POLITECNICO DI TORINO

Corso di Laurea in Engineering & Management

Master's Degree Thesis

Development of a data mart to support decisions in fashion retail store localization

Supervisor

Prof. Marco Cantamessa

Dott. Vincenzo Scinicariello

Candidate

Luca Bregata

Accademic Year 2018/2019

POLITECNICO DI TORINO

Corso di Laurea in Engineering & Management

Master's Degree Thesis

Development of a data mart to support decisions in fashion retail store localization

Supervisor

Prof. Marco Cantamessa

Dott. Vincenzo Scinicariello

Candidate

Luca Bregata

Accademic Year 2018/2019

Alla Mia Famiglia e
alla mia ragazza
per il supporto e l'aiuto
ricevuto durante
questo viaggio.

SUMMARY

INDEX OF FIGURES

INDEX OF TABLES

ABSTRACT

The evolution of business intelligence began decades ago with the first report mainframe, called the system output. They were mainly printed on paper, to then be distributed periodically to the manager. The first queries have sped up the process and made it possible for managers with technical expertise to create customized ad hoc reports, but few managers had the time and the skills to do so. The emergence of the data warehouse has given a great impetus to the BI aggregating all the data in one place, where he could be interrogated interactively without impacting applications with online queries and reports with increasingly easy graphical interfaces use.

The advent of the data warehouse, the data marts and analytic analysis tools have made BI accessible to more operators and allowed managers to obtain information and critical responses efficiently and quickly.

The proposed project will be dedicated to the detailed description of the creation of a data mart dedicated to the sales of the fashion company through an optimal solution of best practices of an ETL process resulting in the Snowflake pattern and the Star scheme, perfect for the date. In addition, using the classification process including both corporate open data, you had the possibility of locating the most effective area to open a new store and to offer an explanation as to why some shops were closed in the recent past.

In conclusions, which will be displayed in Power BI, Microsoft software for data visualization.

INTRODUCTION

As we become a digital society, the amount of data created and collected grows and accelerates significantly. The analysis of this data becomes a challenge for traditional analytical tools that are increasingly more difficult to keep up. It is therefore necessary a constant innovation to bridge the gap between the data generated and the data that can be analyzed in an effective manner.

Large tools and technologies represent opportunities and challenges in the study of profitable way to better understand customer preferences, gain a competitive advantage in the marketplace and make their business grow.

The data management architectures have evolved from traditional data warehousing model to more complex architectures that meet different requirements, such as real-time and batch processing, structured and unstructured data, high-speed transactions, etc.

Mediamente Consulting s.r.l., founded in 2012, it is a consulting firm specializing in the design decision support systems. In particular, it deals with business intelligence projects, data warehouse and advanced analytics in the field of big data.

Thanks to these systems, the customer could view the information and consequently make a more informed decision based on facts. Therefore, Mediamente Consulting supports the customer in the knowledge of their performance and helps to increase them through better decisions.

For the development of my thesis project, I have been part of a team that follows a major fashion company that meets the analytical needs in the field of fashion.

The customer is responsible for managing of various sectors in the field of "Fashion Retail". It is a multinational holding company responsible for sales of products around the world that is developing an innovative form of the strategic and operational analysis which allows to fully seize the growth potential of its brand, within a very competitive global market, where the market segment is in continuously substantial growth.

Below I'm going to present briefly in the chapters that make up my paper. The illustration of the entire work was possible thanks to the help of basic theoretical concepts and the help provided by my colleagues in the workplace.

The first chapter can be seen as an introduction to the fundamental concepts of Big Data and Business Intelligence in general, representing the state of art. In addition, I will explain the mechanisms used and the key concepts of data mining and machine learning.

The central part of my thesis, which corresponds to the chapters 2-3, will be devoted to a description of the analysis of the methodology used and from where I started to develop my thesis project. It will be explained in detail the corporate ETL process resulting in Star Schema, ideal for data visualization, and resulting in Snowflake schema, ideal for the structure of a process Extraction, Transformation & Loading. Using the classification process, including the business data and the open data from ISTAT, I have had the opportunity to locate the most effective area to open a new store and to offer an explanation as to why some shops closed.

In this way we will have an overview of the concepts to do an efficient job.

Finally, in the last chapter, I will going to outline in detail the data obtained through the use of Power BI application on the context in which it is done the project, observing and analyzing the characteristics and choices implemented thus achieving a strategic decision and vision that the company may decide to implement or not.

This chapter is the most important, as will be described the results through reports of the implementation and methodology of the final project, that will help the managers to take a good decision make in term of competitive advantages about competitors.

METHODOLOGY

This section will explain the main stages of the project:

1. **Top-Down Analysis** to understand and analyze fully and effectively all relevant features of the project, asking questions and trying to figure out how to get answers from them, and then get a basic data that will be used to achieve the ultimate goal;
2. **Building the Data Mart** using Talend Open Source software used for Traditional ETL, highlighting all business processes from the staging area of the warehouse, where you will import the source files (CSV, Excel ...);
3. **Development of a Data Presentation Best Practice Model and the Snowflake schema**, obtained by the relations between the surrogate keys of the tables, optimizing performance;
4. **Development of a Presentation Area Best Practice Model and the Star Schema**, Both efficient in the representation of various useful reports for the final customer business decisions;
5. **Development of a module to speed the updating process** tables, grouping the jobs belonging to the same family (Anagrafiche and Movimenti) so you run that at the same time. Thanks to an Error warning system, we have undergone an audit of the problem resulting in sending Email;
6. **Using software from Data Mining and Machine Learning,** extremely linked to the concept of artificial intelligence, independently implemented different algorithms that can combining the data warehouse data and the Italian ISTAT data previously extracted perform in short time difficult analysis in with streamlined process;
7. **Development of a Dashboard Power BI** to have a check on a series of real data on everything regarding the stores and the economic data of the firm. Through this process, you can extract results to be provided to customers in a clear and comprehensive manner, helping them to understand the problems of their production systems and trade and improving their Data Driven Strategy, identified as the choice of effective and new management actions.

TOP-DOWN ANALYSIS OF THE PROBLEM

Attributes successful business in fashion:

- Strategic position;
- Quality price;
- Product Diversification;
- Innovation;

Business goal:

- Making as much as possible by having the minimum inventory cost;
- Increasing Sales.

Competitors:

Fashion Companies who use the same type of market.

Data Mining - Classification and Regression using algorithms that use artificial intelligence:

- Country;
- Type of shop (Retail, Outlet, ...);
- Cause of a closure;
- It will remain open?
- Ranking for geo-positioning to open a new store.

Questions

- Better an elegant or extravagant showcase to attract customers?
- When I have to move the merchandise from a retail store to an outlet?
- How can I make my reassortment?
- How many times my customers come back to buy? How many times they come back to buy it through e-commerce?
- That risk can afford this store? How much merchandise can I leaving unsold?
- What particularly distinguishes me and makes me immediately recognize from competitors?

- Are better open niche stores in big cities or open stores in the outlets in strategic locations?
- When should I create a promotion and for whom? Better short or long term? Should I make deals on a single product or based upon association rules? When my number of sales is below average?
- How can I grow / shrink my brand compared to the sector?
- How can I overtake my competitors?
- Should I focus more on a purely female or male clients?
- As a period of balances has an effect on the number of sales? On the company's overall profit?

Provide the information necessary for analysis

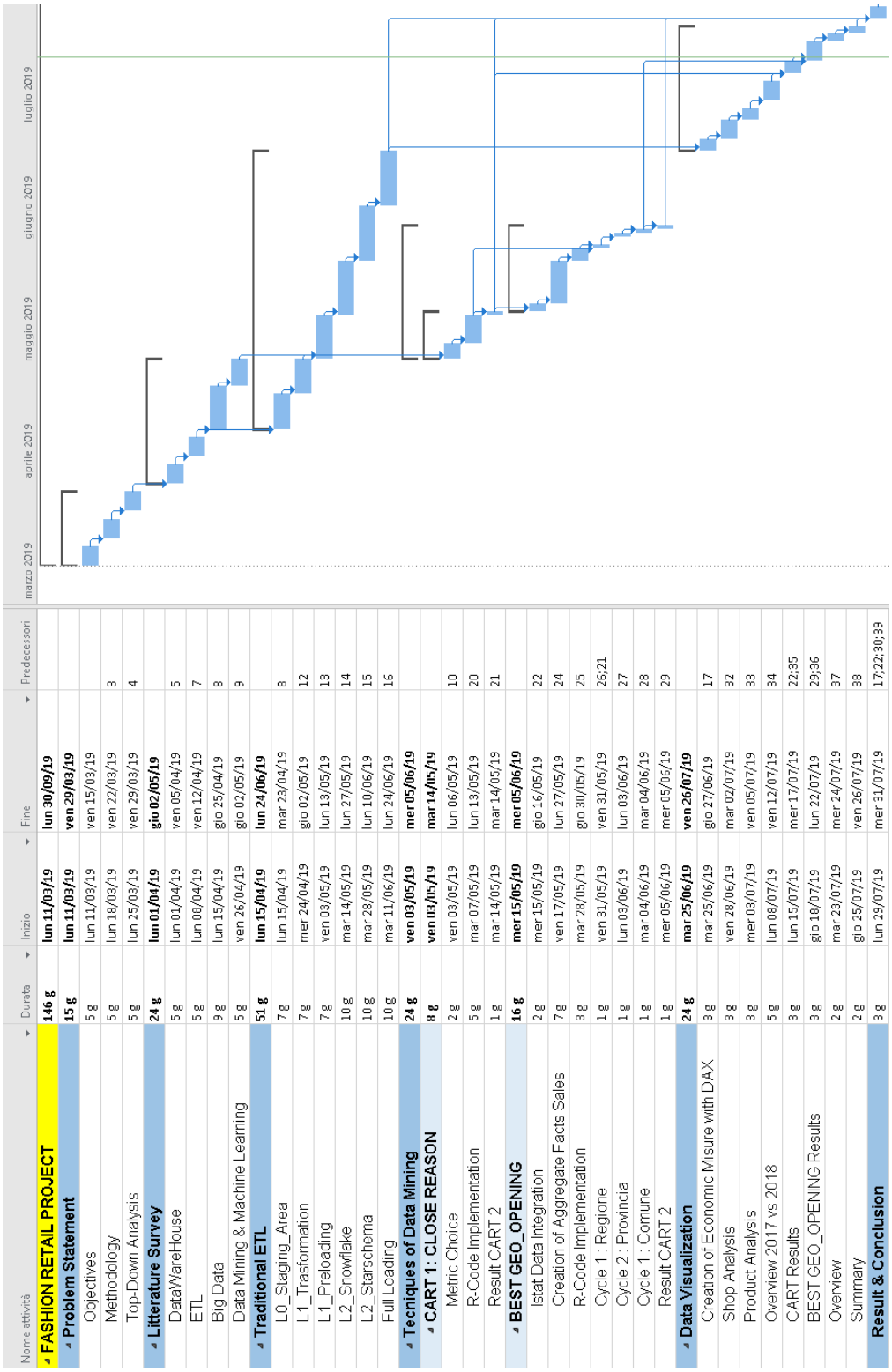
- Product (id, description, category, release date, season, characteristics, gender);
- Stores (id, geographic location, description, type / channel);
- Receipt (id, id_data, product, customer_id, value, discount, quantity);
- Date (id, day, month, quarter, semester, year);
- Open data: budget, ISTAT data about population, tourism, economic indicators.

TARGET

The world of fashion has been extensively studied to understanding how leaders influence their followers by providing valuable information to decision makers on the product marketing strategies. The aim of our study is to implement through the ETL process, the best practices for creating a data warehouse or a data mart for a customer, using the data obtained for make a business analysis on sales, with clear and effective results.

In particular, the project will focus on an in-depth analysis on the Italian territory through the use of algorithms of machine learning and data mining, mainly obtain the causes of closing of the shops them and the best geographical area for a new opening.

PROJECT SCHEDULING



Nome attività	Durata	Inizio	Fine	Predecessori
FASHION RETAIL PROJECT	**146 g**	**lun 11/03/19**	**lun 30/09/19**	
Problem Statement	**15 g**	**lun 11/03/19**	**ven 29/03/19**	
Objectives	5 g	lun 11/03/19	ven 15/03/19	
Methodology	5 g	lun 18/03/19	ven 22/03/19	3
Top-Down Analysis	5 g	lun 25/03/19	ven 29/03/19	4
Litterature Survey	**24 g**	**lun 01/04/19**	**gio 02/05/19**	
DataWareHouse	5 g	lun 01/04/19	ven 05/04/19	5
ETL	5 g	lun 08/04/19	ven 12/04/19	7
Big Data	9 g	lun 15/04/19	gio 25/04/19	8
Data Mining & Machine Learning	5 g	ven 26/04/19	gio 02/05/19	9
Traditional ETL	**51 g**	**lun 15/04/19**	**lun 24/06/19**	
L0_Staging_Area	7 g	lun 15/04/19	mar 23/04/19	8
L1_Trasformation	7 g	mer 24/04/19	gio 02/05/19	12
L1_Preloading	7 g	ven 03/05/19	lun 13/05/19	13
L2_Snowflake	10 g	mar 14/05/19	lun 27/05/19	14
L2_Starschema	10 g	mar 28/05/19	lun 10/06/19	15
Full Loading	10 g	mar 11/06/19	lun 24/06/19	16
Tecniques of Data Mining	**24 g**	**ven 03/05/19**	**mer 05/06/19**	
CART 1: CLOSE REASON	**8 g**	**ven 03/05/19**	**mar 14/05/19**	
Metric Choice	2 g	ven 03/05/19	lun 06/05/19	10
R-Code Implementation	5 g	mar 07/05/19	lun 13/05/19	20
Result CART 2	1 g	mar 14/05/19	mar 14/05/19	21
BEST GEO_OPENING	**16 g**	**mer 15/05/19**	**mer 05/06/19**	
Istat Data Integration	2 g	mer 15/05/19	gio 16/05/19	22
Creation of Aggregate Facts Sales	7 g	ven 17/05/19	lun 27/05/19	24
R-Code Implementation	3 g	mar 28/05/19	gio 30/05/19	25
Cycle 1: Regione	1 g	ven 31/05/19	ven 31/05/19	26;21
Cycle 2 : Provincia	1 g	lun 03/06/19	lun 03/06/19	27
Cycle 1 : Comune	1 g	mar 04/06/19	mar 04/06/19	28
Result CART 2	1 g	mer 05/06/19	mer 05/06/19	29
Data Visualization	**24 g**	**mar 25/06/19**	**ven 26/07/19**	
Creation of Economic Misure with DAX	3 g	mar 25/06/19	gio 27/06/19	17
Shop Analysis	3 g	ven 28/06/19	mar 02/07/19	32
Product Analysis	3 g	mer 03/07/19	ven 05/07/19	33
Overview 2017 vs 2018	5 g	lun 08/07/19	ven 12/07/19	34
CART Results	3 g	lun 15/07/19	mer 17/07/19	22;35
BEST GEO_OPENING Results	3 g	gio 18/07/19	lun 22/07/19	29;36
Overview	2 g	mar 23/07/19	mer 24/07/19	37
Summary	2 g	gio 25/07/19	ven 26/07/19	38
Result & Conclusion	3 g	lun 29/07/19	mer 31/07/19	17;22;30;39

Figure 1: Project Scheduling

CHAPTER 1: STATE OF THE ART

1.1 BUSINESS INTELLIGENCE

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.

In modern business, whose main purpose is to become market leaders, companies are daily compare with realities different from their own. This is accomplished through the analysis of the behavior of competitors belonging to the same sector and studying the market in which they are located.

The BI adoption by firms allows a deeper understanding not only of themselves but also of the market.

In the current period, the "change" is on the agenda, thus being able to read in advance trends of markets can give some huge competitive factors with respect to competitors.

Given the high volume of data generated every day, becomes necessary to find a method that:

- It collects and processes data at high speed (mores often it comes to real-time processes);
- It provides a cleaning service of the data, eliminating dirty data, redundant or incorrect via the ETL processes "Extraction, Transformation & Loading" which pick up the data from the input systems (ERP, Excel spreadsheets, etc.) and carry them in a data warehouse through data quality processes.

This process will be explained in a specific manner in section 2 but in summary, it defines a consolidated and stable system storage for certificates data (data warehouse) and it turn information into a source of knowledge through business analysis on the data, determining new KPI.

The Big Data originate from different sources, both internal and external, they are often including different formats and reside in multiple positions of a legacy system or in other applications. The data can be structured (data stored in relational database, arranged in patterns and rigid tables), unstructured (stored data without any schema as free-form text such as articles and e-mail parts, audio without tags, images and video) or semi-structured (data that contain features of both of those structured that those unstructured; an example is represented by the compiled files with XML syntax for which there are there are no structural limits the insertion of data, but the information is organized according to logical structured).

After that, the data must merge.

The next step is choosing the platform and the technology to be used for big data analytics applications including queries, reports, OLAP system, data mining and visualization, including in all these applications [22].

A central role is played by big data analytics, and by the business intelligence-based technologies such as:

- *CRM & Customer Analytics*: Solutions and technologies that collect, organize and synthesize customer data to help organizations solve business problems related through tools, dashboards, portals and other methods in the areas of Marketing, Sales and Customer Service; Consumers are segmented into groups based on the adopted behaviors, actions to implement customized marketing and general trends;
- *Predictive Analytics*: Advanced Analytics that implement techniques such as regression analysis, predictive models and statistics to analyze data and contents, and answer questions like "What will happen" or "What will most likely happen?";
- *Social Analytics*: Tools that automatically extract, analyze and summarize the content generated by online users;
- *Text Analytics*: Process of extracting information from texts, used for including the summary, finding key content in a large set of data, sentiment analysis or to

determine what drove a particular comment of a person, so, for an explanatory purpose;

- *Web Analytics*: analytical applications used to understand and to improve the online consumer experience, the acquisition of users and the optimization of digital marketing and advertising campaigns. These offer reporting, segmentation, advertising management and integration with other data sources and processes;
- *Prescriptive analytics*: Use optimization technology to solve complex decisions with a very large number of decision variables, constraints and compromises for providing optimal actions to achieve the business objectives.

1.2 DATAWAREHOUSE

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.

Data quality is a critical requirement for the entire information system, because, if the data are dirty, can not only cause a worsening of business performance, but can also lead to take inappropriate decisions, resulting in additional costs and lost opportunities.

The goal of a data warehouse is therefore to support the "Knowledge Worker" (officer, director, manager, analyst) helping him to analyze the data aimed at implementing decision-making and improvement of information assets, to provide a single point of access to all company's data made in a consistent and reliable way through the ETL processes. The data warehouse also ensures a complete historical depth of the data, thus allowing temporal analysis.

A DWH must be carefully designed to have an efficient and effectively manage of the Big Data characteristics.

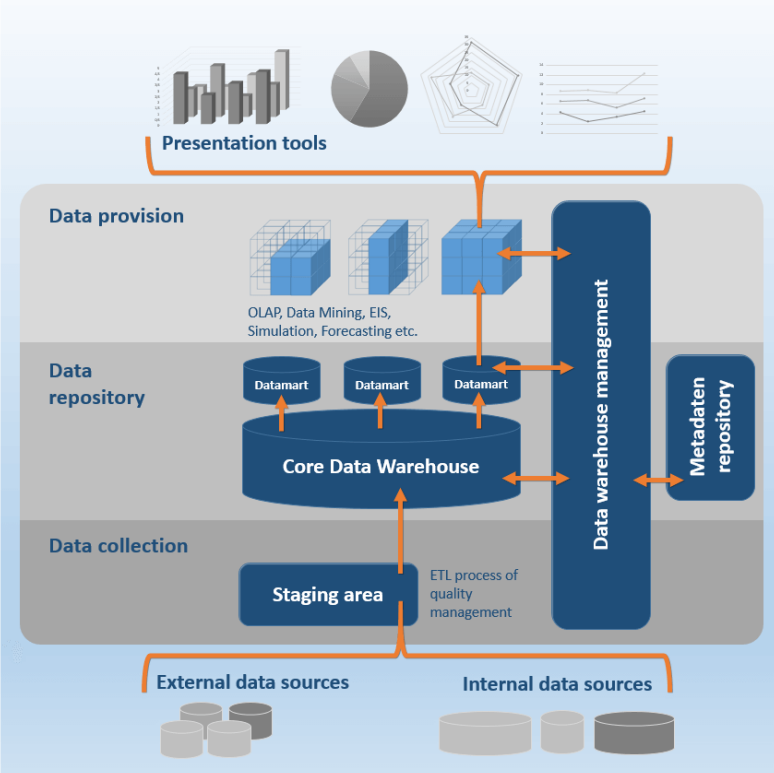


Figure 2: Data warehouse

The Datawarehouse are made as the main tool for the Decision Support System (DSS), that is a system capable of providing clear information to users so they can analyze in detail a situation and take the appropriate decisions on actions to be taken in a easily and quickly way [12].

The DSS relies on data from one or more databases, often organized in different structures with non-homogeneous data.

A system of this type must support the analysis and control of management routines, the research of the causes of a problem (focused search) and complex managerial activities (decision making), besides an easy using to a user with a reduction on time and an improvement towards new technologies (especially in cases in which cannot perceive the benefits in a short time).

Let us describe in detail the features:

- *Subject oriented*: The data warehouse is organized for relevant subjects such as, for example, products, customers, suppliers and the time period, in order to provide all information pertaining to a specific area;
- *Integrated*: The data warehouse must be able to integrate seamlessly with the multitude of standards used in different applications. The data must be re-encoded, in order to be homogeneous from the semantic point of view, and must use the same units of measurement;
- *Variable time*: Unlike the operational data, those of a data warehouse have a very broad time horizon (even 5-10 years), making them reusable in different time instants;
- *Non-volatile*: The operational data is updated continuously; in the data warehouse data are loaded initially with integral processes and subsequently updated with partial loads; data, once loaded, are not modified and retain their integrity over time.

It is possible that a data warehouse is divided into different data marts, each specific to a single business process among those inside the company (orders, sales, customers, marketing, etc.). In Chapter 2 we will see a data mart related to the sales of a fashion retailer.

1.2.1 Data Warehouse Architecture

In the design phase it is essential to determine which types of architecture adopted. Clearly, a DWH must be constructed in accordance with modern principles [10] and the patterns described in this paragraph are still of the bases:

***Model Inmon - Corporate Information Factory**:* The Datawarehouse are constructed in their entirety from the beginning as a single monolithic block; you cannot see how the composition of the DM. It adopted a top-down view.

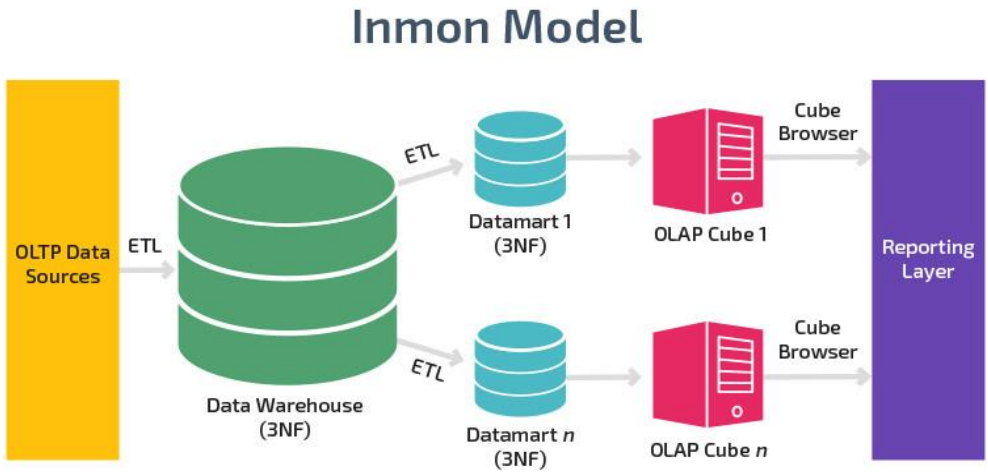


Figure 3: Inmon's Model

***Model Kimball - Dimensional Model**:* Adopts a bottom-up approach in which the data warehouse born from the union of the various data marts that each refer to a specific business area.

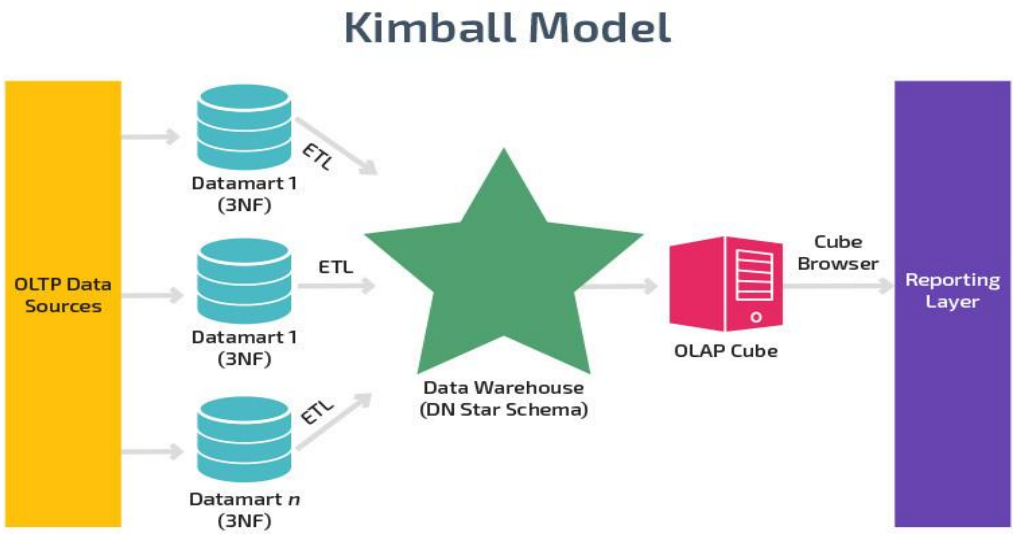


Figure 4: Kimball's Model

It has been shown that the Inmon and Kimball approaches work to successfully deliver data warehouses. But there are excellent organizations where there is a combination of both, in a hybrid model: the data warehouse is created using the Inmon model and the business data mart processes are created using the star schema has been implemented for report creation. We cannot generalize and say that one approach is better than the other; both have their advantages and disadvantages, and both work well in different scenarios. For any approach to be successful, it must be carefully studied and discussed in detail and designed to meet the reporting needs of the organization's BI and should also integrate with the organization's culture.

So, the architect has to select an approach for data warehouse based on several factors;

1.2.2 Extraction, Transformation and Loading (ETL)

The role of ETL tools is to feed a single, detailed and comprehensive high-quality data source that it can in its power once the data warehouse. The operations to be carried out they are often referred to the term reconciliation that, during the data warehouse feeding process is carried out on two occasions: when the DW is populated for the first time and when it is periodically updated. The reconciliation consists of four distinct said processes respectively:

- Extraction or Capture;
- Cleaning or Scrubbing;
- Transformation;
- Loading.

In general, the boundary between cleaning and transformation is quite hazy therefore, for simplicity, we assume that the cleaning operation is essentially aimed at the correction of the data values, while the transformation is concerned more properly of their format.

1.2.2.1 Extraction

The Data Integration consists of extraction and cleanup.
During the first step, the relevant data are extracted from different sources and this operation can be of type:

- *Static:* It is carried out when the DW must be populated for the first time and consists conceptually in a copy of operational data;

- *Incremental:* It is used for the periodic updating of the DW and captures only the changes that have occurred in the data from the last extraction. The basic idea is to use the changes registered at the data level to update the data warehouse. The benefits derivable are the very small volume of data involved from time to time in the operation than extracting static, and that most of the data in the data warehouse remains unchanged and only the data that have changed are analyzed. The Technique that are used is the CDC (Change Data Capture) that allow you to monitor data sources with the goal of identifying the changes at the data level. This technique is particularly important for the maintenance of the data warehouse thanks to the propagation of the changes detected at the level of the source.

The Cleanup, however, is the phase that is concerned to improve the quality of data, going to eliminate "dirty" data due to duplication, inconsistencies, missing data, incorrect values etc.
The main functions of data cleansing found in ETL tools are the correction and homogenization thank to the using of appropriate dictionaries to correct specific errors, recognize synonyms, and cleaning based on regular expressions, which apply their domain rules to establish the correct matches between values.

1.2.2.2 Transformation

It is the central phase of the reconciliation process and has the aim to convert the data from the operational source format to the data warehouse format. Among the features of this level for feeding the reconciled data level you have:

- *Converting and Normalizing*: Operating both in storage size level and at the level of measurement units in order to standardize the data;
- *Matching*: Establishing correspondences between equivalent fields in different sources;
- *Selection*: Reducing, if necessary, the number of fields and records with respect to the sources.

In the phase of feeding data warehouse, we have two substantial differences: The normalization is replaced by the denormalization and the aggregation are introduced, that achieves the appropriate summary of the data.

1.2.2.3 Loading

In this phase the data are uploading to the data warehouse through two alternative methods:

- *Refresh:* the data is rewritten integrally replacing the previous ones. In general, this technique is used only during the initial phase of populating;
- *Update:* They are added to the data warehouse only the changes on the data without overwriting the entire boundary. This technique is used in combination to the incremental extraction for regular updates.

One way to reduce the load time is to parallelize the ETL process. This can occur in two ways: more steps performed in parallel or a single passage running in parallel.

- *Multiple Load Steps*. The ETL workflow is divided into several independent papers loading together. The main objective is the creation of independent jobs to create a process much safer to handle errors;
- *Pipeline*. The database can identify certain tasks that can execute in parallel. For example, the creation of an index may be generally parallel through all available process on the machine.

1.2.2.4 Possible ETL Process Problems

When the ETL system operating, some faults can occur for many reasons.

The common causes of ETL failures include:

- Network errors;
- Database errors;
- Disk errors;
- Memory errors;
- Errors in data quality;
- System updates without notice.

To protect yourself from these failures, you need a solid backup and a restart and restore system. You must plan for fatal errors when loading it happen. The system should anticipate this and provide the recovery of the capabilities, stopping and restarting the crash.

For example, for a loading process should engage relatively small set of records every time to keep track of what has been committed. The size of the set should be adjusted based on the size of the transactions and performance implications of different DBMS.

The recovery and restart system are used, of course, to take a job that is entered in error, stopped or reported it to recover it through the entire backup or a simple restart. This system is significantly dependent on the backup system capacity. When an error occurs, the instinctive initial reaction is groped to save whatever has been processed and restart the process from that point. This requires a solid and reliable ETL tool checkpoint functionality to restart the job in exactly the right point. In many cases, it might be best to get out of all the rows that have been uploaded as part of the process and restart from the beginning.

For this reason, it is recommended to design fact tables with surrogate key columns relative to the dimension tables linked. This surrogate key is a simple integer that is assigned in sequence as the rows created in the dimension tables. With the linkage of the surrogate keys with the fact table, you can easily resume a load that is stopped.

The more an ETL process is long, the more you have to be aware of the vulnerability because of an error. The design of a modular system consisting of ETL processes efficient and resistant to abnormal and unforeseen interruptions arrests, may reduce the risk of a failures resulting in remarkable recovery. Careful consideration must be given to entering physical values by writing the data to the disk, when to insert the accurately prepared recovery points and when to choose the specific date / time of loading of the sequential tables with an appropriate restart logic, to not congest the system of customer.

1.3 OLTP vs OLAP

On-Line Transaction Processing (OLTP)

At the database level, the online transaction processing queries are based on quick and effective multi-access. The main operations performed are INSERT, DELETE, and UPDATE as they modify the data directly. The data are constantly updated and, therefore, require an efficient support for rewrites. A key feature of these systems is the standardization, which provides a quick and effective way to carry out writing into the database.

On-Line Analytical Processing (OLAP)

The On-Line Analytical Processing is a set of technical tools for accelerated analysis and large amounts of data, with the ability to study the problem in different points of view. These systems are very useful for the production of synthetic information, which will support and improve business decision-making. Examples of OLAP tools are the data warehouses and the multidimensional cubes.

The major differences between the two systems are shown in this table [10]:

Table 1: OLTP VS OLAP

	OLTP	OLAP
Purposes	Support in operations	Support in decision-making
How to Use	Prompted by processes	Ad hoc query
The amount of data per elementary operation	Low: hundreds of records for each query	High: millions of records for each query
Quality	In terms of integrity	In terms of consistency
Orientation	To process / application	By Subject
Refresh rate	Continue, through actions	Sporadic, through explicit functions
Time coverage	Current data	Historical
Optimization	To read and write accesses on a data portion	For read-only access to the entire database

According to the data storage, you will have several OLAP architectures, each with their own pros and cons [10]:

- *Relational OLAP (ROLAP):* the data is stored in a relational database as a support to the OLAP engine. The analysis is translated into queries, returning results in a multidimensional form;
- *Multidimensional OLAP (MOLAP):* it has the database and the multidimensional engine. For the Drill-Down operations it is not the ideal system, as it can generate errors;
- *Hybrid OLAP (HOLAP):* It combines the advantages of the two previous systems. In particular, pre-aggregates data in multidimensional systems for efficient and fast analysis, and are useful in a relational database in case of Drill-Down;
- *Desktop OLAP (DOLAP):* the data are loaded into a client system and are calculated by the local engine.

1.4 BIG DATA

The term Big Data refers to a collection of extensive data in terms of volume, velocity and variety that requires specific technologies and analytical methods for the extraction of value and knowledge. The term is used to refer to the ability to analyze, extrapolate or put in relation an enormous amount of heterogeneous, structured and unstructured data, in order to discover the links between different phenomena and to predict future ones.

The size varies in different sectors, from dozens of terabytes to hundreds of petabytes (1000 terabytes), also according to the different available software tools. This size will increase over time due to continuous advancements of technologies.

In this definition, emerge the so-called 5V that characterize the Big Data, which is the volume, speed, variety, authenticity and value [2].

The volume refers to the huge mass of data generated through numerous channels.

The rate refers to the rate at which data is acquired and used thanks to faster and more frequent transactions: companies not only collect data faster, but they try to exploit them as soon as possible, often in real-time.

The veracity regards the data quality and their level of security, where security is a very important challenge. To take advantage of Big Data you must know how to act in order to extract the value and increase the productivity and competitiveness of company for creating an economic surplus for consumers.

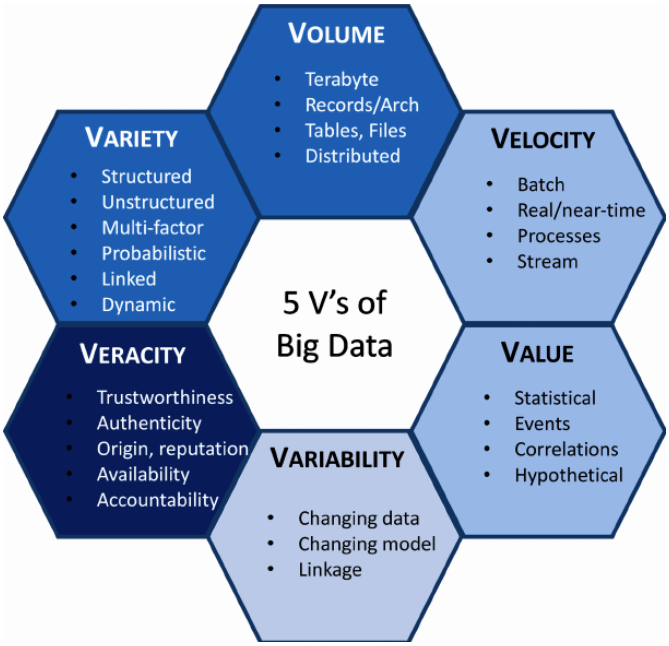


Figure 5: The 5V Big Data

The variety is related to the different types of data available from a growing number of data sources, both structured and unstructured; in particular, it is possible to identify four categories of information that constitutes Big Data:

- Data generated from smartphones and other mobile device, including RFID data (radio-frequency identification) devices that track the product, and data from monitoring devices such as counters for water monitoring or gas;
- Sales and pricing data, generated by loyalty cards and promotional events;
- Computer data log, such as click streams from websites;
- Information from social media such as Twitter and Facebook or from YouTube and similar sites.

The ability to store, aggregate the data and use the results to carry out deep business analysis, is improving thanks to the availability of software tools and the increasingly of sophisticated techniques combined with an increasing computing power. We are seeing a huge change in the ability to create, communicate, share and access to the data due to the increase in the number of people and tools connected to digital networks.

1.4.1 Barriers on the use and extremely beneficial Of Big Data

The Big Data represents a great opportunity for companies and for national economies thanks to several significant benefits:

- *Revealing the variability of performance to improve them:* Creating and storing transactional data in digital form allows companies to have more accurate and detailed data on a variety of performances, from the daily state of the warehouse to staff sickness, all in real time or nearly so. Furthermore, they use the data to analyze and in order to understand the deeper causes of the variabilities;
- *Customize actions*: The Big Data allow you to create specific customer segments, called cluster, to customize products, services, promotions and advertising based on their needs;
- *Improve forecasting and supporting people in the decision-making process*: Using sophisticated analytics on entire dataset can automate and improve decision-making by the predictions of Key Performance Indicators (KPI), useful to minimize risk and discover valuable insights; These benefits cannot be pursued with the analysis and management of small samples of data via spreadsheets, but need a large number of integrated data that only a data warehouse can provide. Retailers, for example, can use algorithms that enable the auto-tuning and optimization of inventories and prices, starting from real-time data on sales.
- *Transparency*: An easy and timely access to big data makes an available larger amount of information and facilitates the sharing of data between the different organizational units of an enterprise;
- *Consumer Profiling*: The availability of near real-time data from smartphones provides detailed characteristics about customers and their complex decision-making process when shopping, identifying consumer behavior patterns and shed light on their intentions;
- *Create new products and services, new types of companies and innovative business models:* Companies can leverage big data to improve the development of future models and to create innovative after-sales services;

- *Increase productivity and profitability of companies:* The exploitation of Big Data can lead to greater efficiency and effectiveness of the companies, which will produce more output using less input, improving the level of quality.

This list of benefits highlights how the investments in Big Data leading to the creation of value for companies, obtaining competitive advantage in the long term.

Despite the opportunities offered by big data is huge, there is still some skepticism within companies on the real benefits due to the lack of results in practice [3].

There are therefore several barriers to consider, which can be classified into six categories:

- *Technical barriers*: Data integration difficulties, low-grade of business influence, poor quality of data;
- *Barriers linked to skills*: Difficulty understanding of the analytical tools and benefit quantification, shortage of talents, difficulty in choosing the suitable tool;
- *Organizational barriers / Management*: Lack of commitment of top managers who are not involved in the Big Data initiatives, towards which show little interest, resulting ineffective;
- *Cultural barriers*: Most companies are not ready and totally open to innovations that could bring big data, since their exploitation would require significant cultural and organizational changes: Inertia.
- *Economic barriers*: The Big Data initiatives require huge expenses in terms of implemented technologies and new professionals to hire or consulting;
- *Barriers related to privacy:* Consumers do not want that their personal information, such as personal location data and electronic data generated by their use of the Internet are used by companies, especially because they do not know where and how these will be exploited by the organizations, which must also consider the laws relating to statements of different countries. Tools that let track any movement of employees and their performance make the interests of organizations but not for worker who see a threat to their privacy.

1.4.2 Techniques for Big Data Analysis

So far, we have talked about the ideology and the value that Big Data can lead to a company. Here, however, they will be listed the main techniques and technologies used to aggregate, manipulate, manage and analyze the data.

- *A / B testing*: Technique in which a control group is compared with the test groups in order to determine what changes and actions will improve one target variable data, such as the response rate to a campaign of Marketing;
- *Crowdsourcing:* Technique used to collect data subjected to a large group of people or a community through, for example, the Web;
- *Data integration*: Set of techniques that integrate and analyze data from different sources in order to develop insight more efficient and accurate than those obtained by examining a single source;
- *Predictive models*: Techniques in which is created or chosen a mathematical model to predict the probability of a result;
- *Data mining*: Set of classification techniques, cluster analysis, regression and association rules, which allows to extract patterns from large datasets by combining statistical methods, machine learning and database management;
- *Machine Learning*: Part of computer science dedicated on the design and development of algorithms that allow computers to identify behaviors based on empirical data and to recognize complex patterns for predicting decisions by means of artificial intelligence;
- *Natural language processing (NLP)*: Set of linguistic science and computer techniques that use computers to analyze human language;
- *Regression*: Set of techniques that make it possible determine how the value of a dependent variable changes when one or more independent variables are changed;
- *Optimization:* Set of numerical techniques used to redesign systems and complex processes in order to improve performance with respect to one or more aspects, including cost, speed and reliability;
- *Sentiment Analysis*: Application of natural language processing and other analytical techniques to identify and extract subjective information from the texts, for example

the "polarity" (positive, negative or neutral) on the characteristics of the products on which the people have expressed an evaluation;

- *Statistics*: The science of collecting, organizing and interpreting data, used to make judgments about the relationships between variables that could have occurred by chance (null hypothesis) and by causal (statistically significant);
- *Data Visualization*: Creation of images, charts, diagrams or animations that let you communicate, understand and improve the results of Big Data.

1.5 BIG DATA PROJECTS IN MARKETING

The exploitation of Big Data in Marketing is a huge potential for the companies that have a great interest in projects that provide for their use in this area. We will face on six type of projects: Direct and Digital Marketing, the Customer Micro-Segmentation, the Location-based Marketing, Price Optimization, In-store Analysis and Cross-Selling / Up-selling.

1.5.1 Direct and Digital Marketing

The Direct Marketing includes all the marketing techniques that allow companies a targeted and personalized way to communicate directly with the customer or end user. The continuous and significant growth of Internet has led to the rapid development of Digital Marketing, which takes the form of advertising, content on Facebook, video clips on You Tube, personalized email and much more. Companies to make Digital Marketing today can rely on the enormous amount of information of users who spend hours and hours on the Internet, sharing their interests, the content of their communications, the purchases they make and more [4].

The Direct Marketing uses many of Big Data techniques, as well as to identify the most profitable customers and those most likely to respond to the market, so they also to predict the behavior of those strangers. They are used both unsupervised learning techniques such as optimization models, neural networks and Bayesian decision trees

and those not supervised, including clustering. For best results, the ideal is to combine the several techniques [5].

The benefits from Big Data to Direct Marketing are, in addition to the personalization of the message, the 360-degree view of the customer, the identification of content, the timing and the most appropriate channel to send the message in real time.

This results in an increase in the conversion rate, like the number of visitors who decide to click some random content or to visit a web site as a result of an action driven, and thus the maximization of Digital Return Of Investment (ROID), the acquisition of new customers and the retention of those who already are client of the company.

1.5.2 Customer Micro-Segmentation

The variety of new types of data and the development of advanced Analytics allows for granular details and a larger number of consumer reports, generating very precise micro-segments, constituted by a small number of people [1]. Traditional segments B2C (Business to Customers) and B2B (Business to Business) based on demographics, psychographics, behavioral and on the size of the companies or the acquisition criteria are obsolete.

The most common criteria are:

- *Activity-Based Data*: Click-stream data from the web, historical purchases, the call center data, mobile data;
- *Profiles of social networks*: Historical activities and membership in groups;
- *Sentiment Data*: Associations with products and businesses (like or follows) and online comments.
- *Traditional data*: Market research and transactional data;

The goal concern into build always more narrow segments. The marketing men can therefore create offers, customized products and services tailored to each cluster, with obvious benefits on returns. This data can also be updated in real-time, thus being able to monitor customer changes and preferences.

1.5.3 Price Optimization

Companies can take advantage from the increasing granularity of sales data and powerful analytics to optimize prices. The amount of information available to them is huge, thanks to the historical demand series, the inventory data, until the current sales level. This database is constantly rising given the explosion of new online sales channels where consumers can compare prices, increasing competition between different brands on the market [6].

From these large amounts of data, through appropriate tools, pricing managers been able to extract insight to define the optimal price almost in real-time that a consumer is willing to pay for each product, based on its characteristics.

The price optimization can consider, for example, the elasticity of demand to price, with specific models that analyze historical sales data to derive insights on the pricing of each unit, which can then be used to make promotions, to reduce prices or to evaluate costs. The benefits that businesses can achieve in this way can increase revenues, margins and market share.

However, is necessary to build a confidence state with the customers, identify the most promising opportunities, for determining what exactly the consumer wants to pay for a given product through customer segmentation and personalized promotions. Particular attention is focused on the correct use of adequate analytics to identify items that are often overlooked, and to determine the driving factors for each customer and product that will lead to price final choice [7].

1.5.4 Location-Based Marketing

The Location-based Marketing is based on the adoption of the growing smartphone and other mobile devices that generate personal location data, which allow you to learn about location and behavior of people in real-time using GPS or Wi-Fi technologies, encouraging the development of a marketing strategy. We need to consider also the habits and fun of worker and not only the consumer preferences. Other sources used

are the signals of the triangulation of cell towers and the payments through credit and debit cards, which, through the point of sale terminal, make available personal information.

What businesses usually they make is called the Geo-Targeted Advertising, advertising or undertaking actions in real time based on the location of its customers. In fact, to get huge benefits, companies use several push notifications, for example, current offers and customized for a specific customer that, for example, walking holding smartphone inside the store. Therefore, exploitation of geo-location data may lead to an increase in sales, with an increase in profits and improved customer experience and therefore, increasing customer loyalty.

However, with respect to this project, companies are faced with two challenges: The privacy and a trade-off, that is, if users wish to receive Mobile offers when you are near to a store.

1.5.5 In-Store Analysis

The in-store analysis includes the analysis of real-time data on the behavior of consumers through the position and location of the customers inside the store are tracked through a variety of technologies: video cameras, Wi-Fi, Bluetooth, systems tools of retail outlets, payment cards, transponder carts, smartphone applications, Path Intelligence and RFID tags on purchasing cards.

By doing this insight are excerpts related to consumer behavior in the store, with an ultimate goal of improving the customer experience.

In particular, the insights obtained are related to how many customers enter the store, how they behave shoppers inside the store and to know the consumer through the attributes like sex, age, if it is the first time that enters the store, if it comes back often, where it comes from and what are his interests.

Companies use these insights to effectively improve the organization, or to optimize the layout of the store, its characteristics, shelf placement and product mix to turn one-time

customers into repeat customers, to increase the frequency of their visits and their expenses by improving the store experience.

1.5.6 Cross-Selling & Up-Selling

The Big Data offer great opportunities to increase the average purchase size of a consumer, providing products or services related to the initial purchase choice improving the actions of Cross- selling and Up-selling. Data such as the demographic characteristics of customers, the real-time location, preferences, history of past purchases are used for this purpose [9].

The benefits that companies have are the increasing sales and profits, and customer loyalty.

Case example is Amazon which collects data from all users, recognizing the trend in people who make purchases through analytics tools, in order to capture potential opportunities and, according to each product or service visited in the site, suggests the client to buy other similar thing available in the website, significantly increasing sales.

1.6 KNOWLEDGE DISCOVERY IN DATABASE (KDD)

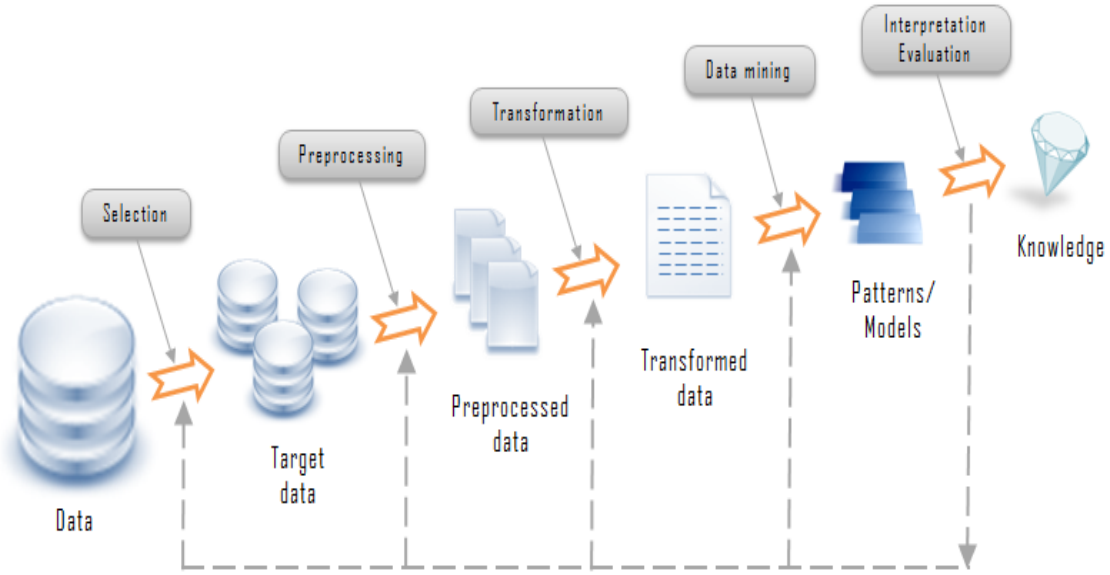


Figure 6: KDD Process

The KDD is an interactive and iterative process that seeks to extract implicit information from the data, a priori unknown but potentially useful.

Let's now analyze the individual steps:

- *Identify your goals*: The goal of this phase is the identification of the objectives to be pursued. It is perhaps the most difficult stage in terms of resource allocation and because it must be determined the criteria for measuring the success or the failure. Is make possible only a partial list of the many aspects that must be taken into account, like the estimated cost of the project and the choice of data mining tools to be used;
- *Selection*: The raw data are segmented and selected according to some criteria in order to create at a subset of data, which represent our target. If the original data is placed in a flat file, creating the target it is very simple. The management systems store and manipulate data transactional database, which enables information systems for making upgrades and extract information quickly. This is due to the

structuring of data using relational models, whose aim are to accelerate the access to information and reduce data redundancy, through the decomposition of individual tables in the relational structures. Often, you also need to put together information extracted from multiple sources, which can make the selection process more difficult because you have to transform the data to ensure consistency in that, for example, data encryption must be equal for all records of the target data, otherwise the analysis is of little use; whose purpose is precisely to bring together data, and do not decompose in order to exploit redundancy. Often you also need to put together information extracted from multiple sources, which can make the selection process difficult because you have to transform the data to ensure consistency in that, for example, data encryption must be equal for all records of the target data, otherwise the analysis is of little use;

- *Preprocessing:* Generally, the available target date should not be analyzed entirely but just extract an appropriate sample, then performing an analysis on a sample basis. Furthermore, the data must be pre-processed, i.e. "clean", treating in a timely manner outliers and missing data. Should be identified incorrect values of the variables; to find errors in categorical data becomes a problem when analyzing very large dataset. The data should also be simplified; these smoothing date techniques are aimed at reducing the number of values for a numeric variable. Some classifiers, such as neural networks, using functions that perform the simplification during the classification process, performing so a date internal smoothing. Two simple simplification techniques are the calculation and rounding of average values;
- *Transformation*: The data to be used, often have to be transformed; this phase may take various forms and may be necessary for various reasons. It can convert data types in other or define new ones, obtained with mathematical and logical operations on the variables, perform normalizations (decimal scaling, normalization min-max or with the z-score) or even eliminate the variables. In general, in fact, the DM algorithms do not work efficiently if the data contains many variables that are not able to predict the class. It is therefore useful to a search and a subsequent elimination of redundant and "unnecessary variables" for the problem in question.
- *Data mining*: Algorithms that study data to give information not trivial or obvious. These are the objectives to be achieved to give an indication of the type of technique to be applied;

- *Interpretation and evaluation*: The purpose of the DM is to determine the validity of the model obtained; in short, is not enough just interpret the results but you will have to understand to what extent this model or results will be useful. This can be done in various ways and through statistical analysis or experimental;
- *Data Visualization*: The ultimate goal is to use what has been learned, creating a technical report on what has been discovered, trying to figure out how to exploit what has been discovered.

From this analysis, it is understandable how the process of extracting knowledge is long and rather complex, therefore, the choices that are made for the treatment of anomalies or errors in the data and the clear identification of the objectives to be pursued are fundamental.

1.6.1 Data Mining vs Machine Learning

Data mining refers to the extraction of knowledge from large amounts of accurate, new and useful data. It is an iterative process of creating a predictive and descriptive model, through the discovery of previously unknown trends and patterns with large amounts of data to support decision making. It can also be defined as the subset of business analysis, similar to an experimental research. The data mining sources are the databases and statistical methods.

The Machine Learning indicates an area of research in artificial intelligence and, thanks to data-driven, involves the study of algorithms that are able to extract information automatically. Two data sources are required: training data and test data. Usually, the machine learning using data mining techniques and another learning algorithm to build models of what is happening behind some data so that it can predict future results.

But we see in the table the various differences:

Table 2: Data Mining vs Machine Learning [13]

	Data mining	Machine learning
Definition	Extract Knowledge from a large amount of data	Introduce a new algorithm to data and past experience
History	Introduced in 1930	Introduced in 1950
Responsibility	Data mining is used to get the rules from existing data.	Automatic learning that teaching to the computer to understand the rules given.
Origin	Traditional database with unstructured data	Existing algorithms and data.
Implementation	Develop models where we can use data mining techniques.	We can use the algorithm of the decision tree machine learning, neural networks, and in some other area of artificial intelligence.
Nature	Manual	Automatic
Application	Used in the cluster analysis	Used in web search, filter spam, credit scoring, fraud detection, computer design
Techniques	Data mining is more than a search that uses methods to give not obvious information.	Auto learning and teaching done by intelligent task.
Purpose	Limited Area	Large area.

1.7 DATA MINING ALGORITHMS

The goal of data mining is to extract new information from existing data. As we shall see, there are two approaches: supervised learning and unsupervised learning [14].

- *Supervised learning*: Machine learning methodology in which it is passed to the machine data containing the original data and the expected result. The task of the machine is to find the rule (or function model) with which to create a relationship between the two in such a way that, at the occurrence of a previously unknown example, can obtain the correct result. The data is previously labeled or assigned to a certain category. The supervised learning is mainly used for classification problems, such as, for example, is used in marketing to classify potential customers and offer products that could be interested on the basis of the profile and history of purchases. Another example is the anti-spam email systems.;
- *Unsupervised learning:* Unlike the previous one, does not use classified data labeled before; we do not know, then, what category they belong. Therefore, to the machine is requested to extract a rule that groups the cases presented in accordance with characteristics which derives from the data themselves. For this, it is also defined characteristics of learning (learning feature). The algorithms in this case seek a relationship between the data to understand if and how they are linked together. Since it contains no preset information, the algorithm is called to create "new knowledge" (knowledge discovery). One of the main applications is the clustering or grouping data into homogeneous groups called clusters. The unsupervised learning, therefore, generally serves to extract information not yet known.

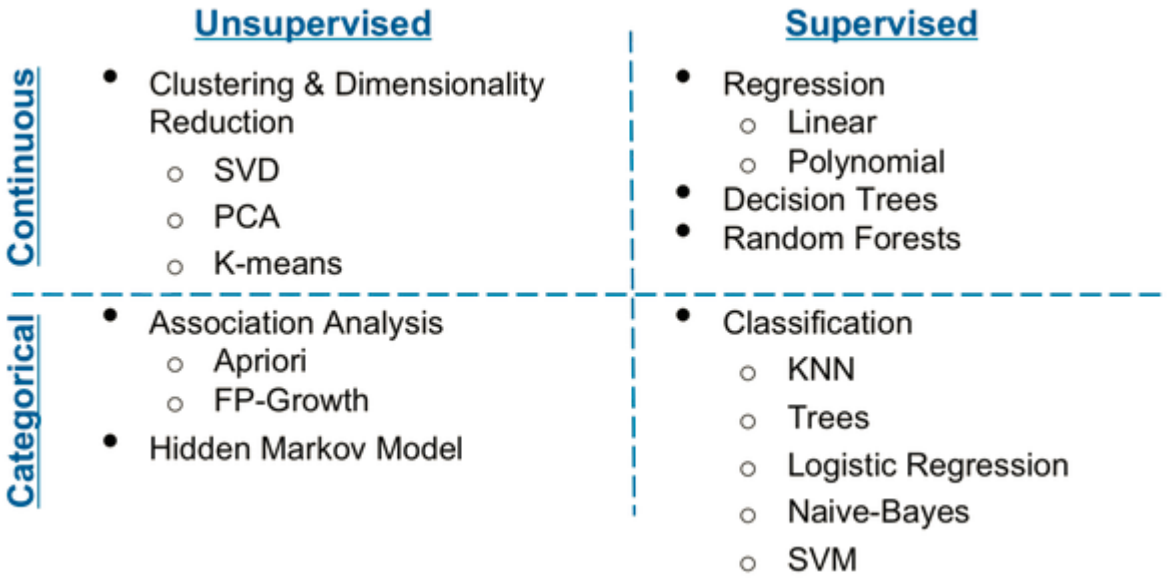


Figure 7: Data Mining Algorithms

Using some of the techniques mentioned above we can create some predictive models. Whatever their application, the predictive models use the experience to give a score, confidence levels and some interesting results for the future. To do this, you have to divide the process into two phases:

The first stage is the creation, in which the model is created using data from the past, while the second is the score of the model created.

We must never forget that is important to get good results (score) in the test data and not in the training data. The overfitting is the situation that occurs when the model explains the training data but cannot generalize to test data.

The innovations that use artificial intelligence and machine learning are the major technology trends in the retail world. They are having a big impact on the industry, particularly in e-commerce companies that rely on online sales, where the use of some form of machine learning is very common today, especially in retail.

Large online retailers such as eBay, Amazon and Alibaba have successfully integrated AI technologies throughout the sales cycle, from storage logistics to customers after-sales service.

Companies using the recommendation systems achieve sales increases as a result of a better customer experience. The recommendations, in general, accelerate research and make it easier to acquire and retain customers by sending e-mails with links to new offers that meet the interests of the recipients and adapt to their profiles.

When the user begins to feel understood, it is more likely to purchase additional products. Knowing what a customer wants and show immediately to him the product, it is less likely that he leaves the platform. This result gives a greater chance of purchase and a decrease in threat of losing a customer that move to a competitor.

By including the offer, seasonality, external events related to your business (such as a concert, a match, a festival), demand forecast and market supply, an automatic pricing system can efficiently adjust prices.

We see in detail the most common by machine learning algorithms used.

1.7.1 Clustering

The goal of clustering is to organize the objects examined in groups that sharing similar properties. Clustering can be considered one of the most important methods of unsupervised learning and, like any method belonging to this category, it does not use certain classifiers prior to guess the possible structure of the data.

There are various forms of clustering [15]:

1) *Exclusive Clustering*: Each element can only belong to a cluster, that is, the intersections between the clusters are always empty datasets; This procedure is also called Hard Clustering;

2) *Inclusive Clustering*: Each item can belong to multiple clusters simultaneously, with an index that determine the degree of membership in each cluster. This procedure that is called Soft and Fuzzy Clustering;

3) *Partitional Clustering*: It uses the concept of distance between the elements, which belong to a particular group based on their relationship with a significant point of the dataset;

4) *Hierarchical Clustering*: It builds a hierarchy of partitions both for aggregation that by division, by means of a tree representation that takes the name of dendrogram. There are other more detailed subdivisions with respect to the Clustering partition, which differ in the evaluation of the distance between the elements and the related cluster [19] creation. This technique is divided into two approaches:

    - Agglomerative: The process begins by considering each point as a cluster, at each step are unified points according to a particular arbitrary function of similarity, to obtain a single cluster and its dendrogram. This approach is based on the development of a Proximity Matrix that take into consideration the function for calculating the similarity between two clusters;
    - Divisive: Complementary case in which one starts from a single cluster that is divided at each iteration, until obtaining a number of clusters equal to the number of points that constitute the data base.

The complexity is in the order of O (N^3), and as in K-Means the presence of outliers create some very important problem to this approach.

Next, it will be shown the main clustering strategies and algorithms used in Fashion.

1.7.1.1 K-Means Clustering

The centroid clustering is based is represented by a prototype called centroid which typically is the average of the distances of the points in the cluster. One of the most popular clustering algorithms in this category is the K-Means that requires you to specify the number of K clusters to be obtained. The algorithm iteratively elects the K centroids of the clusters, and each element is associated with the closest centroid. The algorithm is as follows:

K-MEANS ALGORITHM
1: Function K-Means (clusters K)
2: *Election K centroids*
3: **repeat**
4: *Assignment of each element to the point K closer*
5: *Recalculation of the K centroids* **until** *the centroids do not vary*

Initially, the centroids are chosen randomly while, in later iterations of the algorithm, they typically consist in the average between the distances of the points in the cluster. There are different methods to calculate this distance: Euclidean Distance, Cosine Similarity, Correlation. The algorithm converges to the similarity measures listed. This convergence occurs mainly in the early iterations, followed by a phase of adjustment. In it, in fact, often the stop condition is relaxed, admitting a minimum threshold of change between the centroids.

The choice of centroids is a very sensitive stage, in fact are applied the following techniques to solve, even if not completely, the problem:

- It Performing multiple executions estimating the centroids in different ways or just randomly. After, we must evaluate the quality of the results obtained by means of validation tools that will be described later;
- You use the Hierarchical Clustering procedure to perform subdivisions K and calculate the centroid of the obtained clusters. These are the starting points for the algorithm K-Means;
- It is estimated a number of centroids N> K;
- Postprocessing techniques, such as elimination of small clusters, merge some cluster with very similar between them and breakdown of clusters too large;
- It uses the bisecting algorithm. It consists of a hierarchical approach through which, starting from a single cluster, is divided by a 2-Means algorithm an arbitrary number of times. The iteration that produced the best cluster is taken, and the algorithm is applied recursively until you get the desired K cluster.

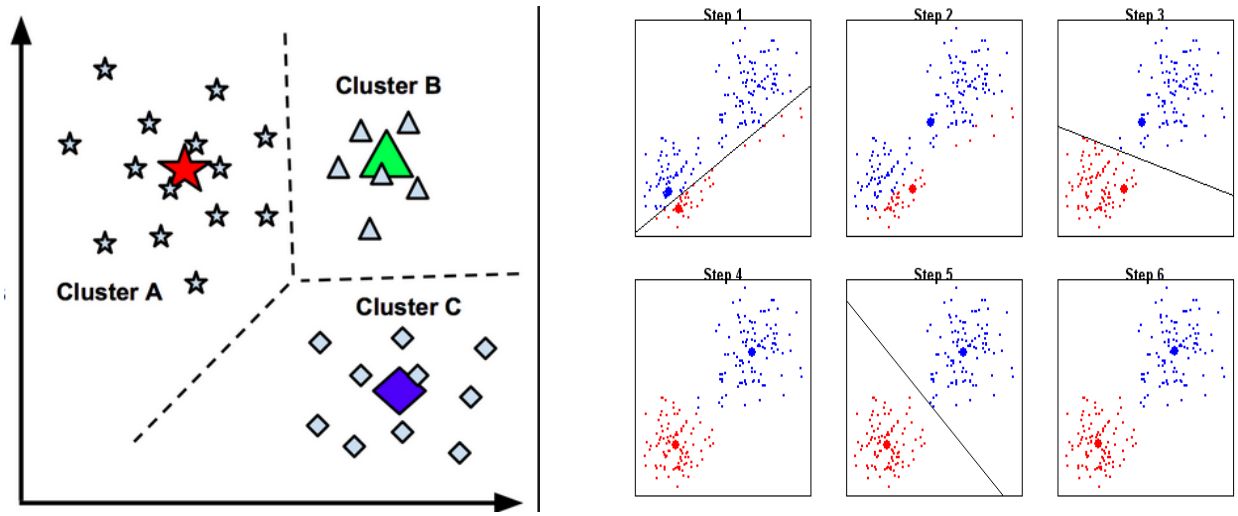


Figure 8: K-Means Algorithm

The complexity of the algorithm is O (n * K * I * d) where n is the number of points, K the number of clusters, I the number of iterations and d is the number of attributes based on which the function for the calculation of the distance.

In conclusion, the K-Means algorithm presents some difficulties in the management of data whose presence of outliers is too high. In fact, are often performed procedures Preprocessing to mitigate the problem. In addition, as mentioned earlier, the choice of centroids is often difficult, especially when it has to do with high-density data. However, K-Means is one of the most widely used algorithms especially with regard to the problem of Customer Segmentation.

1.7.1.2 Density-Based Clustering

The density-based clustering is based on the concept of density. The basic idea is to find clusters defined implicitly by high density regions separated by low-density regions. One of the most famous algorithms of this category is the DBSCAN that uses two parameters to identify dense areas: a ε-range, which is used to identify an area around

a given point, and a minimum number of that must be present at the internal radius of ε. Each point is labeled in accordance with 3 different categories:

- *Core Point*: All the points that exceed the threshold MinPts within the radius ε;
- *Border Point*: All points that do not exceed the threshold MinPts but within their range ε have at least one Core Point;
- *Noise Point*: All the points that are not Core or Border Point.

The algorithm starts from a random point. All points included in the radius ε are calculated and if it contains a MinPts number of points, a new cluster is created otherwise it is labeled Noise-Point. The point could subsequently be found as it is included in the radius ε of a neighbor and consequently be inserted in a cluster.

If a point is associated with a cluster, the points within its radius ε are also inserted in it, and consequently also their neighbors within the radius established. This process continues until all the neighbors have been entered. Each point a cluster is associated with is marked as visited and the algorithm continues by performing the same procedure for a subsequent point that has not yet been visited.

The algorithm has $O(n^2)$ complexity but this can be reduced to O (n log n) by use of structures indexed for querying the neighborhood.

The strength of this approach is given by the good management of outliers and the consequent ability to be able to handle cluster of shapes and sizes very different. However, it is inefficient when dealing with data that are characterized by a density too variable. It is widely used to cluster for Geo-location.

1.7.2 Classification and Regression Trees (CART)

CART is a non-parametric procedure where you do not need to pre-test the normality or other assumptions concern to the statistical distribution of data. The final tree includes only the independent variables that appear to be predictive of the dependent variable; the other independent variables are not predictive have no effect on the final result; CART from this aspect differs from other traditional statistical procedures. With the

classification term is refers to the process that through a collection of records, called Training Set, trying to build a model able to attribute a feature called Class attribute, based on the combination of other properties that characterize the individual of the population. Once you have the template, the structure of a classification tree includes non-terminal nodes (parent nodes), which has two direct descendants (child nodes), and the terminal nodes that do not undergo further bipartitions (terminal nodes). The first node (root node) contains all observations. From the root node descended two "child nodes." Each child node, denoted with the letter "t" contains a subsample of the original sample, in which members share the same characteristics that influence the dependent variable of interest. Each t, in turn, constitutes a parent node potential that can still be divided into two child nodes. The process continues until the tree does not stop its growth.

There are some important steps to follow when building a decision tree with the CART procedure: 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.

To choose the split criterion is generally used a technique of Recursive Binary Splitting.

The method is binary and recursive: binary, since each parent node is divided into two direct descendants, and recursive, since the nodes (non-terminal) born from the splitting of the parent node into two direct descendants that can become, in turn, parent node divided successively into two other nodes.

A decision trees with many nodes and a huge number of divisions may lead to a data overfitting. This means that the model is difficult to interpret, as it becomes inaccurate for later forecasts and needs of the stopping rule. The methods to avoid this problem is to set a minimum number of training data to be used on each leaf node or set the maximum depth of the model, which refers to the length of the path longer from root node to leaf node.

The different existing algorithms differ depending on the strategy employed on individual nodes, for the evaluation of Split. There are in fact different indices for the validation of a classification:

- *GINI INDEX:* Identifies the quality of the split. Considering p_i the relative frequency of the node to the class:

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

- *GAIN INDEX*: It is based on the concept of entropy and identified the homogeneity index relating to the node, obtained by performing a particular split on p_i node:

$$Entropy = \sum_{i=1}^{c} -(p_i) * log_2(p_i)$$

1.7.2.2 Others types of Classificator

- *Based on Instances*: It consists of a family of algorithms which, rather than performing explicit generalizations, compare new instances directly with the analyzed records and properly stored by the training set. Worthy of note, is the Nearest-Neighbor procedure that uses a particular arbitrary metric for the distance calculation and a parameter k representing the minimum number of neighbors to be extracted [15]. For each record that should be classified, it calculates the distance from the training set identifying the k closest records and using the values of their attributes, classify the records;
- *Bayesian Classifier:* It consists of a probabilistic framework for solving the problem of classification. They consider the attributes and the class as random variables based in a strongly relying on the concept of Conditional Probability. Given a record with attributes (A1, A2, ..., An), the goal is to predict the class C. We want to find the value of C that maximizes the probability P (C A1, A2, ..., An). It follows the Bayes theorem:

$$P\ (C|A) = \frac{P(A|C) * \ P(C)}{P(A)}$$

Thanks to Bayes' theorem, you get an equivalent optimization problem that is to find C that maximizes: P (A | C) = P (A, C). There are different ways to estimate this

probability based on data, such as normal distribution, density estimate and Laplace [15];

- *Support Vector Machine (SVM):* The classification is performed by finding the hyperplane that maximizes the margin between the two classes. The vectors (possible attributes of the class) that define the hyperplane are called support vectors. The advantage of this method is that if the data is linearly separable, then there exists a unique global minimum. An ideal SVM produce a hyperplane which completely separates the two non-overlapping classes. Typically, the complete separation is not always possible, but often it gets to obtain a model with too many possible cases that involves an incorrect classification [18].

The validation of these process has a huge importance, since it allows to evaluate the performance of the constructed model and can compare it with other possible modeling. The evaluation measures are based on the Test-Set, the data partition on which to apply the predictive model.

The application of the model on the Test-Set produces the Confusion Matrix, which is a matrix indicating the incidence between the classes and their real value of the records in the Test-Set. You can then determine the following types of prediction:

- *True Positive:* Correct Predictions Positive;
- *False Positive:* Correct Predictions Negative;
- *True Negative:* Wrong Predictions Positive;
- *False Negative:* Wrong Predictions Negative.

This can be applied to any type of attribute, not only to binary classes.

The most commonly used metrics are: Accuracy, Precision, Recall, F-Measure.

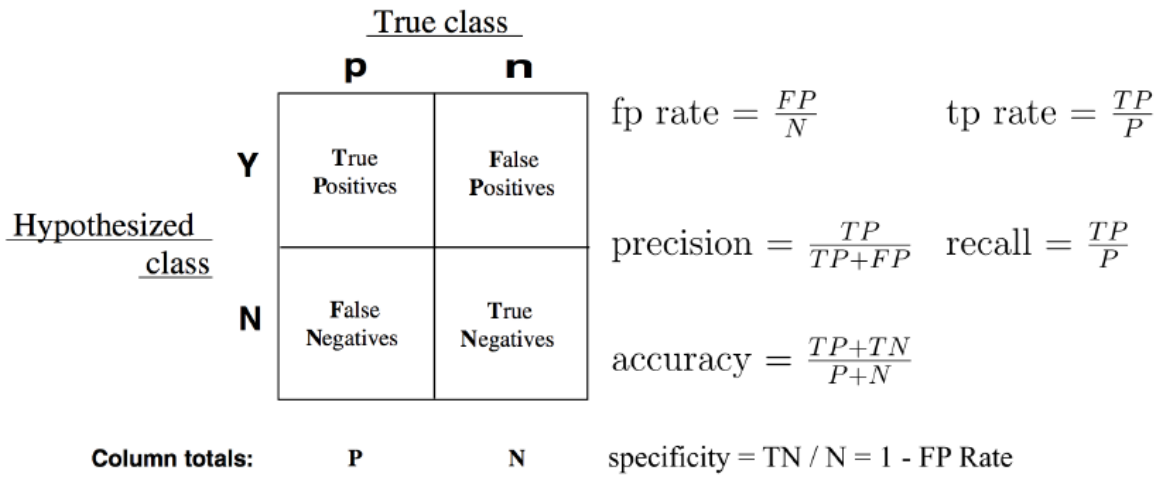


Figure 9: Confusion Matrix

1.7.3 Prediction: Association Rules

The starting point of an association rule algorithm consists of a set of transactions. Each transaction consists of a set of items. The algorithm extracts the Association Rules to predict the occurrence of an item on the basis of other appropriate item, included also in the available transactions.

Is important to define some of the concepts behind this technique:

- *Itemset*: A collection of one or more elements generally defined by means of the parameter k, indicative of its size in k-itemset form;
- *Itemset Support*: Given an itemset I, the support is the fraction of transactions that contain I and denoted by supp (I);
- *Frequent itemset*: All itemset that exceed arbitrary minimum support threshold;
- An association rule is an implication in the form: X → Y with X, Y itemset where X is called premise and Y is called consequence of the rule.

In addition to the support, seen previously, there is another form of validation rule that takes into account both the premise and the consequence: Confidence. Shows the frequency with which a given rule matches, is the ratio between the number of transactions that contain the rules and complete transactions that contain the premise:

$$conf\ (X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Formally the support supp (X ∪*Y*) It can be rewritten as the joint probability *P* ($E_X \cap E_Y$), where E_X e E_Y are all the transactions that contain X or Y respectively. Thus, we can express confidence as the conditional probability *P* ($E_X \mid E_Y$).

Given a set of transactions, the goal is the extraction of all the rules that meet the arbitrary threshold of the support and the confidence. Their extraction cannot be performed with a Brute-force approach, due to the number of rules that can be generated. To reduce the number of possible rules, it uses the Apriori principle.

1.7.3.1 Apriori

This principle is based on the anti-monotonic properties of the support, which allows to establish with certainty that if an itemset is not frequent, then even all itemset that contain it will become not frequent. Such a property is so formalized, with X and Y itemset:

$$\forall X, Y: (X \subseteq Y) \Rightarrow supp\ (X) \geq supp\ (Y)$$

This property is the basis of Apriori algorithm, where, starting from all possible items with cardinality 1, it builds all the itemset of dimension n + 1, with "n" the start dimension of the itemset; at each iteration occurs if the itemset generated is frequent or less.

The anti-monotone property allows to exclude infrequent itemset and therefore all possible itemset arising therefrom.

The steps to which the procedure is made are as follows:

APRIORI ALGORITHM	DESCRIPTION
1: **function** Apriori (*T, s*)	(Set T transactions, minSupport)
2: $L_1 \leftarrow$ {*large* 1 − *itemsets*}	*k* = 1 and Generation itemsets with cardinality 1
3: k=2	
4: **while** L_{k-1} *f* = ∅ **do**	*Generation itemset with cardinality* k + 1.
5: $C_k \leftarrow$ *Generate* (L_{k-1})	
6: **for** *transaction t* ∈ *T* **do**	*Elimination of itemset containing infrequent.*
7: $C_t \leftarrow$ *Subset* (C_{k1}, *t*)	
8: **for** *candidates c* ∈ C_t **do**	*Calculation support itemset generated.*
9: *count*[*c*] ← *count*[*c*] + 1	
10: $L_k \leftarrow \{c \in C_k | count[c] \geq s\}$	*Elimination of itemset containing infrequent.*
11: *k* ← *k* + 1	
12: **return** $\cup_k L_i$	

At the end of this procedure, we get all itemsets that have stood the support threshold. We must proceed with the extraction of association rules from itemsets obtained. The generated rules will be evaluated according to their Confidence (arbitrary threshold), that, generally, does not enjoy the anti-monotonic properties.

Indicating with Conf (X⇒ Y) the confidence of the rule X ⇒ Y will get:

$$Conf\ (ABC\ \Rightarrow\ D)\ \geq\ Conf\ (AB\ \Rightarrow\ CD)\ \geq\ Conf\ (A\ \Rightarrow\ BCD)$$

The algorithm proceeds generating the rules that possess only an item in the result, eliminating all the rules that do not exceed the minimum confidence threshold. On the basis of the remaining rules, it generates and evaluates rules with an additional item until all possible rules have been generated.

The extracted rules are subjected to a further step of post-processing, because the confidence can sometimes be misleading as to the validity index for a rule. This aspect emerges for itemset that are part of the premise of a rule, characterized by high support.

A very frequent itemset tends to raise the confidence index of the rules of which it constitutes the premise, regardless of the fact that the rule is contextually valid.

To have an excellent validation us is based on the following indices:

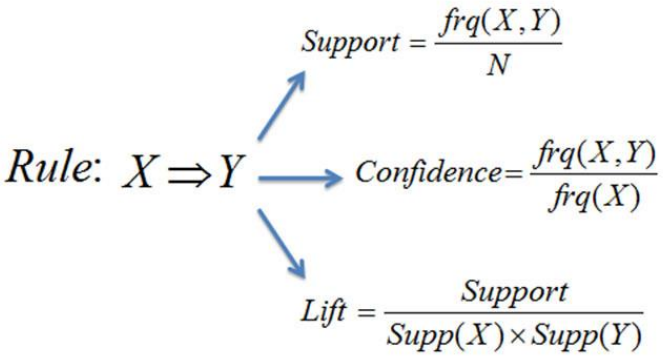


Figure 10: Indicators of Association Rule

1.7.4 Artificial Neural Networks & Deep Learning

The Deep Learning is a specific method of machine learning that incorporates a large number of neural networks together in various layers to learn from the data iteratively.

Neural networks and deep learning are often used in image recognition applications, in speech and in computer vision.

A neural network is particularly useful when trying to study the pattern of unstructured data and are designed to emulate how, through artificial intelligence, computers can be trained to deal with problems that are not well defined [21].

It consists of three or more layers: an input layer, one or more hidden layers and an output layer. Data is ingested through the input level. Then, the data are edited and processed in the hidden layer, obtaining different output levels on the basis of weights applied to individual hidden nodes.

The typical neural network can consist of thousands, or even millions, of simple processing nodes that are densely interconnected.

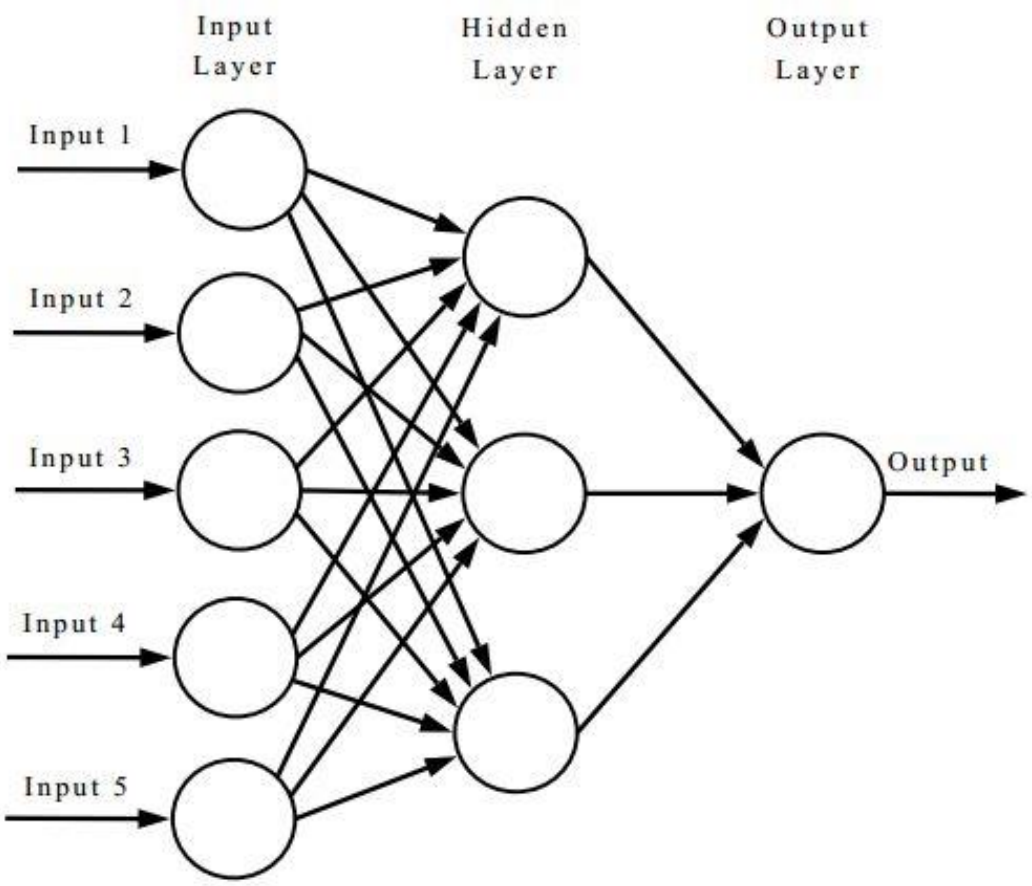


Figure 11: Artificial Neural Network Diagram

In a neural network, the input layer is constituted by the value of the attributes that must be analyzed. The output of this first level of the network remains unchanged, since the output from the input nodes are the same values that are provided for the analysis. In each node belonging to the next levels, hidden layer and output layer, it occurs the actual computation. In fact, the inputs of these levels correspond to the output of previous levels in which, however, must consider the weight associated to the connection between the two nodes and a characteristic value of the node, the offset. Considering a n node between the hidden nodes or between those of output, its In input is given by the following relationship:

$$\mathrm{I_n} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \left(\mathrm{w_{i,n}}\right) * \mathrm{O_i} + \mathrm{offset_n}$$

where wi, n is the weight of the connection between node i of the previous level and node take into account, Oi is the output of node i of the previous level and $offsetn$ is the offset associated with node n considered.

Furthermore, each node applies an activation function on the value that receives as input and sends the output to the next level. When the output from the nodes is generated and if during the learning phase an error occurs between the value of the calculated class and that expected for an instance, this error must be propagated to the previous levels, where the weight and offsets values will be arranged of all the nodes of the layers that make up the neural network..

The term Deep Learning is used when there are multiple levels hidden within a neural network. Using an iterative approach, a neural network adapts and continually makes inferences until a specific stopping point is reached. Practically, it is a machine learning technique that uses the hierarchy of neural networks to learn from untagged and unstructured data through a combination of unsupervised algorithms and supervised algorithms.

Deep Learning is used in Internet of Things (IoT) applications or to predict when a machine will malfunction and is often referred to as a sub-discipline of Machine Learning.

1.7.5 Linear Regression

The regression analysis is a statistical technique used to determine a relationship between a dependent variable and a set of explanatory factors. The dependent variable, referred to as variable Y, is the value we are trying to determine on the basis of the independent variables.

The explanatory factors, referred to as X variables are also called independent variables or simply model factors. Regression analysis helps analysts to find out the sensitivity of the dependent variable to changes in explanatory factors. These feelings are essential for proper risk management.

There are three types of data commonly used in the regression analysis: time series, cross-sections and grouped data.

- Time series: Data collected for a period of time. Are economic and financial data series refer to market returns, prices and asset values, GDP, unemployment rates, interest rates, etc. These data are collected at equal time intervals such as daily, monthly, quarterly, etc.;
- Cross section: Data collected for a family of variables at the same time. For example, fundamental analysis often collects company-specific information such as price / earnings ratio, the book value, the net debt / capital ratio, or the average daily turnover;
- Grouped Data: combination of time series and cross section data.

If we have more explanatory factors, the analysis is called multiple regression model has the form:

$$Y = b_0 + b_1X + b_2X_2 + \cdots b_kX_k + \varepsilon$$

where Y is the dependent variable (what we are trying to predict), X is the independent variable (what we are using to predict), and ε is the random noise (error). In addition, the dependent variable Y, the independent variable X and the error ε are vectors of values of columns.

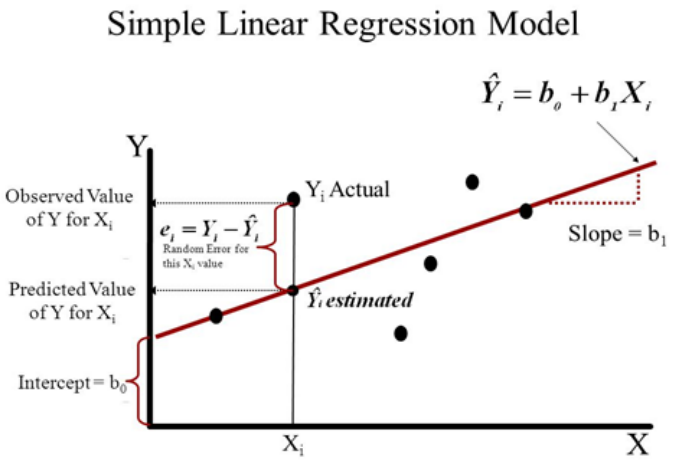


Figure 12: Linear Regression

In the previous equation, b_0 and b_1 are the parameters of the current model which define the exact sensitivity of the dependent variable to the explanatory factors, and ε is the amount of variability that is not explained by the model.

In practice, these exact values are not known with certainty, and must be estimated from the data. To do this you use:

$$\text{Variance} = \text{Var}[\varepsilon] = \sigma_\varepsilon^2$$

$$\text{Expected Value} = E[b\,0] = b\,0 * E[b\,1] = b\,1 * E[\varepsilon] = 0$$

The goal of regression analysis is to determine the set of explanatory factors and corresponding sensitivity that explain as much as possible the employees observed values.

Metrics and Evaluation Assumptions of the model.

When performing regression analysis, the main metrics to analyze are:

- b_k = Parameter of the model refers to the estimated sensitivity of the Y k factor;
- R^2 = Goodness of fit (the percentage of the total variance explained by the model). The linear determination index is defined as the ratio of composition between deviance regression and total deviance measuring in the interval [0,1], explaining how much of the total deviance have the regressors of the model. If we consider the decomposition of the total SST deviance (Sum of Squares for Total Variation) in deviance regression SSR (Sum of Squares two to Regression) and deviance residual SSE (Sum of Squares two to Residual), we show that, with increasing the number of explanatory variables, the deviance of residues decreases and thus the linear determination index increases. Therefore, a high R2 value is not a good fit indicator because it also depends on the number of covariates included in the model.
- F-stat= Critical value for the entire model. Most of the F test arises considering a decomposition of the variability in a data collection in terms of sums of squares. The F-test is the relationship between two scaled sums of squares that reflect different sources of variability. These sums of squares are built in such a way that the statistic tends to be greater when the null hypothesis is not true. Following the F-distribution under the null hypothesis, the data values are independent and

normally distributed with a common variance, so, the sum of the squares should be statistically independent, and each should follow a distribution χ^2 scale [29];

- T-stat= Critical value for the estimated parameter. This statistic is used in a T test when deciding whether to support or to reject the null hypothesis. The higher the T, the greater the evidence that the values are significantly different from the mean. Conversely, a lower value indicates that T is not significantly different from the average [30].

CHAPTER 2: TRADITIONAL ETL FOR THE IMPLEMENTATION OF A DATA MART

The occurrence of the first need to integrated data, has led companies to address the issue internally, as the market did not offer sufficiently flexible and reliable solutions. For address the need for integrated data it was to develop in-house ad hoc software especially, to perform the extraction, transformation and loading of data into a single, integrated environment. Despite recent advances, today most companies use customized ETL solutions to meet the need for integration.

However, the most recent market developments have led to an increased demand for data Integration products, bringing to 60% the percentage of companies using an integration software to carry out activities business intelligence [8].

The recent economic downturn has also led to a decrease of the budget assigned to the development of information technology in enterprises, resulting in increased adoption of open source integration solutions.

One can therefore conclude that the ETL market today is characterized by the coexistence of three types of products:

- *Custom Software: I*ntegration of data internally developed to be able to meet the specific needs of their business scope. With the maturation of the market, this approach has become less and less affordable. In addition, the emergence of SOA and SaaS application architectures is decreeing the end of the products developed “at home”. Today's data integration Suite on the market offer definitely functionality and improved reliability;

- *Software owners:* The data integration products have matured steadily over the years, providing a range of functionality more and more rich and varied making these applications suitable to support the majority of business scenarios. The number of applications on the market today is high, providing suite of products able to cover almost all of the business needs to specialized products in specific business contexts or specific issues;
- *Open Source Software:* the limit of the largest owners' products on the market are the costs necessary for their implementation. To come in against the needs of smaller companies, recently, the first open source products entered in the market. Are products that can support a fair number of features but with a much lower cost than proprietary products (costs license void, reduced infrastructure costs, services paid for with use).

For my thesis, I decided to use an open source software that offers all the features necessary for carrying, possible minimizing costs: Talend Open Studio.

2.1 TALEND OPEN SOURCE

The Talend open source approach provides two products:

- *Talend Open Studio*: Free downloadable suite with an open source license (GPL). Talend Open Studio is presented as comprehensive data integration product and characterized by a wide range of functionality, sufficient for most needs;
- *Talend Integration Suite:* is an enhanced version of the free product that adds advanced capabilities such as collaborative development, advanced monitoring of the project and the Data Masking.

For those who do not have the hardware necessary to support the system, there is a third option consists of Talend On Demand, which is a type of software to offer a Service (SaaS).

The Talend products today offer the following features:

- Development environment user-friendly (based on the Eclipse platform);

- High number of connectors;
- Common Warehouse Metadata;
- Support to collaborative development.
- Data Processing Services;
- Trend monitoring;
- Data Profiling and Data Quality.

Let's see what the strengths of the approach are [7]:

- *No barriers to the adoption*: Free availability of the basic product makes almost immediate the installation of the software. Talend supports the customer through tutorial on the basic use and it is also possible to rely to a large community of users;
- *Fast Learning Curve:* The product is presented graphically user-friendly. Graphical interface is intuitive, and the use of the basic functionality does not require special training;
- *Model of stable and predictable prices:* Proprietary products often involve high costs as that expand the functionality and capabilities of the product. This often makes it difficult to properly costed in the early stages of the project. Talend, however, provides a cost model based on the number of developers and the use of the service, independent of licenses, hardware and quantity of data to be integrated;
- *The importance of supporting communities:* The online community of experts and users of the product is already very broad and today is a very important factor to facilitate the implementation and maintenance of solutions offered by Talend. Forums, wiki, guides and free user contributions represent an added value that only a product of this type can offer;
- *Broad support to different types of data:* With more than 400 preset connections the Talend solution ensures compatibility with a large number of systems, databases, software packages, business applications, web services, etc. No other solution on the market offers a high number of possible connections.
- *Flexibility, versatility and product reuse:* Talend is not limited to a support to standard techniques of ETL but allows the implementation of different integration strategies. The ability to reuse already finalized projects also constitutes another point of strength of open source approach;

- *Functionality and Performance*: The level of functionality offered is comparable to proprietary products. However, there are some gaps in the field of data modeling, data quality and data mining. the research and development team dedicated allows the product to be up to the latest market needs and to propose innovative features;
- *Costs and time-optimized*: The solutions offered by Talend resulting from a 50% to 80% cheaper than traditional products, being less expensive to acquire and maintain and because allow a more rapid development of the integration system.

2.2 IMPLEMENTATION OF A DATA MART

Specifically, 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 [11].

Table 3: Data Warehouse Vs. Data Mart

Definition	Data Warehouse	data mart
Purposes	neutral Centralized Application and shared in the entire company	Specific applications departments or areas
Data	Low de-normalization	High de-normalization
Subjects users	Subjects in many areas	Subjects in a single area
Data Sources	Many External Data Operational	few External Data Operational and DWH
Features	Flexible, extensible Long life, Data-oriented	Rigid, non-extendable, short life, Project-orientation
Implementation Time	9-18 months for the first stage	4-12 months

Deployment can be of two types: top-down, building the data warehouse with a subsequent aggregation and export in various data marts, and bottom-up, build various data marts focusing on specific areas of the business you will to build a data warehouse. In this way you will have a scalable approach.

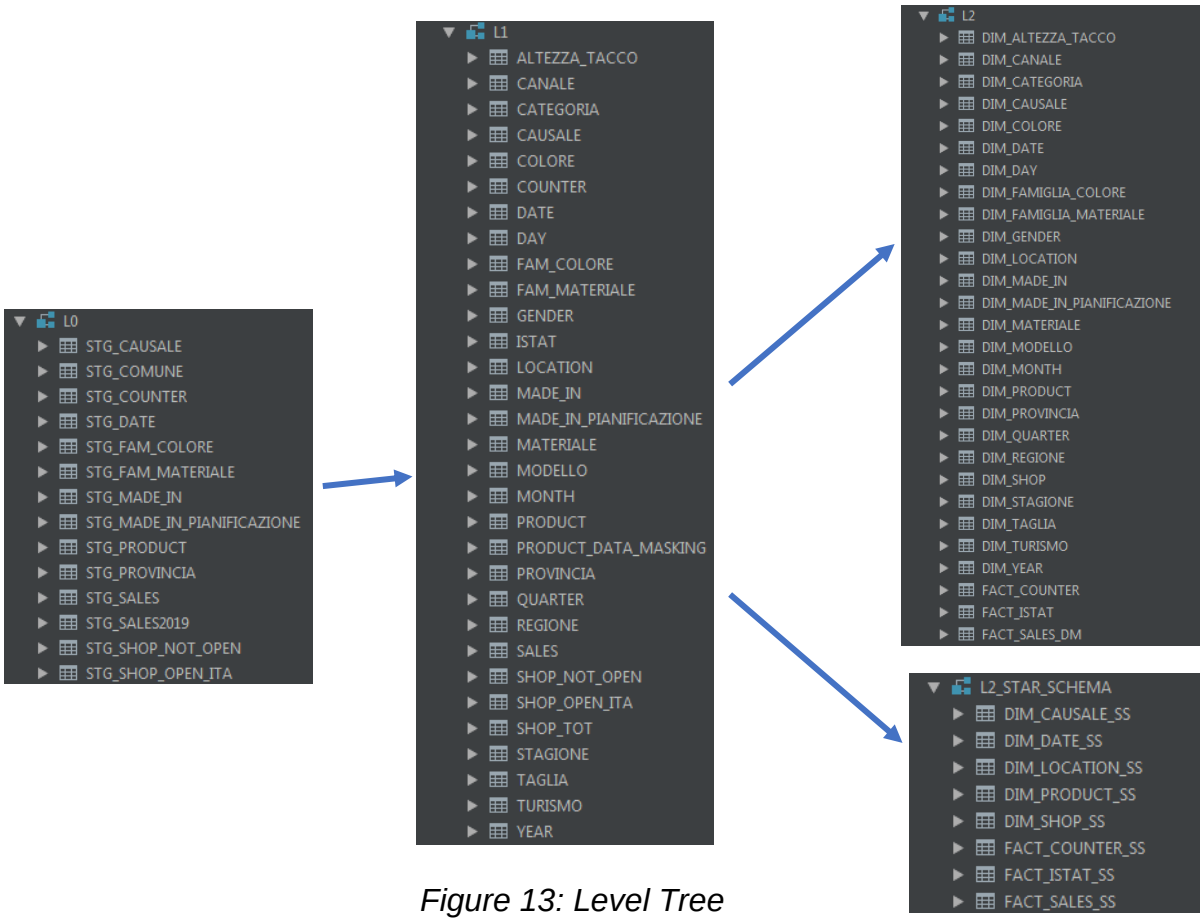


Figure 13: Level Tree

The phase of the ETL for the Fashion Retail project is based on creating a top-down data mart in order to get a fact table of sales with specific data to extract information that currently the customer does not know and that could bring it to produce a competitive advantage or increasing its economic potential. Any data system strictly follows the ACID (Atomicity, Consistency, Isolation and Durability).

To facilitate the recognition of files, tables and any reloading need, we will use a specific nomenclature for each level that make easier to identify the individual files. The tables

created must then reside in a tree structure that allows the historical and conceptual navigation through levels.

Table 4: ETL cores

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.

2.2.3 Delta Of Data

The power of the DWH normally starts with the Initial Load, the process that provides a complete population of data warehouse with consistent data from which you can update and loading new data with delta loads.

The load of delta portion can be done in different ways:

- *CDC (Change Data Capture):* Tables are subjected to a change data capture mechanism that automatically intercepts or replicate, for each table that needs, the delta extraction of new data with respect to the L0 data already loaded

(recommended to replicate the tables very large, so as not having to download moles of unnecessary data);

- *MINUS*: The are not subjected sources tables to any of change detection mechanism and we have to independently extract the amount of data and find the delta performing a minus of the data that we possess with new ones (recommended for the master data tables that contain one mole of note data and do not provide too many changes in the high number of records);
- *FULL*: Daily replication of the table data with small size and low variability of the content.

2.2.2 Historicization

The DLT tables can have a historicizing required to recalculate the data warehouse, to avoid a loss of data or load partial data due to problems, such as server problem.

To have a historicizing of the data we have two choices:

- Creation of partitioned shadow tables by DLT JOB_ID extraction, called HIS. The tables Delta DLT will not be partitioned with the option truncate / insert (truncate the table keeping the initial scheme and insert the new data) and they will contain only the data of the current JOB_ID while the HIS, will insert with partitioning for JOB_ID, with a speed up of the extraction process.
- Have the history directly on the DLTs in insert with the partitioning for JOB_ID, always to speed up the extractions.

You can use only one of these two modes in order to standardize the architecture of the tables to a single model and, depending on the mode used, it will be necessary to differentiate the code of an eventual recalculation.

In the thesis project, the Delta and Historicization tables will not be used because the data come from local csv or Excel file with no possibility of recalculation, not being connected to a source with a constant daily / monthly update.

2.2.3 The Multidimensional Model - Dimensional Fact Model

The ER model (Entity-Relationship Model), spread to design relational information systems, it is not suitable to express and analyze in detail large data sets [10].

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.

To navigate in the multidimensional cube there are different operations that allow you to organize your data in it, through different perspectives [10].

The first is the Pivoting which allows to quickly change the display of data by turning the cube axes and has the purpose to change the point of view from which analyzes the data. The second, however, is the Slice & Dice to select and project the cube data. Specifically, they will extract sub-cubic filtering on a (Slice) or more (Dice) dimensions. Finally, we have the Roll-Up and Drill-Down to help you navigate within a hierarchy, choosing the level of aggregation according to which the user wants to analyze the data.

Specifically, it will rise to a hierarchical level with the roll-up, while you will drop one level with drill down.

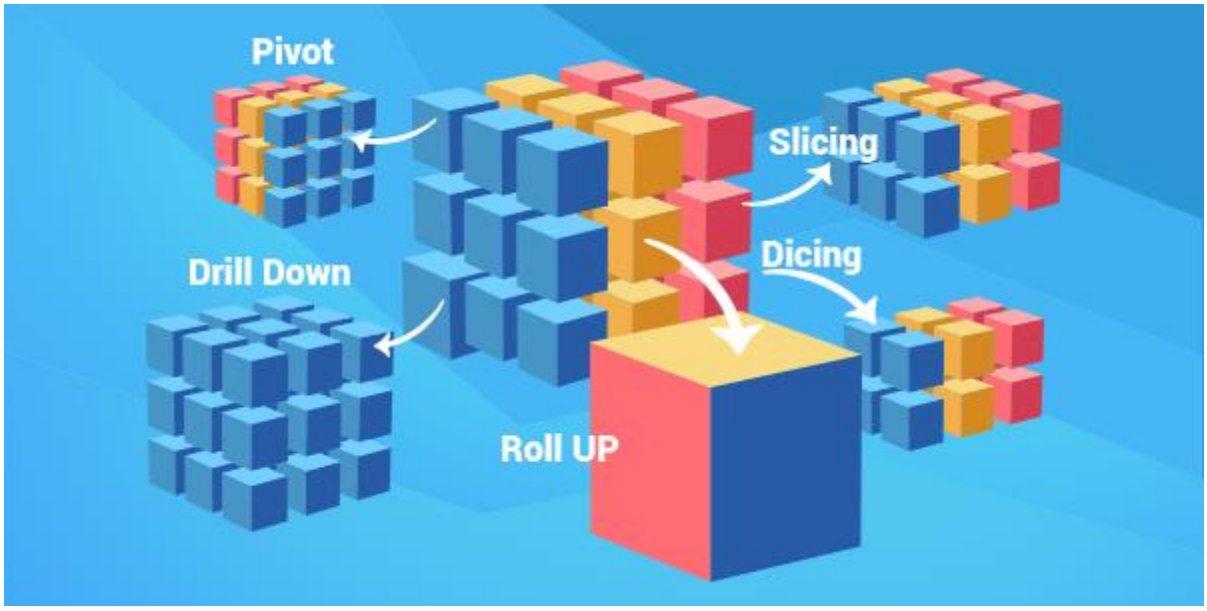


Figure 14: Hypercube OLAP

This system has been idealized to accomplish certain purposes such as provide support to the conceptual design, create an environment where users can make queries intuitively, to interrogate effectively the supplied reports, facilitate communication between designers and users in order to formalize the project requirements, build a stable logic design platform and finally, create and publish a clear and effective documentation.

2.3 L0 LEVEL - DATA INGESTION

The L0 level is the initial level and represent the Extraction phase of the information from source systems for the creation of a data warehouse. The tables are achieved in the Staging Area where the discharge takes place without transformations of data.

The source systems can be of different types but the most common are the operational systems, database or files:

- *From table*: Reading the daily data sets or replication of the entire data set via database;
- *From file*: Extracting information directly from documents.

The data ingestion is the process of acquiring and importing data for the use or the immediate storage in a database.

Data can be transmitted in real-time streaming or ingested into periodic lots of data:

When data is ingested in real time, each data element is imported immediately while it is being emitted by the source. When the data are imported in batches, the data elements are imported into blocks at periodic time intervals. An effective process of data acquisition starts by prioritizing data sources, validating individual files and loads the data items to the correct destination.

If there are several sources of large data in different formats, it can be difficult for companies to acquire data at a reasonable speed and process them efficiently in order to maintain a competitive advantage. To this end, manufacturers offer custom software for specific processing or application environments. When importing data is automated, the software used to run the process, can include data preparation capabilities to structure and organize data; so, they can be analyzed through the Business Intelligence (BI) or, in the specific, through several algorithms of Business Analytics (BA).

Tables that we will create in this level are all prefixed by "STG" as Staging Area, that corresponding to the total import of the source document with no changes to the schema and with only small changes due to the capacity of the database variables of SQL Server, used for the project, to avoid the data truncation errors.

2.3.1 Metadata

The term "metadata", in the context of data warehousing where they play a substantial role, they indicate the sources, the value, the use and functions of data stored in the DWH and describe the altered and processed data during the different levels of architecture.

So, the metadata tables are closely linked to the DWH and its applications are use from both supply side and analysis side.

It is possible to distinguish two categories of metadata, according to various uses

• Internal: Of interest to the administrator, describe the sources, transformations, power policies, logical and physical patterns, constraints and user profiles;

• *External*: Of interest to users, are, for example, the definitions, the quality, the units and significant aggregations.

The metadata is stored in a special container which can access all the other components of the architecture.

It can be classified also about the level that consider:

• *Global*: Contain metadata related to all levels and processes;

• *Process*: We distinguish the metadata depending on the feeding system and the process in which they are involved. The metadata describing the individual process or less related to a particular system (internal sync point between tables, their percentage of fault-tolerant system) must be own for each of them.

The concept of metadata is very critical within the DWH management and is often debated, whether or not, to keep it inside the project or manage it with external logic that release the technology and the product used from the actual purpose of the metadata.

Normally, a metadata is recorded in a table, for usability reasons on the part of the heterogeneous systems, and it has the utility to describe uniquely and precisely in what state is a process, in order to avoid multiple instances launches, launch a process at a specific time, say if the process is completed in correct or with errors or provide the time interval in which the process has been completed or running to extracted the data.

Through the capabilities of ETL software Talend Open Studio it has been possible to import various file types to build a new Database (Data Mart) in SQL Server called FASHION_RETAIL, for evaluating the entire sales department.

To use it on Talend I need to create a database connection for each level, where I will implement the anagrafiche (all the dimensions tables) and the movements (all the fact tables) by importing metadata.

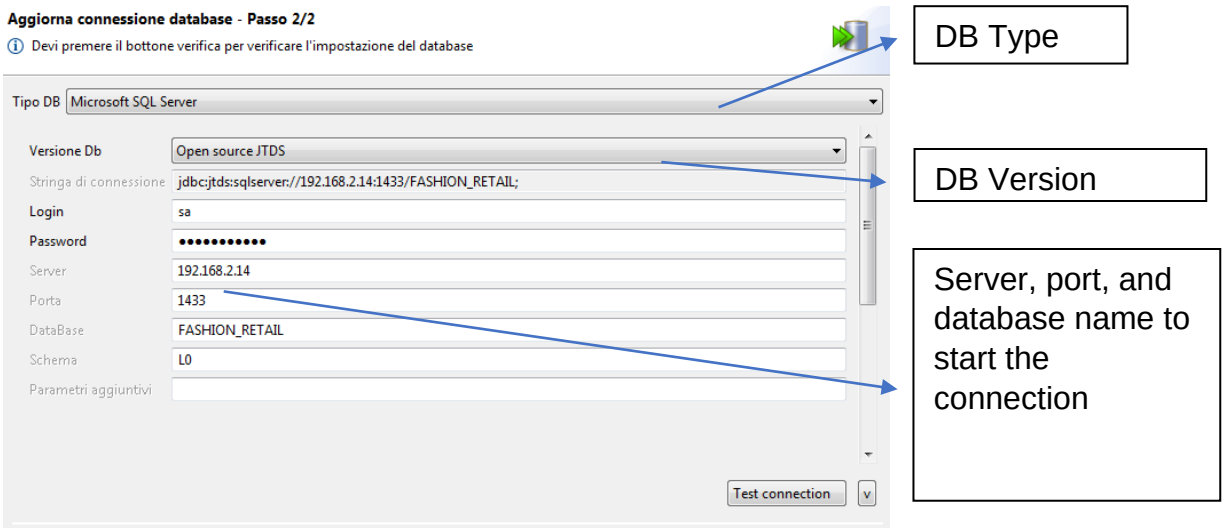


Figure 15: Connecting to the Database

The table below lists the tables created at the L0 level - Data Ingestion with the extracted source files for the project and the mainly types: Excel or .csv (Delimited Files).

Table 5: Metadata

Database Table	Metadata Type	Metadata Name
STG_Causale	File: .Csv	Reasons
STG_Made_In	File: .Csv	Made_In
STG_Made_In_Pianificazione	File: .Csv	Made_In_Pianificazione
STG _Product	File: .Csv	ProdottiS15, Products <S15, Products <S15n2
STG _Shop_Open	File: .Excel	Open Shops
STG _Shop_Closed	File: .Excel	Close Shops
STG_Fam_Colore	File: .Csv	colors
STG_Fam_Materiale	File: .Csv	Materials
STG _Sales	File: .Csv	PIAF, Scontrini2019
STG _Provincia	File: .Csv	province
STG _Counter	File: .Csv	Visitor counter, Contapersone2019

Having a set of files in CSV and Excel, the phase of import in the software development, at the storage level of the ETL process, will incorporate them into folders, one for each type, in the form of metadata.

The project implemented in Talend will have these two types of loading views:

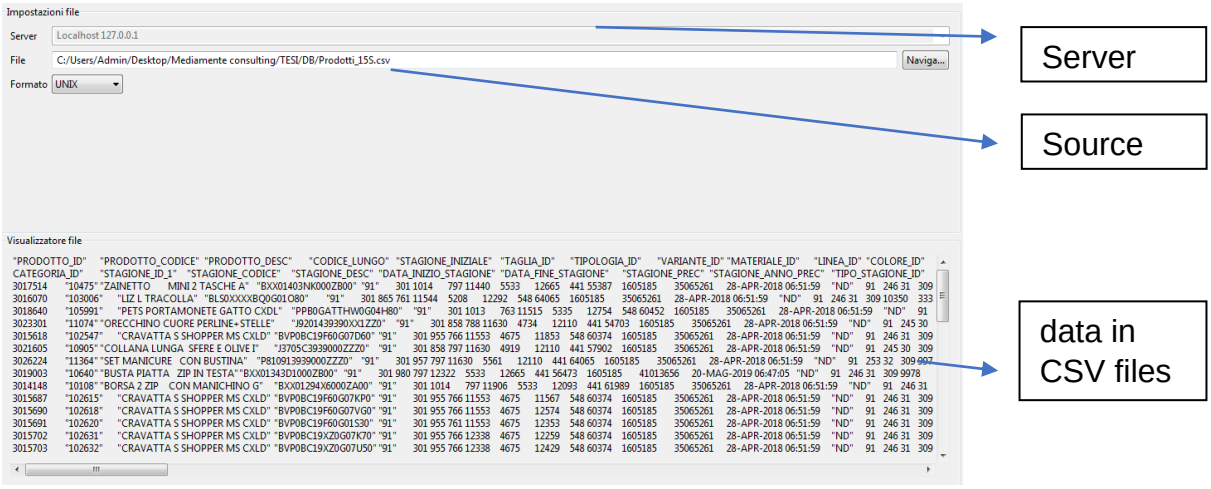


Figure 16: Loading from Delimited File

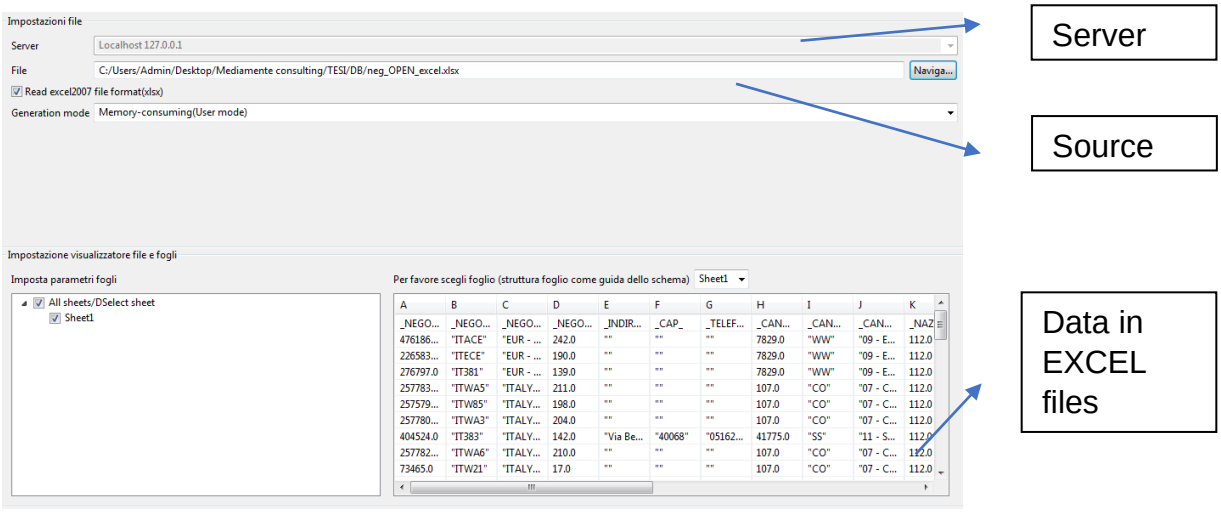


Figure 17: Excel File Upload

Metadata, once imported, must be reworked to create the database on the server. Conceptually, for every job created in Talend, the following four transactions took place:

- Import csv or Excel file;
- Union files via the Tunite instrument, if necessary;

- Change and mapping names, lengths or types of attributes or JOIN tables thanks to primary keys via the TMAP tool;
- Creating a table in SQL SERVER using the tool tDBOutput (tMSSQLOutput).

To perform these tasks, you need to create a new JOB, which will contain the various import methods.

A very significant example is the creation of the STG_PRODUCT, the table in the staging area of the Product dimension.

In it you can observe how the import of various files merged via a Palette called Tunite, which captures the patterns of different source files to create one that suits everyone; if sources have different schemes you will be marked with a warning, even if, the process continues to operate normally.

TMAP tool (2.2.4 data quality) is used to transform the input patterns to optimize outputs (L1) and create join relationships among the various tables.

In this specific case, it helps to identify which products are accepted according to the scheme previously defined and which rejected, respectively, with the creation of the staging area of the product and an excel file with the rejected products.

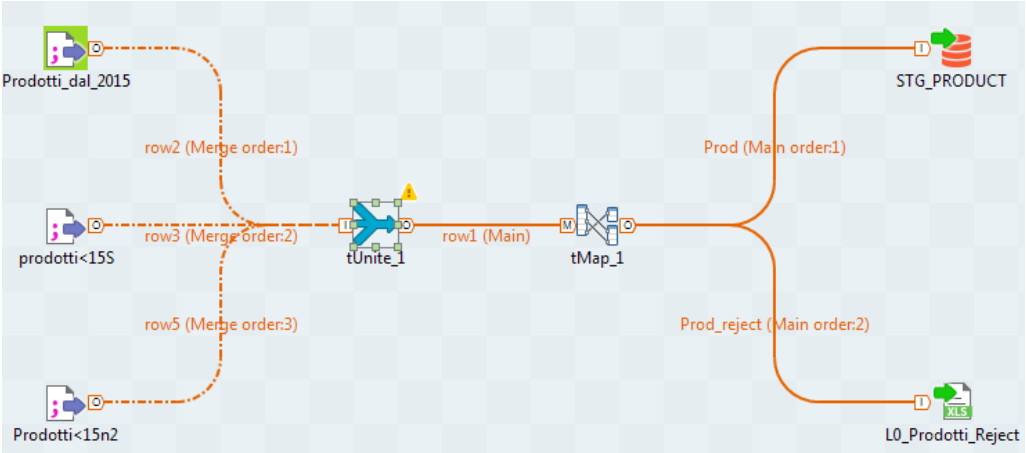


Figure 18: Multi-Loading

The same procedure was carried out for the other tables, with the most common case defined by the direct import of files into the database.

It is important to note that each file is made on a different Job (work area). This is necessary to prevent errors during the data loading or to isolate problems.

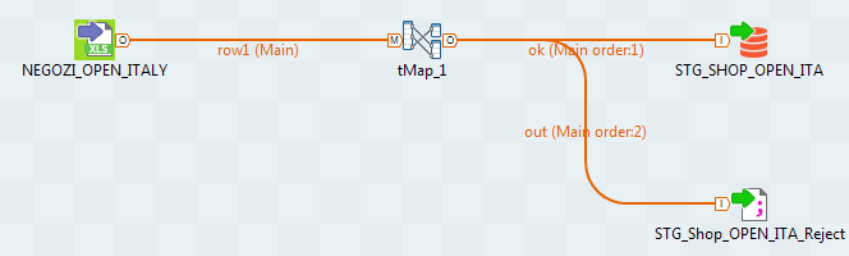


Figure 19: L0 Single Loading

2.4 L1 LEVEL - OPERATIONAL DATA STORE

The L1 level include data quality, data normalization, and all the transformations of the source files, and is definitely the heart of the entire ETL process.

Unlike the L0 extraction level, data are extracted exclusively from the tables created in the Staging Area previously created, to then be subsequently processed and loaded in the same Database FASHION_RETAIL, but in a normalize manner, to enhance the process and have a continuous control over activities taking place.

For tables already in the Staging Area the process is very simple, as you only have to select attributes (mapping) in the TMAP that interest me and make the appropriate changes for the normalization of the data. In the transfer, after a first check of integrity, they create the Primary Key, that will go to uniquely identify the specific attributes of each table, often represented by an ID field.

Important phase is the Pre-Loading, loading of the entities present in the other dimensions to guarantee referential integrity.

Using as input the STG_PRODUCT table, for example, we can observe the possible formation of new dimensions and perform specific queries for each of them, extracting univocally from the source table fields related to the category, thanks to the "SELECT DISTINCT" function of the SQL language.

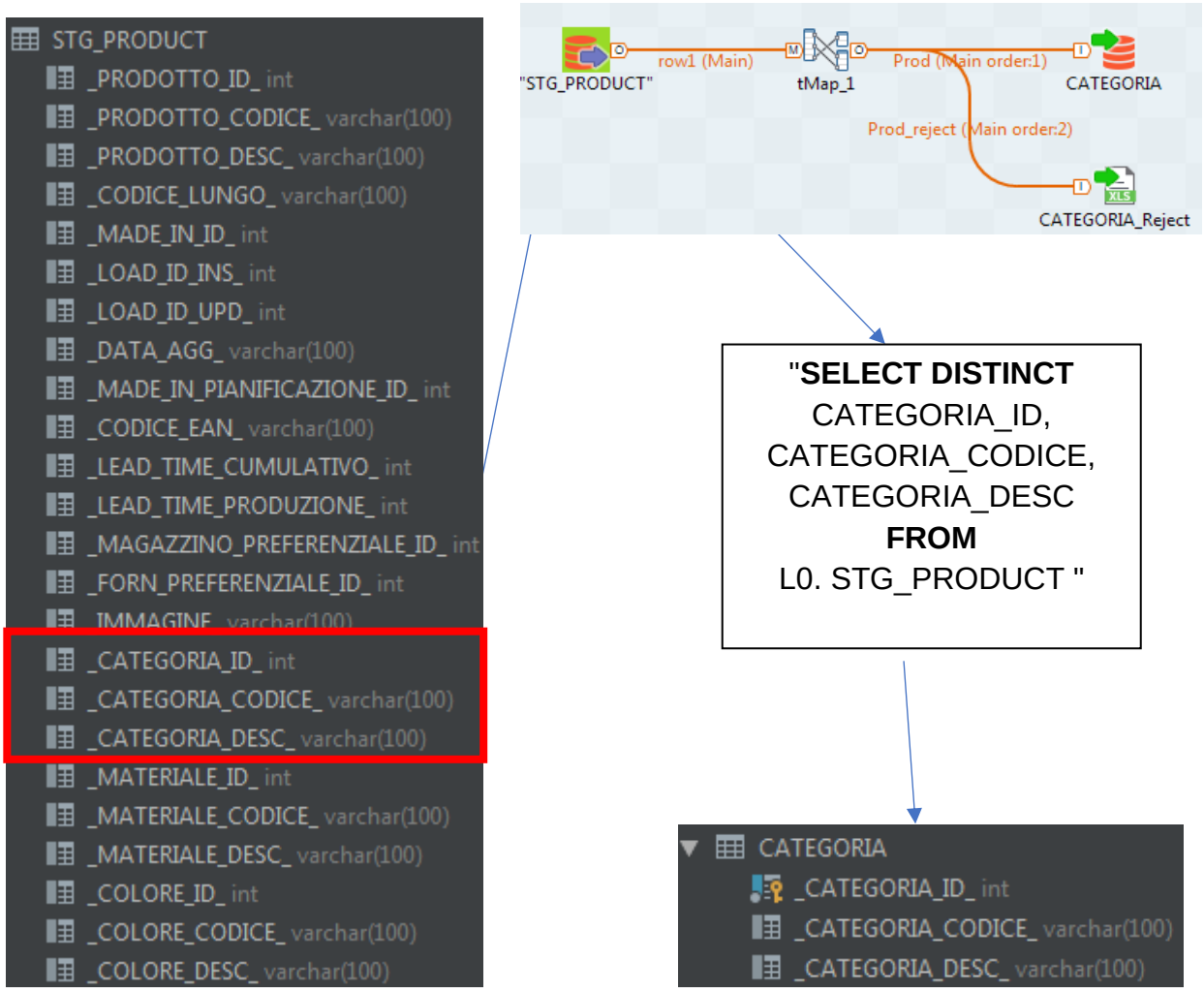


Figure 20: Pre-Loading L1

2.4.1 Data Quality

Data quality is a perception or evaluation of the suitability of the data for a purpose in a given context. Data quality is determined by factors such as accuracy, completeness, reliability, relevance and frequency. Since the data have become more closely linked to the operations of the organizations, the emphasis on data quality has gained more attention.

The check is divided mainly into two parts:

- *Referential integration*: Control through the foreign key checks. It is carried out by means of join with the L1 tables containing the fathers of which verify the relations;
- *Record validation*: Data must undergo checks to reject records that do not meet the requirement not null or other simple and complex conditions (date falling in intervals or text fields of defined length etc...)

The following table shows (one for each type) all the data quality operations performed during the design of the data warehouse, in the TMAP tool.

In some cases, they were sometimes further processing at the metadata level, particularly in the size of fields. These changes have been made to avoid the data truncation, in order to avoid subsequent inconsistencies in the final data.

Poor quality data is often considered as the source of inaccurate reports and strategies in companies. The economic damage due to data quality issues can range from miscellaneous added expenses when the packages are sent to wrong addresses, up to fines of regulatory compliance for improper financial dealings.

Table 6: Data Quality

Data Type Source-Destination	Source	Transformation for Data Normalization	Destination
Datetime-Date	"Dd-MMM-yyyy hh: mm: ss"	Change in the variable type directly in TMAP options	dd-MM-yyyy
String-String	"Piemonte"	Regione.toUpperCase()	"Piemonte"
String-String	"Trentino Alto Adige"	Regione.replace("-", "")	"Trentino Alto Adige"
String-String	Bolzano/Bolzen	Provincia.replace(“/ Bolzen", "")	"Bolzano"
String-String	"Piemonte"	Regione.trim ()	"Piemonte"
Integer - Integer	null	Totale_ Arrivi_2018 == null? 0: Total _Arrivi_2018	0
Integer - Integer	23		23
String-Double	"23"	Double.parseDouble (DISCOUNT)	23.0

2.4.2 TMAP Component: Talend Open Studio

The TMAP component [24] is one of the main components of processing in Talend Open Studio and is used mainly for mapping the input data to output data or a source pattern on a target.

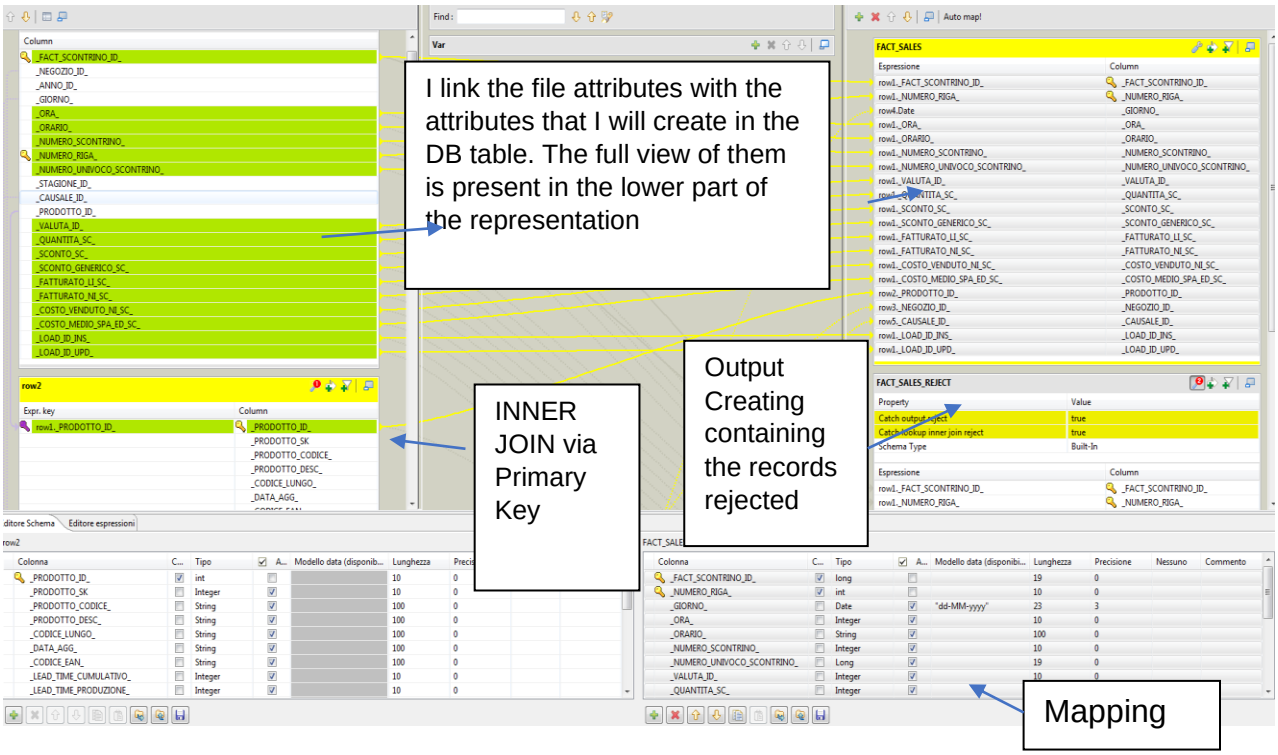


Figure 21: Join & Mapping in TMAP

In addition to perform mapping functions, the Tmap can be also used to merge multiple input tables combining data into a single target table.

All transformations previously listed in the table refers to the Data Quality and they are carried out in Tmap with a further possibility to filter data.

A mapping expression must be done for each columns of the input patterns. For do that you can use the editor of Talend that contains any available Java class or Talend routine.

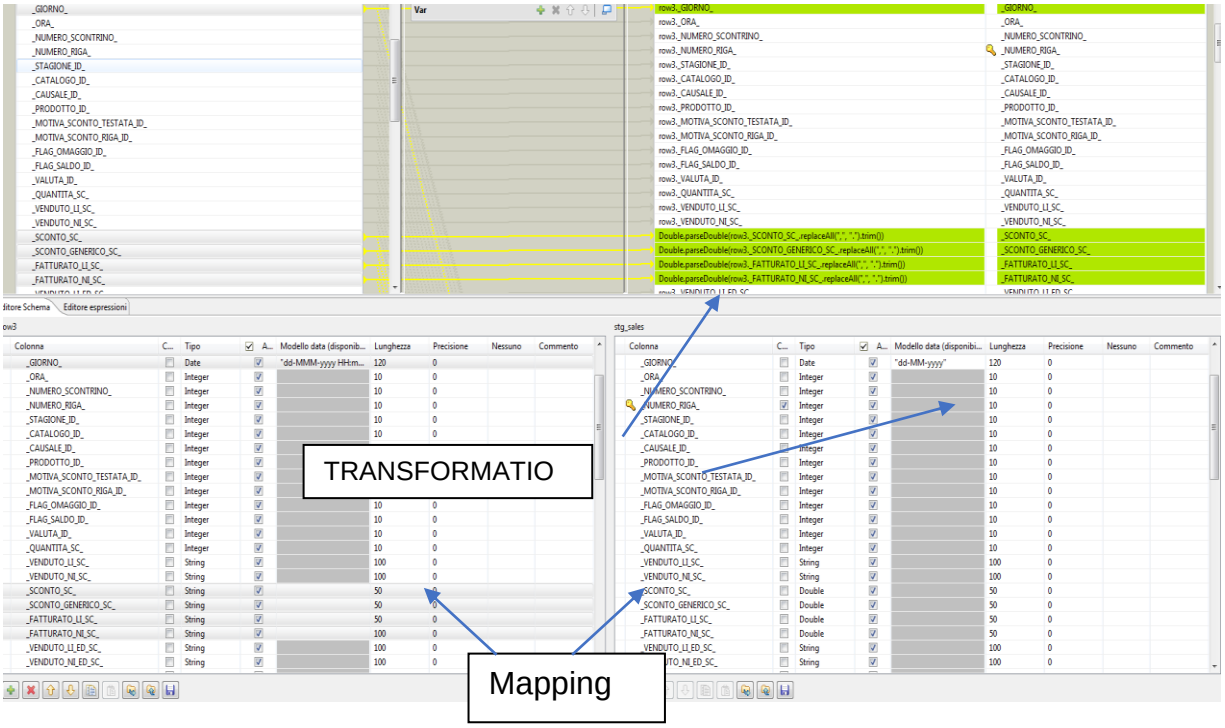


Figure 22: Transformation in TMAP

As you can see from the examples above, each mapping may have a different complexity according to our needs.

In the best practice you should always make sure that the data and the created models will be always available for reuse. In fact, it is very likely that in a recent future, we can extend our ability to a specific attribute of a table, for example by controlling the format. Therefore, to have a code and good process performance is better to change the logic in a single routine, rather than operate several times in the individual mapping expressions.

2.5 LEVEL L2 - DATA PRESENTATION BEST PRACTICE

In the new data analytics projects, one of the best practical implementation of a data warehouse or a data mart is the Snowflake model, mainly developed to have an ETL process effective and fast.

In particular, my goal is to bring a SQL Server relational database, a design level in Snowflake Style:

- Data must be classified and marked appropriately, especially if highly protected;
- Data must preserve its history through audits and verification of data that will have to be validated at the source, if possible;
- Data must be processed in micro-batches;
- The data must be ingested and loaded and then processed according to the rules of ETL or ELT;
- Process the CDC end-to-end data to avoid performance problems;
- Create separately data mart instances according to the specific requirements of the company's business, creating systems of governance.

There are some best practices that apply to the implementation of a Snowflake Database, specific for its unique architectural differences with other relational databases or platforms of big data:

- Use independent DWH Multi-Cluster with storage capabilities and scalability of shared data to optimize processing requirements of different workloads. For example, the Staging Area (Level L0) can be located on a different database than the Core Layer (Level L1);
- Assign a separate virtual data mart for each business to have an optimized data consumption;
- Keep, if possible, the semi-structured data in its original format to increase processing performance data. Often the JSON data is processed faster than those converted into relational tables. When you store semi-structured data,

Snowflake optimizes the storage on the basis of repeated elements within the structure of semi-strings;

- Upload the data into smaller chunks instead of one large file and load them in parallel using multiple nodes. For one customer we were able to charge 24 months telemetry event data in less than two weeks with a small cluster nodes;
- Assign virtual clusters to separate patterns of a data warehouse to optimize the performance
- Consider the clustering of large tables to improve query performance.
- Recluster if performance degrades. Sometimes the cluster / reclustering may cause poor query performance, although it is recommended to analyze the data in the table before you make these changes;
- Snowflake stores the metadata (min and max values, distinct values, etc.) effectively reducing the stock micro partitions needed to scan a query;
- Is recommended running the ingestion of data based on events to allow the chronological order of the data.The data pipeline based on flexible framework must be completely separated from the actual processing structure. This operation is done by the creation of type Integer surrogate keys (SK_KEY) to improve the efficiency allowing quick connection and loading of data for the large tables. This allows minimum changes to the code if there are any changes in the target or in the source schema, still be able to perform all the loading operations, transformation, aggregation and processing of data;
- Building a robust audit of the budgetary control framework that tracks not only the lineage of data and data quality, but optimizes database performance than the processing costs;
- Create database clones for testing or for validation to avoid duplication of data.

Snowflake is a powerful model with unique features that enable rapid implementation of data analysis projects, but as all other databases requires careful planning. Following the best practices just listed, the first part to play in the project is the creation of the Surrogate Keys for all data tables of the level L1, creating parent and child relationships between dimensional tables.

For example, with regard to the pre-loading of the product table, surrogate key will be created with the following SQL:

- ***ALTER TABLE*** *L1.PRODUCT ADD _PRODOTTO_DM_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.ALTEZZA_TACCO ADD _ALTEZZA_TACCO_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.CATEGORIA ADD _CATEGORIA_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.COLORE ADD _COLORE_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.GENDER ADD _GENDER_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.FAM_COLORE ADD _FAMIGLIA_COLORE_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.FAM_MATERIALE ADD _FAMIGLIA_MATERIALE_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.MADE_IN ADD _MADE_IN_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.MADE_IN_PIANIFICAZIONE ADD _MADE_IN_PIANIFICAZIONE_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.MATERIALE ADD _MATERIALE_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.MODELLO ADD _MODELLO_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.STAGIONE ADD _STAGIONE_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.TAGLIA ADD _TAGLIA_SK INT IDENTITY (1,1) NOT NULL;*
- ***ALTER TABLE*** *L1.TURISMO ADD _TURISMO_SK INT IDENTITY (1,1) NOT NULL;*

At this point, through the creation of a connection to the schema L1 of Fashion Retail database in the form of metadata, have need to proceed with the extraction of the tables having the Surrogate Key present in it, that, through the processing in Talend Tmap tool [24], will become the primary key of the new table size.

Specifically, it is seen how the input "row1" defined as the table category of level L1 have as primary key the Id, is set to level L2 with an identical scheme to the previous one, but with primary key, the surrogate key. In general:

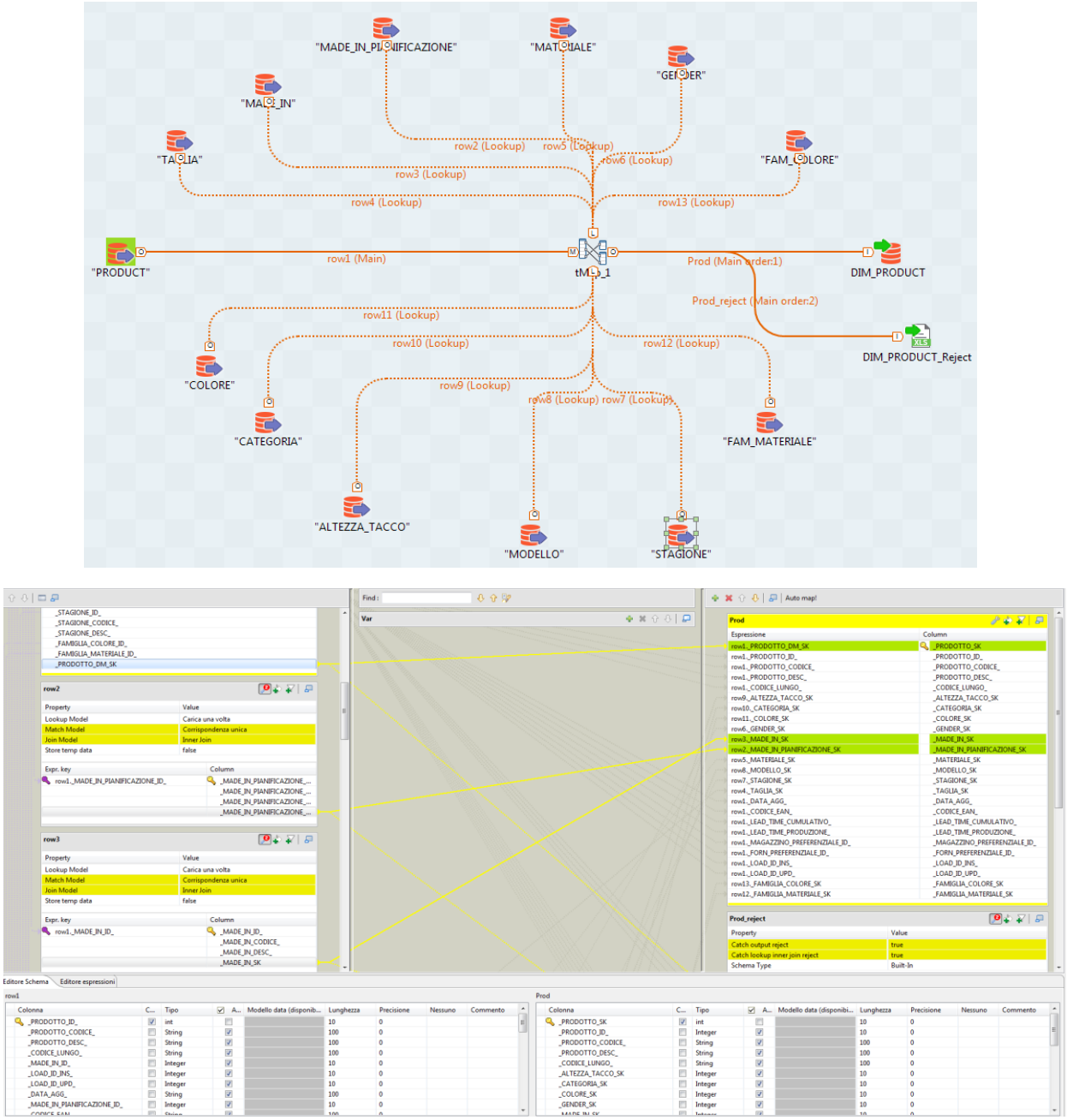


Figure 23: Snowflake Dimension

Moreover, it is very important to be careful in the construction of the fact tables, as, being the father tables of dimension tables, will not have its own surrogate key. In them, it will be shown only the surrogate key of dimension tables that characterize it, ignoring all the information such as, for example, the Id and the description, since, in this process has no need to go into detail in a single table, but it is a tree structured system with a system

of hierarchies, defined precisely by Joins between the surrogate keys of dimension tables.

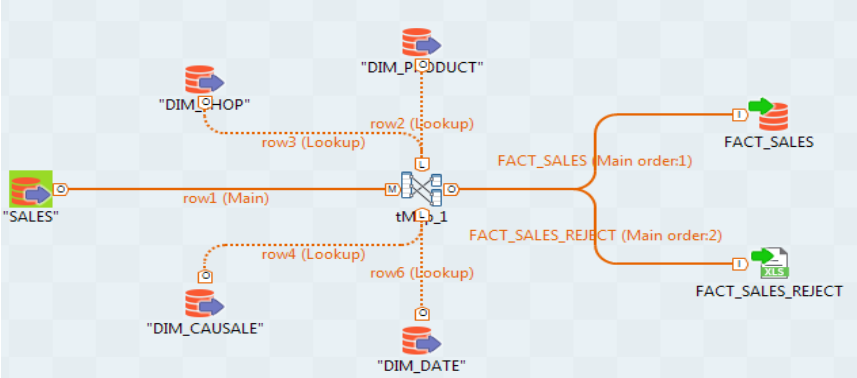


Figure 24: FACT Snowflake

2.6.1 Snowflake Schema

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 schema.

The name comes from the fact that the dimension tables branch and resemble, like a snowflake. Observing the model, is highlighted as a fact table is surrounded by the dimensional tables, with which it will create the branching. Unlike the star schema, tables of dimensions in a snowflake schema may have their own categories. The dominant idea behind the scheme is that Snowflake dimension tables are fully normalized. Each dimension table can be described by one or more lookup tables. This is repeated until the model is not completely normalized.

Obviously, the normalization creates a greater complexity in performing the query of the snowflake schema, as, for example, we will have to dig deeper to get the name of the type of product or the municipality of a store. The structure is based on a series of nested JOIN, where we must add to a simple JOIN another JOIN for each new level within the same dimension. Of course, there is several standard annidation, but depends on the given level that you want to extract. More data is in depth, the more the process of writing query will be complex [23].

Basically, a query against a data mart based on Snowflake schema will run slower than one in Star schema. In most cases, this is not a problem: it does not matter much if we get the result in a second or a millisecond.

In the proposed project, a comprehensive view of the snowflake is as follows:

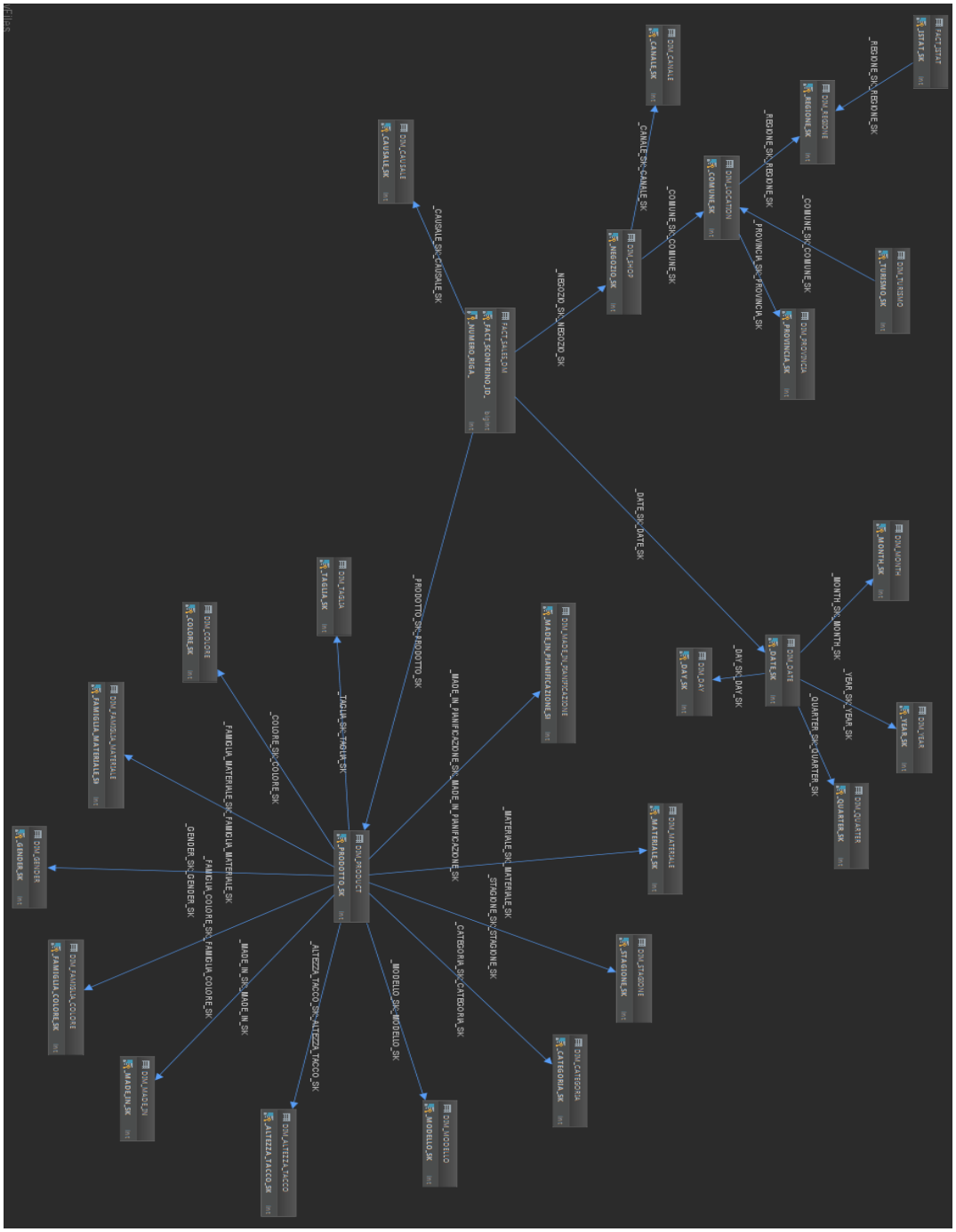


Figure 25: Snowflake Schema

2.6 LEVEL L2 – PRESENTATION AREA BEST PRACTICE

The Relational Database, the most widely used, compared to Snowflake exhibit the same levels of Staging Area and Transformation Area but with a substantial differentiation in the final level L2.

In this case, the goal is not to have an ETL super powerful process, but to have fewer end large tables with more information to facilitate, through data visualization software, reporting to give effective future directive to implement better decision strategies or to provide a simple financial audit on the economic performance of the company.

The Job schema does not change compared to the Snowflake but change the mapping of variables in the Tmap [24].

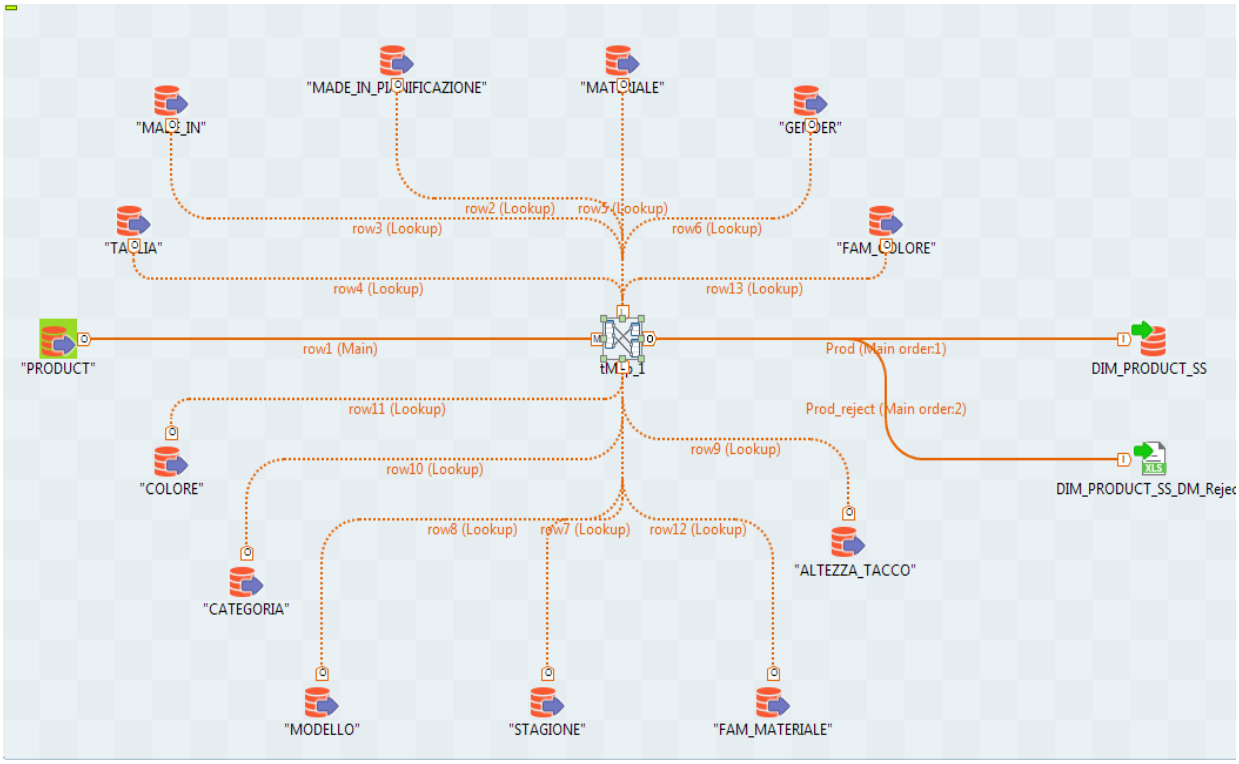


Figure 26: Job Product Star Schema

In fact, the tables of level L2, in this case it can defined like aggregate tables of level L1 tables and it is possible to obtain them with a series of JOIN between tables, no longer connected to the Surrogate Key, but directly to the Primary Key, with the necessary integrity checks.

The final table attributes can also result from different tables, as each attribute of a table is connected to the attribute of the final table through the mapping created by JOIN operation.

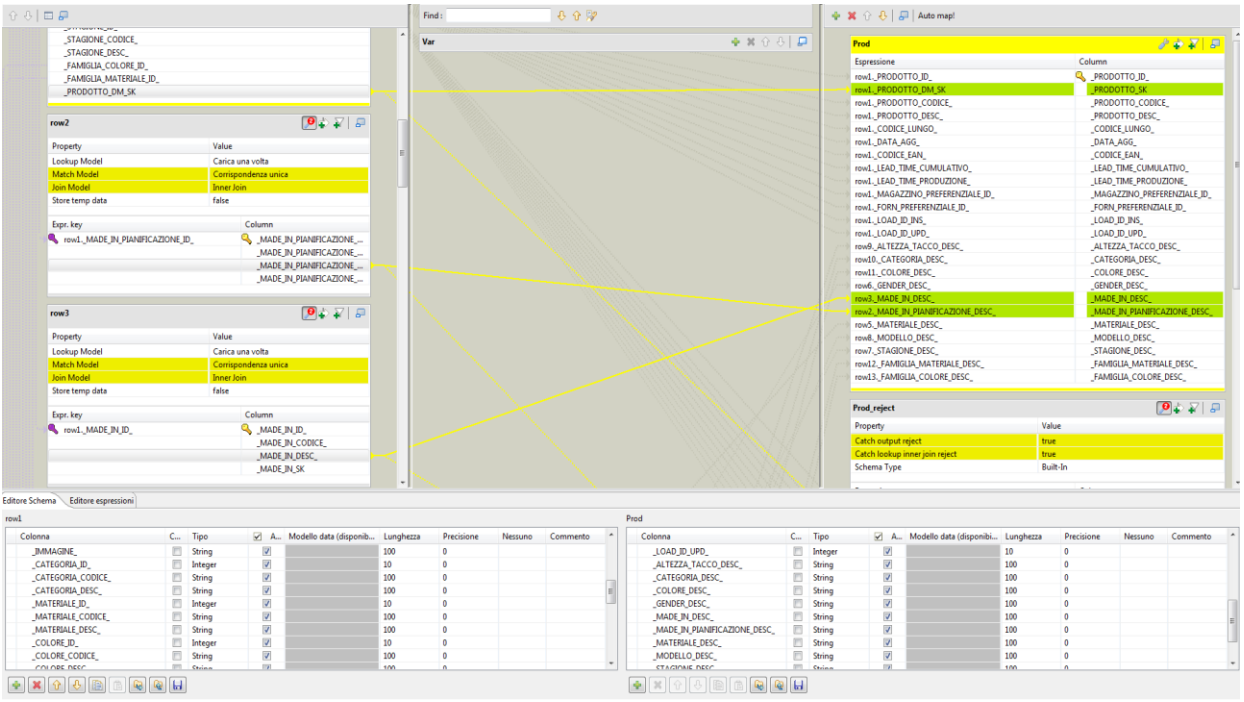


Figure 27: Product TMAP Star Schema

The same procedure was done for the fact tables, remembering to make JOIN not with the primary keys of the dimension tables, but make them directly with the surrogate keys, because the fact tables are in 3NF, like in snowflake schema.

2.6.1 Star Schema

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 many business intelligences tools;
- The fact tables in a star schema is in third normal form, while the dimensional tables are de-normalized [10].

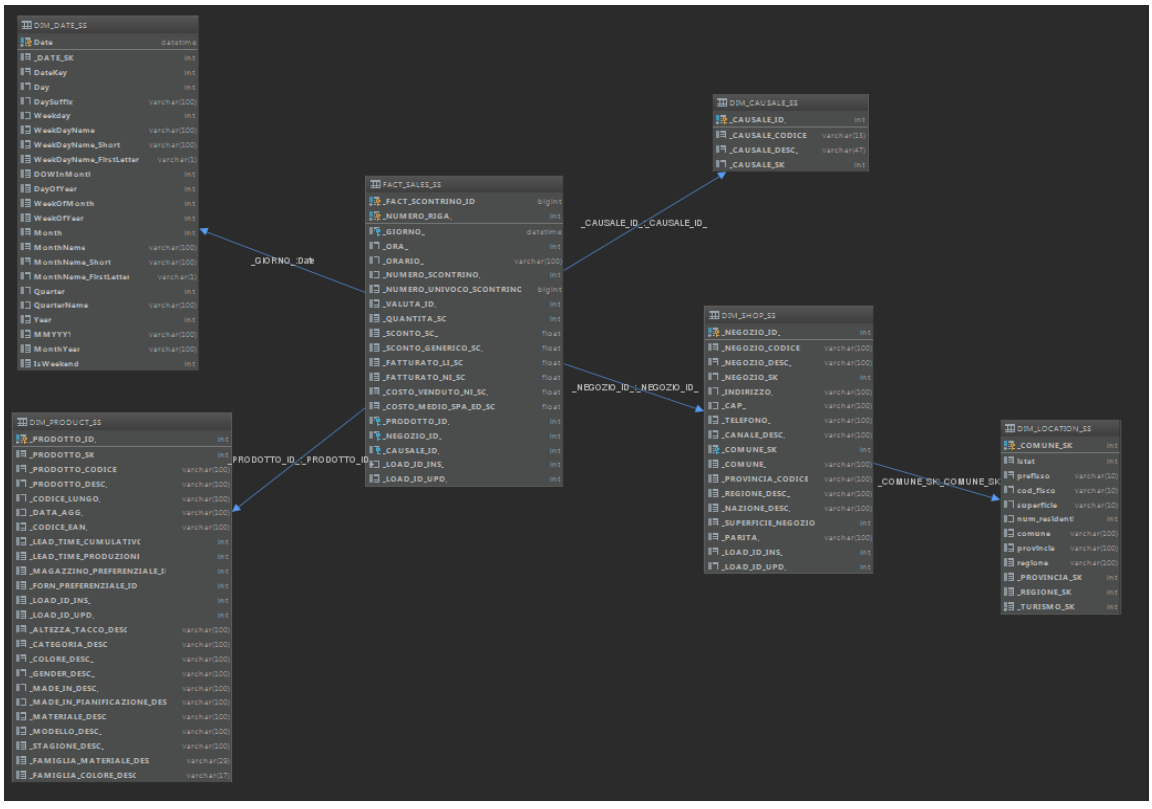


Figure 28: Star Schema

2.7 FULL LOAD ETL

To speed the whole process explained above, the best method is create jobs that contain other job. In this way, I can enclose into subgroups the master data and the movements for each level, and then, run them all at the same time.

For example, the job STG_ANAGRAFICHE in the level L0, where I collect all of the job of level L0 relating to all master data, as shown in the image below.

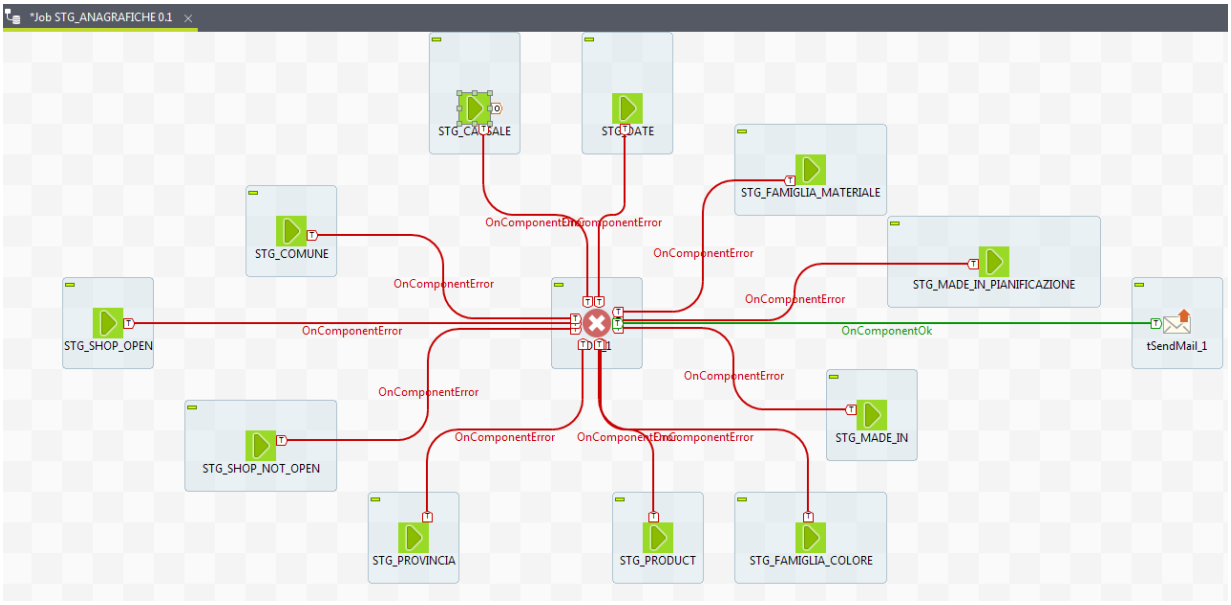


Figure 29: Job STG Anagrafiche

The same procedure is done for the master data and the movements of each level until the L2 level with only two jobs to be joined, respectively, one relating to the master data and to the movements.

As a final step of the ETL process, it must combine the two types of configuration data with all their hierarchy in one final job. The advantage of Full Load ETL is the ability through a unique command to load entirely a database.

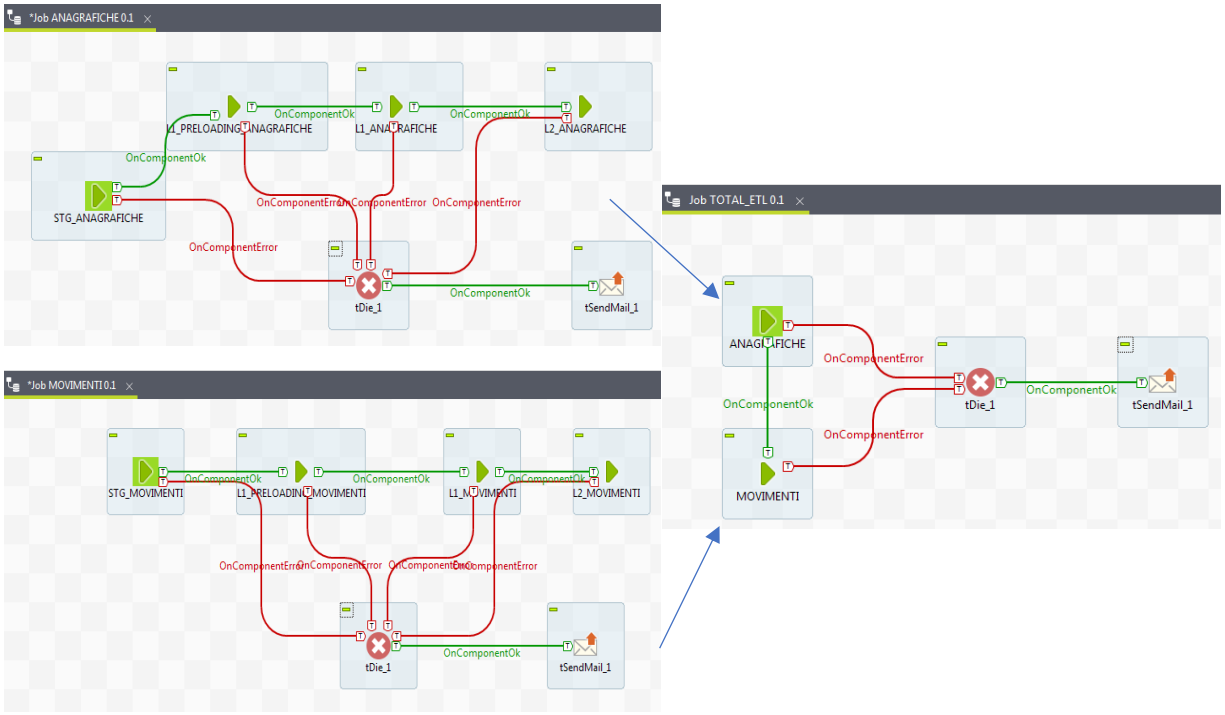


Figure 30: Full Load ETL & Auditing

It is very important to check the flow of data in the various stages of the process. In the figure above because, in addition to the last step is important to focus on two tools: TDIE and tSendEmail. They work together, and when the process is in error, it sent a signal to the TDI e, which by the help of tSendEmail tool, send an email with the attachment in the form of error Script, both the owner and the manager Database.

An example covering the entire explanation process is explained in Appendix A5.

2.7.1 Auditing ETL

The control in a process of extraction, transformation and loading has the aim to satisfy the following objectives:

- Check the data anomalies as well as controlling just the serious mistakes;
- Capture and store an electronic track of any material changes to the data during processing.

The ETL auditing helps to confirm that there are no anomalies in the data even in the absence of errors. A well-designed auditing mechanism also adds to the integrity of the ETL process by eliminating the ambiguity in the transformation logic, trapping and tracing each modification to the data along the path. Even in the most rudimentary ETL architectures, you can check out some high-level metrics to confirm that the loaded data are those provided.

In general, the ETL auditing processes should ensure the following to confirm that correspond to the input output:

- Counting general of the rows;
- Total aggregate (which may include financial amounts or other summary data).

Some processes require a more comprehensive audit. In other cases, it might be necessary to check whether the data are within reasonable limits or if they support these values. Another aspect that we must not forget to check the cases in which no data has been loaded. Unfortunately, it often happens, and the two main causes are due to a source file that contains no data, an incorrectly configured query that returns no rows or a directory of empty originally intended to contain one or more files could lead to the successful completion of ETL process to load but exactly zero rows of data. However, if a given process should always include an uploaded file number other than zero, be sure to add a control step to check it.

The ETL auditing is rarely the most visible element in the architecture, but it is an insurance policy necessary to protect the integrity of the data and the process.

CHAPTER 3: DATA MINING ALGORITHMS FOR FASHION RETAIL

In recent decades, the development of information and communications technologies have given new vitality to the company's marketing. The data to be stored and to be analyzed are increasing at a very rapid pace, probably 1000 times compared to five years ago. However, data and corporate earnings are not directly proportional.

These applications of Data Mining algorithms are referred to a Boundary Science, a variety of scientific theories based primarily on the basic disciplines of Information Technology, Marketing and study of the statistical methods, that is the basis of every possible algorithm. In addition, data mining also refers to literary and behavioral disciplines to better evaluate the characteristics of a customer, such as psychology and sociology [20].

In general, through the extraction, transformation and loading of a large amount of information, we are able to identify the interests, preferences and behaviors of specific groups or individual consumers, but above all, the forecast of consumption, orienting sales for the specific marketing.

Since automation is popular throughout the industry, the companies that manage the processes must have many operational data. The data are not collected for the purpose of analysis but originate from commercial operations. The analysis of these data gives decision-makers the real value of the information, in order to obtain profits.

The commercial information coming from the market through various channels, for example, one of the most popular is the purchase process by credit card where we can collect consumer data, such as time, location, assets or interesting services concerned,

took prices and the level of capacity receipt. In addition, companies can also purchase a variety of customer information from other consulting firms.

The marketing based on data mining can usually create specific sales promotions for the client according to his previous purchase. The most common applications are in banking, insurance, traffic system, retail and commercial matters.

As described in State of the art, technology and marketing analyzes are based on the analysis of the market, such as prediction, segmentation and classification of customer profiling and cross-selling. They can also be used for credit scoring and fraud operations.

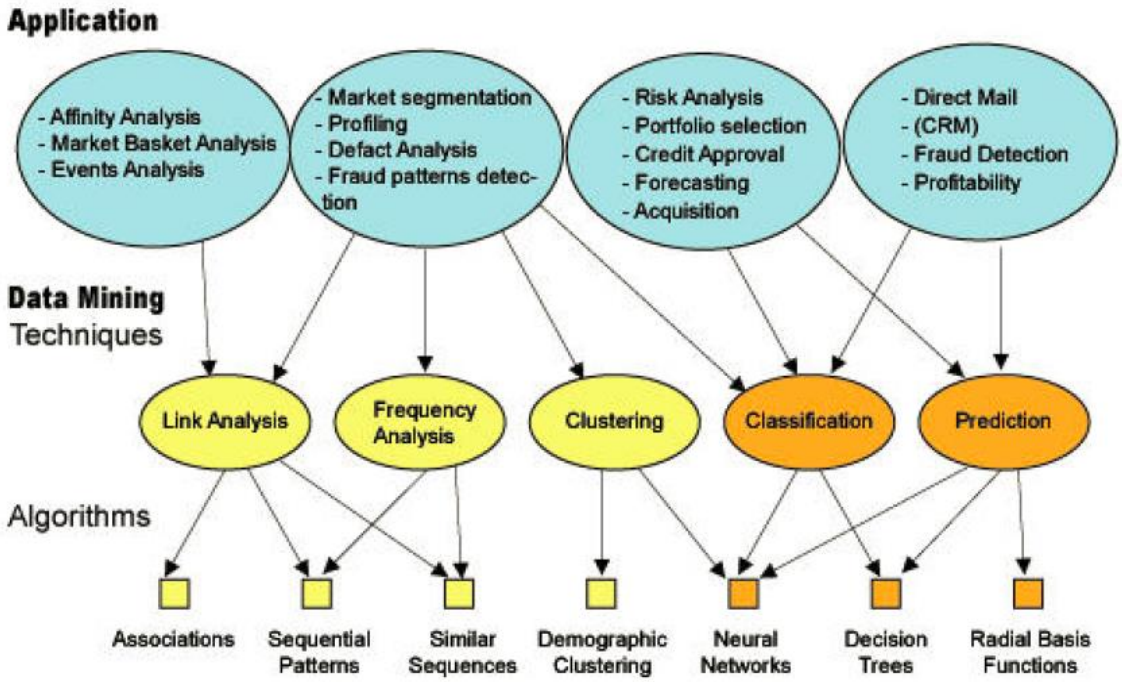


Figure 31: Applying Data Mining in Marketing

The basic process of data mining in marketing shows as follows:

• *Prepare the primitive data:* Includes personal information (age, gender, hobbies, background, profession, address, zip code and income), the earlier shopping experience and customer relationship. The preprocessing of the early data is very important to select potential customers,

• *Establish a certain pattern*: This model can be created by using very traditional technologies mining technologies. However, the problem of these technologies can be solved to identify the best or acceptable market within source of limited information, limited time and limited expenses.

Ultimately, we are using this model to select customers and decide the marketing plan.

In our project we are going to idealize a possible prediction of the ISTAT 2018 Italian data using linear regression, starting from 2007 data until 2017 by the eponymous site [25]. In addition, I try to find a possible best geo location to open a new store. and, in a second time, I will develop a classification CART to understand future expectations of the stores are currently open.

3.1 PREDICTION OF THE BEST GEO - LOCATION TO OPEN A NEW STORE

For a complete and reliable analysis, it is very important the integrity and completeness of the data. After a search in various sites dedicated to open data, I have chosen the data given by the National Institute of Statistics, called ISTAT [25].

The actual data have a horizon starts from 2004 and finish in 2017. For make a good analysis of 2019, the data provided cannot be considered complete. Therefore, it was decided to address the problem by considering a time horizon of ten years, considering the data from 2007 to 2017 to predict also the 2018.

Research to determine the appropriate indicators to the analysis, was held following a matrix process, making a mapping divided into geographical areas and time series covered, inserting all into an explanatory table shown in Appendix A2.

The preference at regional level are expressed with a total of five indicators, one or maximum two for each sector:

• Transportation Sector: operating rail network;

• Family Sector: Average monthly household expenditure on non-food goods and services, average income;

- Macro-Economic Sector: GDP Per Capita;
- Work Sector: Unemployment Rate.

While, at the municipal level, there was evidence the resident population and tourism.

Before the implementation of the code, it was necessary to create worksheets that are unique to each indicator, favoring an efficient data prediction, given the diversity of origin of the same. The subdivision, to avoid complicated mechanisms and complicated process and any copying errors, was made with Talend Open Studio [28], ETL software already widely discussed in the previous chapter.

The job takes Complete data previously loaded into the Staging Area and through the use of queries, each table is interrogated to extract the dedicated Excel spreadsheet, that will be the regional data used to our analysis.

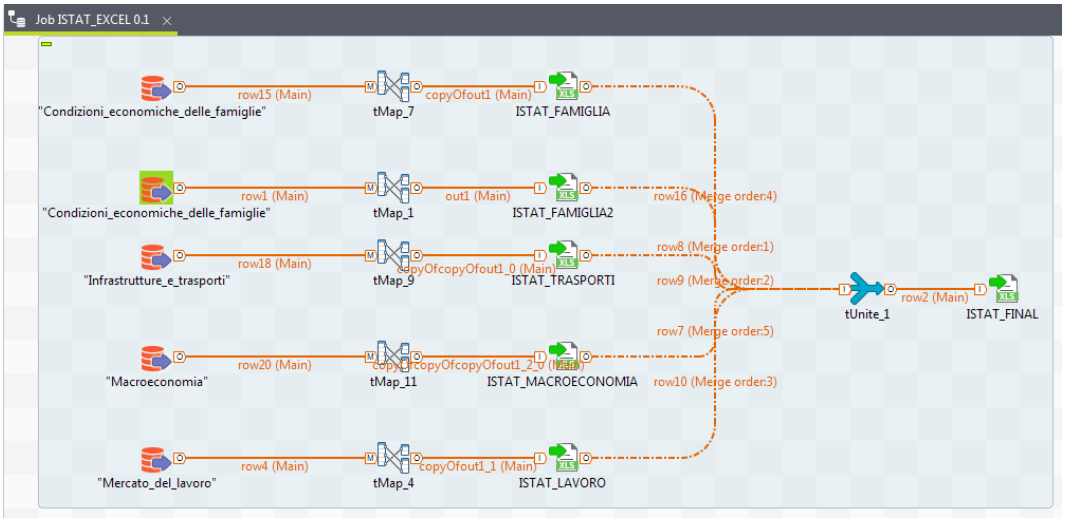


Figure 32: Job ISTAT Excel

For example, to extract only the data for the *'Spesa media mensile familiare per beni e servizi non alimentari'*, it is considered the table dedicated to the economic conditions of families of the ISTAT highlighting the SQL IN clause. The same thing was done to exclude national data or related to membership in areas to be excluding (North Central, Northeast, ...) with a NOT IN.

Of course, the same operations were conducted for the other indicators.

The query used in the case of the average monthly expenditure is as follows:

"

***SELECT** **

```
FROM OPEN_DATA_ITALY.Condizioni_economiche_delle_famiglie
WHERE Indicatore IN ('Spesa media mensile familiare per beni e servizi non
                      alimentari')
 AND Territorio NOT IN ('Nord-ovest',
                        'Bolzano/Bozen',
                        'Trento',
                        'Nord-est',
                        'Nord',
                        'Centro',
                        'Centro-Nord',
                        'Mezzogiorno',
                        'Italia')                    "
```

By creating Excel files unique to each indicator, it is now possible to perform Linear Regression.

3.1.1 Regression of ISTAT data

To perform the prediction, is used "R", a language or a dedicated environment for statistical computing and graphics. It is a GNU project and provides a wide variety of statistical models (linear and nonlinear modeling, classical statistical tests, time series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The R language provides Open Source option that it used in our project. One of the strengths of R is the ease with which you can produce great quality well-designed plots, and mathematical symbols and formulas where necessary. The cons are that there isn't paid great attention to the defaults for the less design choices in graphics, but the user retains full control [27].

R-Studio is an integrated development environment for R, with a console and an editor syntax, that supports the direct execution of code, Monitoring tools, history, debugging and the workspace management [26].

As said before, the objective is to extract a Y predictive value refers to the year 2018, conditioned by Xi variables identified as the years 2007-2017.

The result is the shape of the multi-variable linear regression:

$$Y = b_0 + b_1X + b_2X_2 + \cdots b_kX_k + \varepsilon$$

A complete example of R code used for linear regression is shown in Appendix A4, where the considered indicator is the average monthly expenditure.

To get the actual results of our regression you must use the command "summary (reg)", which will create an output like the one below, which displays all the most important values of the model to assess the real goodness:

summary(reg)

Coefficients:

	Estimate Std.	***Error***	***t value***	***PR(>|t|)***
(Intercept)	***-****64.51987*	*67.42781*	*-0.957*	*0.36363*
Istat_Famiglie$Anno_2007	*0.36274*	*0.15502*	*2.340*	*0.04402 **
Istat_Famiglie$Anno_2008	*0.07351*	*0.10744*	*0.684*	*0.51106*
Istat_Famiglie$Anno_2009	*0.23755*	*0.12889*	*1.843*	*0.09843*
Istat_Famiglie$Anno_2010	*-0.31284*	*0.20341*	*-1.538*	*0.15842*
Istat_Famiglie$Anno_2011	*-0.04140*	*0.12748*	*-0.325*	*0.75278*
Istat_Famiglie$Anno_2012	*-0.82588*	*0.32808*	*-2.517*	*0.03292 **
Istat_Famiglie$Anno_2013	*1.65055*	*0.38821*	*4.252*	*0.00214 ***
Istat_Famiglie$Anno_2014	***-****0.88867*	*0.37681*	*-2.358*	*0.04271 **
Istat_Famiglie$Anno_2015	*0.29580*	*0.34225*	*0.864*	*0.40987*
Istat_Famiglie$Anno_2016	*0.51256*	*0.18976*	*2.701*	*0.02435* *****

Residuals:

Min	***1Q***	***Median***	***3Q***	***Max***
-49.282	*-17.568*	*2.277*	*18.774*	*37.037*

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: *37.12 on 9 degrees of freedom*

Multiple R-squared: *0.9957*

Adjusted R-squared: *0.9908*

F-statistic: *206.5 on 10 and 9 DF*

p-value: *2.185e-09*

The same procedure was carried out for the other indicators, except for average income, where a further prediction of the 2017 data was needed to provide the data of 2018, as missing both.

How reflects the command output summary, the results are very acceptable. In fact, the estimate of the results of a linear regression model could and should contain:

- An adequate number of observations;
- The estimates of the values of β are acceptable parameters;
- The values of the statistic test called T-test of Student are associated with each parameter in order to assess the significance; these statistics are often accompanied by an indication of the error associated standards, as well as the p-value that is considered acceptable only if less than 0.10, 0.05 or 0.01;
- Statistics adapted to evaluate the overall goodness of the model; these may be limited depending on the case in goodness measure of the fitting which R^2 and R^2 Adjustment for degrees of freedom. R^2 range is between 0 and 1: 0 when the model used does not explain at all the data; 1 when the model explains the data perfectly;
- Statistics of tests such as the F-test, namely the F statistic of Fisher are associated with the null hypothesis that all the elements of β, to verify the significance of the entire model. You want to check H0: β1 = 0, . . ., Bk = 0 against the alternative that at least one of the parameters is different from zero. Under the assumption that errors are N (0, σ^2), the total deviance always admits the decomposition SST = SSE + SSR.

3.1.2 ETL process of ISTAT data

Having dealt with the problem of the completeness of the data provided by ISTAT data, resolved with the prediction, you can load the data obtained with a simple ETL process, as done earlier in chapter 2.

The principle is the same. After having copied the data obtained from the regression in the spreadsheet provided by ISTAT, we will go to create the metadata in Talend [24].

This, it will be placed in a job and loaded to L0 level Staging Area, without performing any operations. The data quality will be carried out by the TMAP [25] in the Layer L1, and then be connected to each other via the Surrogate Key to level L2. In the last two processes I will work directly on the database without affecting the original Excel spreadsheet.

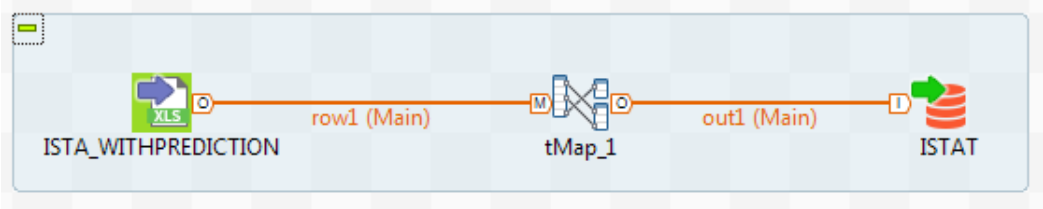


Figure 33: Job ISTAT with Prediction

Of course, the connections will be carried out through the dimension tables refers to the region, which in turn will be linked to the provinces which will be bound to the municipalities, creating a relational level hierarchy. Instead, with regard to the Town table, that already include attribute refer to the population, will be further connected to the table with the data related to tourism. Most important is to note how the table with ISTAT data, will be considered as a fact table, and not as a simple dimension.

A simple visualization of the connections between the various tables is shown in the following figure:

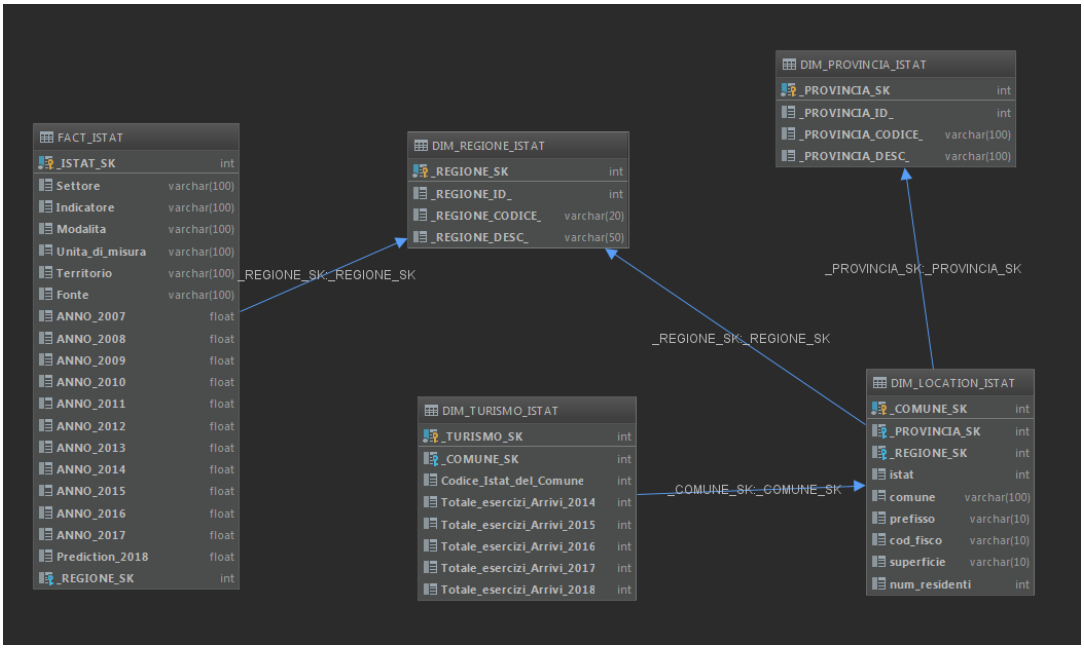


Figure 34: Star Schema ISTAT

3.1.3 Best Geo-Locations

To carry out a complete analysis and obtain a reliable result starting from ISTAT data [25], highlighted and processed via the data quality process, mainly three queries extremely connected to each other have been necessary, with the ultimate goal of finding among the thousands of cities, which are the most economically desirable to open a new store.

Initially, the first and the second queries 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 railway, unemployment rate, Per Capita GDP), while the last, will be extended in depth at the municipal level, filtering for the best three regions obtained through the solution of previous database query.

The first query is a full view of the best and worst three regions for each indicator illustrating Moreover, the corresponding sector and the unit of measurement that characterizes it.

To implement the model, I will create a rank through a partition analytic function, which can be descending, if you treat of positive economic data, or increasing, whether it is data not very favorable to development as the unemployment rate.

In the second query it will instead create a summation factor of various rank grouped by region and sorted by rank descending. In this way, we will get as a result an order of development potential of Italian regions. For it, we will add up the corporate data for the historian of the number of shops closed and still open for each region, and the final value of which will be based our rankings, will take the name of Rank.

As shown in the table below, the three best regions to invest in an economic capital are Trentino Alto Adige, Valle D'Aosta and Friuli Venezia Giulia.

Table 7: Ranking of Regions

Rank	Territory	#Close_Shop	#Open_Shop	Istat_Rank	Final_Rank
1	**Trentino Alto Adige**	6	1	14	21
2	**Valle d'Aosta**	0	0	26	26
3	**Friuli Venezia Giulia**	4	1	37	42
4	**Liguria**	1	0	46	47
5	**Piemonte**	7	5	37	49
6	**Emilia Romagna**	19	7	23	49
7	**Veneto**	11	7	34	52
8	**Umbria**	2	1	51	54
9	**Toscana**	12	5	38	55
10	**Lombardia**	14	17	32	63
11	**Marche**	3	1	59	63
12	**Molise**	1	0	65	66
13	**Abruzzo**	7	1	58	66
14	**Basilicata**	1	0	66	67
15	**Sardegna**	6	0	76	82
16	**Lazio**	23	12	47	82
17	**Calabria**	4	0	83	87
18	**Puglia**	7	4	80	91
19	**Campania**	11	3	87	101
20	**Sicilia**	13	5	90	108

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 deep analyzing based on the number of tourists of the last year and the number of residents for each town.

The solution is the creation of a new indicator obtained through a sum of the number of residents and the tourists, to have an indicative number of potential clients.

The cities with the highest index will be the most interesting.

Table 8: Ranking of Town

Rank	Town	Residents	Tourism 2018	Tot_Possible_Clients
1	**Lignano Sabbiadoro**	6616	691154	697770
2	**Trieste**	201148	414003	615151
3	**Trento**	115540	360388	475928
4	**Riva Del Garda**	16052	428198	444250
5	**Bolzano**	103891	337366	441257
6	**Merano**	37791	328265	366056
7	**Grado**	8434	302626	311060
8	**Castelrotto**	6540	301459	307999
9	**Selva Di Val Gardena**	2657	245928	248585
10	**Badia**	3396	228401	231797
11	**Bressanone**	20921	202598	223519
12	**Pinzolo**	3123	218297	221420
13	**Courmayeur**	2836	205460	208296
14	**Nago-Torbole**	2810	196793	199603
15	**Canazei**	1921	187328	189249

The result shows that Lignano Sabbiedoro, Trieste and Trento are the best three cities to set up a new shop in 2019.

3.2 CLASSIFICATION: CART

The CART algorithm, as previously introduced in the State of the art, 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).

To choose the split criterion is generally used a technique of Recursive Binary Splitting. For the stopping rule, you must pay attention to the type of decision tree that is considered. In fact, the trees with many nodes and split may lead to an overfitting of the data. This means that the model is difficult to interpret, because it becomes inaccurate for later forecasts and, so, needs the stopping rule. The methods to avoid this problem are to set a minimum number of training data to be used on each leaf node or set the maximum depth of the model, which refers to the length of the path longer from root node to leaf node.

3.2.1 Training & Test Set

The very first step to take when it comes to classification is to create an adequate training set for the labeling that we would expect. In the implemented project, we will use the CART process to define and predict which stores will continue to exercise and which stores will close in 2019, having as training set the shops that closed in 2018, give the data from 2017.

The table that will characterize our classification will be an aggregate table for the first three months of each year group by for shops, where you are going to analyze sales, cost of sales and operating margin, as regards the economic and financial aspect, and

will also consider the number of receipts made during the period and the actual label representing the state Closed / Open of the store in the following year.

The aggregated data just listed will be our split criterion, except, of course, the store status which will be the element of our forecast.

To create the aggregate table, it had to refer to the dimension tables previously created in the ETL model (Shop and dates) and the fact table Sales, carrying out a detailed query to obtain the attributes mentioned above.

The totality of the queries is shown in Appendix A5.

3.2.2 Classification and Forecast on the Stores' Causes of Closing

Create the reference tables, you switch to the creation of Classification and Regression Tree, more commonly known as CART. The idea, as explained above, is to take as a starting data stores from 2017 and check if they will be open throughout 2019, pulling out the table directly from the database using the commands shown in Appendix A5.
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.

> summary (CART)

Call:

rpart (formula = FlagOpen ~ ., data = TRAININGSet, method = "class",

control = rpart.control(minsplit = 5))

n= 78

	CP	nsplit	rel error	xerror	xstd
1	0.250	0	1.00	1.0	0.1928198
2	0.100	1	0.75	0.8	0.1783112
3	0.025	3	0.55	0.9	0.1860521
4	0.010	8	0.40	1.0	0.1928198

Variable importance

GAIN	COGS	MARGIN	nRECEIPT
28	26	25	21

 # view results

```
> print (CART)
n= 78
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 78 20 OPEN (0.25641026 0.74358974)
   2) MARGIN< 1235.842 5  0 CLOSE (1.00000000 0.00000000) *
   3) MARGIN>=1235.842 73 15 OPEN (0.20547945 0.79452055)
     6) MARGIN< 17461.73 19  7 OPEN (0.36842105 0.63157895)
      12) nRECEIPT>=113 8  2 CLOSE (0.75000000 0.25000000) *
      13) nRECEIPT< 113 11  1 OPEN (0.09090909 0.90909091) *
     7) MARGIN>=17461.73 54  8 OPEN (0.14814815 0.85185185)
      14) nRECEIPT>=459 43  8 OPEN (0.18604651 0.81395349)
        28) MARGIN< 39091.27 3  1 CLOSE (0.66666667 0.33333333) *
        29) MARGIN>=39091.27 40  6 OPEN (0.15000000 0.85000000)
          58) COGS< 75358.5 6  2 OPEN (0.33333333 0.66666667) *
          59) COGS>=75358.5 34  4 OPEN (0.11764706 0.88235294)
           118) COGS>=161920.7 18  4 OPEN (0.22222222 0.77777778)
             236) nRECEIPT< 2059 2  0 CLOSE (1.00000000 0.00000000) *
             237) nRECEIPT>=2059 16  2 OPEN (0.12500000 0.87500000) *
           119) COGS< 161920.7 16  0 OPEN (0.00000000 1.00000000) *
      15) nRECEIPT< 459 11  0 OPEN (0.00000000 1.00000000) *
```

#Insert TEST set in CART

```
> prevision <- predict (CART, TESTSet, type="class")
```

#accuracy

```
> Gini(prevision)
0.09428571
> mean (prevision == TESTSet$FlagOpen)
0.8571429
```

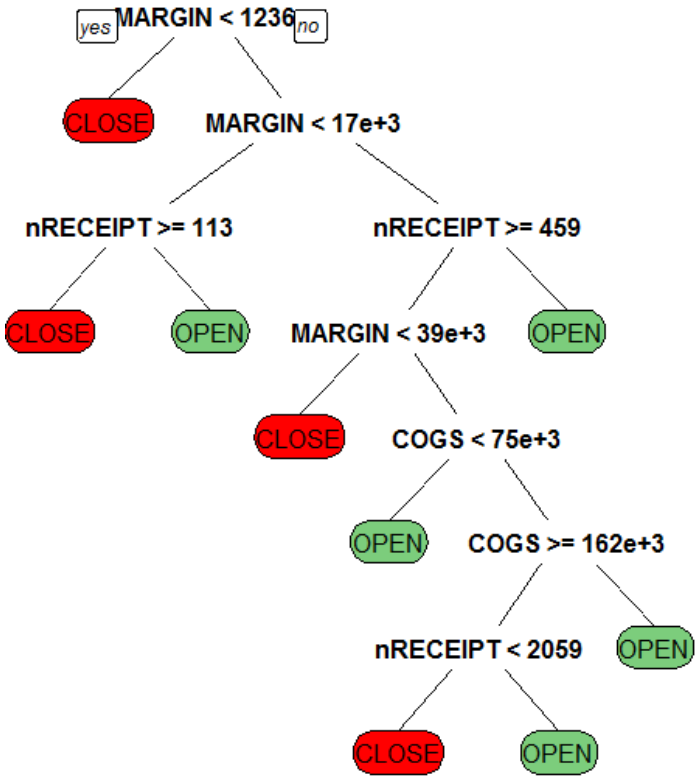


Figure 35: CART

In the obtained tree, are highlighted 4 factors: The split value in colored boxes, the final label, the probability of the input values for the final label and, finally, the probability in which it is verified on the total of the initial data.

One way to avoid the problem of overfitting and bad accuracy is to set a minimum number of splits for the training data to be used on each leaf node. For example, we may use a minimum of 5 to make a decision and ignore any leaf node that requires less than 5 levels.

The "rpart.control control command", do that and shows a minimum number of splits to performed:

CART <- rpart (FlagOpen ~., Data = TRAININGSet,

method = "class", control = rpart.control (minsplit = 5))

Another way to set the pattern, which is take the longest path from the root node to the leaf node.

To improve the performance of the tree I can also use the Pruning technique. Simply, it removes the branches that make use of features that are of little importance. In this way, we reduce the complexity of the tree, and then we increase its predictive power (that in turn reduces the over-fitting).

The simplest of Pruning method starts from the leaves and removes any node with the most popular class in that leaf, stopping before reducing accuracy.

Once the model was developed, we can extract the Test Set represented by the open shops in 2019 directly from the database just created, resulting in the forecast through dedicated command: prevision2019 <- predict (CART, DATASET18, type = "class")

Now, the model will further be evaluated using the Gini index [0.1705464] and accuracy [0.8701299], and finally, compared with the initial results, in order to understand the true efficiency. The obtained result is a prediction of 5 shops that are in a state of risk of closure in a total of 29.

	SHOP_ID	MARGIN	nRECEIPT	GAIN	COGS	FlagOpen	prevision2019
1	4064196	194247.68	5505	681708.61	487460.93	OPEN	OPEN
2	73638	20007.68	337	44479.96	24472.28	OPEN	OPEN
3	2237727	630648.10	7019	1225036.95	594388.85	OPEN	OPEN
4	2286770	175389.83	1656	330031.18	154641.35	OPEN	OPEN
5	3253865	31478.99	473	67379.97	35900.98	OPEN	CLOSE
6	4063154	69378.27	2699	288114.85	218736.58	OPEN	OPEN
7	73647	103459.19	3064	403503.95	300044.76	OPEN	OPEN
8	73569	56877.92	813	116076.54	59198.62	OPEN	OPEN
9	73539	32635.11	525	69497.32	36862.21	OPEN	CLOSE
10	4596364	102839.96	1222	201342.97	98503.01	OPEN	OPEN
11	3694385	292708.76	2945	564322.08	271613.32	OPEN	OPEN
12	3145981	40576.74	1484	172534.32	131957.58	OPEN	OPEN
13	1489952	206888.62	2081	399424.69	192536.07	OPEN	CLOSE
14	2265838	485280.72	8978	1033463.06	548182.34	OPEN	OPEN
15	73830	25486.93	405	56269.82	30782.89	OPEN	OPEN
16	73801	39345.60	568	80020.96	40675.36	OPEN	OPEN
17	73716	32220.88	475	68861.94	36641.06	OPEN	CLOSE
18	2264278	46636.65	701	97573.92	50937.27	OPEN	OPEN
19	2726114	26713.93	402	54045.17	27331.24	OPEN	OPEN
20	73773	12420.89	211	28334.43	15913.54	OPEN	OPEN
21	4064198	336480.19	7787	1196995.75	860515.56	OPEN	OPEN
22	175674	77269.68	1035	162405.14	85135.46	OPEN	OPEN
23	73740	36069.38	529	75713.85	39644.47	OPEN	OPEN
24	73825	72161.05	946	150774.37	78613.32	OPEN	OPEN
25	2022028	4178.04	88	10211.89	6033.85	OPEN	OPEN
26	2167235	445591.73	4230	846926.01	401334.28	OPEN	OPEN
27	2203462	188159.04	5098	659833.14	471674.10	OPEN	OPEN
28	876293	63376.15	891	134117.87	70741.72	OPEN	OPEN
29	2312294	141998.28	4191	549655.07	407656.79	OPEN	OPEN

Figure 36: Results of Prevision

CHAPTER 4: DATA VISUALIZATION

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 report data visualization software such as Microsoft Power BI.

The reporting systems are developed in complex areas which have provided for a data warehouse solution. One of the aims of a DW process is precisely to structure a hardware-software information environment capable of responding to the needs of organizational scenario.

With the growth of the data accumulated available to organizations, the advantages of centralized document processing are revealed in the execution times of individual reports: the particular hardware configuration of the workstations on which resources are physically hosted on the system, allows optimization of requests and decreases the number of activity with respect to the situation in which individual users search information on the system.

The document produced is called report and is presented as a combination of tables and graphs that show important measures for each analyzed topic, disaggregated and de-structured according to the needs of the client.

These measures constitute a common basis for subsequent analyzes. Each report once processed and generated, it is validated by the departments and is distributed (and updated periodically) to customers that will exploit the potential.

A process of implementing a reporting system is generally composed of the following phases, which can be expanded or reduced as a consequence of the particular

development environments and different macro-economic contexts of the organization's activities:

- Identifying information needs and visualization;
- Identification of the information environment and sources;
- Identification of the hardware / software system configuration;
- integration of information resources;
- Preparation of the report;
- Validation of the report;
- System testing phase;
- Operating phase.

These phases are not to be construed as necessarily consecutive, because some may also take place concurrently.

4.1 MICROSOFT POWER BI

The BI of Microsoft Corporation is a complete and integrated suite that helps to reduce the complexity of interaction and organization of information and to obtain competitive advantages for the company through better decisions strategies.

Microsoft provides a number of data warehouse tools and data analysis to drive the enable users to access, understand, analyze, collaborate and act on information when they want and wherever they are. It is used to get a deeper insight for better decisions making and ultimately, to help organizations adopt agile decisions to achieve the goals.

In thesis, I will use Microsoft Power BI, a suite of business analytics tools to analyze data and share information [31].

Power BI dashboards provide a 360-degree view to business users with the most important metrics in one place, updated in real time and available on all their devices. With one click, users can explore the data behind their dashboard using intuitive tools that facilitate the search for answers. Creating a dashboard is very simple, thanks to the hundreds connections with leading enterprise applications and pre-built templates to help

you put to work immediately. Also, you can access to your data and reports anywhere using the Power BI mobile app, that update automatically after any change to the data.

To facilitate the use of the application you were created variables using the DAX, Data Analysis Expressions, that indicates the formula language used in BI applications, even in the background. All the DAX formulas created for data visualization in Power BI can be found in Appendix A7.

4.2 TOOLS USED AND RESULTS OBTAINED THROUGH THE DATA VISUALIZATION

Power BI can be defined simply as a display system divided into blocks that can expand their field of definition dynamically, from a general to a detailed analysis with a simple click, with the aim of creating elaborate and complex reports tailored to the needs of customer.

The blocks that characterize the use are mainly 4:

- Views;
- Report;
- Dashboard;
- Dataset.

4.2.1 Views

When creating or editing a Power BI reports, you can use several types of visual objects. Icons for these visual objects are displayed in the Views pane.

Developers create custom visual objects through SDK. These visual objects enable business users to view data in a way that best suits their business. Users can import files of custom visual objects in reports and use them like any other visual object of Power BI, assuming a leading position. Visual objects can also be filtered, highlighted, modified, shared freely.

The custom visual objects are distributed in three ways:

- Custom file visual objects;
- Visual object organization;
- Marketplace.

In some organizations, custom visual objects are even more important, as might be necessary to communicate data and in-depth information or simply, to bring out to individuals. Therefore, these organizations need to develop custom visual objects, share them in the cloud and make sure they are handled properly.

In summary, the term views in Power BI can be defined as a visual representation of the data. This representation can be as a graph, a map or any other tool that can show your data.

The image below shows some of the views present in Power BI, providing a generic overview of the first quarter 2019:

- The left view (Multiline card), was created to relate a measurement and an attribute of the product table. As can be seen, in fact, for every manufacturing country of origin it has been associated with an equivalent number of total units sold in the first period.

- At the center, we find a quick but effective corresponding value of gain, discount, cost of goods sold and margin until 2019. Each of these values is accompanied by a weekly detailed graph that shows the comparison with the first quarter of the previous year. This chart is very useful to understand where a firm gained and where lost money, and comparing them with other company data, you can also get to the cause of mutative phenomenon.

- On the right, finally, there are two histograms, one vertical and one horizontal, which show, respectively, the total turnover filtered for the period in analysis for a store channel and for the flag weekend, obtained by means of creation of a new Power BI function with

language DAX. The last figure is the percentage of the daily flow of customers at points of sale.

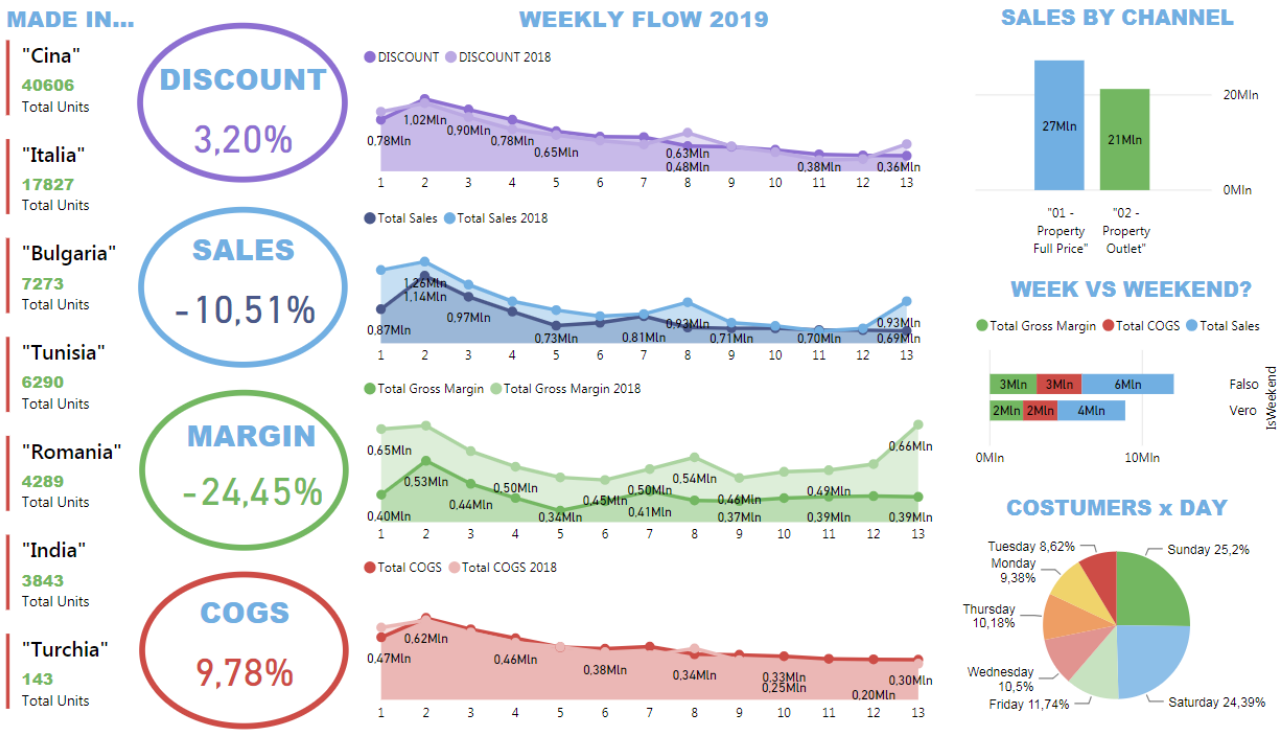


Figure 37: Visualization of 1Q 2019

From this kind of views, the results are clear and hard to confuse. We can say, for example that in this quarter the general performance of the company is below expectations, that most of the products sold are made in China, the sales channel with highest grossing is the Property Full Price and, finally, that the best flow of people occurs on the weekend.

With a small demonstration, you can understand the enormous potential that can bring data visualization, and in general the Business Analytics, for managerial decisions of a company, to improve, for example, over the years in the distribution of goods and of turnover.

4.2.4 Report

A report is a compilation of views displayed together on one or more pages. Reports help you organize your views in a way that tells the story of your data in the way you want. For example, if you want to show the sales of your company within the various sales points of your country, you can have a relationship consists of several charts (pie, line or bar) and maps. In the following example are shown the results obtained thanks to an analysis of the closed stores in Italy from 2017 to the present day of the client analyzed in the thesis project , identifying the cause of the closure of them, classified as a closure for a bad profit margin, or a closure for a covered market in the years or the birth of a neighbor new store. As starting data, we were used the data implemented in the ETL phase of the data warehouse, previously explained in Chapter 2, and the data obtained from CART algorithm illustrated in chapter 3.

The result is visible through the underlying map that encloses some examples for each type of closure. The shops of Serravalle Scrivia and Milano C.so Bueno Aires show a closure caused by new opening, while the shops of Padova and Modena caused by a margin not adequate.

The size of the dots in the figure is proportionate to the total turnover for each point of sale.

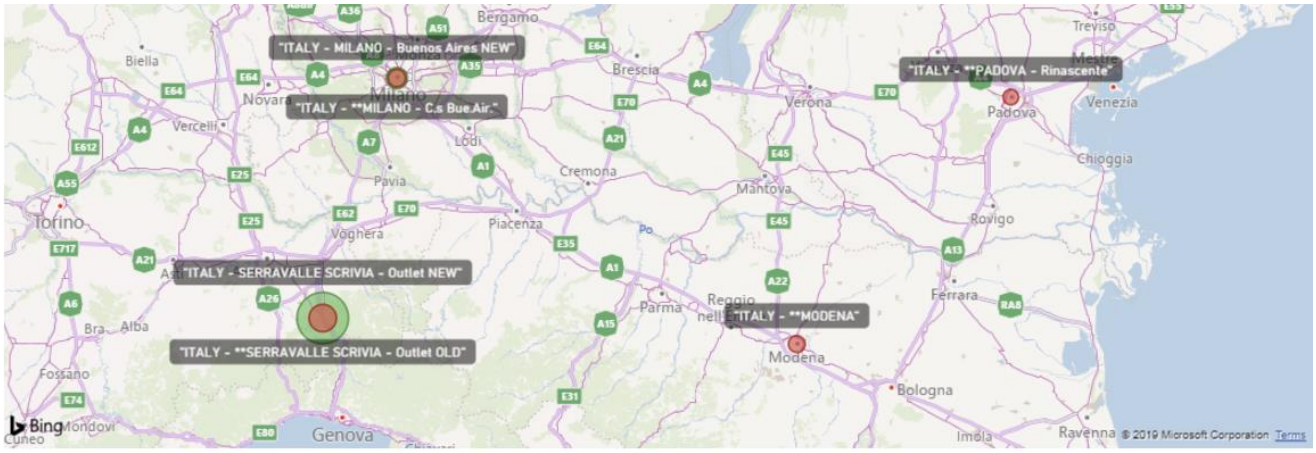


Figure 38: Map of Closing Stores

In connection with the map, an explanatory table was created which highlights the closure type and description for each store.

CLOSE FROM 2017

_NEGOZIO_DESC_	TotalSales2017	Total Gross Margin LY	CloseREASON2017
"ITALY - **MILANO - C.s Bue.Air."	721.213,20	315.535,58	COVERED/NEW SHOP
"ITALY - **ROMA - Outlet"	1.365.206,15	637.587,15	COVERED/NEW SHOP
"ITALY - **SERRAVALLE SCRIVIA - Outlet OLD"	3.628.541,00	1.751.273,21	COVERED/NEW SHOP
"ITALY - **ASCOLI PICENO - C.C. Battente"	895,00	532,00	BAD MARGIN
"ITALY - **BARI"	89.788,52	15.294,87	BAD MARGIN
"ITALY - **CHIETI - C.C. Megal?"	39.346,65	26.852,65	BAD MARGIN
"ITALY - **COMO - La Borsetta"	347.498,50	127.343,25	BAD MARGIN
"ITALY - **ERCOLANO - Donadio"	540,00	324,00	BAD MARGIN
"ITALY - **FORLI - Virgili"	14.244,00	5.626,00	BAD MARGIN
"ITALY - **L'AQUILA "	19.755,50	10.071,50	BAD MARGIN
"ITALY - **NAPOLI - Vomero"	67.064,00	16.168,50	BAD MARGIN
"ITALY - **PADOVA - Rinascente"	102.357,50	46.144,50	BAD MARGIN
"ITALY - **PESARO"	21.213,00	12.325,00	BAD MARGIN
"ITALY - **RIMINI - Virgili"	21.683,40	10.548,40	BAD MARGIN
"ITALY - **TERAMO - c.c. Val Vibrata Colonnella"	3.641,40	1.735,40	BAD MARGIN
"ITALY - **TERMOLI"	9.864,10	4.061,10	BAD MARGIN

CLOSE FROM 2018

_NEGOZIO_DESC_	TotalSales2018	Total Gross Margin TY	CloseREASON2018
"ITALY - **ROMA - Via Nazionale"	456.627,20	287.662,39	COVERED/NEW SHOP
"ITALY - **CATANIA - C.so Italia"	25.093,47	7.790,47	BAD MARGIN
"ITALY - **MILANO - Coin"	146.627,97	95.369,69	BAD MARGIN
"ITALY - **MODENA"	205.376,50	117.997,02	BAD MARGIN
"ITALY - **REGGIO CALABRIA"	8.422,30	2.609,96	BAD MARGIN
"ITALY - **ROMA - P.zza Balduina"	252.484,64	85.815,82	BAD MARGIN
"ITALY -** CERVIA"	157.051,15	71.301,60	BAD MARGIN
"ITALY -** PALERMO - Via Libert? NEW"	238.727,45	158.288,42	BAD MARGIN
"ITALY -** RAVENNA"	240.810,85	115.031,40	BAD MARGIN
"ITALY -** ROMA - Coin"	53.853,69	37.427,34	BAD MARGIN

Figure 39: Table of Closing Stores

4.2.3 Dashboard

A dashboard is a collection of views on a single page, which you can share with others. Although visually like a report, a dashboard must fit on a single page and can be shared with other users who will be able to interact with the data presented in it. By creating and sharing a dashboard for a sales manager, for example, he or she should be able to interact with it and see new information other than that which is clearly visible on the dashboard to start, according to the data.

The following images show an example of a dashboard in Power of BI. The underlying figure, includes a general analysis of sales channels associated with the relative gains of the products, and the made in. It will be the default view for the customer.

The first graph (Bar Chart) includes the view of the total gain of 2019 for each shop channel, divided into Full Price Property and Property Outlet. The second chart is the Pareto 80/20 representation, where evidence that most of the sales derived by products. The last graph, instead, shows the percentage of the made in of the products.

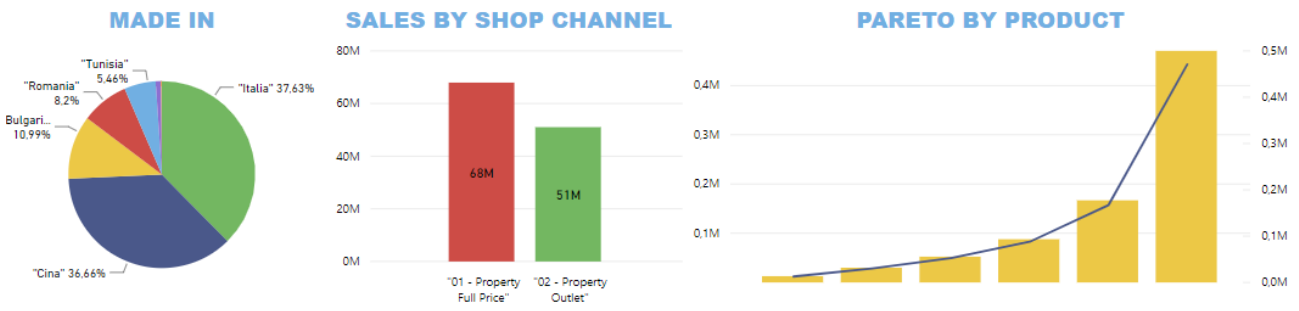


Figure 40: Default Dashboard

Through the use of filters or simple navigation you can go from a very general analysis to a detail analysis in every single aspect. Following, will be explained briefly three different views, each resulting from the default dashboard.

The images explain the variations of products sold and made in thanks to a selection of a specific store channel that you want to consider. It is very important observe how the graphics interact with each other leading to a rapid and effective analysis, chosen according to the need of the customer.

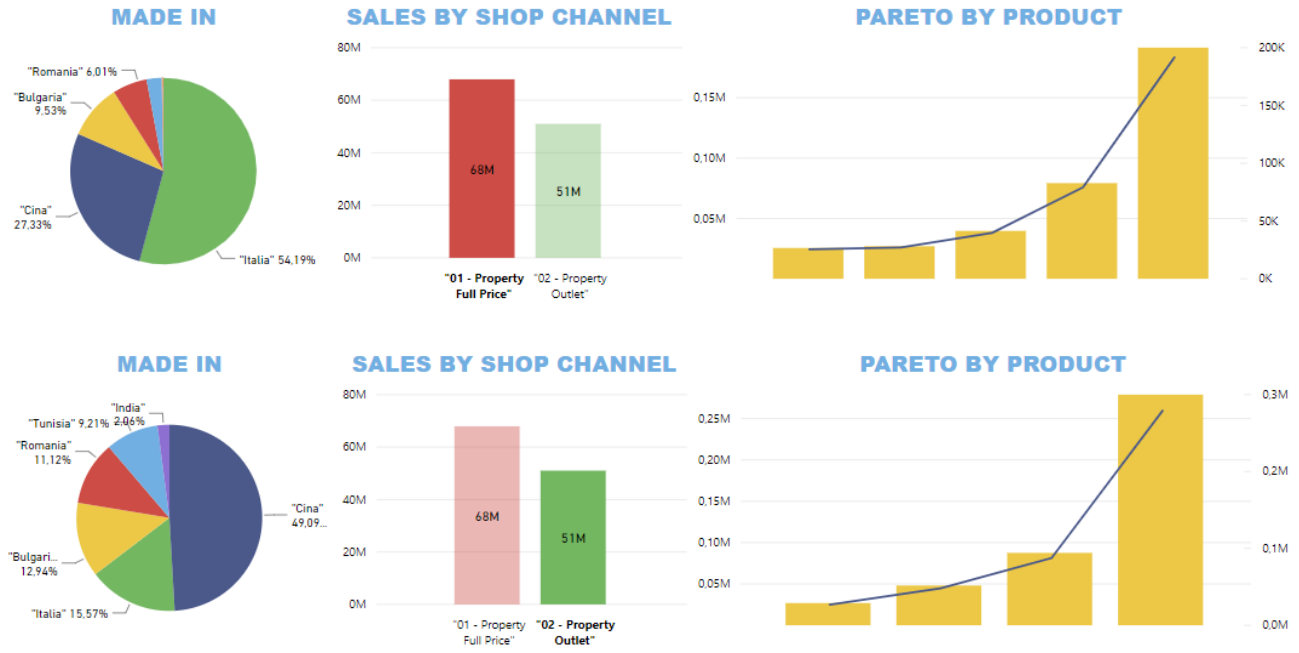


Figure 41: Drill-Through of A Dashboard

4.2.4 Dataset

A set of data, commonly called Dataset, are the data behind a chart or a map in your report. For example, if you have a chart that shows the sold in each month of the year, the data used to produce the graph are known as data records or dataset. It is important to notice, that does not necessarily datasets must be from a single source. Sometimes, it is a filtered collection of combined data from different sources with the goal of producing a unique collection that can be used in Power BI to show a feature that can be useful in the business decisions of a company. This is possible thanks to an impressive number of connectors included in Power BI.

The following image shows a set of applications which considers all economic and financial values of the customer, regarding sales, cost of sales and gross margin of 2017 and 2018 grouped for each month.

Consequently, will be display also the variation in terms of percentage value of 2018 with respect to 2017. The green and the red bar indicate, respectively, conditional formatting values of gain or loss. The last row of the table shows the totals. This Dataset can be compared to a partial income statement of the company.

Month	TotalSalesTY	TotalSalesLY	Total Sales Var %	Total COGS TY	Total COGS LY	COGS Var %	Total Gross Margin TY	Total Gross Margin LY	Gross Margin Var %
1	6.345.706,06	6.381.205,06	-0,56%	3.116.733,50	3.593.597,10	-13,27%	3.228.972,56	2.787.607,96	15,83%
2	4.317.459,62	4.929.981,33	-12,42%	1.999.986,59	2.616.713,96	-23,57%	2.317.473,03	2.313.267,37	0,18%
3	4.120.016,30	4.529.293,17	-9,04%	1.305.991,03	1.978.735,79	-34,00%	2.814.025,27	2.550.557,38	10,33%
4	4.753.446,01	5.526.732,97	-13,99%	1.452.637,74	2.327.580,91	-37,59%	3.300.808,27	3.199.152,06	3,18%
5	4.931.287,17	5.469.303,02	-9,84%	1.343.074,00	2.157.069,57	-37,74%	3.588.213,17	3.312.233,45	8,33%
6	4.837.644,26	5.449.215,32	-11,22%	1.669.463,31	2.427.601,25	-31,23%	3.168.180,95	3.021.614,07	4,85%
7	7.058.923,82	8.484.647,39	-16,80%	3.935.805,15	4.512.366,04	-12,78%	3.123.118,67	3.972.281,35	-21,38%
8	5.117.565,21	5.808.951,24	-11,90%	2.814.755,43	2.677.724,47	5,12%	2.302.809,78	3.131.226,77	-26,46%
9	4.684.542,74	5.133.199,22	-8,74%	2.166.303,97	1.510.705,96	43,40%	2.518.238,77	3.622.493,26	-30,48%
10	4.265.956,80	5.021.728,09	-15,05%	1.957.960,64	1.866.856,54	4,88%	2.307.996,16	3.154.871,55	-26,84%
11	5.300.641,81	4.813.581,63	10,12%	3.415.161,00	2.466.275,59	38,47%	1.885.480,81	2.347.306,04	-19,67%
12	5.453.336,76	5.958.834,88	-8,48%	2.871.214,81	2.533.819,78	13,32%	2.582.121,95	3.425.015,10	-24,61%
Totale	**61.186.526,56**	**67.506.673,32**	**-9,36%**	**28.049.087,17**	**30.669.046,96**	**-8,54%**	**33.137.439,39**	**36.837.626,36**	**-10,04%**

Figure 42: Income Statement Dataset

CONCLUSIONS

RESULTS

The purpose of the thesis project was to implement a data mart with clean and dedicated data concerned the sales of a fashion company, on which were carried out some helpful business analysis for future decisions strategy.

In particular:

1) Creating an ETL best practices data mart to get the best possible performance in the execution of queries and data ingestion made with an entire process of data quality to perform some relevant and efficient business analysis;
2) Data Mining and Machine Learning softwares that using Artificial Intelligence to independently implement different algorithms that can combining data warehouse and open data for studying the causes of store closures and providing the ideal location to open new ones;
3) Creating a dashboard Power BI to have a data visualization helpful to understand the meaning of the data analyzed by placing them in a visual context.

The first goal was achieved because the data mart is currently operating. The performance goal has instead been achieved only in part. Certainly, with respect to the multidimensional analyzes conducted on the operational data, previously performed by the company, we faced with a substantial improvement. The process is optimal, but the software used, being open source, and the fact of working on a local server, slowed much the level of the final performance of the ETL process.

The machine learning algorithms and the prediction have reached a good level of analysis. Through the CART decision tree of paragraph 3.2, we reached a demonstration of the major causes of the client's stores closures, obtaining as optimal value Gini index 0.17, representing the inequality of the distribution of data, and an accuracy of 0.87. Both are acceptable values, because, the Gini index and the ideal accuracy are respectively equal to 0 and 1. The result obtained has identified 5 out of 29 shops risk to closure by the end of 2019.

Instead, as regards the prediction of data ISTAT 2018, as seen above in section 3.1.1, the analysis goodness values are excellent. Using the query and the prediction of open data, it could be concluded that the best location to open a new store, are Lignano Sabbiedoro, Trieste and Trento.

In general, we can conclude that the thesis project achieved effective results for future business strategies, all displayed in smart and intuitive way through the use of dashboards created in Power BI.

FUTURE DEVELOPMENTS: REAL-TIME BUSINESS INTELLIGENCE

While the real-time analytics and big data are both trendy, analysis of big data in real time, which is the combination of them, is the future.

The real-time is often confused with the instantaneously. In fact, the engine in real-time processing is not always able to import the streaming data but can be designed to extract new data has just been placed in the source file. The time between these queries is highly dependent on business needs and can range from milliseconds to hours. For example, the analytical system of a bank would allow several minutes to evaluate the creditworthiness of an applicant and the dynamic price a retailer can take up to an hour to upgrade. However, all these examples are considered in real time.

Unlike traditional models, that examine the historical data for patterns, real-time analysis focuses on understanding the information created to help make faster and better decisions [33].

The real-time business intelligence is the use of analytical and other data processing tools to enable companies to access to the data and the most recent and relevant views. To successfully provide better data using a combination of server-less analysis (where data is transmitted directly to a dashboard or display) and a data warehouse, enabling the dashboard to show data historical and real-time in a complementary way.

For organizations that produce gigabytes or terabytes of data, many of them lose their relevance information once they are stored. The information on inventory levels, customer

needs, ongoing services, and more, can be incredibly useful, but even more so if analyzed as they are generated.

The real-time analysis and BI also allow users to do custom queries, and use the available data, including the ability to perform ad-hoc analysis on existing data or create specific views for new flows, helping to better understand the trends and create more accurate predictive models.

There are several areas where the use of BI can optimize an organization:

- *Customer Relationship Management*: The relationship management suite with clients (CRM) can use real-time data to provide better service to consumers. This includes a better involvement of the services and conversations to the known the preferences of each consumers. A significant example is the Disney company has launched its innovative MyMagicPlus program, after years of testing to Disney World. Now, every guest gets their own MagicBand bracelet, which serves as the key identification, credit card and pass. Customers simply pass it on the band sensors located around the park to entry to attractions or to pay for souvenirs. In this way, Disney giving a large amount of data on where its guests, what they are doing and what they might need.
- *Location Analytics*: The geographic data and locations often hide a lot of useful information and the extraction of these can help a company to optimize their business processes and increase profits with a better resource management. For example, sensors, such as GPS tracking systems linked to the vehicles, periodically emit data on its location. The position analysis can transform this information to detect congestion, delays prediction, detect vehicles inactive, suggest alternative routes, breaking rules and transportation guidelines (speed control), identifying the routes more profitable.
- *Service transformation:* The companies being able to collect data in real time by machinery and production lines and see how they behave, improving both efficiency and productivity and solving maintenance problems before they become an emergency. A case of transformation service doing by technology is to Rolls Royce (aero and marine engines) that has always increased the use of related products, big data and analytics in a systematic and advanced in all the three areas of activity: Product Design, manufacturing and after-sales processes. For over a decade, the

company has changed its business model, developing maintenance plans and innovative services that can connect directly to the performance of its products. The increasing use of these technologies has led Rolls Royce to the launch of Intelligent Insights and to create a set of data collected regarding the operation of engines of customer aircraft. The data are transferred to the cloud and analyzed automatically using the data mining algorithms, with the aim of creating automatically connections from data captured by different sources, allowing preventive and predictive actions [35].

Today, the real-time business intelligence is becoming an increasingly important aspect of the decision-making process of the organizations, implementing the right solution, understanding and collecting the correct data, creating a solid infrastructure and allowing your team to use it, creating a real competitive advantage over competitors.

In conclusion, with the development of new technologies and the passing of the years, we will find ourselves faced with an evolutionary scenario that will no longer consider the use of big data a competitive advantage but will consider them almost indispensable for any company throughout the world.

APPENDIX

A1. SQL - CREATION OF SURROGATE KEYS

ALTER TABLE *L1.CANALE ADD _CANALE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.ALTEZZA_TACCO ADD _ALTEZZA_TACCO_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.CATEGORIA ADD _CATEGORIA_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.CAUSALE ADD _CAUSALE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.COLORE ADD _COLORE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.COMUNI_ISTAT ADD _COMUNE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.LOCATION ADD _COMUNE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.GENDER ADD _GENDER_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.FAM_COLORE ADD _FAMIGLIA_COLORE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.FAM_MATERIALE ADD _FAMIGLIA_MATERIALE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.DATE ADD _DATE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.ISTAT ADD _ISTAT_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.MADE_IN ADD _MADE_IN_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.MADE_IN_PIANIFICAZIONE ADD _MADE_IN_PIANIFICAZIONE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.MATERIALE ADD _MATERIALE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.MODELLO ADD _MODELLO_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.PROVINCIA ADD _PROVINCIA_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.REGIONE ADD _REGIONE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.STAGIONE ADD _STAGIONE_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.TAGLIA ADD _TAGLIA_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.TURISMO ADD _TURISMO_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.SHOP ADD _NEGOZIO_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.SHOP_OPEN_ITA ADD _NEGOZIO_OPEN_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.SHOP_NOT_OPEN ADD _NEGOZIO_NOT_OPEN_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.SHOP_TOT ADD _NEGOZIO_TOT_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.PRODUCT_DATA_MASKING ADD _PRODOTTO_DM_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.PRODUCT ADD _PRODOTTO_DM_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.DAY ADD _DAY_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.YEAR ADD _YEAR_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.MONTH ADD _MONTH_SK INT IDENTITY (1,1) NOT NULL;*
ALTER TABLE *L1.QUARTER ADD _QUARTER_SK INT IDENTITY (1,1) NOT NULL;*

A2. MATRIX SELECTION ISTAT OPEN DATA

ISTAT DATA TABLE	INDICATORS / SECURITIES	AREA					YEARS	
		Italy	area	region	province	town	Data until 2017	2018
Families	Incidence of absolute poverty	■	■	■			■	
	average net household income (Excluding imputed rent)	■	■	■				
	Average monthly household expenditure on goods and non-food services	■	■	■			■	
	Total average monthly household expenditure	■	■	■			■	
Transport ation	Bus	■	■	■			■	
	Cars	■	■	■			■	
	motorway network	■	■	■			■	
	Rail network	■	■	■			■	
	Transport of goods by road	■	■	■			■	
Work	Unemployment rate	■	■	■			■	■
	Youth unemployment rate	■	■	■			■	■
Macroeco nomics	National consumption	■	■	■				
	Gross fixed capital formation	■	■	■				
	Pro-capite GDP	■	■	■			■	
Territory	Density of the population of the municipalities	■	■	■			■	
	Density of the population of the municipalities with a surface area of 1,001 to 2,000	■	■	■			■	
	Density of the population of the municipalities with a surface area of 2,001 to 6,000	■	■	■			■	
	Density of the population of the municipalities with a surface area of 6,001 to 25,000	■	■	■			■	
	Population density of the municipalities with a surface area of up to 1,000	■	■	■			■	
	municipalities with population density area exceeding 25,000	■	■	■			■	
	Building permits - housing in new residential buildings	■	■	■				
	Building permits - m2 useful living space in new residential buildings	■	■	■				
	Resident population average	■	■	■			■	
Tourism	Total arrivals	■	■	■	■	■		

A3. R-CODE: 2018 PREDICTION OF OPEN DATA

*# **Amount of the Library***

library (readxl)

*# **Excel file Amount***

Istat_Famiglie <- read_excel ("C: / Users / admin / ISTAT_FAMIGLIE.xlsx")

*# **Vector of the years***

x <- c (Istat_Famiglie $ Anno_2007, Istat_Famiglie $ Anno_2008, Istat_Famiglie $ Anno_2009, Istat_Famiglie $ Anno_2010, Istat_Famiglie $ Anno_2011, Istat_Famiglie $ Anno_2012, Istat_Famiglie $ Anno_2013, Istat_Famiglie $ Anno_2014, Istat_Famiglie $ Anno_2015, Istat_Famiglie $ Anno_2016, Istat_Famiglie $ Anno_2017)

*# **matrix with the columns of the years and the number rows***

m1 <- matrix (x, ncol = 11)

*#**Call row with regions (territory)***

y <- ISTAT_FAMIGLIE $ Territory

dimnames (m1) <- list (c (ISTAT_FAMIGLIE $ Territory), NULL)

*# **Call columns with years 207-2017***

t <- array (2007: 2017)

dimnames (m1) [[2]] <- c (t)

*# **Mold created the matrix***

m1

*# **Linear regression (lm) data from 2007 to 2017 related to the 'Average monthly household expenditure on non-food goods and services'***

reg <- lm (Istat_Famiglie $ Anno_2017 ~ $ Istat_Famiglie Anno_2007 Istat_Famiglie $ Anno_2008 + + + Istat_Famiglie Anno_2009 Istat_Famiglie $ $ $ Anno_2011 Istat_Famiglie Anno_2010 + + + Istat_Famiglie Anno_2012 Istat_Famiglie $ $ $ Anno_2014 Istat_Famiglie Anno_2013 + + + Istat_Famiglie $ Anno_2015 Istat_Famiglie $ Anno_2016)

*# **prediction 2018***

predict (reg)

*# **Add the column of the prediction to a new array***

m2 <- cbind (m1, predict (reg))

*# **print prediction***

m2 [12]

AOSTA VALLEY PIEDMONT LIGURIA LOMBARDY VENETO TRENTINO

2171.792 2363.691 1978.892 2553.094 2572.198 2305.999

Friuli Emilia Romagna TUSCANY UMBRIA LAZIO MARCHE

2135.363 2530.087 2393.346 1894.259 1871.358 2257.226

ABRUZZO MOLISE CAMPANIA PUGLIA BASILICATA CALABRIA

1779.977 1669.947 1665.452 1707.722 1516.702 1351.419

SICILY SARDINIA

1481.153 1637.064

 # Plot about Liguria

plot (m2 [3], main = "Average Expenses of Liguria", type = "l", ylab = "Euro €" XLAB = "Years", col = "black")

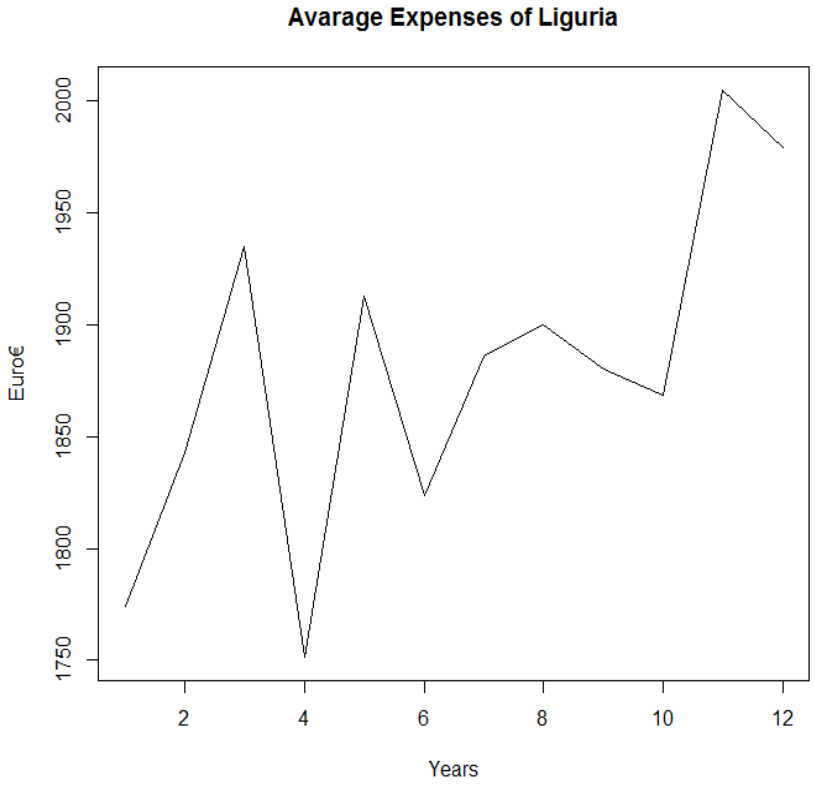


Figure 43: Plot Avg Expenses About Liguria

A4. SQL-CODE: BEST GEO-LOCATIONS

```
-------------------------------------------------------------------------
----------------BEST REGION BY INDICATORS-------------
-------------------------------------------------------------------------
select settore, indicatore, unita_di_misura, territorio, Prediction_2018, rnk
from (Select *,
          case
            when Indicatore <> 'Tasso di disoccupazione'
              then RANK() over (partition by Indicatore order by Prediction_2018 desc)
            else RANK() over (partition by Indicatore order by Prediction_2018 ) end rnk
     from L2_ISTAT.FACT_ISTAT ) t
where rnk < 3 or rnk > 18

-------------------------------------------------------------------------------------
------------------------------BEST REGIONS----------------------
-------------------------------------------------------------------------------------
select t3.territorio, isnull(t2.c, 0) as #closing_shop, isnull(t4.c, 0) as #opening_shop, sum(t3.rank1) Istat_rank,
      sum(t3.rank1 + isnull(t2.c, 0)+ isnull(t4.c, 0)) as Final_Rank
from (select t.Territorio, sum(t.rnk) rank1
     from (Select *,
                case when FACT_ISTAT.Indicatore <> 'Tasso di disoccupazione'
                   then RANK() over (partition by FACT_ISTAT.Indicatore
                  order by FACT_ISTAT.Prediction_2018 desc)
                  else RANK () over (partition by FACT_ISTAT.Indicatore order by
FACT_ISTAT.Prediction_2018
              ) end rnk
           from L2_ISTAT.FACT_ISTAT
          ) t
     Group by t.Territorio) t3
     left join
    (Select DIM_LOCATION_SS.regione, COUNT(DISTINCT DIM_SHOP_SS._negozio_desc_) as c
     From L2_STAR_SCHEMA.DIM_SHOP_SS,
          L2_STAR_SCHEMA.DIM_LOCATION_SS
     where L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_ like '%/*%' escape '/'
       and L2_STAR_SCHEMA.DIM_SHOP_SS._COMUNE_SK =
L2_STAR_SCHEMA.DIM_LOCATION_SS._COMUNE_SK
     group by DIM_LOCATION_SS.regione) as t2
     on t3.Territorio = t2.regione
     left join
    (Select DIM_LOCATION_SS.regione, COUNT(DISTINCT DIM_SHOP_SS._negozio_desc_) as c
     From L2_STAR_SCHEMA.DIM_SHOP_SS,
          L2_STAR_SCHEMA.DIM_LOCATION_SS
     where L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_ not like '%/*%' escape '/'
       and L2_STAR_SCHEMA.DIM_SHOP_SS._COMUNE_SK =
L2_STAR_SCHEMA.DIM_LOCATION_SS._COMUNE_SK
     group by DIM_LOCATION_SS.regione) as t4
     on t4.regione = t2.regione
Group by t3.Territorio, t2.c, t4.c
order by Final_Rank;

-------------------------------------------------------------------------
--------------------------------BEST CITIES-------------------------
-------------------------------------------------------------------------
select t.comune,
       t.num_residenti,
       t2.Totale_esercizi_Arrivi_2018,
       t2.Totale_esercizi_Arrivi_2018 / t.num_residenti measure
```

```
from (select *, RANK() over ( order by num_residenti ) rnk
     from L2_ISTAT.DIM_LOCATION_ISTAT
      Where L2_ISTAT.DIM_LOCATION_ISTAT.regione in ('TRENTINO ALTO ADIGE', 'FRIULI VENEZIA
GIULIA')
          or L2_ISTAT.DIM_LOCATION_ISTAT.regione like '%AOSTA%') as t,
    (select *, RANK() over ( order by Totale_esercizi_Arrivi_2018 ) rnk2
     from L2_ISTAT.DIM_TURISMO_ISTAT) as t2,
    L2_ISTAT.DIM_LOCATION_ISTAT
  Where
    t._COMUNE_SK=t2._COMUNE_SK
  group by
    t.num_residenti,  t.comune, rnk, t2.Totale_esercizi_Arrivi_2018
  order by measure desc;
```

A5. SQL-CODE: AGGREGATE FACT SALES FOR THE CREATION OF THE CART MODEL

```
-----------------------------------------------------------
CREATION OF AGGREGATE TABLES
-----------------------------------------------------------

create table AGGREGATE.CART_YEAR_SHOP_SALES_1718
(
  SHOP_ID  int          not null,
  SHOP     Varchar(1000) not null,
  DISCOUNT float         not null,
  MARGIN   float         not null,
  nRECEIPT int           not null,
  GAIN     float         not null,
  COGS     float         not null,
  FlagOpen Varchar(1000) not null
)
create table AGGREGATE.CART_QUARTER_SHOP_SALES_1718
(
  SHOP_ID  int          not null,
  SHOP     Varchar(1000) not null,
  MARGIN   float         not null,
  nRECEIPT int           not null,
  GAIN     float         not null,
  COGS     float         not null,
  FlagOpen Varchar(1000) not null
)
create table AGGREGATE.CART_QUARTER_SHOP_SALES_19
(
  SHOP_ID  int          not null,
  SHOP     Varchar(1000) not null,
  MARGIN   float         not null,
  nRECEIPT int           not null,
  GAIN     float         not null,
  COGS     float         not null,
  FlagOpen Varchar(1000) not null
)
```

```
-----------------------------------------------------
QUERIES
-----------------------------------------------------

INSERT INTO AGGREGATE.CART_YEAR_SHOP_SALES_1718
select Distinct L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_            as SHOP_ID,
            L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_           as SHOP,
            SUM(_SCONTO_GENERICO_SC_ + _SCONTO_SC_)             as DISCOUNT,
            (SUM(_FATTURATO_NI_SC_) - SUM(_COSTO_VENDUTO_NI_SC_)) as MARGIN,
            count(distinct _NUMERO_SCONTRINO_)                  as nRECEIPT,
            SUM(_FATTURATO_NI_SC_)                              as GAIN,
            SUM(_COSTO_VENDUTO_NI_SC_)                          as COGS,
            case
              when L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_ not like '%*%' then 'OPEN'
              else 'CLOSE' end                                  as FlagOpen
from L2_STAR_SCHEMA.FACT_SALES_SS,
    L2_STAR_SCHEMA.DIM_DATE_SS,
    L2_STAR_SCHEMA.DIM_SHOP_SS
Where L2_STAR_SCHEMA.FACT_SALES_SS._GIORNO_ = L2_STAR_SCHEMA.DIM_DATE_SS.Date
  and L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_ =
L2_STAR_SCHEMA.FACT_SALES_SS._NEGOZIO_ID_
  and L2_STAR_SCHEMA.DIM_DATE_SS.Year in (2018,2017)
Group by
L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_,L2_STAR_SCHEMA.DIM_SHOP_SS._CANALE_DE
SC_,L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_
Order by FlagOpen;


INSERT INTO AGGREGATE.CART_QUARTER_SHOP_SALES_1718
select Distinct L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_            as SHOP_ID,
            L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_           as SHOP,
            (SUM(_FATTURATO_NI_SC_) - SUM(_COSTO_VENDUTO_NI_SC_)) as MARGIN,
            count(distinct _NUMERO_SCONTRINO_)                  as nRECEIPT,
            SUM(_FATTURATO_NI_SC_)                              as GAIN,
            SUM(_COSTO_VENDUTO_NI_SC_)                          as COGS,
            case
              when L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_ not like '%*%' then 'OPEN'
              else 'CLOSE' end                                  as FlagOpen
from L2_STAR_SCHEMA.FACT_SALES_SS,
    L2_STAR_SCHEMA.DIM_DATE_SS,
    L2_STAR_SCHEMA.DIM_SHOP_SS
Where L2_STAR_SCHEMA.FACT_SALES_SS._GIORNO_ = L2_STAR_SCHEMA.DIM_DATE_SS.Date
  and L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_ =
L2_STAR_SCHEMA.FACT_SALES_SS._NEGOZIO_ID_
  and L2_STAR_SCHEMA.DIM_DATE_SS.Year in (2018, 2017)
  and L2_STAR_SCHEMA.DIM_DATE_SS.Quarter = 1
Group by L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_,
L2_STAR_SCHEMA.DIM_SHOP_SS._CANALE_DESC_,
       L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_
Order by FlagOpen;
```

```
INSERT INTO AGGREGATE.CART_QUARTER_SHOP_SALES_19
select Distinct L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_ as SHOP_ID,
        L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_ as SHOP,
        (SUM(_FATTURATO_NI_SC_) - SUM(_COSTO_VENDUTO_NI_SC_)) as MARGIN,
        count(distinct _NUMERO_SCONTRINO_) as nRECEIPT,
        SUM(_FATTURATO_NI_SC_) as GAIN,
        SUM(_COSTO_VENDUTO_NI_SC_) as COGS,
        case
          when SUM(_FATTURATO_NI_SC_) >0 then 'OPEN'
          else 'CLOSE' end as FlagOpen
from L2_STAR_SCHEMA.FACT_SALES_SS,
   L2_STAR_SCHEMA.DIM_DATE_SS,
   L2_STAR_SCHEMA.DIM_SHOP_SS
Where L2_STAR_SCHEMA.FACT_SALES_SS._GIORNO_ = L2_STAR_SCHEMA.DIM_DATE_SS.Date
 and L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_ =
L2_STAR_SCHEMA.FACT_SALES_SS._NEGOZIO_ID_
 and L2_STAR_SCHEMA.DIM_DATE_SS.Year in (2019)
 --and
 and L2_STAR_SCHEMA.DIM_DATE_SS.Quarter = 1
Group by L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_DESC_,
L2_STAR_SCHEMA.DIM_SHOP_SS._CANALE_DESC_,
      L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_
Order by FlagOpen;


-------------------------------------------------------
TRUNCATE
-------------------------------------------------------

truncate table AGGREGATE.CART_YEAR_SHOP_SALES_1718;
truncate table AGGREGATE.CART_QUARTER_SHOP_SALES_1718;
truncate table AGGREGATE.CART_QUARTER_SHOP_SALES_19;


-------------------------------------------------------
CHECK
-------------------------------------------------------

Select COUNT(DISTINCT _negozio_desc_) as OPEN_SHOP, year
From L2_STAR_SCHEMA.DIM_SHOP_SS,
   L2_STAR_SCHEMA.FACT_SALES_SS,
   L2_STAR_SCHEMA.DIM_DATE_SS
where L2_STAR_SCHEMA.DIM_SHOP_SS._NEGOZIO_ID_ =
L2_STAR_SCHEMA.FACT_SALES_SS._NEGOZIO_ID_
 and L2_STAR_SCHEMA.FACT_SALES_SS._GIORNO_ = L2_STAR_SCHEMA.DIM_DATE_SS.Date
 and L2_STAR_SCHEMA.FACT_SALES_SS._FATTURATO_NI_SC_ > 0
group by Year;
-------------------------------------------------------
OPEN_SHOP  year
93      2017
77      2018
29      2019
-------------------------------------------------------
```

A6. R-CODE: CALSSIFICATION AND REGRESION TREE (CART)

library (dbConnect)

library (odbc)

library (DBI)

library (rpart.plot)

library (rpart)

library (tree)

library (gplots)

library (INEQ)

#Extraction of table 17_18

with <- dbConnect (odbc odbc :: (), DRIVER = "SQL Server" SERVER = "192.168.2.14" DATABASE = "FASHION_RETAIL" UID = "" PWD = "")

*query <- dbSendQuery (with, "select * from AGGREGATE.CART_QUARTER_SHOP_SALES_1718")*

*#query <- dbSendQuery (with, "select * from AGGREGATE.CART_YEAR_SHOP_SALES_1718")*

DATE <- dbFetch (query)

DATASET <- DATA [3: ncol (DATE)] 1.2 #elimino columns related to shops (SHOP_ID & SHOP)

 # data Partition

PD17 <- sample (2, nrow (DATASET), replace = TRUE, prob = c (0.8,0.2))

TRAININGSet <- DATASET [PD17 == 1]

Testset <- DATASET [PD17 == 2]

#CART

CART <- rpart (FlagOpen ~., Data = TRAININGSet, method = "class", control = rpart.control (minsplit = 5))

#CART <- rpart (FlagOpen ~ MARGIN DISCOUNT + + + nRECEIPT GAIN + COGS, data = DATASET17, method = "class", rpart.control (minsplit = 5))

#color CART

boxcols <- c ("red", "palegreen3") [CART frame $ $ yval]

#Plot CART

rpart.plot (CART, fallen.leaves = TRUE, box.col = boxcols, uniform = TRUE)

prp (CART, box.col = boxcols)

 # View results

print (CART)

#Insert Test set on CART algorithm

prevision <- predict (CART, Testset, type = "class")

#accuracy of prediction

Gini (prevision)

mean (prevision == Testset $ FlagOpen)

#Estraggo Table 19

con2 <- dbConnect (odbc odbc :: (), DRIVER = "SQL Server" SERVER = "192.168.2.14" DATABASE = "FASHION_RETAIL" UID = "" PWD = "")

*query19 <-dbSendQuery (con2, "select * from AGGREGATE.CART_QUARTER_SHOP_SALES_19")*

DATA19 <- dbFetch (query19)

#delete column refers to the shop's description

DATASET19 <- DATA19 [-2]

#Insert new Data on CART algorithm

prevision2019 <- predict (CART, DATASET19, type = "class")

prevision2019

#accuracy of new prediction

Gini (prevision2019)

mean (prevision2019 DATASET19 == $ FlagOpen)

#statistics

summary (prevision2019)

cbind (DATA19, prevision2019)

---------- -----

OUTPUT ---------- ------

 # View results

> Print (CART)

n = 78

node), split, n, loss, yval, (yprob)

* Denotes terminal node

1) root 78 20 OPEN (0.25641026 0.74358974)

2) MARGIN <1235.842 5 0 CLOSE (1.00000000 0.00000000) *

3) MARGIN> = 1235.842 73 15 OPEN (0.20547945 0.79452055)

6) MARGIN <17461.73 19 7 OPEN (0.36842105 0.63157895)

12) nRECEIPT> = 113 2 8 CLOSE (0.75000000 0.25000000) *

13) nRECEIPT <113 11 1 OPEN (0.09090909 0.90909091) *

7) MARGIN> = 17461.73 54 8 OPEN (0.14814815 0.85185185)

14) nRECEIPT> = 459 43 8 OPEN (0.18604651 0.81395349)

28) MARGIN <3 39091.27 1 CLOSE (0.66666667 0.33333333) *

29) MARGIN> = 39091.27 40 6 OPEN (0.15000000 0.85000000)

58) COGS <75358.5 6 2 OPEN (0.33333333 0.66666667) *

59) COGS> = 75358.5 34 4 OPEN (0.11764706 0.88235294)

118) COGS> = 161920.7 18 4 OPEN (0.22222222 0.77777778)

236) nRECEIPT <2059 2 0 CLOSE (1.00000000 0.00000000) *

237) nRECEIPT> = 2059 16 2 OPEN (0.12500000 0.87500000) *

119) COGS <161920.7 16 0 OPEN (0.00000000 1.00000000) *

15) nRECEIPT <459 11 0 OPEN (0.00000000 1.00000000) *

> Summary (CART)

Call:

rpart (formula = FlagOpen ~., data = TRAININGSet, method = "class",

control = rpart.control (minsplit = 5))

n = 78

CP nsplit rel error xError xstd

1 0,250 0 1.00 1.0 0.1928198

2 0,100 1 0.75 0.8 0.1783112

3 0,025 3 0.55 0.9 0.1860521

4 0,010 8 0.40 1.0 0.1928198

Variable Importance

GAIN MARGIN COGS nRECEIPT

28 26 25 21

#Insert Test set on CART algorithm

> Prevision <- predict (CART, Testset, type = "class")

#accuracy

> Gini (prevision)

[1] 0.09428571

> Mean (prevision == Testset $ FlagOpen)

[1] 0.8571429

> prevision2019

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

OPEN OPEN OPEN OPEN CLOSE CLOSE OPEN OPEN OPEN OPEN OPEN OPEN OPEN OPEN OPEN OPEN CLOSE OPEN OPEN OPEN CLOSE

22 23 24 25 26 27 28 29

OPEN CLOSE OPEN OPEN OPEN OPEN OPEN OPEN

Levels: OPEN CLOSE

#accuracy new prediction

> Gini (prevision2019)

[1] 0.07807417

> Mean (prevision2019 == DATASET1819 $ FlagOpen)

[1] 0.8275862

>

statistics

> Summary (prevision2019)

CLOSE OPEN

4 25

	SHOP_ID	MARGIN	nRECEIPT	GAIN	COGS	FlagOpen	prevision2019
1	4064196	194247.68	5505	681708.61	487460.93	OPEN	OPEN
2	73638	20007.68	337	44479.96	24472.28	OPEN	OPEN
3	2237727	630648.10	7019	1225036.95	594388.85	OPEN	OPEN
4	2286770	175389.83	1656	330031.18	154641.35	OPEN	OPEN
5	3253865	31478.99	473	67379.97	35900.98	OPEN	CLOSE
6	4063154	69378.27	2699	288114.85	218736.58	OPEN	OPEN
7	73647	103459.19	3064	403503.95	300044.76	OPEN	OPEN
8	73569	56877.92	813	116076.54	59198.62	OPEN	OPEN
9	73539	32635.11	525	69497.32	36862.21	OPEN	CLOSE
10	4596364	102839.96	1222	201342.97	98503.01	OPEN	OPEN
11	3694385	292708.76	2945	564322.08	271613.32	OPEN	OPEN
12	3145981	40576.74	1484	172534.32	131957.58	OPEN	OPEN
13	1489952	206888.62	2081	399424.69	192536.07	OPEN	CLOSE
14	2265838	485280.72	8978	1033463.06	548182.34	OPEN	OPEN
15	73830	25486.93	405	56269.82	30782.89	OPEN	OPEN
16	73801	39345.60	568	80020.96	40675.36	OPEN	OPEN
17	73716	32220.88	475	68861.94	36641.06	OPEN	CLOSE
18	2264278	46636.65	701	97573.92	50937.27	OPEN	OPEN
19	2726114	26713.93	402	54045.17	27331.24	OPEN	OPEN
20	73773	12420.89	211	28334.43	15913.54	OPEN	OPEN
21	4064198	336480.19	7787	1196995.75	860515.56	OPEN	OPEN
22	175674	77269.68	1035	162405.14	85135.46	OPEN	OPEN
23	73740	36069.38	529	75713.85	39644.47	OPEN	OPEN
24	73825	72161.05	946	150774.37	78613.32	OPEN	OPEN
25	2022028	4178.04	88	10211.89	6033.85	OPEN	OPEN
26	2167235	445591.73	4230	846926.01	401334.28	OPEN	OPEN
27	2203462	188159.04	5098	659833.14	471674.10	OPEN	OPEN
28	876293	63376.15	891	134117.87	70741.72	OPEN	OPEN
29	2312294	141998.28	4191	549655.07	407656.79	OPEN	OPEN

A7. DAX-CODE: VARIABLES OF POWER BI

Total Sales CALCULATE = (sum ('L2_STAR_SCHEMA FACT_SALES_SS_DM' [_ FATTURATO_LI_SC_]))

Total Sales LY= CALCULATE ([Total Sales]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2017)

Total Sales TY = CALCULATE ([Total Sales]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2018)

Total Sales Change = [Total Sales TY] - [Total Sales LY]

Total Sales Change% = IF ([Total Sales LY] <> 0; [Total Sales Change] / [Total Sales LY]; BLANK ())

Total Units CALCULATE = (sum ('L2_STAR_SCHEMA FACT_SALES_SS_DM' [_ QUANTITA_SC_]))

Total Units Last Year = CALCULATE ([Total Units]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2017)

Total Units This Year = CALCULATE ([Total Units]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2018)

Total COGS = CALCULATE (SUM ('L2_STAR_SCHEMA FACT_SALES_SS_DM' [_ COSTO_VENDUTO_NI_SC_]))

Total COGS LY = CALCULATE ([Total COGS]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2017)

Total COGS TY = CALCULATE ([Total COGS]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2018)

Total Gross Margin = [Total Sales] - [Total COGS]

Total Gross Margin LY = [Total Sales LY] - [Total COGS LY]

Total Gross Margin TY = [Total Sales TY] - [Total COGS TY]

Gross Margin% This Year = [Total Gross Margin TY] / [Total Sales TY]

Gross Margin% Last Year = [Total Gross Margin LY] / [Total Sales LY]

Total Gross Margin var = [Total Gross Margin TY] - [Total Gross Margin LY]

Gross Margin Var% = IF ([Total Gross Margin LY] <> 0; [Total Gross Margin var] / [Total Gross Margin LY]; BLANK ())

Avg $ / Unit TY = IF ([Total Units This Year] <> 0; [Total Sales TY] / [Total Units This Year]; BLANK ())

Avg $ / Unit LY = IF ([Total Units Last Year] <> 0; [Total Sales LY] / [Total Units Last Year]; BLANK ())

Average Unit Price = [Avg $ / Unit TY]

Store Count = DISTINCTCOUNT ('L2_STAR_SCHEMA DIM_SHOP_SS' [_ NEGOZIO_ID_])

Sales for Sq. Ft = ([Total Sales TY] / (DISTINCTCOUNT ('L2_STAR_SCHEMA DIM_DATE_SS' [Month]) * SUM ('L2_STAR_SCHEMA DIM_SHOP_SS' [_ _ SURFACE]))) * 12

REASON2017 = if ([Total Gross Margin LY] <= 200000; "BAD MARGIN"; "COVERED / NEW")

REASON2018 = if ([Total Gross Margin TY] <= 200000; "BAD MARGIN"; "COVERED / NEW")

FLG OPEN = IF (CONTAINSSTRING ('L2_STAR_SCHEMA DIM_SHOP_SS' [_ _ NEGOZIO_DESC]; "~ *") = TRUE; 0; 1)

Count Open = Sum ('L2_STAR_SCHEMA DIM_SHOP_SS' [FLG OPEN])

TotSHOP= COUNT ('L2_STAR_SCHEMA DIM_SHOP_SS' [_ NEGOZIO_ID_])

#OPENshop = Sum ('L2_STAR_SCHEMA DIM_SHOP_SS' [OPEN])

#CLOSEshop = COUNT ('L2_STAR_SCHEMA DIM_SHOP_SS' [_ NAZIONE_DESC_]) - ('L2_STAR_SCHEMA DIM_SHOP_SS' [# Openshop])

Total Discount = Sum ('L2_STAR_SCHEMA FACT_SALES_SS_DM' [_ SCONTO_GENERICO_SC_]) + sum ('L2_STAR_SCHEMA FACT_SALES_SS_DM' [_ SCONTO_SC_])

TOT DISCOUNT TY= Calculate ([Total Discount]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2018)

TOT DISCOUNT LY= Calculate ([Total Discount]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2017)

Total Sales in 2019 = CALCULATE ([Total Sales]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2019)

Total Sales 2019 Var% = IF ([Total Sales TY] <> 0; ([Total Sales 2019] - [Total Sales TY]) / [Total Sales TY]; BLANK ())

Total COGS 2019 = CALCULATE ([Total COGS]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2019)

COGS in 2019 Var% = IF ([Total COGS TY] <> 0; ([Total COGS 2019] - [Total COGS TY]) / [Total COGS TY]; BLANK ())

Total Gross Margin in 2019 = [Total Sales 2019] - [Total COGS 2019]

Gross Margin 2019 Var% = IF ([Total Gross Margin TY] <> 0; ([Total Gross Margin 2019] - [Total Gross Margin TY]) / [Total Gross Margin TY]; BLANK ())

DISCOUNT 2019 = CALCULATE ([Total DISCOUNT]; 'L2_STAR_SCHEMA DIM_DATE_SS' [Year] = 2019)

DISCOUNT 2019 Var% = IF ([Total DISCOUNT] <> 0; ([DISCOUNT 2019] - [DISCOUNT TY]) / [DISCOUNT TY]; BLANK ())

REFERENCES

1) Manyika J., CM (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity.

2) Opresnik D., TM (2015). The value of Big Data in servitization. International Journal of Production Economics.

3) Court D. (2015). Getting big impact from big data. McKinsey Quarterly.

4) E. Hazan, BF (2013). Leveraging big data to optimize digital marketing. McKinsey Quarterly.

5) Country A. (2015). Digital Marketing. Mobile, video, big data and social Internet becomes advertising.

6) Lühr P., MR (2013). Name your price: The Power of Big Data and analytics. McKinsey Quarterly.

7) Talend (2009) The Top 10 Reasons for Choosing Open Source Data Integration.

8) Mark R. Madsen (2009) The Role of Open Source Data Integration, Third Nature Technology Report.

9) T. Baumgartner, HH (2011). Find Big Growth in Big Data. In HH T. Baumgartner, Sales Growth. Five proven strategies from world's sales leader.

10) C. Adamson (2010), Star Schema: The Complete Reference, New York, McGraw-Hill.

11) Rezzani A. (2012), Business Intelligence - Processes, methods, utilization in the company, Milan, Feltrinelli Editore.

12) Vidette P., Patricia K. and Stephen B. (1998), "Building a Data Warehouse for decision support", Prentice-Hall.

13) https://www.educba.com/data-mining-vs-machine-learning/

14) https://www.simonefavaro.it/2017/04/07/introduzione-machine-learning/

15) Michael J. Berry, Gordon Linoff, (2004), "Introduction to Data Mining".

16) Wray Buntine, (1992), "Learning classification trees".

17) https://en.wikipedia.org/wiki/Bayes%27theorem , "Bayes' theorem".

18) Dr. Saed Sayad, "Support Vector Machine - Classification (SVM)". http://www.saedsayad.com/support vector machine.htm .

19) https://en.wikipedia.org/wiki/Hierarchical_clustering "Hierarchical clustering" in 2016.
20) Du Xiaoshan. (2016). Master's Thesis: Data Mining Analysis and Modeling for Marketing Based on Attributes of Customer Relationship.
21) Judith Hurwitz, Daniel Kirsch. (2018). Machine Learning for Dummies, IBM Limited Edition.
22) W. Raghupathi, RV (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems.
23) Emil Drkušić.Star Schema vs. Snowflake Schema 2016https://www.vertabelo.com/blog/technical-articles/data-warehouse-modeling-star-schema-vs-snowflake-schema .
24) Talend Support Center, TMAP. https://help.talend.com/reader/wDRBNUuxk629sNcI0dNYaA/mxzKD~8eLuNFSXH6LMi7qg .
25) Info about 'Istituto Italiano Di Statistica", ISTAT. http://dati.istat.it/ .
26) R-Studio Site. https://www.rstudio.com/ .
27) R. code Definition https://www.r-project.org/about.html .
28) Talend Open Studio. https://help.talend.com/home .
29) F-Stat. Wikipedia.https://en.wikipedia.org/wiki/F-test .
30) T-Stat. https://www.statisticshowto.datasciencecentral.com/t-statistic/ .
31) Microsoft Power BI. https://powerbi.microsoft.com/it-it/ .
32) Alberto Ferrari and Marco Russo. Introducing Microsoft Power BI.
33) Everything you need to know about Real Time Business. https://www.sisense.com/blog/everything-you-need-to-know-about-real-time-business-intelligence/
34) Disney MagicBand. https://disneyworld.disney.go.com/en-eu/plan/my-disney-experience/bands-cards/
35) Amit Roy Choudhury and Jim Mortleman. How IoT is turning into a Rolls-Royce data-fueled business. January 2018.