Project Proposal: Public Transport Data Lake

CS4225 Big Data Systems for Data Science

Group members:
Royce Ho (A0199477A),
Ong Wei Sheng (A0206042X),
Liu Yuhui (A0225308L),
Jodi Choo (A0222236R),
Tan Shee Hui (A0197153Y),
Lim Zi Xuan Jeremy (A0200402J)

Topic Introduction

Traffic congestion has posed a problem to urban areas. Studies have shown and proven the threat of traffic congestion to the economy, as well as public health and well-being. Traffic jams and the longer travel times as a result of it, culminates in a loss of productivity and waste of fuel. The increase in vehicle tailpipe emissions also contributes to air pollution. If left unchecked, the traffic congestion issue slows economic growth and increases public health risks. An increase in population and the rise of various ride-sharing services further increases the number of vehicles on the road, worsening the traffic congestion problem.

Various data collected from multiple avenues can be used in conjunction with data processing methods and analytics to model and predict crowd and traffic distribution. These insights may be used by decision-makers to devise measures that alleviate the negative impacts of traffic congestion. Transport providers can use these insights to improve scheduling plans, increasing the efficiency of public transport to reduce lost time spent in traffic jams, while delivery companies can optimize delivery times based on the traffic predictions to increase the productivity of workers.

A time-series prediction will be conducted using traffic, travel time and crowd density data to predict surges in congestion levels in various areas in Singapore. Given that the amount of traffic on the road can also depend on factors other than the time of day, we have decided to consider the effect of rainfall on traffic congestion as well, by utilizing weather data in our analysis. By considering a greater number of factors, the accuracy of our prediction can be improved.

Other related works and research

Currently, there are some previous works being done that uses public transport data, traffic congestion data etc. and we will go review what each of them does. Firstly, TomTom's Singapore traffic report collects traffic data to visualize and show traffic congestion to users through a dashboard and allows users to select the relevant information that it wishes to view about traffic. However, this is merely just collection and visualization of data as it does not analyze or learn the data obtained, therefore analyzing the data collected to learn some useful trends and insights along with the visualization of the data is one of our goals.

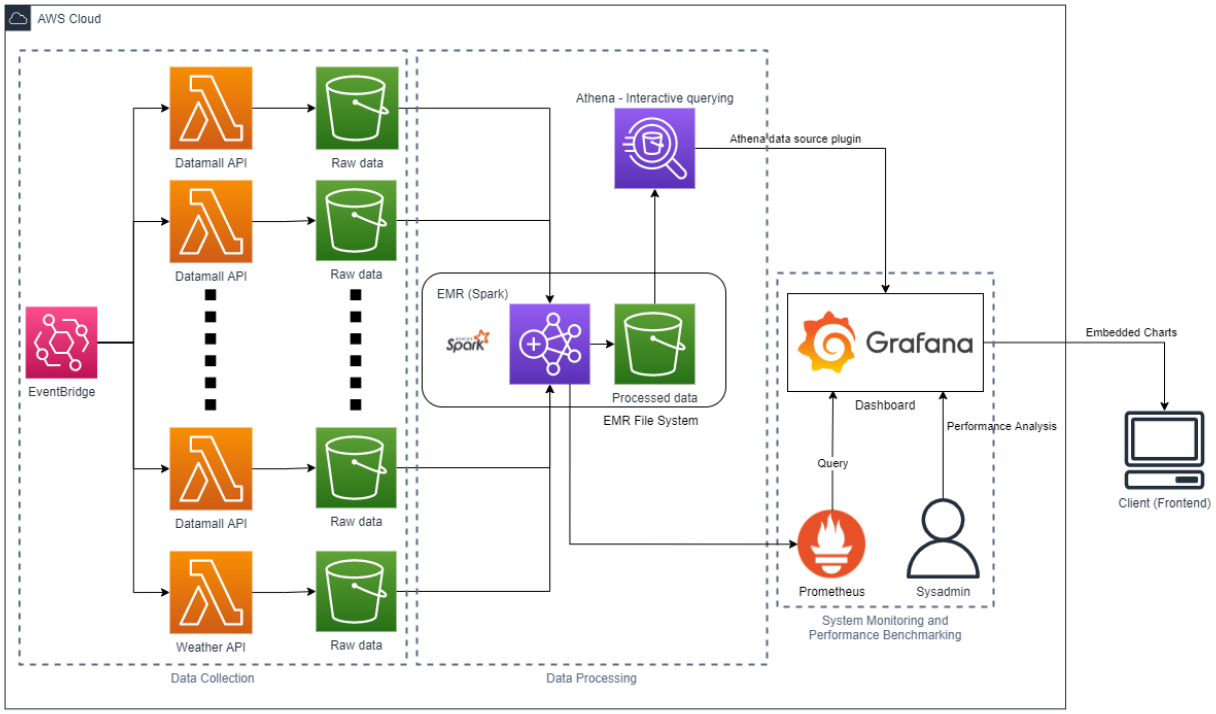
Some of the relevant research articles that analyze traffic data includes a paper called FASTER: Fusion AnalyticS for public Transport Event Response (Sebastien Blandin, Laura Wynter, Hasan Poonawala, Sean Laguna, Basile Dura), which involves methods ranging from statistical machine learning to agent-based simulation to monitor and predict commuter movement in order to effectively respond to unforeseen public transport incidents and to keep up the level of efficiency of public transports. Real time data processing using Spark/Hadoop is used here, alongside Redis for fast operations.

A Data-Drive and Optimal Bus Scheduling Model With Time-Dependent Traffic and Demand (Yuan Wang, Dongxiang Zhang, Lu Hu, Yang Yang, Loo Hay Lee) involves the collection and analyzes traffic data to efficiently determine whether more bus services are required at perhaps

a certain time of the day. The model also takes into account operating expenses to see if the benefits of increasing the frequency of bus services outweigh the negatives of the increase in operating expenses. Our main aim for our project is to combine the interactive front end of current works with insightful analysis of the data being collected.

Experimental Approach

Data Lake Architecture



Data Collection

9 different LTA datamall APIs are used to extract information such as platform crowd density, as well as NEA's weather API for rainfall information. Amazon EventBridge triggers the lambda functions to be called every 5 and 10 minutes, depending on which group they are in. The data which are all in JSON format is then stored in their respective S3 buckets.

Data Processing

Spark is our distributed processing framework of choice. Spark clusters will be created via Amazon EMR and will be used to process raw data from S3, utilizing Amazon's EMRFS, which is an implementation of HDFS that all Amazon EMR clusters use for reading and writing regular files from Amazon EMR directly to Amazon S3, after which data can be queried out via Athena.

System monitoring and performance benchmarking

Apart from visualizing queries from S3, Grafana's other main purpose would be monitoring EMR. Inspired from this blog post, we plan to integrate Prometheus, an open-source systems monitoring and alerting tool, and Grafana, an open-source visualization and analytics tool, to provide an end-to-end monitoring system for EMR clusters.

Data Analytics

We want to provide predictive insights as well, and plan to use time series prediction on historical data to give users a peek on what to expect in the future. We are considering the python libraries Prophet and tsfresh for time series modelling, as well as traditional algorithms such as ARIMA. The insights generated will be sent and rendered on the frontend of our application.

Data sets

As mentioned earlier, 9 LTA Datamall APIs and 1 NEA weather API for rainwall were used. We chose them as they were Real-time APIs without any strict thresholds on the number of calls. A local based data set related to the public also meant that we could work on a topic that involved Singaporeans. Our insights found may also be beneficial for the public good.

1. Carpark Availability
2. Estimated Travel Times
3. Faulty Traffic Lights
4. Platform Crowd Density Real-Time
5. Road Works
6. Taxi Availability
7. Traffic Images
8. Traffic Incidents
9. Rainfall

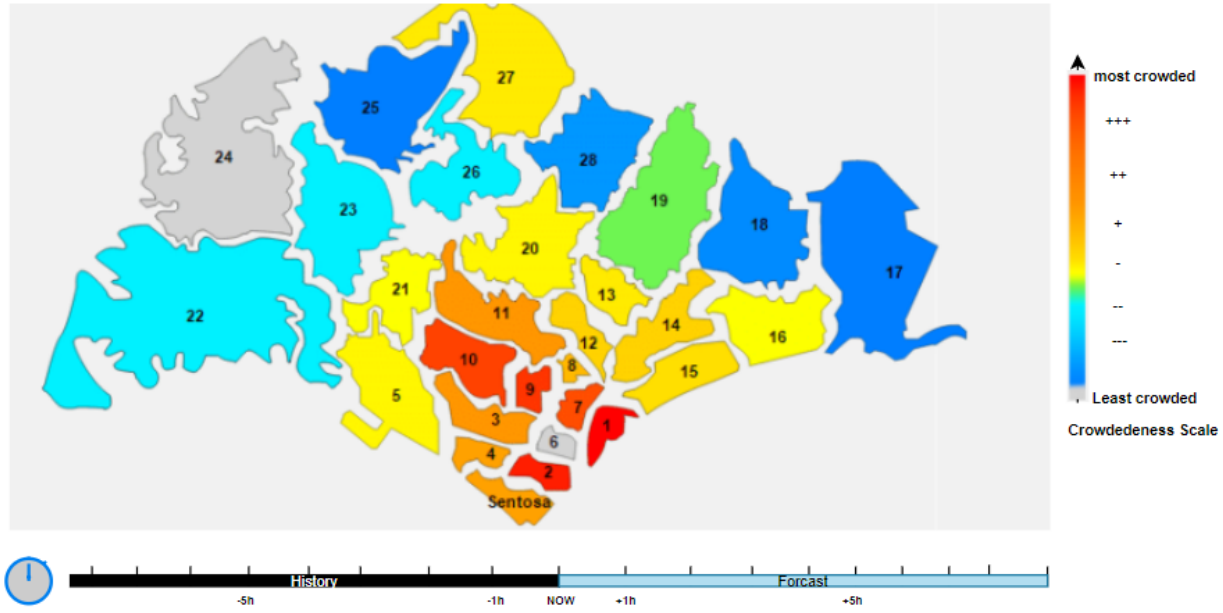
Expected Results

Result Data

The crowdedness distribution of Singapore and its forecast crowdedness.
With Historical accurate data analyzed from our collected data lake and the forecasted data that is predicted by our prediction algorithm.

Result Presentation

Heat map and embedded charts will be used to show the crowdedness of Singapore
Here is a conceptual proposal for the final UI:



In the middle: Singapore map will be segregated into many small districts to show the crowdedness of each district. Districts are labeled in different colors to show the crowdedness scale.

Right part: the scaling indicator to help tell which indicates the most crowded place at that time.
Button: Time selector, you can use it to select the crowdedness map for the historical data and the forecasted data at a specific time.

For the scale, depending on the further work and research, we might also take progressive color to indicate the scale. We will embed additional charts on the frontend, generated from grafana, to display trend predictions and current statistics.

Summary

This project aims to be a product that helps alleviate traffic congestion issues in Singapore and improve transport efficiency by processing historical data to create prediction models for crowd surges. Road related data, such as MRT platform crowd density and taxi availability, is constantly collected through the LTA Datamall API. Weather data is also collected through the NEA weather API to allow us to draw relations between weather and road situations. Data from these two sources are the core of our data lake. Our results and predictive insights (crowd levels and forecasted crowd density) generated from these data will be presented in a meaningful manner, visualized through maps and graphs.

References

Singapore traffic report: Tomtom traffic index. report | TomTom Traffic Index. (n.d.). Retrieved March 9, 2022, from https://www.tomtom.com/en_gb/traffic-index/singapore-traffic/

Passman, D. B. (2013). Fusion analytics: A Data Integration System for public health and medical disaster response decision support. *Online Journal of Public Health Informatics*, *5*(1). https://doi.org/10.5210/ojphi.v5i1.4522

Wang, Y., Zhang, D., Hu, L., Yang, Y., & Lee, L. H. (2017). A data-driven and optimal bus scheduling model with time-dependent traffic and demand. *IEEE Transactions on Intelligent Transportation Systems*, *18*(9), 2443–2452. https://doi.org/10.1109/tits.2016.2644725

Electronic road pricing: Experience & lessons from Singapore. (n.d.). Retrieved March 9, 2022, from https://www.researchgate.net/publication/327280854_Electronic_Road_Pricing_Experience_Lessons_from_Singapore

Tan, D., & Lang, F. (2008, August 10). *Monitor and Optimize Analytic Workloads on Amazon EMR with Prometheus and Grafana*. Amazon. Retrieved March 9, 2022, from https://aws.amazon.com/blogs/big-data/monitor-and-optimize-analytic-workloads-on-amazon-emr-with-prometheus-and-grafana/

Kranc, M. (2017, November 2). *Amazon fills a big data hole with Athena*. Datanami. Retrieved March 9, 2022, from https://www.datanami.com/2017/10/17/amazon-fills-big-data-hole-athena/