Data-Intensive Computing Project Proposal:

Time Series Prediciton On NYC Caps

Group

Dream_Team_v2

- **Xiya Sun**
- **Silvia Pasini**
- **Lijie Li**
- **Davis Siemens**

Institution

- **KTH**
- **Course**: Data Intensive Computing
- **Date**: 13th September 2025

Goal Of Project

Develop a big data pipeline to time series forecast NYC yellow taxi demand using scalable storage, processing, and machine learning tools. We will feature engineer the target variable to be demand per hour for a given location.

Dataset

The dataset can be found on Kaggle (https://www.kaggle.com/elemento/nyc-yellow-taxi-trip-data). It contains information about NYC Yellow Taxi trips, including pickup and dropoff locations, timestamps, and trip distances, for the months Jan 2015, Jan 2016, Feb 2016 & March 2016 (approx. 2GB per Month).

Attributes

- **VendorID:** Provider code (Creative Mobile Technologies or VeriFone).
- **Pickup/Dropoff datetime:** Start and end times of the trip.
- **Passenger_count:** Number of passengers (driver-entered).
- **Trip_distance:** Distance in miles from the taximeter.
- **Pickup/Dropoff coordinates:** Latitude and longitude of trip start/end.
- **RateCodeID:** Fare type (e.g., standard, JFK, Newark, negotiated).
- **Store_and_fwd_flag:** Whether trip was stored before transmission (Y/N).
- **Payment_type:** How passenger paid (card, cash, no charge, dispute, etc.).
- **Fare_amount / Extras / MTA_tax / Surcharge:** Meter fare and surcharges.
- **Tip_amount / Tolls_amount:** Tips (card only) and tolls.
- **Total_amount:** Final charged amount (excl. cash tips).

Tools & Metholodolgy

We propose the following pipeline.

HDFS: Store raw NYC Yellow Taxi CSV files for distributed access.

- **PySpark:** Clean data, filter outliers, extract time/location features, and aggregate demand.
- **Pyspark ML** Train scalable regression models with lag and calendar features.
- **Apache Cassandra:** Save forecasts keyed by zone and timestamp for fast retrieval.
- **Matplotlib or Plotly:** Plot actual vs. predicted demand curves for evaluation.

Presentation of Work

- The coding project will be published in Github and made publicly available after submission deadline.
- Additionally, the project and report will be uploaded on Canvas according to the guidelines.