Waze Traffic Patterns in LA County

Using Hadoop

Authors: Dalyapraz Dauletbak, Wenji Hu, Ketki Vyas, Ritika Joshi

Department of Information Systems, California State University Los Angeles

CIS5200-03 System Analysis and Design

[dmanato@calstatela.edu](mailto:dmanato@calstatela.edu), [whu4@calstatela.edu](mailto:whu4@calstatela.edu), [kvyas2@calstatela.edu](mailto:kvyas2@calstatela.edu), [rjoshi5@calstatela.edu](mailto:rjoshi5@calstatela.edu),

**Abstract:** The paper explains the method and process used for Waze traffic data manipulation and further analysis. The major concentration of the project is to give a clear flow of handling big data files and data cleaning processes using Hadoop and Hive. In addition to that, analysis of this data is conducted using Excel and Power BI, depicting visuals such as maps, timeline and charts on traffic conditions of LA County.Due to limited access to the data, this paper gives a method and a prototype model of analysis on a portion of data, but more insights can be found using full dataset (100GB+) using the same flow of work.

**1. Introduction**

This project uses Hadoop and Hive to keep and process Waze traffic dataset. The dataset is mainly consists of information about jams, traffic incidents, road closure, hazards and etc., reported by users of Waze.

We have chosen this dataset because it reflects the traffic problems in Los Angeles County, which we face in our everyday lives. Since Waze is a popular platform for tracking information on the road using user’s device information and reports written by users, City Departments are interested to use such information in order to understand and improve traffics.[1] Therefore Waze has a lot of partners across the country, including City of Los Angeles, who generously provided portion of data to us upon request from CSULA faculty member, Dr. Jongwook Woo.

**2. Related Work**

Although Waze is a popular application used on roads, there are quite a few works publicly available based on data from Waze. Waze 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 reduce traffic congestion, make data-driven infrastructure decisions and increase the efficiency of incident response.[2]

One of the works that is based on Waze traffic data is available in the form of slides from Summit on Data-Smart Government at Harvard (November, 2017) [3]. 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 analysis [4][5][6]. Obviously, traffic department of Louisville currently have a sustained flow of data and use it on a daily basis, whereas in our work, even if the dataset has the same structure, we were limited to a small portion of data. However our work, apart from analysis of traffics, also explains the flow of big data files management and further interactive map construction using free tools such as Excel and Power BI.

Another work was presented during I-95 Corridor Coalition Summit, an alliance of transportation agencies [7]. During this summit in September 2017, Transcom present their work of Integrating Waze data with Public Agencies’ Data and their work is available in the form of slides online [8]. The work is targeted to explain the data agreement with Waze and also shows some statistics from the data such as amount of jams and alerts for one specific day in New Jersey (Tuesday, 12/6/2016). This work also fundamentally explains the structure of the data and the meaning behind each alert type or jam level. Our work is different from this in the way of deliverables, since we are focusing on interactive visuals and depth of information, giving more insights of traffic pattern and hourly based analysis.

**3. Specifications**

The dataset comprises of the traffic reports related to alerts, jams, and irregularities. Since Waze database is not publicly open and data is shared upon request only, we were authorized to use a portion of the data only. The dataset is of the size 26MB and covers several days in December (Dec-11, 2017 – Dec-13, 2017) in the file of irregularities and one day in January (Jan-3, 2018) in the file of alerts. Although the data is not of a big size, we assume that the same data processes can be applied to bigger dataset (as large as 100GB+). Table 1 shows files and size of the files from dataset.

Table Data Specification

|  |  |
| --- | --- |
| Data Set | Size (Total 26MB) |
| irregularities\_head | 19138 KB |
| jams\_head[[1]](#footnote-2) | 2024 KB |
| alerts\_head | 1999 KB |

The below table show the specification for Oracle cluster we are using and Hadoop specification for our project.

Table H/W Specification

|  |  |
| --- | --- |
| Number of nodes | 5 |
| OCPUs | 10 |
| CPU speed | 2195.084 MHz |
| Memory | 150 GB |
| Storage | 678 GB |
| HDFS Capacity | 147 GB |

**4. Implementation Flowchart**

Initially, the raw dataset, which comprises the detail of traffic conditions on specific days calls to analysis, was downloaded from a trusted source. The whole process of date manipulation is shown in the below flowchart (Figure 1).There are three data logs in csv format were uploaded to the Hadoop File System. After that, HiveQL is used as querying language to create the tables’ schema, clean data, create a summary table and export the results. Once the output file has been downloaded and opened in Excel format, we use Excel’s 3D map and Power BI to obtain the visualizations.

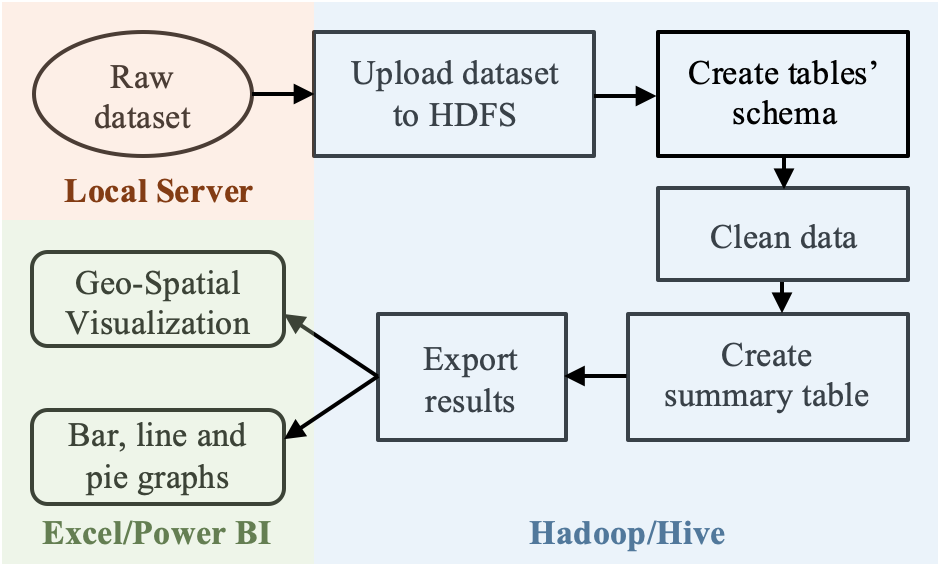


Figure - Implementation Flowchart

**5. Data Cleaning**

Raw files were uploaded and stored in HDFS and then loaded into tables using Beeline Client. Since all files have different information, each was separately cleaned and then exported for further analysis. Data cleaning was conducted using different techniques such as regular expressions, if statements, substrings and joining tables.

Apart from cleaning, since some fields such as “road\_type” or “is\_highway” had only numerical values, which are difficult to process, we created a new table for clarifications of such numerical types (Ex: 1 – Street, 2 – Primary street, 3 – Freeway and etc.). Below is one peace of the code used for cleaning data. Regular expression was used for parsing latitude and latitude from unformatted text. Below is an example of such use:

REGEXP\_EXTRACT(line\_x1,'([-+]?\\d+(\\.\\d+))') as x1,

REGEXP\_EXTRACT(line\_y1,'([-+]?\\d+(\\.\\d+))') as y1,

REGEXP\_EXTRACT(line\_x2,'([-+]?\\d+(\\.\\d+))') as x2,

REGEXP\_EXTRACT(line\_y2,'([-+]?\\d+(\\.\\d+))') as y2

In addition to that we created a summary table to portray basic information about traffic in a smaller aggregated table. Summary table can give insights about number of alerts by types of alerts on a specific date and time in the specific city of LA County.

**6. Analysis and Visualization**

After data cleaning and preparation for further analysis, files were extracted into Excel and Power BI. We used different interactive maps in order to show jams and alerts clearly on the map as well as time line and different charts.

**6.1 3D Map in Excel**

The first visualization (Figure 2), a 3D map, was made in Excel and it is an animated map with a timeline for one day, since the dataset had full cover of one day in January 2018. We used the heat map to show the count of jams, and clustered columns to show the count of accidents (green bar), road closure (blue bar) and weather hazard (yellow cones). By using the time column, we were able to come up with a map with a time flow. Initially, this visualization was in a video format, by playing the video, it is clear to see that the bars grow faster after 5:00 pm of the day, which means that mostly traffic reports are made during that time. Note that weather hazards alerts are almost not seen, since in LA weather is mostly calm.

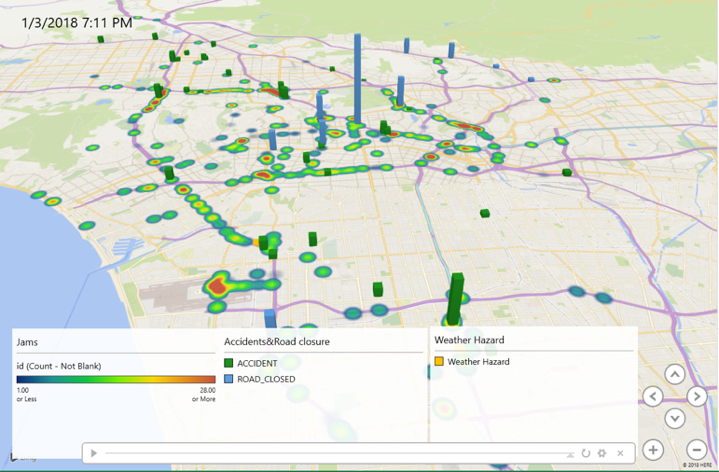


Figure - Counts of Incidents in Excel 3D Map

**6.2 Power BI**

The bubble map from Power BI has been used in the next map (Figure 3), and the location field “City” is being used for geo-coding. The map clearly states that the city of Los Angeles[[2]](#footnote-3) is a top city by reported traffic problems.

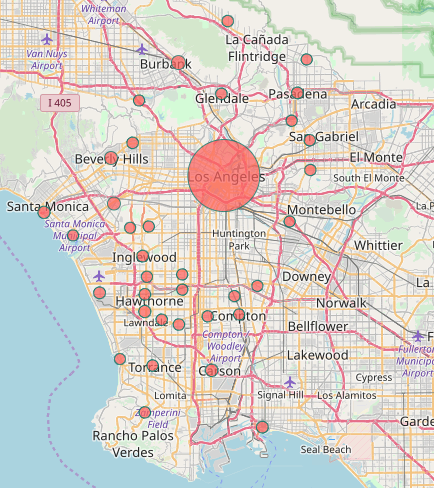


Figure - Counts of Alerts by City

From the time-bar chart shown in Figure 4, we see that the peak-hour of alerts is from 6:00 pm to 7:00 pm and generally, rush hour starts from 4:00 pm and getting better after 8:00 pm.

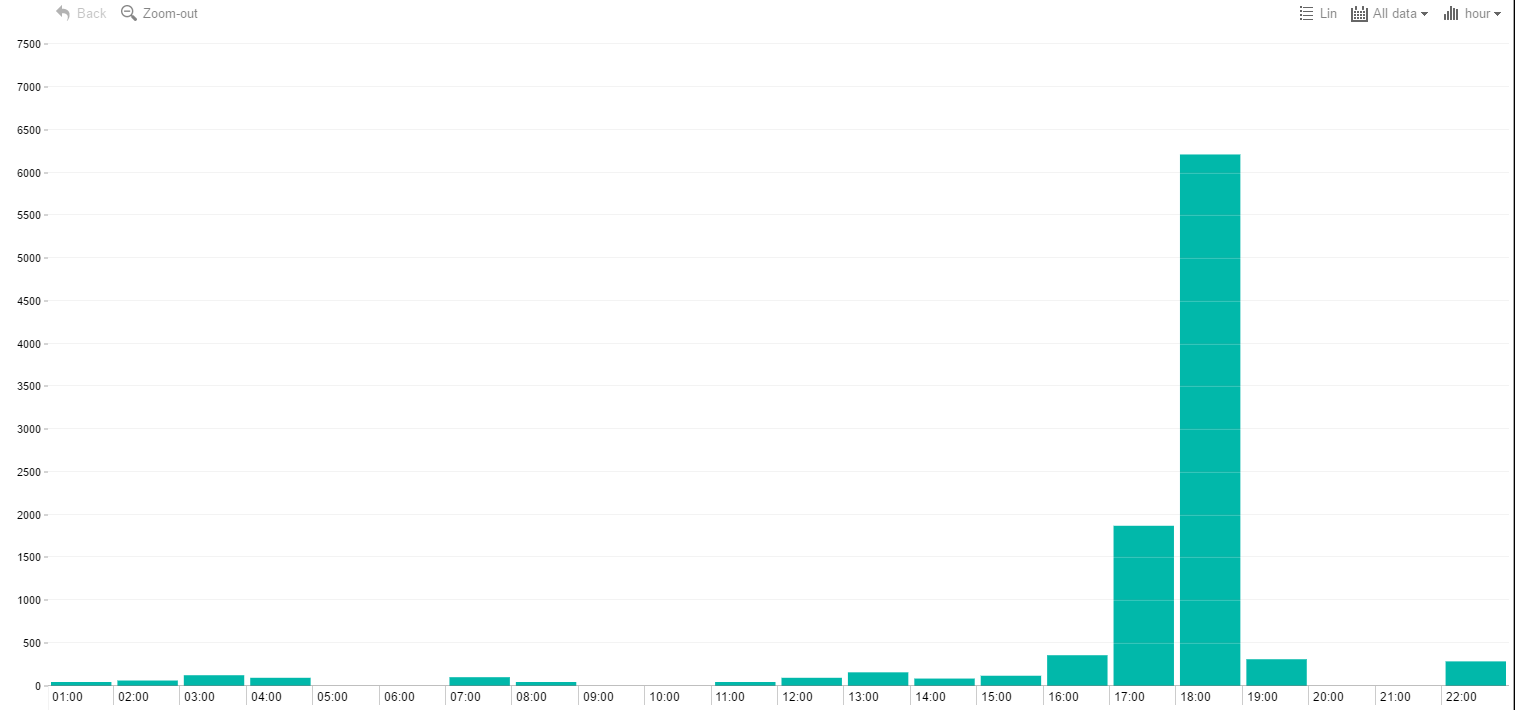


Figure - Number of Alerts by Time

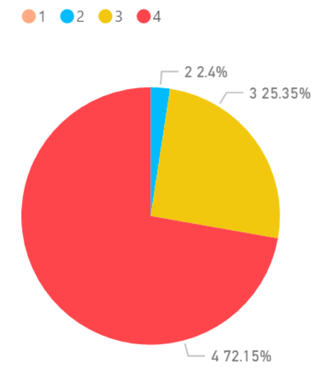
A traffic jam report is categorized in four different levels. The pie chart (Figure 5) shows the count of traffic jams by different levels. Level-1 is the lightest jam and based on the data, is barely reported. Level-2 stands for a moderate jam, which was reported as few as 2.4% of all jams. Level-3 stands for a heavy jam, with the second most percentage at 25.35%. The greatest portion of traffic jams (72.15%) was of Level-4, the standstill jam.

Figure - Amount of Traffic Jams by Levels

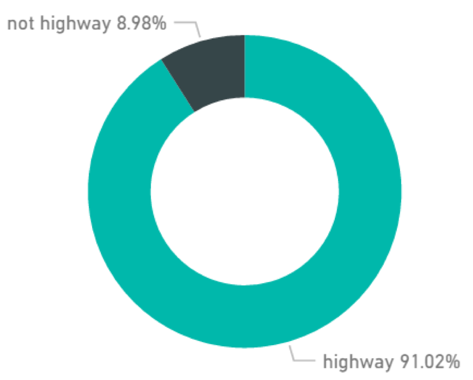
From the next chart (Figure 6) we can see that there is an extreme variation between amounts of traffic jams on highways and not highways. Traffic is mostly concentrated on highways, 91.02% and much less on roads - only ~9%.

Figure - Amount of Traffic Jams by Street Types

The next visualization (Figure 7) is an interactive flow map showing routes of traffic jams on the map. The routes of traffic have 4 different colors for level of the jam: Red for 4 - Huge, Yellow for 3 – High, Blue for 2 – Medium and Green for 1 – Low. We used one pair of longitude and latitude to fetch the traffic jam start location, and another pair for the jam end location.

Among the various areas/cities in Los Angeles, the most concentrated areas with reported traffic jams are Downtown Los Angeles, Santa Monica, and Hollywood with huge (red) and high (yellow) traffic jam level.

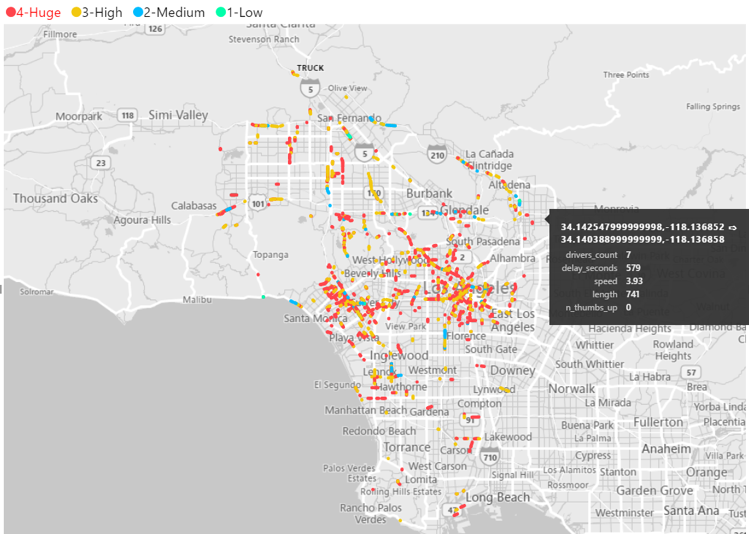
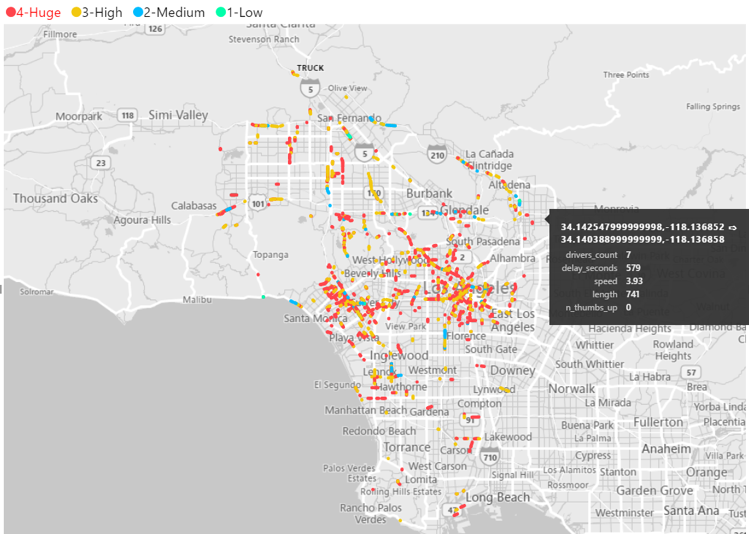


Figure - Traffic Jam Patterns in LA County

The flow map is an interactive map with a ability to filter different levels of jams or see detailed information by pointing, such as drivers count on a specific jam, delay of the traffic in seconds, average speed of a traffic, length of the jam and thumbs up (likes) by users.

Further down on Figures 8 you can see a close up of the map in the area of Santa Monica and clearly heavy traffic on Santa Monica Boulevard and 405 Highway.

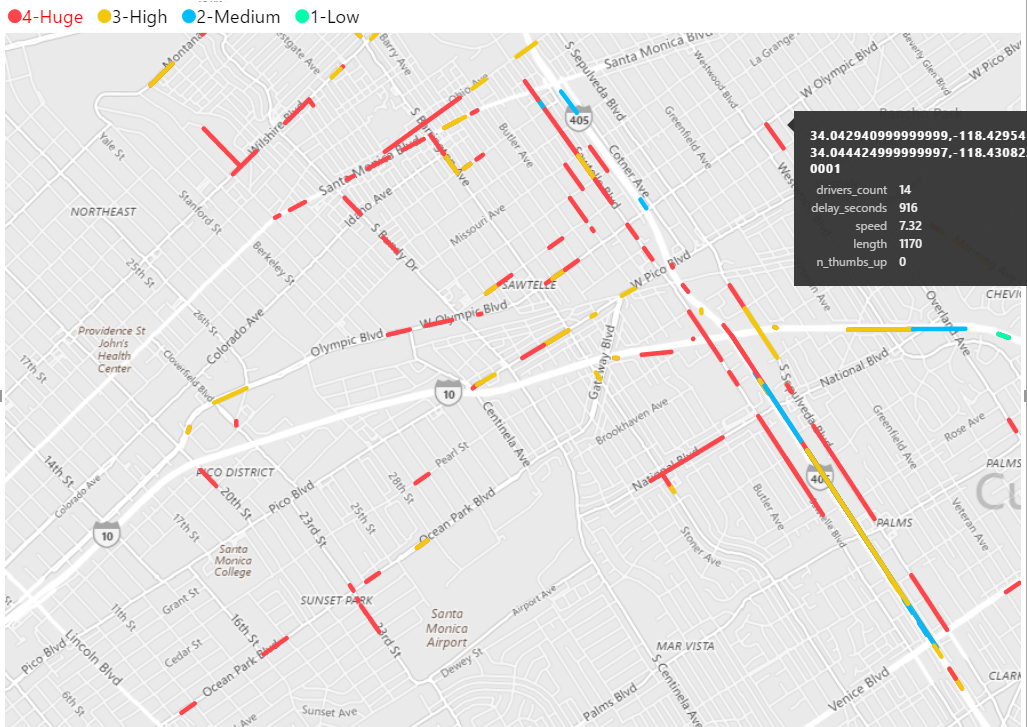


Figure – Close up to Santa Monica

The next maps (Figure 9 -10) are filtered only to huge (red) and high (yellow) levels of jams respectively. We can see that the major concentration of huge (red) level of jams is on 405 and 101 highways, Santa Monica and West Downtown LA. And high (yellow) level of jams is mostly reported in a west part of Downtown LA, which can be explained by the location of most offices along major streets in west part of Downtown LA.

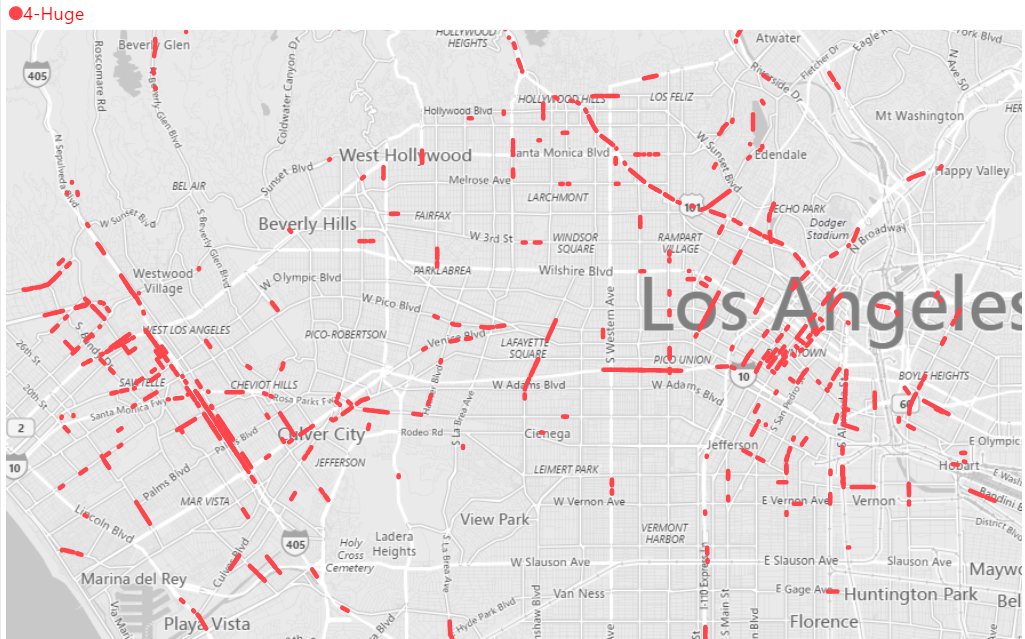


Figure – Huge level of jams filtered

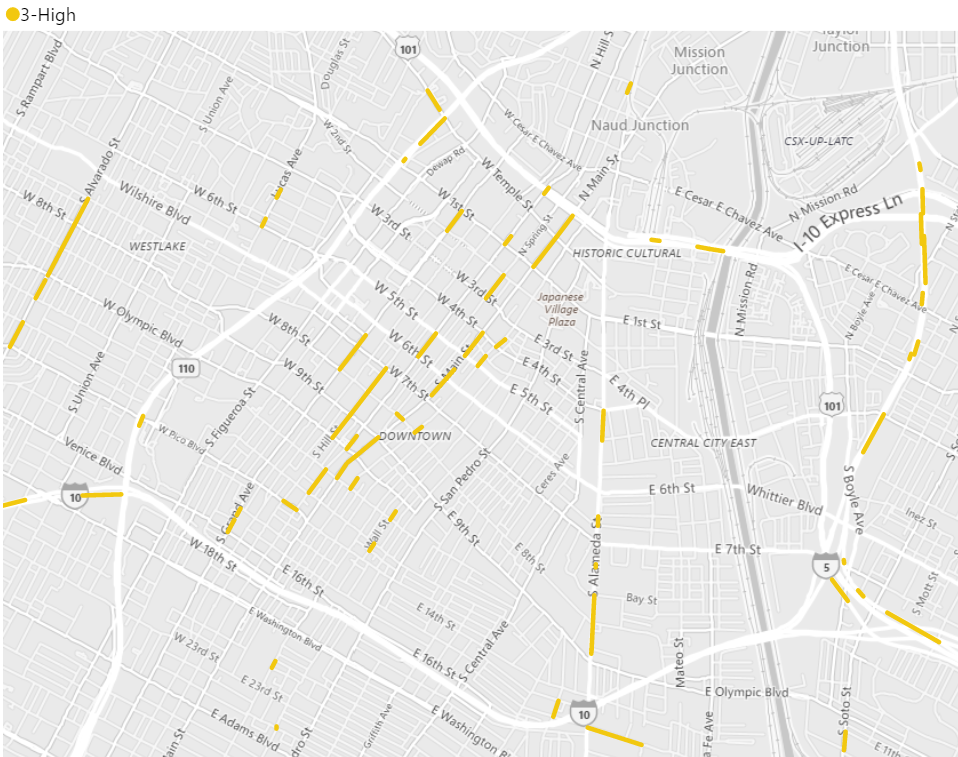


Figure – High level of jams filtered in DTLA

**7. Conclusion**

Finally, summing up all the above work we can conclude the following:

1. Denser traffic on highways;
2. Evening time is the heaviest traffic time with the rush hour from 5:00pm to 8:00pm;
3. Major areas of traffic are in Downtown Los Angeles, Santa Monica, Hollywood and highways.

From the available data, which is limited to few days, we were able to provide an interactive tool for analysis and data manipulation. Further work can be done with bigger dataset in order to find more insights and create a data driven conclusions on LA County traffic situation by using the this framework.

For more information, dashboards and code visit project’s GitHub link[[3]](#footnote-4).

### References

[1] Waze (October, 2016) Connected Citizen Program. Retrieved from <https://wazeopedia.waze.com>

[2] Waze, Connected Citizens by Waze. Retrieved from <https://web-assets.waze.com/partners/ccp/WAZE-CCP-Factsheet.pdf>

[3] Schnuerle, M. (November, 2017) Louisville and Waze: Applying Mobility Data in Cities. Retrieved from <https://datasmart.ash.harvard.edu/sites/default/files/2018-04/FINAL_Harvard_CAN_Waze.pdf>

[4] Louisville Waze Map. Retrieved from <https://louisvillemetro-ms.carto.com/builder/d98732d0-1f6a-4db2-9f8a-e58026bf0d39/embed>

[5] Louisville Waze Map. Retrieved from <https://tinyurl.com/y8ezz9ge>

[6] Louisville Waze Map. Retrieved from <https://cdolabs-admin.carto.com/builder/a80f62bf-98e1-4591-8354-acfa8e51a8de/embed>

[7] I-95 Corridor Coalition. I-95 Facts. Retrieved from <https://i95coalition.org/>

[8] Bamford, R. (September, 2017) Integrating Waze Data with Public Agencies’ Data. Retrieved from <http://i95coalition.org/wp-content/uploads/2017/08/4-XCM-Waze-Presentation-final.pdf?x70560>

1. The jams\_head file was not used, due to absence of date. Jam information was taken form irregularities\_head file instead. [↑](#footnote-ref-2)
2. The area of Los Angeles includes DTLA, Central LA, West LA, Mid LA, Silver Lake, Hollywood etc. [↑](#footnote-ref-3)
3. GitHub Link: <https://github.com/rjoshi5/waze_traffic_patterns_LosAngeles> [↑](#footnote-ref-4)