Faculty of Engineering

Computer Engineering Department

Big Data Proposal

Supervised by

Dr.Lydia Wahid

Presented by

Yousef Mostafa Elmahdy

Gaser Ashraf Fayez

Mohamed Salama

Ahmed Waleed

2023

Problem definition:

the problem definition is to predict the duration of each taxi trip in the test set based on individual trip attributes.

DataSets:

The competition dataset is based on the 2016 NYC Yellow Cab trip record data.
https://www.kaggle.com/competitions/nyc-taxi-trip-duration/data

Planned approach or Proposed solution:

1- Data Preprocessing: The first step would be to preprocess the data, including handling missing values, removing outliers, and performing feature engineering.

2- Feature Selection: Once we have preprocessed the data, we will need to select the relevant features to use in our model. This can be done using techniques such as correlation analysis and feature importance.

3- Model Selection: We will then need to select the appropriate regression model that can predict the taxi trip duration accurately. Some of the models that we can consider include linear regression, decision trees, random forests, and gradient boosting.

4- Model Training via MapReduce: After selecting the model, we will use MapReduce to train the model on a large dataset. MapReduce is a programming model for processing large datasets in parallel across multiple nodes in a cluster. We can use a distributed computing framework such as Apache Hadoop or Apache Spark to implement MapReduce. This will allow us to train our model efficiently on a large dataset.

5- Model Evaluation: Once we have trained our model, we will need to evaluate its performance using metrics such as mean squared error (MSE), root mean squared error (RMSE) and R-squared.