NYC Taxi Demand Prediction

Abstract

The aim of this project is to predict the demand for taxis in New York City (NYC) using large-scale data processing technologies. The project involves collecting the NYC taxi dataset, pre-processing the data, storing it in MongoDB, performing exploratory data analysis (EDA), building a predictive model using Spark and Python, evaluating the model's performance, scaling the computations using Dask, and visualizing the findings using Matplotlib and Seaborn.

Why is this a Big Data problem?

The prediction of NYC Taxi demand is a Big Data problem due to the large volume of data involved, including structured and unstructured data, such as pickup and dropoff locations and weather conditions. The dataset is constantly growing as new rides are added, and a real-time or near-real-time prediction model is required to accurately predict the demand for taxis.

Dataset and tools

The project involves using Big Data technologies such as Hadoop, Spark, MongoDB, Python, and Dask. The dataset contains records of taxi rides in NYC, including the pickup and dropoff location, timestamp, and fare amount. The project aims to predict the demand for taxis based on various factors such as time of day, day of the week, location, and weather conditions.

Conclusion

In conclusion, this project involves using a variety of Big Data technologies to predict the demand for taxis in NYC. By following the steps outlined in this report, we can create a comprehensive analysis of NYC Taxi Demand, which can be used by taxi companies and policymakers to improve the efficiency of the taxi system in NYC.

Team

1. Jaswanth Sai Nandipati - *jn2652*
2. Sai Teja Reddy Parigi - *sp6923*
3. Venu Vardhan Reddy Tekula - *vt2182*