REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies

DEVELOPING SCALABLE DATA WAREHOUSE WITH BIG DATA REQUIREMENTS

Nada Muaad Abdullah ABDULLAH

Master's Thesis

Supervisor

Asst. Prof. Dr. Merve Bulut YILGÖR

Istanbul, 2023

DEVELOPING SCALABLE DATA WAREHOUSE WITH BIG DATA REQUIREMENTS

Nada Muaad Abdullah ABDULLAH

Information Technologies

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

The thesis titled DEVELOPING SCALABLE DATA WAREHOUSE WITH BIG DATA REQUIREMENTS prepared by NADA MUAAD ABDULLAH ABDULLAH and submitted on 26/04/2023 has been **accepted unanimously** for the degree of Master of Science in Information Technologies.

Asst. Prof. Dr. Merve Bulut YILGÖR

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Merve Bulut YILGÖR	Department of Basic Sciences, Altınbaş University	_____
Assoc. Prof. Dr. Hakan KAYGUSUZ	Department of Basic Sciences, Altınbaş University	_____
Asst. Prof. Dr. Elif Segah ÖZTAŞ	Department of Mathematics, Karamanoğlu Mehmetbey University	_____

I hereby declare that this thesis meets all format and submission requirements of a Master's Thesis.

Submission date of the thesis to Institute of Graduate Studies: ___/___/___

I hereby declare that all information and data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Nada Muaad Abdullah ABDULLAH

Signature

ACKNOWLEDGEMENTS

I owe everything to almighty Allah, who is gracious, kind, and supreme. whose grace and glory has provided me with brilliant instructors, loving parents, and supportive siblings. with quivering lips and tearful eyes, we give thanks to the holy prophet Muhammad (p.b.u.h.) for awakening in us the core of faith in Allah and focusing Allah's mercy and compassion on him.

prof. dr. Merve Bulut YILGOR, empathetic demeanour, friendly demeanour, animated directives, observant pursuit, scholarly critique, encouraging perspective.

ABSTRACT

DEVELOPING SCALABLE DATA WAREHOUSE WITH BIG DATA REQUIREMENTS

ABDULLAH, Nada Muaad Abdullah

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Asst. Prof. Dr. Merve Bulut YILGÖR

Date: 04 / 2023

Pages: 73

As the number of systems that create data continues to increase, the ability of algorithms to scale up or down is crucial for the proper management of data. This is referred to by the research and development movement known as "Big Data." The National Institute of Standards and Technology defines big data as "a collection of vast datasets principally in the qualities of volume, variety, velocity, and/or variability that need a scalable architecture for effective storage, manipulation, and analysis" The primary purpose of this thesis is to develop a distributed scale-out storage system-based platform for big data analytics. This will make it possible to achieve low latency, massive scalability, and fault tolerance.

Keywords: Big Data, AI, ML, DL.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

AI	:	Artificial Intelligence
FSK	:	Frequency Shift Keying
IoT	:	Internet of Things
PAN	:	Personal Area Network
OFDM	:	Orthogonal Frequency Division Multiplexing
TDMA	:	Time Division Multiple Access

1. INTRODUCTION

It has been argued that we are now living in a "period of information explosion," which is defined by the continual generation of vast quantities of data that continue to rise in size. Numerous industries gather and analyze this tremendous amount of data. As the number of networked devices increases, the quantity of data that is being generated increases at an exponential pace. It is estimated that by the year 2021, the amount of digital information would have increased by a factor of nine, reaching 35 trillion gigabytes, compared to the amount accessible in 2011. This deluge of information originates from a multitude of sources, the number of which is always increasing. Here are a few increasing instances of very massive datasets.

i. Daily, the New York Stock Exchange produces around one terabyte of data.
ii. Facebook stores around 10 billion images, which occupies approximately one petabyte of storage space.
iii. Ancestry.com holds approximately 2.5 petabytes of data, but eBay creates over 50 terabytes of fresh data daily.
iv. Google handles 40,000 search inquiries every second on average.
v. The entire storage capacity of the Internet Archive rises by around 20 terabytes every month, bringing the grand total to approximately 2 petabytes.
vi. The Large Hadron Collider generates around 15 petabytes of data annually on average.
vii. Every year, 69 petabytes of data are produced by the field of radiology.
viii. the year 2020, between 25 and 50 billion Internet of Things devices will be active, according to certain estimates.
ix. A self-driving vehicle will create around 100 terabytes worth of photos, which is equivalent to 100 million megapixels.
x. Terabytes per second will be produced by the Exascale simulation method. The Square Kilometre Array Telescope can emit one hundred terabits per second (400 exabytes per year).

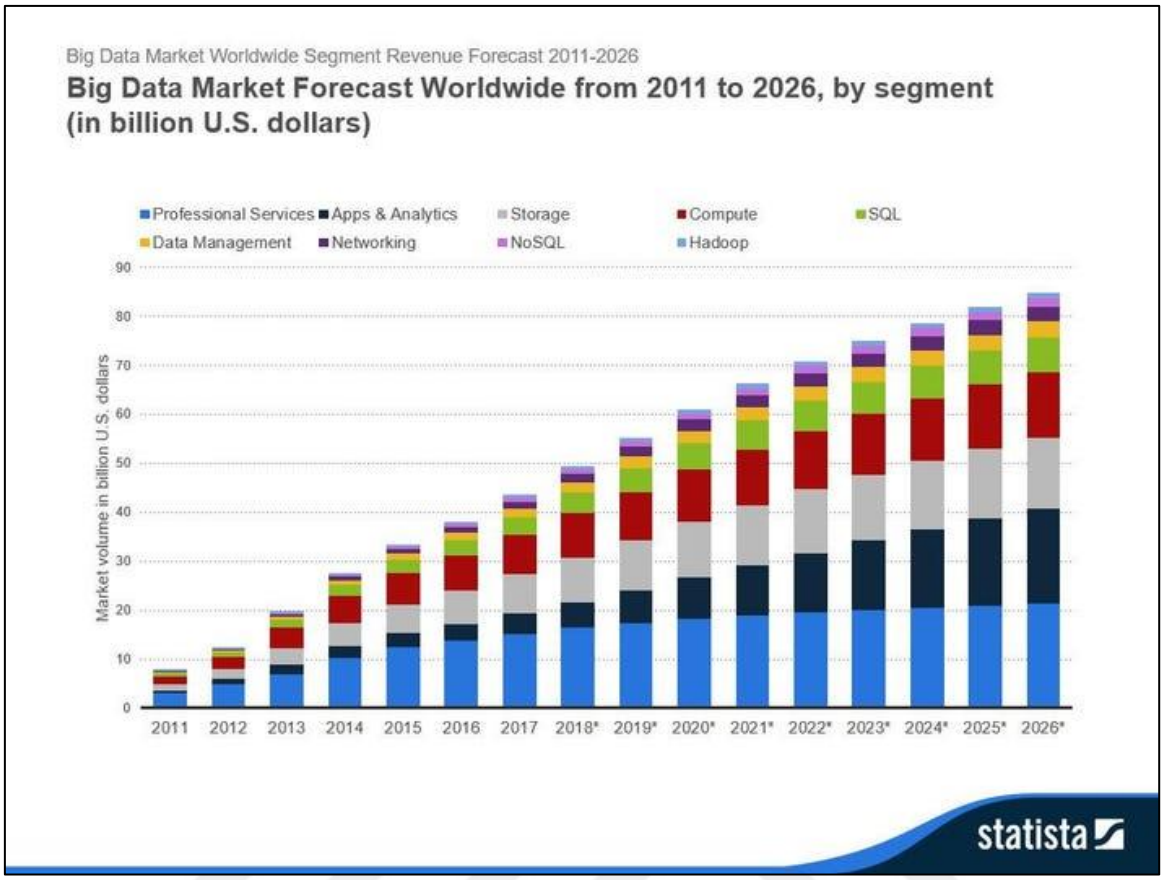


Figure 1.1: Estimated Big Data Market By 2026 [5].

As the number of systems that create data continues to increase, the ability of algorithms to scale up or down is crucial for the proper management of data. This is referred to by the research and development movement known as "Big Data." The National Institute of Standards and Technology defines big data as "a collection of vast datasets principally in the qualities of volume, variety, velocity, and/or variability that need a scalable architecture for effective storage, manipulation, and analysis" (NIST).

a. Big Data's volume refers to the total quantity of information that is being collected. Frequently, the size is indicated in tens of gigabytes or even exabytes. There are 10 billion daily messages received on Facebook, 4.5 billion daily click events, and 350 million daily picture submissions. Not only is it difficult to find a location to store all of this data, but it also requires a large amount of computer power to analyze it. In addition to this being the case, it is tough to locate a site to store all of this data.

b. The term "variety" refers to the many sorts of data structures. Although the advent of "Big Data" technologies led to the production of structured, unstructured, and semi-structured data, 80% of the world's data is now unstructured and cannot be readily entered into relational databases. Despite the fact that "Big Data" technology was responsible for creating all three forms of data, this is the case.
c. "Velocity" refers to the rate at which data is produced, stored, and processed; it is measured in megabytes per second. In this "period of data explosion," it is possible to process and store data incredibly quickly. When we speak to the "Internet of Things," we are referring to the phenomena of a growing number of devices connecting to the internet. This presents additional hurdles when it comes to real-time data processing. Companies in some sectors, such as the telecommunications industry, have been compelled to gather massive amounts of data at irregular periods for a very long time. Enormous Data's horizontal scalability, on the other hand, makes it feasible to effectively manage big data.
d. It is possible for data to undergo several modifications throughout time, including adjustments in volume, structure, and even presentation. There is a high probability that the information included in the data collected from several sources is insufficient. The incomplete, confusing, and diverse origins of the data have a substantial effect on their quality. As a consequence, it is essential that the data processing phase include both the validation of data and the monitoring of its history.
e. The value of data lies in its ability to reveal previously unseen patterns and trends, but only if it can be understood, managed, and integrated into a larger data network. Big Data is the product of several concurrent events, including the fast growth of artificial intelligence, data mining tools, and machine learning algorithms. In addition, it is driven by a process including the analysis of data, the extraction of knowledge, and the pursuit of experience-based values. It is a procedure that use machine learning algorithms to derive commercial value from data.

1.1 PROBLEM STATEMENT

Despite its widespread usage, the term "big data" has not been precisely defined. Data scientists have proposed defining "Big Data" as the "Extraction, Transformation, and Load"

(ETL) process for enormous amounts of data. Big Data is shown here according to five criteria: data volume, data velocity, data variety, data variability, and data value (the 5Vs). As more people, devices, and sensors are linked to the global network, the amount of data and the desire to share, exchange, and access the data will expand proportionally. The volume of data has reached a point where it cannot be processed in the traditional method. According to estimations performed by the Internet Data Center, the volume of data is expected to increase by a factor of 300 between 2005 and 2020, from 130 Exabytes to 40,000 Exabytes. In addition, the discovery of data insights has shifted information management towards a new analytic paradigm called as Data Science. Within this paradigm, applications must expand while ensuring they do not overload the 5Vs (volumes, velocities, varieties, variabilities, or values). This might give an explanation for the "Big Data phenomena." In order to be processed by the most current versions of Big Data technology, data must be partitioned and distributed over several data centers (such as HDFS, MapReduce, Spark, Flink, etc.). Two criteria that influence whether or not data processing is essential are the amount of data to be processed and the location of the data sources. The most challenging obstacle will be reducing the time required to process Big Data. Consequently, efficiency and scalability will be the defining characteristics of the data-intensive paradigm.

Figure 1.2: Top Major Concerns With Big Data Informatics [5].

The trend towards cloud computing is also accelerating the expansion of big data. The scalability of clouds on demand enables their usage for Big Data use cases, and the processing capacity of infrastructure is provided by the large data centers used to store cloud data. As a consequence of the cloud's ability to grow indefinitely, users are relieved of the responsibility of managing complex dispersed equipment. Customers may thus tailor their compilations to the specific processing requirements of their applications by focusing on renting and expanding the scope of their services. However, when there are difficulties with the availability and performance of cloud infrastructure, enormous financial losses are possible. As a result, making performance an explicit objective is an imperative must. Automated failure diagnostics and predictive analytics are essential for cloud service providers to deploy in order to manage their data centers proactively. This dissertation provides both autonomous and scalable analytics. This opens the door to the investigation of vast volumes of data, which may subsequently be utilized for the automated diagnosis of failures and the discovery of abnormalities, both of which contribute to the acceleration of cloud performance and availability. On this dissertation, I will describe various

enhancements that might be done to an existing cloud-hosted Big Data platform. This dissertation presents an approach that has the ability to increase the data center's precision and speed. This is achieved by forecasting the failure of compute nodes and jobs in Hadoop clusters and by recognizing the anomalous behavior of Big Data applications running on cloud architecture. This dissertation concludes with a mechanism for safe deletion that increases the user's perception of the transparency of a large-scale distributed file system.

1.2 OBJECTIVES

The primary purpose of this thesis is to develop a distributed scale-out storage system-based platform for big data analytics. This will make it possible to achieve low latency, massive scalability, and fault tolerance. The following is a list of the primary goals that the proposed platform for big data intends to achieve:

i. For the purpose of doing analysis on enormous amounts of diverse sorts of data.
ii. offering reliable search results in order to assist the process of making knowledgeable selections.
iii. With the aim of enhancing the query performance of a variety of analytical tasks.
iv. In order to get useful insights from very huge datasets, it is necessary to apply statistical methods.
v. The following is required for the proper installation of a big data platform in any storage infrastructure.

1.3 CONTRIBUTION

Following are the two most significant contributions made by this thesis:

Keeping the following aims in mind, we propose constructing a platform for big data analytics based on a distributed scale-out storage system.

i. low latency, scalability, and fault tolerance;
ii. we propose a data rebalancing approach for MATLAB to address the difficulties of excessive file migrations, lengthy file migration durations, and wasteful storage use.

In terms of speed, scalability, and dependability, the field of big data analytics is now confronting its greatest hurdles. The following are the three most significant contributions made to resolve these issues:

a. Combining the elastic hashing technique with the concurrent data processing of the MapReduce programming style may alleviate latency problems.
b. Combining the dispersion of data and programs throughout the Hadoop cluster with the "no metadata server" characteristic produces an astonishing linear scalability.

There are two primary contributions that must be done to tackle these issues:

a. For instance, we might combine a consistent hashing approach with MATLAB Dynamo's virtual nodes to reduce the frequency and duration of file migrations.
b. virtual machine (VM) migration across storage servers for enhanced use of available resources.

1.4 THESIS STRUCTURE

This is the structure of the thesis: In Section 2, we evaluate the evolution and use of smart grids to date. In Section 3, we contextualize each section. Section 4 gives a comprehensive explanation of our concept and its implementation, while Section 5 details the results of simulations using our model. In Section 6, we address the possible ramifications of our results.

2. LITERATURE REVIEW

2.1 INTRODUCTION

Never before in human history have companies had access to as much information as they have today. Every year, the amount of data collected by several businesses increases by a factor of two. The exponential growth will render it difficult for general-purpose hardware and software to keep up with the rising data scale [1]. As a direct consequence of the necessity to handle the issues provided by the big data trend, a wide variety of highly scalable data management systems have emerged. A variety of big data processing methods are now doing research in infrastructure architecture. NoSQL, which stands for "Not Only SQL," is the name of a fascinating phenomenon that is now occurring. This development started in early 2009 and is advancing rapidly.

Every day in the current day, a staggering amount of data from a wide variety of sources is generated (e.g., health, government, social networks, marketing, financial). There are other contributing factors, including the spread of smart devices and the growth of the Internet of Things. The operation of complex systems such as smart grids (Chen et al., 2014a), healthcare systems (Kankanhalli et al., 2016), retail systems such as Walmart's (Schmarzo, 2013), and government systems (Stoianov et al., 2013), amongst others, depends on the existence of dependable backend systems and distributed applications. Before the advent of the Big Data revolution, businesses had the technology in place to securely keep all of their archives for extended periods of time and to handle such massive data sets. In actuality, conventional systems are costly, have restricted data storage capacities, and are confined by inflexible management frameworks. They are incapable of addressing the Big Data requirements for scalability, flexibility, and performance. In reality, Big Data management demands a substantial amount of money, new techniques, and advanced technology. Big Data necessitates the ability to cleanse, process, analyze, secure, and enable granular access to huge volumes of data that are continuously updated. The value of data analysis in maintaining competitiveness, gaining fresh insights, and customizing services to the particular needs of each individual customer is gaining growing attention from businesses and industries throughout the world. Numerous companies in a vast number of nations have begun expansive initiatives due to the potential benefits that may follow from analyzing

enormous amounts of data. The United States of America was one of the first governments to see the potential of big data. The Big Data Research and Development Initiative was launched in March 2012 with $200 million in financing from the Obama administration. (Weiss and Zgorski, 2012). In July of 2012, the development of big data in Japan was given what is now considered a vital position in the government's overall technology strategy. (Chen et al., 2014b). In 2012, the United Nations published a report titled "Big Data for Development: Opportunities and Challenges." It is intended that by exposing the most critical obstacles related with these issues, this would initiate a dialogue about how Big Data might contribute to global progress. Numerous Big Data initiatives conducted in various regions of the globe led to the development of Big Data models, frameworks, and technologies at the forefront of their fields. These developments dramatically expanded the quantity of accessible data storage, facilitated parallel processing, and permitted near-real-time analysis of a broad range of heterogeneous data. In recent years, various solutions have been created to better secure the privacy of people' information. These solutions provide more adaptability, scalability, and efficiency compared to prior types of technology. As a result of continual technical development, the price of a variety of data storage and processing choices continues to decline (Purcell, 2013). The use of a wide range of models, algorithms, softwares, hardwares, and technologies has enabled the extraction of information from vast quantities of data. They want to provide a more precise and reliable result for applications using Big Data. However, in a circumstance such as this, picking the best technical solution from the multiple available options may be lengthy and difficult. There are several factors to consider, including technological compatibility, deployment complexity, cost, efficiency, performance, reliability, and support, as well as possible security risks. Numerous studies on Big Data have been conducted, but the bulk of them have focused on the algorithms and methods used to manage this data rather than the technology used to handle the data itself (Ali et al., 2016; Chen and Zhang, 2014; Chen et al., 2014a).

2.2 RELATED WORK

In this part, we will examine some of the previously conducted research on massive data streaming analytics.

The section [13] provided an overview of several tools, strategies, and techniques for big data analytics. This was accomplished by classifying the literature on big data analytics based on the individual research foci of each article. This paper provides a comprehensive literature assessment of "streaming" analytics for massive data volumes, distinguishing it as an innovative piece of research.

The authors of the paper [19] provide a comprehensive analysis of the use of big data analytics to online purchasing. The research examined the e-commerce environment for big data analytics, including its features, definitions, business values, kinds, and problems. In a similar line, [20] did a research on the use of big data analytics to the administration of technical and organizational resources. The focus of the research was on reviews highlighting big data issues and big data analytics approaches. Although it is not the major emphasis, especially when it comes to the streaming of massive volumes of data, the evaluations are thorough.

The authors of [21] used a systematic mapping technique to provide the current state of empirical research and application topics in the field of big data. A survey of big data technology and machine learning techniques, with an emphasis on anomaly detection, has been published in [22]. This review was comparable to the last one. The authors give a literature review in which they seek to define the boundaries, potential uses, and potential risks of big data analytics in the context of the healthcare industry. This review is available at [23]. The authors of looked into four unique streaming strategies for huge data volumes and reported their results in [2]. This article's research not only examined the tools and technologies that can stream large volumes of data, but also the processes and procedures that may be utilized to analyze such streams. In addition, the authors [2] did not provide a comprehensive description of the selection process's methodology.

3. BIG DATA WAREHOUSES

In order to frame the objectives of this dissertation, this chapter presents the main concepts associated with the various themes addressed in this document. That said, the concept of Big Data and its inherent characteristics, opportunities and problems will be presented. In order to continue the characterization of this phenomenon, a description is made of the different databases, traditional, NoSQL and New SQL. Also in terms of storage, a comparison is made between traditional Data Warehouses and Big Data Warehouses, ending with an approach to the various data models. Finally, the importance of the presence of Data Governance in Big Data Warehouses is addressed, highlighting the importance of Data Profiling tasks and processes for the acquisition and production of metadata.

3.1 BIG DATA

Over the last few years, the volume of data has increased on a large scale due to a number of factors, namely, the increase in data sources and data collection mechanisms that, consequently, produce large volumes of Stored data are increasingly relevant and, in the future, will be an essential component for the success of organizations In certain contexts, namely in Industry 4.0, the analysis of data whose variety and volume are high is seen as one of the most important aspects to produce value for organizations However, the volume, speed and heterogeneity of data have imposed great challenges on traditional technologies, namely in their acquisition, management and processing of data, failing to be sufficiently efficient and effective in their response time Decisions that were previously based on assumptions are now made based on information from data, one of the main challenges for organizations being the need to obtain relevant information at the right time In view of this, due to the fact that traditional technologies present great challenges in relation to factors such as response time, scalability and performance, there is a need to bet on technologies and solutions based on Big Data concepts

Recently, the study carried out by the consultancy Excelacom, shown in Figure 3.1, illustrates the data produced by different sources in just one minute. As can be seen, in

addition to the increase in the volume of data on the various platforms, the potential of information that can be extracted and analyzed in order to make an organization's decision-making process more effective and efficient is also highlighted. Big Data is a topic that poses numerous challenges, not only in the ambiguity regarding its definition, models and architectures, but also in the challenges related to the life cycle of Big Data. Although the term is associated with large volumes of data, this is far from being its only characteristic or its only challenge. The term Big Data began to be used in the late 1970s as a parallel database system, aiming to respond to the increase in data volume. technologies associated with Big Data concepts emerged, namely the GFS.

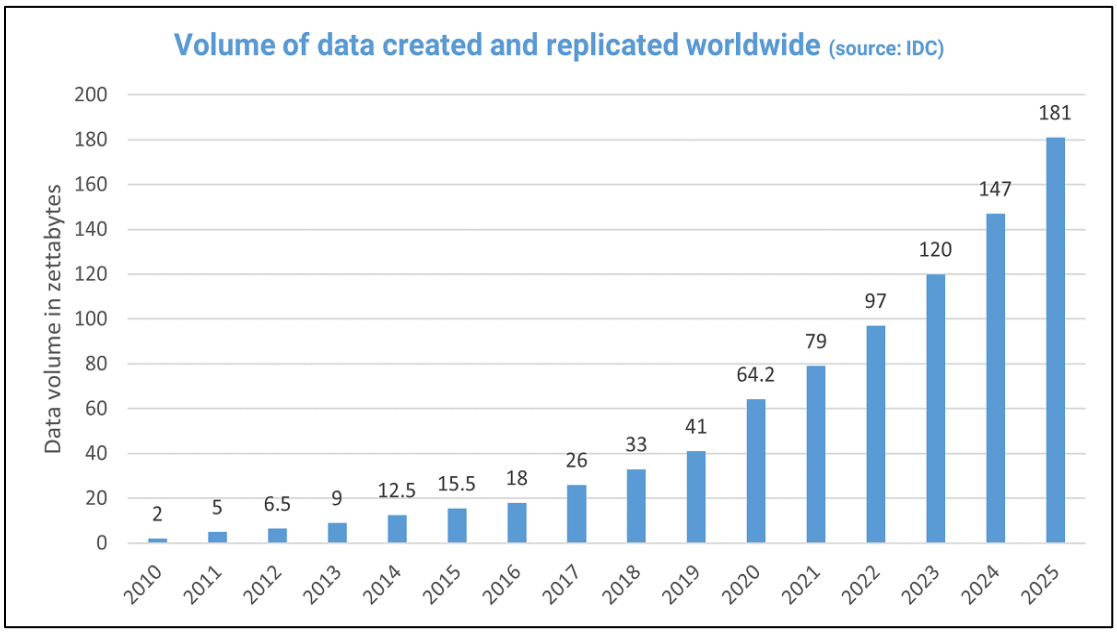


Figure 3.1: Data Volume Growth Rate [12].

3.1.1 The Concept

According to Krishnan (2013) [12], the biggest phenomenon that has captured the attention of large organizations since the emergence of the Internet was Big Data. However, despite its importance being duly recognized, the same is not true when it comes to its definition, usually shrouded in ambiguity in establishing a limit in the classification and quantification of "large volumes of data. the definitions of the various authors differ by the way they are presented, since some authors define Big Data by its characteristics, others are based on the argument of the increase of traditional data

sources with the addition of unstructured data and others through the quantification of data. highlights the importance of Big Data in organizations that currently have access to a vast set of data, but are unable to extract value from them, mainly due to their semi-structured or unstructured format. Given this, the authors define Big Data as data that cannot be processed or analyzed through traditional processes or tools.

Big Data refers to data that exceeds the processing capacity of traditional databases. The author complements his statement by adding that the volume of data is very vast, generated at high speeds and is not suited to the architecture of traditional databases. Therefore, it is necessary to select alternative methods and techniques to process and analyze these data in order to extract value from them. defined Big Data as the vast volumes of data available at various levels of complexity, generated at different speeds and varieties that cannot be processed using traditional technologies, processing methods and algorithms. Big Data allows access to large volumes of data that can be useful to obtain patterns and trends implicit in them, being an opportunity for organizations to increase their effectiveness and efficiency in terms of business processes. the exponential growth of the area in recent years, warning of the lack of rigor in associating Big Data with data storage and analysis. The authors recognize that the lack of rigor is mainly due to the difficulty in quantifying

Big Data, highlighting Intel as one of the few organizations that defines Big Data by its quantification, associating Big Data with an organization that produces an average of 300 terabytes (TB) of data on a weekly basis. However, argue that defining Big Data by the inadequacy of traditional technologies in processing large volumes of data can be relatively dangerous, due to constant advances in technologies, particularly quantum computers. Finally, the authors conclude that Big Data definitions include at least one of the following aspects: volume, complexity and technology. In contrast, the authors even end up including in their own definition of Big Data, the storage and analysis of large and complex data sets using appropriate sets of techniques and methods. Not only researchers, but also organizations, have a point of view on this issue. In 2010, Apache Hadoop recognized Big Data as the dataset that cannot be extracted and processed through conventional computing resources. Accordingly, Oracle centralizes its definition at the infrastructure level, stating that Big Data derives from the value extracted from relational databases with information

about the business, that is, information about transactions from ERP's (Enterprise Resource Planning) with the addition of new unstructured data sources.

there are several different opinions on the subject, however, it is notable that some of them converge in some aspects. faced with the ambiguity of the topic, carried out a study with the objective of proposing a formal definition of Big Data, based on its characteristics and in coherence with the definitions normally used. defines Big Data by its volume, speed and variety of data, motivating the application of specific methods and technologies, with the purpose of creating value for society and organizations. From the perspective of the author of this document, Big Data is associated with the three main characteristics, volume, variety and speed, which exceed the capacity in the extraction, enrichment, transformation and storage of data in tools, traditional architectures and models. Some characteristics mentioned above are not directly considered in this definition, since they are inherent to the use of Big Data (for example, the use of Big Data is due to the complexity that the data present and the purpose of this use will be, clearly, to obtaining value from its implementation).

3.1.2 Main Features

In its beginnings, Big Data appeared mostly associated with large volumes of data, however, its volume is far from being its only characteristic or its only challenge, to the detriment of the growth of the volume of data, defines the challenges and opportunities in its management in three different dimensions, characterizing Big Data by the increase in its volume, speed and variety. Volume refers to the amount of data generated and extracted from the various data sources, the speed at which data enters and leaves the storage system, and the variety addresses the different formats and data sources. Currently, organizations mostly use this definition as a model to classify Big Data, also noting that these are the characteristics mostly used by the community, represented in Figure 3.2.

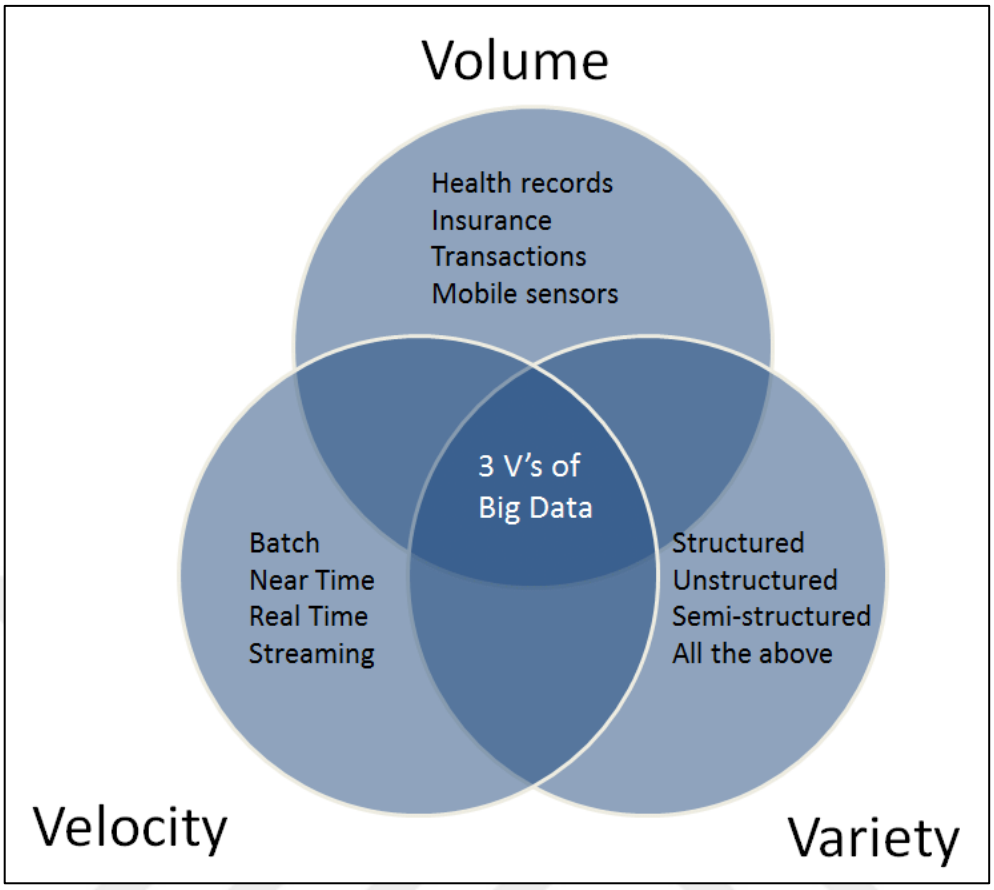


Figure 3.2: Model Of The 3 V's [14].

As a result of the previous statements, each of these characteristics are described and illustrated in Figure 3.2. this characteristic as the biggest obstacle in the processing of large volumes of data, assuming that non-aligned data structures and inconsistent semantics pose significant challenges in their analysis. In order to reduce the complexity in the processing of various data formats, portrays the availability of metadata in order to identify the content of this data, warning that the absence of metadata may lead to delays in processing, from the extraction to storage. Just as the volume and variety of data has changed, so has the speed at which this data is generated, stored and analyzed it concerns both the speed at which data are produced and the speed at which they must be analyzed. The rise of digital devices causes an increase in the rate at which data is generated, driving a need for real-time analytics. this characteristic applies to the speed at which data flows to the repository, stating that the volume and variety of data mentioned above are a consequence of the speed at which they are produced. The data produced come from heterogeneous sources, ranging from batch to streaming. In Big Data contexts, not only the speed at which data is stored

in the repository is important, but also the entire process up to decision making, in order to maximize efficiency in business processes Currently, the data produced cannot be extracted, stored and analyzed using traditional technologies, resulting in the need for a processing system capable of working with scalable speeds and extremely variable sizes in a short period of time Due to what was mentioned above, around the model of the three Vs attributed to Big Data, identifies three new characteristics that arise from the intersection between volume, variety and speed as illustrated in Figure 3.3.

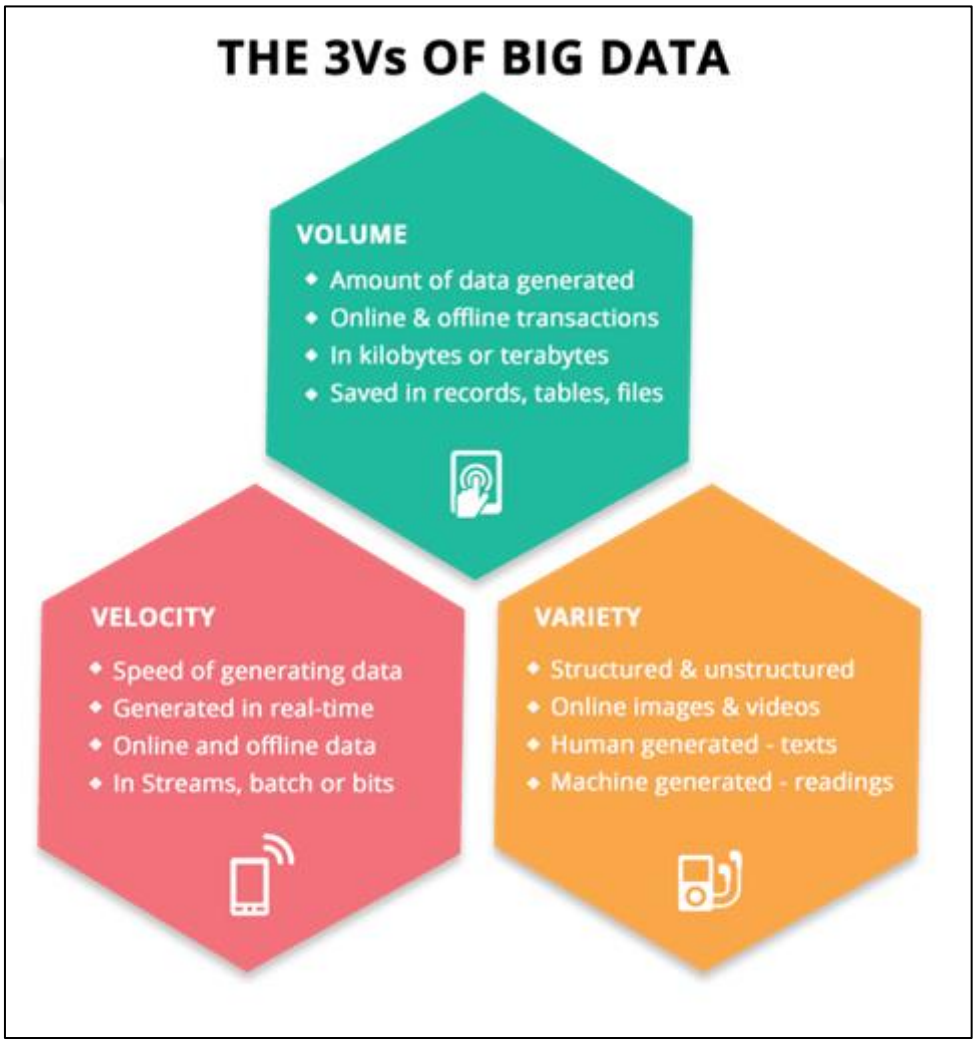


Figure 3.3: 3Vs Model With Additional Features [15].

a. Ambiguity: this characteristic manifest itself between variety and volume. The lack of metadata is one of the main factors for its presence. For example, an “M” or “F” can mean a gender (Male or Female) or it can mean Monday (Monday) and Friday (English: Friday);

b. Viscosity: this characteristic manifest itself between volume and velocity, measuring the resistance to flow in the volume of data, and may be represented in data flows, business rules and technological limitations.
c. Virality: this characteristic manifest itself between variety and speed. Evaluates and describes the speed at which data is shared on the network. For example, re-tweets that are shared from an original tweet are good practice for evaluating a topic or trend.

Over time, several authors have argued that the characteristics presented above are not enough to characterize Big Data. Consequently, characteristics emerge that are constantly mentioned, such as value and veracity. The remaining characteristics identified in the literature, namely, validity, volatility, variability and complexity, not so mentioned by the community, arise through the extension of the 5 V's model proposed. states that this feature has a special reason. Contrary to the previously mentioned features, the author assumes that this feature is the intended result of processing large volumes, speeds and varieties of data, since the interest is to extract the maximum value from this processing. The data that are initially extracted, in their original form (without treatment), typically have a reduced value in relation to their volume. The joining of various types of data has the purpose of extracting knowledge that was previously unknown, with the ultimate goal of obtaining competitive advantages for organizations. veracity refers to the level of confidence associated with the various types of data. Data quality is an important requirement in Big Data, however, even with the best data processing methods it is not possible to remove the unpredictability that is present in some data, such as weather conditions, economic factors and the feelings, which although uncertain by nature, can be valuable in the long run. The speed with which data flows does not allow for excessive periods of time to be spent on its treatment, leading to the need to establish mechanisms to deal with inaccurate and accurate data. Variability refers to the variation of velocity in the data stream, according to different peaks and periodic reductions. Data validity evidences having the same concept of veracity. In fact, validity is framed in the precision and accuracy of the data, in relation to the intended use. In other words, the data may not contain any errors of veracity, however, they may not be considered valid if they are not properly understood. The complexity is

highlighted by the fact that Big Data imposes the challenge of integrating, processing and transforming data from different data. Big Data volatility is associated with an organization's data retention policy, referring to the period of time in which data must be kept and the moment from which it can be destroyed Figure 3.4 summarizes the various features identified throughout this subsection.

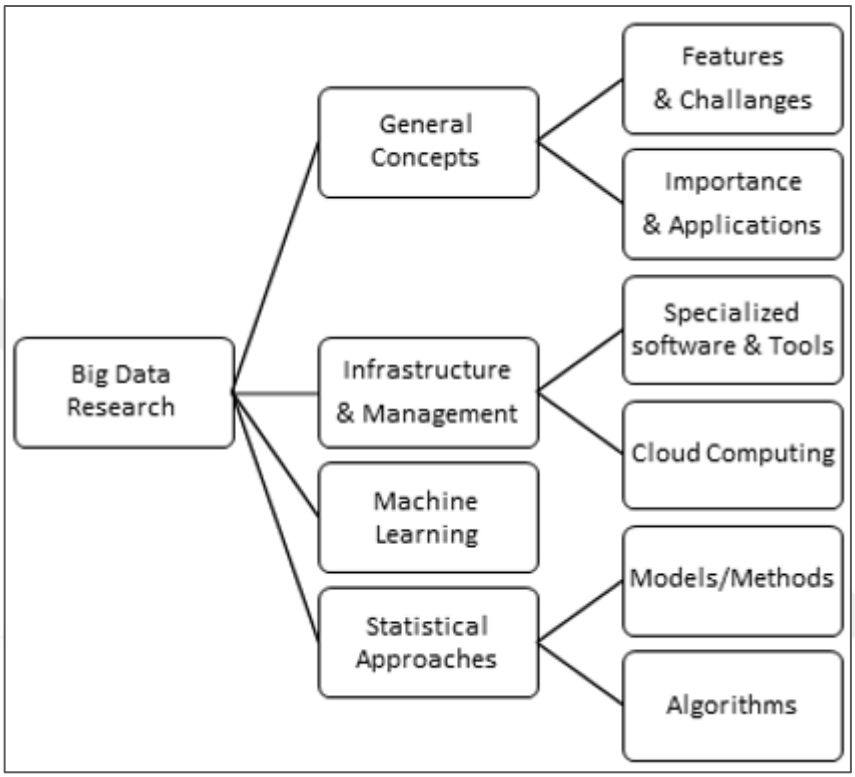


Figure 3.4: Main Characteristics Of Big Data Identified In The Literature [13].

3.1.3 Big Data Opportunities, Problems and Challenges

Recently, organizations have shown significant interest in the potential of Big Data, as a result of its potential and the creation of competitive advantages for organizations that have great plans to accelerate their research. While the amount of large volumes of data is dramatically increasing, a set of opportunities, problems and challenges arise which will be addressed in this subsection. the basis for the emergence of new opportunities is mainly due to the increase in the volume of data. As a result of the volume of data extracted from Big Data, the opportunity arises to provide personalized services, adapting them to consumers. This ability to store large volumes of data, at different speeds and varieties, makes it possible to store the historical data of network traffic,

making it possible to identify the source and destination of an eventual threat, thus reinforcing the security of the internet. in the future, extracting knowledge from data will be a competition between organizations. The authors go on to enumerate a set of advantages that can be obtained through Big Data, as illustrated in Figure 3.5, in which it is possible to obtain a framework of Big Data opportunities at the level of organizational processes. It is noteworthy that 51% of the organizations surveyed believe that Big Data will help improve the efficiency of their operational processes.

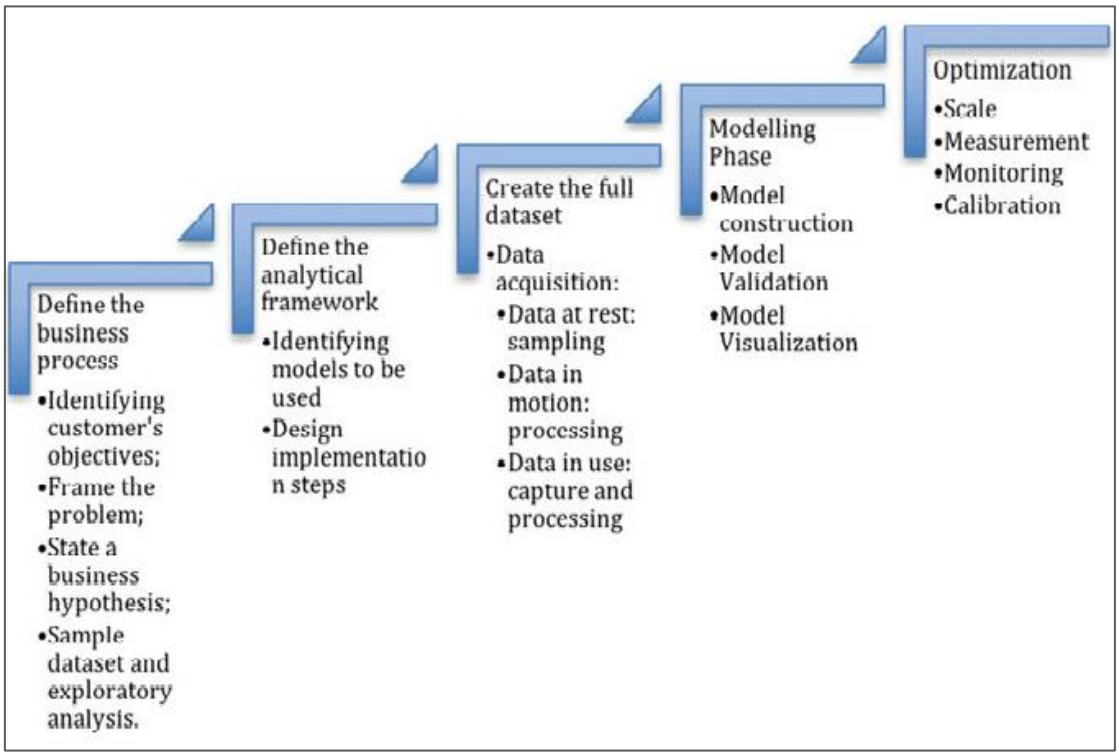


Figure 3.5: Big Data Opportunities In Organizational Processes [22].

Data can play a crucial role. In view of this, through the contributions, Figure 3.6 presents not only the areas in which Big Data can fit, but also shows some examples of contributions that can be performed in them.

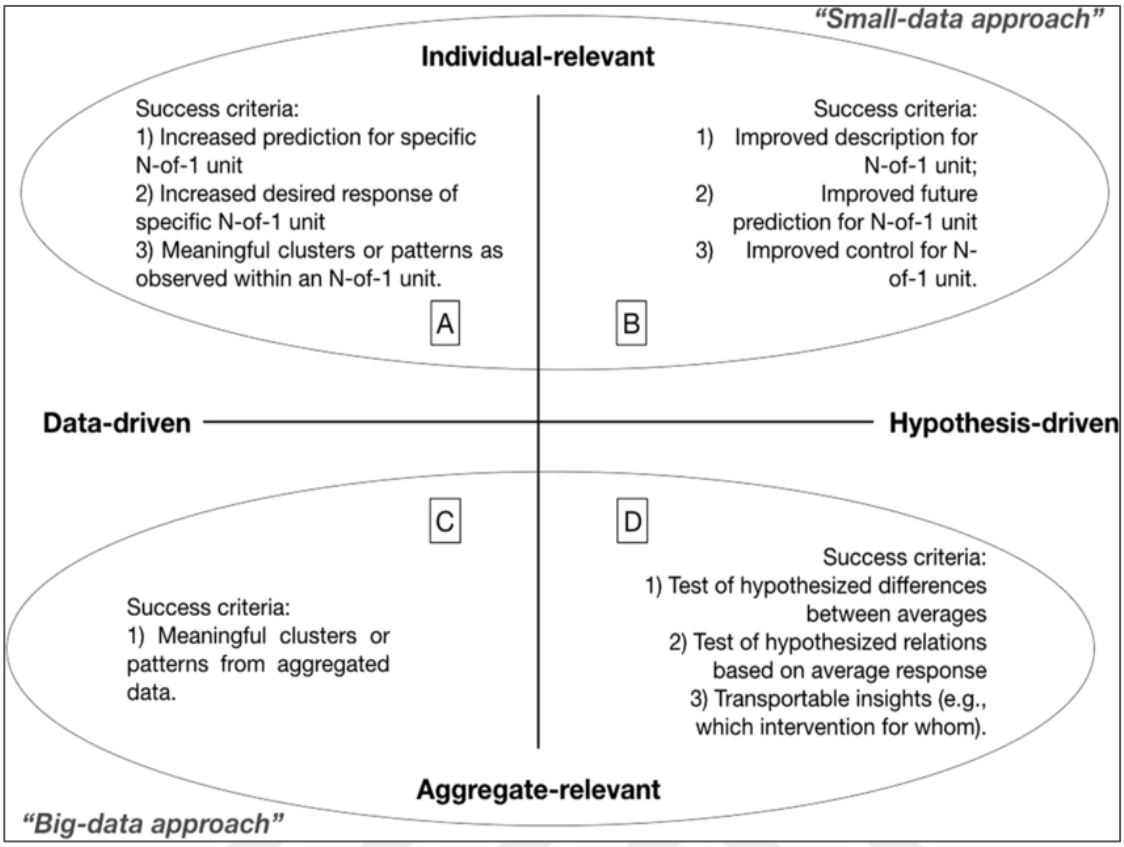


Figure 3.6: Big Data Intervention Areas [17].

However, despite being a domain that provides many opportunities, Big Data also poses several challenges and problems, due to the need to select methods and technologies to extract, store, manage and analyze large volumes of data, at different speeds and varieties, something that cannot be supported by conventional systems. Next, a set of challenges regarding Big Data will be presented, which are divided into four categories, In each of the categories ("General Big Data Dilemmas", "Challenges in the Big Data Lifecycle", "Safe, Private and Monitored Environments in Big Data" and "Organizational Change"), a brief description will be given about it and presented a set of identified challenges and/or problems.

a. General Big Data Challenges

This category includes a set of more general challenges, including the lack of rigor in the definition of Big Data and the lack of standards regarding models and architectures. Such challenges occur because Big Data is not a structurally defined theme and the

existing models do not undergo rigorous verification. Starting with a challenge that is addressed earlier in this dissertation, the concept of big data, is regularly viewed as commercial speculation rather than a topic of scientific investigation. there is a need for a rigorous and holistic definition of Big Data, a structural model and, finally, a theoretical system of Data Science. The authors mention the lack of standardization in Big Data environments, in issues related to the evaluation of data quality and benchmarking mechanisms. From the same point of view, the lack of benchmarking mechanisms for the purpose of comparing different technologies is constantly aggravated by technological evolution in Big Data contexts. An emerging challenge for researchers and users of Big Data is "quality vs quantity". More and more users have access to large volumes of data, believing that with enough data they will be able to explain any Big Data phenomenon. In contrast, a user of Big Data must focus on data quality, in other words, although the data is not available in its entirety, there is an amount of high quality data that can be used to draw accurate and accurate conclusions. of value. Often, when dealing with large volumes of data there are some questions that are difficult to answer, namely, which data are considered irrelevant and which data selection techniques are considered relevant? How is the value extracted from the data estimated? How is the authenticity of data guaranteed? It is regularly debated how Big Data better represents the population compared to an inferior dataset. Obviously, there are issues that vary depending on the context, however, the authors are convinced that more data does not mean better quality. Challenges in the Big Data Lifecycle however, the authors are convinced that more data does not mean better quality. Challenges in the Big Data Lifecycle however, the authors are convinced that more data does not mean better quality.

These challenges refer to technical difficulties in Big Data environments, namely in the mechanisms of extraction, integration, cleaning, transformation, storage, processing, analysis and data governance.

Organizations have been undergoing major technological changes, increasing the number of devices capable of producing data. The increasing amount of data "explodes" each time a new storage medium is created (Kaisler et al. 2013; Katal et al. 2013). As a result, challenges arise in terms of data storage and processing. Philip Chen and Zhang

(2014) state that the process of storing and processing Big Data has undergone a change in terms of storage devices, storage architecture and data access mechanisms. That said, there is a need to rethink these mechanisms in order to make data input and output (I/O) increasingly efficient, since data availability is one of the main priorities for knowledge extraction. Ensuring data quality and adding value to them through their preparation becomes a challenge Data quality influences the use of Big Data as data that is of poor quality is wasting resources in processing and corresponding storage. Data quality is mostly defined by its accuracy, redundancy and consistency. Therefore, there is a need to investigate new methods and techniques in order to make data enrichment mechanisms more effective and efficient When a human being absorbs information, a great deal of heterogeneity is comfortably tolerated. However, the same does not happen with traditional algorithms that can only deal with homogeneous data Heterogeneity comes from multiple sources and different types of data, namely structured, semi-structured and unstructured data. Scalability has become crucial for data storage and analysis. Handling large volumes of data requires redesigning databases and algorithms in order to extract value from the data Previously, traditional data processing systems needed to consider parallelism between cluster nodes, however, the current concern focuses on parallelism within a single node In Big Data environments, data governance is a complex process with regard to control and authority over large volumes of data from different sources. Manipulating data in a heterogeneous environment to plan and ensure data traceability becomes almost impossible without the appropriate governance tools Concluding this category, states that organizations face a large set of challenges in the Big Data lifecycle. Overcoming these challenges depends on a number of factors, namely the maturity state of the organization, the use of traditional systems and the use of incompatible data formats that can impose difficulties in their integration and extraction of value from Big Data.

b. Secure, Private and Monitored Environments in Big Data

Currently, data privacy and security is one of the issues that most concern organizations It is important to mention that organizations must architect Big Data security models in order to achieve a high degree of precision in the prevention and specification of threats states that Big Data solutions face major challenges related to data privacy. This

challenge includes two main aspects, namely, the protection of personal privacy during the data extraction process (habits, personal interests) and protection of data privacy throughout its flows. In this category, data ownership also presents itself as a critical challenge, particularly in scenarios related to social networks. Although large volumes of data are stored on organizations' servers, it does not necessarily imply that they are the owners of the data (Chen et al. 2014). The real owners are the people who create the pages or accounts; however, organizations act as if the data were their property and legislation tend to benefit them, allowing them not to permanently delete the data, even when users require it. Users may not mention some facts about themselves, however there is a fear about the inappropriate use of their data through the crossing of personal information with large sets of high-quality data, That said, sensitive information concerning users is regularly dealt with, namely in matters related to health or finances. Given this information, the application of laws and regulations is relevant. In order to obtain value from the use of Big Data solutions, data must be accessible and usable and, therefore, organizations must comply with these requirements by adhering to this regulation. In compliance, such requirements guide organizations to reflect and make decisions about the selection of data, that is, which data can bring value to the organization, assuming that there are data sets that cannot be used due to regulatory compliance. poses a set of questions that must be considered: What are the rules or legislation that must exist regarding introducing data that contain personal information, from various sources, into a single data repository? Should the rules apply to the entire repository or only to parts that contain relevant information? What rules should exist for the prohibition and storage of data about individuals? in a single data repository? Should the rules apply to the entire repository or only to parts that contain relevant information? What rules should exist for the prohibition and storage of data about individuals? in a single data repository? Should the rules apply to the entire repository or only to parts that contain relevant information? What rules should exist for the prohibition and storage of data about individuals? In summary, regarding this category, it can be considered that the various authors mentioned above feel the lack of defined standards regarding privacy and data security and that it will certainly be a topic in which legal issues must be rethought in terms of simplicity. copying, integrating and using data from different people on a recurring basis.

c. Organizational change

Big Data is one of the most recent trends today and it is certainly a topic of interest to many organizations. However, managers of organizations often do not understand the value arising from this phenomenon and especially how to extract this value One of the main challenges in this category is to understand the value from Big Data, which is mainly due to the lack of knowledge in performing data analysis, presenting as the main obstacle the transformation to a data-oriented organization In many business areas, organizations need to monitor trends in order to obtain competitive advantages over their competitors, however, states that occasionally organizations lack talent, rigor and initiatives to implement Big solutions. Date.

Driving the paradigm shift in the organization so that it is possible to implement a Big Data solution requires the introduction of data analysis in its operational functions and in the main business areas, transforming business processes, providing highlights related to customers, products and services (Costa & Santos, 2017). McAfee and Brynjolfsson (2012) argue that implementing Big Data solutions poses numerous challenges, namely the need for a strong leadership component, data scientists capable of understanding and manipulating Big Data tools and the need to change organizational culture.

3.1.4. Data Processing

Data processing has been a complex topic to address since the early days of computing. According to data processing can be defined as the extraction, processing and management of data for the purpose of providing information to end users. Traditional data processing follows the life cycle in an environment where the data first undergoes analysis and, later, based on the analysis, a data model and a database structure are designed in order to process and store them. In this approach, the extracted data are structured and discrete in relation to their volume, since the entire process is pre-defined based on your requirements. Domains such as data quality and cleanliness are not labeled as problematic, as they undergo treatment in traditional systems However, the lifecycle of traditional data processing typically differs from data processing in big data environments. In the Big Data processing life cycle, represented in Figure 9, there is no

need to start the process by transforming the data so that it "fits" in the relational model. Data is first extracted and loaded, typically to a distributed file system, and later, the metadata layer will be applied to the data and the data structure for its content will be created. Once the data structure is conceived, the data undergoes the necessary transformations. The final result of this processing will allow the removal of relevant analyzes and patterns in the context in which the data are inserted

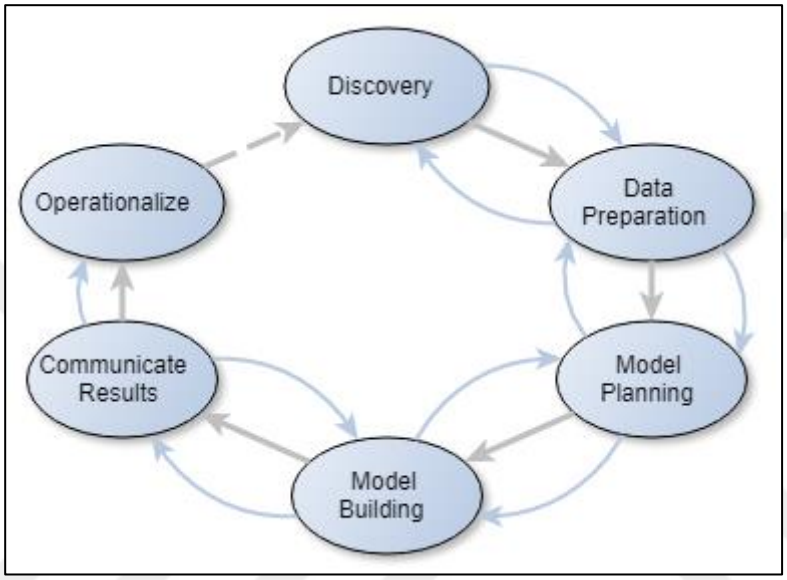


Figure 3.7: Big Data Processing Lifecycle [23].

states that to obtain a positive performance in processing large volumes of data at different speeds and varieties, a file-oriented architecture is necessary. The author continues with his opinion, mentioning that in order to design an infrastructure and efficient processing, there is a need to understand the flow of data to process Big Data. Given the difficulties inherent in data processing in Big Data environments, the author proposes a superficial view of the Big Data processing flow, shown in Figure 3.8.

The following describes each of the four main phases presented by the author for data processing in Big Data environments:

i. collection: at this stage, data are collected from different data sources and loaded into a file system, typically called a staging area.
ii. Loading: in this phase, the data is loaded with the metadata application, applying the data structure for the first time, presenting itself ready for the transformation phase.

iii. Transformation: at this stage, the data is transformed according to the business rules. This phase includes multiple steps for its execution, sometimes complex due to its type of content.
iv. Data extraction: at this stage, the data set can be extracted for various purposes, namely, analysis tasks, operational reports, integration into a Data Warehouse, visualization, among others.

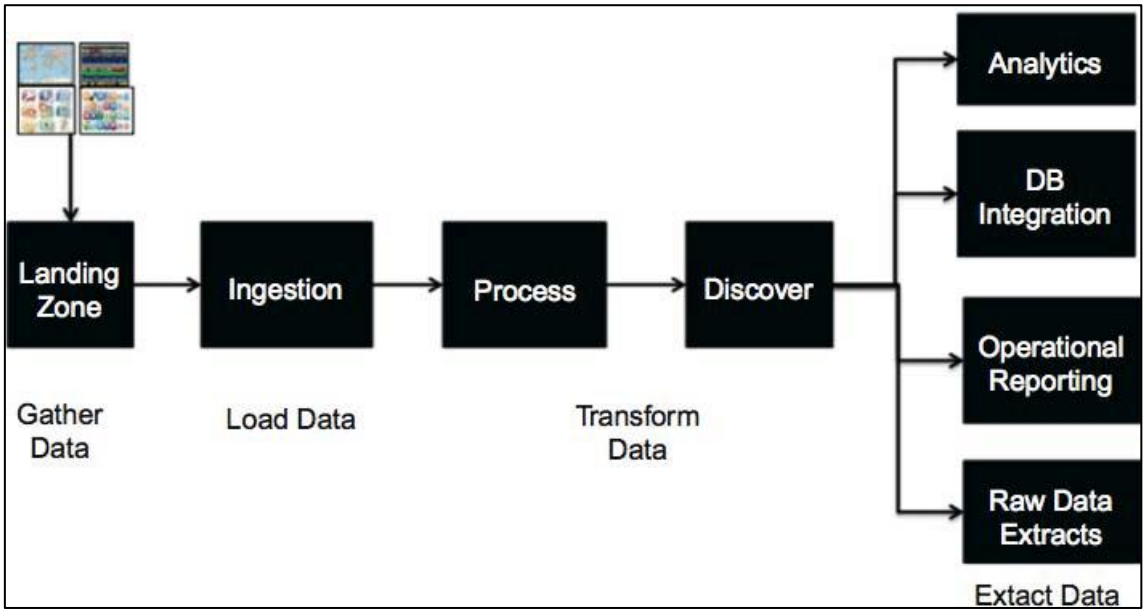


Figure 3.8: Big Data Processing Flow [23].

After completing the main phases of the Big Data processing flow, which allow processing data in a flexible way, it is also important to mention the types of processing that exist in a Big Data environment. data processing can be classified into these three types:

v. Batch Processing: this type of processing is used to process data in batch format, through its extraction, processing and loading to the predefined destination. Data is distributed across the file system, split into smaller files. The disadvantages of this processing model are the inability to perform recursion and interactivity tasks;
vi. Real-time processing: in this type of processing the data is processed as it arrives at the data repository and the results are almost immediate. It is characterized by the ability to execute queries in parallel, instead of sequences in batch formats. Previously, this type of implementation was characterized by its high cost, however,

nowadays the prices of RAM memories are becoming more accessible and this type of implementation is more feasible compared to the past;

vii. Stream processing: in this type of processing, data is continuously processed, obtaining results as new data emerges.

3.2. DATA STORAGE

Organizations are constantly being challenged in different areas and the challenges emerge from stakeholders, technological information, business needs, among others. In the context of information technologies, the concern with the evolution of the ability to extract, store and process large volumes of data has increased the need to develop new environments, using new technologies that require changes in organizations' information systems and their analytical capabilities. In the presence of characteristics such as the volume of data, in Big Data, data storage is one of the most important topics. That said, this section begins with a description of SQL (Structured Query Language), NoSQL (Not Only SQL) and NewSQL databases. Subsequently, the concept of Data Warehouse and the distinction between analytical systems that typically support Data Warehouses and transactional systems (OLTP) that support operational databases follows. Finally, the concept of Big Data Warehouse is approached and ends with an approach to data models.

3.2.1 SQL, NOSQL and NEWSQL Databases

In this subject of SQL, NoSQL and NewSQL databases, there is a great challenge that fits in the selection of the most appropriate database system for a given context. The philosophy of “no one sizes fits all” supported, portrays what was mentioned earlier. That said, given the diversity of available databases, it is difficult to provide a single solution (single database model) that is suitable. presents the main questions that should be considered when selecting the storage system: should you use a SQL, NoSQL or NewSQL system? Within NoSQL systems, what type of database should be used? What operations are the systems intended to perform? among many other questions. The answer to this set of questions reveals a high degree of experience in the area, This diversity is reflected in Figure 11, which represents the separation between relational and non-relational databases.

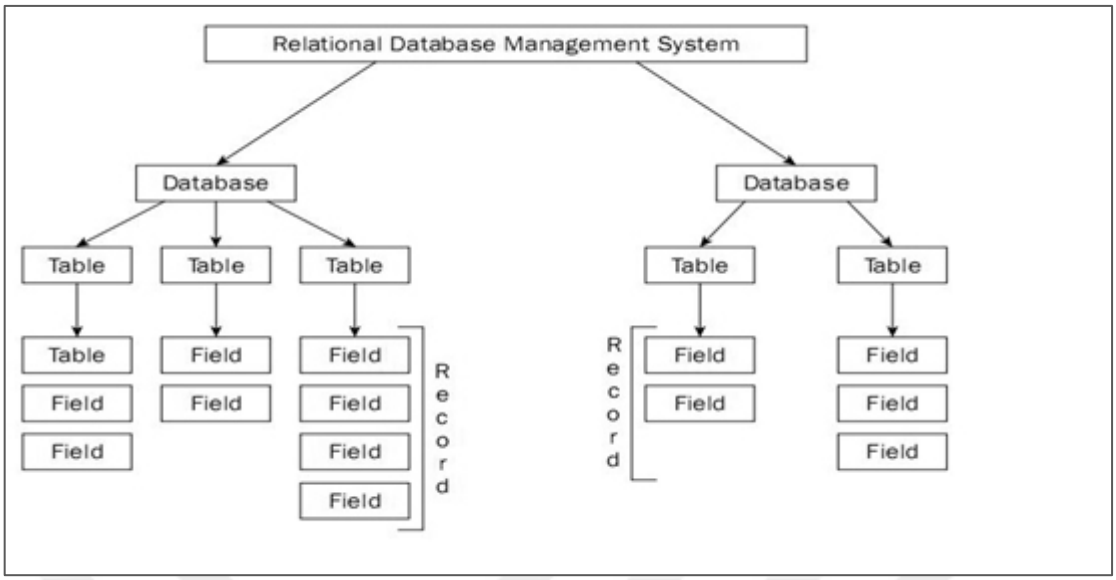


Figure 3.9: Categorization Of The Variety Of Databases Available [24].

There are three main database systems, namely RDBS (Relational Database Systems) or SQL, NoSQL and NewSQL. SQL databases are supported by a relational model and are characterized by the use of several types of keys, namely, the primary key, which is called the most important key in the table due to the identification of each row inserted in the table in a unique way, foreign keys, among others. This database model uses the ACID properties (Atomicity, Consistency, Isolation, Durability) that are the most relevant features of SQL databases Despite the qualities that come from ACID transaction processing, Consequently, NoSQL and NewSQL databases emerged. The popularity of non-relational databases has been increasing due to their characteristics, namely, speed, accessibility and scalability Although NoSQL is treated as “Not Only SQL”, it was initially thought of as a combination of only two words “No” and “SQL”, which would imply a separation from the SQL language, claim that its definition implies separation from the relational model that supports RDBS, rather than from its language. recognize that NoSQL databases offer a set of advantages over traditional databases, which are presented below:

i. Flexibility in data representation: NoSQL implementations present a flexible representation of data. Once defined, the structure is available for possible transformation over time (addition of new attributes to records);

ii. Development time: in some scenarios, the implementation time of NoSQL databases is reduced, since it is not necessary to deal with the complexity of SQL queries, due to not supporting relationships between tables;
iii. Speed: NoSQL databases feature more efficient read and write (I/O) operations and faster data aggregation mechanisms;
iv. costs: most NoSQL databases are open-source;
v. Scalability: Unlike traditional databases, NoSQL databases are characterized by their horizontal scalability. Relational databases require a lot of complexity to guarantee ACID properties.

However, Bonnet (2011) identifies some weaknesses:

i. Data consistency: provide no guarantees on data consistency. Users implementing NoSQL solutions have to keep in mind that the main concern of these databases is not consistency. NoSQL has no ACID transactions;
ii. Data migration: Data migration from SQL to NoSQL systems is not a trivial process, requiring some experience in NoSQL data models in Big Data contexts.

Taking into account the analysis of the characteristics and contributions that NoSQL databases can present in relation to SQL databases, it can be seen that NoSQL databases offer advantages in terms of scalability, availability and accessibility, however, this Database type is not the solution to all problems, particularly in specific contexts where data consistency is a relevant characteristic.

i. ACID Vs BASE

ACID (Atomicity, Consistency, Isolation, Durability) properties are one of the main differences between SQL and NoSQL databases. In some specific scenarios, ACID properties confer the reliability of the database. The transaction must be atomic, in other words, a failure in one part of the transaction causes a total failure of the transaction. The information must be consistent, that is, a transaction cannot compromise the state of the database. Multiple transactions at the same time do not impact each other: this is an isolation requirement and, finally, the information must remain in the storage system even after restarting the application:

However, some authors suggest BASE (Basic Availability, Soft state, Eventual Consistency) over ACID, giving less importance to consistency and serialization, and valuing better performance, scalability and availability. Basic availability ensures that every request gets a response. One of the advantages of NoSQL databases, mentioned earlier in this document, is their flexible state, as their state can change over time without compromising the database. Finally, the system may eventually.

ii. CAP Theorem

In the year 2000, Eric Brewer presented a theorem that consisted of three factors that must be considered when developing and implementing data systems in distributed environments. These three factors are consistency, availability and partition tolerances, represented in Figure 12, being called CAP theorem.

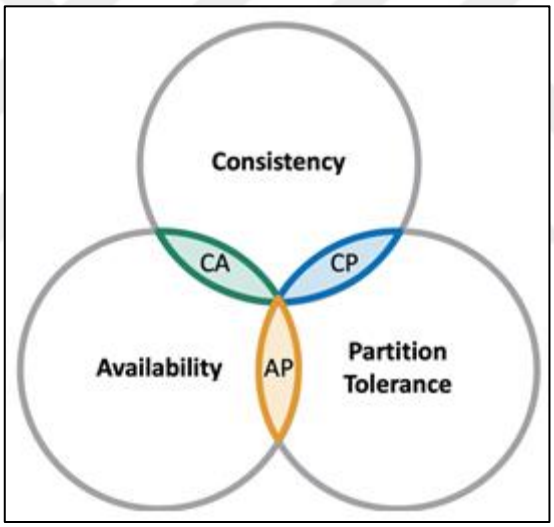


Figure 3.10: Eric Brewer's Cap Theorem [31].

the various NoSQL databases can be classified according to the CAP theorem, however, they are not able to respond to the three properties simultaneously, but they can respond to at least two of the three, which are:

a. Consistency: if a data refresh operation is performed, all users must be able to have the same view on the same dataset. This property matches the atomicity property in ACID transactions;

b. Availability: a system is considered available if it is structured and implemented in a way that allows an operation (writing or reading of data) to be completed even if some cluster node fails or some hardware or software component is under maintenance;
c. Partition Tolerance: the operations must be able to reach their execution in full even if some node is unavailable.

In view of what was mentioned above, through Figure 3.11, it is possible to observe the classification of the different databases according to the CAP theorem.

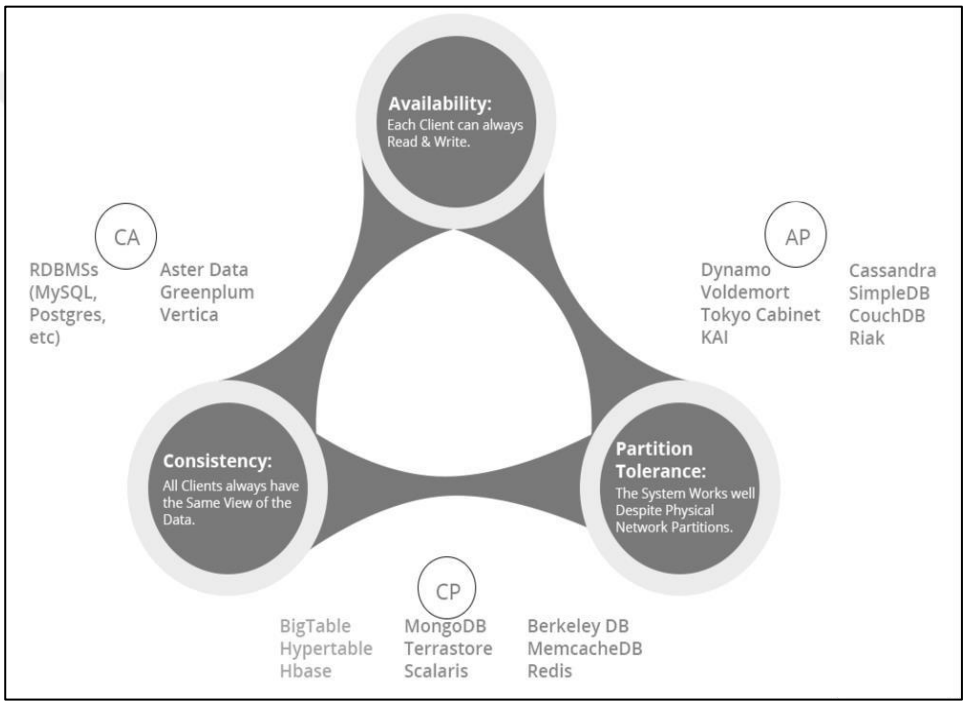


Figure 3.11: Classification Of Nosql Databases According To The CAP Theorem [28].

By analyzing Figure 3.11, it is possible to verify some examples of NoSQL and SQL databases, and verify that some of them manage to guarantee consistency and availability (CA), meaning that the system is limited in terms of scalability. On the other hand, there are solutions capable of guaranteeing consistency and partition tolerance (CP), running the risk of some data not being available if an error occurs in one of the cluster nodes. Finally, some solutions are able to guarantee availability and partition tolerance (AP), although the data returned from the database may sometimes not have the desired level of precision.

NoSQL databases are typically classified into four different types according to their data model, as shown in Figure 3.12.

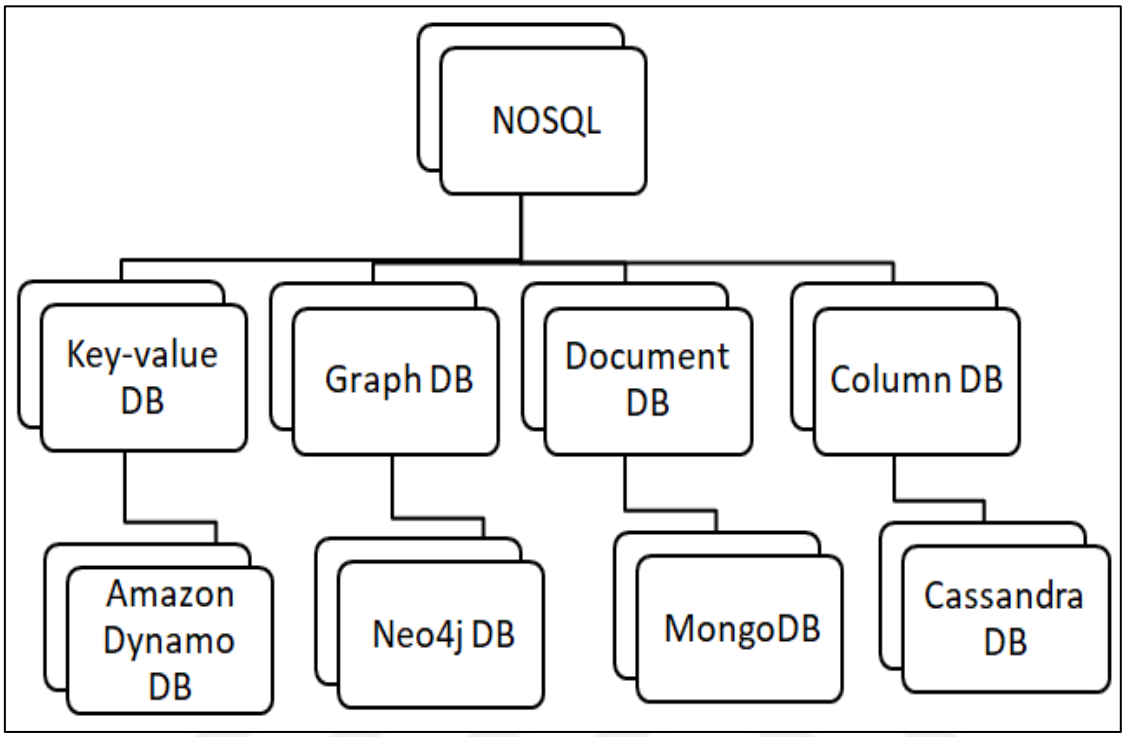


Figure 3.12: Nosql Database Types [28].

i. Key-value databases: this type of database consists of a set of unique key-value pairs, which allow access to the respective value associated with the key. Due to their structure, keys allow the mapping of more complex objects, namely, lists, other key-value pairs, among others. Typically support CRUD operations (Create,Read, Update, Delete), however, do not support aggregation and join operations. Examples: Voldemort, Redis6;
ii. Column-oriented databases: instead of relational databases, column-oriented databases group data into columns. The columns are logically grouped into column families, which contain an unlimited number of columns, making it possible to change the data model by creating new columns over time. Column manipulation allows better performance in computing aggregation functions, particularly in large volumes of data. Examples: Apache Cassandra, HBase;

iii. Document-oriented databases: Document-oriented databases are characterized as a key-value database that restricts data to semi-structured formats, such as JSON formats. Documents are indexed using keys, allowing them to overcome traditional file systems.
iv. Graph databases: represent a special category among NoSQL databases. Graph databases are proposed for certain contexts in which the relationship between entities is relevant. This type of database is not intended to store large volumes of data; however, it allows for the storage of complex data. It is typically composed of a set of nodes and edges, the first representing the entities and the second the type of relationships between entities.

As mentioned earlier, ACID properties portray the main difference between SQL and NoSQL databases. However, recently a new approach has been proposed by the community: NewSQL databases. These databases are promising, conserving and supporting the ACID properties coming from traditional databases and, at the same time, equaling the scalable performance of NoSQL systems. The VoltDB databases and NuoDBare examples of NewSQL databases. Bearing in mind that the focus of this dissertation is on NoSQL databases, more specifically, on column-oriented and graph-oriented databases, it is not considered important to extend the discussion on the differences between NoSQL and NewSQL.

3.2.2 Data Warehousing Systems

This subsection aims to present a summary regarding a Data Warehousing system and how it fits into a Business Intelligence (BI) system. BI systems are approached superficially in this dissertation, since their framework does not fall on them. Data Warehousing is the phenomenon that emerged as a consequence of the storage of large volumes of data and the need to analyze the data in order to improve the decision-making process. a DW is a repository built to consolidate the organization's information in a valid and consistent format, in order to allow data analysis. BI systems are part of the activity of operating the DW system, namely the preparation of reports and consultations that enable the monitoring of the evolution of business indicators. There are two authors that stand out in this theme, approach focuses on a top-down approach that is typically

characterized by the development of the DW based on the company's data model. approach is presented as a bottom-up approach, which is represented by the initial flow between data sources and Data Marts and, later, a data flow between Data Marts and the organizational DW. Data Marts are subsets of data focused on a specific department or repository of the organization, typically characterized as a subset of a DW.

a DW consists of a subject-oriented, integrated, temporally cataloged and non-volatile dataset that supports managers in the decision-making process. It is important to clarify the attributes of its definition, namely:

i. Oriented by theme or subject: in a DW the data is organized around the main subjects of the organizations, such as products, sales, orders or customers, unlike what happens in operational systems, which are dedicated to the processing of daily transactions.
ii. Integrated: a DW is typically built from multiple data sources. Through cleaning and integration techniques, applied to the data, it is possible to obtain a unified view of the total data set of an organization.
iii. Variables in time: the DW should provide a historical analysis perspective on the data. Typically, they are characterized by presenting the data in a time horizon between 5 to 10 years, instead of the operational data that present a relatively short time horizon;
iv. non-volatile: a DW only supports two types of operations: loading (initial loading and refreshing) of the data and accessing the loaded data for query. Thus, data can never be changed or deleted (some exceptions may arise through SCDs (Slowly Changing Dimensions).

In a complementary view, in a DW there are four main components that are represented in Figure 13.

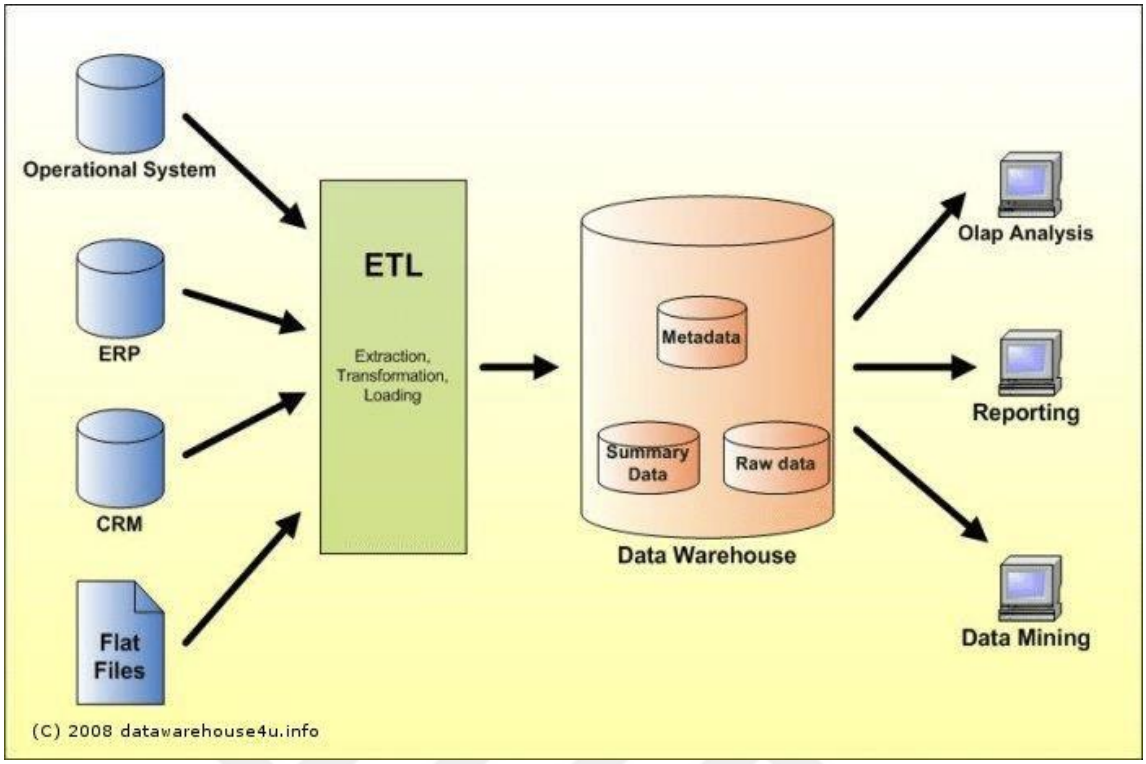


Figure 3.13: Data Warehouse Architecture [19].

a. Data sources: the data comes from multiple data sources, namely ERP's, CRM's (Customer Relationship Management) or other data sources external to the organization;
b. Internship Area and ETL Processes (Extract Transform and Load): the data is loaded to a stage area so that the ETL processes can be carried out later. The extraction phase is carried out by reading and understanding the data for the internship area. The transformation phase requires a set of transformations, namely, data cleaning (correction of invalid values, alteration of attribute format, among others). Finally, the last step concerns loading to the DW;
c. Presentation area: it is the place where the data is organized and stored for possible consultation by end users. This data can later be distributed to the different Data Marts, depending on the organization's needs;
d. Business Intelligence Applications: Set of features that can be provided to users to support decision making.

There are different proposals in the literature for architectures for DWs. Organizations must align the architecture selection with their organizational needs Through Figure 3.14, it is possible to superficially observe several types of architecture taken from however, the author assumes that the organization should choose an architecture that best meets its needs, by way of example, or by implementing of an organizational DW, or by implementing independent Data Marts, or even by implementing Data Marts dependent on the organizational DW (Figure 3.14).

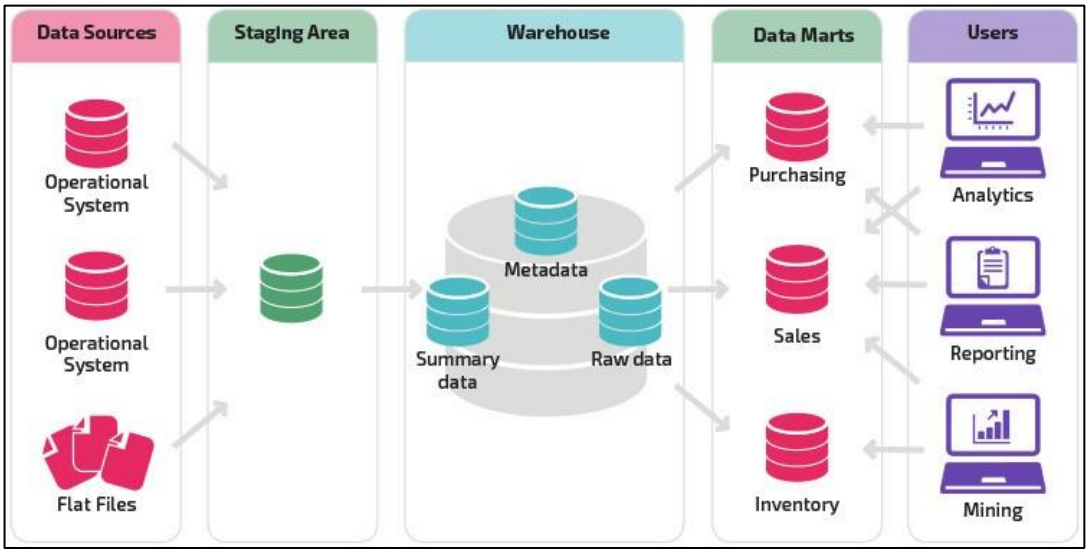


Figure 3.14: Architecture For Implementing A Data Warehouse And Data Marts [34].

3.2.3 Operating Systems Vs Analytical Systems

In an organization it is possible to identify OLTP (Online Transaction Processing) systems, whose purpose is centered on the recording of transactions in real time that occur in its daily operation (registration of new products, sales, orders, among others).

A Data Warehouse, previously characterized, is typically presented as an OLAP (Online Analytical Processing) system, whose purpose is to support the decision making of executives, managers and analysts. it is possible to summarize, through Table 1, the main differences between an operational database and a Data Warehouse.

Table 3.1: Operational Databases Vs Data Warehouses.

Operational Databases	Data Warehouses
operational computer applications	Business analysts and administrators executives
Operational Objectives (OLTP)	Decision making support (OLAP)
Read and write accesses	Mostly read-only accesses
Access to few records at a time	Access to many records at a time
Real-time updated data	Periodic loads of more data
Structure optimized for updates	Optimized structure for processing questions

Along with the synthesis presented in Table 1, it is important to highlight that although Data Warehouses have numerous advantages over operating systems, with the emergence of the Big Data phenomenon, the computation of OLAP cubes with different volumes, speeds and varieties of data has revealed a challenge in this domain, giving rise to Big Data Warehouses.

3.2.4 Big Data Warehouse

Currently, the phenomenon characterized by Big Data has stimulated great challenges for the execution of Data Warehousing, that is, the rules inherent to relational data cannot be applied in images or videos, or in some formats of data coming from sensors. In increasingly competitive and modernized scenarios, the analysis of unstructured data provides valuable information for organizations, leading to competitive advantages. In addition, Open Data allows organizations to produce new opportunities with the aim of

strengthening the economy. For example, However, data analysis with the intrinsic characteristics of Big Data requires the adoption of Big Data Warehouses that allow the extraction and production of valuable information, with the purpose of being used in decision-making processes. Traditional data warehouses are typically characterized according to two methodologies:

i. Data oriented: it consists of defining your multidimensional model in accordance with the models of your data sources.
ii. Requirements oriented: consists of defining your multidimensional model in accordance with your business objectives and needs.

As expected, both traditional methodologies are not able to meet the challenges that arise in Big Data contexts. However, recognizes that each of the methodologies has important advantages. Thus, to support a Data Warehouse in Big Data contexts, the author adopts a hybrid methodology that considers the most important characteristics of both traditional methodologies. Summarizing the various approaches to designing a Big Data Warehouse, the author also mentions an incremental approach, which is based on agile approaches, and an automatic approach that consists of ensuring rapid design, even if new data sources are added Methodologies must adopt new logical models, such as the models used for NoSQL databases, in order to guarantee scalability, flexibility and superior performance. In the same line of work, provide a more detailed definition of Big Data Warehouse, defining it as a scalable, high-performance and flexible storage system, capable of dealing with the increase in volume, variety and speed. of data. Consequently, the concept of Big Data Warehouse emerged as a topic of interest within Big Data systems, providing new relevant features, namely:

i. Flexible storage and integration, including internal and external data to organizations.
ii. High performance with real-time responses.
iii. High scalability at low cost through common hardware.
iv. Highly distributed data processing.
v. Complex workloads (example: ah hoc query queries, running Data Mining models on large volumes of data).

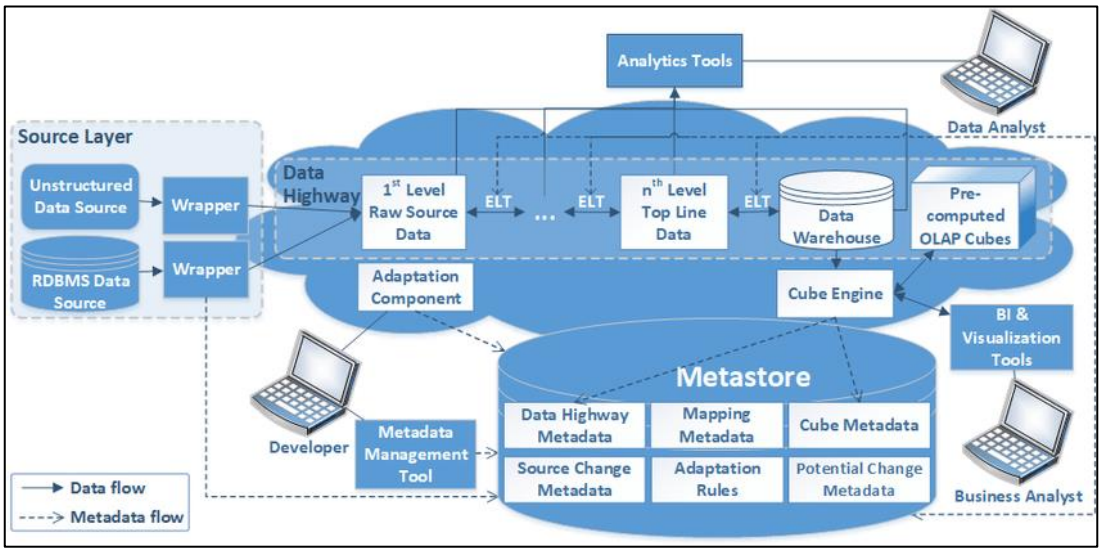
In the methodologies mentioned above, the lack of guidelines from which professionals in this area can orient themselves is sometimes notorious. Occasionally, authors leave their input regarding the type of architecture that should be taken into account, or the method of implementation. Despite being important contributions, it is sometimes necessary to provide logical components and the representation of the data flow between the various components, with the purpose of helping professionals in the implementation of a Big Data Warehouse.

Thus, Figure 3.15 illustrates the architecture of a Big Data Warehouse, proposed by Costa and Santos (2017), in which the various logical components and inherent data flows are highlighted. This architecture is represented in three main layers, namely:

a. Data Extraction, Preparation and Enrichment: the extracted data comes from the data provider component (example: sensors, transactional systems, legacy systems, among others). Its extraction can be carried out in two different ways, namely, streaming and batch mechanisms. The data presented in batch are primarily stored in the distributed file system (sandbox storage), without undergoing any type of processing. Instead, data extracted through streaming mechanisms is immediately processed (prepared and enriched), with the aim of providing more value in the event of being analyzed in real-time. Nevertheless, in specific contexts, data extracted through streaming mechanisms can also be stored in the distributed file system as an optional path.
b. Data Organization and Distribution Platforms: Big Data Warehouse features three different storage components, classified into two different categories: file system and indexed storage. The sandbox storage component is a highly flexible storage area, without the need to define any kind of metadata. In turn, the batch storage component is an indexed storage system to store and process Big Data through batch files. Instead of the sandbox component storage, this component is supported by the definition of metadata to support a structured analysis. Finally, the streaming storage component, like the previous component, is an indexed storage system, which handles the data that arrives in streaming formats. These last two components are indexed to support Big Data Warehouse workloads effectively and efficiently.
c. Analysis, Visualization and Data Access: the main component of this layer corresponds to the distributed query engine element, which is responsible for querying the data stored

in the indexed systems, combining streaming and batch data in the same query. This component is used to explore the data stored in the Big Data Warehouse; however, it also provides data to the data science and data visualization components. Finally, the Data Provider component represents all stakeholders interested in consuming data from the Big Data Warehouse (data analysts, external users, external systems, operational managers or administrative managers).

Table 3.2: Architecture Of A Big Data Warehouse [36].



The focus of this dissertation is framed with the first layer of the architecture shown in Figure 3.15, showing the enrichment process by obtaining the profile of the data that is arriving at the Big Data Warehouse and understanding how the data that are arrive through streaming or batch mechanisms, in various formats, can be integrated with the data that subsists in the second layer of the architecture proposed by Costa and Santos (2017) [36].

3.2.5 Data Models

Data models are the central element in the implementation and development of information systems, in order to ensure that all data needs are ensured. In traditional contexts, relational data models are defined through rules about the requirements that these models must consider. However, in a Big Data context, particularly due to the characteristics of NoSQL databases, data modeling tasks undergo some changes, since

these databases are characterized by being schema-free. The models, in this context, are implemented considering the queries they must answer. there are several data models that are characterized according to the concepts they use to describe the structure of the database. Table 3.2 shows the categories of models that are considered by the authors.

Table 3.3: Data Models And Associated Concepts.

Model of Data	**Description**	**concepts**
conceptual	Also characterized as a high-level model, whose function is to provide concepts that are close to the way stakeholders understand the data.	Entities. Attributes. Relations.
Logical	Also characterized as an implementation model, it provides concepts that can be understood by users, also presenting details of how the data fits into the data model.	Structure of records that depend on database system to be used.
Physicist	Also characterized as a low-level model, detailing how data is stored in the storage system.	Format of records. Order the records.

In line with the implementation of models in Data Warehouses, three levels of abstraction are highlighted, represented in Figure 18, namely the conceptual model, the logical model and the physical model.

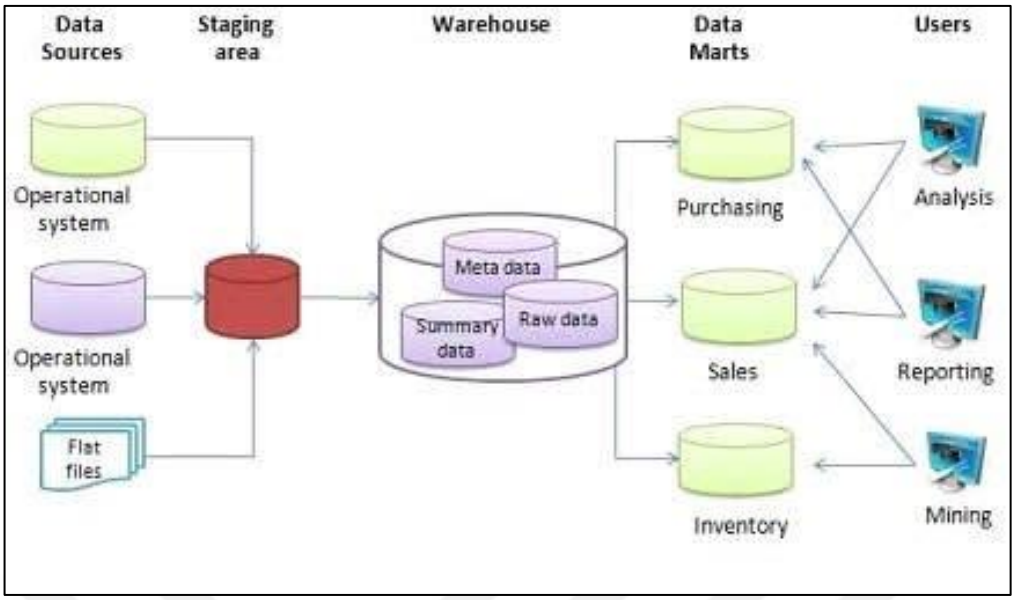


Figure 3.15: Implementation Process In Data Warehouses [37].

the conceptual models refer to the multidimensional model referring to a Data Warehouse. Therefore, according to Figure 3.16, the multidimensional modeling of a Data Warehouse is achieved through the implementation of star, snowflake or constellation schemas. The characteristics of each of these are presented in Table 3.3.

Table 3.4: Multidimensional Modeling Schemes Of A Data Warehouse.

Model	Description
Star	A star schema integrates a single fact table, the center of the star, and multiple dimension tables linked to the fact table. Between dimension tables and fact tables there is typically a one-to-many (1: n) relationship. The fact table typically corresponds to the intended subject.
Model	**Description**
	analyze, usually to a business component (sales, purchases, orders, among others)
flake of Snow	A snowflake schema is a schema whose dimensions are partially or completely normalized. The star schema and the snowflake schema are similar in terms of data content, although the snowflake schema has a more complex structure. This scheme is sometimes associated with the difficulty in its interpretation and loss of performance in the processing of questions due to its normalization.
Constellation	A constellation schema is a schema that integrates multiple fact tables that share common dimensions. It is commonly seen as an integrated set of star schemas.

The connection between the layer referring to the conceptual model and the layer referring to the logical model is performed according to three approaches: ROLAP (Relational OLAP), MOLAP (Multidimensional OLAP), HOLAP (Hybrid OLAP). However, as can be seen in previous sections, the computation of traditional logical models in Data Warehouses is not suitable for large volumes of data present in Big Data contexts. The authors Costa and Santos (2016) state that in Big Data contexts, in which NoSQL databases are typically used, logical data models are characterized by being schema-free, in other words, different rows in a table may contain different columns of data. NoSQL databases are characterized by being schema-free, not being necessary major concerns with data structure, however, in Big

Data contexts it may be necessary to add structure to the data when analytical tasks require it (physical model). Thus, depending on the context, they can be transformed into a graph-based data model, for example, in Neo4J when it comes to storing complex data, or into a column-based data model. Based on the previous statements, in Big Data scenarios, there is a need to add structure to the data at the moments when the analytic tasks are performed. Given this need, in their work, propose a process of structuring a data model in Hive through the transformation of a multidimensional data model, used in a traditional Data Warehouse. The advantage of this transformation is the ability to consider all the needs of the organization, expressed in its operational data model (OLTP) and use this information to identify suitable data schemas to support analytical tasks in Big Data. Within the scope of this dissertation, the focus is related to the access and processing of large volumes of data as quickly as possible, that is, there is a need to execute a set of algorithms for the evaluation of the profile of the data coming from the sources of data, sometimes complex, with the purpose of establishing semantic relationships between the various tables present in the Big Data Warehouse. Thus, it is necessary to select the data model, the database and the appropriate technologies to optimize this type of processing.

4. PROPOSED METHOD

4.1 PROBLEM WITH BIG DATA WAREHOUSING

The problem for which we have been commissioned to try to obtain informative results can be outlined in the following way: we have a system consisting of a power plant that discharges heat to the environment. This heat, in the form of vapor, is collected once its condensation is complete. The heat load discharged from the system constitutes the variability and precisely this parameter influences the correct operation of the machine.
The data provided to us relate to various descriptive variables of the plant in a summer case. Our aim is then to study such data to analyze the relationships of the variables involved.
Professor Sacile suggests that we approach the analysis of this phenomenon by "exploiting" the power of Neural Networks.

4.2 RESOLUTION PROCEDURE ADOPTED

In this section we will try to explain as much as possible the steps that constitute the procedure adopted for solving the problem.

As previously said, the problem is approached through the non-linear regression technique of the neural network. First of all, we studied its theory and, through the use of the MATLAB system and therefore of its toolbox for neural networks, we wrote the codes relating to the approximation of the relative "expected void" and "incondensable" variables. , respectively, to those that were the deliveries of "model 1" and "model 2" provided to us by Engineer Sorce and attached in an appropriate Excel file.
The second step was then to try to alter and / or modify and therefore reduce or restructure the data set to be supplied to the neural network. This is in order to bring out the variables or values selected by them that most affect the output variable estimated by the network.
For the pursuit of this purpose, we have thought of three distinct functions:

i. A sampler
ii. A fashion alterer of one variable
iii. A clusterizer

The first function, after having taken as input the value relating to the sampling interval with which the reduction of the data set to be supplied to the network is to be carried out, proceeds by selecting a row of the matrix (whose columns are the input variables and the output) with a frequency corresponding to the specified sampling interval. (For example: if the interval is relative to the value of 10, then every 10 lines WE take the line).

The selected data are thus saved in an auxiliary matrix which will constitute the reduced data set and the discarded data are in turn saved in a further matrix which will instead be used to carry out the testing on the model generated by the network.

The second function, as the name implies, has the particular objective of altering the fashion of a given variable that is given as input. In fact, this function works by first calculating the aforementioned mode of the variable, then one proceeds by defining a neighborhood of values with the center in the neighborhood in the mode and finally replacing those values with other arbitrary values not belonging to that neighborhood. The choice WE made on the latter arbitrarily selectable values fell on a trivial random set not belonging to that calculated around.

The third and last function, the less trivial of the two, is designed on the basis of what is reported in the chapter on "data preprocessing" (in particular on page 36) of the lecture notes of the data mining course that WE report in the attachment.

It performs a data reduction in order to divide the data set into 'N' partitions (with 'N' number of partitions selected by the user) and, of these, only a subset will then be selected and supplied to the neural network. But how does it carry out the subdivision? In a homogeneous way, which means that it will try to compose the sets so that the respective statistical properties are, among the various partitions, as equivalent as possible. To do this, the function uses a further function to calculate the Euclidean distance between a given row, the matrix of the data set to be reduced, and all the others; then having launched this last function from the starting line, one proceeds by fetching the line whose distance is the minimum among all those calculated from the starting line. Once this is done, go to put this line in a new set. The starting line then becomes this second line and the procedure is iterated.

As you can guess, two lines that have almost equivalent values can never be in the same set. And furthermore, if N is high enough, it will be possible to construct sets where the lines that have almost equivalent values can never be in the same set and, by doing so, WE will “level” the number of lines at high frequencies. (For example in a data set consisting of a

10-row variable x, if the value 8 is equal to three rows of the matrix x, and WE have chosen to partition into N = 3 sets, the three x (we) = 8 will go into different sets and never into a single one. So the frequency of x (we) = 8 if considered in one of the three sets will now be equal to one.)

4.3 MODEL RESULTS

Among the files attached to the project it is possible to find the Excel document relating to the table of results obtained in relation to the fifty simulations / experiments carried out.
In this section we try to give an explanation, as complete as possible, on the relationships of the various experiments.

4.3.1 K-Nearest Neighbor

From the given dataset we imported the training features: X_train, training labels: y_train, and testing features: X_test into matlab and fitted k-nearest neighbor model with the function fitcknn(). We used the value k = 5 and the simple Euclidean distance measure to compute the distance between two samples.

The labels for X_test was predicted using the predict() function. The accuracy was measured with the predicted labels and the y_test data using the classperf() function.

The accuracy for kNN is 98.60%.

4.3.2 Feedforward Artificial Neural Network

From the given dataset we imported the training features: X_train, training labels: y_train, and testing features: X_test into matlab. We transformed the y_train into a sparse matrix of vectors, with 1 in each column, as indicated by ind. ind2vec(ind,N) returns an N -by- M matrix.

We used two methods to train the model.

I. Feedforward: We used the feedforward () function with 25 neurons to train the model. It takes long time to train the model and the accuracy rate is – Feedforward Accuracy = 95.90%.

II. Patternnet: We used the patternnet() function with 25 neurons to train the model. It takes very short time to train the model and the accuracy rate is – Patternnet Accuracy = 98.60%.

4.3.3 SVM

From the given dataset we imported the training features: X_train, training labels: y_train, and testing features: X_test into matlab.

We trained the model using the fitcsvm() function. We used the Kernel function = 'polynomial' and Polynomial order = 2.

The accuracy for SVM is = 99.80%

Advance that the following cases are evaluated:

a. The difference between the results obtained in relation to the standard normalization (therefore with normalized data between the values 0 and 1) and the significant normalization (generating normalized data always between 0 and 1 but using no longer the real minimums and maximums but rather the minimums and maximums which were considered significant by my colleague Ghettini). This analysis is carried out on three distinct sampling intervals: every 100, every 10 and every one, therefore not sampling but taking the data set in full. And, for each of these three, the cases of different clustering (therefore with different number of clusters) are considered once by mixing the data set and once by not mixing. For each of the sampling intervals considered, we perform an obviously distinct clustering.

Specifically: in case 100 we consider the cases:

i. Number of clusters: 0, of which we select: 0
ii. Number of clusters: 25, of which we select: 5
iii. Number of clusters: 1000, of which we select: 200

In the case in which they sample every 10 lines we consider instead:

iv. Number of clusters: 0, of which we select: 0
v. Number of clusters: 1000, of which we select: 200
vi. Number of clusters: 2000, of which we select: 400
vii. Number of clusters: 10000, of which we select: 2000

And finally, in case they don't sample we consider:

viii. Number of clusters: 0, of which we select: 0

ix. Number of clusters: 1000, of which we select: 200

x. Number of clusters: 10000, of which we select: 2000

The above cases we have just talked about are then reanalyzed but this time in relation to the difference between the experiments in which the data are mixed and those in which we do not mix but maintain the temporal order in which they were collected. This time, of course, we will keep the normalization to the standard case fixed in one case and the normalization to the significant case in the other.

we consider a case that we think is interesting. we am going to compare the results obtained, keeping the normalization fixed to standard, in relation to an experiment in which we sample every 1000 rows but we do not cluterise (thus giving the network a data set for training about 90 data) and to a second experiment in which instead we do not sample but cluster the data set into 1000 clusters and taking, among the 1000 clusters obtained, only one cluster to be supplied to the network (once again thus giving the network a data set of 90 data). Let it be clear that each time we consider both the case in which we mix the data and the one in which we keep order.

All in order to evaluate the effectiveness of the clustering algorithm used.

We consider the four models provided by Engineer Sorce using 90% of the total data for testing.

Finally, we are going to apply the frequency alteration function in a neighborhood of the mode of the variable "ambient temperature" on model 1. The experiment modifies the value of about 22 thousand values in a range with the center in the aforementioned mode +/- la theshold.

we carry out the above cases by reporting the respective values from the excel table:

a. we hypothesize not to mix the data (Mixer = 0, reporting first the standard normalization and then the significant one):

b. to) Number of clusters: 0, of which we select: 0

Table 4.1: Neural Networks Training Error Rate.

Cluster (PF3)	Cluster IN (PF3)	NN_Error			Test Err	Accuracy	Train Set Size
		Train	Validation	Test			
0	0	1.135E-03	1.248E-03	1.272E-03	1.132E-03	0.911	930

Table 4.2: Neural Networks Testing Error Rate.

Cluster (PF3)	Cluster IN (PF3)	NN_Error			Data Test_Err	Accuracy	Train Set Size
		Train	Validation	Test			
0	0	1.615E-03	1.489E-03	1.548E-03	1.635E-03	0.962	930

Table 4.3: Number Of Clusters: 25, Of Which We Select: 5.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
25	5	1.505E-03	2.428E-03	1.696E-03	2.478E-03	0.948	188

Table 4.4: Number Of Clusters: 1000, Of Which We Select: 200.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
1000	200	1.583E-03	1.972E-03	2.210E-03	7.415E-03	0.320	280

NOTE: the accuracy is better in the standard in all cases apart from a) (as, probably, the Train Set Size is greater).

Table 4.5: Number Of Clusters: 0, Of Which We Select: 0.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
0	0	1.528E-03	1.692E-03	1.485E-03	1.541E-03	0.968	9307

Table 4.6: Number Of Clusters: 1000, Of Which We Select: 200.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_ Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
1000	200	9.330E-04	8.836 E-04	9.287 E-04	6.130E-03	0.855	1960
2000	400	7.789E-04	7.918 E-04	7.046 E-04	3.336E-02	0.652	1960
2000	400	1.389E-03	9.486 E-04	1.106 E-03	7.431E-03	0.815	1960
10000	2000	5.046E-04	4,989 E-04	4,958 E-04	5,490E-03	0.742	2800
10000	2000	1.085E-03	9.939 E-04	9.784 E-04	1.643E-02	0.731	2800

NOTE: by increasing the number of data supplied to the network, the error on testing is worse in the case of significant normalization.

Table 4.7: Number Of Clusters: 0, Of Which WE Select: 0.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_ Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
0	0	1.590E-03	1.608 E-03	1.614 E-03		0.965	93060

Table 4.8: Number Of Clusters: 1000, Of Which We Select: 200.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
1000	200	1.314E-03	1.380E-03	1.373E-03	1.992E-03	0.952	18620

Table 4.9: Number Of Clusters: 10000, Of Which We Select: 2000.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
10000	2000	9.691E-04	9.565E-04	1.041E-03	5.304E-03	0.855	19600

NOTE: the accuracy is always worse in the case of standard normalization, while the errors of training validation and testing are always greater in the case of significant normalization.

c. we hypothesize not to mix the data (Mixer = 1, reporting first the standard normalization and then the significant one):

Table 4.10: Number Of Clusters: 0, Of Which We Select: 0.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
0	0	1.340E-03	1.681E-03	1.573E-03	1.621E-03	0.968	930
25	5	1.396E-03	1.807E-03	3.010E-03	2.488E-03	0.877	188
25	5	1.091E-03	1.438E-03	2.081E-03	4.000E-03	0.911	188
1000	200	5.071E-04	5.928E-04	7.404E-04	3.141E-02	0.668	280

1000	200	2.332E-03	3.091 E-03	2.064 E-03	1.696E-03	0.869	280
0	0	9.840E-04	1.009 E-03	1.091 E-03	1.006E-03	0.942	9307
0	0	1.567E-03	1.462 E-03	1.696 E-03	1.576E-03	0.966	9307

Table 4.11: Number Of Clusters: 1000, Of Which We Select: 200.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_ Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
1000	200	7.845E-04	8,348 E-04	7.543 E-04	2.728E-03	0.912	1960
2000	400	3.897E-04	3.143 E-04	5,054 E-04	6.625E-03	0.669	1960
2000	400	8.811E-04	8.235 E-04	9.552 E-03	2.575E-03	0.993	1960
10000	2000	7,364E-04	7.035 E-04	7.079 E-04	2.001E-02	0.622	2800
10000	2000	1.733E-03	2.044 E-03	1.664 E-03	9.756E-03	0.799	2800

NOTE: by increasing the number of data supplied to the network, the error on testing is worse in the case of significant normalization. Accuracy is always worse in standard.

Table 4.12: Number Of Clusters: 0, Of Which We Select: 0.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_ Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
1000	200	9.813E-04	9.912 E-04	1.014 E-03	1.304E-03	0.918	18620
1000	200	1.277E-03	1.309 E-03	1.310 E-03	1.952E-03	0.956	18620

10000	2000	5.943E-04	5.981 E-04	5.961 E-04	4.593E-03	0.774	19600
10000	2000	9.697E-04	9.700 E-04	9.331 E-04	5.902E-03	0.851	19600

NOTE: the accuracy is always worse in the case of standard normalization, while the errors of training validation and testing are always greater in the case of significant normalization. we hypothesize not to mix the data (Mixer = 0, reporting first the standard normalization and then the significant one):

Table 4.13: Number Of Clusters: 0, Of Which We Select: 0.

Cluster (PF3)	Cluster_IN (PF3)	NN_Error			Test_Err	Accuracy	Train Set Size
		Train	Validation	Test			
0	0	1.135E-03	1.248E-03	1.272 E-03	1.132E-03	0.911	930

Table 4.14: Number Of Clusters: 25, Of Which We Select: 5.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_ Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
25	5	1.396E-03	1.807 E-03	3.010 E-03	2.488E-03	0.877	188
1000	200	5.196E-04	2.867 E-04	3.862 E-04	1.164E-02	0.724	280
1000	200	5.071E-04	5.928 E-04	7.404 E-04	3.141E-02	0.668	280
0	0	9.739E-04	1.029 E-03	9.515 E-04	9.755E-04	0.948	9307
0	0	9.840E-04	1.009 E-03	1.091 E-03	1.006E-03	0.942	9307

1000	200	4,732E-04	4.727 E-04	5.173 E-04	1.598E-03	0.883	1960
1000	200	6.411E-04	5.906 E-04	7.431 E-04	6,882E-03	0.755	1960
2000	400	7.789E-04	7.918 E-04	7.046 E-04	3.336E-02	0.652	1960
2000	400	3.897E-04	3.143 E-04	5,054 E-04	6.625E-03	0.669	1960
10000	2000	5.046E-04	4,989 E-04	4,958 E-04	5,490E-03	0.742	2800
10000	2000	7,364E-04	7.035 E-04	7.079 E-04	2.001E-02	0.622	2800

Table 4.15: Number Of Clusters: 1000, Of Which We Select: 200.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_ Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
1000	200	9.813E-04	9.912 E-04	1.014 E-03	1.304E-03	0.918	18620
10000	2000	5.567E-04	5.326 E-04	5.419 E-04	3.327E-03	0.810	19600
10000	2000	5.943E-04	5.981 E-04	5.961 E-04	4.593E-03	0.774	19600

NOTE: as the values increase, as known from the statistics, the values tend to be equivalent. we hypothesize not to mix the data (Normalization = STD, reporting first the unmixed case and then the mixed case):

Table 4.16: Number Of Clusters: 0, Of Which We Select: 0.

		NN_Error			

Cluster (PF3)	Cluster IN (PF3)	Train	Validation	Test	Test_Err	Accuracy	Train Set Size
0	0	1.615E-03	1.489E-03	1.548E-03	1.635E-03	0.962	930
0	0	1.340E-03	1.681E-03	1.573E-03	1.621E-03	0.968	930

Table 4.17: Number Of Clusters: 25, Of Which We Select: 5.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
25	5	1.091E-03	1.438E-03	2.081E-03	4.000E-03	0.911	188
1000	200	5.196E-04	2.867E-04	3.862E-04	1.164E-02	0.724	280
1000	200	2.332E-03	3.091E-03	2.064E-03	1.696E-03	0.869	280

Table 4.18: Number Of Clusters: 0, Of Which We Select: 0.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
0	0	1.567E-03	1.462E-03	1.696E-03	1.576E-03	0.966	9307
1000	200	9.330E-04	8.836E-04	9.287E-04	6.130E-03	0.855	1960
1000	200	7.845E-04	8,348E-04	7.543E-04	2.728E-03	0.912	1960
2000	400	1.389E-03	9.486E-04	1.106E-03	7.431E-03	0.815	1960

2000	400	8.811E-04	8.235 E-04	9.552 E-04	2.575E-03	0.993	1960
10000	2000	1.085E-03	9.939 E-04	9.784 E-04	1.643E-02	0.731	2800
10000	2000	1.733E-03	2.044 E-03	1.664 E-03	9.756E-03	0.799	2800
0	0	1.590E-03	1.608 E-03	1.614 E-03		0.965	93060
0	0	1.566E-03	1.587 E-03	1.558 E-03		0.967	93060

Table 4. 19: Number Of Clusters: 10000, Of Which We Select: 2000.

Cluster (PF3)	Cluster IN (PF3)	NN_Error	Test_Err	Accuracy	Train Set Size	Cluster (PF3)	Cluster IN (PF3)
1000	200	1.277E-03	1.309 E-03	1.310 E-03	1.952E-03	0.956	18620
10000	2000	9.691E-04	9.565 E-04	1.041 E-03	5.304E-03	0.855	19600
10000	2000	9.697E-04	9.700 E-04	9.331 E-04	5.902E-03	0.851	19600
1000	1	1.274E-03	8.303 E-04	1.296 E-03	1.684E-03	0.889	93
0	0	8.847E-04	1.992 E-03	2.595 E-03	1.621E-03	0.879	93
1000	1	1.365E-03	1.215 E-03	9.957 E-04	1.549E-03	0.881	93
0	0	3.549E-03	4.712 E-03	1.080 E-02	7.808E-03	0.510	93

NOTE: although the data set supplied to the network is the same size in all four cases, we notice an accuracy of 50% in the case in which we mix but do not perform clustering. While clustering the accuracy is 88%. By not mixing, the accuracy is confirmed on percentages above 85%.

Table 4.20: The Minimum Value In The Case Of Clustering In 1000 Sets And Taking Only One.

	NN_Error			Data Test_Err	Train Set Size
	Train	Validation	Test		
M_2a	7,974E-10	1.036E-09	1.561E-09	1.005E-09	9307
M_2b	1.830E-09	2.393E-09	1.880E-09	1,974E-09	9307
M_2c	1.184E-10	1.341E-10	1.458E-10	1.334E-10	9307
M_1	9.843E-04	8.872E-04	1.032E-03	9.895E-04	9307

NOTE: we immediately notice how the minimum error on testing is in correspondence with the 2C model (being that, as known, neural networks model more accurately if trained on a high set of features).

Table 4.21: Maximum Error Is Observed In The Case Of Model 1.

Mixer (PF0)	Normalization	Tau (PF1)	Cluster_IN (PF3)	P_IN (PF3)	NN_Error			Accuracy	Train Set Size
					Train	Validation	Test		
NO	STD	1	0	0	5.236 E-03	5.276 E-03	5.204 E-03	0.687	93060
			0	0	9.948 E-04	1.008 E-03	9.955 E-04	0.948	93060

NOTE: in the first line, where the experiment with the alteration of the mode is reported, there is a lowering of the precision and a marked worsening in the testing error. To demonstrate how much the model 1 is influenced by this variable.

5. CONCLUSION

The primary purpose of this thesis is to develop a distributed scale-out storage system-based platform for big data analytics. This platform aims to achieve low latency, massive scalability, and fault tolerance. It is argued that we are currently experiencing an "information explosion," characterized by the continuous generation of vast quantities of data that continue to grow in size. Numerous industries gather and analyze this tremendous amount of data. As the number of networked devices increases, the quantity of data being generated grows exponentially. It is estimated that by 2021, the amount of digital information will have increased by a factor of nine, reaching 35 trillion gigabytes, compared to the amount available in 2011. In the first line of experimentation, where the mode alteration is reported, there is a decrease in precision and a noticeable increase in testing error. This demonstrates how much Model 1 is influenced by this variable. We immediately notice that the minimum testing error corresponds with the 2C model, as neural networks tend to model more accurately when trained on a high set of features. Consequently, the maximum error is observed in the case of Model 1. Although the dataset size supplied to the network remains consistent across all four cases, we observe an accuracy of 50% when mixing without clustering. When clustering is implemented, the accuracy rises to 88%. By not mixing, the accuracy is confirmed to remain above 85%. Furthermore, the testing error assumes the minimum value when clustering into 1000 sets and selecting only one

The novelty of this work lies in the development of a distributed scale-out storage system-based platform for big data analytics, designed to provide low latency, massive scalability, and fault tolerance. By experimenting with different models and data processing techniques, the research demonstrates the impact of various factors on the performance of neural networks. The findings reveal that clustering significantly improves the accuracy of these networks, with the best results obtained when clustering into 1000 sets and selecting only one.

REFERENCES

[1] Mavragani A, Ochoa G, Tsagarakis KP. Assessing the methods, tools, and statistical procedures in Google trends research: systematic review. J Med Internet Res. 2018;20(11):e270.

[2] Sun D, Zhang G, Zheng W, Li K. Key technologies for big data stream computing. In: Li K, Jiang H, Yang LT, Guzzocrea A, editors. Big data algorithms, analytics and applications. New York: Chapman and Hall/CRC; 2015. p. 193–214. ISBN 978-1-4822-4055-9.

[3] Qian ZP, He Y, Su CZ et al. TimeStream: Reliable stream computation in the cloud. In: Proc. 8th ACM European conference in computer system, EuroSys 2013. Prague: ACM Press; 2013. p. 1–4.

[4] Liu R, Li Q, Li F, Mei L, Lee, J. Big data architecture for IT incident management. In: Proceedings of IEEE international conference on service operations and logistics, and informatics (SOLI), Qingdao, China. 2014. p. 424–9.

[5] Sakr S. An introduction to Infosphere streams: A platform for analysing big data in motion. IBM. 2013. https:// www.ibm.com/developerworks/library/bd-streamsintro/index.html. Accessed 7 Oct 2018.

[6] Xhafa F, Naranjo V, Caballé S. Processing and analytics of big data stream with Yahoo!S4. In: 2015 IEEE 29th international conference on advanced information networking and applications, Gwangiu, South Korea, 24–27 March 2015. 2015. https://doi.org/10.1109/aina.2015.194.

[7] Marz N. Storm: distributed and fault-tolerant real-time computation. In: Paper presented at Strata conference on making data work, Santa Clara, California, 28 Feb–1 March 2012. 2012. https://cdn.oreillystatic.com/en/assets/1/event/75/Storm_%20distributed%20and%20fault-tolerant%20realtime%20computation%20Presentation.pdf. Accessed 25 Jan 2018.

[8] Ballard C, Farrell DM, Lee M, Stone PD, Thibault S, Tucker S. IBM InfoSphere Streams: harnessing data in motion. IBM Redbooks. 2010.

[9] Joseph S, Jasmin EA, Chandran S. Stream computing: opportunities and challenges in smart grid. Procedia Technol. 2015;21:49–53.

[10] IBM Research (no date) Stream computing platforms, applications and analytics. IBM. http://researcher.watso n.ibm.com/researcher/view_grp.php?id=2531 Accessed 5 Mar 2019.

[11] Gantz J, Reinsel D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the Far East. New York: IDC iView: IDC Analyse future; 2012.

[12] Cortes R, Bonnaire X, Marin O, Sens P. Stream processing of healthcare sensor data: studying user traces to identify challenges from a big data perspective. The 4th international workshop on body area sensor networks (BASNet-2015). Procedia Comput Sci. 2015; 52:1004–9.

[13] Chung D, Shi H. Big data analytics: a literature review. J Manag Anal. 2015;2(3):175–201.

[14] Lu J, Li D. Bias correction in a small sample from big data. IEEE Trans Knowl Data Eng. 2013;25(11):2658–63.

[15] Garzo A, Benczur AA, Sidlo CI, Tahara D, Ywatt EF. Real-time streaming mobility analytics. In: Proc. 2013 IEEE international conference on big data, big data, Santa Clara, CA, United States, IEEE Press. 2013. p 697–702.

[16] Zaharia M, Das T, Li H, Hunter T, Shenker S, Stoica I. Discretized streams: fault-tolerant streaming computation at scale. In: Proc. the 24th ACM symposium on operating system principles, SOSP 2013, Farmington, PA, United States. New York: ACM Press; 2013. p. 423–38.

[17] Fan J, Liu H. Statistical analysis of big data on pharmacogenomics. Adv Drug Deliv Rev. 2013;65(7):987–1000.

[18] Bifet A, Holmes G, Kirkby R, Pfahringer B. Moa: massive online analysis. J Mach Learn Res. 2010;11:1601–4.

[19] Akter S, Fosso WS. Big data analytics in e-commerce: a systematic review and agenda for future research. Electr Markets. 2016;26:173–94.

[20] Sivarajah U, Kamal MM, Irani Z, Weerakkody V. Critical analysis of big data challenges and analytical methods. J Bus Res. 2016;70:263–86.

[21] Wienhofen LW, Mathisen BM, Roman D. Empirical big data research: a systematic literature mapping. CoRR, abs/1509.03045. 2015.

[22] Habeeb RAA, Nasaruddin F, Gani A, Hashem IAT, Ahmed E, Imran M. Real-time big data processing for anomaly detection: a survey. Int J Inform Manage. 2018;45:289–307. https://doi.org/10.1016/j.ijinfomgt.2018.08.006.

[23] Mehta N, Pandit A. Concurrence of big data analytics in healthcare: a systematic review. Int J Med Inform. 2018;114:57–65.

[24] Kitchenham BA, Charters S. Guidelines for performing systematic literature review in software engineering. Technical report 2(3), EBSE-2007-01, Keele University and University of Durham. 2007.

[25] Host M, Orucevic-Alagic A. A systematic review of research on open source software in commercial software product development. 2013. http://www.bcs.org/upload/pdf/ewic_ea10_session5paper2.pdf. Accessed 2 Mar 2018.

[26] J. Gantz and D. Reinsel, ‘‘The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,’’ in Proc. IDC iView, IDC Anal. Future, 2012.

[27] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. San Francisco, CA, USA: McKinsey Global Institute, 2011, pp. 1–137.

[28] K. Cukier, ‘‘Data, data everywhere,’’ Economist, vol. 394, no. 8671, pp. 3–16, 2010.

[29] T. economist. (2011, Nov.) Drowning in Numbers—Digital Data Will Flood the Planet-and Help us Understand it Better [Online]. Available: http://www.economist.com/blogs/dailychart/2011/11/bigdata-0

[30] S. Lohr. (2012). The age of big data. New York Times [Online]. 11. Available: http://www.nytimes.com/2012/02/12/sunday-review/big-datasimpact-in-the-world.html?pagewanted=all&r=0

[31] Y. Noguchi. (2011, Nov.). Following Digital Breadcrumbs to Big Data Gold, National Public Radio, Washington, DC, USA [Online]. Available: http://www.npr.org/2011/11/29/142521910/ the-digitalbreadcrumbs-that-lead-to-big-data

[32] Y. Noguchi. (2011, Nov.). The Search for Analysts to Make Sense of Big Data, National Public Radio, Washington, DC, USA [Online]. Available: http://www.npr.org/2011/11/30/142893065/the -search -foranalysts - to -make-%sense-of-big-data

[33] W. House. (2012, Mar.). Fact Sheet: Big Data Across the Federal Government [Online]. Available: http://www.whitehouse.gov/ sites/default/files/microsites/ostp/big_data%_fact_sheet_3_29_2012.pdf

[34] J. Kelly. (2013). Taming Big Data [Online]. Available: http://wikibon.org/blog/taming-big-data/

[35] Wiki. (2013). Applications and Organizations Using Hadoop [Online]. Available: http://wiki.apache.org/hadoop/PoweredBy

[36] J. H. Howard et al., ‘‘Scale and performance in a distributed file system,’’ ACM Trans. Comput. Syst., vol. 6, no. 1, pp. 51–81, 1988.

[37] R. Cattell, ‘‘Scalable SQL and NoSQL data stores,’’ SIGMOD Rec., vol. 39, no. 4, pp. 12–27, 2011.

[38] J. Dean and S. Ghemawat, ‘‘Mapreduce: Simplified data processing on large clusters,’’ Commun. ACM, vol. 51, no. 1, pp. 107–113, 2008.

[39] T. White, Hadoop: The Definitive Guide. Sebastopol, CA, USA: O’Reilly Media, 2012.

[40] J. Gantz and D. Reinsel, ‘‘Extracting value from chaos,’’ in Proc. IDC iView, 2011, pp. 1–12.

[41] P. Zikopoulos and C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, USA: McGraw-Hill, 2011.

[42] E. Meijer, ‘‘The world according to LINQ,’’ Commun. ACM, vol. 54, no. 10, pp. 45–51, Aug. 2011.

[43] D. Laney, ‘‘3d data management: Controlling data volume, velocity and variety,’’ Gartner, Stamford, CT, USA, White Paper, 2001.

[44] M. Cooper and P. Mell. (2012). Tackling Big Data [Online]. Available: http://csrc.nist.gov/groups/SMA/forum/documents/ june2012presentations/f%csm_june2012_cooper_mell.pdf

[45] O. R. Team, Big Data Now: Current Perspectives from O’Reilly Radar. Sebastopol, CA, USA: O’Reilly Media, 2011.

[46] M. Grobelnik. (2012, Jul.). Big Data Tutorial [Online]. Available: http://videolectures.net/eswc2012_grobelnik_big_data/

[47] S. Marche, ‘‘Is Facebook making us lonely,’’ Atlantic, vol. 309, no. 4, pp. 60–69, 2012.

[48] V. R. Borkar, M. J. Carey, and C. Li, ‘‘Big data platforms: What’s next?’’ XRDS, Crossroads, ACM Mag. Students, vol. 19, no. 1, pp. 44–49, 2012.

[49] V. Borkar, M. J. Carey, and C. Li, ‘‘Inside big data management: Ogres, onions, or parfaits?’’ in Proc. 15th Int. Conf. Extending Database Technol., 2012, pp. 3–14.

[50] D. Dewitt and J. Gray, ‘‘Parallel database systems: The future of high performance database systems,’’ Commun. ACM, vol. 35, no. 6, pp. 85–98, 1992. (2014). Teradata. Teradata, Dayton, OH, USA [Online]. Available: http://www.teradata.com/

[51] Chung D, Shi H. Big data analytics: a literature review. J Manag Anal. 2015;2(3):175–201.

[52] Lu J, Li D. Bias correction in a small sample from big data. IEEE Trans Knowl Data Eng. 2013;25(11):2658–63.

[53] Garzo A, Benczur AA, Sidlo CI, Tahara D, Ywatt EF. Real-time streaming mobility analytics. In: Proc. 2013 IEEE international conference on big data, big data, Santa Clara, CA, United States, IEEE Press. 2013. p 697–702.

[54] Zaharia M, Das T, Li H, Hunter T, Shenker S, Stoica I. Discretized streams: fault-tolerant streaming computation at scale. In: Proc. the 24th ACM symposium on operating system principles, SOSP 2013, Farmington, PA, United States. New York: ACM Press; 2013. p. 423–38.

[55] Fan J, Liu H. Statistical analysis of big data on pharmacogenomics. Adv Drug Deliv Rev. 2013;65(7):987–1000.

[56] Bifet A, Holmes G, Kirkby R, Pfahringer B. Moa: massive online analysis. J Mach Learn Res. 2010;11:1601–4.

[57] Akter S, Fosso WS. Big data analytics in e-commerce: a systematic review and agenda for future research. Electr Markets. 2016;26:173–94.

[58] Sivarajah U, Kamal MM, Irani Z, Weerakkody V. Critical analysis of big data challenges and analytical methods. J Bus Res. 2016;70:263–86.

[59] Wienhofen LW, Mathisen BM, Roman D. Empirical big data research: a systematic literature mapping. CoRR, abs/1509.03045. 2015.

[60] Habeeb RAA, Nasaruddin F, Gani A, Hashem IAT, Ahmed E, Imran M. Real-time big data processing for anomaly detection: a survey. Int J Inform Manage. 2018;45:289–307. https://doi.org/10.1016/j.ijinfomgt.2018.08.006.

[61] Mehta N, Pandit A. Concurrence of big data analytics in healthcare: a systematic review. Int J Med Inform. 2018;114:57–65.

[62] Kitchenham BA, Charters S. Guidelines for performing systematic literature review in software engineering. Technical report 2(3), EBSE-2007-01, Keele University and University of Durham. 2007.

[63] Host M, Orucevic-Alagic A. A systematic review of research on open source software in commercial software product development. 2013. http://www.bcs.org/upload/pdf/ewic_ea10_session5paper2.pdf. Accessed 2 Mar 2018.

[64] J. Gantz and D. Reinsel, ‘‘The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,’’ in Proc. IDC iView, IDC Anal. Future, 2012.

[65] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. San Francisco, CA, USA: McKinsey Global Institute, 2011, pp. 1–137.

[66] K. Cukier, ‘‘Data, data everywhere,’’ Economist, vol. 394, no. 8671, pp. 3–16, 2010.

[67] T. economist. (2011, Nov.) Drowning in Numbers—Digital Data Will Flood the Planet-and Help us Understand it Better.

[68] S. Lohr. (2012). The age of big data. New York Times [Online]. 11. Available: http://www.nytimes.com/2012/02/12/sunday-review/big-datasimpact-in-the-world.html?pagewanted=all&r=0.