Project Requirements Specification Document
MSBX5420
Team La Plata Peak
Spring 2020

Overview (project scope)

We are going to be working with New York City Taxi and Limousine Commission (TLC) Trip Record Data from the Amazon Registry of open data on AWS. We will be ingesting the data into an Amazon S3 bucket. The data is updated as soon as new data is available to be shared publicly. If possible, we would like to ingest only the most recent years worth of data that is available. We are hoping that if the data has been updated recently we could do an analysis of pre and post COVID-19 stay at home order and its effect on New York taxi rides. If data is not available for 2020 we want to calculate some summary statistics regarding the top vendors who gave rides as well as the most common places to be picked up or dropped off. This exploratory analysis can be expanded based on how much time we have left after initially loading the data into our S3 bucket. Once our data is ingested we are planning on using AWS Glue in order to simplify our ETL work and Amazon Athena for querying our now cataloged taxi data. Again, if possible, our end goal would be to look at ride statistics pre and post COVID-19 stay at home order to better understand its effect on the taxi industry. If this data is not available there is plenty of interesting information to work with but we won't know until we start inspecting the data this week.

Functional Requirements

- Ingest data into Amazon S3 bucket
- Basic summary and descriptive statistics
- Basic time analysis, such as which days/times have the highest traffic
- Analysis of pre and post COVID-19 stay at home order

Performance Requirements:

- Horizontal scaling