[7]
- (Yuan et al. 2013), (Ge et al. 2010), (Lee et al. 2004) They work on optimising the availability of empty taxi to clients as soon as possible i.e travel time from taxi to clients on historical data collected from running taxis hence they help taxi companies to optimize their taxi usage. this mobile recommendation system is built to maximize the probability of business success and provide Potential Travel Distance (PTD) function for evaluating each candidate sequence. [6]
- (Jasmina Smailovic, Janez kranjc, Miha Grca), (Martin Znidarsic), (Igor Mozetic)The paper mainly focuses on monitoring the real-time twitter sentiments. For the purpose of achievement of goal some high quality models are used. Feature selection is done in order to maximize the prediction accuracy, binary SVM classifier is also used for feature

extraction its advantage is it is robust to overfitting and uses quite less memory, linear SVM function is used. The paper also focuses on improving the language detection system in tweets as the current twitter API for language detection often can't correctly distinguish between very similar languages so for improving the language detection machine learning models are used. The results shows that the negative sentiments about the political party keeps prevailing even after elections. The difference between the positive and negative tweets for a particular political party is close to election result. [8]

- (Daniel Gayo-Avello),(Panagiotos T.Metaxas),(Eni Mustafaraj) focuses on US elections held in 2010. Predicting election result from social media is quite popular and easy to do thing there are cases when positive results are reported but the principles are not known and there is no proper analysis. In this paper techniques are used on US 2010 congressional elections. After analysis it is found that there is no correlation between the analysis results and election results. Which contradicts the previously available reports. Authors state that people should not accept predictions based on social media as a black box. There should be a proper model which explains the predicting power clearly .[10]
- (Pritee Salunkhe, Avinash Surnar, Sunil Sonawan) This paper emphasis on using the twitter and others social sites for predicting the election result In this paper, various strategies are discussed regarding to political leaning like user graph, twitter specific features, user behaviour for the appropriate prediction of the elections. And also focuses on making election strategies by locating the region wise popularity of candidates from twitter API thus making more efforts on regions where popularity is less . they also shows that collecting data based on the time before the election can make huge difference . they also provide a view related to several features as comparison using user's linguistic content,posting behaviour, reply and concluded that The combination between user profile and linguistic outperforms other features .[14]
- (Alexander Pak), (Patrick Paroubek) uses the twitter platform for the purpose of opinion mining and sentiment analysis.Corpus is collected automatically using the twitter API. Then, corpus analysis is done using frequency of words used mathematical relations for calculation of distance. So far, collected data set is used to extract feature from it so that

we can train our sentiment classifier based on this data. sentiment analysis and opinion mining is done on it. Entropy calculation is also done for dataset in order to improve the accuracy. Thus, a linguistic analysis is done for for the automatically collected corpus and a sentiment classifier is build which is capable of determining positive, negative and neutral sentiments. The techniques used are quite efficient here which are not published before .[9]

IV. METHODOLOGY

From the literature review we concluded that the BigData analysis of taxi ride and transportation system is very helpful in maintaining city traffic and providing various other insights to government as well as other companies related to transport. But, it is also found that the results can't be used as a black box in every region, due to the following reasons – lack of technology leading to unavailability of data in other not so popular places, incomplete data due to not every taxi driver not connected to GPS, so actual result can slightly differ due to the data being incomplete in some places other than New York.

We aim to present here an efficient analysis of New York Taxi Ride data which detects popular locations throughout New York based on count of passenger arrivals. We also compare the popular locations at Day and Night time and present visualization of popular pickup locations for any two fixed cities. At last we present a comparison of popular locations across various months.

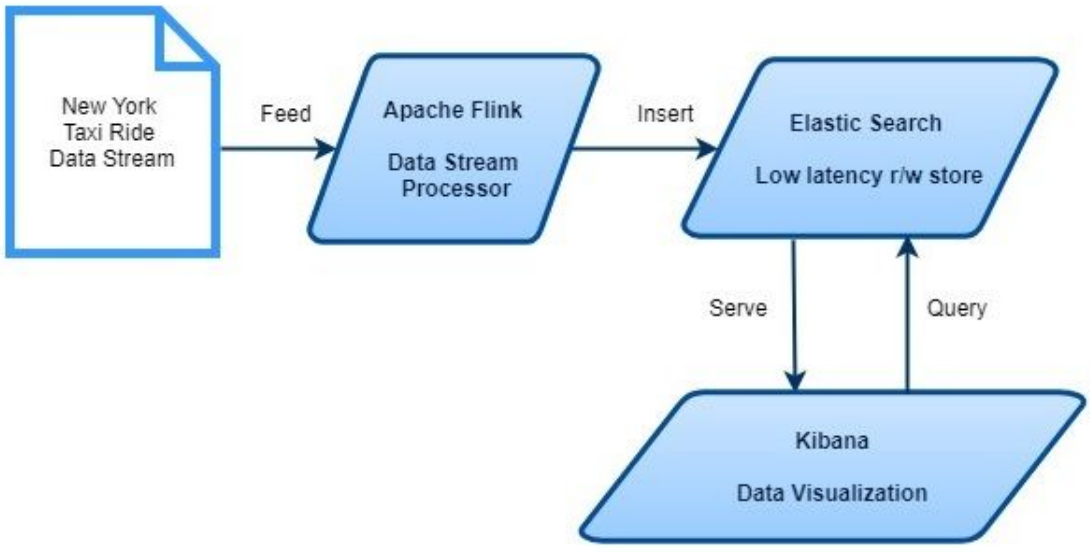

Fig 1. Flowchart representing the modules and data flow in them

V. SOFTWARE/HARDWARE REQUIREMENTS

SOFTWARE REQUIREMENTS:

1) Programming Tools:

    Java JDK, Scala, Apache Flink , Elastic Search, Kibana, IDE Intellij IDEA

2) Dataset:

    NYC Taxi Data [2009-2015]

HARDWARE REQUIREMENTS:

Hardware Specifications:

Architecture: 64 bit system
Operating System: Windows 10
Microprocessor: 2.8 GHz Intel Core i5-6610U
Memory: 8 GB 1600 MHz DDR4L SDRAM (2 x 4 GB)

Minimum Requirement:

Microprocessor: Intel i3 2.0Ghz or above
Memory: 8GB or above
Hard Disk: about 100 GB
Operating System: Windows / Linux

VI. IMPLEMENTATION :

1. The New York Taxi Ride Data is taken from year 2009 to 2015 and converted it into a modified dataset with taxi ride events containing information relevant to our analysis. Events contain the following fields :

   RideId (Long), Time (DataTime), IsStart (Boolean), Location(Geopoint), PassengerCount (Short), TravelDistance (Float), StartLocation (Geopoint).

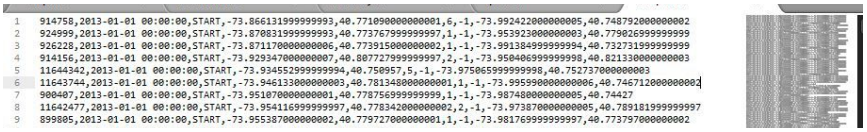

Fig 2. Data set

2. **Apache Flink** takes this dataset as an infinite stream and processes this data with various transformations like map, reduce, fold, window representing timeline to take data, and other transformations such as filter, join, split etc.

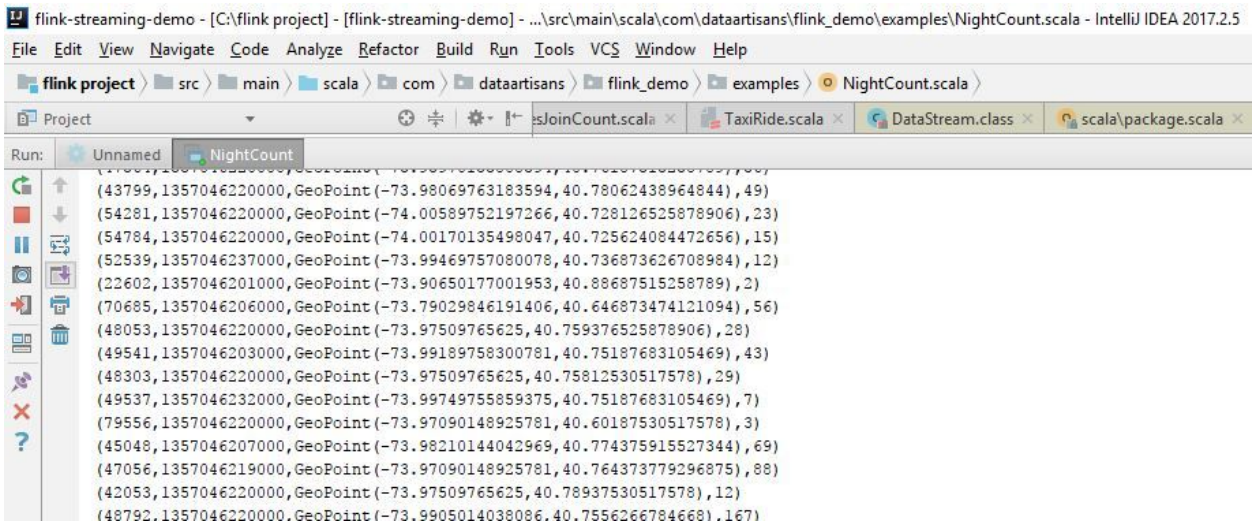

Fig 3. Apache FLink processing data stream

3. The processed data is now feed into **Elastic Search** in a suitable JSON format at localhost port 9300 in a proper index, created beforehand in ES. Our project contains 4 indexes namely - nycidx1,2,3,4 with each index for storing data dedicated to its specific objective.

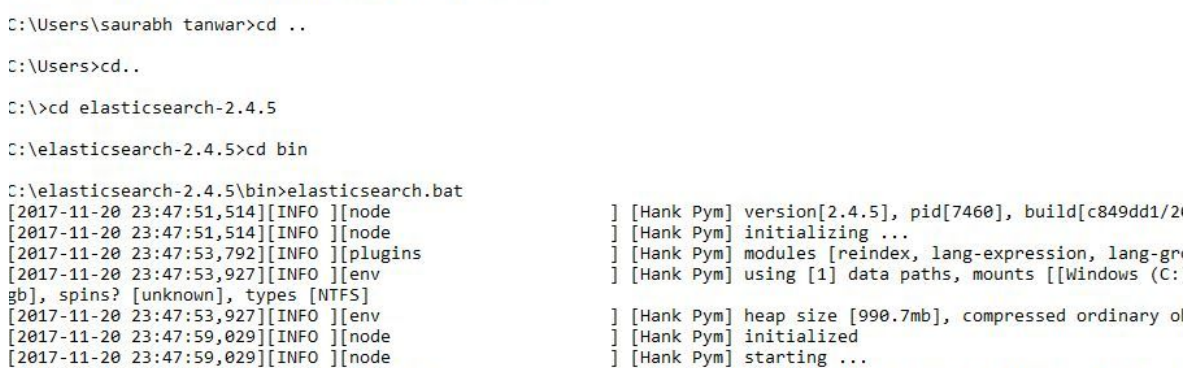

Fig 2. Elastic Search running on command prompt

4. Finally, the data stored in ES is used for visualization using **Kibana** tool. We use tilemap feature of Kibana to visualize popular locations directly in the New York Map and compare them across various timelines.

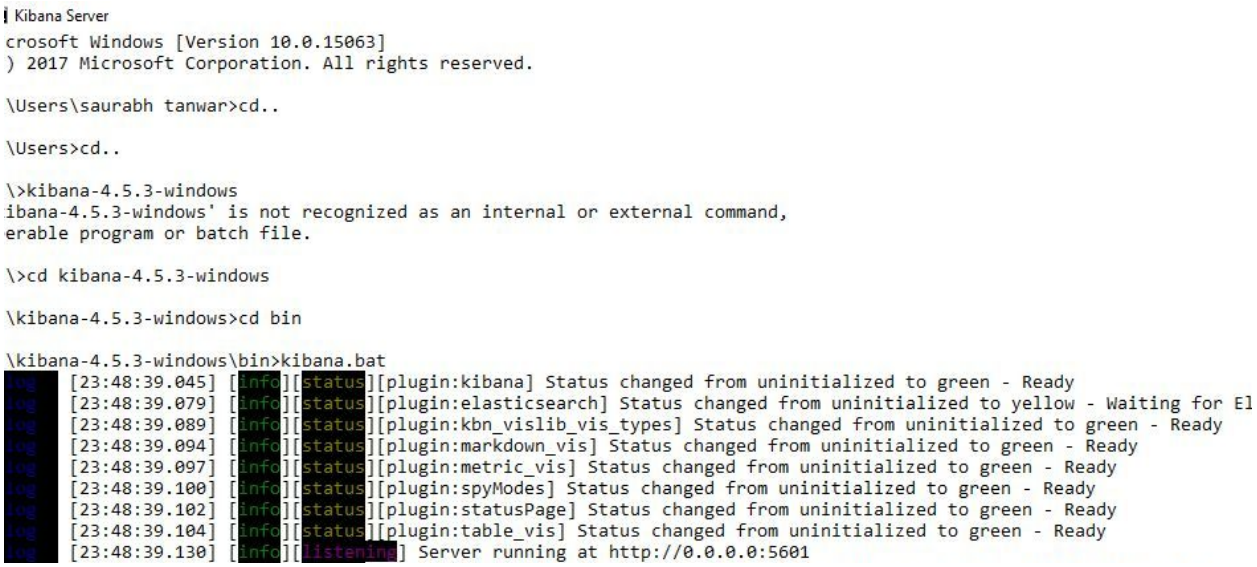

Fig 3. Kibana running on command prompt

VII. RESULTS

1. The figure below represents the most popular locations of the New York city. The colors red,orange, yellow represents most popular, moderately popular and popular cities respectively.

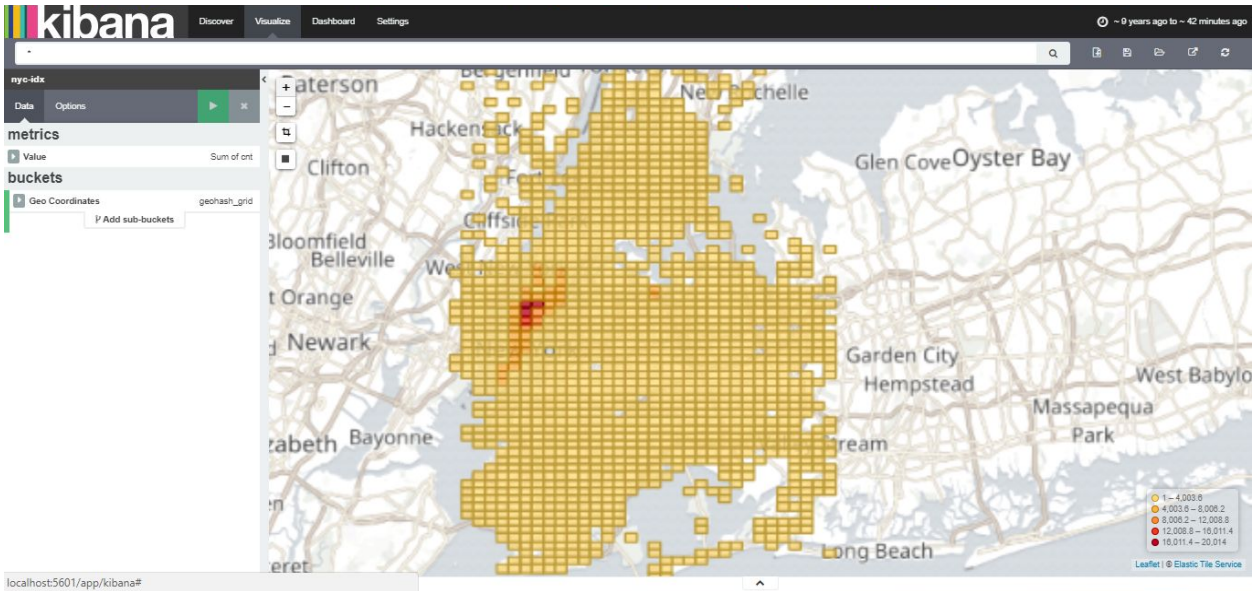

Fig 4. Most popular Arrival locations

2. The figure below represents the most popular pickup locations for two selected coordinates.

1. { -73.978837999999996, 40.740819999999999 } 2. {-73.783680000000004, 40.646217}

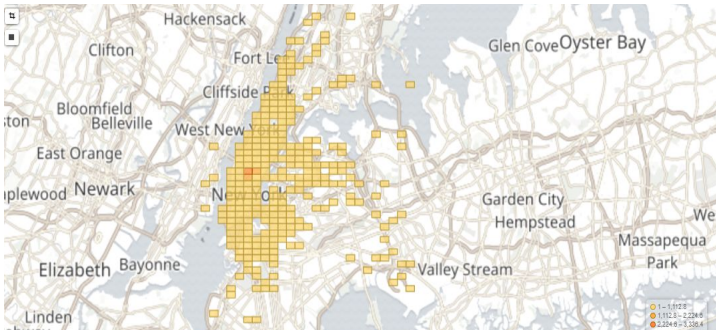

Fig 5. Most popular pickup locations

3. The figure below represents the most popular arrival locations during day-time over a period of 6 years (2009-2015).

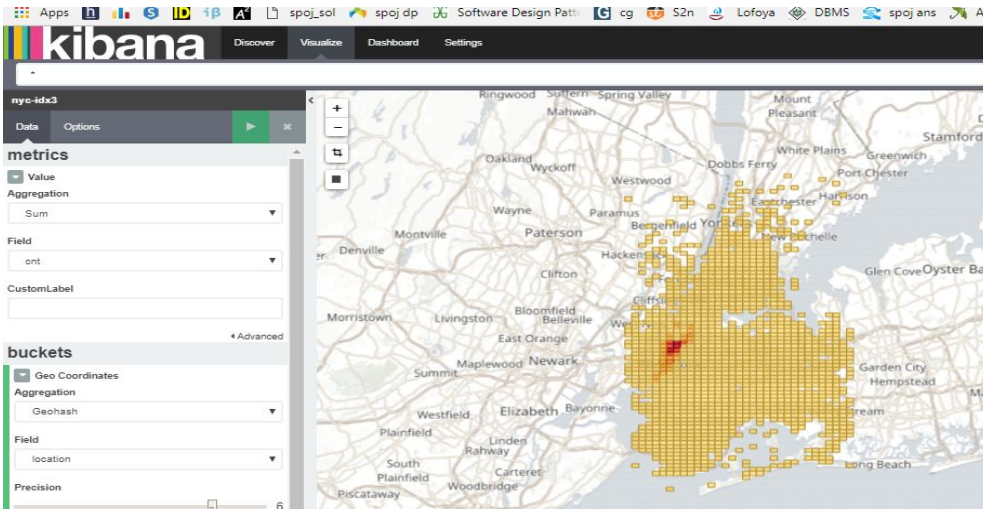

Fig 6. Most popular pickup locations (Day-Time)

4. The figure below represents the most popular arrival locations during night-time over a period of 6 years (2009-2015).

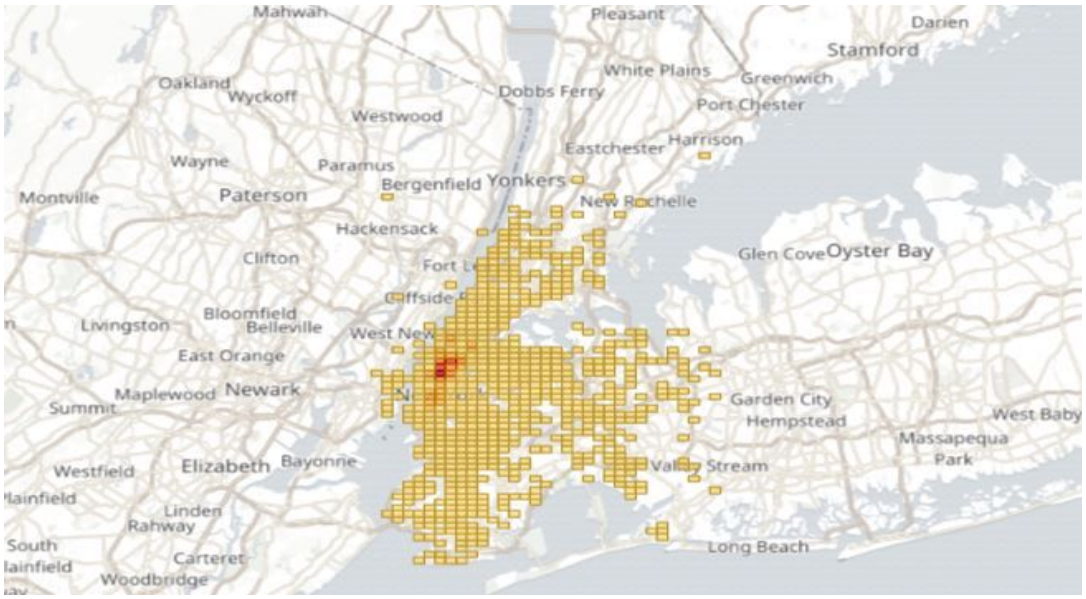

Fig 7. Most popular pickup locations (Night-Time)

5.The below two figures shows the comparison of most popular arrival locations between first and second phase of 2013. It's observed that the first phase there are more locations which are popular while during the second phase the popularity of those locations gradually decreases.

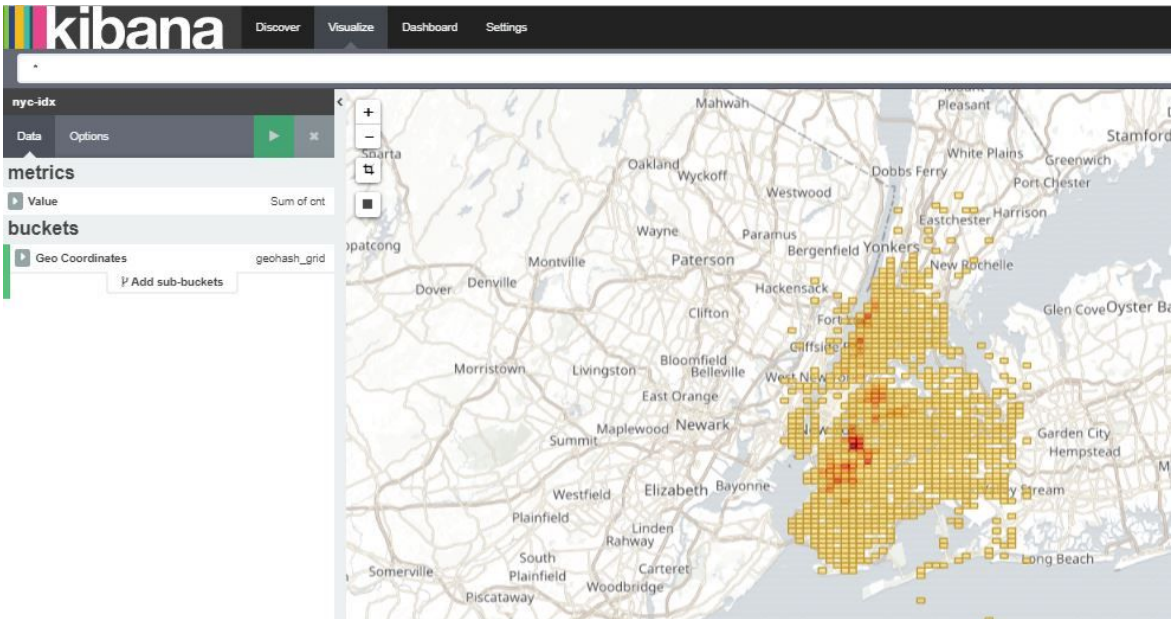

Fig 8. Most popular arrival locations (Jan-Jun (2013))

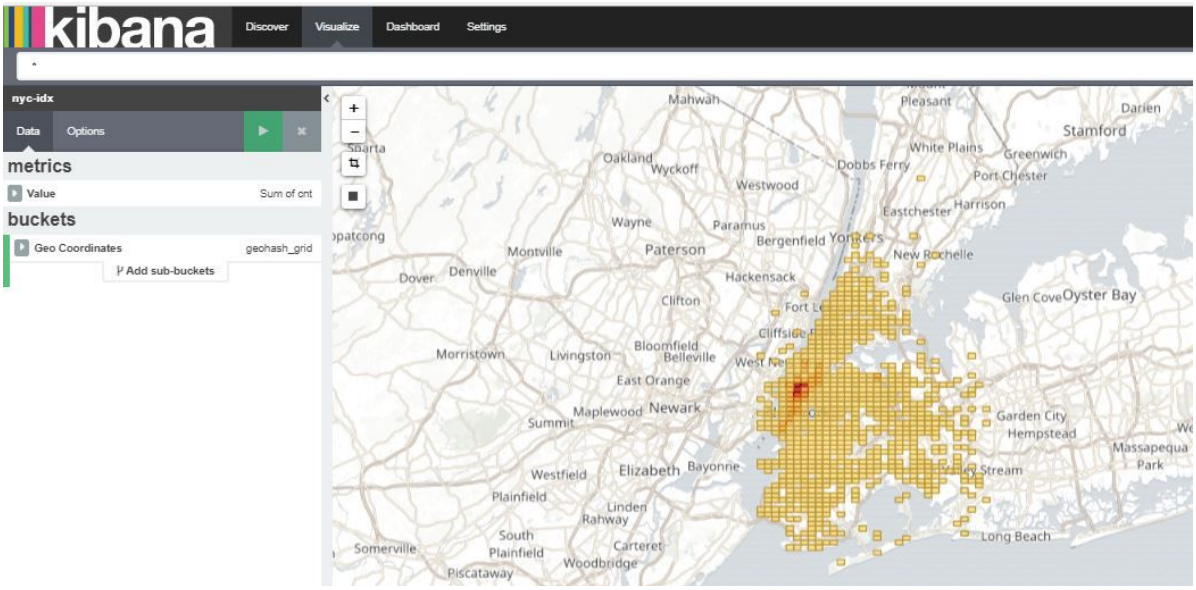

Fig 8. Most popular arrival locations (Jul-Dec (2013))

VIII. REFERENCES

[1].“BigData,” https://en.wikipedia.org/wiki/Big_data [Accessed November , 21/11/17].

[2]. https://en.wikipedia.org/wiki/Apache_Flink [Accessed November , 21/11/17].

[3]. https://en.wikipedia.org/wiki/Elasticsearch [Accessed November , 21/11/17].

[4]. https://en.wikipedia.org/wiki/Kibana [Accessed November , 21/11/17].

[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]. TJasmina Smailovic, Janez kranjc, Miha Grcar, Martin Znidarsic, Igor Mozetic , Monitoring the Twitter sentiment during the Bulgarian elections, Oct. 2015

[9]. Alexander Pak, Patrick Paroubek , Twitter as a Corpus for Sentiment Analysis and Opinion Mining, June 2011

[10]. Daniel Gayo-Avello, Panagiotos T.Metaxas and Eni Mustafaraj , Limits of Electoral Predictions Using Twitter , 2011

[11]. David Anuta, Josh Churchin & Jiebo Luo. . Election Bias : Comparing Polls and Twitter in The 2016 U.S Election ,2016

[12]. Alexandre Bovet, Flaviano Morone, Hern´an A. Makse. Twitter Validation of Twitter Opinion Trends With National Polling Aggregates : Hillary Clinton vs Donald Trump . April 2017

[13]. Nick Beauchamp Northeastern University. Predicting and Interpolating State-level Polling using Twitter Textual Data. September 22, 2013.

[14]. Pritee Salunkhe, Avinash Surnar, Sunil Sonawane. A Review: Prediction of Election Using Twitter Sentiment Analysis , 5 may 2017 .

[15]. Nugroho Dwi Prasetyo ,Claudia Hauff. Twitter-based Election Prediction in the Developing World, 2015

Suggestions by Board Members