**Indian Airline Fare Prediction Documentation**

Internship Project Documentation

# **Project Plan**

Indian Airline Fare Prediction is Internship Project is an Data Science Project comprising of multiple parts. Indian Airline Flight Prediction is based on Real time streaming data on Indian Airline data which is stored in snowflake using Data ETL pipelines. The Data is preprocessed and used in Machine learning . Furthermore an LLM is used to query data and write responses to interact with users in natural language. Lastly a dashboard Is created to visualize hidden trends and patterns in data.

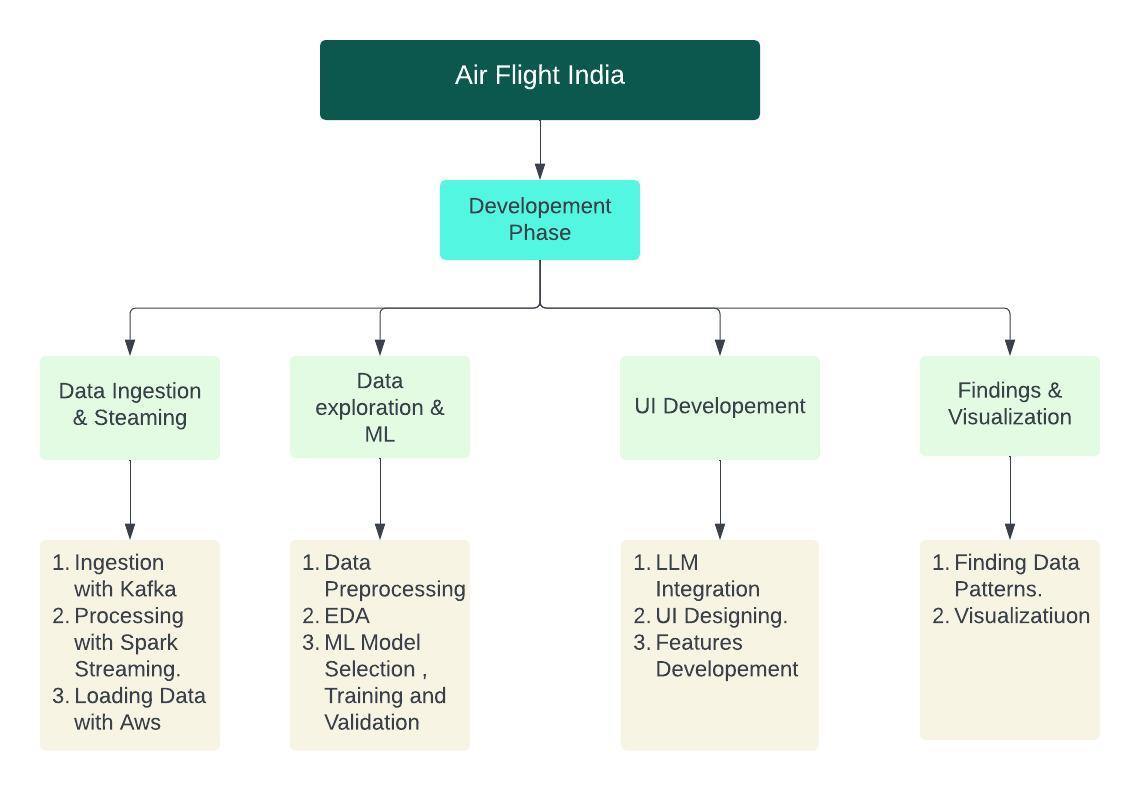
1. Mission Objectives

* Develop a real-time data pipeline for flights data .
* Implement machine learning models for data insights.
* Create a user-friendly interface for natural language querying of data.
* Prepare Visualizations illustrating trends and hidden patterns in data.

2. Management Approach

* Project Management : Assigned Mentor Oversees project milestones and tasks.
* Data Management : Creating a script for ingestion, processing, and loading.
* ML Management : Compare different model performance, training time and size.  
  Document the results and choose one that fits our requirements.
* UI Management: Using Flask querying and visualization interface and deploying it on EC2.

3. Technical Implementation

 Figure 1. Technical implementation

Task 1: Data Pipeline Setup (Weeks 1-2)

* Data Ingestion: Read data from CSV files and write to Kafka topics.
* Data Processing with Spark Structured Streaming: Subscribe to Kafka topics, process data, and save in Amazon S3.
* Loading Data into SQL Database: Use AWS Glue to load processed data from Amazon S3 to Snowflake using a medallion architecture.
* Create script that runs the processed for Data Ingestion and Processing.

Task 2: Data Exploration and Machine Learning (Weeks 3-4)

* Data Exploration: Pre process and analyze data to identify patterns and trends.
* Machine Learning Modeling Train, validate, and evaluate various machine learning models.
* Document various details about modes used and their performance.

Task 3: User Interface Development (Weeks 5-6)

* UI Development:
  + Implement a natural language-based querying interface.
  + Integrate visualization tools for data insights.
* Deploy the model on ec2 providing fulfilling necessary requirements to run model and access it.

Task 4: Findings and Visualization (Week 6)

* Summarize Key Findings: Highlight insights and patterns discovered during data exploration.
* Create Visualizations: Develop interactive dashboards to present key findings.

4. Performance Goals

* Cost Management: Keeping the cost for using AWS services to minimal .
* Schedule Adherence: Complete project phases within the designated timeline.

# **Mission Requirements**

The Airline Flight India is an Internship project made to assess Intern’s performance as well as accomodate intern to industry standard work . There are several mission requirements from Intern as listed below.

* Perform Data Ingestion and make Data Pipelines using unfamiliar technology.
* Perform Machine Learning Operations and work on data from unfamialiar domain .
* Prepare LLM trough Promp Engineering to answer Natural language questions.
* Prepare a Visalisation Illustrating all important findings from the data.

Each Task Will take upto 2 weeks ,except Visualsation which will take 1 week.

**Systems Requirements**

Task 1: Data Pipeline Requirements

The Requirements for Task 1 are -

Ubuntu (Debian based System)

Java JDK version 11

kafka\_2.13 - 3.7.0

pyspark 3.5.1

Aws Lamda

Aws Glue Job

Snowflake Account

The ubuntu based system is a very common requirement through small pcs used in sensors and thus fit for data streaming.Other requirements are up to date libraries for appication.

Task 2: Data Exploration and Machine Learning requirements

The requirements for Task 2 are very less as our work is done on single python notebook and deploying same model -

python version 3.10.12

sklearn version 1.2.2

Ununtu based EC2(t2.small) with same enviroment

Flask framework

For Data Exploration Google Colab is sufficient, but for ML Model interaction. We make use of Flask framework ito ask required data from user to give data to model. Our ML models predicts the price of flight based on given data. Whole application is deployed on EC2.

Task 3: User Interface requirements

Gemini pro LLm API key.

snowflake-connector library & snowflake account

Ununtu based EC2(t2.micro) with same enviroment

Streamlit Framework

The Application is deployed using streamlit framework, this application relies on Gemini pro llm using api reducing the processing power as all processing is done on cloud . The output is generated by parsing query to snowflake-connector .

Task 4: Findings and Visualization

Snowflake Account for data access in Power Bi desktop

Power BI Desktop Application for Data Visualizations

**Data Description**

Airline Flight Data

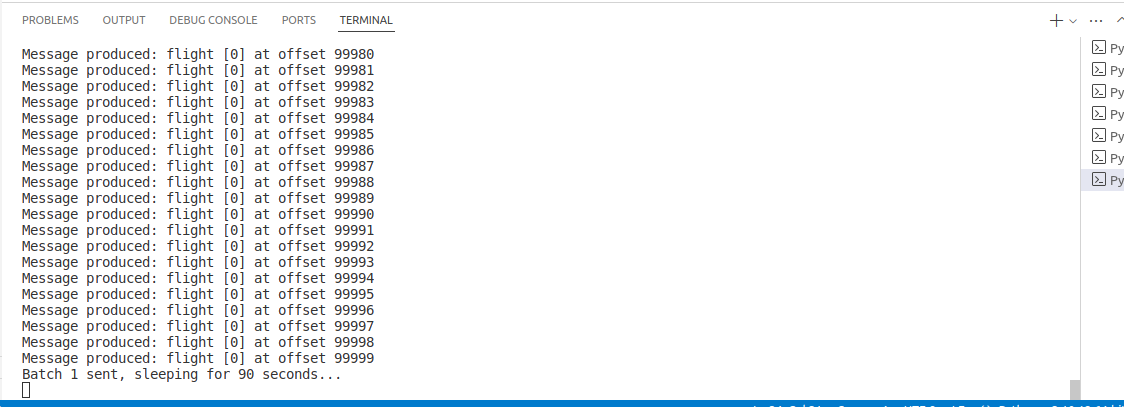
The Project works on flight data set and it provides comprehensive details about various flights, including booking information, flight duration, pricing, stops, and routes. This data is primarily focused on flights within India and this can be interpreted by the name of the cities and for Arrival and departure and Companies. Below is a brief description of each column present in the dataset:

* Date of Booking: This column captures the date when the flight booking was made.
* Duration: The total duration of the flight in number of hours in float type
* Total Stops: Represents the number of stops during the flight, with possible values including 'non-stop', '1-stop', and '2+-stop'.
* Price: The cost of the flight ticket, denoted in Indian Rupees.
* Company: The airline company operating the flight given in data are 'Spice Jet', 'Indigo', , 'Air India', 'Air Asia', 'Missing', 'Vistara', 'Akasa Air', 'Alliance Air', 'Star Air'
* Flight Route: The route taken by the flight, specifying the path from the departure location to the arrival location. Around 200 different routes
* Ticket Class: The class of the ticket, such as economy, business, etc.
* Departure Location: The location from which the flight departs.
* Arrival Location: The destination or arrival location of the flight.
* Journey Date time: The exact date and time when the journey starts.
* Arrival Date time : The exact date and time when the flight arrives at its destination.

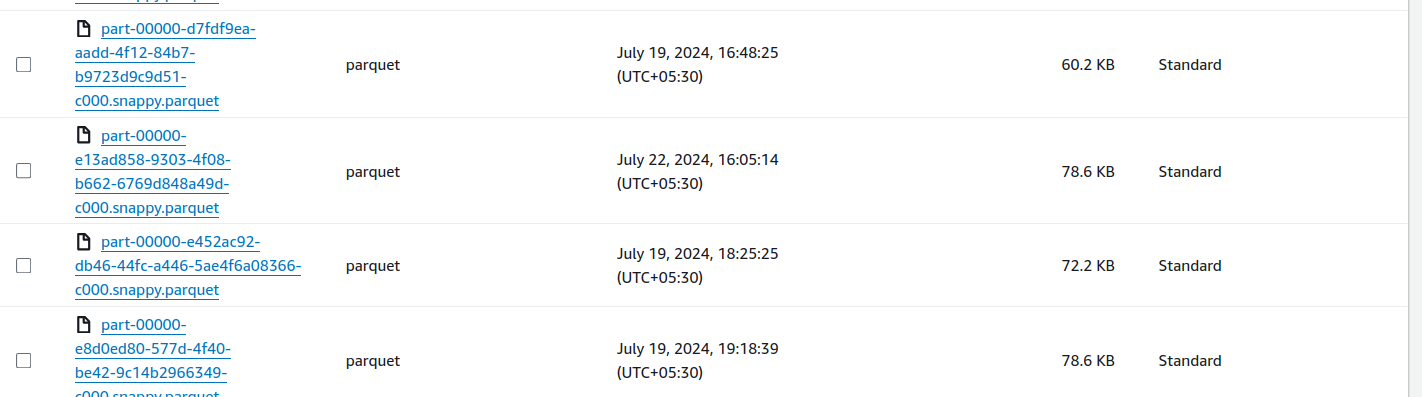
**Validation**

Task 1: Data Pipeline Testing

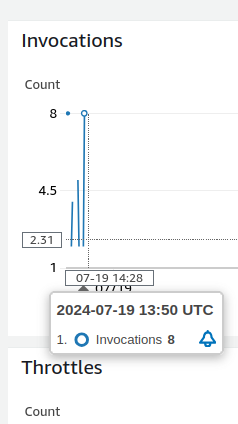
1) Kafka Streaming Data



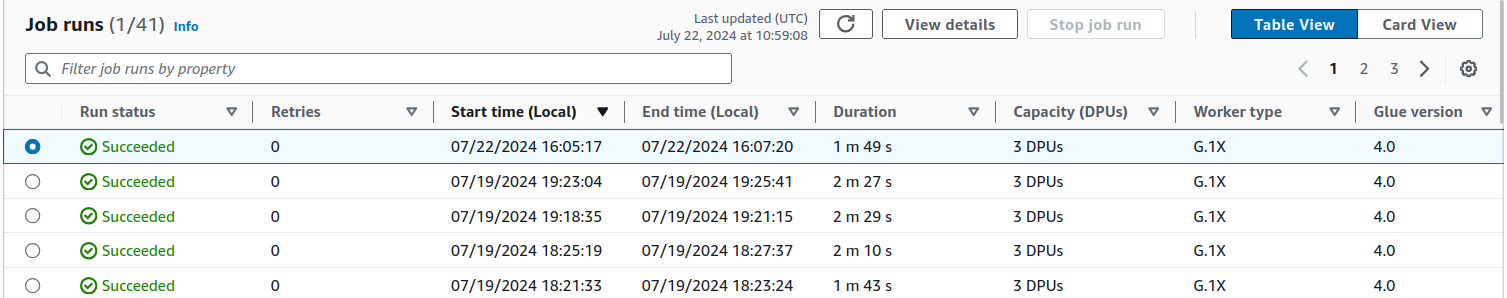
2) Spark Writing Data in S3



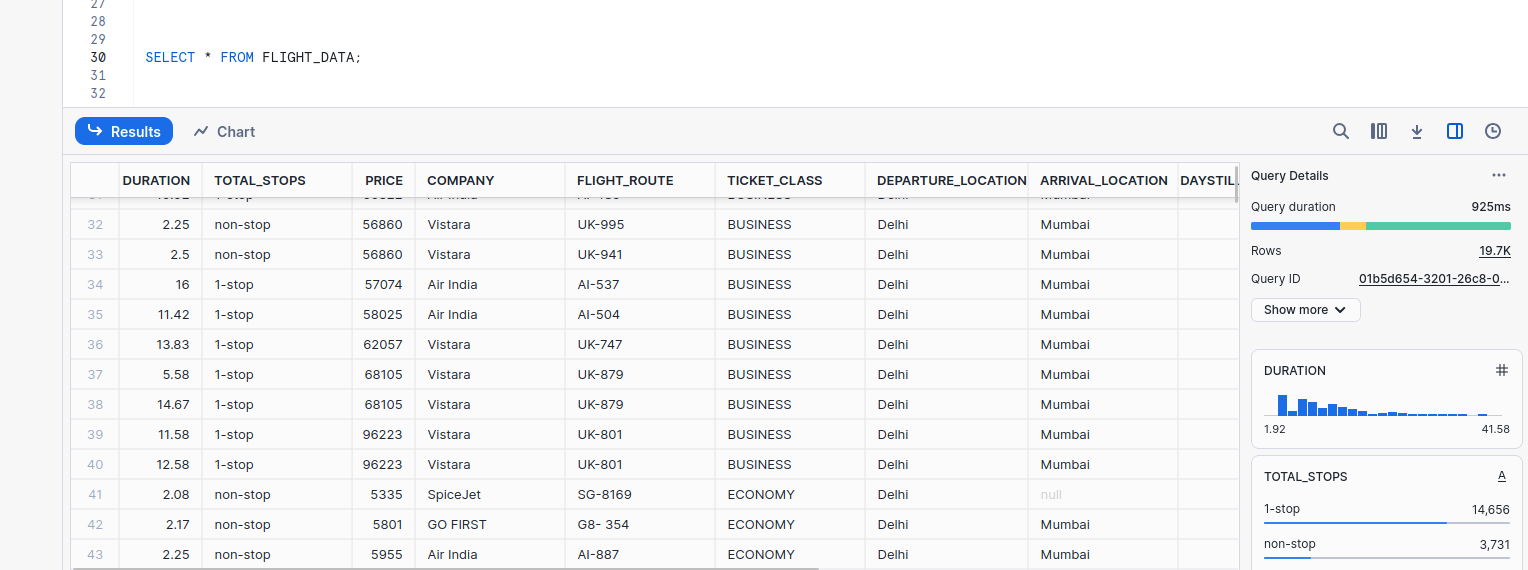
3) Lambda detecting change in S3 and calling Glue Job



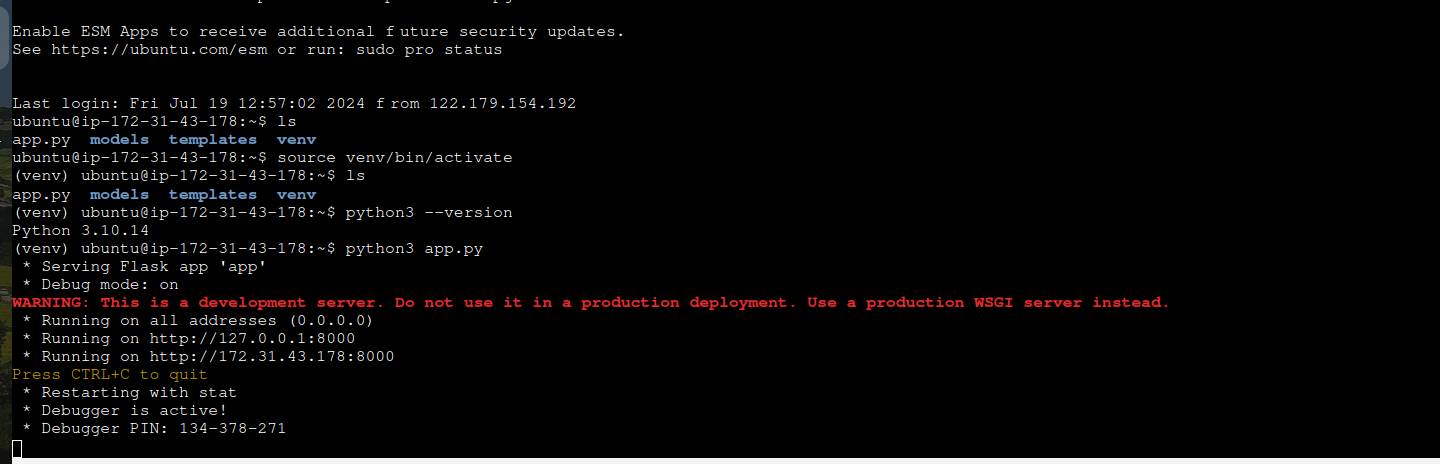
4) Glue Job Running and ETL operations on data and writing in Snowflake

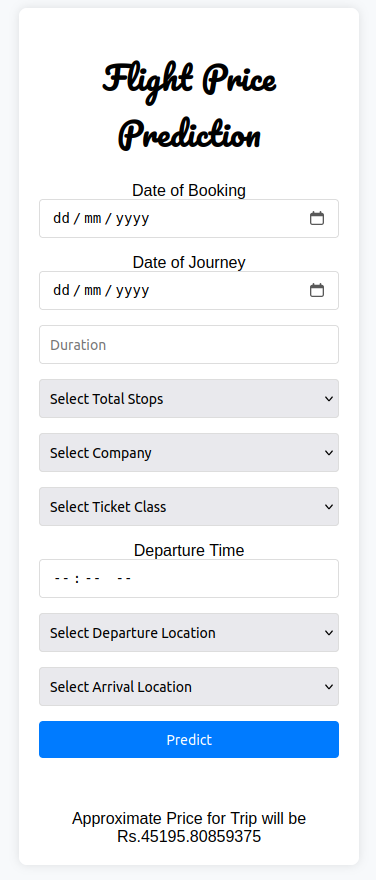
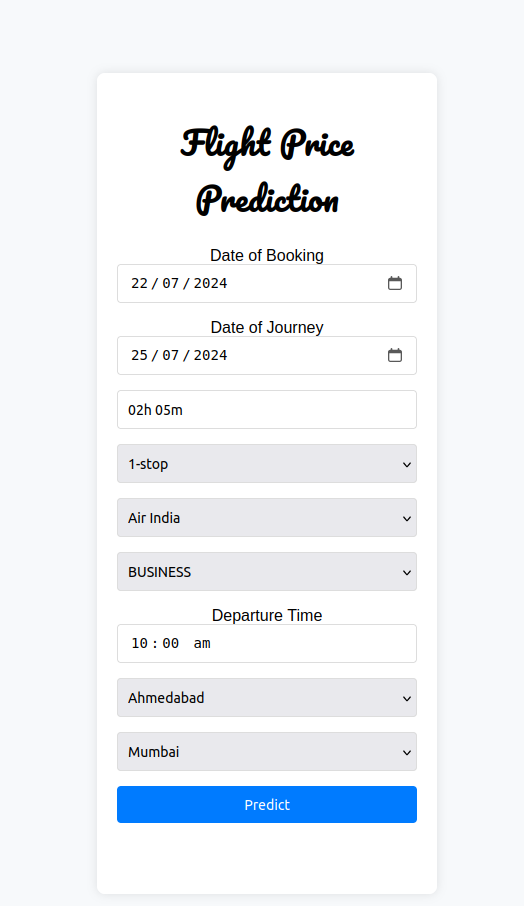


5) Data stored in snowflake



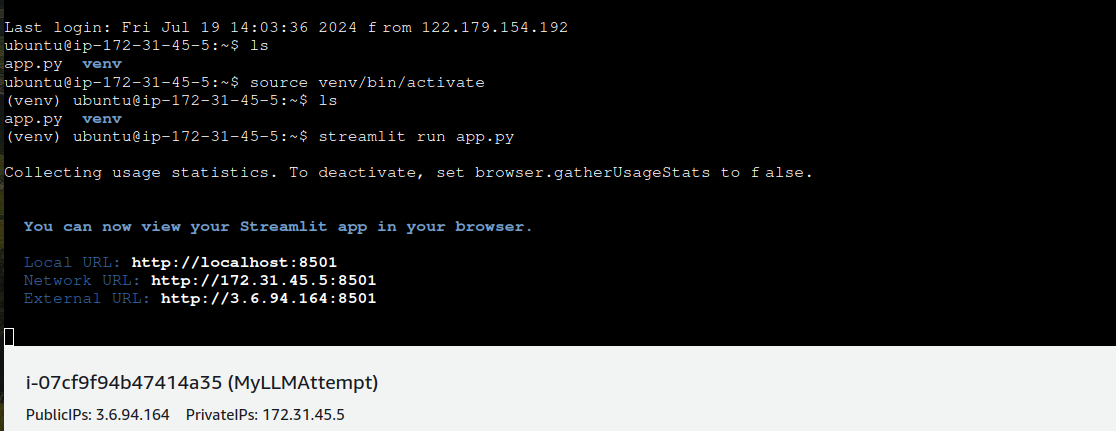
Task 2 Data Exploration and Machine Learning Testing

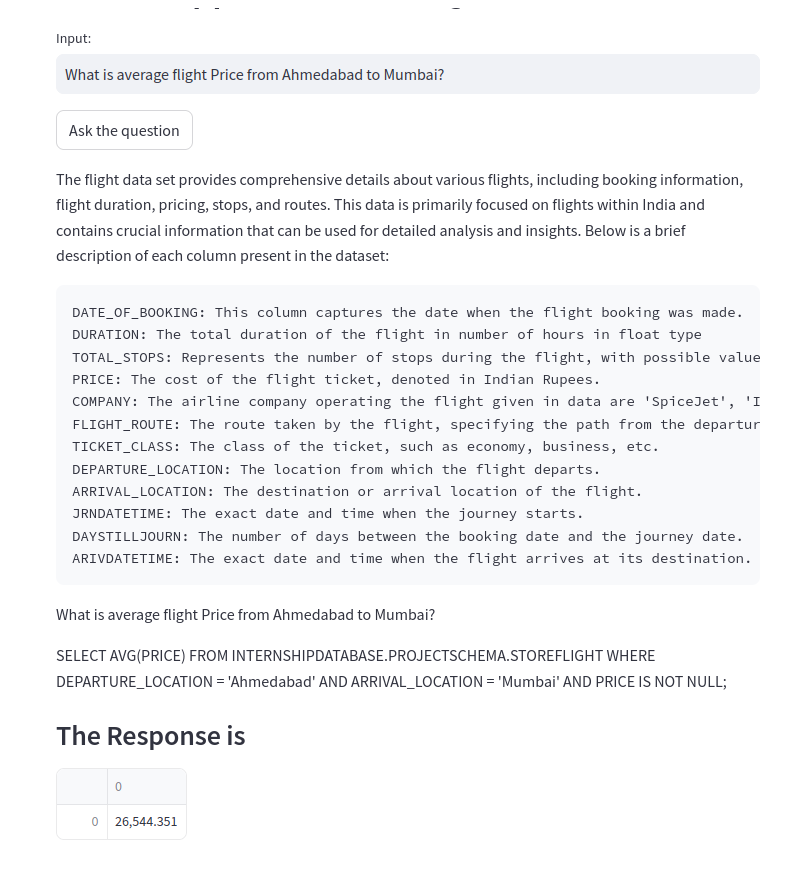


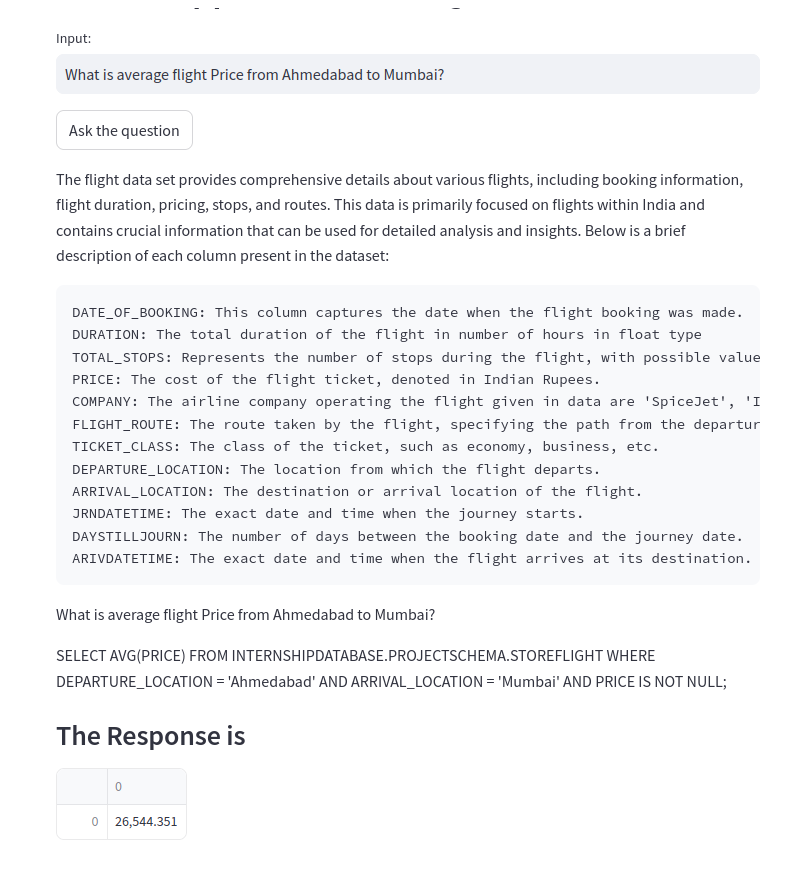


Flight Prediction With ML

Task 3 User Interface



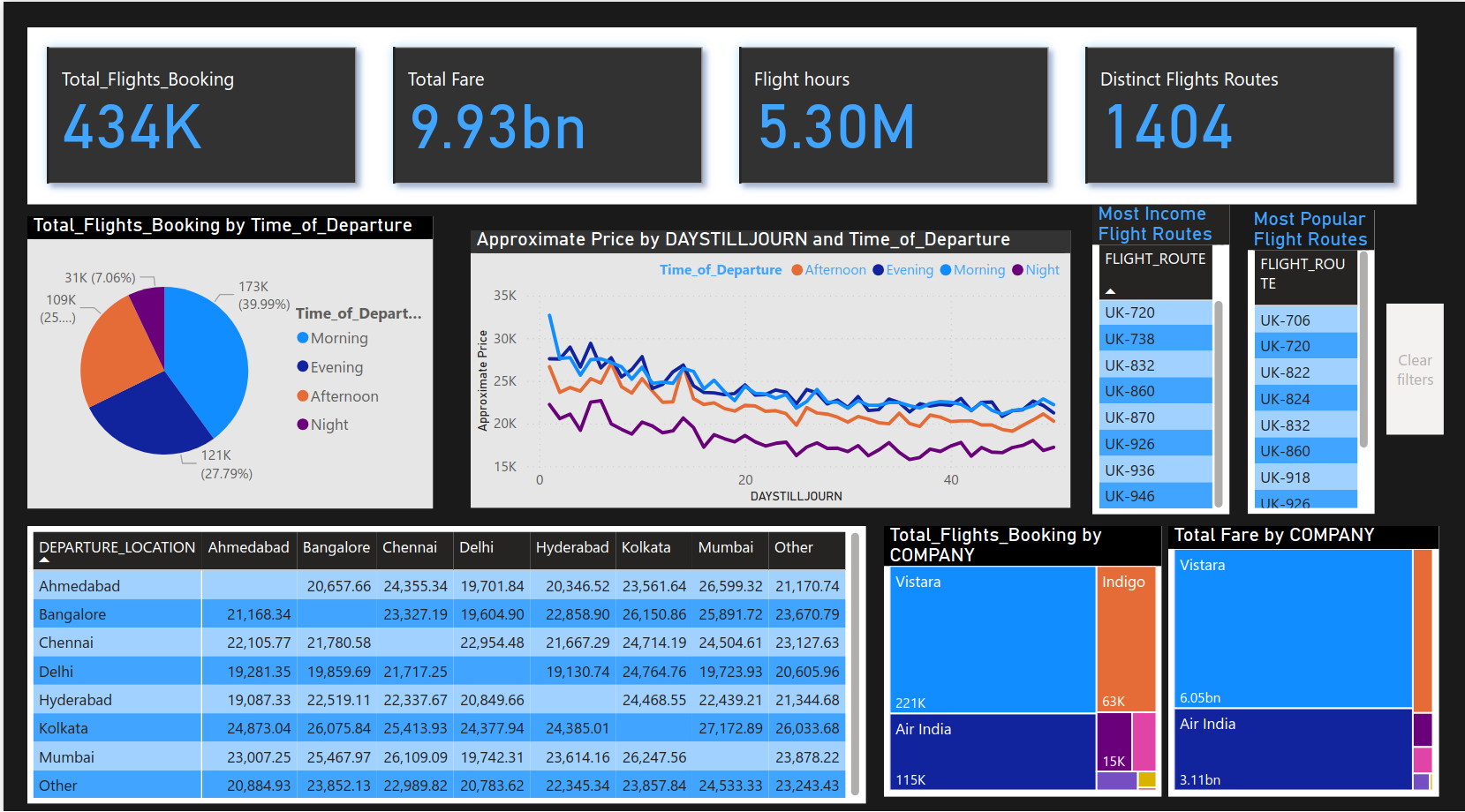


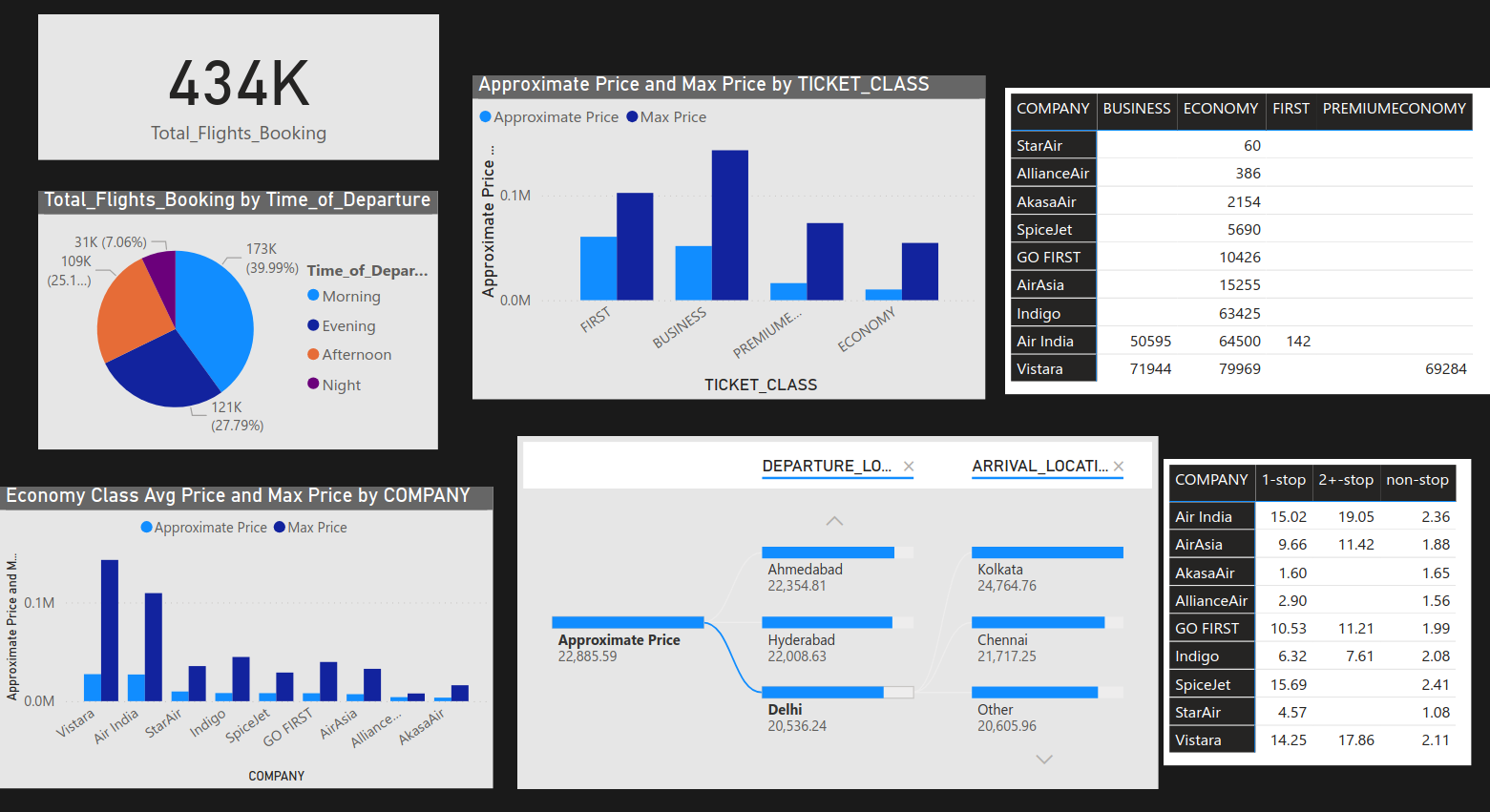


Streamlit Applcation Providing answers to natural language questions and appropriate query.

Task 4 Findings and Visualizations

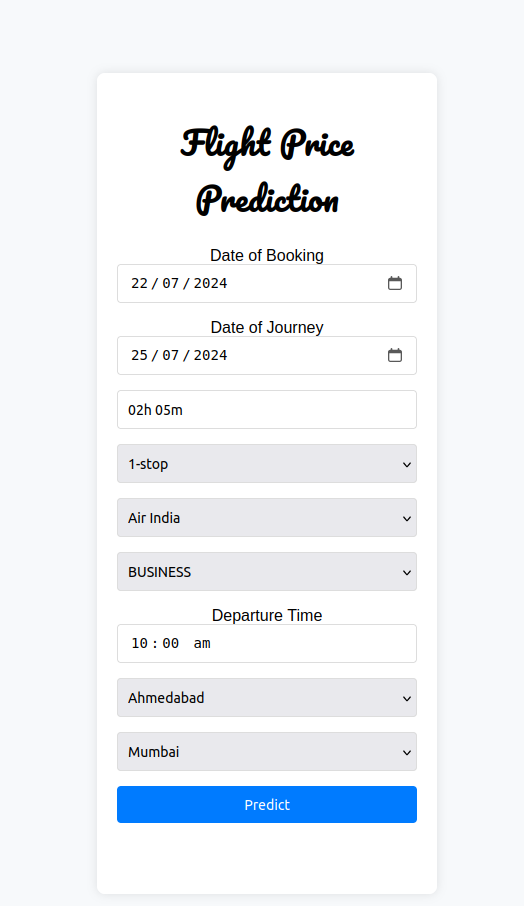
Page 1 Data Visualization

Page 2 EDA Findings

****

**Interface Control**

Task 2 Data Exploration and Machine Learning Testing

****

The Task 4 ML Model UI is strictly taking required data needed for Price prediction.

The Date of Booking and the Date of Journey are required fields to predict the days till journey the user had to wait.

Duration is in ‘Nh Mm’ where N is no of hours and M is number of minutrs

Total Stops , Company and Ticket Class are categorical field selected from limited options given in drop down box.

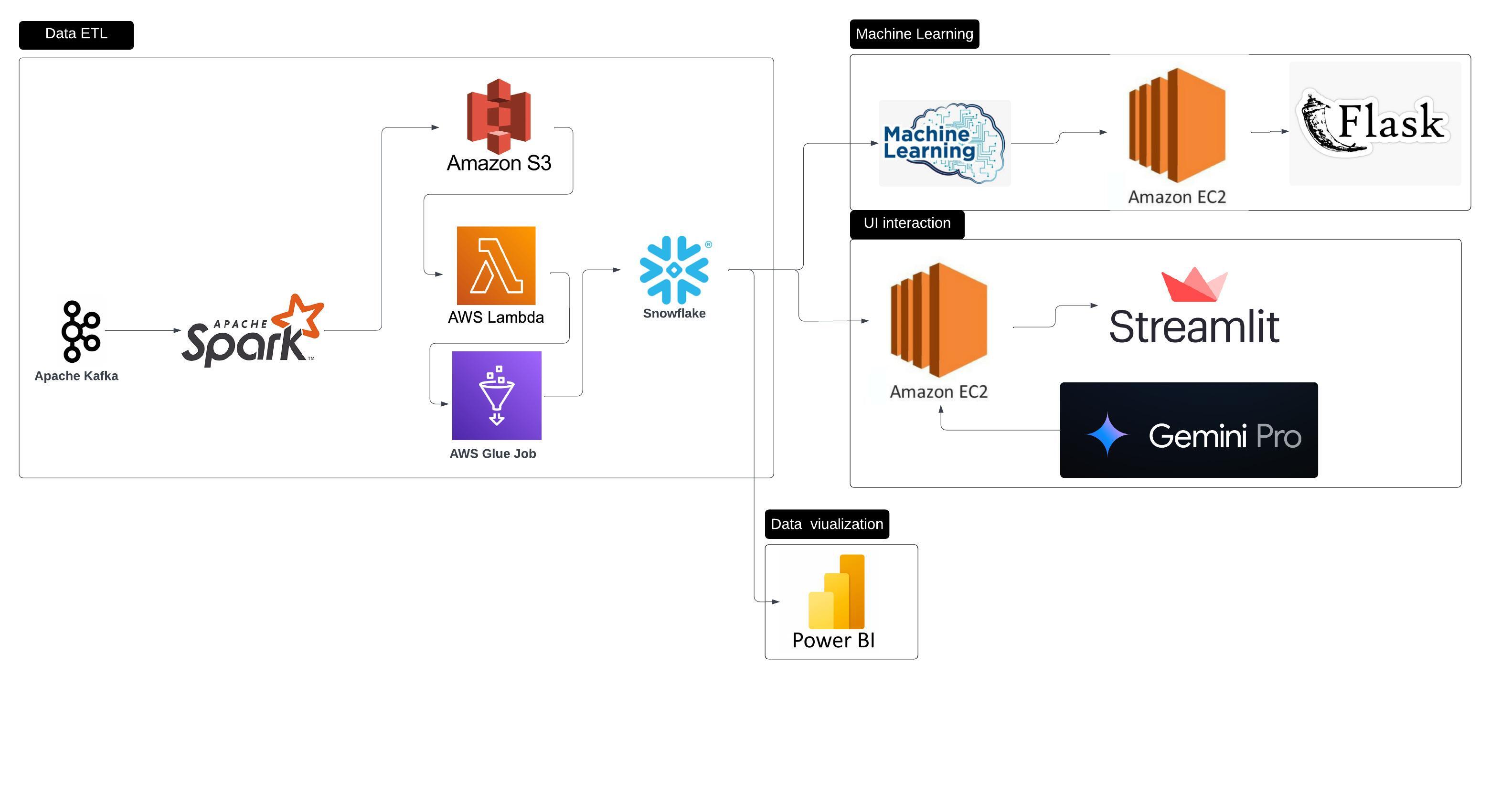
Departure Time is also required time when the flight leaves.

Finally there is Source and destination location.

We have programmed them to not be the same otherwise the model suffers from faulty data input

The Predict Button Runs our program that runs our model on Data.

**Design description**



The Project Task 1 is Data Ingestion and ETL. Kafka is used to stream our data and this data is written in S3 with spark, The next part of ETL is entirely on Cloud .It follows ETL and stores data in snowflake .

In task 2 , we use Snowflake as input and and preprocess the data. Different models are compared and evaluated . The final model selected is XGBoost . XGBoost performs with R2 score of 0.91 and althought it lags behind Random Forest Regressor .Its small size make it excellent for working with small microprocessors and scale up easily.

Our all preprocessing and machine learning models are used in flask application to make a web app that can predict Flight Fare based on User Input.

In Task 3 , we use Gemini Pro api whose task is to convert text to sql . We perform prompt engineering to diirect our model to make right queries. And we pass this query on snowflake. Finally we create a streamlit application doing all mentioned and deploy it on EC2.

In task 4, We perform Visualisation with Power BI Desktop.We make dashboard illustrating information about data and show trends unseen before.