### ****Executive Summary****

**Project Title:** Traffic Optimization for Taxi and Ride-Sharing in NYC using Big Data Platforms  
**Course:** DATA 603 – Platforms for Big Data Processing  
**Term:** Spring 2025

### 1. Business Problem

Urban centers like New York City face immense traffic congestion due to high population density, limited road space, and inconsistent public transportation coverage. Taxis and ride-sharing services, though vital for mobility, often contribute to traffic inefficiencies due to route duplication, idle vehicle time, and poor demand forecasting.

The primary objective of this project is to analyze NYC taxi and ride-share data using big data platforms to:

* Identify congestion patterns
* Discover ride-sharing opportunities
* Optimize routing for better time and cost efficiency
* Enable data-driven recommendations for traffic management

### 2. Dataset Overview

We utilized publicly available data from the **New York City Taxi and Limousine Commission (TLC)**. The dataset spans millions of rides, containing structured information such as:

* **Trip records**: Pickup/dropoff timestamps, location IDs, trip duration and distance
* **Fare data**: Fare amount, tips, surcharges, total payment
* **Geospatial metadata**: Zone-based location IDs with latitude/longitude mapping
* **Taxi Zone Lookup CSV**: Maps zone IDs to boroughs and neighborhoods

The dataset is high-volume (~2GB+ per month) and provides a comprehensive view of urban mobility patterns.

### 3. Methodology

Our analysis followed a pipeline leveraging the **Hadoop ecosystem** and **Apache Spark** to ensure scalability and efficiency. The following steps were executed:

#### a. Data Ingestion & Preprocessing

* Loaded TLC trip data and taxi zone lookup files into **HDFS**
* Converted CSV files to optimized formats for Hive querying
* Cleaned data by removing null, duplicate, and outlier records
* Mapped location IDs to zone names for spatial clarity

#### b. Big Data Processing & Querying

* Used **PySpark** and **Spark SQL** for distributed querying
* Identified peak traffic periods by hour, day, and zone
* Clustered high-demand pickup zones using spatial aggregation
* Computed trip overlap metrics to detect ride-sharing potential

#### c. Analytical Framework

* Used **Hive** to run batch queries on HDFS-stored data
* Built visualizations with **Matplotlib** and **Seaborn** to support findings
* Evaluated ride durations, wait times, and idle periods across boroughs

### 4. Key Findings

* **High-Demand Zones**: Midtown Manhattan, Lower Manhattan, and JFK Airport show the highest concentration of rides during peak hours.
* **Ride-Sharing Potential**: Over **30% of rides** share a common pickup window (±5 minutes) and overlapping destinations, indicating a strong opportunity for pooled ride models.
* **Traffic Trends**:
  + Weekdays (especially Tuesday–Thursday) show the highest congestion.
  + The **5 PM to 7 PM** window consistently experiences the worst traffic delays.
* **Idle Time & Duplicated Routes**: A significant number of taxis travel similar routes within the same time frame, leading to inefficiencies in fuel and traffic flow.

### 5. Recommendations

* **Implement Dynamic Routing Algorithms**: Incorporate live traffic data to optimize trip planning and reduce congested routes.
* **Encourage Ride Pooling**: Provide incentives and in-app ride-matching to promote shared rides, particularly in zones with high overlap.
* **Demand-Aware Pricing**: Introduce surge pricing or discounts during low-demand hours to encourage balanced ride distribution.
* **Zone-Based Resource Allocation**: Deploy vehicles based on predictive demand per borough or zone using historical trends.

### 6. Tools & Technologies

| **Component** | **Tools Used** |
| --- | --- |
| Data Storage | Hadoop Distributed File System (HDFS) |
| Data Processing | Apache Spark, PySpark |
| Querying | Apache Hive, Spark SQL |
| Visualization | Python (Matplotlib, Seaborn), Tableau (optional) |
| Development Env. | Jupyter Notebook, Google Colab |

### 7. Conclusion

This project demonstrates how big data platforms can address real-world urban transportation challenges. By applying distributed processing techniques to a vast dataset, we uncovered patterns in traffic flow, uncovered pooling opportunities, and offered actionable recommendations to optimize NYC’s taxi and ride-share ecosystem.

Our findings support a shift towards smarter, more sustainable urban mobility through data-driven strategies. Future extensions may include real-time traffic integration, predictive analytics using machine learning, and simulation-based evaluation of optimization proposals.