POLITECNICO DI TORINO

Corso di Laurea in Computer Engineering

Master Degree Thesis

E-Commerce monitoring solution for product allocation and marketing planning forecasting

Supervisors
Prof.ssa Tania Cerquitelli
Dott. Vincenzo Scinicariello

Candidate
Antonio Greco

April 2018

Contents

Introduction

E-Commerce Era

Online business constitutes a threat to conventional trade in recent years, but at the same time opens up new opportunities. The constant digitalization development affects the customer expectation and purchasing behavior of an entire generation. The current internet price transparency moves the power from the world of commerce towards the buying public.
Nowadays, the customer takes on the central role in the business. He can shop whenever and wherever he wants. In addition online trade also overrides spatial and temporal restrictions of conventional trade.
All these factors contribute to the remarkable outbreak of online treading that, consequently, is improving considerably its potential.

Supply chain In order to win the other online companies competition, it's necessary to realize a digital supply chain. Supply chain is a system of organizations, people, activities, information, and resources involved in moving a product or service from supplier to customer. Supply chain activities involve the transformation of natural resources, raw materials, and components into a finished product that is delivered to the end customer. It allows to unite the benefits of traditional trading, such as customer proximity and service orientation, with the benefits of online trading that consist in low cost, great market size and a important products availability. When a customer places an order on the web, the order triggers a series of transactions throughout the supply chain. The transaction execution speed represents perhaps the most fundamental form of interaction among supply chain partners. A central information hub, that represents the point of contact of the several supply chain partners, can offer a natural platform to capture the order, coordinate the activities, track the order status and deliver after-sales service. All these applications are called E-Commerce.

Case study

This thesis consists of two main purposes. The first purpose has an academic nature, whereas the other one represents a business goal.

With regards to academic aim, the work is based on the application of web analytics techniques, data mining algorithms and the realization of infrastructure in order to store data coming from both external and internal sources. In particular, with regards to web analytics, the desired objective is represented by the realization of a web crawling infrastructure to capture web sites unstructured data and, successively, storing them in a robust and flexible infrastructure. Thanks to a laborious ELT (extract, load, transform) process, these data will become structured in order to integrate all different data types. In this way, it will be possible to apply them advanced analytics techniques. The data storage infrastructure must be able to contain both structured and unstructured data coming from different sources and to provide scalability adapting to a large amount of data.
Once data will be structured and homogenized, the next step consists in applying them data mining techniques. In this case study, handling especially textual data, text mining will be applied to clean data and extract business information from them. An other data mining technique exploited will be the document clustering that represents the application of cluster analysis to textual documents extracting descriptors that identify a single cluster content.

On the other hand, this work is related to a focused intent of the company, Mediamente Consulting, in which i did my internship.
Mediamente Consulting is a society specialized in the Business Analytics. It is located in Incubator I3P within Politecnico of Turin. The company has realized a modular platform for Big Data and Advanced Analytics, developed for social, web analytics, cross-selling and NBA (Next Best Action) areas of interest. It offers both tailor-made applications and package applications. The company intends to extend the platform adding a module related to e-commerce domain.
The main aim consists in the electronic shop windows monitoring in order to associate it to physical stores supervision. This thesis project matches the business goal of Mediamente Consulting with needs of one of their business clients. This client company works in the fashion industry world and its request concerns the need to monitoring products pricing on its sellers sites, also taking into account products pricing related to its competitors.
The work is characterized by a data-driven approach. In fact, capturing not only pricing values but also all product information from the detail pages of the online stores, this process allows to extract and evaluate key performance indicators (KPIs)

from the product details that can trigger potential trends, driving business strategies. In this way, business progresses are compelled by data, rather than by intuition or by personal experience.
In addition to the monitoring, an other intent consists in implement price forecasting in order to set up business strategies to win over competitors. For this purpose, forecasting algorithms will be exploited and compared.
Lastly, the final part of the thesis concerns the data visualization. Realizing dashboards and reports, through BI tools, allows to visualize the results of all business analysis aiming to provide the business information, extracted from data, to the client company driving its strategies. In this way, highlighting the most important business indicators on reports, the intent is achieving a demand forecast, foreseeing customer demand on the basis of the past events and prevailing trends in the present.

Chapter 1

Advanced Analytics

1.1 Advanced Analytics

Big Data Analytics represents the process of analyzing a number of data sets in order to extract unknown correlations, hidden patterns, market trends, customer preferences and other useful information that allow organizations to obtain insights from data, improving their business decisions.
Big Data Analytics applications enable data scientists, predictive modelers, statisticians to examine growing volumes of data. It has become increasingly important in both the academic and the business communities over the past two decades. Data investigation is continuously triggering significant interest in Business intelligence and analytics, offering a wide range of opportunities. As a data-centric approach, Big Data Analytics has its roots in the well-established database management field.

1.1.1 Analytics 1.0 - the era of "business intelligence"

Data management and warehousing is considered the foundation of analytics 1.0, where data are mostly structured and often stored in commercial relational database management systems(RDBMS). At the beginning, the focus rotated around the storing and the analysis of sales, production processes and customer data.
This was the era of the enterprise data warehouse, used to capture information, and of business intelligence software, used to query and report it. Usually, data sources derived from internal systems and were relatively small. The capture and integration of data is achieved through the employment of the enterprise data warehouse and tools for extraction, transformation and load(ETL). Reporting functions,statistical analysis and data mining techniques are adopted for data segmentation and clustering, classification and regression analysis, association analysis, anomaly detection, and predictive modeling in various business applications.
Readying a data set for inclusion in a warehouse was difficult. Analysts spent much

of their time preparing data for analysis and relatively little time on the analysis itself.
Reporting processes, the core of business intelligence activity,addressed only what had happened in the past; they offered no explanations or predictions.Decisions were made based on experience and intuition.

1.1.2 Analytics 2.0 - the era of big data

The HTTP-based Web 1.0 systems was characterized by the explosion of Web search engines such as Google and Yahoo and E-Commerce businesses such as Amazon and eBay. Consequently, their spread facilitated customer role change considerably. In the 2000s, social network and internet-based firms began to analyze new kinds of information.
Unlike small data, big data didn't represent only company's internal transaction data, but they was also generated from external sources, coming from the internet, audio and video recordings, sevaral sensors and so on.

Unlike Analytics 1.0 technologies, Analytics 2.0 systems are centered on web analytics, social network analysis and text mining. Data extracted through web analytics obligated organizations to deal with unstructured data, starting to make use of a new class of databases known as NoSQL. Furthermore, in order to storage and manage the new remarkable amount of data, firms needed to process data not on a single server but across parallel servers. This aspect provoked the considerable spread of a new open source framework, Hadoop, in order to store and process big data rapidly. Additionally, Machine-learning methods were used to rapidly generate models from the fast-moving data, and other technologies, such as "in memory" databases, were introduces to optimize data processing.
All these aspects triggered the diffusion of a new figure, the data scientist, who possessed both computational and analytical skills.
It was remarkable the spread of a new concept, the Advanced Analytics. "Advanced Analytics is a suite of analytical applications that helps measure, predict, and optimize organizational performance and customer relationships"[2].
While BI was analogous to OLAP (online analytical processing) query-and-reporting tools for many years, many organizations discovered that the effective use of information requires more than reports that show historical data. Effective decision making for competitive advantage is driving the need for such a more comprehensive approach to BI. In the BI evolution, industry leaders are currently transitioning from an operational BI of the past to an analytical BI of the future that focuses on customers, resources, and abilities to drive new decisions everyday.

1.1.3 Analytics 3.0 - the era of data-enriched offerings

Since the number of mobile phones and tablets surpassed the number of laptops and PCs for the first time in 2011, consequently a new research opportunity is emerging in Analytics 3.0. Mobile devices and their applications are continuously transforming different features of society. Every device, shipment, and consumer leaves a trail. Organizations have the ability to analyze those sets of data for the benefit of customers and markets. They attracted viewers to their websites through better search algorithms, recommendations from friends and colleagues, suggestions for products to buy.
Organizations need to integrate small and large volumes of structured and unstructured data, from both internal and external sources, in order to yield new predictive models.
"The most important trait of the Analytics 3.0 era is that every company—not just online firms—can create data and analytics-based products and services that change the game."[11]
While in the Analytics 2.0, firms focused on Hadoop clusters and NoSQL databases, today they tend to combine the classical query approach with Hadoop, exploiting also other solutions such as graph databases.
The challenge in the 3.0 era is to adapt operational, product development, and decision processes to take advantage of what the new technologies and methods can bring forth. The new capabilities required of both long-established and start-up firms can't be developed using old models for how analytics supported the business. Companies that want to prosper in the new data economy must once again fundamentally rethink how the analysis of data can create value for themselves and their customers. Analytics 3.0 is the direction of change and the new model for competing on analytics.

1.2 Web Analytics

The success of an E-Commerce site is, in part, related to its usage facility. Web data are easy to collect but analysis and interpretation are time-consuming. Web analytics may represent an important approach for E-Commerce site managers therefore need to effectively improve the usability of their websites.
"Web analytics is an approach that involves collecting, measuring, monitoring, analysing and reporting web usage data to understand visitors' experiences. Analytics can help to optimise web sites in order to accomplish business goals and/or to improve customer satisfaction and loyalty."[6]
Most web analytics processes come down to four essential stages or steps, which are:

1. Collection of data

2. Extracting information from data processing
3. Analyzing Key Performance Indicators (KPI)
4. Implementing business decisions

There are at least two categories of web analytics; off-site and on-site web analytics.

1. Off-site web analytics refers to web measurement and analysis regardless of whether a company owns or maintains a website. It takes into account the visibility of the website, the opportunities that it offers and customers' feedbacks.
2. On-site web analytics, the most common, evaluates website performances and visitors' behaviors on the website.

There are two common methods used by web analytics tools to collect web traffic data. The first involves the use of server-based log-files, and the second requires client-based page-tagging. Web analytics started with the analysis of web traffic data collected by web servers and held in log-files.

Log file analysis Web servers record some of their transactions in a log file. Initially, only client requests regularity was stored in the log files, since each website often consisted of a single HTML file. However, with the introduction of images in HTML, and websites that spanned multiple HTML files, this count became less useful.
Moreover, the considerable spread of robots, search engine spiders, web proxies and dynamically assigned IP addresses, complicated the identification of human visitors to a website. Log analyzers started to exploit the cookies to track visits, detecting spiders and ignoring their requests.
An other issue concerns the web caches application. If a person revisits a page, the second request will often be retrieved from the browser's cache, and so no request will be received by the web server. This means that the person's path through the site is lost. Caching can be defeated by configuring the web server, but this can result in degraded performance for the visitor and bigger load on the servers.

Page tagging An other method exploited to collect web traffic data is known as Page Tagging. It is realized through the implementation of a "Javascript Tracker" which tags a visitor with a cookie. Usually, through a module written in JavaScript, data are gathered and sent to a central server for the subsequent analysis. Ajax can also be used in conjunction with a server-side scripting language, such as PHP, to manipulate and store data in a database, basically enabling complete control over how the data is represented.

Both methods have their merits. Log file analysis is almost always performed in-house. Page tagging can be performed in-house, but it is more often provided as a third-party service. The economic difference between these two models can also be a consideration for a company deciding which to purchase.

A large part of website traffic comes from non-human visitors, as search engine spiders. Web Server Log files record human and non-human traffic information. They play a crucial role regarding website errors, in order to avoid to lose visitors because of problems such as missing pages or broken links. Unlike Javascript, Log Analyzers provide tracking of all possible site error.
On the other hand, log file is not able to collect some kinds of data such as screen resolution, color depth of visitors' browser. The Javascript tracker has an advantage here.
Some companies produce solutions that collect data through both logfiles and page tagging and can analyze both kinds. By using a hybrid method, they aim to produce more accurate statistics than either method on its own.

1.3 Data Integration

"Today's BI architecture typically consists of a data warehouse (or one or more data marts), which consolidates data from several operational databases, and serves a variety of front-end querying, reporting, and analytic tools."[3]
Populating the data warehouse, through extracted data from heterogeneous sources, is realized by implementing a data integration pipeline. The typical data integration pipeline represents a batch process. It's put in practise through extract-transform-load (ETL) tools.
ETL process consideration has changed considerably over time. In the past, ETL design and implementation was considered only a supporting task for the data warehouse, and was largely ignored by the research community. Perhaps, it appeared as a simple task of data transfer and integration. Conversely, nowadays, ETL represents a labor-intensive activity. Its design and implementation, usually, consumes a crucial fraction of the effort in data warehouse projects.

ETL process ETL covers a process of how the data are loaded from the source system to the data warehouse. The process includes three main steps:

- Extraction: the main objective of this step consists in data extraction from different sources as fast as possible, in order to prepare them to the subsequent transformation process.

- Transform: this step applies a set of rules in order to transform the data from the source to the target. A process of data homogenization is applied so that data can later be joined. Moreover, in this step, other operations are included such as joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules.

- Loading: the target of Load step is usually a database. It is helpful to disable any constraints and indexes before the load and enable them back only after the load completes.

Managing ETL process As with every application, there is a possibility that the ETL process fails. This can be caused by missing extraction from one of the systems, missing values in one of the reference tables, or simply a connection or power outage. Therefore, it is necessary to design the ETL process keeping fail-recovery in mind.
"The decision to store data in a physical staging area versus processing it in memory is ultimately the choice of the ETL architect. The ability to develop efficient ETL processes is partly dependent on being able to determine the right balance between physical input and output (I/O) and in-memory processing."[12]
The issue related to storing data or not depends on two conflicting problems:

- Moving data from data sources to ultimate target as fast as possible

- Failure handling without restarting from the beginning of the process

Staging Area represents the area in which the data are simply dumped, ready to be processed during the next processing stage. It plays a fundamental role in ETL process; in fact, it allows to store intermediate results. Staging area is not accessible to the final business user, because it contains data so far inappropriate for user goals. Only organization's employees can exploit Staging Area content.

ETL tools There are two possible strategies to follow in order to realize a ETL process: Hand-coded ETL and Tool-Based ETL.
Hand-coded ETL provides unlimited flexibility and it represents a crucial requirement. In many instances, a unique approach or a different language can provide a big advantage. Despite this, if the transformation process become more sophisticated during the time or there is a need to integrate other systems, hand-coded solution triggers a significant manageability drop.
On the other hand, market offers many ETL tools. They provide connectors to several data sources like databases, xml, flat files in order to optimize ETL process and they allow to implement operations such as filtering, reformatting, sorting, joining, merging, aggregation.

Until recently, some tools worked almost uniquely with structured and semi-structured data, and other tools were born to manage unstructured data. Thanks to the continuous business technology evolution, nowadays this distinction is gradually fading. Consequently, as time goes by, each tool is able to handle and process all types of data.
Moreover, ETL tools are divided into commercial Tools such as Oracle Data Integrator, IBM InfoSphere DataStage, Informatica and open source ETL Tools, among other Pentaho, CloverETL, Talend.

1.4 Data visualization

Data visualization is the process of helping people understand patterns, trends, and insights by transforming data into a visual context. This process can be realized exploiting column or pie charts, reports, dashboards, pivot table and so on.
Business visualization tools don't provide the initial stages of data transformation and collection. For this reason, it's important to distinguish between data analytics and data visualization.
The first one consists in a platform with tools and algorithms that handle and process all the data to learn and extract hidden knowledge.
On the other hand, data visualization tools focus on reporting data rather than analyzing it, allowing business users' brains to understand a large amount of data through a visual approach. Moreover, the images used in data visualizations can also have interactive capabilities which permit users to manipulate data for query and analysis. Consequently, neither data analytics nor data visualization can stand for the unique component of a powerful Business Intelligence software.

Three crucial strengths of data visualization usage are described as follows:

1. Numerical data allow to simplify complex data understanding, making easier their visual presentation

2. Exploiting letters permits to individuate instantly variables, patterns and outliers

3. Data visualization highlights data changes over a period of time, identifying patterns developments and business indicators efficiently.

The nature of data visualization makes it both an art and a science.

1.4.1 Visual Analytics

Data visualization and visual analytics chase the same purpose that consists in obtaining better insights into data. Though often used interchangeably, these two

terms are inherently different as their approaches are to data.
Visual analytics, too, works towards representing data in clear format, but moves the focus on different aspects, such as interactive visualizations. It also exploits analytical processes such as data mining techniques and statistical evaluation. The design of visual analytics tools is based on perceptual and cognitive principles.
The increasing adoption of new technologies such as mobile business intelligence and location intelligence software also increases the opportunities for accenting visual analytics strengths.

Here are the main features of visual analytics:

1. Visual analytics does not work with raw and unstructured data. It exploits filtered and processed data working on their visualization aspect

2. Visual analytics requires human action when analysis parameters change, in order to make data appropriate to data visualization continually

3. Visual analytics takes into account both scientific and interaction techniques approaches

1.4.2 BI and Visual Analytics Trends

Ease of use Most business users opt for natural language to interact with data in the easiest way and not through reports and code. Since understanding the context is important for many types of BI and analytics, the research moves the focus to optimizing pre-building queries implementation and to prepare data more rapidly using machine learning and artificial intelligence techniques.
Nowadays, the main purpose consists in exploiting natural language queries in order to extract information from the data sets.

Data discovery Data discovery is a user-driven process for patterns, specific items or outliers identification in a data set. It becomes a learning process as the software uses machine learning to understand users' preferences and users can determine more quickly whether the data is useful for analysis.
Data discovery may be divided into three main categories:

1. data preparation

2. guided advanced analytics

3. visual analysis

Data preparation is essential for analysis and helps business users to connect to enterprise and external data sources.

Interactive and new visualization types enable decision-makers to see major trends in an instant, as well as spot outliers. Visual analysis is an important feature for decision-makers to act on data. Visualization make use of our brains' pattern recognition capabilities to digest information.
Guided advanced analytics functions is a challenge provide statistical information on data which users can employ for more sophisticated and pattern oriented data analysis.

Growth in self-service BI Self-service business intelligence (SSBI) allows business users, without a scientific and appropriate background in statistical analysis and data mining algorithms, to access and analyze business data.
The remarkable growth in self-service BI is bringing about an increasing demand for better metadata to manage more effectively. Consequently, business companies to move the focus on other tasks in order to improve their performance, giving custody of data visualization step to the business user. In fact, adopting SSBI approach, end users make decisions based on their own queries and analyses.
Since self-service BI software is used by people who may not qualified, the user interface (UI) for BI must be intuitive, with a dashboard and navigation that is user friendly. Business users should be able to query the data and create personalized reports with very little effort.

Embedded Business Intelligence Embedded BI (business intelligence) represents the integration of self-service BI tools into commonly used business applications. Embedded BI and analytics functionality plays an important role in enabling users to progress faster with data-driven decision making. BI tools permit users to improve their confidence with real-time analytics, data visualization and interactive reporting.
Users may prefer to implement and manage BI algorithms in their own business development environments. Embedded BI may become a defined component of a workflow, allowing users to set and modify parameter sets to develop the decision making towars a new approach.
Nowadays, a number of providers of business applications give value to embedded platforms for their processes.

Big data in motion Business Intelligence and Analytics have traditionally worked on data already integrated in data lake, data warehouse, database or other persistent storage system. These data represent the so-called data "at rest".
Data at rest refers to inactive data stored physically in any digital form. The data analysis occurs separately and distinctly from any action taken on the conclusions of that analysis. Users interact with historical data at rest in these sources. Because

of its nature data at rest is of considerable concern to business organizations.

Nevertheless, high volume, high variety and high velocity of data turn the technology focus toward solving problems of analyzing, managing, and protecting data in motion. Big data in motion refers to a stream of data moving through any kind of network.
The collection process for data in motion is similar to that of data at rest; however, the difference lies in the analytics. In this case, it is applied in real-time as the event takes place. Data in motion analysis follows the principle that current customer actions can offer more business information compared to historical data.
For some operational applications, data in motion can trigger some problems about big data architecture implementation. Latency between systems of engagement and the analytics platforms can be too long. This latency issue arises when quick results are required but the analytical engine needs to populate an entire data set before launching a query.
To succeed in real time data management, organizations may need to exploit different kinds of software technology. A big challenge is represented by sensor data handling. A number of organizations need to take advantage of advanced analytics technique on sensor data. Additionally, sensor data are both at rest and in motion data, requiring more approaches at the same time.
Analyzing data in motion, in addition to traditional data at rest management, allows to capture more indicators for better business insights. However, it requires an appropriate infrastructure and software, optimizing time dependent issues, such as latency.

Data Governance Data Governance describes a revolutionary process for a company. Organizations need to gain control of their data, changing the role of their employees. They need to implement data quality and data management directly. Consequently, they require technology's help to do it, through the employment of metadata. They move the focus on data stewardship.
The objective of the data governance and data steward role is to optimize processes for preventing and correcting issues with data to improve data quality available for decision-making.
The correct choice of the technology is critical. Its role consists in helping and simplifying some aspects of data governance, such as security or data quality processing, in order to ensure that high quality data exists throughout the complete lifecycle.

1.4.3 Traditional vs Self-Service BI

Most BI users use Data Visualization Tools with their products to handle their raw data to create useful and informative data. Data Visualization Tools are the software

tools through which the raw and unstructured data can be used as informative data to increase the Business insights and sales.
It is remarkable the continuous evolution in the Tools application. In fact, actual BI trends are bringing power to the business users. Consequently the future is represented by self-service, data discovery and quick insight.
It is no coincidence that Data Visualization Tools such as Qlik and Tableau hold top position and they are achieving resounding success.
The complaints about traditional BI software are well-known:

- slow
- inflexible
- creating reports is time-consuming

Organizations need to change access time to information in order to manage also real time data efficiently. Moreover, since they want to centralize their role in data management, they need to exploit user-friendly and interactive tools. Consequently, flexible, fast and lightweight tools are required.
On the other hand, some of the newer self-service BI tools can fail to provide the accuracy and the scalability of the traditional BI. Moreover, business users are often unaware of the complexities of data preparation and the risks involved in getting it wrong. Without an authority guaranteeing strong data governance, they may miss mistakes in their own data analysis.

Associative Difference: the modern analytics Traditional BI tools are defined "query-based tools" because SQL is required to pull data from many sources. Their fundamental architecture for modeling data and supporting interactivity depends on query based approaches. These factors provoke a restricted linear exploration and slow performance.
Modern analytics tools fully combine several data types from different sources, even imperfect data, without suffering the data loss or inaccuracy that typically occurs with SQL queries and joins. In fact, the crucial innovation is represented by the fact that business users have access to all their data from all their sources, instead of just the limited result sets returned by queries. On the other hand, query based tools require primary and secondary data sources and then some data will be lost.
With modern analytics tools, such as Qlik's Associative Engine, business user can exploit a total data exploration, without restrictions or boundaries.
Users can immediately spot potential areas of interest, think of new questions, and continue to explore further.

1.5 Big Data in practice

In this section some business models, regarding Big Data Analytics, of well-known companies are described. They show how famous organizations utilize Big Data to predict trends and consumers' behaviour.
The amount of data available in our increasingly digitized world is literally exploding. We have created more data in the past two years than in the entire previous history of mankind. "By 2020, it is predicted that about 1.7 megabytes of new data will be created every second, for every human being on the planet." [7]
Data are coming from millions of messages and emails sent every second via email, Facebook, Twitter, Whatsapp but also from the incredible amount of digital photos and video taken each year. The new role acquired by Instagram in this domain demonstrates how Big Data diffusion is evolving.
An other considerable source of data is represented by the sensors. They may provide a variety of data regarding different kinds of useful business information.

Analyzing well-known companies' approaches to Big Data and business solutions that represent the secret of their remarkable success, may help to develop more informed insights concerning Big Data Analytics methods.

1.5.1 Netflix

Netflix, a streaming movie and TV service, plays a crucial role in Big Data Analytics. Its most important strength consists in prediction what customers will enjoy watching.
At the beginning, analysts owned only few customers information, such as Customer ID, Movie ID, movie watching date, customer's rating. Consequently their studies were limited by the lack of a right amount of data.
Gradually, the considerable data streaming spread allowed Netflix analysts to capture a large amount of data on their customer. This aspect is still enabling Netflix to build predictive models.

Another central element to Netflix's attempt to give us films we will enjoy is tagging. The company pay people to watch movies and then tag them with elements the movies contain. They will then suggest customer watch other productions that were tagged similarly to those they enjoyed.

The need to collect data coming from streaming and customer profiles inspires Netflix to exploit new solutions as Spark for streaming and machine learning in order to create personalized analytics, highlighting customer's needs. In this way, Netflix hope, consisting in improving the number of hours customers spend using that service, can be realize.

The personalization of customers' needs allow Netflix to optimize the customer retention. This aspect represents a large part of its success. Nowadays, Netflix accounts for one-third of peak-time Internet traffic in the US.

With regard to technology, Netflix used Oracle databases initially, but the successive Big Data-driven analysis forced them to exploit different databases, such as NoSQL and Cassandra. Hadoop is used as Netflix data infrastructure, along with in-memory databases to optimize the performance. Business intelligence tools play an important role in Netflix business, such as Teradata and Microstrategy.

Netflix is mentioned because of its innovation ideas about "personalized TV", where each viewer has his own schedule of movies to watch, based on their preferences, examined through big data advanced analytics.

1.5.2 Facebook

Facebook represents the biggest social network in the world. It's free to the end user. Millions of people every day also use it to read news, keep in touch with friends, interact with brands and make buying decisions.

Facebook created a new method to handle and share personal and social data. Its management strategy, concerning end user's individual information, represents its key to success. It captures a large variety of end user's data, such as where he lives, works, his passions regarding culture and sports, how many friends he has and so on. All these kinds of information allow Facebook to carry out efficient advertising strategies and to predict new potential users' interests, also suggesting new friends who manifest similar curiosities.
Facebook, with 1.5 billion active monthly users, has access to far more user data than just about anyone else. Undoubtedly, it holds one of the biggest and most comprehensive databases of personal information ever collated, and it is expanding every second of every day. It uses its own distributed storage system based on Hadoop's HBase platform to manage storage. It is also known that Facebook makes use of Apache Hive for real-time analytics of user data.

Additionally, Facebook expands by buying out external services and data, such as Instagram and Whatsapp service, in order to combine more different kinds of business data, coming from multimedia and streaming domains.

1.5.3 LinkedIn

LinkedIn is the world's largest social network for professionals, with more than 400 million members in over 200 countries. LinkedIn connects professionals by enabling

them to build a network of their connections and the connections of their connections.
People prefer to share their expertise and connect with like-minded professionals to discuss various issues of interest in a platform like LinkedIn, as it allows them to represent themselves formally in a less traditional manner.

Its most important strength concerns the detailed adopted data tracking. Big Data is at the heart of LinkedIn's strategies and decision making ; it tracks every user's action on the site. Much like other social media networks, LinkedIn use data to make suggestions for their users, such as "people you may know".

Its profiles views monitoring allows to implement robust machine-learning techniques in order to offer better suggestions for users. In fact, one of the features that set LinkedIn apart from other social media platforms, like Facebook, is the way it lets you see who has viewed your profile. For instance, a user worked at Company A and Company B in two different time periods in the past. If he almost never click on the profiles of people working at Company A but regularly check out profiles and views related to Company B, this aspect allows LinkedIn's algorithms to prioritize Company B in their suggestions for the user.

The company serve tens of thousands of Web pages every second of every day. All those requests involve fetching data from LinkedIn's backend systems, which in turn handle millions of queries per second. Hadoop, with a number of machines running map-reduce jobs, is used concerning the infrastructure, in order to enable scalability, availability, avoiding single points of failure. Other key parts of the LinkedIn Big Data jigsaw include Oracle, Pig, Hive, Kafka, Java and MySQL.

It's fascinating the comparison between Facebook and LinkedIn approaches to Big Data Analytics. Both these social media exploit and analyze personal data. However, LinkedIn focuses the attention mainly on user's competences. Moreover, it allows people to check who has viewed their profile, helping users to increase their effectiveness on the site.

1.5.4 Google

Undoubtedly, Google symbolizes the core of Big Data Analytics. It plays a crucial role in the domain of the Web Analytics. In fact, Google Analytics , a Google free Web analytics service, is the most used web statistics service. It allows to obtain detailed analysis on visitors of a website. The size of Google's index, that represents its archive of every web page it can find, is estimated to stand at around 100 petabytes.

Google PageRank, developed by Google founders Larry Page and Sergey Brin, stands for its first and crucial search algorithm. The principle is that the more pages link to a particular page, the higher that particular page's "authority" is. Consequently, the algorithm assigns a rank to every page in its index, based on how many other sites use similar keywords linked to it, and also based on how "authoritative" those linking pages are themselves.
Google realizes its index of the Web sending out software robots, often called spiders or crawlers, which capture all kinds of information contained on a website, storing them in the enormous Google's own vast archives.

Google model is mentioned because it is related to the first phase of this thesis project, that concerns the data ingestion theme, implemented through the use of web spiders.

1.5.5 Ralph Lauren

The way the world we live in is increasingly becoming digitally connected is impacting everything, and fashion is no exception. Wearables are expected to become increasingly popular as the Internet of Things takes off.
In the wider fashion world, Big Data is increasingly playing a fundamental part in trend forecasting. Data coming from sales, promotional events, social media, fashion shows, need to be analyzed in order to forecast customers' future demand.
Ralph Lauren represents a big name in high-end consumer fashion. It's interesting to study their Smart Polo shirt model. "Sensors attached to silver threads inside the shirt pick up the wearer's movement data as well as heart and breathing rates, steps taken and number of calories burned."[7]
The key to success consists in the fact that the shirt becomes a real-time data collector. It stores biometric data, like heart rate, and also direction and movement data that will reveal themselves useful business key performance indicators. Successively, these data are sent to a centralized cloud in order to apply advanced analytics on them.
The purpose of their available app concerns the realization of tailored users' workout, through the insights derived from the big data analytics.

Since the case study of this thesis regards a business client coming from the fashion world, Ralph Lauren strategies and models assume a considerable relevance as big data analytics examples.

Chapter 2

Infrastructure Design

2.1 Introduction

A properly designed Infrastructure represents certainly a crucial requirement of a IT network. IT network consists of the entire set of software and hardware components that are designed to connect devices within the organization, as well as the company to other companies and the Internet.
Maintaining a business depends strongly on the strengths and the design of a company's backbone, that's composed of the entire IT network. The available solutions are numerous; the correct architecture's choice is everything except ordinary.
For these reasons, the requisites analysis becomes crucial. Hardware devices,along with software solutions, must guarantee a number of features:

- Connectivity
- Security
- Service and Scalability
- Routing/Switching Capabilities
- Access Control

With the proper design, an organization is able to support the growth of its business without having to redesign the network. In fact, thanks to a robust scalability, a network can change without requiring the infrastructure review.
Without a doubt, a correct infrastructure set-up can considerably improve speeds, productivity and performances.
With regards to data ingestion aspect, one of the key features concerns the data storage. As follows, different data repository solutions are described.

2.2 Data repository solutions

Data lake Data lake is a storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured and unstructured data. The data structure and requirements are not defined until the data is needed.
The idea of data lake is to have a single store of all data in the enterprise ranging from raw data (which implies exact copy of source system data) to transformed data which is used for various tasks including reporting, visualization, analytics and machine learning.
The Hadoop community has popularized it a lot, with the focus on moving from disparate silos to a single Hadoop/HDFS. Its key features allow to avoid to connect to a live production system every time you want to access a record and, moreover, to exploit relatively inexpensive hardware. This eliminates the upfront costs of data ingestion, like transformation. Once data is placed into the lake, it's available for analysis by everyone in the organization.

Data warehouse Data warehouse or enterprise data warehouse represents the previous most common solution. It is considered a core component of business intelligence. They store current and historical data in one single place[2] that are used for creating analytical reports for workers throughout the enterprise.
The data stored in the warehouse is uploaded from the operational systems, such as marketing and sales. The main source of the data is cleansed, transformed, catalogued and made available for use by managers and other business professionals for data mining, online analytical processing, market research and decision support.
The typical Extract, Transform, Load (ETL)-based data warehouse[4] uses staging, data integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing these transformed data in an operational data store (ODS) database. The integrated data are then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchical groups, often called dimensions, and into facts and aggregate facts. The combination of facts and dimensions is sometimes called a star schema. The access layer helps users retrieve data.
The following key features illustrate the different characteristics between a data lake and a data warehouse:

1. The data: a data lake include both unstructured and structured data. Whereas, data warehouse includes a structured and processed data set.

2. The storage: a data lake is designed for low-cost storage. Storage is more expensive for a data warehouse.

3. The processing: a data lake uses a schema on read, whereas a data warehouse uses a schema on write.

4. Agility: a data lake is highly agile with respect to a data warehouse.

5. User perspective: while a data lake, in some sense, tends to be the focus for data scientists, a data warehouse is designed for business professionals.

Data Hub A data hub is a collection of data from multiple sources organized for distribution, sharing, and often subsetting and sharing. Data is ingested in as close to the raw form as possible without enforcing any restrictive schema. It is a hub-and-spoke approach to data integration.
A data hub differs from a data lake by homogenizing data and possibly serving data in multiple desired formats, rather than simply storing it in one place, and by adding other value to the data such as de-duplication, quality, security, and a standardized set of query services. Data lakes do not index and cannot harmonize because of the incompatible forms that will be held. The prime objective of an EDH is to provide a centralized and unified data source for diverse business needs.
A data hub differs from a data warehouse in that it is generally unintegrated and often at different grains.

Why Data Hub? Point-to-point interfaces between pairs of applications represent a possible architectural alternative to data hubs. A point-to-point interface that moves data from one application to another is much simpler to implement than any hub.
However, point-to-point interfaces introduce several issues:

1. Poor control and minimal governance around data

2. They rely on the application pair involved to do things like data integration and transaction integration

3. It's typically difficult to modify their structure

4. They are typically poorly documented and understood

5. They promote the creation of silos that are very difficult to evolve in line with business changes

For these reasons too, data hub represent an attractive architectural solution in opposition to point-to-point approach. Nevertheless, it is possible to poorly implement a data hub.

2.3 Data Hub: common styles

Data Hub approach provides different architectural solutions, as described below.

The Publish-Subscribe Data Hub One or more applications, called publishers, produce data and store them in the hub. Other applications, called subscribers, take specific data sets from the hub. Both the data input and output phases may be implemented through a "pull" or "push" approach, between data hub and the applications. The data hub can coordinate the pushing and pulling of data by recognizing when a publisher is ready to publish, and informing a subscriber when data is available. This can be tied into enforced service level agreements (SLAs). Tipically, governance by the group controlling this kind of hub is often weak. Consequently, the hub may simply improve the growth of virtual point-to-point interfaces. The Figure 2.1 illustrates the mechanism of this solution.

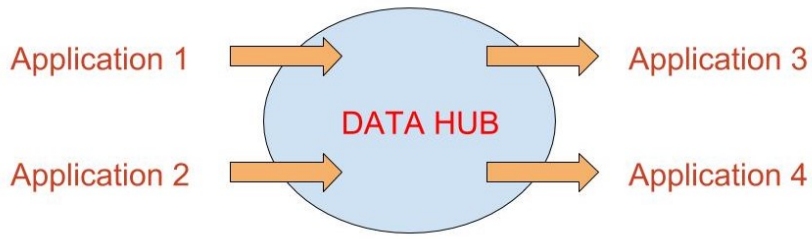


Figure 2.1. Publish and Subscribe Hub

The Operational Data Store (ODS) for Integrated Reporting An important key factor about the ODS is that it does implement integration. This style derived from the need to shift reporting phase out of transactional applications. In fact, initially, the database of transactional applications were replicated and the reports ran off the replicas. It is noticeable that this aspect degraded considerably the performances. Then, it was realized that data from several applications could be integrated into a hub and integrated reporting run from the hub.
A relevant issue of ODS is about how much and what kind of history to keep. Because of this problem, this solution is today at odds with real-time data warehouses and data marts.

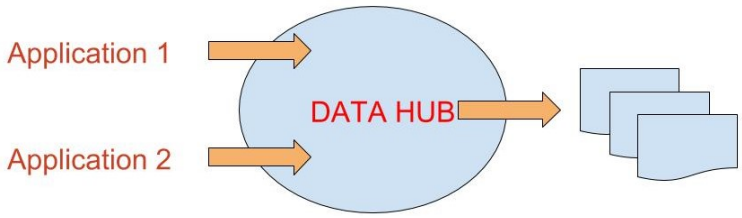


Figure 2.2. ODS for Integrated Reporting

The ODS for Data Warehouses The first integration of transactional applications data is realized in the hub. Further integration of data may occur in the warehouse layer, with additional data coming from applications whose data may not be integrated in the ODS. With regards to new trends, a number of data are sent directly from applications to real-time warehouses and marts. Additionally, an other issue concerns the fact that both data warehouse and data hub contain integration areas. This aspect may entail the duplication of the same processes in different phases of data managing.

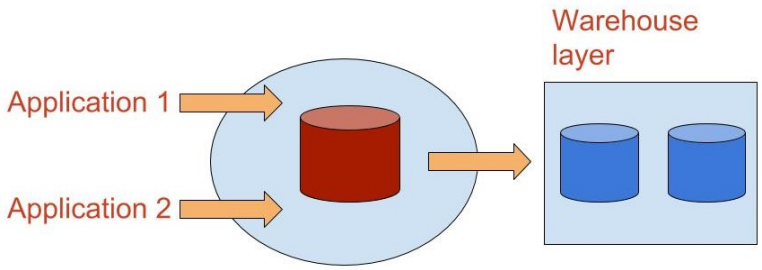


Figure 2.3. ODS for Data Warehouses

The Master Data Management (MDM) Hub The Master Data Management (MDM) hub is a database with the software to manage the master data that is stored in the database and keep it synchronized with the transactional systems that use the master data.
There are three basic styles of architecture used for Master Data Management hubs, the repository, the registry and the hybrid approach:

1. Repository: the complete collection of master data for an enterprise is stored in a single database, including all the attributes required by all the applications that use the master data. The applications that consume, create, or maintain

master data are all modified to use the master data in the hub, instead of the master data previously maintained in the application database.

2. Registry: each source system remains in control of its own data and remains the system of entry, so none of the master data records are stored in the MDM hub.

3. Hybrid approach: Includes features of both the repository and registry models. It allows to replicate the most important attributes of master data records, contained in the application databases, in the MDM hub in order to satisfy a significant number of MDM queries directly from the hub databases. The resulting issue concerns all the conflicts regarding synchronization and replication.

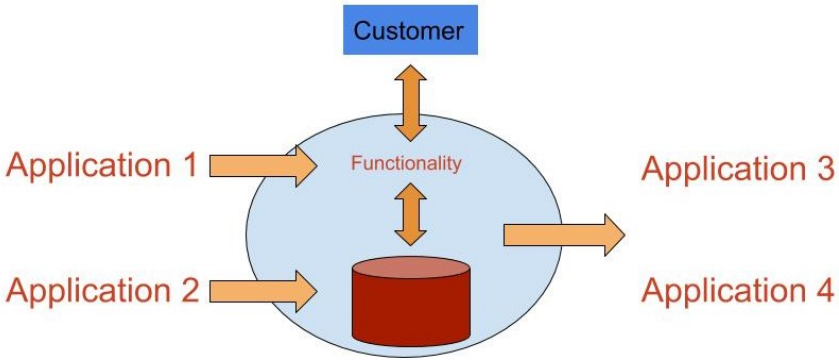


Figure 2.4. Master Data Management Hub

Message Hub Handling real time messages flowing represents the most important feature of this style. As shown in Figure 2.5, the presence of queues, between the several applications and the data hub, may require switching messages from one queue to another, or waiting until a set of messages arrive before processing them as a whole logical unit of work, in order to make them conform to message models corresponding to business needs.

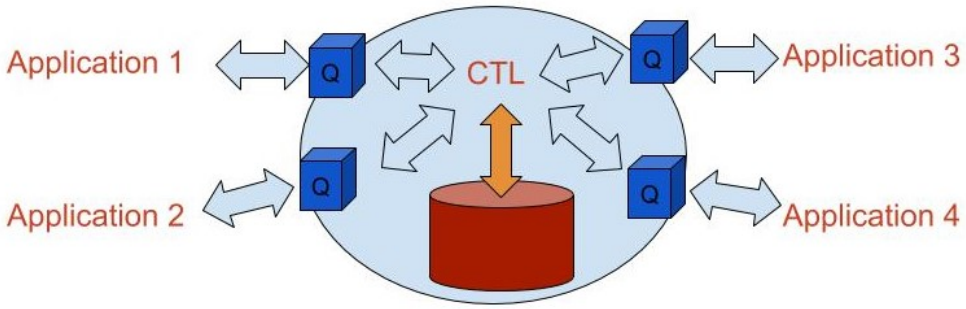


Figure 2.5. Message Hub

Integration Hub Unlike the ODS for Data Warehouses style, this approach aims at realizing integration once in one place. Quite often, transaction applications actually do integration, and often the integration is redundant. In this model, the warehouse layer does not repeat the integration that has already been performed in the hub.

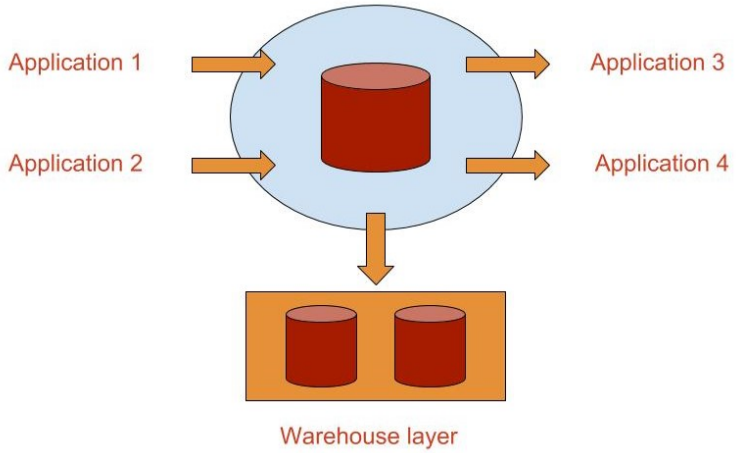


Figure 2.6. Integration Hub

2.4 Cloud computing

Nowadays, the continuous increase in the volume and detail of data captured by organizations has produced an overwhelming flow of data in either structured or unstructured format. In the Big Data World, data are generated and processed rapidly, they are numerous and this aspect doesn't allow to handle them totally through relational databases.
Cloud computing is one of the most significant fast-growing technology in modern ICT industry and business. It represents a promising solution that offers unlimited on-demand storage and compute capacity, minimizing the cost. The advantages of cloud computing include virtualized resources, parallel processing, security, and data service integration with scalable data storage.
Adopting cloud computing has become the main strategy of a number of organizations and individuals in order to satisfy the need to store, process and analyze large amounts of datasets.

2.4.1 Cloud computing and Big Data

The two concepts Big Data and Cloud computing are so connected. Big Data allow users to process distributed queries across multiple datasets and return resultant sets in a timely manner. Cloud computing provides the underlying infrastructure through the use of Hadoop, a distributed platform. It is noticeable that a cloud infrastructure is necessary in order to perform big data processing and analysis.
Furthermore, Cloud computing not only provides facilities for the computation and processing of big data but also offers a service model. Therefore, computing clouds render users with services to access hardware, software and data resources, thereafter an integrated computing platform as a service, in a transparent way:

1. Hardware as a Service (HaaS): As the result of rapid advances in hardware virtualization, users can buy IT hardware or an entire data center as a pay-as-you-go subscription service.

2. Software as a Service (SaaS): Software is hosted as a service to customers across the Internet. This solution allows to avoid both customer software maintenance and the need to install e run applications on the customer's local computer.

3. Data as a Service (DaaS) : Users can handle and pull out data in various formats via services on the network.

As shown in the Figure 2.8, the Cloud computing can deliver the Infrastructure as a Service (IaaS) for users, in addition to the support of the Haas, Saas and

Daas. Users thus can on-demand subscribe to their favorite computing infrastructures with requirements of hardware configuration, software installation and data access demands.

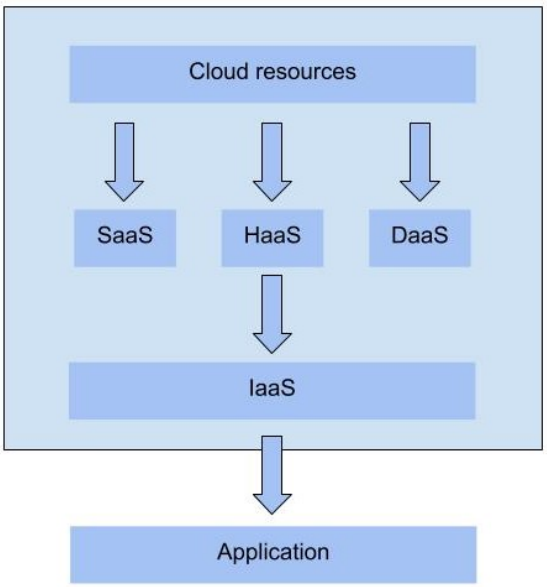


Figure 2.7. Cloud Services

2.4.2 Cloud data storage vs On-premise data storage

The question of whether the optimal deployment route is in the cloud or on-premise provokes often different considerations. Following the main features, that highlight the dispute between the two different approaches, are described:

- Scalability: Scaling up on-premise systems is a time-consuming and resource-intensive task, as it usually entails purchasing and installing new hardware. On the contrary, regarding the cloud, scalability can be accomplished easily and instantly;

- Reliability: For on-premise data storage, reliability is a function of the quality of hardware and staff. It depends on the singular organization. On the other hand, cloud-based data storage are always available;

- Security: it depends on several factors. Security represents one of the biggest challenges for cloud service providers. The recent cloud architectures contain entire and robust modules that take care of cloud data security and encryption. However, if a specific business requires certain security measures, that can not

be met by the security offering of the cloud supplier, on-premise solution can become the best approach;

- Cost: this aspect represents one of the key reasons that encourage organizations to move their data storage to the cloud. In fact, Cloud allows companies to avoid hardware, server rooms, IT-staff issues maintenance. Nevertheless, if transfer rate of data becomes very large, on-premise solution is likely to offer cost advantages;

- Connectivity: in a Cloud infrastructure, it's easy to connect to other cloud services. An important strength is represented by Cloud ETL tools, that allow to extract fastly a number of data from different sources in order to process and manipulating them for analytics. On the other hand, an on-premise approach allows to have a better control over security and connectivity. When this aspect can be crucial, such as in banking enviroments, on-premise solution can represents the best one;

- Speed: Since, often, the distance the data has to travel from the Cloud to the client is considerable, the resulting latency and speed has an unacceptable impact on the business. Consequently, on-premise solution becomes the best one. However, if the specific business user needs to serve multiple locations distributed in several and distant geographic positions, the Cloud is specifically designed to meet these needs. In fact it includes multiple locations for data redundancy.

Analyzing the several aspects listed above, it's noticeable that, since every business is different, there are advantages and disadvantages in both approaches. On the one hand the cloud offers scalability and low entry cost advantages. On the other, there's the security and flexibility that only an on-premise solution can offer.

2.5 Case Study

In this thesis the main goal consists in planning a scalable, flexible and modular architecture.
An important requirement concerns the possibility to ingest both structured and unstructured data in an only one storage repository. A subsequent process of data homogenization allows to realize cross data analysis.

Data hub For these reasons, the choice has fallen upon the employment of a data hub, in particular integration hub. This approach allows to storage both internal and external data of an organization, in the same place. Through a subsequent

process of data transformation, the entire data set can result uniform.

Cloud computing solution Moreover, the aiming to share pools of configurable system resources and high-lever services, triggers the adoption of the cloud computing paradigm. This solution reflects the intent of the organization, in which i worked on for the internship, to create an autonomous and centralized system that can provide resources and services for users on demand.
One of the most important strengths of cloud computing system is undoubtedly the aspect of scalability and flexibility. Scalability concerns geographical locations, hardware performance, software configurations. Whereas, flexibility requires the continuous cloud adapting to various requirements of a potentially large number of users.
Another crucial strength concerns the user-centric interfaces : users can obtain computing cloud platforms with simple methods and, installing only lightweight software components on cloud client, can access to cloud interfaces like Web service frameworks.
On the other hand, a great challenge posed by cloud applications is Quality-of-Service (QoS) management. Guaranteeing continuously a service, along with reliability, availability and performance, can provoke a problem of resources allocation.

Hybrid cloud The Cloud model promotes different deployment models:

- Public Cloud: Cloud infrastructure is available to anyone. An organization sells Cloud services.
- Private Cloud: the Cloud infrastructure is totally exploited for an organization.
- Hybrid Cloud: the entire Cloud infrastructure is composed of two internal cloud infrastructures that can be private or public, bound together by standardized or proprietary technology in order to guarantee data portability.

In this thesis, Hybrid Cloud is adopted. It consists of a centralized Data Hub layer that contains both structured and unstructured data derived from internal and external sources, and a Data Mart layer in which each Data Mart includes only a subset of the entire data set, referred to a specific business user. This type of Cloud can be considered hybrid because each Data Mart can be located entirely within the cloud or can become an On-Premise Data Mart. The Figure 2.8 shows the described infrastructure.
The first strategy, that consists in the total inclusion of the Data Mart Layer in the organization centralized Cloud, allows to manage and process data entirely into the

internal infrastructure. In this way, the clients will visualize just reports and dashboards resulting from the final process of data visualization. In this way, business users entrust the management and maintenance of data totally to the organization that offers this service.
Nevertheless, recently, many companies more and more frequently prefer to assume a relevant governance on the data in order to realize directly several analysis on data. For this reason, the second strategy provides for considering the Data Mart layer as external layer, entrusted to the client company. In this way, the client company may receive the subset of the Data Hub oriented to its specific business strategy and which it is interested in. This solution improves considerably the client governance on data, satisfying the actual need to query directly the data. In fact, it's noticeable the continuous growth in self-service BI that enables business users to work with structured data, without a background in statistical analysis and data mining.

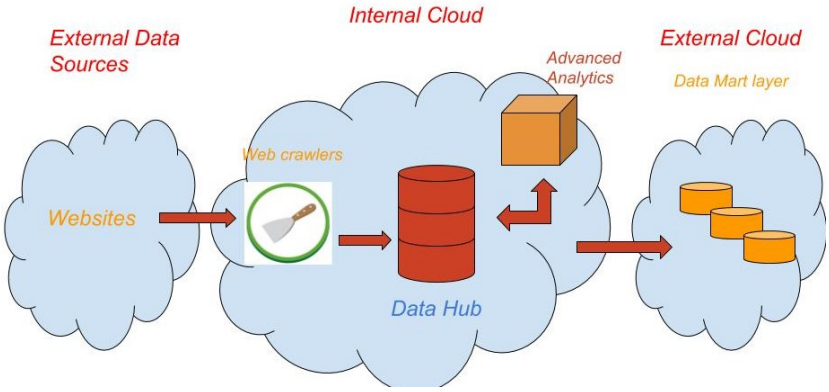


Figure 2.8. Infrastructure Design

Hadoop infrastructure The next step concerns the choice of the adequate database or file system in which to implement the data hub solution. The need to manage Big Data and, in particular, different typologies of dynamic data, in a flexible and scalable way, led to considering Hadoop infrastructure as an important solution for these goals.
Hadoop is an Apache open source framework. It allows to store a large number of data across clusters of computers and reveals different important features:

- Fast data storing and processing due to the presence of distributed computing nodes;
- Flexibility: unlike relational databases, Hadoop allows to store both structured and unstructured data like images,text and videos without preprocessing data;

- Scalability, thanks to the simple adding nodes to the distributed infrastructure;
- Fault tolerance: multiple copies of all data are stored on different nodes in order to avoid bottleneck problems.

Why Hadoop and not RDBMS? First of all, Hadoop is not a database but a distributed file system. It presents two important core components: HDFS(Hadoop Distributed File System) and MapReduce. HDFS represents the storage layer whereas MapReduce is a programming model that processes the data sets by splitting them into blocks of data, distributed across the nodes of Hadoop computer clusters.
The first reason, that leads to choose Hadoop and not RDBMS in this case study, concerns the type of data which have to stored in this data hub. They can be frequently unstructured. RDBMS is a database which is used to store structured and semi-structured data in the form of tables of rows and columns. Consequently, the varying source data nature can represent a problem. Additionally, regarding the main requirements of a Cloud Computing service as this, other crucial aspects are described as follows:

- Data Volume: RDBMS works better with a restricted volume of data (usually Gigabytes). Whereas, Hadoop improves its performance when the data set is numerous.
- Scalability: RDBMS supports a vertical scalability, implemented by adding more machines in the pool of resources. On the other hand, Hadoop realizes the horizontal scalability that allows to realize the falt tolerant, avoiding single points of failure.
- Throughput: it means the total volume of data processed in a defined period of time so that the output is maximum. Regarding throughput, Hadoop performances are considerably better than RDBMS ones.

Apache Ignite on Hadoop A remarkable weakness of Hadoop concerns the data retrieving from data set. In fact RDBMS is faster in recovering the information from the data storage, with lower access time to a particular record. Thus Hadoop is said to have low latency. Consequently, in this thesis, Apache Ignite is adopted in order to optimize access and processing time of a small set of data. Apache Ignite is a memory-centric distributed database, caching, and processing platform for transactional, analytical, and streaming workloads delivering in-memory speeds at petabyte scale. Ignite File System (IGFS) is a plug-and-play in-memory file system, compatible with Hadoop Distributed File System (HDFS) and In-Memory Map Reduce. The key feature of Apache Ignite, that allows to improving the weaknesses of the Hadoop infrastructure, consists in the fact that Ignite's durable memory component

treats RAM not just as a caching layer but as a complete fully functional storage layer.
As a result, the data does not need to be preloaded in-memory to begin processing. Consequently, Apache Ignite adoption allows to improve considerably the performance of the Hadoop infrastructure.

Data model With regards to the different phases of data storage and homogenization in the data hub, a defined data model is adopted. It consists of three different layers :

- Staging Area layer
- Data Factory layer
- Data Mart layer

Staging Area can be designed to provide many benefits, but the primary motivations for its use are to increase efficiency of ETL processes, ensure data integrity and support data quality operations.
Data are extracted from the source systems, by various methods (typically called Extraction) and are placed in Staging Area. Once in this area, data are cleansed, re-formatted and are subjected to a homogenization process, because of their various nature.

Data storage in the Data Factory layer occurs through a data refining process.
Data refining process refines varied data within a common context to improve the understanding of the data, removing data redundancy. It represents a crucial aspect of a data storage repository; in fact, unrefined data may provoke evident errors on statistical output used by business intelligence users. Whereas, through refining process, data are transformed in order to fit business rules and, only at this stage, loaded into Data Factory layer.
With regards to data modeling, data refining occurs when at the conceptual schema development, the semantics of the organization are being described. In this thesis, a snowflake schema is adopted, with the intent to make it dynamic and general as much as possible. The snowflake schema is a variation of the star schema, featuring normalization of dimension tables. In this case, data refining goes into action eliminating unnecessary things to interest, making sure that the structures to hold data are well defined. It also takes place during the database normalization, allowing to avoid data logical inconsistency and minimizing information duplication.
A snowflake schema generic can be obtained making it able to hold both static and dynamic attributes without the need to rethink its structures successively. The normalization of dimension tables represents a crucial aspect to obtain the described purpose.

Each Data Mart in the last layer, the Data Mart layer, is used to get data out to specific business users. A Data Mart contains a subset of the data hub, enabling to isolate the use, manipulation and development of each specific business user data. In this thesis, adopting an hybrid cloud approach, each Data Mart can be within the cloud or on-premise. During the data transfer from Data Factory layer to Data Mart layer, data can become further refined and aggregated in order to response to well-defined user's business needs.

Generic Data Model The Figure 2.9 illustrates the generic data model implemented in the Data Factory layer. As described above, a snowflake schema is adopted.

With regards to dimension tables, the model contains:

- a product dimension, that include the product's features, gathered from the online stores
- a store dimension, that gives information about the specific online store from which data are collected
- a date dimension, concerning the day of the data collection, and also useful information related to the specific day, such as the season to which it belongs
- a time dimension, related to the specific time of day in which data are collected

On the other hand, with regards to fact tables, the model contains two different fact tables:

- Fact eCommerce that includes both product standard pricing and discounted price, gathered from the online stores, and also the product available quantity
- Fact Sellout that contains all product sell-out data supplied by the sellers

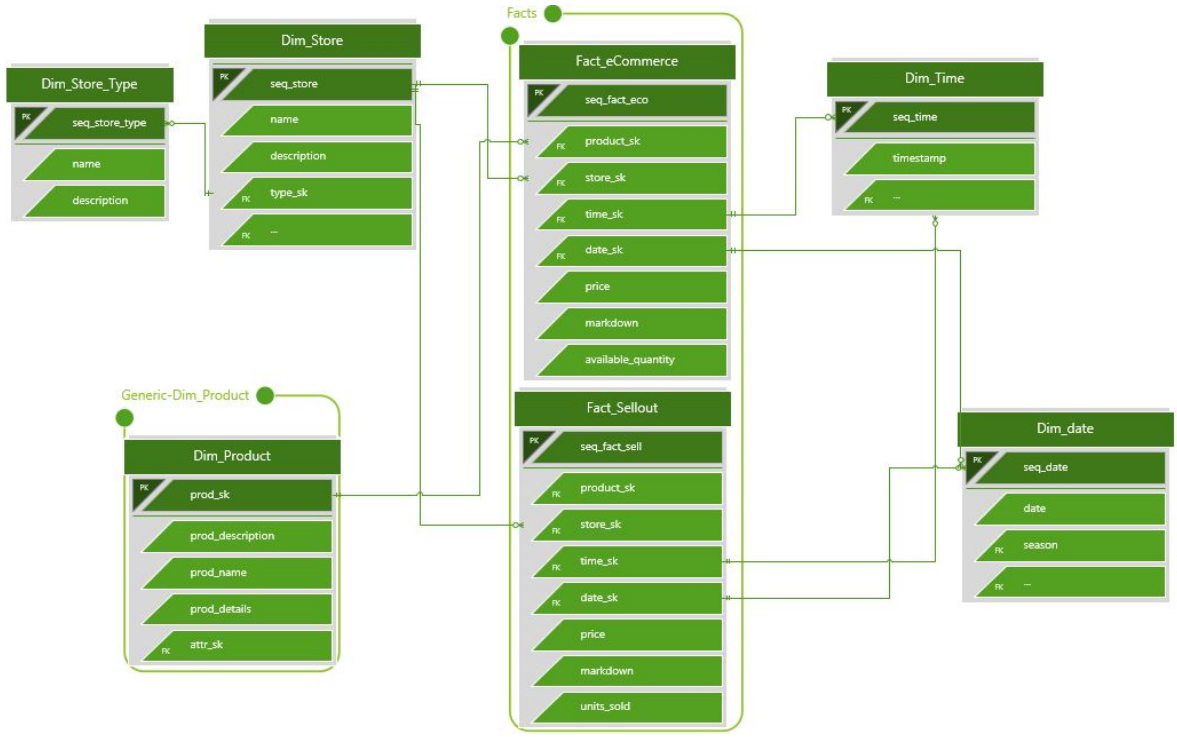


Figure 2.9. Generic Data Model

The choice of the snowflake schema plays an important role in the realization of a generic model. It is due to a specific requirement, that concerns the management of both static and dynamic attributes. In fact, this goal can be reached through the dimension normalization.

The Figure 2.10 shows the normalization process of the product dimension table. In this case study, collected data refer only to luxury shoes. Consequently, a generic product dimension table is connected to a dimension, that contains only the attributes related to the specific analyzed products, the luxury shoes. Taking into account the data model one-to-many relationship, the last dimension tables, illustrated in the hierarchy, are related to the shoes' detailed attributes, such as the colour, the material, the heel.

The described normalization highlights a crucial aspect that allows to handle different types of products in the same generic data model, without changing its structure at a later time. In fact, if data belonging to a different product's type are collected, the generic product dimension table remains equal. It will be connected also to a new different product-driven hierarchy, containing specific product's detailed attributes.

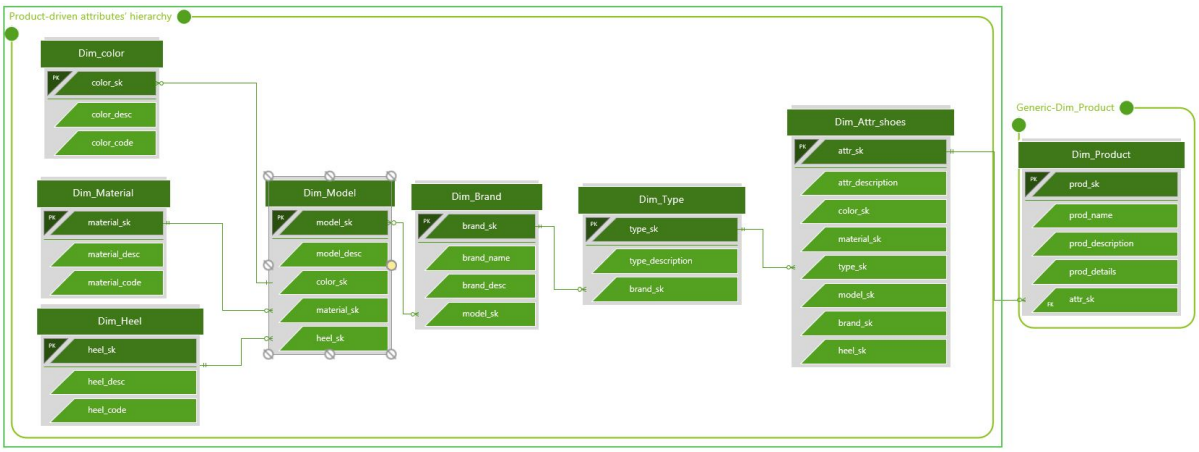


Figure 2.10. Product dimension normalization

Chapter 3

Data Ingestion and Data Integration

3.1 Web Crawling

"Web crawlers are programs that exploit the graph structure of the Web to move from page to page."[10] Data collection systems are required to support the enormous size of the Web.
Crawlers represent an optimum solution since they allow to extract automatically a large amount of content derived from the websites.
"They are one of the main components of web search engines, systems that assemble a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match the queries."[9] Web crawlers, also known as spiders, also permit to undertake the activity of web data mining, which consists in analyzing web pages in order to extract hidden knowledge. Web monitoring services clients can submit standing queries, or triggers, and the services continuously crawl the web and notify clients of pages that match those queries.
The web crawling algorithm is described below:

1. Uniform Resource Locators (URLs) set is made available for the crawler
2. The web crawler downloads all the websites contents regarding each of defined URLs
3. All the hyperlinks, that are present into the contents, are exploited by the web crawlers to collect all data contained in the pages addressed by the extracted hyperlinks,

The process continues to happen until all the linked pages and their contents are extracted. Although the described algorithm appears simple, implementing a web crawler require to deal with a lot of issues, related to HTML pages parsing, network connections, spider traps and so on.

Crawling Infrastructure

As shown in Figure 3.1, the crawler initializes frontier, that consists of unvisited URLs list, with seed URLs. These URLs can be provided by external sources, programs or users. The second step, in the crawling loop, consists in picking the next URL to crawl from frontier. The next stages concern three important operations : fetching the page corresponding to the URL through HTTP requests, parsing the retrieved page to extract the URLs, and finally adding the unvisited URLs to the frontier. Consequently, the crawling loop restarts. It finishes when the crawler has no new page to fetch.

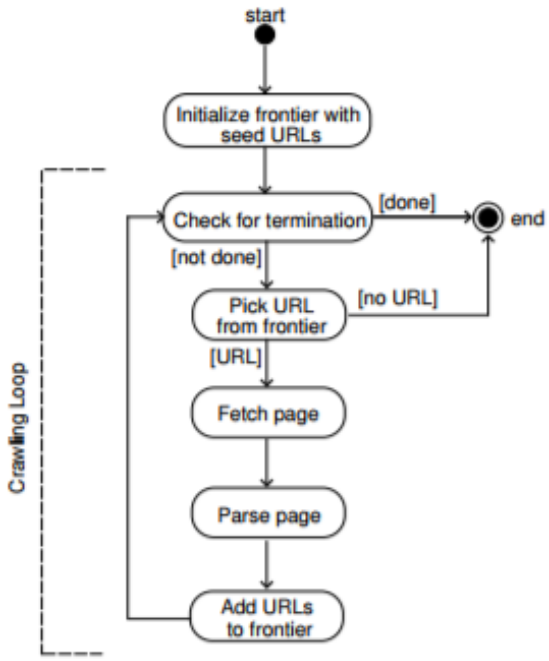


Figure 3.1. Flow of a basic sequential crawler [10]

Frontier The frontier represents the unvisited pages URLs list. It may be implemented as a FIFO queue, adding the new URLs to the tail of the queue, and extracting the next URL that needs to be crawled from the queue's head.
With regards to performance, an issue can be originated from URLs duplicates' monitoring. Two different solutions can be implemented:

1. A separate hash-table is allocated. It stores each frontier URL and must be kept synchronized with the frontier. This method allows to perform a fast lookup

2. Frontier itself represents a hash-table. Frontier URLs become keys and, consequently, the supervision consists in avoiding key duplication

Moreover, if a priority queue is adopted to implement the frontier, an estimated score of unvisited URLs keeps the queue always sorted. In fact, at each step, the best URL is picked from the head of the queue, that can be realized through a dynamic array. URLs are scored following established rules.

Fetching Fetching process consists in sending an HTTP client request in order to receive a page contents that needs to be extracted. The presence of timeout is crucial because it permits to check that an excessive amount of time is spent or server's slowness.
Moreover, the client needs to parse response headers in order to verify status code. This aspect is fundamental because a large amount of issues regarding network connection, such as server availability, can come up.
An other issue concerns the Robot Exclusion Protocol. Web servers administrators write information regarding files that may not be accessed by a crawler in a file, known as robots.txt. This policy forces a crawler to fetch the robot.txt file in the first time, in order to check if a desired URL can be fetched.

Parsing Once fetching stage is completed, page content parsing can be realized to analyze page's content, identifying also new hyperlinks to fetch. Before data extraction, it may be useful to apply text preprocessing techniques to page's content, such as stemming and stoplisting processes. In this way, all the irrelevant words, such as conjunctions, are deleted in order to simplify information extraction process.

3.1.1 Case Study

In this thesis a Web Crawler is realized in Python. With regards to the specific client's needs, its purpose consists in capturing all data from detailed pages of products sold by sellers and competitors daily. These collected data will be stored in the Data Hub successively. Capturing not only purely pricing information but also electronic shop window's details, allows to ingest data that can represent crucial business indicators for consequent marketing decisions.
Launching Web Crawler process twice a day, permits to maintain a refined granularity regarding price monitoring.

Scrapy architecture

Web Crawler is perfomed using Scrapy architecture. Scrapy is a free and open source web crawling framework, written in Python. It is used for web scraping or general purpose web crawler. The figure 3.2 shows an overview of the Scrapy architecture

with its components and an outline of the data flow that takes place inside the system.

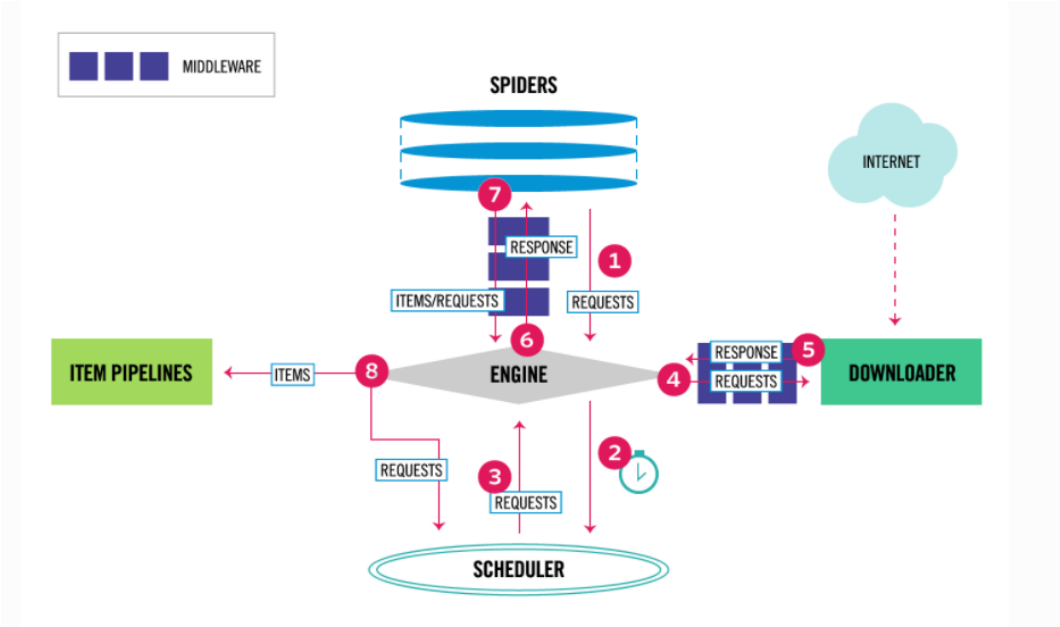


Figure 3.2. Data flow in Scrapy [1]

Spiders The most important component of Scrapy architecture is represented by Spiders. They represent the classes in which crawling logic is implemented manually. Programmer's decisions will be condition the performance of the architecture, including how to manage http requests' scalability about follow links and how to extract data from the product detailed pages.
Each Spider follows a default cycle:

1. Initial Requests to crawl the first URLs is generating, specifying a callback function to be called with the response downloaded from those requests;

2. Parsing the response (the web page content) in the callback function, and return dicts with extracted data;

3. Finally, the items returned from the spider will be typically persisted to a database or written to a file.

Scrapy Engine This component checks the data flow between all the scrapy components. It also activates events in the presence of new actions.

Scheduler The Scheluder receives requests from the engine, inserting them in a queue until the engine requests them.

Downloader The Downloader fetches web pages giving them to the engine.

Item pipeline The Item Pipeline is responsible for processing the items once they have been extracted by the spiders. Tipically, this component stores them in a database.

Downloader middlewares Downloader middlewares sits between the Engine and the Downloader. They process requests when they pass from the Engine to the Downloader, and responses that pass from Downloader to the Engine.

Spider middlewares Spider middlewares sits between the Engine and the Spiders and are able to process spider input (responses) and output (items and requests).

Data flow in Scrapy, illustrated in Figure 3.2, is explained as follows: ”

1. The Engine gets the initial Requests to crawl from the Spider;
2. The Engine schedules the Requests in the Scheduler and asks for the next Requests to crawl;
3. The Scheduler returns the next Requests to the Engine;
4. The Engine sends the Requests to the Downloader, passing through the Downloader Middlewares;
5. Once the page finishes downloading the Downloader generates a Response (with that page) and sends it to the Engine, passing through the Downloader Middlewares;
6. The Engine receives the Response from the Downloader and sends it to the Spider for processing, passing through the Spider Middleware;
7. The Spider processes the Response and returns scraped items and new Requests (to follow) to the Engine, passing through the Spider Middleware;
8. The Engine sends processed items to Item Pipelines, then send processed Requests to the Scheduler and asks for possible next Requests to crawl;
9. The process repeats (from step 1) until there are no more requests from the Scheduler.

”[1]

Splash

In this thesis, Splash module is used in order to improve web crawling performance. Splash is a javascript rendering service. It's a lightweight web browser with an HTTP API. It brings some benefits such as processing multiple web pages in parallel, turning off images or use Adblock Plus rules to make rendering faster, executing custom JavaScript in page context.
Some data are not available through simple http direct requests so that, making use of http redirection to Splash web browser, it allows to capture the entire data content from detail pages.

The Figure 3.3 shows a generic example of Web Crawler implementation related only to the Spider class in the Scrapy architecture:

```python
import scrapy
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class MySpider(CrawlSpider):
    name = 'example.com'
    allowed_domains = ['example.com']
    start_urls = ['http://www.example.com']

    rules = (
        # Extract links matching 'category.php' (but not matching 'subsection.php')
        # and follow links from them (since no callback means follow=True by default).
        Rule(LinkExtractor(allow=('category\.php', ), deny=('subsection\.php', ))),

        # Extract links matching 'item.php' and parse them with the spider's method parse_item
        Rule(LinkExtractor(allow=('item\.php', )), callback='parse_item'),
    )

    def parse_item(self, response):
        self.logger.info('Hi, this is an item page! %s', response.url)
        item = scrapy.Item()
        item['id'] = response.xpath('//td[@id="item_id"]/text()').re(r'ID: (\d+)')
        item['name'] = response.xpath('//td[@id="item_name"]/text()').extract()
        item['description'] = response.xpath('//td[@id="item_description"]/text()').extract()
        return item
```

Figure 3.3. A generic Spider implementation [1]

Online store

The Figure 3.4 shows an example of an online store. In the Scrapy architecture, the implemented spiders contain an initial list of web sites that exhibit the same structure that is illustrated in the above-mentioned Figure.
In particular, an online store shows only the most important information,such as the brand, the product name and the selling price, related to a list of products. In this case study, the products consist in luxury shoes.

Spiders need to identify all the products links that allow to send a client http request to the detail pages, in order to capture all the products itemized information.

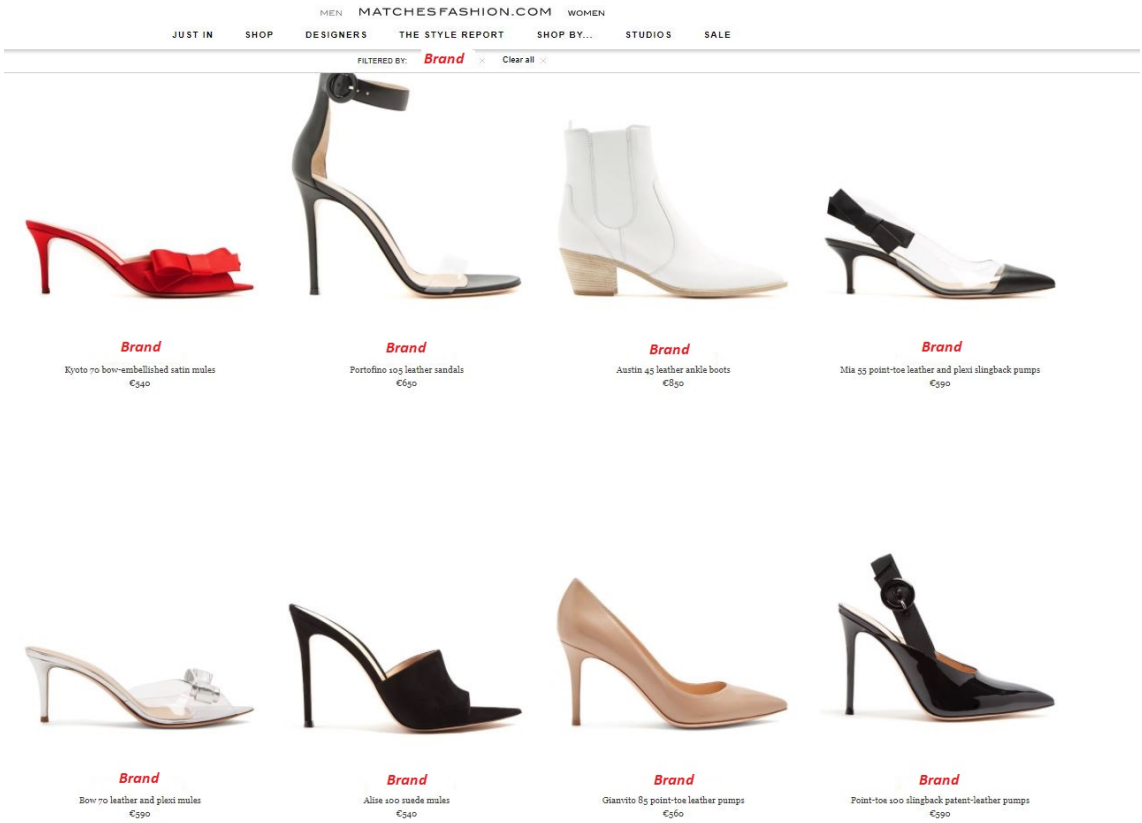


Figure 3.4. An example of an online store

For each product, showed in the online store main page, the spiders capture the link related to their detail page. Successively, they send the http request and all the detail pages data are downloaded through the Scrapy architecture.
The Figure 3.5 illustrates an example of a product's detail page. With respect to the generic information, included in the online store main page, the detail page contains also other data, such as product's details and description. These text fields play a crucial role since they include potential business key indicators, such as the product's material, its category and the heel height.

Figure 3.5. An example of a product's detail page

All the products data, collected in the Scrapy architecture, are subjected to a storage process, through the item pipeline component. They are saved directly in the centralized data hub, without altering their raw and unstructured format.

3.2 ETL

ETL (Extract, Transform and Load) is a process in data warehousing for extracting data of the source systems and placing it into a data warehouse.
The various heterogeneous sources may be represented by databases, flat files, ERP systems, CRM systems, main frame systems. Data are converted into one consolidated data warehouse format which is ready for transformation processing.
Transforming the data may involve the following tasks:

1. applying business rules, calculating new measures and dimensions
2. cleaning
3. filtering
4. splitting a column into multiple columns and vice versa
5. joining together data from multiple sources
6. applying any kind of simple or complex data validation

This manipulated and transformed data are finally loaded into the target data warehouse system or integration system.

The process is often designed from the end backwards, in that the required output is designed first. In so doing, this informs exactly what data is required from the source. A general schema of the different phases of ETL process is shown in the following Figure 3.6.

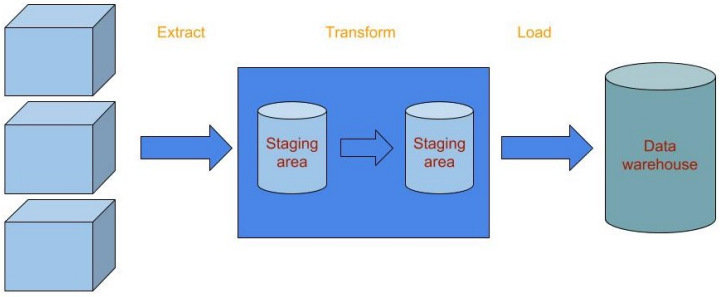


Figure 3.6. ETL process

3.2.1 Case Study

In this thesis, data captured by web crawler are subjected to a subsequent process of cleaning, filtering and data quality. The first important choice concerns the opting for a traditional ETL or ELT process. Considering the adoption of a cloud computing solution with a data hub inside, the choice depends on the requirements of this chosen system. The defined architecture must be able to ingest a large amount of different data types, through the use of Hadoop ecosystem.

ETL vs ELT

ELT(Extract, Load and Transform) consists in a different approach that provides a modern alternative to ETL.
Although data are extracted in the same way as in the ETL approach, successively, they are loaded into the target data warehouse system. Once loaded, the transformations and business logics are applied. The ELT approach leverages the power of the Relational Database Management System (RDBMS) engine.
Both the two approaches show strengths and weaknesses.
With regard to ETL approach, strengths and weaknesses are illustrated as follows.

Strengths:

1. "Designing from the output backwards ensures that only data relevant to the solution is extracted and processed, potentially reducing development, extract, and processing overhead, thus reducing the time to build the solution" [14].

2. Since data, collected in the data warehouse, have already been subject to transformation process, they are ready for their visualization.

3. "ETL can perform more complex operations in single data flow diagrams" [14].

Weaknesses:

1. Low performance regarding data transport; in fact, once data need to be transferred from data sources to ETL server and, again, from ETL server to the data repository

2. Since a dedicate server is required to perform data transformation step, consequently, hardware costs increase considerably

3. A contingent need to manipulate further data may be not satisfied efficiently since data are already filtered and transformed

Whereas, regarding ELT process approach, strengths and weaknesses are now analyzed.

Strengths:

1. Unlike ETL, ELT data transport contains a single step. This aspect permits to obtain a better network management

2. The isolation of transformation process, without exploiting a dedicated server, allows to improve the scalability

3. Flexibility regarding requirements; future transformation needs can be satisfied into the storage structure itself

Weaknesses:

1. A small volume of data may decrease ELT performances

2. Nowadays, ELT is not very common. For this reason, market still offers few ELT available tools

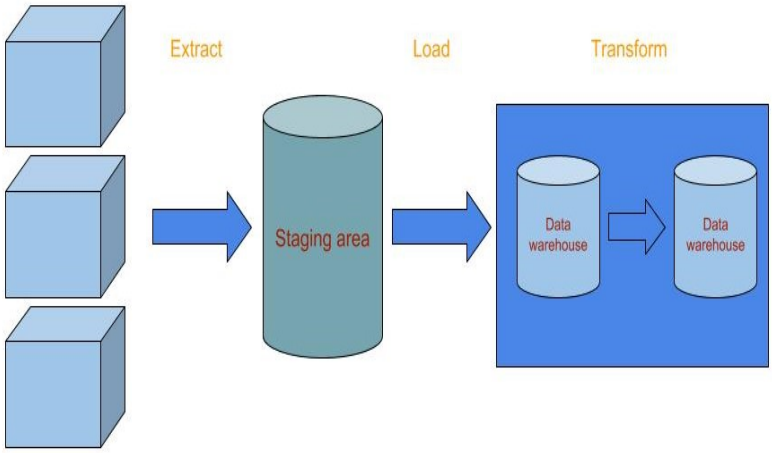


Figure 3.7. ELT process

ELT on Hadoop

It's absolutely remarkable the advent of Hadoop in a number of firms. The majority of enterprises today have one or more Hadoop cluster at various stages of maturity within their organization. Enterprises are trying to cut down on infrastructure and licensing costs by offloading storage and processing to Hadoop.
Hadoop adoption is implemented in warehouse area, since data warehouse hosts the largest amount of data in the enterprise.
Until recently, the data warehouse area has been dominated by RDBMSes and traditional ETL tools. However, the traditional ETL approach, as seen above, is limited by problems related to scalability and cost overruns.
While ETL processes have been solving data warehouse needs, Big Data require several crucial characteristics,the 10 Vs: Volume, Velocity, Variety, Variability, Veracity, Validity, Vulnerability, Volatility, Visualization, Value.
All these needs incite to move to ELT on Hadoop. With Hadoop,in fact, the ELT processes can process semi-structured and unstructured data, in addition to the benefits of cost effectiveness, scalability and flexibility in data processing environment. For these reasons, an ELT approach is adopted in this work. Data are extracted from external sources (web sites) through web crawler's process; successively, they are directly loaded into the centralized data hub and the transformation process is realized in Apache Ignite completely. Finally, structured and refined data can be stored in the Hadoop cluster.

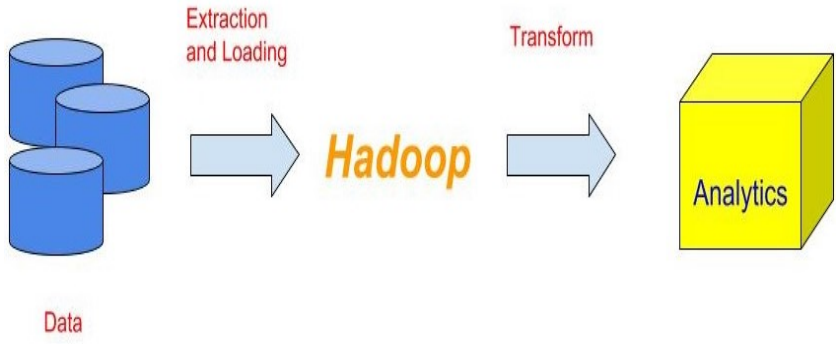


Figure 3.8. ELT process on Hadoop

ELT stages The following figures illustrate the different stages of the ELT process. The Figure 3.9 refers to the first phase of ELT process. The table ITEMS contains the unstructured data coming from the online stores; the web crawlers store them in this table daily. These data are directly transferred to the table `DLT_ITEMS`, which is inside the centralized data hub. In this step the data are not subjected to a transformation process, but they maintain the initial raw format.

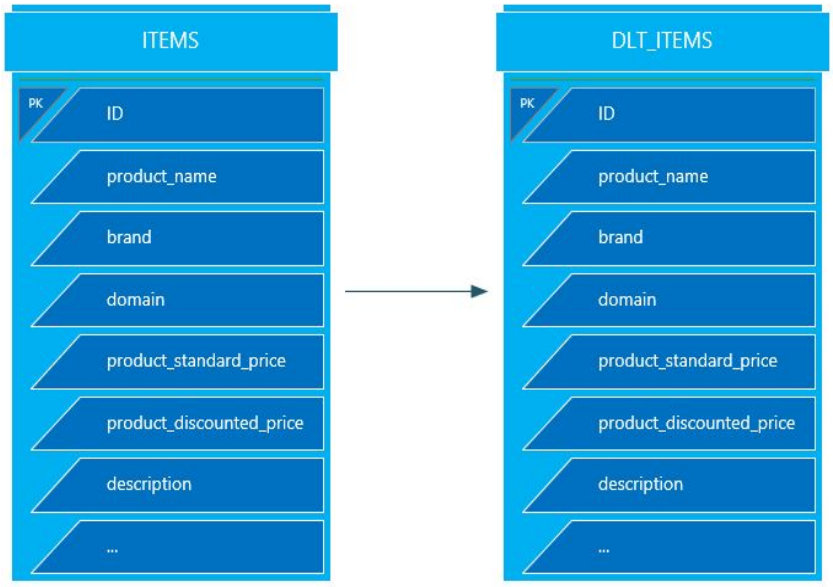


Figure 3.9. Staging Area Layer

Successively, the data undergo a laborious process of data transformation, which include different operations such as data cleaning, filtering, homogenization and refining, in order to make all the data structured and homogeneous in the Data Factory Layer. This process allows also to obtain a satisfying data integration of external data and data coming from the business client.
The Figure 3.10 displays the splitting of the data into different tables. Firstly, they are processed and transformed and, progressively, they are stored in the dimensional tables.
As can be seen in the Figure 3.11, the business client provides the Sell-Out data to the company. Consequently, the necessary and implemented data integration allows to achieve more interesting analysis, extracting more complex business information.

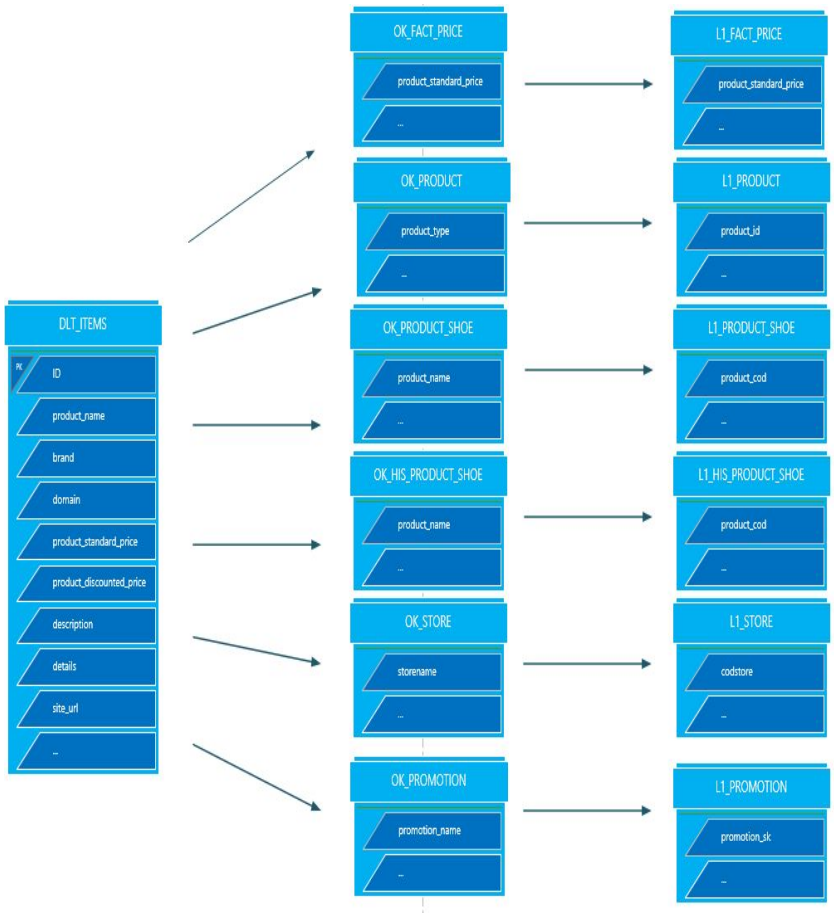


Figure 3.10. Data Factory Layer

Figure 3.11. Client Sell-Out Data

In the final stage of the ELT process, the structured data are further refined, through also a specific data quality process. In this way, they become adequate to be transferred to the Data Mart Layer. Data visualization tool will base its reports, graphs and dashboards on the Data Mart Layer data. The Figure 3.12 shows the data transfer from the Data Factory Layer to the Data Mart Layer.

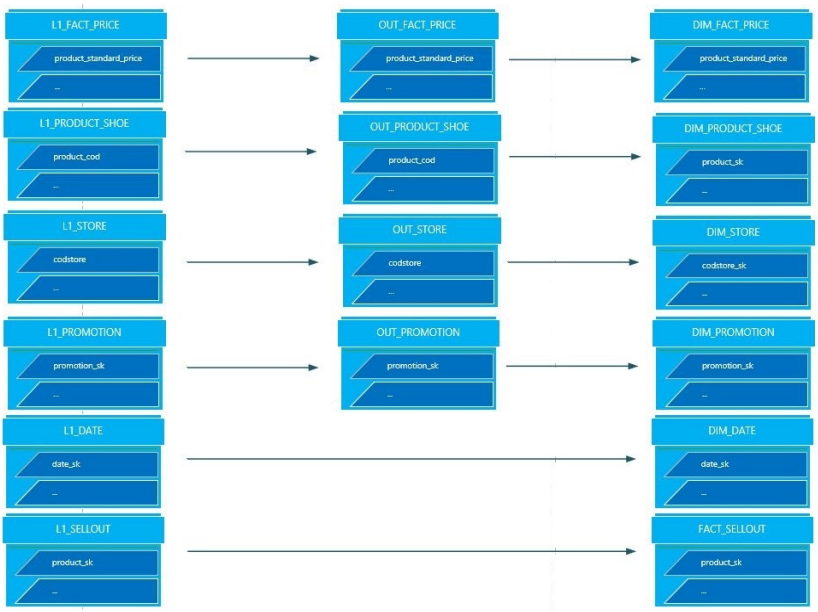


Figure 3.12. Data Mart Layer

The utilized platform for the ELT process is Oracle Data Integrator (ODI). It is a heterogeneous Big Data Integration technology based on an open and lightweight ELT architecture.

One of the most important ODI components is represented by the Mapping. Mappings consist in the logical and physical organization of the data sources, targets, and the transformations through which the data flows from source to target.
The Figures 3.13 and 3.14 illustrate two example of mappings implemented in this work.

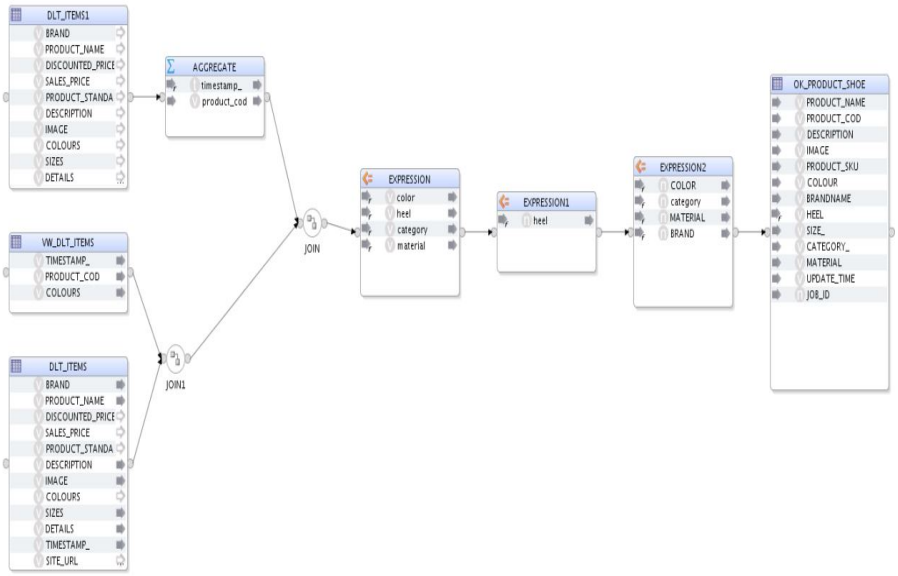


Figure 3.13. ODI mapping

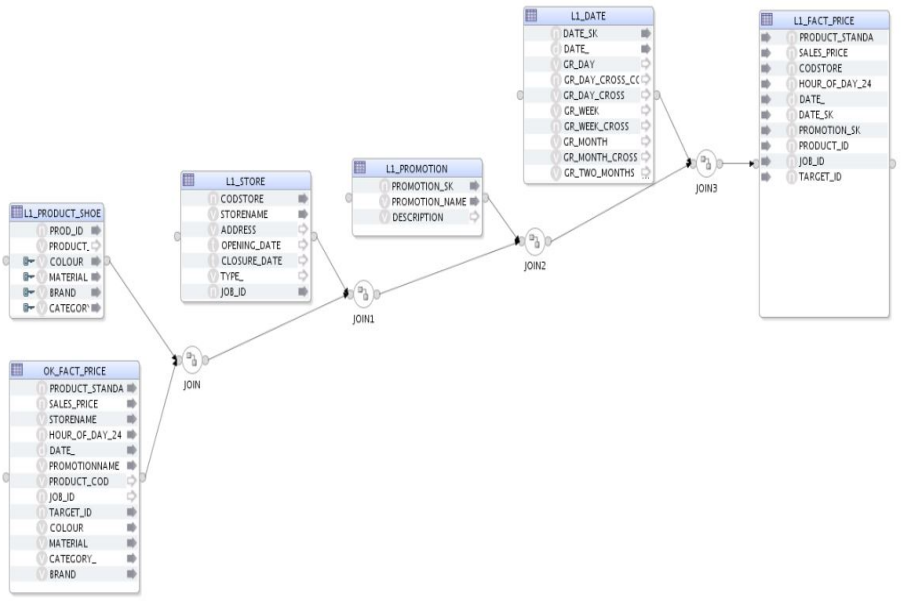


Figure 3.14. ODI mapping

Data homogenization A data hub differs from other storage repositories, such as data lake, by homogenizing data, rather than storing them in one place. In this thesis, a data homogenization work is implemented in order to store all data, coming from different external sources, in the same dimension and fact tables belonging to a generic and dynamic model. Homogenization process difficulty derives from the fact that the initial data are unstructured and data extracted from different online stores are so different among them. Consequently, different transformation implementations are required in order to make different data structured in the same final format,allowing their storage in the same tables.

Data quality Undoubtedly, one of the most laborious procedures treated and implemented in this thesis is represented by the data quality process. It's remarkable that, as data volume increases, the question of internal data consistency becomes crucial.
Data cleaning and filtering are essential in order to ensure data quality. Once data are refined and cleaned, it's necessary that quality requirements will be fulfilled. In this thesis work, a detailed data profiling activity is performed. Data, once they are available in the data hub, are under review in order to discover anomalies, inconsistencies, missing or incorrect values. In this way, the accuracy and the precision can be improved, optimizing the management of unknown values and outliers.
Data quality process assumes a critical role in the entire chain. Since captured data are coming from external web sources, the probability to find inconsistencies

in the data is very high. Without a constant and accurate data quality process, data containing these anomalies will be transferred to data mart layer and they can cause disaster in subsequent data visualization phase. In fact, right in that step, all the specialist and non technical customer will be able to find missing or incoherent values on the reports and dashboards.

With regards to the specific case study, an automated controller is implemented. It's in charge of the data quality dedicated control. In particular, its tasks consist in:

- Monitoring daily that web sites continue to produce data and checking the web crawling action. In fact, sometimes web sites pages are subjected to some changes. Consequently, web crawlers need to be modified in order to continue the data capture related to the specific modified web site
- Supervising the percentage of not classified attributes. If the percentage exceeds the 3 percent, it means that attribute dictionaries need to be improved and optimized.
- Controlling constantly attribute's values: if they reveal themselves meaningless, it means that they represent incorrect value or outliers that need to be managed

Chapter 4

Data mining

Data mining consists in "Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data"[8]. It is an interdisciplinary subfield of computer science. The main goal of data mining is represented by the extraction of information from data.
With the explosive growth of data, traditional data analysis algorithms should be adequately scalable to handle such as tera-bytes of data. Moreover, data show an high complexity because of their varied nature, requiring new and sophisticated algorithms to manage them.
Data mining is divided in two main analysis techniques:

- descriptive data mining: Extract interpretable models describing data
- predictive data mining: Exploit some known variables to predict unknown or future values of (other) variables

As shown in Figure 4.1, data mining represents a confluence of multiple disciplines. This aspect allows to develop several data mining functions and represents a noticeable strength. It becomes necessary in order to handle several kinds of data:

- sensor data
- social network, graph and unstructured data
- time-series data, temporal data
- multimedia data
- textual data
- the World-Wide Web

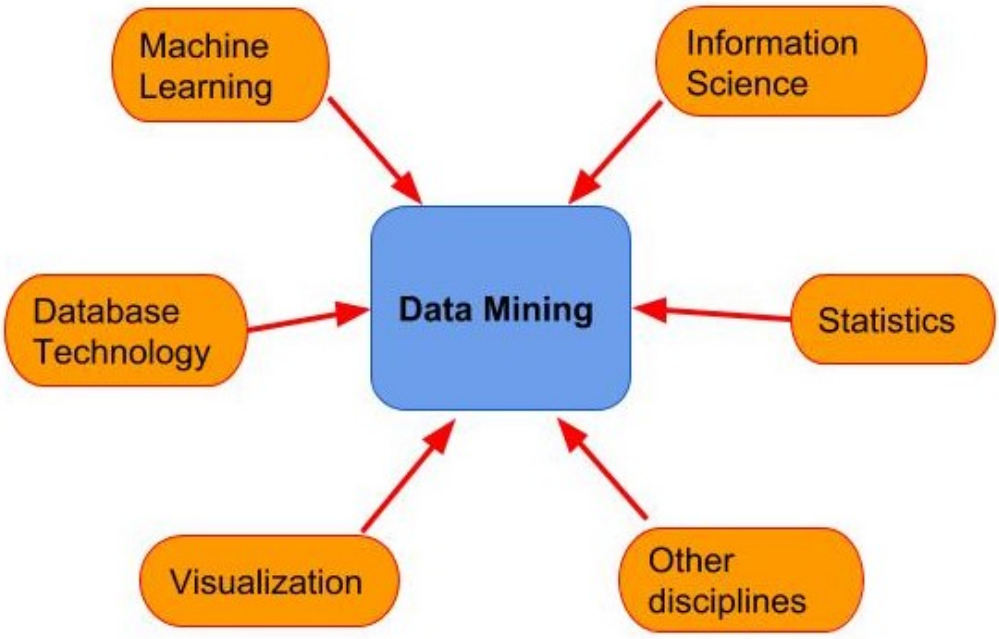


Figure 4.1. Confluence of multiple disciplines

The Knowledge Discovery from Data (KDD) process, shown in Figure 4.2 is commonly defined with the following stages:

1. Selection
2. Pre-processing
3. Transformation
4. Data mining
5. Interpretation/evaluation

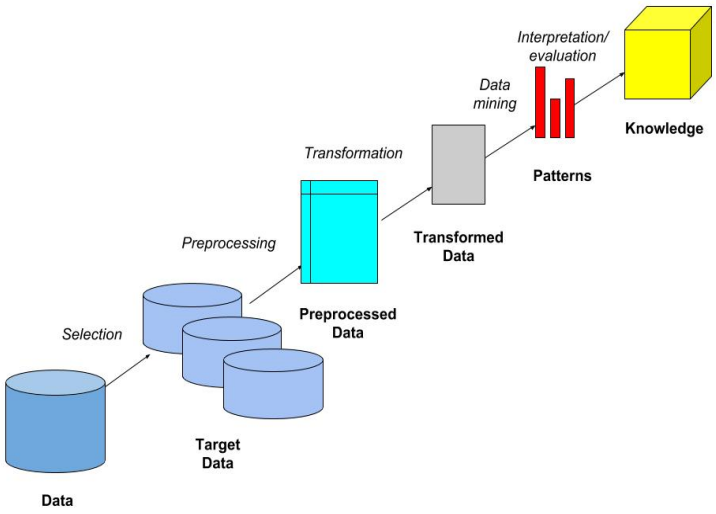


Figure 4.2. KDD process

Data mining involves some common classes of tasks:

- Anomaly detection: identification of data errors or unusual data that need further evaluation
- Association rules: finding relationship between variables
- Classification: a process whose objectives consist in the prediction of a class label and definition of an interpretable model of a given phenomenon
- Clustering: the process of discovering groups "similar" in the data
- Regression: attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets
- Summarization: providing a more compact representation of the data set
- Sequence mining: ordering criteria on analyzed data are taken into account
- Time series and geospatial data: temporal and spatial information are considered

4.1 Text mining

Text mining is the process of deriving high-quality information from text. The expression "high-quality information" refers to the recognition of patterns and trends. Text mining process employs Data Mining techniques to unstructured textual data and, more generally, to any kind of document in order to:

- identify the main thematic groups
- classify the documents in default categories
- find hidden connections
- perform sentiment analysis

Text mining is roughly equivalent to text analytics. Text analytics is the way to unlock the meaning from several unstructured textual data types. It describes a set of linguistic, statistical, and machine learning techniques that allow to reveal customer's needs and wants.
For these reasons, "Text mining plays a significant role in business intelligence that help organizations and enterprises to analyze their customers and competitors to take better decisions."[13] It allows to improve the customer satisfaction an to optimize customer chain management system.

Case study

The work realized in this thesis consists in the study of text content mining for E-Commerce Web Sites. In this digital age, a lot of people have shifted from off line buying to online buying. Online buying helps in understanding user behavior like what are the likes and dislikes of the user, their buying behavior, and many such applications.
Usually, data coming from e-commerce online stores are textual. Consequently, their analysis requires to exploit text mining techniques.
In the first phase, the work focused on finding keyword that can reveal themselves as crucial business indicators. In fact, with regards to analyzed fashion products, in particular luxury shoes, textual data such as brand, category, colours, materials can influence considerably marketing strategies and pricing changes. The extraction of these keywords requires a preliminary laborious text preprocessing. "Preprocessing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process." [4]
In this thesis, the three key steps of preprocessing stage are:

- Regular expressions

- Stop Words elimination
- Stemming

Regular expressions Regular expression represents one of the most important successes in standardization in computer science. It is a language for specifying text search strings. It is used in every computer language, word processor, and text processing tools like the Unix tools grep or Emacs.
Regular expression method may represent one of the best solutions in presence of a pattern to search for and a corpus of texts to search through. In the specific case study, the corpus can be the product name, the details or the product description. In this way, it's possible to extract the most relevant words from the mentioned fields. Additionally, regular expression is applied also to pricing values coming from the different web sites, in order to homogenize them in a unique and uniform format.

Stop Words elimination Stop Words are words which are filtered out before or after processing of natural language data (text). Since they represent a division of natural language, their elimination allows to reduce the dimensionality of term space, making the text look less heavy and more significant for analysts.
Stop words are represented by terms that don't give further documents to the text, such as pro-nouns, prepositions, articles and so on. This process improves the possibility to find keyword in the filtered text.

Stemming Stemming represents the process of reducing inflected words to their word stem, base or root form. "The purpose of this method is to remove various suffixes, to reduce the number of words, to have accurately matching stems, to save time and memory space."[4]
In the case study, stemming is applied to the business keywords that needed to be extracted. For instance, considering a colour, only the singular form is extracted, in order to enable a subsequent match identification between colours of different products.

A subsequent text normalization and text language unification are implemented in order to make possible the extraction and analysis of the most important product features.

4.2 Text Clustering

Cluster analysis consists in "finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups." [5]
Then, clustering is a set of clusters. The main intent of cluster analysis consists in minimizing intra-cluster distances and maximizing inter-cluster distances. There are two main types of clustering:

- Hierarchical clustering: the clusters are organized in a hierarchical tree
- Partitional clustering: it's composed of non-overlapping subsets

With regards to the clustering algorithms, the most important are:

- K-Means and its variants
- Hierarchical clustering
- Density-based clustering

Text clustering is the application of cluster analysis to textual data. It can be used for multiple purposes, such as topic extraction, grouping similar documents, discovering hidden and implicit information in documents. The application of text clustering can be implemented online or offline. The difference between the two approaches concerns mainly efficiency and performance.
Usually, text clustering requires preparatory text mining steps such as tokenization (parsing text data into smaller units, tokens, such as words and phrases), stemming and lemmatization, stop words elimination.

Case study

In this thesis text clustering is used, involving the employment of descriptors and the descriptor extraction. Descriptors consist in a set of words that represent the content within a cluster.
Its employment stems from the concern to be able to identify the same product, belonging to the same brand, sold on the different web sites. In fact, usually, different detail pages of different online stores can show the same product with different product names. Consequently, without an appropriate matching algorithm, the same product with more product names will turn out to be as two different products. Since products will be subjected to business analysis influencing marketing decisions, this evident approximation can't be accepted.

Then, in order to solve the matching issue, text clustering approach can represent a good solution in order to solve the matching issue.

Once text preprocessing is completed, keywords and business indicators can be extracted from the textual data. Text clustering implementation consists in the extraction of common keywords, such as colour, brand, material and category, from all the products identifying every combination of these keywords as the content within a particular cluster. Therefore, in doing so, even if a product shows different name on distinct online stores, exhibiting the same important features, it will belong to the same cluster. The Figure 4.3 shows the content within a single cluster.

Figure 4.3. Product clustering

This matching solution, realized through a text clustering approach, allows to realize advanced analysis on products. In fact, it's remarkable that crucial analysis, such as pricing comparison related to the same product on different web sites, becomes feasible improving considerably the efficiency and the validity of the entire business process.

4.3 Forecasting

Forecasting represents the process of making prediction of the future, relying mainly on data from the past and present and analysis of trends. Usually, it is used for the estimation of some variable of interest at some specified future date. Obviously, variables such as risk and uncertainty assume a centralized role in forecast analysis. The accuracy is one of the most important parameters that distinguish the quality of a forecasting algorithm.
Forecasting methods belong to two main different categories:

- Qualitative methods
- Quantitative methods

Qualitative forecasting methods are based on certain assumptions based on the management's experience, knowledge, and judgment. These estimates are projected into the coming months or years using one or more techniques such as Delphi method, exponential smoothing, regression analysis, trend projection, Box-Jenkins models, moving averages. These type of methods are appropriate when past data are not available and, consequently, key trends and developments are hard to capture.

On the other hand, quantitative forecasting methods use historical data to forecast the future data. In contrast to qualitative methods, these ones are applied to shorter time periods. They exploit statistical models such as ARIMA model, Single and Double Exponential Smooth, back-propagation neural network.

Case study: Price Forecasting Models

In this thesis forecasting algorithms are used in order to make temporal predictions regarding products pricing. Capturing daily pricing values from the online stores, along with available sell-out data concerning the same products sold by sellers, allows to exploit quantitative forecasting algorithms, without applying only analysis based on knowledge and intuition.
It's considerably important to choose the correct parameters as input to forecasting algorithms. In fact, considering other key business indicators and not only pricing values, can represent a crucial factor with regard to the accuracy of the predictions.

Price prediction of e-commerce products concerns time series forecasting. "Time series forecasting is an important area of forecasting in which past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship. The model is then used to extrapolate the time series into the future." [15]

In this work, two different approaches to time series forecasting are proposed: ARIMA and artificial neural network methods. It's difficult in practice to choose the right method for many reasons.
First of all, real-world time series are rarely pure linear or nonlinear. They often show both linear and nonlinear patterns. Moreover, it's also complicated to identify whether a time series is produced from a linear or nonlinear underlying process. Since real-world case studies show several considerable complexities, any single method may not be able to capture patterns completely. Consequently, using an hybrid approach, combining several forecasting methods, represents currently a common practice to improve the forecasting accuracy.

ARIMA Model

ARIMA (Autoregressive integrated moving average) model calculates the future value of a variable as linear function of several past observations and random errors. The process that generates the time series has the following form:

$$x_t = \boldsymbol{\theta}_0 + \boldsymbol{\phi}_1 x_{t-1} + \boldsymbol{\phi}_2 x_{t-2} + ... + \boldsymbol{\phi}_n x_{t-n} + \boldsymbol{\epsilon}_t - \boldsymbol{\theta}_1 \boldsymbol{\epsilon}_{t-1} - \boldsymbol{\theta}_2 \boldsymbol{\epsilon}_{t-2} - ... - \boldsymbol{\theta}_m \boldsymbol{\epsilon}_{t-m} \quad (4.1)$$

where x_t and $\boldsymbol{\epsilon}_t$ represent the actual value and random error at time period t, respectively; $\boldsymbol{\phi}_i (i = 1,2, ...n)$,$\boldsymbol{\theta}_j$ (j=1,2,...m) represent model parameters, n and m are integers corresponding to the orders of the model. Random errors $\boldsymbol{\epsilon}_t$ are assumed to be independently and identically distributed with a mean of zero and a constant variance of $\boldsymbol{\sigma}^2$.

Building ARIMA model means especially model order (p,d,q) determination. When two out of the three terms are zeros, the model may be referred to based on the non-zero parameter, dropping "AR", "I" or "MA" from the acronym describing the model. For example, ARIMA (1,0,0) is AR(1), ARIMA(0,1,0) is I(1), and ARIMA(0,0,1) is MA(1).

"Stationarity is a necessary condition in building an ARIMA model that is useful for forecasting." [15] This property entails that the mean and the autocorrelation remain constant over time. In fact if a time series is generated from an ARIMA process, it should have some theoretical autocorrelation properties.

Results As described above, ARIMA model results depend certainly on the choice of the three non-negative integers parameters (p,d,q). The parameter p represents the number of time lags, d is the degree of differencing and q is the order of the moving-average model.

The parameter d assumes a crucial role to obtain good results. It can be represented only through the range of values 0,1,2. The Figure 4.4 and the Figure 4.5 highlight how the choice of the d influence considerably the final result. The only difference between the two figures concerns the value of the parameter d. It's remarkable that d=2 produces better results, growing noticeably the accuracy of the forecasting algorithm.

In particular, in the two figures, the horizontal axis represents the temporal axis (days), whereas the vertical axis represents the range in which a specific product has been sold on the online stores by the sellers.
The blue line is related to the real price development. On the other hand, the red line refers to the predicted price development.

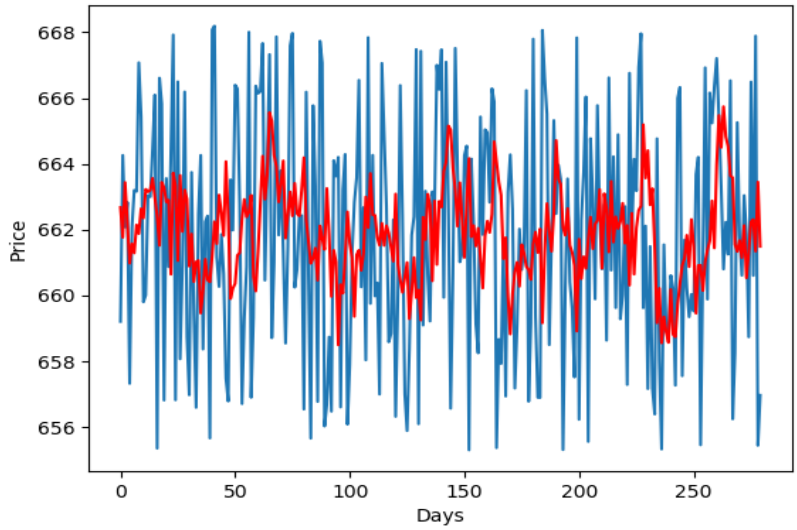


Figure 4.4. ARIMA (5,1,0) example

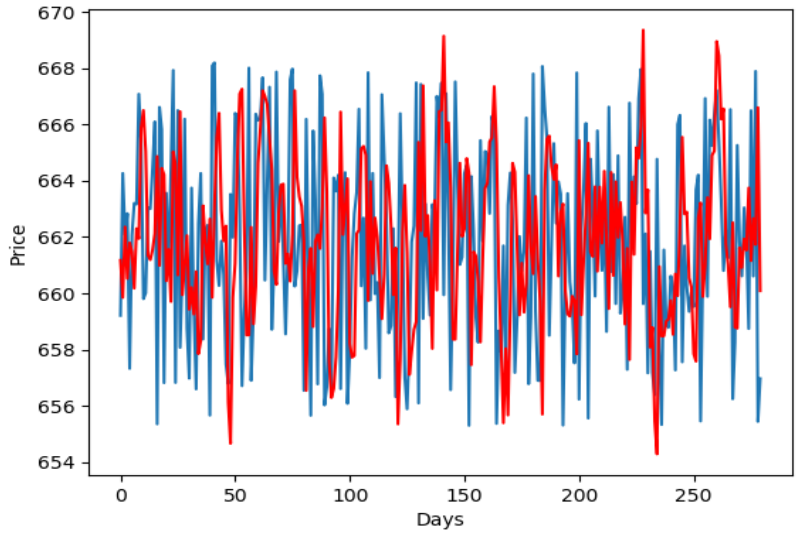


Figure 4.5. ARIMA (5,2,0) example

The Figure 4.6 concerns the same specific product analyzed in the just described figures, changing the p parameter. It's noticeable that the p value alteration doesn't influence considerably the final result. However, taking into account a lower number of time lags, the final accuracy decreases.

The Figure 4.7 and the Figure 4.8 illustrate two different ARIMA model results, taking into account a different product, with respect to the previous figures. An interesting aspect concerns the fact that, also in this case study, the value of the d component plays a crucial role in order to forecast price evolution, obtaining an important accuracy with d=2.

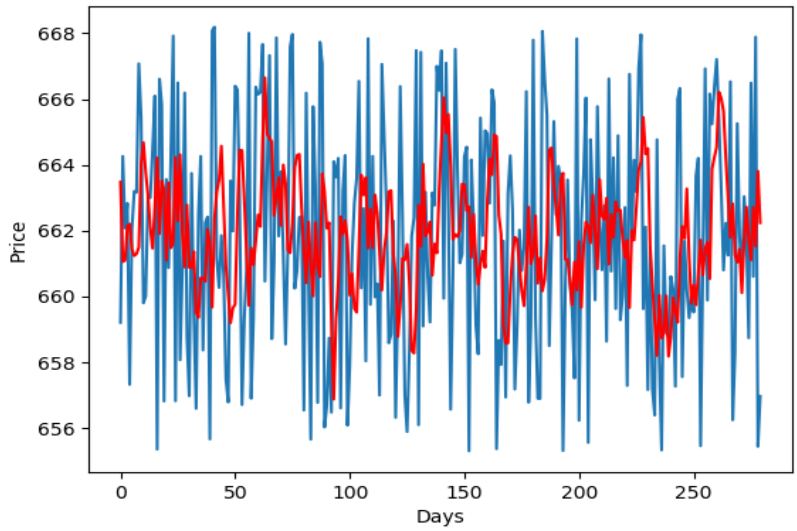


Figure 4.6. ARIMA (3,1,0) example

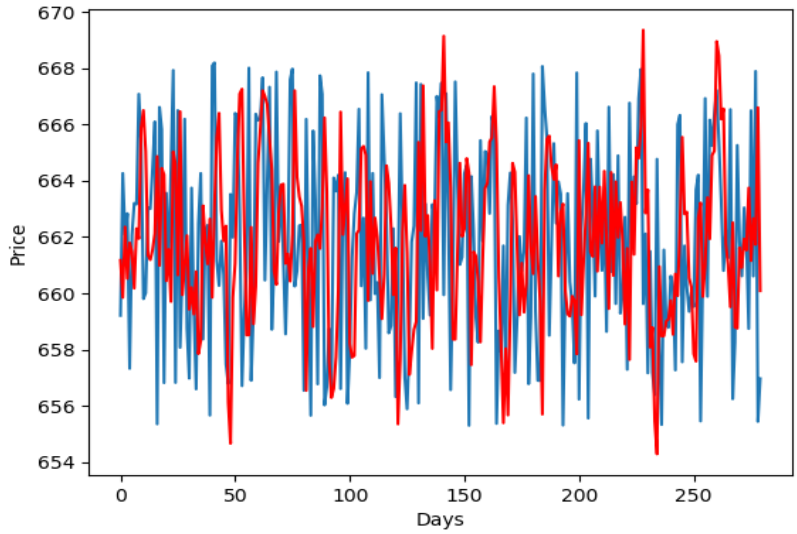


Figure 4.7. ARIMA (5,1,0) example

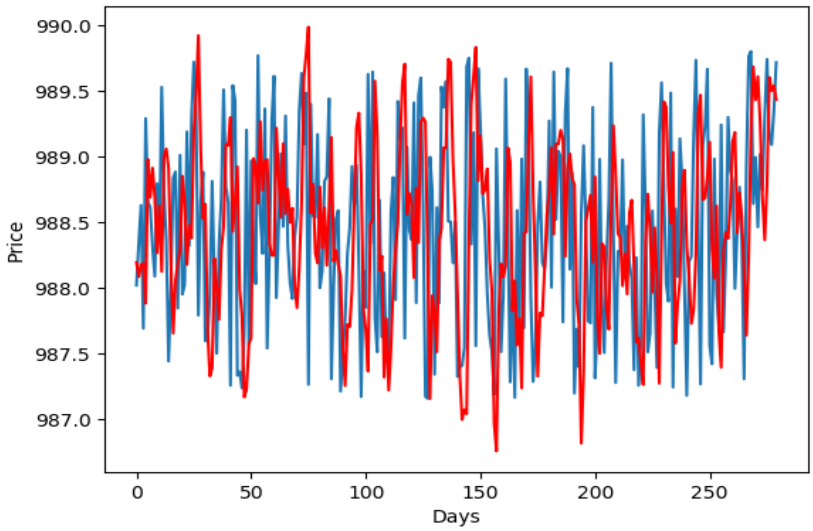


Figure 4.8. ARIMA (5,2,0) example

Artificial Neural Network Forecasting

Time series are often deeply nonlinear. Artificial neural networks are a good nonlinear method and they are able to approximate several nonlinearities in the data. Unlike ARIMA model, artificial neural network don't require prior assumption of the model form. Their model is determined by the characteristics of the data allowing to obtain a high degree of accuracy.
The model shows three o more layers of units connected by acyclic links.

$$x_t = \boldsymbol{\alpha}_0 + \sum_{j=1}^{q} \boldsymbol{\alpha}_j g(\boldsymbol{\beta}_{0j} + \sum_{i=1}^{p} \boldsymbol{\beta}_{ij} x_{t-i}) + \boldsymbol{\epsilon}_t \tag{4.2}$$

where $\boldsymbol{\alpha}_j$ (j=0,1,2,...,q) and $\boldsymbol{\beta}_{ij}$ (i=0,1,2,...p ; j=0,1,2,...q) represent the model parameters, called connection weights. p is the number of input nodes and q is the number of hidden nodes.
The equation 4.2 can be seen as a nonlinear autoregressive model:

$$x_t = f(x_{t-1}, x_{t-2}, ..., x_{t-p}, w) + \boldsymbol{\epsilon}_t \tag{4.3}$$

where w represents a vector containing all the parameters, whereas f is a function that depends on the network structure, the number of layers and the connection weights.
Building an artificial neural network model consists in specify p and q parameters.

In particular, whereas q is data dependent its value doesn't follow any a priori rule, p represents the most important parameter of the model, determining the nonlinear autocorrelation structure of the time series.

Compared to ARIMA model, artificial neural network model may show more likely the problem of overfitting, because ARIMA model is pre-specified, whereas artificial neural network model is determined from the data.
An other crucial difference between the two models concerns the model evaluation. In ARIMA model, model evaluation is realized exploiting the same sample that is used for model identification and estimation. On the other hand, in artificial neural network model, a separate hold-out sample, that is not exposed to the traning process, is used for the model evaluation.
Both the models require a preliminary data trasformation in order to obtain good results and iterative experiments.

Long Short-Term Memory Networks

In this case study, a specific class of the artificial neural network is exploited, the recurrent neural networks (RNNs). RNNs are called recurrent because they perform the same computations for all elements in a sequence of inputs. RNN is particularly appropriate for time series forecasting. In fact, unlike other different class of the artificial neural network, such as the feedforward neural , it can exploit internal memory to process sequence of inputs. Consequently, this important feature allows to show a dynamic temporal behavior for a time sequence.

Long Short-Term Memory (LSTM) Networks consist in a recurrent neural network, composed of LSTM units. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell is responsible for values memorization over arbitrary time intervals.
It's interesting the expression long short-term memory. It is related to a model that provides a short-term memory, that is able to hold only a small amount of information, for a long period of time.
For all these reasons, LSTM Networks show optimal features for time series prediction.

Results Both Basic RNN and LSTM Network are taking into account in order to highlight the comparison between the two models' results.

The most important parameters are:

- Epochs that represent the number of times all of the training vectors are used once to update the weights of the neural network

- Batch size, the number of training data in one forward/backward pass. The higher the batch size, the more memory is required
- Number of iterations that consists in the number of passes, each pass using a number of training data equal to the batch size

One of the problems that occur during neural network training is called overfitting. The network memorizes the training examples but it's not able to generalize to new situations. Consequently, the error on the training set is considerably reduced but it becomes remarkable when new test set is presented to the neural network. The choice of the right number of epochs is important. In fact, too much epochs, that consist in training too much the network, can trigger the overfitting. On the other hand, a small number of epochs can provoke the opposite problem, called underfitting.

The following figures illustrate the results of the model's implementation. In particular, the Figure 4.9 and the Figure 4.10 show two different results, employing the same number of epochs (1000), coming from the application of both Basic RNN and LSTM Network.

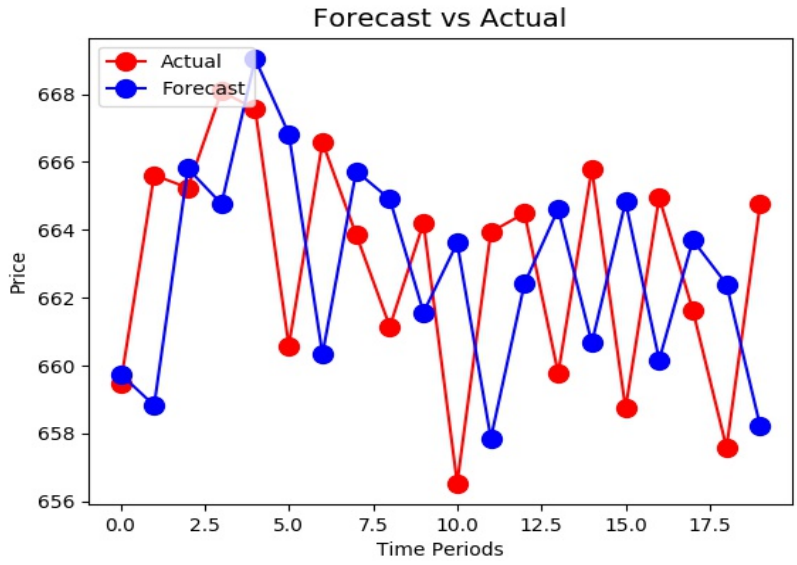


Figure 4.9. Basic RNN example, *epochs*=1000

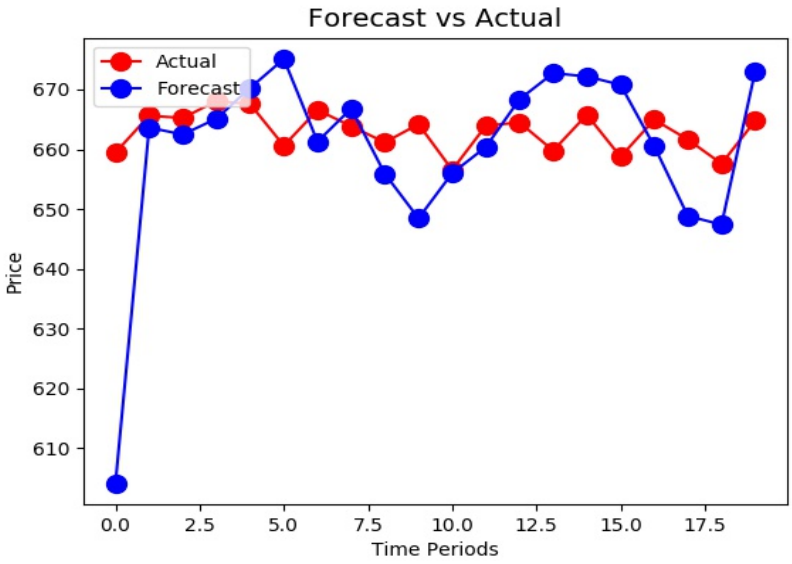


Figure 4.10. LSTM Networks, *epochs*=1000

The following figures exhibit the problems called overfitting and underfitting. Both Figures 4.11 and 4.12 refer to the LSTM Networks, but they differ with regard to the parameter *epochs*.

With respect to the previous Figure 4.10, in which *epochs*=1000, a small number of epochs (300) and a larger number of epochs (3000) are used in order to observe the behavior of the neural network.
As can be seen in the two Figures 4.11 and 4.12, the neural network is less sensitive to small price changes, compared to the Figure 4.10.

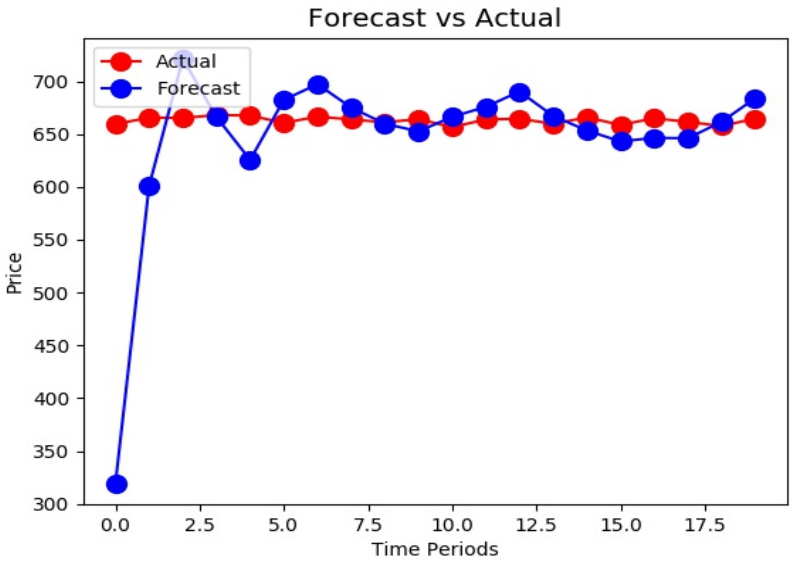


Figure 4.11. LSTM Networks, *epochs*=300

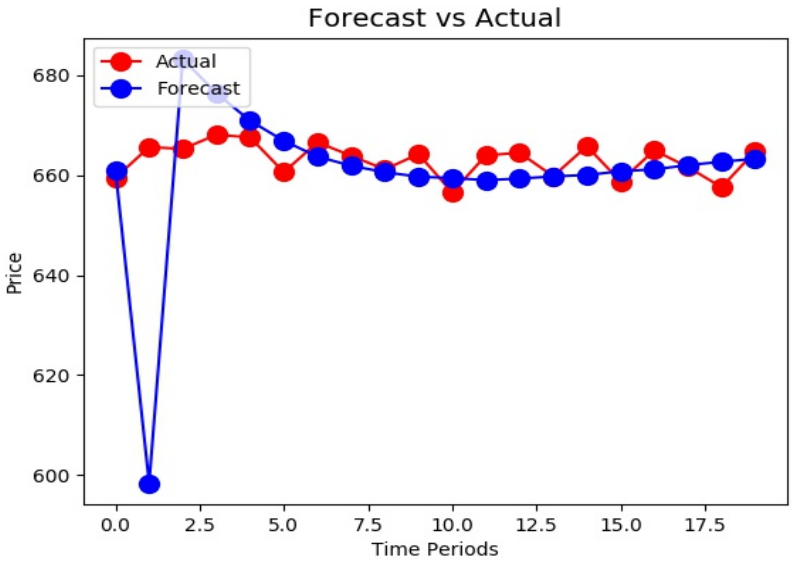


Figure 4.12. LSTM Networks, *epochs*=3000

Chapter 5

Data Visualization and Demand Forecasting

Data Visualization involves the study and the creation of the visual representation of data. Its main purpose consists in transmitting the information extracted from data clearly and efficiently through the employment of plots, graphics, reports and so on.
Data Visualization represents both an art and a science. As discussed extensively in the chapter *Advanced Analytics*, it is repeatedly subjected to changes and developments. It plays a crucial role in the large process of data analysis which includes the main phases of data collection, data storage, data processing and data visualization. The importance of Data Visualization is due to its task to present data to the final business users. The way in which information is presented can influencing considerably the user's final opinion and marketing decisions. As John Tukey affirmed, "The greatest value of a picture is when it forces us to notice what we never expected to see".

The remarkable growth of self-service business intelligence (BI) highlights user's needs to increase the data governance and the data discovery. Nowadays, business users want to access and work with corporate data even though they do not have a background in statistical analysis, business intelligence or data mining.

In this case study, Microsoft Power BI is used for data visualization. It represents a business analytics service, a new visual data exploration and interactive reporting tool. It allows to generate reports and dashboards related to both data residing on-premises and data that are located in the cloud.
One of the most important strengths of Power BI concerns the fact that it takes on an hybrid position between traditional BIs and self-service BIs. In fact, it allows to create data model in order to create reports through queries based on the realized data model. Additionally, it also permits to have access to all the data instead of

just the limited result sets returned by queries.
This described Power BI feature is considerably decisive. The platform, which is implemented in this thesis project, provides for both on-premises data marts and data marts entirely included in the cloud. Consequently, exploiting an hybrid tool, that allows both traditional approach and innovative data governance needs derived from on-premises strategies, may represent an optimum solution.

Demand Forecasting

As described above, the role of business users is subjected to constant developments in the business entities.
Over time, companies are gradually moving the focus from their products to the user's necessities. They need to have a clear understanding of the client's needs and demand curves in their given market. For these reasons, demand forecasting is increasingly becoming a fundamental component in the complex business processes and in the relationship between the companies and final users.
Demand forecasting is the art and science of forecasting customer demand. It includes both informal methods, such as judgments and intuitions, and quantitative methods. It consists in predicting future demand for the product on the basis of the past events and prevailing trends in the present.

It's remarkable the existing connection between the data visualization application and the demand forecasting. Business intelligence solutions, such as reports and dashboards, can be used to track and analyze forecasts in order to better understand and improve demand planning accuracy.

With regards to the specific case study, concerning online fashion sales industry, one of the most important challenges regards the best strategy to adopt between *Planning and Allocation* and *Continuous Replenishment*.
Planning and Allocation represents the process of setting and maintaining future performance goals for different aspects such as sales, inventory, financial metrics and so on. Planning choices are based on historical trends, management insights,promotional events, business strategy shifts.
On the other hand, *Continuous Replenishment* entails the constant stock and sales data exchange between the distributor and the retailers. This approach allows to improve the supply chain efficiency and to reduce Forrester effect. This effect is due to an excessive provisions level, continuous sales prediction inefficiency, constant business shifts. It indicates a demand variability growth from final market to the previous supply chain steps.
Choosing the best strategy is a complex work. A common problem concerns the appropriate amount of stock on hand. Too much stock entails too much costs to

store it and the risk that it remains unsold. Too little stock doesn't allow sellers to make any sales.

Promotional events may constitute a crucial aspect regarding an other key challenge, the pricing and predicting demand for products that a specific retailer has never sold before. When a customer visits a website, he sees several promotional events, each representing products that show common features such as the same designer, or the same category and so on. Usually, each event contain a timer informing about the event remaining time availability. When an event causes the customer interest, he can click on the event which takes him to a new page that shows all of the products for sale in that event.
Monitoring the number of times in which each event is visited, may give crucial business information concerning customer interest for particular brands, categories of products, styles. Consequently, it allows sales managers to obtain business key indicators based on current trends, customers' needs, products' appeal in order to optimize sales pricing and to forecast future demand.

Business intelligence tools may support considerably the just described argument. Highlighting promotional events' information through focused reports, allows to focus the business client attention on data particularly useful to set up new business strategies.
Visual analytics permits to improve considerably a decisive aspect, the information visualization. It brings together several scientific and technical communities from computer science, cognitive and perceptual sciences, information visualization, social sciences, interactive design, graphic design. The appropriate selection of reports' properties, such as shape, colour, data disposition, the choice of the leading information and the related visual accent, enable users to obtain deep insights that directly support assessment, planning, and decision making.

Case study

As described above, the data visualization employed tool is Microsoft Power BI.
In this case study, regarding the online fashion industry, in particular the luxury shoes industry, the business client's main needs consist in monitoring and analyzing the prices at which his sellers sell his products on their online stores. The developed work involves also the price analysis concerning products belonging to the business client's competitors, in order to set up more interesting business strategies to win competitions.

The key results, relating to the presentation of the business information, are reproduced below.

The Figure 5.1 refers to the average price at which the business client's products are sold by the sellers on the different examined online stores. It offers a intriguing overview which allows business client to make easily a comparison between the different sellers strategies.

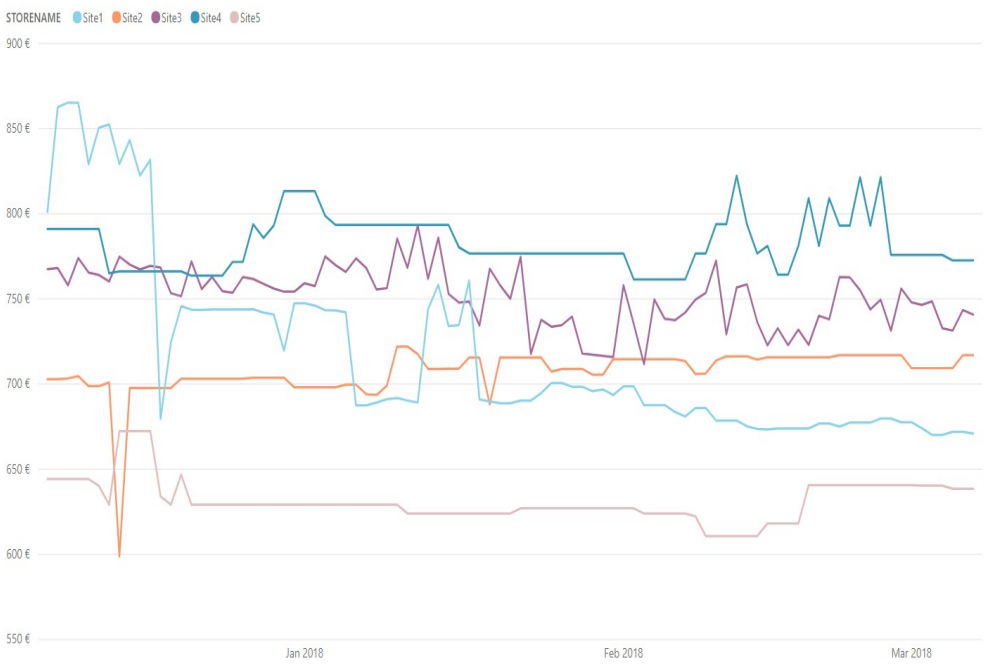


Figure 5.1. Average price on the different online stores

The Figure 5.2 highlights that this work is focused on a specific and particular branch of the fashion industry, concerning the luxury shoes. In fact, as can be seen in the figure, most of the time, the products are sold following their full price and not a discounted price. With regard to some brands, the retailers sell even their products following only their list price.

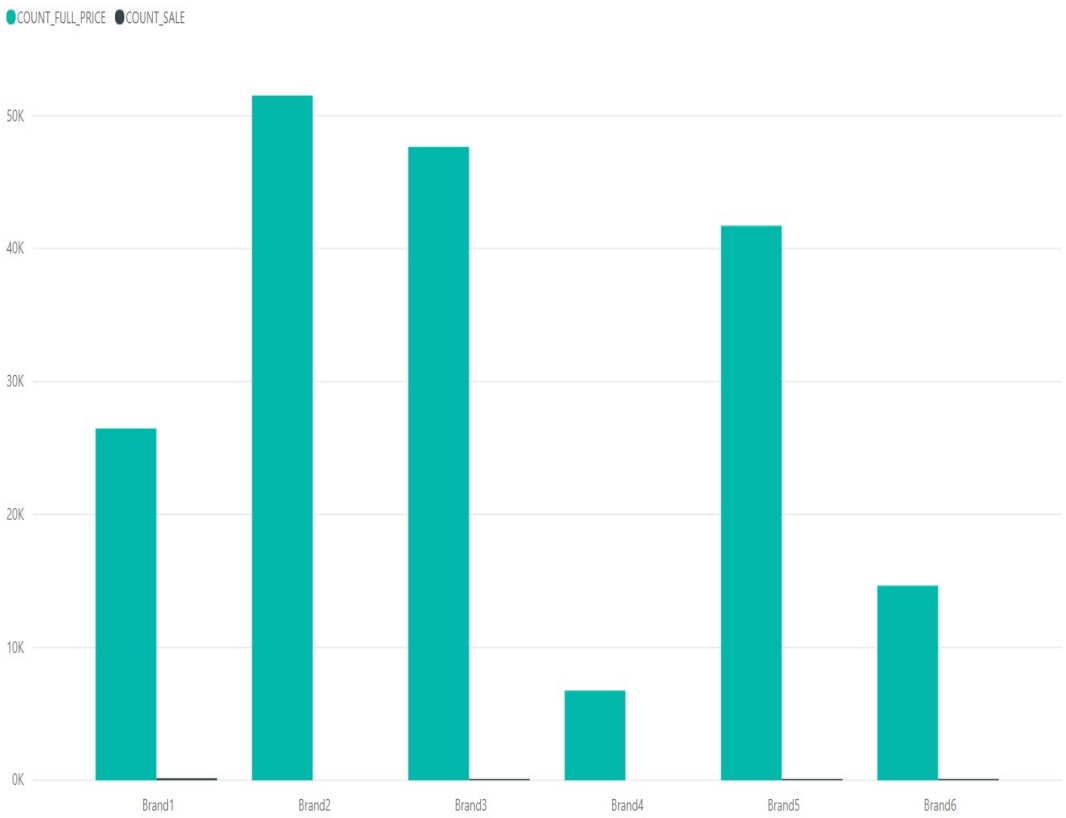


Figure 5.2. Comparison between list price and sales price

The following figures move the focus on the business aspects related to each product. The Figure 5.3 shows a list of products and their business attributes, derived from all the entire text mining stage. Examining the information illustrated in the Figure 5.3, the business client is able to discover the best-selling colours, categories, heels size, materials and the average price of the products.

On the other hand, the Figure 5.4 illustrates all the most relevant business details related to an only one product. In particular, it gives information concerning the specific product's sale on the different online stores, over time.
Moreover, the report highlights the different prices at which the different retailers have been selling the specific product gradually.

PRODUCT_NAME	DESCRIPTION	BRAND	CATEGORY_	HEEL	MATERIAL	COLOUR	Average of PRODUCT_STANDARD_PRICE
Braided Suede 60mm Sandal	[redacted]ebuted his first female footwear collection in 2007 during Milan Fashion Week, garnering praise for his elegant, chic, and surprisingly comfortable designs. With expertise passed down from his father, the legendary shoe designer [redacted] the younger [redacted] melds renowned Italian glamour and craftsmanship with a modern aesthetic that's uniquely his own. Single-sole construction and sculptural shapes are signatures of his collections.	Brand3	sandal	6.80	suede	texas	768.77 €
70 faille ankle boots	This season's boots are perfect for walking, dancing and pulling together countless outfits - [redacted] : pair encompasses all three. Made in Italy from lustrous black faille, they have leather lacing along the front and around the ankle and a sleek point-toe silhouette.	Brand3	boot	7.00	lace	black	850.00 €
70 leather pumps	Made cool by Kate Moss at the height of '90s minimalism, the kitten heel is back for spring and more stylish than ever. This pair by [redacted] has been crafted and hand-finished in Italy from smooth taupe leather and detailed with buckled straps that wrap in contrasting directions around the ankle. We think they look elegant with a midi skirt or dress.	Brand3	pump	7.00	leather	taupe	736.25 €
70 velvet and satin sandals	Glamorous sandals needn't be reserved for dressed-up occasions - they look just as cool styled with raw-edged denim. [redacted] navy pair has been crafted in Italy from plush velvet and fasten with lustrous satin ribbon ankle ties. The supportive back panel keeps them comfortably in place.	Brand3	sandal	7.00	velvet	navy	576.63 €
Foley D'Orsay Velvet Pumps	Add the elegance of velvet to your footwear line-up with the Foley D'Orsay pumps by [redacted] Designed with an ankle buckle closure and almond toe, they sit atop a platform sole perfect for evenings out.	Brand3	pump	7.00	velvet	burgundy	657.40 €
Portofino 70 patent-leather sandals	EXCLUSIVE AT [redacted] , [redacted] 's 'Portofino' sandals are a style every woman should have in their wardrobe - elegant and practical in equal measure. Crafted from glossy white patent-leather, they're set on a slim heel and have a signature round buckle. Make them pop against a black dress.	Brand3	sandal	7.00	leatherpatent	white	590.00 €
Portofino 70 textured-lamé sandals	EXCLUSIVE AT [redacted] . "Once the very definition of a solid, 'safe' shoe, the kitten heel is suddenly something altogether more desirable," says	Brand3	sandal	7.00	texture	pink	553.13 €
Portofino 70 textured-lamé sandals	EXCLUSIVE AT [redacted] . "Once the very definition of a solid, 'safe' shoe, the kitten heel is suddenly something altogether	Brand3	sandal	7.00	texture	red	592.00 €

Figure 5.3. Products' overview

PRODUCT_NAME	BRAND	Average of SELL_PRICE
Portofino leather sandals	Brand3	705.44

DATE_	STORENAME	PRODUCT_STANDARD_PRICE	SELL_PRICE	SALE
12/8/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/9/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/10/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/11/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/12/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/13/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/14/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/15/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/16/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/17/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/18/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/19/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/20/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/21/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/22/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/23/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/24/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/25/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/26/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/27/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/28/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/29/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/30/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
12/31/2017 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/1/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/2/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/3/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/4/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/5/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/6/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%

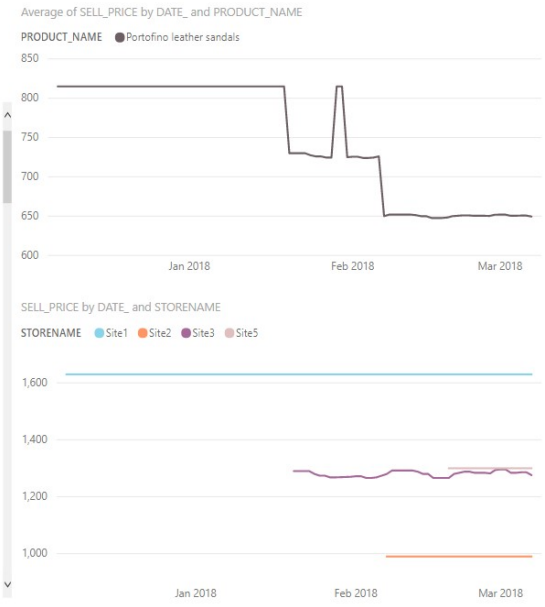
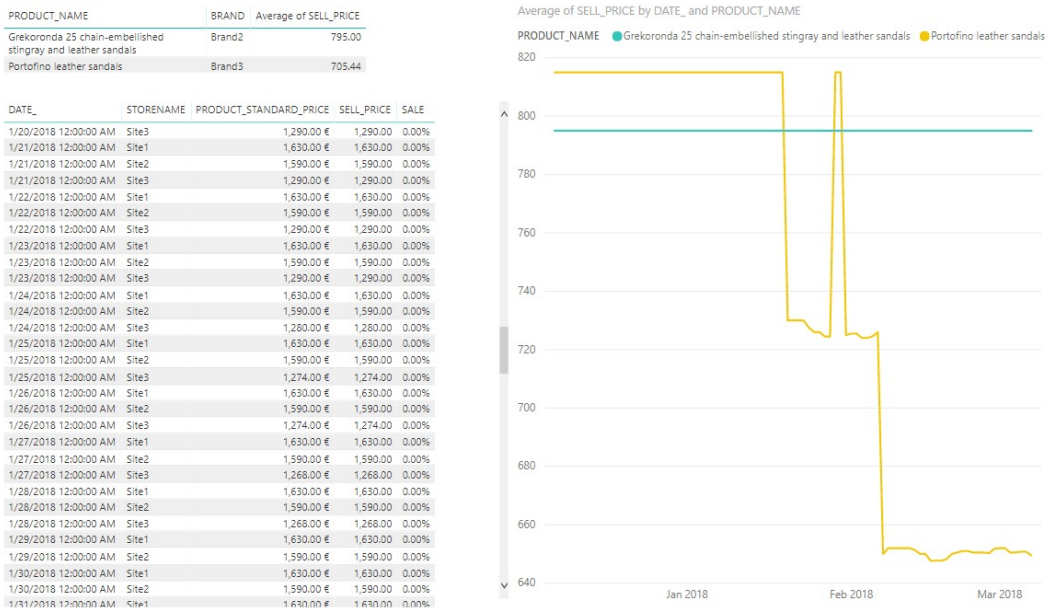


Figure 5.4. Product's details

An other fascinating analysis concerns the comparison between different products, belonging to different competitors (brands), that exhibit similar key business attributes. This investigation allows client business to obtain crucial deep insights regarding potential competitors' business strategies. Furthermore, it stresses not only information regarding pure price value, but it takes into account the entire online store's data, such as category in the example, in order to provide a more deep business knowledge.

The just described analysis is depicted in the Figure 5.5.



PRODUCT_NAME	BRAND	Average of SELL_PRICE
Grekoronda 25 chain-embellished stingray and leather sandals	Brand2	795.00
Portofino leather sandals	Brand3	705.44

DATE_	STORENAME	PRODUCT_STANDARD_PRICE	SELL_PRICE	SALE
1/20/2018 12:00:00 AM	Site3	1,290.00 €	1,290.00	0.00%
1/21/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/21/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/21/2018 12:00:00 AM	Site3	1,290.00 €	1,290.00	0.00%
1/22/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/22/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/22/2018 12:00:00 AM	Site3	1,290.00 €	1,290.00	0.00%
1/23/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/23/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/23/2018 12:00:00 AM	Site3	1,290.00 €	1,290.00	0.00%
1/24/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/24/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/24/2018 12:00:00 AM	Site3	1,280.00 €	1,280.00	0.00%
1/25/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/25/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/25/2018 12:00:00 AM	Site3	1,274.00 €	1,274.00	0.00%
1/26/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/26/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/26/2018 12:00:00 AM	Site3	1,274.00 €	1,274.00	0.00%
1/27/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/27/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/27/2018 12:00:00 AM	Site3	1,268.00 €	1,268.00	0.00%
1/28/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/28/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/28/2018 12:00:00 AM	Site3	1,268.00 €	1,268.00	0.00%
1/29/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/29/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/30/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%
1/30/2018 12:00:00 AM	Site2	1,590.00 €	1,590.00	0.00%
1/31/2018 12:00:00 AM	Site1	1,630.00 €	1,630.00	0.00%

Figure 5.5. Comparison between two similar products' sales

The Figure 5.6 shows the average price for each brand in the x-axis and the consistency of the sample (the number of records) in the y-axis, based on temporal evolution.
With regard to the specific brand analyzed in this figure, as can be seen observing the more coloured portions, the average price remains rather stable over time, introducing slight fluctuations caused by price changes on the retailers' online stores.

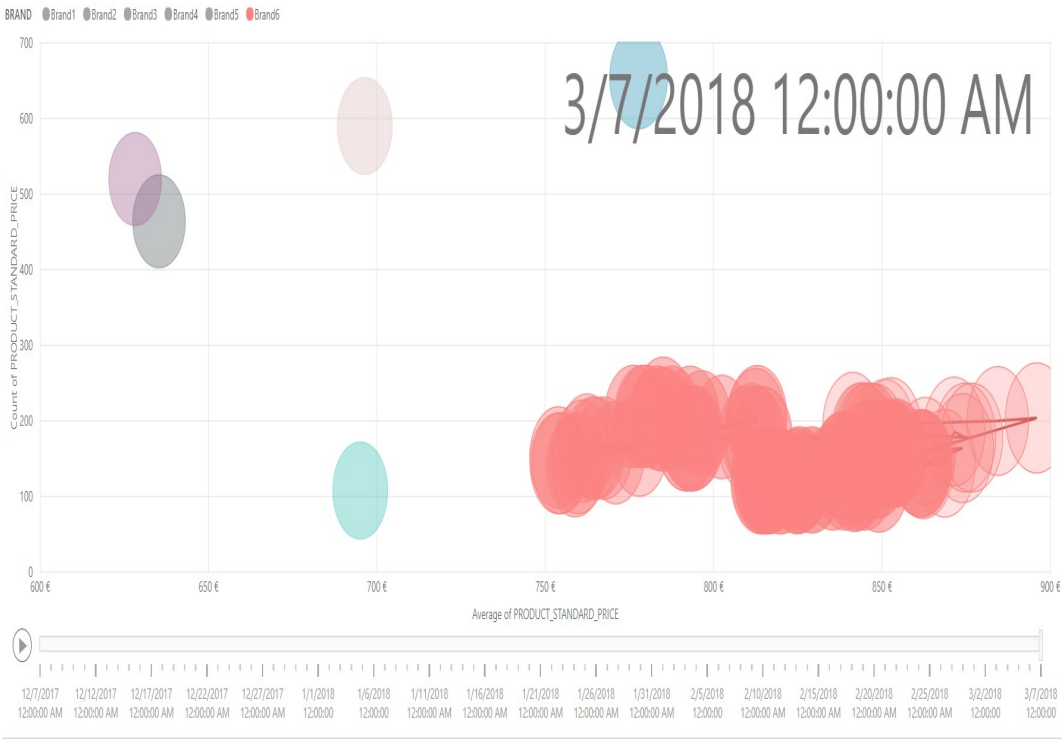


Figure 5.6. Average price development over time

The Figure 5.7 emphasizes one of the most important established goals in this thesis. It concerns the purpose of extracting information not only about price values, but regarding the entire online store. Product detail pages contain a number of secondary data, such as product description's and details' content, that may include potential hidden knowledge. The information extracted from these types of data can influence considerably the marketing planning of the business user, in addition to his future business strategies.
Consequently, attributes such as heel, colour, material, category assume a crucial role in the implemented advanced analytics. As can be seen in the Figure 5.7, an example concerns the business correlation existing between the sell price and the heel size. In fact, the diagram suggests that as the heels size increases, there is a sell price growth.

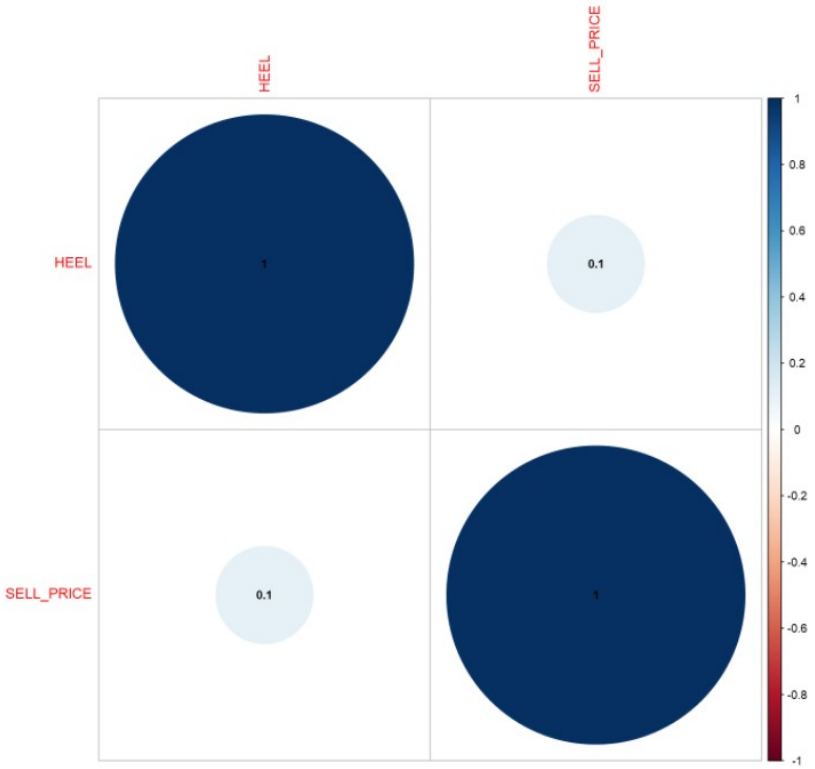


Figure 5.7. Correlations between key business indicators

Chapter 6

Conclusion and next steps

The realized work provides a solution related to the E-Commerce monitoring. It extends the modular platform already implemented by the company Mediamente Consulting. Its initial aim consists in satisfying the business needs of a company's client, regarding the monitoring of his retailers' online stores.
The end-to-end structure of this thesis allows to examine in depth all the main stages of data ingestion, data integration and manipulation, and data visualization.

The first phase concerns the collection of unstructured data coming from external sources, represented by websites in this case study. These raw data are captured and stored in a centralized data hub, through web analytics solution that consists in the realization of web crawlers, exploiting the Scrapy architecture.

In the second step, the saved data are subjected to a process of data integration in order to homogenize these data and the sell-out ones derived from the company's business client. The laborious process, that makes the external data structured, requires the implementation of text mining algorithms and it's realized through a robust ELT process. Other data mining methods, in particular the text clustering and the price forecasting, are implemented in order to extract a considerable number of business information from the collected data.

The third main stage takes into account the data visualization. The filtered, cleaned and structured data are used for the creation of graphs, reports, dashboards that allow business client to obtain a deep insight examining the data. Consequently, he will be able to set up suitable business and marketing strategies to improve his profit and to win over competitors.

Taking into account the entire performed work, feasible next steps may affect different phases of the end-to-end chain.
Regarding the web analytics stage, it would be interesting to capture data coming

from other types of external sources, such as social platforms, involving also data derived from users' comments and reviews. Combining them and online stores data, can trigger the extraction of new fascinating business information.
Moreover, collecting social networks data allows to move the focus also to the users, their needs and interests. This just described aspect promotes the implementation of other modules of advanced analytics, such as the Sentiment analysis, optimizing the data manipulation stage.

With regards to the clustering analysis step, the realization of an image clustering algorithm can considerably improve the performance and the results of the already implemented text clustering. The image clustering algorithm can be exploited regarding the matching test applied to luxury shoes coming from different online stores. It would be applied to the shoes' images, captured through the crawlers, to check if two shoes, having different names, represent the same product sold on different E-Commerce websites.

An other intriguing evolution of this work involves the role of the data visualization step. The use of reports and dashboards by business customers is rapidly evolving, changing continually the dimension of the data visualization process.
The way in which the data are visualized and presented to the client may be optimized in order to increase the business interaction with the client. The reports, implemented in this thesis, aspire not only to provide analysis regarding the client company performance; they try to supply useful recommendations that may influence successfully other aspects such as promotional advertising and marketing planning. Consequently, the business client would have available more information to perform the demand forecasting process, improving the marketing management, in addition to the choice of the appropriate business strategies.

Ringraziamenti

Raggiungere un traguardo importante come questo da soli, è certamente possibile ma molto difficile. Aver l'opportunità di ottenerlo con un gran lavoro di squadra, rappresenta per me un qualcosa di meraviglioso.
La vita continua a regalarmi un team di persone vincenti, e le vittorie di squadra sono sempre le più belle.

Il primo Grazie va ai miei genitori, che mi hanno permesso di intraprendere questo lungo e importante percorso di studi. La loro presenza costante, nei momenti di gioia e, in particolare, nei momenti difficili, è inestimabile. Descrivere in poche righe il loro ruolo è fin troppo riduttivo, sarò loro grato per sempre. Grazie.

Un secondo Grazie va ai miei migliori amici, Simone e Davide, e ai miei due fantastici coinquilini Davide e Alessandro. Quattro persone speciali con cui ho condiviso momenti particolarmente costruttivi che mi hanno dato quello slancio decisivo per superare ogni ostacolo. Sono e resteranno per me dei grandi punti di riferimento, indubbiamente.

Grazie alla mia famiglia, cugini, zii, nonni perchè hanno sempre creduto in me, in ogni istante, e perchè so di poter sempre contare su di loro.

Un Grazie va alla Professoressa Cerquitelli, relatrice della mia tesi, al mio tutor aziendale Vincenzo Scinicariello e alla sua straordinaria disponibilità, e al magico team di persone che ho avuto la fortuna di conoscere durante il mio tirocinio formativo in azienda. Non dimenticherò la loro meravigliosa accoglienza, la loro simpatia e il loro costante e importante supporto durante lo sviluppo del mio lavoro di tesi.

Infine, non posso non ringraziare tutti i miei amici, musicisti e non, che rappresentano elementi indispensabili della mia squadra vincente.

Grazie a tutti.

Antonio

Bibliography

[1] *Scrapy documentation*. https://doc.scrapy.org/en/latest/topics/architecture.html.

[2] Ranjit Bose. Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems*, 109(2):155–172, 2009.

[3] Umeshwar Dayal, Malu Castellanos, Alkis Simitsis, and Kevin Wilkinson. Data integration flows for business intelligence. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 1–11. Acm, 2009.

[4] Ms. Nithya Dr. S. Vijayarani, Ms. J. Ilamathi. Preprocessing techniques for text mining - an overview. 2016.

[5] Tania Cerquitelli Elena Baralis. Clustering fundamentals.

[6] Layla Hasan, Anne Morris, and Steve Probets. Using google analytics to evaluate the usability of e-commerce sites. In *International Conference on Human Centered Design*, pages 697–706. Springer, 2009.

[7] Bernard Marr. *Big Data in practice*. 2016.

[8] What Is Data Mining. Data mining: Concepts and techniques. *Morgan Kaufinann*, 2006.

[9] Christopher Olston, Marc Najork, et al. Web crawling. *Foundations and Trends® in Information Retrieval*, 4(3):175–246, 2010.

[10] Gautam Pant, Padmini Srinivasan, and Filippo Menczer. Crawling the web. In *Web Dynamics*, pages 153–177. Springer, 2004.

[11] Jack Phillips. Analytics 3.0: The era of impact.

[12] Joe Caserta Ralph Kimball. *The Data Warehouse ETL Toolkit*. 2004.

[13] Shaeela Ayesha Ramzan Talib, Muhammad Kashif Hanif and Fakeeha Fatima. Text mining: Techniques, applications and issues. 2016.

[14] Vikas Ranjan. A comparative study between etl (extract, transform, load) and elt (extract, load and transform) approach for loading data into data warehouse. Technical report, viewed 2010-03-05, http://www. ecst. csuchico. edu/˜ juliano/csci693/Presentations/2009w/Materials/Ranjan/Ranjan. pdf, 2009.

[15] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.