Group 18: Urban Mobility and Crime Under Weather Conditions

Team:

- Jasper Zeng - qz2283:
- Hao Huang - sh8313
- Vincent Xu - yx2021

I. Introduction

This project investigates how environmental conditions influence urban mobility and public safety in New York City. By integrating MTA subway ridership, High-Volume For-Hire Vehicle (HVFHV) trip data from Uber and Lyft, weather and air quality data, and NYPD arrest records, we analyze how factors such as precipitation and air pollution shape travel behaviors and relate to crime patterns across time and space.

All datasets are joined through timestamp or location fields to uncover correlations between environmental changes, transportation demand, and crime activity. The system leverages HDFS for distributed data storage, MapReduce for large-scale data processing, and Trino for efficient querying and aggregation. This approach enables scalable analysis of multi-source city data to better understand how environmental stressors affect both mobility and public safety in urban environments.

II. Data Sources

Traffic Data - High Volume For Hire Vehicles (Uber/Lyft)

Link: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Size: 468MB per month

Owner: Shuai Huang

Traffic Data - MTA Subway Data

Link:

https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-2020-2024/wujg-7c2s/about_data

Size: 121M rows

Owner: Vincent Xu

Weather Data - Visual Crossing Weather API

Link:

https://www.visualcrossing.com/weather-query-builder/New%20york/?v=api

Size: TBD - depends the interested time duration (e.g. 356 day, hourly granularity, 8544 rows)

Owner: Jasper Zeng

Crime Data - NYPD Arrests Data

Link:

https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/about_data

Size: 5.99M rows

Owner: Jasper Zeng

III. Technical Design and Approaches

The project is organized into two main phases: data preprocessing and data integration and querying. In Phase 1, raw datasets—including NYPD Arrest Data, MTA Subway Hourly Ridership, HVFHV (Uber/Lyft) trip records, and Weather and Air Quality data—are ingested into HDFS for distributed storage. We run a series of MapReduce jobs to clean, aggregate, and normalize these datasets. This preprocessing step produces intermediate, structured outputs that are aligned either by timestamp or location, ensuring consistency and compatibility across all sources for later analysis.

In Phase 2, the preprocessed intermediate data stored in HDFS is accessed through Trino for efficient querying and integration. Trino acts as the join engine, combining the datasets dynamically based on specific analytical needs—such as linking ride-hailing and subway activity to weather conditions or correlating mobility shifts with arrest patterns. The resulting joined analytical tables support flexible queries and visualization tasks, enabling in-depth exploration of how environmental and social factors jointly influence urban mobility and public safety across New York City.