Research Report on Fetching Real-time Data for Airline Flight Price Prediction

**Introduction**The aim of this research is to set up a pipeline for fetching real-time data related to airline flight prices, processing this data, and storing it in Google Cloud Storage (GCS) for use in a predictive model. This report details the process of utilizing Google Cloud Platform (GCP) to achieve these objectives, along with the specific steps taken to fetch data from APIs and store it in GCS.

**Google Cloud Platform (GCP) Setup**1. Google Cloud Storage (GCS):  
GCS was used as the primary storage solution for the real-time data. Buckets were created to organize and store the data efficiently.  
  
2. API Integration:  
APIs were identified and integrated to fetch real-time flight price data. The data was fetched using scripts that were scheduled to run at regular intervals.  
  
3. Data Processing:  
The fetched data was processed to ensure it was clean and in a suitable format for storage and later use in model training.

**Data Fetching and Storage Process**1. Fetching Data from APIs:  
APIs were utilized to gather real-time data on flight prices. Scripts were written in Python to make requests to these APIs, parse the JSON responses, and extract relevant information such as flight prices, departure times, and destinations.  
  
2. Data Processing:  
The raw data fetched from the APIs often required processing to handle missing values, format inconsistencies, and duplicate entries. Data processing scripts were developed to automate these tasks, ensuring that the data was clean and ready for storage.  
  
3. Storing Data in GCS:  
Processed data was then uploaded to Google Cloud Storage. This involved creating buckets and organizing data in a structured manner. Python's Google Cloud Storage library was used to automate the uploading process.

**Pipeline Design**The pipeline designed for this project involves several key components:  
  
1. Data Fetching:  
Scripts scheduled to run at regular intervals to fetch data from APIs.  
  
2. Data Processing:  
Scripts to clean and process the fetched data.  
  
3. Database System:  
A database system (e.g., BigQuery) to temporarily store processed data before uploading to GCS.  
  
4. Data Storage:  
Structured storage of processed data in Google Cloud Storage, ready for use in model training.

**Conclusion**This research outlines the setup of a robust pipeline for fetching, processing, and storing real-time flight price data using Google Cloud Platform. The steps detailed here ensure that data is accurately collected and stored, providing a reliable foundation for building predictive models. Future work will focus on refining the pipeline, optimizing data processing, and integrating machine learning models to predict flight prices.