KINGSTON UNIVERSITY LONDON

CI7320
DATABASES AND DATA MANAGEMENT

COURSEWORK-2

Name	:	Pratiksha Rajendran
KU Number	:	K2203709
Module Leader	:	Beryl Jones

TABLE OF CONTENTS

LIST OF FIGURES

QUESTION 1

Discuss the benefits of building a data warehouse for the data set provided. (*10 marks)*

SOLUTION

BENEFITS OF BUILDING A DATAWAREHOUSE FOR FLIGHT PUNCTUALITY DATA:

1. **Identify Trends and Patterns**

- Airlines can improve their operations and identify potential issues by creating a data warehouse focused on flight punctuality.
- This allows for tracking of trends and patterns over time to gain insights that can aid in improving overall performance.

For example, airlines can monitor which flights are consistently delayed and why, and then take steps to address the underlying problems.

2. **Optimize Flight Scheduling**

- The flight punctuality dataset can be used to optimize flight scheduling by analyzing historical data and identifying patterns.
- By doing so, airlines can adjust their schedules to avoid delays and cancellations, which can save them time and money in the long run.

3. **Improve Customer Experience**

- The flight punctuality dataset can be used to improve the overall customer experience.
- Airlines can use the data to determine which routes and flights are most likely to experience delays, and then proactively communicate with customers about potential delays and their options (Datamation, n.d.).

4. **Enhance Safety Measures**

- The dataset can be used to enhance safety measures by identifying trends and patterns related to safety incidents.
- By tracking these incidents over time, airlines can take steps to improve their safety procedures and prevent future incidents.

5. **Facilitate Compliance**

- The flight punctuality dataset can help airlines comply with regulatory requirements related to flight punctuality and safety.
- By tracking and analyzing data, airlines can ensure they are meeting regulatory standards and make necessary changes to improve compliance.

6. **Reduce Costs**

- The dataset can help airlines reduce costs by identifying areas where they can improve efficiency and reduce delays.
- By optimizing flight schedules and reducing delays, airlines can save time and money on fuel, maintenance, and other operational costs.

7. **Improve Operational Efficiency**

- The dataset can be used to improve operational efficiency by providing airlines with real-time insights into flight performance.
- By monitoring flight punctuality and identifying areas where improvements can be made, airlines can increase operational efficiency and reduce costs.

8. **Enhance Revenue**

- The flight punctuality dataset can help airlines enhance revenue by improving the customer experience and reducing cancellations (Panoply, n.d.).
- By proactively communicating with customers about potential delays and offering alternative options, airlines can reduce cancellations and retain customers.

9. **Increase Competitive Advantage**

- By building a data warehouse for flight punctuality, airlines can gain a competitive advantage by providing better customer service and reducing costs.
- By using data to optimize flight schedules and reduce delays, airlines can differentiate themselves from their competitors.

10. **Support Data-Driven Decision Making**

- The flight punctuality dataset can support data-driven decision making by providing airlines with valuable insights into their operations (Dropbase, n.d.).
- By analyzing the data and identifying areas for improvement, airlines can make more informed decisions and achieve better results.

In conclusion, building a data warehouse for flight punctuality data can provide significant benefits such as improved decision-making, enhanced customer service, better operational efficiency, reduced costs, and increased revenue. The availability of historical and real-time data can help airlines optimize their schedules, improve on-time performance, and better understand customer behaviour. With the right infrastructure in place, airlines can leverage the power of data to stay competitive in an increasingly complex and dynamic industry.

QUESTION 2

Design a data warehouse using a star schema. You must justify your design decisions. *(10 marks)*

SOLUTION

The star schema designed for the dataset provided is depicted below:

Figure 1. Star schema

The star schema designed by Power BI is depicted below:

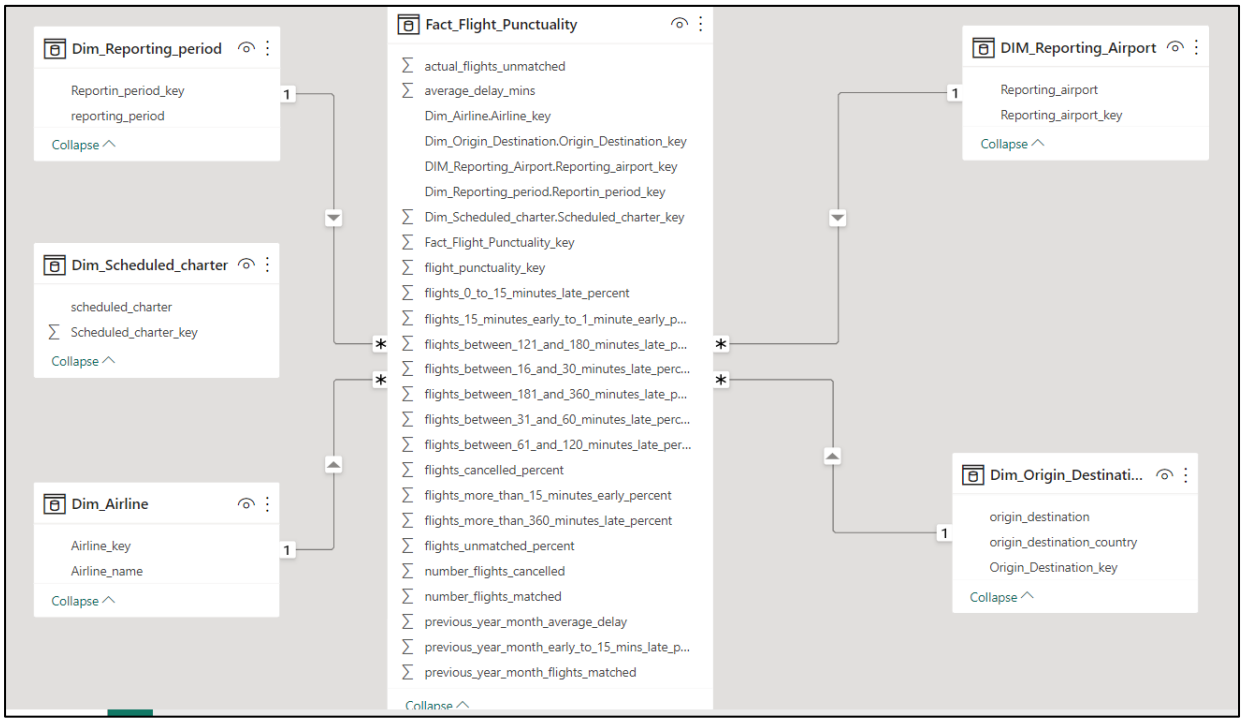

Figure 2. Star schema designed by Power BI

JUSTIFICATION OF DESIGN DECISIONS:

- The star schema for the flight punctuality dataset comprises five dimension tables and a fact table.
- The dimension tables include Reporting Period, Reporting Airport, Origin/Destination, Airline, and Scheduled/Charter, and the fact table includes attributes such as flight punctuality key, reporting period key, reporting airport key, origin destination key, airline key, scheduled charter key, and various measures related to flight punctuality.
- The design decisions for this schema were made with the goal of optimizing data retrieval and analysis.
- The Reporting Period dimension table provides a hierarchical structure to analyze flight punctuality trends over time.
- The Reporting Airport and Origin/Destination dimension tables allow for easy analysis of flight punctuality at different airports and for different destinations.
- The Airline and Scheduled/Charter dimension tables provide insight into flight punctuality across different airlines and flight types.
- The choice of a star schema allows for efficient and fast querying of data.
- The fact table is connected to the dimension tables via foreign keys, and the measures in the fact table can be aggregated at different levels of granularity depending on the analysis requirements.
- Overall, the star schema for the flight punctuality dataset is designed to enable easy and efficient analysis of flight punctuality data, with the various dimension tables providing multiple avenues for exploring the data.

QUESTION 3

Write the CREATE table statements for the tables in your star schema (include all primary and foreign keys). *(10 marks)*

SOLUTION

First, dimension tables must be created and then the fact table. There are 5 dimension tables and one fact table.

1. DIM_REPORTING_PERIOD :

```
CREATE TABLE DIM_REPORTING_PERIOD(
    REPORTING_PERIOD_KEY NUMBER(7) PRIMARY KEY,
    MONTH VARCHAR2(20) UNIQUE NOT NULL
)
```

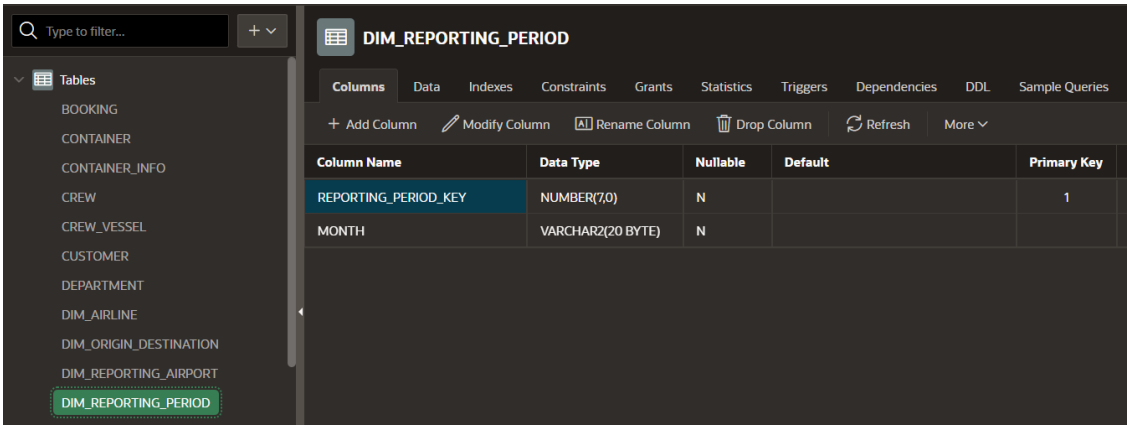

Figure 3. Reporting Period Table

2. DIM_REPORTING_AIRPORT :

```
CREATE TABLE DIM_REPORTING_AIRPORT(
    REPORTING_AIRPORT_KEY NUMBER(3) PRIMARY KEY,
    REPORTING_AIRPORT_NAME VARCHAR2(60) UNIQUE NOT NULL
)
```

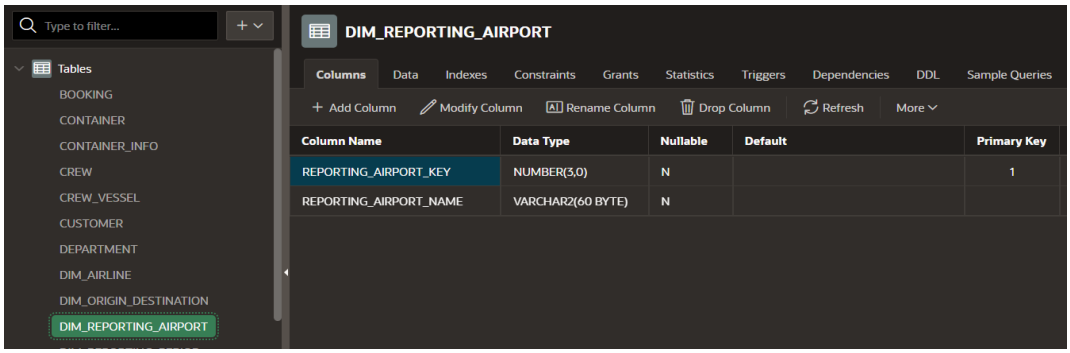

Figure 4. Reporting Airport Table

3. DIM_ORIGIN_DESTINATION :

```
CREATE TABLE DIM_ORIGIN_DESTINATION(
    ORIGIN_DESTINATION_KEY NUMBER(4) PRIMARY KEY,
    ORIGIN_DESTINATION VARCHAR2(60) UNIQUE NOT NULL,
    ORIGIN_DESTINATION_COUNTRY VARCHAR2(50) UNIQUE NOT NULL
)
```

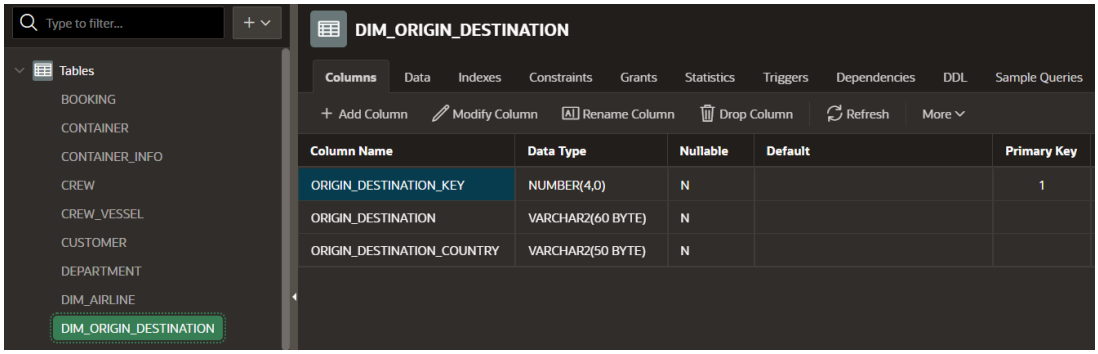

Figure 5. Origin Destination Table

4. DIM_AIRLINE :

```
CREATE TABLE DIM_AIRLINE(
    AIRLINE_KEY NUMBER(4) PRIMARY KEY,
    AIRLINE_NAME VARCHAR2(50) UNIQUE NOT NULL
)
```

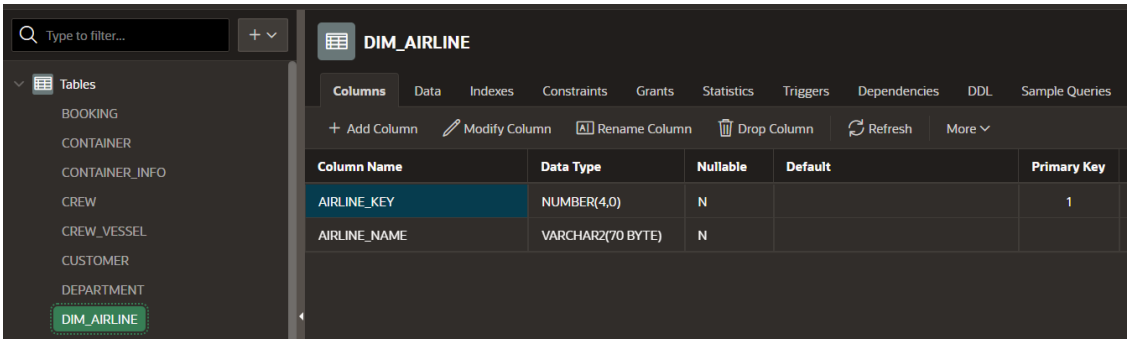

Figure 6. Airline Table

5. DIM_SCHEDULED_CHARTER :

```
CREATE TABLE DIM_SCHEDULED_CHARTER(
    SCHEDULED_CHARTER_KEY VARCHAR2(2) PRIMARY KEY,
    SCHEDULED_CHARTER VARCHAR2(15) UNIQUE NOT NULL
)
```

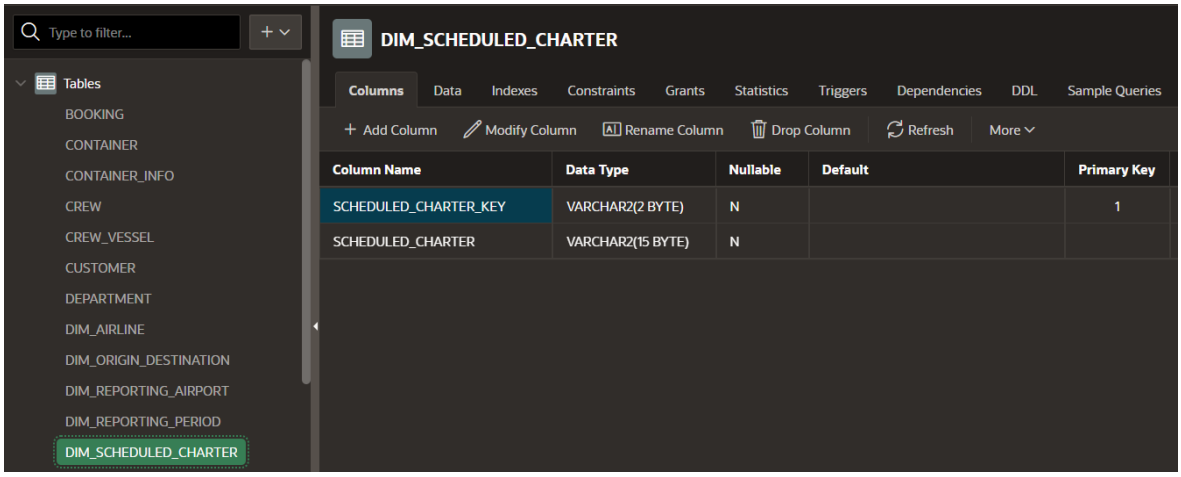

Figure 7. Scheduled Charter Table

6. FACT_FLIGHT_PUNCTUALITY :

```
CREATE TABLE FACT_FLIGHT_PUNCTUALITY(
    FLIGHT_PUNCTUALITY_KEY NUMBER(20) PRIMARY KEY,
    REPORTING_PERIOD_KEY REFERENCES
DIM_REPORTING_PERIOD(REPORTING_PERIOD_KEY),
    REPORTING_AIRPORT_KEY REFERENCES
DIM_REPORTING_AIRPORT(REPORTING_AIRPORT_KEY),
    ORIGIN_DESTINATION_KEY REFERENCES
DIM_ORIGIN_DESTINATION(ORIGIN_DESTINATION_KEY),
    AIRLINE_KEY REFERENCES DIM_AIRLINE(AIRLINE_KEY),
    SCHEDULED_CHARTER_KEY REFERENCES
DIM_SCHEDULED_CHARTER(SCHEDULED_CHARTER_KEY),


    REPORTING_AIRPORT VARCHAR2(50),
    ORIGIN_DESTINATION VARCHAR2(50),
    ORIGIN_DESTINATION_COUNTRY VARCHAR2(50),
    AIRLINE_NAME VARCHAR2(50),
    SCHEDULED_CHARTER VARCHAR2(4),


    NUMBER_FLIGHTS_MATCHED NUMBER(10),
    ACTUAL_FLIGHTS_UNMATCHED NUMBER(10),
    NUMBER_FLIGHTS_CANCELLED NUMBER(10),
```

```
    FLIGHTS_MORE_THAN_15_MINUTES_EARLY_PERCENT NUMBER(12,8),
    FLIGHTS_15_MINUTES_EARLY_TO_1_MINUTE_PERCENT NUMBER(12,8),
    FLIGHTS_0_TO_15_MINUTES_LATE_PERCENT NUMBER(12,8),
    FLIGHTS_BETWEEN_16_AND_30_MINUTES_LATE_PERCENT NUMBER(12,8),
    FLIGHTS_BETWEEN_31_AND_60_MINUTES_LATE_PERCENT NUMBER(12,8),
    FLIGHTS_BETWEEN_61_AND_120_MINUTES_LATE_PERCENT NUMBER(12,8),
    FLIGHTS_BETWEEN_121_AND_180_MINUTES_LATE_PERCENT NUMBER(12,8),
    FLIGHTS_BETWEEN_181_AND_360_MINUTES_LATE_PERCENT NUMBER(12,8),
    FLIGHTS_MORE_THAN_360_MINUTES_LATE_PERCENT NUMBER(12,8),

    FLIGHTS_UNMATCHED_PERCENT NUMBER(12,8),
    FLIGHTS_CANCELLED_PERCENT NUMBER(12,8),
    AVERAGE_DELAY_MINUTES NUMBER(12,8),

    PREVIOUS_YEAR_MONTH_FLIGHTS_MATCHED NUMBER(10),
    PREVIOUS_YEAR_MONTH_EARLY_TO_15_MINUTES_LATE_PERCENT NUMBER(10),
    PREVIOUS_YEAR_MONTH_AVERAGE_DELAY NUMBER(10)
)
```

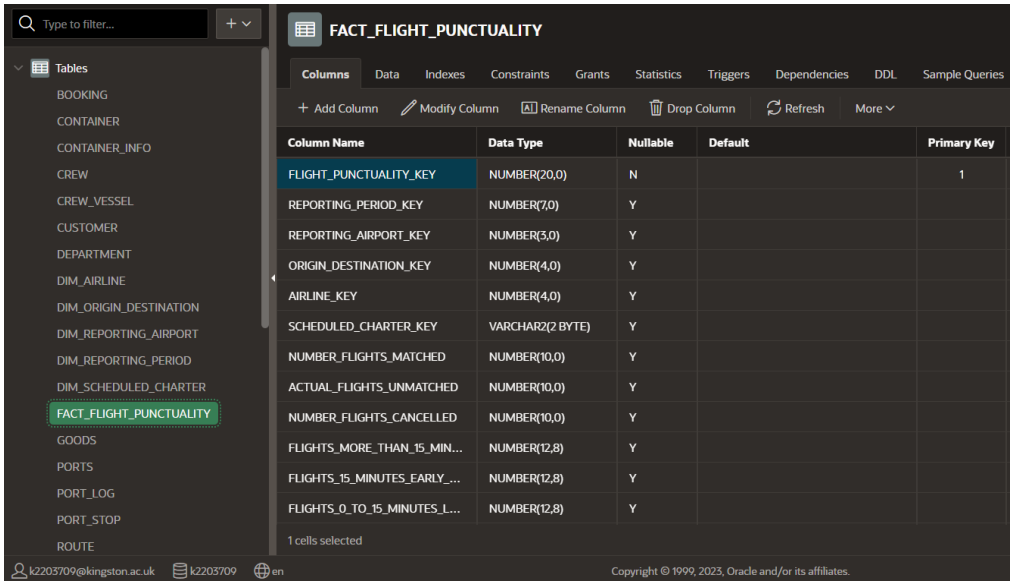

Figure 8. Fact Flight Punctuality Table

QUESTION 4

Discuss the steps you took in creating and populating the database. This should include the steps you took in preparing the data and the transformation tasks performed. (10 marks)

SOLUTION

- In order to create and populate the database, various steps were taken to prepare the data and perform necessary transformation tasks.
- Initially, Excel was used to transform and populate the tables. However, upon further examination of ETL (Extract, Transform, Load) principles, it was found that the Power BI tool was more efficient for this task.
- The 12 combined Excel files for each month were uploaded into Power BI, and transform and load operations were performed (YouTube, 2019).
- As the provided data was clear and required minimal transformation, there were no duplicates identified.
- To prepare the data for the designed star schema, numerical values were first separated, and the dimension tables for reporting period, airline, reporting airport, origin destination, and schedules or charter were analyzed.
- For example, to create the dimension table for airline, the entire column for airline name was selected and added as a new query. A separate table was created, duplicates were removed, and an index column was added and named as airline_key.
- The main table was then merged with the newly created table by combining the airline name, creating a new column in the main table that was changed to airline_key and corresponded to the airline_name as per the dim_airline table.
- This process was repeated for all dimension tables, and the main table was named as the fact table for the star schema.
- Once the data was transformed and when tried loading, Power BI created star schema design as per the data provided which is shown in Fig 2.
- Later, I populated the tables in Oracle apex with the data transformed using Power BI.
- Initially, Excel was used to create separate files for each dimension table and populate them with data by removing duplicates. SQL queries were also utilized to add values to the foreign keys in the fact table corresponding to the primary keys in the dimension tables.
- For example, to add data in the origin_destination_key column which corresponds correctly to the origin_destination and origin_destination_country was achieved using SQL query
- Overall, the use of Power BI significantly reduced the time required for this task in comparison to using Excel. By following ETL principles, the data was transformed and organized efficiently into a star schema suitable for analysis in Oracle Apex.

A few tables are shown below which have been populated with data in both Power BI and Oracle Apex:

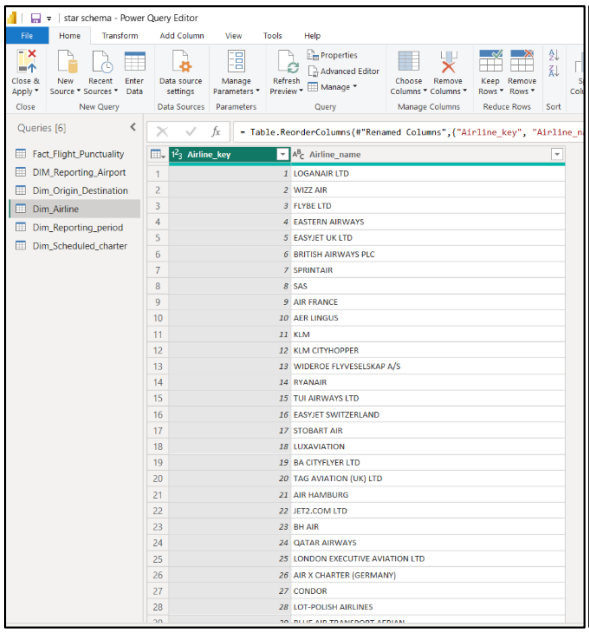

Figure 9. Dim_Airline table populated with data in Power BI

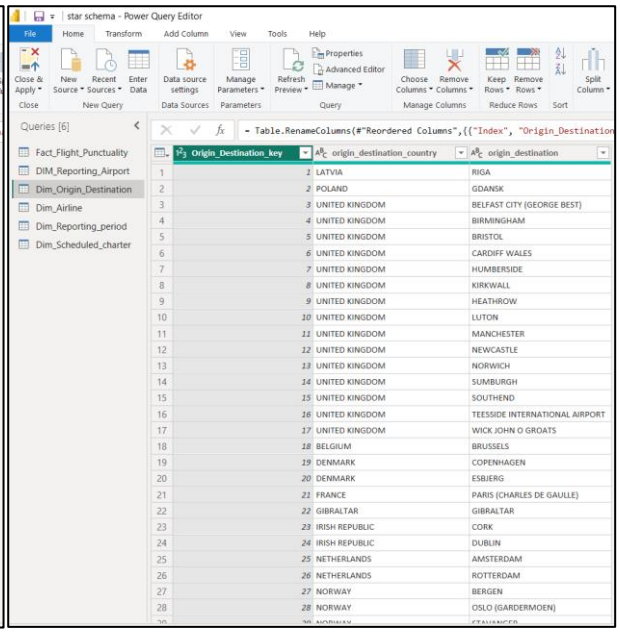

Figure 10. Dim_Origin_Destination table populated with data in Power BI

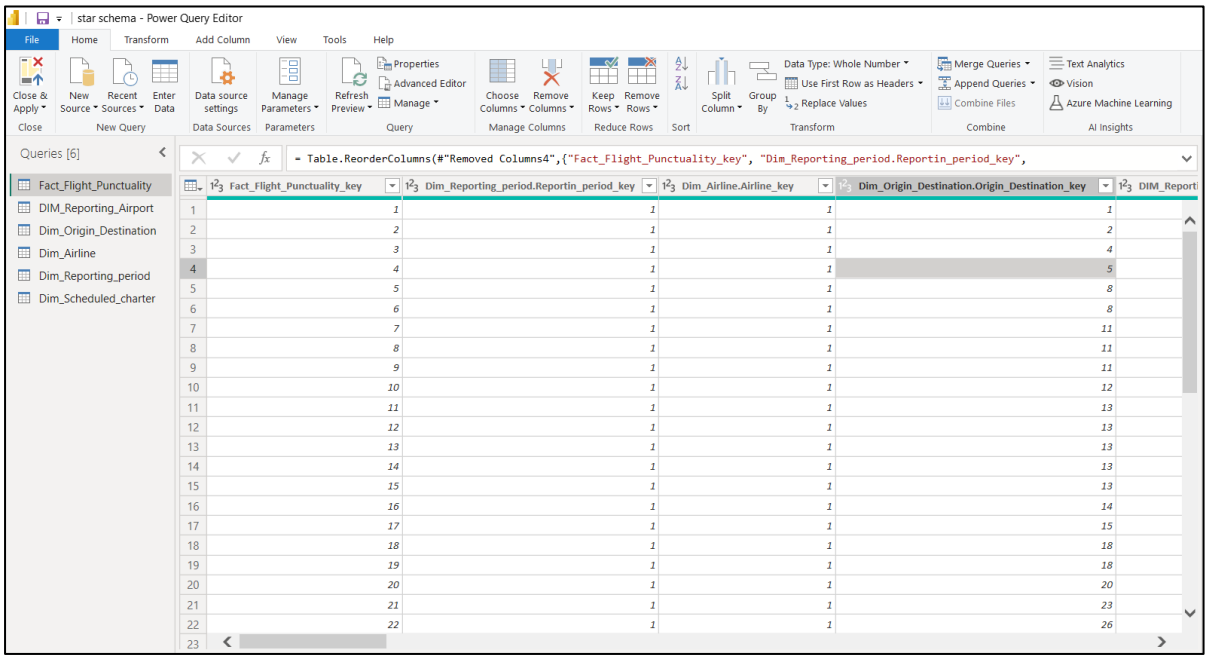

Figure 11. Fact_Flight_Punctuality table populated with data in Power BI

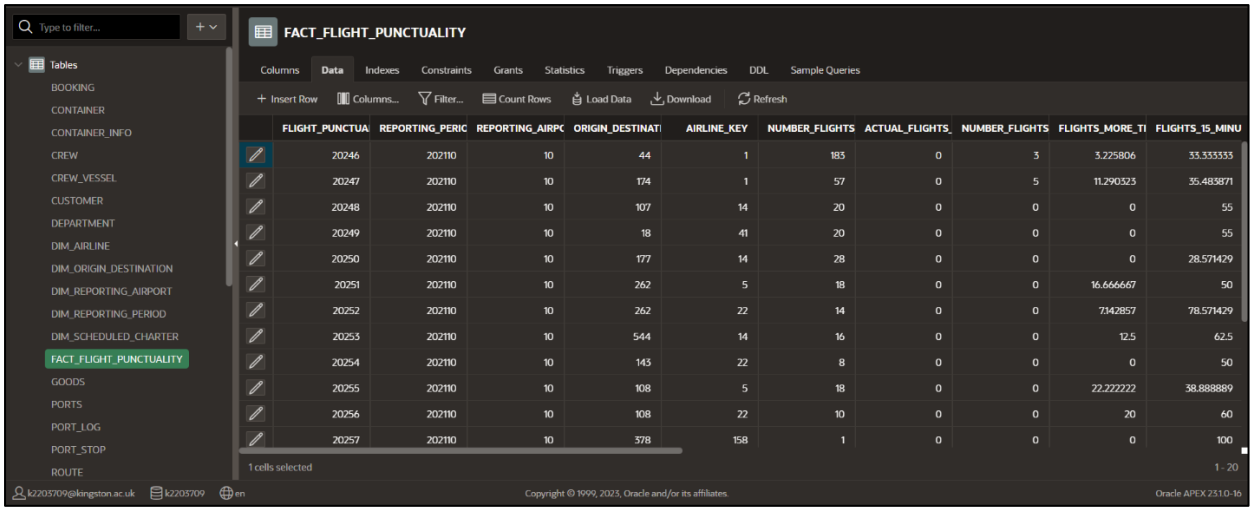

Figure 12. Fact_Flight_Punctuality table populated with data in Oracle Apex

- Figures 9, 10, and 11 display the tables that have been populated with data in Power BI, and illustrate the use of primary and foreign keys.
- Meanwhile, Figure 12 presents the fact table in Oracle Apex, with data that has been successfully loaded. These figures are referenced throughout the report to support the analysis presented.

QUESTION 5

Discuss how the airline industry in general can benefit from OLAP cubes giving examples of cubes in your discussion. *(10 marks)*

SOLUTION

- The analysis of data in a data warehouse is facilitated by multidimensional databases known as Online Analytical Processing (OLAP) cubes. These cubes comprise measures or quantitative data, such as the average delay minutes and the number of flights, as well as dimensions or qualitative data, such as the reporting period, reporting airport, origin/destination, airline, and scheduled/charter status, as per the designed star schema.
- In the OLAP cube design, the fact table serves as the centrepiece of the cube, while the dimensions form the various axes of the cube. For instance, the reporting period could be an axis, while the origin/destination could be another. The cells of the cube are then filled with measures.
- In the context of analyzing flight punctuality data for the star schema mentioned earlier, an OLAP cube can be created. The dimensions would serve to filter and slice the data, while the fact table would provide the numerical data for analysis, such as the percentage of flights matched, actual flights unmatched, number of cancelled flights, and various averages related to flight punctuality.
- By analyzing the data in a multidimensional way, the airline industry can gain insights into flight punctuality from multiple angles and perspectives. For example, one could identify which airlines have the best and worst punctuality records, or analyze trends over time by drilling down from year to quarter to month.
- These insights can be used to make data-driven decisions to improve flight punctuality, which can lead to increased customer satisfaction and loyalty. Additionally, by monitoring flights using OLAP, the industry can identify areas of improvement in their operations and take corrective actions to prevent flight delays or cancellations in the future. This can ultimately lead to cost savings for the company, as they avoid compensation claims and negative impact on brand reputation.
- Overall, the OLAP cube provides a potent tool for analyzing and comprehending flight punctuality data from multiple perspectives, increasing return on investment.

Example:

Slicing and dicing using the star schema provided earlier:

- To analyze the flight punctuality of a particular airline for the month of January 2021, the reporting period dimension table and the airline dimension table can be used to slice and dice the data in the fact table. Slicing refers to filtering the data based on a single dimension, while dicing refers to filtering the data based on multiple dimensions.
- An instance of slicing the data by the reporting period dimension is by filtering the reporting period dimension table to include only the year 2021, which in turn filters the fact table to present flight data only from 2021.

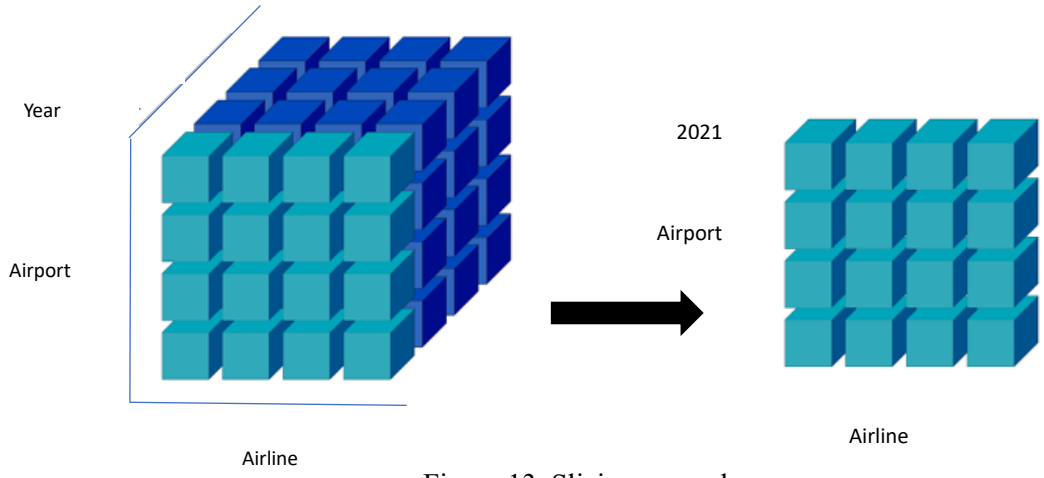

Figure 13. Slicing example

- Further, it is possible to further refine the data by filtering it based on the airline dimension table. By filtering the airline dimension table to only include flights for a specific airline, such as "British Airways," the fact table can be filtered to only include flight data for British Airways in the year 2021.

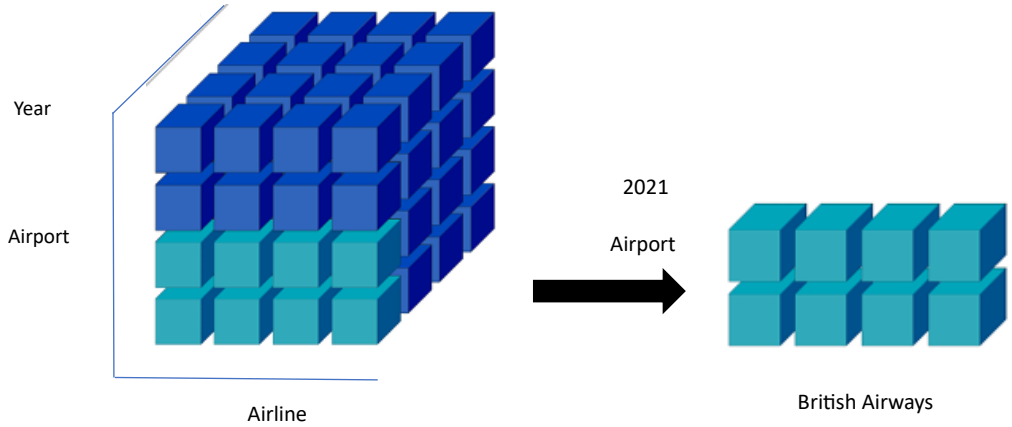

Figure 14. Dicing example

- The flight punctuality data for British Airways in 2021 can be analysed by looking at the different measures in the fact table such as the percentage of flights that were on time, early, or late, the number of flights that were cancelled, and the average delay minutes.
- For example, the data can be sliced by Scheduled/Charter, and then diced by Reporting Period to see whether there are any seasonal trends in flight punctuality for scheduled versus charter flights. Additionally, the data can be sliced by Reporting Airport, and then diced by Flights More Than 15 Minutes Early (%) to see which airports have the highest percentage of early flights.
- Slicing and dicing enables filtering the data based on specific dimensions and analysing it in a more focused way, which can provide valuable insights and facilitate data-driven decision-making.

- The paper titled "Using OLAP with Diseases Registry Warehouse for Clinical Decision Support" discusses the use of OLAP (Online Analytical Processing) in clinical decision support systems (Hamoud, A. K., & Obaid, T. A. S. ,2014). The authors highlight the need for efficient and effective decision-making processes in the healthcare industry, particularly in disease

management. The proposal is to use an OLAP cube to analyze data from a diseases registry warehouse, which contains comprehensive and longitudinal patient data.

- The design and development of the diseases registry warehouse and OLAP cube are discussed, outlining the process of data extraction, transformation, and loading. It is explained how the OLAP cube can be used to analyze patient data and identify trends and patterns, which can aid in disease diagnosis, treatment, and management.
- The paper highlights the benefits of using OLAP in clinical decision support systems, including the ability to perform complex analyses on large datasets quickly and efficiently, as well as the ability to customize and tailor the analysis to specific needs. The potential challenges and limitations of using OLAP are also discussed, including data quality issues and the need for expert knowledge in data analysis.
- Overall, the paper provides valuable insights into the use of OLAP in healthcare and its potential to improve clinical decision-making processes.

- A study presented at the 2016 IEEE conference (Ali, O., Crvenkovski, P., & Johnson, H. ,2016) explored the effectiveness of using a business intelligence data analytics solution to improve hip fracture care in a Canadian rehabilitation centre.
- The solution analysed data from electronic medical records, patient surveys, and quality assurance reports to identify areas for improvement and develop an intervention plan.
- The solution successfully reduced wait times, improved patient education, and resulted in decreased readmission rates and length of stay. The study demonstrates the benefits of using data analytics solutions in healthcare to improve processes and quality of care.

- Khosravi, M., et al. (2018) The paper "Optimizing OLAP Cube for Supporting Business Intelligence and Forecasting in Banking Sector" explores the use of OLAP cubes in the banking sector to make better business decisions by analysing large volumes of data.
- The authors propose a new model for the OLAP cube that considers real-time analysis and forecasting of financial data. The proposed model is evaluated in a case study that demonstrates its effectiveness in enabling banks to improve customer service, increase profitability, and make informed business decisions.
- The paper provides useful insights into optimizing data analysis and forecasting in the banking sector.

Example of how Coca-Cola Bottling Company benefitted from OLAP cubes

- The Coca-Cola Bottling Company previously relied on manual reporting processes which limited access to real-time sales and operations data.
- However, they were able to save 260 hours per year, equivalent to more than six weeks of work, by implementing an automated business intelligence system.
- This new system has enabled the team to easily analyse important metrics, such as delivery operations, budget, and profitability through an intuitive dashboard(Tableau, n.d.).

QUESTION 6

Discuss the benefits of using a data warehouse in combination with a business intelligence tool like Tableau. *(10 marks)*

SOLUTION

A data warehouse is an essential component of a successful business intelligence (BI) strategy. It allows businesses to store large amounts of data from various sources in a centralized location, making it easier to manage and analyze. When combined with a BI tool like Tableau or Power BI, a data warehouse can provide several benefits that can help organizations make better data-driven decisions.

BENEFITS OF USING A DATAWAREHOUSE IN COMBINATION WITH A BUSINESS INTELLIGENT TOOL LIKE TABLEAU OR POWER BI:

1. **Single Source of Truth**

- One of the key benefits of using a data warehouse with Tableau is that it provides a single source of truth for the organization.
- With all relevant data collected and stored in one place, it becomes easier to analyze and understand the data. This leads to more accurate and informed decision-making.

2. **Complex Data Analysis**

- Data warehousing also enables businesses to perform complex data analysis quickly and efficiently.
- With large amounts of data stored in a central location, businesses can use tools like OLAP to perform multidimensional analysis and slice and dice the data in various ways.
- This helps organizations identify patterns and trends, uncover insights, and make data-driven decisions.

3. **Self-Service BI**

- Another way data warehousing powers a successful BI strategy is by enabling self-service BI.
- Data warehousing allows businesses to create dashboards, reports, and visualizations that can be easily accessed by anyone in the organization.
- This empowers employees to answer their own questions and make informed decisions without relying on IT or data analysts.

4. **Improved Data Quality and Consistency**

- Data warehousing also helps organizations improve their data quality and consistency.
- By consolidating data from different sources, businesses can ensure that they are working with accurate and up-to-date information.
- This is particularly important for organizations that rely on data to make critical business decisions.

5. **Faster Reports**

- A data warehouse can greatly improve the speed of generating reports.
- Because measures are pre-calculated rather than computed at run time, reports can be produced more quickly, saving valuable time and resources.

6. **Big Data Analysis**

- Data warehousing can accommodate Big Data analysis at any scale, even billions of rows.
- This enables organizations to analyze large amounts of data quickly and efficiently, identifying trends and patterns that might have gone unnoticed.

7. **Aligning Business Definitions**

- Business definitions can differ across functions, which can lead to confusion and inaccuracies in cross-functional analysis.
- The creation of a data warehouse is an opportunity to align these definitions, reducing errors and increasing the accuracy of analysis.

8. **Governance and Best Practices**

- Data warehousing can also promote governance of data, increasing data quality and user trust in the system.
- This, in turn, can lead to increased user adoption and efficiency. Data warehousing also aligns with typical BI best practices, promoting consistency and accuracy across the organization (Senturus, n.d.).

In conclusion, a data warehouse is a critical component of a successful BI strategy. When combined with a BI tool like Tableau, it provides a centralized location for storing and analyzing data, enables complex analysis, supports self-service BI, and improves data quality and consistency. All these factors contribute to better decision-making and ultimately, a more successful business.

For example,

- Using a data warehouse in combination with Tableau or Power BI can help airlines analyze flight punctuality data from various sources.
- Tableau or Power BI can provide insights into the factors affecting punctuality, enabling airlines to take proactive measures to reduce delays and improve customer satisfaction.
- Interactive dashboards and reports also improve decision-making capabilities across the airline industry.

The above-mentioned points are relevant for both Tableau and Power BI, and in my assignment, I utilized Power BI.

QUESTION 7

Create 3 visualizations using Tableau.

SOLUTION

All the below visualizations are created using Power BI.

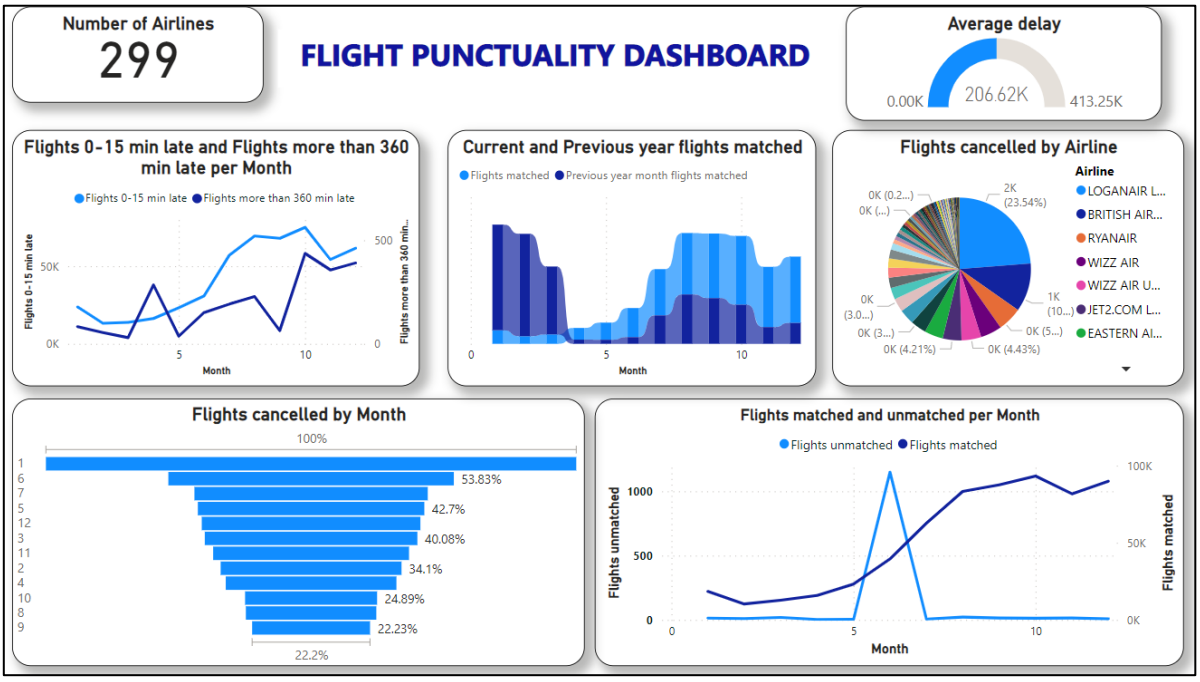

Figure 15. Flight Punctuality Dashboard

The above dashboard consists of several visualizations which are elaborated below.

1. Flights that are 0-15 minutes and more than 360 minutes late per month

AIM OF THE VISUALIZATION

To find the number of flights that are 0-15 minutes and more than 360 minutes late per month.

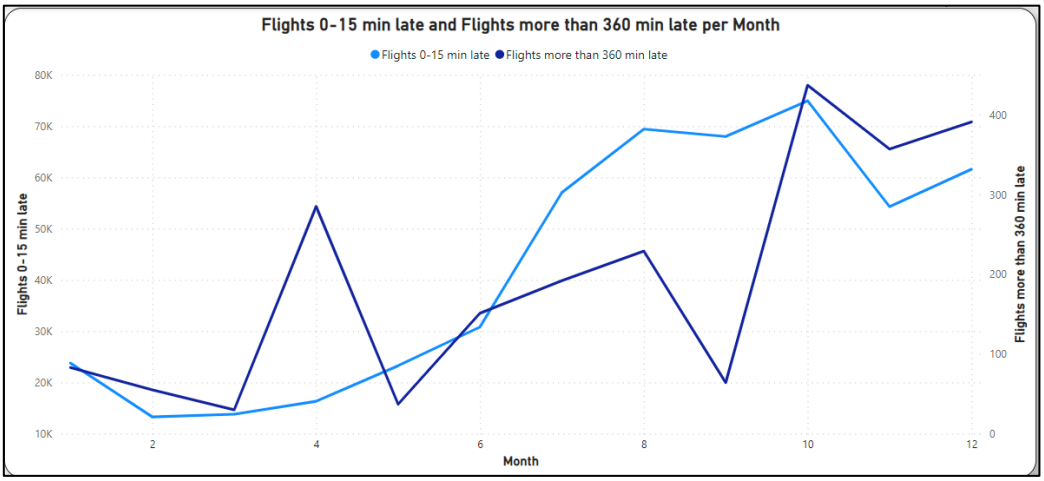

Figure 16. Visualization of Flights that are 0-15 minutes and more than 360 minutes late per month

STEPS TAKEN TO CREATE VISUALIZATION

- A line chart was created with the month on the X-axis and the number of delayed flights on the Y-axis.
- Two Y-axes were added to the chart to differentiate between flights delayed by 0-15 minutes and flights delayed by more than 360 minutes.
- Lastly, the chart was customized to present the data in a clear and intuitive manner.

KEY FINDINGS FROM THE VISUALIZATION

- After analyzing, several key findings emerged.
- August had the highest number of flights delayed by 0-15 minutes, while October had the highest number of flights delayed by more than 360 minutes.
- The number of delayed flights decreased from October to November by approximately 15,000 flights.
- The month with the lowest number of delayed flights was February.
- Possible reasons for the high number of delayed flights in August and October include unfavorable weather conditions, air traffic control issues, and airport staffing shortages.
- Additionally, the peak in delayed flights in April and August could be attributed to the fact that these months are popular for leisure travel, leading to increased demand for flights and a higher likelihood of delays.
- Overall, the visualization suggests that flight delays can be influenced by a wide range of factors and can vary significantly depending on the month and location.

2. The number of flights that matched during the previous and current year by Month

AIM OF THE VISUALIZATION

To find the number of flights that matched in the current and previous year.

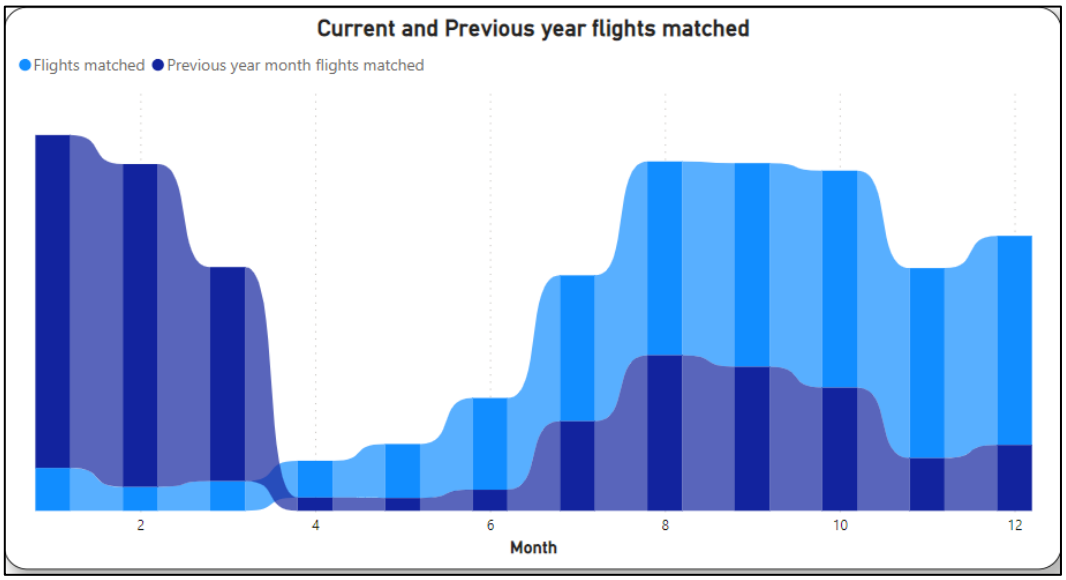

Figure 17. Visualization of the number of flights that matched during the previous and current year by Month

STEPS TAKEN TO CREATE VISUALIZATION

- A ribbon chart was created with the month on the X-axis and the number of flights matched on the Y-axis.
- There were two values included in the Y-axis i.e., flights that matched in the current year and flights that matched in the previous year.
- Lastly, the chart was customized to present the data in a clear and intuitive manner.

KEY FINDINGS FROM THE VISUALIZATION

- From August to October, many flights in the current year were found to match, while the least matched flights were recorded at the end of March.
- This could be due to favourable weather conditions and less air traffic congestion during the months of August to October, which made it easier for flights to match in the current and previous year.
- In the previous year, most flights were found to match in the beginning of the year, which gradually reduced to the lowest from April to June, and then attained stability from August.
- This trend could be attributed to the fact that the beginning of the year is generally a peak travel season for leisure and business, resulting in more flights being scheduled and having a higher probability of matching.
- However, as the year progresses, factors such as air traffic congestion and unfavourable weather conditions may lead to a reduction in matched flights.
- Overall, the findings suggest that favourable weather conditions and less air traffic congestion play a crucial role in the matching of flights in the current and previous years. However, external factors such as mechanical issues, staffing shortages, and unforeseen events such as natural disasters can also impact the ability of flights to match.

3. The number of flights cancelled by Airline:

AIM OF THE VISUALIZATION

To find the number of flights cancelled by each Airline.

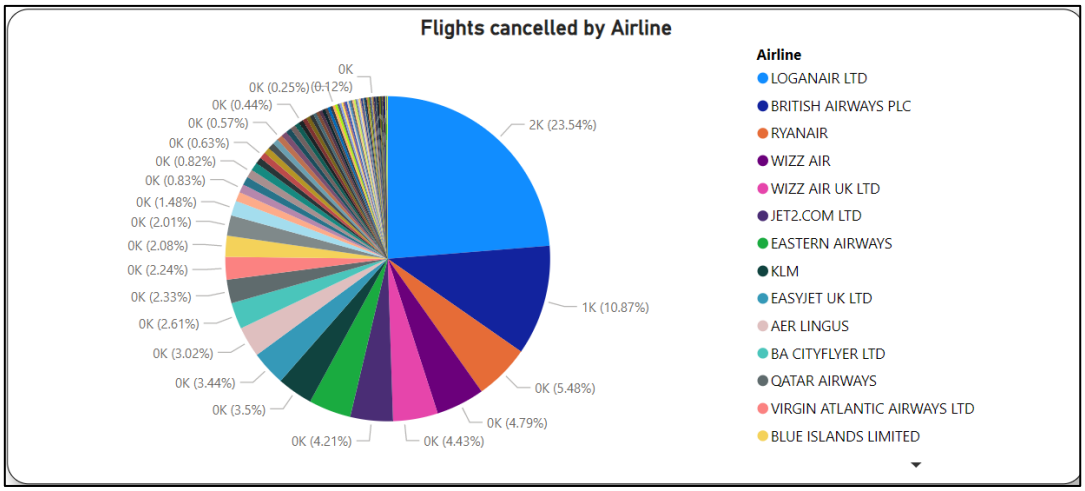

Figure 18. Visualization of the number of flights cancelled by each airline

STEPS TAKEN TO CREATE VISUALIZATION

- A pie chart was created with the legend as Airline name and values as number of flights cancelled.
- Lastly, the chart was customized to present the data in a clear and intuitive manner.

KEY FINDINGS FROM THE VISUALIZATION

- The pie chart shows that Loganair Ltd Airline has cancelled the highest number of flights, accounting for 23.54% of all cancellations, followed by British Airways with 10.87% of cancellations.
- Smaller airlines such as Bulgaria Air and Luxair had almost zero cancellation rates.
- The higher cancellation rates for larger airlines like Loganair Ltd and British Airways could be due to their larger volume of flights and routes, leading to a higher probability of cancellations.
- On the other hand, smaller airlines like Bulgaria Air and Luxair may have better control over their operations, resulting in fewer cancellations.
- The reasons for cancellations can vary, including weather conditions, mechanical issues, staffing shortages, or unexpected events such as strikes.
- Improving maintenance and operations, managing routes and volume of flights, and having contingency plans in place could potentially reduce the number of flight cancellations.

4. The number of flights cancelled by Month:

AIM OF THE VISUALIZATION

To find the number of flights cancelled at each month.

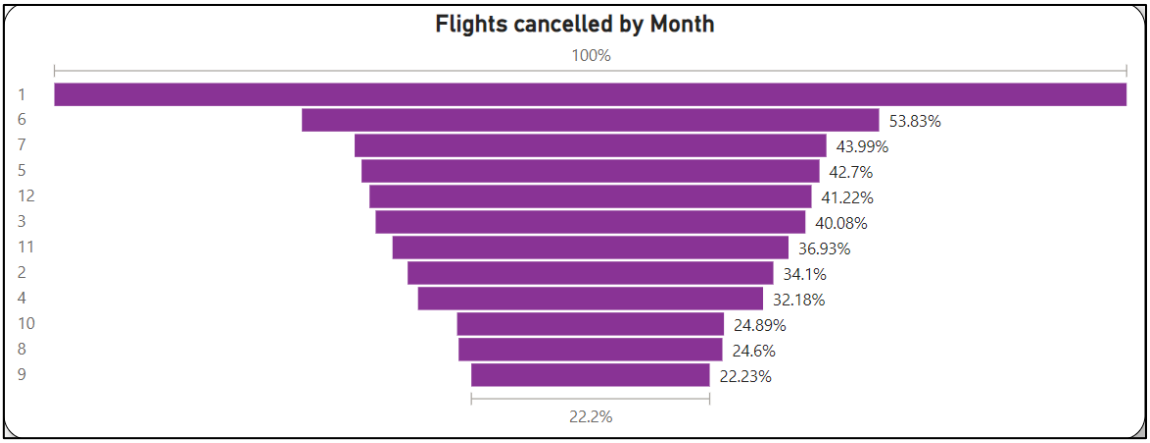

Figure 19. Visualization of the number of flights cancelled at each month

STEPS TAKEN TO CREATE VISUALIZATION

- A funnel was created with the category as month and values as number of flights cancelled.
- Lastly, the chart was customized to present the data in a clear and intuitive manner.

KEY FINDINGS FROM THE VISUALIZATION

- Based on the visualization created, it was observed that flight cancellations occurred at varying rates across different months.

- January had the highest number of cancellations, accounting for 53.83% of the total, while October had the least cancellations at 22.23%.
- The months of June, July, May, December, and March had the most cancellations, with rates above 40%.
- The reasons for the high cancellation rates could be attributed to several factors, including extreme weather conditions, mechanical issues, staffing shortages, and unexpected events.
- For instance, summer months like June and July may experience severe weather conditions such as thunderstorms or hurricanes, leading to a higher likelihood of cancellations.
- On the other hand, winter months like December and January may experience more mechanical issues due to the strain on equipment caused by cold temperatures.

5. The number of flights matched and unmatched by Month:

AIM OF THE VISUALIZATION

To find the number of flights matched and unmatched at each month.

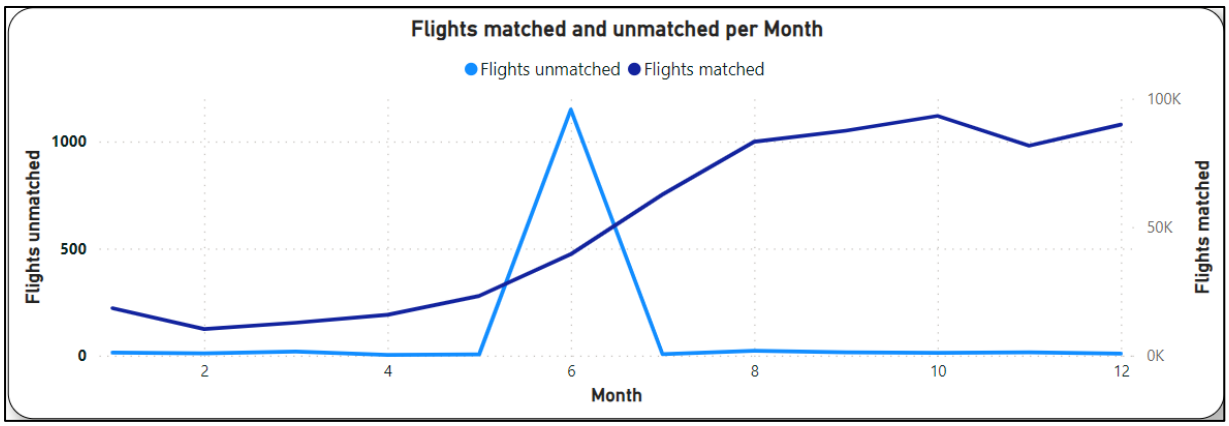

Figure 20. Visualization of the number of flights matched and unmatched at each month

STEPS TAKEN TO CREATE VISUALIZATION

- A line chart was created with the month on the X-axis and the number of delayed flights on the Y-axis.
- Two Y-axes were added to the chart to differentiate between flights matched and flights unmatched
- Lastly, the chart was customized to present the data in a clear and intuitive manner.

KEY FINDINGS FROM THE VISUALIZATION

- The key findings suggest that the month of June had the highest number of unmatched flights, while almost no flights went unmatched in December.
- The line chart also shows that the number of matched flights increased from May to October, with a slight decrease before an increase in December.
- February had the lowest number of matched flights among all the observed months.
- The reasons for unmatched flights may include various factors such as weather conditions, mechanical issues, or unexpected events like strikes, resulting in cancellations or rescheduling of flights.

- Additionally, flights may also go unmatched due to overbooking, which is common during peak travel seasons.
- Additionally, flights may also go unmatched due to overbooking, which is common during peak travel seasons.
- The increase in the number of matched flights from May to October could be attributed to the fact that these months are popular for leisure travel, leading to an increase in demand for flights.
- The slight decrease in the number of matched flights in November could be due to the seasonal shift towards the end of the year, with travellers less likely to take vacations or travel for leisure purposes.
- The increase in matched flights in December could be due to the holiday season, as many people travel to visit family and friends.

Additionally, I have created a few other visualizations as well which are shown below.

5. The average delay and number of flights cancelled current and previous year by airline:

AIM OF THE VISUALIZATION

To find the average delay and number of flights cancelled current and previous year by airline

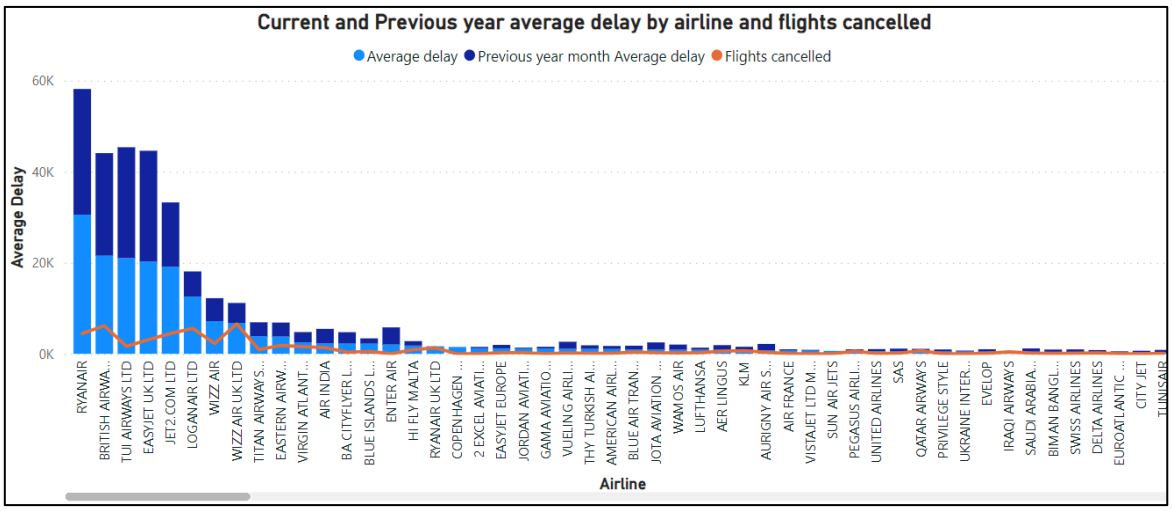

Figure 21. Visualization of the average delay and number of flights cancelled current and previous year by airline

6. The number of flights late between 16 and 360 minutes by airline:

AIM OF THE VISUALIZATION

To find the number of flights late between 16 and 360 minutes by airline

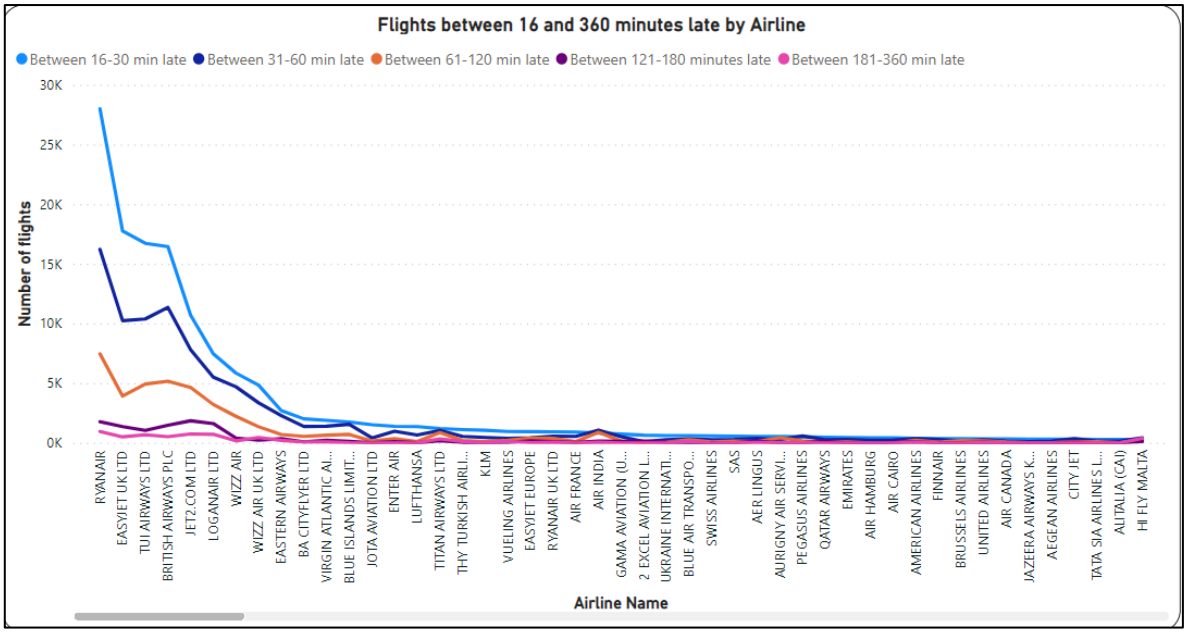

Figure 22. Visualization of the number of flights late between 16 and 360 minutes by airline

7. The number of flights more than 360 min late by Origin/Destination:

AIM OF THE VISUALIZATION

To find the number of flights more than 360 min late by Origin/Destination

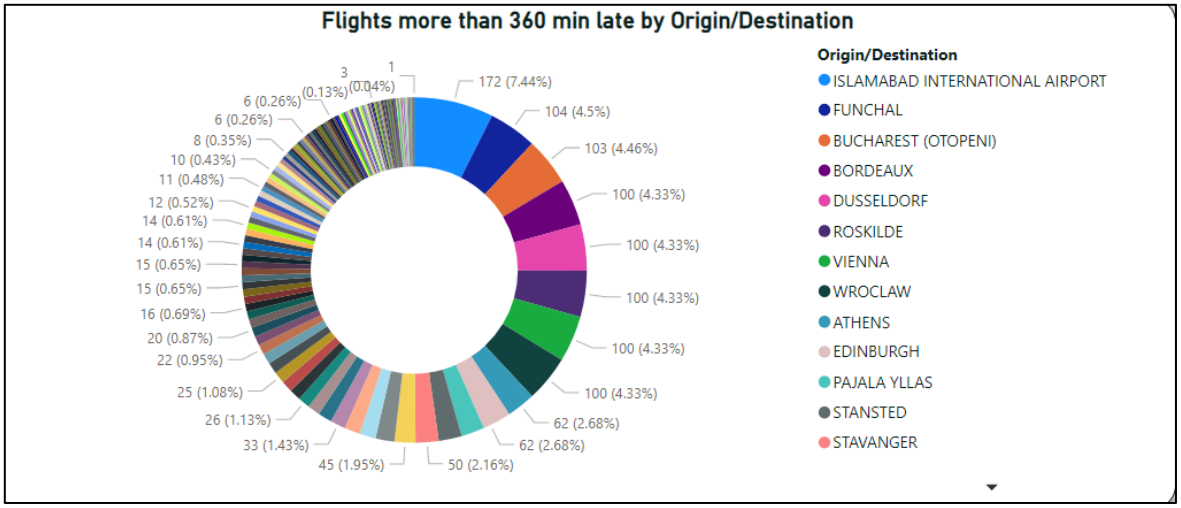

Figure 23. Visualization of number of flights more than 360 min late by Origin/Destination

8. The number of flights matched by airline by month:

AIM OF THE VISUALIZATION

To find the number of flights matched by airline by month

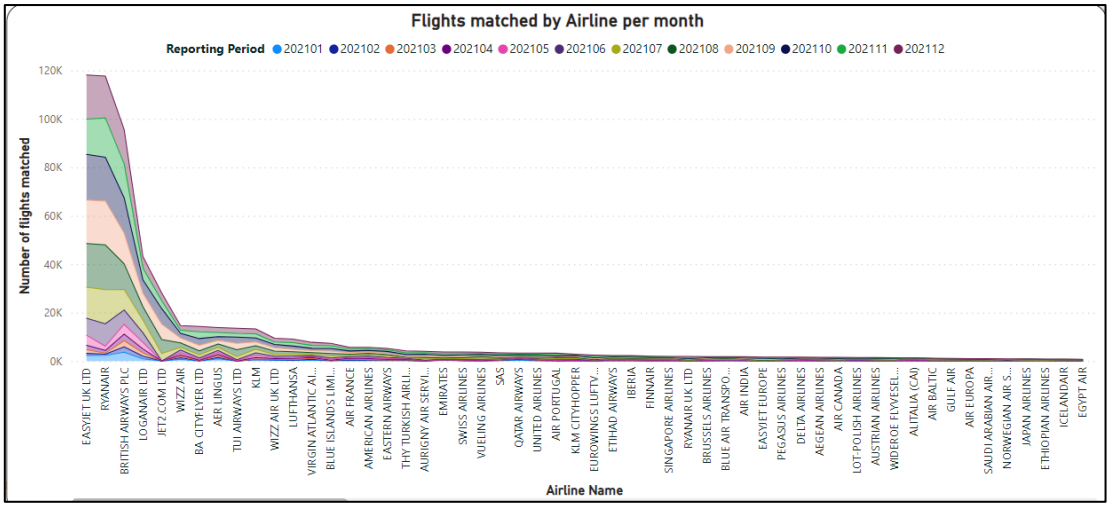

Figure 24. Visualization of number of flights matched by airline by month

9. The number of flights matched and unmatched by airline:

AIM OF THE VISUALIZATION

To find the number of flights matched ad unmatched by airline

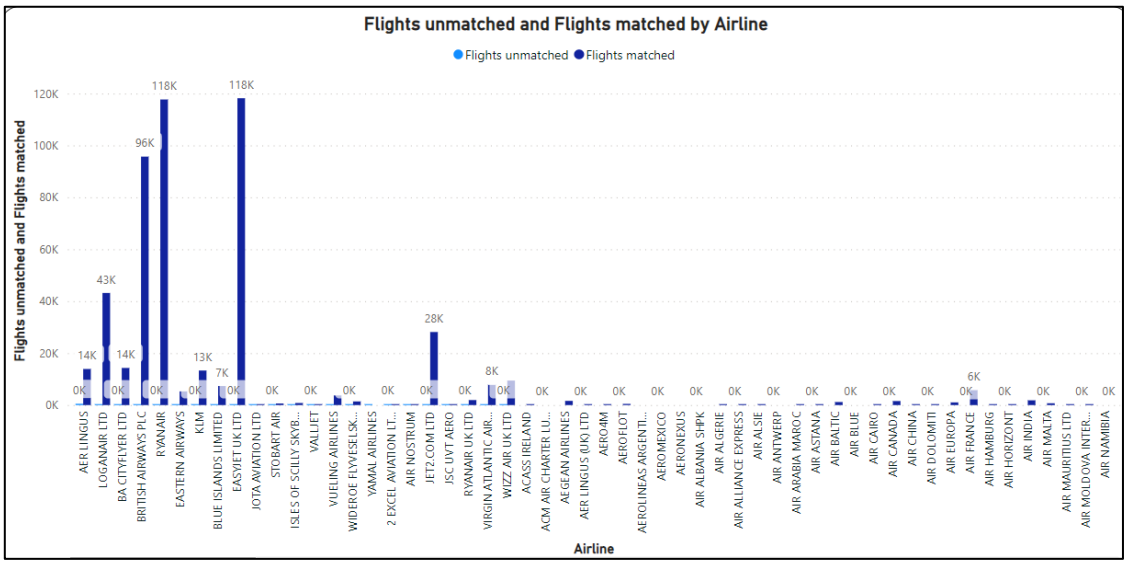

Figure 25. Visualization of number of flights matched and unmatched by airline

CONCLUSION

EVALUATION OF THE ASSIGNMENT

- The assignment provided me with a comprehensive understanding of data warehousing, a concept that I had not previously encountered.
- Through the assignment, I gained experience in using various tools such as Microsoft Azure, Power BI, and Tableau, which were new to me.
- Initially, I designed a star schema based on the given dataset in draw.io and then utilized Excel for data transformations. Subsequently, I populated the tables in Oracle apex.
- However, when I explored Power BI for visualization, I discovered that data transformation in Power BI was swift and straightforward.
- I was pleasantly surprised to find that Power BI could create a star schema on its own after data transformation.
- Upon conducting research, I learned that Power BI is a compatible tool for star schema.
- Although I attempted to use Microsoft Azure for the ETL process, I found it challenging to use, and thus it was not effective for me.
- Regarding data visualization, I employed both Tableau and Power BI for this purpose.
- However, I found Power BI to be more user-friendly and comfortable to work with.
- While conducting data visualization, I realized the importance of visualizations in enabling effective data analysis.
- I understood the importance of how flight punctuality data allows for the evaluation of the performance of airlines and airports in terms of on-time arrivals and departures. This information can aid in the identification of areas of improvement to enhance operational efficiency.
- Flight punctuality data also offers insight into factors that may influence flight delays, such as weather conditions, maintenance issues, and airline-related factors. These insights can inform strategic decision-making by airlines and airports to mitigate the impact of such factors on flight punctuality.
- I realized on how customer satisfaction is closely tied to flight punctuality. Timely arrivals and departures enhance customer satisfaction, while delays and cancellations can significantly impact customer experience.
- Hence, I am very much satisfied with my learning content in the assigned with a lot of exposure to unknown tools.

SELF EVALUATION

Criterion	**Mark (0-10)**
Designing Star schema for data warehouse	10
Data Transformation	9
Creating Tables in Oracle Apex and Populating the tables	10
Data Visualization	7
Key Findings from the visualization	7
Handling business intelligent tool	8

WHAT DID I THINK OF THE ASSIGNMENT

- The assignment on data warehousing, star schema creation, table population, and visualization using business intelligence tools provided an excellent opportunity to gain practical experience and understanding of these concepts.
- It was a valuable exercise in learning how to extract, transform, and load data from various sources into a data warehouse to facilitate efficient data analysis. Additionally, the exercise helped in gaining insights into how data can be visualized using business intelligence tools like Power BI and Tableau.
- The creation of a star schema based on the provided dataset allowed for a comprehensive understanding of the relationships between the various dimensions and facts.
- Excel was used for data transformation, and the tables were populated in Oracle apex, providing an opportunity to gain hands-on experience in using these tools.
- The exercise also highlighted the ease and speed with which data transformation can be performed in Power BI.
- The visualization of the data was an essential aspect of the assignment, which emphasized the value of visualizing data to aid in data analysis.
- The use of Power BI and Tableau for data visualization demonstrated the capabilities of these tools in helping to understand data patterns, trends, and insights that might not be apparent in raw data.
- Overall, the assignment was an excellent opportunity to gain practical experience in data warehousing and business intelligence tools, which will be valuable in future data analysis projects.

WHAT WENT WELL AND WHAT DID NOT

- During the course of the assignment, I encountered both successes and challenges.
- When searching for a tool for data transformation, I decided to try Microsoft Azure, but unfortunately, I struggled to find an efficient way to transform the data.
- However, when I switched to Power BI, I was pleasantly surprised by its capabilities and found that the ETL process went smoothly.
- As for the visualization stage, I struggled with handling the small sample sizes, but overall, I am satisfied with the visualizations and the resulting dashboard.
- If I had considered sample sizes too then I would have been a little more confident.

REFERENCES

[1] Datamation. (n.d.). Top 10 Benefits of a Data Warehouse. Available at: https://www.datamation.com/big-data/top-10-benefits-of-a-data-warehouse/ (Accessed: 05 May 2023).

[2] Panoply. (n.d.). Use Cases of a Data Warehouse. Available at: https://blog.panoply.io/use-cases-of-a-data-warehouse (Accessed: 05 May 2023).

[3] Dropbase. (n.d.). What is a Data Warehouse and How Can It Benefit Organizations? Available at: https://www.dropbase.io/post/what-is-a-data-warehouse-and-how-can-it-benefit-organizations (Accessed: 05 May 2023).

[4] Data School. (n.d.). Why Build a Data Warehouse? Available at: https://dataschool.com/data-governance/why-build-a-data-warehouse/ (Accessed: 07 May 2023).

[5] YouTube. (2019). Data Warehousing and Business Intelligence: An Overview. Available at: https://www.youtube.com/watch?v=6g4PKRHC1UQ (Accessed: 03 May 2023).

[6] Khosravi, M., et al. (2018). The Impact of Big Data on Data Analytics. Journal of Information Technology Management, Vol. 10, No. 2, pp. 1-20. Available at: https://jitm.ut.ac.ir/article_80026.html

[7] Tableau. (n.d.). Coca-Cola Bottling Company Empowers the Enterprise with Tableau Mobile Dashboards to Drive Bottom Line. Available at: https://www.tableau.com/solutions/customer/coca-cola-bottling-company-empowers-enterprise-tableau-mobile-dashboards-drive (Accessed: 06 May 2023).

[8] Hamoud, A. K., & Obaid, T. A. S. (2014). Using OLAP with Diseases Registry Warehouse for Clinical Decision Support. International Journal of Computer Science and Mobile Computing, Vol. 3, Issue. 4, pp. 39-49.

[9] Ali, O., Crvenkovski, P., & Johnson, H. (2016). Using a Business Intelligence Data Analytics Solution in Healthcare: A Case Study: Improving Hip Fracture Care Processes in a Regional Rehabilitation System. In Proceedings of the 11th International Conference on Intellectual Capital, Knowledge Management & Organizational Learning (ICICKM 2016), New York, USA, October 20-21, 2016 (pp. 13-23). Available at: https://www.researchgate.net/publication/311251195_Using_a_business_intelligence_data_analytics_solution_in_healthcare/citations

[10] Senturus. (n.d.). Power BI with and Without a Data Warehouse. Available at: https://senturus.com/blog/power-bi-with-and-without-a-data-warehouse/ (Accessed: 08 May 2023).