NYC Yellow Taxi Trip Record Data Analysis

Shantanu Chaudhari
Mehta Family School of Data Science
and Artificial Intelligence
Indian Institute of Technology
Guwahati, *India*
c.ujwal@iitg.ac.in

Abstract— **The advent of digital technologies has catalyzed a data-driven transformation across various sectors, and the transportation industry is no exception. This study delves into the vast dataset of New York City (NYC) Yellow Taxi trip records to uncover intricate patterns, trends, and behavioral insights. Leveraging advanced data analytics and visualization techniques**

Keywords— Big Data Analytics, Transportation, Data Visualization, Taxi.

I. Introduction

In the bustling metropolis of New York City, the iconic yellow taxis have been an integral part of the urban landscape for decades, weaving through the city streets, carrying passengers to their destinations, and becoming synonymous with the city's pulse. Behind this seemingly ordinary daily occurrence lies a treasure trove of data – a vast repository of information that captures the heartbeat of the city in each trip. With the advent of Big Data technologies, this wealth of data has become more than just numbers and figures; it has become a powerful lens through which we can gain profound insights into the intricate dynamics of urban life.

The project embarks on a journey into this data-rich landscape, leveraging Big Data technologies to uncover hidden patterns, trends, and correlations within the taxi trip records. By harnessing the power of technologies such as Hadoop, Spark, and machine learning algorithms, this analysis delves deep into the vast volumes of data generated by millions of taxi trips.

By deciphering this data, we can unravel the city's commuting habits, discern peak travel times, optimize taxi routes, and even contribute to urban planning decisions. Moreover, this analysis paves the way for data-driven strategies aimed at enhancing the efficiency of taxi services, reducing traffic congestion, and ultimately improving the overall urban living experience.

II. Dataset

The data which will be used in this project is provided by the New York City Taxi and Limousine Commission. This data is released on a monthly basis and include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

III. Implementation

A. Data Storage and Retrival

The data is provided in PARQUET format which makes running queries on data easier. So, the data will be stored as a single or multiple Apache PARQUET files which will be later accessed by the analytics tools down the pipeline.

B. Data Processing

The data retrieved from storage will be processed using Spark SQL, Hadoop MR etc. This data will be processed in batches scheduled every month as the data is coming monthly.

ETL step will also be performed here, along with data enrichment for locations, weather conditions at that time, etc. After this step the cleaned data is stored in a MySQL database for further retrieval

C. Backend

The backend will query the MySQL database to collect the analyzed data and prepare it to be shown in the frontend. It will also be a hub to host various ML models regarding predictions of fares based on geographical location.

It'll transform User inputs into queries which will be used to gather data. The retrieved data will be sent to visualization libraries on frontend to be displayed in form of charts and plots.

D. Frontend

The frontend will consist of an interactive dashboard where the user will be displayed the overall analysis. It will also have various inputs regarding which kind of data is to be displayed, time frame, location etc.

The user can also get a predicted price of their fare using a GUI, where user will put a starting and ending location and will get his fare.

IV. Impact

There can be various impacts of this analysis, some are listed below.

- Predictive analytics can help taxi companies anticipate demand patterns, enabling them to deploy taxis strategically, reducing passenger wait times, and improving overall service efficiency.
- Data analysis can identify areas with high demand, leading to better taxi availability in these locations, enhancing the overall customer experience
- Taxi regulations and licensing policies can be informed by real-time and historical data, ensuring they are aligned with the actual needs of the city.