LUCA BREGATA – ENGINEERING & MANAGEMENT

Creation of a Data Mart evaluate for the closing reasons and the best locations for a geo-fashion retail store

The term Business Intelligence (BI) refers to a series of business processes that revolve around the data, with the collection, processing and the analysis, whose purpose is to produce information to the strategic and tactical management service, which finds support analytical, historical to forecast an efficient Data Driven Strategy. The BI was also placed in the operating subset, as it is playing an increasingly important role in the normal daily activities of the companies.

The Data Warehouse (DWH) is the main business intelligence support tool. They allow you to collect integrated, consistent and certificates data related to all business processes of a company from the operational sources. These data are suitably processed through ETL procedures and controlled through the data quality system.

A Data Mart is an analytical database designed to meet the specific needs of a business. Being logical or physical subset of a data warehouse, larger in size, it follows the same design rules but with aggregated data at various levels of detail, although it may sometimes also be formed in the absence of an integrated data system. Their distribution is based on the type of business that each customer wants to deal with.

The implementation of a Data Mart is often divided into different levels:

SCHEME	L0	L1	L2	L2_STAR SCHEME
Definition	Extraction of data from various types of files without transformation.	Major changes and data quality operations.	Very fast and slender tables. use of surrogate key to connect the various attributes.	Few tables but full-bodied, to have all the necessary fields data for reports.
Area	Staging Area.	Operational Data Store	Presentation Area.	
Primary key	NO.	YES.	YES.	YES.
Surrogate key	NO.	YES.	YES.	NO.
Action on Table	Truncate table but maintain scheme.	Nothing.	Nothing.	Nothing.
Action on data	Insert.	Update / Insert.	Update / Insert.	Update / Insert.

The multidimensional model or DFM (Dimensional Fact Model) is a conceptual model where is possible represent data within a Hypercube whose edges represent the dimensions of analysis, which subsequently will be divided into many "cubes", each identified by a triple of coordinates. Each cube ideally contains the values assumed by the measures for that data triad and is

commonly referred to as "Fact" because it represents the occurrence of an event of interest for the business domain.

A multidimensional model is mainly based on four key concepts:

- Fact: Table that typically models a specific business area (Sales, Orders, Production, etc.) and is characterized by a more measure;
- Measurement: It is the quantitative aspect of the fact and it is of high importance for the analysis. From measures are extracted the KPIs (Key Performance Indicator) that will guide enterprises in their business strategies. Some examples can be the Quantity produced, the profit, and price;
- Dimension: It represents the coordinates of the analysis Done. Among these we can find Date, Product, Shop;
- Dimensional Attribute: It is a logical grouping of some elements of a same size. Are classes of elements that allow the user to select the data for specific characteristics.

Once built the Data Fact Model, the logic diagram must be implemented. It is represented according to a Star Schema, which the center is constituted by a fact table; the points of the star represent instead the dimension tables that branch out from the center. The main features of a Star Scheme are as follows:

- simple structure & easy to understand;
- High performing queries, because they reduce the joins to be made between tables;
- Loading time of the relatively long data, because the data redundancy due to the de-normalization, causes an increase in size of the table;
- Widely supported by a large number of business intelligence tools;
- The fact tables in a star schema is in third normal form, while the dimensional tables are de-normalized.

A database is in 3NF (third normal form) whether all non-key attributes depend on one and only one key, i.e. there are no non-key attributes that depend on other non-key attributes. This normalization eliminates the transitive dependency of attributes from the key and is called SnowFlake scheme.

- Normalized tables;
- Slower than run queries Star Schema;
- Creation of integer surrogate keys (SK).
- Minimal code changes.

To carry out a complete analysis and obtain a reliable result starting from ISTAT data, highlighted and processed via the data quality process, mainly three queries were necessary, all extremely connected to each other, with the ultimate goal of finding which, among the thousands the cities, are the most economically desirable to open a new store (best geo-location).

Initially, the query will be on regional indicators taken from the website of ISTAT (average monthly household spending on non-food goods and services, average income, network operating station, unemployment rate, GDP Per Capita). I create a sum of various factors rank grouped by region and sorted by rank descending or ascending based on the type of indicator. In this way, you will get as a result an order of development potential of Italian regions. On it, you add up your corporate data for the historian of the number of shops closed or still open for each region. The final value will take the name of Rank. The three best regions to invest in an economic capital results Trentino Alto Adige, Valle D'Aosta and Friuli Venezia Giulia.

The last step to be carried out to find the ideal location to open a new store, is to select the municipalities in the three regions previously chosen as the most attractive and go in deep to analyzing the data concerning the number of tourists reported last year and the number of residents for each city. The result shows that Lignano Sabbiedoro, Trieste and Trento are the best three cities to set up a new shop in 2019.

The classification algorithm CART: is a nonparametric procedure that builds a decision tree in order to label an attribute; In fact, the classification term refers to a process, given a collection of records called Training Set, try to build a model able to attribute a feature called Class attribute, based on the combination of other properties that characterize the specific population . Once you have the model, it can be used to predict the class of new records for instances where the class is unknown (Test Set).

The important steps to be followed when you want a decision tree with the CART procedure are mainly two: adopt a criterion of the technical skill with which the nodes are divided from parent nodes to child nodes (split criterion) and establish a stopping rule of tree growth (stopping rule).

In the implemented project, we will use the CART process to define and predict which stores will continue to exercise and what stores will close in 2019, starting with a training set with the data refers to the shops that closed in 2018 based in 2017 data. So, the entire database have an horizon start from 1/1/2017 to 31/3/2019.

These data will be further divided into training and test set thanks to a random 80/20 partition on 100% of the analyzed elements, where you will create your models from training data set to test it on a later test data.
Once performed the test operation, will be verified its inequality distribution via the Gini index, where 0 represents perfect equality, while an index of 100 implies perfect inequality and in addition, the accuracy will be studied through the media.
Optimized most of the template, the database data to analyze and predict the pattern created were introduced, resulting in the forecast. The model was further evaluated by the Gini index [0.1705464] and accuracy [0.8701299], and finally, compared with the initial results, in order to understand the true efficiency. The result led to a prediction of 5 at risk of closure shops on a total of 29.

The data visualization is a generic term describing any attempt to help people understand the meaning of the data analyzed by placing them in a visual context. Patterns, trends and correlations that may not be detected in the text-based data can be exposed and recognized more easily by using the data visualization software such as Microsoft Power BI.

The document produced is called report. In the first example presented, are show the results obtained thanks to an analysis of the shops of the client's project analyzed closed in Italy from 2017, identifying the cause of the closure of them, classified as a closure for a bad profit margin, or as a closure for a covered market in the years or the birth of a new store in the neighborhood. The result is visible through the underlying map that encloses examples for each type of closure. The Serravalle Scrivia shops show a closure of new shop type, while the shops of Padova and Modena a margin not adequate.

The additional images show an example in Power BI of what is meant by a dashboard. The first includes a general analysis of the sales channels associated with the related earnings of products and products. It will be the default view for the client. For each dashboard, the first graph (BarChart) includes the display of the total turnover of 2019 referring to the channel of the store, divided into Outlet and Full Price. The second graphic is the representation of Pareto 80/20, which shows that most sales derive from the stock exchanges. The last graph, instead, shows the percentage of the total origin and the total turnover for each product.

Through the use of filters or simple navigation it is possible to pass from a very general analysis to a much more detailed analysis in each individual report. The initial image explains the variations of the products sold and the made in thanks to a selection of the type of store that you want to consider. The last image, however, use the same idea, but the selection is made for the bags of the product.

It is very important to observe how the graphs interact with each other, leading to a rapid and effective analysis, chosen according to the customer's needs.