**Hrishikesh Sandipan Pawar – A20543095**

**Illinois Institute of Technology**

**ITMD 526-05: Modern Data Warehouse**

**Raj Krishnan**

For this project, I used DuckDB to form datalake and dump the data from my source.

First, Duck DB was installed on the system using ‘pip install duckdb’ in the command prompt.

The Dataset was taken from Kaggle - [Flights](https://www.kaggle.com/datasets/mahoora00135/flights/code)

Overview about the dataset –

* **Time Details**: You get to see when flights are scheduled to leave and arrive, as well as when they actually do.
* **Delays**: It shows if flights are running late, both when they leave and when they land.
* **Flight Info**: You can see which airline is operating the flight and its unique flight number.
* **Route Info**: It tells you where the flight is coming from and where it's going.
* **Flight Length**: You'll know how long the flight takes and how far it travels.
* **Time of Day**: It breaks down flights by hour and minute, so you can see when they're happening.
* **Date and Time**: You get the exact date and time for each flight.

This data is super helpful for airlines and airports to manage their schedules, improve their services, and make sure everything runs smoothly.

**BRONZE LAYER**

* Data is dumped in the duckdb database to form a datalake in bronze layer.

A screenshot of a computer program

Description automatically generated

* This is how the data looks –

A screenshot of a computer

Description automatically generated

**SILVER LAYER**

In this layer, as mentioned in the task, data cleaning was performed using python. All the null values were handled using duckdb queries.

* Null Values

A screenshot of a computer

Description automatically generated

Total of 9430 null values were recorded.

* Since, duckdb doesn’t allow directly to remove/delete null values (entire row). I created new table named – final\_bronze where there were no null values.

A screenshot of a computer error message

Description automatically generated

This step ensures that all the null values are being handled and also there were no duplicate values being recorded.

**GOLD LAYER**

In this layer, I created fact and dimension tables to divide my one whole data set into set columns for facts and dimensions.

* First, I created dimensions table –

A screenshot of a computer

Description automatically generated

Total of 4 tables were created namely – dim\_airports, dim\_flight\_schedule, dim\_airline, dim\_aircraft.

* After creation of these tables, I populated these tables using my final\_bronze table.

A screenshot of a computer

Description automatically generated

Using insert statements, the tables were populated with data from the main final\_bronze table.

* Now, after creation of dimensions table. I created Fact table for the data.

A screenshot of a computer

Description automatically generated

This step ensured that I created fact table named – fact\_flights.

The facts table typically contains quantitative data related to events or transactions.

* Then, I exported all the tables in the form of csv to load in power bi for further visualisation task.

A screenshot of a computer

Description automatically generated

**Data Visualisation using Power BI**

* Created the model using the above mentioned table and formed relationships between them.

A screenshot of a computer

Description automatically generated

The above model showcases relationships between facts and dimension tables.

**Visuals:**

A graph on a computer screen

Description automatically generated

* Top Performers: United Airlines (UA) flies the most distance by far, indicating it likely has many long-distance and international routes. Delta (DL) and American Airlines (AA) follow, suggesting they also have extensive flight networks.
* Market Coverage: The steep drop in distance after the top few carriers shows that only a handful of airlines handle most of the long-distance flights. This suggests that the leading airlines dominate international or cross-country markets.
* Strategic Implications: Airlines flying more distances might focus on global connectivity and frequent long-haul services, which could attract more business travellers or international tourists.
* Environmental Considerations: Airlines covering larger distances could have greater impacts on the environment. This might influence their strategies around sustainability and fuel efficiency.
* Opportunities for Smaller Carriers: Airlines at the lower end of the distance scale might primarily operate regional or short-haul flights. There could be opportunities for them to expand into less competitive long-haul routes or improve their services on existing routes to attract more passengers.

A blue and yellow pie chart

Description automatically generated

* This pie chart displays the distribution of flight origins among three major New York City area airports: JFK, EWR (Newark), and LGA (LaGuardia). Here's a straightforward explanation:
* JFK (John F. Kennedy International Airport) is the busiest of the three, with about 35.88% of flights originating from there. This suggests that JFK handles a significant portion of both domestic and international flights.
* EWR (Newark Liberty International Airport) closely follows, accounting for 33.04% of flight origins. This indicates that Newark also serves a substantial number of flights and possibly serves as a major hub for flights to and from Europe.
* LGA (LaGuardia Airport) handles 31.08% of flights, which is slightly less than JFK and Newark. LaGuardia primarily serves domestic flights and is known for short-haul, regional connections.

A pie chart with many different colored circles

Description automatically generated

* This pie chart represents the total number of flights operated by various airlines. Each slice of the pie corresponds to a different airline, with the size of each slice indicating the proportion of total flights that each airline operates. Here’s a simplified breakdown:
* Southwest Airlines Co. operates the largest portion of flights at 17%, making it the airline with the most flights in this dataset.
* Delta Air Lines Inc. follows closely with 14.35% of total flights, indicating it also has a substantial operational scale.
* United Air Lines Inc. handles 16.18% of flights, showcasing its significant market presence.
* Other airlines like American Airlines Inc., JetBlue Airways, and Alaska Airlines Inc. have smaller but still notable shares, ranging from about 5% to 8% of flights.

A colorful circle with text

Description automatically generated

* This donut chart represents the percentage of total departure delays attributed to different airlines. Each segment's size indicates how much of the total delay time each carrier is responsible for. Here’s a breakdown of what the chart shows:
* EV (ExpressJet) has the largest segment, causing about 24.69% of the total departure delays. This suggests that ExpressJet has significant delays compared to other airlines.
* B6 (JetBlue Airways) and UA (United Airlines) also have substantial portions of delays, 16.99% and 16.9% respectively, indicating that these airlines also frequently experience delays.
* Other airlines like WN (Southwest Airlines) and AA (American Airlines) contribute less to the total delays, with 5.15% and 6.64% respectively.