Student ID: K2283306

Name: Shahesta Nazir Ahmed

Assignment 2- Flight punctuality exercise

1. Discuss the benefits of building a data warehouse for the data set provided.

A data warehouse is a central repository of data comprised of multiple disparate data sources (transactional, legacy, or external systems, applications, and others)- which could be structured, semi-structured or unstructured.

While one may often confuse databases with data warehouses, these are fundamentally different despite both being relational data systems.

Databases use OLTP (online transaction processing) to store and access transactions for ongoing business purposes. On the other hand, data warehouses rely on OLAP (online analytical processing) and can store large volumes of historical data.

A data warehouse once implemented in the airline business intelligence framework will fetch multiple benefits to stakeholders namely improved business intelligence, robust decision support, superior business practices and effective analytics processing. We will discuss some of them below.

1. Harnessing Historical Data:

An Airline may want to access historical data to perform customer trend analysis over the years. This may facilitate marketing strategies to outperform KPIs and to gain actionable insights that can manipulate business decisions backed by factual data rather than instincts. This will help the business stay ahead of cut-throat competition.

We can also tap into historical data to increase ROI (Return on Investment) wherein an airline can determine if their services or purchases have derived increasing profits or reduced expenses. This can help executives evaluate initiatives that have worked beneficially and adjust them to decrease costs, maximize efficiency and increase sales to improve their bottom line.

A typical use case would be if the airline wished to identify which customer groups should be targeted in their newly launched advertisements and which channels should be employed for maximum exposure.

Executives may explore operational costs related to flight maintenance, flight re-scheduling's, fuel costs etc. to get a high-level understanding of company budgets. A centrally stored data warehouse can easily tackle such tasks ensuring data integrity, consistency, reliability, and security.

1. Maintaining Data Quality and Consistency:

A data warehouse is programmed to apply a uniform format to all data collected from multiple inconsistent sources of data. This reduces significant effort when trying to manage duplicate information and preventive methods that employ resources to manage them thus avoiding exorbitant costs for the industry. An airline can consolidate all the information they gather including reservations, flight schedules, crew assignments, operational data, customer data and customer service into an only source of data.

As all departments have access to transformed data, their insights are accurate across various segments of the business- sales, marketing, operations as they all have access to unified data.

For an airline industry this is especially relevant as analysts and strategists save valuable time and can easily focus on business objectives when they have properly structured data. This avoids accessing multiple sources of data, cross-checking and managing different forms of data for different purposes.

For example, we could draw valuable insights from the qualitative data stored in a data warehouse if the airline queried the most frequently flying passengers to offer exciting packages to them. It is quite possible that a customer frequently books a combination of airlines periodically or commutes very often along the same route. On the contrary, we could have customer feedback data and customer data to analyse in-flight experience, or we could speculate the performance of an employee based on other factors.

1. Combining Business Intelligence:

Data warehouses are easily compatible with multiple business intelligence tools. This allows executives and key stakeholders to easily derive insights without overly relying on their technical stack. These advanced tools provide numerous graphs and visualizations that can summarize our findings. We may discover interesting patterns and relationships between two components of data which may have been previously neglected. For example, from a data warehouse we could find an interesting pattern wherein 65% of newly hires quit at the onset of the 1st year. This would then require further investigation of what is fuelling such behaviour and why. Is it because of lack of opportunities for progression in career or incompetent team leaders?

Such insights can then aid in employee retention and reduce training costs.

We may discover that mostly for flights from A to B there are multiple overbookings, which results in dissatisfied customers.

A trend showing multiple cancelled flights on some day may result in re-scheduling's. A quick glance at the data may suggest that harsh weather or un-knowing cases of bird attacks was the reason for such an occurrence.

1. Access to Real-Time data and easy upscaling or downscaling:

Most organizations are shifting towards cloud-based warehousing solutions as they offer a more flexible pay-as-you-go model, which allows easy upscaling or downscaling with respect to the flow of data. Market managers and executives can easily manage these without the need for extensive IT support, which saves time.

Hence, data warehouses that can support real-time and dynamic data are preferred. Airlines have varying prices and time schedules for various flights, which emphasizes the need for timely updates on data. Any airline booking app can predict flight rates for up to 6 months considering all factors that may influence flight prices, which allows pre-booking services to customers. Nevertheless, airline rates are highly subjective and may fluctuate. It is important that an organization has appropriate means and measures to store such information and be able to access it when needed.

1. Hot and Cold Data:

A data warehouse architecture is intricate with three layers. The top tier is the front-end client, the middle tier consists of the analytics engine and the bottom tier is the database server where data is stored and loaded.

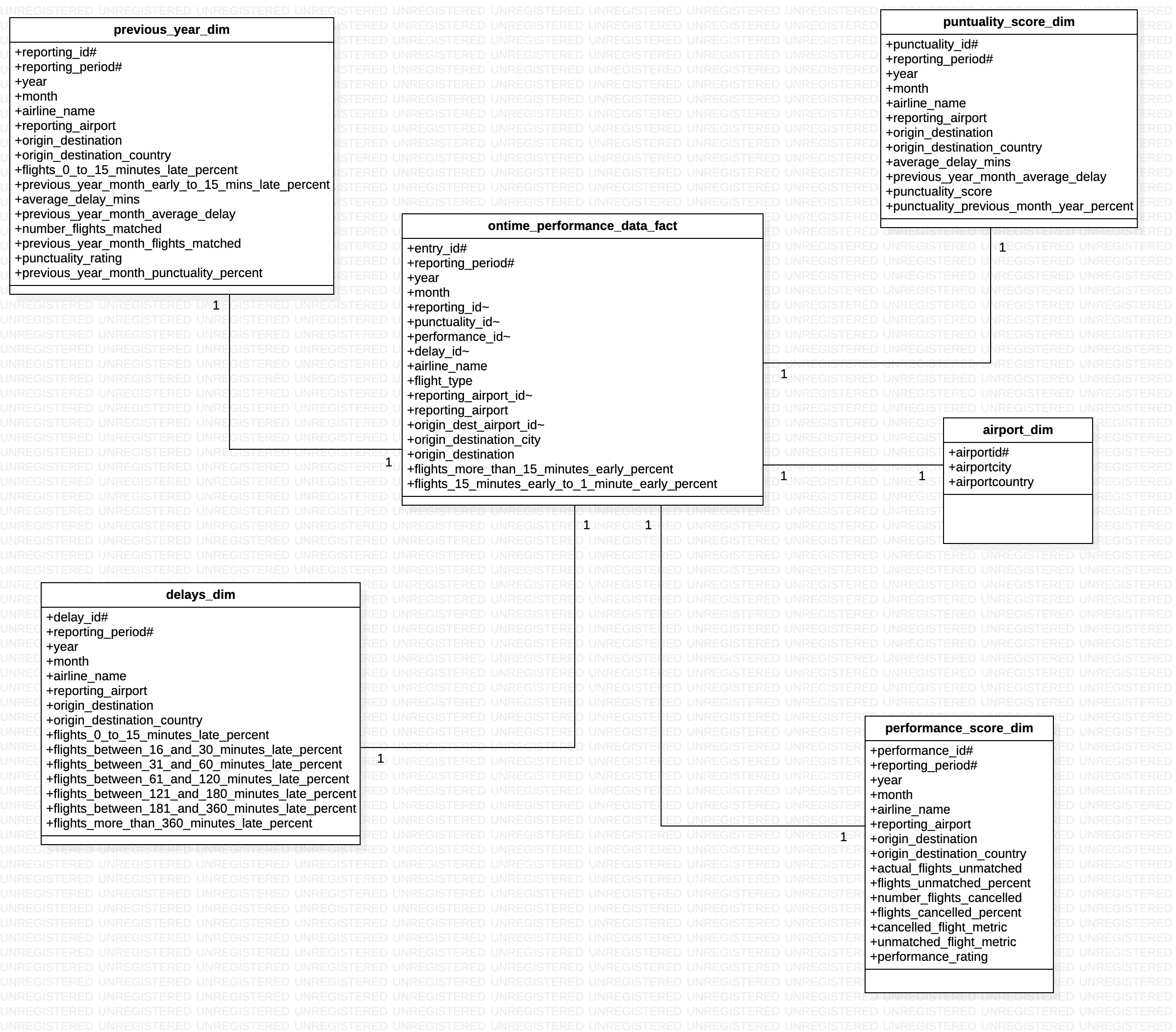
The data is stored in two distinct types of ways 1) data that is frequently accessed (hot data) is stored in fast storage 2) data that is infrequently accessed (cold data) is stored in cheap object store. This ensures that query runtimes are optimized.

Airlines need to comply with various regulations and reporting requirements. All the necessary supporting documentation can always be treated as cold storage and can be accessed when any of these terms need to be updated or when any licence is due to expire.

However, in a scenario where a pilot is suddenly unavailable or any crew is absent, an update to the flight crew should be possible quickly, treating the data as hot.

Processing all data in a similar fashion- such as in transactional databases may be computationally expensive and may incite additional expenses. Hence the use of data warehouses can synchronize industry requirements with business expectations.

2. Design a data warehouse using a star schema. You must justify your design decisions.



Note: #-: Primary Key; ~: Foreign Key

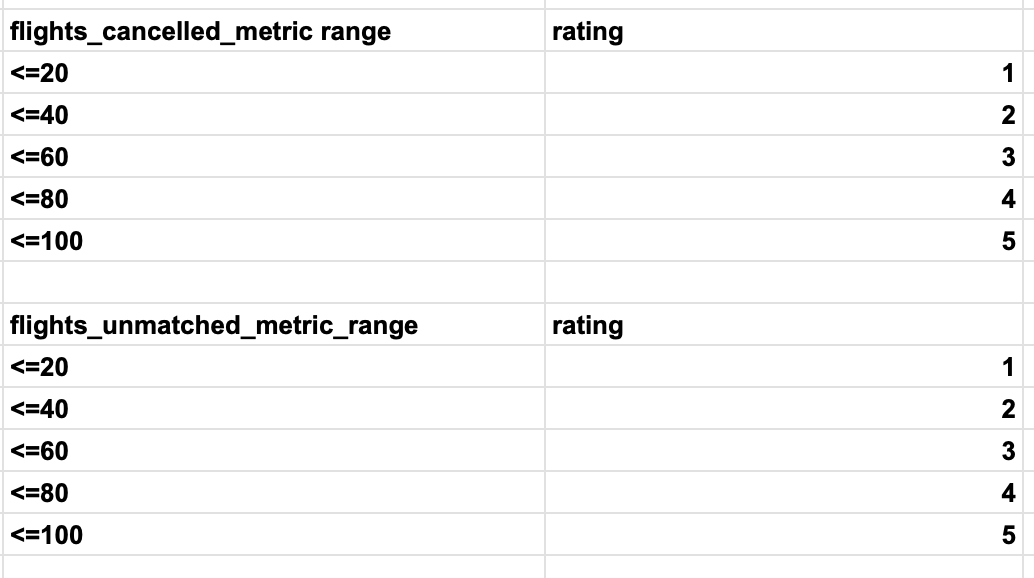
Airlines evaluate their success and popularity using many metrics, out of which Ontime Performance metric is globally accepted as a key KPI in the airline industry. Our data warehouse schema has on time performance data as fact table. It has entry\_id and reporting\_period as a composite primary key and has a one-one relationship with the rest of the dimension tables.

The dimension tables are carefully structured imagining the story each one highlights. Let us shift our attention towards the punctuality\_dim data. We have calculated a punctuality score based on the average\_delay\_times. These are categorised as follows. <=15 as 7, <=30 as 6, <=60 as 5, <=120 as 4, <=180 as 3, <=360 as 2 and >360 as 1

We can analyse airline performance using this score to analyse the punctuality of all flights. It has punctuality\_id and reporting\_period as a composite primary key.

We also have punctuality\_previous\_year\_month\_percent to investigate how flights have performed in the same month for both current and previous years. There are 0 values which are treated as data not available. This metric is calculated by taking a percent difference between the average\_delay\_mins and previous\_year\_month\_average\_delay.

Similarly, performance\_rating highlights the performance of an airline based on the flights\_cancelled\_percent and flights\_unmatched\_percent. We are going to convert these to rates and then take a weighted average to find the overall performance score. In our case we can assume that performance depreciated by cancelled flights is considered a principal factor (0.6) when evaluating performance rating as compared to unmatched flights (0.4)

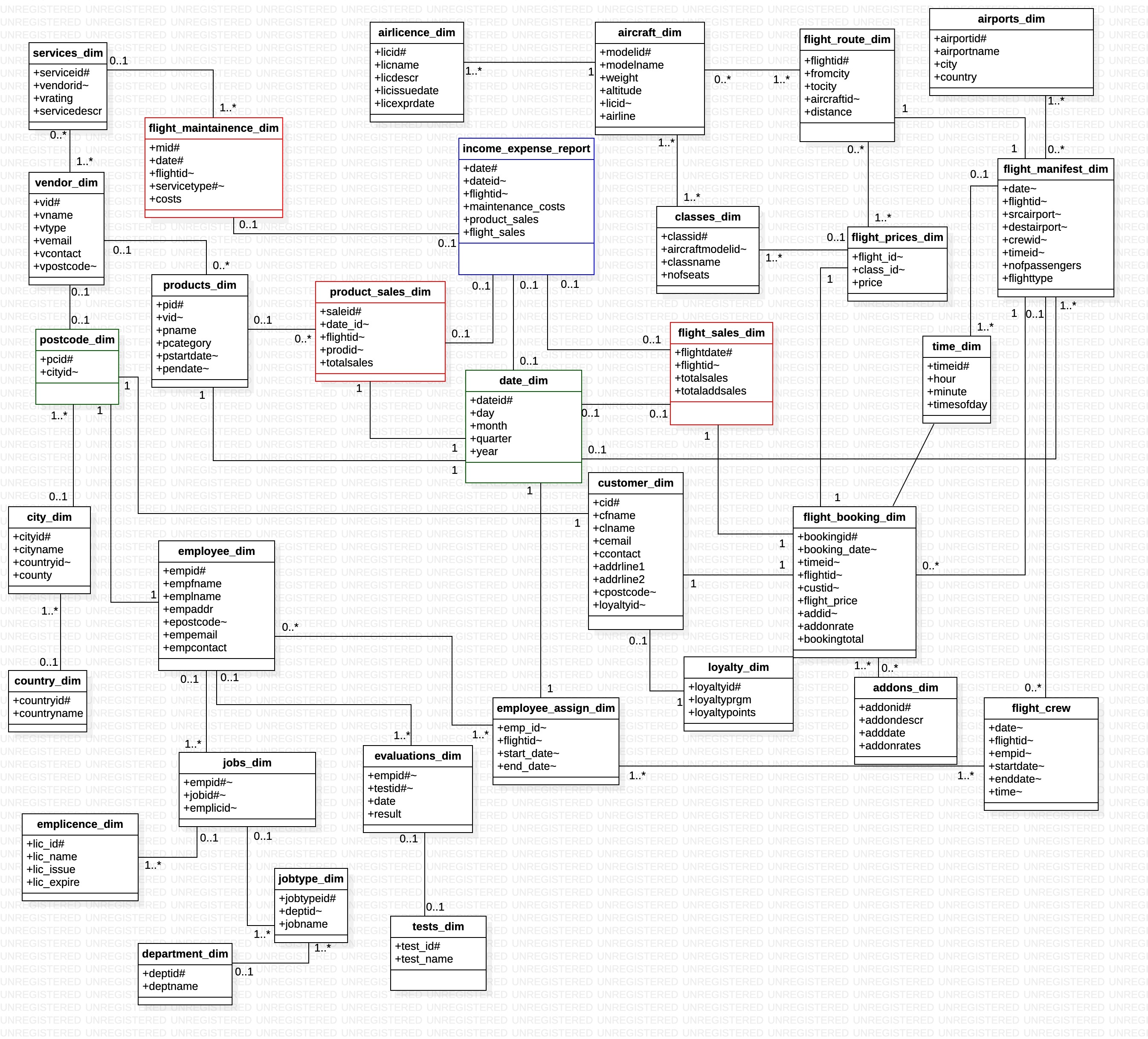


We also have previous\_year dim, which gives a brief insight into current and previous year comparisons with respect to average delay, flights matched and flights that were late between 0 to 15 minutes. This table has reporting\_id and reporting\_period as composite primary key.

Details pertaining to delays is consolidated in the delays\_dim were delay percentages of various ranges staring from 0-15 minutes, 15-30 minutes all the way up to delays >360 minutes is stored together. The delay\_id and reporting\_period together make up the composite primary key for this dimension table.

The overall ontime performance data related to an airline is stored in the ontime\_performance\_data\_fact table, that contains foreign keys to the delays, punctuality and performance tables all of which are one-one relationships. This table also includes flights that are 1-15 minutes early and flights that are more than 15 minutes early.

The diagram below shows a relational schema for an airline industry in general. This diagram clearly shows the delicate relationships that link multiple dimension tables. However, staying true to the business problem that we must solve for this assessment related to the airline industry, a star schema design for a data warehouse seems probable. There are alternatives such as snowflake and galaxy schema that can engage with complex relationships and data, in a more proficient way, however they are difficult to maintain. On the contrary a star schema results in multiple duplicates across data.

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

create table ontime\_performance\_data\_fact

(

entry\_id number(10),

reporting\_period number(10),

year integer,

month integer,

reportingID number(5),

punctualityID number(5),

performanceID number(5),

delayID number(5),

airline\_name varchar2(200),

flight\_type varchar2(1),

reporting\_airport\_id number(10),

reportingAirport varchar2(100),

origin\_dest\_airport\_id number(10),

originDestination varchar2(100),

originDestinationCountry varchar2(100),

flights\_15\_1\_minute\_early\_percent decimal(5,2),

flights\_more\_than\_15\_minutes\_early\_percent decimal(5,2)

)

create table previous\_year\_dim

(

reporting\_id number(10),

reporting\_period number(10),

year integer,

month integer,

airline\_name varchar2(200),

reporting\_airport varchar2(100),

origin\_destination varchar2(100),

origin\_destination\_country varchar2(100),

flights\_0\_15\_minutes\_late\_percent decimal(5,2),

prev\_year\_month\_early\_to\_15\_mins\_late\_percent decimal(5,2),

average\_delay\_mins float,

prev\_year\_month\_avg\_delay float,

number\_flights\_matched number(5),

prev\_year\_month\_flights\_matched number(5),

punctuality\_rating float,

prev\_year\_month\_punctuality\_percent decimal(5,2)

)

create table punctuality\_score\_dim

(

punctuality\_id number(10),

reporting\_period number(10),

year integer,

month integer,

airline\_name varchar2(200),

reporting\_airport varchar2(100),

origin\_destination varchar2(100),

origin\_destination\_country varchar2(100),

avg\_delay\_mins float,

prev\_year\_month\_avg\_delay float,

punctuality\_score float,

punctuality\_prev\_year\_month\_percent decimal(5,2)

)

create table delays

(

delay\_id number(10),

reporting\_period number(10),

year integer,

month integer,

airline\_name varchar2(200),

reporting\_airport varchar2(100),

origin\_destination varchar2(100),

origin\_destination\_country varchar2(100),

flights\_0\_15\_mins\_late\_percent decimal(5,2),

flights\_16\_30\_mins\_late\_percent decimal(5,2),

flights\_31\_60\_mins\_late\_percent decimal(5,2),

flights\_61\_120\_mins\_late\_percent decimal(5,2),

flights\_121\_180\_mins\_late\_percent decimal(5,2),

flights\_181\_360\_mins\_late\_percent decimal(5,2),

flights\_morethan\_360\_mins\_late\_percent decimal(5,2)

)

create table performance\_score\_dim

(

performance\_id number(10),

reporting\_period number(10),

year integer,

month integer,

airline\_name varchar2(200),

reporting\_airport varchar2(100),

origin\_destination varchar2(100),

origin\_destination\_country varchar2(100),

actual\_flights\_unmatched number(5),

flights\_unmatched\_percent decimal(5,2),

number\_flights\_cancelled number(5),

flights\_cancelled\_percent decimal(5,2),

cancelled\_flight\_metric float,

unmatched\_flight\_metric float,

performance\_rating float

)

Create table airport\_dim

(

Aiport\_id number(10),

Airport\_city varchar2(100),

Airport\_country varchar2(100)

)

Primary and foreign key constraints:

alter table previous\_year\_dim add constraint pk\_pyd primary key(reporting\_id,reporting\_period)

alter table PUNCTUALITY\_SCORE\_DIM add constraint pk\_pcd primary key(punctuality\_id,reporting\_period)

alter table Performance\_SCORE\_DIM add constraint pk\_psd primary key(performance\_id,reporting\_period)

alter table delays add constraint pk\_delay primary key(delay\_id,reporting\_period)

alter table airport\_dim add constraint pk\_airport primary key(aiport\_id)

alter table ONTIME\_PERFORMANCE\_DATA\_FACT add constraint fk\_prev\_year\_otp foreign key(reportingID,reporting\_period) references previous\_year\_dim(reporting\_id,reporting\_period)

alter table ONTIME\_PERFORMANCE\_DATA\_FACT add constraint fk\_punctuality\_otp foreign key(punctualityID,reporting\_period) references punctuality\_score\_dim(punctuality\_id,reporting\_period)

alter table ONTIME\_PERFORMANCE\_DATA\_FACT add constraint fk\_performance\_otp foreign key(performanceID,reporting\_period) references performance\_score\_dim(performance\_id,reporting\_period)

alter table ONTIME\_PERFORMANCE\_DATA\_FACT add constraint fk\_delays\_otp foreign key(delayID,reporting\_period) references delays(delay\_id,reporting\_period)

alter table ONTIME\_PERFORMANCE\_DATA\_FACT add constraint fk\_air foreign key(reporting\_airport\_id) references airport\_dim(aiport\_id)

alter table ONTIME\_PERFORMANCE\_DATA\_FACT add constraint fk\_dest foreign key(origin\_dest\_airport\_id) references airport\_dim(aiport\_id)

4. Discuss the steps you took in creating and populating the database. This should include the  
steps you took in preparing the data and the transformation tasks perform.

Most of the data available is in percentages. There are a few shortcomings obviously. Percentages cannot be aggregated, when performing analysis on data. The proposed solution is to distribute percentages across a range which inevitably result in transformations

Transformations:

a) Deleting the very first column run\_date which describe the dates the reports were generated.

b) Using LEFT in excel to slice the reporting period into two separate columns: year and month.

c) Using IFERROR along with IFS in excel to perform the evaluations on percentage values

d) Removing the blank rows by filtering blanks in the dataset and then deeting all blank rows.

e Consolidated month wise data in a single excel sheet.

f) We use VLOOKUP to match the IDs in ontime\_performance\_data from airport\_dim with respect to the airport names.

i) We calculate punctuality\_rating based on the average\_delay in minutes data, by ranking delays <=15 as 7, <=30 as 6, <=60 as 5, <=120 as 4, <=180 as 3, <=360 as 2 and >360 as 1.

ii) We calculate performance rating by taking a weighted average of cancelled\_metric and unmatched\_metric which are evaluated from the flights\_cancelled\_percent and unmatched\_flights \_percent. These metrics are stored as ratings. A weighted average is then estimated to give a performance\_rating.

The rating of cancelled flights is calculated based on cancelled\_flights\_percent and unmatched\_flights\_percent. We then get cancelled\_metric and unmatched\_metric, out of which we obtain the weighted average.

We also create a new column called prev\_year\_month\_avg\_percent which gives a difference between the average delays and previous year average delays.

We do not find obvious relationships in this dataset. Hence, we create additional primary keys in the data. The entire dataset is separated into 6 different excel sheets, each of which has a new additional column marking its unique identifier. This identifier would then help in joining the dimension tables to create a schema like structure in Tableau.

Our fact table is on time performance data table that stores a foreign key reference to all other dimension tables. On successful data transformation and prepping we loaded the data into their respective SQL tables ensuring that columns match the data they are representing. This step completed the creation of the warehouse.

5. Discuss how the airline industry in general can benefit from OLAP cubes giving examples of  
cubes in your discussion.

Traditionally data exists in 2-D, with rows and columns. From Excel to relational databases, data in its simplest form has always been represented this way. One major drawback while accounting for complexity always persists in this method- complex queries require JOINs, GROUP BYs and other SQL concepts to retrieve data.

To overcome this challenge, we introduce OLAP cubes. OLAP cubes as the name suggests is a multi-dimensional array composed of nested lists, that can store data associated with different key areas of business to derive quick insights. For example, you could have sales data, product data and customer data all in a single data cube. To lookup the most selling product in the year 2021 and for which customer group can now be carried out by just one single query. A data cube comprises of dimensions and measures. Dimensions are anything you can group data by, while a measure is a metric or fact that is aggregated by the cube for example total sales or average number.

The rise of data warehouses and the freedom of the cloud has made computing storage cheaper, thus rendering cubes ineffective. Let us consider how a data cube can make an airline business more efficient with examples.

1. Customer Cube:

Assuming we have sufficient data around customer preferences, travelling patterns, in-flight experiences, customer demographics, frequency of flight, premium customers, and other attributes, we can build a data cube varying over the dimensions such as month, year, type of seat (premium/economy/business etc), customer age group, and have measures such as difference is sales in two different quarters, number of times the customer has taken the same flight, number of families flying with the airline, destination, and any relevant factors that could make up the data cube. Every data cube is unique to the business question.

This cube can then be used to answer specific questions or can facilitate executives to deploy innovative marketing strategies to expand customer base, attract a certain customer group, enhance customer experience and much more.

1. Sales Cube

An airline generates revenue not only by flight bookings but many other channels. Airlines advertise products in their pamphlets such as accessories, garments, electronics and others. They also offer cargo services to transport goods. Personalized seat bookings also come at a charge. We need to compare revenue against expenses for any business. Our expenses could be maintenance scheduling costs, fuel costs, employee salaries, taxes and others.

There are factors such as ideal flight times, days, time of month and others. Airlines can leverage such information to capitalize on user demand.

A data cube can have all these attributes in a single place to analyse the Return on Investment in the past 5 years that the business has obtained or to find the total capital gains or expenses in tyre repairs compared against last year.

1. Performance Cube

Due to unforeseen circumstances or obvious reasons, the performance of an airline may not be adequate. A data cube can cluster all related information and then executives can decide the best strategy to solve challenges. Issues related to delays, flight cancellations, poor user experience, environmental factors, operator controller communication can all relay information necessary to deliver actionable insights to the Airline. There is a pattern that there were maximum flight delays in two consecutive years in the month of December, due to extreme weather such as blizzards or stormy rains.

6. Discuss the benefits of using a data warehouse in combination with a business intelligence tool like Tableau.

Tableau is a business intelligence tool. It helps to identify relationships between data from different dimensions so that stakeholders can predict trends or make more informed decisions. A relational database tried to normalize data so that, we can avoid duplication. A data warehouse comes from a data pipeline. We feed in data which corresponds to other dimensions of data. These relationships then guide our analysis which result in eureka moments.

Data warehouses with Tableau fetch the following benefits:

a) In retrospect we are working with data that has about 26,800 rows of data. However, a data warehouse can potentially store petabytes of data. The data is usually complex and has the potential to draw insights from information that is overlooked. By mapping dimensions against each other and using multiple measures we get a visualization of how our data looks and try to modify the visualization based on the numerous in-built graph types in Tableau.

b) Tableau also has a simple user interface. In-fact some professionals who use Tableau come from careers of medicine, finance, and mathematicians. The use of advanced mathematical concepts such as standard deviation, variance, percentile, and several filters make Tableau ideal to experiment with your data. It also helps with its capability to create schemas from the data available if there exist valid relationships. For example, if two tables EMPLOYEE and JOBS have JOBID existing is both tables, it is amazingly easy to make a union and create a relation. Similarly, we can just as easily create data clusters too.

A functionality like above really helps when we have simple Excel Sheets with numerous other sheets which can then be managed as dimension tables. Additionally, Tableau also supports popular Data Warehouse platforms such as RedShift, Azure, Hadoop, MongoDB, Oracle, and others.

c) With many organizations now shifting towards OLAP, Tableau is revolutionizing the way we understand data. It is an effective tool to empower Data Analysts to explain their findings in a more compact and relatable way to their Executives. It is an absolute package, that enables Data Engineers, Data Scientists and Data Analysts share a common vision. This means all professionals share access to the same source when we analyse warehouses using Tableau. This is a welcoming change considering just a few years ago, to answer a business question required data engineers to create an OLAP cube. A data analyst would then process the cube to draw insights. Any mismatch in requirements would result in the process repeating endlessly, eventually wasting valuable time. The era of data warehouses combined with the power of Tableau has made a significant impact on the world of Vizes.

7. Create 3 visualizations using Tableau. For each visualization, you should include the  
following:  
Aim of the visualisation  
The steps you took to create the visualisation  
Key findings from the visualisation

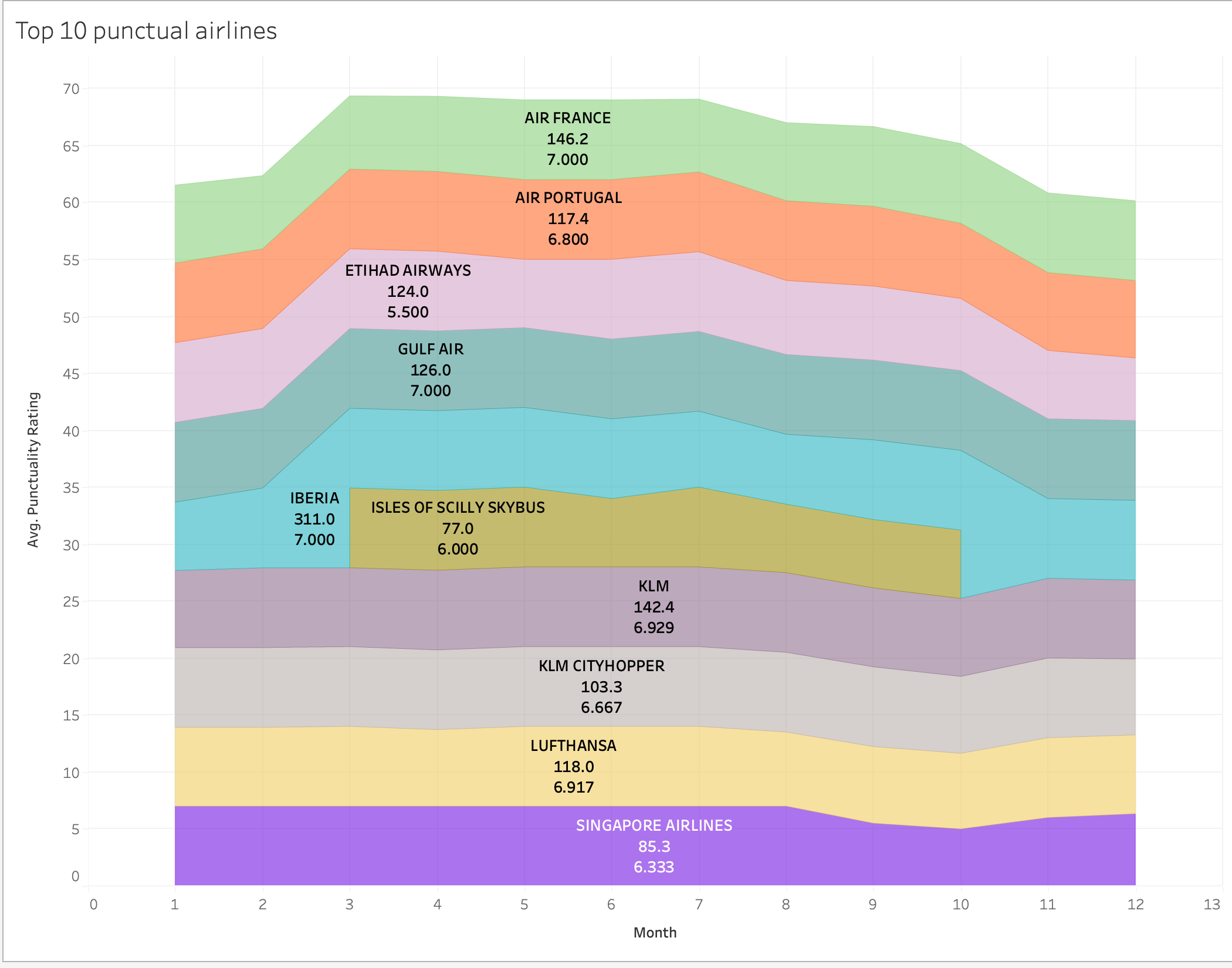
1) Most Punctual Airlines

The aim of the visualization is to fetch a high-level view of the airlines that are most punctual across all months in the year 2021. The area chart below depicts this very clearly.

These are a few insights we can take away from here.

1. It is clearly noticeable that AIR FRANCE is very punctual across all months and has an average number of matched flights as 146.2 which is the highest, which is not so surprising given the number of flights that commute every day from London to other EU countries. Followed closely is AIR Portugal which is ranked second highest in the list.
2. We also have Etihad airways and Gulf Air having parallel performance as they are partners as of 2020. This has significantly boosted their performance.
3. We can see Isles of Scilly Skybus Airline operating from the month of March to the month of October, as they operate seasonal scheduled services from Exeter.

To create this visualization, we first map the punctuality rating (averaged) in rows against the months in columns. We then filter the airline name based on the top 10 category for number of flights matched. We then add the punctuality rating as marks and add the corresponding labels to support the data. We also edit colours to make this visualization more accurate.



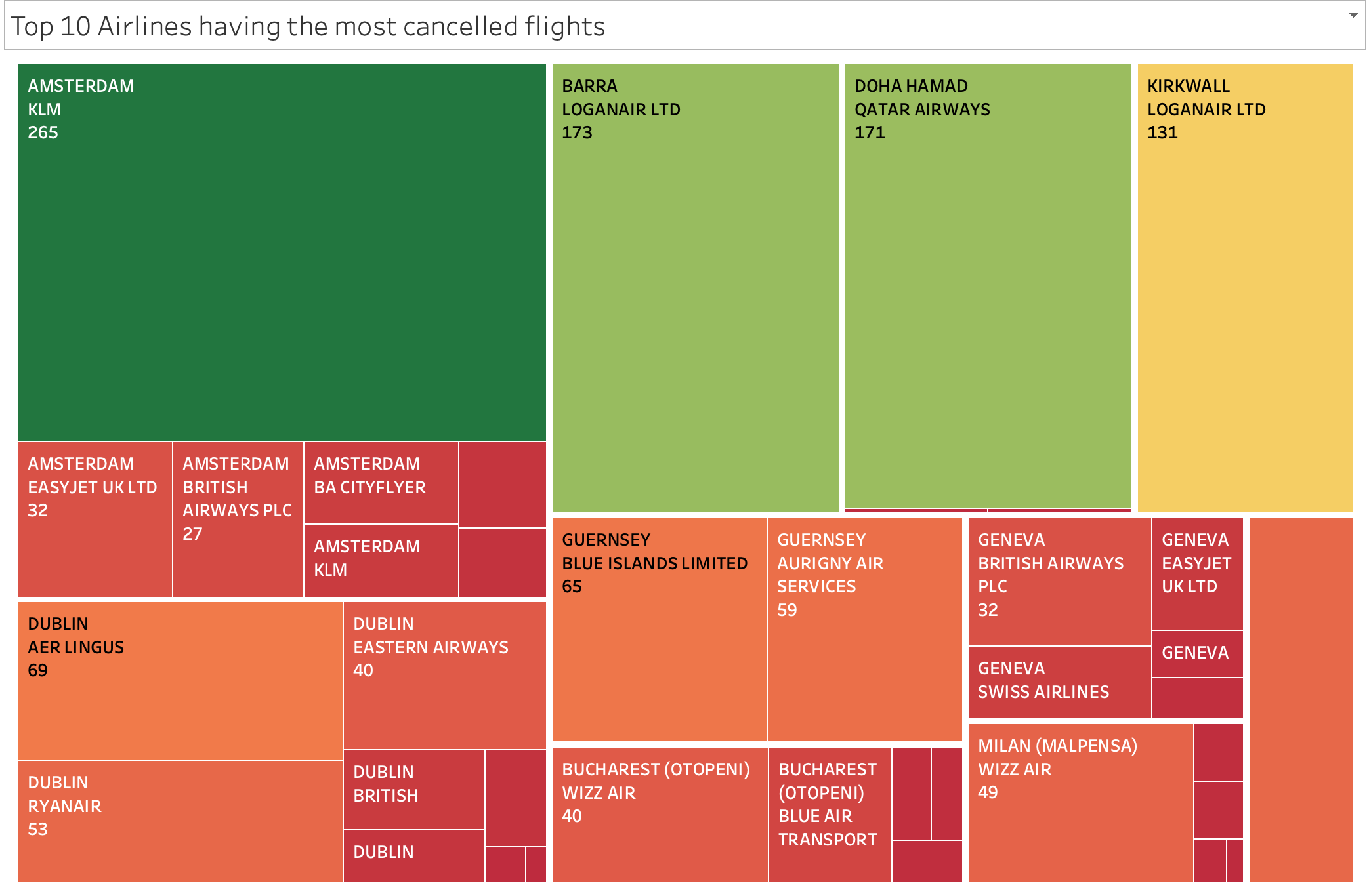
2) Top 10 Airlines having the most cancelled flights

The aim of this visualization is to give the airline company insight into which airlines are not performing so well and if it is conditional on the destination. The following tree map suggests that:

a) We can see that most flights cancel for flights going to Amsterdam, Guernsey, Geneva, Dublin, Bucharest, and Milan.

b) We can also observe that KLM airline going to Amsterdam single handedly has the highest number of flights cancelled going to Amsterdam. This would seem alarming but is normal it seems so. Schiphol airport at Amsterdam is often overwhelmed by security lines, which forces KLM to cancel flights often requested by Schiphol so that they can manage it.

c) Loganair going to Barra and Kirkwall combined gives a whooping 304 cancelled flights. This could potentially be due to forecasts of high winds across Nothern and Western Isles during winter months of 2021.



To create this visualization, we first mark the sum of number of flights cancelled and filter origin destination based on number of flights cancelled. Labels are then defined for the origin destination, the airline and total number of cancelled flights. Colour is defined for the number of cancelled flights.

3) Avg Delay Minutes for Airlines on UK airports.



The aim of this visualization is to view the average delay in minutes across all reporting airports in the UK. The geo map graph presents with the following insights

a) 2 Excel Aviation Ltd is very active across most reporting airports and has the maximum delay in Glasgow at 133 minutes while least delay at 3 in Newcastle. Reports suggest that number of flights cancelled at Glasgow airport are always at an all time high which affects the average delay.

b) The viz gives an overview of all the aiports in UK and their respective average delays in minutes.

The plot had to be matched with geolocations of airports, even though the list was automatically loaded, it had to be done manually. Some airport codes were not recognized.

The latitude and longtitude values are plotted to create a map. And then we mark all the airports alomg with their Avg Avg Delay in mins. We label the avg delay in mins, reporting airport and destination airport.

8. Conclusion

This exercise proved to be challenging for me. Upon understanding the requirenments of the assignment I was skeptical about the design specifications for the warehouse. It was unclear, if a warehouse had to be designed based on existing data or as advocated in lectures.

The liberty to choose Oracle to build a warehouse was very generous. If I had to reattemt this coursework, I would definitely work with Azure Data Pipelines, so that I get hands on experience in that.

Working with Tableau was challenging, quite contrary to the hype around it, that it is easy. Yes after experimenting with a few visualizations, I was more comfortable with Tableau, and the map plots really got me interested. An improvization on my approach here would be to work out how to use percent values. I faced some challenges here. I could not plot percent values as Measures, as they get aggregated as count, avg, stddev, or others. I then used them as dimension values.

When plotting geo plots, it was difficult to match data, as Tableau could not find all values automatically. A manual configuration was needed, which made me wonder if there was a more efficient way to do it.

Plotting charts was easy however plotting meaningful charts was difficult. The dataset did limit the transformations we could do. I wanted to use the percent values of various delays and that was confusing. A percent comparison was just not fitting with any Tableau viz.

It was an interesting learning curve for me, following which the steps are to understand more about Tableau and Excel. I would also focus on Azure/ AWS, so that I can be future ready.