Group Members: Anirudha Tambolkar, Parth Nagori
Unity ID: atambol, pnagori

Data Analytics Pipeline for Taxi Fare Prediction - Project Midreview

Project	Deliverables
List demographic information. 1. Group Number 07 2. Title - Data Analytics Pipeline Taxi Fare Prediction 3. Description - We implement a scalable data pipeline to analyze Taxi fares from New York Taxi and Limousine to predict future taxi fares based on input parameters such as trip duration, location, time etc 4. Team members: Anirudha Tambolkar (atambol), Parth Nagori (pnagori)	The 4 major milestones for this project are: 1. Processing the dataset in parallel which can be directly done in EMR. AWS provides an easy integration between S3 input source and EMR. The output would be dumped in another bucket. This is a change in the project's deliverable as discussed in the proposal. It is acceptable because we can get a simpler design with less number of moving parts. Secondly, the solution we had suggested held true in a non-AWS environment where the developer needs to make all the integrations. AWS provides these integrations by default and we intend to leverage that. 2. ~~Storing the cleaned dataset from EMR on Elasticsearch for faster retrieval.~~ This is no longer required as #1 has changed. We directly feed the data from S3 to Sagemaker. 3. Training machine learning models on the S3 data and then performing hyperparameter optimization using Sagemaker. 4. Provide an API for the user to provide custom input for fare prediction using the trained model.
Status	**Issues**
1. # 1 : In progress - we made good progress but then we found a simpler solution. With this change, we would need to modify/rewrite the logic. 2. #2 : discarded as explained. 3. #3 : In progress - we have explored Sagemaker and its components. 4. #4 : Not yet started	1. Some components in the initial design were not necessary as there exists a simpler way to integrate the flow in AWS. Earlier we had explored Lambda and Kinesis, which for now have been sidelined.