CI7320 – Databases and Data Management
Coursework 2

Kingston University

Kantheti Dhana Shravya - K2247431

Table of Contents

I. INTRODUCTION

The objective of the coursework is to create a data warehouse using a star schema and analyse the provided dataset on flight punctuality using Tableau, with the aim of identifying trends and patterns in the data. The report also explores the potential benefits of designing data warehouses and OLAP cubes for the airline industry. The dataset consists of punctuality statistics for flights operating in various airports across the UK, including details on the origin and reporting airports, airlines, and average delays. The data covers monthly records for the year 2021. A discussion on different columns provided in the dataset is given below:

Column	Description
run_date	The timestamp of generation of the report.
reporting_period	Month and year of the generated report.
reporting_airport	The airport for which the data is reported.
origin_destination_country	The country of origin of the flights.
origin_destination	The origin airports of the flights.
airline_name	Name of the airlines operating the flights.
scheduled_charter	Whether the flights are scheduled flights or charter flights.
number_flights_matched	The number of flights for which data is available.
actual_flights_unmatched	The number of flights for which data is not available.
number_flights_cancelled	The number of flights that were cancelled.
flights_more_than_15_minutes_early_percent	The percentage of flights that arrived more than 15 minutes early.
flights_x_to_y_minutes_late_percent	The percentage of flights that arrived between x and y minutes late. (Multiple columns in the dataset)
flights_unmatched_percent	The percentage of flights for which data is not available.
flights_cancelled_percent	The percentage of flights that were cancelled.
average_delay_mins	The average delay in minutes for all flights.
previous_year_month_flights_matched	The number of flights for the same period in the previous year.
previous_year_month_early_to_15_mins_late_percent	The percentage of flights that arrived between 15 minutes early and 1 minute late for the same period in the previous year.
previous_year_month_average_delay	The average delay in minutes for all flights for the same period in the previous year.

Table 1. Description of columns in the dataset

II. ADVANTAGES OF A DATA WAREHOUSE FOR AIRLINE PUNCTUALITY DATA

A data warehouse is an effective technique for analysing and reporting large volumes of data. Data stored in multiple locations or databases can be accessed and transformed using ETL methods and stored in a single location which is called a “Data warehouse.” Data warehouses are classified as OLAP databases. In contrast to relational databases, data stored in the data warehouse is denormalized and unstructured. Also, allowing data duplication is a benefit as it reduces the amount of time required to fetch records from the database. The main objective of data warehouse design is easy querying so that users can quickly access information from a large source of data for further analysis. Some of the generic advantages of storing large amounts of data in a data warehouse are-

- Relieves pressure on resources.
- Prevents record locks.
- A single point of contact for all the data.
- Allow maintenance of historic data, which an operational database cannot achieve.

Additionally, advantages of storing airline punctuality data in a data warehouse have specific advantages such as-

- **Data Integration** – Data from multiple sources can easily be integrated with data stored in a data warehouse, which can accommodate even more comprehensive analysis. The airline punctuality data can be merged with weather data, passenger data and flight schedules data which will help in understanding the factors that influence flight punctuality.
- **Historical data analysis** – The data warehouse can store historic data of airline punctuality statistics, which can be used to find trends and patterns in the number of delays and the number of flights cancelled over time.
- **Improved data quality** – Storing data in a data warehouse includes pre-processing steps such as cleaning, transforming, and manipulating data, which improves the data quality and adds dimensions to the available data. For example, aggregations and group by functions can be applied to the available columns of data in the datasets and stored in the data warehouse, which will reduce the workload on computational resources to perform calculations each time a record is fetched.
- **Increased scalability** - A data warehouse can be scaled up to accommodate large volumes of data and increased numbers of users and queries. This can help to ensure that the data warehouse remains responsive and efficient, even as the volume of data and complexity of queries increases over time.
- **Potential high returns of investment** – An airline company must commit many resources to ensure the successful implementation of a data warehouse which could include a great deal of money. However, a study by International Data Corporation reported that data warehouse projects delivered an average three-year return on investment of 401%.
- **Increased productivity of corporate decision-makers** – Data warehousing improves the productivity of corporate decision-makers by creating an integrated database of consistent, subject-oriented, and historical data. By transforming data into meaningful information, a data warehouse allows decision-makers of airline companies to perform more accurate, consistent analysis of flight punctuality reports.

Overall, storing the airline punctuality data in a data warehouse can provide a more comprehensive and efficient way to analyse the data, leading to improved insights and decision-making for airlines and other stakeholders in the industry.

III. DATA WAREHOUSE DESIGN

A Data warehouse model is designed using ‘Star Schema’ for the provided dataset. Star Schema is a dimensional model that has a fact table in the centre, surrounded by multiple dimension tables. It is called a star schema because of its physical appearance like a star. Dimensional modelling consists of fact tables and dimensional tables.

- Fact table - Generally store a single measurement of real-world observation, that will not change over time. The information stored in fact tables is mostly numeric data. It also maintains foreign keys of dimension tables. The primary key of a fact table is a composite key made up of all the foreign keys.
- Dimension table - Contains descriptive textual information. Dimension attributes can be used as constraints in data warehouse queries. Star schema can be used to optimize the query performance by denormalizing reference data into dimension tables.

The data warehouse model that is designed using star schema for the flight punctuality data is shown in Figure 1.

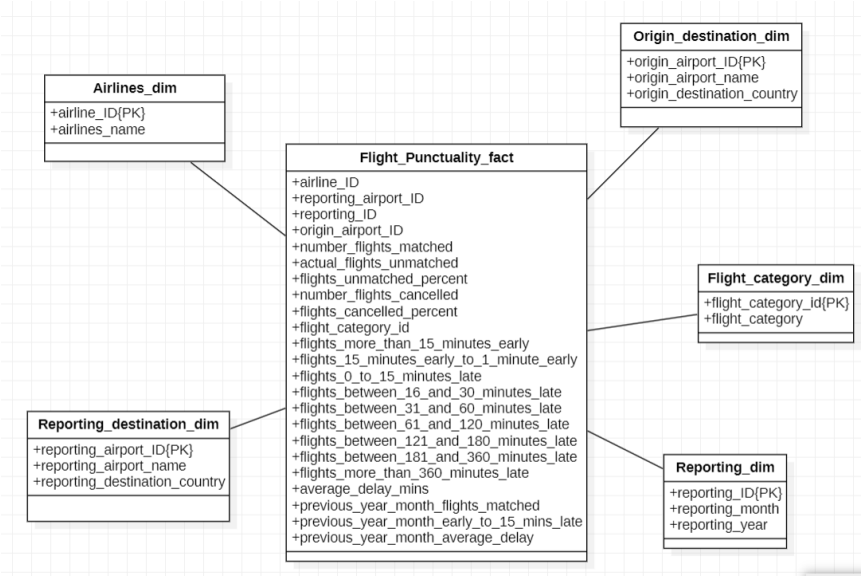


Figure 1. Data warehouse design using star schema

Justification –

- The fact table is designed to store all columns with numeric data, which allows for easy querying and analysis of data.
- Dimension tables are designed to contain additional information about the data stored in the fact table. This information can be used to group and analyze data based on different constraints.
- The fact table also stores foreign keys of dimension tables, allowing for easy joins among the tables. This design enables the retrieval of flight delays and the number of cancelled flight data with respect to the information stored in dimension tables, such as the percentage of cancelled flights grouped by airline names.
- The model facilitates analysis of delays in specific routes or seasonal trends in flight delays.

The above reasons justify that the data warehouse model is designed to provide easy access to data for analysis and allow for the retrieval of specific information based on different constraints.

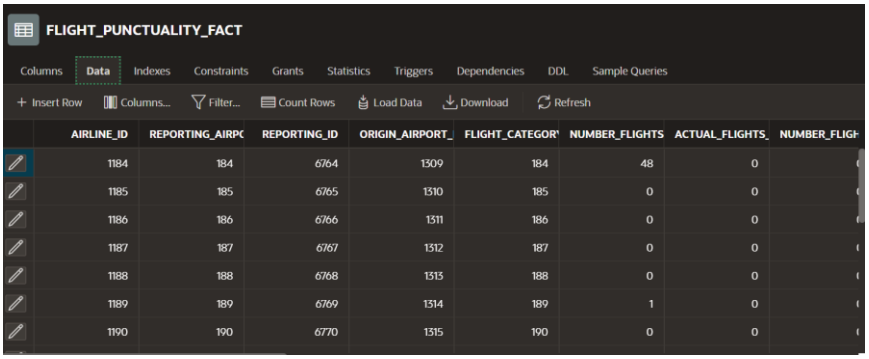
IV. CREATING TABLES

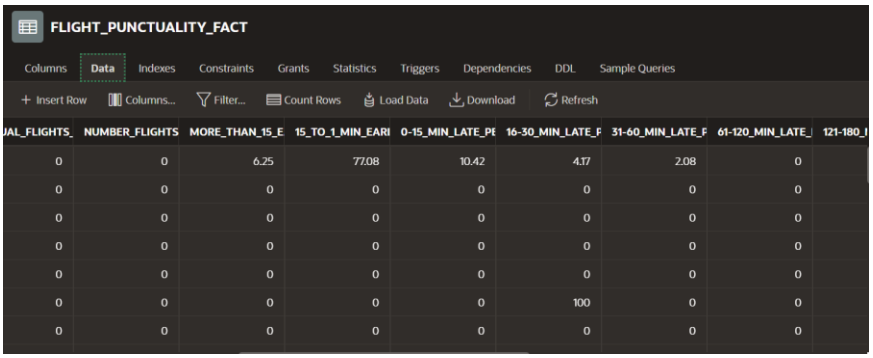
SQL statements to create tables mentioned in the schema are given below –

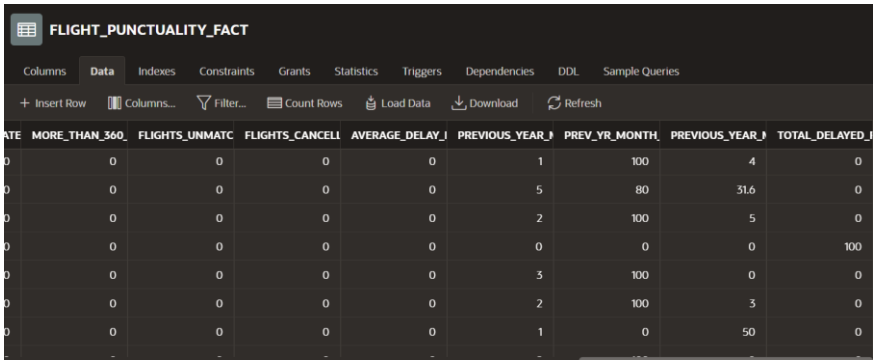
Flight Punctuality (Fact table):

```
CREATE TABLE FLIGHT_PUNCTUALITY_FACT
 ( AIRLINE_ID NUMBER,
   REPORTING_AIRPORT_ID NUMBER,
   REPORTING_ID NUMBER,
   ORIGIN_AIRPORT_ID NUMBER,
   FLIGHT_CATEGORY_ID NUMBER,
   NUMBER_FLIGHTS_MATCHED NUMBER,
   ACTUAL_FLIGHTS_UNMATCHED NUMBER,
   NUMBER_FLIGHTS_CANCELLED NUMBER,
   MORE_THAN_15_EARLY_PERCENT FLOAT,
   15_TO_1_MIN_EARLY FLOAT,
   0-15_MIN_LATE_PERCENT FLOAT,
   16-30_MIN_LATE_PERCENT FLOAT,
   31-60_MIN_LATE_PERCENT FLOAT,
   61-120_MIN_LATE_PERCENT FLOAT,
   121-180_MIN_LATE_PERCENT FLOAT,
   181-360_MIN_LATE_PERCENT FLOAT,
   MORE_THAN_360_MIN_LATE_PERCENT FLOAT,
   FLIGHTS_UNMATCHED_PERCENT FLOAT,
   FLIGHTS_CANCELLED_PERCENT FLOAT,
```

```
    AVERAGE_DELAY_MINS FLOAT,
    PREVIOUS_YEAR_MONTH_FLIGHTS_MATCHED NUMBER,
    PREV_YR_MONTH_EARLY_TO_15_LATE_PERCENT FLOAT,
    PREVIOUS_YEAR_MONTH_AVERAGE_DELAY FLOAT,
    FOREIGN KEY(airline_ID) REFERENCES airlines_dim(airline_id),
    FOREIGN KEY(reporting_airport_ID)REFERENCES
reporting_destination_dim(
    reporting_airport_ID),FOREIGN KEY(reporting_ID) REFERENCES
    reporting_dim(reporting_ID),FOREIGN
    KEY(origin_airport_ID)REFERENCES
    origin_destination_dim(origin_airport_ID),FOREIGN KEY
    (flight_category_id) REFERENCES
    flight_category_dim(flight_category_id)
);
```

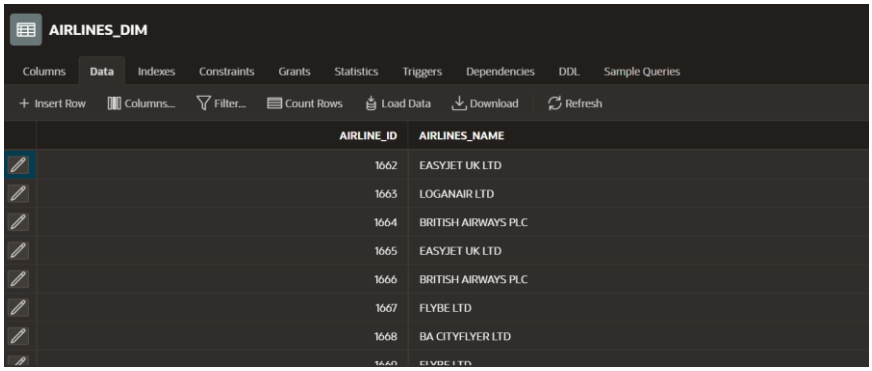






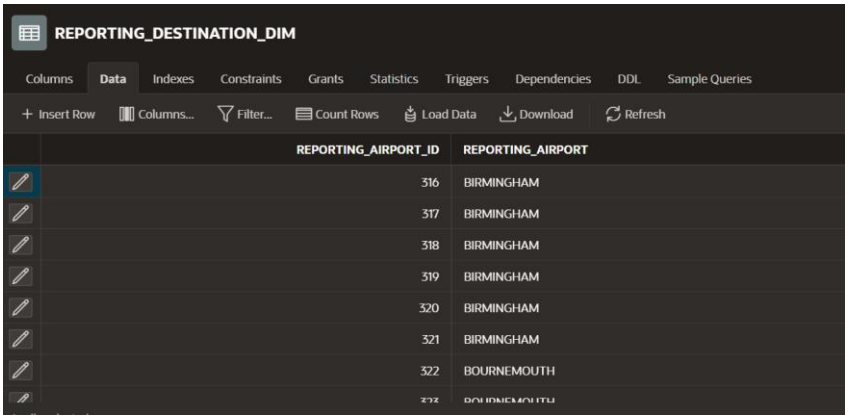
Airlines (Dimension table):

```
CREATE TABLE airlines_dim(
   airline_id number PRIMARY KEY,
   airlines_name varchar(50) NOT NULL
);
```



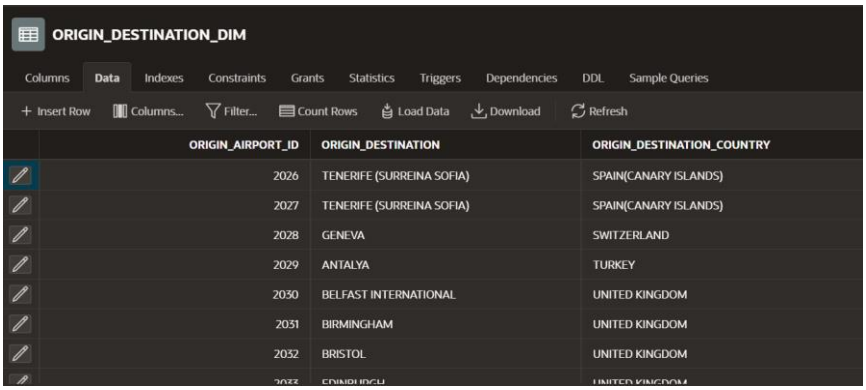
Reporting Destination (Dimension table):

```
CREATE TABLE reporting_destination_dim(
   reporting_airport_ID number PRIMARY KEY,
   reporting_airport varchar(50)
);
```



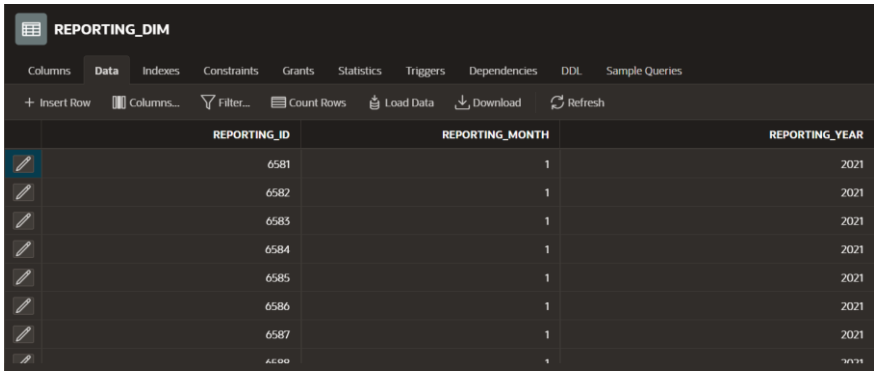
Origin Destination (Dimension table):

```
CREATE TABLE origin_destination_dim(
   origin_airport_ID number PRIMARY KEY,
   origin_destination VARCHAR(60),
   origin_destination_country VARCHAR(50)
);
```



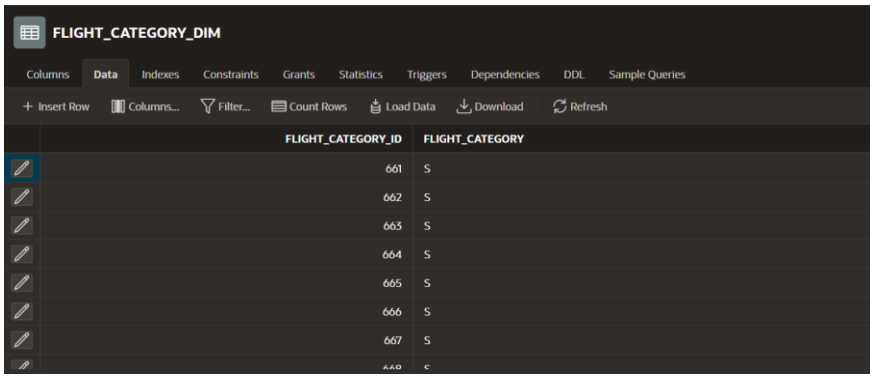
Reporting (Dimension table):

```
CREATE TABLE reporting_dim(
   reporting_ID number PRIMARY KEY,
   reporting_month number NOT NULL CHECK(reporting_month BETWEEN 1 AND
   12), reporting_year number NOT NULL CHECK(reporting_year = 2021));
```



Flight Category (Dimension Table):

```
CREATE TABLE flight_category_dim(
   flight_category_id number PRIMARY KEY,
   flight_category char NOT NULL
);
```



V. DATA LOADING AND TRANSFORMATION

An important step before data is stored in data warehouse is pre-processing step where data is cleaned, manipulated, and transformed if necessary, so that user can fetch valid and meaningful information from the database. This can be dropping few columns, splitting the columns, or aggregating some columns for easy accessibility. Before data is loaded from old tables to newly designed tables, a few transformations are applied to the dataset discussed below –

- 'run_date' column is dropped, as it does not hold any importance to analyse flight punctuality data.
- 'reporting_period' column is split into 2 separate columns – month and year to allow easy accessibility of records.
- The columns that store average values with more decimal places are rounded off to 2 decimal values for easy calculations.
- The columns with longer names are shortened, for better readability.
- Total number of flights delayed more than 15 minutes is calculated – A new column is added to the table which stores the total percentage of delayed flights using the information about flights that are delayed by more than 15 minutes.

```
ALTER TABLE flight_punctuality_fact ADD total_delayed_flights_percent float;

UPDATE flight_punctuality_fact SET
total_delayed_flights_percent = round(("15_TO_1_MIN_EARLY" +
"0-15_MIN_LATE_PERCENT" +
"16-30_MIN_LATE_PERCENT" +
"31-60_MIN_LATE_PERCENT" +
"61-120_MIN_LATE_PERCENT" +
"121-180_MIN_LATE_PERCENT" +
"181-360_MIN_LATE_PERCENT" +
```

```
"MORE_THAN_360_MIN_LATE_PERCENT")/100, 2);
```

This command inserts a new column to the table called 'total_delayed_flights_percent' and calculates the percentage of the total number of delayed flights from the existing data.

Steps to load the data:

1. Pre-processing of data is performed.
2. New Primary keys are created in Excel for each dimension table.
3. The Primary keys are mapped to the corresponding records in the fact table.
4. Month-wise data is captured and stored in a reporting table which shows the month and year of the data stored in the fact table.
5. Tables are created in Oracle Apex along with constraints, references, and validation checks.
6. The data is loaded into the tables from the Excel file that has multiple sheets storing each table's information.
7. SQL commands are used to manipulate the columns that are added by manipulation of existing columns in Oracle Apex.
8. The new data is downloaded to be used for analysis.

VI. OLAP CUBES

OLAP (Online Analytical Processing) cubes are multidimensional data structures used for analytical purposes. They help in aggregating a metric saved in the fact table over different dimensions. OLAP cubes allow for quick and flexible analysis of data from multiple perspectives. The airline industry, in general, can benefit from OLAP cubes in many ways, including:

- Airlines can use OLAP cubes to analyse historical flight data to optimize their flight schedules. For example, they can analyse data on the busiest times of day, routes that generate the most revenue, and the most popular destinations to adjust their schedules accordingly.
- Airlines can use OLAP cubes to analyse data on ticket prices, customer behaviour, and other factors to optimize their revenue management strategies. For example, they can analyse data on customer demographics and purchase history to adjust their pricing strategies and promotions.
- Airlines can use OLAP cubes to monitor their operational performance and identify areas for improvement. For example, they can analyse data on flight delays, cancellations, and other factors to identify patterns and take corrective actions to improve their operational efficiency.

Examples of OLAP cubes in the airline industry include:

- **Flight Performance Cube:** This cube captures flight data such as flight time, delays, cancellations, and revenue. It allows airlines to analyse flight performance by different dimensions such as date, route, and aircraft type. This cube helps airlines to optimize their flight schedules and improve operational efficiency.
- **Customer Analytics Cube:** This cube captures customer data such as demographics, travel history, and purchasing behaviour. It allows airlines to analyse customer behaviour by different dimensions such as location, time, and travel preferences. This cube helps airlines to optimize their revenue management strategies and provide better customer service.
- **Revenue Management Cube:** This cube captures data related to pricing strategies, promotions, and sales performance. It allows airlines to analyse revenue data by different dimensions such as market segment, route, and time period. This cube helps airlines to optimize their pricing strategies and improve revenue management.

In conclusion, OLAP cubes provide a powerful analytical tool for the airline industry. They enable airlines to analyse large amounts of data from multiple perspectives, optimize their operations, and improve revenue management strategies. By using OLAP cubes, airlines can gain a competitive advantage.

Note: The examples and benefits discussed are generic to the airline industry and not specific to the dataset provided.

VII. Business Intelligence and Data Warehouse

Using a data warehouse in combination with a business intelligence (BI) tool like Tableau can offer several benefits to organizations. Here are some of the key benefits:

- A data warehouse acts as a centralized repository that brings together data from different sources and organizes it in a meaningful way. This makes it easy for BI tools like Tableau to access and analyse the data, without the need to manually consolidate data from multiple sources.
- Data warehouses typically include a data cleaning and transformation process to ensure that the data is accurate, consistent, and complete. This leads to improved data quality, which in turn leads to more accurate and reliable insights generated by Tableau.
- A data warehouse is designed for query and analysis, which means it can quickly process large amounts of data and respond to queries much faster than traditional databases. This helps Tableau to generate reports and dashboards quickly and efficiently, leading to faster decision-making.
- A data warehouse provides a flexible data model that can be easily customized to meet specific business needs. With Tableau, users can perform complex data analysis and visualization, exploring data from multiple perspectives to identify patterns and trends.
- Tableau provides a user-friendly interface that enables business users to create their own reports and dashboards without relying on IT support. This helps users to access and analyse data on their own terms, without being restricted by the technical skills required to generate insights.

Overall, using a data warehouse in combination with a BI tool like Tableau can lead to more accurate and timely insights, faster decision-making, and improved business performance.

VIII. Data Analysis And Visualization

Once the data is ready, it is analysed with respect to different dimensions so that patterns and trends in data are identified. Data Analysis is performed using Tableau, a business intelligence tool used to analyse large amounts of data effectively. Some of the visualizations are shown below:

Visualisation 1:

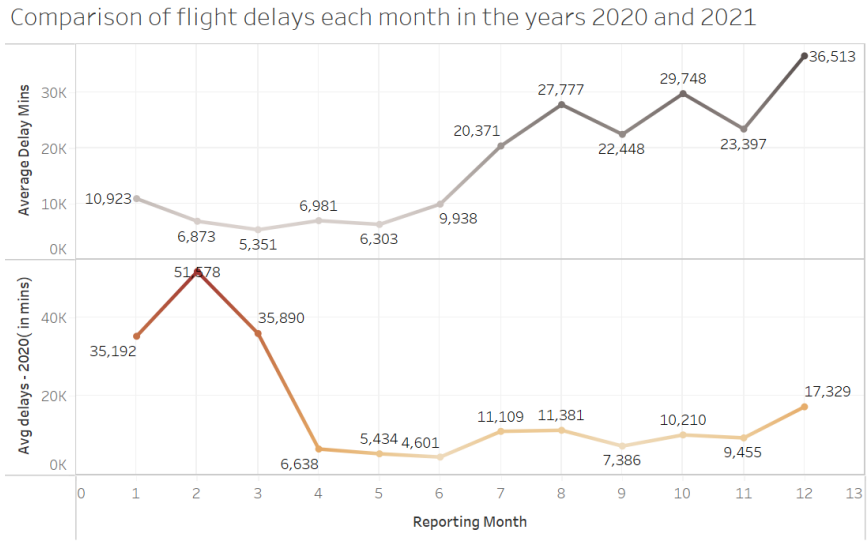


Figure 2. Visualisation 1

Aim:

The comparison of trends in flight delays every month in the years 2020 and 2021 is shown in figure 2. The average delays are shown in minutes.

Steps to Create:

1. Open Tableau and connect to the data source.
2. Drag and drop the relevant fields, 'average delays in the previous year' and 'average delays this year', to the 'Rows' shelf in Tableau.
3. Drag and drop 'Reporting month' to the 'Columns' shelf.
4. Select the line chart type from the 'Show Me' pane.

5. Add 'average delays in the previous year' and 'average delays this year' to markers and colours to display the numbers over the line graph.
6. Use the ‘Sum’ measure to calculate the sum of average delays for both rows.

Key findings:
From the visualisation, it is observed that in 2021, the average delays of flights were the most in December which is comparatively less in the previous year. This could be because December is generally winter and the weather is not very cooperative to operate flights effectively. Also, it is the holiday season which might be a reason for increased air traffic. In 2020, February had the highest number of delayed flights. Overall, the total number of minutes that the flights were delayed increased in 2021 when compared to 2020. Airlines could use this data and work on the reasons behind the increase in flight delays and try to minimize this in the future to avoid causing inconvenience to the passengers.

Visualisation 2:

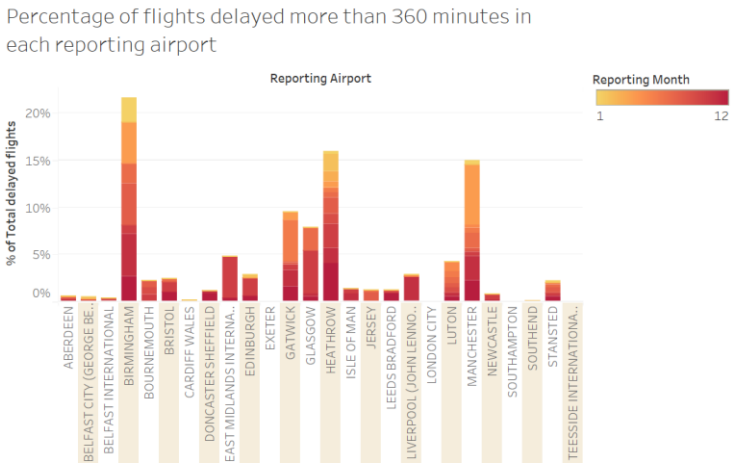


Figure 3. Visualisation 2

Aim:
The visualization presents a monthly comparison of delayed flights for more than 360 minutes among the different reporting airports.

Steps to create:
1. Drag the "Reporting airport" field to the Columns shelf.
2. Drag the "Flights more than 360 minutes" field to the Rows shelf.
3. Add "Reporting month" to the Marks card and set it to "Colour" so that each month is displayed with a different colour.
4. Add a table calculation to the "Flights more than 360 minutes" field on the Rows shelf by right-clicking on it and selecting "Add Table Calculation".
5. In the "Table Calculation" dialogue box, select "Percentage of Total" as the calculation type.
6. The resulting visualization will show the percentage of delayed flights for each reporting airport in different months, with each month being represented by a different colour.

Key findings:
In November, Birmingham Airport reported the highest number of delayed flights. Possible reasons for delayed flights in reporting airports could be related to various factors such as airport infrastructure, operational efficiency, airport management, and government regulations. By obtaining more information about the airport, it may be possible to identify the specific reasons for the high percentage of delayed flights and address them.

Visualisation 3:

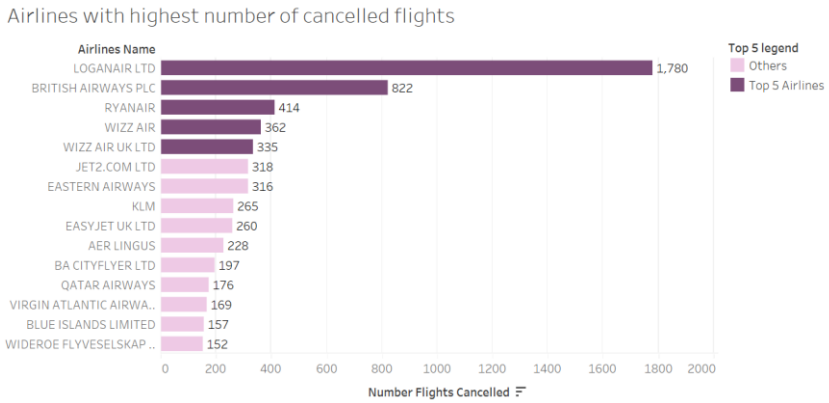


Figure 4. Visualisation 3

Aim:
The airlines with the highest number of cancelled flights are shown in figure 4. The top 5 airlines who have the highest number of cancelled flights are represented using a darker colour. This graph gives a comparison of airlines with respect to the number of flights cancelled.

Steps to Create:

1. Drag 'Airlines Name' to Rows and 'Number of Cancelled Flights' to Columns.
2. The graph is sorted in descending order.
3. Set 'Sum of Number of Cancelled Flights' as a label for the numbers to appear.
4. Create a calculated field to rank the airlines with respect to the number of cancelled flights.
5. Create another calculated field that selects the top 5 airlines with the highest number of cancelled flights and drag it to Color so that those airlines are coloured in a different colour.
6. Create another calculated field to find the airlines whose number of cancelled flights is greater than 116 and add it to Filter so that the graph is filtered accordingly.

Key Findings:
From the graph, it is observed that Logan Air Ltd has the highest number of cancelled flights summing up to 1780. There could be several reasons for flight cancellations, like bad weather conditions, crew unavailability, mechanical issues, security issues, operational issues etc. With access to additional data like weather conditions, or flight conditions of each airline, the reason behind this could further be analyzed. The top 5 airlines with the highest number of cancelled flights are represented in deep purple colour.

IX. CONCLUSION

Airline punctuality data can provide even more valuable insights when information about factors influencing the delays and cancellation of scheduled flights such as weather report in the reporting destination, passenger data and routing information is provided. With the data provided, analysis about the delays, cancellation of flights is made and assumptions are provided for the same reasons due to lack of additional data. This coursework focuses on the research of data warehousing concepts and an overview of how these concepts are applicable to real-world scenarios. Airlines generate huge amounts of data every month, and storing this data efficiently in a data warehouse would be beneficial to the airlines for easy query processing to get useful insights from the existing generated information.
Takeaways from the coursework:

- Designing a star schema data warehouse model for the provided data and the steps involved behind it.
- Transforming, manipulating, and storing the data so that insights are drawn from clean data.
- Using different features of Tableau to visualize large amounts of data effectively.
- Identifying different aspects and dimensions to analyse the provided dataset.

Things that went well:

- The experience of designing, creating, and loading data from the previous coursework was extremely helpful in planning and implementing this coursework.
- Researching about data warehouse concepts and OLAP concepts was beneficial in building the report.
- The book provided as a guide to learning and understanding Tableau was helpful to set up Tableau and use its features like filters, markers, and calculated fields appropriately.

Learnings and challenges:

- There were not many fields in the dataset that require to be transformed, a few transformations were performed on the given data, which may vary when working with even larger datasets. This might be a challenge for future projects to understand different transformations that can be applied to large datasets.
- Although data warehouse concepts were learnt and understood, their practical implementations are of interest and one of the most important skills that data scientists require.

Overall, the coursework provided an overview of how data warehouse concepts can be applied to manage and handle large chunks of data and analyse them so that useful insights can be drawn from the data that are beneficial to companies.