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

Data Analytics Pipeline with Lambda Architecture for Taxi Fare Prediction - Project Overview

Project	Deliverables
List demographic information. 1. Group Number 07 2. Title - Data Analytics Pipeline with Lambda Architecture for 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. Breaking huge csv datasets into chunks for parallel processing. This would involve creating streams of data using kinesis for further feature engineering on a EMR cluster. We would employ the lambda architecture to perform several of these tasks. 2. Storing the cleaned dataset from EMR on Elasticsearch for faster retrieval 3. Training multiple machine learning models on the prepared 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.
Dependencies	**Issues**
1. The projects requires New York Taxi and Limousine’s data dumped into public S3 buckets. 2. The project will be coded in Python and will use common machine learning packages such as Pandas, Scikit, Tensorflow. On the infrastructure side, we will use AWS - Lambda, S3, EMR, Sagemaker. For that we would need an AWS account.	1. We anticipate the choice of attributes for training the ML model would be one of the key factors in getting the solution right. 2. Another issue would be to find the correct ML algorithm. We intend to tackle this using Sagemaker. 3. Making the solution scalable is going to be a big concern. This is going be a design challenge.