POLITECNICO DI TORINO

Master's Degree in INGEGNERIA GESTIONALE (ENGINEERING AND MANAGEMENT)

Master's Degree Thesis

Design and Implementation of Data enrichment modules as support to decision-making processes

Supervisors

Prof. Marco CANTAMESSA

Dott. Vincenzo SCINICARIELLO

Candidate

Stefano LINGUA

DECEMBER 2020

† *To my beloved aunt Angela*

Index

Introduction

Nowadays the majority of companies are able, through computer systems, to collect and store a huge amount of business data. The thesis project aims to model and extrapolate information and metrics that can help the client company in the increasingly complex decision-making process

The project will cover the data life-cycle in its entirety :data extraction, enrichment and visualization.

The study will be divided into qualitative and quantitative analysis, alternating engineering processes with purely management reasoning aimed at an easier and more complete understanding of the business.

Specifically, the project will deal with the "market basket analysis", an area of great interest as it allows the emerge of hidden relationships and ties in the decision-making processes of a consumer who fills his shopping cart.

The project will mainly focus on the description of the *DWH* engineering processes through *ETL* instruments and on the study of business metrics then used in the front end of the process by the customer through the use of views and dashboards.

Environment The project was carried out within the MediamenteConsulting team, a consulting company operating in the field of Business Analytics and BigData based in Turin and Bologna. The project has been carried out inside an existent framework owned by MediamenteConsulting.

This data analysis work is of great value for the company which, thanks to the effectiveness of the metrics obtained, is able to consolidate its reputation and establish competitive advantage over its competitors.

Chapter 1

Data Evolution

This chapter will cover the evolution of data collection and analysis throughout history and then will enter deeper into the meaning of data in a more managerial point of view, covering the most important reasons leading a company to data handling and investments in data management and storage. The chapter will analyze different historical periods, at first considering data collection and then considering data analysis and management both in terms of quality, due to the high level of technicalities that will be implied in further chapters.

–

1.1 From Collection to Analytics

The concept of data is, at least in theory, quite simple to understand. If one would give a definition it may be said that *Data are characteristics or information, usually numerical, that are collected through observation. In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects.*[1]

The need of human beings to store and manage data to obtain useful pieces of

information has been constant throughout history from the beginning of time. What has changed over the course of centuries is the way in which Data are collected other than the technologies available in a given historical period.

The evolution of data consists of course in the systems available to process them, but also in the fact that data became much more than just record of occurrences, they are now a real asset in which also the biggest companies in the world strive to compete for.

1.1.1 Data records: Handwriting and Punch Cards

The first and most simple way in which Data are collected is the handwriting, which is of course a method still adoptable today but very inefficient when facing huge amount of data as it is often the case when considering large company or corporations operating all over the world. Things changed when Herman Hollerit, a former Census Bureau employee, invented the so called *Punch Cards*. Well, the first occurrence in which punch cards were heavily involved was in the 1890 U.S. census. As it is well known, innovation is the solution to the emergence of new needs, either real or perceived. In this case the U.S. census bureau was facing capacity issues, they were collecting more data than they could process. The need of a new way to collect and manage data arose. Hence the bureau held a competition in order to improve efficiency and capacity of data tabulation. Hollerit's solution was astonishingly faster than the two other candidates'. These cards where basically piece of paper with holes with the key characteristic that they were able to be read by a machine. Hollerith would have then founded IBM few years later giving proof that he had a very clear understanding on how to satisfy the increasingly need of better off data collection capacity. In origin the process would go through two main phases constituted by two different machines.

The first step was performed by using a card punch machine, which was in charge

of "punching" holes into a card starting from census schedules (the precursor of this technology was the Hollerith Pantographic Card Punch in the 1890). A bureau's clerk was able to produce on average 500 punched cards per day.

The second step concerned the use of a card reader machine, constituted by two hinged plates operated by a lever. The operator could put a card between plates and then closing them in order to make the upper metal bars go through the holes in the card and reach the mercury wells on the bottom of the machine, completing an electric circuit after which the clerk would transcribe data.

Nowadays, punch cards are of basically no use due to increased obsolescence, in fact new ways to collect data are mostly better and more efficient. However this process was the dominant paradigm throughout almost the entirety of the 20th century and represents the connection between handwriting and computers as we intend them today.

1.1.2 Analytics 1.0 : Business Intelligence

As was explained in the section above, punch cards were a powerful instrument for data transcription and storage basically until 1950. In the 20th century in fact the technology evolved drastically, improving the efficiency of systems in few years, sometimes even in months. What changed was also the way in which Data were looked at. Punch cards and handwriting were merely data collection, hence they were strictly related and used in scenarios where there was the need to collect great amount of data, but that was it, it was just a pure collection process. In the second half of the past century people started to look at data not only as a record but as a way to understand the insights lying beneath them, which is to obtain information. Information may be defined as *the set of data which is processed in a meaningful way according to the given requirement. Information is processed, structured, or presented in a given context to make it meaningful and useful.* [2] This was the first time in

which data regarding sales, customers, cash flows and more were collected in order to provide support for the customers' decision making processes. This was the uprising of *DWH* (Data Warehouse) where customers and transactions were centralised into one repository like *eCDW* (Enterprise Consolidated Data Warehouse). Business Intelligence processes entail for the most part *Descriptive* and *Diagnostic* analysis respectively regarding the description of the fact and the reasons why it happened.

1.1.3 Analytics 2.0 : Big Data

With the advent of the Internet and the establishment of worldwide connections, the way in which data were analyzed moved one step further. If one wants to point out in time the begin of Big Data analytics could set 2009 as a start date. It has to be said that when referring to Big Data one may think that there is a certain size which states if the file or datum taken into consideration can be considered a Big Data. However, the definition of Big Data must be always expressed with respect to a reference system, e.g :

- A 100MB file is Big Data with respect to an email sending system, whereas it is not Big Data if considered in a common DWH environment that can easily handle gigabytes.

We talk about Big Data mainly due to the fact that companies started proposing their business online, it is the birth of the *eCommerce* business model. By going online and relating with customers on websites, companies soon started to understand that the flow of data regarding different spheres of business were increasing rapidly. This kind of analysis entails the deep understanding of data as a source of insights for business strategic and operative decisions. The IT and Business Analysts focus their efforts in *Predictive Analysis* due to the stronger need of companies related to market trends forecast.

One of the biggest changes with respect to Analytics 1.0 is that now it is possible to process data, with the use of *ETL* processes (extract, transform, load), the so called *Un-structured Data*.

- One of the goals of the Big Data Analysis is to collect un-structured data and make sense of it.

The biggest difference between structured and un-structured Data is that the former concerns data in databases in formats such as *.csv* and *.xlsx* whereas the latter regards data that does not have a *meta model* that neatly defines it. For example, image files, audio files and the video files.

Nevertheless the term *BigData* has also another meaning which is widely use. BigData is also a term used to point out that everything that a common visitor does on the internet leave traces behind, and these traces are often use to obtain large scale views about customers behaviours, sometimes even breaking the law from a privacy point of view.

This interesting topic will be analyzed more in detail in the next section of this chapter.

1.2 Current Use of Data

In the previous section has been depicted the evolution of data, considering the infrastructures and the processes in which they go through. The aim of this section instead, is to describe the managerial reasons underneath the use of data in different kind of companies.

1.2.1 Forecast

From the advent of Business Intelligence data has become more than the description of the past, it is now a way to depict the future. Every day that goes by leaves behind

a different world, the change is extremely fast and sometimes it can be disruptive leading to huge losses for the companies. Being able to foresee future behaviours and trends is vital in order to maintain the scale of the business and hopefully to sustain growth. When doing forecast analysis two main components determine the final result

- the quality and quantity of data
- the analyst experience and know-how base

Regarding the quality of data it has to be said that nowadays a firm willing to perform forecast analysis has at his disposal a number of models and algorithms powerful enough to obtain reliable insights on how the *KPI* will behave in the considered time bucket. However, as the insiders well know, these algorithms need to be *trained*. To this end it is also important the amount of data stored in the database, either one uses an algorithm or a regression model the amount of data is very important to obtain any hidden insight or trend lying underneath raw data.

In simple terms one could say that a forecast model is trained when it is tested with different *hystorical series* in order to train it. This lead eventually to a more efficient algorithm and therefore to more precise forecast.

This kind of analysis require not only data quality and quantity but also a good and consolidated managerial experience. This is basically one of the reasons why consulting companies were born in the first place, in particular when talking about data, that is the case of companies in the IT field as MediamenteConsulting.

As it will be explained in detail in next chapters the need of customers companies is to obtain operating insights on their operations and this require not only an articulated system of technologies and machines but also a thorough business analytics experience, and this is why the client firms contact consulting companies, to obtain this kind of useful insights and support decisions. The fields that could be targeted by forecast analysis cover almost every business unit of a given firm, this is another

reason why inferring is such a powerful instrument. Generally, forecast is carried out with the aim of optimize one of the following activities :

- Sales planning
- Financial planning
- Market evaluation

1.2.2 Manipulation

As was introduced in the section regarding Big Data, nowadays the technological infrastructures in which people live entail the intensive use of the web, either with smartphones or personal computers and this yields an outstanding amount of metadata and traces left behind on the internet. This implies a very serious ethical debate, data has become in these scenarios an instrument for mass manipulation and tracking. Unfortunately, the always increasing technological features in smartphones are nothing but a way to make this phenomenon even larger.

For example one may think to vocal assistant software. This is a clear example of unstructured data processing, with the implementation of this feature big companies, may be able and sometimes they actually are, of collecting private pieces of information and possibly use them without the permission of the user.

If instead of a singular point of view, a global perspective is adopted it is clear that the implications and the overall value of Data is literally non quantifiable.

This subsection, even if very distant from the actual core of the thesis project, has been depicted in order to obtain a clear overview of the evolution of data throughout history. From punch cards, when data needed to be collected faster, to Business Intelligence where the data started to be analyzed in database and used for business support decisions, to eventually the Big Data world where not only data are used for support in the decision making process but also to obtain worldwide scale pieces

of information of what people want, desire and like. This has been due also to the advent of social network which had drastically increased the average amount of time people spend on the internet.

Chapter 2

Status of the Art

In this chapter will be described the actual status of the art, regarding the processes and concepts which constitute the core of this project.

By doing so, further chapters which will be much more quantitative, ought to be of an easier understanding.

2.1 Types of Data Analytics

As was depicted in the previous chapter, data has become a fundamental asset for companies. The continuous growth of technological infrastructures and devices yields an outstanding quantity of data collected. If one takes as example a made up database containing a certain amount of data there are various types of analysis which can be performed, depending on what is the goal of the analysis. Generally, these analysis are divided into four main categories, each one will be described in further subsections, they range from the least to the most complex in terms of knowledge, costs, and time.

2.1.1 Descriptive Analysis

This is the most common use of data, it is the description, as the name suggests, of real facts and proceedings and responds to the question

- *What happened?*

This analysis is carried out with the usage of data aggregation and data mining activities, it is not related in any way to the discovering of hidden relationships and ties among data. Most of the times, this analysis is the first step in the analysis process and it is visualized through charts showing means and quantiles. For example one may evaluate a scenario in which an owner of a website is interested in finding out how connections and visits on their website did perform in the last month. This is a clear representation of descriptive analysis since the user is not looking for reasons or causes but for pure facts and numbers. The given user may use for example Google Analytics which is an online tool providing few charts on different key performance indicators (KPI) such as:

- KPI dashboards
- Monthly revenue reports
- Sales overview

2.1.2 Diagnostic Analysis

Diagnostic analysis as the name clearly suggests, refer to evaluating the performance of the KPIs considered in the descriptive analysis, looking for causes and explanations found by studying trends. It responds to the question

- *Why did it happened?*

These analysis are carried out starting with the individuation of relevant anomalies or patterns in the available data. The purpose of this process, is to find out which are

the reasons why a certain *outlier* is present. By doing so, the data scientist is able to depict routines which lead to unexpected results and therefore operate on them, this is of enormous business relevance. In fact through pattern definitions further decision-making processes will be easier and with all probability more effective. A commonly used approach in this field is the drill-down.

Drill-down refers to the evaluation opportunity of a particular set of features. For example, if a grocery store finds out that the sales in a given month are below expectations they may want to see for each region in which they own stores, the actual sales and have a good chance in finding out the reasons why is it so. Furthermore, diagnostic analysis are also valuable when searching for what is good for the business, evaluate the causes and use them for further business decisions. What was good for the past should be good for the future, at least from a routine point of view. The actual performance in the future will be the core of the next two types of analysis.

2.1.3 Predictive Analysis

The predictive analysis are probably the analysis which lead to the major value for the companies performing them. This is due to the capability of evaluating the past in order to depict the future. This type of process is performed to answer the question

- *What it is likely to happen in the future?*

By evaluating past patterns and trends data analysts are able to create a likely scenario of what will happen in the future, this allows the company to plan ahead which is of course a much more effective way of operating rather than to react to facts. This procedures are generally quite complex, as was above mentioned the types of data analysis are presented in order of complexity. However, as it is often the case, the higher the knowledge base and the competencies requested to perform an analysis, the higher the potential benefits for the company. For instance, if one

considers a grocery store, it could be useful to evaluate future trends and sales by stating a certain number of *regressors*. Considering for example seasonality on a certain group of products, a better understanding of future sales can be obtained and therefore decide how to handle inventory rotations. Predictive analysis is also very useful in the risk management field. By evaluating future scenarios the data analyst is able to suggest contingency budget and also operate in order to reduce as much as possible the impact of these risks.

2.1.4 Prescriptive Analysis

These kind of analysis are directly connected to predictive ones, in fact they may be considered as a combined analysis. These analysis are meant to answer the question

- *What is the best course of action to take?*

Prescriptive analysis use both results of descriptive and predictive analysis, with the implication of machine learning activities. Machine learning is a completely computerized way of operating, it entails the use of algorithms and advanced software in order to provide useful support decisions, sometimes these systems are also able to work on their own, putting in place the operative measures obtained. This is the case of AI (artificial intelligence), the frontier of the forecast processes which entails a fully automated job structure. The purpose of these analysis is to map all the possible "routes" which may lead the company to the aimed target. An often cited example of a prescriptive analysis is the one implied by Google Maps. When the user inserts on the platform the information related to the starting point A and the destination point B the tool computes all the possible routes to get there. In a similar way prescriptive data analysis have the role of leading the company to the best route which entails the most effective actions.

2.2 Data Mining

Data mining is the field in which the whole project is built around. It defines the activities that will be implied in order to provide the client with support in the management's decision-making processes. The term clearly reminds to mining, in fact at the same way raw data may be seen as the rock in which the miners constantly work on. It is actually a very precise comparison in fact raw data are generally "dirty" data, meaning that they entail noise, corrupted data and nulls. The uses of data mining are multiples, but they can be redirected to three main categories:

- Marketing strategies
- Increase sales
- Decrease costs

The process takes place inside every phase of the data life cycle, at first the data must be collected. For instance groceries stores and retailers commonly are interested in data mining projects, to do so they offer some exclusive promotion to customers in possess of the loyalty cards. Then data are loaded on datawarehouses, this allows a better consistency of data and therefore the final users are able to organize and transform data in a format which can sustain the enrichment analysis done in the front end. In the following subsections the concepts of datawarehouse and ETL process are depicted, being them the fundamental bricks of every data mining operation.

2.2.1 Database

A database is an organized collection of data, which is generally stored and accessed electronically from a computer system [3]. The term is loosely used referring to one of the three main components of a database that are :

- DBMS (Database Management Systems) which are responsible of the extraction of information upon a query instruction
- The database itself which contains all the data
- Associated applications

When referring to databases the core is in the DBMS. This allows the customer to question the database and the file contained in it, this is exactly why the word query is used in this field. However, databases are also powerful when it comes to cross-referencing capabilities. In fact when the database is used the field of all the tables may be connected and joined with the use of the SQL language (Structured Query Language) which is not a pure programming language but instead a declarative one. Its level of complexity is moderate if compared to programming languages and by being declarative the user is able to traduce with almost no effort what they want into a language which is understood by the database.

This is the case of RDBMS (relational database management systems) which are widely used as they use data in forms of tables constituted by rows and columns. The basic relational building block is called *tuple* and refers to a set of attributes.In the following figure an average table of a relational database is depicted.

Another key feature of a database is the size. There is no best size for a database due to the fact that a database is built with respect to the amount of data that needs to handle. A small database may be hold and managed even from a single personal computer. However, the power of DMBS is that they are able to connect millions even billions of table rows, for that reason often the size of databases is very big. Generally the bigger the company and the amount of data that need to be processed the bigger the size of the database. One example of a huge size database is the one which is adopted by governments and revenue agencies. If something looks odd and the revenue agencies wants to go deep in the transactions made by a citizen, then it

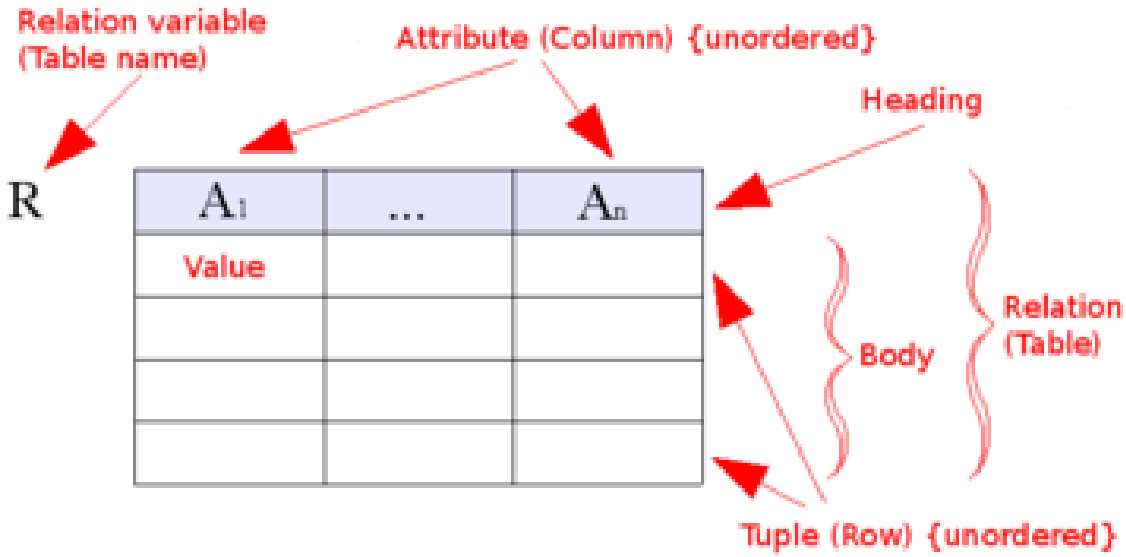

Figure 2.1: Relational model concept

does not only need a huge amount of data but need data from all over the country, this is feasible only by using a huge server database.

2.2.2 Datawarehouse

A datawarehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making. [4]

One of the main objectives of the data warehouse is the permanent reduction of the cost of releasing information. In other words, the datawarehouse appears as a highly efficient process to respond to the ever-increasing information needs of the final user. As the name itself suggests a datawarehouse is a warehouse of data, it is a system of databases which generally are entitled of the collection of operative and transactions data for further use. This systems are needed due to the incompatibility of traditional OLTP systems with online analysis, in fact by doing so the data may result corrupted leading to a great reduction in consistency. Generally

the unique datawarehouse is split into some number of *datamarts*. A datamart is a datawarehouse which collects and manage only data related to a certain department or division. This is often a good way of structuring the data architecture, datamarts are then managed by the business unit which has the competencies and work in this given department. There are two other main types of datawarehouse environment architecture:

- Enterprise Data Warehouse (EDW) is a way of organizing the structure in a centralized way. This approach has the advantage of allowing constant control over the quality of the data but requires more careful planning. It is entitled of tactical and strategic decision support.
- Operational Data Store (ODS) is used when neither an OLTP system nor a common datawarehouse are able to support organizations reporting needs. It is used for operation reporting, controls and decision making.

2.2.3 ETL

The Extract-Transform-Load process is articulated in three phases as the name suggests. The extraction phase consists in taking data from source systems which are commonly :

- OLTP (online transaction processes)
- ERP (enterprise resource planning)
- CRM (customer relationship management)

The transformation phase entails the cleaning and consolidation of data. These are some of the operations that are commonly performed :

- Select only data which are relevant to the system

- Normalize data (for example by removing duplicates)
- Derive new calculated data
- Perform couplings (joins) between data retrieved from different tables

The transformed data are then uploaded on the final synthesis system, commonly the datawarehouse. Furthermore, the granularity of the data need to be correct either from a processing speed point of view and from an analysis point of view. If the granularity is to high in the downstream system then the queries will be efficient but the analysis may be difficult to perform. In the following figure is depicted the flow of data inside a DWH environment.

2.3 Visualization

This field is constantly improving in recent years due to the always growing data amount at disposal of firms and individuals. But why is this step so important one may wonder, why data integration steps are not enough. Well, to respond to this question one may approach the discussion from two different point of view:

- Anatomy of human brain
- Business KPI and ease of use

The former gives us the chance of explaining why, from an anatomic perspective, this visualizations are so effective in communicating a message to the final user. The way in which this occurs in practice is related to two different section of the human brain. The anterior section of the brain is entitled of the "seeing" operations, or as they are commonly called the perception activities. These are computed and understood very efficiently and rapidly from the brain and this is why the images are universally referred to as the universal human language. The other section of the

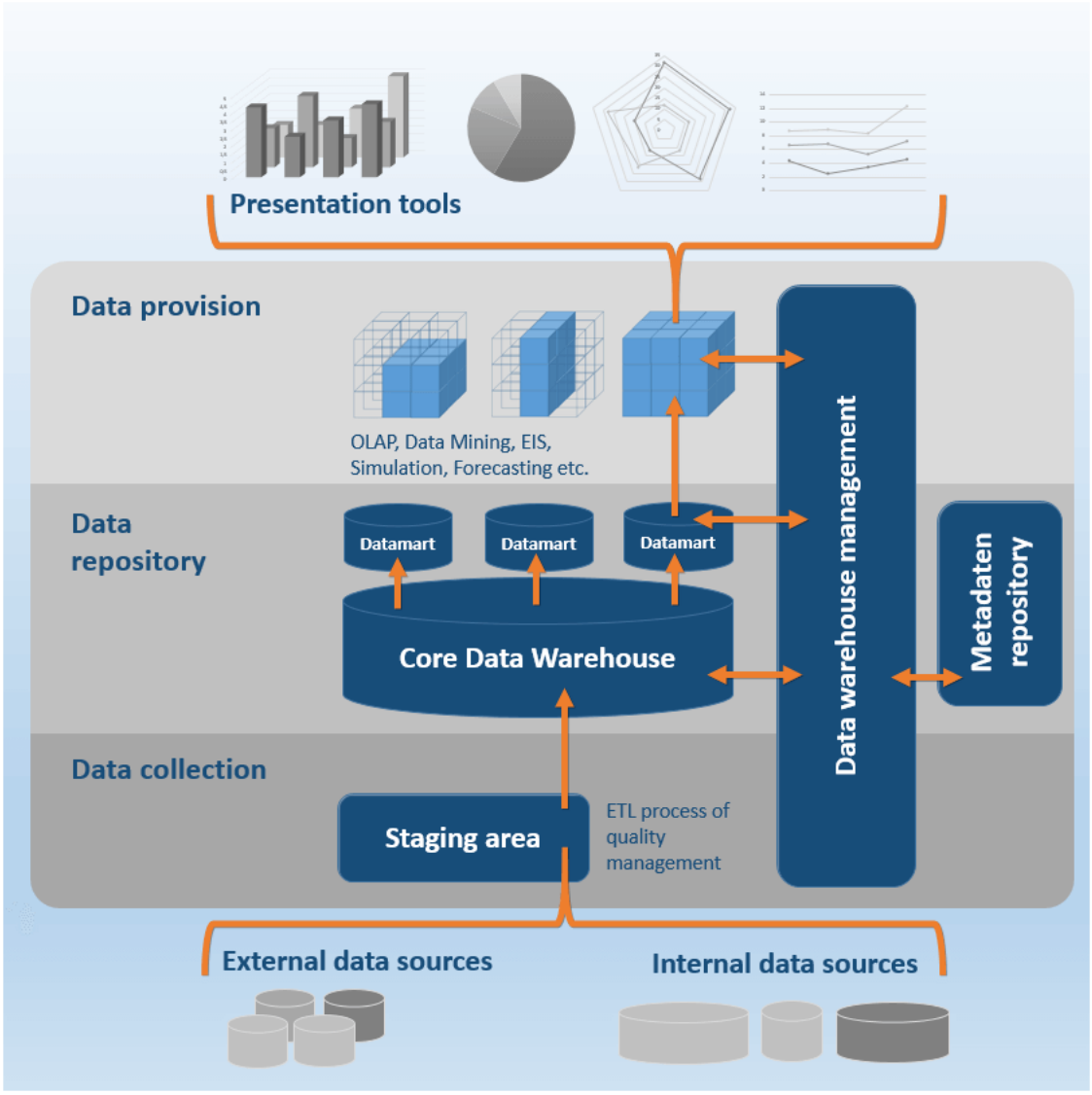

Figure 2.2: Data flow schema in a Datawarehouse

brain involved is the posterior which is entitled of the "thinking" or also cognitive processes. These processes involve deep reasoning and of course take more time to be elaborated and understood by the brain. This is the firs big reason why data visualization is so important, it entails mostly perception instead of cognition, this leads to a faster understanding of the message by the final user. The latter is so

frequent that nowadays it's not even looked at as something innovative, however the use of KPIs is very interesting when referring to visualization processes. This regards the use of measures universally known such as revenues, margins and so on. By doing so the information conveyed is most of the time one number, one chart something very flashy one could say that instantly the final user can understand. Of course by doing so all the processed data that are the basement of the visualization are somehow left behind and not considered properly, but this is not a problem. In fact the aim of the visualization is to balance the ease of understanding with the complexity of data, if the business analyst is able to mediate between these two components the visualization will be of a certain value and be appreciated by the client. Designing a report or a visualization is also a process which highly involves capabilities related to color variations and space filling.

This seems very far from the integration phase, in fact these two phases are generally performed by two different business units inside MediamenteConsulting. This is also a focal point of the project as expressed before, by approaching both business unit it has been possible to fully describe and learn the data life cycle in its entirety. In every visualization activity the business analyst needs to keep in mind two fundamental considerations

- What the final user is looking for
- How can the visualization satisfy client needs

To solve the first problem, the designer of the visualizations must have a good acquaintance of the field in which the customer operates. Related to this project one should know how the large-retail distribution works and which are the main key performance indicators mostly taken in analysis by the final user. This is a pretty straight forward work, hence this leads only to descriptive analysis, representation of what has happened.

In the following sections the most commonly used tool and charts for visualization purposes will be presented.

2.3.1 Kinds of Representation

The way in which data are actually shown is with the use of charts. There are plenty of different kinds of charts, however some of them are used more than others and the reason behind that is they are able to convey information more efficiently. Below will be provided a brief overview on the most common charts used for visualization purposes.[5]

Clustered Bar Chart

These charts are a good solution when the objective is to show the evolution of different values or groups clustered over a given period of time. This solution is possible only if the values are comparable with the same Y axis. There are variants of this chart either horizontal or vertical, in general the horizontal solution is preferred when there are some negative values, the vertical is instead the most common due to the higher level or readability.

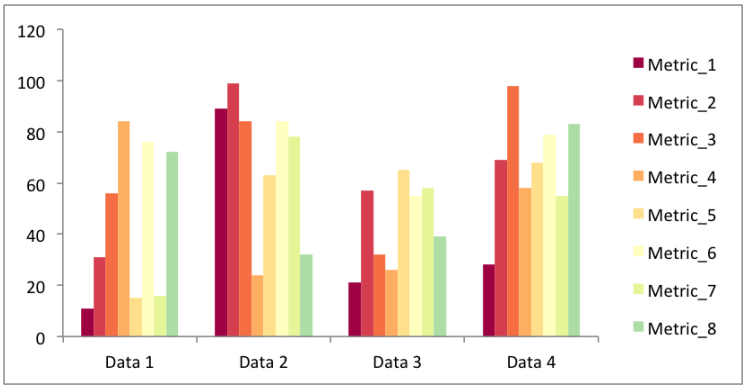

Figure 2.3: Example of Clustered Bar Chart

For example if the sales manager of a sport apparel company wants to overview

the performance of different categories of product sold, then this is a solution that may be satisfying. In general the human brain is trained and used to read from left to right, by using a vertical clustered bar chart this inclination is supported and therefore this yields an higher appreciation by the final user.

Line Chart

Line charts are very simple to use, they are formed by discrete points usually called "markers" and segments or lines which connect consecutive markers. In general they are used for trend purposes, by being discrete representations using almost always time as X axis, entailing one or more variable. This visual representation is useful if used with a single parameter, because the final user instantly understand the trends and the pattern of this given parameter over time. However, they are of greater impact when considering two different values in order to show the trends over time. As shown in figure 2.4 line charts may be used also for competitive comparisons.

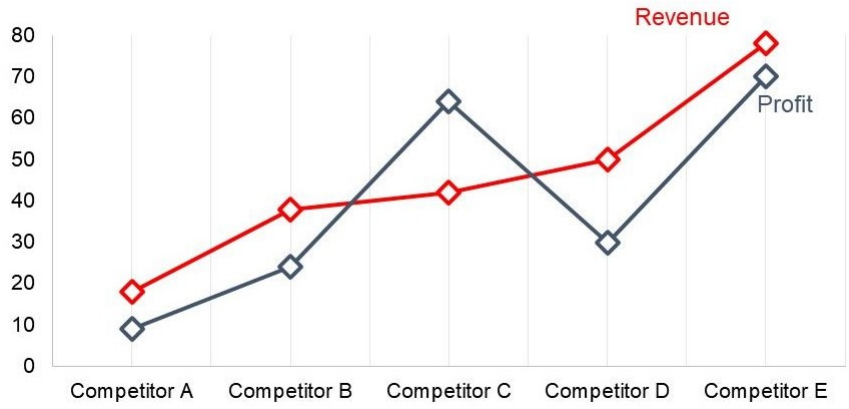

Figure 2.4: Example of Line Chart

In this case on the X axis there is the list of competitors of a given company and then the lines regards the revenue and profit. Of course this does not entail any trend evaluation but it is good for an higher level consideration and may support

the decision making processes of acquisitions or partnerships. In the end line charts are one of the most used visualization instrument due to the high level of flexibility that they entail, and due to the ease of comprehension that they feature.

Pie Charts

Pie charts are one of the most commonly used charts and in fact they are very good at showing a comparison between two categories. However pie charts in particular have the problem of not being good when there are more than two categories to show, this is due to the difficulties of the final user in assessing which is the bigger portion of the chart and therefore understand the relationship among categories. Another issue related to pie charts is that they are generally too big for the information they convey. When designing a report it is important to comprehend which proportions and which location every chart need to feature.

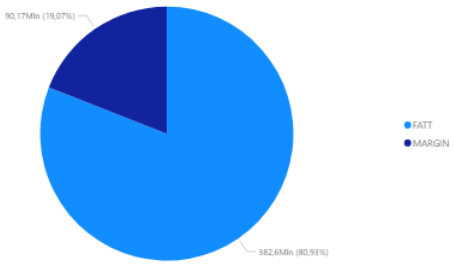

Figure 2.5: Example of Pie Chart

An alternative to pie charts are *Donut Charts*. This visualization is almost the same as the one obtained with a pie chart with the major difference that the circle is empty in the center, as result the chart looks like a donut, that's why the name. This alternative is often preferred to pie charts, in fact the empty center saves space allowing the designer of the visualization to apply further piece of information and percentages related to the sections of the chart.

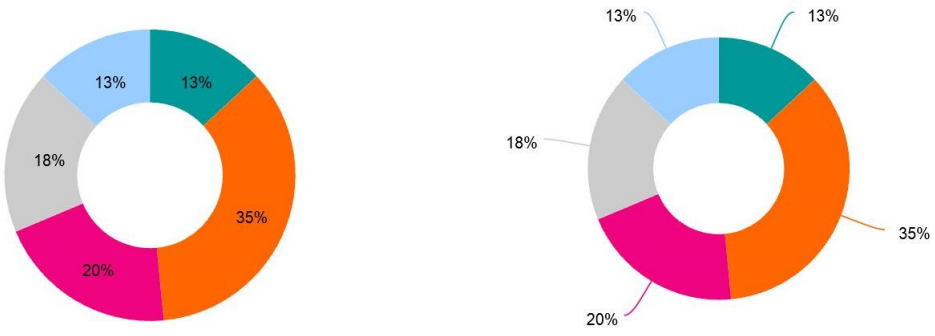

Figure 2.6: Example of Donut Chart

2.3.2 Power BI

This brief subsection describes the business intelligence tool which will be used for visualization purposes and why it has been chosen over other alternatives. Microsoft Power BI is an advanced business intelligence tool owned by Microsoft corporation. This instrument is somehow recent in fact was developed and commercialized in 2014. By connecting data source to Power BI the final users are able to produce useful reports and dashboards which will then be very valuable for decision making support processes, and evaluation of the current status of the business. When referring to Power BI, one actually refers to an umbrella of three different components.

- Power BI Service
- Power BI Desktop
- Power BI apps

The Power BI Service is an online SaaS (Software as a Service) in which the final user may connect and publish the reports produced on the Power BI desktop tool. Of course the designers of the tool were well aware of how important it is nowadays to have a smartphone app able to sustain the operations which are normally carried out online. The two other main features leading Power BI at the top of the visualization

field are:

- Possible source of data connection
- Similarities with other programs

Power BI is designed to support loads of different sources of data connection, from a simple Exel file to SQL Server connections and many others, furthermore Power BI is infused with machine learning capabilities, meaning it can spot patterns in data and use those patterns to make informed predictions and run "what if" scenarios. These estimates allow the users to generate forecasts, and prepare themselves to meet future demand and other key metrics. As it is well known, a tool is not only evaluated for its technical features but also for the ease of being picked up and understood. This is also a point where Power BI shines, in fact being produced by Microsoft implies that the developers could use analogue features of Exel, which is heavily used worldwide. Last but not least when presenting Power BI, one may not forget to mention the share capability of the tool and its ease of use. In fact Power BI reports are very easy to share with other people, however only with the premium version of the tool is possible to obtain a modifiable copy. It has to be mentioned that this service use a language to compute measures and columns which is called DAX (Data Analysis Expressions) very similar to Excel functions and calculations. Again this lead to a major reduction in issues related to learn how to use the tool and therefore better off customer experience. Power BI, however, has some drawbacks. One is for sure related to the fact that when using Power Query, the component which is entitled of data source connections, the tool is not able to accept direct SQL queries as input. Another features that sometimes leads to some disadvantages over other visualization tools, is that even if Power BI is designed to support a very good number of different connections, inside a given dashboard or report it is possible to connect data from a single source.

Chapter 3

Problem Background

This chapter is written with the aim of clearly show which are the problems which led to the beginning of this project of thesis. There will be a careful evaluation of what was needed by the client, and the problems or issues that the consulting company was having in satisfying these needs. After the overview of the problems which need to be solved, it will be provided also a description regarding the approaches that have been chosen to solve them, this is mostly a declaration of intent. The actual work performed in order to technically solve the problems will be reported in further chapters, in particular in chapter 5 and 6.

3.1 Customer Need

The need which is at the basis of all this work, is of course related to the client. The actual field in which the client firm operates will be shown in dept in the next chapter, when the large-retail distribution industry will be over viewed. The client which will be referred to as Alpha in order to respect confidential agreement, after a successful relationship with the consulting company related to a datawarehouse creation project, shown the interest on the possibility to implement a new project

related to the design and implementation of advanced analytics. As first instance, the client communicated to the consulting company the will of obtaining business support on decision making processes, in particular aimed towards a better management of promotional campaigns. The project started before my personal entrance inside MediamenteConsulting in an internship program, however the results obtained showed some critical issues, due to this reason it has been decided to start this work with the aim of solving them and therefore implement and better off the existent advanced analytics framework. Therefore it has to be specified that the real driver of this project is not the customer, whose will and need was already clear to the consulting company. The central goal of the project is instead related to the critical issues, of different nature, which will be described in the further section of this chapter.

3.2 Consulting firm problems

In this section will be presented the major critical issues that the consulting company was having on the advanced analytics project. The solution to this problems is the goal of this work of thesis. As was above mentioned this issues are of different natures, by nature meaning that there are technical issues and managerial issues. The former may be substantially defined by the following statement

- Obtain a consistent framework for the advanced analytics project.

The latter is instead well described by the following statement

- Obtain a measure intelligible by the customer to clearly communicate the results of enrichment modules.

These are the driver of every reasoning that will be entailed in further descriptions and processes. By keeping this issues in mind, it has been possible to move forward with a clear objective and this is a very good way to obtain a valuable solution,

working with a well defined goal and then try to tackle the problems in order to obtain a satisfying end result.

3.3 Solutions

In this section will be described the way in which it is intended to solve the problems above mentioned, either technical or managerial. Since the beginning of the work it was clear that it would be best either from a formation purpose and from a thoroughness perspective, to perform activities related to data life cycle in its entirety. This entails the division of the work in two different but sequential phases, which are normally handled by different business units inside the consulting company.

3.3.1 Integration

This section of the work, may be seen as the basement of the whole project. Here, the goal is to obtain a framework using a data management tool with enough capacity to handle large databases and high complexity computations and operations on raw data. The objective is to obtain a data flow featuring this key particularities:

- Adaptable, removing the need of designing a new one each time
- Efficient, it has to work within a reasonable time window
- Clear, the steps of the integration framework need to be sufficiently self-explanatory of the work performed

This is an approach which entail a very clear formation on the fundamental of database, ETL, SQL and datawarehouse. It has been performed, with the help and supervisions of the consulting company employees, by doing so the endless number of technical issues related to tools and systems has been exceeded.

3.3.2 Visualization

The result of the integration activities is the basement of the project, however it is equally important the way in which the results are shown to the client. This problem, related to obtaining a measure intelligible which would be easily understood by the client, has been performed inside a visualization environment. The objectives of this phase are related to obtaining a some numbers of dashboard, which will then be used by the client to support their decision making processes. This work is also related to more managerial reasoning and therefore, it will be also important to compensate the high level technicalities of the obtained integration result with the most commonly used key performance indicators. To this end, it has to be pointed out that it will constructed an aggregated measure to ease the process towards a solution to the problem mentioned above, the ease with which the final user is able to relate with the dashboards and therefore support their decision making processes.

Chapter 4

Case Study

This chapter will discuss the case study in consideration giving a clear idea of what this project of thesis is aimed towards, yielding an overview of the advanced analytics project, its metrics and business use. As was previously explained, this project is related to research and development activities, starting however from a consolidated framework constantly improved by MediamenteConsulting.

4.1 Large-scale Retail Distribution

The case study refers entirely to a large-scale retail distribution company operating in the south of Italy, this firm will be referred to as Alpha in order to respect privacy and confidential agreements. Alpha is a supermarket holding which manages 4 different brands with a total of more than 200 retail stores. These brands are different retail stores with different peculiarities both in square footage and targeted customers:

1. A brand comprehends the bigger stores with a square footage of 1000 mq or higher. These are the store provided with almost every type of good, they are visited for the most part by customers willing to do considerable shopping.

2. Another brand collects the medium stores featuring a square footage until 800 mq. These are store generally targeted by customers doing the everyday shopping.

3. A brand focusing on convenience formats, targeting convenience price policies and consume local behaviours.

4. A brand addressed to professional selling, featuring stores with a 2000-3000 mq square footage.

The size of Alpha entails a huge amount of available data both fro collection and analysis purposes, as it has already explained data quantity is a very important requirement for useful and meaningful business reasoning and inferences. The main problem that Alpha was facing and which led the company to stipulate a contract with MediamenteConsulting was that its DWH was slow and obsolete, being not able to process data coming from different datamarts, that are smaller datawarehouses collecting data from a specific business unit. This was of course a relevant need, in fact in order to manage correctly such a huge amount of stores all the data must be well organized and usable with an high degree of interconnection among different business units. This first project, required a year of intense work after which, due to the client satisfaction for the new datawarehouse, a new project was signed connected to the need of further analysis on data.

Alpha is clearly well aware of the potential of advanced analytics processes, being so it started a project with the consulting company in order to be provided with useful metrics and indexes mostly aimed to a better understanding of the business status and therefore as a support for decision making processes. The initial request from Alpha concerned the evaluation of promotions, how promotions influence sales and when and which product should be promoted to stimulate sales.

4.2 Advanced Analytics

The advanced analytics project is related to the management and processing of huge amount of data, as was depicted in previous chapters nowadays the urge of using data for forecasts is one of the most value attributive feature of current consulting processes. First it has to be said that these procedures take place inside OLAP (online analysis processes) systems, these systems are the last step of the back end activities, that is where data mining and querying take place. Advanced analytics entail autonomous or semi-autonomous analysis of data through sophisticated technologies, in order to achieve a deeper understanding of information and make predictions useful for the future development of the business.

The main difference between traditional analysis and advanced is that whereas the former is related to historical data analysis only, the advanced analytics collect and process real time data in addiction to data series analysis.

This way of managing data leads to enormous growth potential for the companies which are able to integrate this kind of activity in their business model, furthermore if performed correctly is often bearer of competitive advantage both for the client and the consulting company.

Generally, advanced analytics operations are performed with one of these three purposes :

- Business planning and control optimization
- Identification of new trends
- Improvement of decision making

If one looks at this procedures from a consulting point of view, it is very clear where the potential dwells. It leads the consulting activities to a whole new level providing the customer not only with datawarehouses able to efficiently respond to business

needs but also is provided with metrics, which are thought and implemented according to managerial reasoning jointly with simple and efficient views and dashboard as the final product of the whole process, throughout both integration and visualization activities.

4.3 Market Basket Analysis

The market basket analysis represents the field in which the whole advanced analytics project for Alpha is centered. It is based, at least in principle, on a very simple concept:

- Evaluate and discover hidden relationships among products in a given shopping cart

This is the aim of the project, analyzing and processing large amount of data in order to obtain a clearer view of how and how much different products, of different categories, of difference price levels, are bought together and how much this connection is strong. The value of this analysis is very high due to the impact on sales control and monitoring and also for the decision making processes. These analysis are normally carried out inside a large-retail scenario to improve and increase the Cross-Selling which is related to the impact of selling a certain product on another one or even a couple of products.

4.3.1 Association Rules

The way in which it is possible to state the relationships among products in a check is using what is usually called association rule. However this is a purely qualitative approach, meaning that the results obtained are not useful from the perspective of the client, but they are the basement for all further indexes and metrics that will be presented in the visualization processes. In the used framework this concept refers

to two possible alternatives, a rule featuring two products or three. However the implied product, which is pulled by the purchase of the implicating ones is always unique.

$$A => B \tag{4.1}$$

$$A, B => C \tag{4.2}$$

For the sake of clarity, one may read this formulas with a sentence like this: *the purchase of A drags the purchase of B*, the same is valid for a rule featuring two implying products. One may also consider this formulas from a more probabilistic point of view stating :*the purchase of A increases the probability of buying B*. This is of course a very linear concept, however it is the basement for all further analysis carried out either in the integration processes with ETL instruments or in front-end views and dashboards. The implications of association rules are related to the strength of the tie, in order to evaluate this bond some additional concepts and measures need to be defined. In the following figure its depicted an example of a set of checks and the relative products.

At first must be defined the terminology used to refer to products purchased in a given receipt

- *itemset* as a collection of one or more items
- *k-itemset* as an itemset featuring k items

Now let's consider the measures which allows to understand the actual frequency and the impact of a given association rule

- Support Count: It's the number of occurrences of a given itemset 'X' in all the transactions

$$SupportCount(X) = frequency(X) \tag{4.3}$$

- Support: Fraction of transactions having the itemset 'X' in it

$$Support(X) = \frac{frequency(X)}{N} \tag{4.4}$$

- Confidence: Fraction between the frequency of the itemset and the frequency of the antecedent

$$Confidence(A => B) = \frac{frequency(A, B)}{frequency(A)} \tag{4.5}$$

- Lift: Fraction between the rule's support and the product of single supports

$$Lift(A => B) = \frac{support(A, B)}{support(A)support(B)} \tag{4.6}$$

For a given rule A => B, the support is obtained with

$$Support(A => B) = \frac{frequency(A, B)}{N} \tag{4.7}$$

Confidence may be considered also with a probabilistic approach as :

$$Confidence(A => B) = \frac{P(A \cap B)}{P(A)} \tag{4.8}$$

The most used of these measures are Confidence, Support and Lift. This series of indexes is very helpful when it comes to evaluate and make inferences on association rules. In fact these are the parameters used both in the data flow and in the visualization instances like views and dashboards.

In order to make this concepts clearer they will be now applied to a made-up check.

Let's now assume that A is Bread and B is Milk which is of course a non interesting association rule from a business perspective, it is used for clarification purposes. Applying these measures to the given example yields:

$$SupportCount(Bread, Milk) = 2 \tag{4.9}$$

ID	ITEMS
1	{Bread,Milk,Fries,Wine,Shirt}
2	{Meat, Fries,Candle,Shirt}
3	{Bread,Milk }
4	{Bread,Fries,Tomatoes,Carrots}
5	{Eggs,Milk,Flour}

Figure 4.1: Example of items in a check

$$Support(Bread, Milk) = \frac{2}{5} \tag{4.10}$$

$$Confidence(Bread, Milk) = \frac{P(Bread \cap Milk)}{P(Bread)} = \frac{2}{3} \tag{4.11}$$

$$Lift(Bread => Milk) = \frac{support(Bread, Milk)}{support(Bread) support(Milk)} = \frac{0.4}{0.6 * 0.6} = \frac{10}{9} \tag{4.12}$$

Regarding the values that the Lift may assume three different scenarios are possible:

- Lift > 1 there is a direct and positive relationship among items considered.
- Lift < 1 there is a direct and negative relationship among items considered.
- Lift $= 1$ the items are independent, this method is not able to produce any meaningful rule.

The aim of the whole process is to obtain a relationship among different products, however in this specific example there were only five transaction, in a real database the amount of data that need to be processed is extremely higher, what it's useful is then a standardized procedure, usually referred to with the term algorithm.

4.4 Apriori Algorithm

The Apriori algorithm is used in the advanced analytics framework for the extraction of the association rules starting from a scenario with k-itemsets. It's particularly efficient when referring to transactions data in relational databases. This algorithm entails a bottom-up approach, starting from a single item it will build up a k+1 cycles in order to find all the frequent itemsets,then it prunes the candidates showing an infrequent sub pattern. This word *frequent* is the main feature of the Apriori principle, the main constituent of the algorithm.

This principle states that *Subsets of a frequent itemset are frequent and supersets of a infrequent itemset are infrequent.*

However, there is the need to technically describe what is a frequent itemset.To do so the final user need to "apriori" set two values :

- *minsupport*
- *minconfidence*

The algorithm need this two parameters in order to complete the iterative process, there is no right value for *minsupport* and *minconfidence*, they are strictly related to the environment in which the algorithm is placed, for example the question which leads this decision in the market basket analysis is:

- *With what minimum frequency an itemset must be found to be considered of some business relevance?*

However, the algorithm presents some problems from an operational point of view. First is based on a fixed minsupport and second is very expensive when considering the computational aspect.

It is true that this approach requires an intensive effort in computational operations, however the algorithm is structured keeping in mind that starting from a X number

of transaction the nodes generated are equal to 2^X. For this purpose the apriori principle is very helpful, by pruning all the infrequent supersets the whole process is better off from both a time consuming perspective and from an efficiency one. In the figure 4.2 is represented an example of how the apriori algorithm actually works. For example let's assume that AB has a support of 0,22, if the minsupport decided is 0,3 then the AB itemset is infrequent and applying the apriori principle leads to the exclusion of all further supersets.

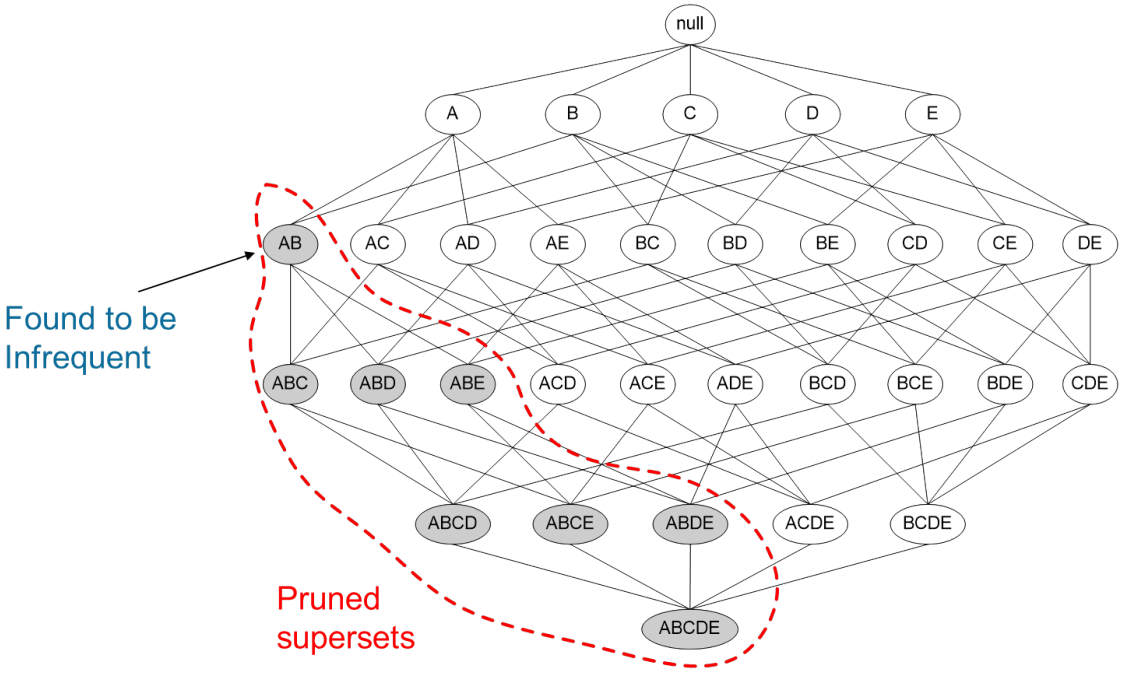

Figure 4.2: Apriori itemset evaluation example

Chapter 5

Data Integration

This chapter will describe in detail the work done inside ETL processes, the aim was to actually define a flow for the advanced analytics project able to produce useful fact tables that will be then used in the visualization phase for the construction of reports and dashboards. The discussion will be extremely quantitative and technical, in the first place an overview of the data available followed by a description of the assumptions made on those data will be provided. Then it will be presented the actual data flow that has been built with the ETL instrument DataStage, a tool owned by IBM. Despite its high level of technicality, the integration process is carried out keeping in mind what kind of data will be needed in the analysis steps, it fills the space between the collection of data from the client and the business analysis. The chosen instrument from the completion of the ETL activities is DataStage. This client is very efficient and very flexible for this type of work so it has been preferred over a more commonly used instrument like VisualStudio SISS.

The main features making DataStage a more efficient instrument are:

- The capability to manage and operate on higher size, higher complexity database
- The number of operations proposed on DataStage is much higher than on other

tools, this leads to a major openness for the user and their data transformation activities. This allows the user to directly operate on the ETL tool rather than on the database.

This work entails the massive use of SQL queries and instructions, hence the most important queries will be inserted in this chapter and fully commented in order to provide a clear description of how the data has been managed and transformed throughout the whole process. Furthermore, it will be also provided a short representation of the tables obtained and their layout.

5.1 Data Preparation

Before the presentation of the work performed, it is necessary to recall the major assumptions and reasoning made on the large scale retail distributor data. First, it has to be specified that the data used for every steps of this project, are related only to the loyal customers, meaning that only those who actually possess a supermarket loyalty card are considered to be ones actually shaping the trends of the supermarket, either from a behaviour perspective or a financial one (how much they buy and spend). From a preliminary examination what turned out was that roughly half of the total customers making shopping from Alpha are customers owners of a shopping card. Therefore the analysis are performed with the aim of discovering hidden insights over purchasing behaviours of these people. The main reason why the customers which do not own a supermarket card have been omitted, is that these are considered as outliers. For example if a customer is just passing by one of the stores of Alpha and decide to stop there and make shopping, their behaviours on purchase should not affect the overall result. In fact when working on the retail store industry a major indicator which is considered by top management is the ratio between customers gained and customers lost and this is a feasible analysis only if one considers the so

called loyal customers, having at disposal the pieces of information regarding their age, the frequency with which they visit the store and how much they spend on average for their shopping. These are all information not available for the "one-visit" customers.

One may debate over the fact that not every customers which does not own a supermarket card is visiting the store only once in a lifetime. This is actually very true, but until the customer does not sign for the fidelity card the supermarket is not provided with the useful data they need to do accurate forecasts and modelling. This is why retail stores offer for free these kind of cards, and provide promotions for the customers who own one. This is another example of how data is valuable nowadays, simply by considering what was mentioned above one may instantly understand the potential benefits that the retail store would be able to obtain if every customers visiting the stores was a fidelity card owner. Another problem that needed to be fixed was the actual relevance of the year 2020 inside the modelling. As we all know, this year has been very different from the others, due to progression of the worldwide pandemic of the Covid-19. This virus brought enormous damages to almost every sphere of the modern world, going from society to economy and of course to health. However, the large retail distribution is one of the very few businesses which, is hard to say, did benefit from the pandemic. This is due to the fact that with the introduction of lockdowns and the social restrictions, people got scared and when there is uncertainty human beings act in order to ensure survival. This reflected into an enormous boost for all primary goods stores and sellers, and of course supermarkets fall in this field. In the end, the analysis have been carried out by keeping in mind that comparisons with this particular year with the past would be of course of huge relevance from a descriptive analysis purpose, but would have been difficult to obtain precise insights on how and how much customers behaviours will change in the future. In the end were considered only data from 2018 and 2019

which are of course more comparable.

The figure 5.1 represents the final data flow obtained by the linkage of all the parallel jobs. This is how the ETL process actually works, from the training of the algorithm to the final fact table.

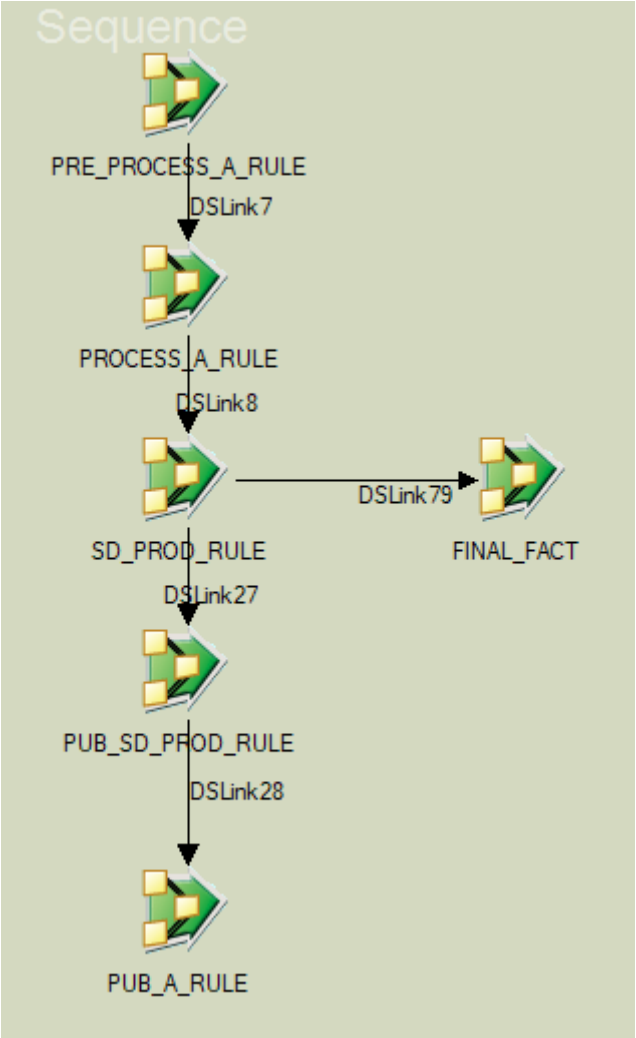

Figure 5.1: Final Data Flow on DataStage

This ETL flow does what it is expected from a common ETL process, it takes raw data and with the usage of enrichment modules and querying, yields fact tables organized in such a way that the further analysis will be feasible. Prior to individual

job analysis, will be now provided a description of the data available for the creation of the above data flow, obtained by the fact table *V FACT VENDITE FILTERED*. This fact was already obtained by previous work performed by MediamenteConsulting employees, it is a filtered fact because it "filters" all the rows and columns related to sales for one single store, entailing a *PDV_SK* equal to 587, as was mentioned before this is the largest store managed by corporate Alpha and therefore by modelling on it, it is obtained a good model of all the stores. This is not done only for simplification purposes, but also because managing data of all the stores would lead to an enormous increase of computational time activities, other than that even Power BI, the visualization tool implied in further steps, would not be able to handle this amount of data.

- *TESTATA* the identification number of every check
- *PDV* point of sale
- *DATE* the date of the transaction
- *PROD* the identification number of every product
- *PROMOTION* tells if the product was in promotion in a given date
- *FATTURATO FINALE* the revenue of the check
- *MARGINE FINALE* the margin of the check
- *CLIENTI CONSUMER* the identification of the consumer doing the shopping
- *ETA*' age of the consumer

5.2 Pre-Processing

In this initial phase, the data regarding sales obtained from Alpha must be prepared in order to match the compliance specified by the Apriori algorithm. In

this initial phase the time bucket is six months, i.e. the amount of time between SALES_START_DATE and SALES_END_DATE is a semester. This is of course a period of time too wide to make punctual analysis, but this issue will be fixed in further jobs.

Apriori expects data with a defined structure which is a table in which every row must entail:

- Check ID number
- Store
- Date in which the purchase was made
- List of item codes, separated by commas

```
01 | SELECT t0.TESTATA_SK ID, t0.date_sk, t0.PDV_SK, t0.LIST, s.
       SALES_START_DATE, s.SALES_END_DATE
02 | FROM
03 |    (SELECT TESTATA_SK, v.date_sk, PDV_SK,
04 |   string_agg(PROD_SK, ';') WITHIN GROUP ( ORDER BY PROD_SK )
       LIST
05 |   FROM V_FACT_VENDITE_FILTERED v
06 |    WHERE v.pdv_sk in (#P_PDV_SK#)
07 |   GROUP BY TESTATA_SK, v.date_sk, PDV_SK) t0
08 |    JOIN l2.DIM_SALE_PERIOD s
09 |     ON t0.DATE_SK between s.SALES_START_DATE and s.
       SALES_END_DATE
```

The expression # P_PDV_ SK # is a variable which has to be enhanced when running the job. However, due to the incredible amount of data considering all retail stores, for each process and analysis it will be used the *587* which is the biggest of all

stores owned by Alpha, hence the obtained results are considered of general validity. Below in figure 5.2 the result of the query is shown.

	ID	DATE_SK	PDV_SK	SALES_START_DATE	SALES_END_DATE	LIST
1	66909422	20181208	587	20180701	20181231	8332;8332;10217;10217;100404;10226;14701;14701;40...
2	66406312	20181125	587	20180701	20181231	94884;9487;14564;19165;440327;95386;128131;93164;...
3	60050549	20180618	587	20180101	20180630	128675;650623;101663;44501;89202;125070;400686;9...
4	68284577	20190113	587	20190101	20190630	73936;87085;39356;112186;82283;82326;122309;8647;...
5	67214348	20181216	587	20180701	20181231	109329;40350;13600;97361;97361;112723;112723;838...

Figure 5.2: Data Structure requested by Apriori algorithm

5.3 Processing

In this phase the Apriori algorithm is launched through a procedure which is stored at the database level. The results are then inserted in a new table which is L2_WRK_A_RULE. When launching the procedures some parameters need to be specified :

- MIN_SUPPORT
- MIN_CONFIDENCE
- MIN_PROD
- MAX_PROD

The first two have been previously described in section 4.4. Regarding instead the last two, it has to be said that these parameters are very important in fact they specify the maximum number and the minimum number of products in an association rule.

Of course these parameters take real value when launching the algorithm. This process was actually a "trial-and-error" step. This is due to the fact that there are no right values to be associated to confidence and support. So a lot of trials were performed in order to evaluate the result from two key point of view

- Amount of row produced
- Business meaning

Without entering too much into detail by presenting all the iterations performed, the final result entail the following values for the requested parameters.

- MIN_SUPPORT = 0.0001
- MIN_CONFIDENCE = 0.4
- MIN_PROD = 2
- MAX_PROD = 3

The Apriori algorithm is built with the key feature that it is able to produce association rules with only one "dragged" product. This translates into a fixed value for the MIN_PROD parameter. However the reasoning behind the decision of associating to MAX_PROD the value 3 was taken by evaluating to which point a rule is able to express a customer behavior. In this case the assumption is that by obtaining a rule with more that two "draggers" products would lead to a mismatch between the increased precision of the rule with the actual effect of the rule on sales and revenues. For the sake of simplicity one may say that using a parameter MAX_PROD higher than 3 would inevitably lead to an enormous amount of rules. In this case the selection of effectively valuable rules in the ocean of results obtained by Apriori would be almost impossible. There are other procedures able to yield rules with more than one product dragged, however the choice has been on Apriori due to its efficiency and also because association rules with more than three products involved may be of a difficult reading, especially from a final user point of view, diminishing the value of the whole process. The result of the algorithm are for a given rule, in a given period of time, support, confidence, lift and count. With count is represented the information related to the metric supportcount. These results are

then inserted in a table called *WRK_A_RULE*. Below an example of how a rule looks like after the completion of the algorithm.

$$\{13510,725321\} => \{13512\} \tag{5.1}$$

However, this is a single cell of the table and the information regarding the products involved in the association rule must be transformed and managed in order to obtain for each rule all the product involved and in which position they fit. Referring to the above rule what need to be obtained is

RULE_ID	PROD_SK	FLAG_POSITION
{13510,725321} => {13512}	13510	0
{13510,725321} => {13512}	725321	0
{13510,725321} => {13512}	13512	1

The column FLAG_POSITION is obtained due to further analysis, in particular when computing the cross index it is primary to have at disposal the information regarding where the product is placed inside the association rule, i.e. if the product drags or it is dragged.

Below the SQL query used in order to obtain the results aimed and explained in this section.

```
01 | SELECT U.RULE_ID,SALES_START_DATE,SALES_END_DATE,support,
       confidence,lift,count, INS_TIME, UPD_TIME,
02 | replace(replace(replace(PROD_SK,'{',''),'}',''),'>','')
       PROD_SK
03 | , u.FLAG_POSITION
04 | FROM (
```

```
05 | SELECT  RULE_ID,SALES_START_DATE,SALES_END_DATE,value PROD_SK
       , FLAG_POSITION, support, confidence,lift,count, INS_TIME,
        UPD_TIME
06 |       FROM (
07 | SELECT RULE_ID, SALES_START_DATE,SALES_END_DATE, value
       rule_string, support, confidence,lift,count, INS_TIME,
       UPD_TIME
08 | , ROW_NUMBER() OVER(PARTITION BY RULE_ID, SALES_START_DATE,
       SALES_END_DATE ORDER BY RULE_ID) -1 AS FLAG_POSITION
09 |   FROM
10 | L2.WRK_A_RULE
11 | CROSS APPLY STRING_SPLIT ( rule_id , '=' )
12 | ) T
13 | CROSS APPLY STRING_SPLIT ( rule_string , ',' ) ) U
14 | ORDER BY SALES_START_DATE, SALES_END_DATE;
```

5.3.1 Rules impact on KPIs

At this point the obtained table shows the information of a given rule, in a given time window and the metrics of support, confidence, lift and count. However additional steps need to be performed in order to obtain the final fact table that will be the input of front end processes that will be analyzed in the following chapter. The main goal of this phase is to get columns related to the fraction of revenue or margin directly related to the rule for a given product. This information is not available from the data obtained by OLTP systems of Alpha, therefore it has to be computed with the help of SQL queries. To do so every ticket in the TESTATA_SK column must be analyzed individually to discover all the association rules in it. This process is mandatory in order to compute the cross index, it has to be found the real impact of the association rule on the KPI. This is probably the most enriching step of all

the data flow, providing data structured in a way that to the front end analysis are given all the instruments to operate and support business decisions. The aim of this process is to obtain these columns:

- *FATT*
- *FATT_PROD*
- *MARGIN*
- *MARGIN_PROD*

The KPIs are two, but the way in which they are computed is exactly the same. For the absolute columns *FATT* and *MARGIN* the result is obtained by the sum of the revenues/margin of a given product, on a given week, on a given store. It is a simple cash flow computation which does not consider at all the association rules involved. These columns are however useful, allowing the final user to easily query the datawarehouse and obtain valuable pieces of information regarding what has been sold and the relative revenues and margins.

The real enrichment contribution however is related to the aggregated columns *FATT_PROD* and *MARGIN_PROD*.

These concepts have been computed with the purpose of understanding the weight of the rule, it is the result of the previous investigations. They both express the KPI in function of the association rules. For the sake of clarity will now be provided a made up example. Let's use again the rule {13510,725321} => {13512} this time paired with another rule {14765,39584} => {11983}.

At this point the information related to the date in which the shopping has been performed must be increased in granularity, from a six months time window to a weekly time window. To do so a join operation has been performed on the *L2.DIM_CALENDAR* dimension.

RULE_ID	PDV_SK	ISO_WEEK
{13510,725321} => {13512}	587	2018-01 Sett
{13510,725321} => {13512}	587	2018-01 Sett
{13510,725321} => {13512}	587	2018-01 Sett

PROD_SK	FATT	FATT_PROD	MARGIN	MARGIN_PROD
13510	34,1937	4,5923	9,983	2,4031
725321	17,8392	3,9846	4,8264	0,9647
13512	9,36492	1,9982	1,8736	0,8832

This table may be commented by saying that all three products involved produced revenues and therefore margins in the first week of 2018. However, a fraction of this absolute measures is related to the revenue and margin produced inside the rule.

Cross Index

As was explained in previous chapters the main target of the Advanced Analytics processes is to obtain hidden piece of information, relationship, and meaning from huge amount of data. This data if considered alone are of no support, they are just a pure description of reality of how the business is performing. The idea behind the Cross Index is to give a quantitative representation of association rules. For example one may wonder how association rules are actually implied in a decision making process. The answer is, without an aggregated measure or insight these association rules are of a difficult use. Therefore, the need of defining a new measure arose. At first the KPI must be decided, one has to evaluate carefully which is the indicator

that better describe the trend of sales, of revenues, of inventory. The first idea was to utilize the revenues. However after a careful evaluation it turned out that with all probability the better KPI was the margin. This is due to the fact that the revenue is better when facing cash flows analysis but does not take into consideration the cost of the items. Therefore, from a managerial perspective and with the constant aim of delivering an index that would have been easy to understand but at the same time efficient in modelling the reality, the choice has been on the margin. Another feature that was good to have on the index, was the visual impact on the final user, meaning that at a first glance even a non expert user may understand how the Cross Index is performing. To this end, the index has been normalized between 0 and 1. However before the actual computation of the index, another key step has to be performed. The calculation of the fraction of the margin related to the rule only. Made 100 the whole margin of a given product on a given week what has been taken out is only the margin of the product related to the rule. Since the goal is to attribute a quantitative value to association rules, the whole margin of the product has been considered as not accurate enough to model the purchasing process of customers. One may see it from another point of view by considering that in the first week of 2018 there was, at least one (probably more) check featuring all three products together. This is the proposed solution to the problem of having a constructed measure that will help the final user approach with the concept of association rule.

Let's compute now the *CROSS_INDEX* for the considered rule.

$$CI = \frac{0{,}8832}{2{,}4031 + 0{,}9647 + 0{,}8832} \tag{5.2}$$

$$CI = 0{,}2077 \tag{5.3}$$

The above result is normalized between 0 and 1. This is a good way to make an aggregated index more understandable from a customer point of view. As it has already been explained the dragged product is the one who mostly influences the

cross index. The closer the cross index to 1 the better it is. The concept of driver and dragged products are hence evaluated in a proper constructed index, this will be of a very relevant value when the front end operations will be performed in the next chapter.

RULE_ID	ISO_WEEK	PROD_SK	CROSS_INDEX
{13510,725321} => {13512}	2018-01 Sett	13510	0,2077
{13510,725321} => {13512}	2018-01 Sett	725321	0,2077
{13510,725321} => {13512}	2018-01 Sett	13512	0,2077

Let's now take into consideration another made up association rule.

$$\{14765,39584\} => \{1983\} \tag{5.4}$$

RULE_ID	PDV_SK	ISO_WEEK
{14765,39584} =>{1983}	587	2018-01 Sett
{14765,39584} =>{1983}	587	2018-01 Sett
{14765,39584} =>{1983}	587	2018-01 Sett

PROD_SK	FATT	FATT_PROD	MARGIN	MARGIN_PROD
14765	12,4567	0	3,6849	0
39584	17,8392	3,9846	4,8264	0,9647
1983	9,36492	1,9982	1,8736	0,8832

This example is proposed with the aim of showing how the *CROSS_INDEX* actually works. As it is shown by the above tables the fraction of revenue generated by the rule, in the first week of 2018 the product 14765 has been bought, as the columns *FATT* and *MARGIN* show. The data are managed in such a way that if one or more product related to the association rule entail a null fraction of revenues or margin in a given week the cross index is *NULL*.

$$CI(\{14765,39584\} => \{1983\}) = NULL \tag{5.5}$$

What comes out of this result is very interesting from a managerial perspective, in fact even if the three products in the association rule have been bought individually, they have not been bought together in any check inside the first week of the 2018. In the next chapter will be analyzed in detail how this information may be considered valuable from the client firm perspective.

RULE_ID	ISO_WEEK	PROD_SK	CROSS_INDEX
{14765,39584} =>{1983}	2018-01 Sett	14765	NULL
{14765,39584} =>{1983}	2018-01 Sett	39584	NULL
{14765,39584} =>{1983}	2018-01 Sett	1983	NULL

This useful fact table *L2.FACT_RULE_NEW*, which is the main table with which the visualization activities refer to, has been obtained by the execution of the following SQL statement.

```
01 | SELECT *,SUM(CASE WHEN FLAG_POSITION = 1
02 |              THEN MARGIN_PROD
03 |              ELSE 0
04 |              END)
05 |               OVER (PARTITION BY RULE_ID, PDV_SK, ISO_WEEK)
```

```
06 |            / CASE WHEN SUM(MARGIN_PROD)
07 |                        OVER (PARTITION BY RULE_ID, PDV_SK,
       ISO_WEEK) = 0 THEN NULL
08 |              ELSE SUM(MARGIN_PROD)
09 |                   OVER (PARTITION BY RULE_ID, PDV_SK, ISO_WEEK
       ) END
10 |                   CROSS_INDEX
11 |                   from(
12 | SELECT RULE_ID
13 | , PDV_SK
14 | , PROD_SK
15 | , ISO_WEEK
16 | , FLAG_POSITION
17 | , PROMO_FLAG
18 | , cnt_dist_rule_id
19 | , SUM(FATT)  FATT
20 | , SUM(MARG) MARGIN
21 | , SUM(fatt_prod) FATT_PROD
22 | , SUM(MARG_PROD) MARGIN_PROD
23 |   from (
24 |    select TESTATA_SK, RULE_ID, v.pdv_sk, v.prod_sk, c.
       ISO_WEEK
25 |    , (DENSE_RANK() over (partition by TESTATA_SK, pdv_sk, V.
       prod_sk, c.ISO_WEEK order by rule_id) + DENSE_RANK() over
       (partition by TESTATA_SK, pdv_sk, V.prod_sk, c.ISO_WEEK
       order by rule_id desc) - 1) cnt_dist_rule_id
26 |    , case when (DENSE_RANK() over (partition by TESTATA_SK,
       RULE_ID, pdv_sk, c.ISO_WEEK order by p.prod_sk) +
       DENSE_RANK() over (partition by TESTATA_SK, RULE_ID,
       pdv_sk, c.ISO_WEEK order by p.prod_sk desc) - 1)
27 |      = cnt_prod then v.FATTURATO_FINALE else 0 end fatt_prod
```

```
28 |     , case when (DENSE_RANK() over (partition by TESTATA_SK,
       RULE_ID, pdv_sk, c.ISO_WEEK order by p.prod_sk) +
       DENSE_RANK() over (partition by TESTATA_SK, RULE_ID,
       pdv_sk, c.ISO_WEEK order by p.prod_sk desc) - 1)
29 |       = cnt_prod then v.MARGINE_FINALE else 0 end MARG_PROD,
30 |       v.FATTURATO_FINALE FATT,
31 |       v.MARGINE_FINALE marg,
32 |       PROMO_FLAG,
33 |       FLAG_POSITION
34 |    FROM (SELECT TESTATA_SK, PDV_SK, PROD_SK, DATE_SK, CASE
       WHEN PROMOTION_SK < 12 THEN 0 ELSE 1 END PROMO_FLAG, SUM(
       FATTURATO_FINALE) FATTURATO_FINALE, SUM(MARGINE_FINALE)
       MARGINE_FINALE
35 |          FROM L2.FACT_VENDITE_FILTERED
36 |          GROUP BY TESTATA_SK, PDV_SK, PROD_SK, DATE_SK, CASE
       WHEN PROMOTION_SK < 12 THEN 0 ELSE 1 END) v
37 |    JOIN L2.DIM_CALENDAR c
38 |      ON v.DATE_SK = c.DATE_SK
39 |    LEFT JOIN (SELECT * FROM L2.PROD_RULES
40 |               WHERE SALES_START_DATE=20170101 AND
       SALES_END_DATE=20170630
41 |       ) P ON (V.PROD_SK = P.PROD_SK)
42 |    WHERE PDV_SK=(#P_PDV_SK#)
43 |      AND C.DATE_SK BETWEEN (#P_DATE_SK#) AND (#P_DATE1_SK#)
44 |          AND c.DATE_SK%2= #P_ITERATION#
45 |          ) f
46 |  WHERE RULE_ID IS NOT NULL
47 |    GROUP BY
48 |       RULE_ID, PDV_SK, PROD_SK, ISO_WEEK, cnt_dist_rule_id,
       PROMO_FLAG, FLAG_POSITION)
49 |  AS V
```

About this complex query there are few main points which need to be explained in further detail.

- *WHEN PROMOTION_SK < 12 THEN 0 ELSE 1 END PROMO_FLAG* this is a constraint for the generation of the column *PROMO_FLAG*. With respect to the data structure, a product is sold with promotion operations if the field of the column *PROMOTION_SK* entail values higher than 11.
- *C.DATE_SK %2= #P_ITERATION#* this is a way to reduce the processing time of the SQL Server tool, in fact when running this statement which entail more than one join operation, the result was that there was not enough space to bring the query to a conclusion. With this datastage ploy, it is possible to actually split the one huge query into two distinct iterations. In this particular case, the first iteration performs the join only on even dates, whereas the second iteration performs the join only on odds dates. The results are then written inside the final fact table.
- *Terminator_Activity* This datastage feature is very helpful in order to check how the work is performing, and then by selecting one of the available options one may obtain error messages. This is very important in scenarios like this one where the query takes a lot of iterations and therefore time. By using the terminator it is possible to better understand the current status of the run and eventually find out which was the problem.
- The first section of the query uses the obtained columns by the inner tables, to compute the *CROSS_INDEX*.This model is actually also flexible, this is due to the fact that even if the chosen KPI for *CROSS_INDEX* is the margin, the obtained columns consider also the revenues (*fatturato*), which is probably one of the overall mostly understood and considered indicator by final users.

In the figure 5.3 below is shown the architecture of the job entitled of the processing

of the association rules, it shows also the tools used in order to obtain the above described loop.

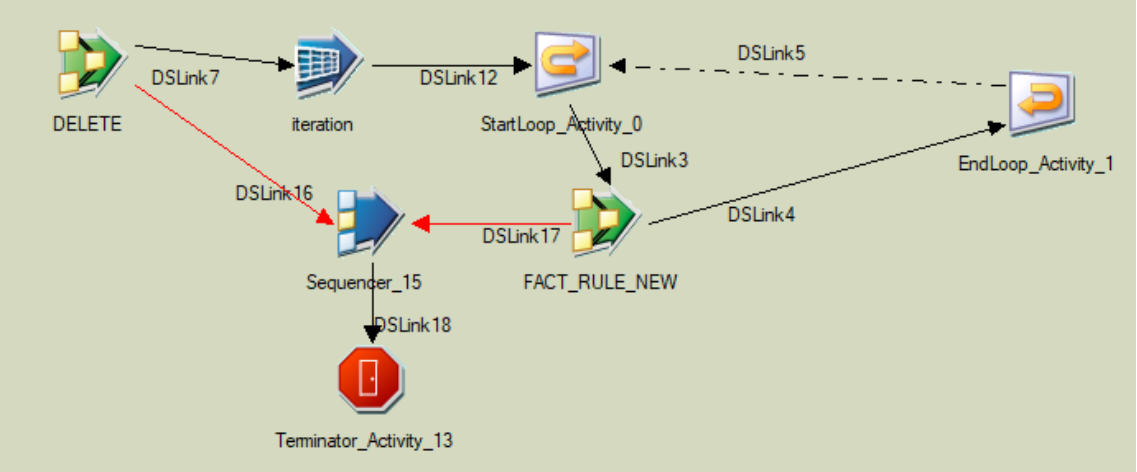

Figure 5.3: Representation of the loop-job

In the following figure 6.2, it is represented the join which causes the computational overload. As it is clearly depicted in the figure, the join is performed also with another fact table which is *L2.FACT_GIACENZE*. By doing so it is obtained also the value of the inventory of a given product, for a given store, in a given week. This particular piece of information will not be used in the following chapter, it has however been calculated because it could have further relevance for different kinds of analysis.

In the last two jobs of the data flow it is performed the publication on the datawarehouse of all the piece of information regarding the technical measures implied by the Apriori algorithm. As was explained before the procedure yield columns which are not very well organized, at least from a business analyst perspective. These two steps are performed with the aim of designing a fact table which will then show

- Support
- Confidence

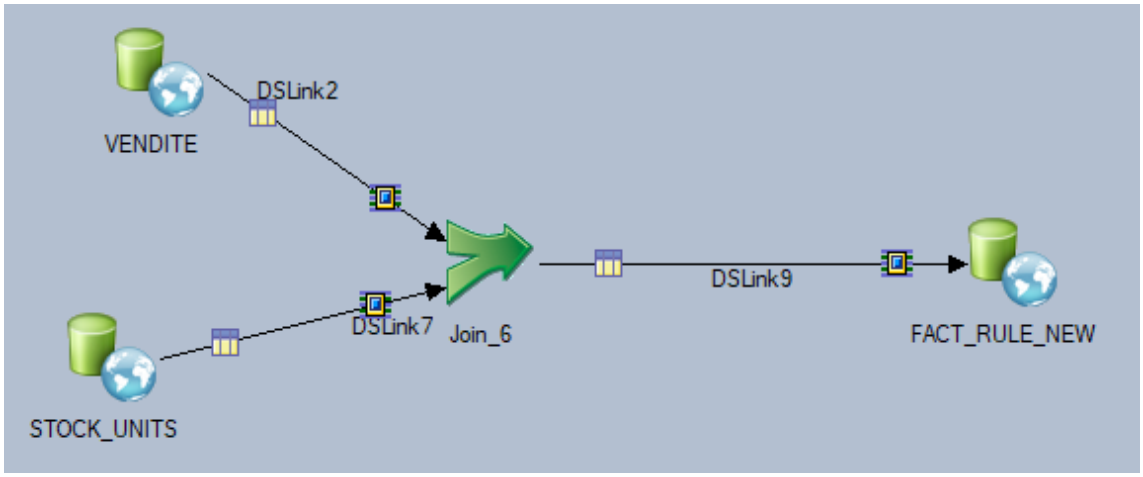

Figure 5.4: Representation of the final job

- Lift

for each rule obtained with the use of the algorithm. It has to be said that these a pure clarification step, with respect to the fact table *L2.FACT_RULE_NEW* this will see much less implications in the visualization activities. In the following figure is represented the job entitled of performing the above mentioned activities.

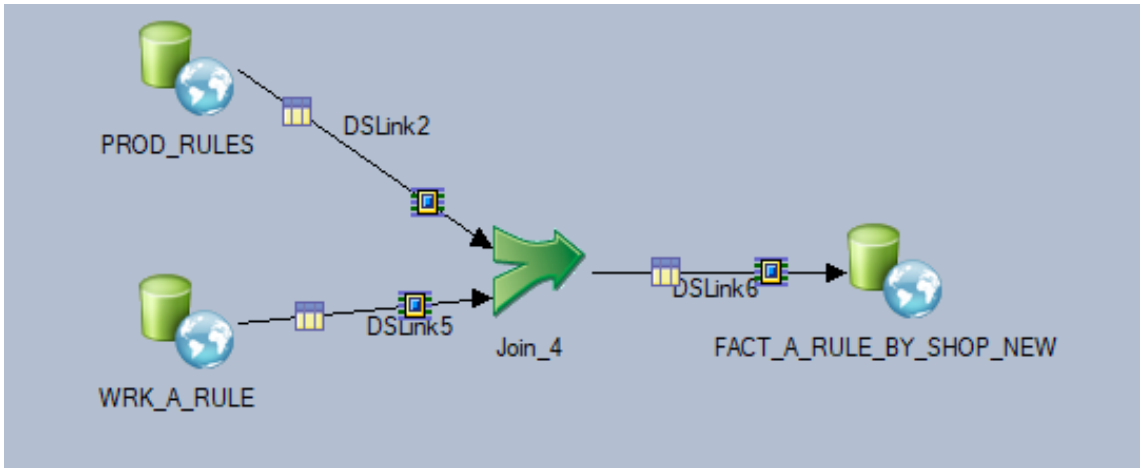

Figure 5.5: Representation of the publication job

Chapter 6

Data Visualization

This is the last phase of the whole data modelling processes. Once the tables are finally obtained with all the data needed for decision making support the data visualization activities take place. This chapter will be organized in two sections. The former will describe in detail the model design process, and the latter will entail the actual results obtained on Power BI, describing the roles for which these views are designed.

The value of advanced analytics resides in the capability of implementing and produce predictive and prescriptive analysis, which, as was carefully depicted in Chapter 2, is the core of data mining, obtaining precious insights which would be absolutely invisible otherwise. The process of visualization design is hence fundamental for every IT consulting company, such as MediamenteConsulting.

This is the interface between the consultant and the client company, in fact this is the result of integration processes which the client does not see, and which is also not interested on being informed about them. Of course, they are willing to obtain practical insights to support their decision making processes.

In the following section will be now provided a careful overview of how data are modelled inside a Power BI environment. This schema, which is designed to support

the scope of the analysis, is central to obtain a good quality result.

6.1 Data Model Design

As was previously announced, this section regards the description of how and with which instrument the logical model between tables is constructed on Power BI. In this visualization tool there are three different areas, entitled of different operations, keep in mind that this overview regards only the Power BI Desktop tool, which is the one used to design dashboards.

- Report
- Data
- Model

The first feature is the core of the desktop tool, in fact is here where all the data are managed and modelled in order to obtain valuable dashboards which could be then uploaded on the online business intelligence tool, or shared with colleagues. The second, refers to a data visualization area. Here it is possible to modify the columns and table imported through data connection operations. These kind of activities may be performed also with the use of the Power Query Editor window which is located in the home page of the tool. Without entering too much in detail, a clarification has to be made over the difference among two concepts which are

- Measure
- Column

A measure is a computed column which is not stored at table level, which means that it is considered as a column but will not be shown inside the Data section of Power BI Desktop. A column instead is constructed to actually modify the structure

of the table itself. One may wonder when the former should be used over the latter. Well, of course there is no iron rule which clearly solve this doubt, but in general if the facts and dimensions are well designed at database level then it will be most likely sufficient to use a measure or even a "rapid" measure which is a function offered by the tool allowing the designer to pick one of the suggested computational operations, such as showing percentage, deltas related to given variables and so on. The last section is the one which will be tackled the most in this overview, is in fact in the model visualization of Power BI desktop where the designer is able to design the connections among tables and therefore their relationships. Before showing the actual result of these processes it is considered to be important a rapid overview over the concept of fact, dimension, schema and keys.

Facts and Dimensions

With the term fact, it is defined a table composed by rows and columns which is constructed to collect all the pieces of information of a given fact. In this particular project the fact table is L2.FACT_RULE_NEW, which has the aim of showing all the useful data regarding the fact which is the association rules. In general a fact table is defined as the measurement of a business process. A good way to describe dimension tables instead is by seeing them as companions of the principal fact table. In dimensions are stored all the pieces of information useful to constrain and query the fact table. Let's make and example with the data obtained from the integration data flow. As it was widely described, the final fact of the integration step entail a column which is called PROD_SK. In order to show also the description of the product, its category and department it is mandatory to create a relationship with the dimension table L2.DIM_PROD which contains all the features of a given product.

Keys

A key is a cell or a group of cells contained inside a given table which uniquely defines a row of the table. Keys are the way with which the relationships among table are actually performed. By referring to the above mentioned example of the column PROD_SK this is a *Foreign Key* for the fact table whereas is a *Primary Key* for the dimension table L2.DIM_PRODOTTO. What this means is that on the fact table they may be found different columns showing the same PROD_SK, however this is impossible in the dimension. This is exactly what primary key means, by sorting data by key it will be rejected one and only one row of the dimension table. In the end, it has to mentioned the presence of hierarchies inside dimensions. This allows the business analyst to show data with different granularity. The most commonly cited example is the calendar, a dimension regarding time is almost always articulated in columns with a hierarchy such as

- Year
- Month
- Week
- Day

Schema

The concept of schema is very important here in the visualization steps, as it is for datawarehouse design purposes. On Power BI, as also on other business intelligence tools, the most commonly used schema types are star schema, snowflake schema and the galaxy schema. For this project has been adopted a star schema which features a central fact table surrounded by dimension tables, in a snowflake schema instead the dimension table related to the central fact are then surrounded themselves by

other dimensions. In general there are some benefits coming from the adoption of a star schema over a snowflake schema, they may be resumed in

- Smaller disk space usage
- Easier to manage when facing new dimensions to be added in the schema
- The query elaboration time is reduced due to the lower number of dimensions involved

As always in this matter, there is no best solution, there are different alternatives. In this case having performed also all the activities related to the integration steps the choice of the schema is a direct consequence of the datawarehouse structure. The so called galaxy schema is a even more complex solution, it is the only featuring two different fact tables connected to dimensions. Of course there is no model featuring connection among fact table, this would be impossible from a technical point of view other than profoundly incorrect conceptually.

6.1.1 Kinds of relationship

After the completion of data connection and insertion on Power BI, the relationships among tables need to be managed. It has to be said, this process is very intuitive and most of the times the tool accomplish a good work on its own. This happens by simply dragging the foreign key of the fact table onto the same primary key on the dimension table. However, it may be useful to keep in mind which are the major kind of relationship connecting tables.

1. Many-to-one (N:1) This is the most common kind of relationship among a fact table at the center of the schema and its dimensions. To many instances of the fact table refer a single instance in the dimension.

2. One-to-One (1:1) In this case, the relationship entails a one-to-one connection among the instance of the fact table and the one of the dimension.

3. One-to-Many (1:N) This is the reflected image of a relationship many-to-one. It is not very common in a star or snowflake schema.

4. Many-to-Many (N:N) This type of connection among tables, is sometimes implied when there is no need to identify a single instance in the dimension, in fact in this case to few specific instances may be related different instances in the dimension.

6.2 Model

Let's see now what the model related to the data actually looks like. It has to be mentioned that the three dimensions directly connected to the fact table have not been designed and implemented in this project of thesis, but they are a part of the "as-is" advanced analytics project. In particular this dimension have been built by MediamenteConsulting as requested by Alpha, asking for a new datawarehouse in order to collect and ease the handling of data and records.In the following figure 6.1 it is offered a peek at the final model. As it is very clear from the chart above there is a single fact table and then three dimensions related to it. The relationship with the fact table are the following:

- L2.DIM_PRODOTTO is connected through a one-to-many relationship. The primary key is PROD_SK

- L2.DIM_PDV is connected through a one-to-many relationship. The primary key is PDV_SK

- L2.DIM_CALENDAR is connected through a many-to-many relationship. The primary key is ISO_WEEK

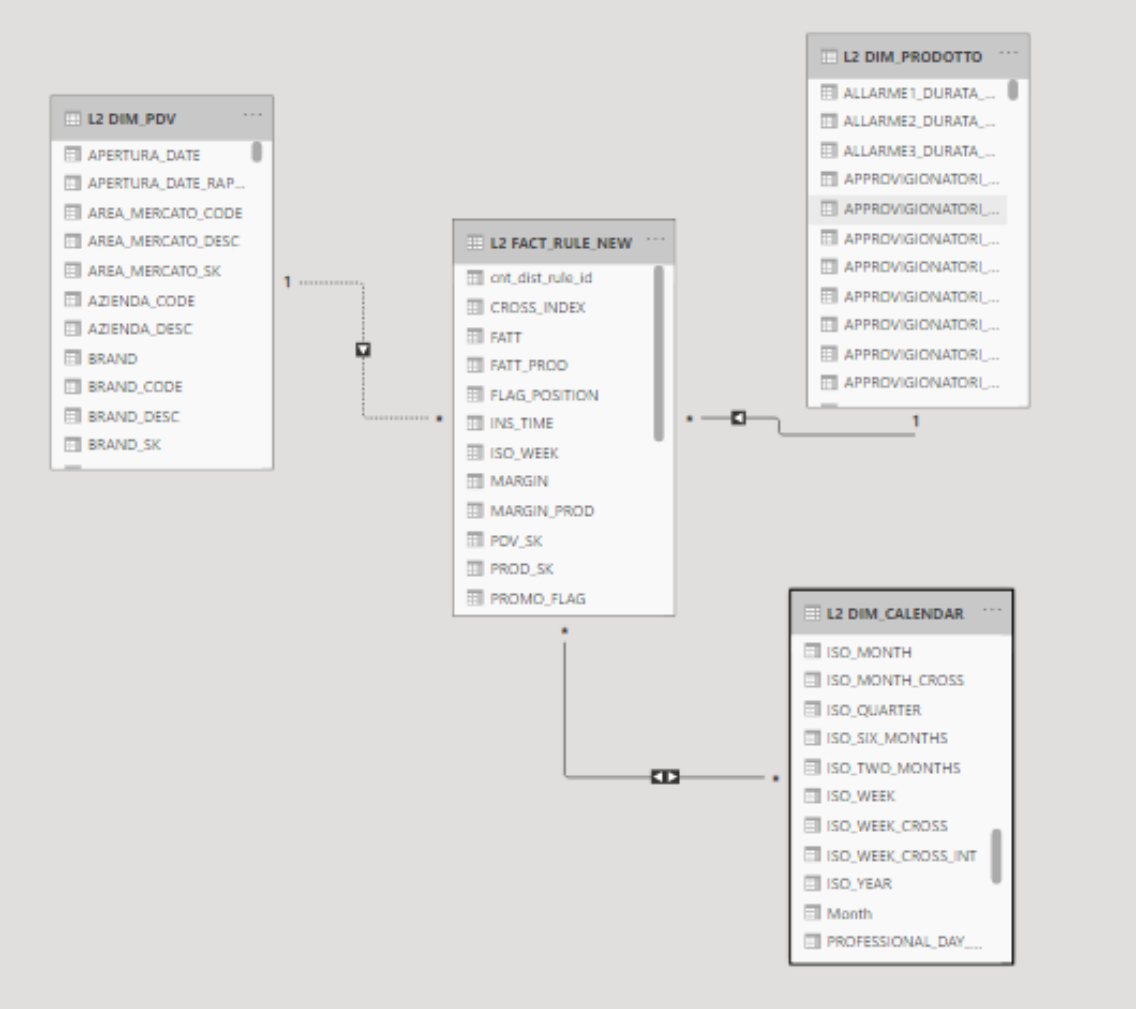

Figure 6.1: Data Model on Power BI

After the description of the model obtained on Power BI, it will be now provided the section related to the results of this process, which are the dashboards.

6.3 Dashboards

When designing dashboards or reports is mandatory to have a clear understanding of the different roles involved and what they are expecting to visualize. For this particular project have been targeted three different roles

- Top Management
- Sales Manager
- Marketing Manager

by keeping in mind that each one of them is entitled of performing different activities, with different levels of detail, the dashboards may or may not see the implications of the association rules obtained from the integration phase. Unfortunately, the representation of the obtained dashboards on paper does not allow the reader to fully comprehend the actual value of this visualizations, this is due to the high level of interactions and drill-down operations that may be performed by the user on Power BI.

6.3.1 Top Management

This is the role which is entitled of managing and leading the company as a whole. In this case one could easily figure that a CEO or a brand manager would probably not be interested in seeing the hidden relationships among products. This roles are mostly entitled of the overview of the current status of the firm, and then if they need lower detailed specifics they generally contact directors below them, being them closer to effective operations. They would be however, highly interested in seeing high level piece of information conveyed by two very well known KPIs which are

- Revenue
- Margin

The first dashboard designed for this level of the company, is for the brand manager. The idea is that this professional figure would be interested in comparing the performance of a given brand with the past year, both in terms of margin and revenues.

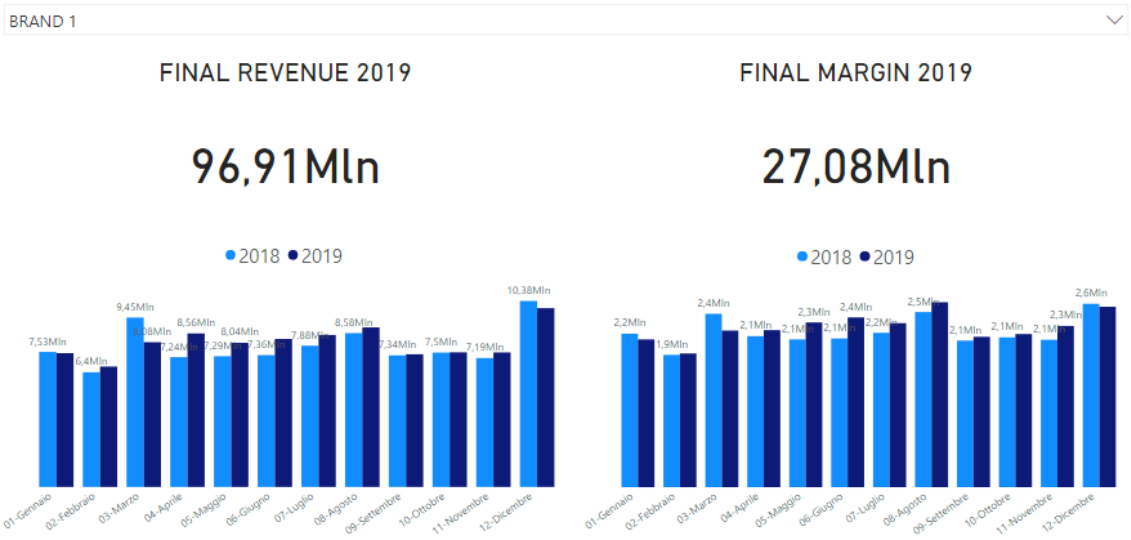

Figure 6.2: Brand Manager Dashboard

The second dashboard is designed for the top manager, one may think to the CEO or a high level director.

In this case the information conveyed is very similar to the first dashboard, the main difference is that here the user is able to obtain an all around picture of the company performance. In this case have been also used a pie chart which as was described in figure 2.5 it is very useful when used with maximum two data labels. By using the filtering menu the manager is able to go through all the four different brands owned by Alpha and therefore compare trends and values also with the help of the clustered column chart which has been built with a monthly time window. These dashboards have been built with the aim of giving a much wider perspective on the role of visualization activities, however the real goal was to obtain meaningful and intelligible dashboards entailing the use of association rules. With respect to the above mentioned figure, further dashboards are used to predictive and prescriptive analysis whereas the former are related to purely descriptive analysis, this should not be considered as a disadvantage. It is in fact in these specific views where lies

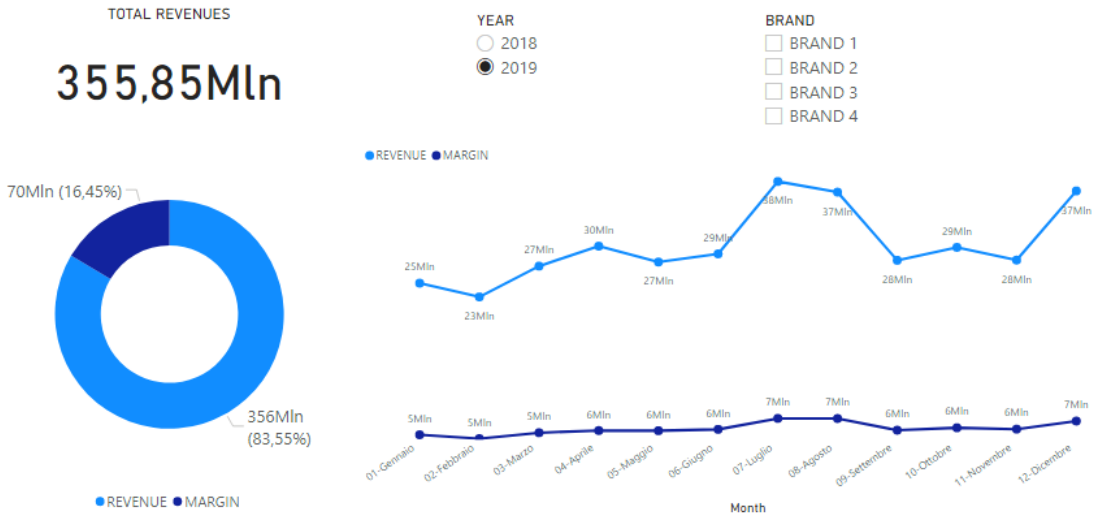

Figure 6.3: Top Manager Dashboard

the need of describing instead of predicting and make inferences. In the figure 6.3 there is also a trend representation which, as already explained, is very valuable when the need is to instantly show the current status of the business. In this case the evolution over the months over a given year is depicted, always using as axes the temporal window, with a monthly time bucket.

6.4 Sales Manager

As was widely described over the presentation of this work, the main goal was to obtain a dashboard which would be ease to relate with, and therefore valuable from the customer point of view. The person covering this role would be interested in understanding which are the hidden relationship among products and therefore discover insights on customers behaviour. First will be presented two similar dashboards, depicting the trends of sectors and departments. This is of course a very valuable descriptive analysis because by showing the delta between 2018 and 2019 the sales manager is able to quickly understand which are the sectors and departments that

are better performing and the ones whose performance have been worse. In the following figure 6.4 it represented a made up example with the segment *freschissimi*, the line chart is also there to help the user understand quickly the variations, month by month, and therefore decide to further investigate on these segments which show a decrease with respect to the previous year. The variation section has been constructed with the help of DAX formulas on Power BI, this indicator's color is defined by a rule. By doing so the final user is able to quickly understand if the variation with respect to the previous year have been positive or negative.

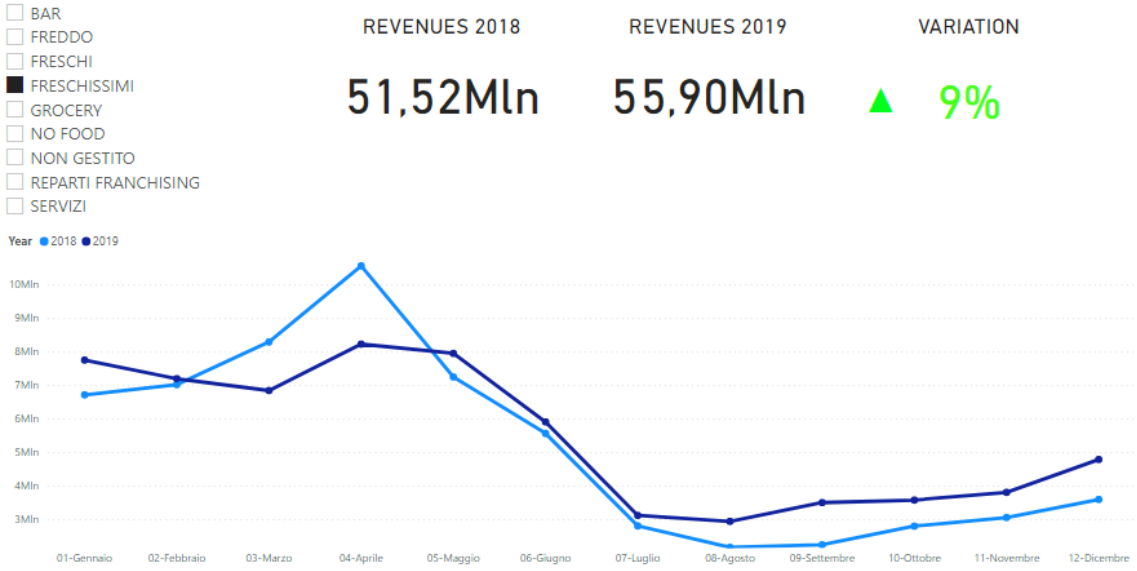

Figure 6.4: Segment revenues trend

The same reasoning is valid for the following figure 6.5. This is a higher level view of course with respect to the previous one and permits investigation over each of the segments of a given retail store. For example in this chart it is possible to see that the department *formaggi e derivati latte serviti* is having a deflection with the respect to the previous year.

Let's now dive in deep into the core of this chapter, the design of dashboard that will ease the understanding of the association rule concept by the final user. This is

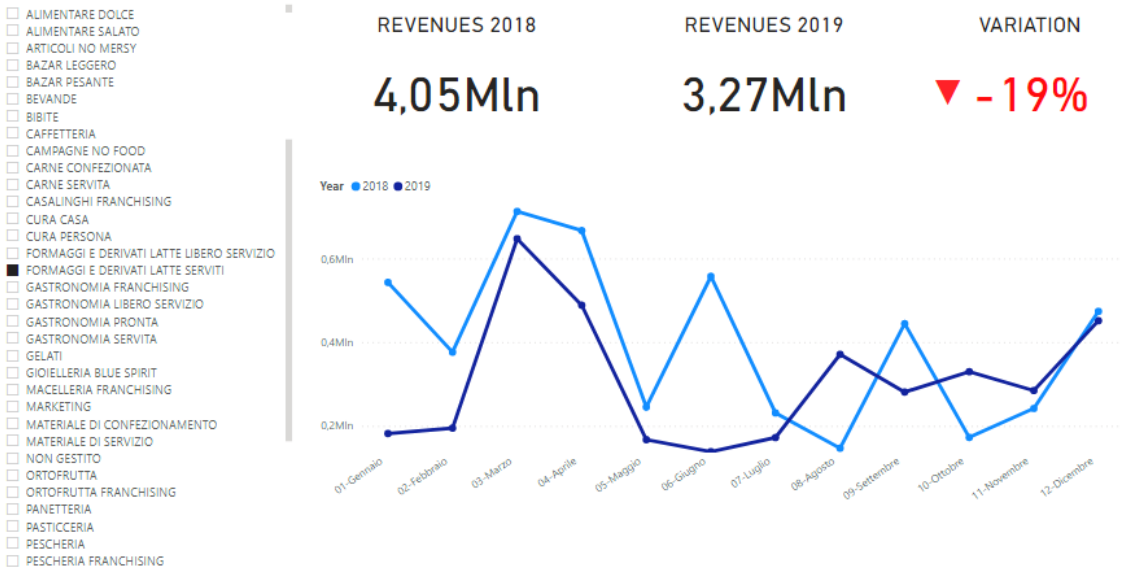

Figure 6.5: Department revenues trend

the step of the process which will see the implications of not only association rules, but also of the CROSS_INDEX. The whole purpose of the chapter 5 was to obtain tables that would have been able to satisfy this need, the need of making the very technical concept of association rule intelligible by the customer. This is, as was explained several times during this thesis, the main reason behind the construction of a new measure. Below there will be the description of two different dashboards, both of them are related to the use of association rules and both of them are designed with the specific aim of helping the sales manager, or the final user, in their decision making processes. The first dashboard represented in the figure 6.6 below

Possible Use

The idea behind the design of this dashboard is that the sales manager may be interested in checking a product and the associated rules related to this product. The dashboard is constructed with the intent of showing, with the help of pie charts and clustered column charts, the impact of the rule on margin and revenues. In

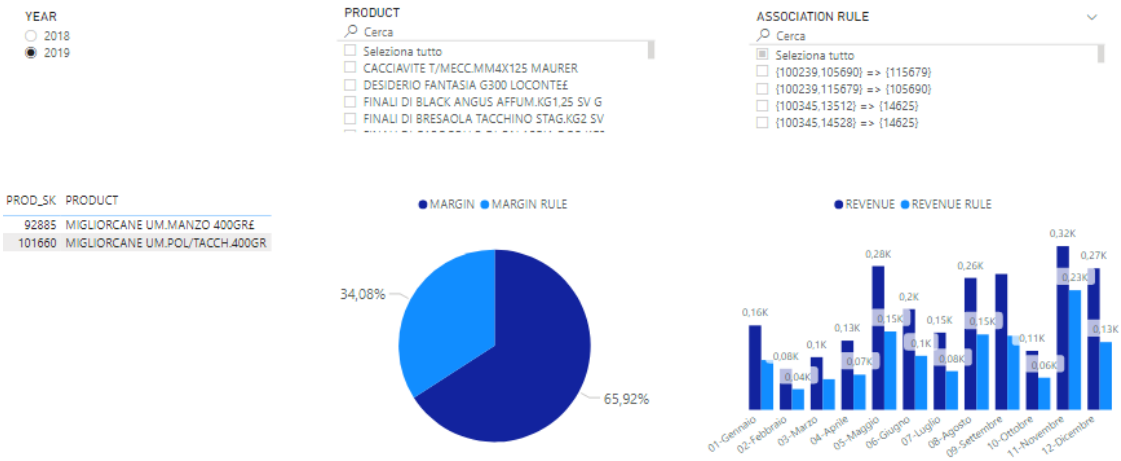

Figure 6.6: Association Rules Dashboard 1

this particular case have been depicted the association rule involving dog meat, as it is shown clearly in the pie chart the portion of the revenues are heavily related to the association rule which entail the purchase of the two different kind of dog meat, in particular chicken or beef. This is of course a very well known rule but it was shown in order to make it more easy to understand the reasoning behind this dashboard. By using a filter key on the year it is also possible to compare the rule over different years, in this case the behaviour of the costumer is very well established this leads to a very little difference on the impact of the rule on the revenue over the years. However, this is true for well know costumer behaviours it is not so obvious for hidden relationships among products. The strength of this dashboard is that it is fully interactive with the user, meaning that it is possible to search either for products or association rules.

Let's now consider the second dashboard designed for the sales manager. In this case the idea is to guide the user with the help of the CROSS_INDEX. As was explained in chapter 5 this measure is computed to evaluate the intrinsic value of a given association rule. By doing so the sales manager is able to discover those

rules which are related to the highest dragging effect. In the following figure 6.7 it is shown a peek of this dashboard.

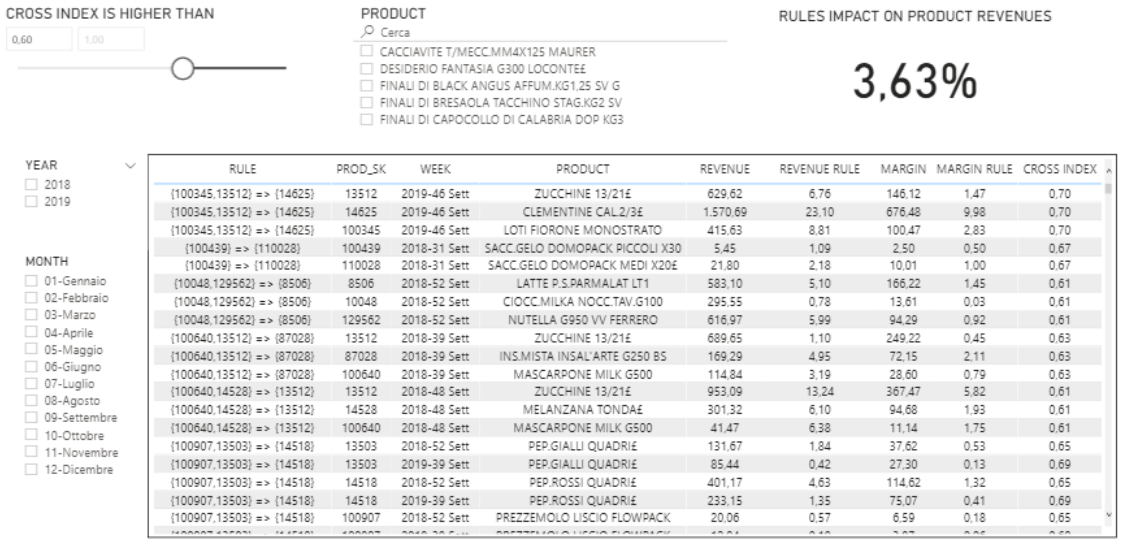

Figure 6.7: Association Rules Exploration Dashboard

Possible Use

The driver idea leading to the design of this dashboard, was to propose a instrument panel related to association rules. In this dashboard in fact, it is possible to operate few filtering actions allowing the sales manager to dive in deep into the association rules. As was explained before the CROSS_INDEX is related to the intrinsic value of the rule, it is been considered however that offering another card visualization related to the impact of the rules on the revenues would add value to the dashboard. This is the case of the card on the top right of the dashboard. This percentage is computed by dividing the revenues coming from the rule by the total revenues of a given week. The result shown in this picture is related to the average impact of the rules considering both 2018 and 2019. This is particularly valuable from a managerial perspective, in fact it is possible to join the information coming from two

different KPIs

- Margin
- Revenues

To obtain a clearer view of how a given association rule is performing, how much it impacts on revenues and margins of the product involved. This allows the sales manager to use this dashboard with two different drivers: the product and the rule. Let's imagine that the manager is looking for implementing sales of a given product, then what they will have to do is simply filter the dashboard by typing the product description inside the filter. On the other way around, the manager may have as objective the increase of the store margin, this is feasible on this dashboard by filtering years and months and of course the CROSS_INDEX starting from values close to 1 and then progressively lowering it. This is probably the dashboard which better represents the solution to the principle aim of this project of thesis, it is designed with the complicated target of letting the user understand the concept of association rules without entering too much in detail of technical measures which are used to obtain them such as confidence and support.

6.5 Marketing Manager

For the marketing manager have been designed two different dashboard as well. The idea here is to obtain pieces of information regarding association rules and promotions, this is why in the integration phase the final fact table have been built by taking into consideration the PROMO_FLAG, by doing so it is now possible to compare rules that entail the presence of product that are sold with promotions campaign and the ones sold without promotions.

The first dashboard is designed with the aim of proposing a clear view of how single products are performing both in terms of margin and revenues. In the following

figure 6.8 it is represented the above mentioned dashboard.

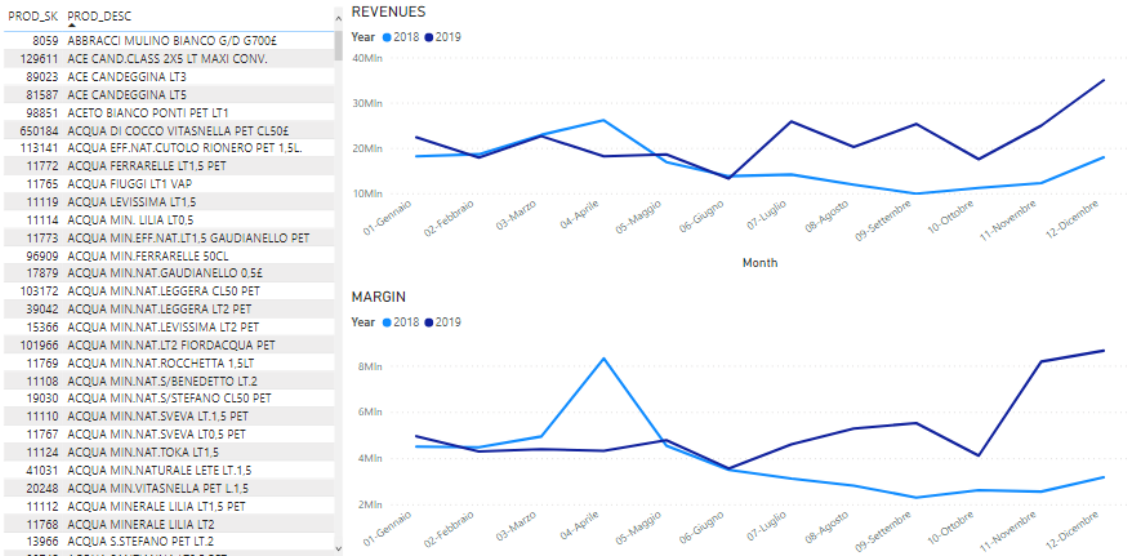

Figure 6.8: Product trends dashboard

This is an instrument panel designed with the aim of helping the marketing manager in a better understanding of how single products are performing and therefore decide, due to several managerial reasons, whether to offer promotional campaign for a given item or not. It has to be said that large scale retail distributors promotion campaigns are influenced by lots of different considerations, not always related to revenue trend. However this is a good way to support decision making processes in these scenarios. In the following figure 6.9 will be now depicted the second dashboard designed for the marketing manager.

Possible Use

This dashboard is designed with the clear intent of helping the marketing manager understanding of the impact of promotions on association rules and therefore on sales. As the ones previously described also this dashboard features different filter keys, in particular referred to time selection. The impact of promotional campaigns over

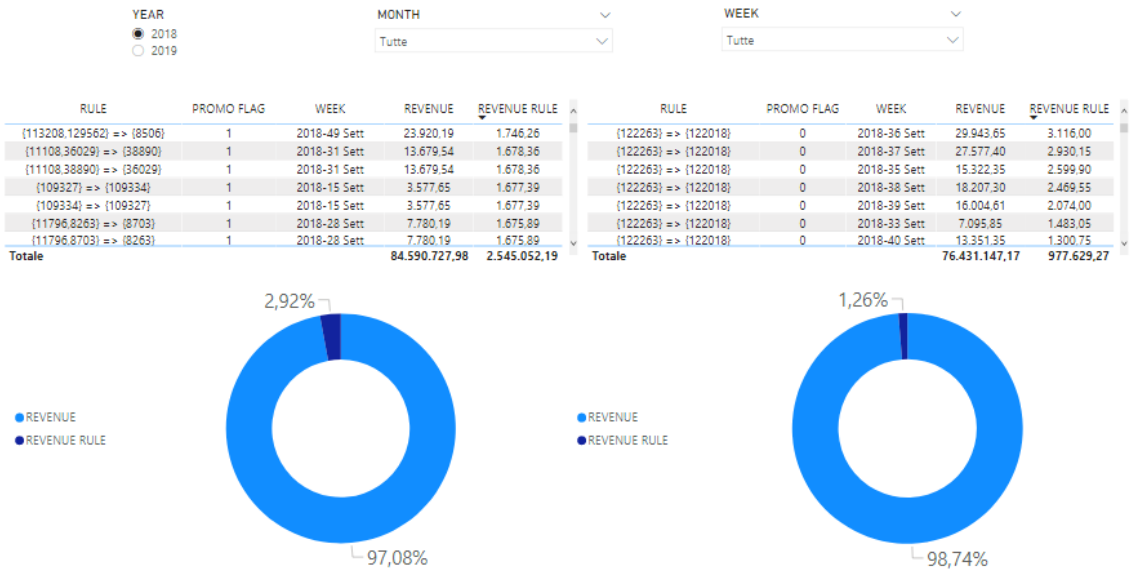

Figure 6.9: Promotions impact on association rule dashboard

association rules is strictly positive. One may imagine that the marketing manager would like to analyze for a given month and given week all the association rules which entail one or more product which is on promotion.

By doing so it will be possible to understand the actual impact of promotions on sales, and therefore this satisfy the initial objective of obtaining support to decision making processes. In particular by using the drill-through functions it is possible to discover which products are discounted and therefore obtain further insights for future promotional campaigns.

In the end the promotional matter is very delicate, as was explained before products are sold with reduction in price or in convenient bundles and by studying the large scale retail distribution industry what turned out was that most of the times promotional campaigns are placed with random patterns, because if customers were aware of when and how the promos will be placed the effect and purpose of these sale initiatives would be highly reduced.

Chapter 7

Conclusions

In this final chapter will be evaluated the results obtained compared to the initial objectives that caused the beginning of this project of thesis. At first it has to be said that the instructions and valuable help received from MediamenteConsulting employees was key to the success of this project, in fact as was carefully explained throughout the whole thesis, the project have been carried out in a consulting firm. Of course the knowledge obtained from the attended course of study were fundamental to approach this matter. However, the theory learnt from college courses must be combined with the experience of professional workers in order to understand the actual work tasks.

Overall the proposed objectives have been reached by this project, in fact the result of the integration phase yield a satisfying procedure which is of high value either from a company perspective or a client perspective. The data flow obtained has been a very good instrument to learn and understand how a complex data analysis tool such as DataStage works in practice, this increases also the perceived value for me as internship student. With respect to more technical aspects of the obtained framework, the strengths are for sure related to the high level of parameters used and flexibility.

In fact by obtaining a flexible structure this highly increases the intrinsic value of the work, it will be indeed possible to re-use this advanced analytics project with future clients, without the need of approaching the project from the starting point. The other main problem that was depicted in chapter 3 has been also solved with a satisfying solution. In fact in the dashboards presented in the above chapter it is clear that the main driver was to obtain a visual solution that would strictly consider the client need of a better understanding related to technical concepts like the association rules, market basket analysis and Apriori algorithm. The creation of an aggregated measure such as the cross index is considered to be a very good solution to this costumer need, furthermore the choice of using Power BI to build the visualizations may be as an additional valuable feature for the client.

The obtained results are also very promising from a future perspective scenario, in fact the data and tables coming from integration steps are constructed in such a way that a lot of different business analysis may be performed. For example reasoning on units and warehouse, clustering and classification of clients, forecasts related to future sales and evolution of the cross index. These are all very valuable analysis that will be, with high probability, carried out in a near future according to client needs and demands.

One may also consider the above depicted potential further analysis and improvements as a strength factor of the work performed in this project of thesis. In fact the obtained results may be seen as a component of a much bigger project placed inside the consulting company, of course being a thesis work entails having clear objectives and satisfying solutions. However, the work performed is also carried out with the precise aim from the consulting company of training the internship students and give them the necessary tools to approach the labour market.

This last chapter is written with a more personal cut as one may have noticed. It is a way to express not only the actual results obtained from a work performed

perspective, but also the overall impressions on the environment and on the internship path carried out inside a consulting company. Regarding this matter i would like to express my complete satisfaction as young student approaching the labour market, MediamenteConsulting personnel really eased this introduction process and maintained the given word on the scheduled planning regardless the extremely uncertainty provoked from the spread of the Covid-19 virus pandemic.

Bibliography

[1] Edgar H Sibley and Robert W Taylor. «A data definition and mapping language». In: *Communications of the ACM* 16.12 (1973), pp. 750–759 (cit. on p. 1).

[2] Dekang Lin et al. «An information-theoretic definition of similarity.» In: *Icml*. Vol. 98. 1998. 1998, pp. 296–304 (cit. on p. 3).

[3] Robert Wrembel. *Data Warehouses and OLAP: Concepts, Architectures and Solutions: Concepts, Architectures and Solutions*. Igi Global, 2006 (cit. on p. 13).

[4] Ralph Kimball and Joe Caserta. *The data warehouse ETL toolkit*. John Wiley & Sons, 2004 (cit. on p. 15).

[5] Alberto Cairo. *The truthful art: Data, charts, and maps for communication*. New Riders, 2016 (cit. on p. 20).