Airline case study: multidimensional design

José Luis Puente Bodoque

January, 2023

Abstract

This report documents the data warehouse design steps for the airline case study. As stated in the wording of the task, we follow the four stages proposed in [Ross and Kimball, 2013]. In addition, we also compare the data actually obtained from the data source to the current design in order to suit the design to the data source. Finally we flesh out the dimensions tables with descriptive attributes and draw the star schema.

Contents

1 Business process

Marketing department want to deeper understand the customers' purchases according to their boarding passes. Thus the business process we are modeling is **flight sales transactions based on the data recorded on boarding passes**.

2 Granularity

The airline captures data at the **level of journey** and distinguishes between **leg** and **trip**. These data is gathered in the passenger's **boarding pass**.

The detail level of the information contained in the boarding pass is such as we manage the origin and departure airports of both leg and trip, departure and arrival hours of the journey, channel and method of payment, airport type or even passengers' birth date.

3 Dimensional hierarchy

Dimensions are enumerated among with dimension attributes in step 3. The figure 1 shows for each dimension the hierarchy levels as a directed acyclic graph (DAG):

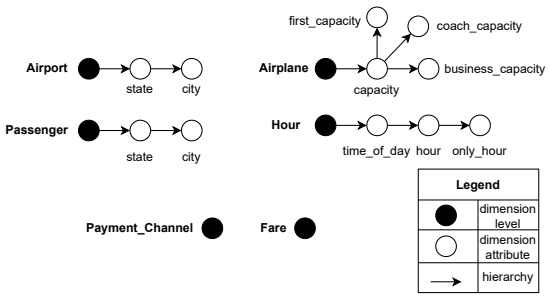

Figure 1: Hierarchical dimensions

Note that payment_channel and fare dimensions have only first-level members. "Each dimension represents all possible descriptions that take on single values in the context of each measurement. Dimensions represent the *who*, *what*, *where*, *when*, *why* and *how* associated with the event".

4 Fact table and measures

"Identifying the facts means identifying measures". In our study case, they're explicitly enumerated in the step 4. Each measure corresponds to the information collected on the passenger's boarding pass for journey. So **boarding pass** represents the fact table and measures are: price, the points obtained (or exchanged), the accumulated delay and time spent in the destination. In spite of what the wording says, true measures are in `frequentflyer.csv` file: **base fare**, **miles**, flight **delay** and **ticket numbers**.

5 Analysis of data source

The data source has a dimensional structure like the data warehouse we're designing. This means that the correspondence between the table header and our dimension attributes are straightforward.

On the other hand there are some issues with `frequentflyer.csv` related to the design:

- "Available seats" measure isn't found but "ticket number" does, so we have replaced this for that.
- "fare_class_keys" column values reach up to 20. Any airline provides basically four classes of seat. We wouldn't like to drop any row so we have modified entirely this column to limit to 4 the possible values.
- The same matter as the previous item for " channel_key_field".
- It lacks the departure and arrival hour keys so we must look them up during the ETL load process.
- We will assign surrogate keys to each row during the ETL load process. They will serve as an immediate identifier of a fact table row without having to navigate through multiple dimensions.

  Note: The surrogate keys won't be part of the primary key.

Data source key	Fact table key	Key type	Part of PK	Identifies...
	id_boarding_pass	SK	N	a journey, boarding pass, fact table row
flown_key	id_trip	FK	Y	a trip
customer_key	id_passenger	FK	Y	a passenger
leg_origin_key	id_leg_origin	FK	Y	an origin airport of a leg
leg_dest_key	id_leg_dest	FK	Y	a destination airport of a leg
trip_origin_key	id_trip_origin	FK	Y	an origin airport of a trip
trip_destination_key	id_trip_dest	FK	Y	a destination airport of a trip
	id_departure_hour	FK	Y	a departure hour at the origin airport
	id_arrival_hour	FK	Y	an arrival hour at the destination airport
flight_key	id_airplane	FK	Y	an airplane
fare_class_key	id_fare	FK	Y	a seat class
channel_key	id_channel_payment	FK	Y	a payment method

Table 1: Fact table keys

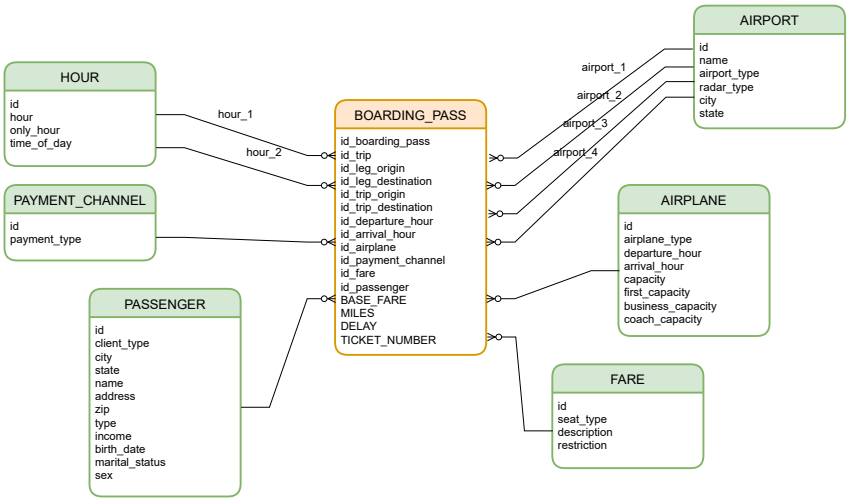

Figure 2: Star schema

6 Star schema

The figure 2 shows the definitive star schema. It suits now both requirements and data source.

Some dimensional tables migrate more than once their primary keys to the fact table:

- hour dimension migrates twice its primary key to identify the **arrival** and **departure** hours

  Note: the task asks for a *date* dimension, not for an *hour* dimension, but there is nothing related to date in the data source provided.

- airport dimension migrates four times its primary key to identify the **origin and destination airports of the leg** and the **origin and destination airports of the trip**

  Note: origin airport of a leg and origin airport of a trip matches when only one flight composes a trip. The same applies to destination airports.

References

[Ross and Kimball, 2013] Ross, M. and Kimball, R. (2013). *The data warehouse toolkit: the definitive guide to dimensional modeling*. Wiley.