NYC Taxi Analysis

Shivesh Ganju
Courant Institute of Mathematical
Sciences, New York University
New York, USA
sg6148@nyu.edu

Ravi Shankar
Courant Institute of Mathematical
Sciences, New York University
New York ,USA
rs6980@nyu.edu

***Abstract*—**

One of the leading global financial hubs in the world and consequently becoming the most densely populated city in the United States, New York City's transportation system has been studied from a multitude of aspects. An average taxi driver in NYC spends 92 hours in a year stuck in the traffic, which comes out to be one of the highest across all cities in the United States. In this paper we use Big Data Analytics to analyse several taxi networks such as Yellow Taxi, Green Taxi, Uber, Lyft, Via and Juno to figure out important trends and features of human society. The correlation between taxi operations and different weather conditions such as snowfall depth and precipitation have been explored in the paper. We also show the emergence of Uber in NYC and how Uber dominated the market share amongst Yellow Cabs, Green Cabs, Lyft, Via as it consistently completed more than 10 million rides starting October 2017. Our results depict that taxi analysis helps unlock key insights into social phenomena before they are visible to the public.

Keywords—Analytics, HDFS, RDD, Spark SQL, API, NYC TLC, Linear regression, Random forest, Tableau, Boroughs, Zones, Yellow Taxi, Green Taxi, FHV vehicles.

I. Introduction

As New York is a leading and one of busiest global centres across the world, the usage of taxis for transportation is immense. It is estimated that on an average people of New York travel more than a million miles using the various taxi services available in New York City [5]. Human transportation on such a level has the potential to unlock several useful insights into human trends and behaviour. Transportation is usually linked with several factors such as crime statistics to housing prices within a city. Hence it is imperative to understand the travelling patterns to explain and predict the social phenomenons. In this paper we aim to analyse the effect of Uber and Lyft entrance into NYC and how they slowly but surely captured the market share as compared to Yellow and Green taxis. Yet another phenomena we observe is how weather conditions including precipitation, snow depth, and snowfall impact the taxi operations. We also wish to establish borough wise pickups and drop offs and visualize the same using heat maps of trip volumes of neighbourhoods across the city.These insights can be leveraged for studying crime statistics and housing prices in the localities of the city. As of 2014 NYC has released a rich dataset with pin-point pick ups and drop-offs across the city. This dataset now includes information about more than a billion taxi rides across the city which shows the data is ripe for network analysis.

II. Motivation

The motivation for the analytics stems from the fact that as NYC is one of the most densely populated cities in the United States, the transportation and the taxi systems plays a pivotal role in shaping the social phenomenons of the city which are intertwined with each other. Through our study we aim to analyse when and how Uber emerged as the goto taxi service in NYC where multiple taxi services co-exist. An interesting topic we want to discern is the correlation between the tipping and the seamless service provided by the driver, thus unlocking the intricacies of human behaviour. We also aim to analyse the vibrant nightlife and the impact of taxi services on posh neighbourhoods and vice-versa. Any person can benefit from the study who wishes to make decisions for places to live based on certain trends such as crime statistics or the serenity of the place or who wants to maximise their savings on transportation cost throughout the year in NYC depending on various factors such as the weather conditions, time of pick ups. Impact of adverse weather conditions on taxi services and it's consequent effect on human behaviour will also be studied. Infact by doing a thorough analysis of the taxi services and the tips offered by customers we can help maintain a healthy ecosystem amongst the taxi drivers of several taxi companies. We compare and contrast between several taxi services such as Yellow, Green, Uber, Via, Lyft taxis.

III. Related Work

We aim to use our analytics to derive various insights such as the correlation of the crime rate in a borough with the late night taxi pickups and dropoffs as well as the socio-economic status of various boroughs. *Correa et al.* [2] have done a spatio-temporal analysis to derive a dependence of taxi demands on their geographical locations and have derived socio-economic and transportation related characteristics through Linear models , spatial error and lag models. The paper aims to study the impact of the app based for hire vehicles on the NYC taxi industry. The authors have used the NYC TLC taxi dataset which has the taxi pick ups and drop offs information of yellow Taxi, Green Taxi and For hire vehicles like uber, lyft, juno etc. The authors have explored the spatio-temporal patterns of the demand for uber taxi and have analyzed the factors that contribute to increasing demand for uber. The authors have first visualized the demand for uber in the entire NYC region through a heat map. Through the heat map they were able to figure out the distribution of the demand through space. They were able to note the insight that the Yellow Taxis are highly concentrated in Manhattan and airports and although the population of Manhattan is less, they were able to generate more demands than other boroughs. The other insight they generated was that unlike yellow taxis, the demand for uber was distributed evenly. The authors have correlated the uber and yellow taxi pickups in various boroughs by drawing a heat map of the pearson's coefficients of correlation. They found out that there is a high correlation between the two in central city areas.

The authors then drew other insights as well like how the increasing pickup trend of uber coincides with the decreasing trend of yellow taxi data for 2015. Major reason for this according to the authors was because ride hailing apps were cheaper and were more flexible. Through the pickup data they were able to find other trends such as the distribution of pickups in the taxis according to the time of the day. Lastly, the authors developed a spatial lag model to quantify special dependence of uber and taxi pickups using Moran's I statistics. They used a linear model and used a spatial error model for spatial dependence. Through the linear model, spatial error models and the spatial lag models they were able to estimate the taxi and uber demands for each neighbourhood using socio-economic and transportation related characteristics. Key factors which affected taxi and uber demands were transit time, length of roadways, vehicle ownership, high income and high education.

For the future work, the authors will also model the interaction between taxi usage of the alternate mode of transportation such as Bikes, subways and other fhv services like Lyft as well as finding out whether the concentration of uber is clustered near subway station or outer boroughs

We also derive key insights using the weather data in conjunction with taxi data sets. *Tang* [6] has done a predictive analysis for the effect of various temperature conditions on the taxi network across NYC and leveraged Machine Learning techniques such as Linear Regression to predict trip amount based on weather states. The correlation between taxi operations and different types of weather, including precipitation, snow depth, and snowfall is discussed in this paper. Specifically the author begins by analysing the trends for a particular month (January) for 2015. Daily pick up amounts and average pick up during each hour were plotted for the month of January. This was used to get an insight about the peak hours during the winter season and taxi pick up trends were studied for the same. Additionally, all observations were extended for monthly trend, weekday trip amount, hourly average trip, and weekday payment were studied too for the year 2015. The impact of weather factors such as snow depth, precipitation, snowfall were done on the number of daily trip amounts in the city. Pearson's correlation was used to check if the variables were linearly dependent on each other and the graph was plotted too using the best weights obtained from the linear regression analyser. As an add on the author describes the differences of impact between snowfall and snow depth and which one has more adverse effect on the daily trip amount in NYC.

Another useful insight drawn by the author is recognising the distribution of taxi networks across the city. By looking at the trips from 8 a.m. to 10 a.m it was found out that certain areas in the city required more taxis during this peak time and thus this data can be used by the taxi services to arrange their taxis accordingly so that it benefits both the customer and the driver by meeting the demand with supply. This worked out to be an important metric as it helps prevent surcharging of fare prices since availability is high at the peak time. The limitation for author's research is the huge volume of data and insufficient computing capacities which limits the research work only for the year of 2015.

Patel et al.[7] have used the dataset provided by NYC Taxi and Limousine commission for their analytics. The authors have carried out their analytics using tools provided by Hadoop ecosystem such as MapReduce, Hive And Pig. The analysis is carried out on the 2014 dataset. Using the tools,they have conducted analysis which recommends the top driver based on factors such as most distance travelled,most fare collected, most time travelled and the most efficient driver.They have further suggested that these parameters might help the TLC in awarding the drivers. The authors have also used geospatial data such as the ride's pickup and dropoff location to conduct analysis on region to figure out the location with the most pickups and drop-offs.Their insight would help in increasing the number of taxis in the locations where the demand is more so as to maximize revenue.They have further visualized the demands acrossNYC using a heatmap.Finally,the authors have analysed the fare to get an estimate of the driver's revenue. They have carried out this analysis using the gross and net revenue of the drivers with the help of Pig.The authors have developed a visual query model which would enable the users to select data slices and use them. The findings would enable city planners, engineers and decision makers information about how people use the mode of transportation. The information could also be used to predict the growth of taxi demands in the city as neighbourhoods evolve as well as develop a model for analysing taxi movement patterns.

Guo et al.[8] explains the various trends as to why Uber is the most popular public transportation choice amongst several modes of transportation available in New York City. The author generates a heat map for Uber pickup request trends across the year for all months and weekdays. It is found out that most rides are in the month of November as the winter kicks in where as it steeply decreases in December owing to the onset of holidays. It was also found out that peak hours of pick up were 7am to 9am in the morning and 5pm to 7pm in the evening during the office commute hours. To compare and contrast the analysis with Yellow and Green taxi services moving weekday averages for requests per hour vs hour were plotted during several months of 2015 and 2017 and it was found that although day hours showed a similar trend for all taxi services, Uber was the most preferred taxi service during the night owing to Uber's freedom given to their drivers to choose their own flexible driving hours which can be during night as well. Another trend which was observed is that requests for pickups are higher in the evenings/nights as compared to early morning maybe due to the fact people are tired after work in the evening and thus prefer easy means of travel as compared to the morning when they are fresh. The author also tries to link the above insights to the area of pick ups across the city, thus a spatial distribution is generated using the gradient of color depth across the map. Heat maps of the taxi services such as Yellow, Green, Uber are plotted. A key insight generated was that although Uber dominates the pickups across the city, density of yellow and Uber taxi services are equal near the airports as people prefer using yellow cabs too owing to their fair fare price strategies from the airport. Also it was found out that green taxi pickups were the least in lower Manhattan as NYC regulations don't allow them to operate in certain regions of the city. Author also studies the effect of rain intensities on the pickups and found out that there weren't significant changes in the amount of pickups with increasing or decreasing rain levels. The limitation of the author's research is lack of rich data available from the Transit App and the huge volume of data and insufficient computing capacities which limits the research work only to the years 2015, 2017.

IV. Datasets

The below datasets were collected from NYC TLC[3] which provided the data for each month and year. The data was collected through API calls and were then collectively put in HDFS. This data was collected statically using the API provided by NYC TLC.

A. Yellow Taxi Dataset and Green Taxi Dataset

Dataset Size for Yellow Taxi - 232 GB

Dataset size for Green Taxi - 9.2 GB

Column Name	**Datatype**	**Description**
VendorID	String	A code indicating theTPEP provider that provided the record
trip_distance	Float	The elapsed trip distance in miles reported by the taximeter
start_latitude	Float	TLC Taxi Zone latitude in which the taximeter was engaged
start_longitude	Float	TLC Taxi Zone longitude in which the taximeter was engaged
end_latitude	Float	TLC Taxi Zone latitude in which the taximeter was disengaged
end_longitude	Float	TLC Taxi Zone longitude in which the taximeter was disengaged
tip_amount	Float	The amount of tip given using either cash or credit
Start_Date	String	The date at which the trip was started. The date is of the form yyyy-mm-dd Range for Yellow taxi dataset - 2009-01-01 to 2019-12-31 Range for Green Taxi DataSet - 2013-08-01 to 2019-12-31
Start_time	String	The time at which the trip was started. The date is of the form HH:MM:SS. This field was later bucketed in with 1 hour ranges starting from 00-01 and ending at 23-00. Eg:- Trip having a pick up time of 02:00:12 lied in 02-03 bucket. 12 categories were made.
End Date	String	The date at which the trip ended. The date is of the form yyyy-mm-dd Range for Yellow taxi dataset - 2009-01-01 to 2019-12-31 Range for Green Taxi DataSet - 2013-08-01 to 2019-12-31
End Time	String	The time at which the trip ended. The date is of the form HH:MM:SS
TotalAmount	Float	The total amount paid in dollars for the taxi trip. For yellow taxi the range is from 2 dollars to 50

		dollars and for green taxi the range is from 2 dollars to 45 dollars
Time_Duration	Float	Difference between pickup time and drop off time is the trip duration.
Speed	Float	Speed is calculated as Trip_distance/trip_duration. Upper limit is set as 50 since it is the speed limit in NYC
Payment Method	Integer	Refers to the mode of payment used. 0 - Card payment, 1 - Cash payment

The above dataset contains information about the yellow taxis that run in New York city. The original schema which was provided by TLC had 17 columns. The columns that were not used in our analytics were dropped off. Moreover each year the data schema was different so the dataset had to be cleaned so as to maintain the above mentioned schema uniformly. Missing value imputation was done by removing the vectors which had a missing value since they would not be helpful for our analytics. The data loss which happened due to the above step was about 0.001% and so it would not affect our analytics.

B. Weather Dataset

This dataset was statically collected from the National Climatic Data center[4] and has the daily data of Central Park,NY from 2009 to 2019. The size of the dataset was 500KB and the purpose of including this dataset is to draw insights about the effect of weather on the taxi dataset

Column Name	Datatype	Description
Date	String	The date for which the weather was collected. The value is of the format yyyy-mm-dd
Precipitation	Categorical variable of type String	The original datatype of this field was float and it ranged from 0 inches to 1 inch. This variable was bucketed with the range of 0.2 with 5 buckets :- 0,0-0.2,0.2-0..4,0.4-0.6,>.6
Snowfall	Categorical variable of type String	The original datatype of this field was float and it ranged from 0 inches to 10 inches. This variable was bucketed with the range of 2 with 5 buckets :- 0,0-2,2-4,4-6,>6

The original dataset contained over 8 columns. The columns which were not required were dropped and only the precipitation and snowfall were retained as our analytic would require the snowfall and the rainfall data to derive a correlation with the taxi fares and pickups.

C. For-Hire Vehicle (FHV) Trip Records

Dataset Size for FHV Taxi - 45 GB

Column Name	**Datatype**	**Description**
Hvfhs_license_num	String	The TLC license number of the HVFHS base or business. As of September 2019, the HVFHS licensees are the following: • **HV0002: Juno** • **HV0003: Uber** • **HV0004: Via** • **HV0005: Lyft**
pickup_id	Integer	TLC Taxi Zone pickup Id in which the taximeter was engaged
dropOff_id	Integer	TLC Taxi Zone pickup Id in which the taximeter was engaged
Base_num	String	The TLC Base License Number of the base that dispatched the trip
Start_Date	String	The date at which the trip was started. The date is of the form yyyy-mm-dd Range for FHV taxi dataset - 2015-01-01 to 2019-12-31

Start_time	String	The time at which the trip was started. The date is of the form HH:MM:SS. This field was later bucketed in with 1 hour ranges starting from 00-01 and ending at 23-00. Eg:- Trip having a pick up time of 04:05:13 lied in 04-05 bucket. 12 categories were made.
End Time	String	The time at which the trip ended. The date is of the form HH:MM:SS

The above dataset includes information about For-Hire Vehicles and High-Volume For-Hire Vehicles i.e categorised on Uber, Via, Lyft and Juno. Apart from the above mentioned column names, the original dataset included SR_Flag too which wasn't required for our analysis. Dataset for years 2015, 2016, 2017 included a substantial amount of missing values for drop off time and place. As most of our trends were based on pick up time and places, which were consistent in the entire data set, thus the missing data rows were retained and missing values were imputed so as to prevent data loss. Net data loss was approximately 0.02%. Different year's datasets had jumbled up schema which were all mapped to the schema shown above.

D. For Hire Vehicles Base Mappings

Dataset size is 45.7 KB

Column Name	**Datatype**	**Description**
Base_num	String	The TLC Base License Number of the base that dispatched the trip
Base_name	String	The corporation that owns the particular taxi in consideration.
Vehicle_category	String	Vehicles are categorised into **Uber, Lyft, Juno, Via or Other**. This is the mapping from Base_num to vehicle type.

This dataset has been utilised in conjunction with the Hvfhs_license_num column of the For-Hire Vehicle dataset to create a mapping from Base_num/Hvfhs_license_num into the four categories of vehicle such as Uber, Lyft, Via, Juno.

E. Taxi Zone Mapping

Column	Datatype	Description
Zone ID	String	Identifier which maps a taxi zone id to the borough. This has been used in the other taxi datasets
Zone	String	This refers to the taxi zone in NYC. Eg:- Lower Manhattan, Newark Airport
Borough	String	NYC Boroughs. Eg:- Manhattan, Queens, Brooklyn,Long Island

This data set has been used to map the pickup ID and drop off ID which are present in the Yellow taxi and green taxi dataset with their respective taxi zone and boroughs which will be later used in analysis

V.Description of Analytic

The size of the dataset is huge which helped us in deriving some keen insight about the market trend for the taxis as well as some insights into the tipping behaviour of people. We were also able to draw some correlation between the weather and taxi pickups and fare. Apart from this, the late night pickup trend also helped us in deriving the areas in New York city which had a vibrant nightlife.

1. Pickup Trends Analysis

The insights were generated in this section by using Spark RDD operations which were used to aggregate the total number of pickups for each year and for each month. The market share was then calculated by using the above generated data along with the total number of pickups in that year

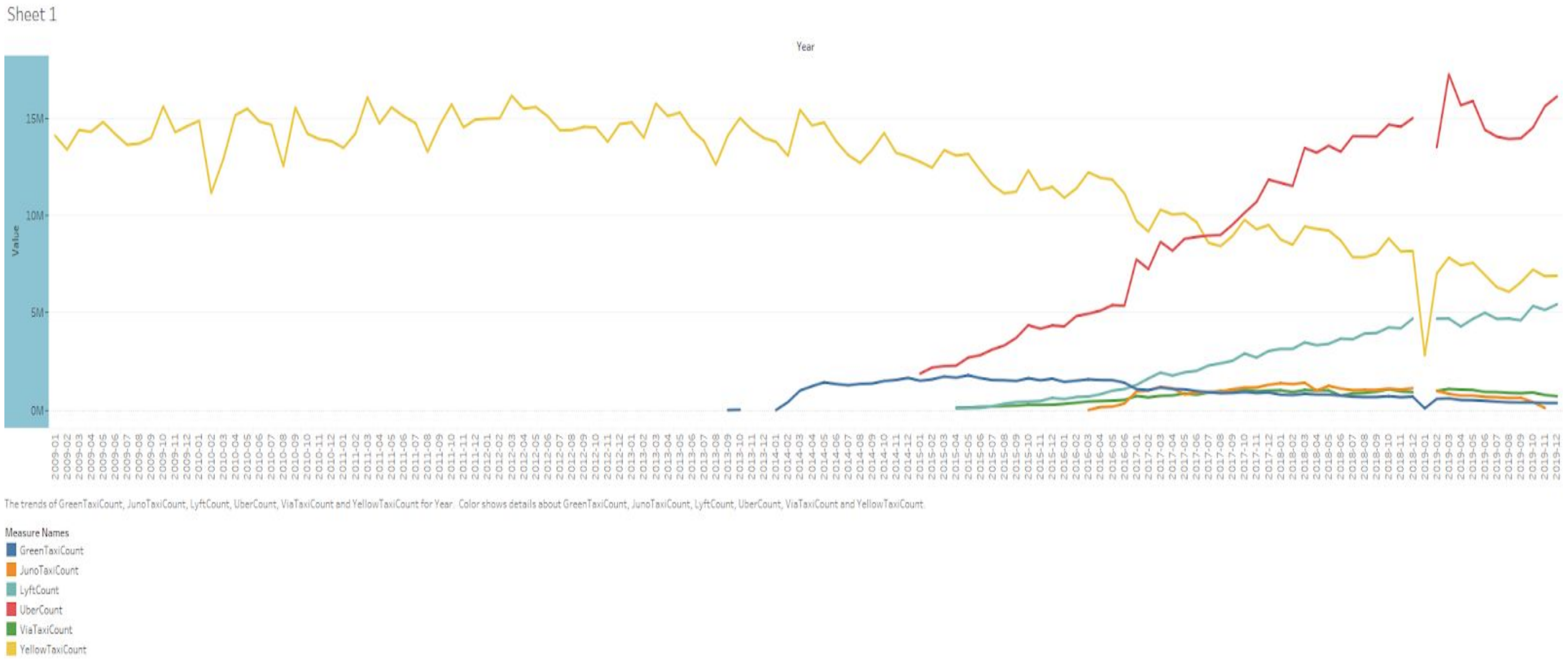

Figure 1: Taxi PickUp Trends

Figure 1 depicts the pickup trends of different types of taxis in different months of the year ranging from January 2009 to December 2019. From the figure we can clearly see that yellow taxis dominate the market initially. Even the introduction of green taxis did not affect the number of pickups made by the yellow taxis. However from the year 2015 we see a downward trend in the pickups for the yellow taxis with a rise in the rides for Uber. The downward trend continues till June 2017, where Uber finally overtakes Yellow taxis in the number of rides made per year. After 2017 there has been a steady rise in the Uber pickups whereas the Yellow taxis as well as the Green taxis have seen a steady decline in the number of rides. Also in the year 2019 Lyft which had overtaken the Green Taxis market share in January 2017, has reached the number of pickups made by Yellow taxi. Looking at the previous trends, Lyft will probably overtake Yellow taxis in 2020. However, the other app based taxis such as Juno and Via don't seem to impact the market as such. This graph shows an interesting change in the behaviour of the people in New York City where people who initially favored traditional taxis are gradually switching to app hailing rides such as Uber and Lyft. This increasing popularity of Uber and Lyft can be attributed to its cheap fare and high availability. In order for the yellow Taxis to make a comeback into the market, they would need to reduce the fair otherwise they would eventually lose their customers as well as drivers to Uber and Lyft.

Figure 2 shows how Yellow taxi gradually lost its market share to Uber. Initially in the year 2015 the market share for Yellow Taxi was 71.6% wherease Uber's was 18.1%. Following this every year the market share for yellow taxis fell by 10%. One interesting insight which can be drawn from the market share is that between the years 2015 and 2016, the customers which majorly shifted to Uber originally used Yellow taxis and not the green taxis since there is a drop in the Yellow taxis market share by 10% whereas Green taxis dropped only by 1%. However with the introduction of lyft, the market share for the traditional based taxis dropped and in the year 2019 Yellow taxi had a market share of mere 34.8% and green taxi became almost non existent with a market of 2% wherease Uber (57%) and Lyft (20%) are taking over the market.

3. Weather Analysis

The shift in New York City's weather conditions is so vast that sometimes quite warm nights in fall are followed by intense snowfall in the morning. Such tilting conditions have an impact on all aspects of the society including taxi services. We attempted to analyse the effect of weather conditions such as snowfall, precipitation, snow depth, airwind etc. on the taxi usage and consequently the average fare collected in harsh conditions. We downloaded the weather data and joined it with taxi services data to initiate the analysis which include the data from 2010-2016. Such huge data helped us form a general and more valid hypothesis rather basing our results on outliers of a few years. It was observed that with the increasing intensities of snowfall, there was an expected decline in the daily pick ups in the city. On the other hand when the daily pickups were compared with precipitation levels in the city,

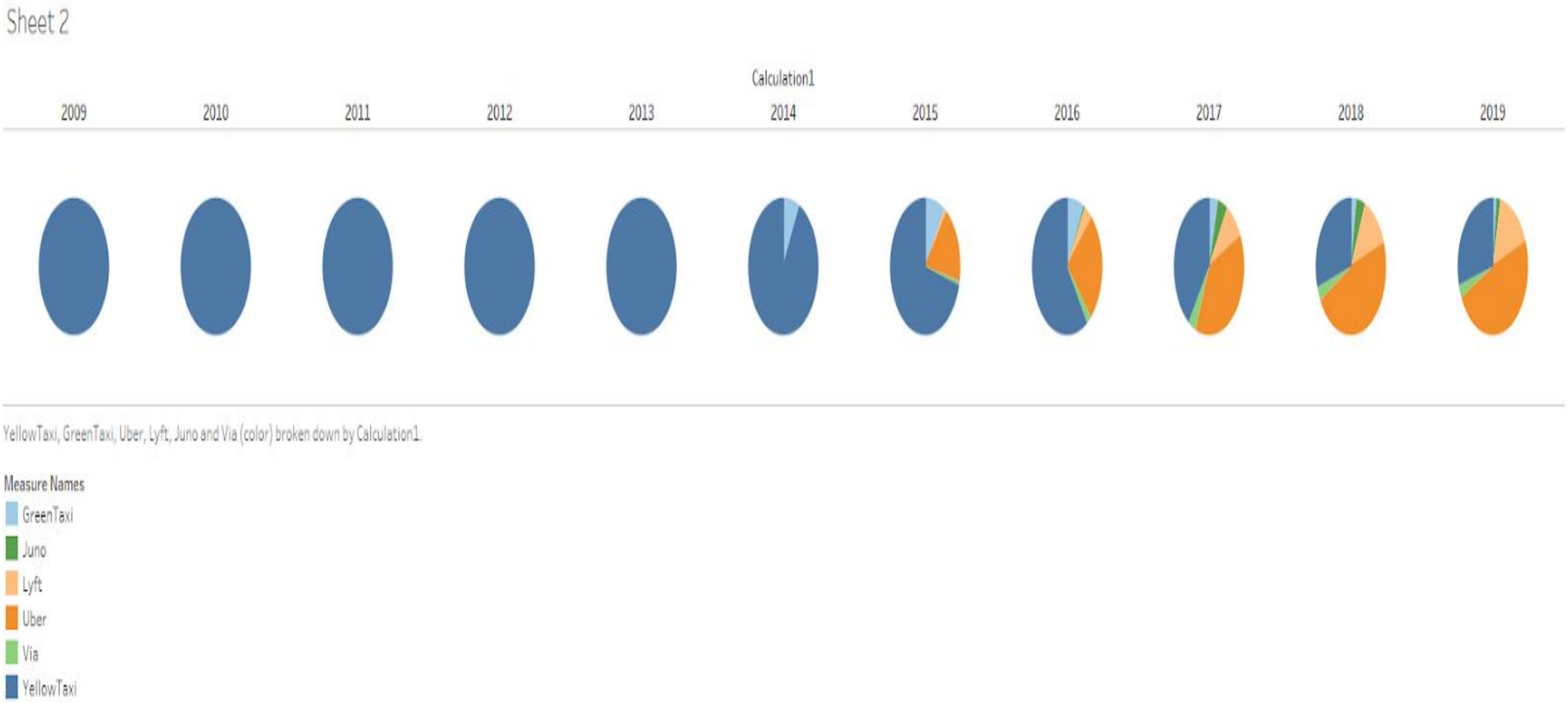

Figure 2 : Market Share

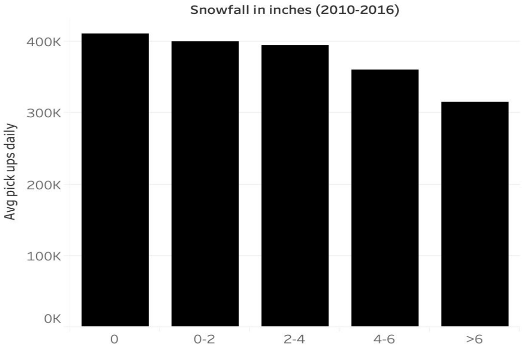

Figure 3: Average daily pick ups with snowfall intensity

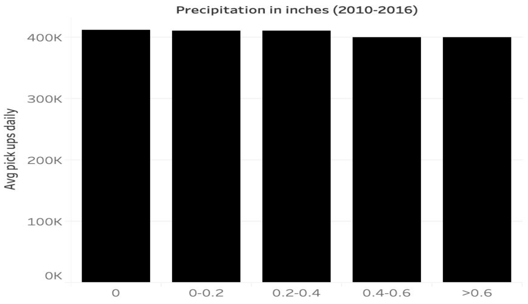

Figure 4:Average daily pickups with precipitation intensity

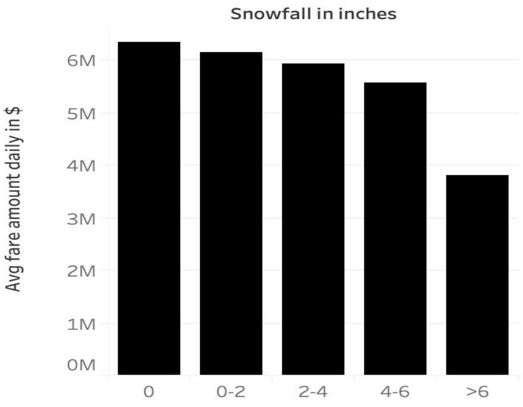

Figure 5: Average daily fare amount with snowfall intensity

level of the bar graphs depicted in Figure 4 almost remains constant. Specifically in Figure 3, we can see that with a relatively lighter snow intensity there is almost the same number of average daily pick ups as compared to precipitation but as soon as the snowfall level intensifies beyond (4-6) or (>6),

there is almost a decline of 25% from the previous pickups. This shows that with extreme snowfall levels people prefer using other means of transportation which include subways when compared to riding a taxi since all transportation means which don’t use roadways are free of snow and thus are preferred. Although Figure 3 also suggests that till a certain degree of snowfall (2-4) level preferred means of transportation is using roadways only as taxi services are well in demand during those times and people are hesitant to use other means of transports maybe due to extra efforts of using a ferry or a subway. This insight was corroborated when we analysed the average daily fare amount collected by taxi services with respect to snowfall intensity levels as depicted in Figure 5. We get a trend similar to Figure 3 where it shows negative correlation between average fares collected and snowfall levels in the city. Likewise to daily pickups trend in Figure 3, there is a constant amount of fare generated during shallow snowfall levels but with the high snowfall level (4-6) there was a steep decline of almost 50% in the fare amount collected during snowfall level (4-6) as compared to snowfall level (2-4). This shows the customer's affinity towards taxi rides until the weather conditions are harsh enough to avoid roadways. The analytics can be used to plan the travels during winters and thus avoid hassle of figuring out the best way to travel during adverse conditions. One can also do cost analysis of the transportation means during harsh weather conditions and thus save some money by optimizing the travelling.

2. Price Analysis

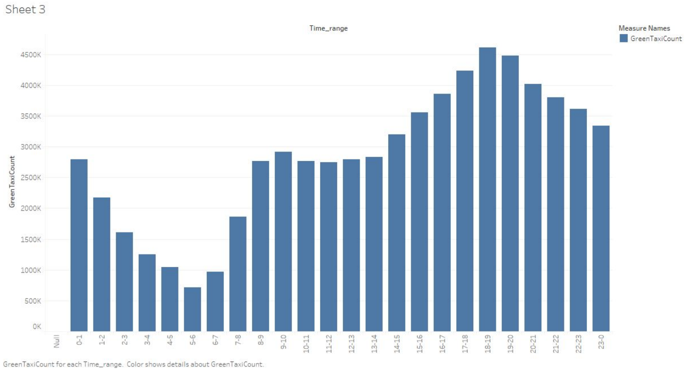

Figure 6: Green Taxi Pickups with time

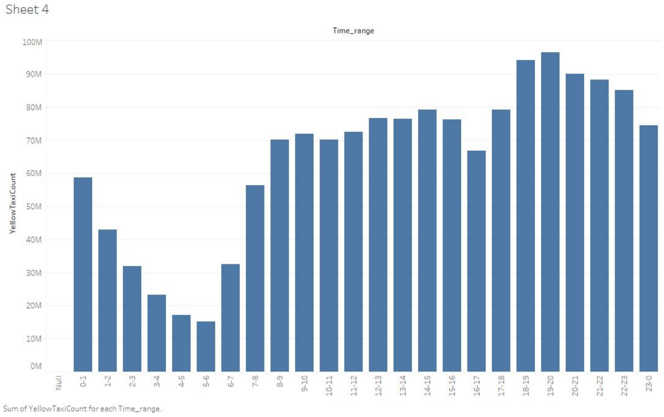

Figure 7: Yellow Taxi pickup with time

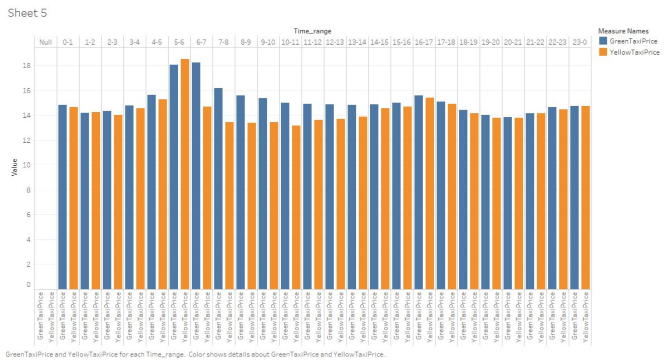

Figure 8: Yellow taxi fare vs Green taxi fare with time

Figure 6 and Figure 7 shows the distribution of pickups with time. From the above two figures it can be clearly seen that there is a surge in the pickups between 6 PM - 8 PM. This can be attributed to the fact that 6 PM - 8 PM is generally the time when people leave their offices to go back to work. One more observation that can be drawn from the bar plots above is that there is a significant drop in the number of pickups in the morning hours. This is supported by the fact that relatively less number of people would take a cab at around 5AM or 6AM. One interesting correlation that can be observed is by combining the observations from figure 6,7 and 5. It can be seen that in the morning both, the yellow taxi and the green taxi charge a relatively higher fare amount than during surge hours. In the morning, the yellow taxi charge around 18.5$ and the green taxi charge around 18$ whereas during the surge hours the fare amount drops by almost 4$ where the yellow taxi and the green taxi charge around 14$ each. This trend shows how with reduced prices, the number of trips for both the taxis increase. As for the customer, almost for all the time ranges yellow taxis seem to be charging less than the green taxis as can be seen in figure 5. This can also be one factor which contributes to the number of pickups being more for the yellow taxis than the green taxis per year. The customer would choose yellow taxis over green because of the less fare and if the green taxis want to increase their customer base they would need to reduce their fares.

Figure 9 shows the relation between the speed of the taxi and the tip given for that trip. The rides where only card was chosen as a payment option were considered for the analytics since many taxi drivers choose not to report any tips when paid in cash, the same fact is defended by a large number of 0 tip trips observed in cash paid trips. Moreover, only the rides where the tip was non zero was taken into consideration. The graph shows a trend in the tipping behaviour of customers. Customers tend to tip high when the speed of the taxi is around 4 miles per hour and tend to decrease as the speed increases. By using a linear regression model in Spark's Mllib we got a relation between the speed and the tip

$$tip \ = \ 16.0097 \ - \ 0.014 * speed$$

A random forest model was also fit on the data to support the same hypothesis. This behavior can be attributed to the fact that the customers recognize the effort put by the drivers during traffic and congestion which is common in NYC which causes the average speed to reduce and pay a higher tip (16%). The customers tend to give a lower tip (15%) amount when the driver is driving on the highways since there is not much difficulty in the trip. From a driver's perspective, in order to maximize earnings they should try to increase the number of tips during surge hour and traffic.

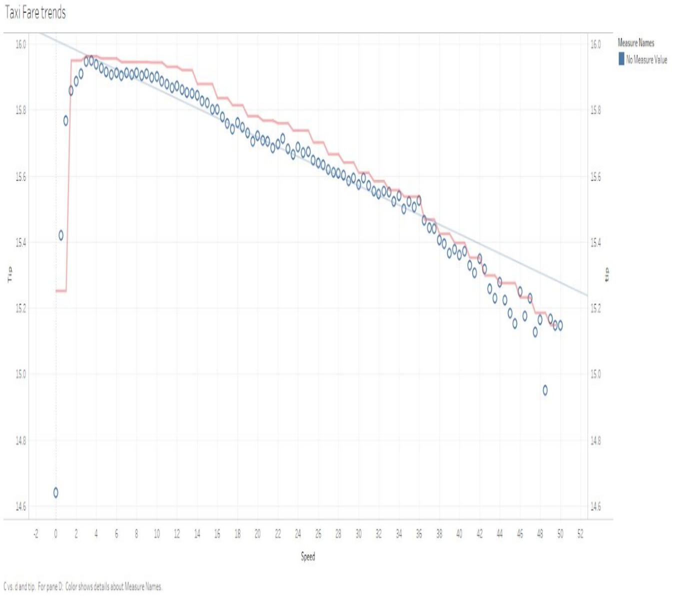

Figure 9: Speed vs Tip

3. Late night pickup analysis

The late night taxi pickup indexes using the heat zone mapping in Figure 10 of the Application Design section has led us to various useful insights. By looking throughout the heat map we can infer which places are the most popular ones during the time 10pm to 5am which we have deemed as late night for this analysis. If we classify the data borough wise , Bushwick in Brooklyn corresponds to the most pickups (~74.5%) during the late night while in Manhattan most of the pickups were dominated in the Lower East side and Meatpacking District. Upon further discerning of the above finding we also found that there is direct correlation between the housing strategies and density of the taxi pickups. According to Propertynest and Wikipedia, places in the Lower East Side in Manhattan and Bushwick in Brooklyn are full of top rated restaurants and have a lively nightlife which directly affects the rent and cost of living for such places. Although rents and cost of living in these areas are high but not as substantially high as in West Village which have amongst the highest average rent in New York City with lesser pick rates. According to Propertynest, West Village mostly comprises expensive family apartments which positively correlates to our finding that West Village has lower density of pickups at night and thus negligible nightlife commotion owing to family places. This data can be useful for people coming to the city from outside and looking for places to live as families won't prefer living in areas with a lot of commotion during the night whereas students will prefer such areas owing to easier food access and affordability in rent as compared to West Village.

VI. Application Design

The below design was implemented keeping in mind the important software design principles and keeping the components as decoupled as possible. Below are the layers in which the application was divided

1. HDFS

HDFS(Hadoop distributed File System) was basically used to store the large datasets. The databases as well as the analyses results were stored in HDFS. Data was accessed and stored in HDFS using Spark APIs.

2. Polling Layer

This layer consists of Scala classes responsible for polling data from respective datasets in HDFS using Spark and passing it to the cleaners in the form of RDDs(Resilient Distributed Dataset). The datasets involve yellow taxi dataset, green taxi dataset, for-hire vehicle dataset, weather dataset and license plate mapping dataset.

3. Cleaning Layer

The Cleaning layer consists of data cleaners for all the datasets which are responsible for cleaning the code and making the datasets uniform following the above mentioned schemas. The cleaning process involves making the data uniform, missing value imputation and fixing the schema for every year (Each year had a different schema) into a uniform schema. The input and output used RDD data structure.

4. Analyzing Layer

This layer used the cleaned up RDDs from the cleaning layer to draw the insights. There are 4 major categories of insights which would be drawn. The Count analyzer would be using the cleaned yellow taxi, green taxi and fhv datasets to analyze the pickup trends and market shares, the price analyzer would be using the same datasets to draw insights related to fare prices, the weather and taxi zone analyzers will draw additional insights related to taxi fare and weather and the above mentioned insight for each NYC Zone. The results from the analyzers will then be stored in the form of Hive tables in HDFS using Spark SQL APIs.

5. Tableau

Tableau will be used for visualising our analytics and insights. Tableau will be connected to NYU's HPC and will have access to the Hive tables generated in the previous layer using Hive thrift servers. Tableau dashboards will be used for plots and heat maps for New York City.

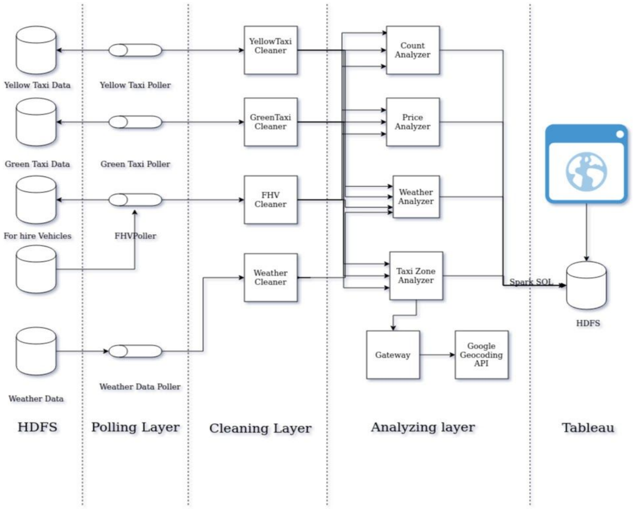

Fig 1 : Design Diagram for the application

6. Visualization Dashboards

Our project will mainly have two types of dashboard. These are mentioned below. These dashboards are created on Tableau.

1. Heat Maps

The heat map will basically show the map of New York City divided into different taxi zones as per the data provided by TLC. Each of the zones will show the name of the zone and the late night pick up percentage that we got through our analytics. For example in figure 10, the zone of Bushwick is shown with a late night percentage of 74.05%.

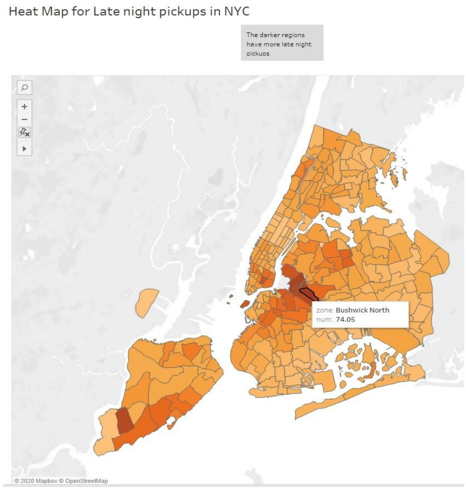
Heat Map for Late night pickups in NYC
The darker regions have more late night pickups
zone: Bushwick North
num: 74.05
© 2020 Mapbox © OpenStreetMap

Figure 10: Heat map of NYC

2. Graph Dashboards

There are three graph dashboards in our project, which are related to the trends drawn from the years 2009-2019. The first dashboard shows the pickup trends for Yellow taxi, green taxi and FHV vehicles along with their market share. The second dashboard shows the relation of weather on taxi pickups and the third dashboard shows the trends of taxi fares with the speed and the pick up count and the time of the day. All the dashboards will look like figure 11.

Figure 11. Dashboard showing taxi fare trend

VII. Analysis

The project was set up using sbt which is an open source build tool for scala and spark projects and is similar to apache maven. Multiple dependencies such as SparkSQL, MLLIB, SparkCSV had to be built in the project. Big data analytics were conducted using Apache spark which is an open source distributed framework for performing big data operations on clusters. Additional libraries like Mllib were used for performing machine learning algorithms like Random forest and Linear regression. HDFS which is an open source distributed storage framework was used for file storage. The entire project was built on the Dumbo cluster which is provided by NYU's HPC. Visualizations were done using Tableau, an interactive data visualization tool.

These were the challenges we faced while developing our project

Datasets for different taxi services such as Yellow taxi services, Green taxi services, For Hire Vehicles (Uber, Lyft, Via, Juno) collected over 10 years posted vast challenges for normalisation.We dealt with filtering data based on several issues such as abnormal average speed, missing values, incorrect latitudes and longitudes, values swapped for different columns, irregular schema within certain months of the same dataset, extremely small journey duration time (order of seconds).

Our final dataset included approximately 300Gbs of Data. While submitting our spark job on the cluster with default parameters the running time of the job exceeded 1 hour. To expedite the running time we tweaked the parameters such as providing more executor and driver memory as well as increasing the number of executors. More memory results in more data processing in-memory thus speeding up the job. To further optimize the running time we increased the number of cores allocated to an executor. The sweet spot for the concurrency comes out to be 5 cores per executor as suggested by Cloudera. This avoids us from facing consequences of fat and tiny executors. It resulted in a high level of parallelism as a single executor was now working on more number of tasks than before. This helped reduce the running time of the job to nearly 30 mins. Deployment on cluster mode helped too as it leads to less network calls as compared to running jobs on client mode. Yet another optimization we did was to write all APIs which used shuffling of data across several worker nodes at the end of chain of commands so as to reduce shuffling during intermediate operations.

Our analysis doesn't yield real time results as we dealt with static data from 2009-2019. A major hurdle we faced was limitations on the number of Geocoding APIs available to us for mapping the latitude and longitudes to the zone Ids for a borough. Beyond a certain threshold (500k) every API call charged a certain fee which summed up to be a huge amount considering we needed to make more than such billion calls. This limited our heat map analysis from datasets corresponding 2017 to 2019 where zone Ids were mentioned in the dataset. Although tableau can generate representations of latitudes and longitudes but those are not as representative as heat maps hence were avoided. Another limitation was lack of rich data values such as fare price, dropoff locations/Ids for Uber, Lyft, Via, Juno dataset which impeded us from doing insightful analytic on FHV dataset.

VIII. Conclusion

Owing to density of population in New York City, the transport system continues to be a vital aspect of shaping the city's ecosystem. Through our extensive studies we have learnt the impact of the taxi services on understanding and predicting the important features of human society.

We found out how Uber dominated the market share in a short duration of time starting 2015, where people who initially favoured the traditional taxi services, gradually switched to app hailing rides such as Uber and Lyft.

Additionally we were also able to figure out how the average speed of the ride affects the tip given to the customer. We processed a huge chunk of data using Big Data analytics and used Machine Learning models to quantify the correlation between the same. Our findings suggest how we can forecast the average tip the driver will get for a journey in NYC.

One of our findings also includes analysis of the effects of adverse weather conditions such as snowfall and precipitation on the taxi services and revenue generated. It was found out that people preferred using roadways until it gets unmanageable with intense snowfall levels.

Finally we also visualised the density of the pick ups by generating a heat map for the city. This was helpful in confirming our hypothesis that family apartments are based out of areas with less commotion, whereas areas such as the Lower East side, Bushwick in Brooklyn have the highest pickup density thus resulting in most vibrant nightlife across the city thus mostly favoring the students.

IX. Future Work

We can use a crime dataset in conjunction with our late night heat map analysis to further categorise places in terms of livilablity by assigning a score to all the zones of NYC. This can be achieved using a deep learning algorithm.

Since the dataset of uber was really unclean and it lacked necessary information for drawing proper comparisons between uber and other taxi services, one extension of the project can be deriving Uber's surge price model using the proper dataset provided by Uber.

If we had more recent data for pick ups during Covid-19 a detailed analysis could have been done over detection of hotspots and sentiments of the drivers during this tough period. The pickup routes of the drivers affected by Covid-19 could have been analyzed to detect which areas could emerge as the next hotspots.

X. Acknowledgment

We would like to thank NYU HPC for providing access to Dumbo and the spark ecosystem which helped us use spark tools to develop our project. We would also like to thank Tableau for providing free access to Tableau Desktop software which was used in our project. We would also like to thank Prof. Suzanne McIntosh for the guidance and support in the course of developing our project.

XI. References

1. T.White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
2. Diego Correa, Kun Xie, Kaan Ozbay Exploring the taxi and Uber demand in New York City: An empirical analysis and spatial modeling Transportation Research Board 96th Annual Meeting Transportation Research Board, 2017
3. NYC Taxi and Limousine Commission Trip Record Data retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
4. Central Park, NY weather data from National Centers For Environmental Information retrieved from https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail
5. "2014 Taxicab Factbook" New York City, n.d. Wed. 20 Oct. 2016
6. Tang, Y.X. Big Data Analytics of Taxi Operations in New York City. American Journal of Operations Research, 9, 192-199.
7. Patel, Umang. (2015). NYC Taxi Trip and Fare Data Analytics using BigData. 10.13140/RG.2.1.3511.0485.
8. Guo, Jing. (2018). Analysis and comparison of Uber, Taxi and Uber requests via Transit. Joint Institute for Computational Science