Airline Company Data Warehouse

TEAM 1

- Mohamed AlGhaly
- Salma Ahmed
- Ahmed Ali

Project Documentation

STEPS

0- Understanding the business.
   - Conduct research on how Major Airline Companies operate.

1- Requirements Gathering.
   - Understand Business Needs.
   - Identify Data Sources.

2- Analyzing Source Systems' Data.

3- Defining Business KPIs & DWH Objectives.

4- Defining Business Processes.

5- Defining Granularity for each business process.

6- Choosing the Optimal DWH Architecture.

7- Deciding the technology stack.

8- Defining Dimensions and Fact Tables.
   - We need also to define the type of each (Fact table, Dimension Table, Measure)

9- Data Warehouse Modeling.
   - Creating the Logical Model (The Schema).

10- Creating the Physical Model.

11- Generating & Populating data.
   - In this step we will simulate data sources' data by generating data using a Python script.

12- Data Integration.
   - Moving data from source Systems to the DWH.

13- Indexing & Partitioning.

14- Data Analysis.

PS:

- The mentioned steps show a more **generalized** DWH Modeling **approach**, some of the steps may not be applicable in our case.
- We have excluded some crucial steps like choosing the technology stack which includes hardware, database software (e.g., Oracle, SQL Server, Snowflake) and choosing between On-Premises and Cloud.
- We have also excluded other steps like choosing the DWH provider, estimating DWH size, defining sparsity of the data, and determining hardware specifications.

- We act as the business owners and we don’t have Data Sources to refer to, so any assumptions made will be from our point of view and we will try to include as detailed description as possible.

STEP1

Defining KPIs & DWH Objectives

DWH Objectives:

The objective of this project is to assist the executive management to analyze their current business processes and expand the company by discovering new opportunities.

We can achieve this by creating a robust and structured Data Warehouse that consolidates data from various sources, facilitating analytical reporting and business intelligence.

Business KPIs & Analysis Scope:

- Enhancing the overall business performance.
- Analyzing the behavior of company’s frequent flyers.
- Measuring the performance of our marketing team.
- Measuring the performance of our Loyalty Program.
- Analyzing the company’s profit.
- Which customers use our services more frequently.
- Analyzing our revenues in each country.
- Defining our main source of revenue.
- Which customer segmentation (gold, platinum, titanium) is most valuable.
- Our most popular booking channels.
- Improving customer satisfaction.
- What are the booking patterns of frequent flyers, and what types of fare classes do they typically book.
- How do customer demographics, such as age or income level, impact travel behavior and preferences.
- Ensure a seamless operation of our loyalty program.

STEP2

Defining Business Processes

1- Ticketing Transactions (Reservations):

- This process captures transactional data related to bookings made by customers. It supports analyses of sales performance, customer booking behaviors, and revenue management.
- This process is mainly concerned with the Finance Team.
- The analysis scope is to analyze company's revenues.

2- Frequent Flyers:

- This process maintains detailed records of frequent flyer mileage transactions, supporting the management of the airline's loyalty program.
- This process is mainly concerned with the Marketing Team.
- The analysis scope is to analyze frequent flyer's behavior.

3- Customer Care:

- This process is mainly focused on achieving higher customer satisfaction.
- It captures customers' inquiries, complaints and keeps their feedback.
- This process is mainly concerned with the Customer Support Team.
- The analysis scope is to define any problems within the company and to ensure a better customer experience.

4- Flights:

- This process records key operational data for each flight conducted by the airline. It supports performance analysis, operational planning, and decision-making processes related to flight operations.
- It mainly focuses on our flight activity.
- This process is mainly concerned with the Upper Management Team.
- The analysis scope is to analyze the overall company's performance.

STEP3

Defining Granularity

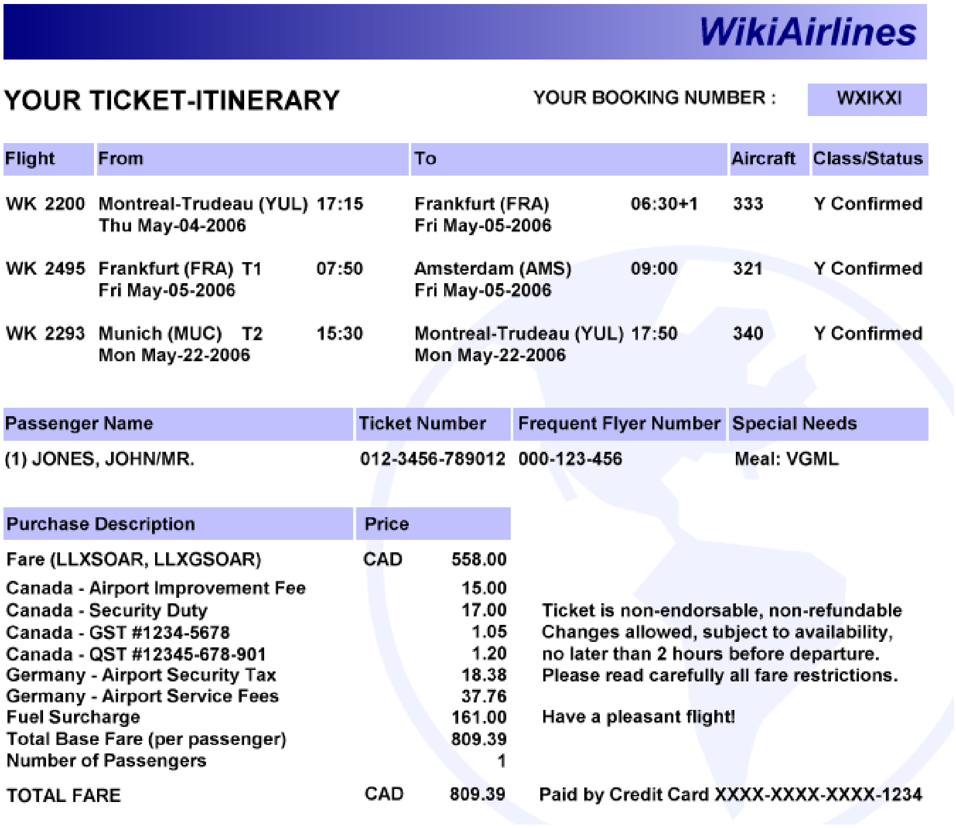
Possible Granularities

The following are the possible grains to consider across defined business processes.

- Transit:
  - The step from one airport to the other.
  - A flight from San Francisco to Minneapolis with a **stop/transit** in Denver will be mapped in **2 rows**:
    - San Francisco to Denver and Denver to San Francisco.
- Flight:
  - The flight from journey to destination.
  - A flight from San Francisco to Minneapolis with a stop/transit in Denver will be mapped in a **single row** holding data about the source and destination only.
- Trip:
  - A return flight that takes you from Paris to London and back from London to Paris is mapped in a **single row**.
- Daily, weekly, etc.

Choosing Granularities for Each Fact Table

Sample Ticket (Trip Level)



WikiAirlines

YOUR TICKET-ITINERARY

YOUR BOOKING NUMBER : WXIKXI

Flight	From		To		Aircraft	Class/Status
WK 2200	Montreal-Trudeau (YUL) Thu May-04-2006	17:15	Frankfurt (FRA) Fri May-05-2006	06:30+1	333	Y Confirmed
WK 2495	Frankfurt (FRA) T1 Fri May-05-2006	07:50	Amsterdam (AMS) Fri May-05-2006	09:00	321	Y Confirmed
WK 2293	Munich (MUC) T2 Mon May-22-2006	15:30	Montreal-Trudeau (YUL) Mon May-22-2006	17:50	340	Y Confirmed

Passenger Name	Ticket Number	Frequent Flyer Number	Special Needs
(1) JONES, JOHN/MR.	012-3456-789012	000-123-456	Meal: VGML

Purchase Description	Price	
Fare (LLXSOAR, LLXGSOAR)	CAD	558.00
Canada - Airport Improvement Fee		15.00
Canada - Security Duty		17.00
Canada - GST #1234-5678		1.05
Canada - QST #12345-678-901		1.20
Germany - Airport Security Tax		18.38
Germany - Airport Service Fees		37.76
Fuel Surcharge		161.00
Total Base Fare (per passenger)		809.39
Number of Passengers		1
TOTAL FARE	CAD	809.39

Ticket is non-endorsable, non-refundable
Changes allowed, subject to availability,
no later than 2 hours before departure.
Please read carefully all fare restrictions.

Have a pleasant flight!

Paid by Credit Card XXXX-XXXX-XXXX-1234

1- Ticketing Transactions:

We will work on the grain of each **ticket** booked by any customer (**Trip Level**).

- A customer books a flight from San Francisco to Minneapolis with a stop/transit in Denver will be mapped in a **single** in this table.
- As shown in the above image (sample ticket), the passenger pays for the trip regardless of any stops/transits.
- A 2/3/.. flights trip containing transits will be mapped to a single row.
- The above sample ticket for example will be represented as a single row in the fact table.

2- Frequent Flyers:

The most detailed grain in this fact table will be of each **flight** booked by any customer **on the loyalty program.**

- We will use this fact table to analyze frequent flyers' behavior and their booking patterns in addition to how they redeem and earn flyer mileages.
- The above Sample ticket will be mapped to 3 rows in our fact table, because it is a single trip with 2 transit steps or flights (flown on 3 legs).
- If the above sample ticket is not reserved by one of our frequent flyers it **Won't** be represented in the fact table.

3- Customer Care:

This most detailed grain is each action taken by a customer in our company, this action can be a complaint, feedback, or inquiry.

4- Flights:

The most detailed grain is each flight organized by the company (flight level).

- The above ticket will be mapped as 3 rows across all passengers flying on the same flight.

STEP4

Choosing DWH Architecture & Technology Stack

- We have chosen a Typical Two-layer Architecture (Ralph Kimball's) approach.
- It will be almost **impossible** to build a 3rd normal form Enterprise Data Warehouse without having a solid understanding of data sources.
- We will implement our DWH on Oracle DBMS.

STEP5

Defining Dimensions & Facts

Step five will be the capstone for the project which involves determining facts, dimensions, and measurements.

FACTS & MEASUREMENTS:

1- Ticketing Transactions (Transactional Fact Table)

a. What?

i. This is the core of our DWH which will help the financial team to analyze company's revenue.

ii. BI Developers will use this fact table to derive ideas to increase sales and improve the company's revenues.

b. Measurements:

i. Total Fare (additive).
- The total charge for the ticket.
  - Back to the sample ticket this numeric value would be 558.00$.

ii. Ancillary Revenue (additive).
- The total fare for using extra flight or airport services.
  - Back to the sample ticket this numeric value would be 37.76$.

iii. Governmental Taxes (additive).
- The total taxes issued by the government not the company.
  - Back to the sample ticket this numeric value would be 1.05 + 1.2$.

iv. Airport Taxes (additive).
- The total taxes issued by the company.
  - Back to the sample ticket this numeric value would be 18.38$.

v. Security Fees
- The security fees issued by the company.
  - Back to the sample ticket this numeric value would be 17.00$.

vi. Baggage Fees (additive).
- Fees charged for extra baggage on the plane.

vii. Discount (additive). => As a numeric value not a percentage (that is why it is additive)

viii. Upgrade Fees (additive).
- Fees charged for Upgrading the class for any of the booked flights.

ix. Other Fees (additive).
- Any other fees issued by the company.
  - Back to the sample ticket this numeric value would be 15$.

c. Dimensions:

i. Passenger
ii. Date
iii. Channel
iv. Class
v. Airport
vi. Passenger Profile
vii. Time
viii. Reservation ID (Degenerate Dimension)
ix. Fare Basis

2- Frequent Flyers (Transactional Fact Table)

a. What?

i. This is a fact table to keep track of each flight booked by a frequent flyer, this fact table will be valuable for the marketing team to measure the benefits of our **loyalty program**, **promotions**, and other marketing campaigns.

ii. This fact table will also be used to analyze the behavior of our **frequent flyers** to help us build brand loyalty.

b. Measurements:

i. Points Redeemed (additive).
- Flyer Miles Redeemed by this frequent flyer in this flight.
  - The miles can be redeemed into discount, extra services, etc.

ii. Points Earned (additive).
- Flyer Miles earned by this frequent flyer from this flight.

iii. Cancelled (additive).
- This is a 0/1 value that represents whether the passenger has cancelled the reservation or not.
  - This is additive as we can sum it to get the total cancellation or average it to get cancellation rate.
  - Analyzing cancellation is a separate business process, we will only partially analyze cancellation patterns for our frequent flyers.

iv. Overnight Stand (additive).

- This represents the hours passengers will wait in transit until the next flight takes off.
- This is a derived attribute that calculates the **lag** between the flight's arrival date and the next flight's departure date, we have decided to add it to the fact table for performance's sake.

Passengers can earn/redeem points from any other activities rather than booking a flight, those activities can even be conducted out of our company, but analyzing those activities are out of the analysis scope for this fact table (Will be a separate business process), so we will only be concerned about points earned from booking a flight or points redeemed as a discount on a ticket! (**When we say points, we refer to flyer miles)**

c. Dimensions:

i. Passenger
ii. Passenger Profile
iii. Class
iv. Date
v. AIRPORT
vi. Promotion
vii. Status
viii. Booking
ix. Flight
x. Fare Basis
xi. Channel
xii. Reservation ID (Degenerate Dimension)

3- Customer Care (Accumulative Fact Table)

a. What?

i. This fact table keeps track of the process of submitting inquiry/complaint/feedback.
ii. It tracks all the steps in this process from submitting an application to getting your issue solved!
iii. This fact table keeps track of actions taken by customers, these actions can be of any type (complaint, feedback, inquiry) and of any severity.

b. Measurements:

   i. Respond Delay (additive).
      - Time taken to respond to the application (in minutes).

   ii. Resolution Delay (additive).
      - Time taken to resolve the issue (in minutes).

   iii. Duration (additive).
      - Time taken by the customer to fill the application.

   - We had an ongoing discussion about whether to put Severity as a measurement or an attribute in the dimension.
   - And because this is a non-additive measure we decided to go with the dimension.
   - There is no point in calculating the total Severity, that is why it is non-additive.
   - BI Developers would use this attribute as a categorical attribute anyway.

c. Dimensions:

   i. Passenger
   ii. Passenger Profile
   iii. Flight
   iv. Date
   v. Channel
   vi. Employee
   vii. Interaction
   viii. Time

NOTE:

- Submitting a complaint or inquiry or whatsoever is an ongoing process that undergoes several steps, so we have decided that the best approach to model this business process will be using accumulative fact table.
- With accumulative fact tables we could track the process from opening an application to closing it!
- We have thought of the process as the following:
   - A customer can fill in an application, it can be filled from the website, mobile app, or a paper application.
   - The user being logged in we get the customer filling the application (for paper applications the customer is asked to provide his name).
   - Then he fills in the application details, he could specify one of his flown flights in the applications (optionally).

- The customer can also fire a complain on a specific reservation channel (optionally).
- The analysis team will keep track of this application from submitting it to resolving the issue to ensure a seamless flying experience and an optimal customer satisfaction.

4- Flights (Transactional Fact Table)

a. What?

i. This fact table keeps track of each flight organized by our company.

ii. This analysis scope for this fact table focuses on flight activity from a relatively higher picture.

iii. We will use this fact to analyze flight's performance and get insights regarding:

- Empty seats, not-showing passengers, fuel consumption, and so on.

b. Measurements:

i. Booked Seats
ii. Passengers Count
iii. Empty Seats
iv. Fuel Consumption
v. Crew Count

c. Dimensions:

i. Airport
ii. Date
iii. Flight
iv. Aircraft
v. Crew Member

Dimensions:

NOTE:

- The attributes of each dimension will be shown in the Model.
- We will only mention here the attributes that need to be described.

1- Date

- This is a typical calendar dimension for any DWH.
  - **Season** can be (Summer, Winter, and so on).
- **Type:** Role-Playing Dimension over for some fact tables and conformed dimension for others.

2- Time

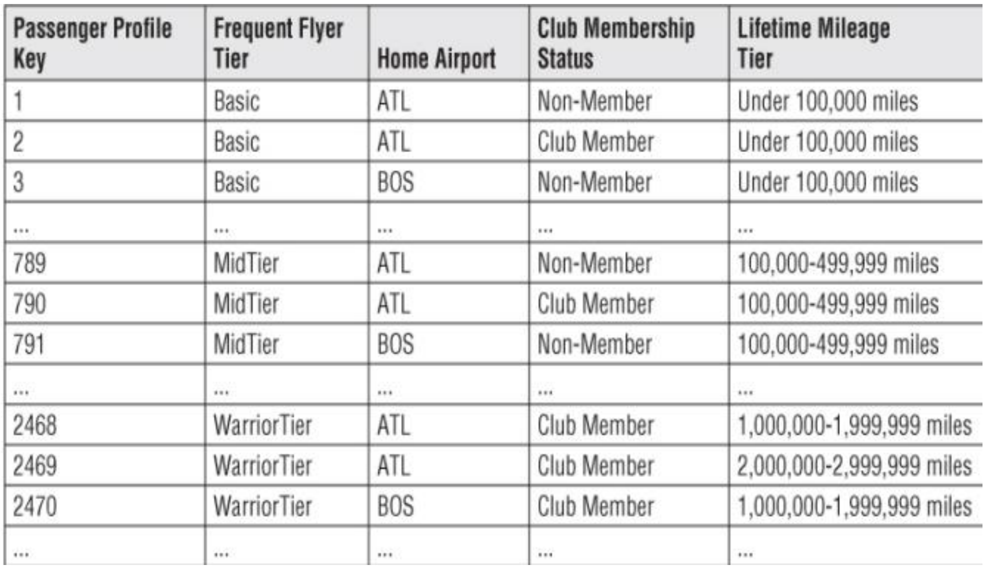
- We have conducted some research on how to present timestamp over different fact tables and we have found the best approach is to go with a Time Dimension (as suggested in The Data Warehouse Toolkit -Ralph Kimpall's Book-).
- **Type:** Role-Playing Dimension over for some fact tables and conformed dimension for others.

3- Passenger

- **Type:** Slowly Changing Dimension.
- We will use the 4th Type SCD as suggested in the toolkit.
  - The size of this table could go up to 10s of millions, so a typical type 2 SCD will be a huge load.
- This table holds the unchanging data for passengers.

4- Passenger Profile

- **Type:** Slowly Changing Dimension.
- This table holds the slowly changing data for each passenger.
- The table goes something like this:



Passenger Profile Key	Frequent Flyer Tier	Home Airport	Club Membership Status	Lifetime Mileage Tier
1	Basic	ATL	Non-Member	Under 100,000 miles
2	Basic	ATL	Club Member	Under 100,000 miles
3	Basic	BOS	Non-Member	Under 100,000 miles
...	...	...	...	...
789	MidTier	ATL	Non-Member	100,000-499,999 miles
790	MidTier	ATL	Club Member	100,000-499,999 miles
791	MidTier	BOS	Non-Member	100,000-499,999 miles
...	...	...	...	...
2468	WarriorTier	ATL	Club Member	1,000,000-1,999,999 miles
2469	WarriorTier	ATL	Club Member	2,000,000-2,999,999 miles
2470	WarriorTier	BOS	Club Member	1,000,000-1,999,999 miles
...	...	...	...	...

  - **Home Airport:** represents the home city's airport for this frequent flyer.
  - **Passenger Profile Key of -1:** Represents a non-frequent flyer.
  - We could also represent non-frequent flyers by putting **nulls** in the passenger profile key.

5- Airport

- **Type:** Role-Playing Dimension.
- A dimension holding data about all airports that the company operates on.

6- Aircraft

- **Type:** Conformed Dimension.
- This table holds data about all the airplanes the company owns.

7- Flight

- **Type:** Conformed Dimension.
- This table holds data about flights organized by the company.
    - At a first glance, one might get confused about the difference between this dimension and the Flight Activity fact table so let’s make sure this point is clear:
        - The fact table keeps records of each single flight organized by our company; it would help us to analyze our flight activity.
        - This dimension holds data for different flights organized by the company; and a flight is recognized by its duration, distance, arrival, and departure time stamp.
        - The dimension table records only flights having unique duration, distance, and schedule.

8- Employee

- **Type:** Conformed Dimension.
- This table holds data about the company’s employees.

9- Channel

- **Type:** Conformed Dimension.
- The channels the company supports for tickets’ reservation.

Column Name	Data Type	Description
ChannelKey	INT	Unique identifier for each reservation channel.
ChannelID	INT	Short identifier or code for the channel.
Name	VARCHAR(250)	Name of the channel, such as "Website" or "Call Center".
Type	VARCHAR(100)	Type of channel, e.g., "Online", "Offline", "Third-party".
Category	VARCHAR(100)	Category to further classify the channel like "Direct", "Indirect", "Partner".
ContactMethod	VARCHAR(100)	Primary method of contact, e.g., "Phone", "Email", "In-person".
Accessibility	VARCHAR(100)	Accessibility of the channel, e.g., "24/7", "Business hours".

10- Promotion

- **Type:** Conformed Dimension.
- This dimension holds data about the promotions the company offers.

Column Name	Data Type	Description
Name	VARCHAR(250)	The name of the promotion.
Type	VARCHAR(100)	Type of promotion, e.g., "Discount", "Bonus Miles", "Upgrade".
StartDate	DATE	The start date of the promotion.
EndDate	DATE	The end date of the promotion.
Terms	VARCHAR(250)	Description of the terms and conditions of the promotion.
Amount	INT	The discount/bonus miles amount provided by the promotion, if applicable.

11- Interaction

- **Type:** Conformed Dimension.
- A dimension holding data about the customers' interactions (Inquiry, Feedback, Complaint)

Column Name	Data Type	Description
InteractionKey	INT	Unique identifier for each interaction record.
InteractionID	VARCHAR(50)	A unique code or ID assigned to the interaction.
Type	VARCHAR(100)	Type of interaction, e.g., "Complaint", "Inquiry", "Feedback".
Description	VARCHAR(250)	Detailed description of the interaction.
Severity	INT	Severity level of the interaction, e.g., 0, 1, 2, 3, 4, 5
When_	VARCHAR(50)	When did the Interaction happen, "After". "Within", "Before", "Not-Related".

12- Booking

- **Type:** Conformed Dimension.
- Holding data about the booking for passengers, like Seat Number, Gate, Confirmation Number.
  - We saw this a context (descriptive) data rather that numeric (measurements).

13- Reservation ID

- **Type:** Degenerate Dimension.
- The Reservation ID For any ticket a passenger books.

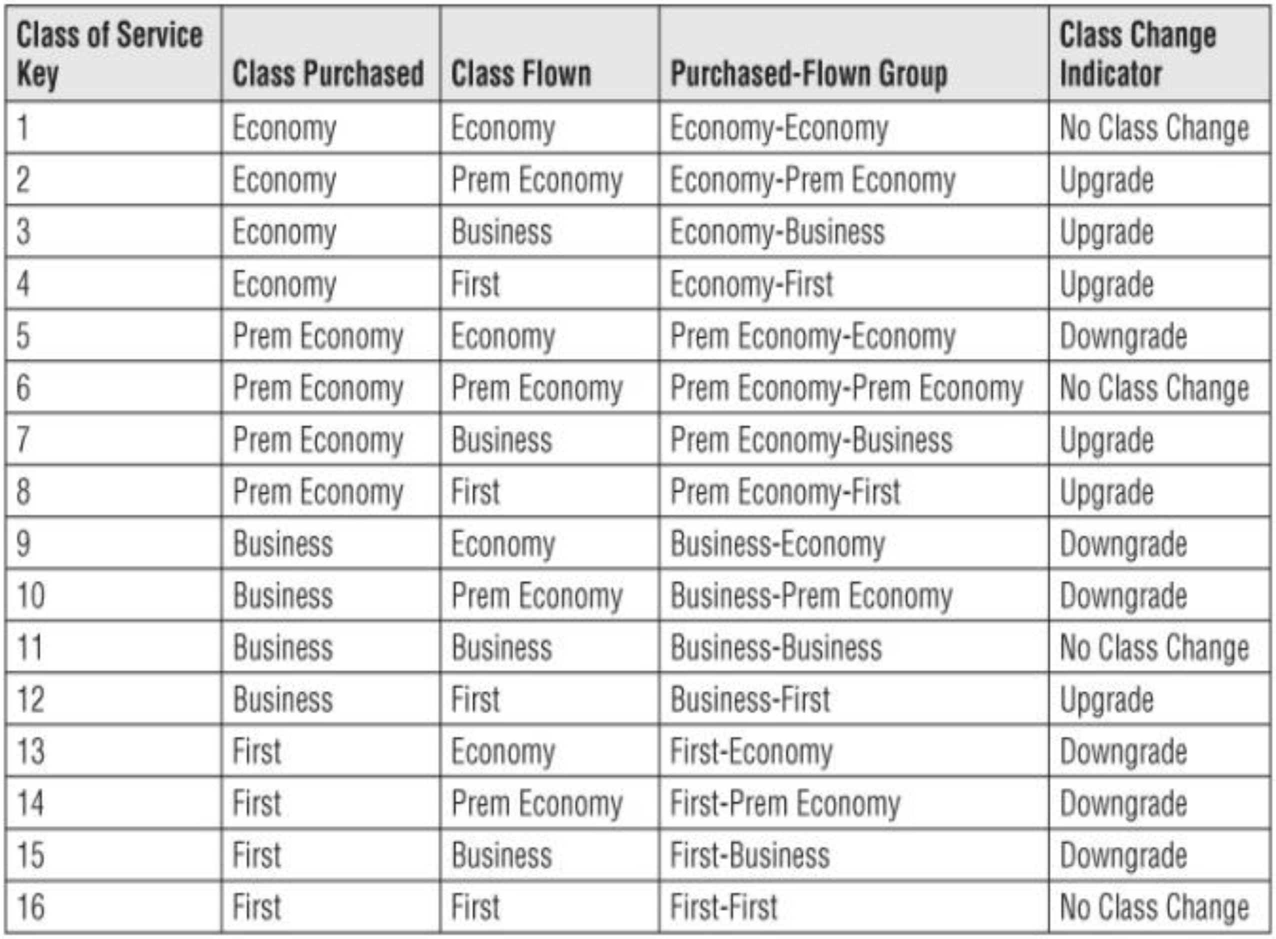
14- Status

- **Type:** Conformed Dimension.
- This dimension holds data for different frequent flyers (gold, platinum or titanium, and so on).

Column Name	Data Type	Description
StatusKey	INT	Unique identifier for each status record.
StatusID	VARCHAR(50)	A unique code or ID assigned to the status.
Name	VARCHAR(250)	The name of the status, such as "gold", "titanium", or "platinum".
Description	VARCHAR(250)	Detailed description of what each status means.
UpgradePriority	INT	Numeric value indicating the priority for upgrades.

15- Class

- **Type:** Conformed Dimension.
- The class of service flown describes whether the passenger sat in economy, premium economy, business, or first class.
  i. We need to keep track of upgrades in fare basis, that is why this table will be structured as follows:



Class of Service Key	Class Purchased	Class Flown	Purchased-Flown Group	Class Change Indicator
1	Economy	Economy	Economy-Economy	No Class Change
2	Economy	Prem Economy	Economy-Prem Economy	Upgrade
3	Economy	Business	Economy-Business	Upgrade
4	Economy	First	Economy-First	Upgrade
5	Prem Economy	Economy	Prem Economy-Economy	Downgrade
6	Prem Economy	Prem Economy	Prem Economy-Prem Economy	No Class Change
7	Prem Economy	Business	Prem Economy-Business	Upgrade
8	Prem Economy	First	Prem Economy-First	Upgrade
9	Business	Economy	Business-Economy	Downgrade
10	Business	Prem Economy	Business-Prem Economy	Downgrade
11	Business	Business	Business-Business	No Class Change
12	Business	First	Business-First	Upgrade
13	First	Economy	First-Economy	Downgrade
14	First	Prem Economy	First-Prem Economy	Downgrade
15	First	Business	First-Business	Downgrade
16	First	First	First-First	No Class Change

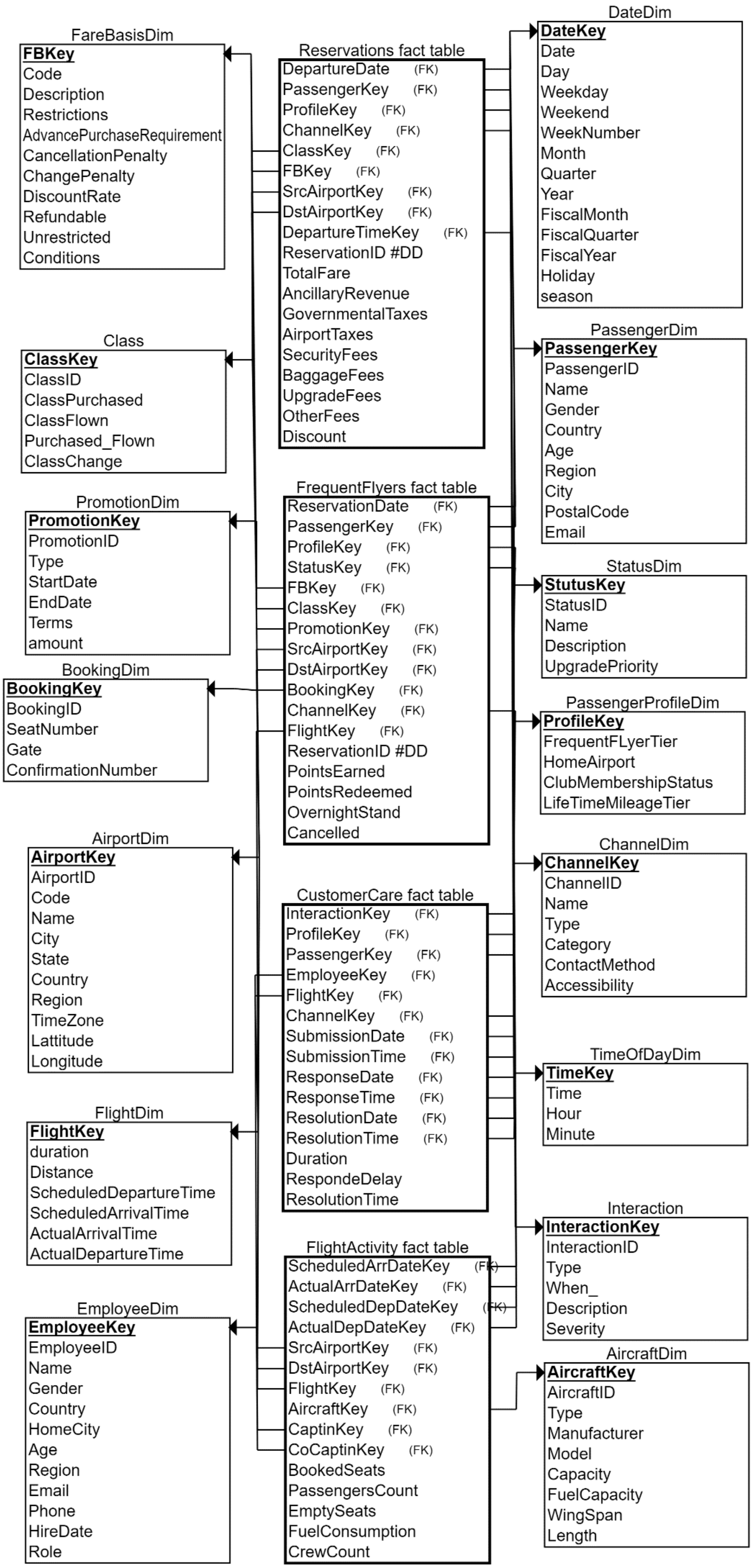
16- Fare Basis

- **Type:** Conformed Dimension.
- The fare basis dimension describes the terms surrounding the fare; It would identify whether it's an unrestricted fare, a 21-day advance purchase fare with change and cancellation penalties, or a 10 percent off fare due to a special promotion.

Dimension Description

Column Name	Data Type	Description
FBKey	INT	Unique identifier for each fare basis record.
Code	VARCHAR(50)	Short code representing the fare basis.
Description	VARCHAR(250)	Full description of the fare basis.
Restrictions	VARCHAR(250)	Describes restrictions such as non-refundable, non-transferable, etc.
AdvancePurchaseRequirement	INT	Number of days in advance the ticket must be purchased.
CancellationPenalty	FLOAT	Penalty for canceling the fare, often a percentage of the fare.
ChangePenalty	FLOAT	Penalty for changing the ticket.
DiscountRate	FLOAT	Any discount rate that applies to the fare, if applicable.
Refundable	INT	Indicates if the fare is refundable (1/0).
Unrestricted	INT	Indicates if the fare is unrestricted (1/0).
PromotionAssociated	INT	Indicates if the fare is part of a promotion (1/0).
Conditions	VARCHAR(250)	Describes other conditions for this fare basis.

FBKey	Code	Description	Restrictions	...
1	U21	Unrestricted fare	None	
2	A21CP	21-day advance purchase with penalties	Non-refundable, Non-changeable	
3	P10SP	10 percent discount due to special promotion	Purchase in advance required	



FareBasisDim
FBKey
Code
Description
Restrictions
AdvancePurchaseRequirement
CancellationPenalty
ChangePenalty
DiscountRate
Refundable
Unrestricted
Conditions

Class
ClassKey
ClassID
ClassPurchased
ClassFlown
Purchased_Flown
ClassChange

PromotionDim
PromotionKey
PromotionID
Type
StartDate
EndDate
Terms
amount

BookingDim
BookingKey
BookingID
SeatNumber
Gate
ConfirmationNumber

AirportDim
AirportKey
AirportID
Code
Name
City
State
Country
Region
TimeZone
Lattitude
Longitude

FlightDim
FlightKey
duration
Distance
ScheduledDepartureTime
ScheduledArrivalTime
ActualArrivalTime
ActualDepartureTime

EmployeeDim
EmployeeKey
EmployeeID
Name
Gender
Country
HomeCity
Age
Region
Email
Phone
HireDate
Role

Reservations fact table
DepartureDate (FK)
PassengerKey (FK)
ProfileKey (FK)
ChannelKey (FK)
ClassKey (FK)
FBKey (FK)
SrcAirportKey (FK)
DstAirportKey (FK)
DepartureTimeKey (FK)
ReservationID #DD
TotalFare
AncillaryRevenue
GovernmentalTaxes
AirportTaxes
SecurityFees
BaggageFees
UpgradeFees
OtherFees
Discount

FrequentFlyers fact table
ReservationDate (FK)
PassengerKey (FK)
ProfileKey (FK)
StatusKey (FK)
FBKey (FK)
ClassKey (FK)
PromotionKey (FK)
SrcAirportKey (FK)
DstAirportKey (FK)
BookingKey (FK)
ChannelKey (FK)
FlightKey (FK)
ReservationID #DD
PointsEarned
PointsRedeemed
OvernightStand
Cancelled

CustomerCare fact table
InteractionKey (FK)
ProfileKey (FK)
PassengerKey (FK)
EmployeeKey (FK)
FlightKey (FK)
ChannelKey (FK)
SubmissionDate (FK)
SubmissionTime (FK)
ResponseDate (FK)
ResponseTime (FK)
ResolutionDate (FK)
ResolutionTime (FK)
Duration
RespondeDelay
ResolutionTime

FlightActivity fact table
ScheduledArrDateKey (FK)
ActualArrDateKey (FK)
ScheduledDepDateKey (FK)
ActualDepDateKey (FK)
SrcAirportKey (FK)
DstAirportKey (FK)
FlightKey (FK)
AircraftKey (FK)
CaptinKey (FK)
CoCaptinKey (FK)
BookedSeats
PassengersCount
EmptySeats
FuelConsumption
CrewCount

DateDim
DateKey
Date
Day
Weekday
Weekend
WeekNumber
Month
Quarter
Year
FiscalMonth
FiscalQuarter
FiscalYear
Holiday
season

PassengerDim
PassengerKey
PassengerID
Name
Gender
Country
Age
Region
City
PostalCode
Email

StatusDim
StutusKey
StatusID
Name
Description
UpgradePriority

PassengerProfileDim
ProfileKey
FrequentFLyerTier
HomeAirport
ClubMembershipStatus
LifeTimeMileageTier

ChannelDim
ChannelKey
ChannelID
Name
Type
Category
ContactMethod
Accessibility

TimeOfDayDim
TimeKey
Time
Hour
Minute

Interaction
InteractionKey
InteractionID
Type
When_
Description
Severity

AircraftDim
AircraftKey
AircraftID
Type
Manufacturer
Model
Capacity
FuelCapacity
WingSpan
Length

DWH Model Discussion

- We have decided to go with a **Star Schema** for each data mart, with common conformed dimensions to unify the analysis.
- We have 4 fact tables each one is unique, unique in its grain, unique in its analysis scope, and each one represents a single business process associated with a single team.
- Each fact table has a wide variety of analysis and insight that can improve the business.
- The analysis scope will be covered in step 11, but now let's demonstrate the assumptions made for each business process.
- We could have more fact tables on top of these ones, but for the first delivery these are enough.

WHY did we choose star schema modeling?

We chose a star schema for the airline data warehouse due to its simplicity and optimal performance in query handling, which is ideal for facilitating fast and intuitive reporting and analysis on specific key subjects, like flights and customer interactions.

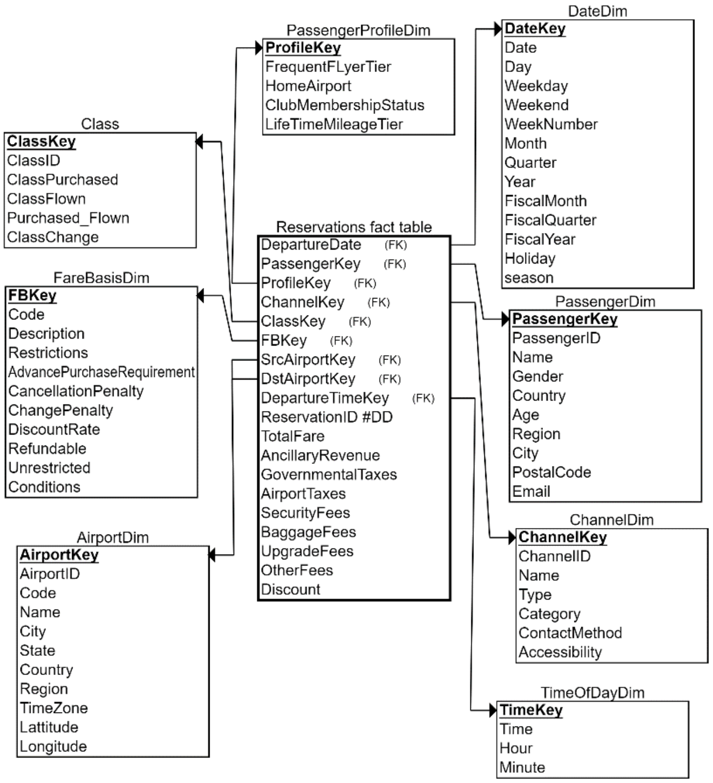
Let's now see a brief description of each fact table and the business assumptions made for the underlying business process.

- After conducting a deep market research on how Airline Companies operates, we have found that each company has its own assumptions and workflow.
- We have discussed the business and decided to go with the most common approach for each case, and follow our point of view for conflicting points.

IMPORTANT NOTE

- There are 2 types of revenue, **Earned Revenue** and **Unearned Revenue** those terms are vital in accounting, but to sum it up when a passenger pays for a service this is an unearned revenue, when he receives it, it becomes an earned revenue.
- We have decided to focus only on earned revenues (for finance team) as suggested on the toolkit, and that is why we are using **Departure Date** instead of ticket reservation date as the date for this reservation's fact table.
- For the marketing team we need to focus also on cancellation patterns, so we have used reservation date as the date for Frequent Flyers Fact Table.

Ticketing Transactions:



- This fact table helps us analyze our revenues.
- We have decided to use transactional fact table to have a control over each component contributing to company's profit.
- Using **Class** Dimension, we can identify which class (1st class, economic, …) contribute the most to revenue.
  We can also analyze impact of class upgrade on revenue.
- Using Fare Basis Dimension, we could analyze how each different component of the fare basis affects our performance.
- Using Airport, we can get insights about which Countries/Cities/… are most profitable for us.
- With this fact table we could get some insights about how our loyalty program is performing (financially).

Why did we choose to work on the grain of the Trip?

- After deep research, we decided to go with the following assumptions:
  - When a passenger pays for a ticket that has more than one flight or transit, this ticket is considered a single line item.

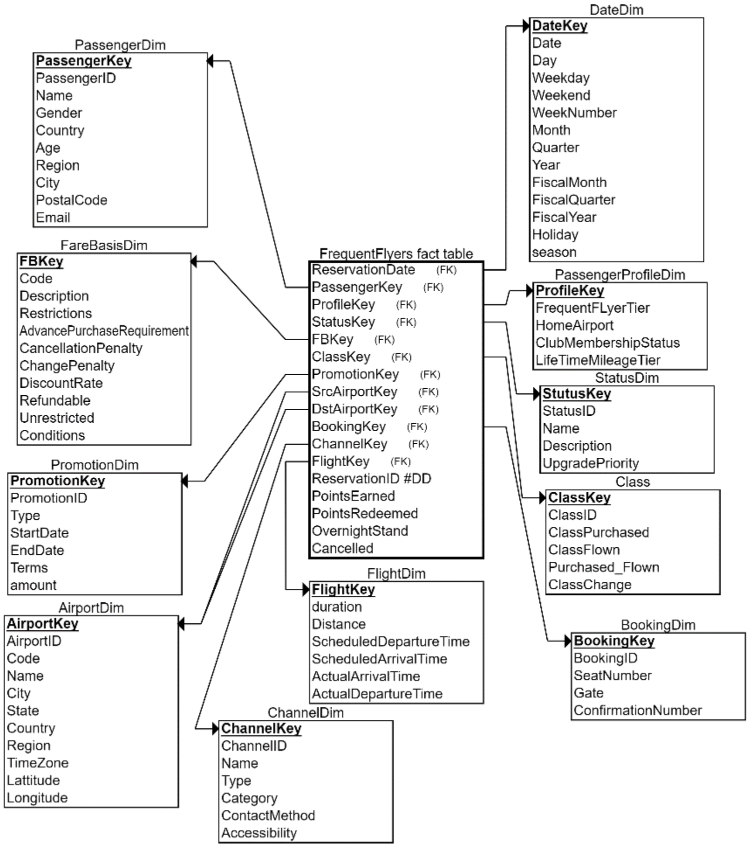
Limitations to this fact table:

- We **can't** calculate profit from this fact table, as we will need integration with other business processes like maintenance, HR, and so on.
- For a sales company calculating the cost for an item is easy, making it possible to calculate profit from sales fact table, but as we are providing a service rather than a product, we can't calculate the cost for each purchased ticket!

Enhancements:

- The first deliverable only focuses on flight activities, we can add new business processes to the next one to cover profit analysis.
- But focusing only on flight activities, we can't analyze our profits.
- We can create more aggregated fact tables on top of this one for performance's sake.
- For example, the Tool Kit suggests building a 90-day periodic fact table on top of this one, but for us this is enough.

Frequent Flyers:



- This fact table helps the marketing team to see the booking pattern for our frequent flyers.
- There are some duplications between this fact table and the previous one, but the 2 tables tackle 2 different business processes concerning 2 different teams and the 2 tables are built on different granularities.
- We have decided to use the reservation date to load this fact table and to add a cancellation measure to the fact table, to indicate whether or not the passenger has cancelled the trip.
- We know that trip cancellation is a separate business process that needs its own analysis, but defining cancellation pattern for frequent flyers would be helpful for the marketing team.

Why did we choose to work on the grain of the Flight?

- After carefully going through the requirements, we have decided that to get an accurate estimation of frequent flyers' behavior we need to analyze each single flight.
- Also, using this grain makes it easy to analyze overnight stand and transit steps for each flight.

Limitations to this fact table:

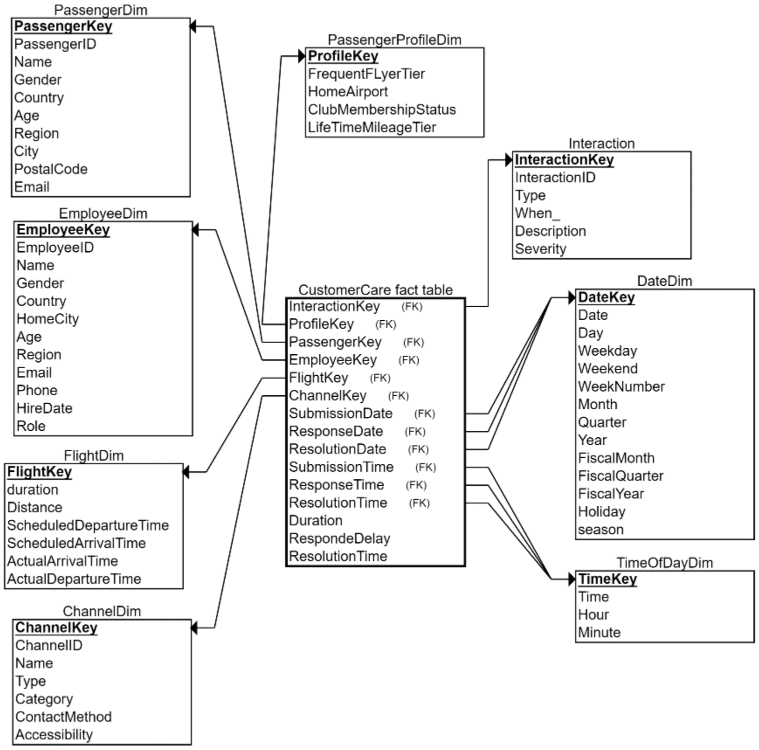
- Passengers can earn points from various channels, not only direct airlines, but also indirect airline partners such as cobranded credit cards, supermarkets, car rental companies, and international hotels and resorts. The points can be either redeemed for flights, flight upgrades or additional services such as extra baggage and airport lounge access.
- As stated on the business requirements (the first deliverable should focus on the flight activity), we are only considered with the flight activity, so we will ignore point earned/redeemed by using our partners' services and we will only focus on point earned/redeemed by traveling with us!

Enhancements:

- The next deliverable can include a new fact table (for a separate business process) modeling loyalty program, including transactions across all partnered companies.

Customer Care:

- This fact table helps us to ensure an optimal customer experience.
- Customers can file a complaint because of an unpleasant flying experience with us.
- They can also make complaints about unpleasant situations while trying to book a ticket.
- It is not limited to complaints, inquiries and feedback are always welcomed.

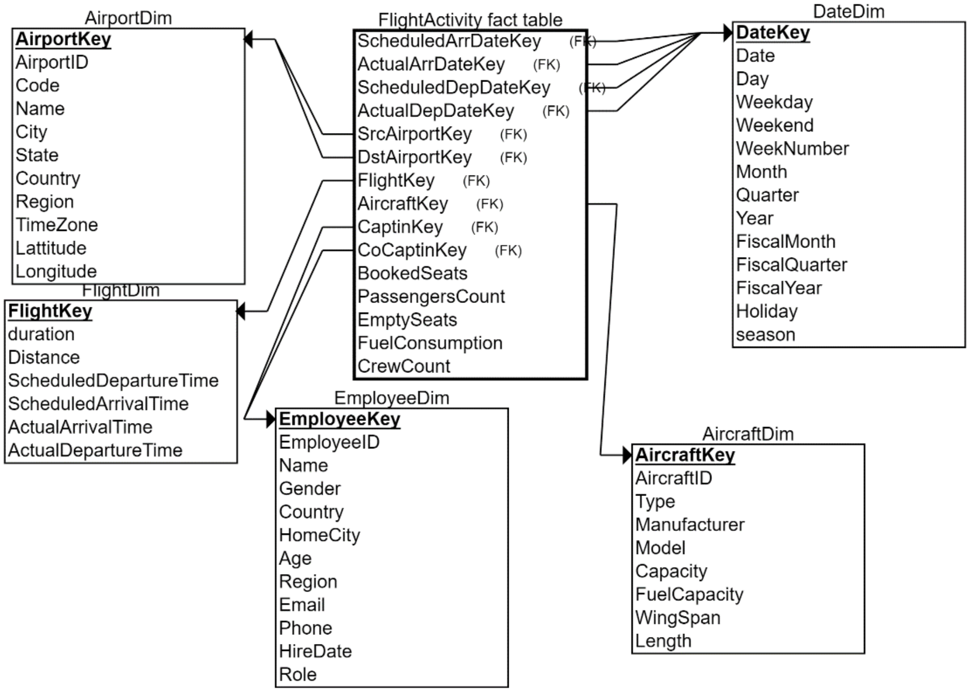


Why did we choose an accumulative fact table?

- We see customer care as an ongoing process, that undergoes some constant steps, from hearing from customers, and communicating with them, to giving our valuable customers what they want.
- So, we need to be able to analyze each step along the way, making an accumulative fact table the best choice!

Flights:

- This fact table will be used to monitor the company's performance and enhance its operation.
- We will use this fact table to gain some insights about booking behavior over different flights, different airports, different aircraft, and so on.
- For the upper management tier this fact table is very crucial, for instance we can know how many flights we organize monthly, and so on.



STEP7

Physical Model

Step nine will be creating the table in both excel sheet for validation and in oracle database for analysis.

The SQL Script for creating the DWH Tables using Oracle SQL can be found in a file named “**Schema Creation.SQL”,** while the excel file containing the tables, columns (name, data type), and indexes is named “**DWH Tables.XLXS”.**

STEP8

Populating Data

The Python script for creating the sample data is in a file named **“Poulate.PY”.**

The SQL Scripts for populating the data are included in Folder named “**Data Population**”.

STEP9

Indexing & Partitioning

- When it comes to Indexing & Partitioning, there are huge differences between a DWH and an operational Database.
- For us, we will index all Foreign Keys columns in each Fact Table.
- Then we will partition each fact table based on the date key.
- That is just a starter, we will analyze frequent queries to add more indexes/partitioning.
- Each unique column will be **CONSIDERED** (A possible candidate) for a unique index, each low cardinality categorical column will be considered for a bitmap index, and each highly queried numeric column will be considered for a balanced tree index.
- The script for indexing and partitioning fact tables is in a file named “**Indexing & Partitioning.SQL**”.
- By default, any DBMS enforce a clustered index on Primary Key columns, and most DBMSs enforce a unique index on unique constrained columns
- And the following are all the indexed and petitioned column alongside a brief explanation on why it is necessary.

let's talk for a little bit about each type of index we used and why did we choose such an index.

NOTE:

- Taking about each index and the algorithm used to implement it would be a lot of fun, but it is way out of the scope of this project and highly dependant on the underlying DBMS, so we will just give a hint on each one.

INDEX TYPE	**WHEN**	**WHY**
CLUSTERED (IMPLICIT)	Primary Key Columns	physically order the data in a table based on the indexed column.
(NON-CLUSTERD)	Categorial Data (With wider value range)	stored separately from the data and contains a copy of the indexed column, along with a pointer to the corresponding data.
B_TREE (NON-CLUSTERD)	Foreign Keys Columns	highly efficient for range-based queries.
UNIQUE	unique columns	This index enforces uniqueness for the indexed column or columns.
HASH (Not supported in Oracle)	Columns used in equality filtering	Uses hash-function to retrieve data in constant time so it is used for equality filtering and join conditions.
BITMAP	Categorial Data (low cardinality)	It uses a bitmap to represent the data, with each bit representing a possible value, so it is used on columns with a small number of distinct values.

- For me, I prefer using a clustered index on the PK columns, non-clustered index (B-tree as an example) on categorial data with wider range of values, bitmap for categorial data with 2-5 values, unique index for any unique column, and the last but not least is hash index which can do magic as it uses a hash function to retrieve a piece of data from a table containing billion of rows in just O(1) time.

- Hash index can be used on any column used a lot in where statement with an equal sign maybe for filtering or joining tables, and it also can be used to perform hash joins.

NOTE:

- Indexes can cause overhead or kill the performance of your database, so use it wisely knowing what you are doing, or just leave it to the DBMS and it will do a great job for you, unless you can do a greater job DO NOTHING.
- The appropriate type of index for a particular situation will depend on a variety of factors, including the size and type of data, the frequency and type of queries, and the overall database design.

Possibly Indexed Columns (Beside Foreign Keys)

- All Dimension keys (FKs inside each fact table would hold a Non-Clustered B-Tree Index as it is a numeric column used in joins and range-based filtering.
- The following possibly indexed columns (to be considered based on frequent queries (All are non-clustered indexes).

Table Name	Column Name	Data Type	Index Type	Why
Reservations	Only Foreign Key Columns			
FrequentFlyers	Only Foreign Key Columns			
CustomerCare	Only Foreign Key Columns			
FlightActivity	Only Foreign Key Columns			
FareBasisDim	Code	String	B-Tree	It will be frequently used in filtering and grouping.
Class	ClassChange	String	Bitmap	It will be frequently used in filtering and grouping.
	ClassPurchased	String	Bitmap	Will be frequently used in range filtering.
	ClassFlown	String	Bitmap	Will be frequently used in range filtering.
PromotionDim	Type	String	Bitmap	It will be frequently used in filtering and grouping.
	StartDate	Date	B-Tree	Will be frequently used in range filtering.
	EndDate	Date	B-Tree	Will be frequently used in range filtering.
BookingDim	BookingID	INT	B-Tree	Will be used frequently in grouping.
AirportDim	Name	String	B-Tree	It will be frequently used in filtering and grouping.
	City	String	B-Tree	It will be frequently used in filtering and grouping.
	Country	String	B-Tree	It will be frequently used in filtering and grouping.
	Region	String	Bitmap	It will be frequently used in filtering and grouping.
EmployeeDim	Gender	String	Bitmap	It will be frequently used in filtering and grouping.
	Role	String	B-Tree	It will be frequently used in filtering and grouping.
DateDim	Date	INT	B-Tree	It will be frequently used in filtering and grouping.
	Day	INT	B-Tree	It will be frequently used in filtering and grouping.
	Month	INT	B-Tree	It will be frequently used in filtering and grouping.
	Year	INT	B-Tree	It will be frequently used in filtering and grouping.

	Quarter	INT	Bitmap	It will be frequently used in filtering and grouping.
	Season	String	Bitmap	It will be frequently used in filtering and grouping.
PassengerDim	Gender	String	Bitmap	It will be frequently used in filtering and grouping.
	Country	String	B-Tree	It will be frequently used in filtering and grouping.
	Region	String	Bitmap	It will be frequently used in filtering and grouping.
	City	String	B-Tree	It will be frequently used in filtering and grouping.
Status	Name	String	Bitmap	It will be frequently used in filtering and grouping.
PassnegerProfile	FrequenFlyerTier	String	Bitmap	Will be used frequently in grouping.
	HomeAirport	String	B-Tree	Will be used frequently in grouping.
	ClubMembership	String	Bitmap	Will be used frequently in grouping.
	LifeTimeMileageTier	String	Bitmap	Will be used frequently in grouping.
ChannelDim	Name	String	Bitmap	It will be frequently used in filtering and grouping.
	Type	String	Bitmap	It will be frequently used in filtering and grouping.
	Category	String	B-Tree	Will be used frequently in grouping.
TimeofDaydim	Time	INT	B-Tree	It will be frequently used in filtering and grouping.
	Hour	INT	B-Tree	It will be frequently used in filtering and grouping.
	Minute	INT	B-Tree	It will be frequently used in filtering and grouping.
Interaction	Type	String	Bitmap	Will be used frequently in grouping.
	When	String	Bitmap	Will be used frequently in grouping.
AircraftDim	Type	String	B-Tree	It will be frequently used in filtering and grouping.

STEP10

Gaining Insights

- All the questions mentioned in the project requirements and way more are addressed in the analysis file "**Analysis.docx**"**.**
- We build a very basic dashboard on top of the populated data you can find it in a file named "**Dashboard.pbix**"
- **Disclaimer:**
    - We don't provide a product, we provide a service; and we don't have any material cost, we only have operational cost.
    - Because of that calculating cost or profit based on only flight activity (first deliverable) was now doable, so we decided to calculate revenue and provide earned revenue-based analysis on the first deliverable, but the second deliverable will focus on the following:
        - Loyalty program analysis: Analyzing frequent miles earned and redeemed across all partner companies and across all our business process.
        - Analyzing more business processes like, maintenance, HR, commissioning, and so on to analyze operational cost and profit of the company.