Project Proposal Outline

Cross Zeigler, Suchada Hapikul, Ewin Hong, Neethu Battula, Sagar Deepakgiri Goswami

Note: This is a skeletal form for the proposal meant to outline ideas for the document, not the fully fleshed out proposal itself

**Cover Page**

* Title
* The Proposal Title
* Author’s Name(s) & Team #
* Course Project Advisor/Professor’s Name(s)
* University Name, Course Number and Name
* Date

**Introduction (Cross)**

* Briefly explain the project and its scope, motivation
  + Explanation: Analyze traffic records to create a model to predict what day(s) experience the most traffic
    - Additionally: Future Considerations: Geospatial analysis, see where most of the traffic is (regions)
  + Scope: New York City, 2017 to present
  + Motivation: Help give suggestions to better infrastructure of NYC to improve driver and pedestrian safety along with improving satisfaction with transportation (public or personal) due to improvements in time (reduction) and ease of access. Advocate for public policy makers to refine infrastructure.
    - Look at specific points of congestion to help with this
    - How many people within a certain area use public transport
* Briefly give some backgrounds and/or related work
  + Two reference articles from Fon
  + List Reference Articles you may have here:

***(Hi team, Please review these related works…… -Fon-)***

In recent years, many researchers have been studying the use of Big Data analytics and machine learning to support the analysis of large amounts of data. Machine learning models can predict traffic situations in advance effectively (Nibareke, T., & Laassiri, J., 2020). Vasudevan et al. (2016) presented a technical approach that combined Apache Spark’s open-source data analytics and machine learning techniques to predict traffic flow patterns using simulated connected vehicle messages. They reported that connected vehicle data can be processed rapidly using Big Data analytics to generate precise predictions of traffic flow regimes (Vasudevan, M., 2016). Hofleitner et al (2012) used a machine learning framework to examine static parameters of the roadways and created forecast travel times through the arterial network using traffic flow theory principles consistent with the physics of traffic (Hofleitner, A., Herring, R., & Bayen, A., 2012). The machine learning technologies are used in the Big Data-driven traffic flow prediction problem. Kong et al (2018) applied the deep learning algorithm based on machine learning to build a traffic flow prediction network. They asserted that real-time traffic flow conditions can be predicted accurately with a large enough sample size (Kong, F., Li, J., Jiang, B., Zhang, T., & Song, H., 2018). Both Yisheng et al. (2014) and Chen et al. (2011) studied the use of machine learning models to create traffic flow prediction by considering a time series equation. While Yisheng et al. (2014) focused on traffic flow prediction by building blocks to represent traffic flow features for prediction (Yisheng, L., Duan, Y., Kang, W., Li, Z., & Wang, F. Y., 2014), Chen et al. (2011) focused on speed measurements by constructing a multi-step speed prediction based on traffic speed data (Chen, H., Rakha, H. A., & Sadek, S., 2011).

**Project Goals (Cross)**

* The project's primary goal is to provide valuable insights into the city's traffic system to improve transportation within NYC for increased satisfaction, safety, and decreased travel times (from the Project Topic Draft file)
  + Focus Areas:
    - Specific points and regions within NYC, specifically those dealing with high volumes of traffic at specific times of day and giving valuable advice to certain audiences
    - Creating a machine learning model based on the previous NYC Average Traffic Speed Data sourced from the New York City Department of Transportation can help determine times/dates and locations that observe a relatively high amount of congestion
    - Audiences: Residents of New York City (travelers/passengers), emergency services, stakeholders in the urban transportation industry, and businesses (transportation agencies)

**Project Requirements (Cross)**

* Function requirements
  + What we need to achieve the project
    - Analysis
      * Python
      * R
      * Excel
    - NoSQL Database:
      * MongoDB
    - Big Data Engine:
      * Spark
* Objectives
  + Briefly describe or list the objectives of the overall project you proposed (see above in Goals and Introduction)

**Proposed Selected Dataset (Ewin)**

<https://data.cityofnewyork.us/Transportation/DOT-Traffic-Speeds-NBE/i4gi-tjb9>

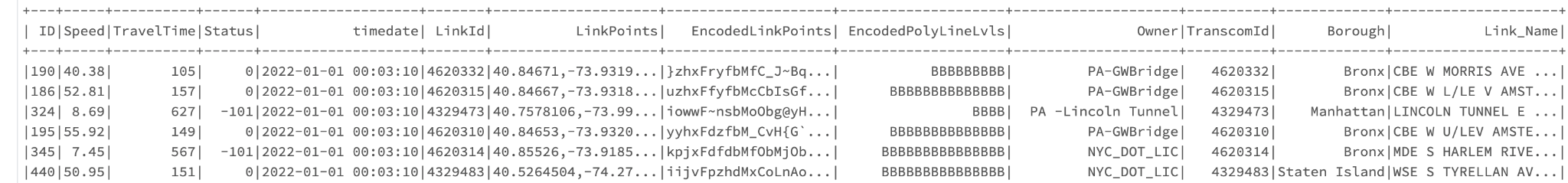
Total Records: 58.8 million records starting on April 17, 2017 to March 9th 2022

13 Features

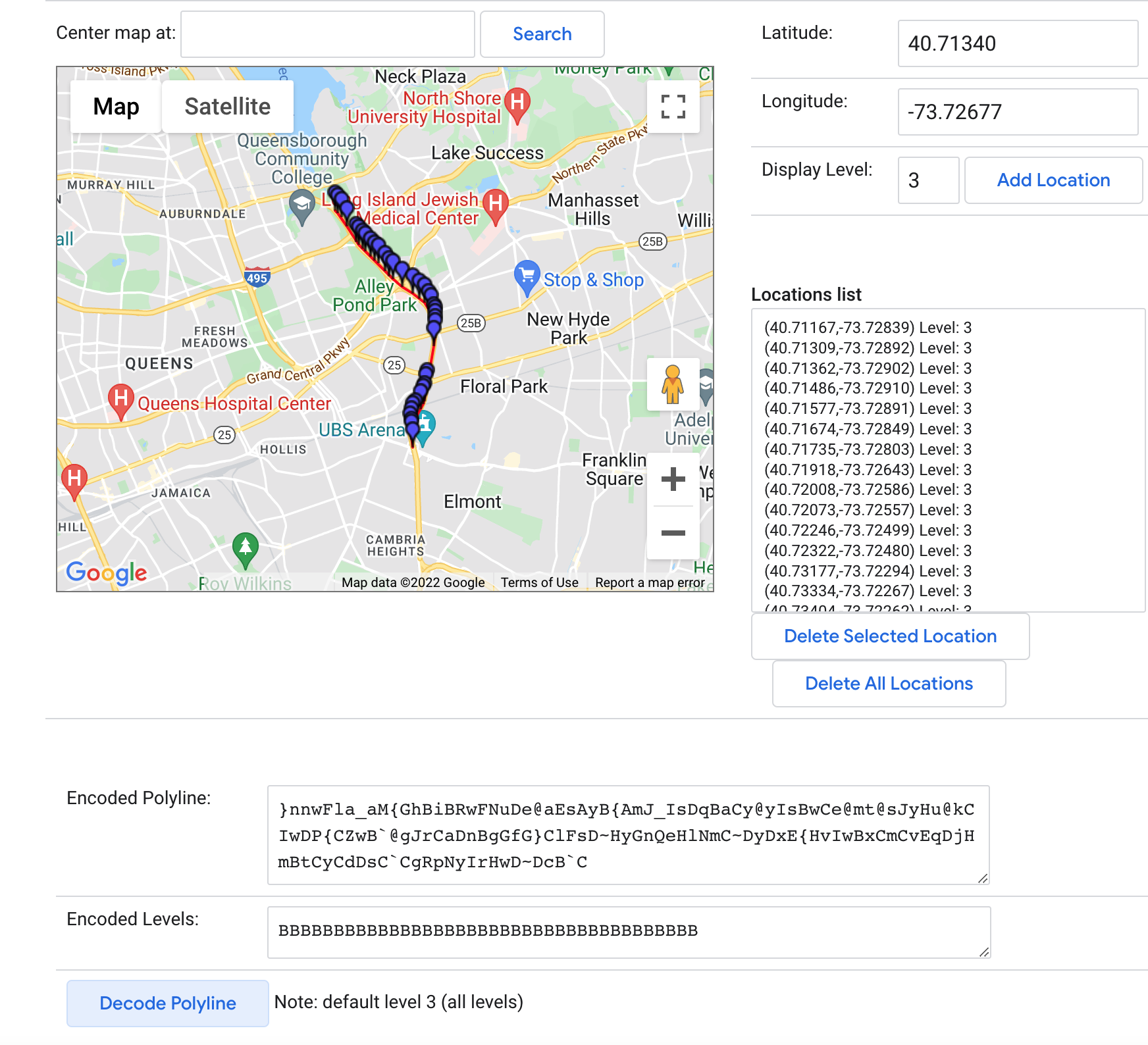
| Name of Feature | Description | Data Type |
| --- | --- | --- |
| ID | Unique Identifier for Sensor within dataset | integer |
| Speed | Average Speed for Sensor at location | double |
| TravelTime | Time Travel in minutes | integer |
| Status | Artifact (not useful) | integer |
| Data\_As\_Of | Date and Time of Day for Sensor Data | datetime |
| Link\_Id | TRANSCOM Link ID | integer |
| Link\_Points | Group of Latitude and Longitude points of Sensor data | List of 2 double (latitude and longitude) points |
| Encoded\_Poly\_Line | Link\_Point representation of Google compatible poly line | string |
| ENCODED\_POLY\_LINE\_LVLS | Encoded representation of Poly Level | string |
| Owner | Owner of Sensor | string |
| TRANSCOM\_ID | Artifact (not useful) | string |
| BOROUGH | Name of Borough Sensor exists | string |
| Link\_Name | Description of Sensor location | string |

With the Link\_Points, the distance of the starting and ending points of a single link point can be calculated. These link points can be evaluated over time and their evolution. Then with the link points, the Boroughs can be added for grouping for additional analysis.

Example:



Explanation of polyline: https://developers.google.com/maps/documentation/utilities/polylinealgorithm

Example of polyline: https://developers.google.com/maps/documentation/utilities/polylineutility?csw=1   
Using this polyline: }nnwFla\_aM{GhBiBRwFNuDe@aEsAyB{AmJ\_IsDqBaCy@yIsBwCe@mt@sJyHu@kCIwDP{CZwB`@gJrCaDnBgGfG}ClFsD~HyGnQeHlNmC~DyDxE{HvIwBxCmCvEqDjHmBtCyCdDsC`CgRpNyIrHwD~DcB`C  


**Description of proposed system** **(Sagar)**

* Draw a diagram of the conceptual system architecture or framework
* Data analytics method
  + You will use or propose (*note that each method should consider the features you will select*) the following methods:
    - data visualization
      * High traffic areas
        + Amount of traffic by hour
        + Amount of traffic by district
    - data analysis
      * Statistical analytics
        + How many sensors do we have

Group of sensors can be categorized to link ID

* + - * + How many locations do we have
      * Advanced analytics (machine learning, deep learning, etc.)
        + Machine learning

Conceptual System Architecture Diagram

* + Data Analytical Methods
    - Data Ingestion and Cleaning
    - Data Exploration
    - Linear Regression Models to predict Congestion  
      (Maybe Geographically Weighted Regression)
      * With respect to location and time
        + Hourly Predictions
        + Daily Prediction
        + Weekly Predictions
        + Monthly Predictions
        + Annual Predictions
    - Random Forest / Feature Extraction
      * Figure out bottlenecks by feature extraction of the average traffic speed for neighboring locations
    - Clustering locations and sensors
      * Based on Individual Sensor values
      * Based on Individual w.r.t. values for neighboring sensors / localities
        + Values of frequencies
        + Values of Congestion Intervals
    - Predictions Dashboard (Heatmap / Bubble Map) (if time allows)
      * Runs based on weekly data
        + Back-end Scheduled Script

Data Extraction | Extracts data from the original dataset source/API to display data

Running the model | Training the model based on the new data

* + - * + Front-end Dashboard Design

Design of the dashboard

File that keeps updating with current coefficients

* + - * Hourly predictions
      * Weekly predictions
      * Monthly predictions
      * Annual predictions
    - Case Studies
    - Documentation and Deliverables
      * Reports
      * Presentation
        + Slide Deck: Cross
        + Share Screen for working demonstration: Cross

**Proposed development platforms (Neethu)**

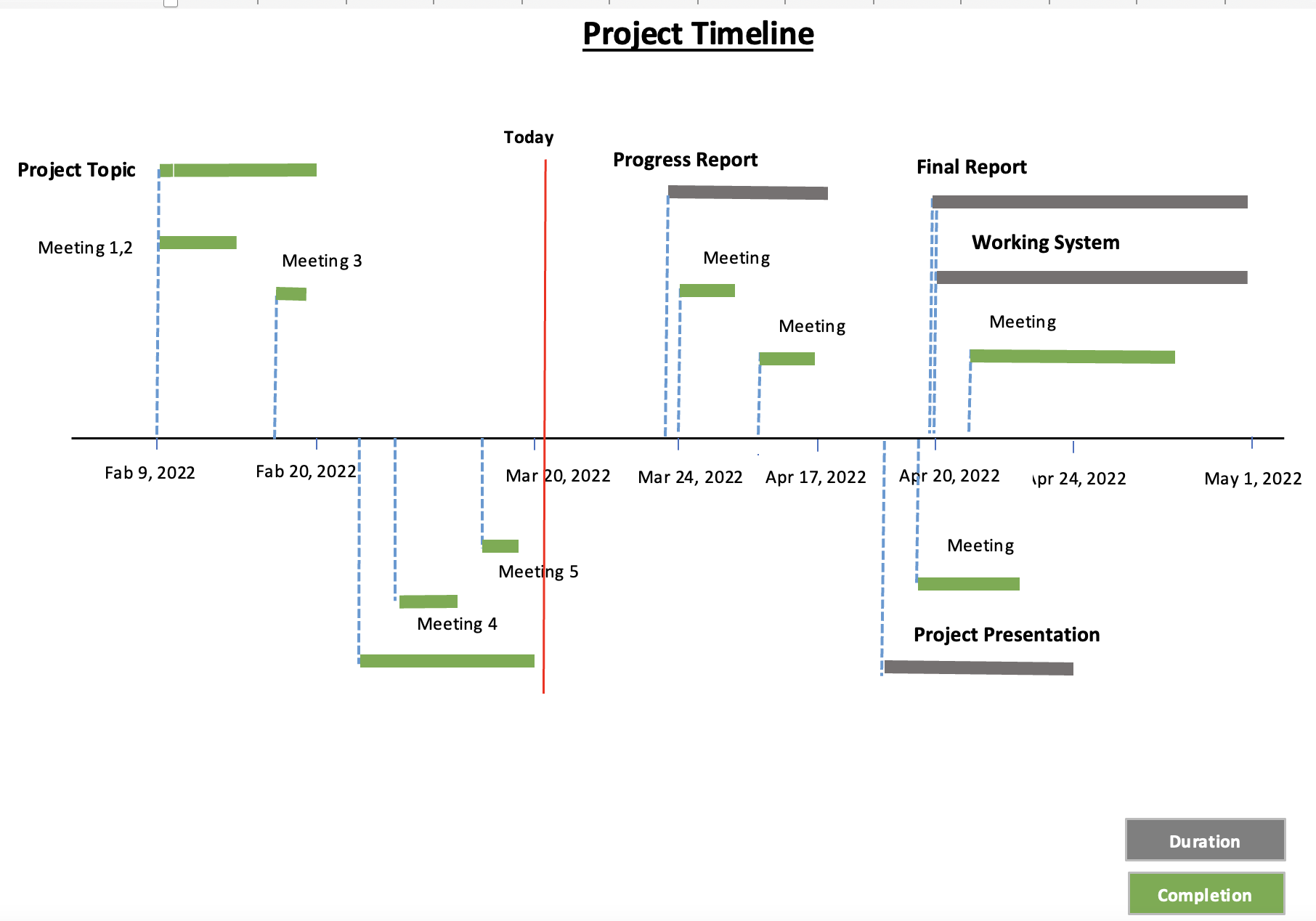
* Briefly describe the software and hardware development platforms
  + OS Everyone is using:
    - Cross: Mac 8GB RAM, 3.1 GHz, Dual-Core Intel Core i-5
    - Ewin: Mac 6core CPU at 3.2GHZ Intel i7 64GB of ram
    - Fon:
    - Neethu: Mac 16GB RAM,2 GHZ intel quad- core intel core i5
    - Sagar: Windows Machine: 16 GB RAM, AMD Ryzen 4800H 8-cores 16-threads, Nvidia RTX 3060 Mobile GPU
  + Databricks (Apache Spark)
  + Python (is quicker) (utilizing a version 2.7 or above)
  + R (3.0 or above)

**Project tasks & deadline (Fon)**

* List tasks/subtasks with assigned team member(s)’ name(s)
  + Cross Zeigler - data analysis, data visualization, linear regression modeling, Dashboard, presentation, project management
  + Suchada Hapikul - data preparation, data exploration, data visualization, timeline project management, presentation
  + Ewin Hong - data ingestion, data preprocessing/cleaning, data visualization in Tableau’s map, presentation
  + Sagar Deepakgiri Goswami - data preparation, data exploration, data visualization, presentation
  + Neethu Battula - data visualization, presentation
  + Tasks: List the people primarily response for the tasks and how long we’ll be doing them for

| **Task** | **Sub-Task** | **Assigned to** |
| --- | --- | --- |
| **Data Preparation**  **(03/01/2022 - 03/30/2022)** | Data Ingestion (batch and API) | Ewin, Sagar |
| Data Wrangling |  |
| Data Cleaning |  |
| Data Exploration |  |
| **Data Analytics**  **(03/20/2022 - 04/10/2022)** | Linear Regression Models |  |
| Random Forest / Feature Extraction |  |
| Clustering Locations and sensors |  |
| **Data Visualization**  **(04/05/2022 - 04/15/2022)** | Back-End Scheduled Script |  |
| Front-End Predictions Dashboard |  |
| **Case Studies**  **(04/11/2022 - 04/20/2022)** | Locations with highest congestion |  |
| Locations that cause highest congestion |  |
| **Deliverables**  **(04/17/2022 - 04/24/2022)** | Project Presentation | Everyone |
| Project Video | Everyone |
| Final Project Report | Everyone |

* + - Data Preparation: Fon, Ewin, Sagar (3/20/2022 - 3/27/2022)
    - Data Analysis: Cross, Fon, Sagar (3/27/2022 - 4/3/2022)
    - Visualization: Everyone (4/3/2022 - 4/17/2022)
    - Presentation: Everyone (4/17/2022 - 4/24/2022)
    - Case Study: Everyone (4/24/2022 - 5/1/2022)
    - Report: Everyone (4/24/2022 - 5/1/2022)
* Follow the overall project deliverables to allocate your time for each of your proposed tasks/subtasks.
  + Project Topic (2/20/2022)
  + Project Proposal (3/20/2022)
  + Progress Report (4/17/2022)
  + Project Presentation (4/24/2022)
  + Final Report (5/1/2022)
  + Working system (5/1/2022)
  + Data Preparation:
* Create your project management timeline, and determine the time required to complete the project.

~~~~

**References:**

Nibareke, T., & Laassiri, J. (2020). Using Big Data-machine learning models for diabetes prediction and flight delays analytics. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00355-0

Vasudevan, M. (2016). *Big data analytics: predicting traffic flow regimes from simulated connected vehicle messages using data analytics and machine learning.* US Department of Transportation. Retrieved March 14, 2022, from https://rosap.ntl.bts.gov/view/dot/32616

Hofleitner, A., Herring, R., & Bayen, A. (2012). Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. *Transportation Research Part B: Methodological*, *46*(9), 1097–1122. https://doi.org/10.1016/j.trb.2012.03.006

Kong, F., Li, J., Jiang, B., Zhang, T., & Song, H. (2018). Big data‐driven machine learning‐enabled traffic flow prediction. *Transactions on Emerging Telecommunications Technologies*, *30*(9). https://doi.org/10.1002/ett.3482

Yisheng, L., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2014). Traffic Flow Prediction With Big Data: A Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 1–9. https://doi.org/10.1109/tits.2014.2345663

Chen, H., Rakha, H. A., & Sadek, S. (2011). Real-time freeway traffic state prediction: A particle filter approach. *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. https://doi.org/10.1109/itsc.2011.6082873

* Provide appropriate citations and references.
* Be sure to include a citation and link(s) for the dataset(s)
* See <http://infoguides.gmu.edu/citingdata>

**Appendix**

* Show a piece of data

Other content should be added as necessary