Data Warehouse Modelling Workgroup

IE Masters Big Data & Business Analytics - Business Intelligence and Data Warehousing

Group E
Maria Joyce, Oday Almajed, Ana Chavarri Ceballos, António Teixeira De Sousa Crespo Carvalho, Maximiliano Franco Martin, Mark Fleming-Williams

Dataset - Campaign Data

Contents **1**

Data Analysis

i) **Analyse the data set**
The data set represents data about an advertising campaign which took place in 2016. There is sales data about various products which were sold at shopping centres in Catalonia, in the north-east of Spain. The shopping centres are divided amongst six cities, within which three are based in the periphery and three in urban centres. Also included are the total customer visits to each shopping centre on each day. The data set contains advertising information which is divided into two advertising points in each shopping centre - big screen and information point screen. The data set contains sales and advertising points data for 15 products from five product families – there are three types each of pizza, television, trousers, shoes, and (picture) frames.

Inspecting the data more closely the impacts data all fits between values of 1 and 12. Within this range there is also more granularity: the three product types in a product family are divided within distinct ranges, for example in one shopping centre the first quarter saw one pizza showing values between 1 and 3, the second pizza between 4 and 8, and the third between 9 and 12. These mini-ranges change every quarter. These numbers likely represent the amount of advertising that was obtained for each product during the campaign, with higher numbers representing more intensive exposure. A marketing representative armed with this advertising data and the corresponding sales data should be able to extract actionable insights related to the efficacy of the advertising spend. In this case the user requirement for this data would be to discover the impact the advertising campaign had on sales.

ii) **Detect potential data quality issues**
When processing the campaign source data during ETL, any data quality issues can be detected by configuring and then applying some rules about the expected structure and format of the data. These rules should be applied as processing steps before adding the data into the Data Warehouse.

See the Appendix at the end of this document for analysis carried out on the data that has been extracted into the Source System Tables. Each of the columns in those Source System Tables are documented with information about its datatype, nullability and set of possible values.

During the Transformation Stage of the ETL process, it is possible that several iterations of the transformation will be required, before the data quality specifications have been met. It is also possible that the values in the source system will need to be cleaned in order to conform. If a value does not conform to the specifications defined in the Tables, then the bad records should be logged along with supporting information about the nature of the error. An option for this could be to log the contents of the full bad record to a text file. These files should then be used when providing feedback about data quality issues to the owner of the source system. It is important to rectify these issues as early as possible.

Below are some general rules for the three data types that have been identified for the campaign dataset:

1. No null values allowed
2. For String/VARCHAR Columns,
    a. No misspelt or unknown values. For example, some VARCHAR columns can only contain certain values, these are documented in the Appendix Tables.
3. For Date columns:
    a. Date format is: DD/MM/YYYY
    b. YYYY must equal 2016
    c. Note that 2016 is a leap year

iii) **Detect duplicates**
Detecting duplicates can be facilitated by specifying whether a column value or a combination of column values (as a composite) can be unique or not during the transformation process. Suitable columns for uniqueness are documented in the Appendix Tables. Again, it is important to rectify these duplicate issues as early as possible. If any records do not match the uniqueness constraints, then the record will be logged to a suitable table and feedback will be sent to the source system owners and then corrected.

Combining data
The source data includes pairs of data sheets made up of:

IMPACTS and SALES per Shopping Centre.

For example:

IMPACTS *GRAN JONQUERA SC* and **SALES** *GRAN JONQUERA SC*

One proposal to reduce the number of Sheets in the source system dataset would be to just have one overall **IMPACTS** Sheet and one **SALES** Sheet.
This Sheet would contain all of the IMPACTS and SALES records for all Shopping Centres merged into one sheet.
To facilitate this, a new column, SHOPPING CENTER would be added to the **IMPACTS** Sheet and the **SALES** sheet.
This column would then specify the shopping centre that the IMPACTS and SALES relate to.
There would then be ten fewer sheets in the source system dataset.
The approach for the merge of the IMPACTS and SALES per Shopping Centre data, has been included in the Appendix Tables documentation.
The new column is called SHOPPING_CENTRE.

Data Warehouse Approach

The approach selected for the data set was a star schema. The star schema data warehouse model was chosen for two reasons. The first is that it permitted the highest granularity level possible when considering the data, i.e it did not constrain the data at all. An example of this would be the analysis of sales units per product in a certain city in a particular quarter. Achieving the highest level of granularity enables the marketing manager to make decisions based on detailed analysis, identifying key points in the current campaign and using the resulting information to increase performance in a subsequent campaign. The second reason for choosing the star schema is that it is the simplest model to retrieve data. This efficiency can improve the time and ease with which a manager or an analyst can access the right data for analysis.

Before choosing the star schema model, the snowflake schema and the datavault model were also considered as alternatives. The snowflake schema was not selected because it was not appropriate for the provided dataset. While sales can be analyzed by their different dimensions and its correspondent attributes such as product name, this is the highest granularity level one can obtain when interpreting a fact (sales units) described by an dimension´s attribute (product name). While a fact can be analyzed by its different dimensions and attributes, such as total sales quantity per product name or total impacts per shopping center type, this is the maximum granularity level one can obtain when retrieving data about any of the facts. Taking this example, the snowflake schema would fit the data set if the product name or or shopping center type had different sub-attributes, extending the data warehouse model by creating multiple tables describing previous attributes. Even though there is no minimum number of levels required to build a snowflake, it is clear that any marketing manager using this data would not benefit from this option.

With regards to the datavault option, one reason it was discarded was that it would add unnecessary complexity in terms of accessing and retrieving data, considering the relative simplicity of the dataset. The main reason relates to the purpose of the data set. The data describes the results of a single advertising campaign. Since this is a unique event in time, this data warehouse model is not likely to need to evolve over time in order to include new facts and dimensions beyond those already in the model. A model for a growing company, for example, might benefit from the flexibility afforded by a data vault, but a single advertising campaign is unlikely to see much evolution.

The ability to respond to all types of questions by using the designed star schema thus confirms the appropriateness of this model, when considering the dataset.

Dimensional Design Process

For the dimensional design process there are four key steps to follow: identify the Business Process, Grain, Dimensions and Facts.

Starting with the business process, an inspection of the data reveals that the business process is as follows:

Visits ➔ Impacts ➔ Sales

The above model represents the logic of how the business is oriented. Visits to the shopping centre represent the total potential customers that can be targeted by all advertisements on that day. Advertisements are displayed at the impact points (the big screen and information impact point), and in theory these advertising ‘impacts’ could then cause the influenced customers to proceed to the advertised shops and purchase the represented products.

The next step is to identify the grain of the model that is designed, this is the level of detail the fact table is going to have. Inside the fact table is where the grain of the process is revealed, as it is shown in sales, marketing and visits.

- There is one row per sale unit per product per shopping centre per date.
- There is one row per Information Point Screen impact per product per shopping centre per date.
- There is one row for Big Screen impact per product per shopping centre per date.
- There is one row for visits per shopping centre per date.

Next the dimensions of the fact tables need to be identified. Looking at the data, a dimension for the date is needed because the campaign has a start and an end, while the sales, marketing and visits are analyzed by day of the year. There is also a need for a dimension related to the shopping centre, since each sale and impact is registered by the location at which it occurred. In addition a dimension is needed for the products that are being advertised and sold, because each has its own name and family type.

As the last stage in the process, all the facts are identified by answering the question ‘What is the business process measuring?’. As discussed above, in the business process visits enable impacts which lead to sales, all of which are measured on a date-related basis; as a result visits are considered a fact, while for the second fact sales and impacts are placed together because they have the same granularity. These are the columns that ultimately make up the two fact tables.

Appendix

Data Types Table:

Datatype Name	Description
DATE	A date consisting of a day, month and year.
INTEGER	Positive and negative whole numbers
VARCHAR	A String of Alphanumeric characters up to 45 characters long.

Sheet: POINTS
All of the columns must be used to identify unique records.

Column Name:	Column DataType	Nullable (Y/N)	Possible Values
ZONE	VARCHAR	N	‘NORTH’ or ‘CENTER’’
CITY	VARCHAR	N	Figueres or Girona or Barcelona or Lleida or Tarragona or Reus
SHOPPING CENTER	VARCHAR	N	Gran Jonquera Outlet & Shopping or Girocentre Or La Maquinista Or Mercat del Pla Or Les Gabarres Or La Fira
TYPE	VARCHAR	N	Periphery or Urban
IMPACT POINT	VARCHAR	N	‘Big Screen’ OR ‘Information Point Screen’

Sheet PRODUCTS
All of the columns must be used to identify unique records.

Column Name:	Column DataType	Nullable (Y/N)	Possible Values
FAMILY	VARCHAR	N	Food Electronics Clothing Sports Home
PRODUCT	VARCHAR	N	PIZZA TDRL PIZZA DROE PIZZA CUSN TV SMRT42SNG TV SMRT42PHI TV SMRT42SNY TROUSERS LVS TROUSERS MNG TROUSERS LEE SHOES NKE SHOES ADS SHOES ACS FRAME 18X10 CK FRAME 18X10 CH FRAME 18X10 MD

Sheet: VISITS
The date column can be used to identify unique records

Column Name:	Column DataType	Nullable (Y/N)	Possible Values
DATE	DATE	N	01/01/2016 -> 31/12/2016
Gran Jonquera Outlet & Shopping	INTEGER	N	All Integers >= 0
Girocentre	INTEGER	N	All Integers >= 0
La Maquinista	INTEGER	N	All Integers >= 0
Mercat del Pla	INTEGER	N	All Integers >= 0
Les Gabarres	INTEGER	N	All Integers >= 0
La Fira	INTEGER	N	All Integers >= 0

Sheet: IMPACTS
Columns DATE, IMPACT POINT, SHOPPING CENTRE can be used to identify unique records.

Column Name:	**Column DataType**	**Nullable (Y/N)**	**Possible Values**
DATE	Date	N	01/01/2016 -> 31/12/2016
IMPACT_POINT	VARCHAR	N	‘Big Screen’ OR ‘Information Point Screen’
PIZZA TDRL	INTEGER	N	1 -> 12
PIZZA DROE	INTEGER	N	1 -> 12
PIZZA CUSN	INTEGER	N	1 -> 12
TV SMRT42SNG	INTEGER	N	1 -> 12
TV SMRT42PHI	INTEGER	N	1 -> 12
TV SMRT42SNY	INTEGER	N	1 -> 12
TROUSERS LVS	INTEGER	N	1 -> 12
TROUSERS MNG	INTEGER	N	1 -> 12
TROUSERS LEE	INTEGER	N	1 -> 12
SHOES NKE	INTEGER	N	1 -> 12
SHOES ADS	INTEGER	N	1 -> 12
SHOES ACS	INTEGER	N	1 -> 12
FRAME 18X10 CK	INTEGER	N	1 -> 12
FRAME 18X10 CH	INTEGER	N	1 -> 12
FRAME 18X10 MD	INTEGER	N	1 -> 12

SHOPPING_CENTRE	VARCHAR	N	Gran Jonquera Outlet & Shopping or Girocentre Or La Maquinista Or Mercat del Pla Or Les Gabarres Or La Fira

Sheet: SALES
Columns DATE and SHOPPING CENTRE can be used to identify unique records.

Column Name:	**Column DataType**	**Nullable (Y/N)**	**Possible Values**
DATE	Date	N	01/01/2016 -> 31/12/2016
PIZZA TDRL	INTEGER	N	All Integers
PIZZA DROE	INTEGER	N	All Integers
PIZZA CUSN	INTEGER	N	All Integers
TV SMRT42SNG	INTEGER	N	All Integers
TV SMRT42PHI	INTEGER	N	All Integers
TV SMRT42SNY	INTEGER	N	All Integers
TROUSERS LVS	INTEGER	N	All Integers
TROUSERS MNG	INTEGER	N	All Integers
TROUSERS LEE	INTEGER	N	All Integers
SHOES NKE	INTEGER	N	All Integers
SHOES ADS	INTEGER	N	All Integers
SHOES ACS	INTEGER	N	All Integers
FRAME 18X10 CK	INTEGER	N	All Integers

FRAME 18X10 CH	INTEGER	N	All Integers
FRAME 18X10 MD	INTEGER	N	All Integers
SHOPPING_CENTRE	VARCHAR	N	Gran Jonquera Outlet & Shopping or Girocentre Or La Maquinista Or Mercat del Pla Or Les Gabarres Or La Fira