**I-535 Management Access using Big Data**

**Topic: Analysis of Average Itinerary Fares for Domestic Airlines**

**Author: Kirthivasan PN (kirpandu@iu.edu)**

**1. Introduction:**

Gaining insight into the travel habits and inclinations of various consumer segments requires an understanding of domestic carriers' typical itinerary fares. We can determine pricing and trends in demand through analyzing these fares; this will enable airlines to adapt their offerings to suit the demands of a variety of passengers plus optimize their pricing policies. Airlines can manage resources more effectively and increase revenue by using these data points to determine the routes and destinations draw higher-paying consumers.

Cloud computing platforms, such as Google Cloud Platform (GCP), provide many benefits for data analysis and insights. Airlines can swiftly and effectively process enormous amounts of data via GCP's scalable infrastructure and robust analytical tools. They can extract useful insights from the massive volume of ticketing and itinerary information by utilizing sophisticated data mining algorithms and data visualization approaches. Airlines can increase operational efficiency, boost customer satisfaction, and ultimately increase their bottom line by utilizing the capabilities of GCP to inform their decision-making process.

**2. Background:**

To comprehend the dynamics of air travel throughout the extensive network of airports in the United States, average itinerary fares for domestic airlines must be analyzed. There is a plethora of alternatives available to passengers while planning their travels, with approximately 20,000 airports around the nation, ranging from small rural airstrips to big international hubs. Examining airfare and flight data can provide important new information about how much flying costs and what affects prices at various locations. We specifically want to help travelers make well-informed decisions about their travel arrangements by identifying the most and least expensive airports to fly out of.

Examining the average rate at the major airports—not to mention the most popular 10 airports by passenger traffic—is one of our main areas of concentration. Travelers should be aware of the costs involved in traveling from these busy airports, and airlines should use this knowledge to enhance their revenue management tactics. We'll also look at how average rates are distributed among various airports to expose pricing discrepancies and provide airlines a better understanding of how to charge competitively and transparently. In the end, this analysis seeks to provide data-driven insights that will enable airlines and passengers to negotiate the challenges of domestic flight more skillfully.

**3. Methodology:**

Considering the above factors, a pipeline was orchestrated to extract, transform, and load the raw data. The below steps are implemented:

Setting up a Virtual Private Cloud (VPC) is essential for creating a secure and isolated network environment where resources such as clusters for data processing can be deployed. After configuring the correct properties for VPC, we can set up the resources for the ETL pipeline.

**Data Acquisition:** Collected the data from Bureau of Transportation and Statistics and uploaded the data file in Cloud storage buckets containers in Google Cloud Storage. (GCS)

**Data Processing:** This is performed with the help of Wrangler in Data fusion which helps to explore, clean, and transform data. This is a vital step for building pipelines.

**Data Visualization:** The visualization code was written in Colab to create charts from BigQuery results, providing better insights from the data.

**Automation:** Data processing and loading into BigQuery are automated using Airflow, triggered by file uploads to Cloud storage buckets.

Detailed steps are mentioned below along with code snippets:

1) At the beginning, I started a project in the GCP environment and turned-on billing (this time, through an educational institution coupon voucher). For this project, access to Google cloud services like Big Query was necessary.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**2)** Loading the downloaded Aviation dataset to cloud storage bucket along with the files to automate the dataflow pipeline process.

**A screenshot of a computer

Description automatically generated**

The CSV file from the Bureau of Transportation was loaded into cloud bucket storage to facilitate the data pipeline process. This was automated using Airflow, where DAG (Directed Acyclic Graph) code was written for orchestration.

3) The composer primarily aims to orchestrate workflows by defining tasks and their dependencies. Another essential feature is that it is built on Apache Airflow, an open-source platform for workflow automation. Additionally, it is beneficial in monitoring and alerting the progress of pipelines, task execution, and receiving notifications in case of failures or issues. These features were researched on how the automation can be performed and learned while working on the project.

A screenshot of a computer

Description automatically generated

4) Dataproc was implemented to dynamically provision clusters with the necessary compute resources for data processing tasks. In this project, various data processing frameworks, including Hadoop, were supported by Dataproc for building data pipelines. Batch processing was involved in the project. This setup allowed me to use Data for ingesting, processing, analyzing, and visualizing the uploaded data.

A screenshot of a computer

Description automatically generated

5) This crucial step involves the implementation of the data pipeline through Data Fusion, which seamlessly executes the Extract, Transform, and Load (ETL) process into BigQuery. By leveraging Data Fusion's capabilities, data is meticulously preprocessed and transformed, ensuring that only relevant and refined data is efficiently loaded into BigQuery.

**A screenshot of a computer

Description automatically generated**

**Steps for creating a pipeline:**

**Step1**: Data Fusion Studio is accessed from Google cloud platform, and the Google Cloud Storage is added and configured according to the bucket where the CSV file has been uploaded. This will be the first step in the pipeline to perform the extraction.

A screenshot of a computer

Description automatically generated

**Step 2:**

The transformation step is crucial as it enables the conversion of raw data into a structured and usable format. Through Wrangler in Google Cloud Data Fusion, data can be transformed by applying various operations such as filtering, aggregation, and cleansing. This transformation process enhances the quality of data, making it suitable for decision-making purposes and visualize the data. In the image below, I implemented the filter option by state to compare the average fare and the busiest airport.

A screenshot of a computer

Description automatically generated

Step 3: In the ETL process, loading data into BigQuery as the final step, the aviation dataset becomes easily accessible and can be analyzed alongside massive volumes of structured data using SQL queries. Additionally, BigQuery supports real-time analytics this creates a holistic process for deriving insights from the pipeline.

**A screenshot of a computer

Description automatically generated**

**6) Querying in Big Query:**

**Query1: Selecting the higher and lower fare in dataset**

**A screenshot of a computer

Description automatically generated**

**Query 2: Finding the mean and median of average price.**

**A screenshot of a computer

Description automatically generated**

**Query 3: Total passengers for each state in Q3 of 2023**

**A screenshot of a computer

Description automatically generated**

**Query 4: Average price along with state wise comparsion:**

**A screenshot of a computer

Description automatically generated**

**Automating the Pipeline:**

**A screenshot of a computer

Description automatically generated**

The code written to perform the DAG operation plays a crucial role in orchestrating the data pipeline. By defining the sequence of tasks and their dependencies within the Directed Acyclic Graph (DAG), the code ensures the systematic execution of data processing steps. This automation streamlines the pipeline, eliminating the need for manual intervention and reducing the risk of errors. The airflow detects the necessary code uploaded from the bucket as the code is configured that way.

**Code to process the CSV file:**

A computer screen shot of a program code

Description automatically generatedA screen shot of a computer program

Description automatically generated

**Code to trigger the pipeline after the necessary imports:**

**A screen shot of a computer program

Description automatically generated**

**Code to check the status of the pipeline after the necessary imports:**

**A screen shot of a computer program

Description automatically generatedA screen shot of a computer program

Description automatically generated**

**4. Results:**

**A graph of a number of people

Description automatically generated**

The above image illustrates the average fare data for the top 10 busiest airports during the fourth quarter of 2024. It presents a clear depiction of the relationship between the airport codes and their corresponding average fares this correlation is presented with help of bar chart.

A green and red graph

Description automatically generated

The above image depicts the most expensive airports on the left and most affordable airports and the fares according with help of bar chart.

A graph of passengers in california

Description automatically generated

Above image gives an overall idea about California Airports and ordering them city-wise. Bar chart is utilized to visualize the data.

A graph of a bubble chart

Description automatically generated

The Bubble chart above indicates the intensity of fares in each city in California. This information is top 5 cities. Moreover, considering city-wise analysis gives us an idea about busiest airport located in which part of the city.

A graph with blue circles and white lines

Description automatically generated

The above bubble chart gives the correlation between Average fare and Total Passengers. The intensity of bubble represents the fare distribution for large and small airports.

**5. Discussion**

The horizontal bar chart in the first image compares the average airfares at the top 10 busiest airports, revealing that LAX has the highest average fare while SFO has the lowest. The chart suggests a disparity in pricing strategies or cost factors affecting flight prices at these key airports. Moreover, it indicates that passenger volume, as represented by busyness, does not necessarily correlate with higher average fares.

The second image, a comparison of two sets of images, reveals intriguing insights. The more expensive airports, with average fares ranging from about $200 to nearly $1600, likely reflect their geographical location, limited competition, or specialized routes. In contrast, the affordable airports, with average fares less than $140, may benefit from high competition, greater flight availability, or service by low-cost carriers. The significant variation in fares among the most expensive airports could indicate a broader diversity of service levels or destination access.

The third image, a bar chart, provides a comprehensive view of passenger distribution in California. Unsurprisingly, Los Angeles leads with over 1.2 million passengers, highlighting its status as a major travel hub. San Francisco and San Diego follow, indicating their smaller hub status. The graph also shows a rapid decrease in passenger numbers as we move to smaller cities, providing a clear idea of the passenger volume for each city in California.

For the fourth image, the bubble chart compares the top 5 cities in California by average airfare and passenger volume. Los Angeles has a giant bubble, placed lowest on the fare axis, indicating it has the highest passenger volume and one of the higher average fares, which could suggest a high-efficiency, high-volume economy of scale at play. In contrast, San Jose has a smaller bubble, placed highest on the fare axis, suggesting fewer passengers and comparatively affordable fares. San Francisco's bubble is notably lower on the fare axis than San Jose's, implying higher prices than Los Angeles. Sacramento and San Diego fall between these extremes, showing a mid-range fare and passenger volume.

In the final image, the bubble chart depicts average fares versus total passengers in millions for the top 10 states. California (CA) and Texas (TX) have the most enormous bubbles, indicating they handle the highest passenger volumes among the depicted states, with CA seeing over 4.5 million and TX around 3.5 million passengers. Florida (FL) also has a giant bubble, indicating a high passenger volume above 2.5 million, but with a lower average fare of around $350, reflecting a more affordable destination or origin than TX and CA, which hover around $400 and $450, respectively. New York (NY) and Illinois (IL) have moderate-sized bubbles and average fares in the $400-$450 range, with passenger volumes of around 2 million. Despite their smaller passenger volumes than NY and IL, Georgia (GA) and Colorado (CO) are higher on the fare axis, suggesting pricier air travel. New Jersey (NJ) shows the lowest average fare, below $300, and a relatively small passenger volume, indicating more affordable travel options but less traffic. North Carolina (NC) and Washington (WA) are clustered, suggesting similar passenger volumes and average fares.

**Challenges faced:**

1. This project took me a good amount of time understand the orchestration on how to build a data pipeline. VPC was constantly failing I had to understand properly the concept of master and worker nodes.

2. To understand resources that is required a data pipeline required intense research to understand how it plays a major role in each part of the data pipeline.

3. Even after finding the data fusion the platform to implement this, transformation I faced major issues on how to filter the data and didn’t understand the concept of wrangler until I did few days of research.

4. Automating the pipeline was the most challenging part as code I had to create separate function to start the pipeline in data fusion.

**6. Conclusion**

The research demonstrated the importance of using powerful cloud computing platforms like GCP for processing and querying large datasets. With the help of the above information implemented by using the technologies and skills I acquired through this course gave a base to start and explore the tools and implement the data on the lifecycle and pipeline approach and produce insightful data that will aid the Aviation industry in making well-informed decisions regarding the average prices state and city wise. The pipeline can now extract insights from the Bureau of Transportation Statistics data file for improved utilization. The study also illustrated the need to handle information (preprocessing and transforming) with appropriate data types and the necessity of data visualization to enhance comprehension of patterns and trends.

**7. References:**

1. <https://www.usatoday.com/money/blueprint/travel-insurance/flight-statistics/>

2. <https://www.bankrate.com/finance/credit-cards/domestic-airfare-average-cost/>

3. <https://cloud.google.com/data-fusion/docs/create-data-pipeline>

4. <https://medium.com/@erkajalkumari/flight-price-prediction-a-regression-analysis-using-lazy-prediction-7fc8157284e3>

5. <https://airflow.apache.org/docs/apache-airflow-providers-google/stable/operators/cloud/datapipeline.html>

GitHub link: <https://github.com/KirthivasanPN-hash/Itinerary_fares_analysis_big_data.git>