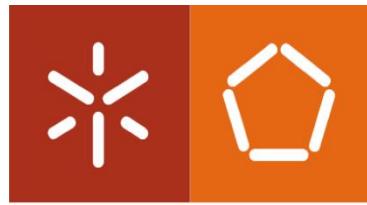


Universidade do Minho
Escola de Engenharia

Luciana Cristina Barbosa de Lima

**Big data for data analysis in financial
industry**

October 2014



Universidade do Minho
Escola de Engenharia

Luciana Cristina Barbosa de Lima

Big data for data analysis in financial industry

Master Thesis

**Integrated Master in Engineering and
Management of Information Systems,
University of Minho**

Supervised by:

Professor Doctor Carlos Filipe Portela

October 2014

Perseverance is the hard work you do after you get tired of doing the hard work you already did.— Newt Gingrich

Acknowledgements

Finally finished! One step further in my path that opens the door to a new world. All the effort and commitment is dedicated to my family, hoping to one day be able to repay all that they have done and still do for me.

Father, mother, Fabio, I dedicate this project to you with the utmost gratitude for all the support, love and patience you have had. You are my pillars, inspiration and above all you are part of me and what I am today. The pride you have shown has made me a fighter, always wanting more and more.

Thank you! I will continue to strive to never disappoint you.

I will take this opportunity to thank all those who supported me through these years and that somehow helped me grow.

To Filipe Portela, thank you for all the support and patience that you have shown. The way you dedicate yourself to your work and supervised students is unique, so I feel very fortunate to have had you as my advisor.

To Cátia, my friend, adventures and study mate. Someone who came into my life and helped me, without ever asking for anything in return. She was always there to support me, sometimes even in a motherly way. Thanks for everything and I hope that our friendship is even stronger as time goes by.

To Laureano, the brother of my heart, I thank him for always guiding me back "home" in the hardest of times, by wisely comforting me without ever asking "what is wrong with you?" and finally by sharing with me moments of joy and fun.

To Patricia, an acquaintance who happened to have a very special place in my heart. I thank her for entering into my life, even unexpectedly. We do not share moments of work, but somehow, we share so many moments that made me a happy and positive person. Thanks for always being available to listen to my complaints about life and for helping me in this project.

To Brian, my "breath of fresh air," I thank him for all the affection, for listening to me and for the attempts to make me forget the difficulties and concerns. The way he sees me, yet

overly proud, made me strive and never give up. He will certainly be a person who I will never forget and I hope to follow his successes as he followed mine.

To Pedro, Antonio, Patrick and all my "kids" thank you for helping me somehow, directly or indirectly, to overcome barriers, academic or even personal challenges.

To all of you I wish the best of luck in the world and thank you for composing my happy memories. We shall meet again, I promise!

Ending these long thanks I will want to thank Deloitte for supporting me in this project.

To the University of Minho and especially to the Department of Information Systems, a house of knowledge where I had the opportunity to learn and design the path of my future career.

Abstract

Technological evolution and the consequent increase of the society and organization dependency was an important driver for the escalation of volume and variety of data. At the same time, market evolution requires the capability to find new paths to improve the products/services, client satisfaction and avoid the cost increase associated with it.

Big Data comes up with huge power, not only by the ability of processing large amounts and variety of data at a high velocity, but also by the capability to create value for the organizations that include it in their operational and decision making processes.

The relevance of Big Data use for the different industries and how to implement a Big Data solution is something that raises many doubts and discussion.

Thus, this paper comes with a business orientation so it will be possible to understand what Big Data actually means for organizations. The project follows a top-down approach which is done in the first instance, an overview on what defines Big Data and what distinguishes it. As it evolves, it directs the focus to the Big Data contribution at an organizational level and the existing market offers. The decomposition of the problem closes with two main contributions. A framework that helps to identify a problem: Big Data and a trial of a case study that identifies the correlation between financial news articles and the change in the stock exchange. The outcome of this trial was a platform with analytic and predictive capabilities in this new Big Data context.

Resumo

A evolução tecnológica e consequente aumento da dependência da sociedade e organizações levou, nos últimos anos, ao crescimento exorbitante do volume e variedade de dados existentes. Ao mesmo tempo, a evolução do mercado exige às organizações a capacidade de encontrarem novas formas de melhorarem os seus produtos/serviços, satisfazer os seus clientes e evitar o aumento de custos para atingir esses objetivos.

O Big Data surge em grande força, apresentando um elevada capacidade de processar a alta velocidade grandes quantidades e variedade dos dados. Este conceito tem evoluído pela sua capacidade de gerar valor às organizações que o incluem nos seus processos operacionais e na tomada de decisão. A pertinência da utilização do Big Data pelas organizações dos mais diversos sectores e a forma como se poderá implementar uma solução Big Data é algo que ainda suscita várias dúvidas e alguma discussão

Desta forma, o presente documento, surge com uma orientação ao negócio de modo a que seja possível entender o que o Big Data representa na verdade para as organizações. O projecto segue uma abordagem top-down onde é feito, numa primeira instância, um síntese sobre o que define o Big Data e o que o distingue. À medida que o projeto evolui, o foco direciona-se para o contributo do Big Data a nível organizacional e quais as ofertas existentes nos mercado. A decomposição do problema culmina com dois principais contributos. Uma framework que ajuda à identificação de um problema, como sendo Big Data e experimentação de um caso de estudo que identifica a correlação entre artigos de notícias financeiras e a variação da bolsa de valores. Como resultado desta experimentação foi desenvolvida uma plataforma com capacidades analíticas e preditivas neste novo contexto do Big Data.

Table of Contents

1.	Introduction.....	1
1.1.	Scope and Motivation	1
1.2.	Goals	4
1.3.	Structure of the report	4
1.4.	Organization Overview (partial).....	5
A1)	Purpose	6
A2)	Environment	6
A3)	Activities	7
A4)	Handled Objects	7
A5)	Organs.....	8
B)	Size of company (Characterization of firm size).....	9
2.	Methodologies.....	10
2.1.	Research Methodologies.....	10
2.1.1.	Artifact	11
2.1.2.	Problem Relevance	11
2.1.3.	Research Rigor	11
2.1.4.	Design as a Search Process and Evaluation, Research Contributions and Communication.....	12
2.2.	Practical Methodologies.....	12
2.2.1.	SCRUM	12
2.2.2.	Big Data analysis pipeline	14
3.	Literature Review.....	17
3.1.	Big Data.....	17
3.1.1.	Definitions and Concepts	18
3.1.2.	Characteristics and Challenges	19

3.2.	Big Data and Hadoop	21
3.2.1.	Apache Hadoop Core System Components	23
3.3.	Big Data Architecture.....	25
3.3.1.	Context.....	25
3.3.2.	Big Data Architectures	26
3.3.3.	Big Data Solutions	30
3.4.	Value Creation.....	36
3.4.1.	Big Data business opportunities	38
3.4.2.	Financial Services Industry.....	40
3.5.	Text Mining	43
3.5.1.	Concept.....	43
3.5.2.	Text Mining Process Framework.....	43
3.6.	News Articles Influence on Stock Market.....	45
3.6.1.	Concept.....	45
3.6.2.	Approach Overview	46
4.	Towards Big Data project.....	49
4.1.	Big Data vs Business Intelligence.....	50
4.2.	Big Data Complexity Framework	55
4.3.	Having the best of two worlds	59
5.	Big Data Solutions – Comparative Summary	61
6.	Development.....	64
6.1.	News Articles Influence on Stock Market (Use Case)	64
6.1.1.	Architecture	65
6.1.2.	Data Acquisition.....	68
6.1.3.	Extraction/Cleaning/Annotation	70
6.1.4.	Integration/Aggregation/Representation.....	74

6.1.5. Analysis Model.....	77
6.2. Discussion and Conclusions	90
7. Conclusion	91
7.1. Project Remarks.....	91
7.2. Final Remarks	92
7.3. Future work.....	93
7.4. The Profession Acts.....	93
8. References.....	96
9. Appendix.....	102
9.1. Publications	102

Table of Figures

Figure 1 - Deloitte Environment	7
Figure 2 - Deloitte Handled Objects	8
Figure 3 - Deloitte Organic Structure.....	9
Figure 4 - Methodology Design Science	10
Figure 5 - Scrum Methodology	13
Figure 6 - The Big Data Analysis Pipeline	14
Figure 7 - Hadoop Environment.....	22
Figure 8 - HDFS Architecture.....	24
Figure 9 - MapReduce engine	24
Figure 10 - Oracle Big Data Architecture.....	27
Figure 11 - IBM Big Data Architecture	28
Figure 12 - Teradata Unified Data Architecture	30
Figure 13 - Big Data Technology Panel adapted.....	31
Figure 14 - Prominent Cloud Vendors providing Big Data Solutions	33
Figure 15 - Open Source Big Data Technologies Panel.....	34
Figure 16 - Text Mining Process Tasks	44
Figure 17 - Workflow of the proposed automatic news based trading approach	47
Figure 18 - BI Architecture Layers	50
Figure 19 - Enhanced BI architecture	59
Figure 20 - Use Case Architecture.....	66
Figure 21 - Data acquisition Workflow.....	69
Figure 22 - Hbase dataset.....	69
Figure 23 - Teradata dataset	70
Figure 24 - USA cities Dictionary	71
Figure 25 - Foreign Countries Dictionary.....	71
Figure 26 - Company Dictionary	71
Figure 27 - Government and Society Dictionary	71
Figure 28 - Positive Dictionary.....	71
Figure 29 - Negative Dictionary	71
Figure 30 - Content Extractor	72

Figure 31 - Structured Data update	73
Figure 32 - Tagging Workflow	73
Figure 33 - Text Integration Approach.....	75
Figure 34 - Data Integration	75
Figure 35 - Final DataSet	77
Figure 36 - Final Analysis Model Workflow	78
Figure 37 - Confusion Matrix.....	80
Figure 38 - Knime Accuracy Statistics	81
Figure 39 - ROC Curve	81
Figure 40 - Balance Distribution	82
Figure 41 - Sentiment Category Distribution	83
Figure 42 - Market Category Distribution	83
Figure 43 - Company Category Distribution	84
Figure 44 - Accuracy Statistics for J48 algorithm Scenario 1	85
Figure 45 - ROC Curve J48 algorithm scenario 1 (two classes)	85
Figure 46 - Predicted DataSet Scenario 1 J48 algorithm	86
Figure 47 - Predicted Balance Distribution.....	86
Figure 48 - Average of stock variation Distribution.....	87
Figure 49 - Market category Distribution	87
Figure 50 - Sentiment Category Distribution	88
Figure 51 - Balance Prediction Distribution per Market and Sentiment Categories	88
Figure 52 - Balance Prediction Distribution per Company and Sentiment Categories	89
Figure 53 - Balance Prediction Distribution per Company and Market Categories	89

Table Index

Table I - Analytic Open Source Tools - Summary	34
Table II - Data Source Layer	51
Table III - ETL Layer.....	52
Table IV - Data Analysis Layer	52
Table V - Technology Landscape Analysis	53
Table VI - Big Data Complexity framework	56
Table VII - Big Data Complexity framework (calculation)	57
Table VIII - Big Vendor Solutions - Summary	61
Table IX - Big Data Commercial Evaluation	63
Table X - Big Data Architecture Phases, Goals and Expected Results	67
Table XI - Test Scenarios.....	79
Table XII - Classification Algorithm Results.....	84
Table XIII - Project Goals and Results	92
Table XIV - Domain Analysis and Requirements Engineering	94
Table XV - Design and Construction of Computer Solutions	95

1. Introduction

1.1. Scope and Motivation

The new trends of information and communication technologies and the number of users from this kind of services and products is now bigger than a few years ago. The simple use of internet for daily actions creates a trail, in an involuntary way, of our “personal” information and information related with our activities.

The McKinsey Global Institute has estimated a data volume growth of 40% per year and between 2009 and 2020 this growth will be 44 times more (Dijcks, 2012). They also assume that each 2 years the amount of data in the world doubles and in 2015 the total of data will closely reach 7.9 zeta bytes (Manyika et al., 2011).

We are now facing the “Information Era”, where information became a prized and powerful resource. The production and consumption of data about a product/service and its customers is essential to achieve the most diverse segment market, meet client’s needs and their satisfaction levels. It’s up to the organizations to be capable of transforming this information into knowledge giving them the possibility of acting in a preventive way and of shaping their business according to their target.

In this business model, both clients and organizations have advantages. On the one hand, a customer who’s interested in a service/product could obtain it in a personalized form; on the other hand, organizations could avoid risk situations, always a step forward and sustaining a competitive approach.

The factors above mentioned are responsible for the appearance of new data acquisition sources (sensors, mobile and social networks) and new data format (semi-structured, unstructured). Now, it is fundamental to understand how to deal and transform all of this data into useful information

In the current market, the process of discovering new information is so important for the business that organizations have invested in Business Intelligence as a tool for supporting the decision-making process. However, the business is changing, the consumer’s

behaviour is changing faster than ever and the market is even more unpredictable, which leads to the necessity to collect and analyse all available information.

The organizations in the future will succeed if its executives can reliably forecast the future demand. Based on that, they will effectively assess the alternative business strategies and implement them with optimal technology business solutions". (Dr. Santhosh Baboo & P Renjith Kumar, 2013).

Therefore it raises a new challenge, organizations are facing a huge amount of data but they don't know how to get value from it. Most information is obtained from raw or unstructured data, being hard to know how to recognize what is relevant and how to interpret it (Lima & Calazans, 2013).

Big Data represents a new information Era, dataset whose size and complexity is beyond the ability of conventional tools of managing, storing and analysing data (Manyika et al., 2011).

It is considered as the biggest phenomenon that has captured the attention of modern computing industries since the internet era (Krishan, 2013). It is defined as a huge volume of data available in different levels of complexity and ambiguity created at different velocities, being their development limited by the processing associated to technologies and traditional algorithms.

It will be wrong to think of this new concept just as a big dataset whose machines are incapable of processing. Big Data must be also known by its big velocity and variety of data that requires costs and innovative forms of information processing (Baboo & Kumar, 2013).

Although Big Data definition may be seen from different perspectives, it is well defined concerning its relevance and capability of solving problems for the majority of organizations. Following this idea, some solutions were developed by companies that needed to quickly deal with large quantities of data, provided from the web (eg. social network, blogs, web sites), resulting from scientific/business simulations or available from other kind of information sources (Maier, Serebrenik, & Vanderfeesten, 2013).

For that reason those companies developed some technologies with the main purpose of helping them to process and analyse this “Big Data”.

These solutions, by their iterative and exploratory features, allow users to do analysis without defined business indicators (Zikopoulos, Eaton, & deRoos, 2011) providing a great flexibility and robustness.

The Big Data could bring other advantages, because when “distilled” and analysed in combination with operational traditional information, provides a deeper understanding about business, which leads the organization to innovation processes, to productivity growth and to a competitive and strong market position (Maier et al., 2013).

However, even with the shown interest for this concept, it is still hard to persuade organizations that Big Data solutions bring value for them and it should also be a strategic priority. Big Data carries a lot of advantages but also a huge set of challenges related with collection, storage and analysis of data.

This project has the main focus on the application of Big Data concepts and technologies on the analysis of data in Financial Industry Services. The development phase was supported by Deloitte Consultores, S.A, a Portuguese company member of Deloitte Touche Tohmatsu Limited (DTTL) that provides audit, tax consultancy, and consulting and corporate finance services.

Once again, the project was embraced by Financial Services Industry, with focus on the bank sector. This is one of the candidates to get more benefits from Big Data since it is the one of the biggest data producers.

As many other companies have been doing, Deloitte could also explore some business opportunity offering technologic consulting in a Big Data perspective and exploring another set of use cases, from financial services, through healthcare and even agriculture or real state.

Understanding how and when it is possible to produce valuable information from the available data on the network overcomes all obstacles and issues underlying Big Data by preserving the organizational competitiveness in the market following the technological evolution

1.2. Goals

The chosen business-oriented scope was a guideline to define the project goals and challenges.

In a high-level, the main purpose was to get knowledge about Big Data and apply it to building a Big Data platform with analytic and predictive capabilities. For experimentation reasons and because the scope is the Financial Services Industry, the platform must be able of collect, store and process data in order to find patterns between news articles and the stock exchange variation.

Regarding the Big Data literature review, it was deemed necessary to get more insight about the current business opportunities and which offers we can find in the Big Data market.

There are so many options, so many decisions that even the organizations can lose their focus. Taking this into consideration we believe that the organizations must set its ideas about what Big Data really means and how they can get value from it. For that reason, the objective of creating a framework to help defining a Big Data complexity level for the organizations business concerns was established.

1.3. Structure of the report

The result of these six months of development is divided in several artefacts.

This document combines all the several outputs that go from the Big Data overview, from a broader spectrum to a closed scope that ends with a use case experimentation, which is a very specific branch of big data.

Trying to provide an easier way to explore the content of the project, this section is directed towards exposing the structure of the document.

The document is organized in major chapters being each one composed by sections and subsections for a better understanding of concepts and contents.

The first chapter contains the project introduction, where it is defined the scope of the theme, the predominant inspirations for the theme choice and contributions for the target organization and society.

The second presents the chosen and applied methodologies for the management and developing components of the project

The following chapter is dedicated to the literature review with researches on the definitions, concepts and particularities of Big Data, a survey about the value created by Big Data and the sectors where Big Data represents a business opportunity.

To end this chapter the research narrows a little more the scope and it is focused in concepts such as Big Data in financial services, text mining techniques and finally a review about a study that enforces the premise that news articles influence the stock market.

In order to study the project viability it was necessary to evaluate if the problem could be exclusively solved by a Big Data solution or implemented with other options.

Therefore, the fourth chapter presents a Framework capable of mapping a specific problem based on a complexity scale that justifies to the organization the necessity (or not) of a Big Data solution.

Following this, once decided the path of Big Data, it becomes imperative to choose the best solution that best suits the needs of the organizations. This way the fifth chapter presents a comparison of some of the most popular vendors.

The sixth chapter is all dedicated to the developed practical use based on the premise claimed in the Literature Review. This chapter details the entire workflow, data analysis and work conclusions.

Finally, the last chapter presents a mapping of the project through the Acts of Engineering, project conclusions and future perspectives.

1.4. Organization Overview (partial)

To describe Deloitte Consultores, SA, this section partially uses an organization overview guideline in order to provide a systemic description of the business. The guideline helps the understanding of the business through the detail of relevant aspects. This kind of document is usually used in order to perform and interventions related to adoption, exploitation and management of ITI.

A - Elements of the definition of general system
(Purpose, environment, activities, handled objects and organs)

A1) Purpose

Deloitte vision is "to be the standard of excellence" (Deloitte, 2013)

A2) Environment

The established relationships by Deloitte Consultores, SA (Figure 1) are defined in five big groups: clients, suppliers, regulators, stakeholders and Deloitte Touche Tohmatsu Limited (DTTL).

Deloitte provides services for clients from several industries such as financial services, manufacturing, energy, telecommunications, public sector, utilities and others. In the suppliers are included the services, software products, technologic support (outsourcing) and logistic.

Regarding regulators we are talking about government, quality regulators such as ISO and audit regulators, CMVM, Banco de Portugal.

Stakeholders are people with interest in the company success and it includes partners, investors and employees.

Finally the Deloitte Touche Tohmatsu Limited (DTTL), which is a UK private company limited by guarantee.

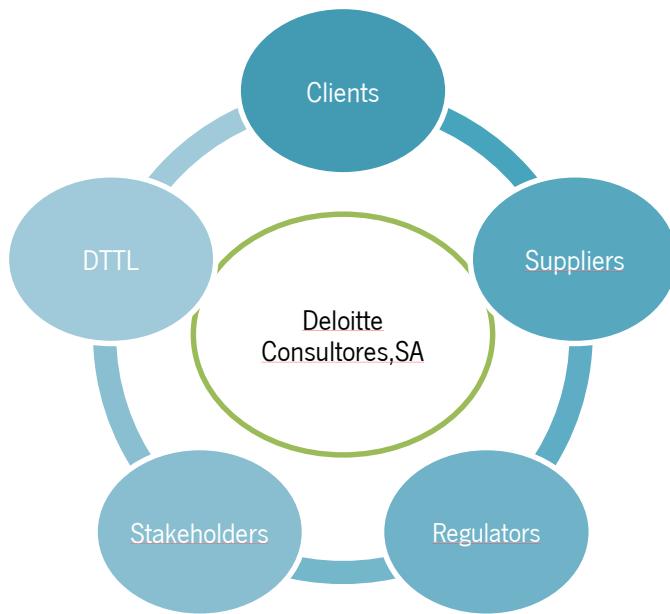


Figure 1 - Deloitte Environment

A3) Activities

Deloitte stands out for the variety of services that they can provide to their clients. The main purpose is to help them to regulate their business, find opportunities, define goals, mitigate flaws and improve the business process. Basically it helps the organization to tackle new challenges and it takes the organization to another level of competitiveness.

The services are as follows:

- Audit.
- Business Process & Applications Solutions
- Consulting.
- Enterprise Risk Services.
- Tax.
- Corporate Finance.
- Outsourcing
- Finance Advisory.

A4) Handled Objects

The handled objects (Figure 2) present the business model and the main inputs/outputs within the environment.

The flow between Deloitte and the environment is an exchange of services, information and payments for those services and products. The information exchange between DTTL, stakeholder and regulators helps the organization to align its own goal and regulate their activities to the legislation and quality standards. Looking at the clients and suppliers, the flow represents a business to customer relationship where the first provides some service or product and the second pays the associated costs. Also there is a cash-flow between Deloitte, stakeholders and DTTL that may occur in investments cases.

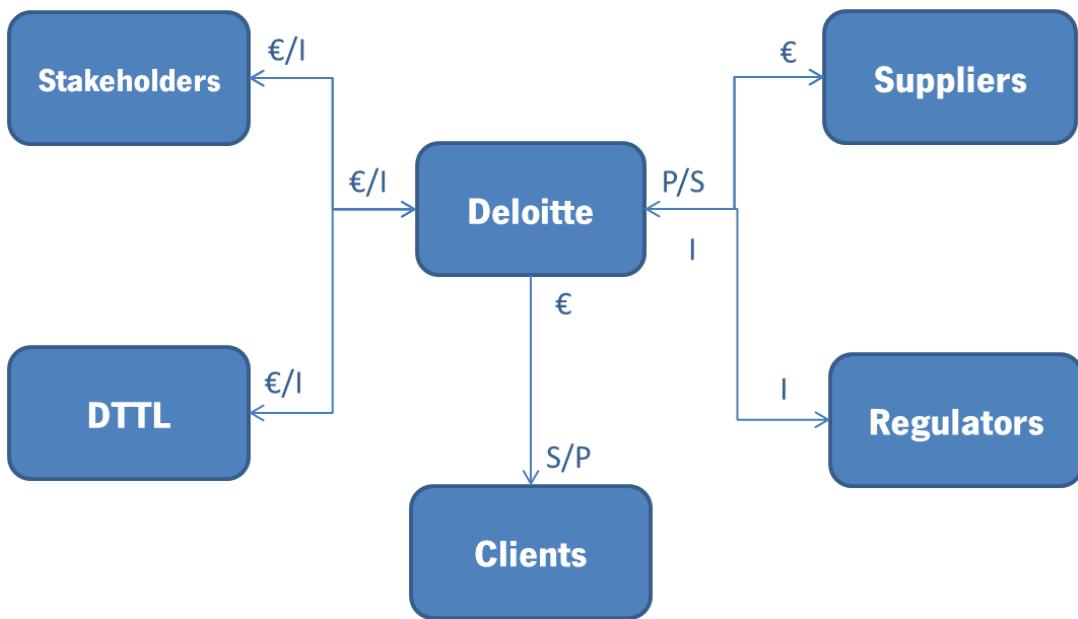


Figure 2 - Deloitte Handled Objects

A5) Organs

Deloitte defines a multidimensional organizational structure (Figure 3) that merged its business units, practice functions and Corporate Functions. Deloitte is organized by industries that provides a set of practice functions and is supported by corporate functions such as human resources, financial services, legal, etc.

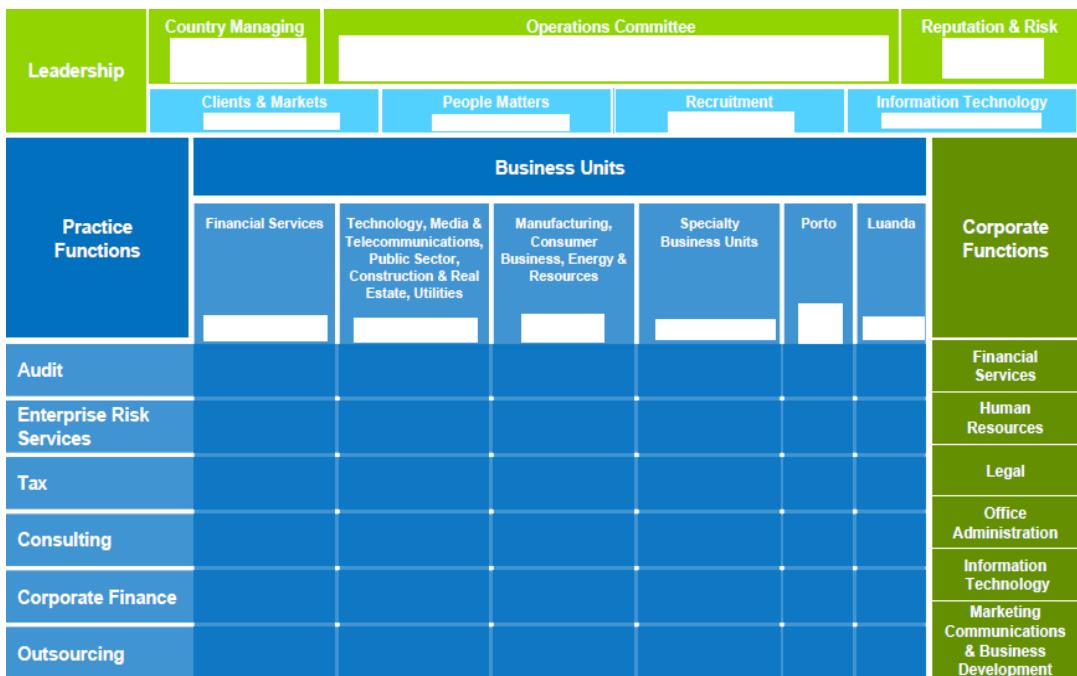


Figure 3 - Deloitte Organic Structure adapted from (Deloitte, 2013)

B) Size of company (**Characterization of firm size**)

Deloitte Consultores, SA has in Portugal and Angola 2000 professionals. Globally Deloitte has 200,000 professionals in more than 150 countries. For the fiscal year of 2014 the revenues reached the amount of US\$34.2 billion.

2. Methodologies

This chapter contains the methodologies used both in research and practice. Although this is a practical project, it follows a research methodology to give greater credibility to the scientific work.

2.1. Research Methodologies

To carry out this investigation, the Design Science Research (DSR) approach is the more appropriate because it was focused in the development and validation of an artifact, being in this case a framework and a big data platform with analytical and predictive capabilities. As shown in Figure 4, the DSR approach is composed of five main phases: Awareness of problem, Suggestion, Development, Evaluation and Conclusion (Hevner, March, & Park, 2004)

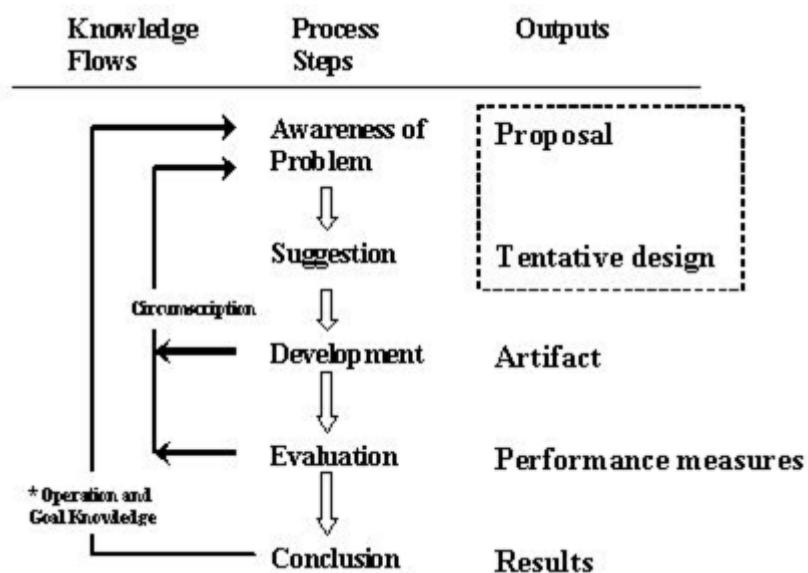


Figure 4 - Methodology Design Science from (Vaishnavi & Kuechler, 2004)

The first phase of the process reflects the awareness of the problem, where it is presented the knowledge gap, increasing the familiarity around the problem. The result of this phase is the description of the topic under investigation.

The following step concerns the suggestion of the solution. After deepening the knowledge in the area, the result obtained upon completion of the literature review, allows the researcher to overcome the shortcomings of existing knowledge in the initial stage, serving as theoretical support for research. The next phase is the development of the solution and where it is centered the most of the research work, having as the final result the defined

artifact. The review stage concerns the validation of the artifact. Finally, the conclusion is the last stage and it is where the design cycle ends, determining the impact of quality and artifact produced.

In order to accomplish this investigation, a DSR approach was undertaken focusing on the seven guidelines provided by Hevner (Hevner et al., 2004) artifact, problem relevance, research rigor, design as a search process, design evaluation, research contributions and research communication.

2.1.1. Artifact

This project presents a functional framework Big Data complexity level and a big data platform with analytical and predictive capabilities. This framework helps organizations to perceive the problem they have in hand: if it is a Big Data problem or not, avoiding making overvalued investments. The platform is a set of integrated tools that collect structured and unstructured data, process this information to be later analyzed through data mining techniques. These techniques will set a training model based in the data patterns with the purpose of predict future stock exchange variations

2.1.2. Problem Relevance

The speed of data creation, accompanied by the dynamism of the market makes the life cycle of the organizations decrease. If they are not able to adapt to changes, they will be forgotten. The Big Data comes with the purpose of helping the company staying a step ahead by making use of the most important resource, information.

2.1.3. Research Rigor

The proposed methods, frameworks and platform, were tested using real datasets in order to correctly evaluate its utility and efficacy. To verify the proposed framework and platform real scenarios were simulated with consultation of experts in the field. All results and methods were also verified several times before any conclusions were taken.

2.1.4. Design as a Search Process and Evaluation, Research Contributions and Communication

All frameworks and platforms produced in this project will be submitted to open-source repositories in order to contribute for further research in the matter. From this output scientific publications will be developed and will be submitted to journals and peer review conferences. Furthermore, the result of this work will be also available at the Minho University library.

2.2. Practical Methodologies

2.2.1. SCRUM

It defines a set of management practices and roles that must be adopted to guarantee the project success (Bissi, 2007).

Scrum is a development process to small teams to build a product in an incremental and interactive way when it is complex and unpredictable.

The product development occurs in small pieces, where each part is created above the previous one. This guarantees that the product complies with the requirements.

In the Scrum methodology (Figure 5) it is defined three roles, project-owner, project master and team member.

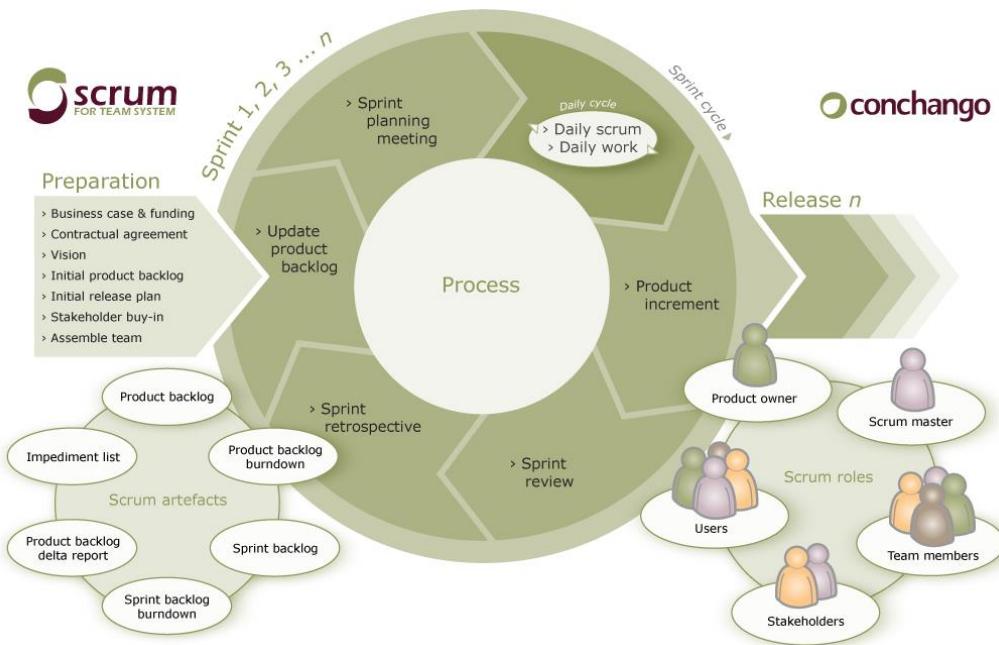


Figure 5 - Scrum Methodology from ("Scrum features," 2010)

Project-Owner: Represents the customer needs and communicates those requirements. It is responsible for aligning the development team with the vision and objectives of the project and guaranteeing that the team delivers as planned.

Project Master Responsible for removing any obstacles that are obstructing the team from achieving its sprint goals

Team Member: The responsible plays the role of a "Completer". "For software projects, a typical team includes a mix of software engineers, architects, programmers, analysts, QA experts, testers, and UI designers." (Danube, 2008).

A project based Scrum consists of three phases:

Pregame

- Planning: Definition of a new project based on the actual known backlog along with scheduling and cost estimative, leaders and evolved resources distribution.
- Architecture: Defining and designing the backlog items that will be development. The development tools and infrastructure are defined. This phase includes also the system architectures changes.

Game

Sprint: Developing of a functionality whose tasks are defined also by a backlog. It must respect the time variables, requirements, quality, costs and competitiveness. The interaction between those variables defines the end of this phase. Throughout the whole project it exists a lot of multiple interactive sprints that are used to develop the system.

Postgame

Closure: Preparation for the product release, including final documentation, systems and integration test, preparation of the marketing material, etc. This phase is started when all the quality, time, requirements, cost and competitiveness aspects are accomplished.

Sprint: Developing of a functionality whose tasks are defined also by a backlog. It must respect the time variables, requirements, quality, costs and competitiveness. The interaction between those variables defines the end of this phase. All through the project there is a range of multiple interactive sprints that are used to develop the system.

2.2.2. Big Data analysis pipeline

The white paper “Challenges and Opportunities with Big Data” (Agrawal et al., 2012) presents a pipeline called “The Big Data analysis pipeline”(Figure 6) that is a set of multiple and distinct phases.

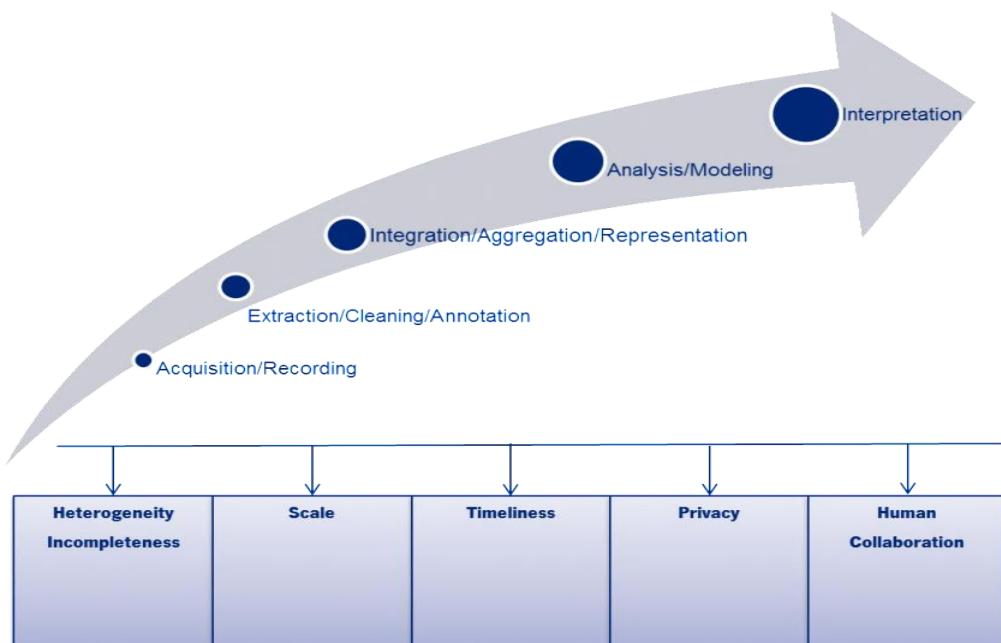


Figure 6 - The Big Data Analysis Pipeline adapted from Agrawal (Agrawal et al., 2012)

The phases are as follows:

Data Acquisition and Data Recording: Data is collected from several sources. Much of this data has no interest and it can be filtered up front. Filters must be well defined to avoid discarding useful information and defined with chosen business requirements and information sources. After collecting the data it is important to record information about it. It is fundamental to generate the right metadata to describe what data is recorded, how data is recorded and what is the data origin, which could be a big challenge.

Extraction/Cleaning/Annotation: The data needs to be extracted in a format suitable for analysis. This extraction is often highly application dependent.

Integration/Aggregation/Representation: Data will be integrated because with the heterogeneity of the flood of data is not enough merely to record it into a repository. Related data from different sources and types will enrich the database and improve de search. For effective large-scale analysis all of this has to happen in a completely automated manner. The data structure must be prepared to be expressed in forms that will be easy to understand.

Analysis/Modeling: Perform methods for querying and mining Big Data. This process is different from tradition analysis because it needs a major effort to integrate, clean and turn data trustworthy and efficiently accessible, a declarative and mining interface, scalable mining algorithms and Big Data computing environments. Also, here data mining can be used to help test and improve the quality and trustworthiness.

Interpretation: Interpretation of the analyzed data results. It involves examining all the assumption made and retracing the analysis. Here are provided not only the results, but also supplementary information that is important to justify and clarify those outputs. Also the created infrastructure provides to end users the ability to interpret the results from different perspectives and scenarios.

Besides, in this pipeline, each phase introduces a Big Data challenge.

The challenges are:

Heterogeneity and Incompleteness: Machine analysis algorithms expect homogeneous data and cannot “read” the nuances or richness of natural language that usually provides valuable information.

Scale: Managing large and fast data growth of volumes has become an issue since data scales are faster than the capability of processing resources.

Timeliness: Designing a system capable of collecting and analyzing data almost instantly is something difficult to do. Producing results near-real time by scanning that entire total is unfeasible. Index structures could be the solution but we still have the challenge to create those structures that must be always prepared to receive the new classes of criteria which come out with new data.

Privacy: The paradigm of privacy is a big struggle in the Big Data context and managing privacy is both a technical and a sociological issue. Today many applications require shared personal information but the users of such social networks don’t even know what really is “share private data”. However, the reality is that we don’t have control over our own information. Similarly users can provide private information without knowing it, by leaving a trail with their web history comments. The question and the challenge is to define what is or not private and how we can share private data while limiting disclosure, at the same time, ensuring sufficient data utility in the shared data.

Human Collaboration Ideally, analytics for Big Data will not all be computational, but rather designed to have human in the loop. The challenge is to support all the human resources working in team in a distributed way and the shared exploration of results. This kind of collaboration is already solved through the concept of crowd-sourcing but the issue is that in some cases the information provided could be false. Here the technologies must be capable of detecting and correcting those errors. Also, it is needed a framework to use in the analysis of crowd-sourced data with conflicting statement. The machine should be prepared to perform “human-like” tasks ,to classify texts in the similar way that humans can evaluate reviews as positive or negative.

3. Literature Review

This chapter presents a review about the Big Data concept and its particularities. The review of concepts and features will be followed by the comparison with Business Intelligence (BI) technologies and value creation associated to Big Data.

During the literature review questions will be raised about when to implement a project of this kind and what options (tools and solutions) are now available in the market.

3.1. Big Data

New forms of data creation have been merged in the last few years. There are more and more people and organizations that are using the technologic advances to improve their “lives” either in the hardware/software level, both within the network (internet) and services.

The way society and the people communicate with each other changed the way data is produced and consumed. The development of mobile/ubiquitous and pervasive computing and the emergence of technologies such as smartphones and tablets, leads the societies into a process of connection even more universal, in such a way that being connected became an preponderant factor for our everyday (Lima & Calazans, 2013). In fact it makes the society more dependent of having the information available anywhere, anytime.

In an organizational point of view, the information while an important asset, is generated and stored in amounts that round Zetabytes. Organizations collect large volumes of information of their clients, suppliers, operations and millions of sensors linked to the network and implemented in the physic world in devices such as mobile phones, computers, cars, etc., (Manyika et al., 2011).

According to Lima (Lima & Calazans, 2013) every minute, 57 of new web sites are generated and 204.166.667 new email messages are sent. Mobile network gets 217 new users and Google receives more than 2 million searches. YouTube receives 48 hours of new videos, 684.478 contents are published on Facebook, 3.600 photos are shared on Instagram and more than 100 000 tweets are sent through twitter. The trend is that these numbers keep growing exponentially and growing around 40% times more.

The observed data explosion occurs due to four main factors such as innovation, transformation of business models, globalization and service personalization (Krishan, 2013a). In addition the innovation verified in the last years changed the way the organization commits to their businesses, to the client services delivery and measuring the yield and value associated.

The appearance of Web 2.0 changed the way information is shared between friends and family, the approach of organizations to managing their business and the manner they interact with their customers (Krishan, 2013a).

Due to the globalization, which changed the world market (Zikopoulos, Eaton, & deRoos, 2011), new business models presented a more service oriented side where the value of organization is measured by the service efficiency and not by the product usefulness.

These and other factors proved that data explosion is the inception of a new Era called “Big Data Era” or “The Age of Big Data”. A period referred to the huge volume of generated, stored and consumed data. An Age that has been a target for several publications and discussions but it still is a concept with a not well interiorized main purpose.

For that reason it makes sense for the next subsections to explain Big Data’s definition, particularities, advantages, challenges, the evolution stage and even the market presence.

3.1.1. Definitions and Concepts

There are many definitions about what Big Data should be but all of them essentially talk about huge amounts of existing data with great complexity, where traditional machines are enabling to process it.

A common description is Big Data represented by a challenging large dataset to store, search, share, visualize and analyse (Sun & Heller, 2012).

To Cuzzocrea (Cuzzocrea, Song, & Davis, 2011) this concept is related to a large quantity of non-structured data produced by devices of high-performance in different scenarios.

According to Baboo (Baboo & Kumar, 2013) Big Data is everywhere, from sensors to monitor and manage the network traffic, until “tweets” and “likes” on Facebook.

Also Manyika (Manyika et al., 2011) declares that Big Data is a dataset with a size and complexity that is beyond the ability of conventional tools of managing, storing and analysing.

On the other hand, the International Data Corporation (IDC) defines Big Data technologies as a new generation of technology and architectures designed to economically extract value from a huge amount and large variety of data acquired, improved by a high speed capture, discovery and analysis (Maier, Serebrenik, & Vanderfeesten, 2013).

Finally, Hurwitz (Hurwitz, Nugent, Halper, & Kaufman, 2013) refers that it isn't just a single technology but a combination of older and new technologies that help organizations to obtain practical knowledge. Big Data is the ability to manage a large amount of dissimilar data, with a considerable velocity and between a short time frame improving the real-time analysis and reaction.

The above definitions are very similar; some authors have thought about it more thoroughly than others. However, there is a consensus when it is claimed that traditional tools are not suitable for these new concepts. The first two definitions are more focused on the amount and presentation of data in a non-structured form. The remaining others include several and different sources, format, velocity, storage and analysis' complexity.

From this research it is possible to associate more key-words to Big Data. We do not think of volume as the only concern. We can now talk about a set of technologies capable of collecting, processing and getting value from data provided from several sources in real-time. This data could be structured, semi-structured or with no structure at all.

3.1.2. Characteristics and Challenges

Authors such as the ones referred above claim that Big Data possesses some main features that summarizes this entire concept. Those features are grouped by "V's" that in the ending will represents a set of requirements to support a Big Data infrastructure (Demchenko, Grosso, De Laat, & Membrey, 2013).

Initially this set was composed by three V's: Volume, Variety and Velocity (Zikopoulos et al., 2011). However, some authors preferred to go further and joined others like Veracity and Value (Sridhar & Dharmaji, 2013).

Although Sridhar (Sridhar & Dharmaji, 2013) and others considered five V's, in this project we will consider only the first three V's (Volume, Variety, Velocity). We believe that the Veracity and Value are also important for Big Data but they are not exclusive for the new way to process data. In fact, to guarantee the Veracity and Value of data for the business is something that already is done in Business Intelligence infrastructures.

Therefore, those are the three V's that define Big Data:

a) Volume

It leads to an amount in terabyte scale created by several sources and devices. The amount of data is so huge that dealing with it becomes an important challenge. The data arrives at high velocity, with different formats and needs to be processed almost in real-time. The traditional analysis is not enough, requiring analysis' methods more complete (Zikopoulos et al., 2011). The author also sustains that the volume cannot be considered as the only feature to represent a Big Data challenge. It must be related with the other V's, so together they can require a large computation capacity and process complexity.

b) Velocity

Data is created at high velocity and needs to be processed (near) real-time. An example of that high velocity is the data created by sensors in monitoring devices.

According to Maier (Maier et al., 2013), the challenge of the velocity is related to the capacity of processing the huge amount through "small operations" and, at the same time, keeping some consistence and persistence. The author presents as a solution the use of filter applied in the collection and storage, recording only what is important. This suggests a technology capable of making those filters in an intelligent and automatic way without losing vital data, which implies investing on resources and time.

More than how fast technologies can process and analyse all the data that arrives, it is crucial to take into account the reaction to those analysis. Applying the knowledge about the situations reported is necessary to the survival of the business.

c) Variety

The presence of structured, semi-structured and non-structured (text files, audio, video) data is provided by the use of several sources such as social network, sensors, mobile devices, GPS, and others.

This variety is challenging in what concerns the storage, processing and management of data. The Relational Database Management Systems (RDBMS) could not be suitable to store all the different types.

The technological development on this area allows the existence of “extractors” capable of collecting entities, relations and other information in text format (Maier et al., 2013). However, as referred, in Big Data there is more than this type of data and also it will be necessary to extract its information. Here the major problem is not only the difficulty in collect information from non-structured data, but also the integration of this information with other data, structured or not.

Knowing how to get value from data and indicators until now excluded, will open new opportunities and eradicate some existent flaws in the business models leading the organizations to the top of the market.

3.2. Big Data and Hadoop

When we talk about Big Data architectures and technologies the concept of Hadoop appears almost as a synonym. It is part of Big Data processing, classified as a distributed cluster along Massive Parallel Processing (MPP).

The Hadoop is an implementation in Java and Open Source of distributed computing used for processing and storage of data in large scale. This apache works by dividing workloads across three thousands of servers or machines, each one offering local computation and storage.(Deloitte, 2014)

A Hadoop environment (Figure 7) is composed by its core components (MapReduce, Hbase and Hadoop Distributed file systems) related components, tools/projects developed on top of Hadoop and outside technologies.

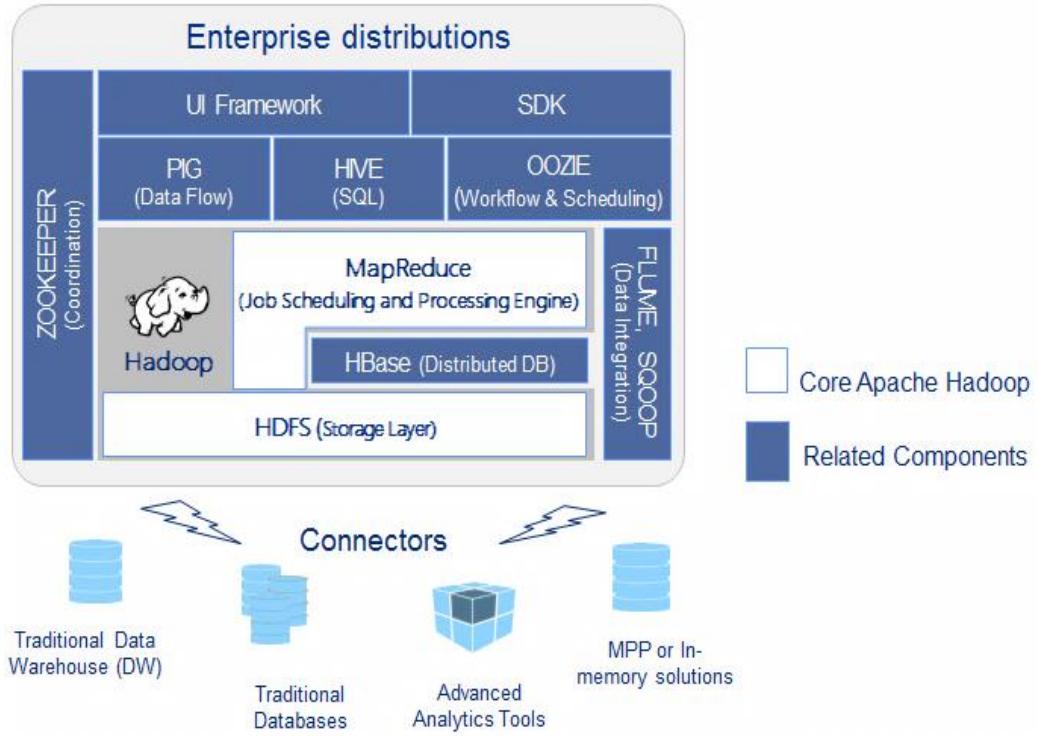


Figure 7 - Hadoop Environment from (Deloitte, 2014)

Companies seek deeper insights from the massive amount of structured, unstructured, semi-structured, and binary data at their disposal in order to dramatically improve business outcomes (Gualtieri & Yuhanna, 2014). Hadoop can help by:

Capturing: and storing all data for all business functions: Hadoop has made it possible for enterprises to capture, store, and analyze a lot more data in a much more cost-effective way.

Supporting: advanced analytics capabilities: Hadoop does not just involve supporting large volumes of data; it also has to have the compute horsepower to perform the advanced statistical and machine learning algorithms that data scientists use. And it is designed to scale out from a single server to thousands of servers for optimized performance.

Sharing customer: Data quickly and generously with all those who need it: Hadoop can be used to create a “data lake” — an integrated repository of data from internal and external data sources.

Continuously: Accommodating greater data volumes and new data sources: Hadoop can scale linearly to accommodate any amount of data, making it a future-proof solution for the enterprise.

3.2.1. Apache Hadoop Core System Components

a) Hadoop Distributed File System

Distributed file system that simplifies the management of related files through machines and provides high throughput access to data. It is not the final destination for the file. "Is designed for use of MapReduce jobs that read input in large chunks of input, process it, and write potentially large chunks of output." (Venner, 2009)

Hadoop Distributed File System (HDFS) is intended to reliably store very large files across machines in a large cluster. It stores each file as a sequence of blocks; all blocks in a file, except the last block, are the same size. The blocks of a file are replicated for fault tolerance. (Apache Foundation, 2014)

This file system stores metadata on a dedicated server called the NameNode. Application data are stored on other servers called DataNodes. All servers are fully connected and communicate with each other using TCP-based protocols (Chansler, Kuang, Radia, Shvachko, & Srinivas, 2014).

HDFS (Figure 8) works by breaking large files into smaller pieces (blocks) stored on data nodes. NameNode is responsible for knowing what block on which data nodes make up the complete file. Moreover it is responsible for managing all access to the file including reading, writing, creating, deleting and replication of data blocks (Hurwitz et al., 2013).

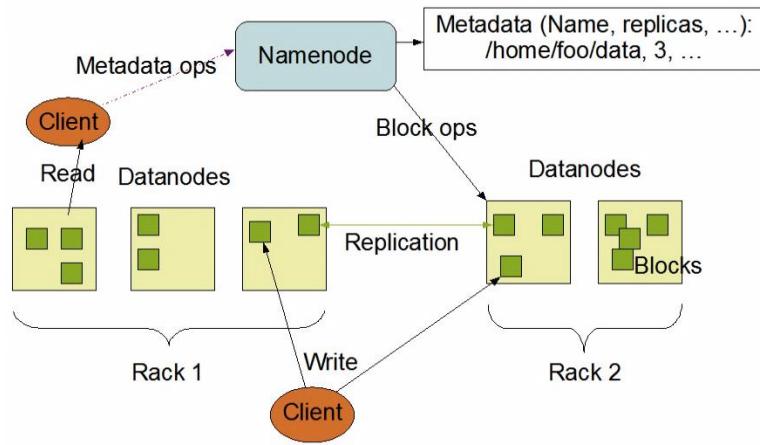


Figure 8 - HDFS Architecture from (Apache Foundation, 2014)

b) MapReduce

MapReduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes.

MapReduce is at the head of Hadoop and consists of two separate and distinct tasks – Map and Reduce (Figure 9). The map job takes a set of data and converts it into another set of data where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output of map job and combines data tuples into a smaller set of tuples.(Deloitte, 2014)

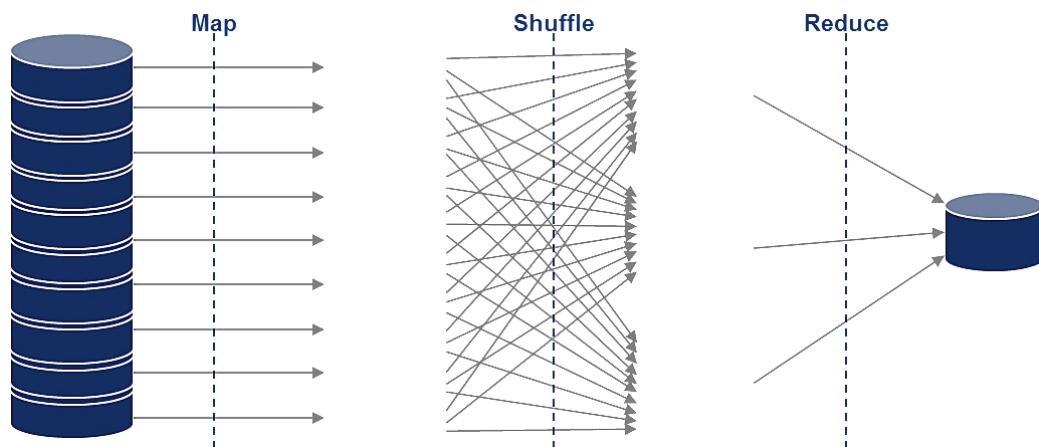


Figure 9 - MapReduce engine from (Deloitte, 2014)

3.3. Big Data Architecture

Whatever approach is presented, Big Data architect aims to solve problems through storage, process, and analysis of these new data.

The section below presents an overview of some of the existing architecture, followed by a technological analysis.

3.3.1. Context

Big Data architecture is seen as a cluster, which means it is based on a distributed processing. In a cluster architecture the processing machines are connected to a network architecture, either software or hardware. Each machine performs its task locally to process data or computational requirements and store every result in the main server.

Processing this way improves the availability to fault tolerance and reduces the time and costs. However, when dealing with high data dependency and at low complexity levels, leads to unnecessary redundant task distribution due to excess of resources.

Big Data processing differs from the data from traditional BI architectures (Krishan, 2013).

The traditional architectures first analyze the data and define a set of requirements that lead to the exploitation of data and to the creation of a data model. From an architecture design's point of view, this seems to be more efficient because the database is created for a data schema already defined.

In contrast, the data in Big Data architecture is first collected and loaded onto a platform where metadata is set to be classified and categorized. Then a data structured is built for the collected content.

Once the structure is identified, the data is transformed and analyzed. The knowledge is created from the results, according to the collected data and the business context.

Krishan (Krishan, 2013) indicates that because of the volume and data complexity, the processing will be more efficient with a file-driven architecture, where the storage is made in files, in a file system with a programming language interface.

3.3.2. Big Data Architectures

In this section we will be looking at some architecture approaches used by the current developers and suppliers in the Big Data market. We found it interesting to look for what the database suppliers have been doing in this area, something that we must take in account when implementing Big Data architecture in our organizations.

a) Oracle

Oracle presents two technological strategic for real-time or batch processing.

To real-time processing, Oracle does key-value storage such as NoSql ("Not only SQL") and for batch processing they apply a MapReduce technique that filters data according to a well-defined strategy. After that, the data can be directly analyzed, loaded onto non-structured data bases or included in a traditional Data Warehouse (DW) and merged with relational data (Sun & Heller, 2012). We must beware that the non-structured data cannot be stored directly in a DW. However, the "results" of the MapReduce process may be stored in DWs, ready for visualization, analysis, report, and others.

By doing this, the software optimization using programming procedures is no longer necessary, which was complicated and expensive.

As referred before, Oracle also defends the best idea is to combine BI with platforms capable of dealing with Big Data (Sun & Heller, 2012).

The Oracle Big Data Architecture (Figure 10), shows a proposal for a capability map that integrates traditional and Big Data components.

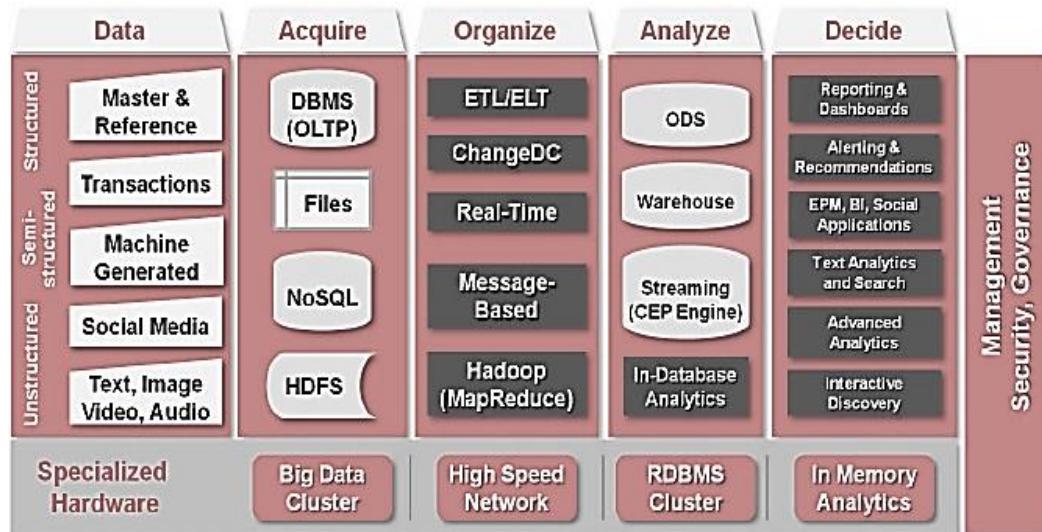


Figure 10 - Oracle Big Data Architecture from (Sun & Heller, 2012)

In the data layer all types of collected data are represented and they go from the structured to the no-structured. Data is later stored into RDBMS, files, and it distributes systems such as NoSql and the Hadoop Distribute File Systems (HDFS).

In the organization layer, the data is processed, cleaned, filtered, organized and categorized. In this case, tools need to be capable of both traditional and new ways of data acquisition. Also, they must guarantee that they can operate with both high amounts and frequency of data or lower volumes and large periods to collect it and process it.

It is important that these tools show the flexibility to live in an integrated environment, which means, to be capable of integrating Hadoop/MapReduce with a DW in a bidirectional way.

After, in the analysis layer, the results of de reduction are loaded onto Operational Data Store (ODS), DW, Streaming – Complex Event Processing (CEP) or In-Database Analytics that in a simplified way allow the analysis to be applied directly onto a database and not analysis servers.

Finally, the decision layer includes statistical and visual analysis to support the decision making. The main difference in terms of what exists in BI architectures is the possibility of performing Interactive Discovery triggered by the in-Database Analytics.

With the way this architecture is prepared, users can not “see” the frontier between Big Data and the transactional information in the company that simplifies change management activities (Sun & Heller, 2012).

b) IBM

IBM (IBM, 2013) presents a similar architecture to the Oracle architecture. It organizes its components in four layers (Figure 11): Big Data sources (structured, semi-structured, non-structured data), data processing and storing, analysis and consumption

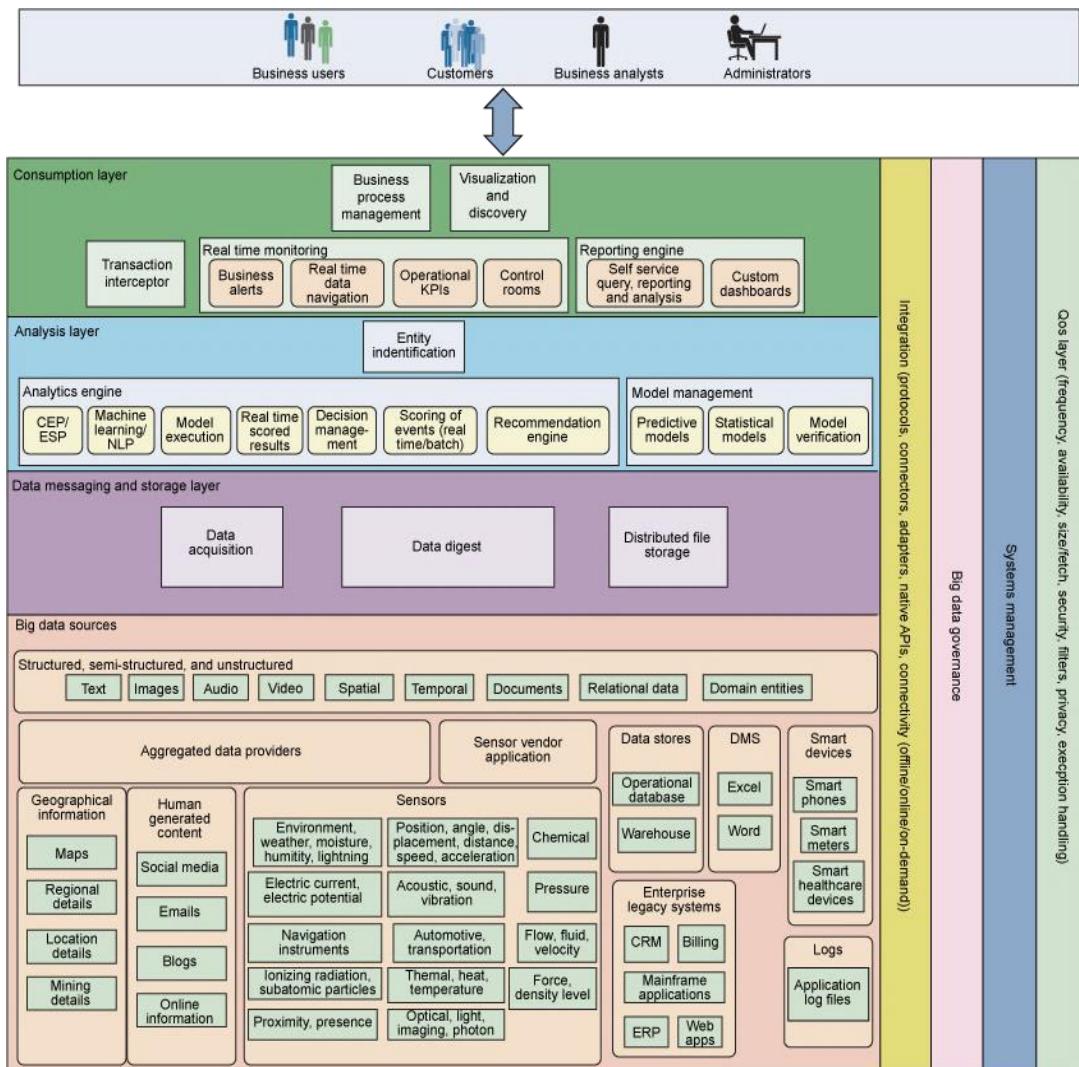


Figure 11 - IBM Big Data Architecture from (IBM, 2013)

c) Teradata

Teradata presents the Big Data architecture (Figure 12) as a combination of three of its own platforms, Terada Portfolio for Hadoop, Teradata Aster Discovery Platform and Teradata Integrated Data Warehouse (IDW).

The first one loads any type of data without performing any kind of preprocessing. It is in this portfolio that the data is reduced and prepared to be passed on to further discovery analysis (Schrader & Dietz, 2013). The next platform, Aster Discovery Platform, can allow the processing and analysis using the MapReduce engine or SQL queries. This platform can also combine analysis by applying several analytic technologies. This way, users can explore, mine and perform interactive analytics against any kind of data, including large volumes and different types from various sources, such as any data from Hadoop (Schrader & Dietz, 2013).

The last platform provides operational and strategic analytics for the BI, adding the benefits of Big Data Analytics.

The similarities between most of the architectures in the market are visible, in fact, they all suggest the use of distributed computing through Apache Hadoop and provide the possibility of an integration between transactional and Big Data. We can also see in more detail in the (Figure 12), the flexibility and multiplicity of data workflows. All layers can play, at the same time, as a source and destination of data processing/analysis. The insight can be obtained by complex processing and combining of several data sources, in order to enrich the existing operational data. Or it can be simplified by analytical discovery, in order to search for new business opportunities or to find some performance gaps that the organization cannot see.

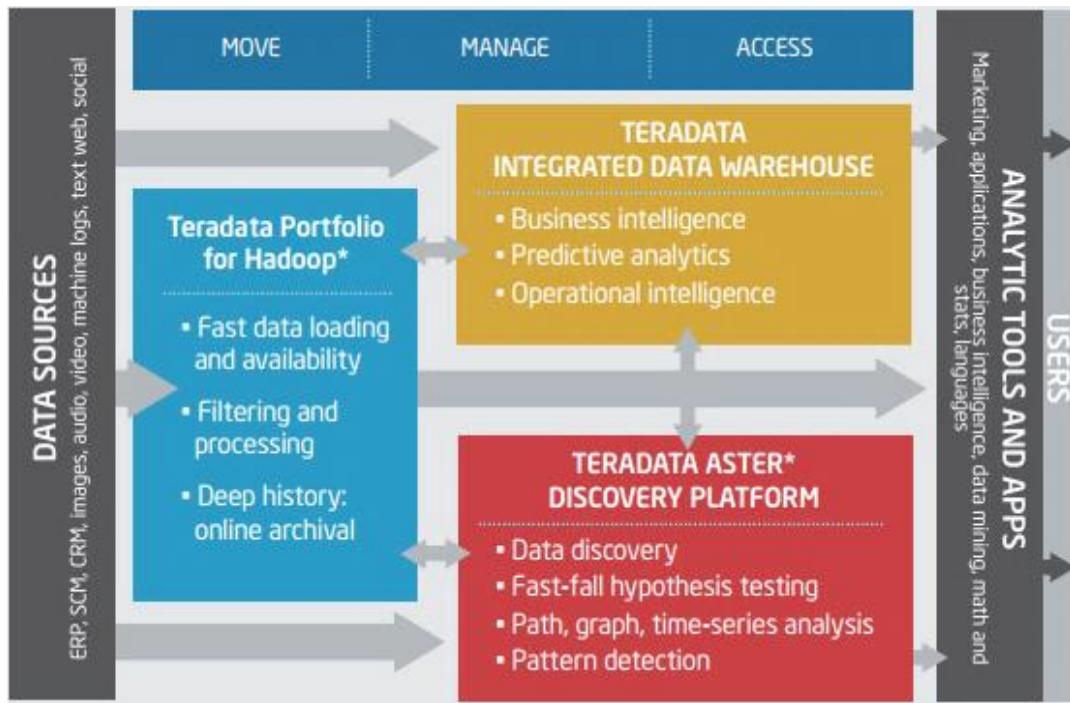


Figure 12 - Teradata Unified Data Architecture from (Schrader & Dietz, 2013)

3.3.3. Big Data Solutions

Since the appearance of the Big Data buzzword some companies made their own solutions to solve their problems with complexity and volume .The success of these solutions sells Big Data as something very useful and unique. Nowadays, companies are invaded by a lot of ideas and tools by all kind of suppliers.

For that reason, this chapter tries to display and categorize this panel of technologies with the purpose of simplifying the choice.

At this point, all the technologic offers are divided into Commercial, Cloud and Open Source Tools. Some suppliers have their offers in commercial and open source (sometimes free) distribution that justify their presence on both panels. Moreover, on both deployment options (Commercial, Open Source) a short summary of some technologies is provided. The Commercial section presents Big Data Solutions known as Massively Parallel Processing (MPP) tools.

Finally, in the open source section we can find the summary of Analytic Open Source Tools.

a) Commercial

The commercial or vendor-provided software is a software tool with property rights. The software is designed for sales purposes and it satisfies commercial needs. It is the model where the software developed by a commercial entity is typically licensed for a fee to a customer (either directly or through channels) in object, binary or executable code (GOH, 2006).

To the customers, this kind of offer is (usually) related as a higher quality, secure and trustworthy software.

The Big Data Technology Panel (Figure 13) exhibits commercial tools distributed in a Big Data Ecosystem that is composed by four big classes.

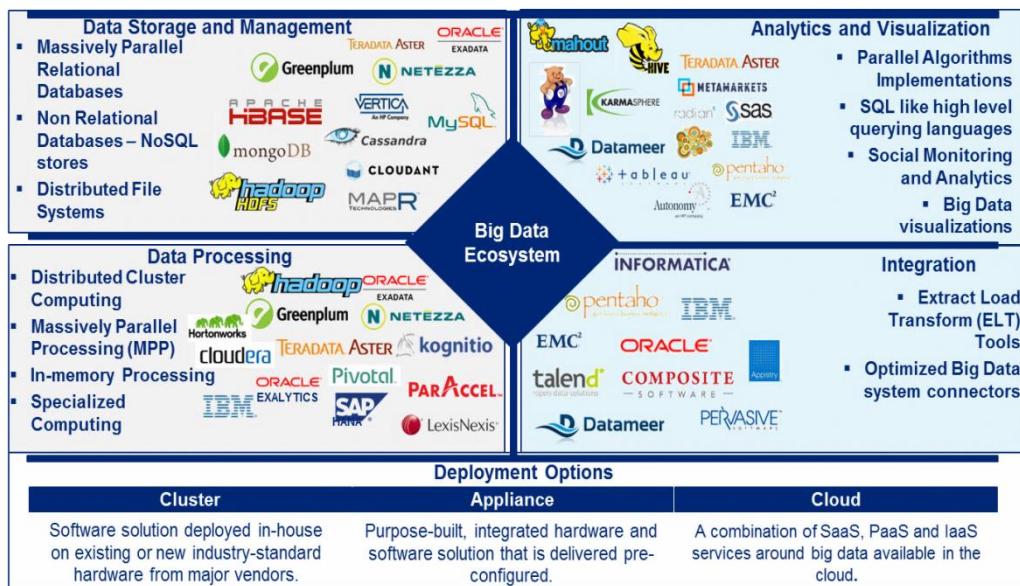


Figure 13 - Big Data Technology Panel adapted from (Deloitte, 2014)

Data Storage and Management: Technologies that have the capacity to store large volumes of different types of data. These technologies are also capable of managing and automating their process to maximize and improve the performance of their storage resources. The data storage can be in the Massively Parallel Relational Databases, Non-Relational Databases (NoSQL stores) and Distributed File Systems.

Data Processing: Technologies that are known for collecting and manipulating data to produce meaningful outputs, suitable to be analyzed. The process could be the Distributed

Cluster Computing environment, the Massively Parallel Processing (MPP) and the In-memory and Specialized Computing.

Analytics and Visualization: Technologies which are capable of providing to end-users the advanced mechanisms to explore analyze and present data in a pictorial or graphical format. In this class we found tools with Parallel Algorithms Implementations, SQL like high level querying languages, Social Monitoring and Analytics and a Big Data visualizations feature.

Integration: Technologies well suited to combine data from different sources to give an integrated view of valuable data. This process embraces extraction, cleaning, transformation tasks. The existing offer consists of Extract Load Transform (ELT) tools and Optimized Big Data system connectors.

b) Cloud

Cloud Computing it is another theme that is on top of the organizations' minds .As a delivery model for IT services, cloud computing has the potential to enhance business agility and productivity while enabling greater efficiencies and reducing costs (Intel IT Center, 2014).

Although Cloud Computing is in an evolution process, it continues to mature and a growing number of enterprises are building efficient and agile cloud environments, as cloud providers continue to expand service offerings.

To manage Big Data challenges like flexibility, scalability and cost to data access, the clouds are already deployed on pools of servers, storage and networking resources and they can scale up or down according to the needs . Cloud computing offers a cost-effective way to support big data technologies and the advanced analytics applications that can drive business value.

It will take a while for organizations to see the cloud as valid, secure and completely prepared for their needs.

As presented in the cloud vendor map (Figure 14), there is a wide variety of cloud computing service models to provide storage, management and processing for Big Data.

Organizations can leverage SaaS, PaaS or IaaS solutions depending on the needs (Deloitte, 2014).

Cloud Solution		Amazon	Microsoft	Google
Storage	Big Data Storage	Amazon S3	HDFS	Cloud Storage (GFS)
	NoSQL Store	DynamoDB	Table Storage API	AppEngine Datastore
	Relational Store	MySQL or Oracle	SQL Azure	Cloud SQL
Processing	Hosting Service	Amazon EC2	Azure Compute	AppEngine
	Map Reduce Service	Elastic MapReduce	Hadoop on Azure	AppEngine Mapper API
	Big Data Analytics	Elastic MapReduce	Hadoop on Azure	BigQuery

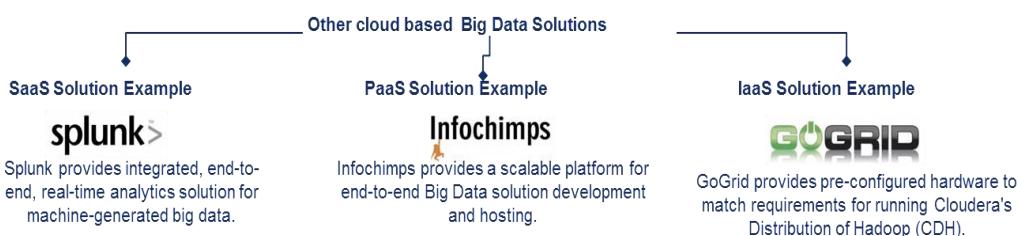


Figure 14 - Prominent Cloud Vendors providing Big Data Solutions from (Deloitte, 2014)

c) Open Source

Open Source is a kind of offer defined as source code which must be available for everyone and all modifications made by its user can also be returned to the community.

These technologies aren't necessarily free. Anyone who wants to can sell an open source program however, its prices will be low whereas the development to achieve new markets will be fast.

Similar to the earlier sections of this chapter, we provide an open source technologies panel (Figure 15) that is supplied by "native" open source companies or by renown Big Vendors who provide some tools in an Open Source Model.



Figure 15 - Open Source Big Data Technologies Panel from (Hague & Netherl, 2014)

The table Analytic Open Source Tools (Table I) describes some of the technologies in the panel (Figure 15). For now the table focuses only on the analytic tools because they are best known and easily accepted by companies

Table I - Analytic Open Source Tools - Summary

Key Tools	Summary	Class
Knime	<ul style="list-style-type: none"> • Data analytics platform that allows you to perform sophisticated statistics and data mining on your data to analyze trends and predict potential results. • Its visual workbench combines data access, data transformation, initial investigation, powerful predictive analytics and visualization • The open integration platform offers over 1.000 modules. KNIME is also the open source data analytics platform (Knime.org, 2010) 	Data Integration Data Mining

Key Tools	Summary	Class
WEKA	<ul style="list-style-type: none"> WEKA stands for “Waikato Environment for Knowledge Analysis” and it is a collection of machine-learning algorithms in order to solve data mining problems. It is written in Java and thus runs on almost any modern computing platform. It supports different data mining tasks such as clustering, data pre-processing, regression, classification, feature selection as well as visualization.(Hague & Netherl, 2014) 	Data Mining
RapidMiner	<ul style="list-style-type: none"> RapidMiner offers data integration and analysis, analytical ETL and reporting combined in a community edition or enterprise edition. It comes with a graphical user interface for designing analysis processes. The solution offers a metadata transformation, which allows inspecting for errors during design time.(Hague & Netherl, 2014) 	Data Integration Data Mining
Talend	<ul style="list-style-type: none"> Open source software developed by Talend has developed several big data software solutions, including Talend Open Studio for Big Data, which is a data integration tool supporting Hadoop, HDFS, Hive, Hbase and Pig. The objective is to improve the efficiency of data integration job design through a graphical development environment. Next to open source tools, Talend also sells other commercial products.(Hague & Netherl, 2014) 	Data Integration
Jaspersoft	<ul style="list-style-type: none"> Jaspersoft has developed several open source tools, among others a Reporting and Analytics server, which is a standalone and embeddable reporting server. The Open Source Java Reporting Library is a reporting engine that can analyze any kind of data and produce reports in any format. Jaspersoft ETL offers a data integration engine, powered by Talend. They claim it is the world's most used business intelligence software. 	Data Integration

Although commercial software is usually related to higher quality and trustworthy software, Open Source Software has attracted substantial attention in the last years (GOH, 2006).

The same happens with cloud solutions, they represent an attractive path when we talk about resource costs but they still haven't won the trust of the companies, in terms of the information systems security.

Organizations must evaluate these three options with the purpose of finding the one that best fits their commercial needs.

Be it commercial or open source tools, deployed in a cluster, cloud or appliance, the organizations must take into consideration the costs of acquisition, implementation, maintenance, upgrading and, also, the flexibility of the architecture to integrate other tools with different natures.

Finally, the security that truly depends just as much on how well the software is deployed, configured, updated and maintained, including product vulnerabilities discovered and solved through appropriate and timely updates

3.4. Value Creation

Now, more than ever, organizations must look beyond the delivery of the best products or services. "To succeed they must uncover hidden customer, employee, vendor and partner trends and insights. Organizations need to anticipate behavior and then take proactive action and empower the team with intelligent next steps to exceed customer expectations" (Baboo & Kumar, 2013)

In his study, Manyika (Manyika et al., 2011) identified five ways to prove that Big Data has potential to create value and change the way organizations should be designed, organized and managed.

- Creating transparency

Big Data provides more available information for all stakeholders thus reducing the exploration and processing times.

Each organization can create and store more transactional data, they can access more consistent and detailed information about their performance in real-time. For that reason, by integrating different sources they are now able to launch a product in a shorter period of time and with more quality.

- Enabling experimentation to discover needs, expose variability and improve performance

With the possibility of creating and storing transactional data in a digital form, organizations will get, in near real-time, detailed data performance from everything that happens inside the organization. By automating processes, it will be possible to set up controlled experiments, analyze variability in performance, and then identify the causes. This will allow the leader to find new ways to improve the performance of the organization.

- Segmenting populations to customize actions

Organizations will be able to approach their products and services in a concept of personalization. This will only be possible because Big Data allows organizations to create more specific population segments and develop their productions/services according to the population's needs.

- Replacing/supporting human decision making with automated algorithms

More sophisticated analytical tools improve the decision making and minimize the risks. The decision does not need to be automatic, but Big Data could make this process faster and more reasoned.

This is possible because Big Data reaches a large panel of data sources that can be merged with organizational data, which allows us to get other types of analysis.

On the other hand, automatic algorithms can reduce the reaction time of the organization, when facing a new problem.

- Innovating new business models, products and services with big data

Big Data will make the creation of new products and services creation possible, improving or even creating a new business model. This could be done through the use of information about customers, consumption and product utilization.

It is clear that Big Data could be a great benefit and it will be valuable for many industries in many ways, as presented in the section below.

However, it is not enough to collect all the available data and if the companies are not capable of extracting some meaning from that information, then the entire process will be a waste of resources and time. A balance between the quantity, the associated cost and the business interest of the collected data is required, so that the return of the investment is guaranteed.

3.4.1. Big Data business opportunities

Thanks to Big Data, sectors such as health care, agriculture, education, financial and insurance services, manufacturing, retail, energy, public services and administration have revealed shown great improvements to their services and also to their own financial systems. For these sectors, the possibility of collecting profile, behavior and transactions, presents a set of valuable business opportunities for personalized services with better quality.

However, enterprises need to be cautious and ensure that this in-depth collection and mining of personal data does not result in privacy and compliance lapses.

Studies as, “Let’s Play Moneyball: 5 Industries That Should Bet on Data Analytics” (Horowitz, 2013). ,“Business opportunities: Big Data” (IDC, 2013) and “Big data: The next frontier for innovation, competition, and productivity.”(Manyika et al., 2011) challenge and prove the relevance of using Big Data technologies in some market sectors.

Manufacturing and retail constantly need new ways to increase sales, through the relative analysis of their customers, product/services quality and also the analysis of the sales where the “seller” performance is as an input. From this information, it is possible to know what the population likes, its emotions and needs. As we know, the young ages are the ones that diffuse more information through social networks and this could be seen as a new market to invest.

A product might not result in the expected sales if customers are not motivated to buy. Beyond the marketing efforts, it is important that sellers know how to encourage the client. Thus, e-commerce is growing and it is a great information source about customers, their behavior and needs. However, there are still many consumers who choose to buy directly from points of sales where they can understand if the product really satisfies their needs.

In both cases, the important thing to do is to design the target customer's profile and adapt the sales approach.

In retail, the amount of data increases more and more and retailers not only store every transaction and operation from the client, but also the new data sources, like chips that can track the product, the consumer's behavior and the online sentiment.

The agriculture sector can also do some analytical processes. Information about the tillage quantity produced and the logistics and transport costs are important factors for the final cost equation. The producers will be able to use the data to decide which location for the facilities will represent less waste of money and resources(Horowitz, 2013).

Still, the proliferation of mobile devices and the electronic payment for this kind of products will help the Government monitoring the sector activities. The knowledge created in this context can provide more information about the financial services which this sector needs. This way, the Government will be able to identify the regions in "crisis" and correctly direct the support needed, thus avoiding lack of production, depopulation and importation decrease.

Inside the sector of financial services, insurance is seriously depending on the available data. The economy and demographics are changing some parts of this industry, just like the increase in life expectancy is (StackIQ, Inc., 2012).

With more and different indicators it is possible to define a specific profile of the insured, thus personalizing their products and minimizing the risks. Here, the technologies include behavior models based on the profile of the clients, combined with other indicators which are relevant for each product, such as geography, economics and security. Also, the analysis of data from multiple channels could help the insurers manage their products, create new ones, based on the needs of the social media users, and even detect some new fraud indicators.

Another opportunity comes with the analysis of claims stored in text or even in voice format. Big Data allows us to understand the sentiment toward the product or the company and, in case of a standard between the claims or suggestions, to understand what is wrong with the offers. The identification of these problems enables a fast reaction

and avoids the loss of client interest. Finally, through the access of data in real time, the insurer will be able to predict dangerous weather conditions and to act accordingly minimizing the costs associated to the claim payments.

In the health care and education sectors, the use of different devices that create new data sources could help the analysis of information as a catalyst for knowledge in investigation activities. Concerning the education, new types of data could improve the public sector's understanding of the educational needs and flaws in the teaching-learning relation. (World Economic Forum, 2012).

Regarding medical activities, the diagnosis could be more assertive and reliable in the decision making. Today, all indicators can be tracked and processed and the analytics predictions can be taken more seriously. The medicine administration or treatment can be done in a personalized way, guided by the detailed profile of the patient and with a lower risk margin.

The Big Data can also help in the investigation and testing of new drugs. Based on every available information on certain pathologies , on the treatments already performed and on the results, it will be possible to test hypotheses before applying them to the patients (World Economic Forum, 2012).

Lastly, in the public and administration sectors the Big Data interaction relates to cost reduction and increased productivity. The variety of areas, functions and budgets leads to the need for efficient management, not necessarily meaning an intensification of the resources. Big Data will then be proficient in expanding the transparency levels and in applying advanced analytics techniques. That will offer a set of strategies and solutions to increment productivity, to reach high levels of efficiency and effectiveness, leading to a highly significant reduction of costs (Manyika et al., 2011).

3.4.2. Financial Services Industry

The Financial Service sector is one of the most data-driven industries and most of the data that exists within a bank's datacenter, such as call logs, weblogs, emails and documents, are not analyzed (Mathew, Halfon, & Khanna, 2012).

In addition, there are other relatively new sources of unstructured data which continuously spew digital exhaust and contribute to what is known as ‘Big Data’ – blogs, online news, weather, Twitter, YouTube and Facebook, to name a few.

The regulatory environment, that commercial banks and insurance companies operate from within, requires these institutions to store and analyze many years of transaction data, and the pervasiveness of electronic trading has meant that Capital Markets firms both generate and act upon hundreds of millions of market related messages every day.(Mathew et al., 2012)

The business of banking and financial management is rife amongst transactions, conducting hundreds of millions daily, each adding another row to the industry’s immense and growing ocean of data.(Turner, Michael, & Rebecca, 2013)

In order to better understand what is forcing Big Data technology adoption in Financial Services, the Oracle white paper(Mathew et al., 2012) detailed some drivers that increased the need of a Big Data architecture:

Costumer Insight – The consumer’s bank was the primary source of information to get the consumer’s identity for all financial, and many non-financial, transactions. Gaining a fuller understanding of the customer’s preferences and interests is a pre-requisite to ensure that banks can address customer satisfaction and to build more extensive and complete propensity models. Banks must therefore bring in external sources of information, information that is often unstructured. Consequently, Big Data technologies play a pivotal role in enabling customer centricity in this new reality.

Regulatory Environment - The spate of recent regulations is unprecedented for any industry. For example, Dodd-Frank alone added hundreds of new regulations which affect banking and securities industries.

Explosive Data Growth - Perhaps the most obvious driver is that financial transaction volumes are growing, leading to explosive data growth in financial services firms.

Technology Implication - Faster growth in structured and unstructured data from both internal and external sources requires better utilization of existing technologies and new technologies to acquire, organize, integrate and analyze data. Pre-defined, fixed schemas

may be too restrictive when combining data from many different sources. The faster changing needs imply that the schema changes must be always allowed. Traditional Business Intelligence systems work extremely well when the questions to be asked are known. But business analysts frequently don't know all the questions they need to ask. Finally, whether it is a front office trader or a back office customer service rep, business users demand real-time delivery of information

Looking at this, the “Business opportunities: Big Data” (IDC, 2013) presents a list of processes that can be executed with the support of a Big Data infrastructure.

- Algorithmic trading.
- Fraud prevention and detection in banking.
- Customer insights from the integration of transactional data (from CRM, credit card payments, account transactions) and unstructured social media feeds.
- Portfolio and Risk exposure assessment.
- Customer profiling, targeting, and optimization of offers for cross-selling.
- Influencer analysis.
- Customer center and call center efficiency.
- Sentiment analysis and brand reputation.
- Correlation of social media sentiment with stock returns to support investment decisions.
- Underwriting and loss modeling.
- Real-Time stock exchanged analysis and prediction.
- Risk prediction.

Financial services firms are leveraging Big Data to transform their processes, their organizations and soon, the entire industry. Big data is especially promising and differentiating for financial services companies. With no physical products to manufacture data, the source of information became one of their most important assets.

For consulting companies the interest of the industries means new business opportunities. It is up to them to take a step further, understand their customers' needs, and also be prepared to change their strategies.

3.5. Text Mining

By now, we have realized that Big Data is known for the velocity in which it collects and processes large amounts of data of several types. “Approximately 80 percent of the corporate data is in an unstructured format. The information retrieval from unstructured text is very complex as it contains massive information which requires specific processing methods and algorithms to extract useful patterns” (Sumathy & Chidambaram, 2013).

For that reason, it is very important to be aware of the existing methods and technologies and to be able to create an interesting insight for the organization.

This chapter provides an overview of this knowledge discovery engine and explains all that is necessary to do on the text processing.

3.5.1. Concept

Text mining which is also known as text data mining or knowledge discovery from textual databases, is the process of discovering hidden useful and interesting patterns from unstructured data (Tan, 2000).

This concept is very similar to the concept of data mining which tries to find patterns and trends on large relational databases and enables the automation of the analysis of text data and it could be used directly into production systems.

Each process incorporates data mining, web mining, information retrieval, information extraction, computational linguistics and natural language processing (Sumathy & Chidambaram, 2013).

3.5.2. Text Mining Process Framework

According to Ah-Hwee Tan, text mining can be visualized as consisting of two phases, text refining and knowledge distillation (Tan, 2000)

Text Refining - Transforms free-form text documents into a chosen intermediate form.

Knowledge distillation - Deduces patterns on knowledge from the intermediate form (IF).

- The intermediate form can be semi-structured like a conceptual graph representation, or structured like a relational data representation.
- The IF can be document-based, where each entity represents a document, or concept-based where each entity represents an object or concept of interests in a specific domain.
- Mining a document-based IF deduces patterns and relationships across documents.
- Mining a concept-based IF derives patterns and relationships across objects or concepts.

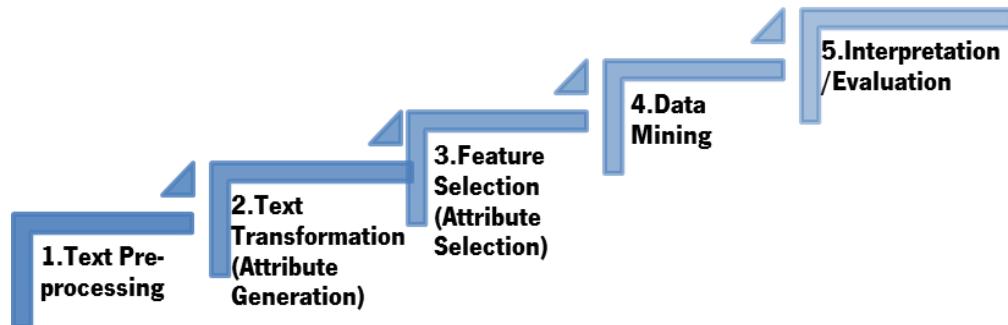


Figure 16 - Text Mining Process Tasks adapted from (Tan, 2000)

1. Text Pre-Processing:

Text Cleanup - The removal of any unnecessary or unwanted information, such as removing ads from web pages(html code), normalizing converted text from binary formats or dealing with tables, figures and formulas.

Tokenization - Achieved by splitting the text into white spaces and punctuation marks which do not belong to the abbreviations identified in the preceding step.

Parts of Speech Tagging (POS) - Tagging means assigning word class to each token. Its input is given by the tokenized text. Taggers have to manage unknown words and ambiguous word-tag mappings. Rule-based approaches like ENGTWOL operate on:

- a) Dictionaries containing word forms together with the associated POS labels and morphological and syntactic features.

b) Context sensitive rules to choose the appropriate labels during application.

- 2. Text Transformation (Attribute Generation)** - A text document is represented by the words (features) it contains and their occurrences. Two main approaches of document representation are Bag of words and Vector Space.
- 3. Feature Selection (Attribute Selection):** Is the process of selecting a subset of important features for use in model creation. The main assumption when using a feature selection technique is that the data contain many redundant or irrelevant features.
- 4. Data Mining** - At this point the Text mining process merges with the traditional Data Mining process. Classic Data Mining techniques are used in the structured database which resulted from the previous stages. Using text data to add news features do structured data increases the power of the data mining techniques.
- 5. Evaluation** – To evaluate the result after the evaluation, the result can be discarded or the generated result can be used as an input for the next set of sequence.

3.6. News Articles Influence on Stock Market

In order to explore one of the Big Data's applications in Financial Services Industry, this project has its focus on the correlation between daily news articles and market variations. It is necessary to be less reactive and more preventive in relation to the market since a brutal variation of its conditions could be very harmful for the companies. The best way to know what goes on around is through news articles.

This chapter presents an overview of the master thesis "Text Mining of News Articles for Stock Price Predictions" that served as a base for the practical use case of the topic that will be detailed in the development chapter.

3.6.1. Concept

Social media can have an enormous influence on society and markets. News articles written about companies serve the purpose of spreading information about the companies, and this information then influence people either consciously or unconsciously in their decision process when trading in the stock market (Aase, 2011).

In an abstract perspective, general subjects like politics, health, governance, war or even climacterics changes or tragedies could also bring some impact for the different industries.

The information shared almost instantly by the social media (newspapers online, blogs, social networking) is capable of determining the decisions of the stock traders and the entire market. This impact it will be bigger when a given information is unexpected by the companies and investors.

It is important to analyze this information as fast as possible so it can be used as help for trading decisions by traders before the market has had time to adjust itself to the new information.

According to Kim-Georg Aase (Aase, 2011) it is possible to predict stock price changes after news articles publications. When the stock trading is done from signals generated from sentiment analysis of news articles, then the profit is better compared to what a random trader gives. A training set of news articles for the sentiment classifier might be automatically created and labeled by looking at how the price for the related company changes after the article is published.

3.6.2. Approach Overview

This study provides a workflow for an automatic news based trading approach (Figure 17). The workflow contains five main processes: data acquisition, new sentiment labelling by price trends, preparation of dataset, label refining, classifying and training.

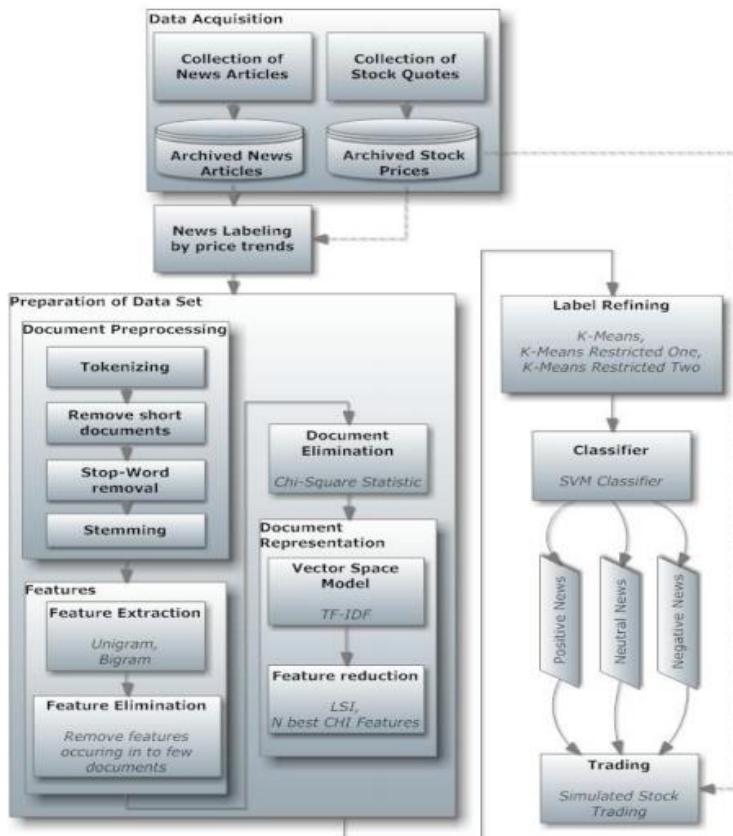


Figure 17 - Workflow of the proposed automatic news based trading approach from (Aase, 2011)

Data Acquisition: The acquisition of news articles from the web page Netfonds that gathers news articles from several financial news sites. It is also the collected stock prices from the Oslo Stock Exchange (OSE).

News Sentiment Labelling by price trends: News sentiment labelling as positive (uptrend), neutral (flat) or negative (downtrend). This step is performed before any news article documents are prepared, that is, this step only needs the news article's id, publication date and time and the stock prices, not the content of the news.

Preparation of DataSet

- Document: Preprocessing: it is the news processing through the use of tokenization, stop-word removal and the removal of documents with short content.
- Features: to extract features (i.e. words) from the news documents and remove the features which occur in a number of times smaller than the given number of documents.

Document Elimination: To eliminate documents that is not used for the classifier. Only the documents that contain at least a significant feature to distinguish between the three categories are kept.

Document Representation: It represents documents through a vector space model to define the feature's relevance.

Label Refining: The basic method (using price trend directions) is very likely to result in a labeled dataset with a significant degree of noise. A document with a given label has more in common with other document with the same labels than it has with documents with different labels. Clustering method should fix the noise by rearranging the documents in these three groups of labeled documents.

Classifier Learning: The classifier method for the learning algorithms is the Support Vector Machine classifier (SVM) that is used to find the relationship between the contents of news and price trends. K-Means is one of the simplest learning algorithms that is refined by K-Means Restricted to assure some rules of documents "movements" trough clusters

4. Towards Big Data project

As much as any other technology, choosing a Big Data supplier and considering the Big Data complexity of their “issue”, are two critical tasks for any organization. Nevertheless the most important thing to do is to ask ourselves: “Is This a Big Data problem?”; “Do I really need these Big Data tools?”

Most organizations, especially European, are not ready for Big Data in a technical and business perspective and don't see significant “value” in Big Data investments (IDC, 2013).

A survey made by International Data Corporation (IDC) revealed that in a short percentage (in Europe) the awareness and readiness were higher among very large companies but even those are still a minority. Also “short-term intentions to increase usage of Big Data technologies were concentrated mainly among very large companies, and in sectors that were already at the vanguard of adoption” (IDC, 2013). Moreover they conclude that in most of the cases, Big Data is used as an extension of what already exists.

For Enterprises, Big Data means that they need to acquire new resources that have some expertise not only in development, implementation and system integration but also specialization on statistic and analytic processes.

Many IT executives see open source Big Data technologies, such as Hadoop, as immature or unstable and carrying significant security and retooling risks when compared to proprietary tools.

Big Data technologies enable detailed tracking and analysis of consumer profiles and behaviors, from non-traditional data sources such as social networking sites, mobile device applications, and sensors. This generates valuable business opportunities for more targeted/personalized services and cross selling.

Considering this, we already know that Big Data is defined by 3v's: Volume, Velocity and Variety.

Could volume represent all Big Data concept? We only call it Big Data when data meets all those features? Where can we set the frontier between BI and Big Data?

Those are the questions that companies need to think about because the possibility of solving current problems with known, cheaper and easier (implementation and utilization) Business Intelligence tools may exclude all Big Data options.

On the other hand, being resistant to change may lead the organization into a constant investment on infrastructure with a short scalability or an unsatisfactory return. Identifying what kind of data needs a Big Data solution could be the first step to support the decision with an absolute measure.

4.1. Big Data vs Business Intelligence

Business Intelligence (BI) is described as a set of technologies, applications and tools that create knowledge and help in the decision making. It also includes a life cycle that goes from data acquisition to analysis of results (Maier et al., 2013).

This discussion has the purpose of understanding where BI solutions are different from Big Data solutions and why traditional tools are not enough to deal with Big Data. The approach to this evaluation is inspired on the layers of a BI architecture (Figure 18) where each technology is described according to common dimensions.

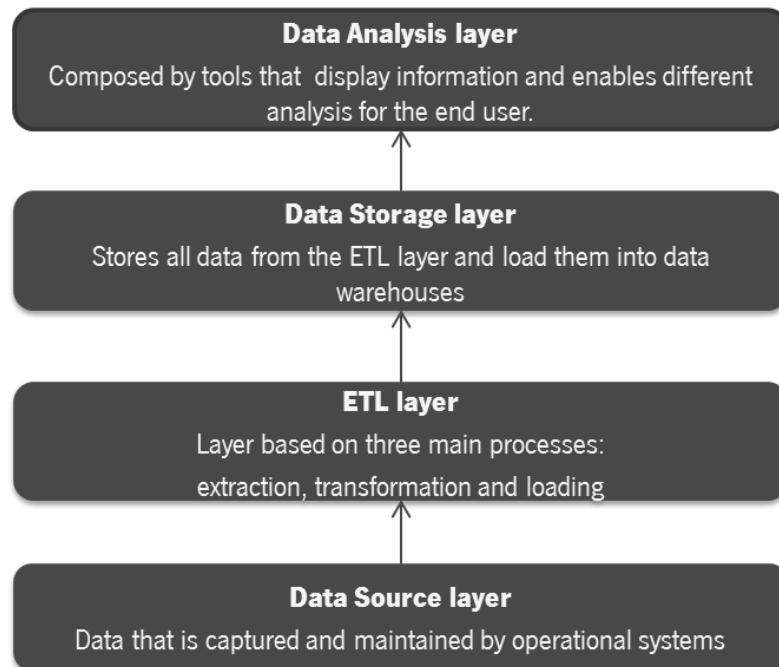


Figure 18 - BI Architecture Layers

According to the architecture (Figure 18) the following tables present the difference between both architectures layer by layer

In the first layer (Table II) both architectures are compared by source, format, volume, and process dimensions. Here Big Data stands out by embracing more data sources, types and volume showing different capabilities to process all the information.

Table II - Data Source Layer

BI	Dimension	Big Data
<ul style="list-style-type: none"> • ERP • CRM • Tables • Files • Web Site Traffic • Core systems 	Source	<ul style="list-style-type: none"> • Image and Video • Docs, txt and XML • Real time streaming (web logs, sensors and devices, Events) • Social Graph feeds • Spatial, GPS • DB Data • DW
<ul style="list-style-type: none"> • Structured 	Format	<ul style="list-style-type: none"> • Structured • Semi-Structured • Unstructured
<ul style="list-style-type: none"> • Gigabytes to Terabytes 	Volume	<ul style="list-style-type: none"> • Terabytes to Petabytes
<ul style="list-style-type: none"> • Batch 	Process	Batch, Real-time or near real-time and Streaming
<ul style="list-style-type: none"> • Traditional ETL tools and hand code scripts 	Data Source Connectors	<ul style="list-style-type: none"> • Social listening tools • HDFS tools • Sqoop • Apache Flume (real-time streaming) • Traditional ETL tools and hand code scripts

In the next layer (Table III) the architectures are reviewed according to the dimensions, transformation and storages mechanisms, process, and schema. The main difference of this table is that Big Data not only processes an Extract-Transform-Load process (ETL) but an Extract-Load-Transform-Load (ELTL). This happens because Big Data databases can store data with no-schema, raw data is stored in the same format that is collected. If necessary, the data is then transformed and stored again into non-structured or structured database systems.

Table III - ETL Layer

BI	Dimension	Big Data
<ul style="list-style-type: none"> Extract data from staging area storage (RDBMS); Transformation and integration of data to fit business goals Load into a single target Database (DW, DM, etc.) 	Transformation and Storage mechanism	<ul style="list-style-type: none"> Extract data from different types of sources Load to database systems such NoSQL and HDFS Transform reduce and filter the Big Data Load the transform results into unstructured and structured databases
<ul style="list-style-type: none"> Centralized and batch 	Process	<ul style="list-style-type: none"> Distributed and batch real-time or near real-time
<ul style="list-style-type: none"> Analytics can be limited as the data model and domain is predefined 	Schema	<ul style="list-style-type: none"> Open scope and flexibility for analytics as the data model is flexible

Data Analysis Layer (Table IV) is more focused on the analytic features and approaches. The fact that Big Data provides a Discovery Analysis changes the way organizations get insight from the collected data. Another difference is the chosen approach that, in the case of Big Data, allows a Bottom up where the users can perform analysis and decision making in real-time

Table IV - Data Analysis Layer

BI	Dimension	Big Data
<ul style="list-style-type: none"> Based on descriptive Analytics 	Analyze	<ul style="list-style-type: none"> Discovery Analysis (direct to native data sources-HDFS; NoSQL)
<ul style="list-style-type: none"> Batch 	Analytic Process	<ul style="list-style-type: none"> Batch, Real-time or near real-time
<ul style="list-style-type: none"> Top-down putting data in the hand of executives and managers who are looking to track their business on the big-picture (Vinnakota, 2012) 	Approach	<ul style="list-style-type: none"> Bottom-up- When properly harnessed, it empowers business end user in the trench, enabling them to carry out in-depth analysis to inform real-time decision-making (Vinnakota, 2012)

The last layer (Table V) compares both architectures looking at security, enterprise capability to deal with implementing those technologies and when each one is needed.

Table V - Technology Landscape Analysis adapted from Deloitte Resources (Deloitte, 2014)

BI	Dimension	Big Data
<ul style="list-style-type: none"> Highly secure with established methods and processes for processing and storage 	Security	Security can be challenging to enforce at granular level
<ul style="list-style-type: none"> Proven track record of enterprise readiness 	Enterprise-Ready	Non-relational stores have been widely adopted in internet based companies and thus were designed to be web friendly
Best when system needs : <ul style="list-style-type: none"> Transaction Processing (OLTP) Analytics on structured data Enterprise level Service Level Agreements (SLAs) 	When?	Best when system needs : <ul style="list-style-type: none"> High Availability and scalability Offline reporting with large datasets Unstructured Data Processing

Despite all these different dimensions in comparison, the main difference lies in the use, or not, of Data Warehouse (DW).

A Data Warehouse is a software tool responsible for integrating, organizing, preparing and storing data to support and optimize the analytic applications.

With the Big Data appearance, the DW begins to show some weakness. The problem appeared because, until now, the DW was constructed on RDBMS, and systems cannot deal with these new types of data.

Therefore data warehousing as we know it could disappear as Big Data solutions enter the markets.

For some authors such as Dijcks (Dijcks, 2012), the reengineering of DW and inclusion of components that help the communication between organization data and Big Data could be the solution that mitigates the DW limitations.

A study made by Wikibon, Financial Comparison of Big Data MPP solution and Data Warehouse Appliance (Foyer, 2013) suggests that DW projects are more expensive, need more time to be developed and have a less attractive return of investment (ROI).

However, because it is a recent topic and it makes use of recent technologies, Big Data is less “mature” and represents a major risk on the investment.

Besides a financial point of view, this study presents a comparison between both approaches looking at processes and activities.

A DW project requires a major effort on data definition and its transference between every system from the distributed network.

In many cases, the data sources are incomplete, they don't use the same description and the availability is not always total. Under those conditions, copying and organizing the data on centralized system is extremely hard. Building and providing a visualization and analysis experience for the client is a complex process which increases time and resources.

The opposite happens in a Big Data project where data is extracted from several sources using traditional and recent tools and stored in non-relational data bases.

Since the process is done locally where the data is allocated, it provides a load and analysis faster than in DW context. Here the services of analytics experiences specific of the client are used as an interactive part of the process.

In conclusion, Big Data seems to bring more advantages over Data Warehouses but this does not mean that we must substitute the data warehousing process for the new solution. In fact there are more advantages when we mix both “technologies”. They can work independently and, at the same time, cooperate with each other.

It is up to organizations to decide what kind of approach to follow and which better satisfies their needs.

4.2. Big Data Complexity Framework

This project presents a proposal for a “Big Data Complexity framework” as a way to reach an answer for the main question in this section. The framework aims to help find a cluster that better defines their “issue” according to the dimensions of the 3v’s and the corresponding complexity level. It is also the base for the calculation that will position the “issue” in a Big Data spectrum.

The result will be an absolute measure that could simplify the decision making and help the organizations to better support the decision making.

The build framework (Table VI) assumes a form of matrix with two main features:

Complexity Level (CL) – Defines a scale of complexity for each dimension.

Five levels were defined for this framework, ranging from the less complex dimension (CL1) to the most complex and harder in the scale (CL5).

In other words, the more data is collected and processed the complexity increases, in a shorter period of time with complex types of data.

Contrarily, simplest types of data with large periods of time to collect and process “small” amounts of data describe projects that are well known by organizations that already implement BI architectures.

The complexity levels of each dimension were distributed according to the knowledge about the existent technologies and projects about each theme.

Dimension – According to earlier investigations, Big Data is defined by three dimensions which are related to each other:

- Volume: Volume of data that is collected in a specific period of time.
- Velocity: Period of time that the data is collected/refreshed processed and analyzed.
- Variety: Types of collected data which are directly related to the complexity of process and analysis.

The combination of the three dimensions with the respective complexity level defines a Big Data issue.

As mentioned above, we must take into account that the volume is not the volume of the existing data in the organization (historical data) but the amount of data that is currently collected.

Table VI - Big Data Complexity framework

Dimension/CL	CL1	CL2	CL3	CL4	CL5
Volume	<1000GB	5TB-50TB	50TB-500TB	500TB-000TB	>2PB
Velocity	Batch	Intra-day	Hourly-refresh	Real-time	Streaming
Variety	Structured Data	Docs: XML; TXT, JSON	Web-log; sensors and device events	Image; social graph feeds; geospatial information	Video; Voice

In order to avoid an inconclusive definition of the issue, we tried to provide an absolute measure for the combination of the three dimensions in its respective CL. The proposal consists on the sum of each product between dimension and CL value (1-5).

Because some dimensions are more important than others, we also defined weights for each one and the volume has the higher weight.

The matrix features and weight distribution were based on the earlier literature review and on the opinion of Deloitte specialists from the Analytics technologies.

The features of the matrix were defined according to the research which made possible to find what was already done for the data processing activities. Processing 1000GB of structured data in a batch procedure is something that is executed every day and it does not represent difficulties. Nevertheless, the difficulty grows as the volume increases and the several types of data need to be processed faster than ever.

With respect to the weights, we agreed that what is important is not the exact value of each variable, but the scale that separates each one.

In our view, this is a possible distribution, however, each organization should adapt and attribute the weights according to the size and capacity of the current infrastructure.

We believe that the volume is not the only characteristic of Big Data but the most challenging one and it is the one that is more easily recognized from the technology's or infrastructure's perspective. The volume is more significant in the choice between Big Data technology and the decision of invest in the increase of the space and process capabilities.

Variety comes next, since it represents the challenges for the organizations that have to deal with data types which were never processed, in order to provide value for them. Here we talk about learning new kinds of technologies that "understand" unstructured data and acquire knowledge from data that seems "meaningless".

Finally, Velocity presents the minor weight because we believe that velocity challenge is solved with the increment of process capabilities, like parallelism and hardware investments.

Once the weights are distributed it is possible to perform the calculation and map the problem.

The equation is a simple sum of the Big Data dimensions, each one multiplied by the CL.

Table VII - Big Data Complexity framework (calculation)

Dimension/CL	CL1 (1)	CL2 (2)	CL3 (3)	CL4 (4)	CL5 (5)
Volume (10)					
Velocity (1)					$CL = (WVolume * CLx) + (WVelocity * CLx) + (WVariety * CLx)$
Variety (5)					

[100-200] – Traditional BI Issue.

[200-300] – BI Issue near Big Data challenge

[300-400] – Big Data Issue

[400-500] – Complex Big Data Issue

Traditional BI issue – The Company is facing a problem that could be solved with a relative investment on storage capability and/or simple text processing tools. The period of data refresh and process is not a threat.

BI Issue near Big Data challenge - Defines a problem that might evolve to a Big Data issue. It could be solved, for now, through some advanced analytics tools and a system capable of scheduling tasks (Intra-day, Hourly-refresh). Problems that fit in this class must complement its analysis with expected evaluation and consider the need to advance from the beginning to a Big Data project.

Big Data Issue - Need for investment in Big Data architecture, through a comfortable process skill. Once we reach a Big Data problem, its characteristics (Volume, Velocity and Variety) are not a big issue because Big Data tools are prepared for these conditions.

Complex Big Data Issue – In this case, it is even more urgent to invest in a Big Data project. There is no possibility for the BI to be enough to support such a huge and instantaneous data flow with this complexity. Even the Big Data suppliers have to prove that they are capable of dealing with this problem because not all offers presented on the market will fully serve the customer's needs.

4.3. Having the best of two worlds

It is true that both projects are limited when they have to perform the opposite task, i.e., Big Data technologies are not ideal to deal with the structures and traditional volumes and the BI/Data Warehousing technologies are not even capable of dealing with large amounts of unstructured data.

Nonetheless, if Big Data architecture is capable of integrating and extending the traditional BI for the analysis of more complex data, it will provide new capabilities:

- The conventional BI for the mainstream business allows the users to do ad hoc queries and reports. Now, it is possible to complement that effort with a Big Data analytics environment, optimized to handle a flow of unstructured data.
- BI tools help the analysts to filter and analyze according to the business requirement.
- Big Data can operate from the business requirements and the context of the problem.
- The Existing ETL process for loading data warehouse can be improved by the power of Hadoop.

As shown in the architecture (Figure 19), there are many multiple connections between the Business intelligence architectures and the new Big Data components.

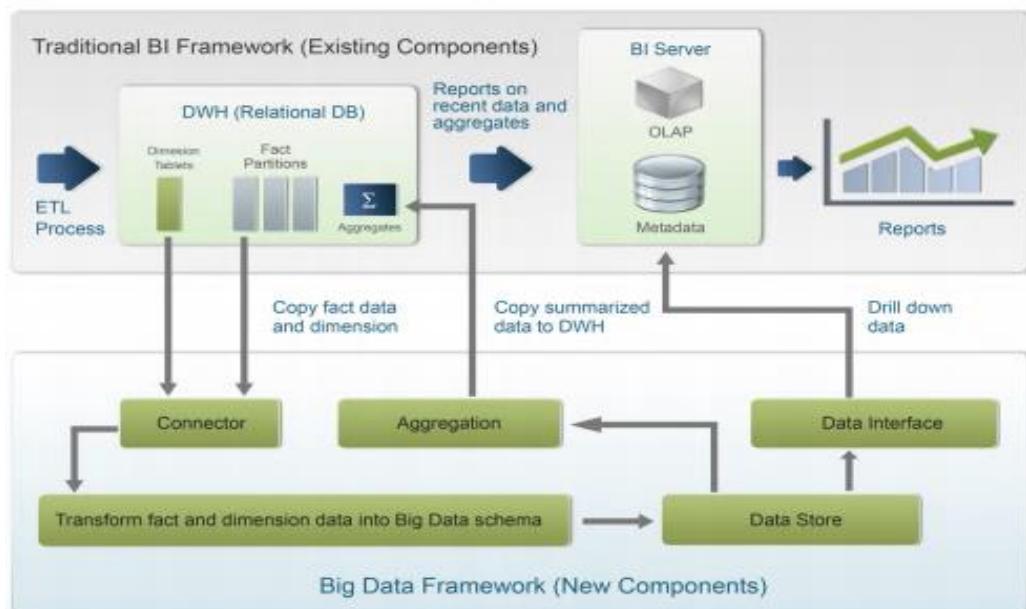


Figure 19 - Enhanced BI architecture from(Persistent Systems, 2013)

The paper of the Persistent Systems (Persistent Systems, 2013) details some data suggestions which can be done and get the maximum benefit of the integration of both architectures :

- Adding a Big Data infrastructure to support Data Warehouse (DW) which enables the transaction of the enterprise data to the Big Data tables, with connectors augmenting the performance of relational data bases which no longer need to save all the historical data.
- Changing BI metadata to work with Big Data allows the reports to switch between DW and Big Data. Also, Big Data tools can perform some sql instructions and present some offline reports (even though less diversified).
- Summarizing Big Data and the copy to DW. In other words, Big Data can be processed, extracted and transformed and the result of this process (summarized data) could be stored into DW for reporting purposes, increasing the performance of the reporting process.

5. Big Data Solutions – Comparative Summary

Once the investment in a Big Data project is approved and finally gathered the information about the market offers, organizations must choose the best solutions that better fits their needs. The large number of options makes it hard for organizations to decide what is best and why. This happens because each Big Data supplier sells its solution as the best one, the most feasible, robust and scalable, amongst other features, and the organizations cannot base their decisions solely on the kind type of characteristics.

Nowadays, sharing information about it is the best “counseling”. Tools trial reports, vendor surveys and white papers, benchmark reports from specialists and the opinion of non-professionals (from the social media), these are the best sources to sustain our decisions. More available information simplifies the choice.

Therefore, the following chapter presents a short summary (Table VIII) and an objective comparison of the most popular vendors.

Table VIII - Big Vendor Solutions - Summary

Key Vendors	Offers	Hardware Appliance	Connectors
EMC Greenplum	<ul style="list-style-type: none"> • MPP Database • Hadoop distribution • Chorus – search, explore, visualize, analyze • Command Center 	<ul style="list-style-type: none"> • Optional- Greenplum Data Computing • Appliance (DCA)- single rack expandable in quarter rack increments up to 12 racks) • Optional – gNet connector 	<ul style="list-style-type: none"> • Connects to most traditional EDWs and BI / analytics applications (SAS, MicroStrategy, Pentaho)
IBM Netezza	<ul style="list-style-type: none"> • IBM Netezza DW Appliance • IBM Netezza Analytics 	<ul style="list-style-type: none"> • S-Blade servers contain multi-core Intel CPUs and IBM Netezza's unique multi-engineFPGAs. 	<ul style="list-style-type: none"> • Data Integration with most IBM and 3rd party solutions. • Working with Cloudera to bring in

Key Vendors	Offers	Hardware Appliance	Connectors
		<ul style="list-style-type: none"> • Configuration in single and multiple racks 	Hadoop connectivity
Oracle Exadata	<ul style="list-style-type: none"> • Oracle Exadata Database Machine • InfiniBand High Speed Connectivity • Oracle Data Integrator 	<ul style="list-style-type: none"> • Exadata Storage Server X2-2 • Exadata Database Machine X2-2 • Exadata Storage Expansion Racks • Exadata X2-2 Memory Expansion 	<ul style="list-style-type: none"> • Data Integration with Hadoop, NoSQL and other relational database sources. Special Integration to R. Connectivity to 3rd party solutions. include Cloudera Hadoop distribution and management software
Teradata	<ul style="list-style-type: none"> • Aster Database 5.0 with SQL-MapReduce • Teradata Database 14 • Hadoop Integrator 	<ul style="list-style-type: none"> • Aster MapReduce Appliance • Active Enterprise Data Warehouse • Extreme Performance Appliance • Data Warehouse Appliance • Extreme Data Appliance • Data Mart Appliance 	Data Integration with Hadoop, NoSQL and other relational database sources. Special Integration to SAS. Connectivity to 3rd party solutions.
Cloudera	<ul style="list-style-type: none"> • Hadoop platform distribution (CDH) • Data Integrator • Automated Cluster Management 	<ul style="list-style-type: none"> • CDH (Cloudera's Distribution including Apache Hadoop) • Cloudera Express 	<ul style="list-style-type: none"> • Data Integration with Hadoop, NoSQL and other relational database sources.

Key Vendors	Offers	Hardware Appliance	Connectors
	<ul style="list-style-type: none"> • Search, explore, visualize and analyze engines • Web applications that enable you to interact with a CDH cluster(HUE) 	<ul style="list-style-type: none"> • Cloudera Ent 	<ul style="list-style-type: none"> • Connectivity to 3rd party solutions. • Connectors for Netezza, Teradata,Tableau, Microstrategy

Table IX - Big Data Commercial Evaluation

	IBM	Oracle	SAP	Teradata	Cloudera
Data Integration	✓	✓	✓	✓	✓
Data Visualization	✓			✓	
Big Data Analytics	✓			✓	
Interactive Search	✓	✓	✓	✓	
Text Analytics	✓		✓		
Real Time	✓	✓	✓	✓	✓
Batch	✓	✓	✓	✓	✓

6. Development

6.1. News Articles Influence on Stock Market (Use Case)

After the extended overview about Big Data it makes sense explore this Big Data paradigm in practical perspective.

To provide knowledge enrichment it was felt the need of experiment some technologies and concepts such as text mining. Although, the project was support by Deloitte Consultores SA, the target for the study was more generic, because the access to the Financial Industry transactional data was considered limited. The intention was to use the available in the internet so we can experiment some of the concepts above.

For that reason it was decided develop a workflow that will analyze news articles and their impact over the stock market. The idea is to follow some of the opportunities suggested for the industries apply some of the insight from the earlier investigation.

News Articles Influence on Stock Market was the chosen scope and is based in the Kim-Georg Aase study described above in the section 3.6. The main purpose is to find patterns between the financial news articles and the stock market variations

In this chapter it is firstly provided the used architecture, workflow execution and reached results based on the methodology for the process of Big Data analysis

6.1.1. Architecture

In order to support this use case, it was defined an architecture based on the literature review and the analysis of the existent tools. These are the used tools that composed the architecture (Figure 20).

The choice of the tools was based on the comparative study done earlier. Still, this choice was mainly “forced” by the possibility of integration with other tools or free deployments.



CDH integration with Hadoop. This tool represented a no structured data store through Hbase.



TDExpress14 performed as data store for structured data.



Open source tool used to perform Web Crawler, Data integrator, Data Mining and Reporting.



Plugin for Data Mining installed on Knime



Business Discovery Platform used to perform some analytic reports.

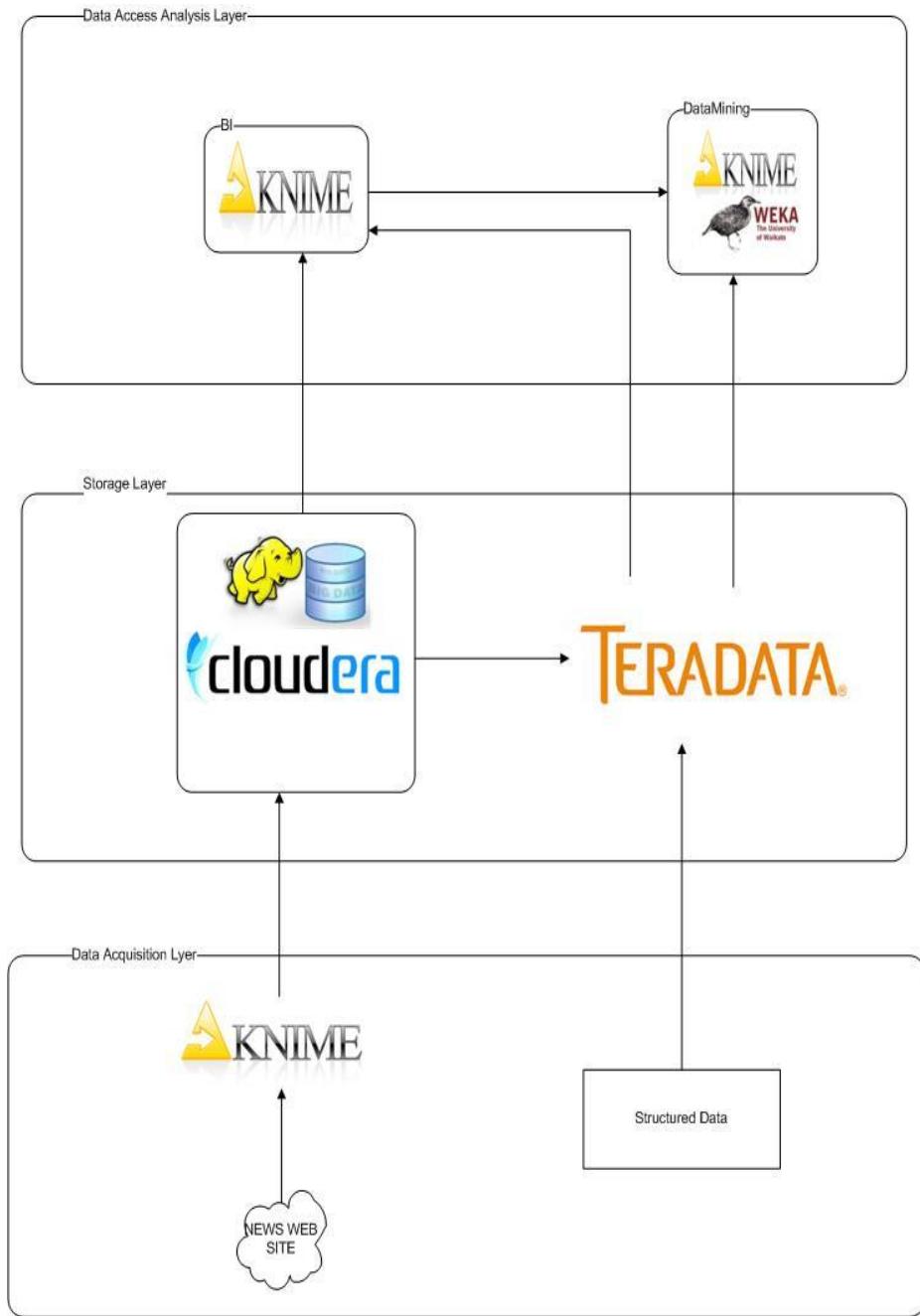


Figure 20 - Use Case Architecture

According to the methodology earlier explained (Figure 1) and the use case architectures (Figure 20), the development of our use case will be represented phase by phase (Table X).

The goals for this use case are as follows:

Table X - Big Data Architecture Phases, Goals and Expected Results

Phase	Goals	Expected Results
Data Acquisition	<ul style="list-style-type: none"> Crawl news web sites; Collect Stock Exchange history 	<ul style="list-style-type: none"> News articles stored into Hbase Stock Exchange History stored into Teradata Database
Extraction Cleaning	<ul style="list-style-type: none"> Extract interest content from the news articles Tag keywords for text classification 	<ul style="list-style-type: none"> Processed documents classified as local/foreign, company/generic subject, positive/negative/neutral sentiment
Annotation	<ul style="list-style-type: none"> Create Dictionary of keywords Classify text according defined categories(Market, Company, Sentiment) 	
Integration Aggregation	<ul style="list-style-type: none"> Integration of processed documents (unstructured) with Stock Exchange history 	<ul style="list-style-type: none"> Dataset stored into Teradata Database joined by date of publication of the news
Representation		
Analysis Model	<ul style="list-style-type: none"> Execution of Data Mining Classification Algorithms Models' evaluation 	<ul style="list-style-type: none"> Results of trained classification models Performance Measure
Interpretation	<ul style="list-style-type: none"> Choose ideal model Perform data analysis 	<ul style="list-style-type: none"> Chosen model presentation Analysis Report Conclusion

6.1.2. Data Acquisition

Taking into account the business requirement, it is to be predicted stock variation based on news articles. The data sources, filters and technique to use were defined.

Our unstructured sources are the economics and general news web sites such as the ones extracted using the Web Crawl technique:

- <http://www.reuters.com/>
- <http://www.washingtonpost.com/>
- <http://www.businessweek.com/>

The news websites mentioned above were chosen by their popularity combined with their releases, economics and general subjects.

The structured source, also from the web, was the NYSE Stock Exchange which provides the stock value history from every company quoted on the stock exchange.

Because we decided to do a specific analysis, we chose the USA market, Bank and Energy Industries.

a) Workflow

The data acquisition workflow (Figure 21) shows the data collection process. In the first flow is used the HttpRetriever that collects data from web via http. In the second it is used the WebSearcher node that collects data according to Google's news API (in this case). This node not only pulls out the HTML code but it collects the title, date, URL and summary of a news article. After that, in both cases the HTML is parsed into XML and stored into HBASE.

In "HbaseWriter" node we can define different schemas for the same table.

In the case of the historical stock exchange, data was extracted and manually imported into Teradata database.

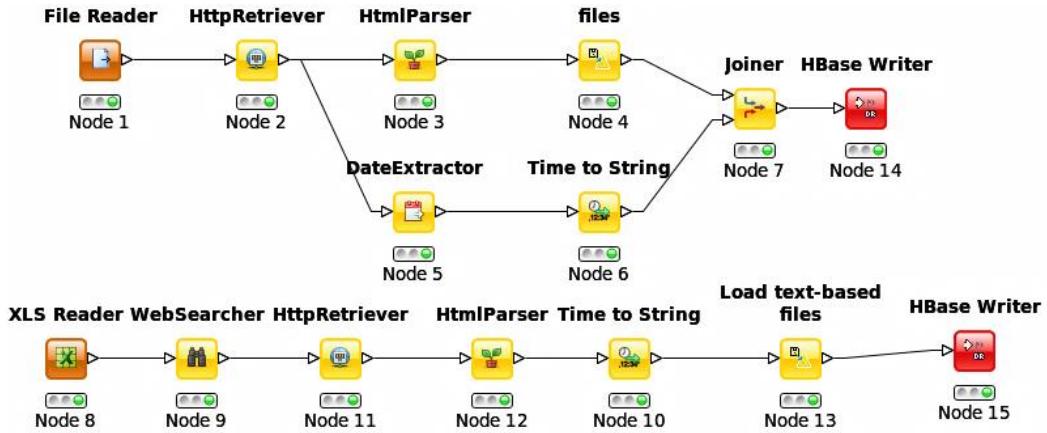


Figure 21 - Data acquisition Workflow

The result part of this workflow data stored into Hbase (Figure 22) is column-oriented and presents the different schemas defined on Knime. The attributes in the column-family File of News_Crawl table are as follows:

Summary: Summary of the article retrieved according to Google's API criteria.

Date: Date of publication.

HTML: Html code from the web page collected.

Title: Title of the news article

URL: Source of the article

Row0				
FILE: Summary	FILE: Date	FILE: HTML	FILE: Title	FILE: URL
In its latest bid to compete with online king Amazon .com, Wal-Mart is rebuilding its website to make shoppers' experience simpler, faster and mo...	2014-08-04	<!DOCTYPE html><html><head> <meta name="viewport-uuid" content="2259bb3a-1b4-11e4-9b6c-12e30cbef86a3" /> <meta http-equiv="X-UA-Compatible" co...	A glance at new changes at Walmart.com	http://www.washingtonpost.com/business/a-glance-at-new-changes-at-walmartcom/2014/08/04/2259bb3a-1b4-11e4-9b6c-12e30cbef86a3_story.html
Row0_Row0				
FILE: Date	FILE: HTML	FILE: URL		
2014-07-29	<!DOCTYPE html><!--[if lt IE 7]> <html class="ie6 no-js light_layout" lang="en" > <![endif]-->!--[if IE 7]> <html class="ie7 no-	http://www.businessweek.com/articles/2014-07-29/mit-researchers-have-made-a-thing-that-reads#r=hp-list		

Figure 22 - Hbase dataset

At the same time, structured data was imported from csv files with the stock exchange history and has the following attributes (Figure 23):

ID: Key of the company.

Symbol: Symbol of the company.

DateSE: Date of the stock exchange release.

OpenSE: Daily opening value of the stock exchange.

HighSE: Highest stock value of the day.

LowSE: Lowest stock value of the day.

CloseSE: Daily closing value of the stock exchange.

VolumeSE: Daily total volume of stock

	ID	Symbol	DateSE	OpenSE	HighSE	LowSE	CloseSE	VolumeSE
1	74.308	ZB-F	2014-08-07	28,0500	28,0500	27,7100	27,7100	1.500
2	74.287	XNY	2014-08-07	0,8300	0,8500	0,8000	0,8200	21.500
3	74.247	WPT	2014-08-07	18,5500	18,5900	18,1000	18,2100	35.600
4	74.226	WIW	2014-08-07	12,0600	12,1400	12,0600	12,0900	80.300
5	74.165	VRS	2014-08-07	3,2000	3,2900	3,1600	3,2900	193.100
6	74.144	VLT	2014-08-07	16,1100	16,1100	16,0100	16,0800	8.600
7	74.104	USM	2014-08-07	36,4700	37,1000	36,2300	36,8100	182.300
8	74.083	UMH	2014-08-07	9,9900	10,0200	9,9400	10,0200	48.600
9	74.043	TVC	2014-08-07	24,0500	24,0500	23,9000	23,9600	13.300
10	74.022	TS	2014-08-07	42,6500	42,7600	41,8800	42,0300	973.300
11	73.961	TIME	2014-08-07	23,7500	23,8700	23,5000	23,7200	831.100
12	73.900	TAP	2014-08-07	71,4900	71,5000	69,5500	69,7000	1.742.300
13	73.879	SXCP	2014-08-07	31,4500	31,5100	31,2000	31,3900	6.900
14	73.839	STJ	2014-08-07	65,2800	65,2800	63,3400	63,4400	2.588.200
15	73.818	SSW-D	2014-08-07	26,8000	26,8000	26,5400	26,6000	25.000
16	73.778	SNV	2014-08-07	22,9100	22,9800	22,5400	22,6100	1.482.900
17	73.757	SLH	2014-08-07	64,3100	64,3300	63,7800	64,0100	181.200
18	73.696	SCE-H	2014-08-07	25,8200	25,8200	25,5100	25,6000	96.500
19	73.675	SAN-F	2014-08-07	.050.0000	.050.0000	.050.0000	.050.0000	3.000
20	73.635	RPT-D	2014-08-07	61,5500	61,5500	61,5500	61,5500	100
21	73.614	RMD	2014-08-07	49,8300	50,3500	49,6800	49,9700	963.400
22	73.574	RES	2014-08-07	21,5500	21,7200	21,3500	21,5400	730.000
23	73.553	RBS-P	2014-08-07	23,6800	23,7500	23,6500	23,6700	22.600
24	73.492	PSB-U	2014-08-07	22,8000	23,0300	22,8000	22,9900	18.500
25	73.431	PMT	2014-08-07	21,8000	22,1000	21,8000	21,9500	941.400
26	73.410	PKD	2014-08-07	6,5800	6,6100	6,3100	6,3200	992.100
27	73.370	PES	2014-08-07	15,0300	15,0500	14,4700	14,6400	877.600
28	73.349	PCI	2014-08-07	22,6200	22,6700	22,5000	22,5400	309.800
29	73.309	ORA	2014-08-07	25,3400	25,5100	24,8600	25,0000	176.300

Figure 23 - Teradata dataset

6.1.3. Extraction/Cleaning/Annotation

In this phase the objective is to extract the interest content. In other words we must extract all the article text, clean it from all the HTML and JavaScript code and register the categories of the particular news. Those categories are:

Market - If the news refers a local (USA) (Figure 25) or foreign market (Figure 24). The idea is to identify a relation between news about the national and foreign market. Small businesses may be more susceptible to variations influenced by “local” news. On the other hand, big companies are influenced by local and foreign events.

S	Description
	South Carolina
	Kentucky
	Maine
	Iowa
	Pennsylvania
	North Dakota
	Oklahoma

Figure 25 - USA cities Dictionary

S	Country
	El Salvador
	Macau
	Sierra Leone
	Italy
	Central African Rep.
	Morocco
	Philippines
	Lithuania

Figure 24 - Foreign Countries Dictionary

Company- If the news refers to a specific company (from our chosen list) (Figure 26) or other subjects such society and government (Figure 27). This indicator suggests the existence of news articles that can be labelled as specific (about a particular company) or labelled as a generic subject such as government, society and environment. These events may have different impacts for the different companies.

S	Description
	Consol Energy Inc
	Royal Bank Scotland
	Banc of California Inc.
	Nextera Energy
	Banco Bilbao Viscaya Ar...
	Zion Bancorporation

Figure 26 - Company Dictionary

S	Description
	congressman
	legal
	economy
	culture
	employ
	law
	deputy
	social order

Figure 27 - Government and Society

Sentiment - If the news could be classified as Positive, Negative or Neutral. The sentiment is defined by the main transmitted idea. By common sense, news articles with positive message may have a good impact. Consequently, a negative message means a negative impact and neutral has no impact at all.

S	Word
	ACQUITTALS
	MISMANAGES
	CONSPIRATORIAL
	SEVERELY
	DISSOLUTION
	PREMATURELY

Figure 29 - Negative Dictionary

S	Word
	BRILLIANT
	PROFITABLE
	SUCCEEDING
	ENABLES
	HONORABLE
	PLEASANT

Figure 28 - Positive Dictionary

a) Workflow – Prepare Data

Here we need to prepare the collected data to be analyzed in “Text Mining” context (Figure 30). With the Xpath node are defined queries for selecting nodes from the XML documents. Here the article content is extracted from the xml documents where the entire code is cleaned from de files. Finally data is converted into documents so the Text Processing nodes will be able to read it (Strings To Document nodes).

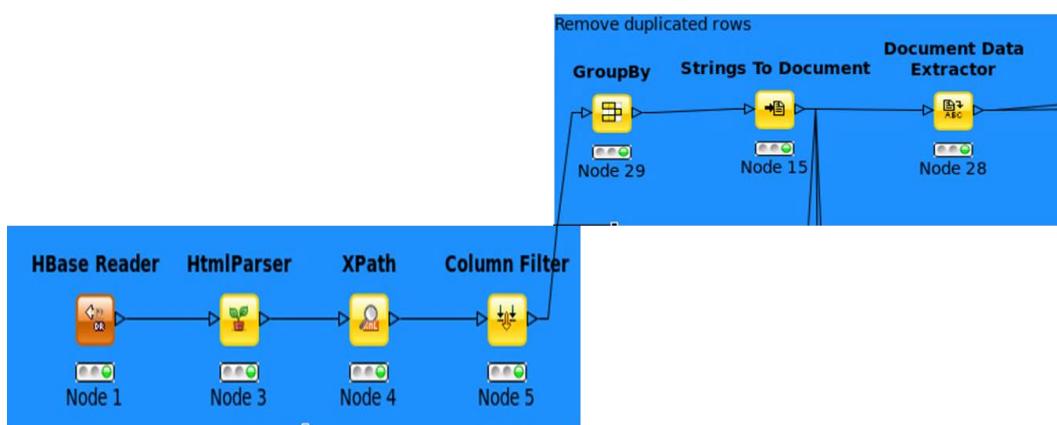


Figure 30 - Content Extractor

The Data collected to Teradata also needed some iteration. In this case it was necessary to add three columns as shown on the figure (Figure 31):

Balance: Defines, on the same day, if the value of the stock rises, decreases or is maintained.

PercentDay: – Percentage of daily variation.

VarDay – Average of PercentDay variation grouped by Symbol.

ID	Symbol	DateSE	OpenSE	HighSE	LowSE	CloseSE	VolumeSE	Balance	PercentDay	VarDay
74.308	ZB-F	2014-08-07	28,0500	28,0500	27,7100	27,7100	1.500	-	0,9879	0,999
74.287	XNY	2014-08-07	0,8300	0,8500	0,8000	0,8200	21.500	-	0,9880	0,999
74.247	WPT	2014-08-07	18,5500	18,5900	18,1000	18,2100	35.600	-	0,9817	0,999
74.226	WIW	2014-08-07	12,0600	12,1400	12,0600	12,0900	80.300	+	1,0025	0,999
74.165	VRS	2014-08-07	3,2000	3,2900	3,1600	3,2900	193.100	+	1,0281	0,999
74.144	VLT	2014-08-07	16,1100	16,1100	16,0100	16,0800	8.600	-	0,9981	0,999
74.104	USM	2014-08-07	36,4700	37,1000	36,2300	36,8100	182.300	+	1,0093	0,999
74.083	UMH	2014-08-07	9,9900	10,0200	9,9400	10,0200	48.600	+	1,0030	0,999
74.043	TVC	2014-08-07	24,0500	24,0500	23,9000	23,9600	13.300	-	0,9963	0,999
74.022	TS	2014-08-07	42,6500	42,7600	41,8800	42,0300	973.300	-	0,9855	0,999
73.961	TIME	2014-08-07	23,7500	23,8700	23,5000	23,7200	831.100	-	0,9987	0,999
73.900	TAP	2014-08-07	71,4900	71,5000	69,5500	69,7000	1.742.300	-	0,9750	0,999
73.879	SXCP	2014-08-07	31,4500	31,5100	31,2000	31,3900	6.900	-	0,9981	0,999
73.839	STJ	2014-08-07	65,2800	65,2800	63,3400	63,4400	2.588.200	-	0,9718	0,999
73.818	SSW-D	2014-08-07	26,8000	26,8000	26,5400	26,6000	25.000	-	0,9925	0,999

Figure 31 - Structured Data update

b) Workflow - Enrichment and Tagging

Several lists were also stored on Teradata Database to tag the articles documents. Here it is processed an entity recognition using a Dictionary Tagger and Bag of Words (Figure 32). After that, the entities are filtered and pass through the Keygraph Keyword extractor. The tags are converted into string and pass through an inference rule engine that defines the tags from a dictionary with -1 and the other with 1.

Once again Data is extracted from those documents converted into string and the rows tags are grouped by URL and Date of the row. Calculating the mean of the tags score, the result is a dataset containing each article categorized with the mean of its correspondent scored tags. Most of the values are absolute defining the document as positive and negative. However some of the documents are inconclusive, with the decimal values close to 0. Those values were round to 0 representing neutral documents.

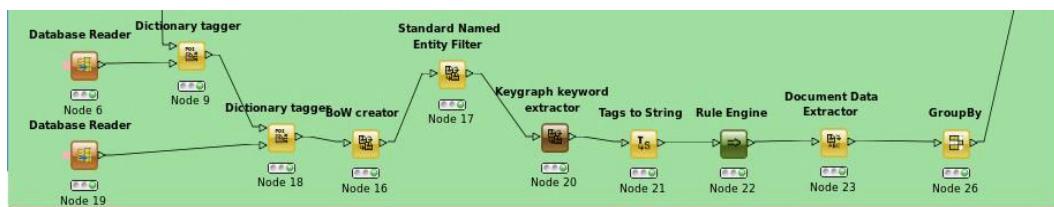


Figure 32 - Tagging Workflow

Inference Engine Rules:

Keygraph Keyword extractor node analyses documents and extracts relevant keywords using the graph-based approach described in "KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor" by Yukio Ohsawa". (For more information about this algorithm consult: Yukio Ohsawam et al, 1998." KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor" and Y.Matsou, 2003. "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information").

Market- 1 USA cities; -1 Foreign Country.

Company – 1 listed Company name; -1 government and society subject.

Sentiment – 1 Positive; -1 negative.

6.1.4. Integration/Aggregation/Representation

Because the main goal is to analyze and predict the impact of news articles in a set of company's stock exchange history we need to integrate these two information sources. This is one of the particularities of Big Data, to be able to make analysis mixing several types of data. As seen before in this document, to merge different types of data will improve the analysis of data and make it more interesting. The approach for the integration is collect and transform, separately, structured (stock exchange historic) and unstructured data (news articles) (Figure 33). From each source it is stored the processed information. Finally the both data types were merged and are now ready to be analyzed.

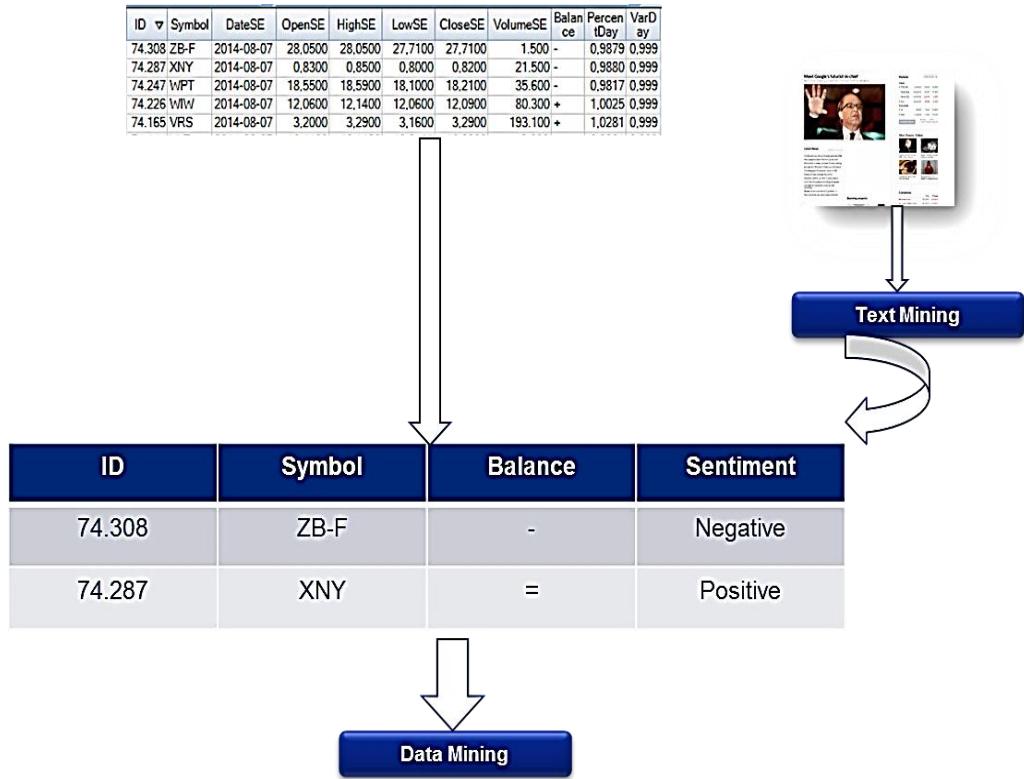


Figure 33 - Text Integration Approach

a) Workflow – Integrate Data

In this phase it was integrated the stock exchange history with news articles structured information (Figure 34). Here the final dataset will contain all the stock results where the date of release matches with the date of news publication. This dataset was stored on Teradata Database. The merge of these two types of data was based in the data of the news article and release date and date of the stock value register.

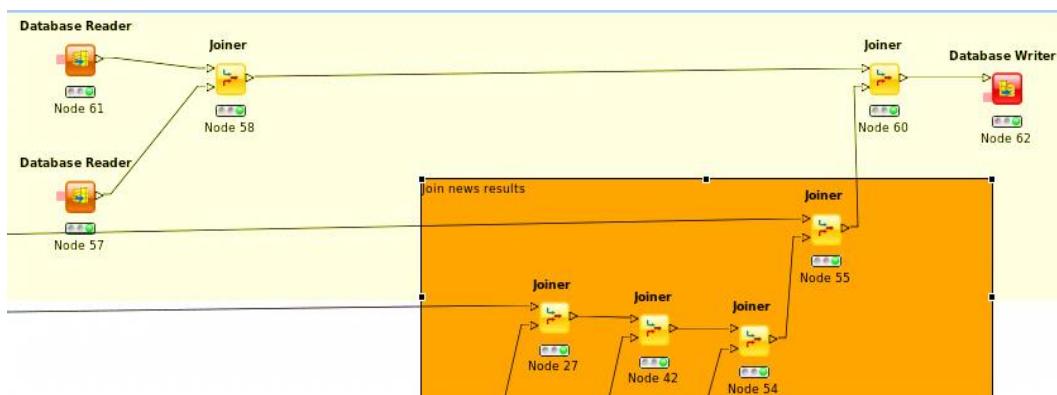


Figure 34 - Data Integration

Through that process (Figure 34) is presented the final dataset resulting from the connections performed before (Figure 35). Because the dataset represent a table that does not fit in this document is was cropped presenting only the important columns.

The table has the following scheme

ID: Key of the company

Symbol: Symbol of the company

DateSE: Date of the stock exchange release

Sector: Industry of the company

SubSector: Sector inside of industry

First (URL): URL of the news article

VarDay – Average of PercentDay variation grouped by Symbol

OutOfRange: Identifies which company in certain day has its Percent of variation different from the average of VarDay

Market: Text Mining Category

Company: Text Mining Category

Sentiment: Text Mining Category

Integrating the processed information the results are as follows:

D	ID	S	Symbol	S	DateSE	S	Balance	D	Percent...	S	Description	S	Sector	S	SubSector
3,487	ATO	2014-07-...	-	1.003	Atmos Energy Co...	Energy	Gas								
19,719	BCH	2014-07-...	-	1.013	Banco De Chile	Bank	Bank								
3,487	ATO	2014-07-...	-	1.003	Atmos Energy Co...	Energy	Gas								
19,719	BCH	2014-07-...	-	1.013	Banco De Chile	Bank	Bank								
3,487	ATO	2014-07-...	-	1.003	Atmos Energy Co...	Energy	Gas								
19,719	BCH	2014-07-...	-	1.013	Banco De Chile	Bank	Bank								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
19,719	BCH	2014-07-...	-	1.013	Banco De Chile	Bank	Bank								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
19,719	BCH	2014-07-...	-	1.013	Banco De Chile	Bank	Bank								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
54,059	POM	2014-07-...	-	1.008	Portland General...	Energy	Electricity								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
54,059	POM	2014-07-...	-	1.008	Portland General...	Energy	Electricity								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
54,059	POM	2014-07-...	-	1.008	Portland General...	Energy	Electricity								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
54,059	POM	2014-07-...	-	1.008	Portland General...	Energy	Electricity								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
54,059	POM	2014-07-...	-	1.008	Portland General...	Energy	Electricity								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
54,059	POM	2014-07-...	-	1.008	Portland General...	Energy	Electricity								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
54,059	POM	2014-07-...	-	1.008	Portland General...	Energy	Electricity								
21,126	LGP	2014-07-...	-	1.006	Lehigh Gas Partn...	Energy	Gas								
54,059	POM	2014-07-...	-	1.008	Portland General...	Energy	Electricity								

Figure 35 - Final DataSet

6.1.5. Analysis Model

Now it is time to perform some data mining trying to discover if there is a pattern between the text mining results performed to the news articles and the stock exchange variation. For this use case we performed classification algorithms to predict the value Balance. Classification, a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

a) Workflow

As mentioned above, although Knime has its own data mining nodes we chose to select the nodes from Weka plugin. Where it is possible we verify all the algorithms selected. For data analysis the dataset was divided, 70% for training and 30% for testing. For classification nodes we also used the Weka predictor that uses the model trained to predict the Balance value on the data saved for testing (Figure 36).

After prediction Knime has nodes to automatically score the model and present the ROC curve.

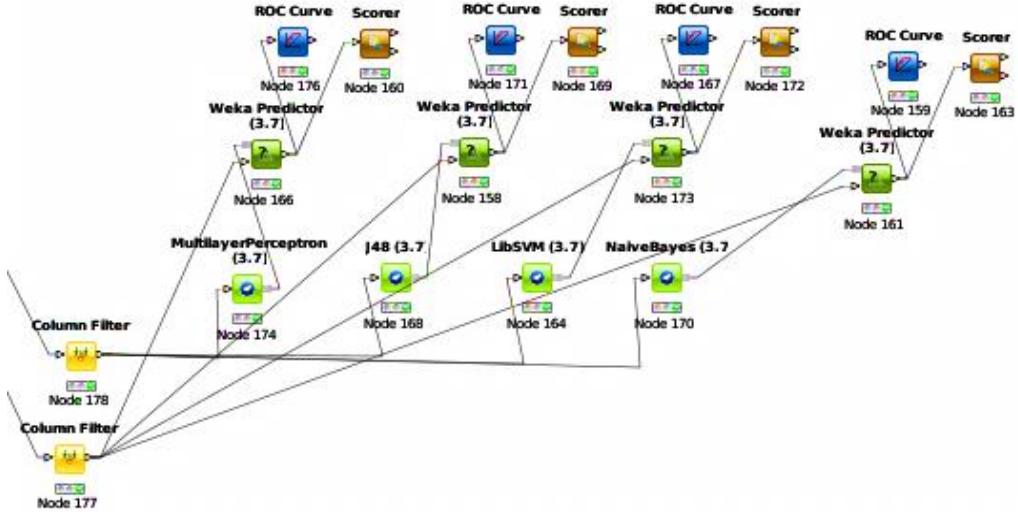


Figure 36 - Final Analysis Model Workflow

To better understand, this workflow contains the algorithms nodes for each scenario that is filtered from the training dataset. After the algorithm node it is connected the Weka prediction node that will perform the insight into the test dataset. After the prediction it is used a scorer to analyze the confusion matrix and accuracy statistics.

b) Classification Algorithms

By option, we trained our data with four classification algorithms provided by Weka, these four algorithms are:

J48: open source implementation of the C4.5 algorithm that builds decision trees from a set of training data. For each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other (“C4.5 algorithm,” 2014).

MultiLayerPerceptron: Multi-Layer perceptron (MLP) is a feed forward neural network with one or more layers between input and output layer. This type of network is trained with the back propagation learning algorithm. MLPs are widely used for pattern classification, recognition, prediction and approximation. Multi-Layer Perceptron can solve problems which are not linearly separable (Source Forge 2011).

LibSVM: Integrated library for support vector classification and regression and supports multi-class classification.

NaiveBayes: Simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model", a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature ("Naive Bayes classifier," 2010)

c) Scenarios

To perform classification algorithms, some scenarios with different variables combination are defined. The Balance, the variable data which indicates either the stock value increased or decrease, is the target of the classification. The other variables will help the algorithm find a pattern between the news article indicators and the target. For testing five scenarios presented are defined in the following table (Table XI)

Table XI - Test Scenarios

	Scenario	1	2	3	4	5
Variables	Target	Balance	Balance	Balance	Balance	Balance
	ID	X	X	X	X	X
	DateSE	X				
	Balance	X	X	X		
	Percent	X				
	Description					
	Sector	X			X	
	SubSector	X	X			
	Varday	X				
	Market	X	X	X	X	
	Company	X	X	X	X	
	Sentiment	X	X	X	X	X

d) Performance Measure

For classification algorithms, the performance is evaluated through the analysis of the capability that the model has to predict a value correctly.

The confusion matrix (Figure 37) is the base (or basis) for several measures and presents the number of correct classifications *vs* predicted classifications for each class.

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives
-	12589	292	7596	0
+	7308	4	13034	131
=	268	16	20012	181

Figure 37 - Confusion Matrix

According to the relation between True/False Positives and True/False Negatives it is possible to measure the performance as follows:

Accuracy: Percentage of positive and negative samples correctly classified in the universe:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Sensitivity: Percentage of positive samples correctly classified within the total of positive samples:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{Positive}}$$

Precision: Percentage of positive samples correctly classified within the total of samples classified as positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Specificity: Percentage of negative samples correctly classified within the total of negative samples:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{TN}}{\text{Negative}}$$

F-measure: Weighted average of Sensitivity and Precision:

$$F = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Row ID	TruePo...	FalseP...	TrueN...	FalseN...	Recall	Precision	Sensiti...	Specificity	F-meas...	Accuracy	Cohen'...
-	12480	141	7747	109	0.991	0.989	0.991	0.982	0.99	?	?
+	7306	129	12909	133	0.982	0.983	0.982	0.99	0.982	?	?
=	417	4	20024	32	0.929	0.99	0.929	1	0.959	?	?
Overall	?	?	?	?	?	?	?	?	?	0.987	0.973

Figure 38 - Knime Accuracy Statistics

Classification ROC curve

To complement the evaluation above, the Receiver Operating Characteristic (ROC) curve (Figure 39) was designed. ROC can be plotted as a curve on an X-Y axis. The false positive rate is placed on the X axis. The true positive rate is placed on the Y axis (Oracle, 2011). The area under the ROC curve (AUC) measures the discriminating ability of a binary classification model. The larger the AUC, the higher the likelihood .That is, an actual positive case will be assigned a higher probability of being positive than to an actual negative case. The AUC measure is especially useful for datasets with unbalanced target distribution (one target class dominates the other).

In this plot it is possible to confirm the high precision of the model because the red line is near the absolute value in the y axis. The straight gray line in this plot shows what you would expect from randomly picking compounds.

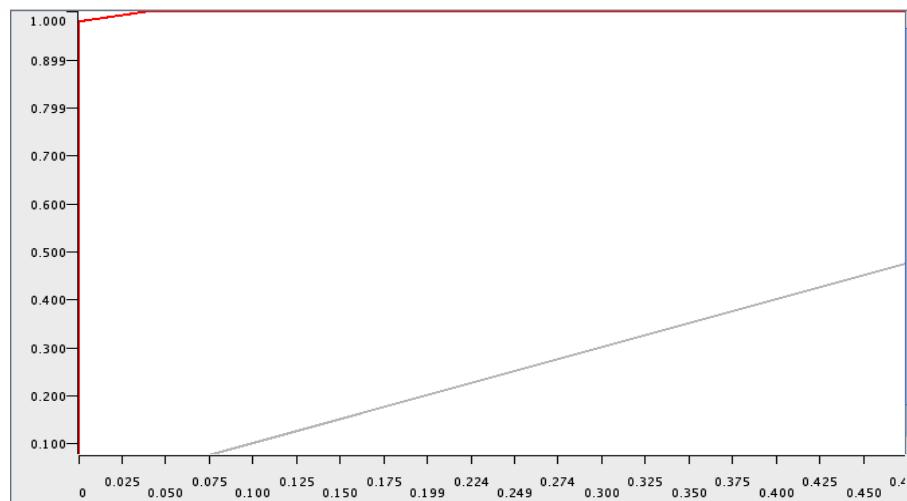


Figure 39 - ROC Curve

6.1.6. Interpretation – DataSet Analysis

This phase includes the analysis, evaluation and interpretation of the results already achieved. In order to produce knowledge, it is important to understand the data that is presented to us.

Before we look at the predicted data we must know well our dataset and their particularities. After that, the stage that follows is to evaluate the results of the classification algorithms to find the best model that is capable of predicting the kind of variation that will prevail on the stock market in that day. The objective is not only to find the model but also to understand its characteristics.

As we know, the dataset used for study is a merge between the stock exchange history data and the news articles text mining results. This data visualization is provided by the QlikView plattform.

At this we conclude the following:

More than half of the records show that in the end of the day the value was lower than at the opening (Figure 40).

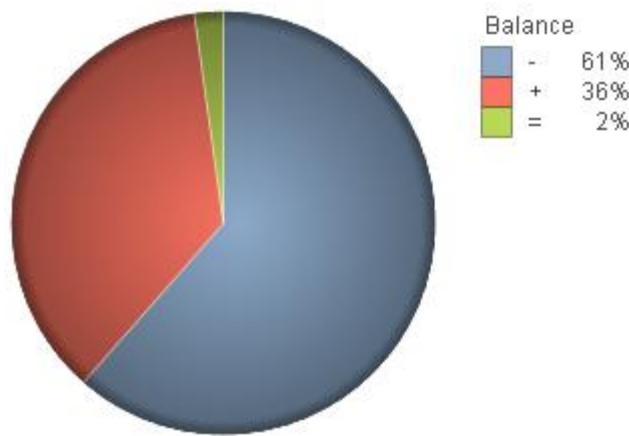


Figure 40 - Balance Distribution

In the distribution of the news articles sentiment (Figure 41), most of the news, 73%, were classified as negative because the words with negative meaning are more expressive and easier to “find”.

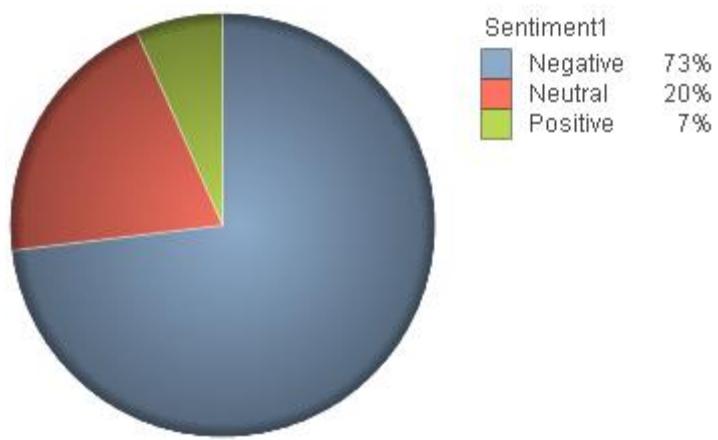


Figure 41 - Sentiment Category Distribution

This chart shows 51% related to news from USA cities (Figure 42). The 35% value (or data) is related to news about foreign countries and the rest of the data concerns inconclusive results about the location.

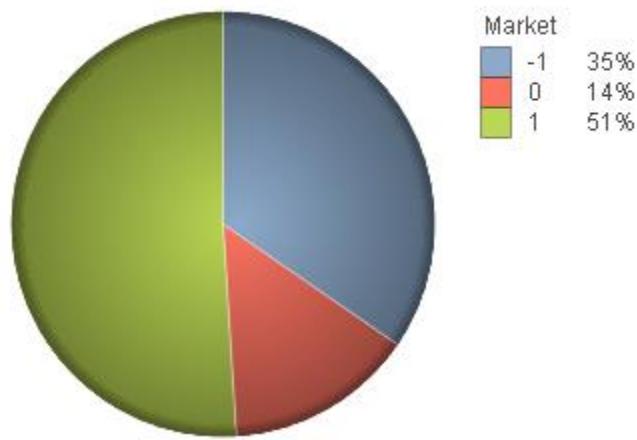


Figure 42 - Market Category Distribution

In the Distribution per Company Category (Figure 43) almost 100% of the news doesn't identify any listed companies. This 99,80% shows that the subject has something related to government and society issues.

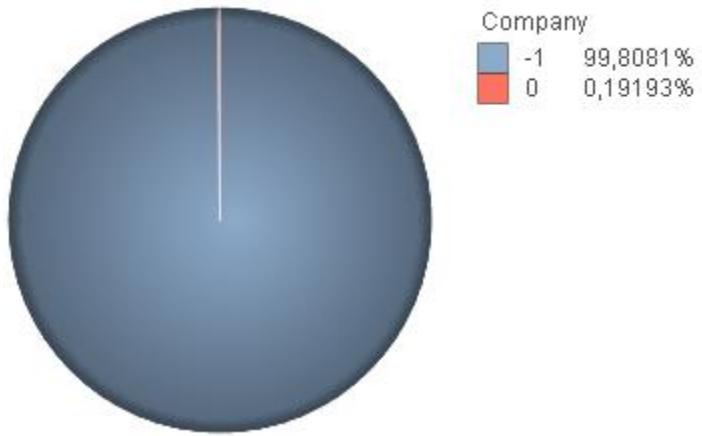


Figure 43 - Company Category Distribution

a) Algorithm Evaluation - Results

After executing each Weka node in the Knime interface, here are the top 3 of the classification results (Table XII).

Table XII - Classification Algorithm Results

Scenario	Algorithm	Class	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's Kappa
1	J48	"+"	1	1	0,999	1	1	0,999
		"+"	1	0,999	1	1		
		"=="	1	0,998	1	0,999		
1	LibSVM	"+"	0,989	0,991	0,982	0,99	0,987	0,973
		"+"	0,983	0,982	0,999	0,982		
		"=="	0,99	0,929	1	0,959		
5	LibSVM	"+"	0,999	0,999	0,997	0,999	0,998	0,997
		"+"	0,998	0,998	0,999	0,998		
		"=="	1	0,98	1	0,998		

By analyzing the measures above, we see that the model that is capable of predicting the three classes with a balanced precision is **the classifier J48 for the scenario 1** which presents accuracy with the highest value. Just a reminder, the scenario 1 predicted de balance target using all existent variables.

Although there are great models with strong precision in some predictions, this is the model that will be used to perform data analysis with the new prediction.

The ROC curve, because it performs on two classes, we can't present the scenario with the current three classes (Figure 44), however, in an ideal representation, the curve would be something very similar to the ROC curve shown (Figure 45).

TruePo...	FalseP...	TrueN...	FalseN...	Recall	Precision	Sensiti...	Specificity	F-me...	Accuracy	Cohen'...
12587	5	7883	2	1	1	1	0.999	1	?	?
7435	2	13036	4	0.999	1	0.999	1	1	?	?
448	0	20028	1	0.998	1	0.998	1	0.999	?	?
?	?	?	?	?	?	?	?	?	1	0.999

Figure 44 - Accuracy Statistics for J48 algorithm Scenario 1

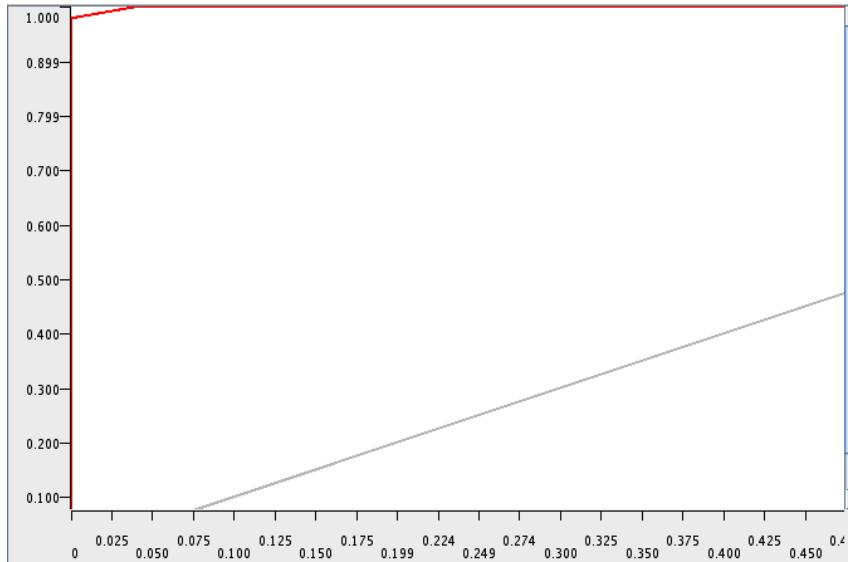


Figure 45 - ROC Curve J48 algorithm scenario 1 (two classes)

b) Data Analysis (Prediction)

After submitting the training algorithm we can see the resultant prediction made to a test dataset (Figure 46). Here the objective is to produce some statistics, in order to get to know the model better.

D ID	S DateSE	S Balance	D PercentDay	S Sector	S SubSec...	D VarDay	D Market	D Compa..	D Sentim...	S Prediction (Balance)
3,487	2014-07-09	-	1.003	Energy	Gas	0.997	1	-1	-1	-
19,719	2014-07-16	-	1.013	Bank	Bank	0.997	1	-1	-1	-
3,487	2014-07-09	-	1.003	Energy	Gas	0.997	1	-1	-1	-
19,719	2014-07-16	-	1.013	Bank	Bank	0.997	-1	-1	-1	-
21,126	2014-07-16	-	1.006	Energy	Gas	1.003	1	-1	0	-
21,126	2014-07-16	-	1.006	Energy	Gas	1.003	0	-1	-1	-
21,126	2014-07-16	-	1.006	Energy	Gas	1.003	1	-1	0	-
13,479	2014-07-14	-	1.002	Energy	Oil and Gas	0.999	1	-1	0	-
54,059	2014-07-30	-	1.008	Energy	Electricity	0.999	-1	-1	-1	-
13,479	2014-07-14	-	1.002	Energy	Oil and Gas	0.999	-1	-1	0	-
13,479	2014-07-14	-	1.002	Energy	Oil and Gas	0.999	1	-1	0	-
54,059	2014-07-30	-	1.008	Energy	Electricity	0.999	-1	-1	1	-
54,059	2014-07-30	-	1.008	Energy	Electricity	0.999	1	-1	1	-
13,479	2014-07-14	-	1.002	Energy	Oil and Gas	0.999	1	-1	-1	-
54,059	2014-07-30	-	1.008	Energy	Electricity	0.999	-1	-1	-1	-
13,479	2014-07-14	-	1.002	Energy	Oil and Gas	0.999	-1	-1	-1	-
13,479	2014-07-14	-	1.002	Energy	Oil and Gas	0.999	-1	-1	0	-
36,685	2014-07-23	+	0.997	Bank	Bank	0.999	1	-1	0	+
54,059	2014-07-30	-	1.008	Energy	Electricity	0.999	-1	-1	-1	-
36,685	2014-07-23	+	0.997	Bank	Bank	0.999	-1	-1	0	+
36,685	2014-07-23	+	0.997	Bank	Bank	0.999	1	-1	1	+
54,059	2014-07-30	-	1.008	Energy	Electricity	0.999	0	-1	-1	-
57,138	2014-07-31	-	1.014	Energy	Oil	0.999	1	-1	-1	-
54,059	2014-07-30	-	1.008	Energy	Electricity	0.999	1	-1	0	-
57,138	2014-07-31	-	1.014	Energy	Oil	0.999	1	-1	1	-
54,059	2014-07-30	-	1.008	Energy	Electricity	0.999	1	-1	0	-
57,138	2014-07-31	-	1.014	Energy	Oil	0.999	1	-1	1	-

Figure 46 - Predicted DataSet Scenario 1 J48 algorithm

In the Histogram bellow (Figure 47) it is evident the tendency for the daily closed stock value to be lower when compared with the open stock value.

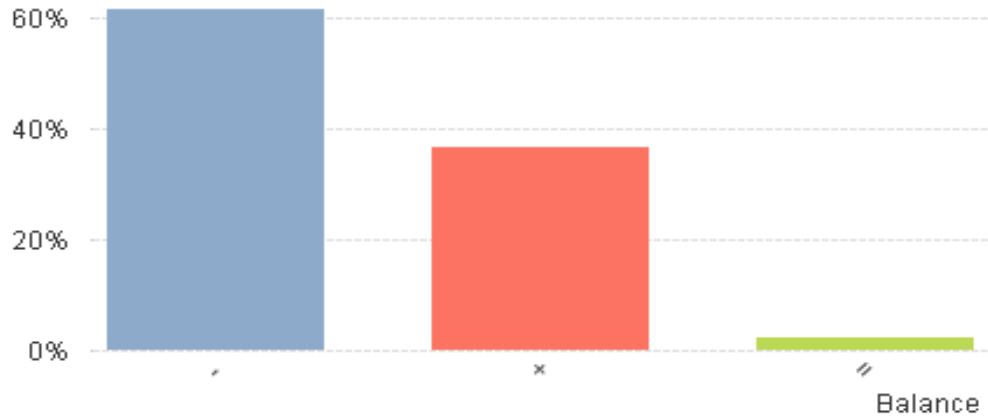


Figure 47 - Predicted Balance Distribution

Also, we can see that the average of variation itself (Figure 48) does not present its values dispersed and the bank, Oil and Gas companies are the companies shown to be more sensitive to the market variations.

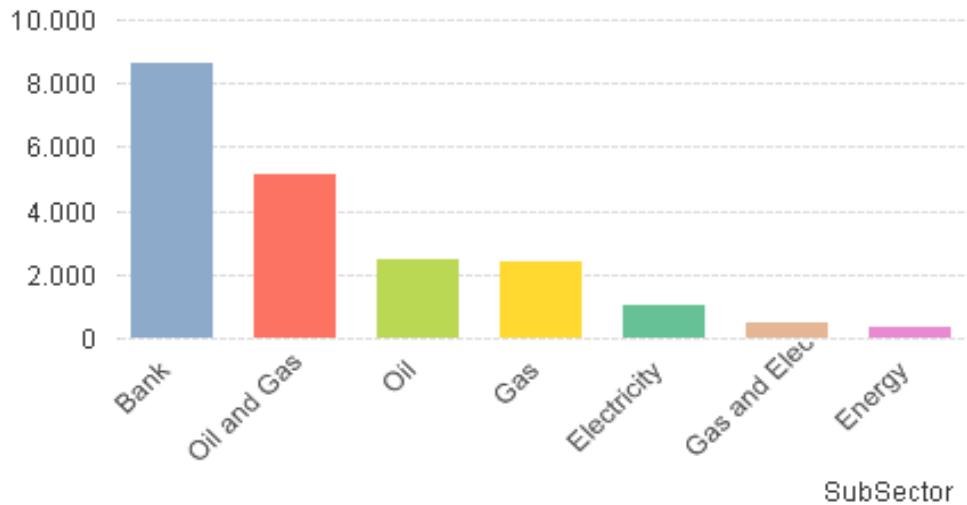


Figure 48 - Average of stock variation Distribution

Through the analysis of the market category (Figure 49), we can see that the subject which involves different countries is predominant with 51% and most of the news does not refer the name of the company, 35%. This means that the article has a more generic subject and the impact cannot be considered direct but indirect on several industries.

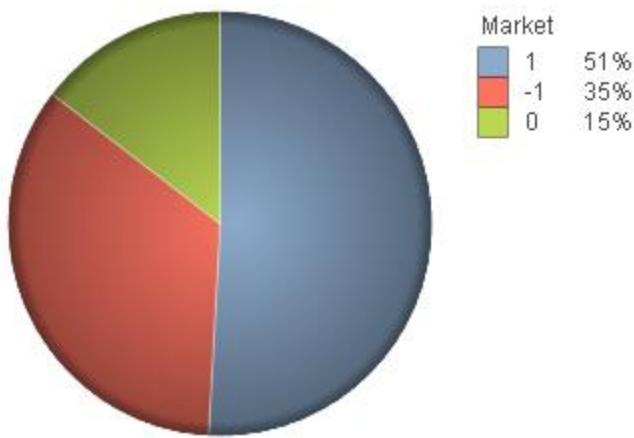


Figure 49 - Market category Distribution

The final category that which defines the news article is the sentiment (Figure 50). Also, we can see the predominance to classify the article as negative with 73%.

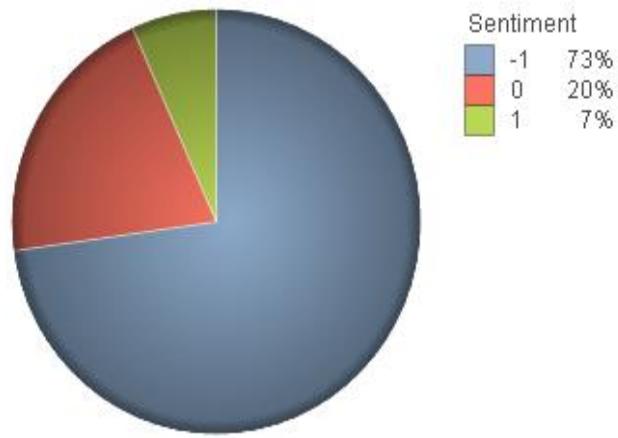


Figure 50 - Sentiment Category Distribution

To conclude, the analysis to our model enables us to group the occurrences of some categories which classify the news article by the balance prediction. The difference between the averages of each category is very low but this precision is what provides the model efficiency.

In the Grid-Chart below (Figure 51) it is possible to see the Balance distribution from a combination of Market and Sentiment values. According to this, the negative balance has more presence in a combination of negative sentiment and a local market (USA).

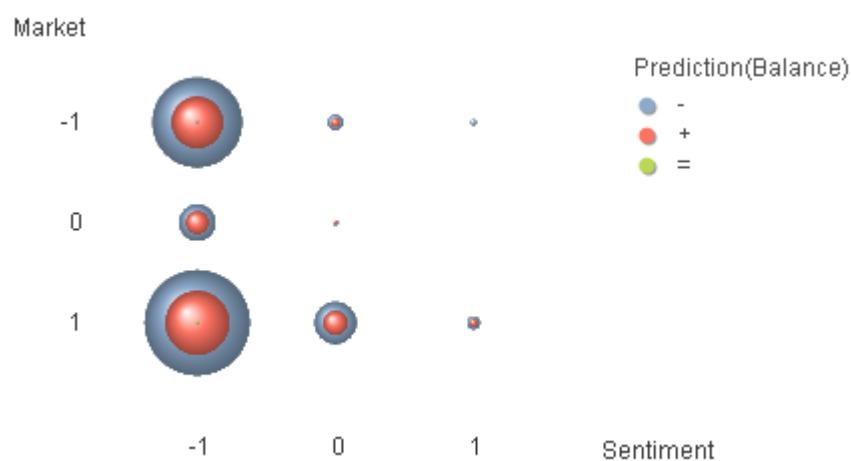


Figure 51 - Balance Prediction Distribution per Market and Sentiment Categories

This Grid-Chart (Figure 52) presents the combination between Company and Sentiment where the predominance is the negative sentiment but the company is not mentioned. This may occur because it is not every day that a news article about a specific set of companies is released.

In fact, the bank and energy sector are two industries which are very sensitive to subjects related with economy, society and politics.

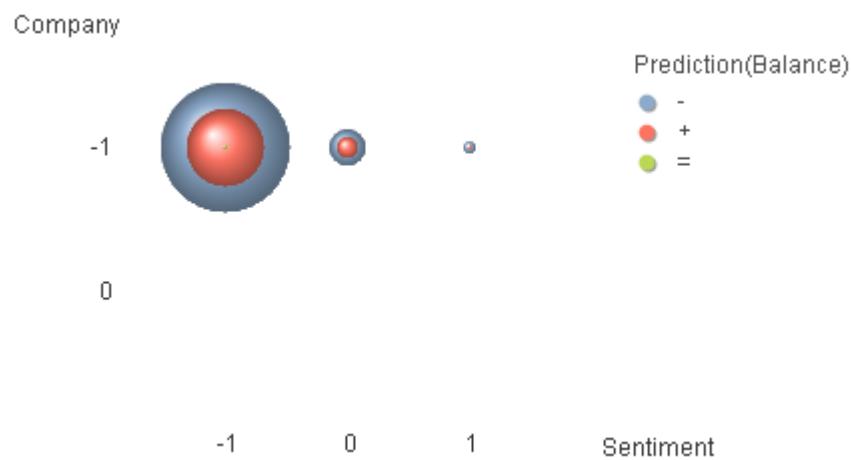


Figure 52 - Balance Prediction Distribution per Company and Sentiment Categories

The final Grid-Chart (Figure 53) presents the combination between Company and Market where it is evident the tendency of fall when the news article. Although not mentioning the company, communicates something that happens in the local market (USA).

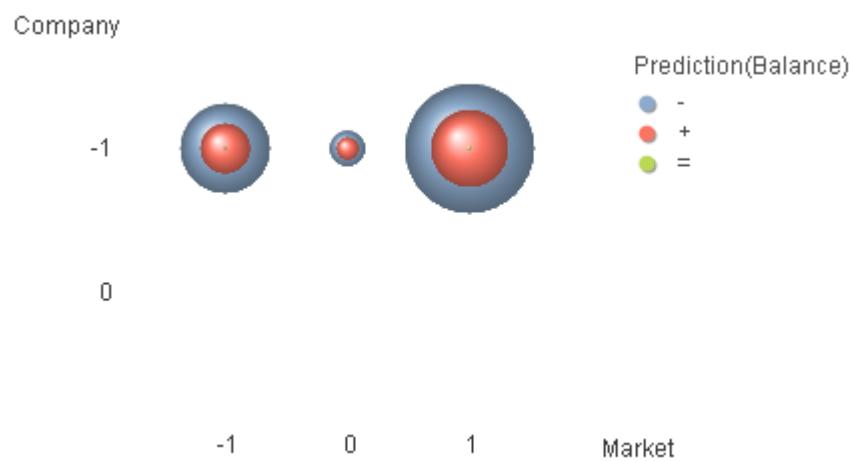


Figure 53 - Balance Prediction Distribution per Company and Market Categories

6.2. Discussion and Conclusions

This use case was a small sample of the capabilities attached to Big Data. The objective was to be aware of the existence of other types of data, with the challenge of volume and velocity required.

At the end of this use case, it is possible to conclude that, actually, merging the sources and types makes the data analysis more interesting and complete.

The analysis of text content is very complex, it requires an extra effort for the machine to identify the characteristics in a document, such as sentiment with the same sensitivity of a human (being).

In our classification of the news, according to its sentiment, we defined a set of words and combinations of positive and negative words. Luckily, in most cases, those words were enough to classify the sentiment but, at the same time, the occurrence of “undefined” classification and the false classification indicate that our dictionaries must be more embracing and capable of detecting the particularities of natural language like context and double meaning and, in a more advanced form, they should be capable of detecting irony.

For that reason, we believe that the stock exchange prediction is better when combined with the history stock exchange, not only with the sentiment of an article but also with other indicators that could be extracted and, in some manner, determinant for the market variations.

7. Conclusion

7.1. Project Remarks

The Big Data has been a much sought after topic, not only by researchers but also by the organizations. The fact that this project has been developed in a company as Deloitte, made it more oriented to business.

The entire development was done in a more general context of financial industry because the cases in which the data was "private" they would realize a process of partnership with other organizations (data sources). This would not be possible for the academic scope of the project and for the lay in this matter.

Some difficulties with this project focused on the choice and integration of the tools. Being an academic project, it was limited to a few licenses.

Another challenge was the text mining used to define a document based on its keywords. In the practical case a dictionary was used to find, not only positive and negative words, but also keywords that would provide the definition of the location and the subject of the news document. However, it was still hard to classify the documents with the same precision a human being would have, and also to identify its impact on sensitive areas such as the stock market exchange.

Despite all difficulties, there is a great satisfaction on the results achieved because this way, we get to know the best of both worlds, "large data" in "different types of data". On the other hand, there was also a call for attention to some important issues that may facilitate or restrict the entry of Big Data in the companies.

7.2. Final Remarks

At the beginning of this project a set of main goals were defined to guide the entire development of the activities.

To better understand the relation between what was achieved, a table is presented (Table XIII) mapping the goals and results.

Table XIII - Project Goals and Results

Goal	Results
A Big Data platform with analytic and predictive capabilities	<ul style="list-style-type: none">• Platform integrating Big Data tools.• Collected, processed and stored unstructured data (news articles)• Merged unstructured and structured data• Data Mining processing <p>Choice of an optimal model for data prediction</p>
Identification of a Big Data problem	<ul style="list-style-type: none">• Big Data Complexity Level Framework
To get an Insight about the current business opportunities and which offers can we find in the Big Data market.	<ul style="list-style-type: none">• Overview about the business opportunities associated to Big Data• Summary and comparison of Big Data technologic solutions

As shown in the project, the market was invaded by several commercial products and open source projects. They claim to be the best solution for the Big Data issues that some organizations are facing right now.

Although organizations have the propensity to choose commercial tools as a secure and high quality product, the emergence of Open Source tools has proved their quality and attracted substantial attention.

To better perform the decision making, managers have to balance the pros and cons of choosing a commercial/open source tool and deciding which deployment option fit “better” their needs. No matter the choice made, it is important to take into account the security, cost and flexibility of the product, as well as the resources and time needed to do this project.

By exploring the Big Data market, it was possible to understand why everybody talks and wants to be part of this new age. Big Data has proved its concept and value, promoting big opportunities in several industries. With Big Data, companies, as the financial services industry, will be able to know their customers and competitors better, through analyzing data from new sources like the social media, and to prevent some market changes by analyzing the news online.

7.3. Future work

Even though there is a lot of information about Big Data's technology, benefits and successful cases, the organizations have further questions about what really defines Big Data and outlines a project of this type. A theme like this needs to be developed and the organizations must be ready and motivated to enter this new era.

In terms of future work, once the value of Big Data is proven in industries such as the financial one, a proposal would be exploring new business opportunities using the social media and "public" transactional data.

In Deloitte's case, the financial industry is a growing area and the use of data for fraud detection, prediction and mitigation of risks may be some of the consulting services that can use Big Data.

7.4. The Profession Acts

The characterization of the acts of the profession is an essential reference for any professional order sustaining and regulating the scope of the professional activity which identifies, classifies and contextualizes the acts of the profession that the National Council of the College of Computer Engineering of the Order of engineers considers pertinent to be adopted within the Computer Engineering. This framework aims to be a catalyst for a conscious and competent performance with a modern and comprehensive perspective of professional Computer Engineering to the service of humanity, society and the economy. "

Given the fact that this is a project, it makes sense to fit the same inside of the acts of the profession of Engineers.

In the tables below, there is a mapping which acts on the profession that this project, as a whole, can cover.

Table XIV - Domain Analysis and Requirements Engineering

Area	Act Group	Project	Obs.
1.1.1.	Modeling applicational domains of information systems	✓	
1.1.3.	Identify and describe functional requirements of information systems	✓	
1.1.4.	Identify and characterize non-functional requirements of information systems	✓	
1.1.5.	Analyze and validate requirements for information systems	✓	
1.2.1.	Specify information requirements in the business perspective	✓	
1.2.2.	Specify requirements for interoperability between information systems	✓	
1.2.3.	Specify interactions with people in information systems	✓	
1.3.1.	Define and model the processes of acquisition, processing and storage of information	✓	
1.3.2.	Define and model information systems architectures	✓	
1.3.3.	Perform cost / benefit of information systems	✓	
1.4.	Specify requirements for IT solutions	✓	
1.4.1.	Specify functional requirements for IT solutions from the perspective of the user	✓	
1.4.2.	Specify requirements for interoperability between software solutions	✓	
1.4.3.	Specify user interfaces in computer solutions	✓	
1.4.4.	Specify (other) non-functional requirements of software solutions (eg, performance, security)	✓	

Table XV - Design and Construction of Computer Solutions

Area	Act group	Project	Obs.
2.1.1.	Analyzing and validating software solutions requirements (including, eg, identification, characterization and evaluation of the technical risk associated with requirements)	✓	
2.2.2.	Define and model software solutions architectures	✓	
2.3.1.	Identify and select platforms and tools to support the construction and maintenance of IT solutions	✓	
2.3.3.	Identify, characterize and assess the risk of making changes to software solutions (including, eg, impacts on compliance with the requirements and technical characteristics)	✓	
2.4.2.	Integrating IT solutions (including, eg, inter-operate prebuilt software solutions)	✓	

8. References

- Aase, K.-G. (2011). *Text Mining of News Articles for Stock Price Predictions*. Retrieved from <http://www.diva-portal.org/smash/get/diva2:440508/FULLTEXT01.pdf>
- Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Han, J. (2012). *Challenges and Opportunities with Big Data. A community white paper developed by leading researchers across the United States.* [2012-10-02], <http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf>.
- Apache Foundation. (2014). HDFS Architecture Guide. Retrieved from http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
- Bissi, W. (2007a). Metodologia De Desenvolvimento Ágil. *Campo Digital*, 2(1). Retrieved from <http://www.revista.grupointegrado.br/revista/index.php/campodigital/article/view/312>
- C4.5 algorithm. (2014, July 15). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=C4.5_algorithm
- Chansler, R., Kuang, H., Radia, S., Shvachko, K., & Srinivas, S. (2014). The Hadoop Distributed File System. Retrieved from <http://www.aosabook.org/en/hdfs.html>
- Cuzzocrea, A., Song, I.-Y., & Davis, K. C. (2011). Analytics over large-scale multidimensional data: the big data revolution!. ACM Press. Retrieved from <https://webvpn.uminho.pt/http/0/dl.acm.org/citation.cfm?id=2064695>
- Danube (2008). Scrum Methodology & Agile Scrum Methodologies..Retrieved from <http://scrummethodology.com/>
- Deloitte. (2013). Deloitte Resources. Retrieved from <https://pt.deloitteresources.com/practice/peopledev/Pages/New-to-Deloitte.aspx>
- Deloitte, R. (2014). *Big Data: Understanding the Technology Landscape*. Retrieved from <https://pt.deloitteresources.com/practice/peopledev/Pages/New-to-Deloitte.aspx>

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). Addressing big data issues in Scientific Data Infrastructure Presented at the 2013 International Conference on Collaboration Technologies and Systems (CTS).

Dijcks, J. P. (2012). Oracle: Big Data for the Enterprise. *Oracle White Paper*. Retrieved from <http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf>

Dr. Santhosh Baboo & P Renjith Kumar, L. (2013a). Next Generation Data Warehouse Design with Big data for Big Analytics and Better Insights. *Global Journal of Computer Science and Technology*, 13(7). Retrieved from <http://computerresearch.org/stpr/index.php/gjcst/article/view/1473>

Foyer, D. (2013, December 2). Financial Comparison of Big Data MPP Solution and Data Warehouse Appliance. Retrieved from http://wikibon.org/wiki/v/Financial_Comparison_of_Big_Data_MPP_Solution_and_Data_Warehouse_Appliance

GOH, S. H. (2006). Open Source and Commercial Software. An In-depth Analysis Of The Issues. Gualtieri, M., & Yuhanna, N. (2014). *The Forrester Wave™: Big Data Hadoop Solutions, Q1 2014*. Forrester.

Hague, & Netherl. (2014). BigData Startups - the big data knowledge platform. Retrieved from [http://www.bigdata-startups.com/open-source-tools/,%2012-08-2014/](http://www.bigdata-startups.com/open-source-tools/)

Hevner, A. R., March, S. T., & Park, J. (2004). Design Science in Information Systems Research. Retrieved from <http://em.wtu.edu.cn/mis/jxkz/sjkx.pdf>

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big Data For Dummies*. Retrieved from <http://it-ebooks.info/book/2082/>.

- Horowitz, Alan S. (2013). Let's Play Moneyball: 5 Industries That Should Bet on Data Analytics. Retrieved from <http://plotting-success.softwareadvice.com/moneyball-5-industries-bet-on-analytics-1113/>
- IBM. (2013, October 15). Big data architecture and patterns, Part 3: Understanding the architectural layers of a big data solution [CT316]. Retrieved October 26, 2014, from <http://www.ibm.com/developerworks/library/bd-archpatterns3/>
- IDC. (2013). Business opportunities: Big Data. Retrieved from http://ec.europa.eu/enterprise/dem/sites/default/files/page-files/big_data_v1.1.pdf
- Intel IT Center. (2014). *Big Data in the Cloud: Converging Technologies*. Retrieved from <http://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/big-data-cloud-technologies-brief.pdf>
- Knime.org. (2010). Knime.org. Retrieved from <http://www.knime.org/knime>
- Krishan, K. (2013a). *Data Warehousing in the Age of Big Data*. Retrieved from <http://www.sciencedirect.com/science/book/9780124058910>
- Lima, C. A. R., & Calazans, J. de H. C. (2013). Pegadas Digitais: "Big Data" E Informação Estratégica Sobre O Consumidor. Retrieved from http://gitsufba.net/anais/wp-content/uploads/2013/09/13n2-pegadas_49483.pdf
- Maier, M., Serebrenik, A., & Vanderfeesten, I. T. P. (2013). Towards a Big Data Reference Architecture. Retrieved from http://www.win.tue.nl/~gfletche/Maier_MSc_thesis.pdf
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Mathew, S., Halfon, A., & Khanna, A. (2012). *Financial Services Data Management: Big Data Technology in Financial Services*. Retrieved from <http://www.oracle.com/us/industries/financial-services/bigdata-in-fs-final-wp-1664665.pdf>

Naive Bayes classifier. (2010). In *Wikipedia, the free encyclopedia*. Retrieved from http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Naive_Bayes_classifier.html

Oracle, (2011, December 3). Data Mining Concepts. Retrieved November 23, 2014, from https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004

Persistent Systems. (2013). *How to Enhance Traditional BI Architecture to Leverage Big Data*. Retrieved from http://sandhill.com/wp-content/files_mf/persistentsystemswhitepaperenhance_traditionalbi_architecture.pdf

Schrader, D., & Dietz, J. (2013). *Teradata Unified Data Architecture - Gives Users Any Analytic on Any Data*. Retrieved from <http://www.intel.com/content/dam/www/public/us/en/documents/success-stories/big-data-xeon-processor-teradata-solution-success-story.pdf>

Scrum features. (2010). Retrieved from <http://www.scrummethodology.org/scrum-features.html>

Source Forge (2011). Multilayer Perceptron. Retrieved from <http://neuroph.sourceforge.net/tutorials/MultiLayerPerceptron.html>

Sridhar, P., & Dharmaji, N. (2013, August). A Comparative Study on How Big Data is Scaling Business Intelligence and Analytics. *International Journal of Enhanced Research in Science Technology & Engineering*.

StackIQ, Inc. (2012). Capitalizing on Big Data Analytics for the Insurance Industry. Retrieved from
http://cdn2.hubspot.net/hub/173001/file-18488782-pdf/docs/stackiq_insuranceind_wpp_f.pdf

Sumathy, K. L., & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues –An Overview. Retrieved from
<http://research.ijcaonline.org/volume80/number4/pxc3891685.pdf>

Sun, H., & Heller, P. (2012). Oracle Information Architecture: An Architect’s Guide to Big Data. Oracle, Redwood Shores. Retrieved from
<http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf>

Tan, A.-H. (2000). Text Mining: The state of the art and the challenges. Retrieved from
http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf

Turner, D., Michael, S., & Rebecca, S. (2013). *Analytics: The real-world use of big data in financial services.* Retrieved from http://www-935.ibm.com/services/multimedia/Analytics_The_real_world_use_of_big_data_in_Financial_services_Mai_2013.pdf

Vaishnavi, V., & Kuechler, W. (2004, January). *Design Research in Information Systems*. Retrieved from <http://desrist.org/design-research-in-information-systems/>

Venner, J. (2009). *Pro Hadoop*. Apress.

Vinnakota, S. (2012). Is Big Data eclipsing traditional BI? Retrieved from
<http://www.thebusinesstechnologyforum.com/2012/12/is-big-data-eclipsing-traditional-bi/>

World Economic Forum. (2012). Big Data, Big Impact: New Possibilities for International Development. World Economic Forum. Retrieved from

<http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development>

Zikopoulos, P., Eaton, C., & deRoos, D. (2011). *Understanding Big Data: Analytics for Enterprise*

Class Hadoop and Streaming Data. Retrieved from

<http://dl.acm.org/citation.cfm?id=2132803>

9. Appendix

9.1. Publications

Are in submission phase the following publications:

- Big Data for Stock Market by means of Mining techniques (World Conference on Information Systems and Technologies)
- Why Big Data - A project assessment framework (European Journal of Information Systems)
- Big Data - A Global framework using open-source (Information and Software Technology Journal)