Uma imagem com gráfico

Descrição gerada automaticamente

Table of Contents

[2. Subject description 3](#_Toc132059299)

[3. Planning 5](#_Toc132059300)

[3.1 Dimensional bus matrix 5](#_Toc132059302)

[3.2 Dimensions 5](#_Toc132059305)

[3.3 Facts dictionary 8](#_Toc132059306)

[3.4 Scope 9](#_Toc132059307)

[4. Dimensional data model 11](#_Toc132059308)

[5. Data sources selection: Extraction, transformation and loading 15](#_Toc132059309)

[6. Querying 17](#_Toc132059310)

[7. Data Analysis 18](#_Toc132059311)

[8. Reflection 20](#_Toc132059312)

[9. Conclusion 21](#_Toc132059313)

[10. Next Steps 22](#_Toc132059316)

[11. Referencies 23](#_Toc132059317)

Table INDEX

[Table 1 – Passenger 6](#_Toc132058976)

[Table 2 – Airline 7](#_Toc132058977)

[Table 3 – Aircraft 7](#_Toc132058978)

[Table 4 – Airport 7](#_Toc132058979)

[Table 5 – Date 8](#_Toc132058980)

[Table 6 – Time 8](#_Toc132058981)

[Table 7 – Schedule 8](#_Toc132058982)

[Table 8 – StatisticsPeriod 9](#_Toc132058983)

[Table 9 – Flight\_Class 9](#_Toc132058984)

[Table 10 – Flight 9](#_Toc132058985)

[Table 11 – Booking 10](#_Toc132058986)

[Table 12 – Flight\_Statistic 10](#_Toc132058987)

[Table 13 – Scope 11](#_Toc132058988)

FIGURE INDEX

[Figure 1 – Data Warehouse Architecture 4](#_Toc132059009)

[Figure 2 – Feasibility versus Impact 5](#_Toc132059010)

[Figure 3 – Dimensional Bus Matrix 6](#_Toc132059011)

[Figure 4 – Dimensional Data Model 16](#_Toc132059012)

[Figure 5 – Loading Data 17](#_Toc132059013)

[Figure 6 – Descriptive Statistics 17](#_Toc132059014)

[Figure 7 – Create Booking 18](#_Toc132059015)

[Figure 8 – Load Airport 18](#_Toc132059016)

[Figure 9 – Load Booking 19](#_Toc132059017)

[Figure 10 – Distance Function 19](#_Toc132059018)

[Figure 11 - PBI Report 20](#_Toc132059019)

[Figure 12 - PBI Dimensional Model 21](#_Toc132059020)

1. Introduction

Organizations constantly strive to improve their decision-making processes in order to maintain a competitive advantage and improve business processes in a global environment. Even though businesses accumulate and store a significant amount of data on a daily basis, this data comes from a variety of disparate sources, making it challenging to review and aggregate. However, in today's highly competitive environment, businesses require immediate access to complex, timely analysis on a consolidated view of high-quality data. The development of data mining and data warehousing, both of which have the potential to support an organization's efforts to effectively utilize their data, has been facilitated by advancements in technology and business processes, improved data management, an increase in the availability of information, and a decrease in the costs associated with storage.

Data mining is the automated process of extracting data, analyzing it from various angles, and then putting it all together into useful information. Organizations can use data mining to connect a large amount of disparate data and gain knowledge they can use to predict future trends and behaviors, reduce costs, boost revenues, and enhance processes. On the other hand, data warehousing is the process of managing and retrieving data from a central location. It helps turn a lot of data into information that is useful and reliable that businesses can use to stay competitive and efficient. As a result, data warehouses not only serve as the foundation for effective data analysis methods like data mining, but they also give users the information they need to make decisions.

As it does for many modern organizations, data warehousing also plays a crucial role in the efficient management and analysis of data in modern flight reservation systems. As the airline industry generates massive amounts of data from various sources, including booking information, passenger details, and flight schedules, a well-designed data warehouse becomes a critical component for consolidating, storing, and analyzing this data in a unified and organized manner.

A data warehouse for a flight reservation system can serve as a central repository that integrates data from multiple airports, airlines, and other sources to provide a comprehensive view of the system's operations. It enables airlines and airports to gain valuable insights into booking trends, passenger demographics, flight schedules and other key metrics which can be used for strategic decision-making and operational optimization.

# Subject description

The aim of this project is to design a comprehensive data warehouse for a flight reservation system that consolidates and manages data from various sources, including web-based information on bookings, passengers, and flights from multiple airports and airlines. The data warehouse will be designed to efficiently store, retrieve, and analyze large volumes of data to support real-time decision-making and enable effective business intelligence for the flight reservation system.

The project will involve designing a robust data model that can handle complex data relationships. This will require data extraction, transformation, and loading (ETL) processes to clean, validate, and transform data into a unified format suitable for analysis.

The data warehouse will provide a comprehensive and holistic view of the flight reservation system, enabling analysis and reporting on various aspects such as bookings, passengers, flight schedules, airline performance, and airport utilization. Advanced analytics and reporting capabilities will be implemented to provide insights and trends for strategic decision-making, operational optimization, and customer relationship management. The data we are working with is extracted from the web on two different sources: Kaggle.com and Openflights.org.

Uma imagem com diagrama

Descrição gerada automaticamente

*Figure 1 – Data Warehouse Architecture*

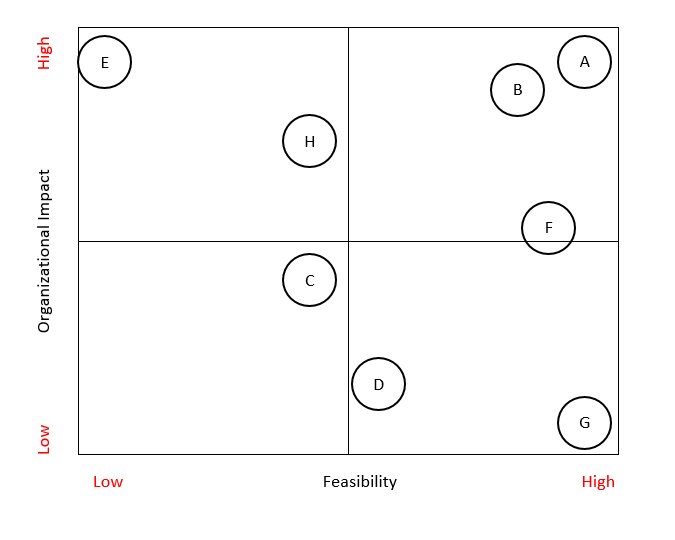
The project will start in the initial planning phase where a dimensional bus matrix is built based on the data mart, dimensions and facts recognized from the data set. Later, it will be implemented and the data will be analyzed with the use of tools such as MySQL, Power BI and Python. The final deliverables of the project will be a well-designed and documented data warehouse architecture, ETL processes, and data models that meet the requirements of the flight reservation system. Below we have listed the main KPIs and requirements recognized for the project.

KPIs:

* 1. How many passengers per flight were transported;
  2. How many stops for a passenger;
  3. Number of airports;
  4. Number of operating Airlines;
  5. Ability to detect delays.

Requirements:

* 1. The number of flights to and from each airport;
  2. Records of delays;
  3. Include passengers’ samples from at least two years;
  4. Comparison between airlines based on number of passengers transported;
  5. Ordered airports in terms of total stops;
  6. Number of bookings made for an airline;
  7. Number of aircrafts for each airline;
  8. Cities with most flights.



*Figure 2 – Feasibility versus Impact*

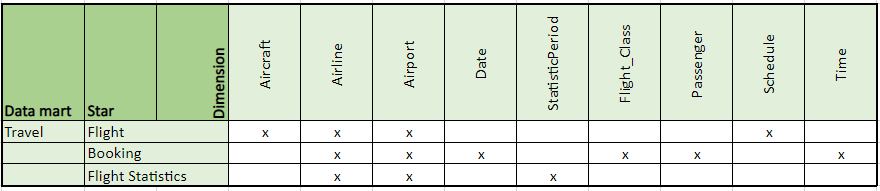
# Planning

The planning phase of a data warehouse is a critical step in designing a robust and effective data warehousing solution. It involves several key components, including designing a bus matrix, creating a dictionary of dimensions, and developing a dictionary of facts. These components provide a solid foundation for designing a well-structured and effective data warehouse, ensuring alignment between business requirements and technical implementation, and facilitating efficient data integration, data modeling, and data analysis.



## Dimensional bus matrix

Below figure 1 represents the dimensional bus matrix. As can be seen from the graph, the data mart *travel* has three different stars: flight, booking and flight statistics. The 9 dimensions recognized from the data set are connected to one or more of these stars based on connectivity and necessity for a certain star. The bus matrix helps to identify the data sources, as well as the dependencies between them.



*Figure 3 – Dimensional Bus Matrix*



## Dimensions

The below tables 1 - 9 present the dictionaries of the dimensions in the travel data mart. Descriptions of each dimension, its attributes as well as the type and size are presented. They serve as a reference guide for designing the dimensional model of the data warehouse.

*Table 1 – Passenger*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Name*** | ***Description*** | ***SCD*** | ***Version*** | 1.0 | ***Date*** | 28.3.2023 |
| **Passenger** | Passenger of a flight | type 1 | ***Hierarchy*** | Passenger | | |
| ***Attribute*** | ***Description*** | ***Level*** | ***Key*** | ***Type*** | ***Size*** | ***Precision*** |
| Passenger\_id | Passenger surrogate | Passenger | PK | ID |  |  |
| First\_name | First name of passenger | Passenger |  | Varchar | 50 |  |
| Last\_name | Last name of passenger | Passenger |  | Varchar | 50 |  |
| Age | Age of passenger | Passenger |  | Integer |  |  |
| Gender | Gender of passenger | Passenger |  | Char | 1 |  |
|  |  |  |  |  |  |  |

*Table 2 – Airline*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Name*** | ***Description*** | ***SCD*** | ***Version*** | 1.0 | ***Date*** | 28.3.2023 |
| **Airline** | Airline of a flight | type 1 | ***Hierarchy*** | Airline | | |
| ***Attribute*** | ***Description*** | ***Level*** | ***Key*** | ***Type*** | ***Size*** | ***Precision*** |
| Airline\_id | Airline surrogate | Airline | PK | ID |  |  |
| Name | Airline name | Airline |  | Varchar | 50 |  |
| IATA | IATA code | Airline | UK | Varchar | 10 |  |
| ICAO | ICAO code | Airline | UK | Varchar | 10 |  |
| status | Airline activity status | Airline |  | Varchar | 1 |  |
| Callsign | Airline callsign | Airline |  | Varchar | 50 |  |

*Table 3 – Aircraft*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Name*** | ***Description*** | ***SCD*** | ***Version*** | 1.0 | ***Date*** | 28.3.2023 |
| **Aircraft** | Aircraft of a flight | type 1 | ***Hierarchy*** | Aircraft < Airline | | |
| ***Attribute*** | ***Description*** | ***Level*** | ***Key*** | ***Type*** | ***Size*** | ***Precision*** |
| Aircraft\_id | Airline surrogate | Aircraft | PK | ID |  |  |
| Name | Name of an aircarft | Aircraft |  | Varchar | 100 |  |
| IATA | IATA code | Airline | UK | Varchar | 10 |  |
| ICAO | ICAO code | Airline | UK | Varchar | 10 |  |

*Table 4 – Airport*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Name*** | ***Description*** | ***SCD*** | ***Version*** | 1.0 | ***Date*** | 28.3.2023 |
| **Airport** | Airport | type 1 | ***Hierarchy*** | Airport < City < Country | | |
| ***Attribute*** | ***Description*** | ***Level*** | ***Key*** | ***Type*** | ***Size*** | ***Precision*** |
| Airport\_id | Airport surrogate | Airport | PK | ID |  |  |
| Name | Name of Airport | Airport |  | Varchar | 50 |  |
| IATA | IATA code | Airport | UK | Varchar | 10 |  |
| ICAO | ICAO code | Airport | UK | Varchar | 10 |  |
| longtitude\_direction | Longitude direction of airport | Airport |  | Varchar | 1 |  |
| longtitude\_decimal\_degrees | Longtitude of airport | Airport |  | Double |  |  |
| latitude\_direction | Latitude direction of airport | Airport |  | Varchar | 1 |  |
| latitude\_decimal\_degrees | Latitude of airport | Airport |  | Double |  |  |
| Altitude | Altitude of airport | Airport |  | Double |  |  |
| City\_id | City surrogate | City |  | ID |  |  |
| city\_name | Name of city | City |  | Varchar | 100 |  |
| Country\_id | Country surrogate | Country |  | ID |  |  |
| Country\_name | Name of country | Country |  | Varchar | 50 |  |
| iso\_code | Iso code | Country |  | Varchar | 5 |  |
| dafif\_code | Dafif code | Country |  | Varchar | 5 |  |

*Table 5 – Date*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Name*** | ***Description*** | ***SCD*** | ***Version*** | 1.0 | ***Date*** | 28.3.2023 |
| **Date** | Date | type 1 | ***Hierarchy*** | Date | | |
| ***Attribute*** | ***Description*** | ***Level*** | ***Key*** | ***Type*** | ***Size*** | ***Precision*** |
| Date\_id | Date surrogate | Date | PK | ID |  |  |
| Year | Year | Date |  | Integer |  |  |
| Month | Month | Date |  | Integer |  |  |
| Day | Day | Date |  | Integer |  |  |

*Table 6 – Time*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Name*** | ***Description*** | ***SCD*** | ***Version*** | 1.0 | ***Date*** | 28.3.2023 |
| **Time** | Time | type 1 | ***Hierarchy*** | Time | | |
| ***Attribute*** | ***Description*** | ***Level*** | ***Key*** | ***Type*** | ***Size*** | ***Precision*** |
| Time\_id | Time surrogate | Time | PK | ID |  |  |
| Hour | Hour | Time |  | Integer |  |  |
| Minute | Minute | Time |  | Integer |  |  |
| Second | Second | Time |  | Integer |  |  |

*Table 7 – Schedule*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Name*** | ***Description*** | ***SCD*** | | ***Version*** | ***1.0*** | ***Date*** | ***28.3.2023*** |
| **Schedule** | Schedule in a flight | type 1 | ***Hierarchy*** | | Schedule | | |
| ***Attribute*** | ***Description*** | ***Level*** | ***Key*** | | ***Type*** | ***Size*** | ***Precision*** |
| Schedule\_id | Schedule surrogate | Schedule | PK | | ID |  |  |
| Departure\_Date | Sheduled departure date | Schedule |  | | Date |  |  |
| Departure\_Time | Sheduled departure time | Schedule |  | | Time |  |  |
| Actual\_Departure\_Date | Actual date of departure | Schedule |  | | Date |  |  |
| Actual\_Departure\_Time | Actual time of departure | Schedule |  | | Time |  |  |
| Arrival\_Date | Scheduled arrival date | Schedule |  | | Date |  |  |
| Arrival\_Time | Scheduled arrival time | Schedule |  | | Time |  |  |
| Actual\_Arrival\_Date | Actual date of arrival | Schedule |  | | Date |  |  |
| Actual\_Arrival\_Time | Actual time of arrival | Schedule |  | | Time |  |  |

*Table 8 – StatisticsPeriod*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Name*** | ***Description*** | ***SCD*** | ***Version*** | 1.0 | ***Date*** | 28.3.2023 |
| **StatisticsPeriod** | Time period of a statistic | type 1 | ***Hierarchy*** | StatisticsPeriod | | |
| ***Attribute*** | ***Description*** | ***Level*** | ***Key*** | ***Type*** | ***Size*** | ***Precision*** |
| StatisticsPeriod\_id | StatisticsPeriod surrogate | Statistics-Period | PK | ID |  |  |
| Year |  | Statistics-Period |  | Integer |  |  |
| Month |  | Statistics-Period |  | Integer |  |  |

*Table 9 – Flight\_Class*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Name*** | ***Description*** | ***SCD*** | ***Version*** | 1.0 | ***Date*** | 28.3.2023 |
| **Flight\_Class** | Flight class of a flight | type 1 | ***Hierarchy*** | Flight\_class | | |
| ***Attribute*** | ***Description*** | ***Level*** | ***Key*** | ***Type*** | ***Size*** | ***Precision*** |
| Class\_id | Class surrogate | Flight class | PK | ID |  |  |
| Desription | Descrption of Class | Flight class |  | Varchar | 20 |  |

## Facts dictionary

The below tables 1-3 describe the facts’ dictionaries of the travel data mart. The three tables, flight, booking and flight statistic, have information on the dimensions of each star as well as their measures. Here the granularity of each star is also defined. For the flight statistic star, the granularity is statistic of a flight per airport and airline. For the booking star, the granularity is the process of one booking and for the flight star the granularity is on a flight level.

*Table 10 – Flight*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Star** | **Flight** | **Version** | 1.0 | **Date** | 28.3.2023 |
| **Granularity** | Flight | | | | |
| **Dimensions** | | | | | |
| Aircraft\_id | Aircraft | | | | |
| Airline\_id | Airline | | | | |
| Origin\_Airport\_id | Airport (Origin) | | | | |
| Arrival\_Airport\_id | Airport (Destination) | | | | |
| Schedule\_id | Schedule | | | | |
| **Measures** | | | | | |
| Miles | Total miles of a flight | | | | |
| Number\_Passenger | Total number of passengers of a flight | | | | |

*Table 11 – Booking*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Star** | **Booking** | **Version** | 1.0 | **Date** | 28.3.2023 |
| **Granularity** | Process of booking | | | | |
| **Dimensions** | | | | | |
| Time\_id | Time | | | | |
| Date\_id | Date | | | | |
| Class\_id | Flight Class | | | | |
| Passenger\_id | Passenger | | | | |
| Origin\_Airport\_id | Airport (Origin) | | | | |
| Arrival\_Airport\_id | Airport (Destination) | | | | |
| Airline\_id | Airline | | | | |
| **Measures** | | | | | |
| Seat | Flight seat of a booking | | | | |
| N\_Stops | Total number of stops in a booking | | | | |
| Duration\_stops | Total duration of stops in a booking | | | | |

*Table 12 – Flight\_Statistic*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Star** | **Flight Statistic** | **Version** | 1.0 | **Date** | 28.3.2023 |
| **Granularity** | Statistic of flight per airport and airline | | | | |
| **Dimensions** | | | | | |
| Airline\_id | Airline | | | | |
| Airport\_id | Airport | | | | |
| StatisticsPeriod\_id | StatisticsPeriod | | | | |
| **Measures** | | | | | |
| nr\_delays | Number of delays | | | | |
| nr\_arrival\_delays | Number of delays on arrival | | | | |
| nr\_departure\_delays | Number of delays on departure | | | | |
| nr\_on\_time\_departure | Number of times departure was on time | | | | |
| nr\_on\_time\_arrivals | Number of times arrival was on time | | | | |

## Scope

Having a well-defined scope is crucial for the success of any data warehouse project. It defines the boundaries of the project and sets expectations for what will be included and excluded from the data Warehouse.

The following table has the purpose, exclusion, success criteria, risk and mitigation of the Flight Reservation System.

*Table 13 – Scope*

|  |
| --- |
| DW Project – Scope document |
| Context |
| The goals are:   1. Data Integration: The data warehouse should integrate data from various sources, such as flight bookings, customer information, flight details, and other relevant data, to provide a unified and consistent view of the data for analysis and reporting purposes. 2. Data Storage and Management: The data warehouse should efficiently store and manage large volumes of data related to flight reservations in a structured and organized manner. This may involve data partitioning, indexing, and optimizing data storage for fast and efficient data retrieval. 3. Data Analysis and Reporting: The data warehouse should support advanced analytics and reporting capabilities to enable data-driven decision-making. Building the data warehouse based on a structured dimensional model allows complex queries to take place. 4. Security and Data Privacy: The data warehouse should implement robust security measures to protect the sensitive data related to the customer bookings and flights. This may involve implementing authentication, authorization, and encryption mechanisms to ensure data privacy and security. |
| Scope of phase 1 |
| 1. This phase is completed by 10.04.2023 2. Duration: 3 Weeks 3. Date Spam: Two years (2017 – 2019) 4. Develop:   A. The number of flights to and from each airport.  B. Records of delays.  F. Number of bookings made for an airline. |
| Exclusions from phase 1 |
| Following requirements have been excluded from the phase:   1. Functionalities such as customer service or airport security management have not been included in the phase due to the lack of data regarding these operational services. 2. The different data sources may limit the amount and type of data they provide. For example, the different airlines don’t provide customer sensitive data due to privacy regulations. 3. The budget and resources limit the scope of this project phase to only include certain airlines and geographies. In future stages, expanding to more airlines and airports could be possible. 4. Requirements C, D, E, G and H will not be implemented in this phase. |
| Success criteria |
| 1. Successful integration of the data 2. Dimensional data model built 3. The dimensional data model is successfully implemented in MySQL 4. Being able to implement requirements A, B and F |
| Risks and mitigation plans |
| The risks in phase 1 are:   1. Data quality risks: Poor data quality, such as data inconsistencies, inaccuracies, and duplications, can impact the integrity and reliability of the data in the data warehouse. This can result in incorrect or unreliable analytics, reporting, and decision-making. 2. Security Risks: Dealing with sensitive flight data, data security risks must be taken into account. Data security risks may involve unauthorized access to sensitive data, data breaches, or data privacy violations. Ensuring proper security measures such as access controls, data encryption, and compliance with data protection regulations is essential to mitigate such risks. 3. Data Integration Risks: In the project, the data is extracted and integrated from multiple sources, which can be challenging. Data integration risks may result in delays, errors, or incomplete data in the data warehouse. |

* 1. Cost & Benefits

**Benefits:**

* Data centralization.
* Data integration and analysis.
* Performance and scalability.
* Decreased response time.
* Reduced storage costs as a result of improved forecasting.
* Process dynamics.

**Costs:**

* Implementation costs.
* Data Integration cost.
* Maintenance and operational costs.
* Consistent support.
* Unexpected expansion.

# 

# Dimensional data model

The below figure represents the dimensional data model of the travel data mart. It is built based on the dimensional bus matrix presented in the previous chapter. Below are some explanations regarding the model.

**Facts:**

* + Flight

This fact table has the granularity of each flight were the primary key (PK) is the Flight\_id which is an auto incremented integer, and not one foreign key (FK) that is Aircraft\_id and refer to the aircraft operating the flight, Airline\_id refers to which airline, Origin\_Airport\_id and Arrival\_Airport\_id refers from where the flight began and to which destination it went to, Schedule\_id refers to the schedule of the flight.

* + Booking

This fact table has the granularity of each booking made with seven foreign keys that are airline\_id which refers to the airline in the booking, Passenger\_id refers to the passenger the booking was for, Origin\_Airport\_id and Arrival\_Airport\_id refers from where to where the booking is made, Class\_id refers to the class the passenger choose, Time\_id refers to which hour the booking was made in, and Date\_id refers to which year, month and day the booking was made in, this table has measures Seat which is the number of the seat, Nr\_Stops refers to the number of stops in the booking, Duration\_stops is the amount of time that the passenger will do during the stops.

* + Flight\_Statistic

This fact table has the granularity of each airline and airport as well as the year and month with FK Airline\_id referring to the airline, Airport\_id refers to the airport and StatisticsPeriod\_id refers to the year and month. This table have the measures nr\_delays which is the total number of delays, nr\_arrival\_delays is the total number of arrival delays, nr\_departure\_delays are the number of departure delays, nr\_on\_time\_arrivals is the total number of arrivals without delays, nr\_on\_time\_departure is the total number of departures without delays.

**Dimensions:**

* + Aircraft

This represents each aircraft or airplane with PK is a number referring to the id of the airplane, aircraft\_name is the name of the airplane, International Air Transport Association (IATA) and International Civil Aviation Organization (ICAO) are codes used for each airplane.

* + Passenger

This table represents each passenger with Passenger\_id as PK, the table also has the first name, last name, age and gender of the passenger.

* + Flight\_Class

This table represents each class with Class\_id as PK and description which has the name and more details about the class.

* + Date

This table represents of each date having Date\_id as PK, the table contains year, month, and day of each date.

* + Airline

This table represents each airline company with Airline\_id as PK, this table also contains the name, IATA, ICAO, status of the company and Callsign which is a unique identifier used by airlines to identify themselves when communicationg with air traffic control and other aviation authorities. The callsign is usually a combination of letters and numbers and its different from airlines name or IATA code.

* + Airport

This table represents airport with Airport\_id as PK, this table also has the name, IATA, ICAO, the direction of the longitude, the longitude in decimal degrees, the direction of the latitude, the latitude in decimal degrees and altitude of the airport, we also have as a Level Key (LK) City\_id which is and id of the city that the airport is in and it could be another dimension but in our case it fits better to ass it to the airport table and to get more information about the city we also have the city name, same goes for the country since we have a hierarchy of having aiport in a city and city in a country, which leads to have LK Country\_id, country name and as user key (UK) iso\_code and dafif\_code.

* + Time

This table represents each time in each hour having Time\_id as PK and from this table we can get the hour, minute and second of the time.

* + Schedule

This represents each schedule for the flight with Schedule\_id as PK, we also have the departure\_date which is the planned date of departure, departure\_time which is the planned time of departure, actual\_departure\_date which is the actual departure date, actual\_departure\_time which is the actual departure time, arrival\_date which is the planned date of arrival, arrival\_time which is the planned time of arrival, actual\_arrival\_date which is the actual arrival date, actual\_arrival\_time which is the actual arrival time.

* + StatisticsPeriod

This table represents each period which is a combination of year and month to get a statistic about this period, having StatisticsPeriod\_id as PK, and this table also have year and month.

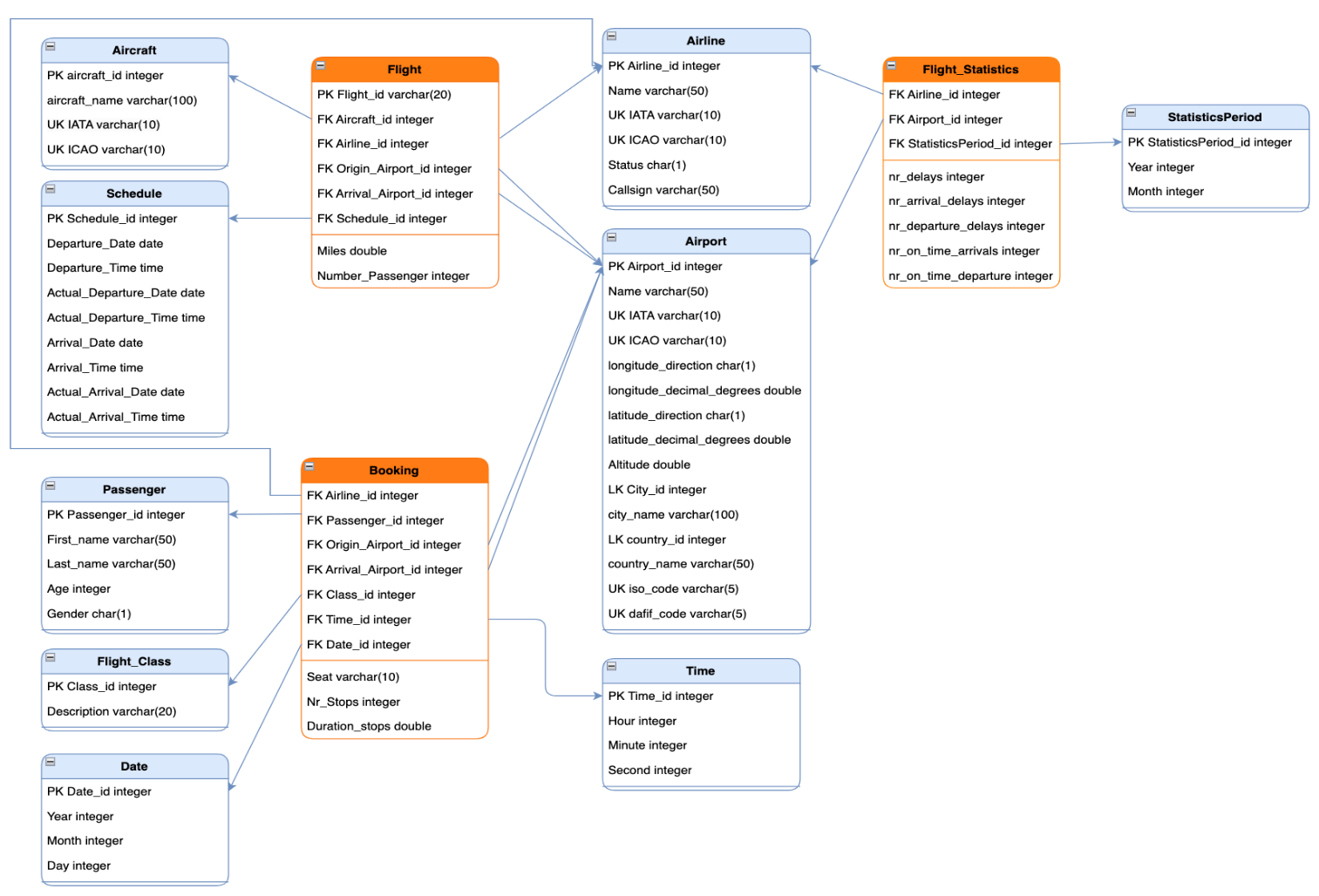
**Measures:**

1. Additive measures

In Flight fact table Miles is an additive measure. This means that when the miles flown by multiple flights are added together, the total miles flown is meaningful and useful. For example, to get the total miles flown of an airline by all its flights in a month, it can simply add up the miles flown by each individual flight to get the total.

1. Semi Additive measures

In Booking fact table Nr\_Stops is a semi additive measure that could be aggregated in many dimensions but not in the Time dimension for example because it does not make sense to have the aggregation of number of stops in a specific hour in our case. Also, the number of passengers in Flight fact table is a semi-additive measure. This means that the total number of passengers across different dimensions may not be meaningful or useful in certain contexts. For example, to get the total number of passengers flown on all its flights in a month of an airline, it cannot simply add up the number of passengers on each individual flight because some passengers may have flown on multiple flights. However, if the goal is to know the total number of passengers for a specific flight or route, then adding up the number of passengers across those dimensions would be meaningful and useful.



*Figure 4 – Dimensional Data Model*

# Data sources selection: Extraction, transformation and loading

The process of Extract, Transform, Load (ETL) is a critical step in the data warehousing process, and it involves extracting data from various sources, transforming it into a unified format, and loading it into a database for analysis. In this case, the data was extracted from two different sources: Kaggle and Openflight. The data was altogether 5 sets of Excel workbooks containing different information regarding the flight travel system.

After extracting the data, a transformation process was performed in the raw data, deleting unwanted rows that were not relevant to the analysis and columns. This step is crucial to ensure that the data is accurate and of high quality, which is essential for making informed decisions based on the data. The transformation process also included transforming the data into unified and easily readable format. For example, the date column was transformed into a standard and unified format to not face any future problems when analyzing it.

Also, the data was loaded into Python to have a better understanding on the data statistics. Below there is an example of loading the global airport dataset.

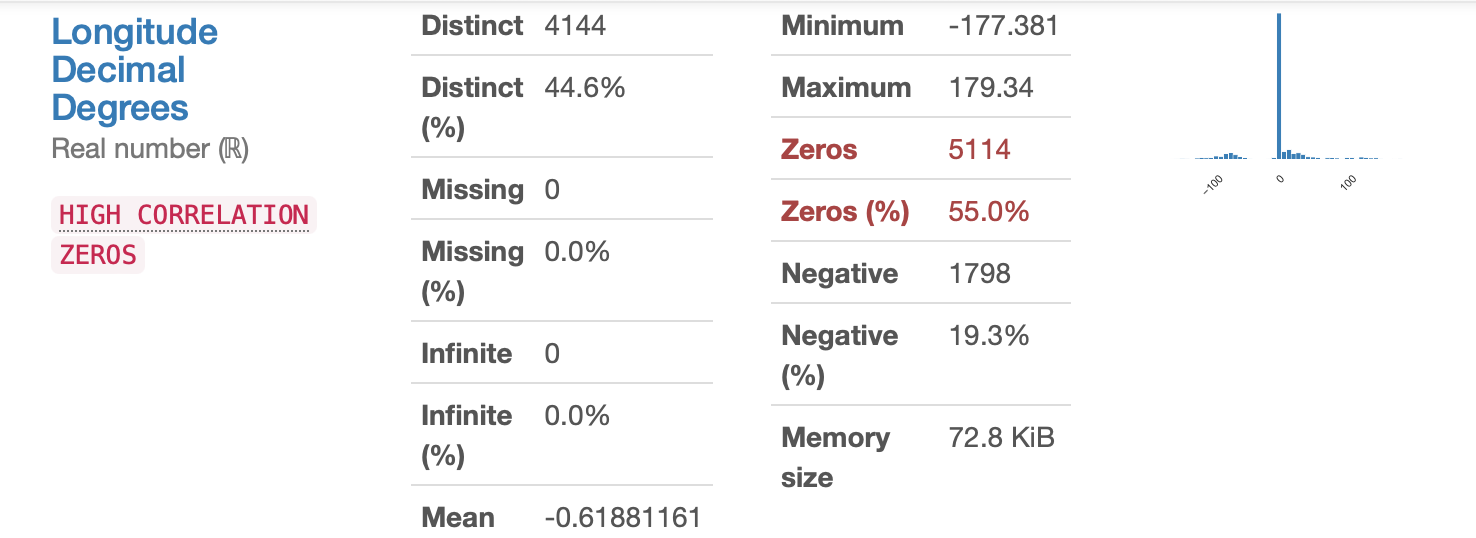
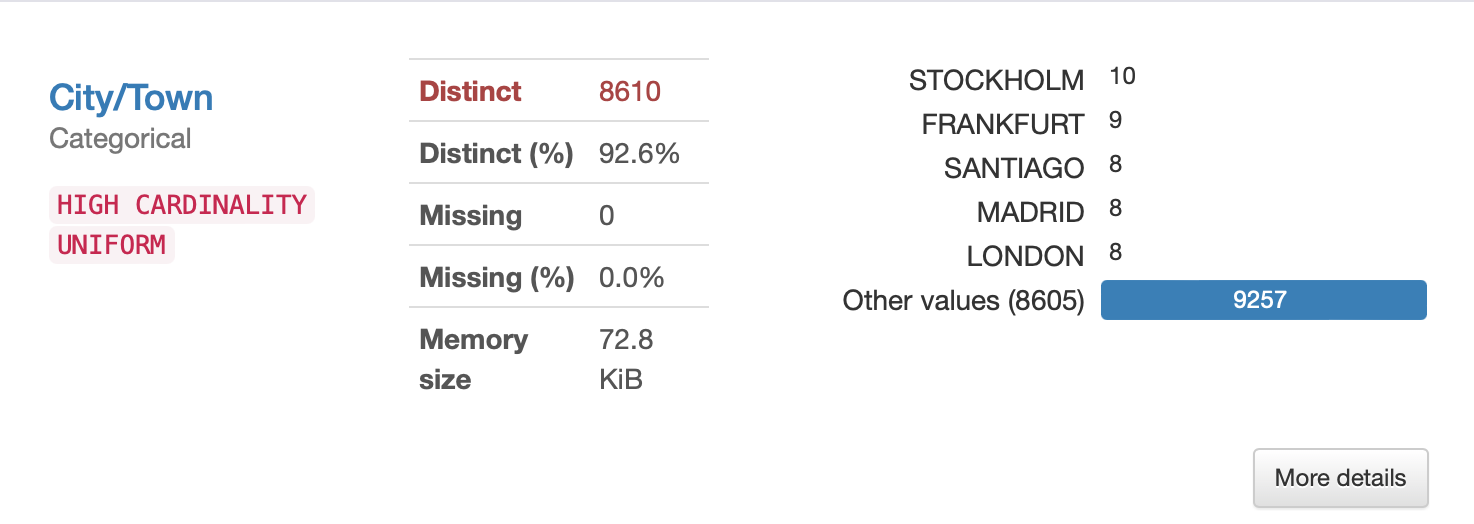
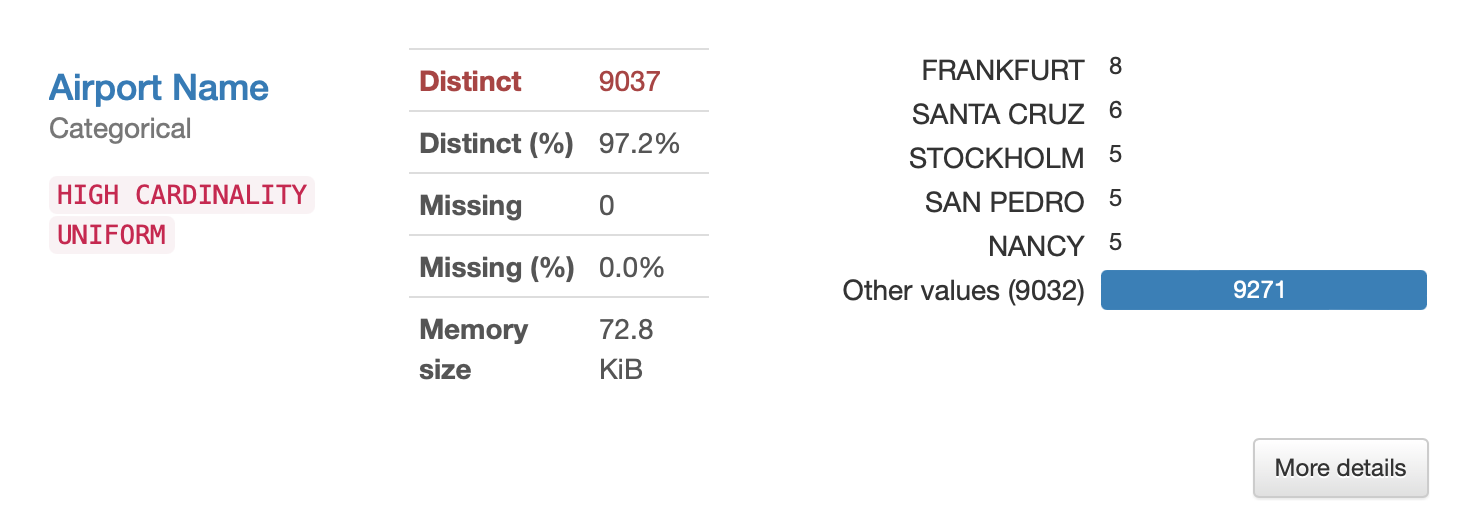
Uma imagem com mesa

Descrição gerada automaticamente

*Figure 5 – Loading Data*

For all the dataset the statistics description was calculated. Below is an example of airport attributes.

Uma imagem com texto

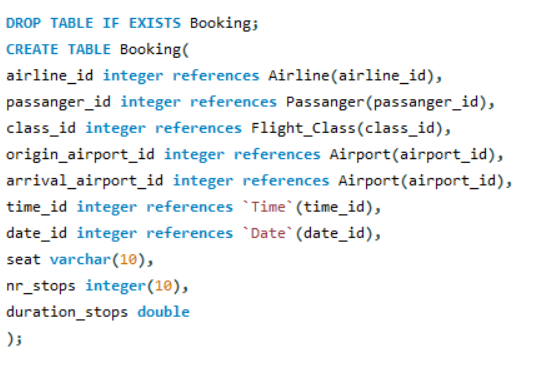
Descrição gerada automaticamente

*Figure 6 – Descriptive Statistics*

After the data was transformed into a unified form, the next step was to create tables according to the dimension model. It is a crucial step in the ETL process, as it helps to ensure that the data is organized in a way that makes it easy to analyze and understand. The table creation was implemented in MySQL, which after the data was loaded into the tables. Later, the dimension tables were joined to the fact tables in MySQL using the join clause.

# Querying

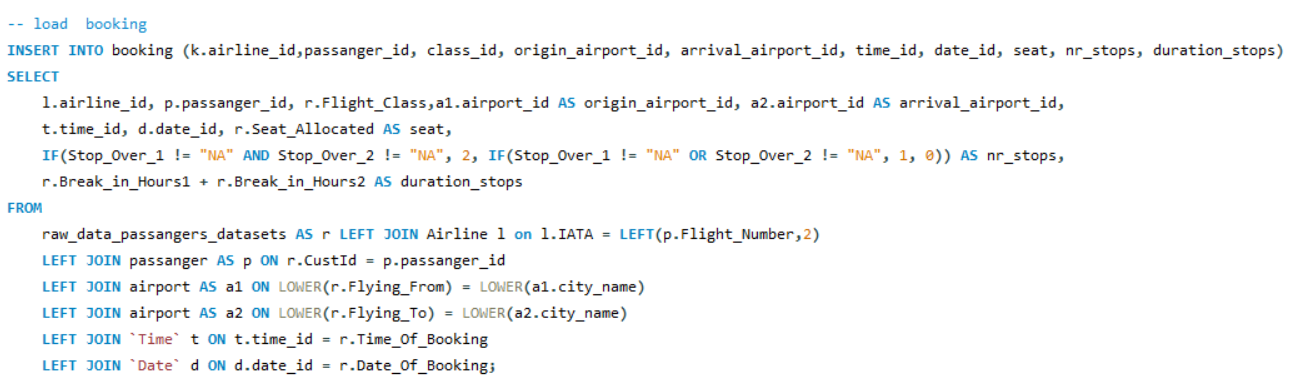
The source files were imported to MySQL Workbench with the importing wizard. Next all the tables from the dimensional model were created and populated. The figures below are an example of the queries performed in MySQL.



*Figure 7 – Create Booking*

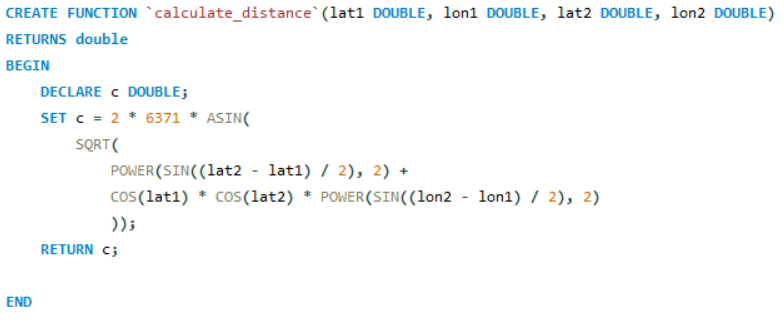


*Figure 8 – Load Airport*



*Figure 9 – Load Booking*

In order to calculate the miles on the fact table ‘Flight’ it was necessary to create a fuction to performe this task. It calculates the distance in miles given the longitude and latitude of two between two airports.



*Figure 10 – Distance Function*

# Data Analysis

Data analysis is important in many different industries and applications, including in flight reservation systems. It allows to do:

* Better decision-making - Data analysis helps decision-makers to make more informed decisions based on the insights derived from the data. In flight reservation systems, it can help airline companies to better understand customer behavior and preferences, which can inform decisions about pricing, scheduling, and other aspects of the business;
* Improved efficiency: Data analysis can help to identify inefficiencies and areas for improvement, which can lead to cost savings and increased productivity. It can help to optimize flight schedules and routes, reducing the amount of time and resources needed to operate the airline.
* Enhanced customer experience: It can help to improve the customer experience by providing personalized recommendations and insights based on their preferences and past behavior. It can help airlines to offer tailored flight options and services to their customers, improving customer satisfaction and loyalty.
* Competitive advantage: It can provide insights that give companies a competitive edge in the marketplace. In the airline industry, where competition is fierce, data analysis can help airlines to differentiate themselves by offering unique and personalized services to their customers.

Biased on this, a Power BI report was developed presenting important KPIs for the Flight Reservation System including the analysis for the flight biased on the Schedule.

In the figure below it is possible to see some KPI and graphics to support the the data analysis:



*Figure 11 - PBI Report*

The following image is the Dimensional Model of PBI. In this it’s possible to see all the relation between the tables.

Uma imagem com diagrama

Descrição gerada automaticamente

*Figure 12 - PBI Dimensional Model*

# Reflection

Operational databases are the foundation of any modern organization's information management system. They serve as repositories of real-time data that support day-to-day operations and facilitate decision-making. However, like any technology, operational databases have both advantages and shortcomings that organizations must carefully consider.

Same advantages of the Flight Reservation System operational database are:

* Increased efficiency: Operational database of flight reservation system can greatly increase the efficiency of the booking process by allowing agents to quickly access and modify customer information, flight schedules, and seat availability in real-time. This can help to reduce wait times for customers and improve the overall booking experience.
* Improved data management: Operational database provide a centralized location for storing and managing customer and flight data, which can help to reduce data redundancy and inconsistencies. This can help to ensure that all data is up-to-date and accurate, which is critical in the airline industry where safety is a top priority.
* Real-time updates: Operational database can provide real-time updates on flight availability and scheduling changes, allowing airlines to quickly respond to changing conditions and minimize disruptions to travel plans.
* Better decision-making: It can provide airlines with insights into customer behavior and preferences, allowing them to make better decisions about scheduling, and other aspects of the business.

Shortcomings:

* Data security risks: Operational database can be vulnerable to data breaches, which can result in the loss or theft of sensitive customer information. It is important to take appropriate measures to ensure that the operational databases are secure and protected from cyber attacks.
* Reliance on technology: The operational database is reliant on technology, which can sometimes fail or experience glitches. This can result in delays or errors in the booking process, which can negatively impact the customer experience.
* Complexity: It can be complex and require a significant amount of maintenance and management. This can be costly and time-consuming for airlines, particularly smaller ones with limited resources.
* Lack of flexibility: The operational database can sometimes be inflexible, making it difficult to make changes or adapt to new circumstances. This can be particularly challenging in the fast-paced airline industry, where conditions can change rapidly.

In summary, while operational database offers numerous advantages to flight reservation systems, it also comes with some potential shortcomings. To maximize the benefits of operational database and minimize the risks, it is a must to carefully manage and maintain the database, while also ensuring that appropriate security measures are in place to protect customer data.

# Conclusion

After building this project it was possible to conclude that a data warehouse can be a valuable tool for a flight reservation system, providing numerous benefits to airlines and their customers. By consolidating data from various sources and transforming it into a format that is optimized for analysis, a data warehouse can help airlines to gain a better understanding of customer behavior and preferences, optimize flight schedules and routes, and make more informed decisions about pricing and other aspects of the business.

In addition, a data warehouse can help to improve the efficiency and accuracy of the booking process by providing real-time access to relevant information, reducing wait times for customers, and minimizing errors. Furthermore, a data warehouse can enable airlines to provide a more personalized and tailored experience for their customers, enhancing customer satisfaction and loyalty.

However, there are also potential challenges and drawbacks in the implementation of the data warehouse. These includes issues to find data available. For this project, while the airport, airline and aircraft data were easy to find, passenger in other hand wasn’t. The only passenger data found was of India flights. The main reason that it was so hard to find is because they are personal data. It was also found other data, but they can only be accessed by having a submission fee.

The decision to implement a data warehouse in a flight reservation system needs always to be carefully considered, taking into account the specific needs and goals of the an airline and its customers, as well as the costs and benefits of the system. When implemented effectively, a data warehouse can be a powerful tool for airlines looking to gain a competitive edge in the market and provide a better experience for their customers.



# Next Steps

We believe that the dimensional model in order to be more realistic, information about boarding pass, boarding gate, terminal, airline capacity, bags registration, payment, booking site, cabin crew management and service during the flight (meal service, convenience).

Cabin crew management for example, will be another data mart. This data mart is to be populated with data that is relevant to the cabin crew, such as crew schedules, flight manifests, crew performance metrics, and safety and security information.

The purpose of a cabin crew data mart is to provide the cabin crew department with quick and easy access to the information they need to perform their jobs effectively. By consolidating relevant data into a single location, the data mart makes it possible for cabin crew managers and supervisors to quickly identify patterns and trends, monitor crew performance, and make informed decisions about scheduling, training, and resource allocation.

In addition to providing operational support to the cabin crew department, a data mart can also be used to support strategic decision-making by providing insights into passenger behavior, crew preferences, and other factors that can impact the overall performance of the airline. By leveraging the insights gained from a cabin crew data mart, airlines can improve their operations, enhance the customer experience, and gain a competitive advantage in the marketplace.

# Referencies

1. T. K. Das and A. Mohapatro, "A Study on Big Data Integration with Data Warehouse," International Journal of Computer Trends and Technology*,* vol. 9, pp. 188-192, 2014.
2. D. Martins, C. Ramos, J. Rodrigues, P. Cardoso, R. Lam and F. Serra, "Challenges in Building a Big Data Warehouse Applied to the Hotel Business Intelligence," in 6th Int. Conf. on Applied Informatics and Computing Theory (AICT’15), in Recent Research in Applied Informatics, Salerno, Italy, 2015.
3. C. Huda, R. P. Sari, M. H. Haekal and T. Agustaria, "The Development of the Application of Data Warehouse at PT JKL," Communication and Information Technology (COMMIT)*,* vol. 6, no. 1, pp. 8-19, 2012.
4. R. Kimball, M. Ross, W. Thornthwaite, J. Mundy and B. Becker, The Data Warehouse Lifecycle Toolkit, New Jersey: Wiley, 2007
5. <https://www.flightdatasystems.com/?utm_term=flightdataservices&utm_campaign=Competitors&utm_source=adwords&utm_medium=ppc&hsa_acc=7438970472&hsa_cam=12608127976&hsa_grp=118668857654&hsa_ad=508970230770&hsa_src=g&hsa_tgt=kwd-1947127078920&hsa_kw=flightdataservices&hsa_mt=b&hsa_net=adwords&hsa_ver=3&gclid=CjwKCAjwiOCgBhAgEiwAjv5whKXsP_C_bTDAfv3LUx8wnVmux9kTEiam5_UNzsplsFU2sa6IXEVVbRoC24MQAvD_BwE>
6. <https://www.anac.gov.br/acesso-a-informacao/dados-abertos/areas-de-atuacao/seguranca-operacional>
7. <https://www.kaggle.com/datasets/ianalyticsgeek/flight-passangers-data>
8. <https://www.eurocontrol.int/our-data>
9. <https://www.flightradar24.com/54.12,-4.44/8>
10. <https://www.faa.gov/air_traffic>