1. Introduction

Project: **Distributed Analysis of NYC Taxi Trip Data Using Hadoop and Apache Spark**

The project focuses on analyzing real taxi trip data from New York City. The data includes millions of records with pickup/dropoff times and locations, trip distances, and durations. The goal of the project is to process and analyze this data using distributed systems technologies to uncover patterns in taxi usage across different areas and at various time intervals.

2. Dataset Description

The dataset is sourced from **NYC Yellow Taxi Trip Data** and includes data about trips made by taxis in New York City. Each dataset contains the following information:

- Pickup and dropoff time
- Pickup and dropoff coordinates (latitude, longitude, zone)
- Trip distance and duration
- Fare amount, payment method
- Number of passengers

The dataset size is **1.71 GB.**
Source: NYC Yellow Taxi Trip Data on Kaggle

---

3. Project Objectives

1. Load the taxi trip data into **HDFS** for distributed storage.
2. Clean and filter the raw data using **Hadoop MapReduce**.
3. Perform analysis using **Spark** to identify traffic congestion, popular routes, etc.

4. **Present visualizations** with analytical results (e.g., heatmap for trip density, graphs for congestion analysis by time of day).
5. **Output analysis results**, such as:
    - Top 20 most popular routes
    - Congested routes and their time intervals
    - Repeated congested routes
    - Most popular payment types
    - Distribution of passengers

---

4. Methods and Technologies

- **Hadoop HDFS**: For distributed data storage.
- **Hadoop MapReduce**: For data cleaning and preprocessing.
- **Apache Spark (SQL, DataFrames)**: For more complex analytical queries and analysis.
- **Python**: For programming and data processing.
- **Google Colab/Matplotlib/Seaborn**: For data visualization.

5. Data Processing Steps

1. **Data Loading**:
   Taxi trip data was loaded into **HDFS** for further processing.
2. **Preprocessing with MapReduce**:
    - Data cleaning (removing trips with zero fare or duration).
    - Parsing timestamps for proper data handling.

Before: (10906858, 19)
After: (10660661, 19)
~250K rows

3. **Data Transformation**:

- Calculated metric average trip duration.
- 13.23 minute

4. **Advanced Analysis with Spark**:

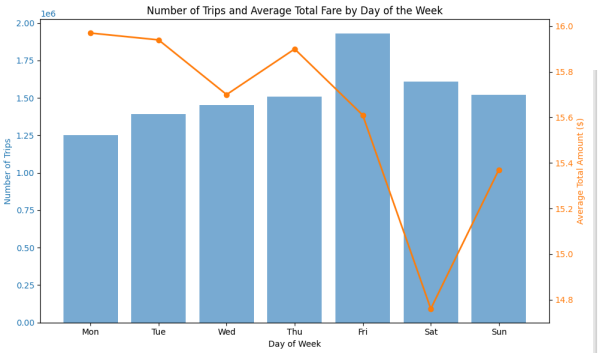
   https://drive.google.com/file/d/1D9lxX8kpUSH2TMKBL3ffn9BVXAoW-E3T/view?usp=drive_link

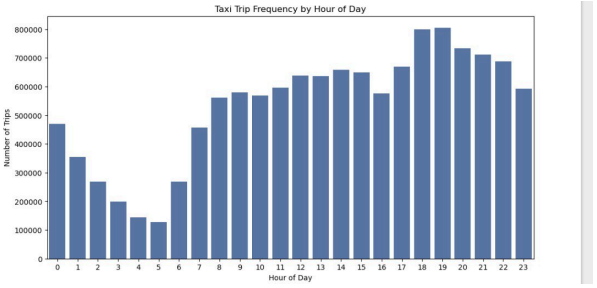
   - Identify the top 20 most frequent routes based on pickup and dropoff coordinates.
   - Analyze congestion by hour, showing the number of trips and average speed for each hour.
   - Calculate the average trip distance and the number of trips per hour.
   - Show the distribution of payment types used for trips.
   - Calculate the average fare and number of trips per day of the week.
   - Analyze the distribution of trips based on passenger count.
   - Group congested routes by both hour and coordinates to analyze congestion patterns.

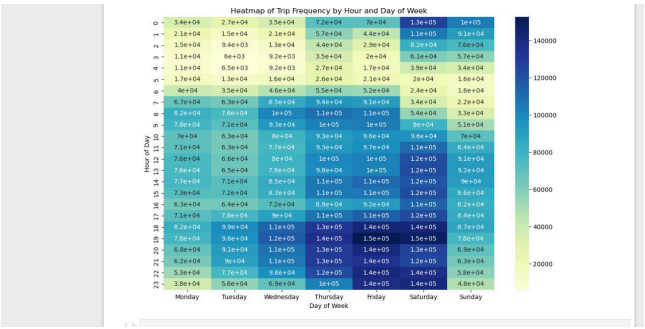
5. **Saving Results**:

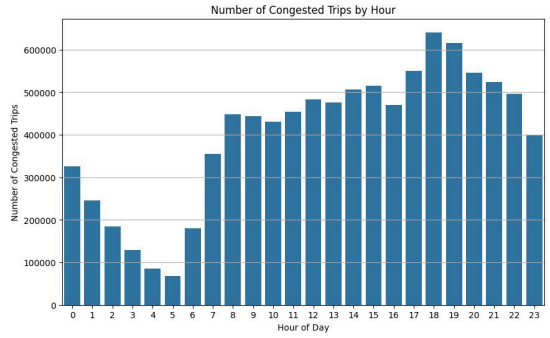
   - All results were saved to **HDFS** for further analysis and use.

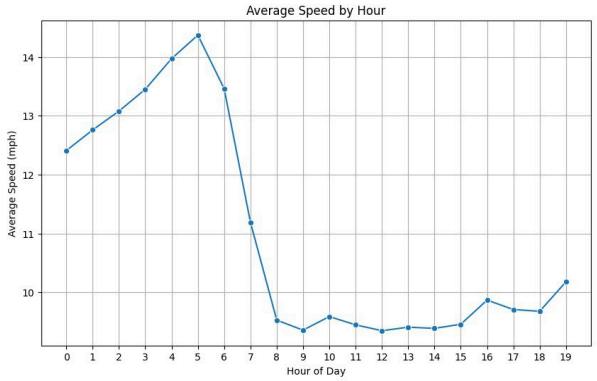
6. Results of visualization

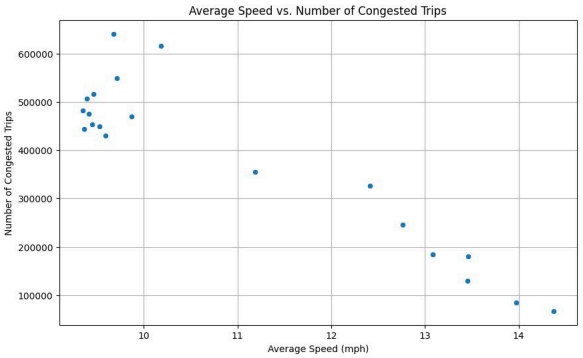


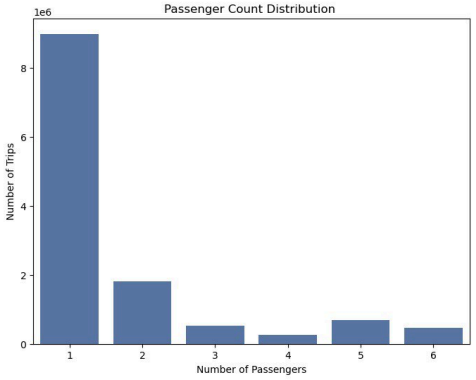


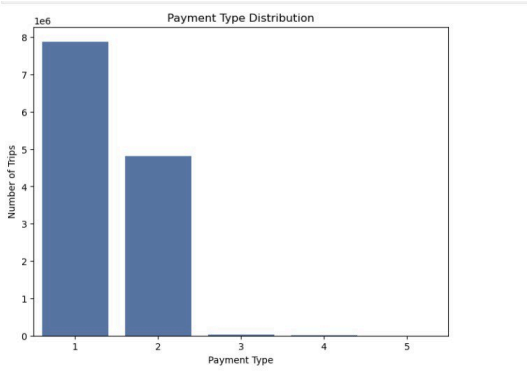


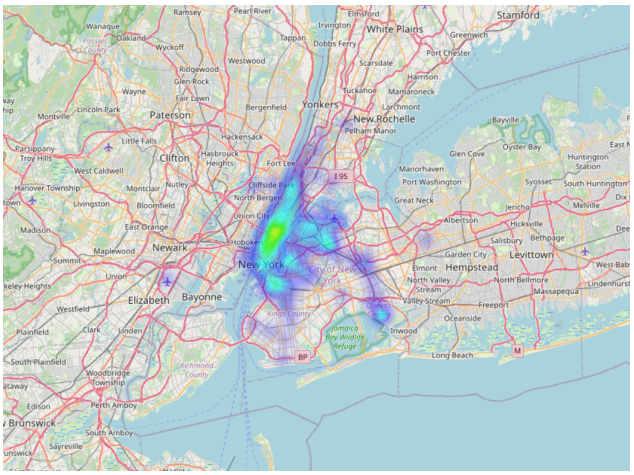


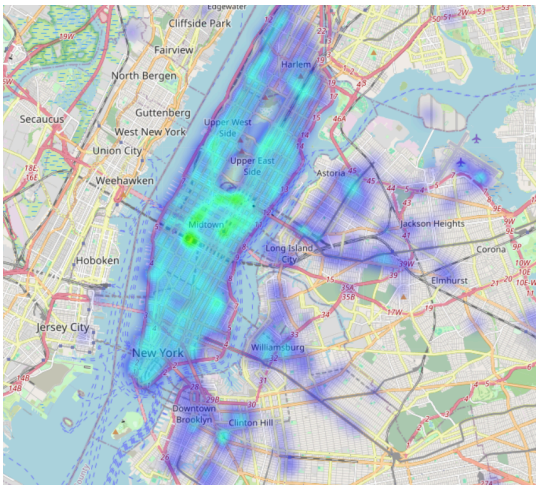












7. Analysis

1. Number of Trips and Average Total Fare by Day of the Week

- Friday sees the highest trips (~1.9 million), likely due to weekend preparations.
- Monday has higher fares (~$15.97), likely from business-related trips.
- Saturday and Sunday show fewer trips, with slightly lower fares, suggesting more local or leisure travel.

2. Heatmap of Trip Frequency by Hour and Day of the Week

- Peak hours (4-7 PM) show the highest number of trips, especially on Thursday and Friday.
- Morning rush (7-9 AM) peaks on Monday and Tuesday.
- Weekends show more evenly distributed trips, with Saturday having higher activity than Sunday.

3. Number of Congested Trips by Hour

- Peak congestion occurs between 5:00 PM and 7:00 PM, corresponding to the evening rush hour.
- Morning congestion begins around 7:00 AM but is not as severe as in the evening.
- Late-night congestion significantly drops after 10:00 PM

4. Taxi Trip Frequency by Hour of Day

- The highest trip frequency is between 5:00 PM and 7:00 PM, mirroring the congestion data.
- Midday (11:00 AM to 3:00 PM) sees moderate trips, likely due to business and tourist activities.
- Late-night trips decrease after 10:00 PM, reflecting reduced demand.

5. Payment Type Distribution

- Credit cards dominate payments, with over 7 million trips, showing preference for digital payments over cash.
- Cash payments are a small portion of the total, reflecting modern payment trends in urban transport.

6. Passenger Count Distribution

- Single-passenger trips account for the majority (~8 million trips).

- Multi-passenger trips are rare, with only a small portion of trips carrying more than one passenger.

7. Central Manhattan Congestion

- Central Manhattan (Midtown, Upper East and West Sides) sees the highest congestion due to heavy commuter and tourist traffic.
- Brooklyn and Downtown Manhattan also show significant traffic, particularly in areas like Williamsburg.
- Queens (Long Island City) shows moderate congestion, influenced by residential and commercial activity.

Conclusion

This project analyzed NYC taxi trip data using Hadoop and Apache Spark to uncover insights into traffic, popular routes, and passenger behavior.

- **Traffic Patterns:** Peak congestion occurs during the morning (7-9 AM) and evening (5-7 PM) rush hours, driven by commuter traffic
- **Popular Routes:** The top 20 routes were identified, with **credit card payments** being the most common
- **Passenger Trends:** Most trips are with **single passengers**, and **Fridays** see the highest trip volume
- **Speed Analysis:** The highest average speed is at **4 AM**, with the lowest speeds at **7 AM** due to heavy morning traffic