Task 1: Multidimensional design

Cristian Torres Ortega, Pablo Gamarro Lozano

En el ejercicio se pide que se realice el diseño multidimensional para el caso de estudio de una aerolínea. Pues bien, antes de iniciar con el mismo se debe observar de qué información se parte, para así poder abarcar la actividad con éxito.

Como en el propio ejercicio se indica, se parte tanto de información escrita facilitada por el departamento de marketing como de un conjunto de archivos del tipo “csv” repleto de datos aportados por la aerolínea. De acuerdo con esta información, para el diseño del Data Warehouse se va utilizar una técnica de diseño híbrida, donde primero se analizará los requerimientos del cliente (reporte del departamento de marketing) y luego se completará el diseño analizando el conjunto de datos aportado por la compañía de vuelo.

Comenzando con el análisis del reporte proporcionado por el departamento de marketing, se debe de buscar o establecer en primer lugar el hecho o hechos existente, es decir, el objeto del análisis. En este caso, se observa de manera explícita que se requiere analizar los vuelos realizados por los pasajeros más recurrentes, más concretamente, se deben de analizar las tarjetas de embarque para poder analizar de manera completa estos vuelos, ya que dichas tarjetas de embarque recogen las medidas de hechos que son importantes para el departamento de marketing, por lo que el análisis debe girar alrededor de ellas. Esto se puede observar, atendiendo al texto subrayado, en la siguiente imagen:

The marketing department of an airline wants to analyse the flight activity of each member from their frequent flyer program. This department is interested in analysing which flights choose the frequent passengers of the company, which airplanes they travel, what rate they pay, how they earn and exchange their points, what is the duration of their residence in destination, etc.

The process of developing a data warehouse to meet the needs of this marketing department is detailed below. This process is determined by the four stages proposed by Kimball for the design of a data warehouse.

Step 1: select the business process to model
Selects the management flights made by frequent passengers of an airline. We do not focus on flights reserve that do not result in the boarding of a frequent plane passenger.

Step 2: Select the granularity of the fact that will represent the business process
In this study case, the airline captures data at the level of journeys. One route represents an airplane that takes off from an airport and lands at another airport without intermediate stops. We know information about the date of make it, the time of exit, the time of arrival, the origin and destination airport, the base price of the journey, the points obtained (or spent) on the way and, finally, the delay accumulated by the journey. We also know the airplane that makes a trip and the base rate, as well as, the available seats of each type on each flight.

The airline also collects data from the itinerary, which is equivalent to the reservation or flight ticket with its corresponding number of reservation and a reservation date. An itinerary consists in many routes. For example, if a client reserves an itinerary to travel from Alicante to Malaga and then back to Alicante, this itinerary can be like the following routes: Alicante-Madrid, Madrid-Málaga, Málaga-Alicante. Actually the passenger only wants to go from Alicante to Malaga and return without worrying about the connections (Madrid). However, intermediate path data are needed for more complete analyses.

Each fact will correspond to the information collected on a boarding pass of each frequent passenger.

Una vez detectado el único hecho que contiene el reporte, el siguiente paso es comenzar con la búsqueda de las dimensiones. Las dimensiones son características o variables que se utilizan para describir y comparar los datos (el hecho). De manera general, la forma de proceder es analizar la información de manera detallada, subrayando todo aquello que sea relevante para nuestro hecho, y posteriormente analizar toda la información recolectada diferenciando entre dimensiones, atributos de dimensiones y medidas del hecho.

Esta separación es sencilla, si observamos que la palabra clave subrayada es un término general, que hace referencia a una posible agrupación de los datos, se trata de una dimensión. Si la palabra clave parece una especificación dentro de una dimensión, se tiene un atributo de una determinada dimensión. Por último, si la palabra clave es un valor numérico y de importancia para el negocio, se tiene una medida del hecho.

La clave está en entender que una dimensión es un término general que agrupa y organiza los datos, mientras que las medidas de los hechos son valores numéricos que ayudan a cuantificarlos y son de por sí, valores de mucha importancia para el negocio. Por lo tanto, se comienza con el proceso de detección de información clave:

Step 2: Select the granularity of the fact that will represent the business process

In this study case, the airline captures data at the level of journeys. One route represents an airplane that takes off from an airport and lands at another airport without intermediate stops. We know information about the date of make it, the time of exit, the time of arrival, the origin and destination airport, the base price of the journey, the points obtained (or spent) on the way and, finally, the delay accumulated by the journey. We also know the airplane that makes a trip and the base rate, as well as, the available seats of each type on each flight.

The airline also collects data from the itinerary, which is equivalent to the reservation or flight ticket with its corresponding number of reservation and a reservation date. An itinerary consists in many routes. For example, if a client reserves an itinerary to travel from Alicante to Malaga and then back to Alicante, this itinerary can be like the following routes: Alicante-Madrid, Madrid-Málaga, Málaga-Alicante. Actually the passenger only wants to go from Alicante to Malaga and return without worrying about the connections (Madrid). However, intermediate path data are needed for more complete analyses.

Step 3: choose the dimensions that will be applied to each fact

If the information on the fact corresponds to the data of a boarding pass, as stated above, the dimensions are the follows:

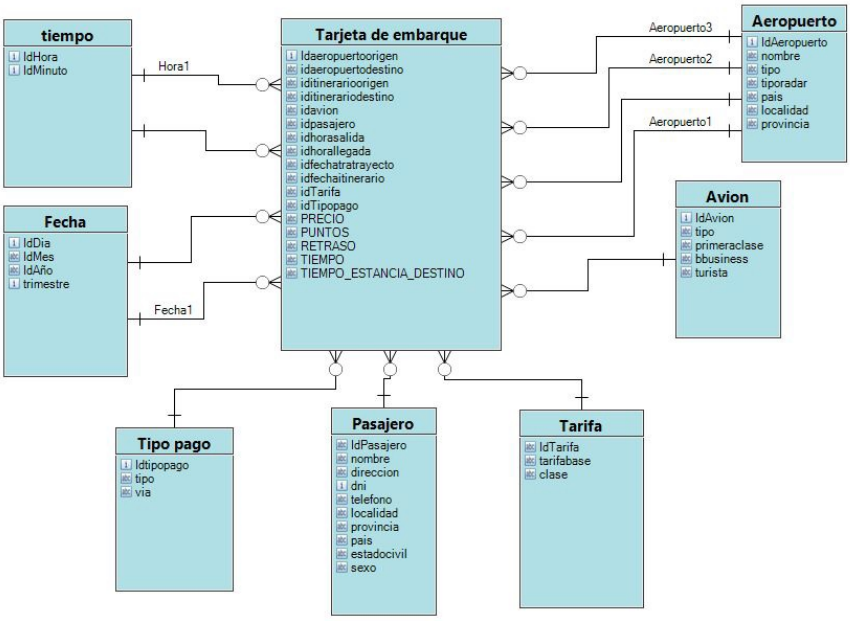
The type of payment identifies how the itinerary was purchased (telephone, Internet, etc.).

- The origin and destination airports for each route (including information on the name, type and type of radar).
- For frequent passengers, there is all the valuable information for the airline, including the type of frequent passenger.
- For each route, the plane that performs it is known.
- The base fare for each journey is also known.
- The different dates must be added as time dimensions.

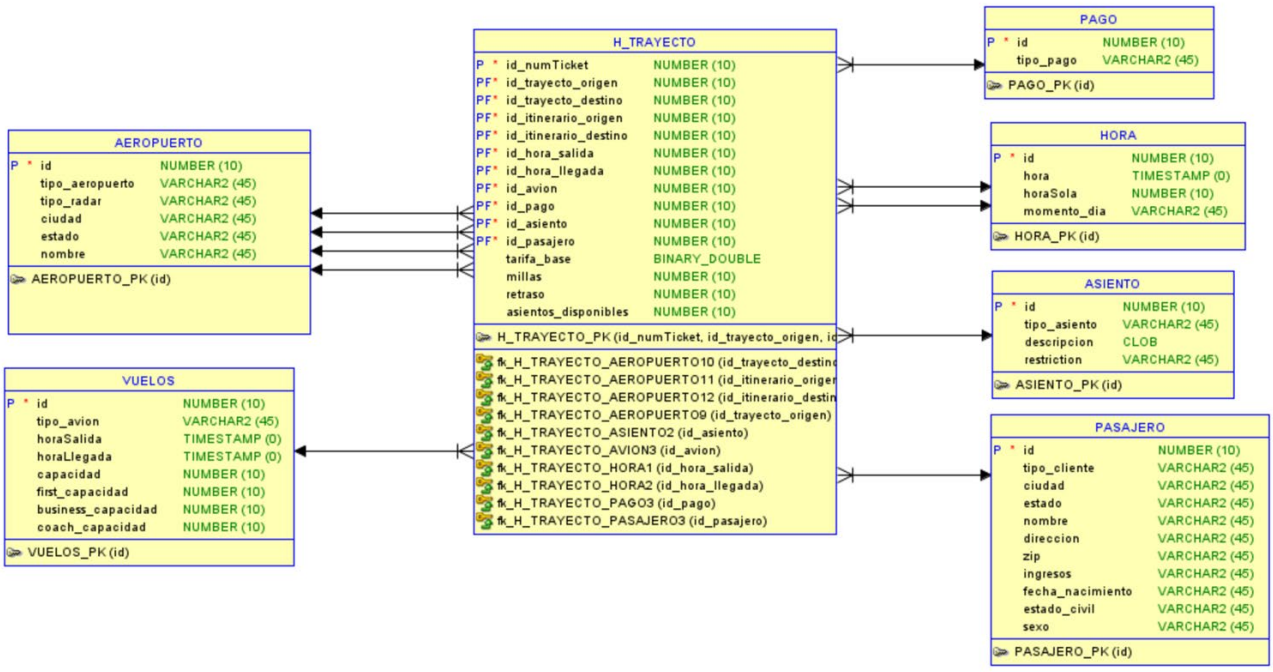
Step 4: identify the measures in the fact table

The measures for each journey are: the price of each journey, the points obtained (or used, depending on the type of base fare) per journey and the accumulated delay for each journey. The derived measures can be: total delay of each itinerary, total price of the flight, points of each itinerary, duration in minutes of each route, etc.

A continuación, se procede a mostrar un esquema lógico de toda la información que se ha recolectado. Como el fin es el análisis de los datos de manera multidimensional, y tenemos un hecho relacionado con múltiples dimensiones, se va a optar por adoptar el esquema estrella, ya que proporciona rapidez en las consultas debido a su estructura y un acceso a los datos muy intuitivo. De tal modo que, el hecho se situará en la parte central y estará conectado por las "foreign keys" con cada una de las dimensiones:



Ahora bien, como se ha indicado en el inicio del proyecto, se ha aplicado un diseño híbrido, por lo tanto, ahora se debe de analizar la información de los conjuntos de datos proporcionados por la aerolínea en formato de “csv”, en busca de nueva información clave o detectar información que haya que eliminar. Este paso se realiza con el fin de ayudar a perfilar de una manera aún más realista el esquema estrella:



Como se observa, ha cambiado la estructura, ya que se ha pasado de tener siete dimensiones a seis y se han realizado cambios sustanciales en los diferentes niveles de las dimensiones o atributo:

1) Se ha eliminado la dimensión fecha, esto se debe principalmente a que en los datasets aportados de la aerolínea no se recogen datos sobre las fechas. En su lugar se ha creado una dimensión hora. Por otra parte, se observa que en la dimensión vuelos se tiene almacenadas también las horas. De este modo, almacenar la información de las horas de salida y llegada en la dimensión vuelo puede ser útil para realizar agregaciones o filtrados por vuelo específico y tener acceso a las horas asociadas sin necesidad de acceder a la dimensión hora. En cambio, almacenar esa información en la dimensión hora puede ser útil para realizar agregaciones o filtrados por horas específicas y tener acceso a los vuelos asociados sin necesidad de acceder a la dimensión vuelo. Todo depende de las preguntas sobre el negocio que se requieran hacer. Además, que esta información sea redundante no aumenta la complejidad ni el tamaño de la base de datos de forma crítica en este caso, pues la cantidad de líneas de vuelos en un aeropuerto nunca es una cantidad de registros demasiado grande para una base de datos.

2) La dimensión avión pasa a llamarse vuelo, donde a parte de los atributos de la anterior dimensión avión, también almacena las horas de salida y llegadas de los vuelos, para realizar análisis simples sobre ellos. Esta dimensión también recoge la capacidad total, así como la capacidad en las diferentes clases.

3) Aeropuerto y pasajero, prácticamente siguen similares.

4) La dimensión tarifa se incluye en las medidas del hecho directamente, como “tarifa_base”, ya que es un valor numérico y de importancia para el análisis.

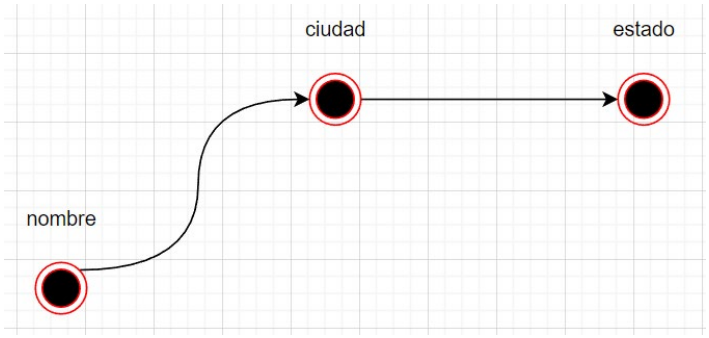
5) La dimensión tipo de pago pasa a llamarse pago, recogiendo en una variable de tipo “varchar” el canal de pago.

6) Se ha creado una nueva dimensión asiento que recoge el tipo de asiento, una descripción y la posible restricción.

7) Se ha modificado las medidas de hechos en la tabla hecho en función a los datos disponibles.

Así se obtiene un esquema estrella ajustado a la realidad de los datos aportados. Por último, queda pendiente realizar las jerarquías entre los diferentes niveles de las dimensiones:

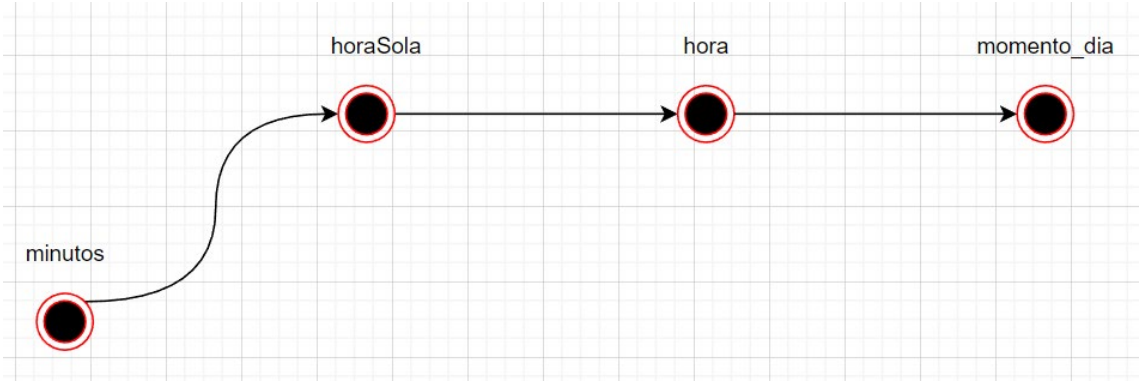
AEROPUERTO:



VUELOS: No tiene jerarquía

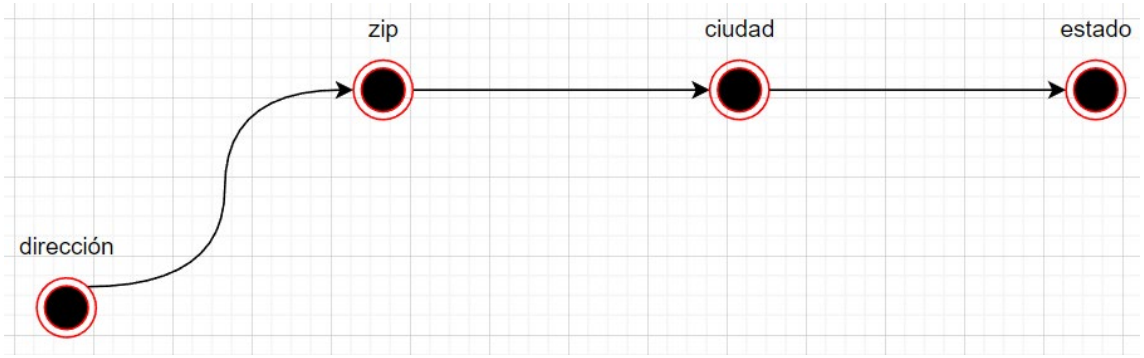
PAGO: No tiene jerarquía.

HORA:



ASIENTO: No hay jerarquía.

PASAJERO:


zip
ciudad
estado
dirección