POLITECNICO DI TORINO

Facoltà di Ingegneria

Corso di Laurea in Ingegneria Gestionale

Tesi di Laurea Magistrale

Prediction model for drinking water network failures

Data Science

Relatori
prof. Victor Leiva Sanchez
prof.ssa Elisabetta Raguseo

Candidato
Daniele Coppola

Anno accademico 2018-2019

Abstract

The large amount of data nowadays available for the companies creates wide business opportunities. Data Science meets application in different domains. Data science processes leverage internal or external resources, such as big data, from which useful insights and predictive models can be obtained, to create value for the company through the significant support that can be provided to the decision making through these data analytics tools. This research traces the foundations of the Data Science approach, Business Intelligence and Big Data. In this context, is placed the PME Project, a Data Science project that proposes a solution for the management of pipeline failures in the drinking water networks. In particular, this project is based on the search for a predictive model of failures in a drinking water network, a study that will allow a better understanding of the complexity of the network management. However, the research focuses on predicting and its purpose is to support decision making in the context of renovations and repairs. In this way, the aim is to reduce the uncertainty about network management, reduce costs related to emergency purchases, avoid fines and improve six-monthly renewal planning.

Keywords. Business Intelligence, Big Data, Big Data Analytics, Data Science, Decision Making.

Acknowledgements

I would like to thank the profesor Victor Leiva Sanchez and Pablo Zuniga who supported my work. I am also grateful to the Pontificia Universidad Catolica de Valparaiso that gives me the opportunity to attempt classes and develop the thesis.

I would like to thank the professoress Elisabetta Raguseo who supported me for the development of the thesis.

I would like to thank my friends and my family for supporting me writhing this thesis and in my life.

Last but not the least, I would like to thank Chile and all the people who have shared with me this experience.

In particolare vorrei ringraziare i miei genitori e mio fratello che mi hanno sempre sostenuto sia nel mio percorso di studi che nella vita dimostrandomi costantemente il loro amore per me e spronandomi a migliorarmi sempre di piu' per poterli rendere fieri e orgogliosi di me.

Vorrei ringraziare Emiliana, entrata nella mia vita come un uragano a piccoli passi e che, nonostante la distanza, ha saputo riempire tutti i miei giorni e darmi tutta la forza, l'appoggio e l'amore che potessi desiderare per affrontare le esperienze che la vita mi ha posto davanti, con o senza di lei, facendomi sentire di poter contare sempre su di lei.

Vorrei infine ringraziare tutta la mia famiglia, nonna, zii e cugini, i miei fratelli Obe, tutti i miei amici e le persone che ho conosciuto e che hanno arricchito la mia esperienza.

Sono felice e orgoglioso di avervi nella mia vita e poter condividere momenti come questo con voi, Vi amo!

”Dio ti prego salvaci da questi giorni
Tieni da parte un posto e segnati sti nomi”

Contents

List of Figures

Chapter 1

Introduction

Business Intelligence and Analytics (BI&A) and the field of Big Data analytics have emerged as an area of research on decision support systems (DSS), with growing interest among academics and researchers.[1][2] Sector studies have highlighted this significant development. For example, based on a survey of over 4,000 IT professionals from 93 countries and 25 industries, the IBM Tech Trends Report (2011) identified Business Analytics as one of the four major technology trends of the 2010s. In a Bloomberg Businessweek Business Analytics Status Survey (2011), 97% of companies with revenues in excess of $100 million were found to use some form of Business Analytics. A study by the McKinsey Global Institute states that the United States alone faces a shortage of 140,000 - 190,000 people with deep analytical skills and 1.5 million managers and analysts to analyze Big Data and make decisions based on its results.[3] In the age of Big Data, Business Intelligence and Analytics can help improve organizational performance as a result of improved business decision-making.[1][4] Business Intelligence and Analytics born from the success of Business Intelligence (BI) in the 1990s and the introduction of Business Analytics (BA) in the 2000s, a key element of data analysis in Business Intelligence.[4] Thus, Business Intelligence and Analytics can be defined as techniques, technologies, systems, practices, methodologies and applications that analyze critical business data to help a company better understand its business and market and make timely business decisions.[1]

Howard Dresner of the Gartner Group introduced the term Business Intelligence in 1989, describing a set of concepts and methods to improve business decision-making using fact-based support systems.[5] Although Dresner's original Business Intelligence definition, as well as the more recent definitions of analysts such as Gartner, Forrester and TDWI, most professionals associate the term with a narrow set of capabilities, such as extraction, transformation and loading (ETL); data warehousing; on-line analytic processing (OLAP); and reporting.[4] We use Business Intelligence to understand the company's available capabilities, the state of the art, future market trends and directions, technologies, the regulatory environment in which the company competes, the actions of competitors and the consequences of their actions. The focus of these uses of Business Intelligence is on analyzing historical data. More recently, Big Data and Big Data Analytics have been used to describe data sets and analytical techniques in applications so large, from terabytes to exabytes, and complex like sensors and social media data, that they require advanced and unique data storage, management, analysis and visualization technologies. Big data is the current buzz among researchers, who are trying to use the huge amount of data to come to a conclusion that improves not only their decision making, but also profitability, productivity and the power to peek into the past to look to the future to make it much better and more efficient. The concept of Big Data has become so popular because it has different applications in various fields: (1) e-commerce and market intelligence, (2) e-government and policy 2.0, (3) science and technology, (4) smart health and wellness and (5) security and public. The reason for these claims is that Big Data can change competition by transforming processes, altering business ecosystems and facilitating innovation[6]; unlocking the business value of the organization by unlocking new organizational capabilities and value[7]; and making it easier for companies to address key business challenges.[8] The impact of Big Data offers enormous potential for competition and growth for industries and private companies, and the proper and intelligent use of Big Data increases productivity, innovation and competitiveness of entire sectors and economies.

In particular, researchers and professionals use the term Big Data to refer to the continuous expansion of data in terms of 3‘V’: volume (how large the volume of data can be in petabytes), variety (the different data formats that can be processed together) and speed (how fast data are generated and managed).[9] More recently, value which means doing something valuable with the data is important and veracity, which refers to the certainty and accuracy of the data have emerged as integrating ‘V’.[10] Others argue that Big Data is not a revolution but an evolution of traditional Business Intelligence.[1][11] According to this vision, Big Data Analytics expands the scope of Business Intelligence, which focuses on the integration and reporting of structured data residing in the company’s internal databases, trying to extract value from semi-structured and unstructured data from data sources such as the web, mobile devices and sensor networks external to the company.

The fact that Big Data Analytics tools and platforms are relatively accessible is significant because Big Data is not just for large companies. Many small and medium-sized businesses, especially those involved in digital processes for sales, customer interactions or the supply chain, also need to manage and leverage Big Data. A survey conducted by MIT Sloan Management Review and the IBM Institute for business value of nearly 3,000 executives, managers and analysts from over 30 industries and 100 countries found that the most performing organizations make decisions based on rigorous analysis at more than double the rate of organizations with lower performance and that in such organizations analytical understanding is used to guide both future strategies and day-to-day operations.[12]

Characterized by volume, variety, speed and value[1], Big Data Analytics is seen as a disruptive technology that will reshape Business Intelligence to gain business insights to improve decision-making and is believed to be the next blue ocean in cultivating business opportunities. Big Data Analytics is the place where advanced analytical techniques operate on Big Data sets. So, Big Data Analytics is actually about two things: Big Data and Analytics. Put them together and you get Big Data Analytics, the new hottest Business Intelligence practice. It is clear that domain knowledge and analysis cannot be separated. Thus, academics and professionals must have both analytical skills and knowledge and understanding of business and management. What differentiates today's Big Data Analytics applications from traditional Business Intelligence applications is not only the breadth and depth of the data processed, but also the type of questions they answer. While Business Intelligence traditionally focuses on using a consistent set of metrics to measure past business performance[4], Big Data applications emphasize exploration, discovery and forecasting.

In this context, is placed the PME Project, a Data Science project that uses the big volume of data available on the drinking water network to propose a solution for the management of pipeline failures, with an emphasis on data analysis and management and on the various steps of a Data Science project. In particular, this study will allow a better understanding of the complexity of the network management. The project arise from the need to better manage the renewal of the drinking water network, as the company, it is currently the company that pays the highest rate of fines at national level. This project consists in the development of a predictive model based on Cox's regression model, from which the probability of survival of a pipeline section in a certain pressure sector is obtained. With the output of this model, a ranking is made to predict (in time and space) which pipelines and sectors are more incline to fail. With this predictive model, the aim is to support the decision making in the context of six-monthly renewals and network repairs, so as the company should be able to improve its network management process and reduce fines and costs related to emergency purchases.

Chapter 2

Business Intelligence

Business Intelligence (BI) can be defined as the process of transforming data into information and then into knowledge. Information has become an important competitive factor in today's business world, so providing the right people with the right information at the right time is critical to an organization both to achieve and maintain its competitive advantage.

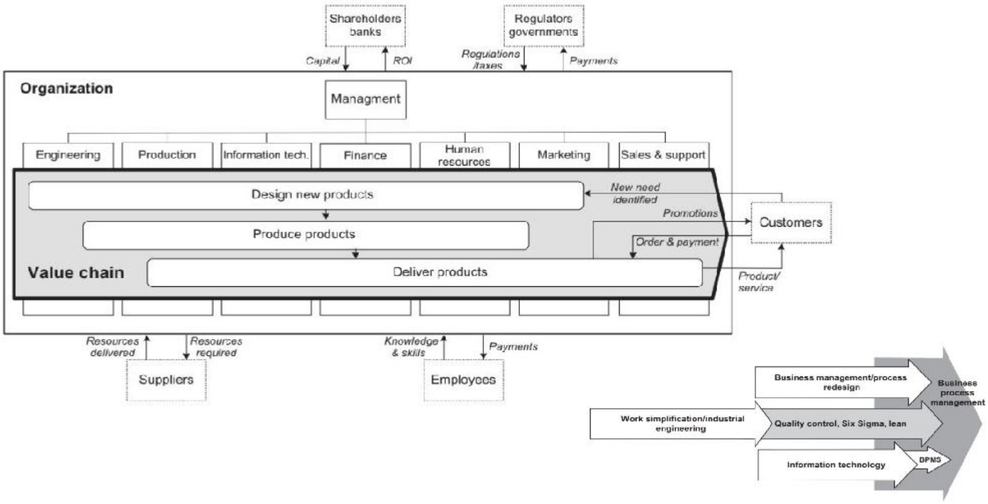


Figure 2.1. Information System to gain Competitive Advantage[84]

Information and communication technologies aimed at providing an organization's management with the relevant information to stay ahead of the competition have undergone fundamental changes since the 1960s. Several terms have been introduced such as decision support systems, management information system and management support system. Finally, in the mid-1990s, the term Business Intelligence was coined by analysts and consultants from Gartner, Inc. to define the application of a set of methodologies and technologies, such as Web Services, XML, data warehouse, OLAP, Data Mining, representation technologies, etc., to analyze an organization and its competitive environment, to improve the effectiveness of business operations and support management/decision to achieve competitive advantages. Business Intelligence systems are complemented by specialized IT infrastructures including data warehouses, data marts and Extract Transform & Load tools, necessary for their effective use.

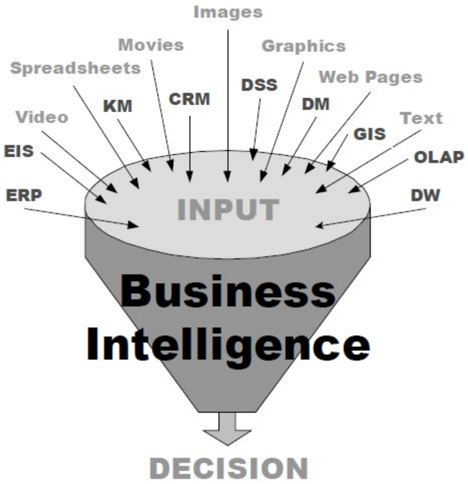


Figure 2.2. Business Intelligence to support the Decision Making[19]

Business Intelligence is the process of taking large amounts of data, analyzing that data and presenting a set of high-level reports that condense the essence of that data into the basis of business actions, enabling management to make fundamental day-to-day business decisions.[13] Business Intelligence is a way and method to improve business performance by providing powerful assists to executive decision makers to enable them to have usable information at their fingertips. Business Intelligence tools are seen as a technology that enables the efficiency of business operations by providing added value to business information and thus the way in which this information is used.[14] Business Intelligence is the process of collecting, processing and disseminating information that has a goal, reducing uncertainty in making all strategic decisions. Experts describe Business Intelligence as a business management term used to describe the applications and technologies that are used to collect, provide access to analyze data and information about a business, in order to help them make more informed business decisions.[15] Business Intelligence is defined as comprising an effective data warehouse and also a reactive component capable of monitoring time-critical operational processes to allow tactical and operational decision makers to adjust their actions according to business strategy.[16]

Business Intelligence has two different basic meanings related to the use of the term intelligence. The first is the ability of human intelligence applied to business/activities. The second refers to intelligence as information evaluated for its currency and relevance. The term implies having a complete knowledge of all the factors that influence the business. It is imperative that companies have a thorough understanding of factors such as customers, competitors, business partners, the business environment and internal operations to make effective and good quality business decisions.

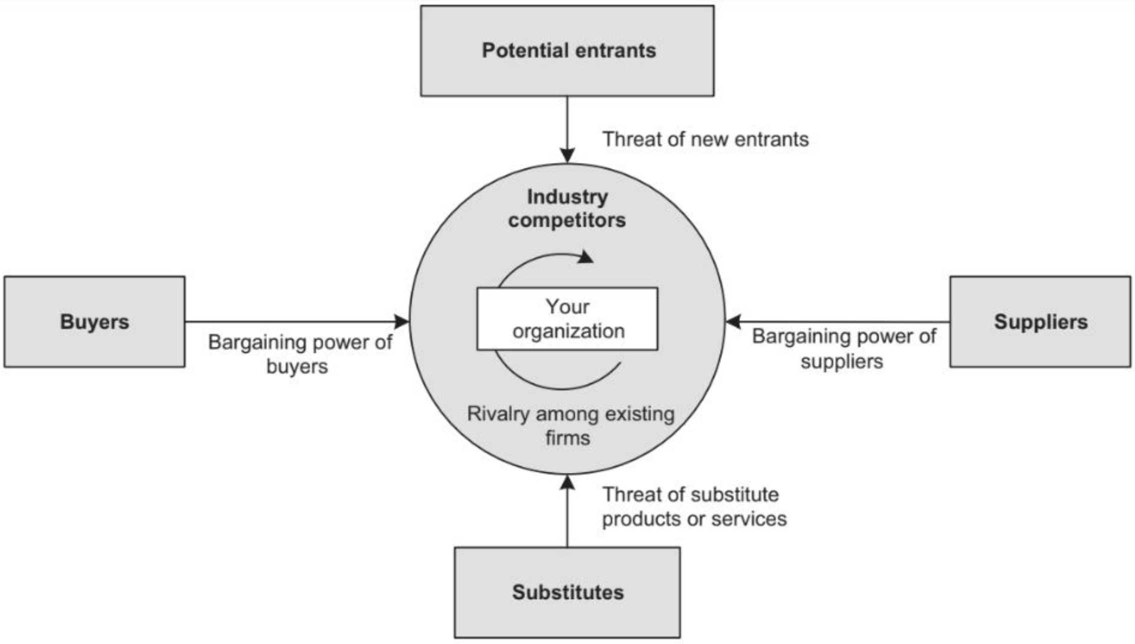

Potential entrants
Threat of new entrants
Industry competitors
Buyers
Bargaining power of buyers
Your organization
Rivalry among existing firms
Bargaining power of suppliers
Suppliers
Threat of substitute products or services
Substitutes

Figure 2.3. Competitive Environment[84]

Business Intelligence is both a process and a product. The process is composed of methods that organizations use to develop useful information, or intelligence, that can help organizations survive and thrive in the global economy. The product is information that allows organizations to predict with some degree of certainty the behavior of their competitors, suppliers, customers, technologies, acquisitions, markets, products and services, and the overall economic environment. Consequently, Business Intelligence is mainly aimed at providing top management with relevant information to support strategic decisions such as planning, control and coordination of activities based on internal and external data.[17]

These definitions of Business Intelligence have two important implications: first that, often, approaches to Business Intelligence are marked by supported functions, systems or types of systems.[18] Second, Business Intelligence aims primarily at providing the management circle of an organization (at all levels of management and support staff functions) with analytical information relevant to the decisions supporting their management activities.[18] Operational tasks such as the execution of business processes and support processes are poorly supported by Business Intelligence processes or systems. In many cases, the analytical information needed to support management decisions is extracted from a common database, the so-called data warehouse (DWH). A DWH is a object-oriented, integrated, non-volatile, non-volatile and time-variable data collection to support management decisions. As a result, many, if not all, of today's Business Intelligence implementations are mainly data centered, for example they are focused on the analysis of data from an organization's DWH with the ultimate goal of generating reports and providing management information systems with relevant aggregated information to support management processes.

Business Intelligence systems combine data collection, data storage and knowledge management with analytical tools to present complex and competitive information to planners and decision makers. The first objective of Business Intelligence systems is to improve the timeliness and quality of inputs to the decision-making process. Data is treated as a corporate resource and transformed from quantity to quality. In this way, actionable information can be provided at the right time, in the right place and in the right form to help decision makers improve their decision-making.[19] Many companies are adopting Business Intelligence tools and systems to learn from the past and predict the future. The impact on the performance of Business Intelligence systems can be seen on at least two levels[20]: (1) improve the efficiency and effectiveness of the organizational structure and business processes (internal strategy), and (2) outperform the performance of other organizations in the industry (competitive strategy). The measure at the organizational level is an evaluation tool that informs managers if the company has realized the benefits of organizational performance. While process level performance measurement is a diagnostic tool, it will inform the management about why/why not organizational performance has improved.

2.1 Components of Business Intelligence

In this paragraph we will describe the main components of a Business Intelligence system.

2.1.1 OLAP (On-line Analytical Processing)

Online Analytical Processing (OLAP) provides multidimensional and summary views of companies' data and is used for reporting, analysis, modeling and planning to optimize the business process and strategy. OLAP techniques and tools can be used to work with data warehouses or data marts designed for sophisticated Enterprise Intelligence systems that process queries to discover trends and analyze critical factors. Business Intelligence tools are used to store and analyze data, such as data mining and data warehouse, decision support and forecasting systems, document warehouse and document management, knowledge management, mapping, information visualization and dashboarding, management information systems, geographic information systems, trend analysis and software as a service (SaaS).

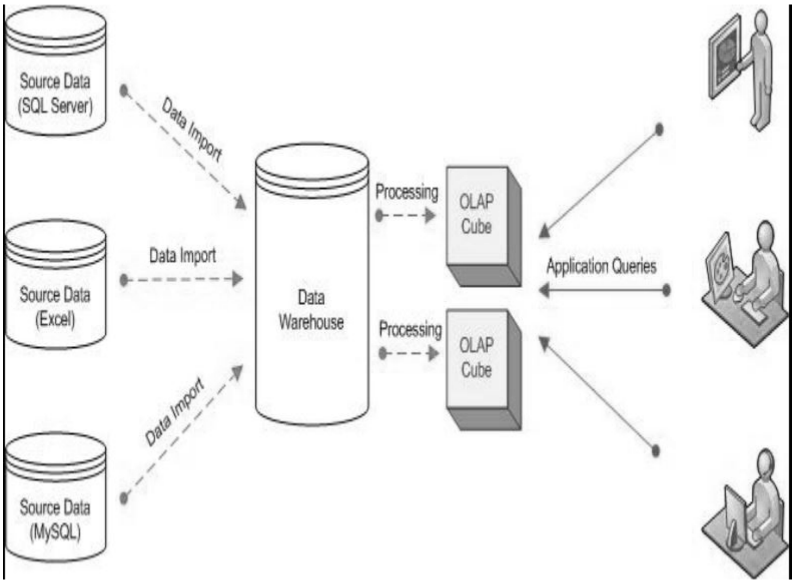


Figure 2.4. OLAP[85]

2.1.2 Advanced Analytics

These are data mining, forecasting or predictive analytics. Advanced Analytics uses statistical analysis techniques to predict or provide measures of certainty about the facts. Analytics helps us discover how facts have changed and how we should react. It is a collection of techniques and types of related tools, which usually include predictive analysis, data mining, statistical analysis and complex SQL (Structured Query Language). We could also include data visualization, artificial intelligence, natural language processing, and database capabilities that support analytics. In other words, with advanced analytics, the user is typically a business analyst who is trying to discover new business facts that no one in the company knew before. To do this, the analyst needs large volumes of truly detailed data. This is often data that the company has not yet used for analytics.

2.1.3 Corporate Performance Management (Portals, Scorecards, Dashboards)

It is a general category that usually provides a container for several pieces to be connected so that the aggregate tells a story.

2.1.4 Real-time Business Intelligence

Allows real-time distribution of metrics through email, messaging systems and interactive displays. Business users have always requested more recent data. Today, business information integration (EII), business application integration (EAI) and real-time data warehousing technologies make it possible to provide decision support data that is literally in real time or, if not, at least almost in real time. This development is profoundly changing the face of decision support, making it possible to influence the current decision-making process, operational business processes and customer-facing applications.

2.1.5 Data Warehouse and Data Marts

The data warehouse is the significant component of business intelligence. It is subject-oriented, integrated. The data warehouse supports physical data propagation by managing numerous business records for integration, cleaning, aggregation and querying. It can also contain operational data that can be defined as an updatable set of integrated data used for tactical business decisions in a particular subject area. It contains real-time, non-instantaneous data and maintains a minimal history. Data sources can be operational databases, historical data, external data or information from the existing data warehouse environment. A data mart is a collection of themed areas organized for decision support based on the needs of a given department. Finance has its own data mart, marketing has its own, and sales have their own and so on. Perhaps most importantly, individual departments have the hardware, software, data and programs that make up the data mart. Each department has its own interpretation of the data mart and the data mart of each department is peculiar and specific to its own needs. Like data warehouses, data marts contain operational data that help business experts implement strategies based on trend analysis and past experiences. The fundamental difference is that the creation of a data mart is based on a specific and predefined need for a certain grouping and configuration of selected data. Within an enterprise, there may be multiple data marts. A data mart can support a particular business function, process or unit.

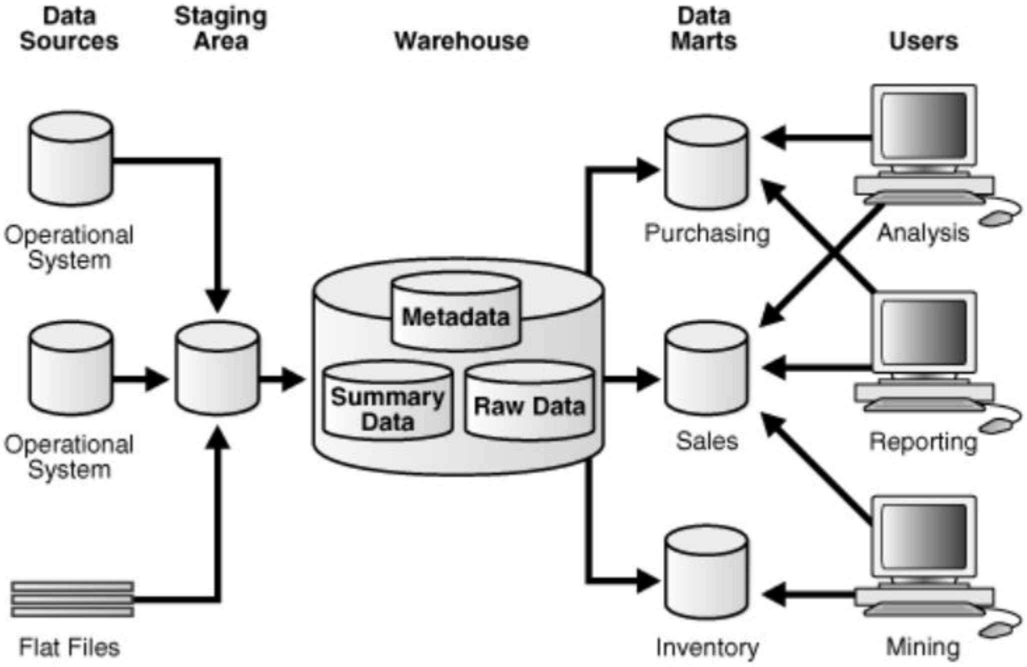


Figure 2.5. Data Warehouse and Data Marts[86]

2.1.6 Data Sources

Data sources can be operational databases, historical data, external data, or information from the existing data warehouse environment. Data sources can be relational databases or any other data structure that supports the enterprise application line. They can also reside on different platforms and can contain information.

2.2 Benefits and Costs

An organization can use Business Intelligence to: (1) improve management processes (planning, control, measurement, monitoring and change) to increase revenue and reduce costs, (2) improve operational processes so that the business can increase revenues and reduce costs.

In other words, the business value of Business Intelligence depends on its use within management processes that affect operational processes that in practice generate revenue or reduce costs, and in its use within operational processes themselves.

2.2.1 Benefits

There are many reasons for greater integration between Business Intelligence and process-oriented management:

- Many operational processes generate transactional data that are integrated and analyzed in a DWH as in the use of processes and systems of Business Intelligence.
- Many operational processes require analytical information as input for execution. Operational processes often include steps that are highly or entirely related to Business Intelligence tasks, therefore, a model of such a process must contain Business Intelligence services.
- Business Intelligence deals with the integration and consolidation of raw data into key performance indicators (KPIs) that represent a fundamental input for business decisions during the execution of the process. Therefore, operational processes provide the context for data analysis, information interpretation and appropriate measures.
- Once decisions have been made on the basis of KPIs, useful context information such as specific KPI values, corresponding decisions, consequences and reactions to decisions, etc. can be added to a data repository dedicated to experience. Thus, step by step, the data repository is filled with indications for making decisions. This leads to the acquisition of know-how and possibly improvements in KPI definitions and recommendations on whether or not to take a particular action when KPIs take on certain threshold values.

Business Intelligence allows organizations to make well-informed business decisions and can therefore be the source of competitive advantage, especially when companies are able to obtain information from external environment indicators and make accurate predictions about future trends or economic conditions. Business Intelligence offers many advantages to companies that use it. It can eliminate many guesswork within an organization, improve communication between departments by coordinating activities, and enable companies to respond quickly to changing financial conditions, customer preferences, and supply chain operations. Business

Intelligence improves the overall performance of the company that uses it. The ultimate goal of Business Intelligence is to improve the timeliness and quality of information to know what actions are best taken for the benefit of the company and how quickly they respond and adapt to changes. Information is a very important resource that a company has, so when it can make decisions based on timely and accurate information, the company can improve its performance; this also accelerates the decision-making process.

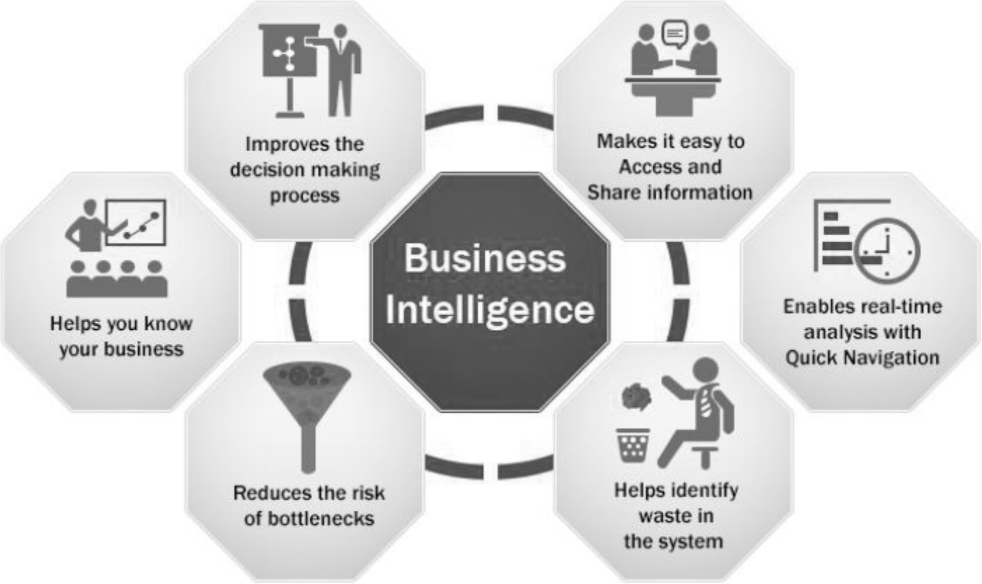

Improves the decision making process
Makes it easy to Access and Share information
Helps you know your business
Business Intelligence
Enables real-time analysis with Quick Navigation
Reduces the risk of bottlenecks
Helps identify waste in the system

Figure 2.6. Business Intelligence Benefits[][87]

Companies have recognized the importance of business intelligence to the masses has arrived. Some reasons are listed below:

- With superior Business Intelligence tools, employees can now also easily convert their business knowledge through analytical intelligence to solve many business problems.
- With Business Intelligence, companies can identify their most profitable areas and identify future opportunities.
- Analyze click-stream data to improve e-commerce strategies.
- Quickly identify issues reported under warranty to minimize their impact.

- Discover criminal money laundering activities.
- Analyze customers' potential growth profitability and reduce risk exposure through more accurate financial credit scoring of their customers.
- Determine which combinations of products and lines of services customers can purchase and when.
- Analyze clinical trials for experimental drugs.
- Set more profitable rates for insurance premiums.
- Reduce downtime by applying predictive maintenance.
- Determine with friction and churn analysis why customers leave for competitors and become customers.
- Detect and deter fraudulent behaviour, such as peaks in use in the event of theft of credit cards or telephone cards.
- Identify promising new molecular compounds.

The benefits that could be realized within an organization such as improvement of process performance and quality, improvement of services provided to internal stakeholders, acceleration of process execution and efficiency gains in resource use and process execution outweigh the benefits that are external to the organization such as improvement of services provided to external stakeholders, increase in customer profitability, satisfaction and loyalty. The advantages that can be realized within an organization seem to be more prominent and important than the advantages that are rather external to the organization, so organizations implement the Process Centric Business Intelligence to become more efficient in the execution of processes and allocation of resources and to improve the service to internal stakeholders because these are the areas that generate the highest return on investment in projects and processes of Business Intelligence.

2.2.2 Costs

Business Intelligence projects are not exempt from the growing pressure on companies to justify return on investment. Surveys show that the return on investment (ROI) for Business Intelligence installations can be substantial, in fact it is between 17% and 2,000% with an average ROI of 457%. Most of the benefits of Business Intelligence are intangible before the fact. An empirical study for 50 Finnish companies found that most companies do not consider cost or time savings as the primary benefit when investing in Business Intelligence systems. The size of the Business Intelligence market can be seen from the published forecasts. For example, a study estimates that the current Business Intelligence market should grow from 15.64 billion dollars in 2016 to reach 29.48 billion dollars by 2022.[21] However, the Business Intelligence budget and ROI were not correlated; the challenge is to try to assess ROI before installation. The traditional model of designing, building and integrating Business Intelligence systems is long (at least six months) and expensive (2-3 million dollars). Therefore, many companies opt for pre-built analytical applications to achieve lower total cost of ownership, faster implementation, and a rapid return on investment, while achieving a basic framework for performance, scalability, and flexibility. An IDC study of OLAP investment over 5 years indicated an investment of $2.1 million in building OLAP solutions within the company, with a 104% ROI. The same study indicated that an investment of $1.8 million in the purchase of pre-constructed OLAP solutions led to a ROI of 140%, which means that Business Intelligence solutions cost less and bring a higher ROI.[22] Achieving the expected return on investment for Business Intelligence is a difficult issue. Like most information systems, the initial costs of Business Intelligence are high, as is maintenance. Unfortunately, although it is possible to predict cost reductions of information systems in terms of efficiency, savings in terms of efficiency are only a small part of the payoff, in fact it would be rare for a Business Intelligence system to repay itself through cost reduction. The implementation of a Business Intelligence system includes a variety of costs that the company has to face during the execution of the process. In particular, we have:

- **Hardware costs**: These costs depend on what is already installed. If a data warehouse is in use, then the main hardware needed is a specific data mart for Business Intelligence and, perhaps, an update for the data warehouse. However, additional hardware may be required to perform additional tasks.

- **Software costs**: Typical Business Intelligence packages can cost $60,000 and subscriptions to various data services must also be taken into account.

- **Implementation costs**: Once the hardware and software have been acquired, a major expense is implementation, including initial training. Training is also a continuous cost because new people are introduced to use the system and because the system is constantly updated. In addition, annual software maintenance contracts usually cover 15% of the purchase costs.

- **Personnel costs**: Personnel costs for Business Intelligence and IT support staff must be fully taken into account to take into account salary and overheads, space, IT equipment and other infrastructure for individuals. Sophisticated cost analysis also takes into account the time spent reading Business

Intelligence results and the time spent searching the Internet and other sources of Business Intelligence.

2.3 Critical Success Factors

Although Business Intelligence has the potential to improve a company's performance, a significant number of companies often fail to realize the benefits expected from Business Intelligence and sometimes consider the Business Intelligence project a failure in itself.

2.3.1 Business Intelligence Implementation

Implementing a business intelligence system is a complex and resource-intensive enterprise. When implementing a Business Intelligence program the company should ask some questions and make some decisions such as:

- **Questions about the alignment of objectives**: The first step determines the short and medium term objectives of the program. To which strategic objectives of the organization will the program refer? What mission/organizational vision does it refer to? An elaborated hypothesis needs to detail how this initiative will eventually improve results/performance.
- **Basic questions**: Current competence needs for information gathering and competence needs assessment. Does the organization have the capacity to monitor important sources of information? What data does the organization collect and how does it store it? What are the statistical parameters of these data?
- **Questions about costs and risks**: The financial consequences of a new Business Intelligence initiative should be estimated. It is necessary to evaluate the cost of ongoing operations and the increase in costs associated with the Business Intelligence initiative. What is the risk of the initiative failing? This risk assessment should be converted into a financial metric and included in the planning.
- **Questions from customers and stakeholders**: Determine who will benefit from the initiative and who will pay. What types of customers/interested parties will benefit directly from this initiative? Who will benefit indirectly? What are the quantitative/qualitative benefits? Is the specified initiative the best way to increase the satisfaction of all types of customers or is there a better way? How will customer benefits be monitored?
- **Metric questions**: These information requirements must be made operational in clearly defined metrics. You have to decide which metrics to use for each piece of information collected. Are these the best metrics? How do we

know? How many metrics should be tracked? If it is a large number (usually it is), what kind of system can be used to track them? Are the metrics standardized so that they can be compared with the performance of other organizations? What industry standard metrics are available?

- **Measurement method questions**: A methodology or procedure should be established to determine the best (or acceptable) way to measure the required metrics. What methods will be used and how often will the organization collect the data? Are there industry standards for this purpose? Is this the best way to take measurements? How do we know?
- **Questions about the results**: Someone should monitor the Business Intelligence program to ensure that the objectives are achieved. Adjustments to the program may be necessary. The program should be tested for accuracy, reliability and validity.

2.3.2 Organizational, Process and Technological Dimension

There are a number of critical success factors that influence the success of implementation that take into account two key measures: the performance of the infrastructure, for instance, the quality of the system and the output standard and the performance of the process, for instance, how well the process of implementing a Business Intelligence system went. Infrastructure performance is parallel to the three main success variables of the IS known as system quality, information quality and system use, while process performance can be assessed in terms of timing and budgetary considerations.[23] In particular, the quality of the system relates to the performance characteristics of the computer system itself, which include ease of use, functionality, reliability, flexibility, integration and response time. Information quality means the accuracy, timeliness, completeness, relevance, consistency and usefulness of the information generated by the system. The use of the system is defined as consumption of the product of an information system.

A survey conducted with the Delphi method shows the existence of a set of CSFs for a good implementation of a Business Intelligence system and we can group them into three dimensions: organizational dimension, process dimension and technological dimension.[24]

(1) Organizational dimension

- **Committed Management Support and Sponsorship**

  The commitment of management and sponsorship has been widely recognized as the most important factor in the implementation of the Business Intelligence system. All Delphi participants agreed that consistent support and sponsorship from corporate executives makes it easier to secure the necessary operational resources, such as funding, human competencies and other requirements throughout the implementation process. This observation is reasonable and expected because the entire effort of implementing the Business Intelligence system is an expensive, time-consuming and resource-intensive process. In addition, Delphi's experts have argued that the implementation of the Business Intelligence system is a program of continuous information improvement to leverage decision support. As a result, the Business Intelligence initiative, especially for the entire enterprise scale, requires consistent resource allocation and top-management support to overcome organizational issues.

- **Clear business vision and a well-defined case**

  Since a Business Intelligence initiative is business-driven, a strategic business vision is needed to guide the implementation effort. Delphi participants indicated that a long-term vision, especially in strategic and organizational terms, is needed to enable the creation of Business Intelligence business cases. The business case must be aligned with the corporate vision, as it could have an impact on the adoption and outcome of the Business Intelligence system. Otherwise, they will not receive the executive and organizational support necessary for their success.

(2)Process dimension

- **Business-Centric Championship and balanced project team composition**

  Most Delphi experts believed that having the right sample from the commercial side of the organization is critical to successful implementation. According to them, a champion who has excellent entrepreneurial acumen is always important as he will be able to predict organizational challenges and change course accordingly. More importantly, this business-centric sample would see the Business Intelligence system primarily in strategic and organizational perspectives, unlike one who might overly focus on technical aspects. All respondents also agreed that the composition and skills of a Business Intelligence team have a great influence on the success of the implementation. The project team must be cross-functional and composed of staff with technical skills and staff with a strong business background. A Business Intelligence system is a business-driven project to provide advanced management decision support, so an adequate mix of information technologies skills is needed to implement the technical aspects, while the reporting and analysis aspects must be under the area of corporate personnel. In addition, most experts assume that the Business Intelligence team should include business domain experts, especially for activities such as data standardization, requirements engineering, data quality analysis and testing. Many also agreed with the critical role played by external consultants, especially in the initial phase. They believed that the lack of experience and expertise could be complemented by external consultants who spent most of their time working on similar projects.

- **Business user-oriented change management**

  Having an adequate user-oriented change management effort was considered critical by Delphi participants. Better participation of users in the change effort can lead to better communication of their needs, which can help ensure successful implementation of the system. This is particularly important when the requirements of a system are initially unclear. A significant number of respondents said that formal user participation can help meet the demands and expectations of various end users.

- **Business driven methodology and project management**

  According to Delphi experts, proper project definition and planning allows the Business Intelligence team to focus on the best opportunities for improvement. In particular, scoping helps to establish clear parameters and develops a common understanding of what is in the scope and what is excluded.

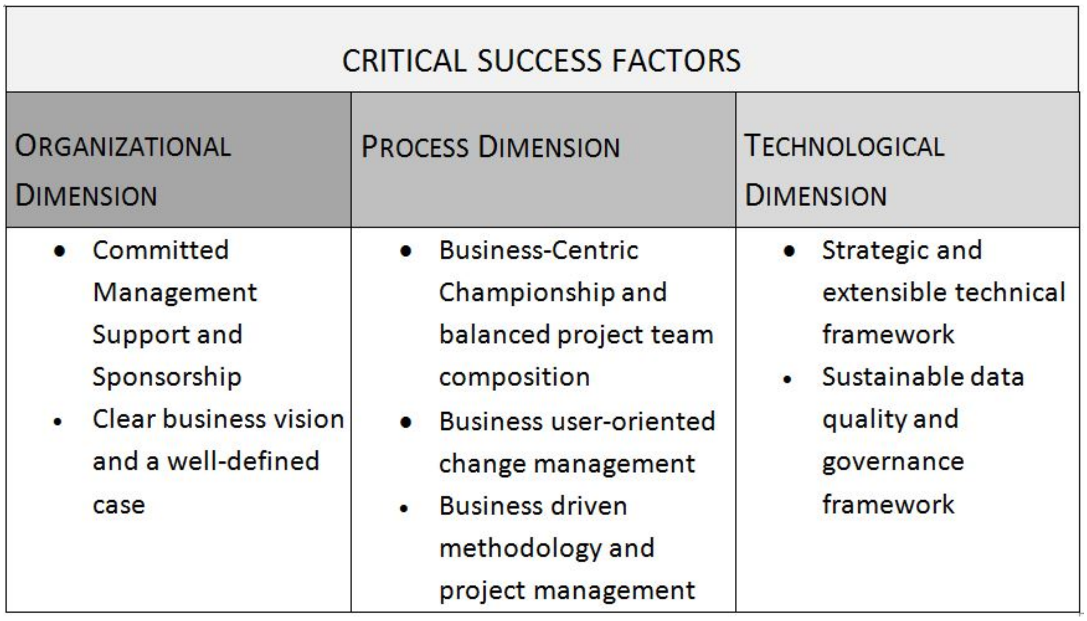
(3) Technological dimension

- **Strategic and extensible technical framework**

  In terms of strategic and extensible technical structure, most experts stated that stable source/back end systems are crucial for the implementation of a Business Intelligence system. A reliable back-end system is critical to ensure that data updates work well for the extraction, transformation and loading (ETL) processes in the staging area. In this way, data can be transformed to provide a consistent view into quality information for better decision support. It is therefore essential for the Business Intelligence team to assess the stability and consistency of the source systems before undertaking a Business Intelligence effort. Otherwise, after the implementation of the system, the cost of changes in terms of time and money can be significant. Another key element was that the technical structure of a Business Intelligence system must be able to meet the requirements of scalability and extensibility. Having an integrated strategic vision in the system design, this scalable system structure could include additional data sources, attributes and dimensional areas for fact-based analysis, and could incorporate external data from suppliers, contractors, regulators and industry benchmarks. It would then enable a long-term solution to be built to meet the incremental needs of enterprises.

- **Sustainable data quality and governance framework**

  The quality of the data, in particular of the source systems, is essential if a Business Intelligence system is to be successfully implemented. According to the respondents, a primary purpose of the Business Intelligence system is to integrate silos of data sources within the enterprise for advanced analysis, so as to improve the decision-making process. Often, many data issues within back-end systems are not discovered until the data is populated and interrogated in the Business Intelligence system (Watson, 2004). In this way, business data can only be fully integrated and exploited to achieve greater business value if its quality and integrity are guaranteed. Management is also invited to initiate data governance and stewardship initiatives to improve data quality in back-end systems because unreliable data sources will have a knock-on effect on Business Intelligence applications and later on decision outcomes.



CRITICAL SUCCESS FACTORS		
ORGANIZATIONAL DIMENSION	PROCESS DIMENSION	TECHNOLOGICAL DIMENSION
• Committed Management Support and Sponsorship • Clear business vision and a well-defined case	• Business-Centric Championship and balanced project team composition • Business user-oriented change management • Business driven methodology and project management	• Strategic and extensible technical framework • Sustainable data quality and governance framework

Figure 2.7. Critical Success Factors

2.4 Business Intelligence and Analytics

Business Intelligence and Analytics (BI&A) has emerged as an important area of study for both professionals and researchers, reflecting the breadth and impact of data issues in contemporary business organizations.[1] Business Intelligence and Analytics can help improve organizational performance as a result of improved business decision-making. Business Intelligence and Analytics enables companies to improve existing organizational applications by providing business-centric practices and methodologies that could provide a competitive advantage. Opportunities associated with data and analysis in different organizations have helped generate significant interest in Business Intelligence and Analytics, which is often referred to as techniques, technologies, systems, practices, methodologies and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions. In addition to the underlying analytics and data processing technologies, Business Intelligence and Analytics includes business-centric practices and methodologies that can be applied to various high-impact applications such as e-commerce, market intelligence, e-government, healthcare and security. Figure 2.8 shows the evolution of Business Intelligence and Analytics, applications and emerging analytical research opportunities.

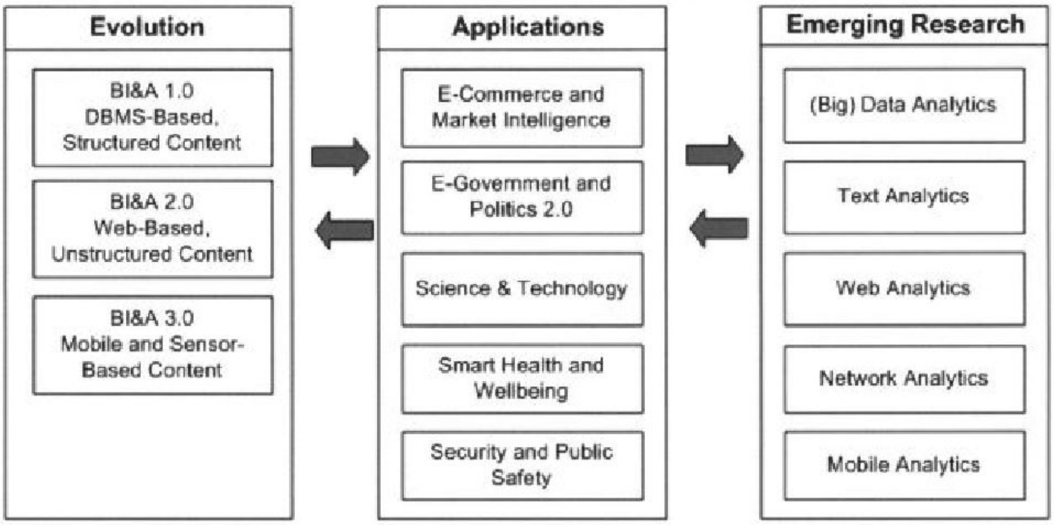
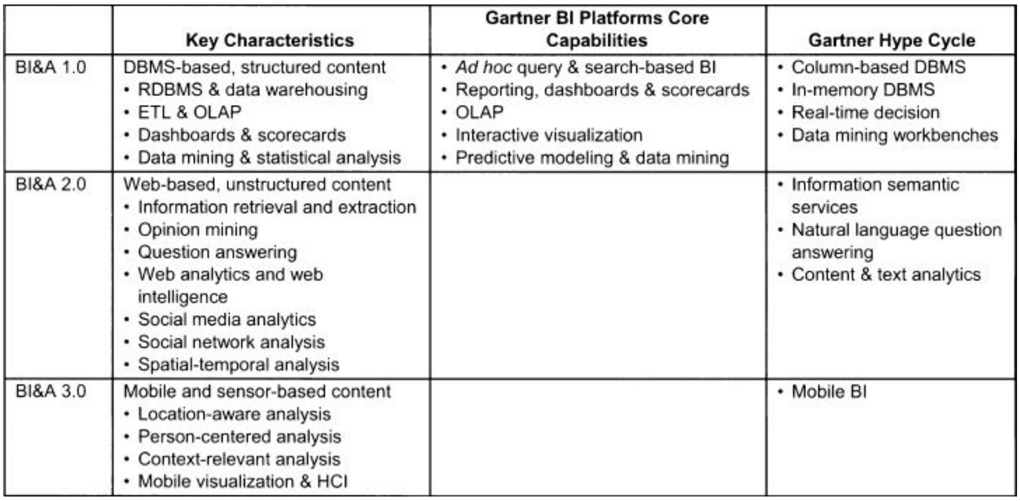

Evolution
BI&A 1.0 DBMS-Based, Structured Content
BI&A 2.0 Web-Based, Unstructured Content
BI&A 3.0 Mobile and Sensor-Based Content
Applications
E-Commerce and Market Intelligence
E-Government and Politics 2.0
Science & Technology
Smart Health and Wellbeing
Security and Public Safety
Emerging Research
(Big) Data Analytics
Text Analytics
Web Analytics
Network Analytics
Mobile Analytics

Figure 2.8. Business Intelligence and Analytics evolution, applications, research opportunities[1]

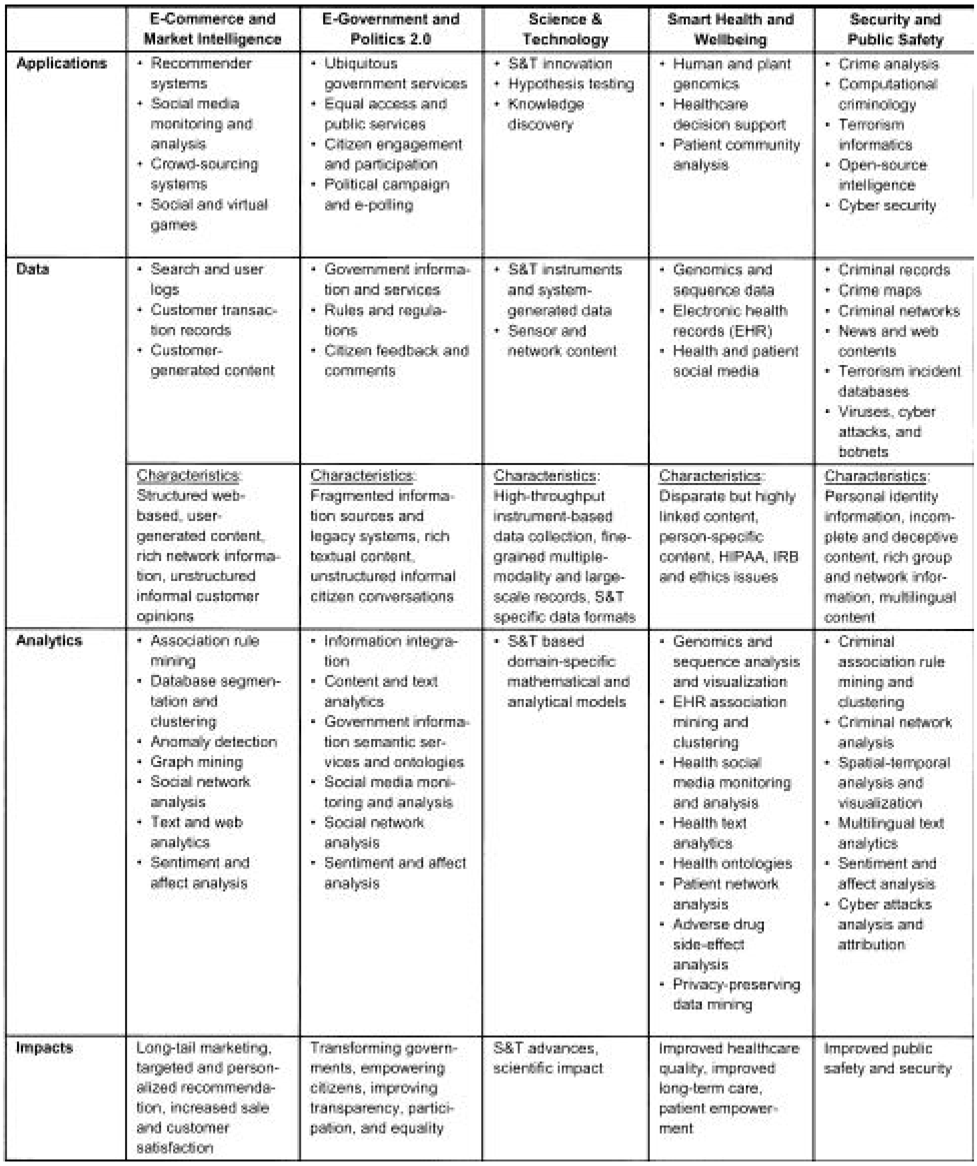
Business Intelligence and Analytics, as a data-centric approach, has its roots in the field of long data-centric database management. It is based on different technologies of data collection, extraction and analysis. We can analyze the evolution of Business Intelligence and Analytics technologies and applications by highlighting three different phases.[1] Business Intelligence and Analytics 1.0 could be considered as the phase in which data is mostly structured, collected by companies through various legacy systems, and often stored in commercial relational database management systems. Data management and storage is considered the foundation of Business Intelligence and Analytics 1.0. The design of data marts and tools for extraction, transformation and loading (ETL) are essential for the conversion and integration of enterprise-specific data. Database querying, online analytic processing (OLAP), and intuitive yet simple graphic-based reporting tools are used to explore important data features. Business Performance Management (BPM), which uses scorecards and dashboards, helps to analyze and visualize a variety of performance metrics. In addition to these proven business reporting functions, statistical analysis and data mining techniques are adopted for association analysis, data segmentation and clustering, classification and regression analysis, anomaly detection and predictive models in various business applications. Among the 13 features considered essential for Business Intelligence platforms, according to Sallam's Gartner report (2011), we consider Business Intelligence and Analytics 1.0: reporting, dashboards, ad hoc queries, search-based Business Intelligence, OLAP, interactive visualization, scorecards, predictive modeling and data mining. Web intelligence, web analytics, and user-generated content collected through Web 2.0-based social and crowd-sourcing systems drove a new era of Business Intelligence and Analytics 2.0 research in the 2000s, focusing on text and web analytics for unstructured web content. Except for the basic search and query capabilities, no advanced text analysis for unstructured content is currently considered in the 13 capabilities of Gartner's Business Intelligence platforms. New opportunities in Business Intelligence and Analytics 3.0 are emerging. The ability of these Internet-enabled and mobile devices to support highly mobile, location-aware, person-centered, and context-sensitive operations and transactions will continue to offer unique research challenges and opportunities throughout the years 2010. Figure 2.9 summarizes the key features and capabilities of Business Intelligence and Analytics 1.0, 2.0 and 3.0.



	Key Characteristics	Gartner BI Platforms Core Capabilities	Gartner Hype Cycle
BI&A 1.0	DBMS-based, structured content • RDBMS & data warehousing • ETL & OLAP • Dashboards & scorecards • Data mining & statistical analysis	• *Ad hoc* query & search-based BI • Reporting, dashboards & scorecards • OLAP • Interactive visualization • Predictive modeling & data mining	• Column-based DBMS • In-memory DBMS • Real-time decision • Data mining workbenches
BI&A 2.0	Web-based, unstructured content • Information retrieval and extraction • Opinion mining • Question answering • Web analytics and web intelligence • Social media analytics • Social network analysis • Spatial-temporal analysis		• Information semantic services • Natural language question answering • Content & text analytics
BI&A 3.0	Mobile and sensor-based content • Location-aware analysis • Person-centered analysis • Context-relevant analysis • Mobile visualization & HCI		• Mobile BI

Figure 2.9. Key characteristics and capabilities of Business Intelligence and Analytics 1.0, 2.0, and 3.0[1]

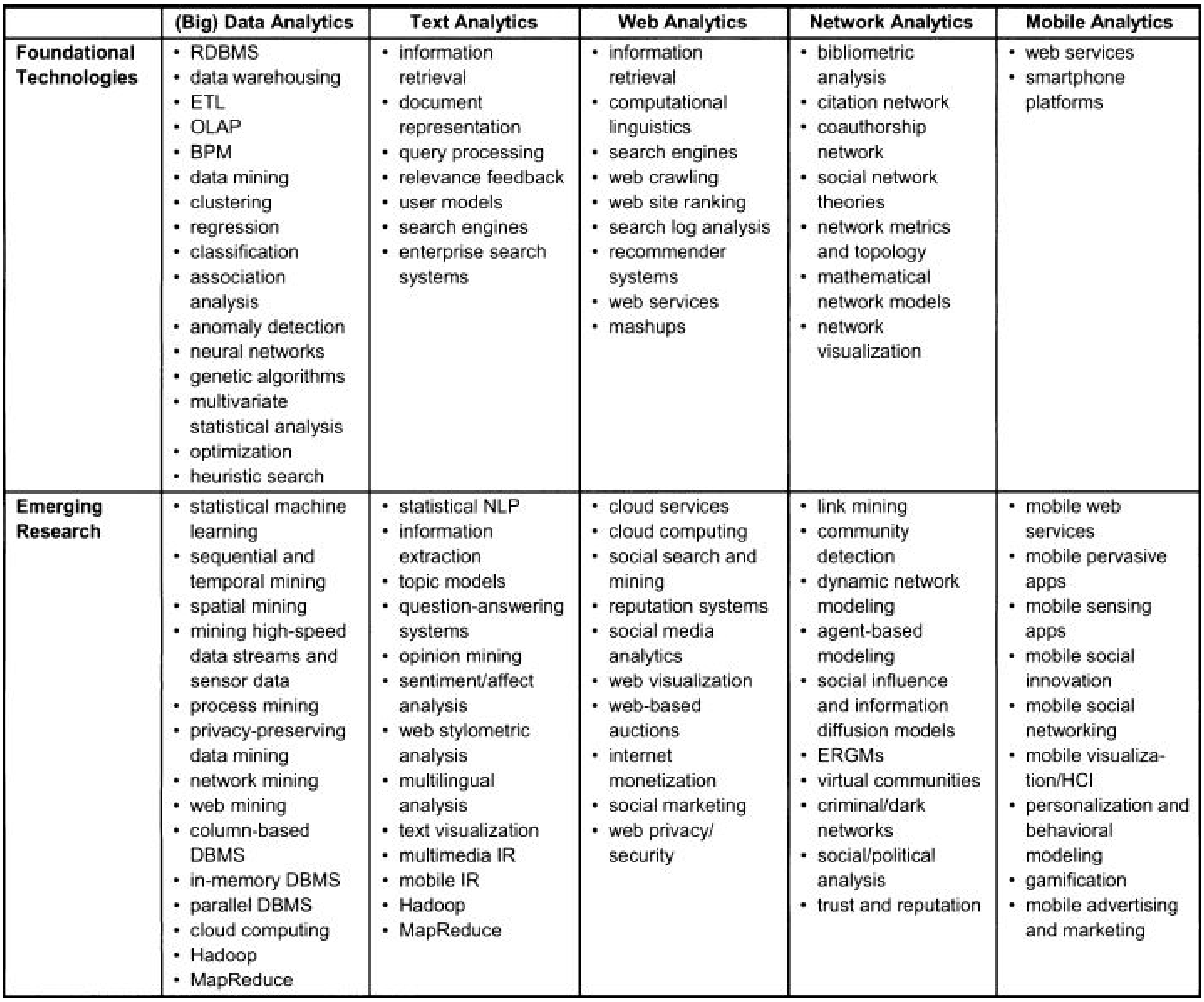
In addition to being data-based, Business Intelligence and Analytics is highly applied and can take advantage of the opportunities offered by the abundance of domain-specific data and analysis needed in many critical and high-impact application areas. Many of these high-impact Business Intelligence and Analytics applications are listed below, with a discussion of data characteristics and analysis, potential impacts, and selected illustrative examples or studies: (1) e-commerce and market intelligence, (2) e-government and policy 2.0, (3) science and technology, (4) smart health and wellness, and (5) security and public. Figure 2.10 summarizes Business Intelligence and Analytics applications, data characteristics, analysis techniques and potential.[1]



	E-Commerce and Market Intelligence	E-Government and Politics 2.0	Science & Technology	Smart Health and Wellbeing	Security and Public Safety
Applications	• Recommender systems • Social media monitoring and analysis • Crowd-sourcing systems • Social and virtual games	• Ubiquitous government services • Equal access and public services • Citizen engagement and participation • Political campaign and e-polling	• S&T innovation • Hypothesis testing • Knowledge discovery	• Human and plant genomics • Healthcare decision support • Patient community analysis	• Crime analysis • Computational criminology • Terrorism informatics • Open-source intelligence • Cyber security
Data	• Search and user logs • Customer transaction records • Customer-generated content	• Government information and services • Rules and regulations • Citizen feedback and comments	• S&T instruments and system-generated data • Sensor and network content	• Genomics and sequence data • Electronic health records (EHR) • Health and patient social media	• Criminal records • Crime maps • Criminal networks • News and web contents • Terrorism incident databases • Viruses, cyber attacks, and botnets
	Characteristics: Structured web-based, user-generated content, rich network information, unstructured informal customer opinions	Characteristics: Fragmented information sources and legacy systems, rich textual content, unstructured informal citizen conversations	Characteristics: High-throughput instrument-based data collection, fine-grained multiple-modality and large-scale records, S&T specific data formats	Characteristics: Disparate but highly linked content, person-specific content, HIPAA, IRB and ethics issues	Characteristics: Personal identity information, incomplete and deceptive content, rich group and network information, multilingual content
Analytics	• Association rule mining • Database segmentation and clustering • Anomaly detection • Graph mining • Social network analysis • Text and web analytics • Sentiment and affect analysis	• Information integration • Content and text analytics • Government information semantic services and ontologies • Social media monitoring and analysis • Social network analysis • Sentiment and affect analysis	• S&T based domain-specific mathematical and analytical models	• Genomics and sequence analysis and visualization • EHR association mining and clustering • Health social media monitoring and analysis • Health text analytics • Health ontologies • Patient network analysis • Adverse drug side-effect analysis • Privacy-preserving data mining	• Criminal association rule mining and clustering • Criminal network analysis • Spatial-temporal analysis and visualization • Multilingual text analytics • Sentiment and affect analysis • Cyber attacks analysis and attribution
Impacts	Long-tail marketing, targeted and personalized recommendation, increased sale and customer satisfaction	Transforming governments, empowering citizens, improving transparency, participation, and equality	S&T advances, scientific impact	Improved healthcare quality, improved long-term care, patient empowerment	Improved public safety and security

Figure 2.10. Business Intelligence and Analytics applications, data characteristics, analytics techniques, and potential[1]

Emerging analytical research opportunities can be classified into five critical technical areas: (big) data analytics, text analytics, web analytics, network analytics and mobile analytics.[1]



	(Big) Data Analytics	Text Analytics	Web Analytics	Network Analytics	Mobile Analytics
Foundational Technologies	• RDBMS • data warehousing • ETL • OLAP • BPM • data mining • clustering • regression • classification • association analysis • anomaly detection • neural networks • genetic algorithms • multivariate statistical analysis • optimization • heuristic search	• information retrieval • document representation • query processing • relevance feedback • user models • search engines • enterprise search systems	• information retrieval • computational linguistics • search engines • web crawling • web site ranking • search log analysis • recommender systems • web services • mashups	• bibliometric analysis • citation network • coauthorship network • social network theories • network metrics and topology • mathematical network models • network visualization	• web services • smartphone platforms
Emerging Research	• statistical machine learning • sequential and temporal mining • spatial mining • mining high-speed data streams and sensor data • process mining • privacy-preserving data mining • network mining • web mining • column-based DBMS • in-memory DBMS • parallel DBMS • cloud computing • Hadoop • MapReduce	• statistical NLP • information extraction • topic models • question-answering systems • opinion mining • sentiment/affect analysis • web stylometric analysis • multilingual analysis • text visualization • multimedia IR • mobile IR • Hadoop • MapReduce	• cloud services • cloud computing • social search and mining • reputation systems • social media analytics • web visualization • web-based auctions • internet monetization • social marketing • web privacy/ security	• link mining • community detection • dynamic network modeling • agent-based modeling • social influence and information diffusion models • ERGMs • virtual communities • criminal/dark networks • social/political analysis • trust and reputation	• mobile web services • mobile pervasive apps • mobile sensing apps • mobile social innovation • mobile social networking • mobile visualization/HCI • personalization and behavioral modeling • gamification • mobile advertising and marketing

Figure 2.11. Business Intelligence and Analytics research framework: Foundational technologies and emerging research in Analytics[1]

Business Intelligence and Analytics is the science of data in business. Jobs looking for data scientists and business analytics specialists abound these days. There is a clear shortage of professionals with the deep knowledge needed to handle the three V's of the big data: volume, velocity and variety. There is also a growing demand for individuals with the deep knowledge needed to manage the three perspectives of business decision-making: descriptive analytics, predictive and prescriptive. Training in Business Intelligence and Analytics should be interdisciplinary and cover the analytical and critical IT skills, business and domain knowledge, and communication skills required in a complex, data-centric business environment. To provide useful insights and decision support, Business Intelligence and Analytics professionals must be able to understand business issues and design appropriate analytical solutions. Business Intelligence and Analytics professionals need business knowledge ranging from general familiarity with the areas of Accounting, Finance, Management, Marketing, Logistics and Operation Management, to the domain knowledge required in specific Business Intelligence and Analytics applications. To support this culture, Business Intelligence and Analytics professionals need to know not only how to transform raw data and information through analytics into meaningful and workable knowledge for an organization, but also how to properly interact with the organization's business and domain experts.

2.5 Impact on Decision Making

To assess the joint effects of the use of Business Intelligence and Analytics and the organizational decision making process on organizational performance (how the use of Business Intelligence and Analytics affects organizational decision making processes and how it affects the use of Business Intelligence and Analytics) we analyze three phases of use of Business Intelligence and Analytics to achieve performance gains[25]: (1) the data at the insight phase, (2) the insight at the decision phase and (3) the decision at the value phase.

2.5.1 Data to Insight

Although current technologies provide analysts and managers with a large amount of structured and unstructured data, insights do not automatically emerge by mechanically applying analytical tools to the data. Rather, insights emerge from an active process of engagement between analysts and business managers who use the data and analytical tools to discover new knowledge. More importantly, these engagements take place within existing structures and decision-making processes. A better understanding of the insight generation process is important to understand how the use of Business Intelligence and Analytics leads to improved performance. The process of generating insight from data involves multiple actors from different areas of the organization. Team composition and structure is the result of management decisions made within existing decision-making routines that may allow or limit teams' ability to generate insights. There is a complex relationship between data, analytical tools and human sense formation. Business Intelligence and Analytics give rise to a process called datafication that allows analysts and managers to use data and analysis as a means of understanding the phenomena that data represent.[26] Lycett also argues that, despite the data-driven nature of analytic-based sense making, existing frameworks of analysts and managers have an important influence on which data elements are selected to describe the phenomena and which models and relationships linking the data are derived from the data. These insights are then used by managers and analysts to weave a narrative that gives meaning to the world and then to build action repertoires that make these interpretations explicit. Although business analytics tools make it easy to identify statistical models, trends and relationships, the next critical step in understanding the causes of these models is still important to take action that generates value. Probably, automated learning algorithms can detect patterns and even improve their performance over time.

How do structures and decision making affect the ability of insight generation teams to generate insight from the use of Business Intelligence and Analytics?

2.5.2 Insight to Decision

Just as it is essential to generate meaningful insights, it is equally essential for the company to transform insights into value-creating decisions. Intuitions, which refer to a deep and intuitive understanding of phenomena, must be exploited by analysts and managers in strategic and operational decisions to generate value. Good understanding leads to better decisions and Big Data leads to big impact.[1] The use of Business Intelligence enables a better understanding of business issues and opportunities through the analysis of current operations that can lead companies to discover new sources of revenue or achieve cost savings. While it is reasonable to expect that there is a relationship between the use of Business Intelligence and Analytics and better insights and decisions, it is not clear under which conditions these best results could be observed. First, there is no individual correspondence between an intuition and a specific course of action to exploit that intuition. Intuitions, including those based on understanding trends, operations, customers and suppliers, probably suggest multiple options to leverage them and convert them into value. Some options may be obvious, while others may be the result of a more creative process. The second issue is that organizational decision-making processes have an important influence on how insights are converted into decisions. Good insights do not necessarily have to translate into good decisions, and wrong decisions are also possible. In particular, complex circumstances, limited time and inadequate mental processing power have been found to have an impact on the quality of decisions.

Can organizations use Business Intelligence and Analytics to compensate for the limitations of their managerial and organizational decision-making processes and, if so, how?

2.5.3 Decision to Value

If the quality of decisions can be improved through the use of Business Intelligence and Analytics, then the question of how organizations can create value from these decisions is trivial. We can highlight two uncertainties associated with converting decisions into value: (1) the uncertainty of successfully implementing decisions and (2) the uncertainty associated with the success of strategic actions. Business Intelligence, Analytics and resource allocation processes can mitigate these uncertainties. Although high quality decisions can be a good starting point, it is not certain that such decisions will be successfully implemented. Two criteria characterize good decisions. One criterion refers to the quality of the decision, for instance, whether the decision is able to achieve its objectives; the other refers to the acceptance of the decision, for instance, its acceptance by the subordinates and other stakeholders responsible for the successful implementation of the decision. Decision-making processes have an important influence on the acceptability of decisions, in particular, the level of influence and participation that subordinates and key stakeholders have on a decision has an important influence on its acceptance and, presumably, on its correct implementation. The use of Business Intelligence and Analytics can help to improve the quality of decisions, but it is unclear whether it can be used to somehow improve the acceptance of decisions. The information generation and decision-making processes associated with the use of Business Intelligence and Analytics often do not involve the key stakeholders in the functional areas that will be responsible for implementing these decisions.

How do decision making processes affect the implementation of decisions arising from the use of Business Intelligence and Analytics and how can they be used to improve decision acceptance?

An organization's search and selection capacity and its ability to orchestrate assets have an important influence on its performance. While it is clear that Business Intelligence and Analytics can improve an organization's search and selection capabilities, it is unclear how it can affect its asset orchestration capabilities. Organizational assets and resources are typically governed by formal or informal structures and managers will typically have to negotiate across organizational boundaries to access the assets they need to implement their strategies. Managers face uncertainty about the availability of resources to implement strategies, reflecting the important role of asset orchestration in the implementation of strategic actions. The key role of asset orchestration capabilities suggests that governance structures may need to evolve as organizations move toward greater confidence in the use of Business Intelligence and Analytics to support strategic decision-making. In general, strategy implementation is a business unit or functional responsibility. However, Business Intelligence and Analytics strategies supported by Business Intelligence and Analytics can increasingly rely on the use of IT resources and resources even during implementation. The roles of the CIO, the IT function, and the heads of business and functional units will need to evolve to address the blurring of institutionalized roles and structures. Organizations may need to focus more on information governance than on the governance of computer artefacts.[27]

A second source of uncertainty in converting decisions into value stems from the uncertainty of results. This refers to the uncertainty surrounding the results as a result of organizational actions. Organizations generally take strategic action in the hope of successful outcomes. However, actual results often deviate significantly from expectations and uncertainty of results is often taken into account in the decision-making process. Much of this uncertainty is beyond the control of the actors and the organization. It is not clear whether decisions supported by Business Intelligence and Analytics would be affected by uncertainty in results but, the effects of the use of Business Intelligence and Analytics on the quality and acceptance of decisions, could have an independent effect on reducing uncertainty in results.

A model to understand the joint effects of Business Intelligence and Analytics use and organizational decision-making processes on organizational performance is represented in the figure 2.12.

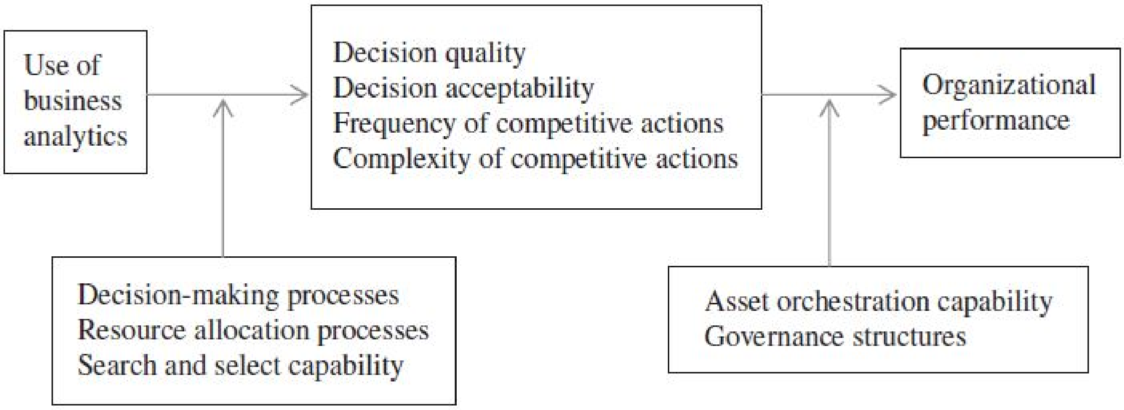

Use of business analytics
Decision quality
Decision acceptability
Frequency of competitive actions
Complexity of competitive actions
Organizational performance
Decision-making processes
Resource allocation processes
Search and select capability
Asset orchestration capability
Governance structures

Figure 2.12. A model to understand the joint effects of Business Intelligence and Analytics use and organizational decision-making processes on organizational performance[25]

Chapter 3

Big Data

Data is the basic resource of Business Intelligence. Probably, it is the increasing availability of data, the so-called Big Data, that provides the impetus to Business Intelligence. Over the years, companies have developed new and more complex methods that allow them to see the evolution of the market, their position on the market, the efficiency of the offer of their services and products, etc.. To be able to do this, you need a huge volume of data to be mined to generate valuable insights. Every year the data transmitted on the Internet grows exponentially. Cisco estimates that global annual IP traffic will reach 3.3 ZB (ZB; 1000 Exabytes [EB]) by 2021. In 2016, global IP traffic was 1.2 ZB per year or 96 EB (one billion Gigabytes [GB]) per month. By 2021, global IP traffic will reach 3.3 ZB per year, or 278 EB per month. Global IP traffic will increase almost three times in the next five years and will have increased 127 times from 2005 to 2021. Overall, IP traffic will grow at a compound annual growth rate (CAGR) of 24% from 2016 to 2021. The challenge will be not only to speed up Internet connections, but also to develop software systems that can handle large data demands in optimal time.[28] With the concept of Big Data, we refer to data that does not conform to the normal structure of the traditional database. Big Data consists of several types of key technologies such as Hadoop, HDFS, NoSQL, MapReduce, MongoDB, Cassandra, PIG, HIVE and HBASE (machine learning) working together to achieve the ultimate goal of extracting value from data. According to a recent market report published by Transparency Market Research, the total value of the large data was estimated at $6.3 billion in 2012, but by 2018, it is expected to reach the staggering level of $48.3 billion which is almost a 700 percent increase.[29]

Big Data comes from multiple sources and involves not only traditional and structured relational data, but all paradigms of unstructured data sources. Big Data is a data analysis methodology made possible by recent advances in technologies that support high-velocity data acquisition, storage and analysis. There is no uniform definition of Big Data. As described below, different stakeholders have provided different definitions:

Big Data: data captured by sensors, posts on social media sites, digital photos and videos, recordings of purchase transactions, GPS signals for mobile phones, etc..[30]

Big Data: extremely large set of data related to consumer behavior, social network posts, geotagging, sensor outputs.[31]

Big Data: data from everything, including clickstream data from the Web to genomic and proteomic data from biological research and medicine.[7]

Big Data: data sets with dimensions that go beyond the ability of typical database software tools to acquire, store, manage and analyze.[3]

Big Data: description of the massive amount of unstructured and semi-structured data that a company creates or data that would take too much time and money to upload to a relational database for analysis.[32]

Big Data: data that cannot be easily managed and processed.[33]

Big Data: data that cannot be uploaded to the running computer.[34]

Big Data: data that is too large to be inserted into a relational database and analyzed with the help of a desktop statistics/display package - perhaps, whose analysis requires parallel software on tens, hundreds or even thousands of servers.[35]

Big Data has three main features of Big Data: the data itself, the data analysis and the presentation of the analysis results. Then there are the products and services that can be wrapped around one or all of these elements of Big Data.[36]

Big Data: a cultural, technological and scientific phenomenon based on the interaction of (1)Technology: maximizing computing power and algorithmic accuracy to collect, analyze, link and compare large sets of data. (2) Analysis: drawing on large data sets to identify models in order to make economic, social, technical and legal claims. (3) Mythology: the widespread belief that large data sets offer a superior form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity and accuracy.[37]

The most common approach to defining Big Data is the notion of 'V'. Big Data in terms of 3Vs[38] [39] [40] [41]: volume or the large amount of data consuming a huge amount of data or involving a large number of records; velocity, which is the frequency or speed of data generation and the frequency of data transmission; and variety, to highlight the fact that data is generated from a wide variety of sources and formats, and contains multidimensional data fields that include structured and unstructured data.[41] Another 'V', value, is includedto highlight the importance of extracting economic benefits from available Big Data, which introduced a fifth dimension, veracity, to highlight the importance of data quality and the level of trust in various data sources.[42] [43] [44]

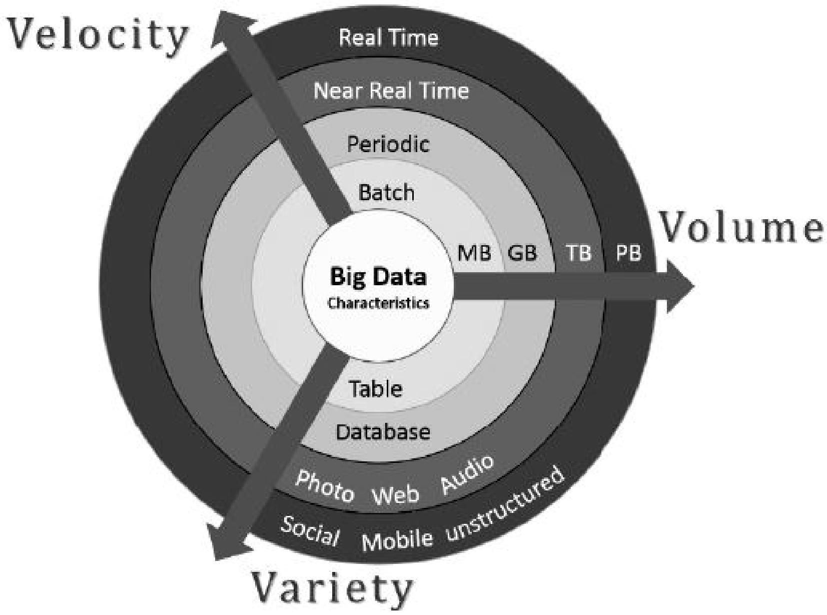


Figure 3.1. Dimensions of Big Data[][52]

These dimensions are described below:

Volume: Refers to the size of Big Data, which is in the order of terabytes to petabytes and above, which is manipulated and analyzed by companies to achieve the desired results. The volume of data used to play an important role in storage and processing. However, the problem of storage capacity has become less pressing due to the rapid advancement of storage technologies and falling prices per gigabyte. Modern storage technologies enable users to implement cost-effective storage solutions for internal deployments. In addition, organizations can also outsource their storage needs. Data volume, however, plays an important role in processing considerations because data volumes have increased faster than the computing power of processing systems.

Velocity: Refers to the rate of data generation and the speed of analysis and intervention. The diffusion of digital devices such as smartphones and sensors has led to an enormous rate of data creation and is leading to an increasing need for real-time analysis and evidence-based planning.

Variety: Represents the type of data stored, analyzed and used. Variety refers to different types of data such as structured, semi-structured and unstructured data. The data available in spreadsheets or relational databases in the form of tabular data are structured. Images, audio, video, video, text, chat messages are not structured. Extensible Markup Language (XML), text language used for data exchange on the web is semi-structured.[45] From an analytical point of view, data variety could be considered the biggest obstacle to the effective use of large volumes of data due to incompatible data formats, misaligned data structures and inconsistent data semantics that make all this data unreadable by all users accessing it and allow you to create ambiguous results.

Veracity: Refers to the unreliability inherent in some data sources. The need to address inaccurate and uncertain data is another aspect of large data, which could be managed using tools and analyses developed for the management and extraction of uncertain data.

Value: Refers to the quality of stored data and its further use. Data received in its original form usually has a low value compared to its volume, but this value can be extracted by analyzing these huge volumes of data.

SAS has introduced two additional Big Data dimensions, variability and complexity:

Variability: Variability refers to the variation of data rates. Often, the great speed of the data is inconsistent and has periodic peaks and troughs.

Complexity: Complexity refers to the fact that large data are generated through a myriad of sources. This imposes a critical challenge: the need to connect, compare, compare, purify and transform data received from different sources.

There are no universal benchmarks for volume, variety and speed that define Big Data. The limits that define the limits depend on the size, sector and location of the company and these limits evolve over time. It is also important that these dimensions are not independent of each other. When one dimension changes, another dimension also changes. However, there is a 'three-V turning point' for each company beyond which traditional data management and analysis technologies become inadequate to achieve timely intelligence. The three-V tipping point is the threshold beyond which companies start dealing with Big Data. Companies should therefore compensate the future value expected from Big Data technologies with implementation costs.[46]

Large data are aggregated from internal and external data sources which we should verify for reliability. Large data can also be dirty (inaccurate, improper or incomplete), so this should be considered to build a strategy for cleaning dirty data, for example, when it should be cleaned or not. Sometimes we need to analyze the streaming data in real time, which involves many complexities about storage and processing.

Big Data also refers to the way information is handled. To store and process large amounts of data you need a set of tools that can explore and sort large volumes of data. The methods of sorting and processing data differ from one type of data to another.

Data analytics is about extracting knowledge and insights from Big Data. Big Data has no value if it cannot be used to obtain information for decision making. Thus, there is a need to process the huge volume of rapidly evolving diversified data into something that is meaningful to extract insights. The overall method of processing Big Data to obtain insight is divided into five stages in which each stage has two main sub-processes: Big Data management (acquisition, storage and preparation of data for analysis) and Big Data Analytics (analysis and acquisition of intelligence from data). The first stage concerns the identification of the databases that will enable the structured data set to be analyzed and stored in the data warehouse. The second stage is the way to form the data set stored in the data warehouse, which will be used to perform the analysis through the data processing and to generate knowledge to be discovered in the databases. This stage requires the use of a concept known as Extract, Transform and Load (ETL). The third stage is the construction of the data warehouse. So, once we have it, the fourth stage is to perform data analysis. The fifth and final stage is to make decisions that support the business process based on the information obtained from the data (quantitative aspect) and the experience of the decision-makers (qualitative aspect). The analysis of large data can have two perspectives: decision oriented which is also traditional business intelligence where we analyze selective sub-sets and representations of large data sources and then we try to apply the results obtained to make business decisions and action oriented used for a rapid response, when some specific types of data or models are detected and some actions are necessary.[47]

3.1 Characteristics of Big Data

In this paragraph we will describe the main characteristics of Big Data.

3.1.1 Quality vs Quantity

An emerging challenge for large data users is quantity vs quality. Users acquire and have access to more data, but often want even more because they believe that, with sufficient data, they will be able to perfectly explain any phenomenon of interest to them. On the contrary, a large data user can concentrate on quality, which means not having all the data available, but having a very large amount of high quality data that can be used to draw precise and high value conclusions.[48] The data that is stored is not always totally useful or is not already ordered. To become useful, the data must be ordered so that it can be analyzed later and can be of any value. Sorting and cleaning data is a difficult challenge, so companies usually hire people who have the skills to manipulate the data that managers and managers will use further. Lack of Big Data sorting skills is likely to lead to wrong results. There is another problem: data quality. To achieve this, the architecture of the source that collects the data must be able to sort it logically, in a way that is understandable to the software that orders it. Various data characteristics can affect data quality, such as completeness, accuracy and timeliness. Completeness refers to missing values in the data. Data without missing values are complete. Characteristics such as unavailability or imperceptibility can cause missing values in the data. Data accuracy refers to the precision with which data are expressed.[49]

3.1.2 Speed

Time-to-information is critical when considering near real-time processes that generate almost continuous data. Data rate has two components: inflow rate and outflow rate. The data rate indicates the data acquisition rate or the maximum speed of the input channel. The data flow rate indicates the data output rate from a system or the maximum speed of the output channel. Data rate is usually measured in bits or bytes per second.[49 Different data types may require different data rates.

3.1.3 Privacy and Security

Privacy and security have many implications and affect both individuals and businesses. Individuals have the right to control information that may be disclosed about them. This information can be used by companies to develop their marketing strategies and extend their services. Individual privacy is still a sensitive issue that can only be resolved by allowing people to choose which information about them to disclose or not. For companies, the privacy issue is more related to the sensitive data they work with. Companies can store their information on cloud systems, in-house systems or a hybrid solution. Data storage on cloud systems is more cost-effective for businesses in terms of the cost and speed of processing the operations required. Data security still remains an important issue here, and to address this issue, some companies choose to build their own infrastructure to store and manipulate the data they hold. For smaller businesses this may be a solution, but in most cases the costs are high. In addition, maintaining this type of system requires trained staff and the more the company grows, the more it will need to be added to the infrastructure. The only advantage of this solution is privacy. Information sharing is one of the most valuable features of development. Every person and every company has a large amount of information at their disposal that they can use for their own purposes. As far as people are concerned, there is a difference between what is personal and what can be made public. The question of what is personal and what is public lies primarily in the point of view of the services they use. As far as companies are concerned, the reason why they don't want to share their Big Data warehouse is more related to their competitiveness and the sensitive data they have at their disposal. Otherwise, if this line of demarcation is crossed, each company will have more data that it can analyze to obtain more accurate results and better planning is possible. If companies share the information in their possession on the current market situation and on possible customers and approach strategies, the degree of development will be drastically reduced and they will start to focus on how to maintain their current customers. A more transparent representation of the current information a company holds will benefit everyone. In this way, the type of information and the way in which it is structured can help the further development of software systems that can be standardized and can work with all types of data imported from various sources.[50]

3.1.4 Sensitivity

Data sensitivity is crucial for most organizations. It expresses whether the data contains sensitive information such as personal information, confidential inside information, confidential information, non-disclosure information, etc., or whether it is not sensitive. Sensitivity determines the requirements for data processing. If the data contains personally identifiable information, organizations must comply with regulatory requirements. Similarly, if the data contains confidential internal information such as internal know-how or financial data, it is in the interest of the organization to adequately protect it from external exposure and compromise. Such data must be stored within the company and not on external cloud providers. In addition, its internal exposure must be protected by measures such as encryption and its access must be limited only to relevant members of the organization.[49]

3.1.5 Diversity

Data diversity refers to the spectrum of different types of elements within the data. The data can be of various types such as audio, video or multimedia data, text data, location data, time data, etc.. Different types of data generally require different management and impose different requirements in terms of allocation capacity, processing speed and other issues. Data diversity may or may not be advantageous in different cases, because data diversity adds wealth, but also adds complexity to processing and maintenance.[49]

3.2 Big Data types

Big Data consists of structured, semi-structured and unstructured data as shown in figure 3.2.

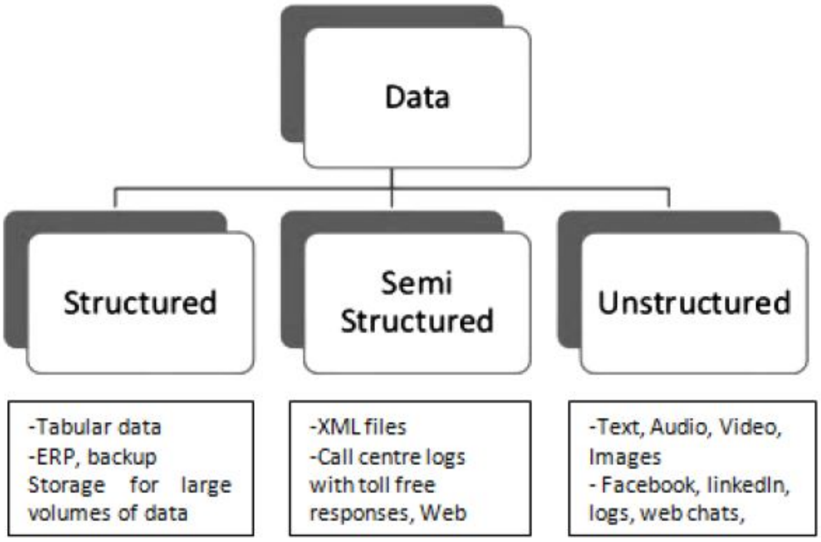

Data
Structured
Semi Structured
Unstructured
-Tabular data
-ERP, backup Storage for large volumes of data
-XML files
-Call centre logs with toll free responses, Web
-Text, Audio, Video, Images
- Facebook, linkedIn, logs, web chats,

Figure 3.2. Types of Big Data[52]

This is referred as structural heterogeneity in Big Data. Data structure underlines the degree of data organization. Structured data has a high degree of organization, whereas unstructured data lacks sufficient degree of organization. Unstructured data is more suitable for human processing-for instance, text document or e-mail. Structured data is suitable for machine processing. Structured data can be seamlessly included into conventional databases and managed with database tools. Unstructured data is generally more difficult to process since it usually requires pre-processing before conventional tools can process it. Only 5% of existing data is structured.[51] Structured data is tabular data which exists in the form of relational database and spreadsheets. Data having no particular format, schema or structure is said to be unstructured, it can be in any form such as text, audio, images and video for examples, word files, PDF files, content of blogs, forums, tweets, emails, web pages, audio files, video files, images etc. Major contributors to unstructured data are social networks and sensors. In between structured and structured data, there exist semi-structured data which has no strict standard for example Extensible Markup Language (XML) files, web logs, sensor logs etc. Unstructured data adds complicacy to the analytics process but in order to have a competitive advantage over other organizations, the potential information locked in unstructured data needs to be extracted by organizations. The real value of unstructured data can be leveraged by deconstructing it and converting it into semi-structured content which can be used to gain insights.[52] For decision making, the decision makers cannot rely completely on structured data because they would miss the vast amount of information available on the open source unstructured web content.[51] For predictive analytics, both structured and unstructured data are required. Without their integration and analysis, a complete picture cannot be obtained. A wealth of information is stored in unstructured data which if leveraged properly can make a big difference to business. One major challenge in unstructured data analytics is that unstructured data is noisy, so it needs to be cleaned first and then it is analyzed and integrated with structured data.[53]

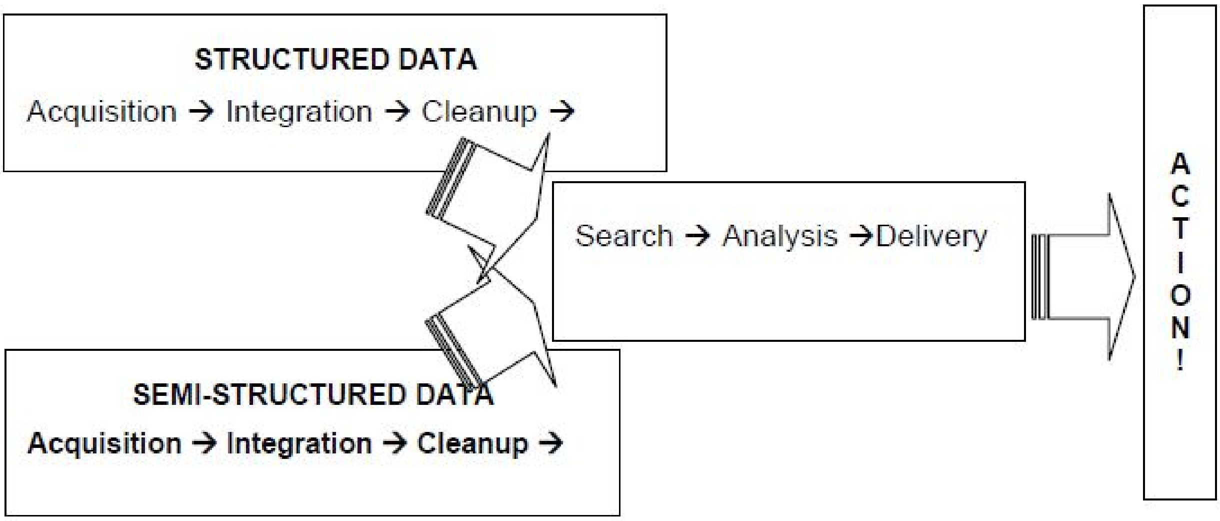
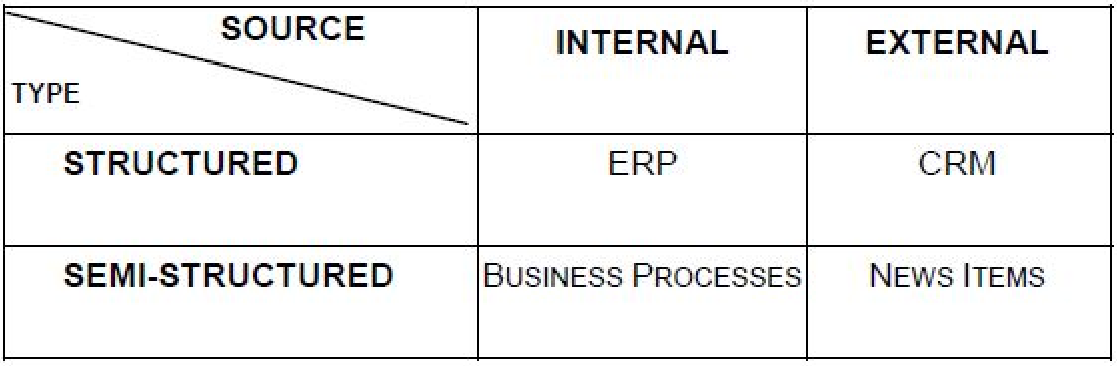

STRUCTURED DATA
Acquisition → Integration → Cleanup →
Search → Analysis →Delivery
SEMI-STRUCTURED DATA
Acquisition → Integration → Cleanup →
ACTION!

Figure 3.3. Business Intelligence Data Framework for Structured and Unstructured Data[19]

Figure 3.3 shows a framework that integrates the structured and semi-structured data required for Business Intelligence. Semi-structured data are equally important, if not more, as structured data for taking action by planners and decision makers. Another implication is that the process of acquisition, cleanup, and integration applies for both structured and semi-structured data. To create valued information, the integrated data are searched, analyzed, and delivered to the decision maker. In the case of structured data, analysts use Enterprise Resource Planning (ERP) systems, ETL tools, data warehouses, data-mining tools, and on-line analytical processing tools. But a different and less sophisticated set of analytic tools is currently required to deal with semi-structured data.[19]

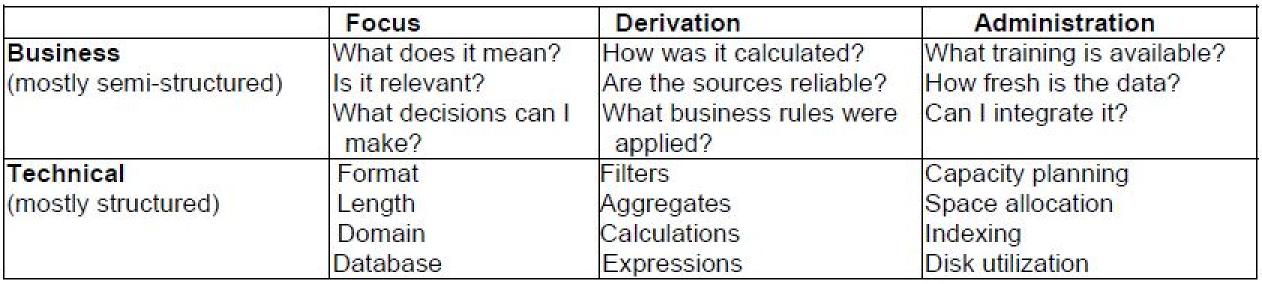
Structured and semi-structured data types can be further segmented by looking at the internal and external data sources of the organization.[19] These two dimensions,data type and data source,are illustrated in Figure 3.4. The transition between structured and semi-structured data types and between internal and external data sources is not defined severely.



SOURCE TYPE	INTERNAL	EXTERNAL
STRUCTURED	ERP	CRM
SEMI-STRUCTURED	BUSINESS PROCESSES	NEWS ITEMS

Figure 3.4. Data Type/Source Matrix[19]

Since it must deal with both structured and semi-structured data simultaneously, Business Intelligence's data architecture is business rather than technically oriented. While technical data architectures focus on hardware, middleware, and DBMSs, BI data architecture focuses on standards, metadata, business rules, and policies.[54]



	Focus	Derivation	Administration
Business (mostly semi-structured)	What does it mean? Is it relevant? What decisions can I make?	How was it calculated? Are the sources reliable? What business rules were applied?	What training is available? How fresh is the data? Can I integrate it?
Technical (mostly structured)	Format Length Domain Database	Filters Aggregates Calculations Expressions	Capacity planning Space allocation Indexing Disk utilization

Figure 3.5. Architecture for Structured Data[19]

Typical Business Intelligence architecture for structured data centers on a data warehouse. The data are extracted from operational systems and distributed using Internet browser technologies (Figure 3.6). The specific data needed for Business Intelligence are downloaded to a data mart used by planners and executives. Outputs are acquired from routine push of data from the data mart and from response to inquiries from Web users and OLAP analysts. The outputs can take several forms including exception reports, routine reports, and responses to specific request. The outputs are sent whenever parameters are outside pre-specified bounds.[19]

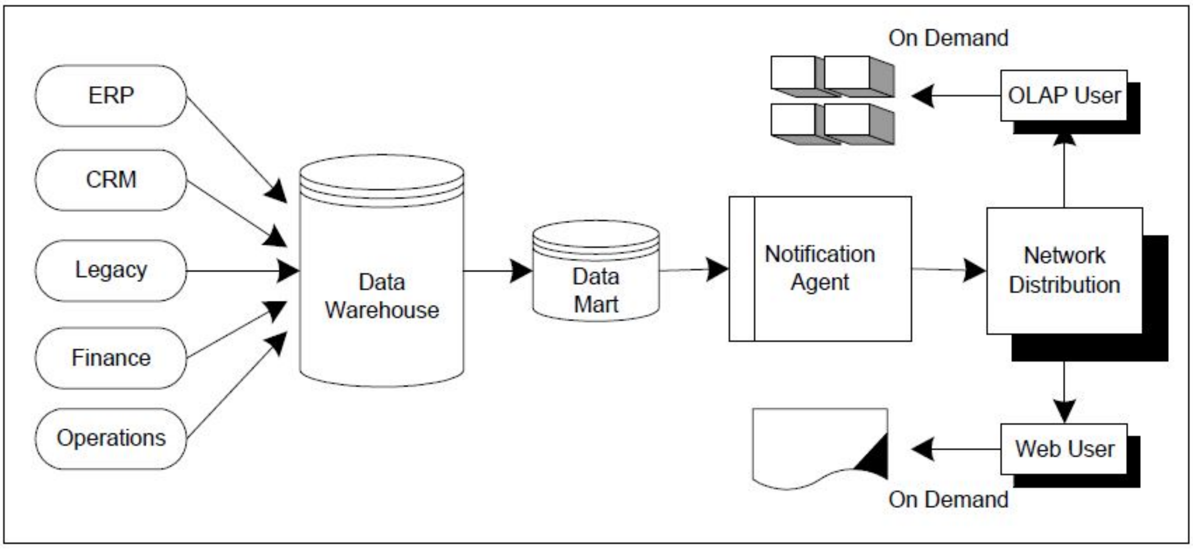


Figure 3.6. Architecture components for Structured Data[19]

Business Intelligence architecture for unstructured data (Figure 3.7) includes business function model, business process model, business data model, application inventory, and meta data repository.[54]

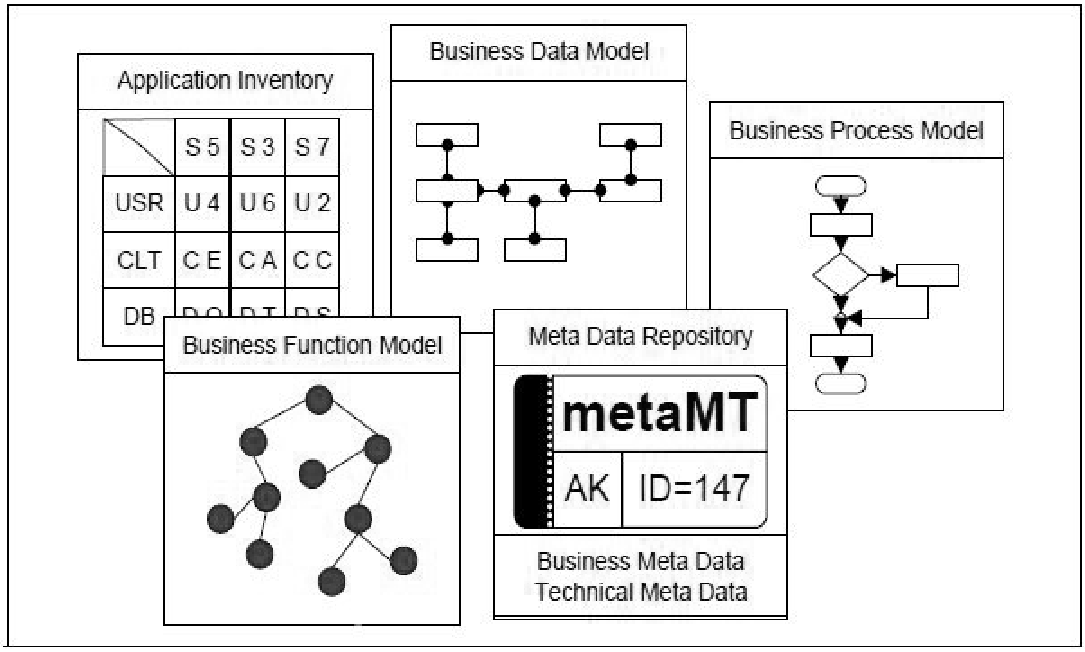
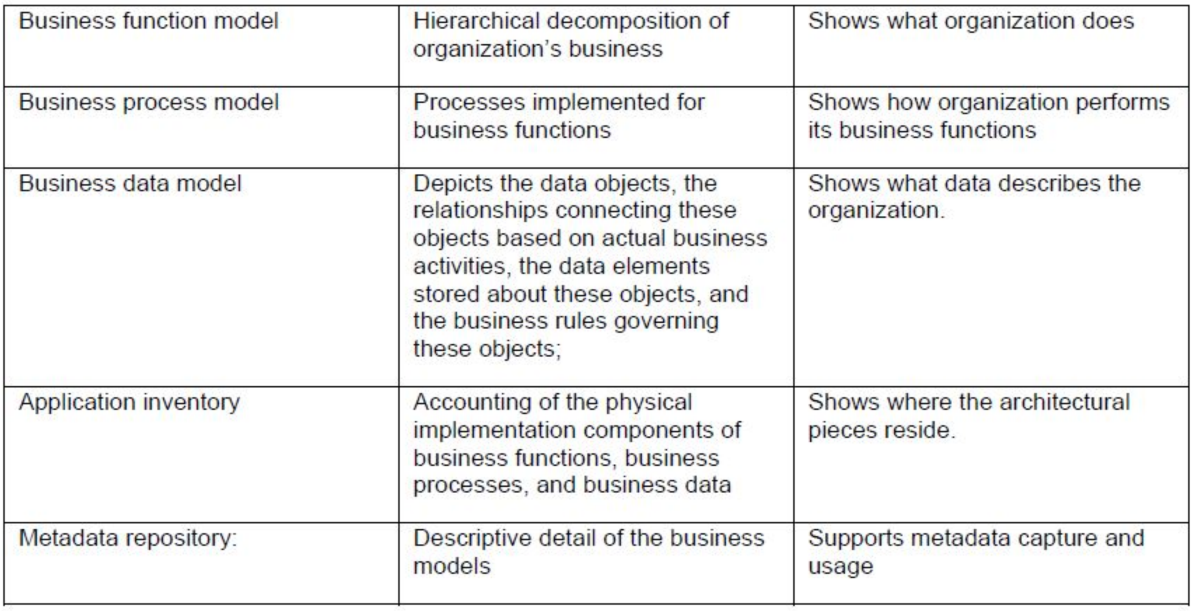


Figure 3.7. Architecture for Unstructured Data[19]

Figure 3.8 describes the five components.



Business function model	Hierarchical decomposition of organization's business	Shows what organization does
Business process model	Processes implemented for business functions	Shows how organization performs its business functions
Business data model	Depicts the data objects, the relationships connecting these objects based on actual business activities, the data elements stored about these objects, and the business rules governing these objects;	Shows what data describes the organization.
Application inventory	Accounting of the physical implementation components of business functions, business processes, and business data	Shows where the architectural pieces reside.
Metadata repository:	Descriptive detail of the business models	Supports metadata capture and usage

Figure 3.8. Architecture components for Unstructured Data[19]

3.3 Technologies

To obtain good insights and mine the amplitude of information that Big Data make available, users had to develop tools capable of creating the expected results. A common implementation that handles Big Data is MapReduce. MapReduce consists of two things: mapping and reducing. Mapping a certain dataset means to restructure it into a different set of values. Reducing is a process that takes several mapped outputs and forms a smaller set of tuples. The most popular technology that is able to mine and sort data is Hadoop. Being open source software, Hadoop is the most implemented solution for handling Big Data. It has enough flexibility to work with multiple data sources, or even assemble multiple systems to be able to do large scale processing. Hadoop also use HDFS (Hadoop Distributed File System) that has the role to split data into smaller blocks and distribute it throughout the cluster. Large companies use Hadoop as a starting point in order to deploy other solutions. DB2 from IBM is a fast and solid data manipulating system. It has feature that reduces the cost of administration by doing an automated process that increases storage efficiency and improves performance.[50] Oracle, on the other hand, comes with a complete system solution for companies. It starts from the basic ideas of Big Data sources, which can be traditional data generated by ERP systems, sensor data and social data, defined by the feedback that the company receives from customers and other sources. The solution given by Oracle is to create a system from top to bottom, based on NoSQL. A NoSQL database is capable of handling various types of data that traditional relational databases are unable to handle and lose data consistency. NoSQL derives from ‘Not only SQL’, which means that it allows regular SQL queries to be executed. Oracle’s solution is presented in 4 steps: Acquire, Organize, Analyze and Decide.[55] All the steps combine different solutions like HDFS, NoSQL, Oracle Big Data connectors and Cloudera CDH. CDH or Cloudera’s Distribution Including Apache Hadoop, offers batch processing using MapReduce and HDFS, interactive SQL which allows users to execute SQL queries and interactive search.[56] All these key features that Cloudera offers are solutions that allow users to navigate through clusters and retrieve data that they need. SAS offers multiple solutions to overcome Big Data mining and analysis. It also tries to cover all that is necessary for a company to create value from stored data. One solution is SAS DataFlux which is a data management solution that can provide users the right tools for integrating data, mastering data and data quality. It also allows access and use of data across company and also provides a unified set of policies in order to maintain data quality. SAS also provides high-performance analytics solution that is providing the company good insights from analyzing data in a structured, easy to read, report. This is basically one of the main goals when working with Big Data, to get best insights from quality data. Also SAS provides analytics solution that is based on a drag-and-drop system which can provide easy to understand and customized reports and charts. SAS is more oriented in providing software solutions to help companies benefit from data that they have stored.[57] The problem of handling Big Data, it also has a great impact over hardware systems and their capability of processing. So, the two of them, software and hardware solutions, create a complete Big Data system that can be viable and will produce the expected outcome.

3.3.1 Hadoop System

Hadoop is the name of a software system developed to process Big Data. It is open source software written in a high level programming language (Java) and the source code is freely available for download from the Internet. The development of Hadoop was funded by Apache Software Foundation. Hadoop system is designed to be used with large clusters of commodity processors each with both local main memory and large secondary memory, normally disks. It consists of two major parts: one part for storing files to be processed and the other to process in parallel the data stored in these files. They are called Hadoop Distributed File System and MapReduce respectively.[58]

3.3.2 Hadoop Distributed File System (HDFS)

HDFS is designed to store large files in which data is written once and read many times. Files are split and stored as 64 bytes or 128 bytes blocks. Three sets of blocks of the same file are stored in three different computers, which may be in different racks, to tolerate faults. The HDFS system works in a master-slave mode.[58] The master process is executed in a node (a computer) called name node which maintains information about files so that it has a global view of the file system. If the name node fails, the whole system will stop functioning. The slave nodes, called data nodes, obey commands sent by the name node on data blocks stored locally. HDFS is used to store both input and output of MapReduce jobs and intermediate data if any.

3.3.3 MapReduce

MapReduce processes data stored in HDFS in parallel. The fundamental concept of MapReduce is to divide problems into two parts: a map function that processes source data into sufficient statistics and a reduce function that merges all sufficient statistics into a final answer. The map tasks and reduce tasks can both be run in parallel, but reduce has to follow map run to combine the results of the map phases.[59] MapReduce provides the fastest, most cost-effective and most scalable mechanism for returning results. Today, most of the leading technologies for managing Big Data are developed on MapReduce.

3.3.4 Implementation of MapReduce in Hadoop

There are two types of processes which control users' programs in Hadoop. One is called a job tracker that runs on the master node and the other is called a task tracker that runs on slave nodes. The job tracker splits a job submitted to the system into map tasks and reduce tasks. These are scheduled to task trackers. The scheduler tries to assign map tasks close to the node where the requisite data is located. The job tracker monitors the tasks that run on slave nodes. When a task tracker has no work to do, the job tracker picks a new task from the queue and assigns it to that task tracker. When a reduce task is to be performed it is assigned to a task tracker which has completed the appropriate map task. The job tracker waits for a response from a task tracker, called a heart-beat signal, before any more tasks are assigned to it. Due to the fact that inexpensive nodes are used as task trackers in the cluster, there is high probability that one of them would fail. Thus, there must be a mechanism to handle a failed task tracker. The system does it by transferring a task assigned to a failed Task tracker to a healthy node. The fact that a node has failed is detected by the job tracker if it does not receive a task completion signal from the task tracker to which it had assigned the task. If a task tracker, for some reason, performs an assigned task slowly, the run time system of Hadoop sends the same task to another node, called a Backup task. The job tracker takes the results from the task tracker which finishes first and aborts the other. The complexity of fault handling is hidden from the user.[58]

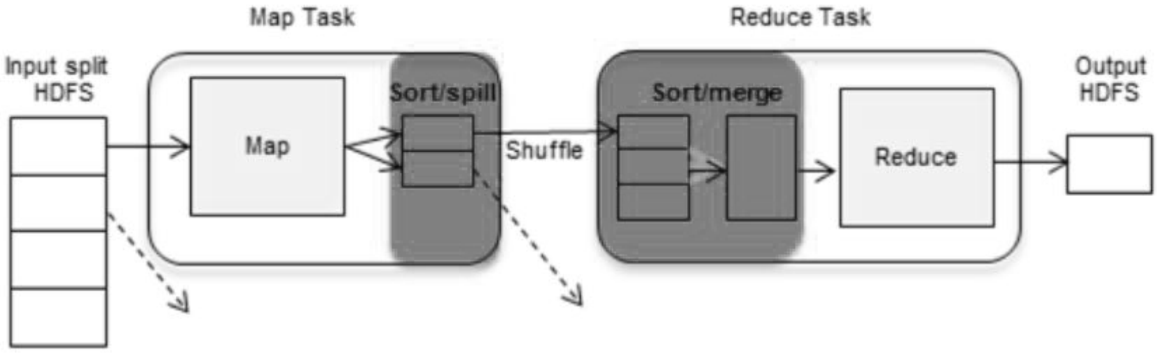


Figure 3.9. MapReduce in Hadoop[88]

3.4 Big Data Analytics

Big Data Analytics reflect the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods.[59] Organizations now routinely generate data of unprecedented scope and complexity. Analytics has become vital to realize the full value of Big Data to improve the companies' business performance and increase their market share. The tools available to handle the volume, velocity, and variety of Big Data have improved a lot over the years and in general are not prohibitively expensive, and much of the software is open source. We define Big Data Analytics as technologies (database and data mining tools) and techniques (analytical methods) that a company use to analyze large scale, complex data for various applications intended to leverage its performance in various dimensions. Big Data Analytics is the process of collecting, storing and analyzing huge volumes of high velocity diverse data to extract hidden patterns and meaningful insights. Its output can be used for optimize the decision making by providing meaningful insight. Big Data Analytics can help in making discovery and research in various areas such as space science or medicines, in analyzing various trends to make predictions, in predicting customer demand, in making future predictions for sales etc. Big Data Analytics enables organizations to create real-time intelligence from Big Data.[52] Devices such as smartphones and sensors have accelerated the rate of growth of data generation and the need of real-time analytics.

The process of extracting information and insights from Big Data can be divided into five stages.[60] The first stage is the identification of the database with the acquisition and recording of data that will allow the structured data set to be analyzed and stored in the data warehouse. The second stage is the way of forming the data set (stored in the data warehouse), which will be used to generate knowledge discovering in databases (KDD). This stage requires the use of a concept known as extract, transform, and load. The third stage is integration, aggregation and representation to build the data warehouse. Once we have it, the fourth stage is modeling and analysis. Finally the fifth stage is the interpretation of the output by making decisions based on the information obtained from the data (quantitative aspect) and the experience of the decision makers (qualitative aspect). These five stages form two main sub-processes: data management and analytics. Data management involves processes and supporting technologies to acquire and store data and to prepare and retrieve it for analysis. Analytics, on the other hand, refers to techniques used to analyze and acquire intelligence from big data.

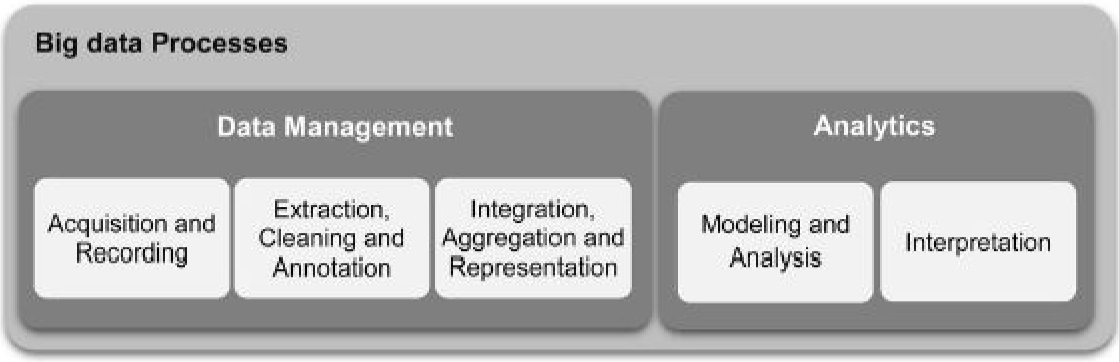

Big data Processes
Data Management
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Analytics
Modeling and Analysis
Interpretation

Figure 3.10. Processes for extracting insights from Big Data[46]

3.4.1 Big Data Analytics Categories

Analytics can further be categorized into Descriptive Analytics, Predictive Analytics, Exploratory or Discovery Analytics and Prescriptive Analytics.[52]

- **Descriptive Analytics** is connected with Business Intelligence and is based on historical data.[58] Practically it tells what happened in the past and what is happening in the present in an understandable form. Data gathered is organized as bar charts, graphs, pie charts, maps, scatter diagrams, etc., for easy visualization which gives insight into what the data implies. A typical example of descriptive analytics is presentation of population census data which classifies population across a country by sex, age groups, education, income, population density and similar parameters.

- **Predictive Analytics** uses statistical data and is meant for making predictions for future scope based on the data.[58] Predictive Analytics is the use of past/historical data to predict and to forecast on future trends. This analysis uses of the statistical models and machine learning algorithms to identify patterns and learn from historical data. Predictive Analysis can also be defined as a process that uses machine learning to analyze data and make predictions.[59] It extrapolates from available data and tells what is expected to happen in the near future. The tools used for extrapolation are time series analysis using statistical methods, neural networks, and machine learning algorithms. One major use of predictive analytics is in marketing by comprehending customers' needs and preferences.

- **Exploratory**or **Discovery Analytics** finds unexpected relationships among parameters in collections of Big Data. Collection of data from a variety of sources and analyzing them provides additional opportunities for insights and serendipitous discovery.[58] One of the major applications is discovering patterns in customers' behaviour by companies using their feedback, tweets, blogs, Facebook data, emails, sales trend etc..

- **Prescriptive Analytics** is used to find out the optimized solution for the concerned problem having a set of constraints.[61] This identifies, based on data gathered, opportunities to optimize solutions to existing problems. In other words, the analysis tells us what to do to achieve a goal. One of the common uses is in airlines' pricing of seats based on historical data of travel patterns, popular origins and destinations, major events, holidays, etc., to maximize profit.[58]

Handling heterogeneity in Big Data Analytics is difficult due to heterogeneity, complexities arise in Big Data Analytics because of other factors as well such as noise accumulation, spurious correlations, scale etc.. A sample of data set is selected and relationships are found out of the sample. The conclusion is then generalized to the entire population but in case of Big Data, the sample is massive and represents the entire population.[52]

3.5 Big Data Analytics Techniques

In the following section we are going to discuss some of the relevant techniques for analyzing Big Data (structured and unstructured):

3.5.1 Text Analytics

It refers to the techniques used for extracting or retrieving information or insights from text based data such as emails, documents, advertisements, forums, blogs, news content, social network content, website content, call center logs, customer comments and reviews, tweets etc. It involves statistical analysis, computational linguistics and machine learning.[52] Text analytics enables to gain meaningful insights to support decision making in businesses. There are various methods to approach text analytics; Information Extraction or IE techniques convert unstructured text into structured content. Information extraction processes have two sub tasks namely entity recognition and relation extraction.[62] ER or Entity Recognition finds name information in text data and classifies the information into categories. RE or Relation Extraction finds and retrieves the semantic relationships between entities such as persons, company, etc from the text data. Another method used in text analytics is text summarization.[63] Text Summarization enables to produce a summary from a single or multiple text sources. There are two different approaches to the summarization methods, extractive approach and abstractive approach. The Extractive approach creates summary by extracting original text units or sentences from the source document. This approach requires no understanding of the document. On the other hand, the Abstractive approach extracts semantic information from the document. The summaries generated by this approach do not contain the original text units necessarily. This approach uses advanced Natural Language Processing Techniques or NLP techniques to analyze the text and produce the summary. Another method is the Question Answering techniques or QA techniques which provide answers to questions formulated in natural language. This kind of systems are usually applied in various fields such as finance, academics, healthcare and marketing. They are based on complex natural language processing techniques. Question answering techniques have three different approaches: information retrieval, knowledge-based, hybrid.[52] In the Information Retrieval or IR approach, first there is a processing of the questions to create an appropriate query from the question, then there is a processing of the documents to extract the relevant pre-written content from the existing documents, after that there is the processing of the answers to retrieve the candidate answers which are ranked and the answer at the top of the ranking is returned as a solution or output. The knowledge-based approach produces the semantic information of the question and then is used for querying. This approach is suitable for restricted domains without large volumes of pre-written content. The hybrid approach semantically analyzes questions using the Knowledge-based approach while the Information Retrieval approach generates candidate solutions. Sentiment Analysis or Opinion Mining methods are those that analyze opinionated texts that consist of opinions or points of view of people towards entities such as individuals, products, brands, companies or events.[64] It is applied in areas such as marketing, political and social sciences and finance. Sentiment analysis is performed at the document, sentence and appearance levels.[52] The techniques at the document level deduct whether there is a positive or negative feeling about a single entity throughout the entire document. In techniques at sentence level, the single feeling is determined by the sentence size. This is more complicated than document-level techniques. Aspect-based techniques determine all the entity-specific feelings about the different aspects of the entity in the document. This is useful when customers' opinions about different product characteristics are examined.

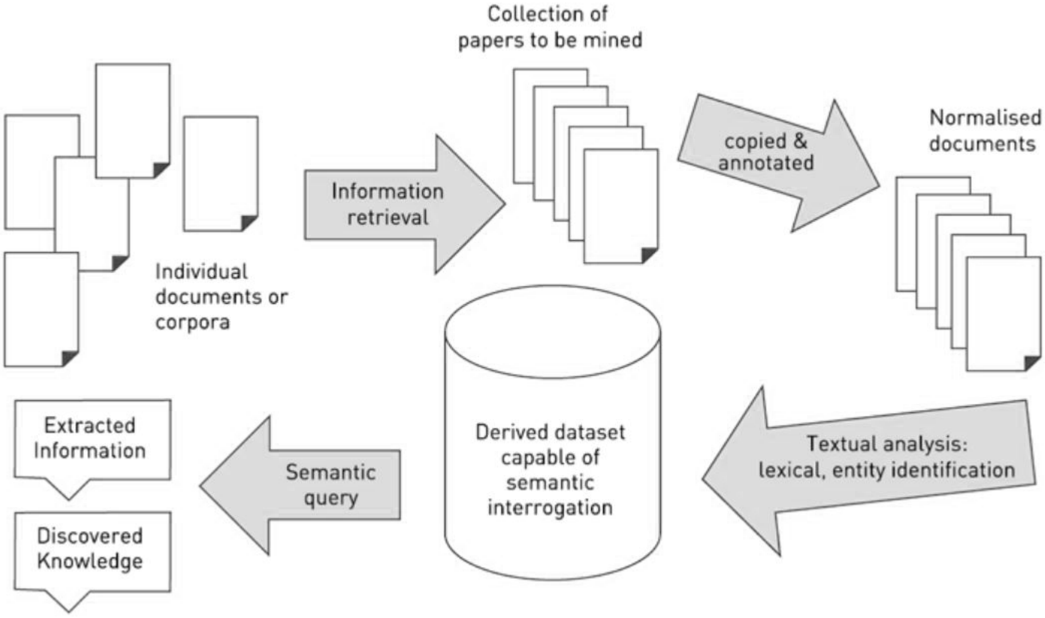

Collection of papers to be mined
Individual documents or corpora
Information retrieval
copied & annotated
Normalised documents
Textual analysis: lexical, entity identification
Derived dataset capable of semantic interrogation
Semantic query
Extracted Information
Discovered Knowledge

Figure 3.11. Text Analytics[89]

3.5.2 Audio Analytics

It is the process of analyzing and extracting or retrieving information from unstructured audio content or data. It is known as Speech Analytics when applied to human speech. It is implemented in areas such as customer call centers to analyze the content of recorded customer calls and healthcare where is useful for diagnosing and treating diseases that affect patients' communication patterns such as depression, cancer and schizophrenia.[65] There are two approaches to language analysis: the transcript approach and the phonetic approach. The Transcript approach, also known as the Large Vocabulary Continuous Speech Recognition (LVCSR) approach, is further divided into two phases: indexing and searching. In the Indexing phase, Automatic Speech Recognition (ASR) algorithms are used to match sounds with words that are identified with the help of a predefined dictionary. If the system fails to do so, the closest word is returned. The output of this system is a file containing the sequence of words spoken in the speech. In the Searching phase, standard text-based methods are implemented to find the search term. The Phonetic approach works with sounds or phonemes that help to distinguish one word from another. This approach can also be further divided into two phases which are phonetic indexing and searching. In the Indexing phase, the input speech is translated into a sequence of phonemes. In the Searching phase, search terms are phonetically represented by searching for the result obtained from the previous phase.[52]

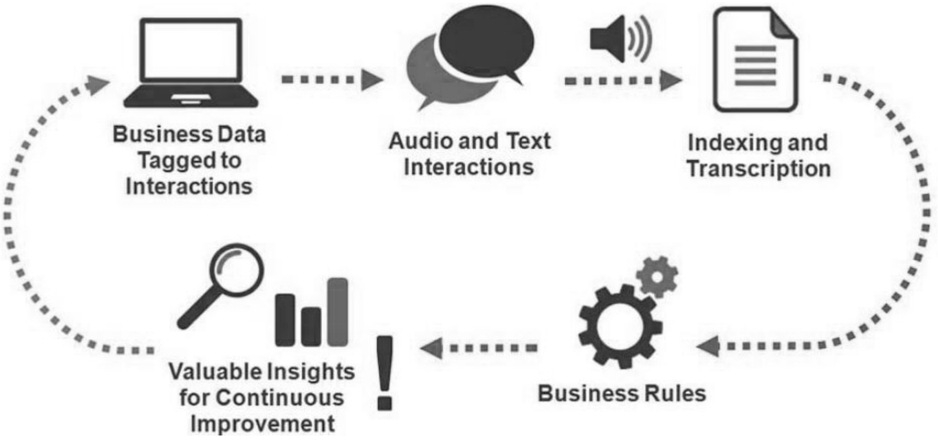

Business Data Tagged to Interactions
Audio and Text Interactions
Indexing and Transcription
Business Rules
Valuable Insights for Continuous Improvement

Figure 3.12. Audio Analytics[90]

3.5.3 Video Analytics

It is the process of monitoring, analyzing and acquisition of significant insights from the video streams. It is also called Video Content Analysis (VCA)[66]. There are techniques for processing real-time and pre-recorded video. The main contributors to the video data are Closed-Circuit Television (CCTV) cameras and video sharing websites such as YouTube. One of the main challenges in video analysis is the huge size of video data. Over two thousand pages of text data equals one second of high-definition (HD) video.[3] The high volume of video content is a challenge, but also an opportunity with the help of large data technologies. Intelligence can be gained from thousands of hours of video content. Video analysis has been applied in areas such as automated security, monitoring and surveillance systems to detect violations, identify thefts, detect waste areas, control suspicious activities or objects left unattended. When such activities are detected, security agents can be alerted in real time or the necessary automatic actions can be taken, such as turning on audible alarms, blocking doors or turning on lights. Work-based surveillance systems are expensive and less effective than automatic systems.[67] The content of the CCTV cameras located at the point of sale can be analyzed to extract information. This information can help improve marketing and operations management. Another video analytics application is Automatic Video Indexing and Retrieval for easy video search and retrieval. For video search and retrieval, there is a metadata-based approach in which relational database management systems are used. In the approach of soundtracks and transcripts, video indexing can be performed by applying audio analysis and text analysis. There are two different approaches to system architecture in video analysis: server-based architecture and edge-based architecture.[68] In the server-based architecture, a centralized server performs video analysis on the video captured by each camera. The limit to this approach is that the bandwidth is limited and therefore the videos are compressed by reducing the image resolution and frame rate. And therefore, the accuracy of the analysis is compromised due to the loss of information. But maintenance is easier in this approach and also provides economies of scale. In edge-based architecture, video analysis is applied to video content captured by the camera locally. In this approach there is no loss of information and content analysis is more effective. The maintenance of such systems is expensive and the processing power is lower than with server-based systems.

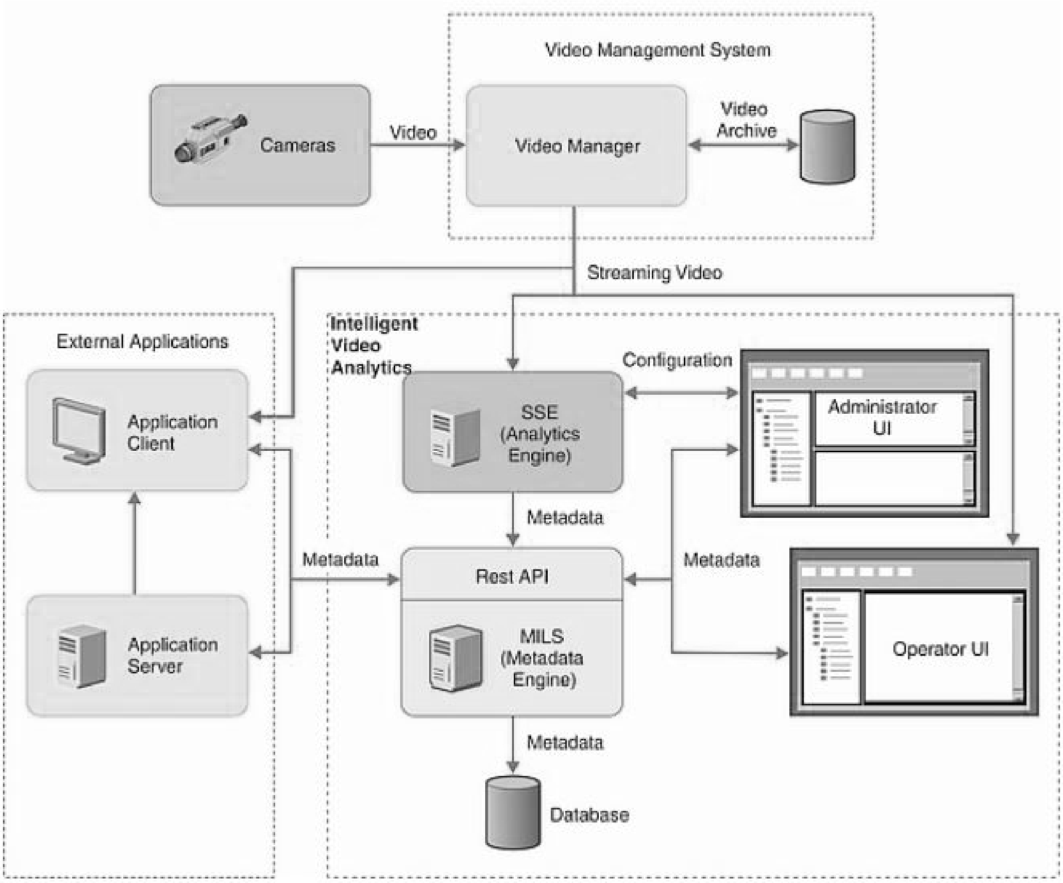


Figure 3.13. Video Analytics[91]

3.5.4 Social Media Analytics

It is the analysis of structured and unstructured content of social media channels such as social networks, social news, blogs, micro blogs, media sharing, wikis, social bookmarking, question and answer sites and review sites.[69] Many mobile applications also facilitate social interactions and are therefore social media channels. Social media research is carried out in various disciplines such as psychology, sociology, computer science, mathematics, economics, physics and anthropology. In recent years, the results of social media analysis have been useful for marketing due to the widespread adoption of social media by users worldwide.[70] In social media, the two sources of information are content (images, audio, customer feedback, product reviews, videos, bookmarks, bookmarks, feelings, etc.) generated by users and the relationships between network entities (people, organizations, products, etc.). Social media analysis can be divided into two parts: content based analysis and structure based analysis.[71] In content-based analysis, the analysis is performed on content published by users on social media platforms. Such content is large in volume, unstructured, noisy and dynamic. To extract insights from these data, text, audio and video analysis techniques can be applied. The challenges of data processing are faced by large data technologies. In structure-based analysis, the focus is on the structural attributes of social networks. Intuitions are extracted from the relationships of entities. A social network is represented by a graph consisting of a set of nodes representing the participants and a set of borders representing the relationships between the entities or participants. Network charts can be further categorized into two types of charts: Social charts and Activity charts.[72] In social charts, borders represent the existence of bonds of friendship between the respective entities. Communities or hubs may be determined by such data. In activity networks, the actual interactions between entities are represented by borders. There is an exchange of information between related entities, such as preferences and comments. Activity charts are more preferable than social charts from an analytical point of view because active relationships are more informative. Community detection or discovery is a technique that extracts the implicit communities of a network. Communities are subnets in which entities interact more with each other than with the entire network. Behavioural patterns and network properties can be extracted from such data. It has various application areas such as marketing and the World Wide Web (WWW).[73] It is useful for developing better product recommendation systems. Social influence analysis analyses the influence of entities and connections in a social network. It is based on the assumption that the behaviour of one entity or participant is influenced or influenced by others. These data give an idea of the influence of the actors, the strength of the connections and the patterns of influence in the network. This technique is useful in marketing to increase awareness of brands and products and their adoption. Link Prediction is a technique that predicts possible future links between existing entities in the social network.[71] A social network continues to grow over time and new edges and nodes continue to add up. The goal is to understand and predict possible interactions, collaborations or influences between social network participants over a given period of time. Through link forecasts, recommendation systems such as Facebook's 'People You May Know', YouTube's 'Recommended for You' and Netflix and Amazon's reccomendation engines are developed.

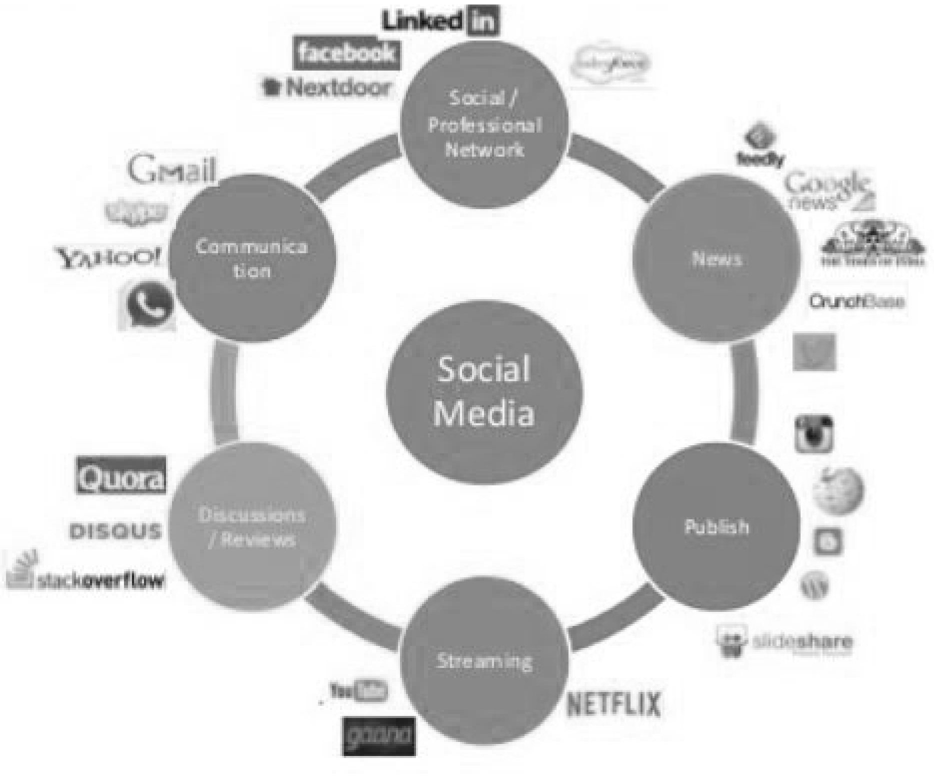


Figure 3.14. Social Media Analytics[92]

3.5.5 Predictive Analytics

Predictive analysis includes a variety of techniques that predict future results based on historical and current data. In practice, predictive analysis can be applied to almost all disciplines, from predicting the failure of jet engines based on the flow of data from several thousand sensors, to predicting customers' next moves based on what they buy, when they buy, and even what they say on social media. At its core, predictive analysis seeks to discover patterns and capture relationships in data. Predictive analysis techniques are divided into two groups. Some techniques, such as moving averages, attempt to discover historical models in result variables and extract them into the future. Others, such as linear regression, aim to capture the interdependencies between result variables and explanatory variables, and use them to make predictions. Based on the methodology below, techniques can also be categorized into two groups: regression techniques (multinomial logit models) and machine learning techniques (neural networks). Another classification is based on the type of result variables: techniques such as linear regression address continuous result variables, while others such as random forests are applied to discrete result variables. Predictive analysis techniques are mainly based on static methods. Several factors require the development of new statistical methods for large data. First, conventional statistical methods are rooted in statistical significance: a small sample is obtained from the population and the result is compared with the case to examine the significance of a particular relationship. The conclusion is then generalized to the whole population. On the contrary, large data samples are massive and represent most, if not all, of the population. Consequently, the concept of statistical significance is not so relevant for large data. Secondly, in terms of computational efficiency, many conventional methods for small samples are not able to scale up to large data. The third factor corresponds to the distinctive characteristics inherent in large data: heterogeneity, noise accumulation, spurious correlations and accidental endogeneity.[46]

Figure 3.15. Predictive Analytics[93]

Chapter 4

Data Science

Data science is a blend of various tools, algorithm development, and machine learning in order to extract knowledge and insights from the raw data to solve analytically complex problems with data and generate business value.[74] Data Scientists, with data exploration, try to understand hidden patterns or characteristics within the data such as trends, inferences and complex behaviors to help companies make better business decisions. They use quantitative techniques such as inferential models or time series forecasting to deeply analyze the data in order to understand what data are saying. Data science enables the creation of data products. A 'data product' is a technical asset that: (1) utilizes data as input, and (2) processes that data to return algorithmically-generated results. The classic example of a data product is a recommendation engine, which ingests user data, and makes personalized recommendations based on that data. Some examples of data products are Amazon's recommendation engines suggest items for you to buy, determined by their algorithms. Netflix recommends movies to you. Spotify recommends music to you. Gmail's spam filter is data product, an algorithm behind the scenes processes incoming mail and determines if a message is junk or not. Computer vision used for self-driving cars is also data product, machine learning algorithms are able to recognize traffic lights, other cars on the road, pedestrians, etc. A data product is different from a data insight which provide advices to the executives to support the decision making. A data product is a technical function that encapsulates an algorithm and is designed to integrate directly into core applications.[74] Data science is a blend of skills in three major areas:

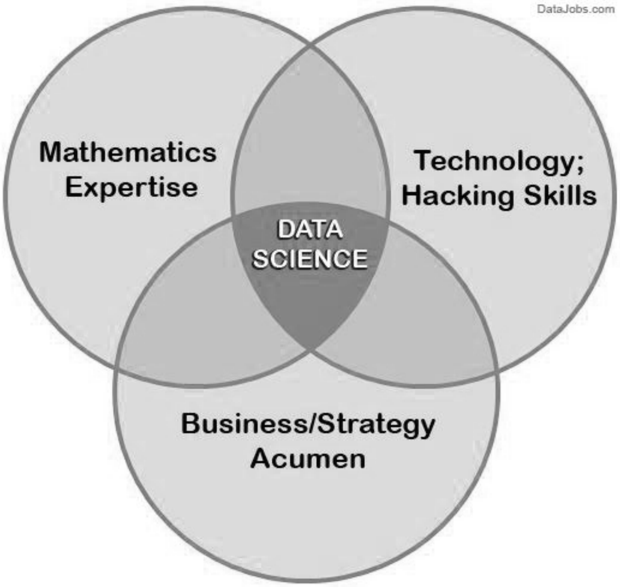


Figure 4.1. Data Science Skills

4.1 Data Science to make predictions

Data Science is mainly used to make decisions and forecasts using causal predictive analysis, prescriptive analysis (predictive plus decision science) and machine learning.[75]

- **Causal Predictive Analysis**: If you want a model that can predict the possibilities of a particular future event, you must apply predictive causal analysis. For instance, if you are providing money on credit, then the likelihood that customers will make future credit payments on time is a matter of concern to you. Here, you can build a model that can perform predictive analysis on the history of customer payments to predict whether future payments will be on time or not.

- **Predictive analysis**: If you want a model that has the intelligence to make your own decisions and the ability to modify it with dynamic parameters, a prescriptive analysis is certainly needed. This relatively new field consists of providing advice. In other words, it not only predicts but suggests a series of prescribed actions and associated results. The best example for this is Google's self-guidance. Data collected from vehicles can be used to train cars that drive alone. You can run algorithms on this data to bring intelligence. This will allow your car to make decisions such as when to turn, what path to take, when to slow down or accelerate.

Machine Learning is a term closely associated with Data Science. It refers to a wide class of methods that revolve around data modeling to (1) make predictions algorithmically and (2) algorithmically decipher data models.

- **Machine Learning to make predictions**: If you have transactional data from a financial company and you need to build a model to determine the future trend, then machine learning algorithms are the best solution. This is part of the supervised learning paradigm. It is called supervised because you already have the data on the basis of which you can train the machines. The central concept is to use tagged data to train predictive models. Labelled data means observations where the basic truth is already known. Training models means automatically characterizing labeled data so that tags for unknown data points are predicted. For example, a fraud detection model can be trained using a historical record of fraudulent purchases.

- **Machine Learning for model discovery**: If you don't have the parameters against which you can make predictions, you need to find the models hidden within the data set in order to make meaningful predictions. This is just the unsupervised template because you don't have predefined labels for the grouping. The most common algorithm used for pattern discovery is Clustering, which algorithmically detects which are the natural groupings that exist in a data set. For example, clustering can be used to programmatically learn the segments of natural customers in a company's user base.

4.2 Data Science Lifecycle

Data Science lifecycle consists of six stages: Discovery, Data Preparation, Model Planning, Model Building, Operationalize and Communicate Results.[75]

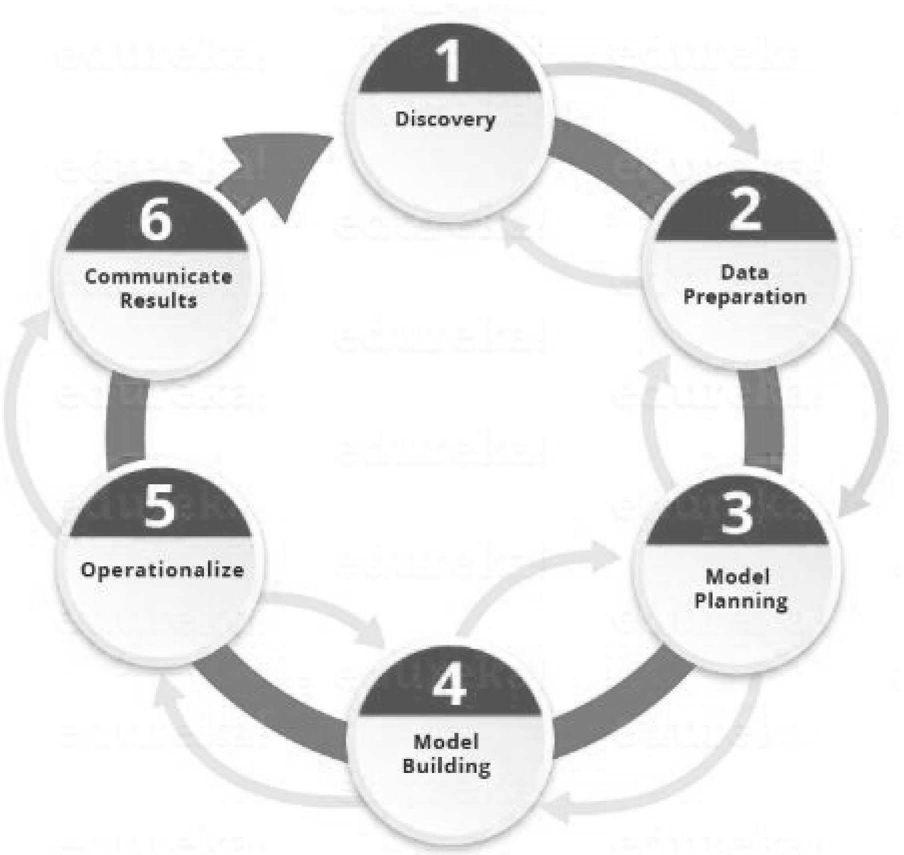


Figure 4.2. Data Science Lifecycle[75]

4.2.1 Discovery

Before starting the project, it is important to understand the various specifications, requirements, priorities and budget required. You must be able to ask the right questions. Here, assess whether you have the necessary resources in place in terms of people, technology, time and data to support the project. At this stage, you also need to frame the business problem and formulate initial assumptions to be tested.

4.2.2 Data Preparation

At this stage, you need an analytical sandbox where you can perform analysis throughout the project. You need to explore, pre-process and condition the data before modeling. In addition, you will run the ETL to get the data in the sandbox. You can use R for cleaning, transforming and displaying data. This will help you locate outliers and establish a relationship between variables. Once you have cleaned and prepared the data, it is time to do exploratory analysis on it.

4.2.3 Model Planning

Here, you determine the methods and techniques for drawing relationships between variables. These relationships will form the basis for the algorithms that you will implement in the next step. Exploratory Data Analytics (EDA) will be applied using various statistical formulas and visualization tools.

Some model planning tools are:

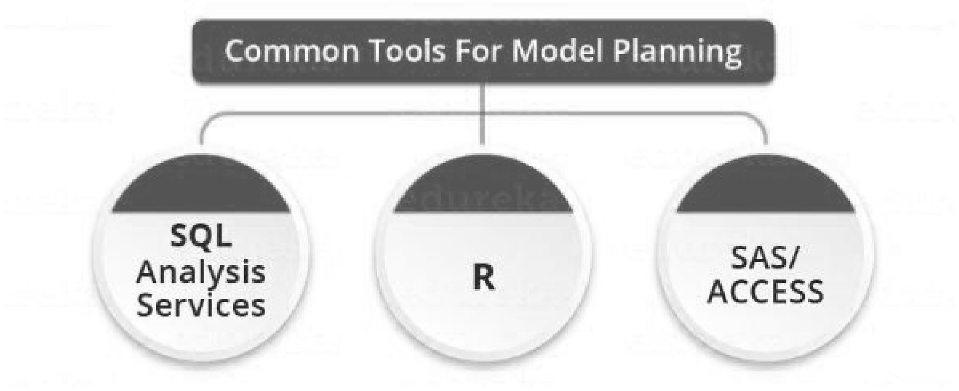


Figure 4.3. Model Planning Tools[75]

- R has a full set of modeling capabilities and provides a good environment for building interpretive models.

- SQL Analysis services can perform analysis in a database using common data mining functions and basic predictive models.

- SAS/ACCESS can be used to access data from Hadoop and is used to create repeatable and reusable model flowcharts. Although many tools are on the market, but R is the most commonly used tool.

4.2.4 Model Building

In this step, you will develop data sets for training and experimentation. You will evaluate whether existing tools are sufficient for modeling or whether a more robust environment such as fast, parallel processing is needed. Various learning techniques such as classification, association and clustering will be analyzed to build the model.

The model building can be achieved through the following tools (figure 4.4).

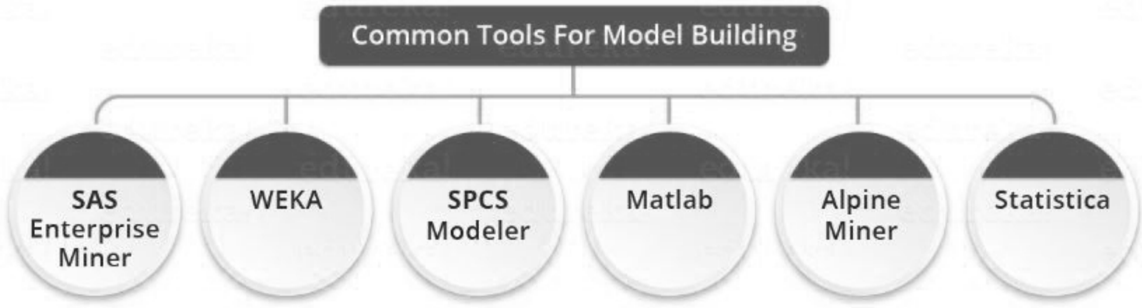


Figure 4.4. Model Building Tools[75]

4.2.5 Operationalize

In this phase, final reports, briefings, code and technical documents are delivered. In addition, sometimes a pilot project is also implemented in a real-time production environment. This will provide you with a clear picture of performance and other related constraints on a small scale before full deployment.

4.2.6 Communicate Results

Now it is important to assess whether you were able to achieve the goal you planned in the first phase. Then, in the last phase, you identify all the key results, communicate them to your stakeholders and determine whether the project results are a success or a failure based on the criteria developed in first stage. The result of the project is a success or failure.

4.3 Companies' Readiness to Data Science

When a company decides to switch to data-driven, first of all, it is important to understand at what stage of maturity regarding the adoption of data and analytics companies are in: the crawl stage, the walking stage, or the running stage.[76] The Crawl stage is where the company is still determining how it wants to use analytics specifically. They are in the process of identifying which business areas can produce data to use in this way. The Walking stage is when the company is already dealing with a data-based business, so there is a need to implement processes and hire qualified staff. The company is very familiar with data management and analysis, and the decision-maker team should know what is going on in the company's business by simply reading a report that tracks the data. They should be able to look at these reports and see the things that will help capture the best return on all investments. The final stage is the Race. At this point, companies can use their insights to make accurate predictions that lead to business growth. All data that has been collected and interpreted can now be used to build models or forecasts. The return on investment that the company will see is enormous. From here on out, the only ads that the company invests will be the ones that bring the most revenue. This is where the company's business has finally reached the data-driven status. The most important thing is that the company will have to carefully consider how data science will eventually benefit its organization either in terms of increased revenue and reduced costs, or in other ways that support its organization's mission.

To create business value, a company needs not only data, but also the right technology, talent and culture: new data streams create business opportunities but at the same time increase ambiguity, noise and complexity. Companies are often faced with a situation described as data overload, doubts about the process of understanding and using data in the right business context: in many cases, employees and teams are not ready to discover, interpret and use data in decision making processes. To make sense of the data and systematically leverage valuable insights to better manage the business, companies also need the right technologies and a strategic program to establish the analytical culture and ensure sufficient willingness to use the data and insights in real business problems.[77]

The key elements for a successful transition to the analytics culture are data, technology and mindset. An organization needs as much data as possible, powerful analytical tools, creative people and the right culture. This will gradually allow teams to explore, analyze and model data, to inherit a data-driven decision making approach, to conduct strategic analytic initiatives and data-driven experiences.

We can identify some key factors of readiness that organizations should consider: strategic, domain, cultural, technological and operational. Ensuring readiness in all factors is important for organizations that enter the analytic, because many of these factors can then become important analytics success factors.[78]

4.3.1 Strategic readiness: making the business case

Companies must first have an idea of what they are trying to measure. Does the business have a clearly defined use case for the use of analytics? What business problem are they trying to solve? Connecting analytics to organizational strategy is crucial to success. Part of the strategic readiness factor is also the willingness to make investments in analytics, but in such a way that these investments follow a clearly identified business case. Strategic readiness involves the clear definition of objectives for analytics, the problems the organization hopes to solve with analytics, and the type of questions that can and cannot be answered.

4.3.2 Domain readiness: skills, tools and data

Has the organization invested in the infrastructure to easily access and analyze the data? Has the organization invested in analytical tools and software to enable these functions? Has the organization invested in training, or recruiting, dedicated staff who have the technical training and ability to exploit the data? Domain preparation involves identifying key competency gaps and investing in the required talent. In addition to new talent, training of existing employees is also required to enable users to learn new analysis tools and ensure proper use. Domain readiness also involves understanding the complex landscape of organizational data. In this environment, multiple sources with multiple properties and data quality issues, data integration and data governance are abundant. Organizations should have a good understanding of what data is needed to obtain the desired information. Data quality is really important: to establish a data-based culture, it is necessary to ensure clean, accurate, reliable and active data that reflects all major business activities. Teams need to be able to trust data and feel confident that they can use it in important data-based decisions. Lack or limited quality of data can be the only point of failure. Data engineering teams must ensure that only clean and reliable data is available to the users. Once identification and scoping have been done, the question of availability arises: Is the data available and has the organization considered how the different sources can be integrated? The following is an assessment of the quality of the data, to ensure consistency and to understand the degree of cleanliness required. Finally, data governance requires an understanding of ownership and access rights, especially with sensitive and regulated data such as financial or health data. Beyond data, investment in the appropriate analytics infrastructure, tools and technologies, is also needed.

4.3.3 Cultural readiness: data-based decision making

Is the organization's leadership ready to make data-based decisions as opposed to intuition or prejudice? Does the company have internal processes and an appropriate culture to embrace the insights generated by the analysis team and leverage the work? In particular, the company has to engage in the use of the analytical results even if the whole organization is not analytical, which means to use analytics as the primary driver of decision making. In fact, for many organizations the use of analytics is a disruptive cultural change that needs to be addressed. Creating a culture of analytics soon can help organizations become better with analytics. When there is a 'culture of analytics', employees use naturally data and insights to support their decisions and improve business performance. Teams, look for data to test their assumptions and evaluate business outcomes. Managers optimize product and services development efforts by combining multiple signals, quantitative and qualitative insights.

4.3.4 Operational readiness: planning for analytics

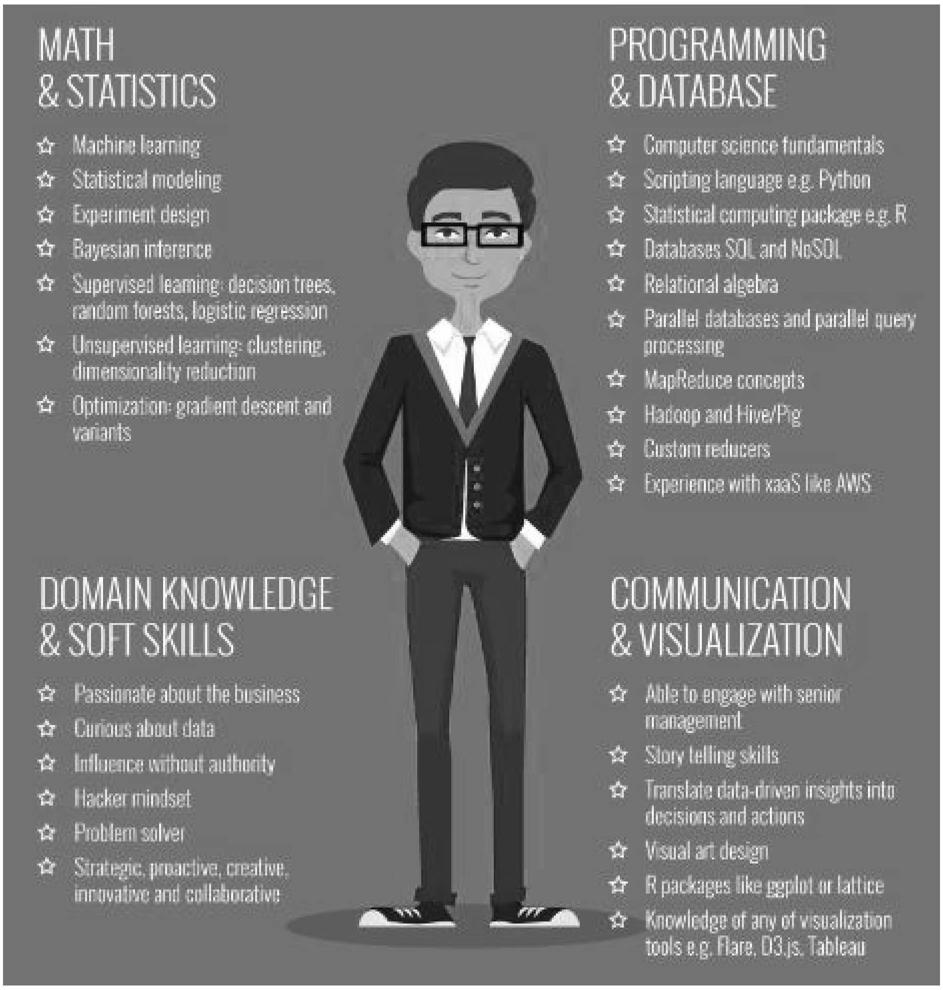
Develop an hypothesis of how the results of the project will generate value to the organization and define the method for defining that potential value. Proper planning, timing, budgeting, and clearly defined milestones would make the difference between success and failure. Two interesting factors fall under the domain of operational readiness. First, the organization should be ready to measure success and identify key performance indicators and metrics to evaluate progress towards the organizational goals with analytics. Second, the organization should develop a clear dissemination plan, a way to bring analytics insights to those who need them. Finally, the need to benchmark and collaborate with industry partners in order to leverage analytics. Learning about best practices and success stories from other institutions is important.

4.3.5 Technological readiness

The volume and complexity of data requires a stack of advanced technologies, including high-performance data stores, reliable and fast data processing pipelines, powerful analytical models, flexible integration layers and interactive visualization systems. At the data layer, companies typically use combinations of data warehouses, data marts, document stores and distributed file systems. At the modelling layer, a range of data mining, statistical modelling and machine learning technologies are typically used to model data, identify trends, patterns and generate insights. At the exploration and presentation layer, Business Intelligence platforms and visualization systems provide access to data models and insights, typically trough reporting, interactive dashboards or OLAP. As analytical systems become proactive, they will know who, when and how to notify based on events, trends, outliers and known or newly discovered patterns.

4.4 Data Scientist Skills

Data Scientist is known as the sexiest job of the 21st century. A Data Scientist requires a mix of multidisciplinary skills including mathematics, statistics, information technology, communication and business and is responsible for collecting, analyzing and interpreting large amounts of data to identify ways to help a company improve operations and decisions and gain a competitive advantage. The role of Data Scientist includes the use of advanced analytics technologies, including machine learning and predictive modeling, to provide insights that go beyond statistical analysis. In recent years, the demand for expertise in Data Science has grown significantly as companies seek to obtain useful information from the amount of structured, unstructured and semi-structured data. The mix of personality characteristics, experience and analytics skills required for the role of Data Scientist is considered difficult to find and, as a result, the demand for qualified Data Scientists has exceeded supply in recent years. Basic responsibilities include collecting and analyzing data and using various types of analysis and reporting tools to detect patterns, trends and relationships in data sets. Data Scientists typically work in teams to extract large data to obtain information that can be used to predict customer behavior and identify risks and business opportunities. The Data Scientist must be able to communicate how to use analytics data to guide business decisions that may include changing course, improving a process or product or creating new services or products. A Data Scientist needs a wide range of skills in the areas of Mathematics and Statistics, which means being able to view data in a quantitative perspective and propose solutions that express the correlations within the data in a mathematical way; the area of Computer Science, which means being able to use technology to manage large data sets and work with complex algorithms; the area of Business Domain Knowledge, which means being a tactical business consultant using data insights. In particular, we can distinguish two classes of skills, Technical and Non-Technical skills.[79]



MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Figure 4.5. Data Scientist skills

4.4.1 Non-Technical skills

- **Education**

  Data Scientists are highly qualified - 88% have at least a master's degree and 46% have a PhD - and developing the depth of knowledge required to be a Data Scientist usually requires very strong training and educational background.

- **Intellectual curiosity**

  Curiosity is the desire to acquire more knowledge. As a Data Scientist, you need to be able to ask questions about data, because Data Scientists spend about 80% of their time discovering and preparing data. This is because the field of Data Science is a rapidly evolving field and there is needs to learn more to keep up with the pace. Curiosity will allow you to sift through the data to find answers and more insights.

- **Business acumen**

  To become a Data Scientist you need a solid understanding of the industry you are working in, and know what business problems your company is trying to solve. In terms of Data Science, being able to discern which problems are important to solve for your business is critical, as well as identifying new ways in which your business should exploit its data. To be able to do this, you need to understand how the problem you solve can have an impact on your business.

- **Communication skills**

  A Data Scientist must be able to communicate his technical discoveries clearly and fluently to a non-technical team, such as Marketing or Sales departments. As a Data Scientist, you need to know how to create a storyline around the data to make it easy for anyone to understand. When communicating, it is important to pay attention to the results and values that are embedded in the analyzed data, because most business owners do not want to know what you have analyzed, but are interested in how this can have a positive impact on their business.

- **Teamwork**

  A Data Scientist can't work alone. He will collaborate with members of his team to develop use cases in order to know the business goals and data that will be needed to solve problems. As a Data Scientist you need to know the right approach to dealing with use cases, the data needed to solve the problem and how to translate and present the result into what can be easily understood by everyone involved. You will literally have to work with all the members of the organization, including your clients.

4.4.2 Technical skills

- **R Programming**

  R is designed specifically for the needs of Data Science. You can use R to solve any problem you encounter in Data Science. However, R has a steep learning curve.

- **Python Coding**

  Python is the most common coding language required in the roles of Data Science, along with Java, Perl or C/C++. Due to its versatility, Python can be used for almost all phases of Data Science processes. It can require various data formats and you can easily import SQL tables into the code.

- **Hadoop platform**

  As a Data Scientist, you may come across a situation where the volume of data you have exceeds the memory of your system or you need to send data to different servers, that is where Hadoop comes in. You can use Hadoop to quickly transmit data at various points in a system. You can also use Hadoop for data exploration, data filtering, data sampling and synthesis.

- **SQL Database/Coding**

  SQL (structured query language) is a programming language that can help you perform tasks such as adding, deleting and extracting data from a database. It can also help you perform analytical functions and transform database structures. SQL is specifically designed to help you access, communicate and work with data.

- **Apache Spark**

  It is a framework for Big Data computation just like Hadoop. The only difference is that Apache Spark is faster than Hadoop. Apache Spark is specifically designed for Data Science to help it run its complicated algorithm faster. It also helps the Data Scientist to handle complex unstructured data sets. Apache Spark makes it possible for Data Scientists to prevent data loss.

- **Machine Learning and AI**

  Artificial intelligence is about systems or machines that imitate human intelligence. Often used interchangeably with its subfields, including machine learning and deep learning, artificial intelligence has become a generic term for applications that perform complex tasks that once required human input.

  Machine Learning is a branch of artificial intelligence focused on building systems that learn or improve performance based on the data they consume. Because machine learning models can adapt without being explicitly programmed, they are ideal for quickly transforming large data sets into highly accurate predictions of future results.

  Deep Learning is an automatic learning technique freely based on the structure of a biological nervous system. Data is passed through interconnected layers of nodes, each of which performs an operation before passing the result to the next node. Typically used in applications that require complex deductions, including facial recognition.

- **Data Visualization**

  The data must be translated into an easy-to-understand format. People naturally understand images in the form of charts and graphs more than raw data. As a Data Scientist, you need to be able to view data with the help of data visualization tools like ggplot, Matplottlib and Tableau. These tools will help you convert the complex results of your projects into an easy-to-understand format. Data visualization gives organizations the opportunity to work directly with data. They can quickly grasp insights that will help them act on new business opportunities and keep up with the competition.

4.5 Data Science Teams

The role of the Data Scientist is often confused with that of the data analyst, but while there is overlap in many of the competencies, there are significant differences.[80] Although the role of a data analyst varies from company to company, in general, these professionals collect data, process it and perform statistical analysis using standard statistical tools and techniques. Analysts also identify models and correlations in data sets to identify new opportunities for improving business processes, products or services. In some cases, data analysts design, build and maintain large relational and data database systems. Data Scientists are responsible for these and many other tasks. These professionals are equipped to analyze large data using advanced analysis tools and are expected to have the research background to develop new algorithms for specific problems. They may also be tasked with exploring data without a specific problem to be solved. In this scenario, they need to understand data and business well enough to ask questions and provide insights to business executives with the goal of improving business operations, products, services or customer relationships.

There are generally two main types of Data Scientist:

- Type A Data Scientists (analysis) focus on making sense of the data through statistical analysis.

- Type B Data Scientists (building) develop predictive models and algorithms to feed data products.

In particular, we can identify several key roles in the Data Science industry.[81]

4.5.1 Data Science Team Leader

It is a key leadership role, the CAO. It is a corporate translator who bridges the gap between Data Science and domain experience, acting both as a visionary and a technical leader.

Preferred skills: Data Science and analytics, programming skills, domain expertise, leadership and visionary skills.

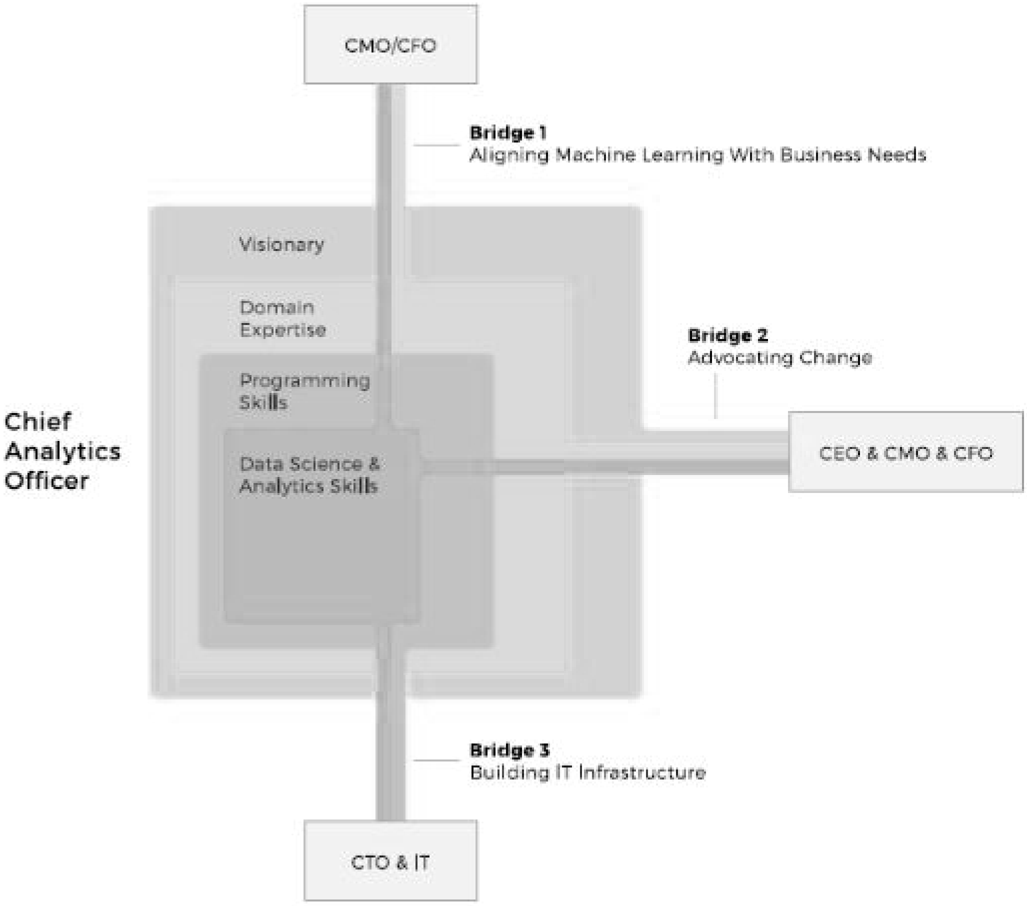
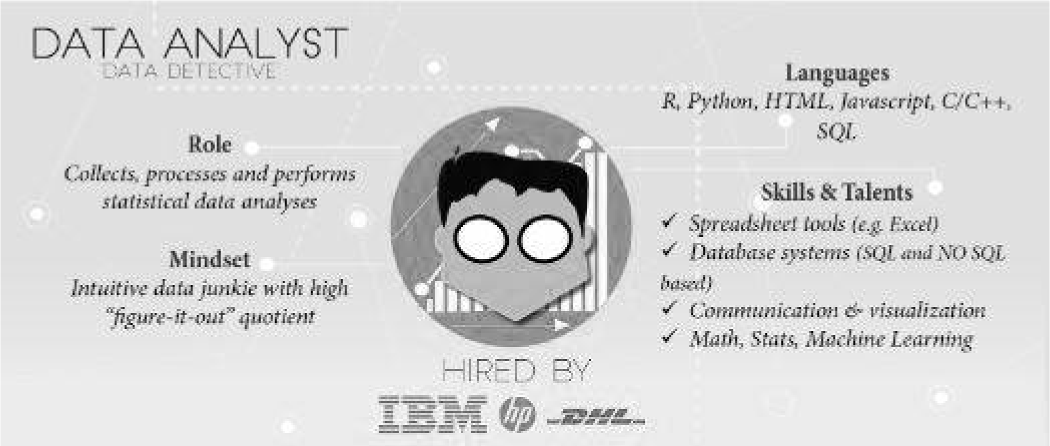


Figure 4.6. Data Science Tam Leader

4.5.2 Data Analyst

The role of the data analyst implies adequate data collection and interpretation acivities. An analyst ensures that the data collected is relevant and comprehensive, while interpreting the results of the analysis. Some companies, such as IBM or HP, also require data analysts to have visualization skills to convert alienating numbers into tangible insights through graphics.

Preferred Skills: R, Python, JavaScript, C/C++, SQL



DATA ANALYST

DATA DETECTIVE

Role

Collects, processes and performs statistical data analyses

Mindset

Intuitive data junkie with high "figure-it-out" quotient

Languages

R, Python, HTML, Javascript, C/C++, SQL

Skills & Talents

- ✓ *Spreadsheet tools (e.g. Excel)*
- ✓ *Database systems (SQL and NO SQL based)*
- ✓ *Communication & visualization*
- ✓ *Math, Stats, Machine Learning*

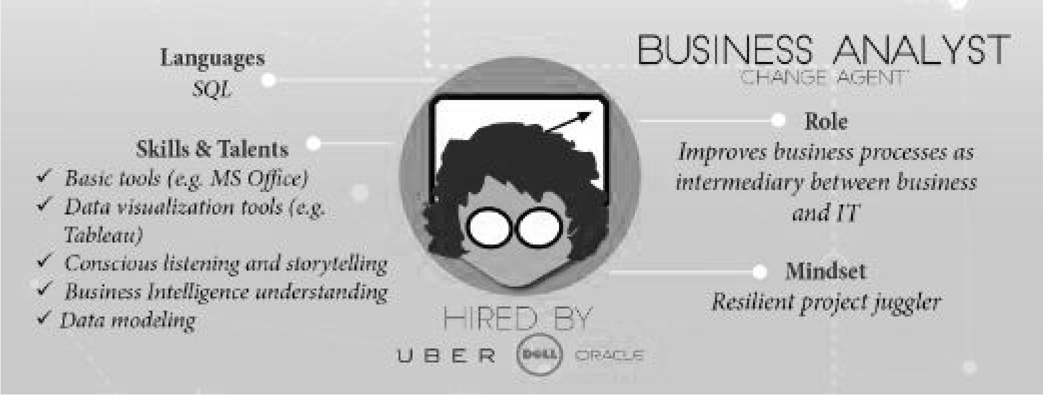
HIRED BY

Figure 4.7. Data Analyst

4.5.3 Business Analyst

A business analyst performs the functions of a CAO but at an operational level. This involves converting business expectations into data analysis. If your core Data Scientist has no industry experience, a business analyst bridges this gap.

Preferred skills: data visualization, business intelligence, SQL



BUSINESS ANALYST

"CHANGE AGENT"

Languages

SQL

Skills & Talents

- ✓ *Basic tools (e.g. MS Office)*
- ✓ *Data visualization tools (e.g. Tableau)*
- ✓ *Conscious listening and storytelling*
- ✓ *Business Intelligence understanding*
- ✓ *Data modeling*

Role

Improves business processes as intermediary between business and IT

Mindset

Resilient project juggler

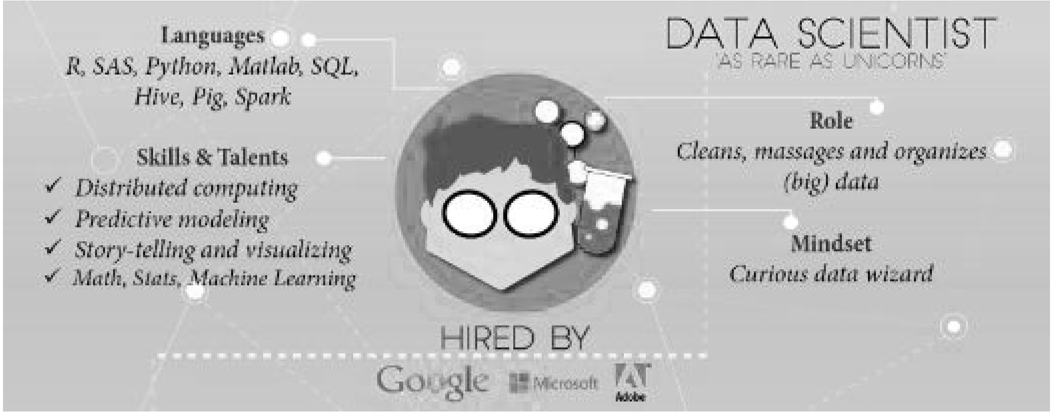
HIRED BY

Figure 4.8. Business Analyst

4.5.4 Data Scientist

What does a data scientist do? A data scientist is a person who solves business tasks using machine learning and data mining techniques. This role can be reduced to data preparation and cleaning with further training and model evaluation.

Preferred Skills: R, SAS, Python, Matlab, SQL, noSQL, Hive, Pig, Hadoop, Apache Spark.



DATA SCIENTIST
'AS RARE AS UNICORNS'

Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents
✓ Distributed computing
✓ Predictive modeling
✓ Story-telling and visualizing
✓ Math, Stats, Machine Learning

Role
Cleans, massages and organizes (big) data

Mindset
Curious data wizard

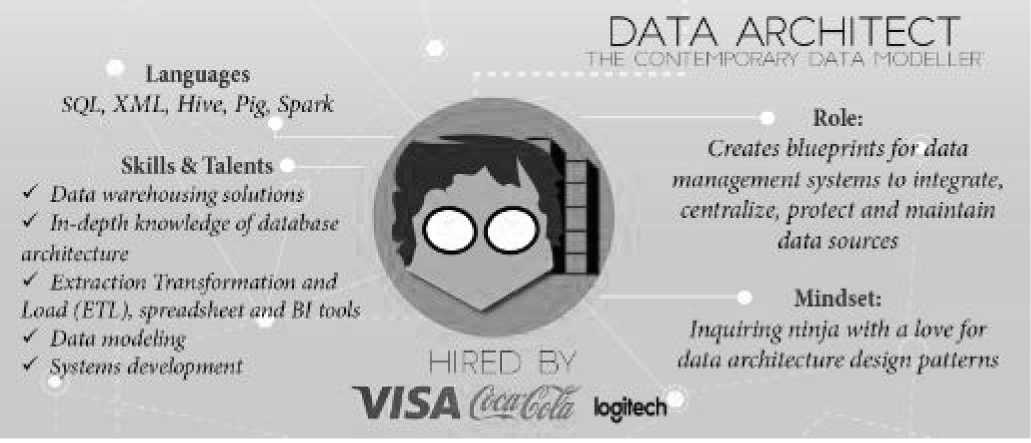
HIRED BY
Google Microsoft Adobe

Figure 4.9. Data Scientist

4.5.5 Data Architect

This role is essential for working with large amounts of data. This role is essential to store data, define the database architecture, centralize data and ensure integrity between different sources. For large distributed systems and large data sets, the architect is also responsible for performance.

Preferred Competences: SQL, noSQL, XML, Hive, Pig, Hadoop, Apache Spark



DATA ARCHITECT
'THE CONTEMPORARY DATA MODELLER'

Languages
SQL, XML, Hive, Pig, Spark

Skills & Talents
✓ Data warehousing solutions
✓ In-depth knowledge of database architecture
✓ Extraction Transformation and Load (ETL), spreadsheet and BI tools
✓ Data modeling
✓ Systems development

Role:
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset:
Inquiring ninja with a love for data architecture design patterns

HIRED BY
VISA Coca-Cola logitech

Figure 4.10. Data Architect

4.5.6 Data Engineer

Engineers implement, test and maintain the infrastructure components that data architects design. Realistically, the role of an engineer and that of an architect can be combined into a single person. The set of skills is very close.

Preferred skills: SQL, noSQL, Hive, Pig, Matlab, SAS, Python, Java, Ruby, C++, Perl

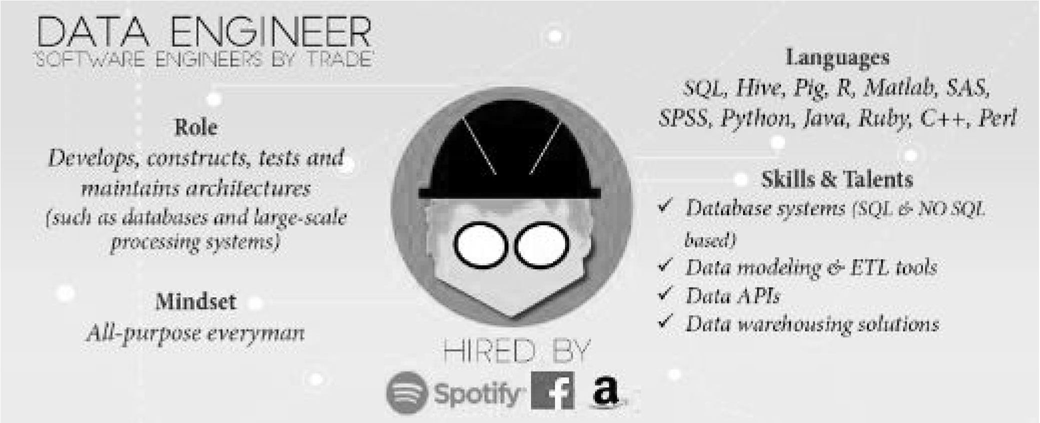
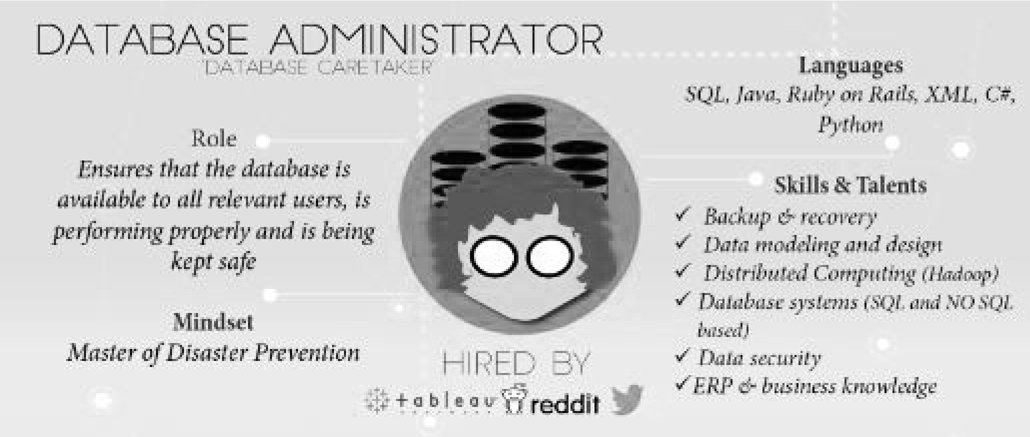

DATA ENGINEER
"SOFTWARE ENGINEERS BY TRADE"
Role
Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)
Mindset
All-purpose everyman
Languages
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl
Skills & Talents
✓ Database systems (SQL & NO SQL based)
✓ Data modeling & ETL tools
✓ Data APIs
✓ Data warehousing solutions
HIRED BY
Spotify
f
a

Figure 4.11. Data Engineer

4.5.7 Database Administrator

This role ensures that the database is available to all interested users, works properly and is kept safe.

Preferred skills: SQL, noSQL, Hadoop, ERP, data modeling and data design



Role
Ensures that the database is available to all relevant users, is performing properly and is being kept safe

Mindset
Master of Disaster Prevention

Languages
SQL, Java, Ruby on Rails, XML, C#, Python

Skills & Talents
- ✓ *Backup & recovery*
- ✓ *Data modeling and design*
- ✓ *Distributed Computing (Hadoop)*
- ✓ *Database systems (SQL and NO SQL based)*
- ✓ *Data security*
- ✓ *ERP & business knowledge*

Figure 4.12. Database Administrator

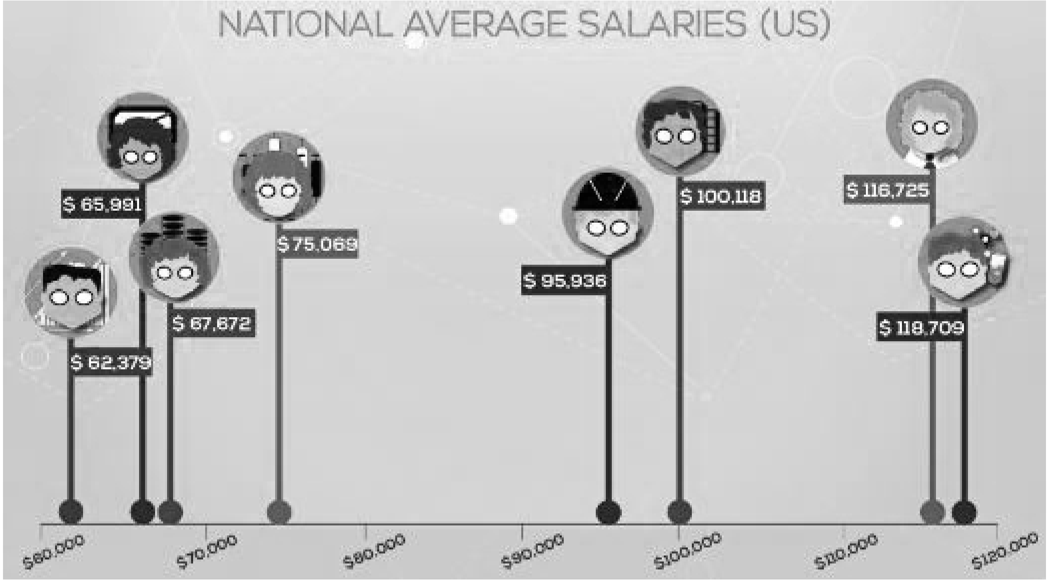


Figure 4.13. Data Scientist Salaries

4.6 Teams' Structure

To understand how a Data Science project will be implemented, it is important to know how Data Science teams are trained and structured based on the particular industry, the particular type of machine learning process that the company is facing and the skills available to the company. In terms of team structure, we can consider three basic team structures: IT-centric structure, integrated structure and department specializing in information technology.[81]

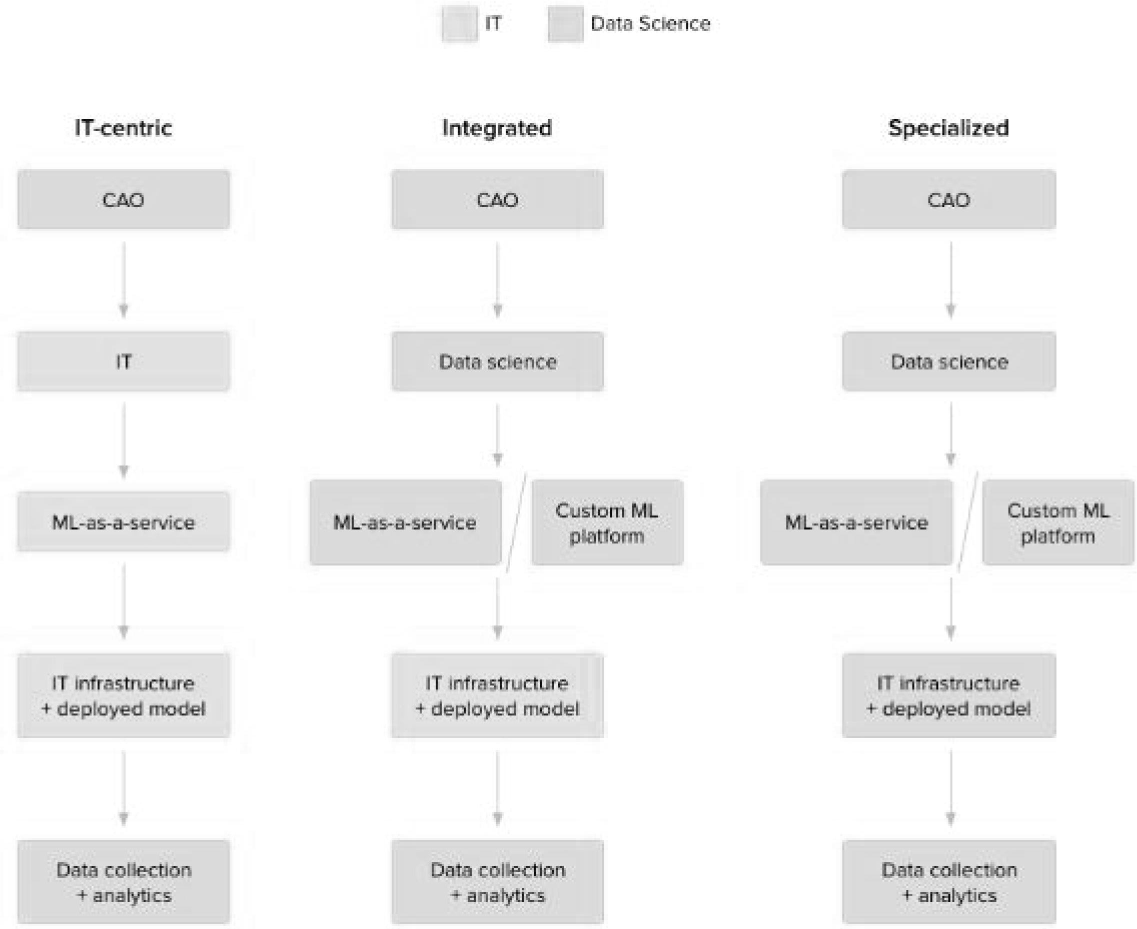


Figure 4.14. Data Science Team Structure[81]

4.6.1 IT-centric structure

Sometimes, hiring data scientists is not an option, and you must exploit the talent that is already within the company. The main role of analysis and leadership would be that of a business translator, usually referred to as chief analytics officer (CAO) or chief data officer (CDO). Everything else - data preparation, training models, creation of user interfaces and distribution of models within a corporate IT infrastructure - can be largely managed by the IT department. This approach is rather limited, but can be achieved using MLaaS (Machine Learning As A Service) solutions. MLaaS solutions have their limitations in terms of automated learning methods and costs. All operations, from data cleaning to model evaluation, are priced separately. And considering that the number of iterations to form an effective model cannot be estimated in advance, working with MLaaS platforms involves some budget uncertainty.

Pros of IT-centric structure:

- Leveraging new investments with existing IT resources.
- The IT infrastructure is provided and maintained by an external service.
- Internal specialists can be trained to further realize the potential of predictive analysis.
- Cross-sectional management is reduced as all operations are carried out within the IT department.
- Less time-to-market for relatively simple machine learning tasks that require one or a few models.

Cons of IT-centric structure:

- Limited methods of automatic learning and data cleaning procedures that these services provide.
- Training, testing and forecasting of models should be paid for. This leads to uncertainty of the final cost per forecast as the number of iterations needed cannot be estimated in advance.

4.6.2 Integrated structure

With the integrated structure, a data science team focuses on preparing the data set and training the models, while IT specialists take care of the interfaces and infrastructure supporting the implemented models. Combining machine learning skills with IT resources is the most viable option for consistent and scalable machine learning operations. Unlike the IT-centric approach, the integrated method requires the presence of an experienced data scientist in a team and an elaborate recruitment effort in advance. This ensures better operational flexibility in terms of available techniques.

Pros of Integrated structure:

- Leverage existing IT resources and investments.
- Data scientists focus on innovation.

- Exploiting the full potential of ML applications both as a service and as a custom application.
- It starts with one or two data scientists, then trains and on board other local experts.
- Use of combinations of customized models (ensemble models) to obtain better or wider forecasts.

Cons of Integrated structure:

- The IT infrastructure is required in case of custom ML use.
- Cross-syllable management requires considerable effort.
- Significant investments in data science talent acquisition.
- The commitment of talents in the field of data science and preservation challenges.

4.6.3 Specialized Data Science Department

To reduce management effort and build a complete auto-learning framework, you can run the entire auto-learning workflow within an independent data science department. This approach involves the highest costs. All operations, from data cleaning and modeling training to front-end interface construction, are performed by a dedicated data science team. In the case of large organizations, data science teams can integrate different business units and operate within their specific fields of analytical interest.

Pros of specialized data science department:

- Centralized data management and increased problem-solving capabilities.
- Realize the full potential of ML applications both as a service and as a custom application.
- Solve complex forecasting problems that require in-depth research or the construction of segmented model factories (which automatically operate between different segments and business units).
- Creation of a fully functional data science playground to promote innovation.
- Increased scalability potential.

Cons of specialized data science department:

- Building and maintaining a complex computational infrastructure.
- Substantial investment in data science talent acquisition.
- The commitment of talents in the field of data science and preservation challenges.

4.7 Teams' Organization

According to Accenture's classification, there are six options for organizing a Data Science team: Decentralized, Functional, Consulting, Centralized, Centre of Excelence, Federated.

The model can change and evolve according to your business needs. In general, to create a high-performance Data Science Team could be some tips:

Spend less time hiring people for each title and focus on understanding the roles that a single data specialist can play. For startups and smaller organizations, responsibilities do not need to be strictly clarified.

Encourage cross-functional collaborations. Designers, marketers, product managers and engineers all need to work closely with the DS team.

Practical incorporation. As we said above, recruiting and retaining talent in data science requires some additional work. One of these is integration: placing data scientists to work in business-focused departments to have them report centrally, collaborate better and help them feel part of the overall picture.

Establish the team environment before hiring. This means that your product managers need to be aware of the differences between data and software products, have adequate expectations and process differences in results and deadlines. PMs must have sufficient technical knowledge to understand these specificities. Alternatively, you can start looking for data scientists who can perform this role immediately.

4.7.1 Decentralized

This is the least coordinated option where analytical efforts are used sporadically across the organization and resources are allocated within each group's function. This is often the case in companies where experience in data science has appeared organically, which often leads to silos' efforts, lack of standardization of analytics, and decentralized reporting.

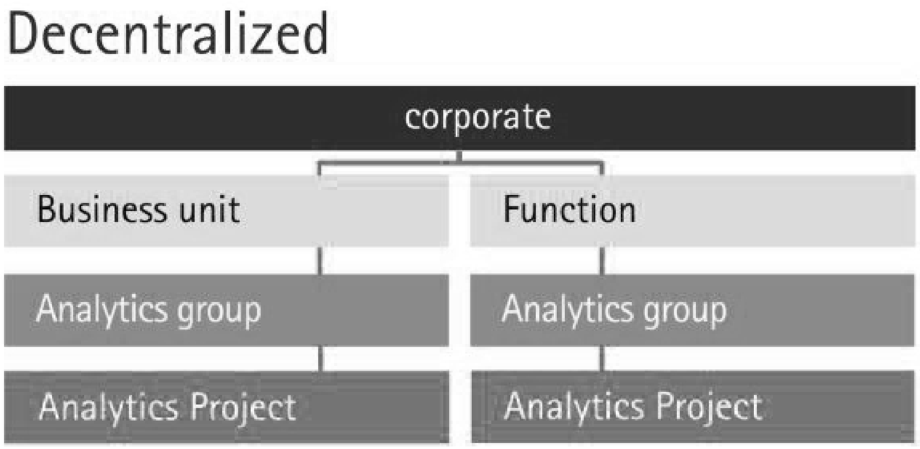


Figure 4.15. Decentralized[81]

4.7.2 Functional

Here most analytics specialists work in a department where analytics is most relevant: often it is marketing or supply chain. This option also involves minimal or no coordination and the expertise is not used strategically at the enterprise level.

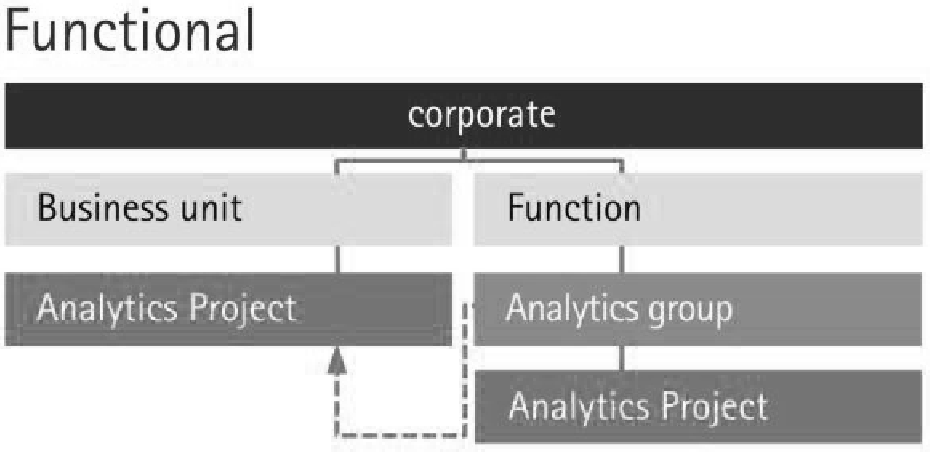


Figure 4.16. Functional[81]

4.7.3 Consulting

In this structure, analytical people work together as a single group, but their role within an organization is consulting, which means that different departments can take over them for specific tasks. This, of course, means that there is almost no allocation of resources - either specialists are available or not.

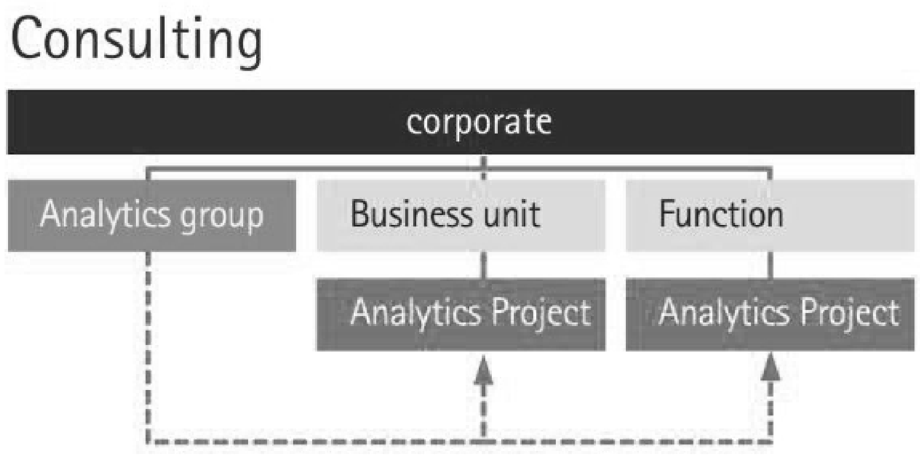


Figure 4.17. Consulting[81]

4.7.4 Centralized

This structure finally allows you to use analytics in strategic tasks - a single data science team serves the entire organization in a variety of projects. Not only does it provide a DS team with long-term funding and better resource management, it also encourages career growth. The only trap is the danger of turning an analysis function into a support function.

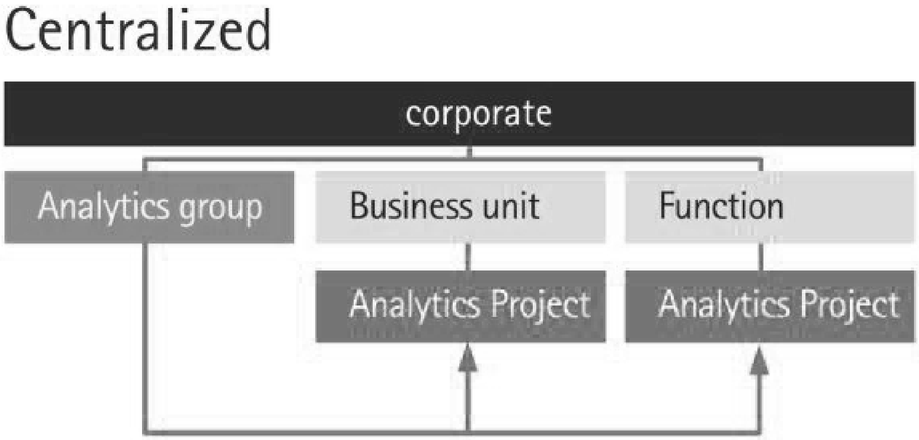


Figure 4.18. Centralized[81]

4.7.5 Centre of Excellence (CoE)

If you choose this option, you will still maintain the centralized approach with a single business center, but the data scientists will be assigned to different units in the organization. This is the most balanced structure - the analysis activities are highly coordinated, but the experts will not be removed from the business units.

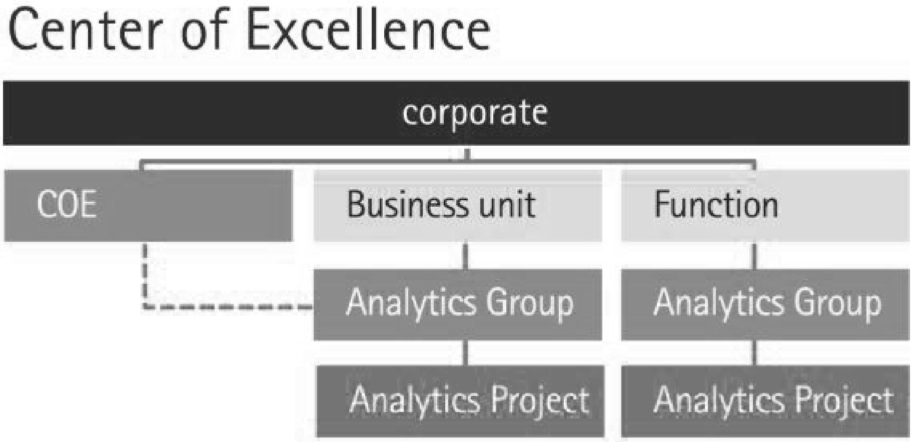


Figure 4.19. Centre of Excellence[81]

4.7.6 Federated

This model is relevant when there is a strong demand for analytical talent across the enterprise. Here, a sort of SWAT team is employed - an analysis group working from a central point and dealing with complex cross-functional tasks. The rest of the data scientists are distributed as in the model of the Centre of Excellence.

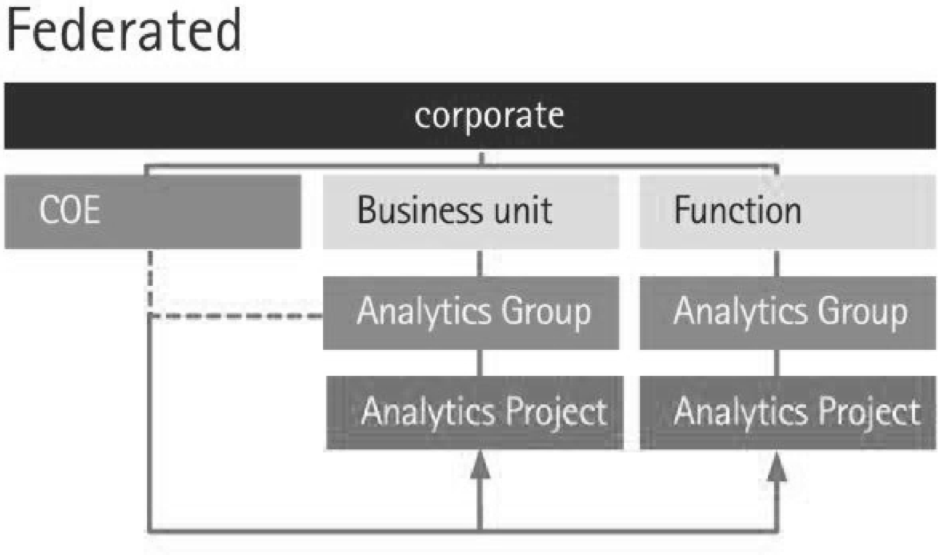


Figure 4.20. Federated[81]

4.8 Data Science Applications

Today more companies than ever grow Data Science teams and build open source-centric technology stacks to find measurable value in large data. Data Science is changing the way many companies do business. Highly customized customer experiences, operational efficiency and improved employee productivity are just a few of the benefits companies see when implementing data science applications.

Now is the time for companies to act on their plans to implement Data Science applications. In the next decade, companies that do not adopt intelligent technologies will have, among other problems, a blurred competitive advantage. The main advantage of integrating Data Science into an organization is empowerment and facilitation of decision-making. Organizations with data scientists can take into account quantifiable, data-based evidence in their business decisions. These data-based decisions can ultimately lead to increased profitability and improved operational efficiency, business performance and workflows. In customer-facing organizations, data science helps identify and refine the target audience. Data Science can also help recruitment: Internal application processing, aptitude tests and data-based games can help an organization's human resources team make faster and more accurate selections during the recruitment process. The specific benefits of Data Science vary depending on the company's purpose and industry. Sales and marketing, for example, can extract customer data to improve conversion rates or create one-to-one marketing campaigns. Banks are extracting data to improve fraud detection. Streaming services such as Netflix extract data to determine what its users are interested in and use it to determine what television programs or movies to produce. Data-based algorithms are also used in Netflix to create custom recommendations based on a user's viewing history. Shipping companies such as DHL, FedEx and UPS use Data Science to find the best routes and delivery times, as well as the best modes of transport for their shipments. There are a lot of domains where Data Science is being applied. Data Science is still an emerging field within your business because identifying and analyzing large amounts of unstructured data can be too complex, costly and time-consuming for businesses. During my studies in Chile, thanks to the professor Victor Leiva from the Pontificia Universidad Catolica de Valparaiso, I had the possibility to study and analyze some Data Science applications that will be described below.

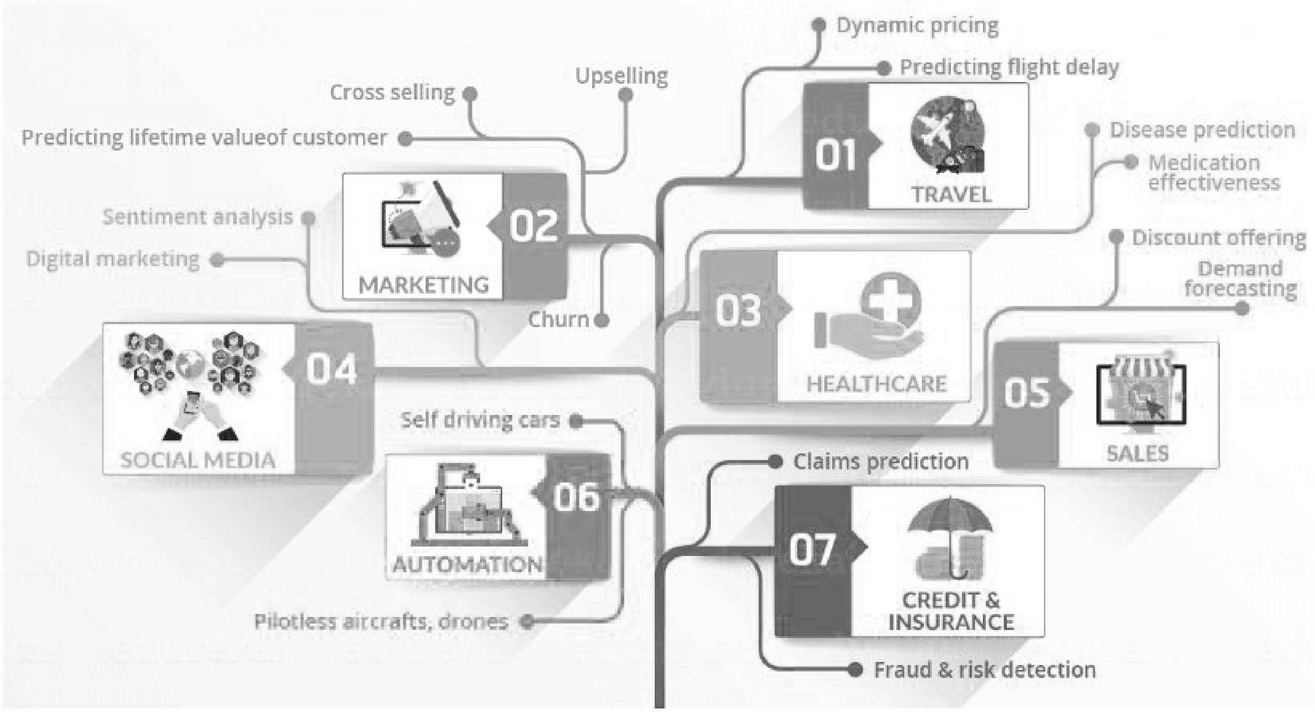


Figure 4.21. Data Science Domains[75]

4.8.1 Mining

This project seeks to define a predictive model to answer the question: what will be the future electricity consumption of Codelco-Chuquicamata's A2 grinding plant?

In the first place, as a previous step to the realization of the project, it is very important for the correct management of the project, to assure the accessibility, availability and integrity of all the documentation, files and articles related both to the project itself and to the area in which it is developed. For this reason, the team had to make a data warehousese, to later enter the ETL stage.

The data obtained was provided by the Interdisciplinary Engineering Center, which gave the team the data corresponding to the A0, A1 and A2 grinding plants of Codelco Chuquicamata. However, after analyzing the quality of the data, it was decided to limit the scope to the A2 plant, due to the fact that the data of the remaining plants were critically incomplete. It should be noted at this point that the A2 plant data correspond to ball mills 16-A, 16-B, 17-A and 17-B, as well as SAG-16 and SAG-17.

As a next step in this stage, a .csv archive was created with the data of the different mills of the plant in a grouped form for later use. This file has the composition of 25 columns (variables) and 2136 rows (including a header).

Once the dataset with which the project is developed was obtained, we proceeded to analyze the composition of the different cells that made up the archive. Thus, it was possible to visualize that many of these cells contained incorrect data; boxes with text (Arc Off-Line and No data) and boxes with zeros (it was consulted and confirmed with the Interdisciplinary Engineering Center that it was not a valid data). Subsequently, these incorrect data were cleaned and the different cells containing these values were left blank.

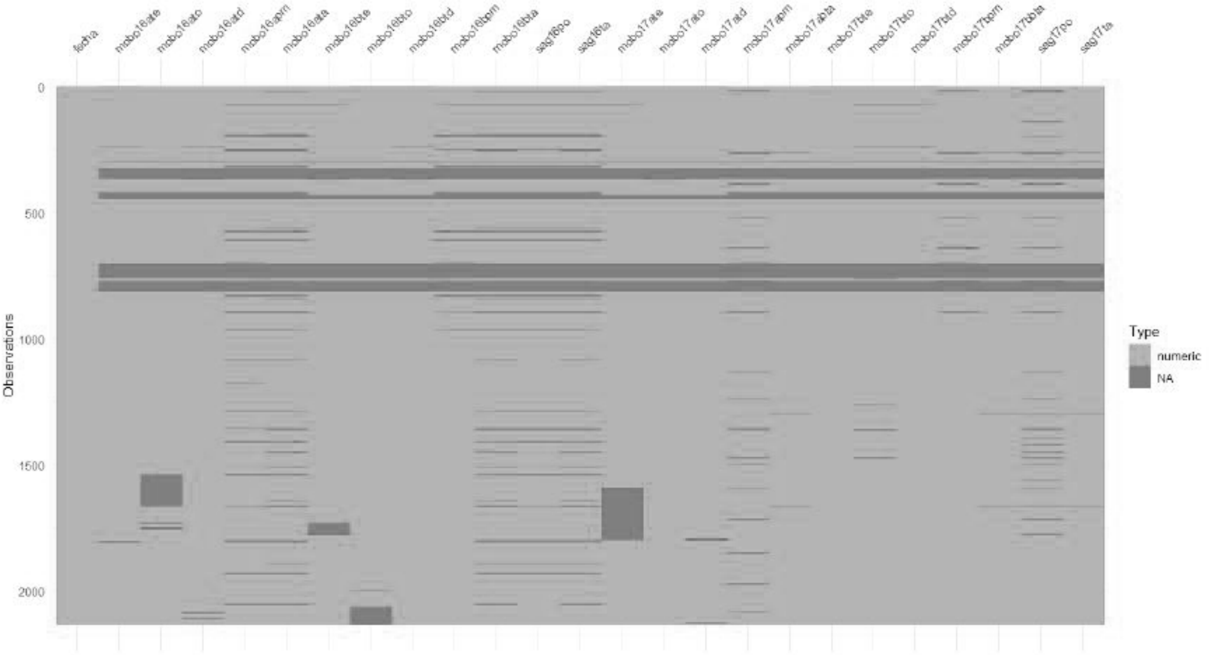


Figure 4.22. Missing Data Visualization

As can be seen through the visualization, the missing data, in gray, is not presented completely randomly, but there are some continuous rows of missing data and other isolated areas. The amount of missing data was 10.3% of the total, so there were two alternatives, to impute the missing data or eliminate the rows that have some missing data.

Once the previous step had been completed, the course of action to be followed was determined. It was decided to perform an autocorrelation analysis on the subset corresponding to the 16-A ball mill, in order to determine if the variables of the mills behave as a time series to determine if it would be necessary to perform a data imputation. Thanks to the information obtained from the autocorrelation analysis, it was possible to conclude that the imputation of data is indeed necessary since it is necessary to use a time series. It was determined that the package that best adapted to the problem was the Amelia in R, because most of the relative errors are less than 50% and there are also few cases with relative errors greater than 100%.

Finally, we proceeded to group the different subsets that were separated for the correct imputation of data in a single .csv file, which contains the final dataset with the extract and transform stages performed.

In terms of the data loading stage, the final file was uploaded to the data warehouse.

In terms of establishing the criteria to consider in the management of the different rows and columns of the dataset, it is necessary to remember that the different columns that make up the dataset correspond to the variables for each of the mills in the plant A2. Thus, it was established that in order to determine the possible relationships between these variables and the problem to be solved by means of the analysis carried out in the project, all these variables will be kept within the final data set with which the project will be carried out.

Once the ETL stage was completed, it was possible to carry out an exploratory analysis on the data obtained through graphics support and data visualization.

First, the team performed a descriptive statistical analysis of the data in order to better understand the characteristics of the corresponding descriptive statistics for each of the quantitative variables.

Once the descriptive statistics analysis was completed, the team performed an analysis of the variables through the R's Pysch package. The package allowed the generation of graphs for each mill separately, which are composed of:

- The middle line represents the histogram of each variable (in light blue) and its respective adjusted distribution line.

- The upper right triangle shows the numerical correlation between each combination of positions.

- The lower left triangle shows the scatter plot between each combination of positions.

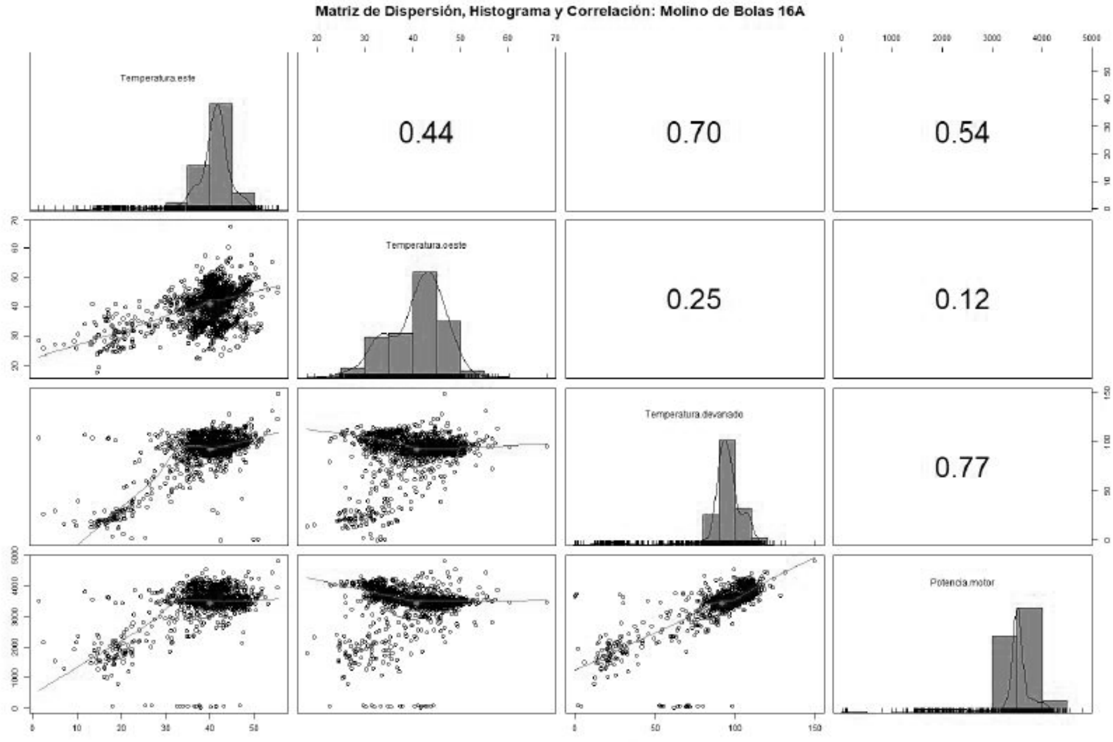


Figure 4.23. Graphical Summary

In addition, the team developed through the DataExplorer package of R, a matrix of correlations between the different variables.

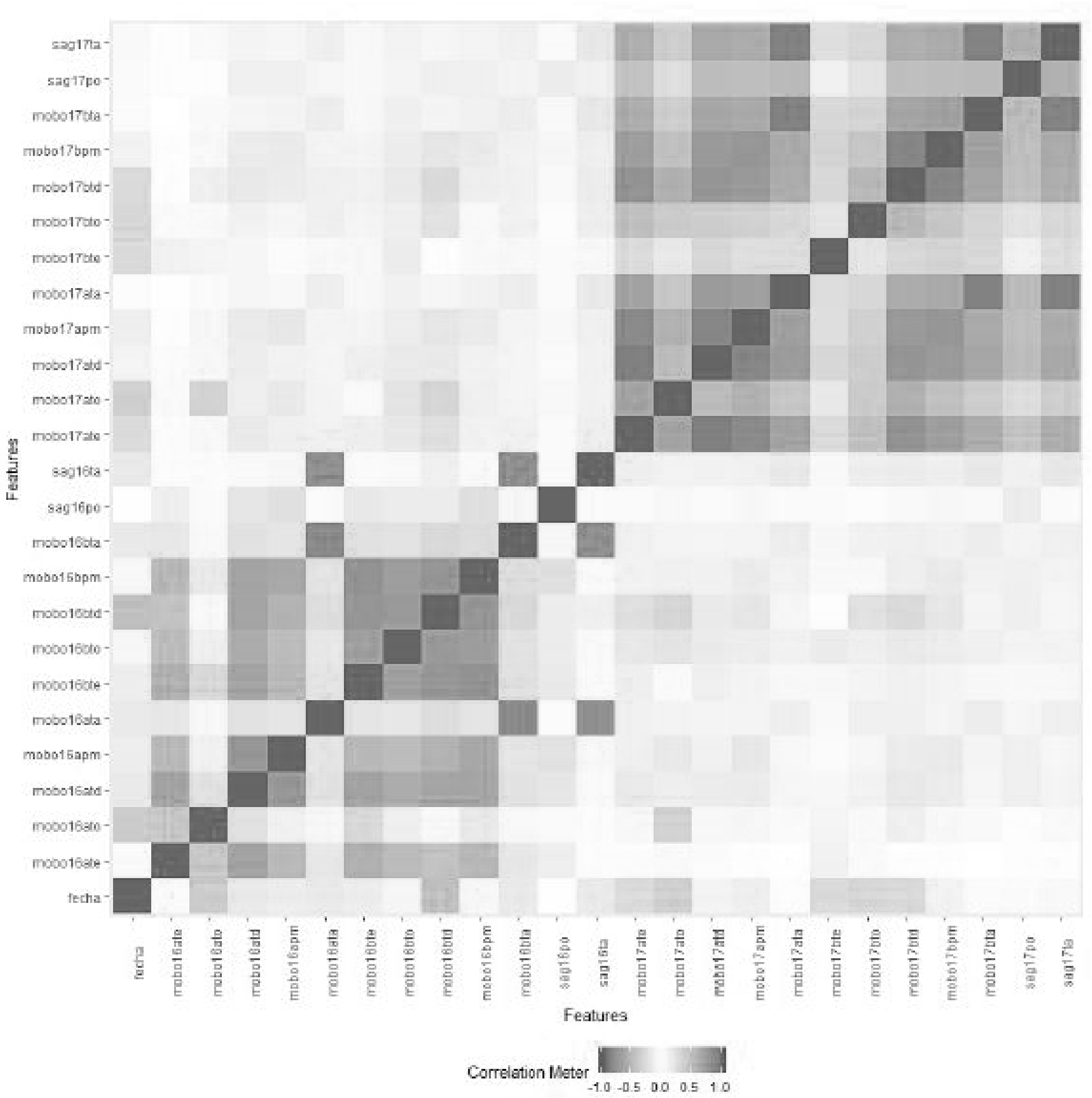


Figure 4.24. Correlation Matrix

Finally, since the variable of interest to predict is the power of the different mills, graphs of the time series were also made.

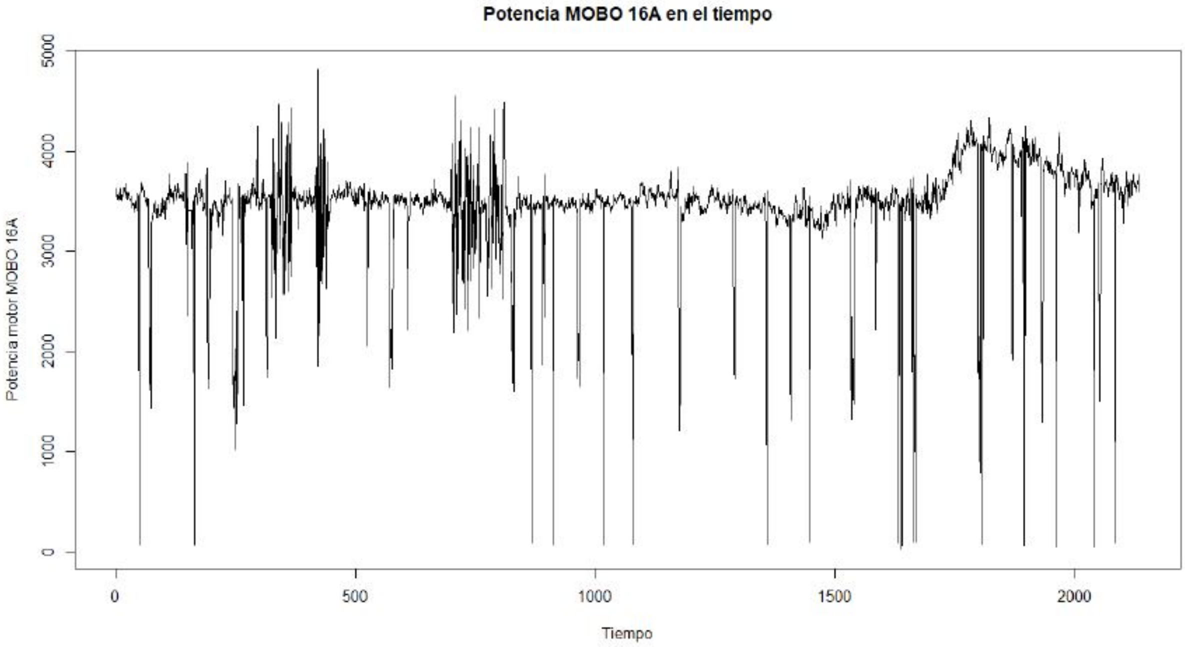


Figure 4.25. Power Time-Series

From the results obtained in the AED it was possible to extract various information about the data set available for the development of the project.

After the AED, predictive models were developed to answer the aforementioned question for each of the mills that make up the plant A2.

The work team chose to use a GARMA predictive model, which is a moving average (MA) autoregressive time series (AR) that incorporates predictor variables (x) using the idea of generalized linear models (GLM). This last component allows the model to use any distribution of the exponential family. As it was proved in the AED, through autocorrelation tests, the power of the mills has an auto-regression behaviour, so this model adjusts to the characteristics of the problem and also allows the use of a wide range of distributions. The inputs of the garmaFit function of the gamlss.util package in R are the order of the ARMA(p,q) model, the distribution and the tail. These inputs were determined by an analysis of the AIC coefficient. It is understood that a lower AIC coefficient is a positive feature of the model.

Thus, the statistical model for the power of the 16A Ball Mill is presented below:

$$\begin{aligned}\widehat{\eta_t} = x_t{}^T\hat{\beta} + (9{,}98e^{-1}) * \left(y_{t-1} - x_{t-1}{}^T\hat{\beta}\right) &+ (3{,}763e^{-1}) * (y_{t-1} - \widehat{\eta_{t-1}}) \\ &+ (2{,}131e^{-1}) * (y_{t-2} - \widehat{\eta_{t-2}}) \\ &+ (1{,}033e^{-1}) * (y_{t-3} - \widehat{\eta_{t-3}})\end{aligned}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}intercepto \\ \hat{\beta}tempdevanado \\ \hat{\beta}tempoeste \end{bmatrix} = \begin{bmatrix} 3.473\mathrm{e}^{+3} \\ 2.941\mathrm{e}^{-3} \\ 8.497\mathrm{e}^{-3} \end{bmatrix}$$

$$x_t = \begin{bmatrix} 1 \\ tempdevanado_t \\ tempoeste_t \end{bmatrix}$$

To replicate the model to the other mills, a similar procedure was performed.

The GARMA model proves to be a capable and robust model to be used in the prediction of electricity consumption in mining, and therefore, as a tool in the decision-making process.

The management recommendations that the team proposes for Codelco-Chuquicamata were the following:

- Improvement of the PI-System for data collection, which was shown as offline for several consecutive months.
- Training of operators regarding the importance of keeping the system working optimally, given that, as mentioned in this paper, there were registration errors due to human intervention.
- Make use of predictive models such as those proposed by the work team in order to make better decisions regarding electricity consumption and production planning.

4.8.2 Citizenship

Are there patterns of occurrence of traffic accidents that allow to improve their prevention programs in the Region of Valparaiso? This is the question that has been sought to answer with the use of the Data Science methodology.

It began with obtaining a database of traffic accidents from CONASET, which contains the variables obtained through the collection of data by the policemen who go to the site of the occurrence of a traffic accident. As it is a raw database, it does not have any type of data cleaning work. In order to be able to use these data to try to answer the established question, the team had to carry out an arduous work of selection and transformation of data on the raw data base, to be able to obtain a dataset on which one could work in a better way during the development of the project. It should be noted that this process was carried out throughout the project and was not a one-off process. This work was carried out mainly with the help of Microsoft Excel and R software.

The extraction, transformation and loading of data is an activity that had to be carried out transversally and constantly throughout the development of the project, because it is what allows the database to be prepared for any type of analysis. In order to perform a correct segmentation analysis, it is necessary to constantly retake this process in order to answer the different questions that arise. The data transformation resulted in the creation of attributes without the use of any statistical criteria, that is to say, only by convention of the research team. These were in particular two attributes: causing, light vehicles and heavy vehicles. About the cause, this classification is based on the grouping of the 51 causes of accidents in five, according to the main responsible for the occurrence of the accident. There are five categories: Drivers, pedestrians, mechanics, passengers and others. About light and heavy vehicles, this classification is a grouping of each of the vehicles involved in the accident according to their size and weight.

First, was made a review of the columns of the raw database obtained from Conaset. After a long process of transformation of this database, in the last stage of this process we have added the following attributes to do the segmentation work:

- **Deceased Drivers**: number of drivers killed in an accident.
- **Serious Drivers**: number of seriously injured drivers in an accident.
- **Drivers Less Serious**: The number of drivers injured in a less serious accident.
- **Drivers Light**: Number of lightly injured drivers in an accident.
- **Affected Drivers**: The number of drivers affected or involved in an accident, equivalent to those involved who did not necessarily experience any physical

damage to property.

- **Passengers Dead**: number of passengers killed in an accident.
- **Passengers Serious**: number of passengers seriously injured in an accident.
- **Passengers Less Serious**: number of passengers less seriously injured in an accident.
- **Passengers Lights**: number of passengers injured in a minor accident.
- **Affected Passengers**: number of passengers affected in an accident, equivalent to those involved and who did not necessarily experience any physical damage.
- **Involved**: total number of passengers involved in an accident.
- **Time range**: hourly range. Four ranges were created to segment the time of occurrence of the accidents.
- **Injuries**: number of people unharmed by accident.
- **Light Vehicles**: number of light vehicles involved in an accident.
- **Heavy Vehicles**: number of heavy vehicles involved in an accident.
- **Vehicle Sum**: total number of vehicles involved in an accident.
- **Classification**: segmentation of the data according to the weighting of each accident in the value of the severity index. The aim is to classify the severity of the accident according to the number of deaths, serious injuries, less serious injuries and minor injuries. Which take a value of 1 to 5 respectively.
- **Cause**: classification that indicates who caused an accident. This classification is based on the grouping of the causes of accidents according to what each of them is due to.

Once the first useful versions of the dataset were obtained, the stage of exploratory analysis of the data was carried out.

At a global level, the main reason why each of the attributes is being added is to generate the inputs to use the severity index and subsequent classification of each one of the accidents according to this indicator. In the first place, a bibliographic analysis of different severity indices used to qualify accidents at a global level was carried out. After the analysis of each indicator was sought to make a similar crash severity index (CSI), based on the final dataset and create a classification of each accident. The classification of each accident, seeks to be a categorization of the accident according to its severity, for it was carried out a treatment of the data based on the study of statistical criteria.

For the creation of the hourly range, an histogram was made to analyze the frequency of accidents in the Region of Valparaiso and how these were distributed throughout the day. The hours with the highest and lowest frequency of accidents were detected.

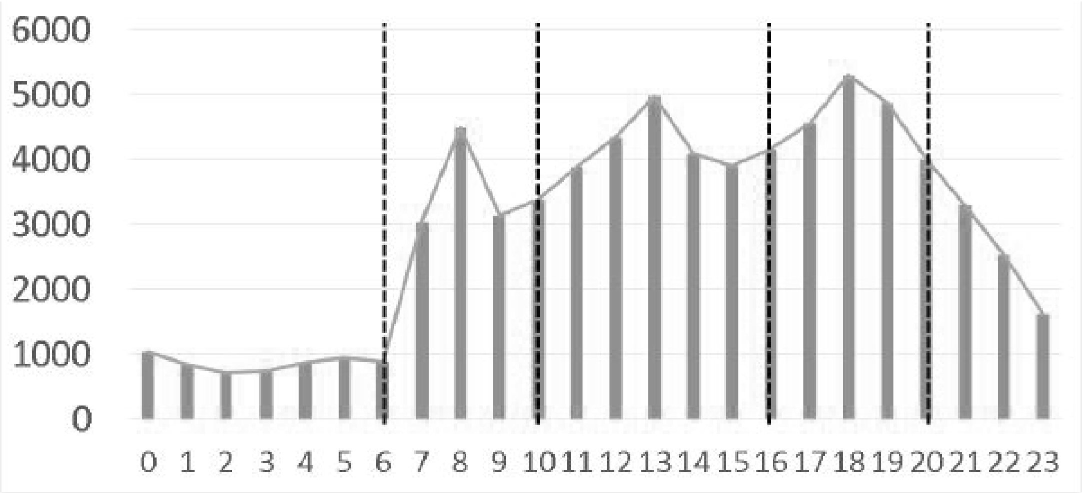


Figure 4.26. Frequency of traffic accidents according to the time of the day

The severity index is the indicator used for the severity of a set of accident occurrences. The set of accidents is taken under various criteria (temporality, causes, location, etc.) and is calculated as the quotient between the number of deaths and the total number of accidents. This indicator only shows the proportion of deaths of the total number of accidents, but does not say anything about the other types of people affected (serious, less serious and slight injuries, are the CONASET criteria), so if there are no deaths in any of these data sets, the severity index will be 0, even if they were all seriously injured. Certainly not considering other types of affected in the measurement, the team working on this project argues that this index is not a consistent indicator for the severity of accidents. It is for this reason that a proposal is presented below for a new, more robust severity index that takes into account more factors that are considered relevant when classifying the occurrence of accidents. From a literature review of accident classifications and calculations for the severity index around the world, it was discovered that, generally, traffic accidents are classified within the KABCO scale, which is based on the judgment of the officer attending the site of the event and is a subjective scale based on the perception of injuries resulting from an accident.

The scale has 5 levels as follows: K is the level when there is at least one fatal victim, A is the level when there are disabling injuries, but not mortality, B is the level when there are people with significant but not disabling injury results, C is the level when injuries are minor, and O is the level when there is only property damage. In Chile such a scale is not used at the time of data collection in the occurrence of a traffic accident and, therefore, there is no classification that indicates the level of severity of a particular accident. Within the CONASET database that manages the equipment, there are variables such as deaths, serious injuries, less serious injuries and minor injuries, which show the number of people that resulted within each of these levels of classification in an accident. From these data, a classification of traffic accidents was generated, which indicates the severity of each occurrence and seeks to be a comparison of the KABCO scale, but generated from quantitative data. The classification and each level used in the classification for each accident are explained below:

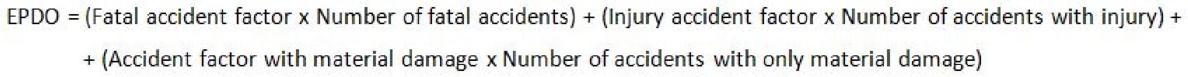
- **Fatal**: there is at least one deceased as a result of the accident.
- **Serious**: there is at least one serious injury in the accident, but no deaths.
- **Less serious**: there is at least one less serious injury and no serious injury or death.
- **Mild**: there is at least one minor injury but no less serious, serious or deceased injury.
- **Property damage**: there is no one physically affected in the accident.

Following a literature review, it was also determined that a widely used worldwide severity index is EPDO (Equivalent Property Damage Only), EPDO assigns a weight factor to the amount of each of the KABCO scale levels for traffic accidents and results in the cost equivalent to material damage from any set of accidents. The procedure for calculating this index is shown below:

- Compilation of the costs incurred for each type of accident: This cost was obtained from calculations previously made by CONASET and MIDEPLAN, which show the costs incurred for each type of involved based on the previously established scale (deceased, severe, less severe, mild). In addition, the costs for damages to light vehicles and heavy vehicles are part of these data. These costs are important when determining the cost for accidents where there is only material damage.

- Calculate the weight of severity factors as a function of the costs of accidents with only material damage: Based on the costs obtained and represented in the tables above, a cost is calculated for each dataset accident by multiplying the sum of the costs per row in Table 1 by the respective number of injured/dead and adding to this the multiplication of the costs per vehicle and per corresponding type of accident by the number of each type of vehicle involved. This gives a total cost for each accident. After this, an average cost is calculated for each of the accident classification levels. Accidents of the serious, less serious and minor type are grouped together in an accident with injury. Once the costs for each level have been obtained, the weights corresponding to each type of accident are calculated according to the classification of fatal accident, accident with injury and accident with only material damage.

- Calculation of the EPDO for each subset of accidents: Once the weight factors for each classification level are obtained, they are incorporated into the formula for the calculation of the equivalent cost for each subset where it is applied:



EPDO = (Fatal accident factor x Number of fatal accidents) + (Injury accident factor x Number of accidents with injury) +
+ (Accident factor with material damage x Number of accidents with only material damage)

The application of cost equivalent to material damage as a severity index gives a value that indicates how severe a set of accidents is based on the costs associated with each level of accident classification. This value is easy to obtain and compare between each of the values of a specific criterion.

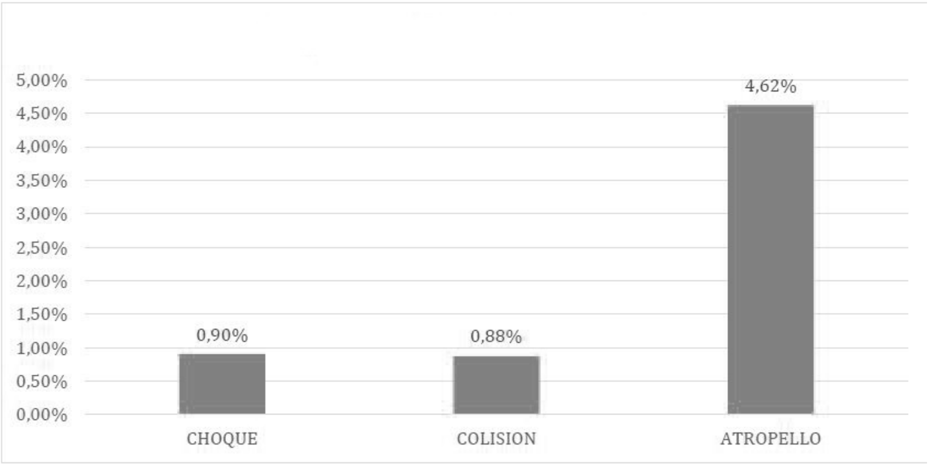


Figure 4.27. Proportion of fatal accidents for each accident

The project focused on two lines of work. First, the project seeks to identify patterns of accident occurrence in the Region of Valparaiso, using segmentation trees based on the CHAID methodology. Second, and complementary to the behavior patterns, a new classification for accidents and a new severity index for sets of accidents were developed. This makes it a more robust indicator than the one currently used by CONASET and allows a clear classification and interpretation for each grouping criterion used. As a result of the work carried out, it was possible to obtain various conclusions from the project and the question itself, as well as from the institutions and their handling of the data:

- Indeed, it is possible to determine accident occurrence patterns with the data obtained. Identifying these patterns can mean significant improvements in preventive measures by focusing on the identified risk factors.

- Accident data collection by policemen is sometimes deficient. This results in a lot of data being lost and unusable for a research project such as this. Data collection and management is tremendously relevant to improving management in any organization. When it comes to public agencies that may allow improvements in public policies for the welfare of citizens are even more relevant. In addition, according to the revised bibliography, it is important to add variables to the dataset that provide us information on the environment at the time of an accident, such as climatic conditions, infrastructure conditions, etc.

- The segmentation made and the patterns found in determining the level of classification of accidents represent risk factors for such a level. Although this project does not contemplate a more elaborate work than finding these patterns, it is understood that these risk factors can be integrated as covariates to a multinomial logistic regression whose dependent variable is the accident classification. This can be used for future work in determining the incidence of each risk factor in the result of an accident.

With respect to the management, it is suggested to make a review of everything that implies a rural area, because as seen in the study, the rural area is much more fatal compared to the urban, this review should be made from all factors that are present at the time of a traffic accident, this involves infrastructure, climatic conditions, road culture, and so on. This in order to decrease the proportion of fatal accidents that occur in the rural areas in the future.

Chapter 5

Case study: PME Project

PME Project is a classic example of a Data Science project which involves all the steps of Data Science, from the data acquisition to the support in the decision making. This project is based on the research for a predictive model for a drinking water network failures. This study will allow to better understand the complexity of the network management. However, the research focuses on predicting failures, and its purpose is to support decision making in the context of renewals and repairs by seeking to reduce the uncertainty about network management, reduce costs related to emergency purchases, avoid fines, and improve six-monthly renewal planning.

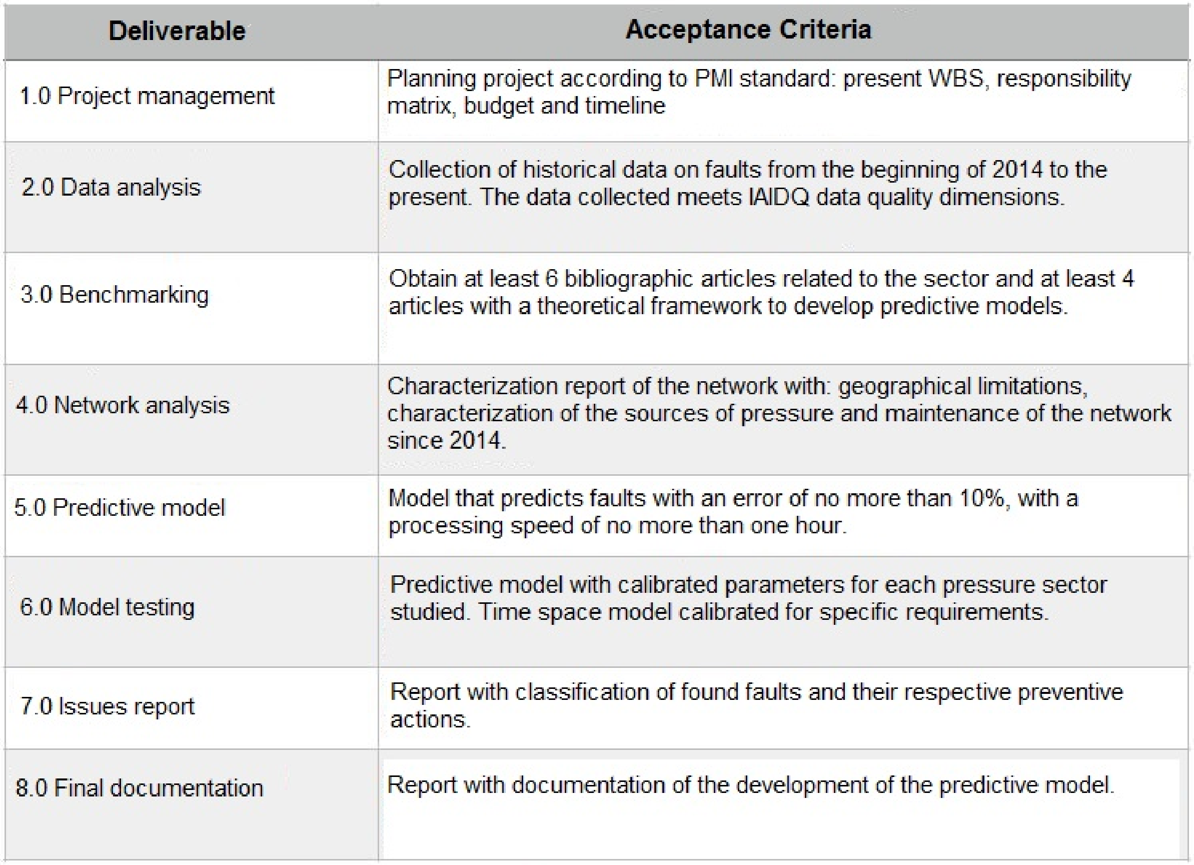
The output is a predictive model for two pressure sectors that allow the identification of pipelines sections that are prone to failure. The greatest risks obtained in the tested sectors correspond to pipelines sections that have a high probability of rupture. The result was validated with new failures observations concluding that the models work. For the first sector, 3 out of 5 possible failures are predicted, and in the second, 2 out of 3 failures are predicted.

Both the time-space model and the predictive model developed by the team allow us to know where there is a greater concentration of failures and which pipelines sections have a greater propensity to fail. This saves: time, since with the time-space model it is no longer necessary to search for failures concentration case-by-case; money, since with the predictive model we pass from an emergency scenario to a preventive scenario, which reduces the costs related to emergency purchases and associated fines.

The use of agile methodologies in the project development, as well as the software development, contributed to the execution of the predictive model for drinking water network failures.

5.1 Project definition

The company in charge of the drinking water network management has more than 603,000 customers, corresponding to 1.9 million of end users, which are supplied with drinking water through an hydraulic network of 5,300 [km] of pipelines divided into more than 109,000 sections. This company is under constant control by the regulatory and supervisory organism of the concession companies that provide drinking water and sewerage services. Through sixthy-monthly audits, the company is inspected for compliance with the requirements imposed to ensure the quality of life of people. These requirements are chemical, such as the hardness or amount of chlorine present in the water, as well as service levels, in terms of continuity of service with respect to the number of failures and affected, as well as the time of suspension of service in case of failure. The company internally is a strongly hierarchical and functional company, in whose, at the top of the hierarchy is the directors, which delegates the administrative function to a general manager, who in turn divides and functionally assigns responsibilities to different managers, who assign responsibilities by geographical area to different sub-managers. Finally, at a final division level, there are the chiefdoms, which respond directly to their corresponding sub-management. The PME team works intensively with the network management department, which aims to propose preventive management actions to reduce failures in the matrices, reducing the number of failures and obstructions in the network as well as to standardize the pressures in order to ensure continuity of service, in the drinking water and sewage systems. The objective of the project is to create a predictive model for drinking water failures to improve the drinking water network management at a cost of UF760. The unit of support (UF) is a unit of account used in Chile, readjustable according to the inflation. Actually, one UF corresponds to 41,30 american dollars. The project arises from the company's need to better manage the drinking water network renewals, since it is currently the company that pays the highest level for fines, at national level. The change in the internal management indicators by the company that regulates the sector, generates the urgency to improve the current processes through which the renewals and repairs of the drinking water network are managed. The renewals correspond to the supply of part of the pipelines as part of a planned renewal plan, while repairs, are modifications to the pipelines in response to an unexpected failure in a section of the drinking water network. This project consists of the development of a predictive model based on Cox's regression, from which is obtained the survival probability of a pipeline in a certain pressure sector based on covariates of interest, such as the diameter and the number of repairs. With the output of this model we perform a ranking to see which pipelines and sectors are most prone to fail. With this predictive model, it is expected that the company can improve the decision making related to the determination of the scheduled renewals and with it the consequent improvement of the indicators associated to the failures in the drinking water network assures the continuity of the service and the decrease of the fines. To realize this predictive model, were used computational tools and forecasting models for failures based on probability distributions over complex networks subject to material fatigue. The expected results, acceptance criteria and scope of the project were modified after meetings between representatives of the company and the PME team.



Deliverable	Acceptance Criteria
1.0 Project management	Planning project according to PMI standard: present WBS, responsibility matrix, budget and timeline
2.0 Data analysis	Collection of historical data on faults from the beginning of 2014 to the present. The data collected meets IAIDQ data quality dimensions.
3.0 Benchmarking	Obtain at least 6 bibliographic articles related to the sector and at least 4 articles with a theoretical framework to develop predictive models.
4.0 Network analysis	Characterization report of the network with: geographical limitations, characterization of the sources of pressure and maintenance of the network since 2014.
5.0 Predictive model	Model that predicts faults with an error of no more than 10%, with a processing speed of no more than one hour.
6.0 Model testing	Predictive model with calibrated parameters for each pressure sector studied. Time space model calibrated for specific requirements.
7.0 Issues report	Report with classification of found faults and their respective preventive actions.
8.0 Final documentation	Report with documentation of the development of the predictive model.

Figure 5.1. Acceptance criteria

5.1.1 Stakeholders System

The system under study corresponds to the whole drinking water network and to all the stakeholders who take part in the decision making process regarding renewals and repairs in the network. In particular, it is composed of 6 components:

- the **network management department**, in charge of proposing preventive management actions to diminish failures in the matrices.
- the **network department**, responsible for the reactive management of failures in the network.
- the **losses department**, responsible for the management of water losses of the network.
- the **company** in charge of supervising the sanitary companies.
- **maintenance department**, responsible for the maintenance operations of the drinking water network.
- **zonal administration**, who provide detailed information on the network.

5.1.2 Drinking Water Network division

The drinking water network being analyzed has about 5,300[km] of installed pipelines. The company currently divides its drinking water network as follows:

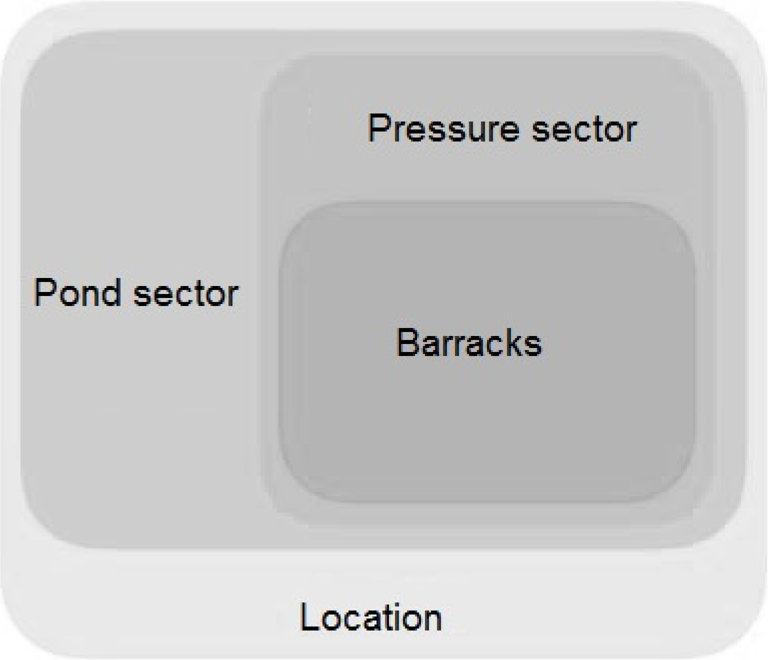


Figure 5.2. Drinking water network structure

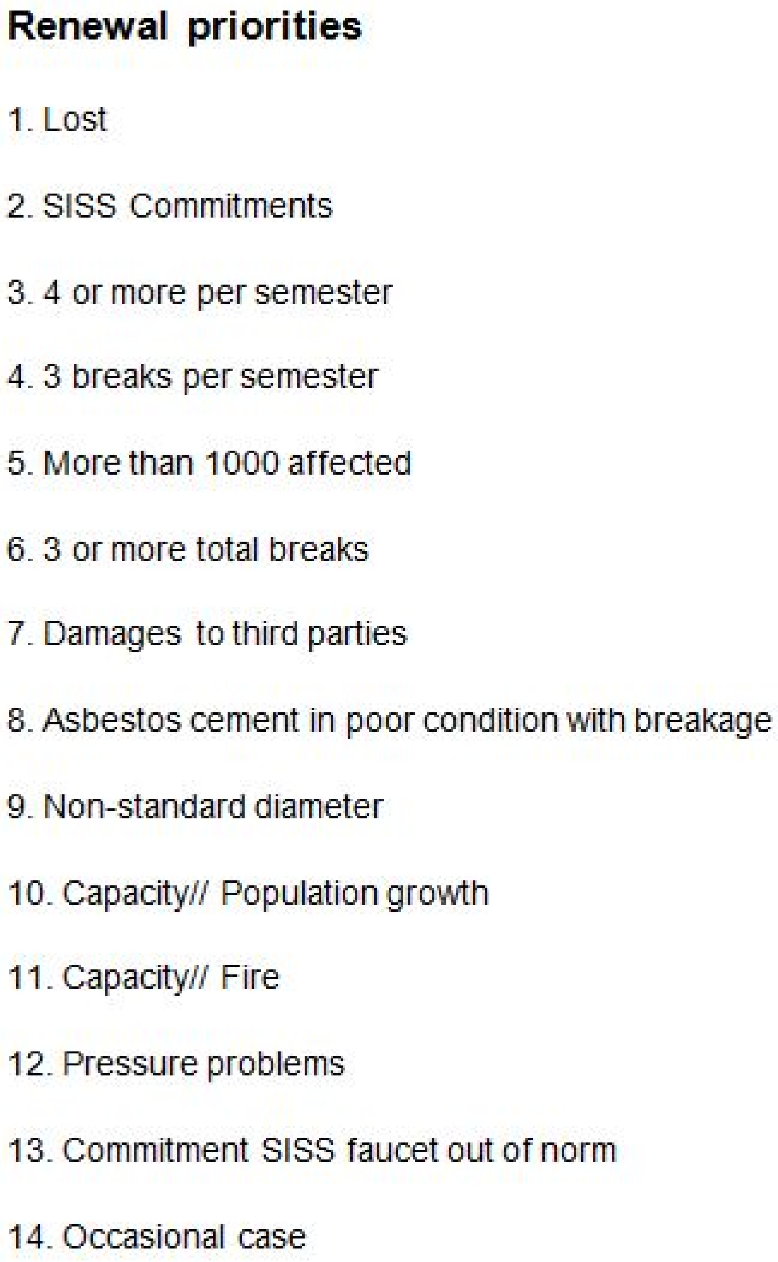
- **Location**: corresponds to the communes, this is the first division that is made to the drinking water matrix. Each locality is divided into several pond sectors.

- **Pond sector**: corresponds to the sector covered by the same pond, the which can be buried, half-buried or correspond to a cup pond. Each pond sector has one or more pressure sectors.

- **Pressure secto**r: these are located within the pond sector and the main characteristics of these sectors are that they are fed by the same source of pressure, for instance they are sectors that are fed by the same pond, pressure reducer or booster. This division however is not exact, given the complex and extensive network, situation that generates that there are geographical areas that are covered by more than one pressure sector.

- **Barracks**: barracks correspond to divisions within pressure sectors and they are the sectorization with the highest resolution that the company has in mind. The barracks is characterized by being a geographic sector that is completely surrounded by shut-off valves that allow, in case of emergency or failure, interrupt the supply of drinking water to a sector without disrupting its operation from the rest of the network. This division, as in the pressure sectors, is not rigorous, a situation that generates the existence of areas that are not within anyone barracks and, on the contrary, areas that are in more than one barracks at the same time.

For the management of renewals, the company has a budget for the renewal of 50[km] per year, at the rate of $200,000/m, that is, approximately $10 billion annually for renovation. Of those 50[km], 10[km] are of exclusive responsibility of the losses department, taking into account as the main reason for the losses of drinking water that are generated along the network. Losses are calculated as the difference between the water the company provides and the water it uses is actually invoiced. The remaining 40[km] are the responsibility of the network management department, which for renewal is governed by the document of renewals prioritization. This document, in the form of a ranking, shows the priorities with which renewals should be carried out in the drinking water network. This document arises from the need for the company to manage the failures that occur in the network.



Renewal priorities

1. Lost
2. SISS Commitments
3. 4 or more per semester
4. 3 breaks per semester
5. More than 1000 affected
6. 3 or more total breaks
7. Damages to third parties
8. Asbestos cement in poor condition with breakage
9. Non-standard diameter
10. Capacity// Population growth
11. Capacity// Fire
12. Pressure problems
13. Commitment SISS faucet out of norm
14. Occasional case

Figure 5.3. Renewal prioritization

5.2 Execution Methods

The project is based on statistics, mathematical modeling, data processing and software development so the PME team adheres to the use of agile methods for its development. High uncertainty projects have high rates of change, complexity and risk. These characteristics may present problems for traditional predictive approaches that aim to determine most requirements in advance and to control changes through a process of requesting change. Instead, agile approaches have been created to explore feasibility in short cycles and adapt quickly based on evaluation and feedback.[82] The four values of the agile manifesto highlight the need to discover better ways of developing software through which we came to evaluate: individuals and interactions on processes and tools, software working on a complete documentation, collaboration with the client in the negotiation of the contract, respond to change by following a plan. Thus, the agile methods provide a flexible and robust methodology as a framework for the activities carried out for the execution of the project. In the context of engineering projects, the greater the relevance of the agile methods the greater the frequency and intensity of a software use and development. To propose these methods, among other outstanding elements, there is the need of an iterative and incremental reference for the project execution. Agile methods are iterative because its use supposes a constant revision and execution that is given to a major frequency than that observed in a conventional project. On the other hand, the agile methods are incremental because each iteration is aimed at delivering a functional contribution that allows to approach marginally the defined objective.

During the execution of the project, the PME team uses the following practices approaches: Kanban, SCRUM and Extreme Programming, which frame the task of the team during the project execution. These approaches deliver dynamism at the same time than flexibility to project development. The most salient aspects of each approach used during the implementation phase are described below[82]:

5.2.1 Kanban

Kanban in lean manufacturing is a system for planning inventory control and replenishment. The word Kanban is literally translated as ‘visual sign’ or ‘paper’. Physical Kanban panels with tabs allow and promote visualization and workflow through the system for everyone to see. This information radiator (large display) consists of columns that represent the states that the work must cross to be able to do so. The Kanban method is used and applicable in many settings and allows a continuous flow of work and value for the customer. The Kanban method is less prescriptive than some agile approaches and therefore less disruptive to implement as it is the original ‘start where you are’ method. Organizations can begin to apply the Kanban methods relatively easily and progress towards full implementation of the method if this is what they deem necessary or appropriate. Unlike most agile approaches, the Kanban method does not prescribe the use of timed iterations. Iterations can be used within the Kanban method, but the principle of continuously pulling individual elements through the process and limiting ongoing work to optimize the flow should always remain intact. The Kanban method can best be used when a team or organization needs the following conditions: flexibility, focus on continuous delivery, increased productivity and quality, increased efficiency, attention of team members, variability of workload, reduction of waste. In the Kanban method, it is more important to complete the work than to start a new job. There is no value derived from the work that has not been completed so that the team works together to implement and comply with the work in progress (WIP) and to ensure that each piece of work through the system is done. During the project execution was used the control of progress of the activities through the use of matrices to do / doing / done, system that provided visualization of work done at any time at the same time as a tool of coordination among the members of the PME team.

5.2.2 SCRUM

Scrum is a single team process framework used to manage product development. The framework consists of roles, events, artifacts and Scrum rules and uses an iterative approach to provide a working product. Scrum runs on timeboxes of 1 month or less, with constant sprint durations, where a potentially releasable increase in product is produced. The Scrum team consists of a product owner, a development team, and a mix master. The product owner is responsible for maximizing the value of the product. The development team is a cross-functional and self-organizing team composed of team members who have everything they need within the team to provide a functioning product without depending on others outside the team. The scrum master is responsible for ensuring that the scrum process is maintained and works to ensure that the scrum team adheres to the practices and rules and trains the team in removing obstacles. In addition to being Lean, this methodology is Agile. Its agility is given by the incremental development of projects that adhere to its use. The contribution of SCRUM during the execution of the project is given by the increments typically weekly that were given for the project development, about the software development as well as the stages of data processing and mathematical modeling. The sprint - long cycle, period from 1 to 4 weeks - will be considered the period of iteration with the client, manifested through coordination meetings and alignment of objectives and expectations, while the Stand Up - short cycle, typically daily - corresponds to the iteration that is given among the members of the PME team.

5.2.3 Extreme Programming

eXtreme Programming (XP) is a software development method based on frequent cycles. The name is based on the philosophy of distilling a given best practice in its purest and simplest form and applying it continuously throughout the project. eXtreme Programming corresponds to the style of programming that is adapted to changing requirements. During the development of the project was faced the need to modify the development of the project and, therefore, of the generated code for these purposes. The concepts: simplicity, communication, feedback, courage and respect proper to the extreme programming were referents for project members during the various stages of the project implementation.

5.3 Execution Stages

The execution of the project takes place through a series of activities with new iterations of the same activities recursively, as detailed below:

5.3.1 Data Acquisition

Together with representatives of the company, it is investigated the operating and use of the SIGEC and the information contained therein is analyzed. Starting with this research and this analysis, is defined a conceptual and convenient configuration to obtain data contained in SIGEC in such a way that these contribute significantly to the development of the project. Within the SIGEC reporting possibilities, the required fields are consulted, which are extracted using a *.xlsx file.

5.3.2 Extract, Transform, Load (ETL)

ETL is a type of data integration that refers to the three steps (extract, transform, load) used to merge data from multiple sources. It is often used to build a data warehouse. During this process, data is taken (extracted) from a source system, converted (transformed) into a format that can be analyzed and stored (loaded) in a data warehouse or other system. When used with a company data warehouse (resting data), ETL provides a deep historical context for the company. By providing a consolidated view,

ETL makes it easier for business users to analyze and report data related to their initiatives. ETL can improve data professionals' productivity by encoding and reusing processes that move data without requiring technical expertise to write code or scripts. ETL has evolved over time to support emerging integration requirements for things like data streaming. Organizations need both ETL and ELT to gather data, maintain accuracy, and provide the auditing typically required for data warehousing, reporting, and analysis.

ETL's goal is to produce clean, accessible data that can be used for analysis or business operations. Raw data must be extracted from a variety of sources. The application of an adequate ETL allows for the expeditious use and manipulation of data through a software. Once an ETL has been carried out, the data is free of any undesired effects generated by: white cells, syntax errors, format and mathematical formulation of the information system. The extraction (E) of the data corresponds to the immediately previous stage, already described. The extracted data is then sometimes placed in a destination such as a data warehouse. The ETL transformation stage is where the most critical work is done. The most important thing about transformation is to apply all business rules to the data to meet the reporting requirements. Transformation changes raw data into correct reporting formats. If the data is not cleaned, then it becomes more difficult to apply the business rules for reporting. Transformation is achieved through a series of rules and regulations that are outlined. Standards that ensure the quality and accessibility of data during this stage should include: Standardization, Deduplication, Verification and Sorting. These transformation steps reduce what was once a mass of unusable material into a product data that you can present in the final stage of ETL, the loading stage.

The transformation (T) is the stage of data manipulation in terms of: (1) structure, (2) semantics, (3) content and, (4) format. The detail of this ETL stage, in these terms, is as shown below:

- **Structure**: data are characterized, identifying the quantity and semantics of the existing fields (142 columns). In this point you can also characterize the data integrity looking for empty, faulty, inconsistent format and the existence of mathematical formulations that could prejudice the future manipulation of the data.

- **Semantics**: each field is redenominated by editing each element of the header, the latter, typically located in the first row of the archive. This procedure must generate a header: (1) complete, (2) semantically adequate and (3) compatible with the data reading stage through software.

- **Content**: based on the work carried out on the data structure, the following shall be to manage empty cells, faulty records, inconsistencies of format and the existence of existing mathematical formulations in the data, for each field. This procedure is carried out using filters, conditional format and MSExcel pivot tables. The extent of this stage will depend on the amount of data processed, the proportion of data they require manipulation and the complexity of the format inconsistencies and of the mathematical formulations existing in the data.

- **Format**: it is necessary to generate files containing the data, that respond to some standard of reading files by the software, question which involves the use of certain extensions, typically *.csv and *.txt, which are often different from the file extensions that are used to perform the ETL. At this point, it was decided to generate datasets in files *.csv from the *.xlsx files obtained in the ETL process.

Finally, the last step of a typical ETL process is to load this extracted and transformed data. There are two typical ways to load data into a data warehouse: full load and incremental load. The load (L) alludes to the set of activities that allow the reading of data through a software. At this stage the use of RStudio was considered for the generation of code that allows the reading of the available data.

R is a statistical software. The R source code is written in C language and some routines in Fortran language. So, R can be recognized as a language of computer programming. R is an interpreted language such as Java and not a compiled language such as C or Fortran. This means that the commands typed on the keyboard are run directly without the need for a compiler. R is an object-oriented matrix language: this means that variables and functions, and data and results, etc., are stored in the RAM memory in the form of objects with a specific name. The user can modify or manipulate R objects with operators (arithmetic, logical or comparative) and functions (which in turn are also objects). R has a very simple data management and is very versatile to build graphs. R has some advantages like : free distribution software, lightweight and easy to install, any R user can create their own custom functions and analysis, great capacity to store and manipulate Big Data, access to a great source of information on the Internet (blog, forums, manuals, among others) on programming in R, available in Linux, macOS and Windows; and disadvantages because it is not easy to use for those who do not have basic knowledge in programming, but there are interactive alternatives. R is able to extract, transform and load the data to be analyzed. As R is free software, packages were created for each part of ETL with Business Intelligence applications. Some of these packages are: Extract and Load and Transform. This reflects the various options that R offers to solve the ETL problem. There is no single module for each ETL stage. So, R is able to assume the role of process engine related to Business Intelligence. RStudio is a free R plug-in that uses the computer's graphics memory to facilitate interactions with R.[83]

5.3.3 Exploratory data analysis (EDA)

Exploratory Data Analysis (EDA) is an approach for data analysis that uses a variety of techniques (mostly graphical) to maximize the understanding of a data set, discover the underlying structure, extract important variables, detect abnormal values and abnormalities, test the underlying assumptions, develop thrifty models and determine optimal factor settings. Most EDA techniques are graphical in nature with some quantitative techniques. EDA emphasizes graphic techniques, while classic techniques emphasize quantitative techniques. In practice, an analyst typically uses a mix of graphic and quantitative techniques. The particular graphic techniques used in EDA are often very simple, consisting of various graphic techniques: plotting the raw data; plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data; positioning such plots so as to maximize our pattern-recognition abilities. Once the data has been loaded through the software, we continue with the set theory analysis associated with the fields and with the quantitative analysis of the values they take. The set theory associated with the fields refers to the possible values taken by the different fields, record by record, and the combinations of possible values that are given between the values that take different fields. Quantitative analysis represents the analysis that is made, field by field, of its values for effects of the characterization. In this context, pivot tables and graphs are made, which allow for graphically characterize the available data. Then, an univariate analysis is performed using R for the variables of interest. A multivariable analysis is then carried out by calculating correlations between variables of interest. Finally, the variables are adjusted to probability distributions via RStudio.

5.3.4 Forecasting model definition

Depending on the results obtained in the EDA, we proceed to define a set of statistical models that serve as a reference for making failures forecasts in the drinking water network. At the same time, the variables of interest are selected for the effects of the statistical analysis; for simplicity of computational processing and software development, subsets of data are generated with those variables representative for the selected statistical models. Among the selected variables, the most important are the ones of mechanical and temporal connotation. Through the values they take, the mechanical variables describe the material wear that occurs in the drinking water network pipelines. Temporal variables correspond to variables that characterize the moments in which occur the network failures. It is key at this point the conversion of the temporal variables to time between failures, because this family of variables has the same characteristics of probability distributions such as Weibull or Exponential.

Once the statistical model has been defined and the data to be used in it, we proceed to use code through R libraries that allows the processing of the data loaded in statistical terms for the failures prediction. With the model execution, the numerical and graphical outputs of the model are generated. Then, we look for those elements of the output that are susceptible of improvement by calibrating the parameters of the model used. Subsequently, the survival analysis of the drinking water network is carried out by the use of the Kaplan-Meier estimator for the observed failure time, that is the time elapsed from a time reference to the occurrence of each date. Finally, we proceed to the implementation of the Cox's regression model.

5.3.5 Forecasting model test

From new data on the drinking water network failures, a new ETL is made, with the same structure used in the previous ETL to get a new set of contrast data respect to the initial execution of the predictive model. The initial prediction provided by the model is compared with the new failures recorded. From this contrast you get the detail of those correct predictions as well as the of type I (false positives) and type II (false negatives) errors.

5.3.6 Documentation

The work carried out is documented highlighting the relevant aspects in terms of the methodology used, the characterization of the drinking water network and the main mathematical and statistical topics addressed. In terms of software, the final versions of the graphic outputs generated by this code are documented, while autonomously executable code versions are generated from external hardware.

5.4 Project execution and Results

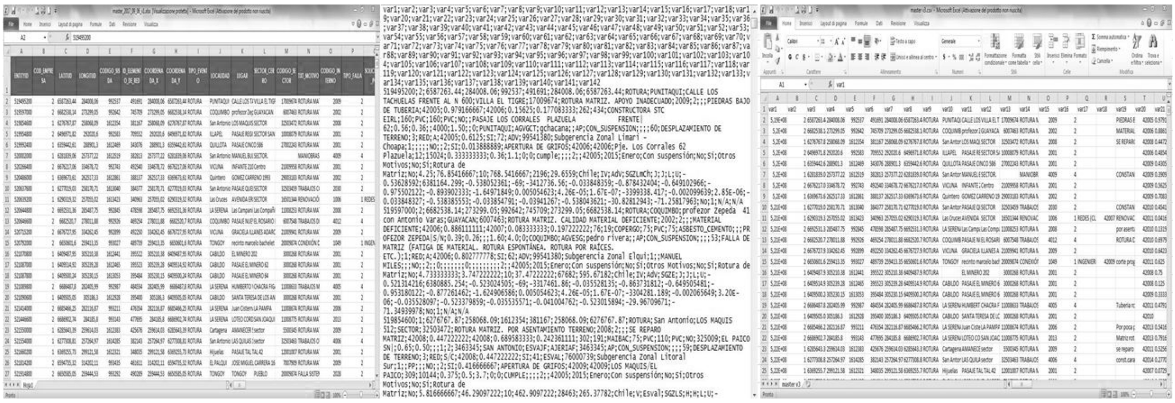
For the project development, SIGEC has been considered as the only information source, as it contains centralized and updated data: (1) that allow characterizing the drinking water network, (2) storing the data associated to the events that affect the network and the actions that are executed on it and, (3) preventive actions that are executed on the drinking water network within the context of the six-monthly planning of material renewal. Four documents from SIGEC were used as information source and we described them below:

- **Event Master**: detail of the events that took place in the drinking water network.
- **Crossroads Network Cutting Sector**: detail of the network sections by cutting sectors
- **Pressure and Cutting Sector**: detail of the sections in the pressure sector with the following cutting sector.
- **Event Master II**: detail of the events that took place in the drinking water network.

From these four archives was created a data set that allowed its use in R to perform exploratory data analysis and the Cox's regression model. The data extracted in the Event Master from the SIGEC constitute the records of events occurring in the drinking water network. The first effort made with the data is the homologation of data for purposes of use, while a second effort addressed the characterization of these in terms of: (1) rows management, (2) columns management, and (3) semantic.

5.4.1 Data characterization

The data was provided in *.xlsx files, which is the reason why it was processed in order to allow its adequate loading to the software in use. The header was standardized and a *.csv dataset was generated using MS Excel.



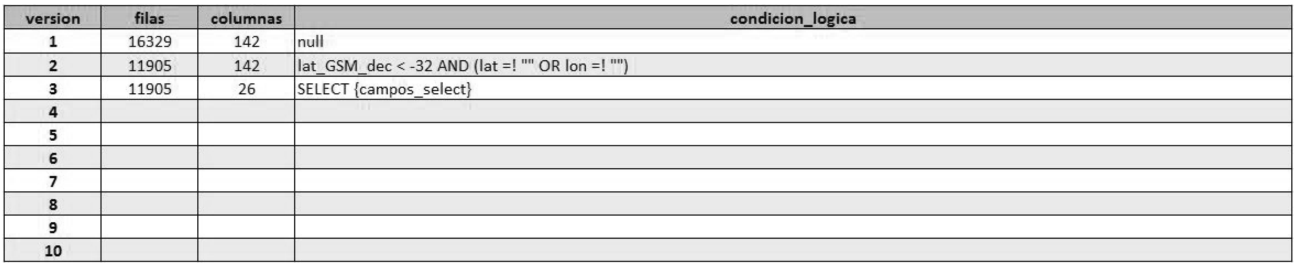
```
var1;var2;var3;var4;var5;var6;var7;var8;var9;var10;var11;var12;var13;var14;var15;var16;var17;var18;var1
9;var20;var21;var22;var23;var24;var25;var26;var27;var28;var29;var30;var31;var32;var33;var34;var35;var36
;var37;var38;var39;var40;var41;var42;var43;var44;var45;var46;var47;var48;var49;var50;var51;var52;var53;
var54;var55;var56;var57;var58;var59;var60;var61;var62;var63;var64;var65;var66;var67;var68;var69;var70;v
ar71;var72;var73;var74;var75;var76;var77;var78;var79;var80;var81;var82;var83;var84;var85;var86;var87;va
r88;var89;var90;var91;var92;var93;var94;var95;var96;var97;var98;var99;var100;var101;var102;var103;var10
4;var105;var106;var107;var108;var109;var110;var111;var112;var113;var114;var115;var116;var117;var118;var
119;var120;var121;var122;var123;var124;var125;var126;var127;var128;var129;var130;var131;var132;var133;v
ar134;var135;var136;var137;var138;var139;var140;var141;var142
519495200;2;6587263.44;284008.06;992537;491691;284008.06;6587263.44;ROTURA;PUNITAQUI;CALLE LOS
TACHUELAS FRENTE AL N 600;VILLA EL TIGRE;17009674;ROTURA MATRIZ. APOYO INADECUADO;2009;2;;;PIEDRAS BAJO
DE TUBERIA;42005;0.979166667;42006;0.15625;0.177083333;262;434;CONSTRUCTORA STC
EIRL;160;PVC;160;PVC;NO;;PASAJE LOS CORRALES  PLAZUELA           FRENTE|
62;0.56;0.36;;4000;1.50;;0;PUNITAQUI;AGVGCT;gchacana;;AP;CON_SUSPENSION;;;;60;DESPLAZAMIENTO DE
TERRENO;3;RED;A;42005;0.6125;SI;72;ADV;99541380;Subgerencia Zonal Limari -
Choapa;1;;;;;NO;;2;SI;0.013888889;APERTURA DE GRIFOS;42006;42006;Pje. Los Corrales 62
Plazuela;12;15024;0.333333333;0.36;1.1;0;0;cumple;;;;2;;42005;2015;Enero;Con suspensión;No;Sí;Otros
Motivos;No;Sí;Rotura de
Matriz;No;4.25;76.85416667;10;768.5416667;2196;29.6559;Chile;IV;Adv;SGZLmCh;J;J;L;U;-
0.53628592;6381164.299;-0.538052361;-69;-3412736.56;-0.033848359;-0.878432404;-0.649102966;-
0.975502122;-0.893902333;-1.64971849;0.005054623;4.26E-05;1.67E-07;-3399338.417;-0.002099639;2.85E-06;-
0.033848327;-0.538385553;-0.033854791;-0.03941267;-0.538043621;-30.82812943;-71.25817963;No;1;N/A;N/A
519597000;2;6682538.14;273299.05;992642;745709;273299.05;6682538.14;ROTURA;COQUIMBO;profezor Zepeda  41
con Antonio Varas;GUAYACAN;6007463;ROTURA MATRIZ. CALIDAD MATERIAL DEFICIENTE;2002;2;;;MATERIAL
DEFICIENTE;42006;0.886111111;42007;0.083333333;0.197222222;76;19;COPERGO;75;PVC;75;ASBESTO_CEMENTO;;;PR
OFEZOR ZEPEDA|S/N;0.39;0.26;;;1.60;4.0;0;COQUIMBO;AGVESG;pedro rivera;;AP;CON_SUSPENSION;;;;53;FALLA DE
MATRIZ (FATIGA DE MATERIAL. ROTURA ESPONTÁNEA. ROTURA POR RAÍCES.
ETC.);1;RED;A;42006;0.802777778;SI;62;ADV;99541380;Subgerencia Zonal Elqui;1;;MANUEL
MILES;;;NO;;2;;0;;;;;;;0;;;;;;;;;2;;42005;2015;Enero;Con suspensión;No;Sí;Otros Motivos;No;Sí;Rotura de
Matriz;No;4.733333333;3.747222222;10;37.47222222;67682;595.67182;Chile;IV;Adv;SGZE;J;J;L;U;-
0.521314216;6380885.254;-0.523024505;-69;-3317461.86;-0.035528135;-0.863731812;-0.649505481;-
0.953180122;-0.877261462;-1.624906586;0.005054623;4.26E-05;1.67E-07;-3304281.189;-0.002065649;3.20E-
06;-0.035528097;-0.523379859;-0.035535571;-0.041004762;-0.523015894;-29.96709671;-
71.34939978;No;1;N/A;N/A
519854600;1;6276767.87;258068.09;1612354;381167;258068.09;6276767.87;ROTURA;San Antonio;LOS MAQUIS
512;SECTOR;32503472;ROTURA MATRIZ. POR ASENTAMIENTO TERRENO;2008;2;;;SE REPARO
MATRIZ;42008;0.447222222;42008;0.689583333;0.242361111;302;191;MAIBAC;75;PVC;110;PVC;NO;325009;EL PAICO
SN|;0.65;0.50;;;1;2;3463345;SAN ANTONIO;ESVAJF;AJERIAF;3463345;AP;CON_SUSPENSION;;;;59;DESPLAZAMIENTO
DE TERRENO;3;RED;S/C;42008;0.447222222;SI;41;ESVAL;76000739;Subgerencia Zonal Litoral
Sur;1;;PP;;;NO;;2;SI;0.416666667;APERTURA DE GRIFOS;42009;42009;LOS MAQUIS/EL
PAICO;309;10144;0.375;0.5;3.7;0;0;CUMPLE;;;;2;;42005;2015;Enero;Con suspensión;No;Sí;Otros
Motivos;No;Sí;Rotura de
Matriz;No;5.816666667;46.29097222;10;462.9097222;28463;265.37782;Chile;V;Esval;SGZLS;H;H;L;U;-
```

Figure 5.4. From *.xlsx to *.csv

In order to manage the rows, a filter was applied with respect to the latitude of the location of events. Only those latitudes less than -32º were considered, in order to consider only the records of events belonging to the Region we are considering. The management of the columns was reflected by grouping the 142 columns of the Event Master in the following categories: event classification, date, drinking water network, identification and location, each of them being a set of variables of the same semantics. The description of each of the categories used is as follows:

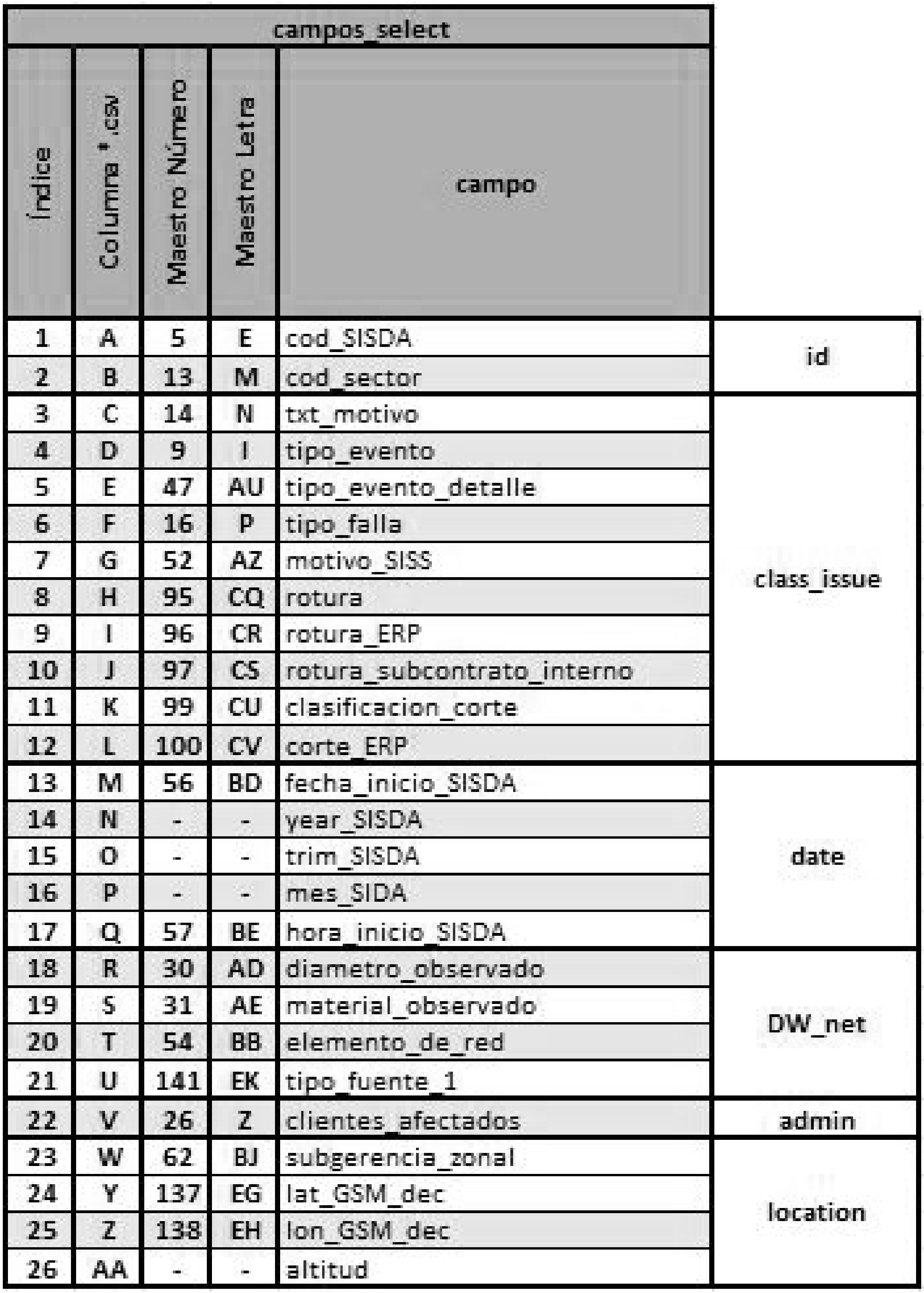
- **Classification of event**: fields that categorize internally or according to the standards.
- **Date**: temporary references to year, month, day, hour and minute.
- **Drinking water network**: mechanical characteristics of the material of the drinking water network at the point of occurrence of the event or other interventions.
- **Identification**: codes and identifiers for network components and indices identifiers of the events registered in the Event Master.
- **Location**: geographic connotation fields that allow to geolocate the events recorded in the Event Master.

For the selection of fields of interest 2 types are differentiated: the fields used for the creation of the Cox's model and the fields used to determine the distribution function of time between failures of the pressure sectors. The fields of interest for the Cox's model creation correspond mainly to those covariates that are believed to be influential in drinking water network matrix failures. These covariates that are believed to be influential arise a priori from the expert opinion of the client, who has worked for years analyzing the network, and from the academic literature that exists on drinking water network worldwide. These fields correspond to the diameter, material, age and length of the pipelines sections. For the calculation of the Kaplan-Meier non-parametric estimator, the exact date of failure and the number of times it has failed, are also considered. Finally, identification columns, such as the section and sector id, are considered. The fields of interest for determining the distribution function of time between failures was only the column time between failures, which arises from the difference between the exact date of the failure ordered from the oldest to the last one registered. Thus, for the calculation of the time between failures we have (n - 1) records for a set of n failures.



version	filas	columnas	condicion_logica
1	16329	142	null
2	11905	142	lat_GSM_dec < -32 AND (lat =! "" OR lon =! "")
3	11905	26	SELECT {campos_select}
4			
5			
6			
7			
8			
9			
10			

Figure 5.5. Master history



campos_select					
Índice	Columna *.csv	Maestro Número	Maestro Letra	campo	
1	A	5	E	cod_SISDA	id
2	B	13	M	cod_sector	
3	C	14	N	txt_motivo	class_issue
4	D	9	I	tipo_evento	
5	E	47	AU	tipo_evento_detalle	
6	F	16	P	tipo_falla	
7	G	52	AZ	motivo_SISS	
8	H	95	CQ	rotura	
9	I	96	CR	rotura_ERP	
10	J	97	CS	rotura_subcontrato_interno	
11	K	99	CU	clasificacion_corte	
12	L	100	CV	corte_ERP	
13	M	56	BD	fecha_inicio_SISDA	date
14	N	-	-	year_SISDA	
15	O	-	-	trim_SISDA	
16	P	-	-	mes_SIDA	
17	Q	57	BE	hora_inicio_SISDA	
18	R	30	AD	diametro_observado	DW_net
19	S	31	AE	material_observado	
20	T	54	BB	elemento_de_red	
21	U	141	EK	tipo_fuente_1	
22	V	26	Z	clientes_afectados	admin
23	W	62	BJ	subgerencia_zonal	location
24	Y	137	EG	lat_GSM_dec	
25	Z	138	EH	lon_GSM_dec	
26	AA	-	-	altitud	

Figure 5.6. Fields of interest

The variables used are variables resulting from various stages of ETL and manipulation (and some algebra) of the resulting data to obtain variables that made sense and were a contribution to the Cox's regression model.

The variables are as follows:

- **id_tramo**: identifier of each section of the pipeline in the drinking water network.
- **fecha_exacta**: date containing the effect of year, month, day, hour and minutes.
- **diametro_cc**: diameter of the pipe in millimetres.
- **ranking_material**: label for pipeline material. 1 for asbestos, 2 for iron (steel), 3 for 'others', 4 for PVC and 5 for HDPE.
- **asbesto**: binary variable about the material. 1 If the pipeline is made of asbestos, 0 if not.
- **fierro**: binary variable about the material. 1 If the pipeline is constructed of iron (steel), 0 if not.
- **otros**: 0 if the pipeline is made of asbestos, iron (steel), PVC or HDPE. 1 if not.
- **PVC**: binary variable about the material. 1 If the pipeline is constructed of PVC, 0 if not.
- **HDPE**: binary variable about the material. 1 If the pipeline is constructed of HDPE, 0 if not.
- **antigua**: binary variable. Separates old materials (1) from new materials (0).
- **nueva**: binary variable. Separates old materials (1) from new materials (0).
- **largo_tramo**: length of the pipeline in metres.
- **cod_sector_presion**: code of the sector in which the pipeline section is located.
- **cantidad_fallo**: number of failures of the section.
- **fallo_binario**: binary variable. 1 if the section has already failed, 0 if not.
- **tiempo_observado**: time of failure.

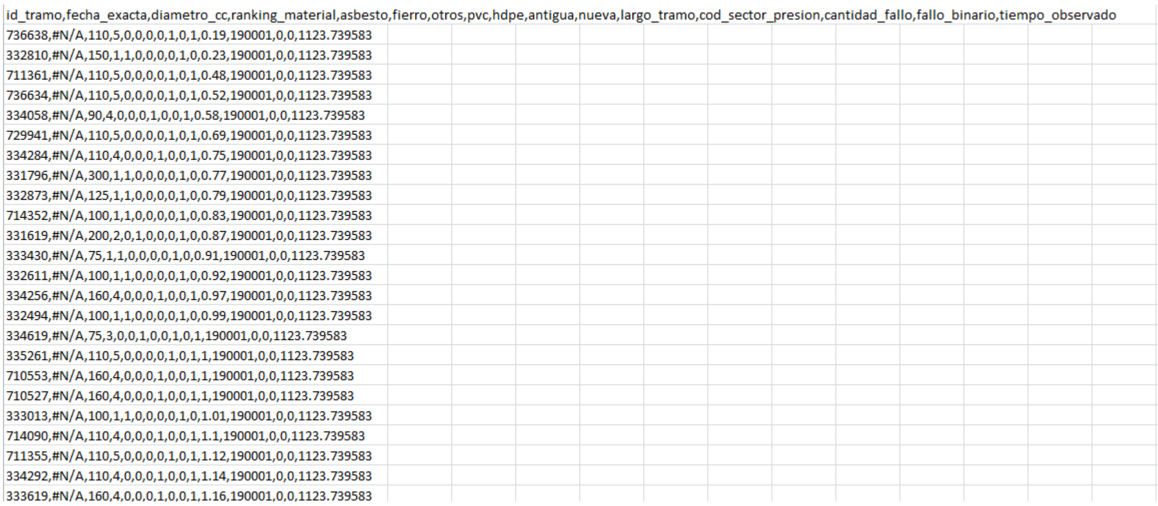
Since this project was executed with a non-trivial and non-sequential sequence of ETL on the original data, there is no a single set of independent variables from which (and through a function) dependent variables are calculated. Said that, in part of the project, this question can be answered in two ways, (in other stages of the project you can answer in many other ways).

First set of variables

- **Independent variables**: the 142 (not strictly all) initial variables.
- **Dependent variables**: the variables that I previously defined, obtained as resultant from the original 142 variables

Second sets of variables (for Cox's regression)

- **Dependent variables**: instantaneous risk (h(t,X1,...Xn))
- **Independent variables**: for the implementation of the Cox regression, a dataset was used, which was intentionally constructed in order to make use of the functions that exist in R for the Cox's regression. For instance, binary variables like cantidad_fallo, fallo_binario and largo_tramo (although were discarded that ones showing a cross in the graphs, which means that the variable is not proportional to the risk of failure).



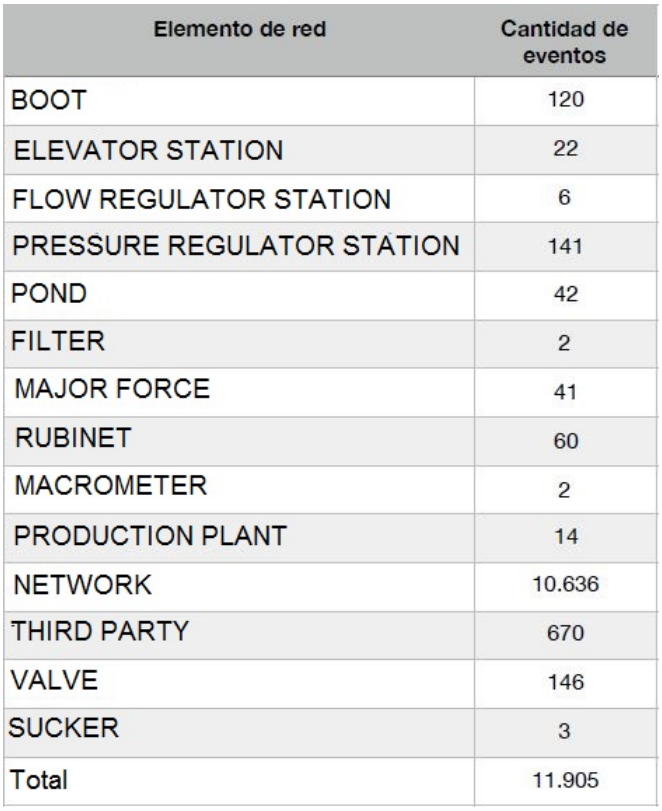
```
id_tramo,fecha_exacta,diametro_cc,ranking_material,asbesto,fierro,otros,pvc,hdpe,antigua,nueva,largo_tramo,cod_sector_presion,cantidad_fallo,fallo_binario,tiempo_observado
736638,#N/A,110,5,0,0,0,0,1,0,1,0.19,190001,0,0,1123.739583
332810,#N/A,150,1,1,0,0,0,0,1,0,0.23,190001,0,0,1123.739583
711361,#N/A,110,5,0,0,0,0,1,0,1,0.48,190001,0,0,1123.739583
736634,#N/A,110,5,0,0,0,0,1,0,1,0.52,190001,0,0,1123.739583
334058,#N/A,90,4,0,0,0,1,0,0,1,0.58,190001,0,0,1123.739583
729941,#N/A,110,5,0,0,0,0,1,0,1,0.69,190001,0,0,1123.739583
334284,#N/A,110,4,0,0,0,1,0,0,1,0.75,190001,0,0,1123.739583
331796,#N/A,300,1,1,0,0,0,0,1,0,0.77,190001,0,0,1123.739583
332873,#N/A,125,1,1,0,0,0,0,1,0,0.79,190001,0,0,1123.739583
714352,#N/A,100,1,1,0,0,0,0,1,0,0.83,190001,0,0,1123.739583
331619,#N/A,200,2,0,1,0,0,0,1,0,0.87,190001,0,0,1123.739583
333430,#N/A,75,1,1,0,0,0,0,1,0,0.91,190001,0,0,1123.739583
332611,#N/A,100,1,1,0,0,0,0,1,0,0.92,190001,0,0,1123.739583
334256,#N/A,160,4,0,0,0,1,0,0,1,0.97,190001,0,0,1123.739583
332494,#N/A,100,1,1,0,0,0,0,1,0,0.99,190001,0,0,1123.739583
334619,#N/A,75,3,0,0,1,0,0,1,0,1,190001,0,0,1123.739583
335261,#N/A,110,5,0,0,0,0,1,0,1,1,190001,0,0,1123.739583
710553,#N/A,160,4,0,0,0,1,0,0,1,1,190001,0,0,1123.739583
710527,#N/A,160,4,0,0,0,1,0,0,1,1,190001,0,0,1123.739583
333013,#N/A,100,1,1,0,0,0,0,1,0,1.01,190001,0,0,1123.739583
714090,#N/A,110,4,0,0,0,1,0,0,1,1.1,190001,0,0,1123.739583
711355,#N/A,110,5,0,0,0,0,1,0,1,1.12,190001,0,0,1123.739583
334292,#N/A,110,4,0,0,0,1,0,0,1,1.14,190001,0,0,1123.739583
333619,#N/A,160,4,0,0,0,1,0,0,1,1.16,190001,0,0,1123.739583
```

Figure 5.7. Data set for the Cox's regression model

5.4.2 General data analysis

The general data analysis generated knowledge about the drinking water network through a preliminary exploration in Excel in order to identify relevant variables, to make a diagnosis of the drinking water network and to know how the department of network management administers the data on the drinking water network.

The drinking water network is considered a complex system because it presents several elements or equipment, including: shut-off valves, pressure reducing valves, booster or bombs, ponds, others. Of these elements the most damaged is the pipeline, this one presents 8,189 failures related to material fatigue or wear causing breakage, 530 failures due to contractor work and 478 renewals. One of the reasons why the equipment focuses on the pipeline element is due to its high rate of intervention and occurrence of failures. In fact, the ratio between the number of events in the pipelines element versus total events in the considered region was 89.34%, with 10.66% being failures in the other equipment. Furthermore it is verified that 10,636 events occur in the pipelines elements. The events include: failures, renovations, maintenance, other interventions.



Elemento de red	Cantidad de eventos
BOOT	120
ELEVATOR STATION	22
FLOW REGULATOR STATION	6
PRESSURE REGULATOR STATION	141
POND	42
FILTER	2
MAJOR FORCE	41
RUBINET	60
MACROMETER	2
PRODUCTION PLANT	14
NETWORK	10.636
THIRD PARTY	670
VALVE	146
SUCKER	3
Total	11.905

Figure 5.8. Events per network element

The antiquity of drinking water networks in the world and technological progress are two reasons why the drinking water network is composed of more than 15 types of different materials. In this drinking water network we can observe the presence of 19 different materials, among which we can classify them in two families: new and old pipelines. The first family is made up of: PVC and HDPE, which correspond to plastic pipelines. The second consists mostly of: steel, cast iron and asbestos cement, which are called metal pipelines. The material with the greatest presence in the drinking water network is asbestos cement, with 39% of the total, followed by PVC with 33% and finally HDPE with 19%. It is possible also note that of the 5,398[km] corresponding to the total drinking water network, the total length per family of pipelines is: 2,824[km] of new pipelines and 2,574[km] belonging to the old pipelines.

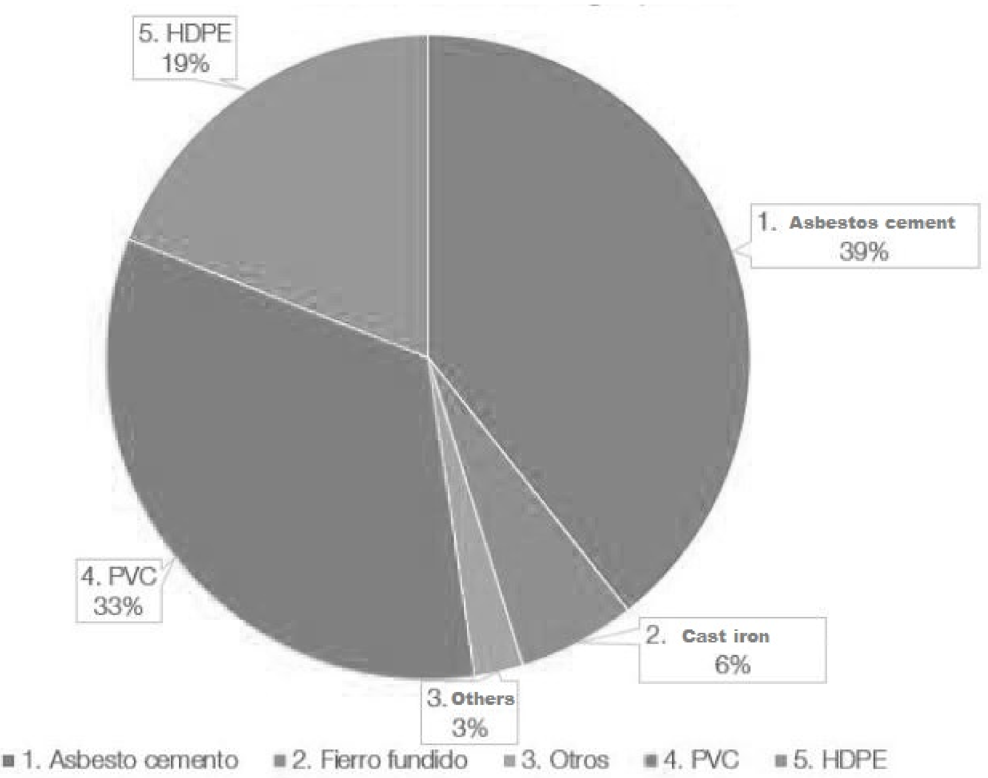


Figure 5.9. Drinking water network materials

Every time the company identifies an event within the network, registers it in the SIGEC. These events are classified into four types of events: programmed, not programmed, of majeure force and third party actions. In addition to this first classification, there is a level of additional detail for each type of event, detail that is achieved through the rupture classification that is reported as the reason for the event. According to these classifications the team makes the decision to analyze the programmed and not programmed events, excluding events caused by actions of third parties, floods or earthquakes. This decision implies working with more than 90% of the events registered. However, not all of the not programmed events can be analyzed in the same way, this is due to 2 effects: first, in this set there are failures that do not occur in pipelines, but in equipment, such as boosters or pressure reducers, which due to their physical-mechanical characteristics must be analyzed independently. Second, there are effects, such as earthquakes and human factors, that are difficult to predict. There is a large differential between the number of programmed events versus the number of not programmed events. This means a priori that the drinking water network deteriorates faster than it renews itself.

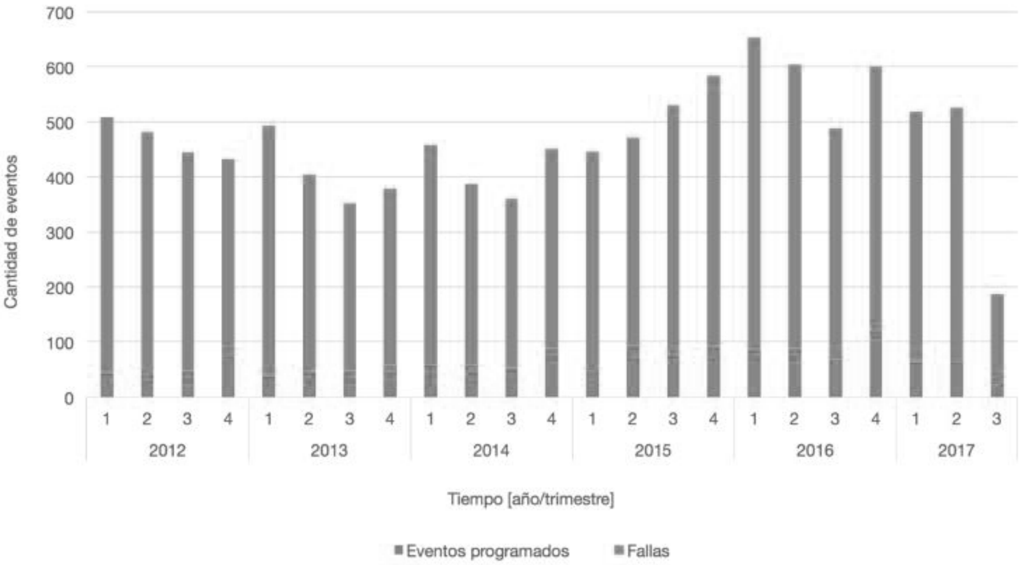


Figure 5.10. Failures number for failures type

The most frequent reasons for failures are: matrix rupture, internal contractor and equipment failure. It is decided to use not programmed events that have as a reason the rupture. This can be caused by material fatigue or natural deterioration of the network.

One of the important aspects to consider is the number of events that presents a pressure sector with the elements that this implies, since these will be the critical sectors for the company and given the amount of failures data they have, they will be the main candidates for the realization and testing of the predictive model. The PME team identifies the existence of a large number of events that do not have assigned a pressure sector, which means that for 5,360 events it becomes unreachable to identify the pressure sector. Basing on this criterion, the PME team decides to study the two pressure sectors that present the most number of events, which are are: 190001 and 270001.

5.4.3 Exploratory data analysis in R

Exploratory data analysis is a necessary process that allows us to understand the behaviour of the variables and the relationships between them, to distinguish trends and adjustments of distributions that are adapted to empirical data and to analyze the possible models that can be applied. This data analysis is more sophisticated and deep than the general data analysis. In this new exploration the objective is to work exclusively with failures that are matrix rupture and that have as a reason the wear and the fatigue of the material and natural variables typical of the network.

In order to understand the location and concentration of failures it is necessary to generate a data visualization, so a data set that contemplates latitude and longitude fields is configured, the data is loaded and a plot is generated with limits in the coordinates to have a view of the Region. Plotting a map is important because the location and concentration data, together with the temporal data, allow to answer to the question where? and when? a failure is occurring.

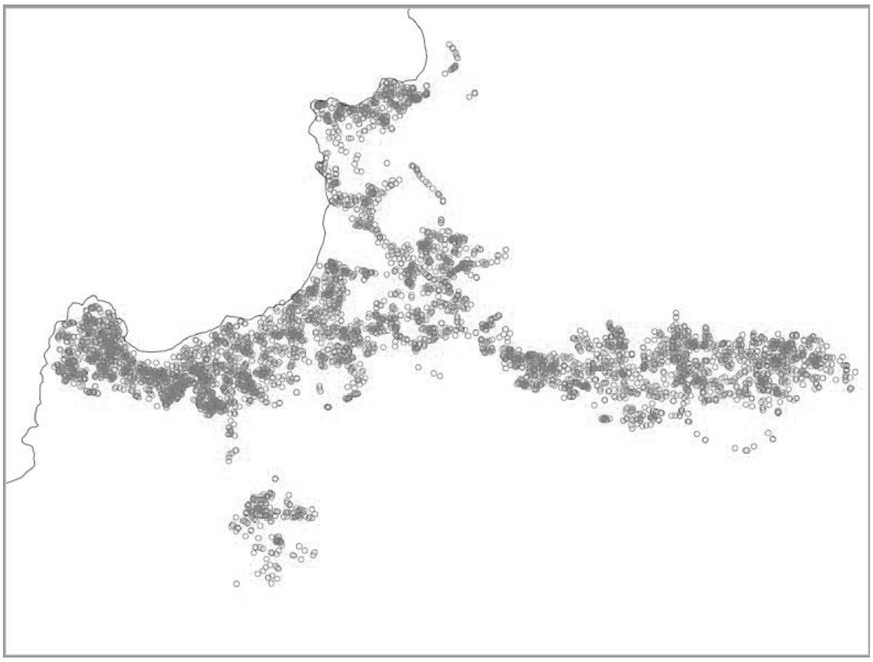


Figure 5.11. Failures location and concentration - Map

The software R is used to iteratively run the exploratory analysis of data for each of the selected pressure sectors. The first command to be executes is summary() to the dataset used for the creation of the Cox model for a pressure sector. In the sector 190001, in this summary it is possible to identify a greater number of failures in pipelines with small diameters, on average 110[mm]. This is due to the fact that the sector is mostly composed of small pipelines, on average 122[mm]. An average length of 94[m] is obtained for the length of the pipelines. However, it is not clear what happens with the type of material and the number of failures nearby, so the commands table(ranking_material) and table(cantidad_fallo) are executed, where ranking_material is a categorical variable with 5 possible values corresponding to the material, with the following detail: asbestos cement (1), cast iron (2), others (3), PVC (4), HDPE (5); and cantidad_fallo is the number of failures that have occurred in the pipelines section. With this analysis we obtain the frequency of the categorical variables.

Using the scatterplotMatrix() command, we obtain the correlations between the variables of interest, in addition to providing the numerical correlation between all pairs of covariates. From this analysis very weak relationships are observed, of which the highest direct relationship between diameter versus length of the pipeline stands out. This analysis provides different correlations for both pressure sectors, which tells us that the two sectors behave differently.

5.4.4 Prediction model

The PME team develops a script in R language, and uses mathematical models related to statistics: Weibull and Exponential distributions, survival function using the non-parametric Kaplan-Meier estimator, and Cox's regression model.

In the software R and through an iterative process, we obtain the parameters of the Weibull and exponential distributions, which were the most adjusted distributions to the times between failures in both sectors. Particularly for sector 270001, it happens that both the Weibull distribution and the Exponential distribution are properly adjusted, however, given the academic literature on the failures in the drinking water network, it is that the PME team decides to keep a Weibull adjustment for that sector. Once the distribution adjustment has been made, it can be concluded that the probability of a failure occurring within a 10-day time window is approximately 80% with 95% of confidence. This frequency is because there is a large concentration of time between failures in the interval [2,7] days. The distribution adjustment of the sectors can also be supported graphically.

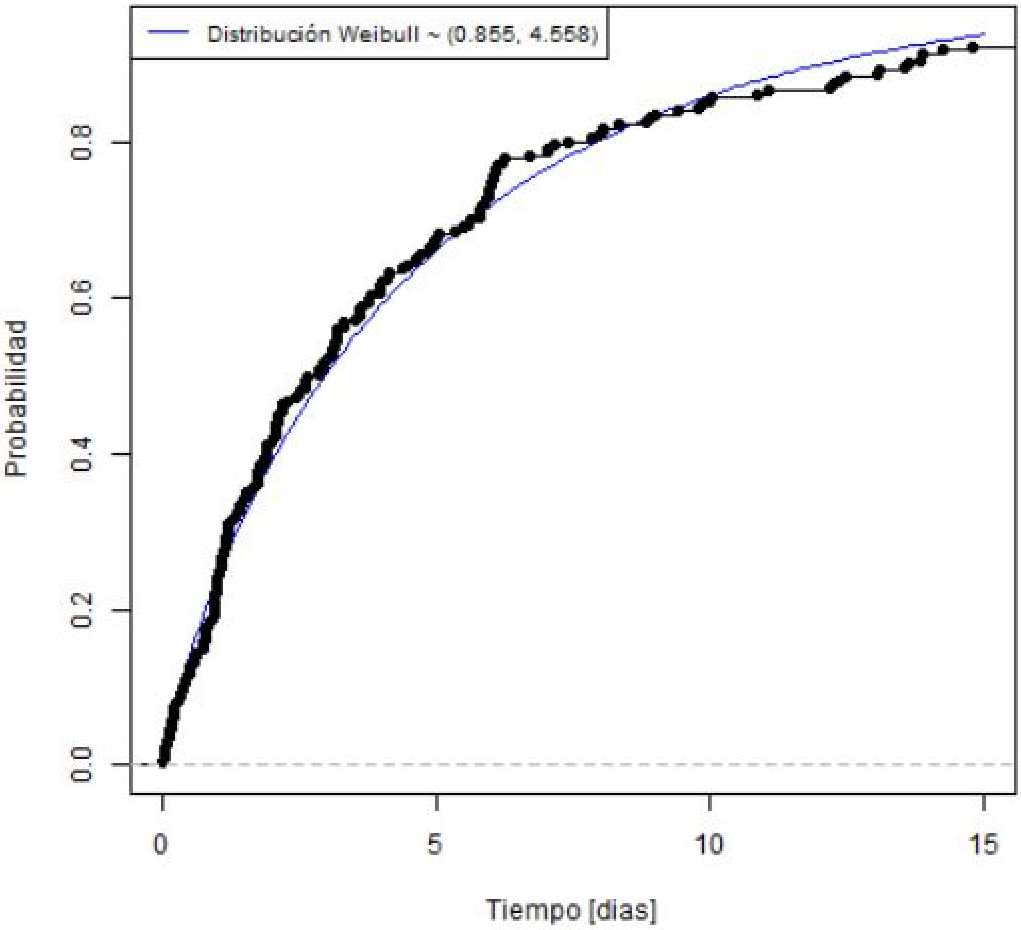


Figure 5.12. Distribution adjustment sector 190001 - Cumulative probability function with adjusted Weibull distribution

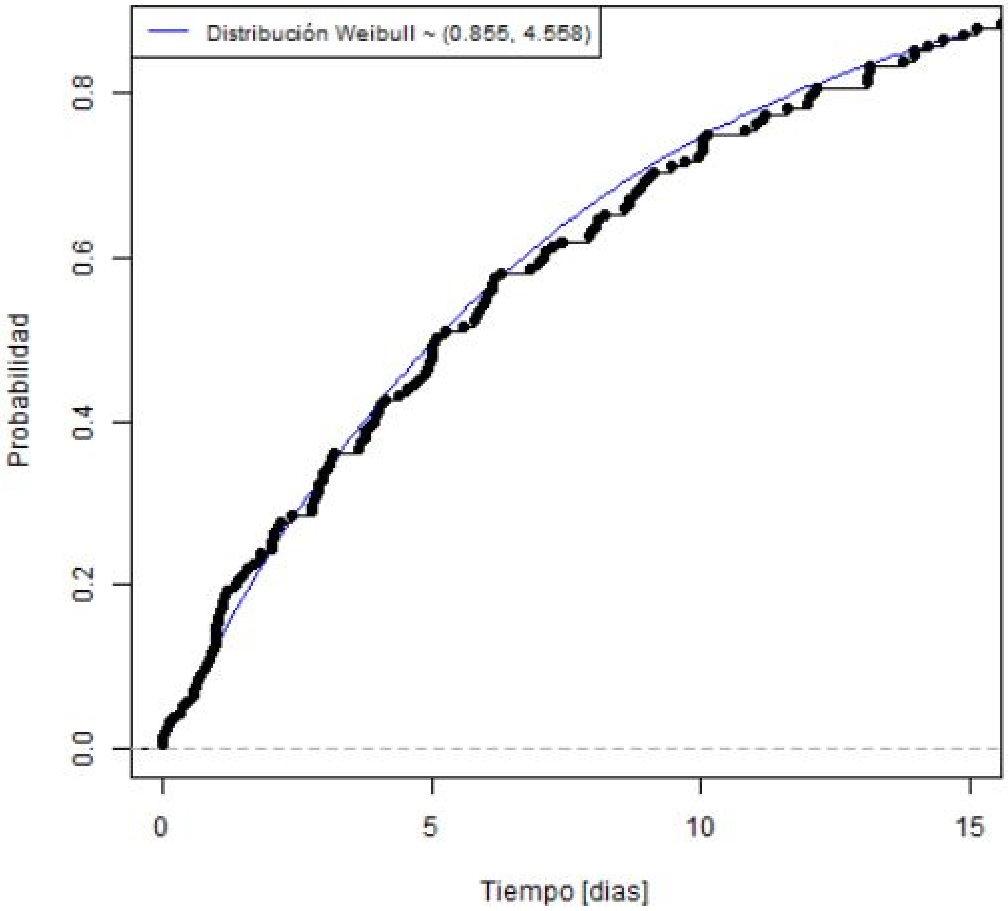


Figure 5.13. Distribution adjustment sector 270001 - Cumulative probability function with adjusted Weibull distribution

- **Time-space model**

  The time-space model is a model developed by the PME team that takes the form of a script that allows to quantify the closeness, in time and space, between failures with the aim of obtaining their concentration. The applicability of this model is in the context of the company's sixthy-monthly renewal decision making. The way in which they make these decisions is based on the criteria established in a document called Renewals Prioritization. In this six-monthly renewal planning process a pressure sector is examined and if it has 4 or more six-monthly ruptures, the concentration of the historical failures must be identified. With this information the renewal of the sector is requested. The task to identify the concentration of failures is done by a person within the network management department and is done through an exhaustive failure-for-failure review procedure, it is at this point that the time-space model plays a supporting role by providing rapid information to the decision making of the renewal. The output of the time-space model delivers the amount and location of nearby failures, in time and space, which allows a quick and efficient verification of the criteria imposed by the prioritization of renewals.

- **Survival Model**

  The survival model is characterized by non-negative random variables, so that the random variable T corresponds to the failure time of an item, in this particular case, relates to a section of the pipelines in a given sector. The survival model requires contemplating pipelines sections that have suffered failures and those not, called uncensored and censored data respectively. With the definition of covariates, establishing a time frame between the first and last failure of the event master, which corresponds to a censure by the right, and the failure times the estimated survival function is obtained (S(t)), this is a steps function that decreases immediately after each observed survival time. The non-parametric Kaplan-Meier estimator is used to generate this function.

  Figure 5.13 shows the graphical results of the survival analysis for the sector 190001 and its associated risk function. In Figure 5.14 we can see that the survival for the sections of sector 190001 is close to 0.93. With the survival function we analyze each of the covariates of interest for the sector. For example, in Figure 5.14 we can see the behavior between those individuals who have not suffered failures in the period of observation, and those individuals who have suffered 1 or more failures. This result reflects the great influence that have historical ruptures in the survival of the pipelines sections. Presenting a number of failures bigger than or equal to 1, pipelines sections' survival decreases.

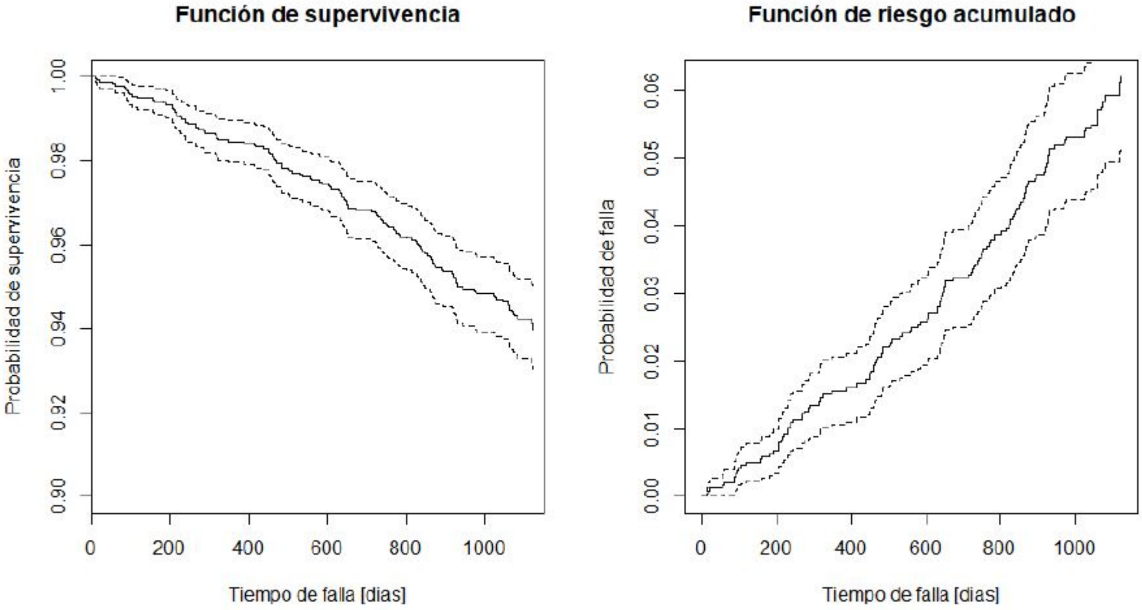


Figure 5.14. Survival function - Accumulated risk function

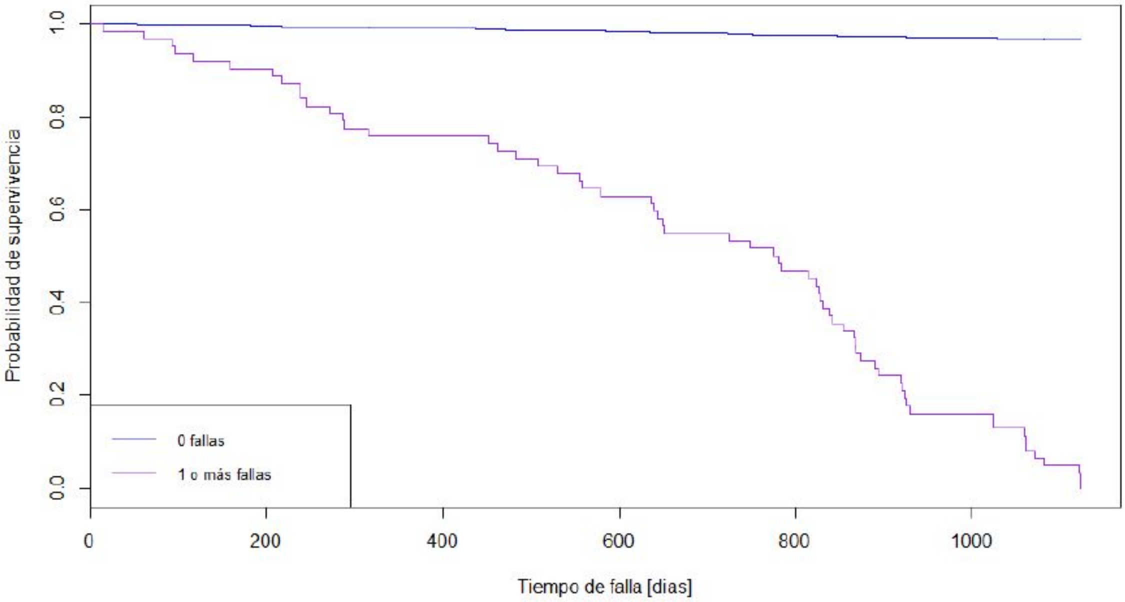


Figure 5.15. Survival function for failures number

This same analysis is performed for X_{largo} and X_{material} and it is concluded that the longer the section and the older the pipeline, the less will be its survival probability.

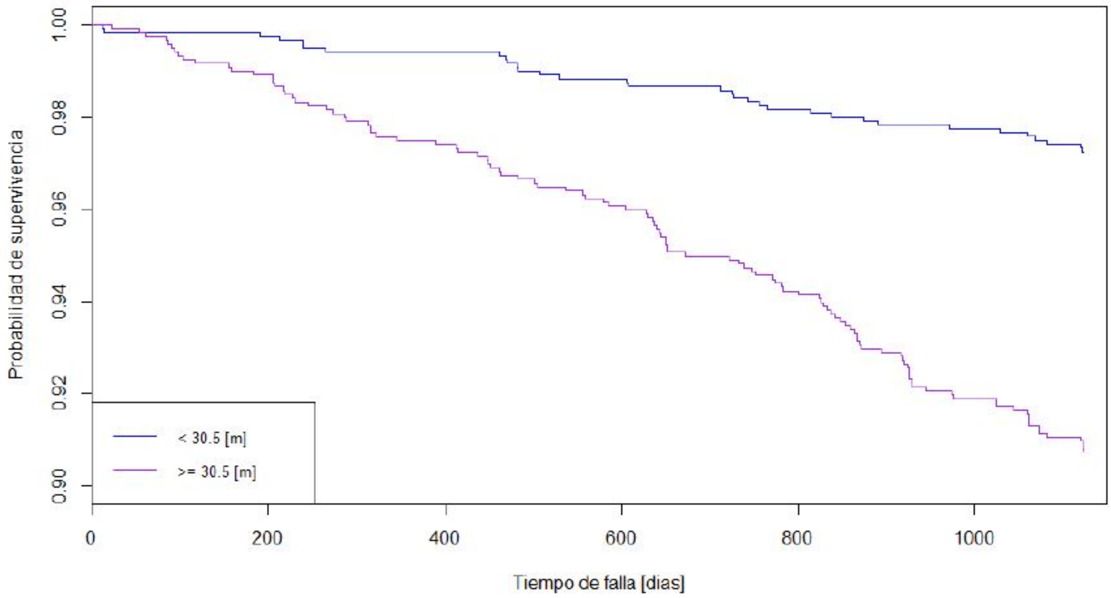


Figure 5.16. Survival function for section length - Sector 190001

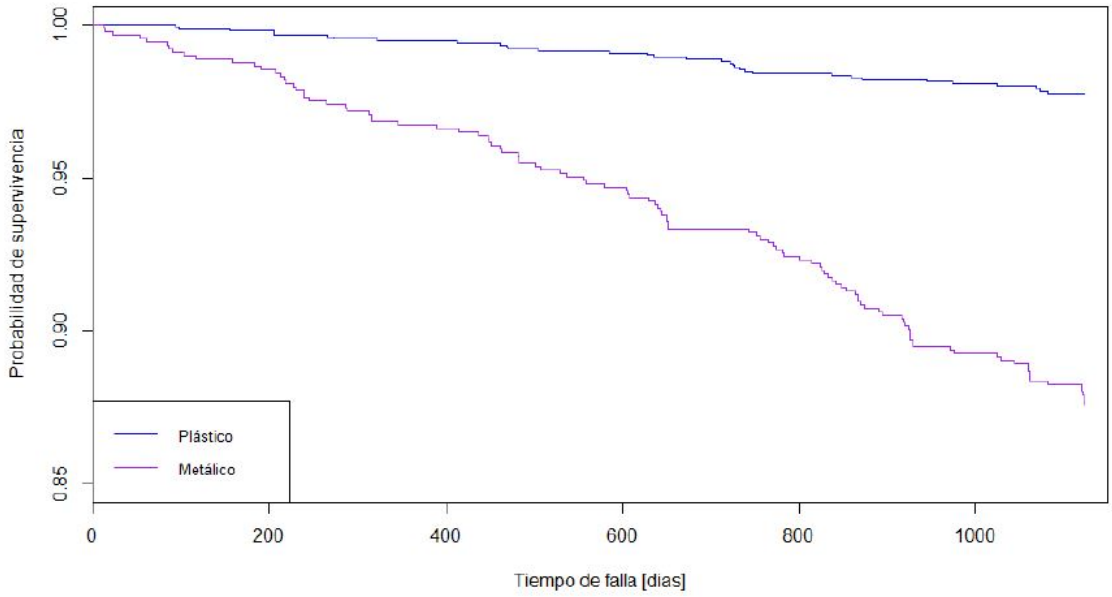


Figure 5.17. Survival function for material - Sector 190001

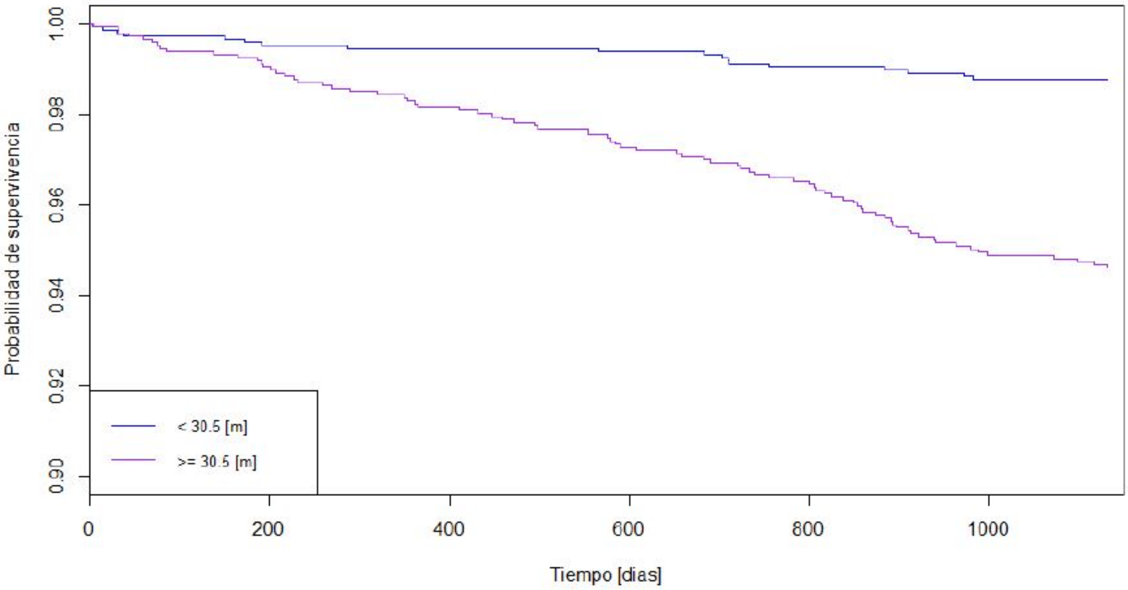


Figure 5.18. Survival function for section length - Sector 270001

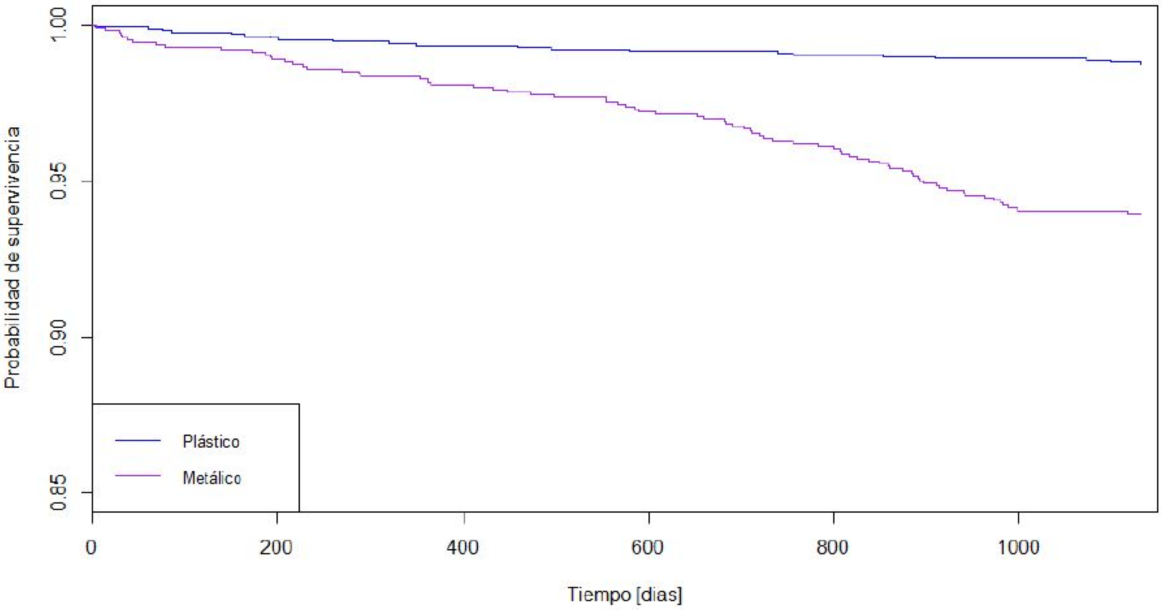


Figure 5.19. Survival function for material - Sector 270001

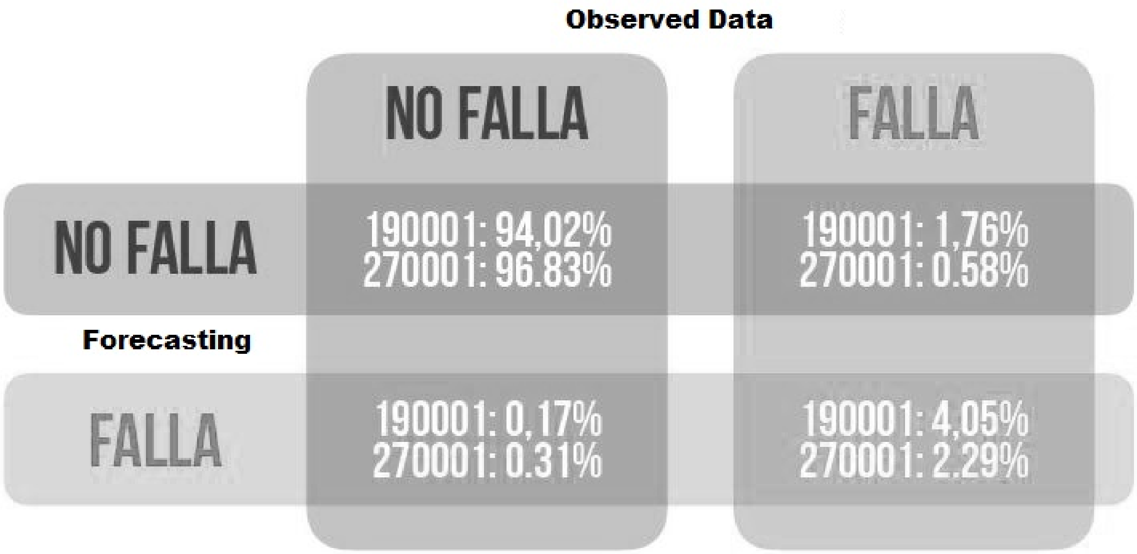
- **Cox's Regression Model**

  The Cox's regression model also known as the Proportional risk model

  $h_i(t,X_1,...,X_n)=h_0(t)*\exp(sum(i,1,n,\beta_i*x_i))$

  generalizes the situation where the risk of failure in a particular moment depends on the values $x_1,x_2, ...,x_n$ of n explanatory variables $X_1,X_2,...,X_n$. Assuming that the values of these variables have been recorded at the origin moment of the study. The estimation or adjustment of the proportional risks model given the expression prior to an observed set of survival data, involves the estimation of unknown coefficients of the explanatory variables $X_1,X_2,...,X_n$ in the linear component of the model, $\beta_1,\beta_2,...,\beta_n$. In certain cases it may also be necessary to estimate the initial risk function, $h_0(t)$. But these two components of the model can be estimated separately. The coefficients are estimated first and these estimators are used to construct an estimator of the initial risk function. This means that to make inferences about the effects of n explanatory variables on the relative risk, $h_i(t)/h_0(t)$, we do not need the $h_0(t)$ estimator. In the software R the regression coefficients estimators of the Cox's proportional risks model are obtained by the function coxph(). This function in its simplest form, requires information about the failure time, the type of censorship of the individuals being studied and information on the covariates of each individual. The process to find the appropriate covariates for the model of each individual sector is iterative, it begins by analyzing each covariable of the proposals a priori separately and then select the significant ones according to their p-value. Different results are obtained for each pressure sector, and therefore, different initial risk functions.

  The results of the survival model provide a first approach to the results expected by Cox's regression model. In order to obtain the regression coefficients of the Cox's proportional risks model, we use the command coxph. For the particular case of the sector 190001 we obtain that the coefficients $b_{material}$, b_{fallo}, and b_{largo} are respectively 0.8954, 0.8921 and 0.0068. It should be noted that the exponential of these coefficients is the impact that each explanatory variable contributes to the survival curve. Finally, the command predict() is executed, obtaining the predictors for each section, and the relative risk value is predicted for each of them. These risks are transformed into a probability of failure for each pipelines section and it is considered that a section will fail when the probability is 80% or higher. After Cox's regression model is verified by seeing how many of the failures present in the event master, the model is able to predict, how many are false positives and how many are false negatives. In the particular case of the sector 190001, 98% of failures are correct and the remaining 2% corresponds to the sum of false positives and negatives.



Observed Data

Forecasting	NO FALLA	FALLA
NO FALLA	190001: 94,02% 270001: 96.83%	190001: 1,76% 270001: 0.58%
FALLA	190001: 0,17% 270001: 0.31%	190001: 4,05% 270001: 2.29%

Figure 5.20. Error Type I - Error type II

For the failures contained in the event master II, which means those failures that were not used for the model creation, the model was capable of predict 3 failures out of a total of 5. It is important to note that there are failures that the model was not able to predict because these sections were not present in the sections that were considered in the Cox's regression model.

The main deliverable of the project is defined as a failures predictive model for the drinking water network. This model is configured as a consequence of the results obtained, allowing the prediction of failures in the pipelines sections for a specific sector.

$h_{\text{pipeline}}(t, X_{\text{fallai}}, X_{\text{largoi}}, X_{\text{materiali}}) = h_0(t) * \exp(\beta_1 * X_{\text{fallai}} + \beta_2 * X_{\text{largoi}} + \beta_3 * X_{\text{materiali}})$

Sean:

X_{fallai} : number of faults in section i

X_{largoi} : length of section i

$X_{\text{materiali}}$: 1 if material = old, 0 and o.f.

In this case, the Cox's proportional risks model for the pressure sector 190001, is defined in the following expression:

$h_{190001i}(t, X_{\text{fallai}}, X_{\text{largoi}}, X_{\text{materiali}}) = h_0(t) * \exp(0{,}8921 * X_{\text{fallai}} + 0{,}0068 * X_{\text{largoi}} + 0{,}8954 * X_{\text{materiali}})$

5.4.5 Results

- The management of the columns was reflected by grouping the 142 columns of the Event Master in the following categories: event classification, date, drinking water network, identification and location, each of them being a set of variables of the same semantics.

- The fields of interest for the Cox's model creation correspond mainly to those covariates that are believed to be influential in drinking water network matrix failures. These covariates that are believed to be influential arise a priori from the expert opinion of the client, who has worked for years analyzing the network, and from the academic literature that exists on drinking water network worldwide. These fields correspond to the diameter, material, age and length of the pipelines sections. For the calculation of the Kaplan-Meier non-parametric estimator, the exact date of failure and the number of times it has failed, are also considered. Finally, identification columns, such as the section and sector id, are considered.

- The drinking water network is considered a complex system because it presents several elements or equipment, including: shut-off valves, pressure reducing valves, booster or bombs, ponds, others. Of these elements the most damaged is the pipeline. One of the reasons why the equipment focuses on the pipeline element is due to its high rate of intervention and occurrence of failures.

- The antiquity of drinking water networks in the world and technological progress are two reasons why the drinking water network is composed of more than 15 types of different materials. In this drinking water network we can observe the presence of 19 different materials, among which we can classify them in two families: new and old pipelines. The first family is made up of: PVC and HDPE, which correspond to plastic pipelines. The second consists mostly of: steel, cast iron and asbestos cement, which are called metal pipelines.

- There is a large differential between the number of programmed events versus the number of not programmed events. This means a priori that the drinking water network deteriorates faster than it renews itself.

- The output of the time-space model delivers the amount and location of nearby failures, in time and space, which allows a quick and efficient verification of the criteria imposed by the prioritization of renewals. We can note that the majority of the failures don't have a failure near in time and space according to a certain time and space interval. Furthermore, if a an event has some events near in time and space, it is about a few quantity of events.

- The survival model requires contemplating pipelines sections that have suffered failures and those not, called uncensored and censored data respectively. With the definition of covariates, establishing a time frame between the first and last failure of the event master, which corresponds to a censure by the right, and the failure times the estimated survival function is obtained (S(t)), this is a steps function that decreases immediately after each observed survival time. The non-parametric Kaplan-Meier estimator is used to generate this function. The historical ruptures have a great influence in the survival of the pipelines sections. Presenting a number of failures bigger than or equal to 1, pipelines sections' survival decreases. This same analysis is performed for X_{material} and the other covariates and it is concluded that the older the pipeline, the less will be its survival probability and that the other covariates are not proportional to the risk of failure because of showing a cross in the graphs.

- For each section is predicted the relative risk value. These risks are transformed into a probability of failure for each pipelines section and it is considered that a section will fail when the probability is 80% or higher.

- After Cox's regression model is verified by seeing how many of the failures present in the event master, the model is able to predict, how many are false positives and how many are false negatives.

5.4.6 Assumptions

For the creation of the Cox's regression model, some assumptions have been made.

- **Geospatial limitations**: this study is limited to a region and specifically to two pressure sectors. The analysis of these sectors implies high computational efforts to transform, and load the data related to pipelines sections for each of the pressure sectors, so the integration of the data must be done from different archives with different update dates. It should be noted that this work can be extrapolated to various pressure sectors.

- **Failures filtered by reason**: the events that are studied are the failures that are associated with a natural cause or wear due to the fatigue of material. The reason for this decision is because of issues of scope, since the failures that occur for other reasons relate to the reliability of specific equipment and events no predictable.

- **Variables choice**: according to the benchmarking study and meetings with the head of the network management department, is validated that the main variables are: diameter, length of section, number of failures, static pressure and material. From the study it is excluded the static pressure, as there is not enough detail to carry out the assignment of this variable to a pipelines section. In any case, the incorporation of this variable is possible in the future through an effective ETL.

- **Data adaptation to the Cox's regression model**: the homologation of *.xlsx files to *.csv, fields modification and new fields incorporation, allows to work with pipelines sections (subject of interest), failure time and censorship.
- **Pipelines characteristics**: the pipelines section is considered to be a subject that has the following properties: it is straight and flat and it is associated to the same type of soil and humidity.
- **Assumptions about materials**: in the used files, the empty cells correspondents to the failures were considered asbestos cement, since the opinion of the company's experts indicated that at the time of the creation of the SIGEC, it was carried out a massive digitization of the network's own characteristics, a process that generated the existence of pipelines sections with no defined material, therefore the experts assume that is asbestos cement because of the age of empty cell records.
- **Types of material**: two groups of materials are individuated: plastics or new materials and metallic or old materials. These categories allow us to evaluate a categorical variable that has multiple numerical values related to each type of material, turning it into a binary variable. This decision simplifies Cox's regression model decreasing the amount of covariates.

5.5 Conclusions

In order to achieve a potential cost decrease and less impact due to underground work in public areas, it is necessary a certain grade of communication between the company and all those organizations, public or private, linked to underground civil works in areas where there is drinking water supply and also an adequate integration of the underground activities executed by these organizations.

For the execution of a project of this complexity and size, the incorporation, collaboration or participation of professionals from different areas is indispensable. In particular, it concerns with the following areas: reliability, mechanical engineering, hydraulic engineering, mathematics, statistics and programming.

The execution of the project generates a set of future opportunities, due to the complexity of the network, the number of situations to be resolved and the diversity of the aspects to be considered on the drinking water network, both technically and professionally. For example, it could be possible to apply the Cox's regression model in other drinking water networks, to reform the data structure in order to answer to other kind of questions the company could be interested in, or consolidate the current database to make more detailed studies on the drinking water network.

5.5.1 Technical conclusions

- **Time-space model**: this model allows analyzing the failures concentration. It is possible to achieve this through the inspection, failure-to-failure, and through the incorporation of restrictions of time - temporal window - and space - radial distance - calculating with it the number of failures that are close to each other failure in terms of time and space defined by the user. Time and space constraints are quantified through a series of parameters that company's personnel can calibrate by their discretion. Finally, the output of the model provides two files: the first one is a detail of the close failures in time and space for each failure. The second file shows the frequency for the number of nearby failures in time and space. This model shows more accurate results when applied individually to the cutting sector.

- **Cox's regression model results**: when using the Cox's regression model with a predictive purposes, the aim is to quantify the relative risk of the pipelines sections of the drinking water network. This indicator is obtained through the Cox's regression model for the pipelines sections in a defined pressure sector, with which also it is obtained the failures probability. In this case, it is considered high probability all the probability equal to or greater than 80%. During its management, the company will apply its criteria and their experience in determining the probability from which they will consider high probabilities of failure.

- **Distribution Adjustment**: the distribution adjustment made to the time between failures in a defined sector gives to the company's staff a better understanding about the importance of the failure probability as time passes. This analysis serves as an indicator within the process of the sector analysis, in order to reduce the emergency purchases.

- **Integration of statistical results in the renewal process**: the renewal process begins with the compilation of the pressure sectors with events or specific requests; the sector is then analyzed in terms of types of failures, concentration and affected clients; and if it is not possible to manage the pressure, is necessary a zonal consultation, sections are proposed and validated for renewal, and the renovation plan is finally carried out. The moment in which the sector analysis is carried out, is where get value the results obtained during the project. Part of the sector analysis is carried out through the time-space model, the Cox's regression model and the distribution adjustment.

Short and medium term considerations

- **Soil type research**: according to the literature and expert's opinion from the company, the amplifying effect of seismic waves increases the susceptibility to failure and this effect is different for all sectors, since it depends on the type of soil in which the tested drinking water network is located. In the context of a seismic country like Chile, this variable gets a greater importance

- **Inclusion of static pressure**: one of the reasons why static pressure is not included in this analysis is because of its information is in different files for which there is no certainty, definition or access

- **Real-time control system**: data management through the visualization of them in real time would allow the company to have a power of anticipation against the drinking water network failures

The final scope of the project is the creation of a predictive model to predict drinking water failures in order to better understand the complexity of the drinking water network management and to support the decision making in the context of renewals and repairs. The Cox's regression model allows to prioritize, for each sector, the pipelines sections which need intervention. In this way the drinking water service interruption is reduced, because it will be easier and preventive to know where and when there is an high rupture probability for a defined section and the nearby ones. At the same time are reduced also the potential economic sanctions and the associated operational costs. The company's have to incorporate the use, of their decisions.

Chapter 6

Conclusions

The thesis goal is to investigate about the paradigms and the literature concerning the value that the companies can achieve by taking advantage from the enormous quantity of internal and external data available, which analysis can create useful insights to support business decision making. With this researching I had the opportunity to get closer to that concept which are the base for a Data Science project development and that a good Data Scientist must know. I learned how a company can successfully implement Data Science project and create value from the available data. First of all, if a company wants to work with and get profit from the data, is fundamental to be prepared in terms of personal of the organization,technologically and, most of all, in the company vision. Key aspects for a success are the clear vision of the company and the good communication within the different levels of the organization and within the components of the teams who work to the project. Then, but not less important, the organization has to know with what kind of data they are working, the characteristics and the contents, to be able to understand which data are useful for the company according to its scope. In particular, I had the opportunity to find all these theoretical concepts applied in a real case study, such as the PME Project. By following the project development, I saw how important was for the team project to have a clear goal. To create the predictive model, first of all was necessary to be well informed about the characteristics of the data available and furthermore about the sector in which the company operates. The clear subdivision of the roles and the good communication of the results within the team components also was fundamental to respect the project time table and goals.

With the PME Project I saw how a Data Science project could help a company to create value from the available data. In this case, the available data was useful to create a prediction model with the aim of being supporting the decision making about the renewals plan and to reduce the costs due to the renewals and the fines for the interruption of the service when failures happen.

The objective of the project to create a predictive model to predict in time and space the failures of the drinking water network has been achieved from the results obtained with the time-space model and with the Cox's regression model that facilitate the identification of the sections of the pipeline that will suffer failures and thus supporting the decision making in the creation of plans for the renewal of the network every six months. Furthermore, all the analysis carried out thanks to the available data, favours a better understanding and knowledge of the functioning and management of the drinking water network in order to reduce the huge costs related to fines and emergency purchases that the company is currently forced to bear because the drinking water network deteriorates faster than it renews itself.

The output of the time-space model delivers the amount and location of nearby failures, in time and space, which allows a quick and efficient verification of the criteria imposed by the prioritization of renewals. Instead, the survival model requires contemplating pipelines sections that have suffered failures and those not, to see which covariates impact on the risk of failure and which ones not. For each section is predicted the relative risk value. These risks are transformed into a probability of failure for each pipelines section and it is considered that a section will fail when the probability is 80% or higher. After Cox's regression model is verified by seeing how many of the failures present in the event master, the model is able to predict, how many are false positives and how many are false negatives. In this particular case, 98% of failures are correct and the remaining 2% corresponds to the sum of false positives and negatives. So we can conclude that the model is working correctly.

During its management, the company will apply its criteria and experience to determine the probability from which it will be considered high probabilities of failure. It is at the moment in which the sector analysis is performed that the results obtained in this project make sense; part of the sector analysis is carried out through the time-space model and the Cox's regression model. The management of data through the visualization of them in real time would allow the company to have a power of anticipation to the failures of the drinking water network that, for the moment, it has not achieved.

Thanks to the professor Leiva, I also had the opportunity to see how Data Science could find application in a wide range of domains and with different goals. It could be to create predictive models to support the decision making or, by a more statistical and descriptive approach, to better know the competitive environment of the company to be able to exploit the opportunity and to transform their weakness in their strengths.

Bibliography

[1] Chen H, Chiang R, Storey V. (2012): “Business Intelligence and Analytics: From Big Data to Big Impact", *MIS Quarterly*, 36(4):1165-88.

[2] Kowalczyk M, Buxmann P, Besier J, editors (2013): “Investigating business intelligence and analytics from a decision process perspective: a structured literature review", *21st European Conference on Information Systems*, Utrecht, Netherlands.

[3] Manyika J, Chui M, Brown B et al (2011): “Big data: the next frontier for innovation, competition, and productivity", pp 1-156

[4] Davenport (2006): “Competing on Analytics", *Harvard Business Review*, 84:1-12.

[5] Power D (2007): “A brief history of decision support systems".

[6] Brown B.,Chul M.,Manyika J.(2011): “Are you ready for the era of 'Big Data'?", *McKinsey.*

[7] Davenport T.H.,Barth P.,Bean R.(2012): “How Big Data is different", *MIT Sloan Manag.*, Rev.54(1),43-46.

[8] Gehrke J. (2012): “Quo vadis, data privacy?", *Ann.NewYorkAcad.Sci.*, 1260(1),45-54.

[9] Laney D. (2001): “ 3D data management: controlling data volume, velocity, and variety", *META Gr.*

[10] IBM (2012): “Analytics: the real-world use of big data" .

[11] Golden B. (2013): “Does big data spell the end of business intelligence as we know it?", *CIO.*

[12] LaValle S, Lesser E, Shockley R et al (2011): “ Big data, analytics and the path from insights to value big data, analytics and the path from insights to value", *MIT Sloan Manag.*, Rev 52:21-31.

[13] Stackowiak R., Rayman J. and Greenwald R. (2007): “Oracle Data Warehousing and Business Intelligence Solutions", *Wiley Publishing, Inc, Indianapolis.*

[14] Cui Z., Damiani E. and Leida M. (2007): “Benefits of Ontologies in Real Time Data Access", *Digital Ecosystems and Technologies Conference, DEST '07*, pp. 392-397.

[15] Zeng L., Xu L., Shi Z., Wang M. and Wu W. (2006): “Techniques, process,

and enterprise solutions of business intelligence", *2006 IEEE Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan, Vol. 6,*, pp. 4722.

[16] Golfarelli M., Rizzi S., Cella L. (2004): "Beyond Data Warehousing: What's next in Business Intelligence?", *Proceedings of DOLAP-04, Washington, DC, USA. Retrieved May 17 2006 from www.acm.org.*

[17] Chamoni P., Gluchowski, P. (2004): "Integrationstrends bei business-intelligence-systemen-empirische untersuchung auf basis des business intelligence maturity model", *Wirtschaftsinformatik, Vol. 46 No. 2*, pp. 119-128.

[18] Kemper H.G., Mehanna W., Unger C. (2004): "Business Intelligence", *Grundlagen und praktische Anwendungen, Vieweg, Wiesbaden.*

[19] Negash S. (2004): "Business intelligence", *Communications of the Association for Information Systems, 13*, pp. 177-195.

[20] Elbashir M.Z., Collier P.A., Davern M.J. (2008): "Measuring the effects of business intelligence systems: The relationship between business process and organizational performance", *International Journal of Accounting Information Systems 9 (2008)*, pp. 135-153.

[21] Orbis Research (2017): "Orbis Research", *https://www.reuters.com/brandfeatures/venture-capital/article?id=4403*.

[22] Morris H. (2003): "The Financial Impact of Business Analytics: Build vs. Buy", *DM Review, (13)1*, pp. 40-41.

[23] Delone W., McLean E.(2003): "Information Systems Success: The Quest for the Dependent Variable", *Journal of Information System Research, 3(1)*, pp. 60-95.

[24] Yeoh W., Koronios A. (2010): "Critical Success Factors for Business Intelligence Systems, *Journal of Computer Information Systems, 50:3*, pp. 23-32.

[25] Sharma R., Mithas S., Kankanhalli A. (2014): "Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organizations", *European Journal of Information Systems (2014) 23*, pp. 433-441.

[26] Lycett M. (2013): "'Datafication': making sense of (Big) data in a complex world", *European Journal of Information Systems 22(4)*, pp. 381-386.

[27] Tallon P.P. (2013): "The information artifact in IT governance: toward a theory of information governance", *Journal of Management Information Systems 30(3)*, pp. 141-177.

[28] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf .

[29] Transparency Market Report. (May, 2015).Big Data Applications in Healthcare likely to Propel Market to US$48.3 Bn by 2018. Retrieved June 26, 2015, from http://www.transparencymarketresearch.com/pressrelease/big-data-market.htm .

[30] IBM, 2012b.WhatisBigData? (http://www-01.ibm.com/software/data/bigdata/) (retrieved25.02.13) .

[31] Johnson J.E. (2012): "BIG DATA + BIG ANALYTICS = BIG OPPORTUNITY", *Financ.Exec. 28 (6)*, pp. 50-53.

[32] Rouse M. (2011): "Big Data", .

[33] Fisher D., DeLine R., Czerwinski M., Drucker S. (2012): "Interactions with big data analytics", *Interactions19(3), 50*.

[34] Havens T.C., Bezdek J.C., Leckie C., Hall L.O., Palaniswami M. (2012): "Algorithms for very large data", *IEEE Trans. 20 (6)*, 1130-1146.

[35] Jacobs A. (2012): "The patologies of big data", *Assoc.Comput.Mach.Commun.ACM*, 52 (8), 36.

[36] IDC (2013): "Big Data in 2020", *IDCiView(Ed.):IDC*.

[37] Boyd D., Crawford K. (2012): "Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon", *Inf.Commun.Soc.15(5)*, 662-679.

[38] Gartner (2012): "Big Data", *http://www.gartner.com/it-glossary/big-data/*.

[39] Kwon O., Sim J.M., (2013): "Effects of data set features on the performances of classification algorithms", *ExpertSyst.Appl.40(5)*, 1847-1857.

[40] McAfee A., Brynjolfsson E., (2012): "Big data: the management revolution", *Harv.Bus. Rev.October*, 61-68.

[41] Russom P., (2011): "The Three Vs of Big Data Analytics", *TDWI*.

[42] IDC, (2012): "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East", .

[43] Oracle, (2012): "Big Data for the Enterprise", *RedwoodShores,CA:Oracle*.

[44] Forrester, (2012): "The Big Deal About Big Data For Customer Engagement Business: Leaders Must Lead Big Data Initiatives To Derive Value", .

[45] Saravanakumar, Nandini, (2017): "A Survey on the Concepts and Challenges of Big Data: Beyond the Hype", *ISSN 0973-6107 Volume 10, Number 5 (2017)*, 875-884.

[46] Gandomi A., Haider M., (2015): "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management 35*, 137-144.

[47] Vashisht P., Gupta V., (2015): "Big Data Analytics Techniques: A Survey", .

[48] Kaisler S., Armour F., Espinosa J.A., Money W., (2012): "Big Data: Issues and Challenges Moving Forward", *IEEE 2012*.

[49] Geczy P., (2014): "Big Data characteristics", *The Macrotheme Review 3(6)*.

[50] Tole A.A., (2013): "Big Data challenges", *Database Systems Journal vol. IV, no. 3/2013 31*.

[51] Cukier K., (2010): "A special report on managing information", *The Economist, February 25*.

[52] Tanwar M., Duggal R., Kumar Khatri S., (2015): “Unravelling Unstructured Data: A Wealth of Information in Big Data", *IEEE 2015*.

[53] Andriole S., (2015): “Unstructured Data: The Other Side of Analytics", *Forbes*.

[54] Moss L.T., (2003): “Nontechnical Infrastructure of BI Applications", *DM Review, (13)1*, 42-45.

[55] Oracle “Oracle Big Data strategy guide", *http://www.oracle.com/us/technologies/big-data/big-data-strategy-guide-1536569.pdf*.

[56] Cloudera: “Cloudera’s 100% Open Source Distribution of Hadoop", *http://www.cloudera.com/content/cloudera/en/products/cdh.html*.

[57] SAS: “SAS Enterprise Guide", *http://support.sas.com/software/products/enterprise-guide/index.html*.

[58] Rajaraman V., (2016): “Big Data Analytics" .

[59] Zakir J., Seymour T., Berg K., (2015): “Big Data Analytics", *Issues in Information Systems 2015*, 81-90.

[60] Labrinidis A., Jagadish H.V. (2012): “Challenges and opportunities with big data", *VLDB Endowment, vol. 5(12)*, 2032-2033.

[61] Blackett G., (2013): “Analytics Network", *O. R. Analytics*.

[62] Jiang J., (2012): “Information extraction from text", *C. C. Aggarwal,& C. Zhai (Eds.), Mining text data, Springer*, 11-41.

[63] Hahn U., Mani I., (2000): “The challenges of automatic summarization", *Computer, vol. 33(11)*, 29-36.

[64] Liu B., (2012): “Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies, vol. 5(1)*, 1-167.

[65] Hirschberg J., Hjalmarsson A., Elhadad N., (2010): “You’re as sick as you sound: Using computational approaches for modeling speaker state to gauge illness and recovery", *Neustein (Ed.), Advances in speech recognition, Springer*, 305-322.

[66] Panigrahi B.K., Abraham A., Das S., (2010): “The challenges of automatic summarization", *Studies in Computational Intelligence, Springer, vol. 302*.

[67] Hakeem A., Gupta H., Kanaujia A., Choe T.E., GundaK., Scalon A., (2012): “SVideo analytics for business intelligence", *Springer*, 309-354.

[68] Hu W., Xie N., Li L., Zeng X., Maybank S., (2011): “A survey on visual content-based video indexing and retrieval", *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE*, 797-819.

[69] Barbier G., Liu H., (2011): “Data mining in social media", *C. C. Aggarwal (Ed.), Social network data analytics, Springer*, 327-352.

[70] He W., Zha S., Li L., (2013): “Social media competitive analysis and text mining: A case study in the pizza industry", *International Journal of Information Management, vol. 33(3)*, 464-472.

[71] Charu C., Aggrawal, (2011): “An Introduction to Social Network Data Analysis", *Springer*.

[72] Heidemann J., Klier M., Probst F., (2012): “Online social networks: A survey of a global phenomenon", *Computer Networks, vol. 56(18)*, 3866-3878.

[73] Parthasarathy S., Ruan Y., Satuluri V., (2011): “Community discovery in social networks: Applications, methods and emerging trends", *Springer*, 79-113.

[74] Lo F., (2018): “What is Data Science? What is analytics? What is a data scientist?", *https://datajobs.com/what-is-data-science*.

[75] Sharma H., (2018): “What Is Data Science? A Beginner’s Guide To Data Science", *https://www.edureka.co/blog/what-is-data-science/*.

[76] Schott J., (2018): “How to Hire Data Scientists Based on Your Company Readiness", *https://www.datascience.com/blog/how-to-hire-data-scientists-based-on-company-readiness*.

[77] Krasadakis G., (2018): “Establishing a culture of Analytics. How to transform your company using data and analytics", *https://medium.com/innovation-machine/establishing-a-culture-of-analytics-947c7947af*.

[78] Nevo D., (2016): “Thinking About Analytics Readiness", *https://www.kdnuggets.com/2016/06/thinking-domain-readiness.html*.

[79] KD Nuggets, (2018): “Must have skills data scientist", *https://www.kdnuggets.com/2018/05/simplilearn-9-must-have-skills-data-scientist.html*.

[80] Rouset M., (2017): “What is a data scientist?", *https://searchenterpriseai.techtarget.com/definition/data-scientist*.

[81] AltexSoft, (2018): “How to Structure a Data Science Team: Key Models and Roles to Consider", *https://www.altexsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/*.

[82] Agile Alliance, (2017): “Agile Practice Guide, PMI", *Global Standard*.

[83] Leiva V., (2017): “Business Intelligence and Big Data: Background and case studies", *Pontificia Universidad Catolica de Valparaiso*.

[84] Harmon P., (2014): “Business process change: a Business Process Management guide for managers and process professionals", *Morgan Kaufmann, 3rd edition*.

[85] MBA Group, (2017):, *https://mbagroup1blog.wordpress.com/2017/07/13/question-5-advantages-and-disadvantages-of-a-manager-being-the-direct-user-of-an-olap-tool-rather-than-providing-an-intermediary-to-operate-the-olap-tool-on-behalf-of-the-manager/*.

[86] Rozenberg T., (2018): “Is Data Warehousing Dead?", *https://medium.com/@tamiro/is-data-warehousing-dead-727757b0c424*.

[87] Nubicus, (2017):, *http://nubicus.com/business-intelligence/*.

[88] SyncSort, (2013):, *https://blog.syncsort.com/2013/02/big-data/hadoop-mapreduce-to-sort-or-not-to-sort/*.

[89] University of Cambridge, (2018): “Text & Data Mining: What is TDM?", *https://libguides.cam.ac.uk/tdm/definitions*.

[90] VPI, (2018): “Text & Data Mining: What is TDM?", *http://www.vpi-corp.com/speech-analytics-softwarecall-centers.asp?no_redirect=true*.

[91] IBM, (2018): “Product architecture", *https://www.ibm.com/support/knowledgecenter/en/SS88XH.6*

[92] Freepng, (2018):, *https://www.freepng.es/png-mh18ha/*.

[93] Shri Shyam Cargo, (2018):, *http://www.shrishyamcargo.in/2018/08/29/3-case-studies-of-predictive-analytics/*.