

Flight Punctuality data warehousing and analysis



Table of Contents

[Benefits of building a data warehouse for the Airport Punctuality data: 2](#_Toc134567534)

[Data Warehousing using Star Schema and supportive design decisions: 3](#_Toc134567535)

[CREATE Table statements for the above schema design: 4](#_Toc134567536)

[Creating and populating the database: 5](#_Toc134567537)

[Benefits from using OLAP cubes in Airline Punctuality database: 7](#_Toc134567538)

[Benefits of using a data warehouse in combination with Tableau: 7](#_Toc134567539)

[Visualizations: 7](#_Toc134567540)

[Visualization 1: 7](#_Toc134567541)

[Visualization 2: 8](#_Toc134567542)

[Visualization 3: 8](#_Toc134567543)

[Visualization 4: 8](#_Toc134567544)

[Conclusion: 9](#_Toc134567545)

# Benefits of building a data warehouse for the Airport Punctuality data:

Building a centralized repository for any database not only airport data helps us with data management and visualization in general. This report will go over into details of benefits of having a warehouse for the airport punctuality data.

1. Centralizing the data from different data sources: Airport data is one such data which comes from different sources such as official airport websites, aviation authorities and also some airlines which make their data available for promotion of their on-time performance. This data may have different kinds of measures, data types and sometimes different formats which makes it difficult to compare it all together before enhancing it. This process of enhancing can also be called as warehousing, warehousing is a process which involves gathering the information or data, extracting it into spreadsheets, transforming the data which could fit into a structure (modeling), querying this data for analysis and finally making the repository able for future maintenance.
2. By doing all of the above mentioned steps to our data, we are improving the quality of overall database hence making it more accurate and consistent. For example there is a single information source that is given on time information for this data, which is the column called reporting period. This column has year and month information of data all stored in a single continuous number ‘202101’. Here the first 4 digits represent year and last year the month ‘01’. By the base of this information, we have defined the months involved in the data in a different table and linked it to the fact table to make meaning full analysis on the data which is further discussed in the visualization part of the report.
3. Warehousing’s major benefit is business intelligence and analytics. By providing a single source of data, we enable analysts and companies to draw meaningful insights into their business and make efficient choices which supports the overall growth of the business. For example in this report we look at different visualizations which help the airlines understand which is doing better in terms of booking and what might be the possible effecting attributes (average delay, cancellations or locations) for the same.
4. While performing the process of warehousing we have to keep in mind the scalability and growth of business. For example, in this data warehouse, we have created a separate table for time all together in the main warehouse design though it is only a single column field in the main source. The table only has 3 columns and 12 rows in the database as of now, but as the airlines or aviation companies keep a track this data, it is ever increasing especially overtime. Hence, this design supports a lot in terms of future entries in form of dates which will help the users manage the data more efficiently.

# Data Warehousing using Star Schema and supportive design decisions:

A diagram of flight performance

Description automatically generated with low confidence

Fig 1

1. The above design for the data warehouse signifies star schema, star schema has a fact table and dimension tables all linked to one fact table. We do this to store data separately in different tables. Here the raw database has month wise division of all the columns above mentioned in one single file. By dividing them into tables like this we have kept all the flight performance defining metrics in flight performance table and this is the fact table as it has all the primary information needed for the analysis and this also has all foreign keys from dimension table primary keys. We have combined all months and put it separately into Time\_details table and all airlines\_info into airline\_info table and all airports data into airports table. By doing the we have reduced on a lot of data duplication as the things like date airports and airlines used by passengers may get repeated a lot of times resulting in repetition of this data and names a lot of times resulting in inaccuracy, also if we have to change a particular detail like the name of airline we just need to change in one column in the dimension table which automatically changes the data in fact table hence supporting data consistency and optimization.
2. These design decisions also improves the ease of maintenance of the database. Star schema has comparatively few tables and fewer relationships hence making it easy to make changes and optimize the performance.
3. This design also reduces the use of joins as they are all already connected to one single table which has the most attributes necessary for visualization.

# CREATE Table statements for the above schema design:

1. Fact table:

CREATE TABLE "FLIGHTPERFORMANCE"

( "PERFORMANCEID" NUMBER,

"AIRLINEID" NUMBER,

"AIRPORTID" NUMBER,

"TIMEID" NUMBER,

"REPORTINGPERIOD" NUMBER,

"SCHEDULEDFLIGHTS" NUMBER,

"UNSCHEDULEDFLIGHTS" NUMBER,

"CANCELLEDFLIGHTS" NUMBER,

"EARLYFLIGHTSPRCT" NUMBER,

"ONTIMEFLIGHTSPRCT" NUMBER,

"SLIGHTDELAYFLIGHTSPRCT" NUMBER,

"MODERATEDELAYFLIGHTSPRCT" NUMBER,

"SIGNIFICANTDELAYFLIGHTSPRCT" NUMBER,

"VERYDELAYFLIGHTSPRCT" NUMBER,

"EXESSIVEDELAYFLIGHTSPRCT" NUMBER,

"EXTREAMDELAYFLIGHTSPRCT" NUMBER,

"INFINITEDELAYFLIGHTSPRCT" NUMBER,

"UNSCHEDULEDPERCENTAGE" NUMBER,

"CANCELLEDPERCENTAGE" NUMBER,

"AVGDELAYMINS" NUMBER,

"PREVYYMMSCHEDULEDFLIGHTS" NUMBER,

"PREVYYMMLATEPRCT" NUMBER,

"PREYYMMAVGDELAY" NUMBER,

"SCHEDULEORCHARTER" VARCHAR2(100),

PRIMARY KEY ("PERFORMANCEID")

) ;

ALTER TABLE "FLIGHTPERFORMANCE" ADD CONSTRAINT "FLIGHTPERFORMANCE\_CON" FOREIGN KEY ("AIRLINEID")

REFERENCES "AIRLINEINFO" ("AIRLINEINFOID") ENABLE;

ALTER TABLE "FLIGHTPERFORMANCE" ADD CONSTRAINT "FLIGHTPERFORMANCE\_CON1" FOREIGN KEY ("TIMEID")

REFERENCES "TIMEDETAILS" ("TIMEID") ENABLE;

ALTER TABLE "FLIGHTPERFORMANCE" ADD CONSTRAINT "FLIGHTPERFORMANCE\_CON2" FOREIGN KEY ("AIRPORTID")

REFERENCES "AIRPORTS" ("AIRPORTID") ENABLE;

1. Dimension Table 1:

CREATE TABLE "AIRPORTS"

( "AIRPORTID" NUMBER,

"REPORTINGAIRPORT" VARCHAR2(100),

"ORIGINDESTINATIONCOUNTRY" VARCHAR2(100),

"ORIGINDESTINATIONCITY" VARCHAR2(100),

"ROUTES" VARCHAR2(100),

PRIMARY KEY ("AIRPORTID")

) ;

1. Dimension Table 2:

CREATE TABLE "TIMEDETAILS"

( "TIMEID" NUMBER,

"REPORTINGPERIOD" NUMBER,

"MONTH" VARCHAR2(10)

PRIMARY KEY ("TIMEID"),

) ;

1. Dimension Table 3:

CREATE TABLE "AIRLINEINFO"

( "AIRLINEINFONAME" VARCHAR2(100),

"AIRLINEINFOID" NUMBER,

PRIMARY KEY ("AIRLINEINFOID")

) ;

# Creating and populating the database:

* The database provided was a month wise split (12 files for each month) containing most of the columns mentioned in the fig 1 into a single file. This had everything you need but with a lot of redundancy and inconsistency, by creating the tables mentioned in the code by the basis of design in fig 1, a skeleton of database has been created.
* Once that has been done the column names have been changed to names which are more easy to understand and they are mentioned below:

|  |  |
| --- | --- |
| **Original name of the column** | **Changed name of the column** |
| reporting\_period | reportingperiod |
| reporting\_airport | reportingairport |
| origin\_destination\_country | origindestinationcountry |
| origin\_destination | origindestinationcity |
| airline\_name | airlineinfoname |
| scheduled\_charter | scheduleorcharter |
| number\_flights\_matched | Scheduledflights |
| actual\_flights\_unmatched | unscheduledflights |
| number\_flights\_cancelled | Cancelledflights |
| flights\_more\_than\_15\_minutes\_early\_percent | earlyflightsprct |
| flights\_15\_minutes\_early\_to\_1\_minute\_early\_percent | ontimeflightsprct |
| flights\_0\_to\_15\_minutes\_late\_percent | slightdelayflightsprct |
| flights\_between\_16\_and\_30\_minutes\_late\_percent | moderatedelayflightsprct |
| flights\_between\_31\_and\_60\_minutes\_late\_percent | significantdelayflightsprct |
| flights\_between\_61\_and\_120\_minutes\_late\_percent | verydelayflightsprct |
| flights\_between\_121\_and\_180\_minutes\_late\_percent | exessivedelayflightsprct |
| flights\_between\_181\_and\_360\_minutes\_late\_percent | extreamdelayflightsprct |
| flights\_more\_than\_360\_minutes\_late\_percent | infinitedelayflightsprct |
| flights\_unmatched\_percent | unscheduledpercentage |
| flights\_cancelled\_percent | cancelledpercentage |
| average\_delay\_mins | avgdelaymins |
| previous\_year\_month\_flights\_matched | prevyymmscheduledflights |
| previous\_year\_month\_early\_to\_15\_mins\_late\_percent | prevyymmlateprct |
| previous\_year\_month\_average\_delay | preyymmavgdelay |

* By doing this step the data is more understandable now.
* It was observed from these columns that there are three major categorization of information of data. It can be further divided into flight’s performance details, airport details, time details and airline information First the table flight performance is populated with corresponding columns which are :

1. Reportingperiod
2. Scheduledflights
3. Unscheduledflights
4. Cancelledflights
5. Earlyflightsprct
6. Ontimeflightsprct
7. Slightdelayflightsprct
8. Moderatedelayflightsprct
9. Significantdelayflightsprct
10. Verydelayflightsprct
11. Exessivedelayflightsprct
12. Extreamdelayflightsprct
13. Infinitedelayflightsprct
14. Unscheduledpercentage
15. Cancelledpercentage
16. Avgdelaymins
17. Prevyymmscheduledflights
18. Prevyymmlateprct
19. Preyymmavgdelay
20. Scheduleorcharter

* Then an identifying column was created with the name PerformanceID and constrained as primary key. Now, the second tables airports was populated with data from all 12 files in the columns reportingairport, origindestinationcountry, origindestinationcity an identifying primary key column has been created called airportID. Next, the third table time\_details has been populated the same way with column values with reportingperiod and an identifying primary key column has been created called timeID. Similarly, the table airlineinfo has been populated with column airlineinfoname and an identifying primary key column as airlinrinfoID. The primary keys from the tables Airports, time\_details and airlineinfo are made into foreign keys in the fact table flights\_performance to establish a relation.
* In the source files the columns containing airport destinations, dates and airline information was redundant hence needed to be normalized. After populating the reportingairport, origindestinationcountry, origindestinationcity in airports table the values in these cells are concatenated into a single cell and a condition is applied to the column that the identifying value should increment if the cell value is different and if not repeat the value. By this we uniquely identified the repeated values and placed them into the corresponding foreign key value of fact table and deleted these duplicates. Similar logic has been applied to airlineinfoname column and redundant values here are also deleted.
* To make visualizations more understandable, a new column has been created in the time\_info table which is giving the names to the corresponding months, timeID 1- January and so on. Another new column Routes has been created in the aiports table which is a a concatenation of reportingairport and origindestinationcity column values.
* First column, in the raw data is information about run-date of the file hence insignificant to analyse anything. It has been deleted.

# Benefits from using OLAP cubes in Airline Punctuality database:

OLAP basically stands for online analytical processing cubes which are predominantly used for analyzing data multidimensionally. They are usually a replacement for complex querying where once wants to explore data with large attributes and cell values. OLAP cubes are built using warehousing also follow ETL process on the data and load them into the cubes.

The benefits of using these cubes for airline punctuality data is as follows:

1. Delays of flights can be analyzed over the time period of months and years.
2. Comparison of different airlines over the attributes such as schedules, delays and cancellations can be done. This can enable the airline business to make informed business decisions and changes to improve their performance.
3. Also, patterns of delays and cancellations over particular regions or countries can be observed and future delays and cancellations can be predicted.
4. The effect of any natural calamity or a global pandemic like Covid on the airline industry can be recognized by the number of schedules and bookings of these flights.

All these insights can help the airline companies make data driven decisions to promote optimization of their flight functioning. The above mentioned analysis can be made by using excel or any other OLAP reporting tools and also create tables and charts using the same.

# Benefits of using a data warehouse in combination with Tableau:

Data warehousing allows the data to be extracted transformed and loaded in form of a database and make it entirely ready for analysis to draw business related insights on data. Tableau on other hand helps out the latter process mentioned above that is drawing insights by visualizing the data once the data is ready.

By warehousing the data, the quality of the data is improved as we remove duplicates and validate the data and then load it. This makes sure that the data is clean and accurate which in turn helps Tableau in visualizing accurate charts and graphs in turn help making insightful reports to take business decisions on.

When the data is warehoused into proper schemas, there is a symphonic structure to the centralized data. One advantage of this is that it can be used to query the data for needed information as the data is organized into one single source that from multiple sources. Other advantage is that when the data is already into a schema and when it is loaded into tableau in form of tables with defined and valid relationships it makes it easy for tableau to play around and visualize data from any column in the entire database to any other column in the entirety again without having to worry about relationships between them.

One thing to keep in mind when designing a data warehouse is its future scalability. When we divide the tables and put them into a schema we are making it easy for the organization to add more details in future with out compromising on data accuracy. Accurate data always helps in making accurate reports and visualizations on Tableau.

Overall, using data warehousing for ETL, and Tableau to visualize the data helps organizations to effectively manage data and efficiently take data driven decisions by gaining knowledge from the insights.

# Visualizations:

The aim of all visualizations discussed below is to analyze data at an optimized level using as many attributes as possible to help in making optimized decisions for upliftment of airline business in terms of the punctuality.

## Visualization 1:

A screenshot of a computer

Description automatically generated

1. The above visualization is a heat map which shows information on routes named in the cubes. The aim of this visualization is to understand what are the most booked/scheduled routes and what is the average delay in them. By doing this the airlines/ aviation authorities can increase the flight availability for these locations and also by understanding delay in these they can make business decisions to reduce delay increasing more business in these routes as they already have a lot of bookings flowing in.
2. The first step taken in this visualization is that the routes column is selected to display as text in boxes for the heat map and a filter is applied on this field which would compare this column with scheduled flights column, sum the values and display the top 20. Once that is done, under marks, average delay time is added to understand the delay percentage in these routes, the gradient for the map is from orange to blue where darker orange represents less percentage delay and the colour grows lighter, reaches blue and darker blue as this increases. The text under marks card here is defined as routes.
3. The key findings for these visualizations is that, the busiest route of all is Heathrow to New York (JF Kennedy). Also, the origin country of all routes is mostly Heathrow and this is might be the most busiest airport of all in UK as most of the flights are scheduled from here. Though the busiest route is this, the scheduled flights are 5051 and the average delay is 11.59 mins. The next busiest route is Heathrow to Glasgow with almost half the schedules of what the above route had but only a minute difference in delay. Hence, may be the aviation authorities might have to focus on this route to improve delays than the first.

## Visualization 2:

A screenshot of a computer

Description automatically generated

1. The second visualization is a global map shaded in regions the aim is to understand the destination countries scheduled flights, average delay in minutes and cancelled flights. Since the above visualization has more focus on origin countries, this focusses on destination country.
2. The first step taken to create these visualization is that the column origin destination is dragged to the rows and columns and when the map is selected Tableau split the values into longitude and latitude and generated the corresponding values for the countries. The columns that are added in marks are average delay which is a gradient that goes from light to darker shade of blue when the time increases in minutes. The sum of scheduled and cancelled flights are also added in these marks to understand the performance of these airports. The delay time is averaged as it is already an averaged value. The filter for all these three values is set for the most.
3. The key findings from this visualization is that country Zimbabwe has the most delayed flights as its darker in colour, with the value at 51 mins. The scheduled flights are just 2 and no cancellations hence this airports must work on schedules as well as delays to make the average better. Pakistan also has a closer darker colour and when the mouse is hovered on it, it shows that average delay is 39 mins, when scheduled flights are 1642 and cancelled flights are 66. Though Pakistan has lesser delays when compared to Zimbabwe, the performance of the airport is still poor as the cancellation flights are relatively high in number.

## Visualization 3:

A screenshot of a computer

Description automatically generated

1. The third visualization is a stacked bar chart, the aim of the visualization is to compare the overall business of aviation over the months for the year 2021 and previous.
2. In order to achieve this the month column is dragged from the data and taken as a row value. And there are two column values taken here where pink represents the sum of prevyymmscheduledflights column and purple represents scheduledflights column. The axis values instead of displaying the column name are edited to displaying a meaningful content ‘number of scheduled flights’ and the title of the legend is edited in edit option at the legend. The mark here is counting the data.
3. The major findings from the visualization is that the starting months in the year 2021 had very less schedules compared to same month’s previous years data. This could be because of first wave in covid. However, the count gradually increased over the last months. By hovering the mouse over the bars we can look at exact values, as we defines month, prevyymmscheduledflights and scheduledflights columns.

## Visualization 4:

A screenshot of a computer

Description automatically generated

1. The fourth visualization is a highlights table where the aim is to understand the airlines performances in terms of different delay levels, cancelled and scheduled flights.
2. The above aim has been put into implementation by selecting airlineinfoname as a row value and a filter is applied to this which displays the top 5 popular airlines by number of scheduled flights. Here in order to compare the top airlines performances, the columns earlyflightsprct, ontimeflightsprct, slightdelayflightsprct, moderatedelayflightsprct, significantdelayflightsprct, verydelayflightsprct, exessivedelayflightsprct, extreamdelayflightsprct and infinitedelayflightsprct are selected to compare how much percentage of flights were delayed by how many minutes. Also, cancelledflights and scheduledflights columns are selected to understand the performance. All are averaged as most of them are percentages, except for scheduledflights and cancelledflights which are a sum. The colour gradient is given by averagedelay column as the delay increases the colour grows darker.
3. The key findings from the visualization is that the best performing airline according to all comparative values is Easy Jet UK ltd, as it has the least average delay value with maximum number of scheduled flights also excessive, extream and infinite delays are always less than 1 and you can find these values by hovering over the coloured box. The least performing is the British Airways PLC as it has the most average delay value though the number of schedules are not as high as best performing. The delay value of Loganair LTD is closer to British Airways PLC however, it is as bad as the British Airways PLC because the schedule to cancellation ratio is pretty much the same.

# Conclusion:

This report concludes on a lot of information on how to manage and design a data warehouse for airport punctuality data and also visualize it using Tableau to make meaningful insights.

Undertaking this project helped me understand and a get a grip on following things:

1. This project helped me enhance my technical skills as I got to experience a hands on warehousing and analyzing real data and work on tools such as Tableau and SQL which are well sought in market right now.
2. By analyzing the visualization derived from Tableau helped me understand how business decisions are made by data driven approaches and also improved my analytical thinking overall.
3. By framing and creating this project I have gained good communication skills in terms of analytical which are a great add on to any aspiring data analyst. Now I know how to report my insights on a particular analysis of a business meaningfully for which I struggled in past.
4. Hands on experience is something I would want to reiterate as its giving me real work experience on how to work on real time data and this particularly helped me understand what would my job flow be as a data analyst, what come under my responsibility and what don’t in terms of data management and visualizing.