**Taxi Data Analysis using GCP**

Data regarding taxis ranging from hours travelled, number of passengers etc can be analyzed on a day-to-day basis

In this project, will be focusing mainly on the below problem statements

* Categorizing number of trips a particular cab can make in next year or month
* Fares for a ride are planned based on the most common distance travelled by most of passengers

Hence could help the company to understand and plan their strategies better and earn more profit than it earns at present

**Technology Stack:**

1. Dataflow GCP(Data processing)
2. BigQuery (Data Transformation)
3. GCP ML (Predictive Analytics)
4. Cloud storage (to store Data) and Python for scripting

**Architecture:**

**Diagram

Description automatically generated**

**Process:**

* I will be ingesting batch and stream data into BigQuery tables using Dataflow job and perform aggregations and joins etc. next will visualize using various charts and perform ML algorithms (Linear Regression, Clustering, classification etc.) to predict Fares etc.

**Steps:**

* Created **Cloud storage bucket**

![Graphical user interface, text, application, email

Description automatically generated]()

* Created **Dataflow job** for streaming data from Pub/sub topic to Big query table. while running this job, we can publish messages to topic and data will be inserted in Big Query Table

![Graphical user interface, application, email

Description automatically generated]()

**Pub/sub Topic**

**![Graphical user interface, text, application, email

Description automatically generated]()**

* Applied Big query ML forecasting models (linear reg) by considering latitude and longitude within NYC
* Written Python script to publish multiple messages to Pub/sub topic
* Connected Big query client in Jupiter notebook and done with exploratory analysis by accessing tables and attached notebook in github
* Identified and removed Outliers using IQR(Q3-Q1)
* Applied Random Forest algorithm and multiple linear regression
* RMSE is 1.69