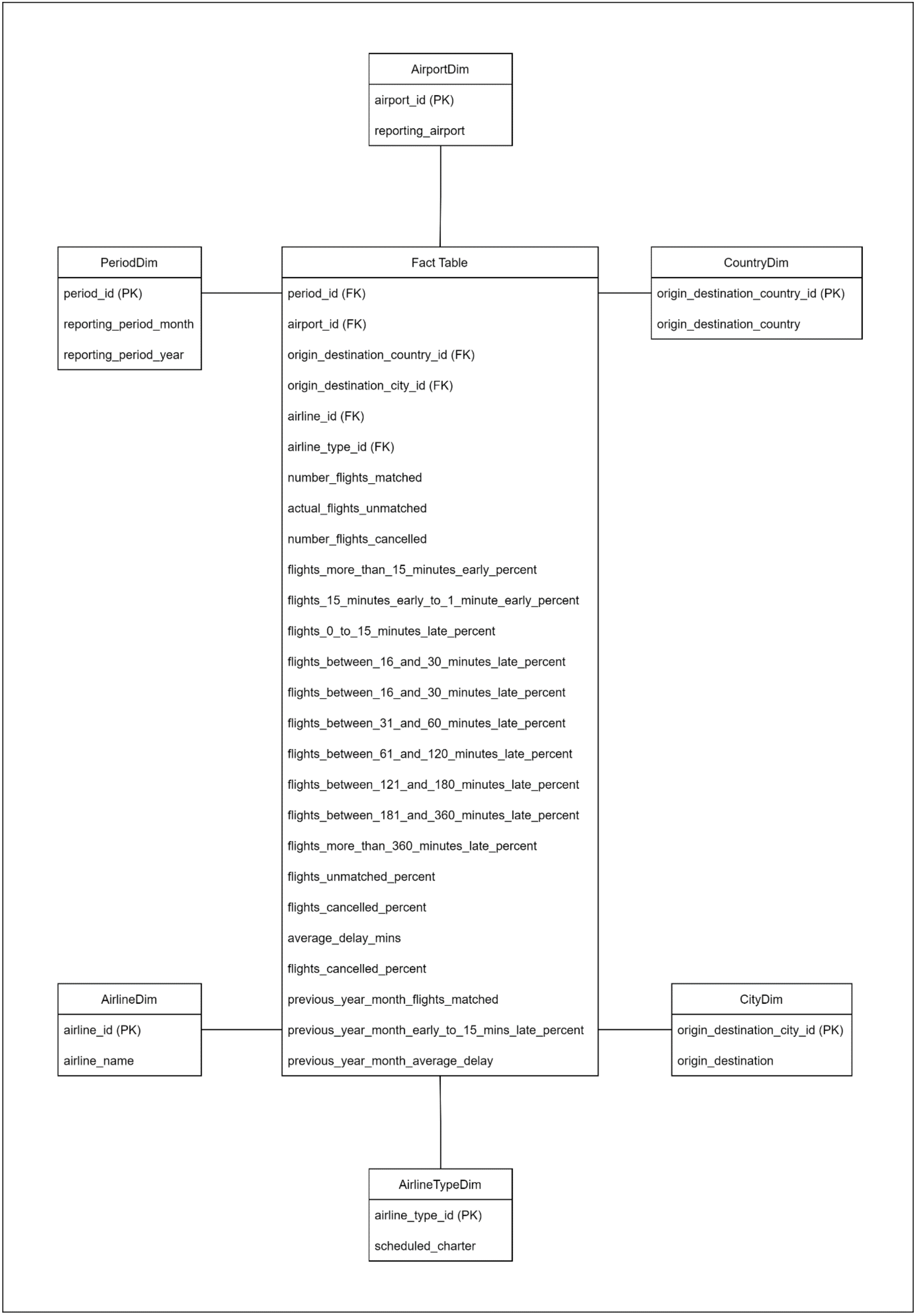
1. Benefits of constructing a data warehouse for the provided dataset

- Centralized Data Storage: A data warehouse provides a centralized repository for storing and managing large volumes of structured and standardized data from multiple sources. By consolidating data in a single location, it becomes easier to access, query, and analyse the dataset efficiently.
- Improved Data Quality: Data warehouses typically undergo a data integration and cleansing process, which helps improve data quality. Inconsistent, redundant, and inaccurate data can be identified and resolved during the data preparation phase, ensuring that the dataset used for analysis is reliable and accurate.
- Enhanced Data Analysis Capabilities: Data warehouses are specifically designed for analytical processing. They incorporate optimized data structures, indexing techniques, and query optimization strategies that enable fast and efficient data retrieval and analysis. This allows for complex queries, aggregations, and joins to be performed swiftly, facilitating in-depth analysis of the dataset.
- Streamlined Reporting and Business Intelligence: With a data warehouse, organizations can establish standardized reporting processes and create customized business intelligence (BI) dashboards. Users can easily access predefined reports or build their own ad-hoc reports using intuitive BI tools. This empowers decision-makers to gain actionable insights quickly and make informed decisions based on timely and accurate information.
- Long-Term Historical Analysis: Data warehouses facilitate historical analysis by storing historical data over an extended period. This allows for trend analysis, pattern identification, and comparison of data across different time periods. Such long-term analysis can uncover valuable insights into the dataset, providing a broader perspective on performance, trends, and anomalies.
- Support for Advanced Analytics: Data warehouses provide a solid foundation for implementing advanced analytics techniques such as data mining, predictive modelling, and machine learning. These techniques can uncover hidden patterns, perform forecasting, and generate predictive models to support decision-making processes.
- Scalability and Performance: Data warehouses are designed to handle large datasets and support scalability. As data volumes increase, data warehouses can accommodate the growth without sacrificing performance. They can be optimized for efficient data loading, query performance, and resource utilization, ensuring timely access to data even as the dataset expands.
- Data Governance and Security: Building a data warehouse enables organizations to establish robust data governance practices. Data access controls, data lineage tracking, and data security measures can be implemented to ensure data integrity, compliance with regulations, and protection against unauthorized access.

2. Designing a Data Warehouse Schema.

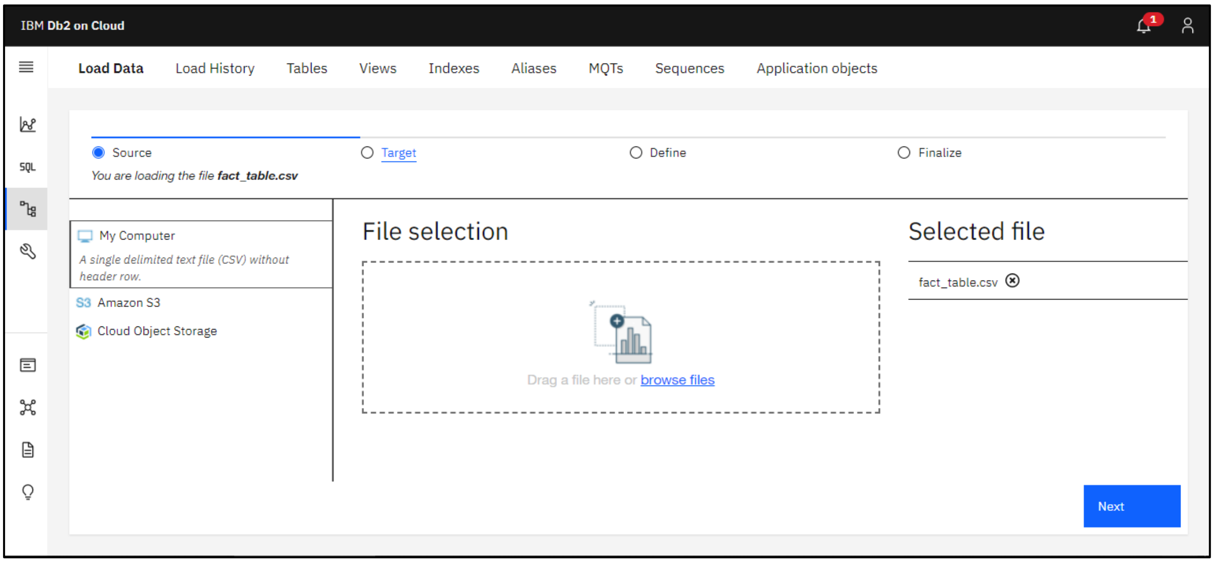

AirportDim
airport_id (PK)
reporting_airport
PeriodDim
period_id (PK)
reporting_period_month
reporting_period_year
Fact Table
period_id (FK)
airport_id (FK)
origin_destination_country_id (FK)
origin_destination_city_id (FK)
airline_id (FK)
airline_type_id (FK)
number_flights_matched
actual_flights_unmatched
number_flights_cancelled
flights_more_than_15_minutes_early_percent
flights_15_minutes_early_to_1_minute_early_percent
flights_0_to_15_minutes_late_percent
flights_between_16_and_30_minutes_late_percent
flights_between_16_and_30_minutes_late_percent
flights_between_31_and_60_minutes_late_percent
flights_between_61_and_120_minutes_late_percent
flights_between_121_and_180_minutes_late_percent
flights_between_181_and_360_minutes_late_percent
flights_more_than_360_minutes_late_percent
flights_unmatched_percent
flights_cancelled_percent
average_delay_mins
flights_cancelled_percent
previous_year_month_flights_matched
previous_year_month_early_to_15_mins_late_percent
previous_year_month_average_delay
CountryDim
origin_destination_country_id (PK)
origin_destination_country
AirlineDim
airline_id (PK)
airline_name
CityDim
origin_destination_city_id (PK)
origin_destination
AirlineTypeDim
airline_type_id (PK)
scheduled_charter

- The above given star schema is simply made by appending all the tables in chronological order and then allotting IDs to all of the dimensions and creating separate tables for them whose primary keys are referenced as foreign keys in the central fact table. The fact table contains foreign keys from the rest of the dimension tables and the measures those were already present in the provided datasets.
- There was only one way in which I found these tables to be created in a relevant manner. As there are more than 26000+ records in the consolidated data warehouse, adding any of the column that contains a numeric value as a column in a dimension table would create consistency issues. Also, there are multiple values of numeric data for a single column of dimensional table which would render the dimension table useless. Therefore, this was the only possible combination to set up the measures and the dimensions distinctively.

3. Data Preparation and Transformation.

- In the primary phase, the data was stitched together by appending all the 12 datasets one below the other. A coding approach was not used to avoid increasing the complexity of the problem
- Next, IDs were added to all the dimensions in the table. The first column, which was deemed irrelevant to the solution, was discarded. The data was then sorted column-wise in alphabetical order, and numerically incremented values were assigned to each dimension column as keys. A formula was utilized to accomplish this task due to the large dataset size of over 26,000 rows.
- The formula employed a basic IF condition to check if the value in the adjacent column cell was equal to the value in the previous cell of the same column. If a change occurred, the value was incremented; otherwise, it remained the same. This process allocated IDs to all six-dimension columns.
- Following that, the dimension table keys and values were extracted from the consolidated database to form the dimension tables. Only the values were deleted from the consolidated table, resulting in the creation of the fact table for the problem
- For the period dimension, the basic numeric data was replaced by splitting it into separate columns for "Month" and "Year".
- Once all the dimension tables and the fact table were formed, a basic schematic structure for the tables was created in IBM DB2. The primary keys from the dimension tables were linked to the fact table as foreign keys, thereby establishing a star schema.
- Subsequently, the provided CREATE TABLE statements were executed to create the tables in IBM DB2.
- Finally, the data files were uploaded to DB2, following the steps demonstrated below:

   a. Step 1 – Select and load the source data file

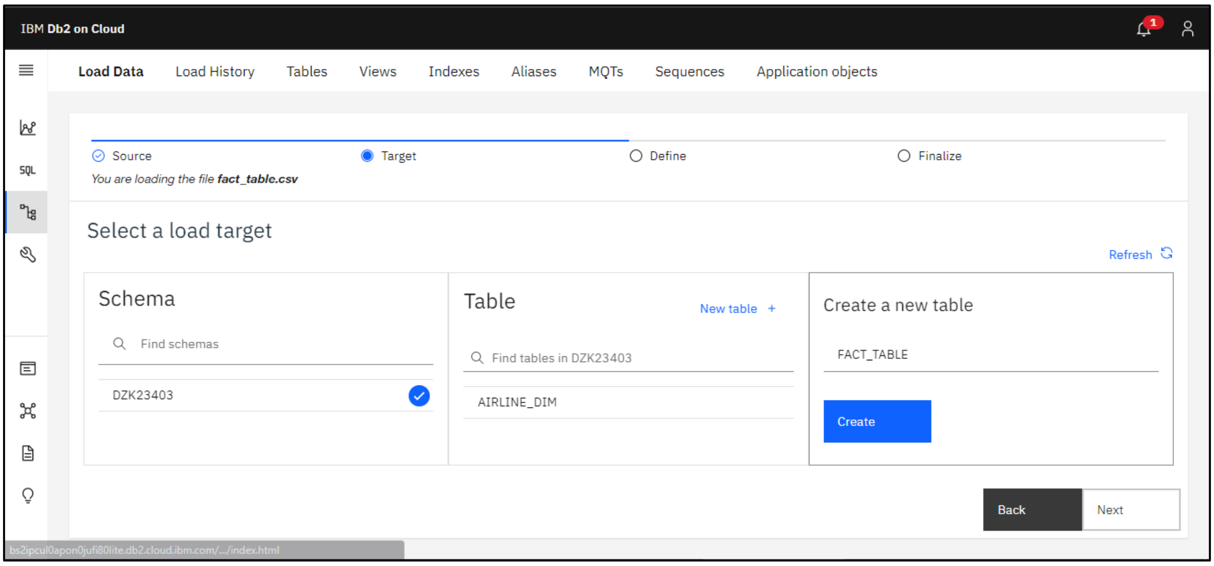


You are loading the file ***fact_table.csv***

A single delimited text file (CSV) without header row.

Selected file: fact_table.csv

   b. Step 2 – Create the Target table which would automatically map the columns to its respective data types



*You are loading the file **fact_table.csv***

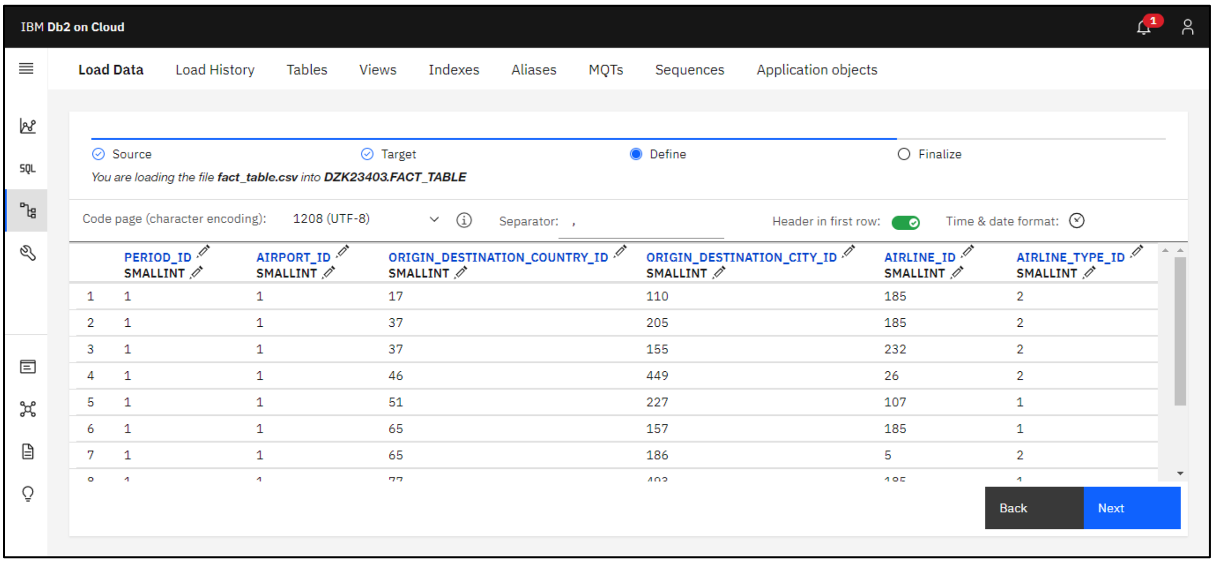
Select a load target

Schema: DZK23403 ✓

Table: AIRLINE_DIM

Create a new table: FACT_TABLE

c. Step 3 – Check if all the data and data types are relevant and matched properly



*You are loading the file **fact_table.csv** into **DZK23403.FACT_TABLE***

Code page (character encoding): 1208 (UTF-8)

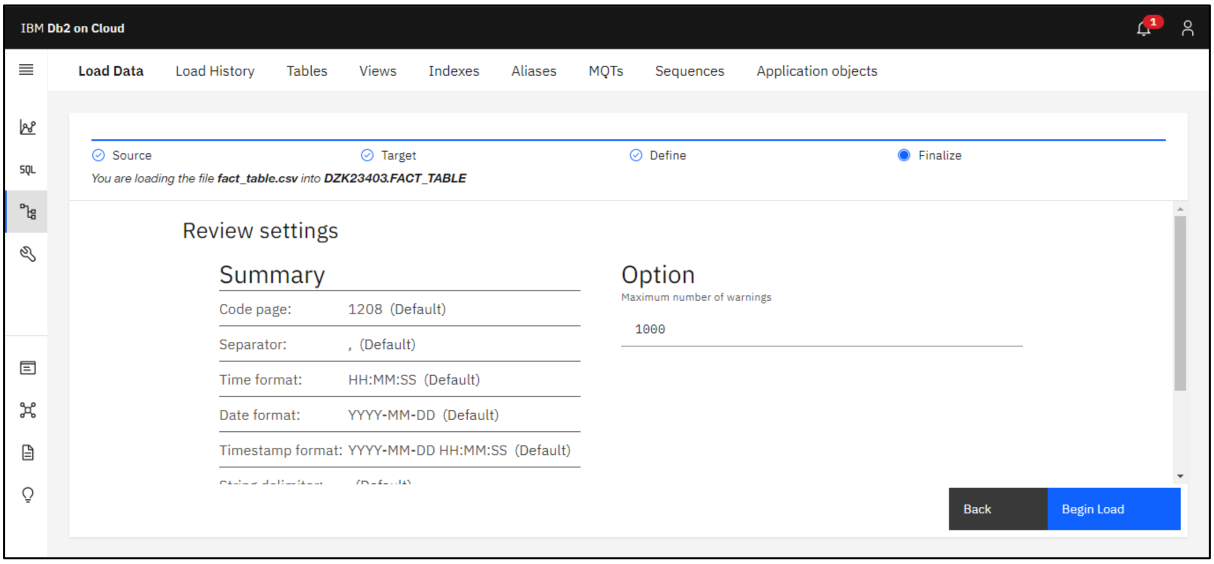
Separator: ,

Header in first row: ✓

Time & date format:

	PERIOD_ID SMALLINT	AIRPORT_ID SMALLINT	ORIGIN_DESTINATION_COUNTRY_ID SMALLINT	ORIGIN_DESTINATION_CITY_ID SMALLINT	AIRLINE_ID SMALLINT	AIRLINE_TYPE_ID SMALLINT
1	1	1	17	110	185	2
2	1	1	37	205	185	2
3	1	1	37	155	232	2
4	1	1	46	449	26	2
5	1	1	51	227	107	1
6	1	1	65	157	185	1
7	1	1	65	186	5	2
8	1	1	77	403	185	1

d. Step 4 – Finalize the table to load the data



*You are loading the file **fact_table.csv** into **DZK23403.FACT_TABLE***

Review settings

Summary

Code page: 1208 (Default)

Separator: , (Default)

Time format: HH:MM:SS (Default)

Date format: YYYY-MM-DD (Default)

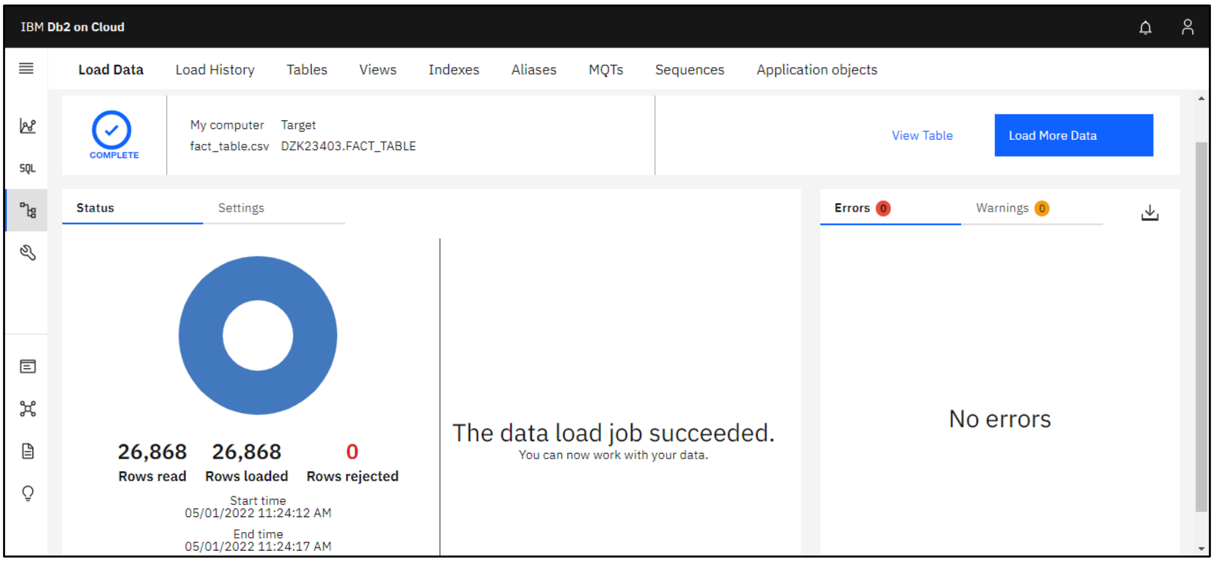
Timestamp format: YYYY-MM-DD HH:MM:SS (Default)

String delimiter: (Default)

Option

Maximum number of warnings: 1000

e. Step 5 – Confirmation of data upload



My computer	Target
fact_table.csv	DZK23403.FACT_TABLE

26,868	26,868	0
Rows read	Rows loaded	Rows rejected

Start time
05/01/2022 11:24:12 AM

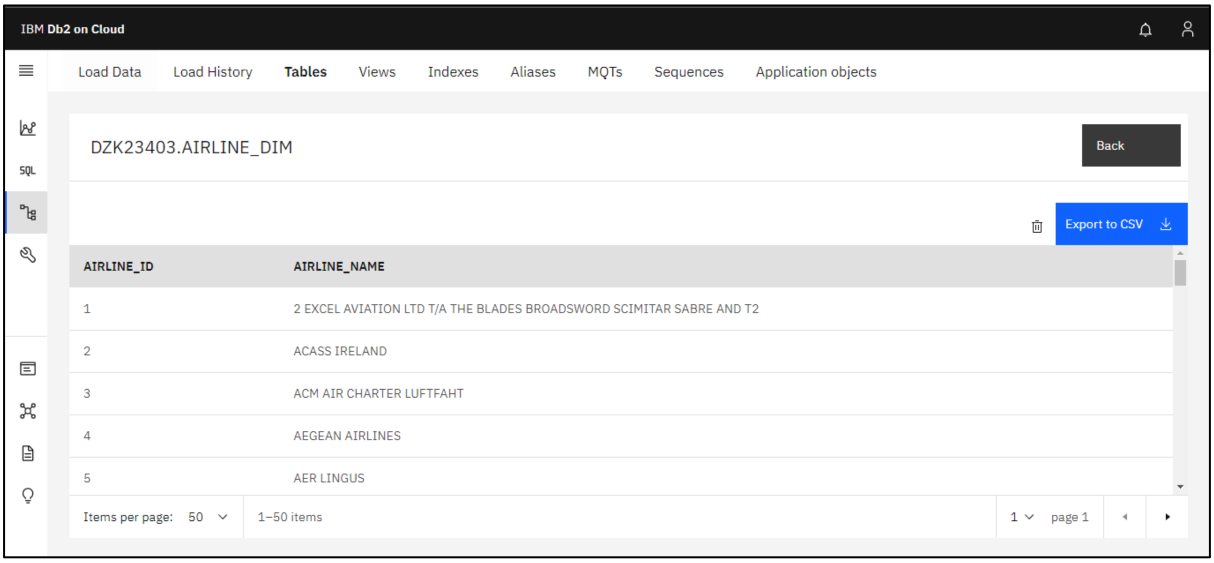
End time
05/01/2022 11:24:17 AM

The data load job succeeded.
You can now work with your data.

No errors

- We can similarly upload the data for all the tables.
- Snapshots of data for all the dimension tables and the fact table

a.

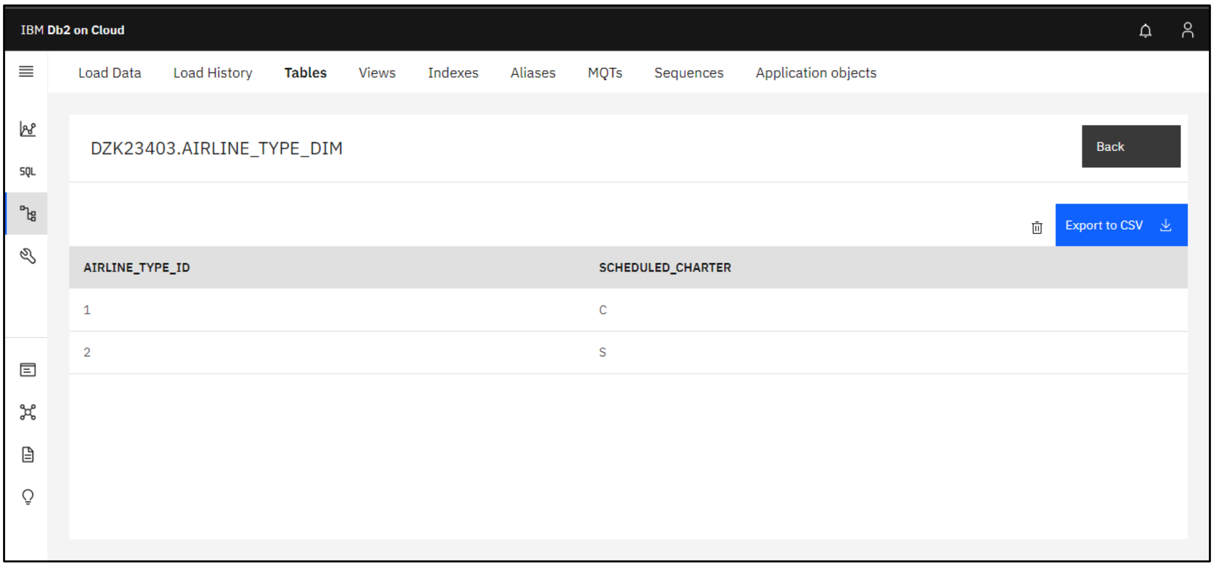


DZK23403.AIRLINE_DIM

AIRLINE_ID	AIRLINE_NAME
1	2 EXCEL AVIATION LTD T/A THE BLADES BROADSWORD SCIMITAR SABRE AND T2
2	ACASS IRELAND
3	ACM AIR CHARTER LUFTFAHT
4	AEGEAN AIRLINES
5	AER LINGUS

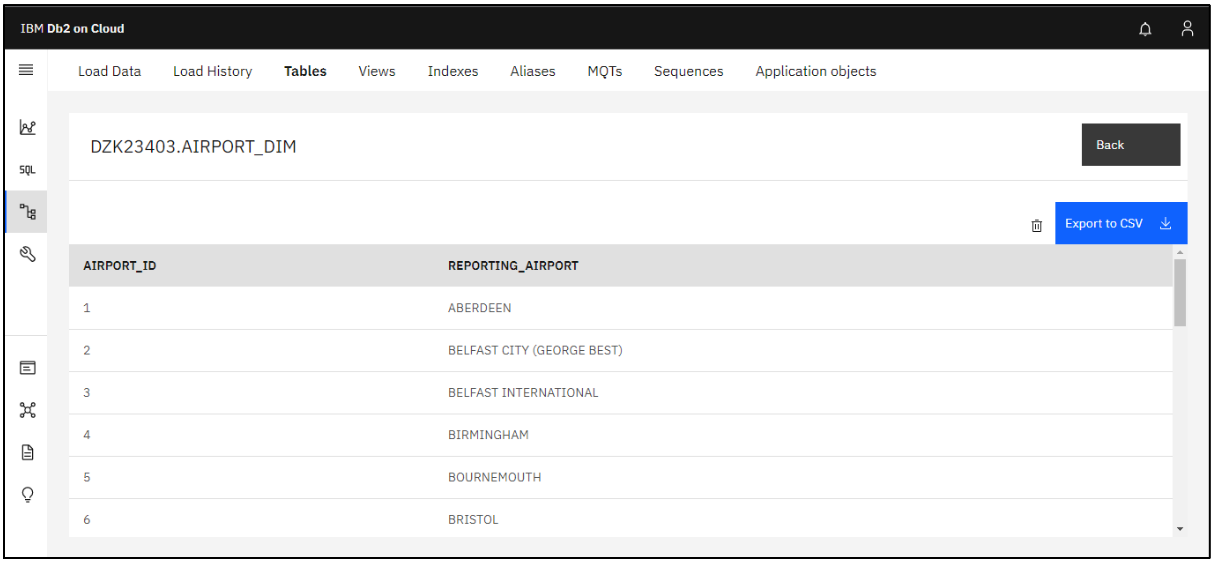
Items per page: 50 1–50 items 1 page 1

b.



DZK23403.AIRLINE_TYPE_DIM

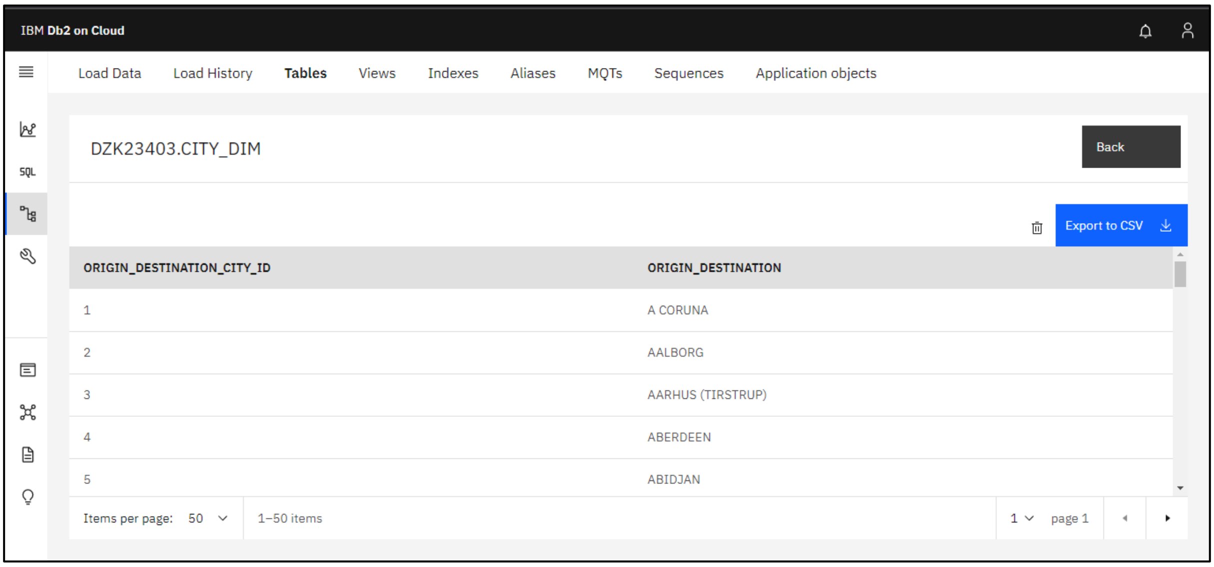
AIRLINE_TYPE_ID	SCHEDULED_CHARTER
1	C
2	S



DZK23403.AIRPORT_DIM

AIRPORT_ID	REPORTING_AIRPORT
1	ABERDEEN
2	BELFAST CITY (GEORGE BEST)
3	BELFAST INTERNATIONAL
4	BIRMINGHAM
5	BOURNEMOUTH
6	BRISTOL

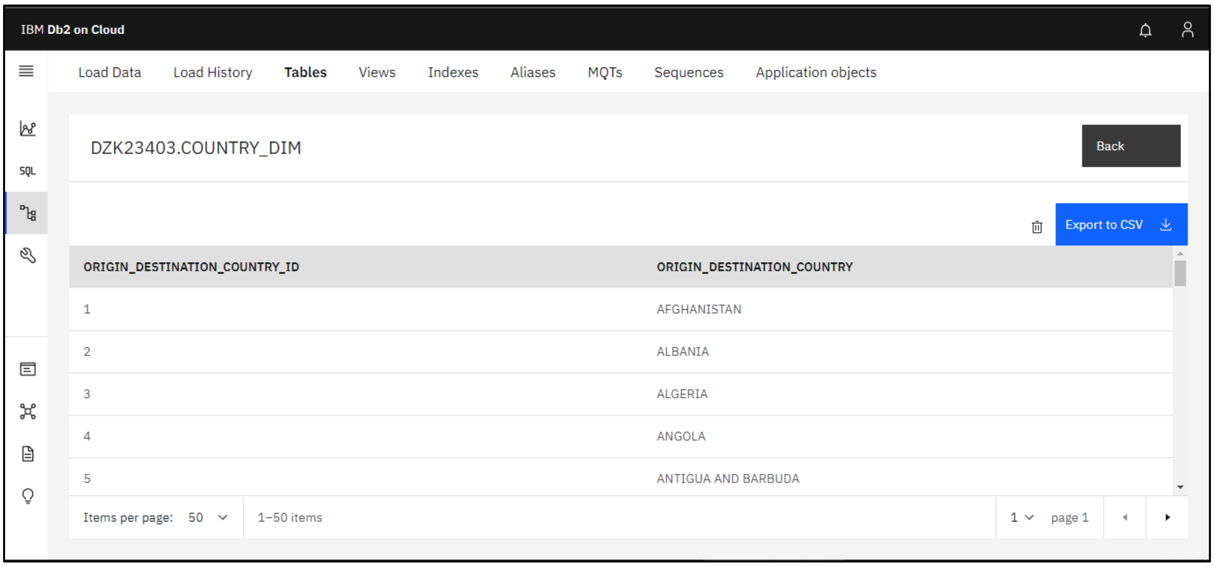
c.



DZK23403.CITY_DIM

ORIGIN_DESTINATION_CITY_ID	ORIGIN_DESTINATION
1	A CORUNA
2	AALBORG
3	AARHUS (TIRSTRUP)
4	ABERDEEN
5	ABIDJAN

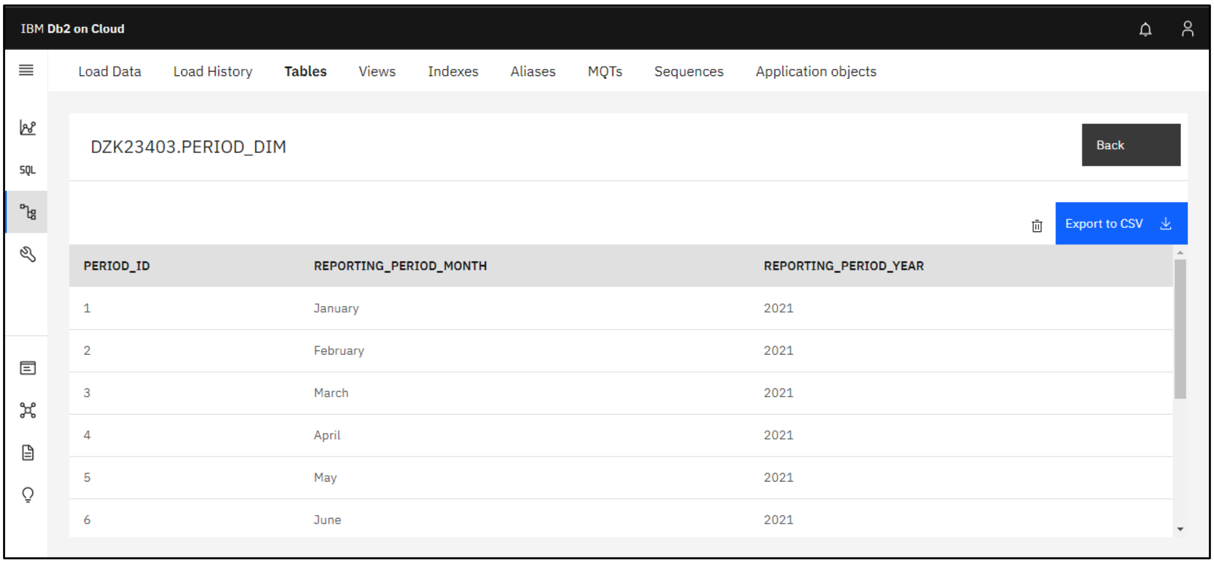
d.



DZK23403.COUNTRY_DIM

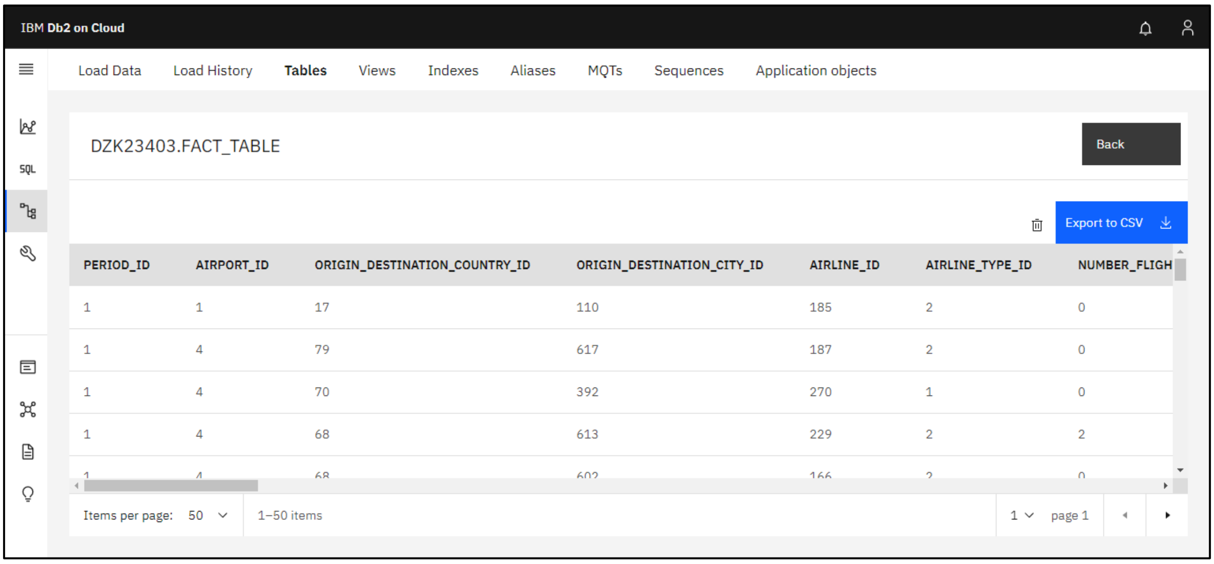
ORIGIN_DESTINATION_COUNTRY_ID	ORIGIN_DESTINATION_COUNTRY
1	AFGHANISTAN
2	ALBANIA
3	ALGERIA
4	ANGOLA
5	ANTIGUA AND BARBUDA

e.

f. 

DZK23403.PERIOD_DIM

PERIOD_ID	REPORTING_PERIOD_MONTH	REPORTING_PERIOD_YEAR
1	January	2021
2	February	2021
3	March	2021
4	April	2021
5	May	2021
6	June	2021

g. 

DZK23403.FACT_TABLE

PERIOD_ID	AIRPORT_ID	ORIGIN_DESTINATION_COUNTRY_ID	ORIGIN_DESTINATION_CITY_ID	AIRLINE_ID	AIRLINE_TYPE_ID	NUMBER_FLIGH
1	1	17	110	185	2	0
1	4	79	617	187	2	0
1	4	70	392	270	1	0
1	4	68	613	229	2	2
1	4	68	602	166	2	0

Items per page: 50 ⌄ 1–50 items 1 ⌄ page 1

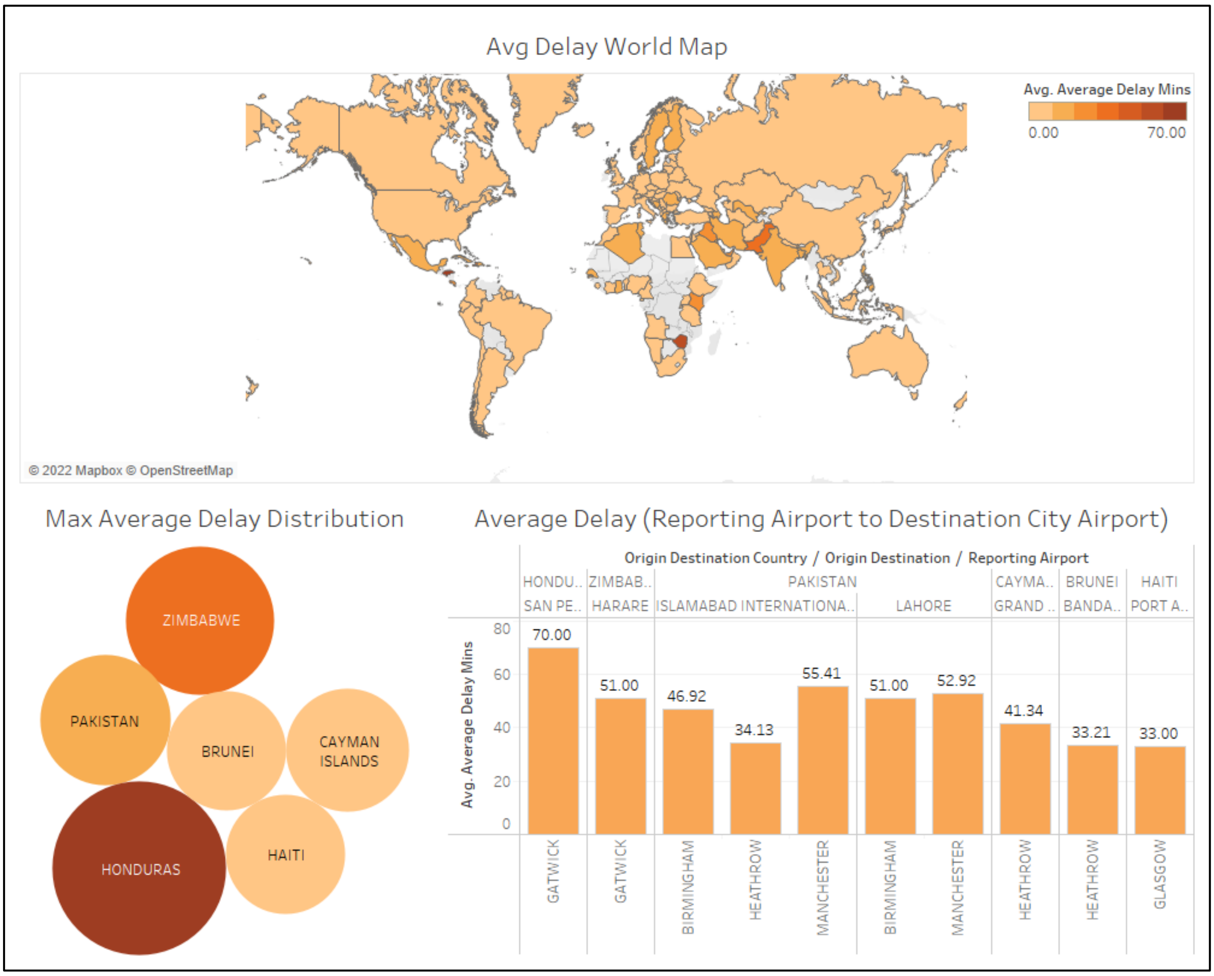
4. Benefits of Data Warehouse and Tableau.

- Centralized Data Repository: A data warehouse serves as a central repository that consolidates data from various sources, integrating them into a unified and structured format. By connecting Tableau to the data warehouse, organizations gain access to a single source of truth, ensuring data consistency and eliminating data silos. This enables users to access comprehensive and reliable data for analysis and reporting.
- Improved Data Accessibility and Visualization: Tableau provides intuitive and interactive visualizations that make it easier to explore and understand complex data. By connecting to a data warehouse, Tableau can directly access large volumes of data and leverage its capabilities to create dynamic dashboards, charts, and graphs. Users can visually explore data, drill down into specific details, and uncover actionable insights quickly.
- Real-Time and Near-Real-Time Reporting: With a data warehouse feeding data to Tableau, organizations can generate real-time or near-real-time reports and dashboards. This enables stakeholders to monitor key performance indicators (KPIs), track business metrics, and make timely decisions based on up-to-date information. The combination of a data warehouse and Tableau empowers users with live data analysis and reporting capabilities
- Advanced Analytics and Data Exploration: Tableau offers a wide range of advanced analytics features, such as statistical analysis, forecasting, and predictive modelling. By connecting to a data warehouse, Tableau can leverage the enriched and structured data to perform in-depth analysis and derive meaningful insights. Users can identify patterns, trends, and correlations, enabling data-driven decision-making and strategic planning.
- Self-Service Analytics: Tableau's self-service capabilities allow users across the organization to access and analyse data without heavy reliance on IT or data experts. With a data warehouse as the underlying data source, Tableau provides a user-friendly interface that empowers business users to create their own reports, perform ad-hoc analysis, and answer their own data-related questions. This self-service approach fosters a culture of data-driven decision-making throughout the organization.
- Scalability and Performance: Data warehouses are specifically designed to handle large volumes of data and support scalable data processing. By integrating Tableau with a data warehouse, organizations can efficiently retrieve and analyse vast amounts of data without compromising performance. The combination enables seamless scalability as data volumes increase, ensuring fast and responsive analytics and reporting.
- Data Governance and Security: Data warehouses often implement robust data governance practices, including data quality management, data lineage tracking, and access controls. By utilizing a data warehouse with Tableau, organizations can enforce data governance policies and ensure data security. Tableau's integration with the data warehouse allows for controlled access to sensitive data, ensuring compliance with regulations and maintaining data privacy
- Collaboration and Sharing: Tableau provides collaboration features that enable users to share insights, reports, and dashboards with others. By connecting to a data warehouse, teams can collaborate on data analysis, annotate visualizations, and share findings across departments. This promotes data-driven decision-making and facilitates knowledge sharing within the organization.

5. Tableau Visualizations containing:
    - Aim of the visualisation
    - The steps you took to create the visualisation
    - Key findings from the visualisation

A. Average Delay (World)

Aim – To analyse which countries have the maximum average delay and to visualize average delays from specific reporting airports to destination cities in the countries with maximum average delays



Steps – This is a dashboard prepared using 3 different types of visualizations.

1. Starting with the World Map, I used the Origin Destination Country from the Country Dimension Table which was automatically plotted to a world map using the latitudinal and longitudinal data that Tableau has.
2. After that I added AVG of Average Delay Mins to Colours so that it distinctively shows which countries have the highest average delay.
3. Then I used Packed Bubbles to visualize the countries which have average delays between 30 mins and the max delay i.e., 70 mins using filters. After that, I added AVG of Average Delay Mins to Colours in order to colour code the countries.

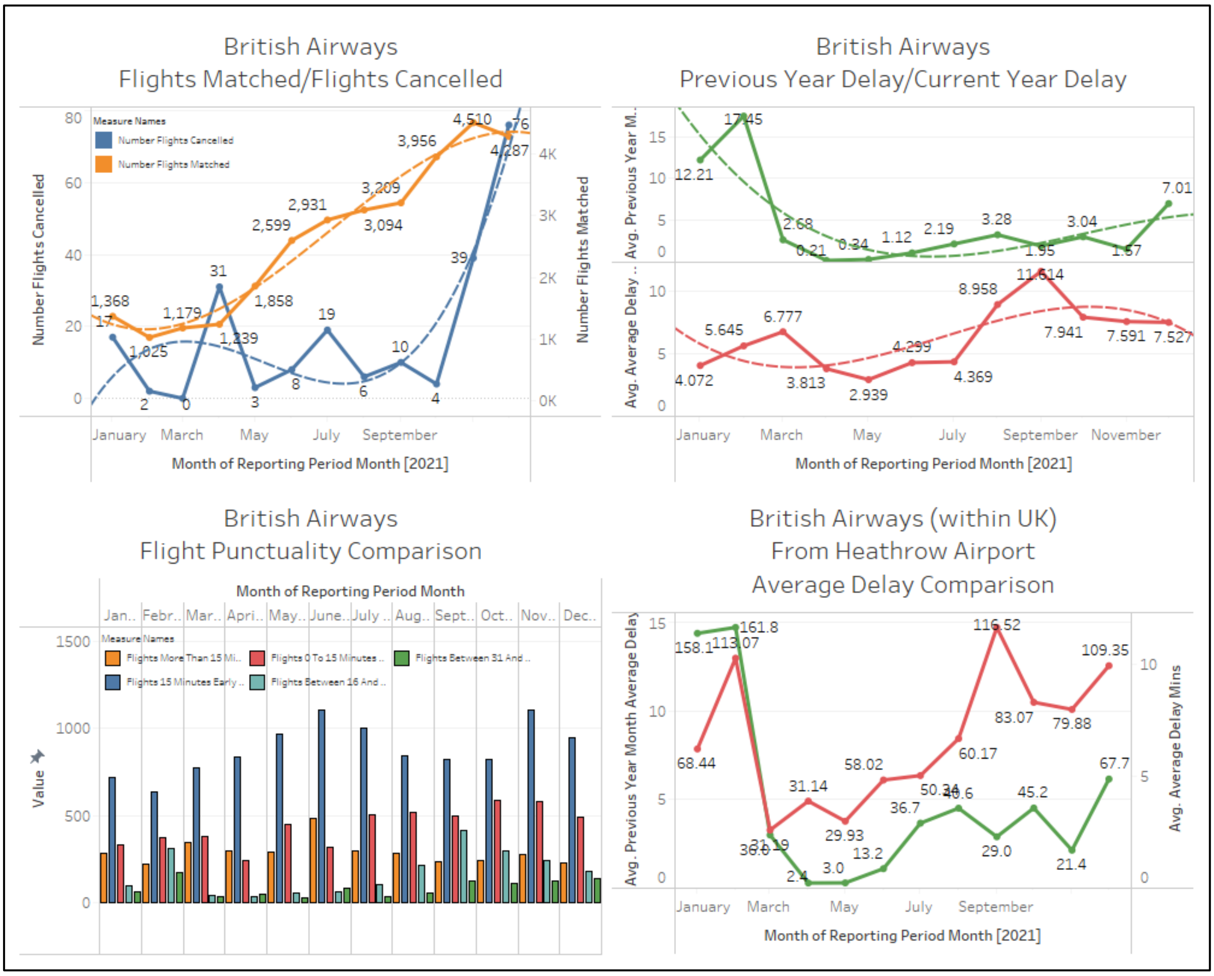
4. Analysing data from the Packed Bubbles, I created side-by-side bar graphs which displayed the average delay from each reporting airport to the origin city for the countries with max average delay. I used filters for the origin city and average delay.
5. At the end, I consolidated all of these 3 visualizations into 1 dashboard which effectively demonstrates the story of average delays throughout the world and then goes on to minimalizing the dimensions to get us the actual average delays from and to specific places in the countries having more than 30 minutes of average delay.

Key Findings – After analysing the visualisations we can come to a conclusion that the countries from the Caribbean Sea feature in 50% of the data chunk taken out for visualisation. Also, we can see that Heathrow airport features the greatest number of times as the reporting airport with max average delays. Heathrow is closely followed by Gatwick, Manchester and Birmingham being in the mix. We can also observe that the average delay from the Heathrow airport is in the 30's to early 40's which shoots to 50's for Manchester airport but for Birmingham airport it is in the mid 40's to early 50's. For Gatwick as we can see the average delay is quite high but for Glasgow in contrast is the lowest.

I have used 3 different kinds of visualisations in this dashboard starting from the bigger picture and then granularizing it to the most analysis-worthy format.

B. British Airways Analysis

Aim – The aim of this dashboard is to analyse the punctuality statistics for British Airways in specific and contains various comparisons on various measures along with analysing the trends.



Steps – All the visualisations in the above dashboard are made with the same filter of Origin Destination Country being set to the United Kingdom and Airline Name set to British Airways in order to monitor and visualise the operations of British Airways in the UK.

1. For the 1st visualisation, I added the SUM of Number of Flights Cancelled and Number of Flights Matched in the rows and the Reporting Period Months in the columns. For the labels, I just added the same fields added to the rows to labels. Then, I selected dual lines to make the visualisation appealing and comparative. Also, from the Analytics tab I used trend lines to plot the trend on the dual lines.
2. For the 2nd visualisation, I kept the columns field same, I just added the measures of Previous Year Average Delay Month and Average Delay as AVG and used Discrete Lines as a visualisation and used dotted trend lines from the Analytics tab.
3. For the 3rd visualisation, keeping the columns same, I added the SUM of measures for flights – early than 15 minutes, between 15 minutes to 1 minute early, 0 to 15 minutes late, 16 to 30 minutes late and 31 to 60 minutes late in the rows field. As this created a new variable Measure Value, I added it to Colour to have distinctly recognizable colours for the side-by-side bar graph.
4. For the 4th visualisation, again keeping the timeline same, I added a filter for Heathrow airport to compare the delay from previous year to this year.

Key Findings – Using multiple visualisation techniques in this dashboard, I was able to draw out multiple observations. For the 1st visualisation, we can see the trend for the number of flights matched dipping at the start but starts escalating from the month of April with minor disturbances to the exponential growth coming in the month of September where it takes a dip and towards the end of the year whereas the number of flights in the same period is on a rollercoaster ride right from the start. But the important thing to observe is that the number of flights cancelled towards the end of the year sky rocket from 4 in October to 76 in December.

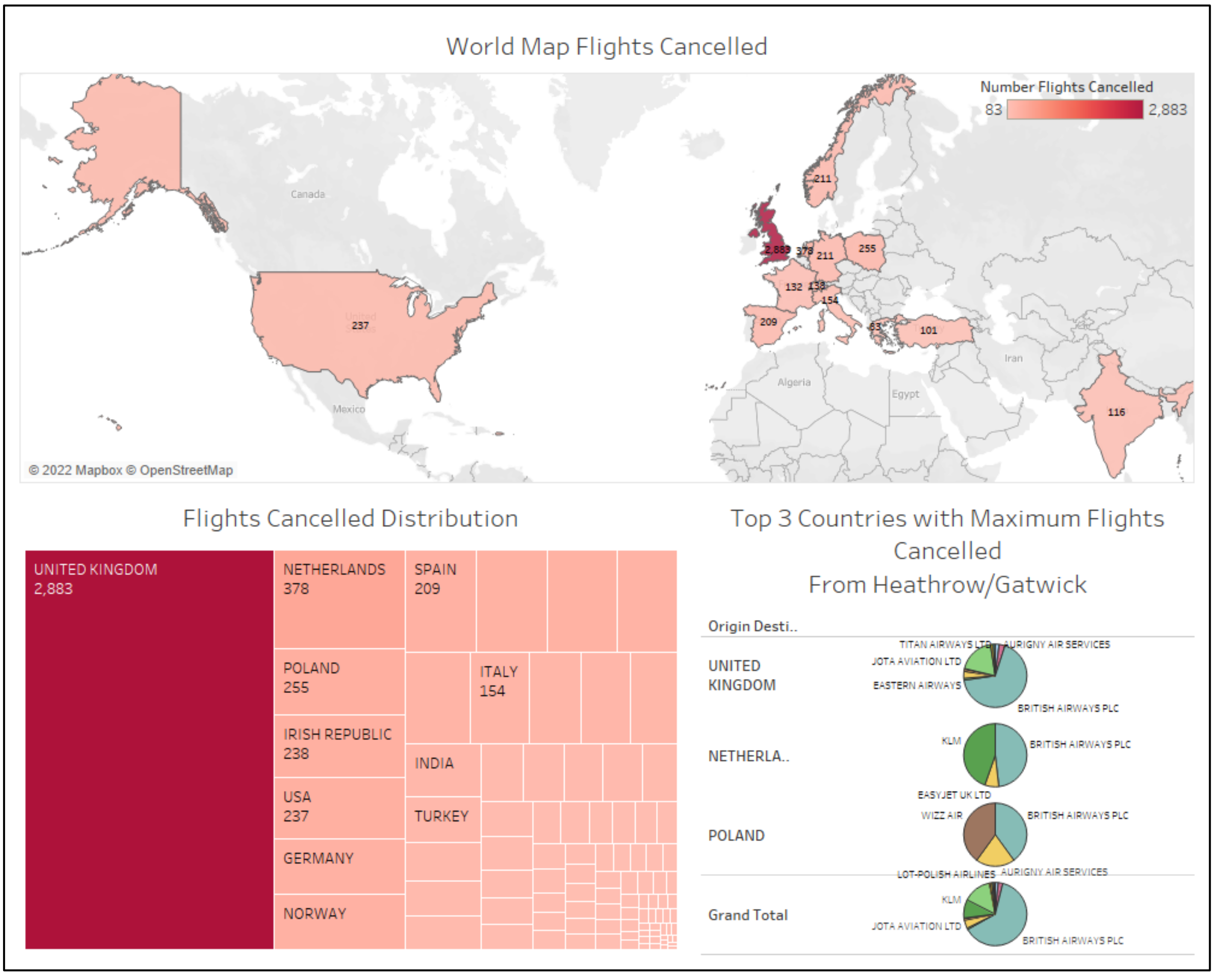
For the 2nd visualisation, it is interesting to see contrasting trends in delays of previous year compared to this year. For the previous year, the trend is seen to dip drastically after starting the year on a high and reaches a global minimum in April from where on it takes on an upward trajectory. In comparison for this year, the general trend is the same for the first quarter though not having a drastic dip. On similar lines, after reaching the global minimum in May, the trajectory is observed to be going upwards reaching the global maximum in September.

For the 3rd visualisation, it is clearly observed that the flights that are 15 mins to 1 min early are the highest in number throughout the entire year. Also, the common trend observed is that the flights that are 0 to 15 minutes late are the 2nd highest throughout the year followed by flights that are 15 mins or more early. Another clear trend is observable between the flights which are late from 16 to 30 minutes and 31 to 60 minutes where the latter is always lesser in number than the former with some tiny exceptions in April and June where it is vice versa. From this we can strictly say that majority of the flights take off in the time slot of ±15 minutes from their take off time.

For the last visualisation, we can see some similarities at the start of the year where delay reaches a high point in February in the previous year as well as this year but falls impressively in the month of March. But then we see some growth from there with minor hiccups taking place every now and then with dips and rises throughout the year. We also observe that when comparing both the years, the trend from October to December has some resemblance where it dips from October to November and then shoots up from November to December. This can be used to analyse what factors create such sort of distortions and what measures can be implemented to address those issues. We can also term it as a gift of the Christmas vacations where the people travel more often leading to more delays.

C. Flight Cancellation Statistics

Aim – To visualise the cancellation of flights throughout the world and gain insights from them.



Steps – For the 1st visualisation, I have used the measure of Number Flights Cancelled and filtered the data from 80 flights cancelled to the maximum number of flights cancelled. I used Origin Destination Country from Country Dimension Table to plot the world map in the column field which was automatically mapped to its latitudinal and longitudinal coordinates by Tableau. Also, I added the number flights cancelled to Colour in order to have a distinctively visualisable distribution. In addition to that, I added the filtered data of number of flights cancelled as a label to the visualisation to make it self-explanatory.

For the 2nd visualisation, I used the same data and filters from the first visualisation but in a Tree Map in order to demonstrate its distribution and the contribution of each country to the cancellation of flights.

For the 3rd visualisation, I used the same filters but added a couple of restrictions on them. Firstly, I added the reporting airport filter and filtered only the values corresponding to Heathrow and Gatwick airports and added a filter of airline name and filtered the airlines by having number of cancelled flights set to greater than 0.

Key Findings – From the 1st visualisation, we filtered the countries with number of flights cancelled which were greater than 80 and it was observed that the United Kingdom had a gigantically humongous number of flights cancelled when compared to its counter-parts. It was also observed that the majority of the countries with major flight cancellations belong to the continent of Europe with the outsiders just being USA and India.

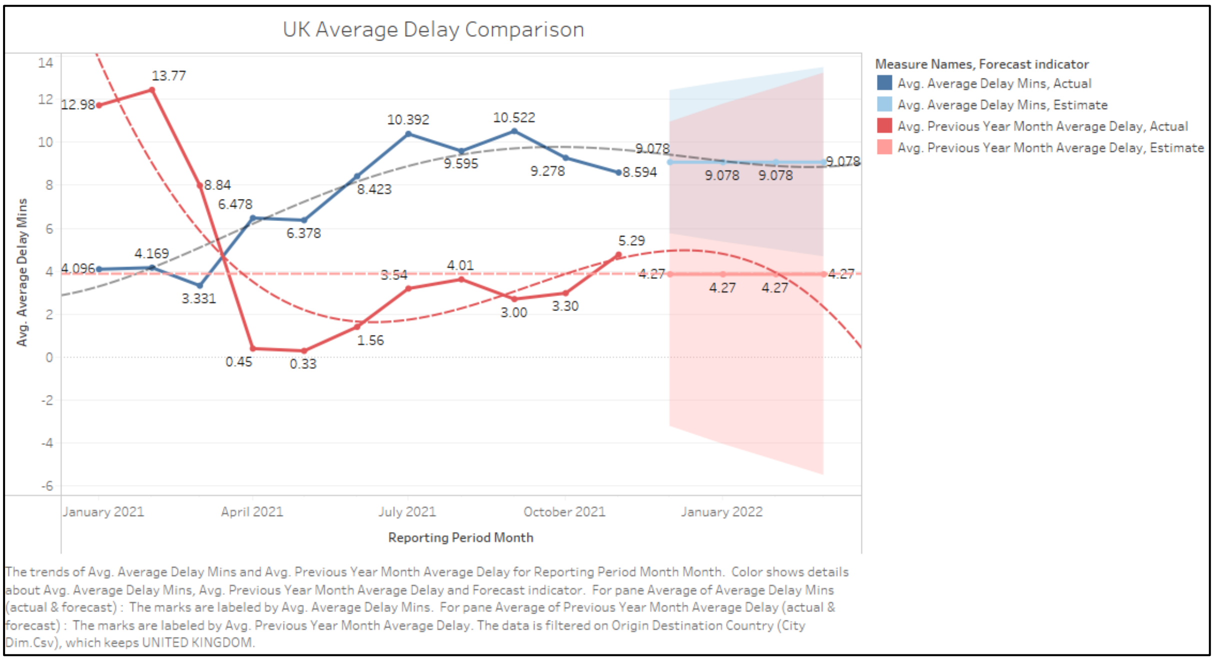
As observed from the 1st visualisation, the tree maps effectively demonstrate what was quite clear about the contribution of the UK to the flight cancellation statistics. The 2nd being Netherlands followed by Poland and the Irish Republic. As the tree maps also confirm the observation from the world map about the European countries being the major contributors to flight cancellations statistics.

When we move on to the 3rd visualisation, we see that the top 3 countries with maximum flights cancelled from Heathrow/Gatwick airports along with the airline names. We see British Airways to be the major contributor with the maximum number of flights being cancelled collectively throughout the top 3 countries. In the grand total, we can see that British Airways collectively results towards a higher contribution in the pie chart.

I have used tree maps and pie charts as they effectively demonstrate the distribution of data and also their contribution to the visualisation.

D. UK Average Delay Comparison

Aim – The aim of this visualisation is to have a comparative analysis for the average delay of flights in the UK from the past year to this year.



The trends of Avg. Average Delay Mins and Avg. Previous Year Month Average Delay for Reporting Period Month Month. Color shows details about Avg. Average Delay Mins, Avg. Previous Year Month Average Delay and Forecast indicator. For pane Average of Average Delay Mins (actual & forecast) : The marks are labeled by Avg. Average Delay Mins. For pane Average of Previous Year Month Average Delay (actual & forecast) : The marks are labeled by Avg. Previous Year Month Average Delay. The data is filtered on Origin Destination Country (City Dim.Csv), which keeps UNITED KINGDOM.

Steps – I initially added a filter for Origin Destination Country to the United Kingdom. Thereafter, in the columns section I added reporting period month to have a timeline generated in which having a trend and a forecast line would be possible. After that, I added the AVG values of Average Delay Mins and Previous Year Average Delay Month to the rows field to generate the dual line graph which is useful to compare the data without any ambiguity. Post that, I added the same values added in the rows to the labels field to annotate the line graphs with values they demonstrate. Once all of this was set up, I added a trend line to observe the trend in delay throughout the year. After that, I added a forecast line to have a tentative idea of the prediction for the future months.

Key Findings – We can clearly observe that at the start of the previous year the average delay was a sky touching point. It started dipping since February and almost reached 0 in April. Since then, it has just gone upwards with some minor dips throughout the year. Whereas, for the current year the trend is much more like a rollercoaster ride and keeps on rising and dipping at regular intervals. One of the most important insights to draw from this is the forecast from Previous Year Delay (Pink) suggests that the following year (current year) would have somewhat tentative readings of 4.27 until April. As we delve deeper towards the start of the current year's line (Blue) we see that the readings for the months of January and February are pretty close to the ones predicted by the pink line which is 4.096 and 4.169 respectively. We can clearly infer from this that the prediction made by the forecast line was approximately accurate and would in turn provide us with useful data and insights which can lead us to make data-driven informed decisions.

6. Conclusion

In this project, a number of different processes were employed, ranging from data transformation to obtaining visualizations. A thorough understanding of data warehousing and OLAP techniques was gained throughout the project. The process began by merging the data to create a data warehouse with a substantial number of rows. To extract insights from this extensive data warehouse, various ETL processes were conducted, including data cleansing. The project then progressed to designing a schema that could effectively store and correlate the data, leading to the adoption of a Star Schema. The Star Schema facilitated organizing the data in a presentable and accessible manner. Recognizing a specific pattern in the arrangement of columns within the fact table and the data warehouse as a whole, dimension tables were created accordingly. Subsequently, SQL statements were utilized to create tables in IBM DB2, followed by uploading the data to these tables using the CSV upload option.

Next, the project focused on linking the database to Tableau for further analysis and visualization. Through Tableau, the processed yet raw data in the data warehouse was visualized to extract important insights related to various parameters. Different types of data visualization techniques were explored, understanding how they could be effectively employed to interpret the data. The project also highlighted the significance of data pre-processing in generating effective and easily understandable visualizations. By transforming and grouping the data in a suitable manner, relevant visualizations were designed to convey meaningful insights. Tools such as Tableau and data warehousing options like IBM DB2 offered flexibility and compatibility across different mediums, addressing the requirements of the project.

Tableau, as a highly useful business intelligence tool, played a crucial role in the project. Business intelligence involves utilizing data analysis, information retrieval, data visualization, analytics tools, and best practices to aid organizations in making data-driven decisions. By having a comprehensive understanding of the organization's data and leveraging that information, contemporary business intelligence allows for driving change, eliminating inefficiencies, and swiftly adapting to market or source changes. The project demonstrated the capabilities of Tableau in exploring and analysing the data warehouse to extract valuable information presented in a visual format. It provided insights into various aspects of the given data warehouse, such as average delay analysis, flight cancellations, and a comparison between historical and current data. By identifying and showcasing trends within the data, the project aimed to guide organizations in making data-driven decisions that would contribute to their long-term success.