Big Data Analysis and Prediction of Traffic in Los Angeles

Dalyapraz Dauletbak[1*] and Jongwook Woo[2]
[1]Department of Information Systems,
California State University Los Angeles,
Los Angeles, CA 90032, US
[e-mail: dmanato@calstatela.edu]
[2] Department of Information Systems,
California State University Los Angeles,
Los Angeles, CA 90032, US
[e-mail: jwoo5@exchange.calstatela.edu]
* Corresponding author: Dalyapraz Dauletbak

Received Septebmer 5, 2019; revised November 18, 2019; accepted December 4, 2019; published February 29, 2020

Abstract

The paper explains the method to process, analyze and predict traffic patterns in Los Angeles county using Big Data and Machine Learning. The dataset is used from a popular navigating platform in the USA, which tracks information on the road using connected users' devices and also collects reports shared by the users through the app. The dataset mainly consists of information about traffic jams and traffic incidents reported by users, such as road closure, hazards, accidents. The major contribution of this paper is to give a clear view of how the large-scale road traffic data can be stored and processed using the Big Data system - Hadoop and its ecosystem (Hive). In addition, analysis is explained with the help of visuals using Business Intelligence and prediction with classification machine learning model on the sampled traffic data is presented using Azure ML. The process of modeling, as well as results, are interpreted using metrics: accuracy, precision and recall.

***Keywords*:** Traffic Analysis, Traffic Prediction, Big Data, Machine Learning

A preliminary version of this paper was presented at APIC-IST 2019, and was selected as an outstanding paper. This version contains an improved prediction accuracy.
This study was supported by Isaac Engineering and Oracle Cloud Innovation Accelerator. Oracle Big Data Cloud Service was used in the data analysis.

http://doi.org/10.3837/tiis.2020.02.021 **ISSN : 1976-7277**

1. Introduction

It is known that US Governments turn to Advanced Traffic Management Systems in order to solve traffic congestions and adopt new transport management plans and utilize transport resources [1]. Unfortunately, major cities are still waiting for traffic to be resolved, where Los Angeles is ranked number one city in US with major problems in it [2]. City Departments are interested in improving traffic situation and therefore adopt information from popular navigation platforms in order to understand and analyze current situation. In our case City of Los Angeles provided traffic data set from one of the famous navigating app companies in the US for research purpose. In this paper we have conducted analysis of traffic jams in Los Angeles County area.

Although traffic analysis attracts enough attention of researchers and predicting of traffic patterns is a goal for major businesses, prediction of road traffic can be divided into two major areas: short-term and long-term traffic prediction [3]. Short-term studies aim to predict traffic conditions using the real-time data, while car is driving, developing precise algorithms to capture speed and time for alternative routes and predicting road traffic several minutes to several hours ahead. Whereas, long-term studies dive into historical data and predict behavioral traffic conditions for weeks and months. In this paper we have covered long-term study and prediction of traffic patterns.

2. Related Work

A preliminary version of this paper was presented at the 14th Asia Pacific International Conference on Information Science and Technology (APIC-IST 2019) [4]. The updated version contains an improved traffic prediction accuracy and more detailed analysis of traffic jams with the use additional geo-map tool ArcGIS.

In 2006 the U.S. Department of Transportation launched the Integrated Corridor Management (ICM) initiative to support new technologies that can operate to improve transportation corridor [5]. Such growing interest in traffic prediction systems was launched in order to support traffic operators in city's decision-making tasks. Therefore, several widely used navigation platforms became willing to help government improve traffic by participating in numerous studies.

Traffic for London in collaboration with Google Cloud arranged hackathon of traffic simulation using London traffic dataset, where one of the teams (companies) came with data flow for processing, visualizing and predicting traffic speed in London [6]. Our work adopts Big Data and Machine Learning for data flow of analysis and prediction of traffic patterns in Los Angeles. Our work is different in the way of deliverables, since we are focusing on interactive visuals, depth of information, giving more insights of traffic pattern analysis and prediction of traffic congestions.

Another major navigator, Waze company, has a special program for those who are interested and willing to connect with it for better community - The Connected Citizens Program (CCP), and through such program partners can exchange data with Waze to make data-driven infrastructure decisions and increase the efficiency of incident response [7]. One of the works that is based on Waze company traffic data is available in the form of slides from Summit on Data-Smart Government at Harvard (November 2017) [8]. This study focuses on collaboration of Waze and Louisville City and points out major insights from such partnership. The outcome of this work is analysis of data in the form of animated maps and Excel tables of

hot spot traffic [9,10]. Traffic department of Louisville currently have a sustained flow of data and uses it on a daily basis. However, our work, apart from analysis of traffics, also explains the flow of big data files management, dynamic geo-maps and further prediction of traffic jams using machine learning.

Another study was conducted in New Haven County, Connecticut. In this research GPS data set was gathered from MapMyRun traffic website and further processed and analyzed using R [11]. The author used sampled small data set for analysis, whereas we present a framework that can be applied to bigger data sets. Also, this work concentrates on clustering the condense areas of traffic, however, our work gives insights to traffic patterns with interactive geo-maps and prediction of jams using classification model.

3. Method

3.1 Dataset Specification

The raw dataset, which comprises the details of traffic conditions in Los Angeles County, was provided by Information Technology Agency of Los Angeles City Department for study purposes. The dataset consisted of 5,858 JSON files covering information reported by app users (accidents, jams, road closure etc.) and information captured from users' devices (location, speed, time deviation from original route). This database is not publicly open, and data is shared upon request only, therefore we were authorized to use a portion of the data only, which is of size 1.8 GB and covers nine days (Dec 31, 2017 – Jan 8, 2018). However, the data was captured with millisecond difference and is considered a raw dataset from a navigation app.

After parsing JSON files into readable CSV format, two major files are rendered: *alerts* (information reported by users) and *jams* (information captured from users' devices). Total number of rows (records) for alerts and jams are 2,170,694 and 16,058,236 rows respectively. Since alerts and jams files have different information (one has information reported by app users, such as jams, road closure, hazards, car accidents; and the other has information tracked from users' devices, such as location, speed, time deviation from original route) each was separately cleaned and then exported for further analysis. The workflow is explained in the next section (3.2 Workflow).
After cleaning and removing irrelevant fields, the attributes and metadata of jams, which are generated passively from device's GPS, are the following (**Table 1**):

Table 1. Jams attributes

location_x	X-coordinate of location
location_y	Y-coordinate of location
pub_date	UTC Time of the publication of traffic report
date_pst	Pacific Time of the publication of traffic report
month	Month number of the publication (1-12)
day	Day of the publication (1-31)
hour	Hour of the publication (0-23)
min	Minute of the publication (0-59)
sec	Second of the publication (0-59)
weekday	Day of the week of the publication (Monday - Sunday)
level	Jam level, where 1 – almost no jam and 5 – standstill jam

speed	Driver's captured speed in mph
length	Length of the traffic ahead in the route of user in meters
delay	Time deviation from the original time in seconds

And the attributes and metadata filtered for alerts, which are reported by users, are the following:

Table 2. Alerts attributes

location_x	X-coordinate of location
location_y	Y-coordinate of location
street	Street name
city	City, administrative division of LA County (LA County has up to 88 cities [12])
country	Country (US);
road_type	Road type (Ex: Street, Primary street, Freeway and etc.)
report_description	Small text describing the traffic event written by user
type	Type of reported traffic event (road_closed, jam, accident, hazard)
pub_date	UTC Time of the publication of traffic report
date_pst	Pacific Time of the publication of traffic report
month	Month number of the publication (1-12)
day	Day of the publication (1-31)
hour	Hour of the publication (0-23)
min	Minute of the publication (0-59)
sec	Second of the publication (0-59)
weekday	Day of the week of the publication (Monday - Sunday)

In addition, a summary table was created to portray basic information about traffic in a smaller aggregated table. Summary table can give insights about amount of jams by time, days, and level of the traffic jam.

3.2 Workflow

Firstly, the raw dataset of 5858 JSON flat files are parsed into readable tables before data cleaning. This is be done with Python using Pandas library – "pandas.io.json.json_normalize" [13] and extracted data is then exported in two csv files (alerts and jams). Further, files are uploaded to Hadoop Big Data system. Big Data is defined as non-expensive frameworks, mostly on distributed parallel computing systems, which can store a large-scale data and process it in parallel. A large-scale data means a data of giga-bytes or more, which cannot be processed or expensive using traditional computing systems [14]. Hadoop is one of the popular Big Data platform and Hive is one of ecosystems for Big Data analysis.

HiveQL is used as a querying language to create the tables' schema, clean data, create summary table for analysis and sample dataset for prediction and output the results. Data cleaning was conducted using several techniques such as regular expressions, conditional statements, substrings, joining tables with detailed info, date and time formatting and time conversion from UTC to PST time zone (Pacific time zone). Final tables' schema and metadata, after cleaning and removing irrelevant fields, is explained in the previous section (3.1 Dataset Specification).

Once the output files have been downloaded on local machine, Excel's 3D map, Power BI and ArcGIS can be used to obtain the Geo-Spatial visualization of reported traffic events and traffic jams. Further, the sampled 100,000 rows from jams file (traffic information captured by user's device), which were randomly pulled from the whole dataset are further used for prediction in Microsoft Azure Machine Learning Studio. Traffic jams prediction can be divided into further major steps of uploading the sample dataset, applying data transformation required for accurate modeling, splitting dataset to train a machine learning model and evaluate prediction accuracy. This process will be explained in detail further in this paper (5.1 Machine Learning Flow).

The whole process of date processing shown in the below flowchart (**Fig. 1**).

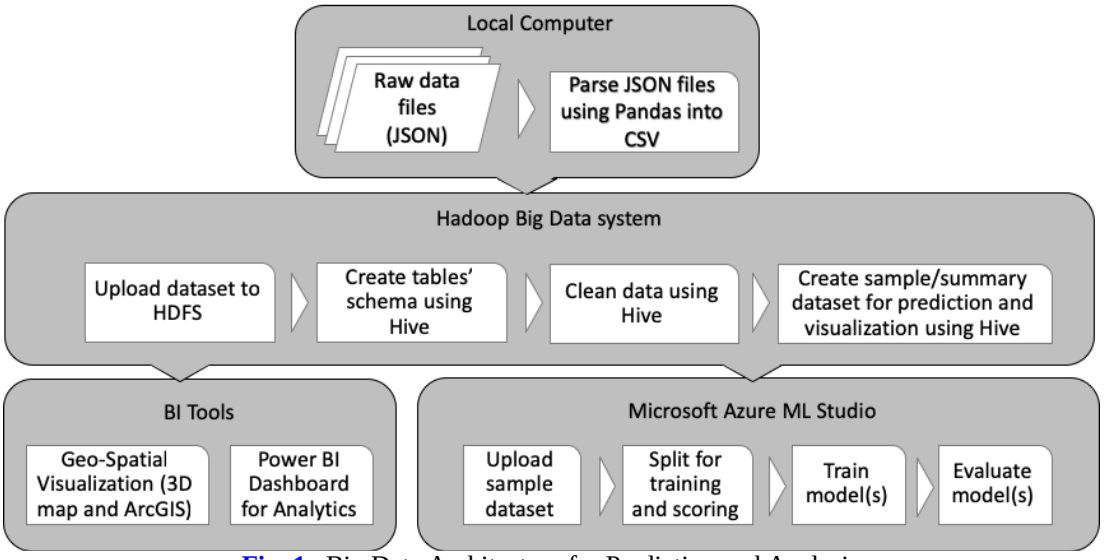

Fig. 1. Big Data Architecture for Prediction and Analysis

The same data processes can be applied to much bigger dataset (as large as 70GB+ annually) as Hadoop system is linearly scalable.

The below table shows the specification for Oracle cluster we were using for our study (**Table 3**):

Table 3. H/W Specification

Number of nodes	6
OCPUs	12
CPU speed	2195.196MHz
Memory	180 GB
Storage	682 GB

4. Analysis and Visualization

We used different interactive visuals in order to show traffic events (including jams) clearly on the map as well as time dependent patterns and different sliced information. After data cleaning and preparation for further analysis, files were extracted into Microsoft Excel, Power BI and ArcGIS.

The geo-map, (**Fig. 2** and **Fig. 3**) was made in add-in Excel tool 3D-map, which can be used for animated map with a timeline. This map shows a sampled day (Friday, Jan 5, 2018) from a full dataset giving an insight into traffic events reported by users and traffic jams captured from the user's device. We used the heat map to show the amount of traffic jams and clustered columns to show the amount of reported accidents (red bar) and reported road closure (yellow bar). By using the time filed, we were able to build up a dynamic geo-map changing over time, showing timeline flow of traffic on a map. Originally, this visualization consists of two 52-second videos, showing the full day span (Friday, Jan 5, 2018) of traffic patterns in LA County and traffic incidents reported by users.

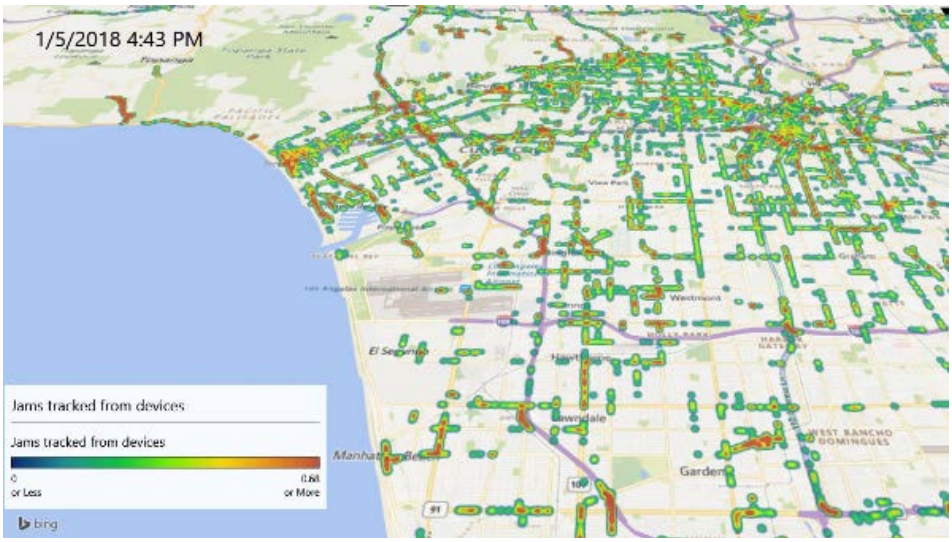

Fig. 2. Jams tracked from users' devices

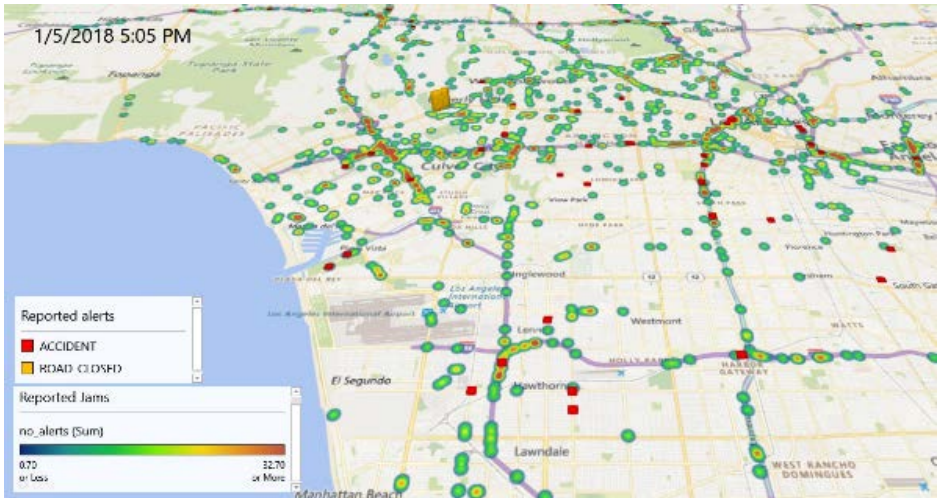

Fig. 3. Jams and other traffic incidents reported by users

From the geo-map of jams (**Fig. 2**), we were able to see condensed traffic on highways - 101, 405, 10; Downtown LA - west area (major concentration of business centers); Santa Monica – area close to pier (tourist place); Beverly Hills – along major streets, such as Santa Monica Blvd. Also, the most condensed traffic hours appeared from 3 pm to 6 pm, although morning hours (7 am - 9 am) are heavy as well, with less intension, this can be also seen on Power BI bar chart (**Fig. 4**). Another interesting insight is huge traffic in Topanga on Topanga Canyon Blvd and Tuna Canyon Blvd.

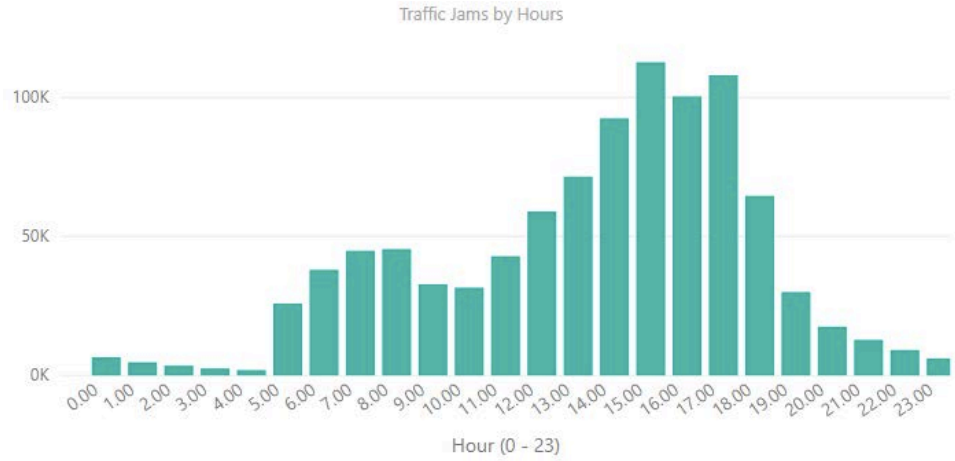

Fig. 4. Traffic Jams by Hours

As you can see both maps (**Fig. 2** and **Fig. 3**) depict traffic situation in Los Angeles at the same time around 5 pm, whereas top figure (**Fig. 2**) shows amount of traffic that platform was able to identify, and bottom one (**Fig. 3**) shows the amount of traffic that users of the app reported. Clearly users tend to report less traffic jams than their devices can capture. However, most condensed traffic seems to be reported in the same pattern.

Let's have a closer view to the specific areas of Los Angeles using ArcGIS tool, which has a map layer with the clear street view. Interestingly, heavy hours around the airport appeared from 5 am to 8 am and from 7 pm to 10 pm. On the map below (**Fig. 5**) we can locate condensed traffic inside the airport LAX and on the highway leading to the airport (Highway 405) from 6:30 am to 7 am (Friday, Jan 5, 2018) that can be explained by numerous flights arriving at this time [15].

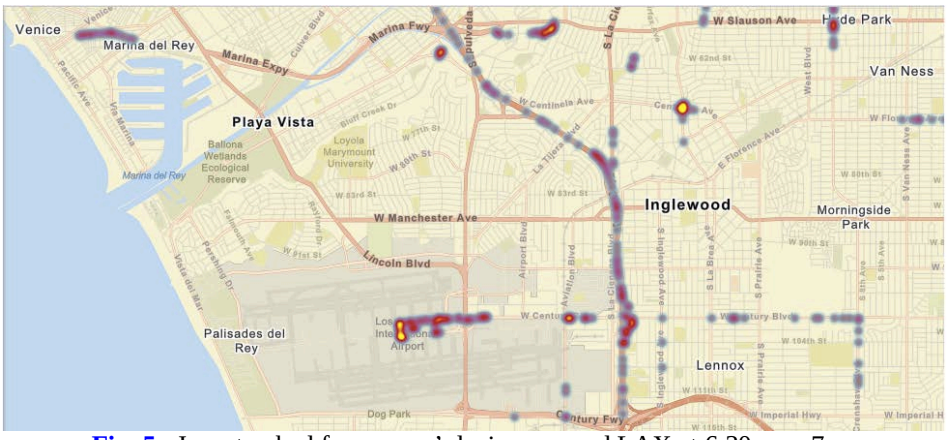

Fig. 5. Jams tracked from users' devices around LAX at 6:30am – 7am

A closer map of jams at 7 am around Down Town LA, specifically on Freeways 101, 110, is captured on the map below (**Fig. 6**). Clearly this can be explained by high density of business centers in DTLA and business day starting at 8 am. Interestingly, there is also traffic in western part on La Cienega Blvd on the entrance to 10th Freeway.

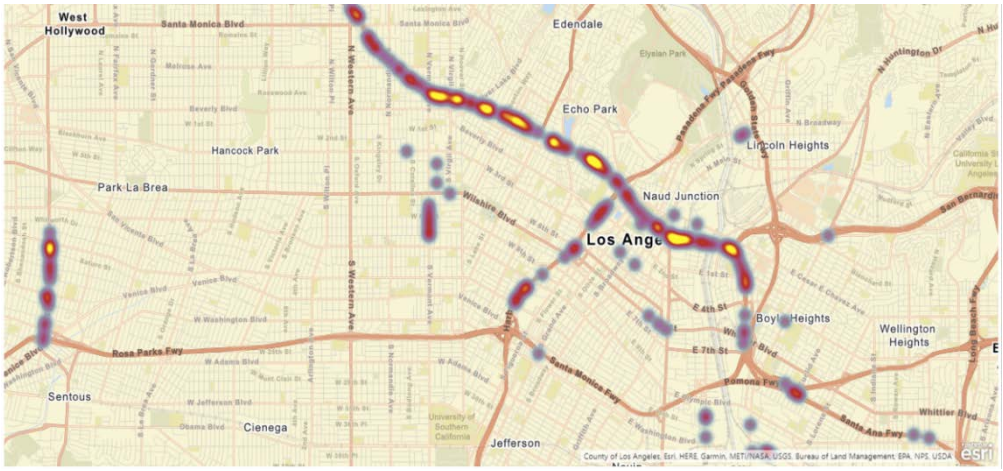

Fig. 6. Jams tracked from users' devices around DTLA a 7:10am – 7:20am

The sectioned line chart (**Fig. 7**) shows percentage portion of traffic jams by days of week. It can be clearly seen that the most congested days are Monday and Friday, while Sunday is least one. Traffic jam report is categorized in five different levels. The chart (**Fig. 8**) shows the count of traffic jams by different levels. Level-1 is considered as almost no jam and based on the data, is barely captured. Level-2 stands for a light jam, which was captured as few as 30% of all jams. Level-3 stands for a moderate jam, with the most portion of traffic jams almost of 50%. Less than 15% consist of Level-4 traffic, which stands for heavy jam and less than 0.5% for Level-5, the standstill jam. So, we can see that our data is slightly skewed to Level – 3 jam, whereas heavy traffic condition is more of an interest, specifically for further prediction, since an accurate prediction of heavy jams helps better planning of city infrastructure.

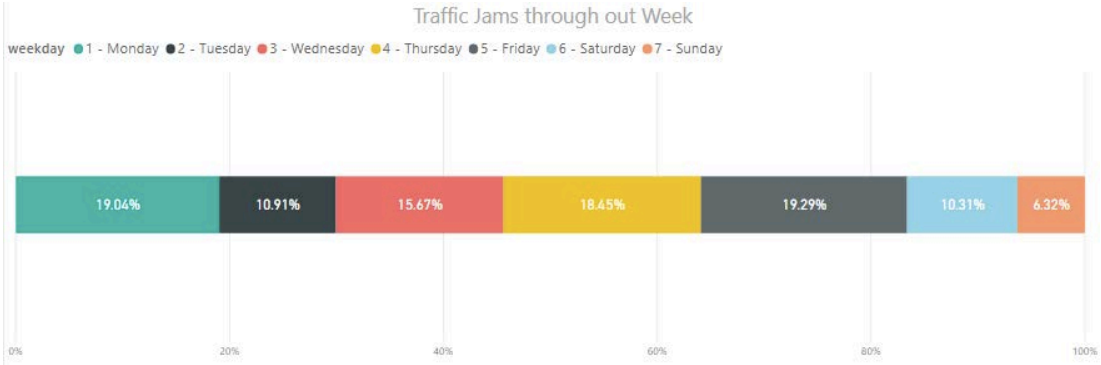

Fig. 7. Traffic Jams by Days of Week

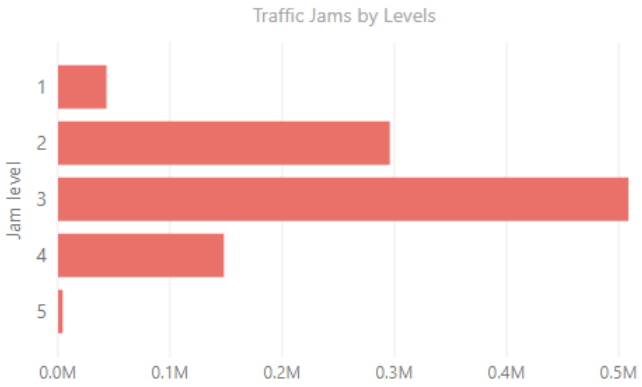

Fig. 8. Traffic Jams by level of traffic

5. Prediction with Machine Learning

5.1 Machine Learning Flow

As mention before, traffic jams dataset, which was passively captured from users' devices GPS, has more than 16 million rows of data. This data is huge for training machine learning model without appropriate high computational speed. Microsoft Azure ML Studio is adopted for predictive analysis and the sampled dataset is uploaded to build a machine learning model, which is a GUI-based integrated environment for constructing Machine Learning workflow [16]. Sampled dataset of size 10 MB (100,000 rows) was randomly selected using HiveQL from HDFS in a csv file format and then uploaded to for further prediction modeling.

The workflow of machine learning process is pictured on **Fig. 9**. Overall, the process is iterative until evaluation result of model is satisfying. The process is explained further in details.

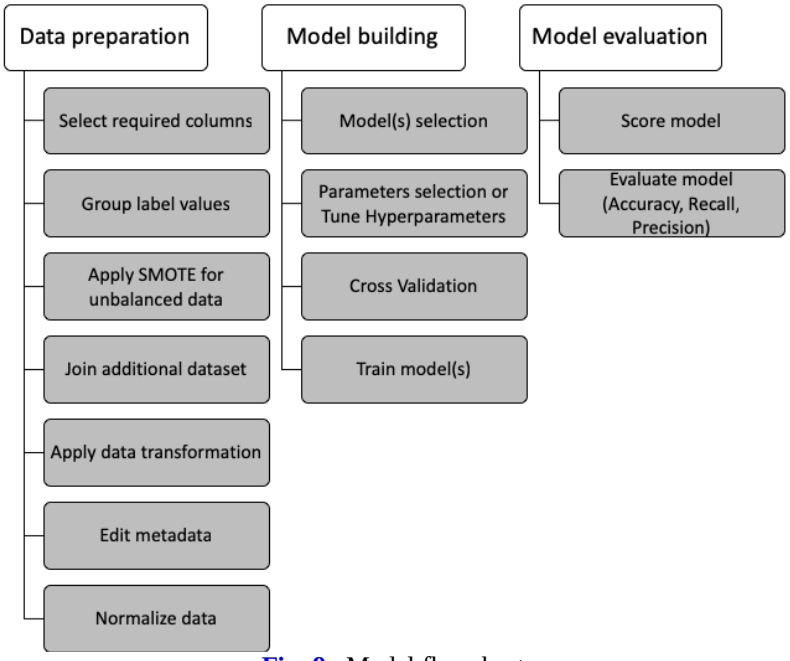

Fig. 9. Model flowchart

5.2 Data preparation

In order to begin modeling, first a label column is selected – **level**, which indicates jam level from 1 (almost no jams) to 5 (stand still jam). This field will be used for classification model building. However, the dataset is critically unbalanced, with the following percentages of a dataset: level 1 – 4.2%; level 2 – 30%; level 3 – 51%; level 4 – 15% and level 5 – 0,44%. This means that based on this dataset our model might fail to predict level 5 (stand still jam), which is very critical in our case, since predicting the heaviest jam is more important than others. In order to balance data, we grouped five categories into three groups: *1 (low)* – jam level of 1 and 2, *2 (medium)* – jam level of 3, *3 (heaviest)* – jam level of 4 and 5. In this case we assume that difference in original *levels 1 and 2* (low levels of jam) is insignificant, as well as for *levels 4 and 5* (high levels of jams).

Although, grouping the categories help to balance data, unfortunately this is not enough, as data is still biased to the medium level, which can affect model's prediction accuracy for other levels, particularly for the highest level. To overcome such imbalance, we used Synthetic Minority Oversampling Technique (SMOTE), which helps statistically increase the number of under-sampled records in a dataset [17]. SMOTE is applied to the newly grouped level – 3 jams (*heaviest*).

Since it is natural to expect less traffic during holidays, additional dataset was joined "national holidays", with the dates for national holidays in US for 2018. And further new fields were created for model: "*is_holiday*" (1 – holiday, 0 - non-holiday) and "*is_weekend*" (1 – weekend, 0 – not a weekend).

Adding the field for depicting the nature of busy hours help to improve the model as well. So, as we concluded from analysis part, we expect rush time to be between 7 am and 9 am, as well as between 3 pm and 6 pm, therefore, a new field "*is_rush*" (1 – if time between 7 - 9 am or 3 – 6 pm) is added.

Such fields as hours, minutes, seconds, weekday number and etc., have a nature of cyclical behavior, which means that numbers do not increment by one in every case and neighborhood of 0 to the maximum value is common. For example, 23rd hour incremented by one gives 0 or after 59th second comes 1st. Such cyclical nature of the field should be transformed to the appropriate representation. This can be done by converting features from Polar coordinate system to Cartesian, applying trigonometric functions. For such cyclical field there is two fields of x and y:

$$\begin{aligned} x &= \sin\varphi \\ y &= \cos\varphi, \\ where\ \varphi &= k\frac{2\pi}{n} \end{aligned} \tag{1}$$

where k is the original value of the field and n is a number of possible values in the field, assuming that all the values are discrete.

Example*:* *hour* field (0-23) transforms to *SIN(hour*(2*PI()/24))* as *sin_hour* and *COS(hour*(2*PI()/24))* as *cos_hour*.

Last step in data preparation is normalizing data using MinMax method, in order to rescale numerical data to one range.

5.3 Model Building

This paper is aimed to predict the appearance of three different levels of traffic jam (1-3) and clearly multi-class classification model is a good fit in this case. Azure ML has some of it: Multiclass Logistic Regression, Multiclass Decision Jungle and Multiclass Decision Forest,

with the best performance in our case. It has to be mentioned that Decision Jungle and Decision Forest models are expected to be a good fit for our case, since both are based on assembled decision tree algorithm and are appropriate for classification with non-linear parameters, however Decision Jungle is compact and powerful discriminative model for classification, whereas Decision Forest has no limitation on paths and depths of tree structure. As for Logistic Regression, we expect it to run with severely less cost (time), however with optimal accuracy, since it does not require any linearity in the variables. We might expect tuning the L1 regularization weight and L2 regularization weight, since probability threshold might be not clearly established.

Model training is conducted after dataset is split into training set and testing set, we chose 70% and 30% of set respectively. After several iterations of model training/testing and by calculating the weight of various columns, we excluded columns that have no value for traffic jams prediction and do not improve the model's performance. We also used Cross Validation and Tune Model Hyperparameters, which helps determine the optimum parameters for selected model. By evaluating performance of several multiclassification models - the best model with the highest measures of performance is Multiclass Decision Forest with the following parameters: *Number of decision trees – 50, Maximum depth of the decision trees – 32, Number of random splits per node – 300 and Minimum number of samples per leaf node – 1*.

5.3 Model Evaluation

There are several metrics to evaluate performance of the multiclass classification model as follows [18]:

- *Classification Accuracy* (overall and average) - percentage of total records classified correctly, which can be defined by the following ratio:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

  where TP, FN, FP and TN represent the number of True Positives, False Negatives, False Positives and True Negatives, respectively.

- *Precision/Sensitivity* (micro and macro) – ratio of correctly identified records as positive out of total records identified as positive:

$$\frac{TP}{TP + FP}$$

- *Recall* (micro and macro) – ratio of correctly identified records as positive out of total actual positives:

$$\frac{TP}{TP + FN}$$

Since our data has imbalance issue the best metrics for model validation in our case are micro-averaged methods, where separate true positives and false negatives are summed up for different sets and then applied for accuracy and recall calculations [19].

An improved Multiclass Decision Forest model predicts the class of traffic jam with Micro-average Recall of *0.692109*. Confusion matrix (**Tables 4**) shows that model performs best at predicting 3rd class of label, the heaviest traffic jam level, with the accuracy of 78.4%.

Although model is less successful at predicting the lowest class of traffic jam with the same accuracy (accuracy 56.5%), it still has a low critical false classifying into the highest level of jam, which is only 6.8%. This means that model is good at predicting the heaviest traffic jams and probability of misclassifying any actual heavy congestions into low level is insignificant as well as probability of misclassifying any no-jam (low level) records into heavy traffic. (**Tables 4** and **Table 5**).

Overall micro-averaged accuracy of the model is *79.4%*, which proves that our model has almost 80% chance of predicting the right traffic in Los Angeles. And Recall of *69.2%* proves that correctly identified traffic patterns were almost *70%* of the actual traffic congestion.

Table 4. Confusion Matrix

		Predicted Class		
		1	2	3
Actual class	1	**56.5%**	36.7%	6.8%
	2	16.6%	**69.7%**	13.7%
	3	4.7%	16.9%	**78.4%**

Table 5. Model Metrics

	Macro - averaged (overall)	**Micro-averaged**
Accuracy	0.692109	**0.794739**
Precision	0.691629	0.692109
Recall	0.681826	**0.692109**

6. Conclusion

From the above we can conclude that our study has revealed several insights of traffic pattern in Los Angeles County. Most of the traffic is condensed on highways/freeways and the busiest are highways 101, 405, 10. Although morning rush hours from 7 am to 9 am produce a lot of traffic, the heaviest traffic time starts from 3pm and gets better after 6pm. Major areas of traffic are: Downtown Los Angeles, Santa Monica, Hollywood, and highways. There are also traffic congestions observed near LAX airport from 5 am - 7 am and after business hours at 7 pm - 10 pm.

We can also conclude that traffic jam prediction is possible in a long-term perspective. Location, date and time can be useful in order to classify the existence of a jam in LA County. Prediction can be performed using machine learning algorithm, multi-class classification model - Decision Forest. The accuracy of traffic prediction is 78.4% for the heaviest traffic jam and overall accuracy of the model is 79.4%, which is an improved performance from the previous version of this work, that was presented during the conference APIC-IST 2019.

In this paper we presented Big Data platform and architecture that allows storing and analyzing giga-bytes of data set – possibly more datasets as the system is linearly scalable. From the available data in Hadoop, which is limited to few days, we were able to provide an interactive tool for analysis, data manipulation and data prediction. Further work can be done with bigger dataset and more classification models in order to find more insights and create a data driven conclusions on LA County traffic situation by using this framework.

References

[1] J. Barbaresso, G. Cordahi, D. Garcia et al., "USDOT's Intelligent Transportation Systems (ITS) ITS Strategic Plan 2015- 2019," 2014.

[2] L. Abrams, City Watch "Traffic Congestion in Los Angeles will get Worse," https://www.citywatchla.com/index.php/2016-01-01-13-17-00/los-angeles/17537-traffic-congestion-in-los-angeles-will-get-worse. Accessed September 3, 2019

[3] M. Heiskala, J. Jokinen, and M. Tinnilä, "Crowdsensing-based transportation services — An analysis from business model and sustainability viewpoints," *Research in Transportation Business & Management*, Vol 18, pp. 38-48, 2016. Article (CrossRef Link).

[4] D. Dauletbak, J. Woo, "Traffic Data Analysis and Prediction using Big Data," in *Proc. of KSII The 14th Asia Pacific International Conference on Information Science and Technology (APIC-IST) 2019*, pp127-133, 2019.

[5] "Integrated Corridor Management," *Intelligent Transportation Systems - Integrated Corridor Management*, www.its.dot.gov/research_archives/icms/. Accessed April 14, 2019.

[6] J. Kestelyn, "Real-Time Data Visualization and Machine Learning for London Traffic Analysis," *Google Cloud*, 2016, cloud.google.com/blog/products/gcp/real-time-data-visualization-and-machine-learning-for-london-traffic-analysis. Accessed April 14, 2019.

[7] "Connected Citizens by Waze," *Waze*, www.waze.com/ccp. Accessed April 14, 2019.

[8] M. Schnuerle, "Louisville and Waze: Applying Mobility Data in Cities," *Harvard Civic Analytics Network Summit on Data-Smart Government*, 2017.

[9] Louisville Metro. "Thunder Jams, 2017 Traffic Delays," *CARTO*, louisvillemetro-ms.carto.com/builder/d98732d0-1f6a-4db2-9f8a-e58026bf0d39/embed. Accessed April 14, 2019.

[10] Louisville Metro, "Pothole Animation," *CARTO*, cdolabs-admin.carto.com/builder/a80f62bf-98e1-4591-8354-acfa8e51a8de/embed. Accessed April 14, 2019.

[11] E. Necula, "Analyzing Traffic Patterns on Street Segments Based on GPS Data Using R," *Transportation Research Procedia*, Vol. 10, pp. 276–285, 2015. Article (CrossRef Link).

[12] United States, Chief Executive Office County of Los Angeles, "Cities within the County of Los Angeles," *lacounty.gov*. Accessed April 14, 2019.

[13] "Pandas.io.json.json_normalize," *Pandas.io.json.json_normalize - Pandas 0.24.2 Documentation*, pandas.pydata.org/pandas-docs/stable/reference/api/pandas.io.json.json_normalize.html. Accessed April 14, 2019.

[14] J. Woo and Y. Xu, "Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing," in *Proc. of International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), Las Vegas.*, 2011.

[15] The Wall Street Journal, "Los Angeles International Airport: Insiders' Tips for Flights to and From LAX," *Travel*, https://www.wsj.com/articles/los-angeles-international-airport-insiders-tips-for-flights-to-and-from-lax-1542204004. Accessed September 3, 2019

[16] Garyericson, "What Is - Azure Machine Learning Studio," *Microsoft Docs*, docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio. Accessed April 14, 2019.

[17] N.V. Chawla, et al., "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002. Article (CrossRef Link).

[18] A. Tharwat, "Classification Assessment Methods," *Applied Computing and Informatics*, 2018. Article (CrossRef Link).

[19] M. Sokolova and L. Guy, "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing & Management*, Vol. 45. No. 4, pp. 427–437, 2009. Article (CrossRef Link).

Dalyapraz Dauletbak received her bachelor's degree in Mathematics from Nazarbayev University and master's degree in Information Systems from California State University, Los Angeles. She is a member of Big Data AI Center (BigDAI): High Performance Information Computing Center (HiPIC) at California State University, Los Angeles; She Loves Data in Los Angeles; International Data Engineering and Science Association (IDEAS). She has data analysis and consulting experience at KPMG. She has received 2019 Analytics Challenge Honorable Mention at Teradata Universe Conference.

Dr. Jongwook Woo received his Ph.D. from USC and went to Yonsei University. He is a Professor at CIS Department of California State University Los Angeles and has served as a Technical Advisor/co-founder of Big Data AI Center at Isaac Engineering, Council Member of IBM Spark Technology Center and as a president at KSEA-SC. He has consulted companies in Hollywood: CitySearch, ARM, E!, Warner Bros, SBC Interactive. He published more than 70 papers and his research interests include Big Data Analysis and Prediction. He has been awarded Teradata TUN faculty Scholarship and received grants Amazon, IBM, Oracle, MicroSoft, DataBricks, Cloudera, Hortonworks, SAS, QlikView, Tableau. He is a founder of Hemosoo Inc and The Big Link.