INFO I535 Course Project: Big Data Concepts and Implementations

International Flight Insights: Crafting a Data Pipeline for Global Travel Trends

Submitted by:

Ravi Sharma (ravishar)

Table of Contents

Introduction

In this project, the aim was to practice and implement what I learned in this class and combine it with my previous experiences in a manner that suits my learning style and experience. The project focuses on building a data processing pipeline and performing analyses using big data technologies on the Google Cloud Platform on one of the suggested datasets i.e. Bureau of Transportation Statistics Flight Performance, I particularly went ahead with the Air Carriers : T-100 International Segment dataset that had like 8500+ rows and 43 features and contained all necessary information like origin city, destination city, carrier, number of passengers, distance and so on about international flights that either had their origin or destination in the United States.

The project seeks to uncover valuable insights into international flight patterns, airlines, passenger traffic, and international hotspots through the utilization of cutting-edge managed and serverless services offered by Google Cloud Platform like cloud storage, Dataproc, BigQuery and Google Looker studio in the end to create interactive and informative visualizations.

Background

I am particularly drawn towards this topic as I have a keen interest in the Aviation industry, and I belong to the group of people who admire these flying machines. This dataset combines the best of my interests i.e. Aviation and Data Science. Additionally, the transport industry in general is not given the attention and credit it deserves, the amount of data-driven decisions that these companies make is humungous, and data is involved everywhere right from pricing decisions to scheduling and it just requires small mistakes that disrupt the entire workflow of the industry. With modern avionics and next-generation auto-pilot systems, the aviation industry is highly dependent on data in some or the other way.

Methodology

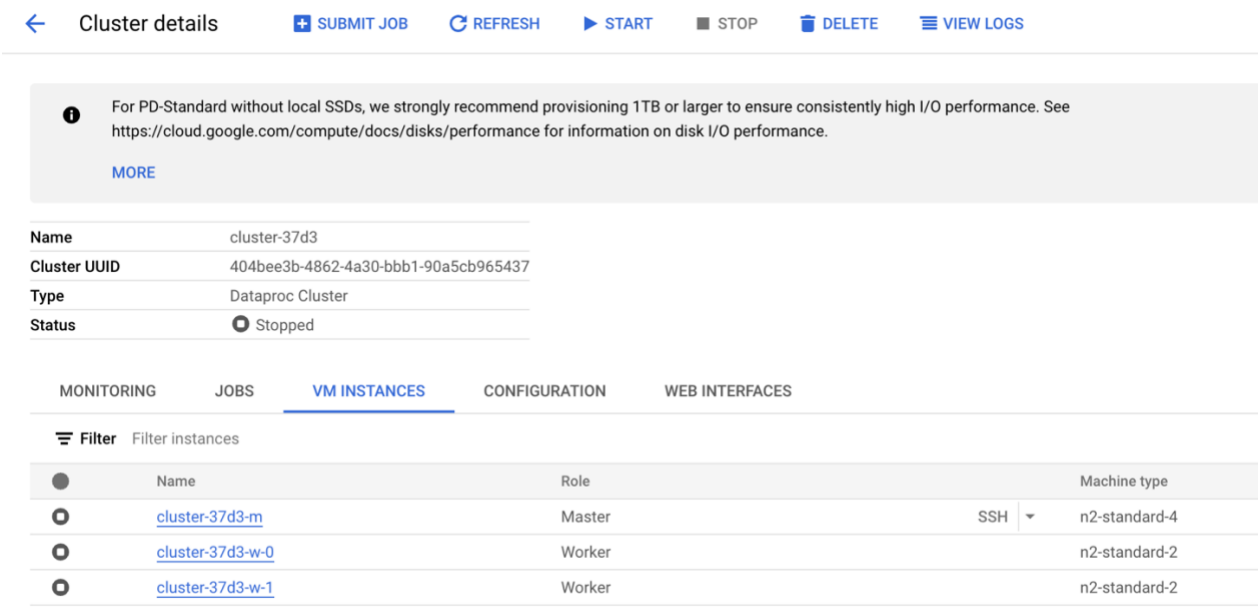
What you will need?

1. Cloud Storage
2. Dataproc Cluster
3. BigQuery
4. Looker Studio

The workflow for this project is as follows :-

1) Data Acquisition: The data used for this project is one of the suggested datasets i.e. Bureau of Transportation Statistics Flight Performance. The dataset for the first quarter of 2024 was obtained by downloading the CSV file from the given link and then stored in the Google Cloud Storage bucket for further use. The dataset provided by the Bureau of Transportation Statistics is well structured and has a documentation file along with it that provides information about all the features present in the dataset. The dataset has a total of 8641 rows and 43 columns, out of which a lot columns had redundant meanings and the dataset indeed needed some cleaning like removing columns containing city ID, airline ID, airport ID, etc. which were practically of no use in this case as we also have corresponding names that makes our lives much more easier.

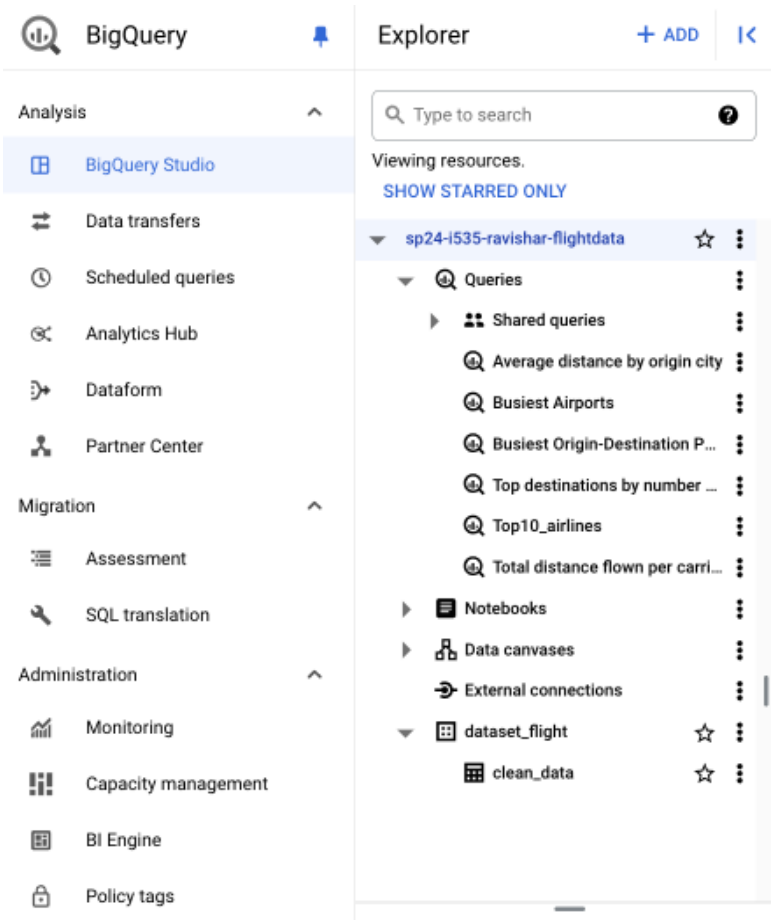
2) Data processing, cleaning, validation, and modelling using Dataproc: The hero for this project has to be the Dataproc cloud service. Google Cloud Platform's dataproc is a fully managed cloud service that provides users with a fast, easy and cost-effective way to run Apache Spark and Hadoop clusters i.e. leveraging the power of cloud computing using distributed storage and computing. Dataproc simplifies the process of deploying, managing and scaling big data processing infrastructure and helps us focus on just writing the logic of the process rather than managing the underlying infrastructure. The users still have control over the size, storage and processing power of the clusters. In this case, keeping in mind the resource quota I used a Master node with n2-standard-4 and two worker nodes with n2-standard-2 with 50GB storage on all of them. The reason why dataproc is well suited for this job is because of the Google Cloud ecosystem that lets us seamlessly connect other powerful services offered by GCP like BigQuery and BigTable.



IMG 1 : Screenshot with cluster details

I used the PySpark service of Dataproc to load, validate, clean, transform and write desired data into BigQuery for further analysis. The Python file, in short, reads the file from the cloud storage and validates that values such as distance, payload and passengers are non-negative and verifies that essential fields such as origin, destination, distance, etc. are not null. Then, the code proceeds to remove duplicate records and logically just keeps the columns that actually make sense and then checks if any columns have the same value for all the rows which in that case raises a ValueError. After all these preprocessing and validation steps, the final dataset is then written into a BigQuery dataset which had to be initialized before the execution of our python file. As DataProc levrages distributed computing and storage all these tasks were done on this large dataset just in 1min 2secs taking into consideration that it takes approximately 60 seconds for initialization of the spark cluster. Though all this may sound very simple and straightforward I faced a lot of challenges regarding networking and setting up the dataproc cluster in the cloud environment which I would like to discuss in the discussion section.

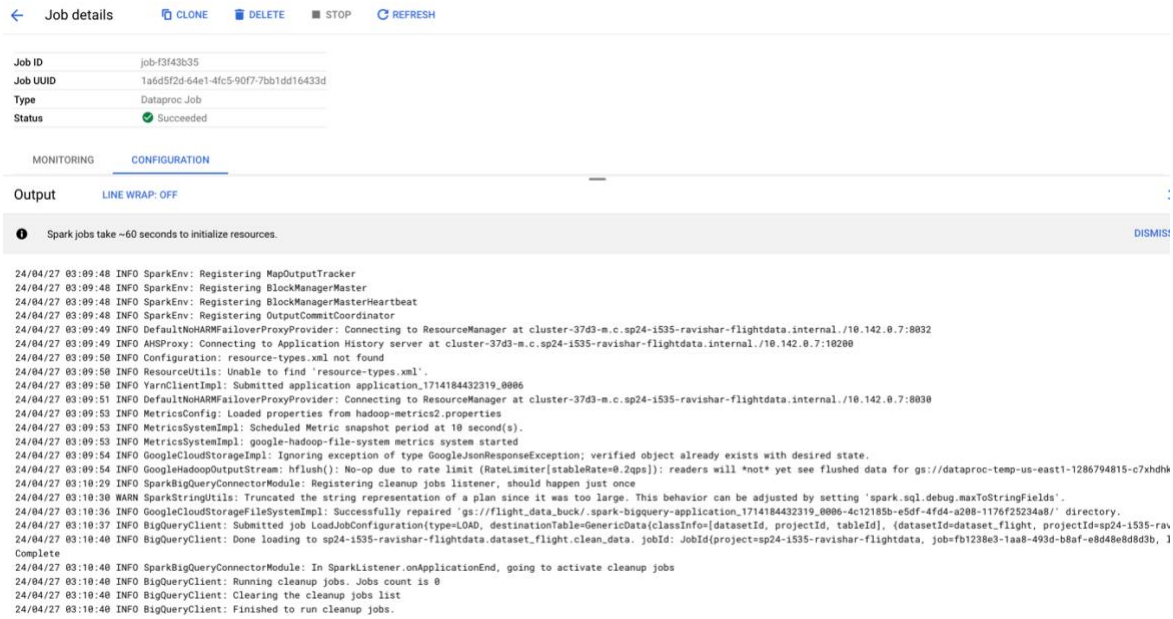
3) Analysis and Visualization: The tools that I utilized for analysis and visualization are BigQuery and Looker Studio. BigQuery is a fully managed and serverless data warehouse solution by GCP, it has been designed with the ability to handle massive datasets with superior scalability. BigQuery makes it really easy to execute SQL queries rapidly across distributed computing resources and supports a wide range of data formats. And connecting Lookers Studio to BigQuery is just a two-click process. Lookers Studio allows us to create interactive visual reports/dashboards like PowerBI and Tableau in GCP. I managed to write a lot of interesting SQL queries on the dataset using BigQuery and create a couple of visually stunning and informative dashboards on Lookers Studio.



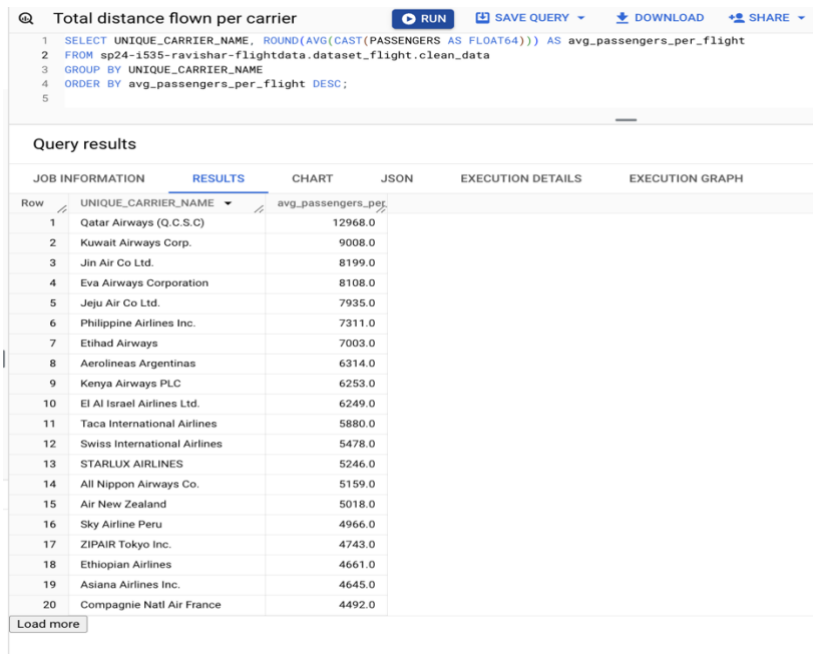
IMG 2 : Screenshot of BigQuery console with dataset, table and queries

Results

After all the data validation and cleaning, what truly matters is how well we managed to extract useful insights from the data. Here are some selected screenshots of SQL queries and reports created using BigQuery and Looker Studio (All files can be found here).



IMG 3 : Screenshot of Succeeded Dataproc Job



IMG 4 : SQL Query to select total distance per carrier

Busiest Origin-Destination Pairs by number of fli... RUN SAVE QUERY

```
SELECT ORIGIN_CITY_NAME, DEST_CITY_NAME, COUNT(*) AS num_flights
FROM sp24-i535-ravishar-flightdata.dataset_flight.clean_data
GROUP BY ORIGIN_CITY_NAME, DEST_CITY_NAME
ORDER BY num_flights DESC
LIMIT 10;
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION

Row	ORIGIN_CITY_NAME	DEST_CITY_NAME	num_flights
1	Bogota, Colombia	Miami, FL	28
2	Miami, FL	Bogota, Colombia	26
3	Los Angeles, CA	Tokyo, Japan	21
4	Tokyo, Japan	Los Angeles, CA	21
5	Miami, FL	Sao Paulo, Brazil	19
6	Chicago, IL	Cancun, Mexico	18
7	Seoul, South Korea	Anchorage, AK	18
8	Hong Kong, Hong Kong	Anchorage, AK	18

Load more

IMG 5 : SQL Query to select busiest origin-destination pairs

Busiest Airports RUN SAVE QUERY DOWNLOAD

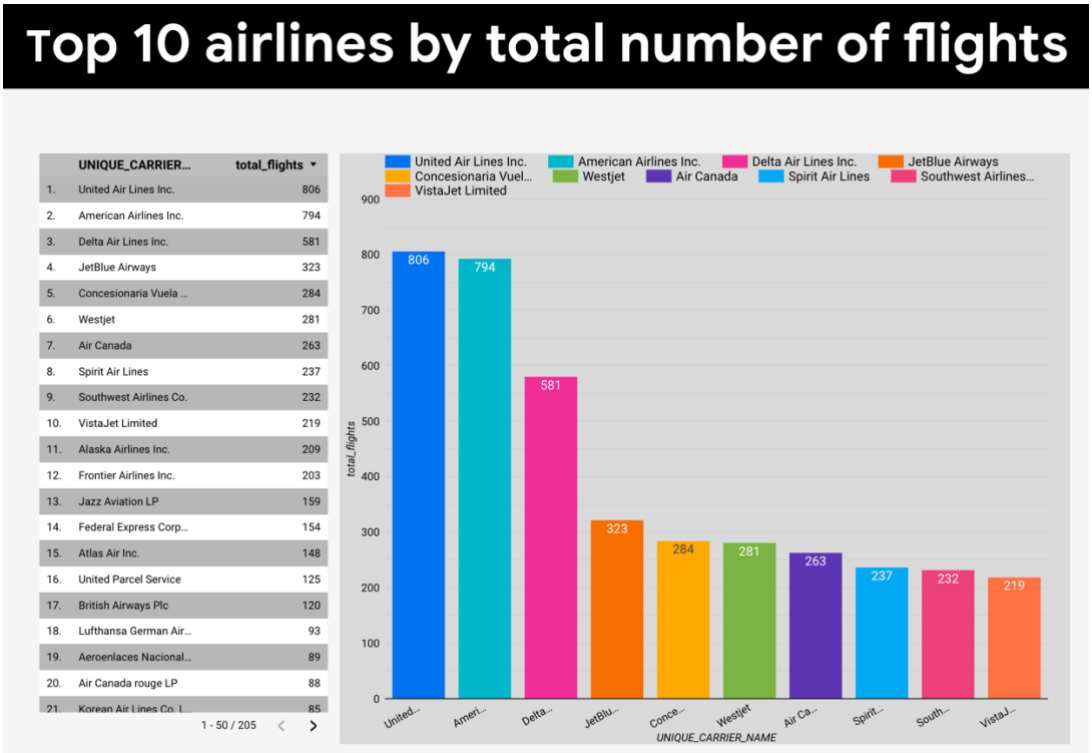
```
WITH city_counts AS (
    SELECT ORIGIN_CITY_NAME AS city, COUNT(*) AS total_departures
    FROM sp24-i535-ravishar-flightdata.dataset_flight.clean_data
    GROUP BY ORIGIN_CITY_NAME
    UNION ALL
    SELECT DEST_CITY_NAME AS city, COUNT(*) AS total_departures
    FROM sp24-i535-ravishar-flightdata.dataset_flight.clean_data
    GROUP BY DEST_CITY_NAME
)
SELECT
    city,
    SUM(total_departures) AS total_count
FROM
    city_counts
GROUP BY
    city
ORDER BY
    total_count DESC
LIMIT 10;
```

Query results

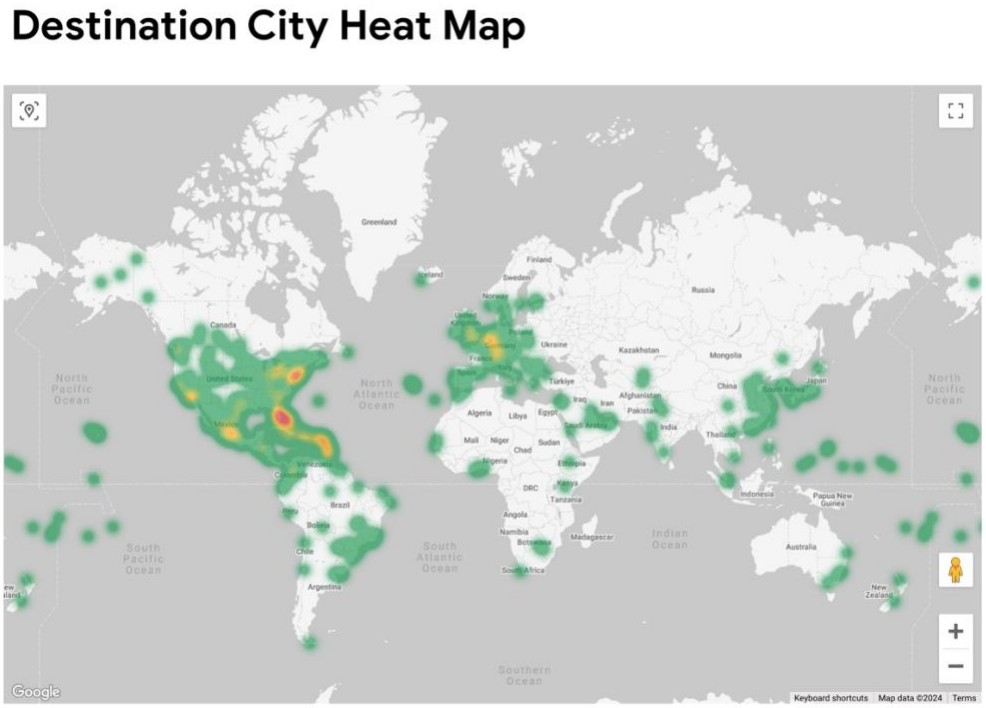
JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS

Row	city	total_count
1	Miami, FL	956
2	Los Angeles, CA	641
3	New York, NY	639
4	Toronto, Canada	496
5	Cancun, Mexico	466
6	Chicago, IL	439
7	Houston, TX	430
8	Fort Lauderdale, FL	342
9	Atlanta, GA	313
10	Newark, NJ	305

IMG 6: SQL Query to select busiest airport

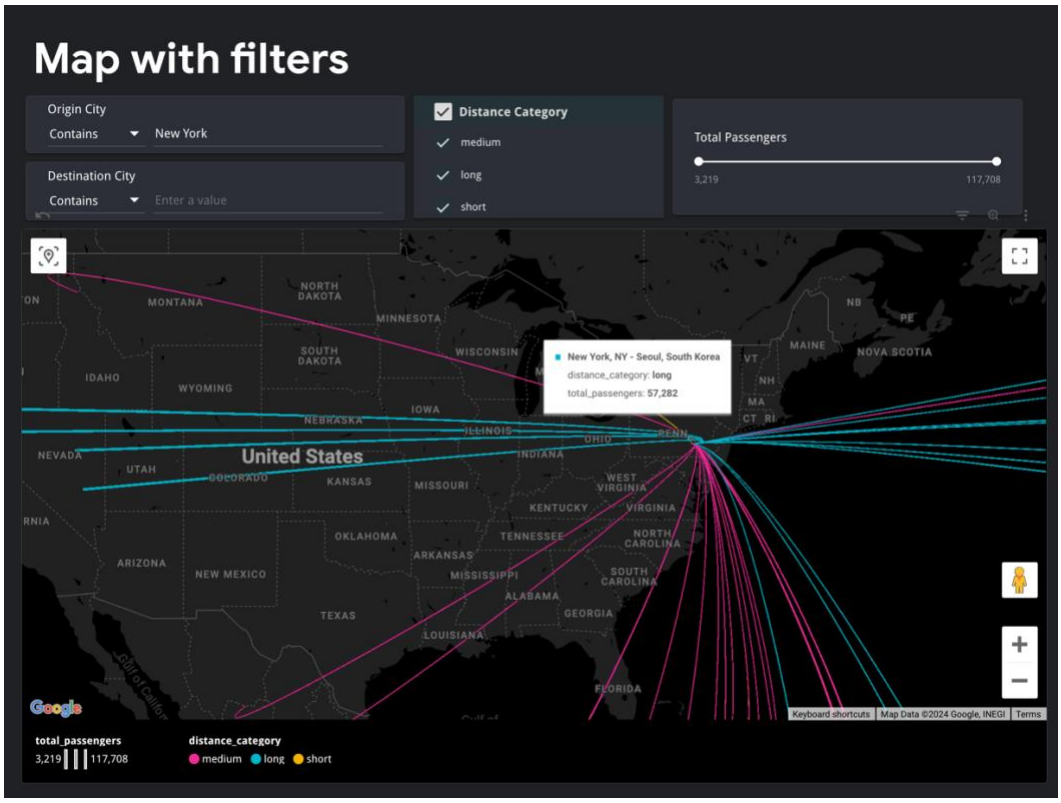


IMG 7 : Looker Studio Report of top 10 airlines by total number of flights



IMG 8 : Looker Studio Report of Destination City Heatmap

IMG 9 : Looker Studio Report of a map with filters showing all the flights with their distance category



IMG 10 : Looker Studio Report of a map with origin country set as new York showing total no. of passengers

IMG 11 : Looker Studio Report of top 10 busiest airport with heatmap

Discussion

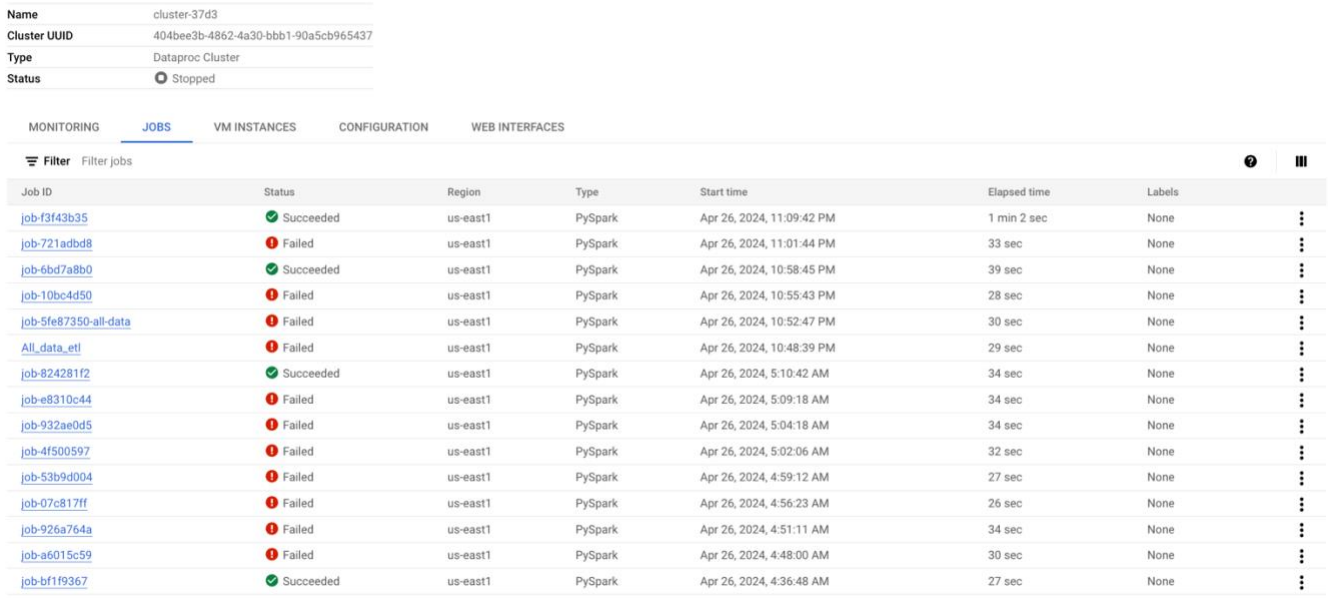
By employing cloud-based infrastructure and distributed computing frameworks, this project showcased how we extract insights from the data using SQL queries as well as interactive dashboards such as Looker Studio. From the amount of analysis, I can interpret that Miami, FL is the busiest airport for 2024 for international flights, with Bogota, Colombia to Miami, FL and Miami, FL to Bogota, Colombia being the routes with the most number of flights. I can also make out that Qatar Airways handles the most amount of passengers than any other airline as far as international flights related to the US are concerned, in contrast, United Airlines is the airline with the most number of international flights. Looking at the destination heat map report, we can see that many flights go to different places worldwide, with Mexico, Germany, and the United Kingdom being some of the most popular spots. Additionally, I noticed something interesting: no flights are operating directly to Ukraine, Russia, and their neighboring countries, likely due to current tensions and sanctions. On the other hand, I observed a lot of flights going to the islands west of the US, possibly around Hawaii.

In this project, I applied several technologies and skills learned from the course to effectively design and handle a data pipeline to help analyze international flight data. I utilized cloud-based infrastructure which was an integral part of the course with a couple of labs revolving around the Google Cloud Platform which indeed taught me how to use Dataproc as a tool to leverage concepts from distributed storage and computing.

I did face a lot of challenges while setting up this data pipeline, the failures began from not being able to create and provision a dataproc cluster because of multiple reasons that took a lot of time and effort. I spent more time correcting those errors rather than analyzing the data. Problems began when I could not even create a simple instance because the default vpc that is usually present in startup labs was missing. After all, the project was under the IU organization and the default vpc option was turned off because of security purposes and it took a lot of time to figure things out.

But, even after that the challenges didn't end, as I was constantly facing an error saying the firewall rules for specified network or subnetwork would likely not permit sufficient communication in the network or subnetwork for Dataproc to function properly, and even after following all the steps in the guide document I couldn't find a solution to this problem for a very long time.
In the end, I managed to overcome this problem by modifying the firewall rule to allow specific ports on TCP, UDP and ICMP as specified by the Dataproc networking document.

After finally getting my cluster up and running, I tested a demo Python code, which worked seamlessly. However, when I ran my actual code, I encountered a strange error with the message "The billing account for the owning project is disabled in state closed." It turned out that the error was not related to billing at all; in fact, it was simply due to an error in writing the correct bucket name. However, the screenshot I've provided clearly illustrates my struggle with this misleading error message.



Name	cluster-37d3
Cluster UUID	404bee3b-4862-4a30-bbb1-90a5cb965437
Type	Dataproc Cluster
Status	Stopped

MONITORING | JOBS | VM INSTANCES | CONFIGURATION | WEB INTERFACES

Filter: Filter jobs

Job ID	Status	Region	Type	Start time	Elapsed time	Labels
job-f3f43b35	Succeeded	us-east1	PySpark	Apr 26, 2024, 11:09:42 PM	1 min 2 sec	None
job-721adbd8	Failed	us-east1	PySpark	Apr 26, 2024, 11:01:44 PM	33 sec	None
job-6bd7a8b0	Succeeded	us-east1	PySpark	Apr 26, 2024, 10:58:45 PM	39 sec	None
job-10bc4d50	Failed	us-east1	PySpark	Apr 26, 2024, 10:55:43 PM	28 sec	None
job-5fe87350-all-data	Failed	us-east1	PySpark	Apr 26, 2024, 10:52:47 PM	30 sec	None
All_data_etl	Failed	us-east1	PySpark	Apr 26, 2024, 10:48:39 PM	29 sec	None
job-824281f2	Succeeded	us-east1	PySpark	Apr 26, 2024, 5:10:42 AM	34 sec	None
job-e8310c44	Failed	us-east1	PySpark	Apr 26, 2024, 5:09:18 AM	34 sec	None
job-932ae0d5	Failed	us-east1	PySpark	Apr 26, 2024, 5:04:18 AM	34 sec	None
job-4f500597	Failed	us-east1	PySpark	Apr 26, 2024, 5:02:06 AM	32 sec	None
job-53b9d004	Failed	us-east1	PySpark	Apr 26, 2024, 4:59:12 AM	27 sec	None
job-07c817ff	Failed	us-east1	PySpark	Apr 26, 2024, 4:56:23 AM	26 sec	None
job-926a764a	Failed	us-east1	PySpark	Apr 26, 2024, 4:51:11 AM	34 sec	None
job-a6015c59	Failed	us-east1	PySpark	Apr 26, 2024, 4:48:00 AM	30 sec	None
job-bf1f9367	Succeeded	us-east1	PySpark	Apr 26, 2024, 4:36:48 AM	27 sec	None

IMG 10 : Screenshot of failed jobs

Conclusion

To conclude, this project was all about using my love for aviation and the skills I acquired in my coursework, focusing on analyzing the Bureau of Transportation Statistics International Flight dataset using advanced services on the Google Cloud Platform, leveraging distributed storage and processing capabilities I utilized Dataproc and BigQuery to handle the large dataset with ease. I ensured the dataset's integrity and relevancy for further analysis by applying various validation checks and transformations like restructuring and cleaning. The data architecture followed in this project involved designing scalable data pipelines using cloud-based services for storage, processing, warehousing and visualization. Additionally, I am maintaining a well-documented **repository** that contains all the code files along with links to the Looker Studio reports. Through this project, I not only uncovered valuable insights into international flight patterns but also demonstrated and utilized the power of cloud-based/virtual machine technologies in modern data science.

References

[1] GitHub repository link

[2] Dataproc Qwik Start lab link

[2] Dataproc cluster network configuration link

[3] Dataproc-BigQuery connector link

[4] Debugging – stackoverflow link

[4] BigQuery for data analysis link

[5] BigQuery to Looker Studio link