CSCI-GA.2436 Realtime and Big Data Analytics
Final Project Proposal

Team Members:
- Yueyan Lu (yl6211)
- Guanshi Wang (gw2310)
- Wenbo Bao (wb2128)
- Haowei Tu (ht2397)

Project Objectives

Through analyzing taxi, for-hire vehicle (uber/lyft), city bike and real-time traffic speed data, we would like to track how traffic speeds fluctuate throughout the day and identify peak and off-peak periods. Next, we aim to draw some insights on transportation patterns, user behavior and preference in choosing transportation mode, the correlation between traffic velocity and demand for ride-hailing services. Ultimately, our goal is to possibly provide some optimization suggestions for the transportation network.

Proposed Work

- Through data cleaning and profiling, we would like to determine peak hours and popular locations for each mode of transport in NYC.
- Explore how traffic congestion (or possibly other factors) affects the usage of different modes of transportation.
- Understand which transportation mode is preferred for specific routes or distance.
- Based on the historical data, we would like to suggest improvements in transportation infrastructure

Data Sources Description

- **Taxi & Limousine Commission (TLC) trip record data**
  - Link: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
  - Profile & clean by: yl6211
  - Data is stored in parquet format containing important information such as pickup/dropoff datetime, location and trip fare etc.
  - The size of the yellow taxi trip records is approximately 50MB per month, and our intention is to process data from the preceding 2 years, resulting in a total of 1.2 GB.
  - Big Data tools to use: HDFS, Hive

- **NYC Real-Time Traffic Speed Data**
  - Link: https://www.kaggle.com/datasets/aadimator/nyc-realtime-traffic-speed-data/data

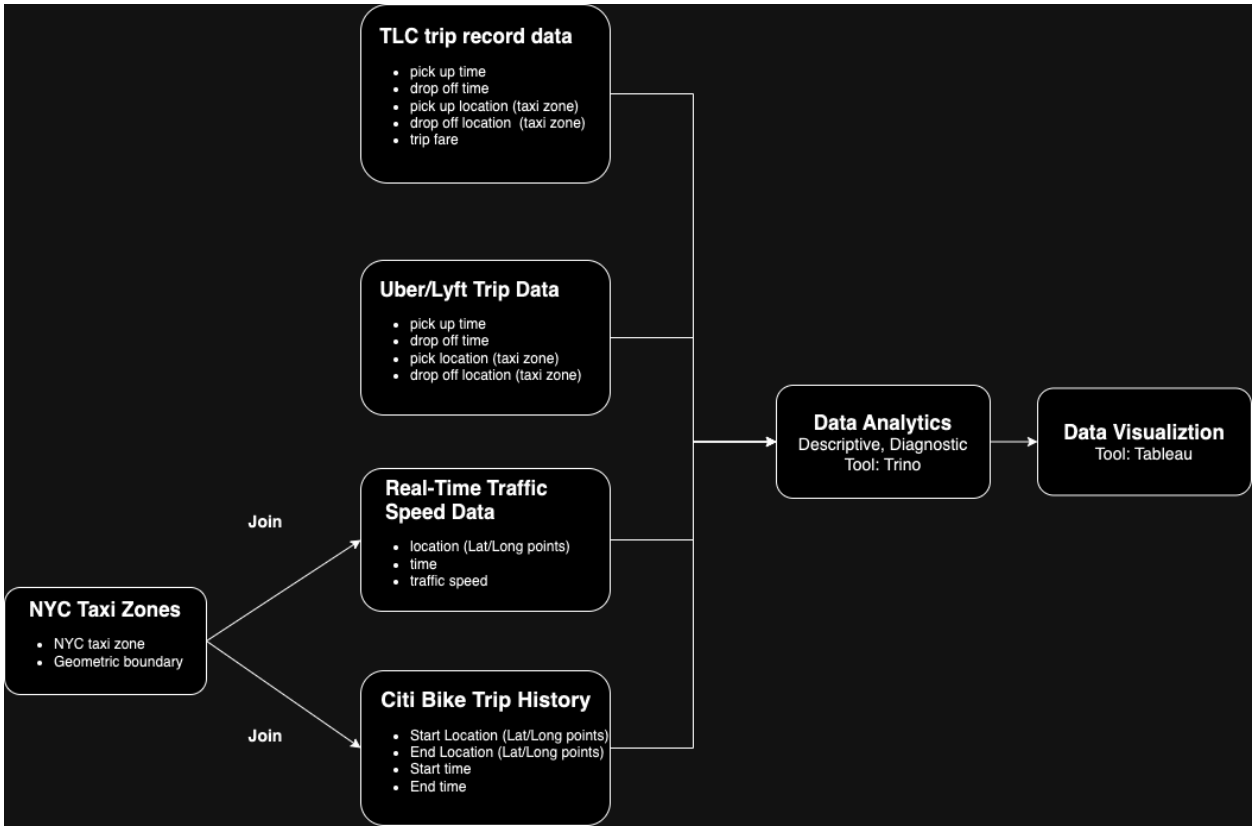
    - Profile & clean by: gw2310
    - Data is stored in a csv file, and after unzip the size is ~27G. We may need the following data: **LINK_POINTS** (Sequence of Lat/ Long points, describes locations of the sensor links), **DATA_AS_OF** (Last time data was received from link), **SPEED** (Average speed a vehicle traveled between end points on the link in the most recent interval)
    - For LINK_POINTS, we simply take an arithmetic average to get a Lat/ Long point, and then map these to a specific taxi zone through joining NYC Taxi Zones.
    - Big Data tools to use: HDFS, Trino

- **NYC FHV (Uber/Lyft) Trip Data**
    - Link: https://www.kaggle.com/datasets/jeffsinsel/nyc-fhv-uberlyft-trip-data-simple-2015-2022
    - Profile & clean by: ht2397
    - Data is stored in parquet format and the size is 5.39 GB
    - It contains the date and time of the trip pick-up, the date and time of the trip dropoff, TLC Taxi Zone in which the trip began, and TLC Taxi Zone in which the trip ended
    - Big Data tools to use: HDFS, Hive

- **NYC Taxi Zones**
    - Link: https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc
    - Profile & clean by: gw2310, wb2128
    - This is a small csv helper dataset for our project. We can map Lat/ Long points to taxi zones by joining it to our big datasets.

- **Citi Bike Trip Histories**
    - Link: https://citibikenyc.com/system-data
    - Profile & clean by: wb2128
    - Data is stored in CSV and size is about 7.2 GB
    - It contains the date and time of the citi bike pick-up and drop-off dock location in latitude and longitude, the date and time of the trip
    - Big Data tools to use: HDFS, Hive

Initial Design Diagrams



https://drive.google.com/file/d/14BTB4SbVg3VVIMA2V62AOsE9t5dPmlPg/view?usp=drive_link