INTERNATIONAL
HELLENIC
UNIVERSITY

Taxonomy of (Big) Data tools & Technologies for all phases of data lifecycle for data-driven governments_A Survey

Valerios Tsompanidis

UNIVERSITY CENTER OF INTERNATIONAL PROGRAMMES OF STUDIES

SCHOOL OF SCIENCE AND TECHNOLOGY

A thesis submitted for thesis submitted for the degree of

Master of Science (MSc) in e-Business and Digital Marketing

January 2021

Speyer –Germany

Student Name:	Valerios Tsompanidis
SID:	3305190053
Supervisor:	Prof. Vasileios Peristeras

I hereby declare that the work submitted is mine and that where I have made use of another's work; I have attributed the source(s) according to the Regulations set in the Student's Handbook.

January 2021

Speyer –Germany

Abstract

This dissertation will be written as a part of the MSc in e-Business and Digital Marketing at the International Hellenic University.

By this Master Thesis we will try to prepare an analysis for one of the most important fields in business and knowledge economy, big data tools and technologies.

Big Data is a relatively new and upcoming concept. However, in the recent past, it has become a major topic for discussion and gained the attention of the industry since big data is used to provide insights for transparent and simpler products analyzing and predicting behavior through data.

A literature review will be conducted on the topic of big data tools and technologies which will be deducted by analyzing and discovering available tools and technologies for all the phases of the complete data lifecycle. We will explore the tools mostly for data-driven EU governments and will conclude with a taxonomy of these tools. We will conduct a qualitative research for the detailed analysis of the key (big) data tools throughout their lifecycle.

This Master Thesis based on the proposed roadmap will be accomplished under the supervision and kind guidance of Dr. Vasilios Peristeras, assistant Professor at the International Hellenic University, School of Science and Technology and the guidance of Syed Iftikhar Hussain Shah, PhD Scholar at the International Hellenic University, School for Science and Technology in Thessaloniki.

Contents

1. Introduction

Big Data is defined as large amount of data, produced very quickly by high number of diverse sources [1]. According to the European Union, data can either be created by people or generated by machines such as sensors gathering climate information, digital pictures and videos, purchase transactions records etc. Generating value at the different stages of the data value chain -data lifecycle- is at the center of the knowledge economy. Good use of data could transform Europe’s service industries by generating a wide range of innovative information, products and services. Also in the public sector the efficiency would be increased as well as the productivity and the economy through improved business intelligence.

As stated in the book [2], Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processed. To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data. The upcoming sections explore a specific data analytics lifecycle that organizes and manages the tasks and activities associated with the analysis of Big Data. From a Big Data adoption and planning perspective, it is important that in addition to the lifecycle, considerations should be made for issues of training, education, tooling and staffing of a data analytics team.

The Big Data analytics lifecycle can be divided into the following nine stages: Business Case Evaluation, in which the Big Data analytics lifecycle begins by defining a case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis. The Data Identification stage, which is dedicated to identifying the datasets required for the analysis project and their sources. Data Acquisition & Filtering, at this stage, the data is gathered from all of the data sources that were identified during the data identification. The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives. Data Extraction stage, which is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can be used for the purpose of the data analysis. During the Data Validation & Cleansing step, complex validation rules are being established and known invalid data are being removed. Later on at the Data Aggregation & Representation stage, multiple datasets are being integrated together in order to arrive at a unified view. Some of the most crucial lifecycles step is the Data Analysis stage, in which the actual analysis task is being carried out. It typically involves one or more types of analytics. This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered. At the Data Visualization stage, data visualization techniques and tools are used, to graphically communicate the analysis results for effective interpretation by business users. Is one of the most important and promising stage for the whole lifecycle. Even more and more data visualization tools are being developed and taxonomy of them is also part of this research. Last but not least, there is the Utilization of Analysis Results stage where is determined how and where processed analysis data can be further leveraged.

2. Literature Review

The appearance of a continuously increasing amount of data from sources such as Open Data on the Web, data from mobile applications and social network data, creates a demand for new data management strategies which can cope with these new scales of data. Big data is an emerging field where innovative technology offers new ways to reuse and extract value from information [1]. The ability to manage information, extract knowledge, cleanse the information and visualize the data by using tools and technologies is a competitive advantage for many organizations including also data-driven governments.

This chapter examines definitions and concepts related to (big) data, (big) data lifecycle stages, (big) data tools and technologies. Apart from that, a comparative review and research will be conducted in order to document existing papers books or reports regarding or topic, meaning the taxonomy of big data tools and technologies for phases of the data lifecycle.

2.1 Definitions, Concepts and Theoretical Approaches

2.1.1 What is Big Data

Several definitions of big data have been proposed over the last years. A table with definitions of big data is depicted below and was prepared by [1] .

Big Data Definition	Source
"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"	[2] [3]
"When the size of the data itself becomes part of the problem and traditional techniques for working with data run out of steam"	[4]
Big Data is "data whose size forces us to look beyond the tried-and true methods that are prevalent at that time"	[5]
"Big Data technologies [are] a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis"	[6]
"The term for a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications"	[7]

“A collection of large and complex data sets which can be processed only with difficulty by using on-hand database management tools”	[8]
“Big Data is a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies.” By extension, the platform, tools and software used for this purpose are collectively called “Big Data technologies”	[9]
“Big data can mean big volume, big velocity, or big variety”	[10]

It was really challenging to define one definition about what big data is, although the predominant definition involves the three specific characteristics of big data which are the following [11] :

Data Volume – measures the amount of data available to an organization. The organization does not necessarily have to own all of the data as long as it can access it. As data volume increases, the value of different data records will decrease in proportion to age, type and quantity among other factors. In short, it is the act of dealing with large scales of data within data processing.

Data Velocity – measures the speed of data creation, streaming and aggregation. Ecommerce has rapidly increased the speed and richness of data used for different business transactions. Data velocity management is much more that a bandwidth issue; it is also an ingest issue. Summarizing, it is the act of dealing with streams of high frequency incoming real-time data.

Data Variety – measures the richness of the data representation – text, images, video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl. It actually describes the dealing with data using different syntactic formats, schemas and meanings.

2.1.2 Big data importance

Processing and using big data offer a lot of benefits to many industries, examples of which we will see later on. By using big data businesses can utilize outside intelligence while taking decisions, they can improve the customer service by analyzing the data and are able to identify potential risks to specific products or services. In general terms, big data analysis and usage offers better operational efficiency.

In order to utilize big data to the fullest, it is not only important for businesses to have much data but rather to collect, process, clean and analyze the available data. This way big data become the mean for the business to grow and to achieve higher objectives.

According to the key performance indicators (KPIs) that every company has defined, big data analyses can help in some of the following. By using big data tools and cloud based analytics, this can result cost savings to the company especially when large amounts of data are to be stored. With these tools it is possible to identify more efficient ways in performing business. Another improvement that usage and analysis of big data brings is to use them as a driver of innovations and product development. It is possible to retrieve marketing insights and support for advertisement reasons the business. By using big data analysis it is easier to identify the customer's expectations and thus change the marketing plans in order to attract more potential customers. Furthermore by analyzing big data it offers you a better understanding on the market conditions and this helps in thriving among the competitors. Last but not least, a business can get advantage of big data analysis for time reductions purposes. The best the data are used, the quicker it enables businesses to either make adjustments or make decisions that will increase their revenue.

Big data can serve to deliver benefits in various areas both in the public and private sector. Therefore big data is key for governments in order to improve their services for the residents but also to become innovative and exceptional in those digitalized times. Big data is used from real-life scenarios such as guest services and entertainment to targeted advertising, education and of course massive industries (healthcare, insurance, manufacturing or banking). Let's see some detailed examples bellow on how industries can benefit from what big data analysis and tools can offer to them.

Big data in insurance industry

The insurance industry is not only important for individuals but also business companies. Insurance holds a significant place, since it supports people during times of adversities. The data collected from insurance sources are collected in various formats and alter at massive speeds. Insurance companies gather big data from various sources in order to understand their customers, their needs and to offer them insurance packages effectively at a calculated cost. Using big data customer insights experience and needs can be gained and utilized in the benefit of the industry. Also big data can be used for security purposes in order to detect fraud or cyber security threats.

Big data in the education industry

Some fields in the education industry that have been transformed by big data are the following. Big data are used in developing customized and dynamic learning programs; also they are used for improving grading systems and for supporting students in order to predict their career paths. Last, the analysis of big data can be used in reframing for example course material.

Big data in the banking sector

The banking sector is one of the most rapidly evolving sector in regards to big data study and analysis. Big data can help in detecting the misuse of credit and debit cards, can provide business clarity and the ability to identify money laundry cases, this way risk mitigation can be succeeded and better service quality and security can be ensured.

2.1.3 Big data in the Government Sector

Big data in government can have an enormous impact at a local, national but also global level. As it is commonly known, governments have to deal with many complex issues every day. They have to analyze and interpret all the information that they are receiving and make vital decisions that will affect millions of people. Big data applications in the government sector are described below. According to the research of [12] in the same way that companies study, analyze and use big data in order to pursue profits, governments use it to promote and ensure public good. In the last years even more and more governments are using big data in order to redefine the landscape of data management, from extract, transform and load or processes to new technologies for cleansing and organizing unstructured big data applications. The public sector has started deriving insight in order to support decision making in real time using fast growing data from multiple sources, such as web, industrial sector, social communications and so on. An advantage that governments use of big data is to overcome challenges such as better citizen’s service, decreasing health costs and create more job positions. In order to succeed and use big data at its fullest governments needed to develop new capabilities and adopt new technologies in order to transform it into useful information.

Historical Review

Let’s see in more detail some big data applications across leading e-government countries all over the world and how it started.

The first attempt of a government to manage real high streaming data was made from United States in 2002, where in collaboration with IBM they started developing a scalable infrastructure. Their goal was to manage real time analysis of very high volume data. They succeeded to create two platforms which were used from government agencies and organizations in order to visualize the information which they received from millions of sources. Later on, in 2009 United States brought to life Data.gov, which is a warehouse containing datasets, for data transparency. In 2010, a big-data strategy were designed by US which then in 2012 started to deep dive into research and development initiatives for using big data technologies in various fields. Since then and until now in US, even more and more initiatives come to life and contribute in creating single source of information about the citizens.

In the European Union, that will mostly concern us in this dissertation, the initiative “Digital Agenda for Europe” started in 2010 which aimed to develop interoperable Internet applications that will benefit the EU citizens from single digital market point of view. In 2012, the European Commission made big-data part of the strategy focusing their attention on the economic potential of having public data in data centers of public agencies. Of course they took into consideration the data protection and the increase of individuals’ trust. They started for example developing the internet of things between devices without direct human intervention and they reassured data exchanges in a secure way. In the Action Plans of European Union many companies are leading digital changes that enhance their e-government characteristic. For EU we will present more details in the following paper.

The United Kingdom government was one of the EU countries that have started implementing big-data digital programs already since 2004. Their first attempt was to use big-data technologies in order to deal with cross-departmental and cross-disciplinary challenges. Later on in 2011 in the U.K. they used big-data programs in order to visualize the drastic change on climate and its effect on food and water availability. Worth mentioning is the project that was launched by the collaboration of U.K., the Netherlands, Switzerland and 17 more EU Countries with the support of IBM, which was called DOME. DOME, created a supercomputing system with the help of which, it would be able to investigate technologies for data transport, data storage and analysis. It would provide the necessary means in order to store and analyze raw data at a daily basis.

Apart from EU and US though, it is interesting to see the effort that Asia countries put on big-data usage. Countries such: as South Korea, Singapore and Japan have shown Big-Data initiatives and strategies on national levels. The aim is always to establish pan-government big-data networks systems which goal the harmonization between government and private sectors. The usage of large-scale data for various fields Education, Culture, Science and Technology is the top priority for e-governments, since solutions and value added results can be succeeded.

Main big-data trends

From the evaluation of big-data projects and initiatives through various e-governments all over the world, we conclude that there are three main characteristics that are notable.

First, we see that the majority of the data that were used in government projects, are not only big- data and unstructured but also they share structured databases, meaning they do not always use real time data. This is a challenge that countries need to overcome in order to benefit fully from the advantages of big-data evaluation, Secondly, governments select to use big-data technologies in order to serve the citizens in better manners and to offer them digital applications and opportunities for their daily interactions with public services. Of course, projects have been initiated also for optimization in sectors such as economy, health care and so on but the focus is on the digitalization of public services. Last, it should be noted, that big data applications are an fast evolving stage, even more and more governments are being involved in projects and even more projects and initiatives are being launched.

Concluding, it is remarkable how fast are governments valuating the positives of big data analysis and usage and hoe eager they are to collaborate with various countries and with private businesses in order to receive the knowhow. The world is changing even faster the amount of unstructured, large scale data is inconceivable and if this is used for the benefit of the government and the citizens only positives it has to offer.

Let’s define below the complete data lifecycle stages, what needs to be done in order to have data ready for usage and let us see technologies that have been developed for the most important stages.

2.2 Data Lifecycle

In this section we will explain the (big) data lifecycle phases, such as data collection, data integration, data cleansing, data analysis, data visualization, data access, data security, and functions that needs to be performed to transform data, and information into knowledge for efficient and data-driven decision making.

First of all we need to define what exactly data lifecycle is. The data lifecycle contains all stages that big data need to go through from its creation for a study to its distribution and reuse. The data lifecycle starts by developing a concept of data study. Once the study is developed, data are then collected and managed. More detail will be reported later on.

Over time, more and more companies, businesses and countries handle large amounts of unstructured data. In order to handle big data many data lifecycle models have been introduces with the aim to process big data at the most efficient and effective way.

The management and analysis of the data is a challenging task for the interested party, since according to the type of knowledge that big data need to deliver. Every company, business or government should adjust to the data stages that apply more to their needs. Below we will examine some data lifecycles models that have been presented in the work of [13].

Data should be handled and managed from the very first moment where the decision was taken to use big data until they are no longer needed. Data management is such a crucial task since according to the company's need, the data phases should be defined accordingly and with precision. Effective data management means better performances and more added values for the companies or even the government. Poor preparation of data can lead to high risks for the data management, data loss, data inconsistency and bad quality results are some of the negative effects. Therefore it is important to define the correct data management stages according to the needs.

The more the data are, the more complex the task is to transform them into knowledge and into value for the business.

Following, some of the data lifecycles models that have been created over the years will be presented.

2.2.1 Data Lifecycle models

DIKW pyramid

One of the first presented data lifecycle model was the DIKW pyramid (Data, Information, Knowledge and Wisdom) which was reported in 1989 already more information we find in [14]. This model tries to display the transformation processes in which raw data are transformed into information. The information later is being transformed to the indicators that compose knowledge for the interested company.

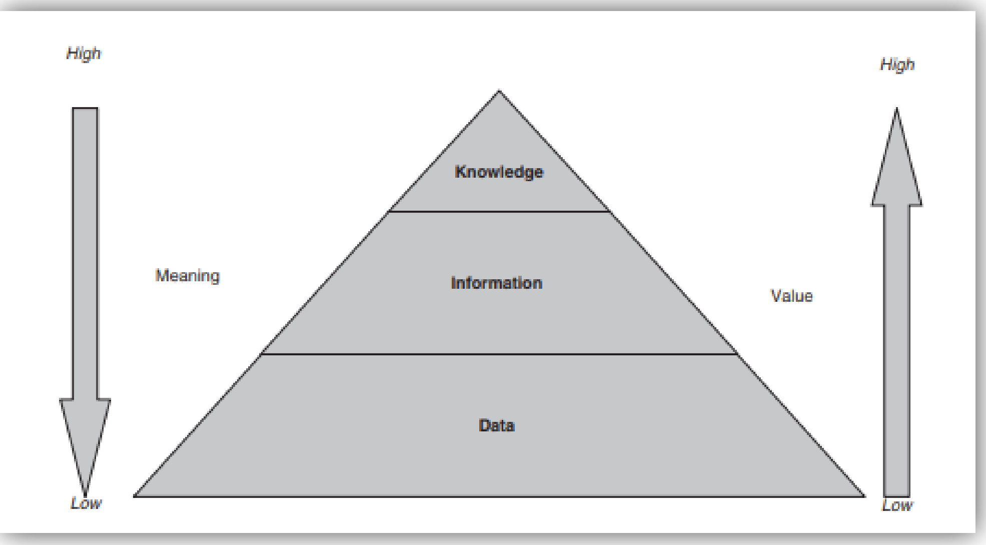


Figure1.DIKW Pyramid

According to [15] in the figure and the DIKW pyramid, data are defined, as symbols, products of observation that are not useful until they are transformed in to a particular form. Information is contains the answers which are inferred from data. Information systems generate, store, retrieve and process data. Knowledge is what enables the transformation of information into instructions. And wisdom is the ability to increase effectiveness. Wisdom adds value.

Concluding, the very large data at the base of the pyramid need to produce information and more advanced indicators that will be transformed into knowledge and wisdom.

This first data lifecycle model was the base for the development of further models.

DataOne lifecycle

The DataOne is a lifecycle model specific to the field of scientific research and has been adopted in [16]. This data lifecycle model is developed dependent on the way data moves through eight unique stages in order to succeed management and preservation of data for use and reuse. The eight phases of this data lifecycle model are the following.

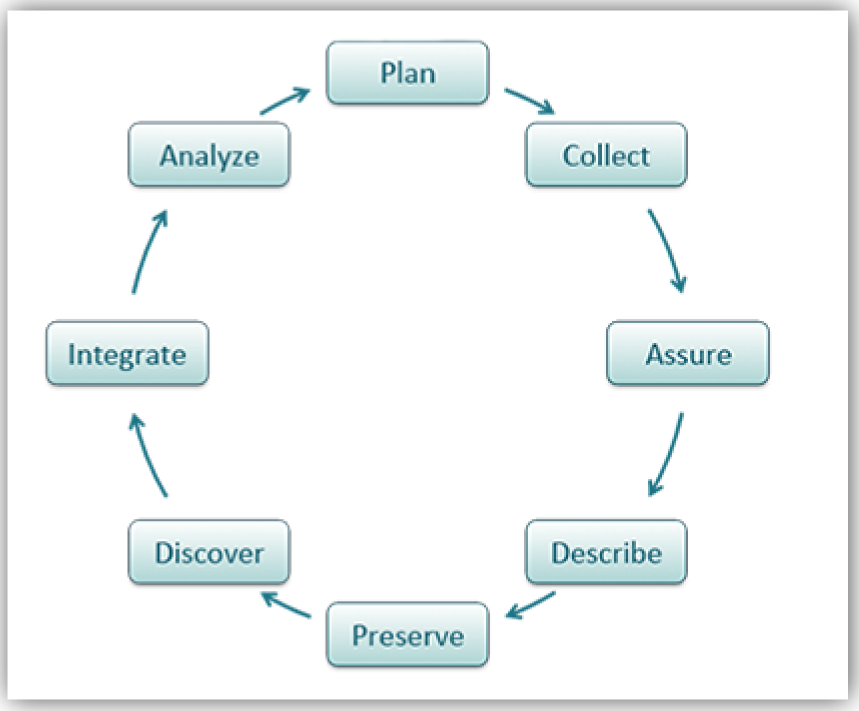


Figure2.DataOne Lifecycle Model

The lifecycle starts with **Plan**, at this phase, a description of the data will be deducted and the way on how data will be managed and made accessible will be considered. At the stage of **Collect** the relevant data are being observed and collected in a digital format. After the data are in an acceptable format, stage **Assure** follows, where controls and inspections with regard to the quality of the data are carried out. Later at **Describe**, the data are accurately described using metadata. At the **Preserve** stage, the data are already discoverable and are stored in a long term archive. After that, potentially useful data are located and obtained, together with useful information at the **Discover** stage. Continuing in the model the next stage is **Integrate**, where data from various sources are being combined to an equivalent set of data. Last comes the later stage of **Analyze**, where all data gathered are being analyzed. This analysis helps later on with interpretations and conclusion making.

The DataOne model we see that it gives further details for the data lifecycle with its many stages, although a lot of manual work needs to be done, therefore it is preferred when the volume of the data is not very large.

Information Lifecycle

This data lifecycle model is designed for the intelligent management of Big Data which focuses more on data security throughout the process of the data. It is an interesting model which consists of seven phases and according to [17] they are the following.

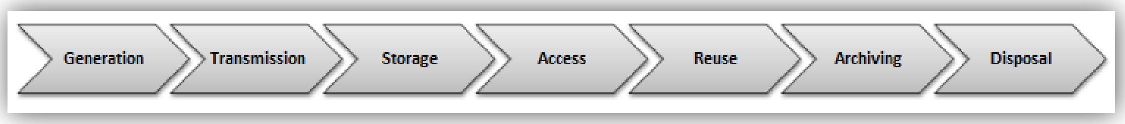


Figure3.Information Lifecycle Model

Data Generation, at this phase, the data is being specified according to the user's requests. **Data Transmission**, the liability of the data is being verified and a secure transmission channel for the data is being established. At the next phase, **Data**

Storage, the data is being stored following agreements and regulations. Then at **Data Access** it is being secured that only users with the necessary credentials can access the data gathered. At the stage of **Data Reuse,** the data is being integrated as in the traditional lifecycles while at the stage of **Data Archiving** the data is being archived in such a way that it is possible to retrieve those any time. Last at the stage of **Data Disposal**, all necessary data are moved to the cloud and anything unnecessary is being completely removed.

Data Documentation Initiative (DDI) Lifecycle

DDI Lifecycle model is an international standard model that is widely used for metadata analysis. DDI was first invented in1995 where data professionals started working on this model the latest version was updated on 2015. As described by [18] this lifecycle model involves eight steps as shown below [19].

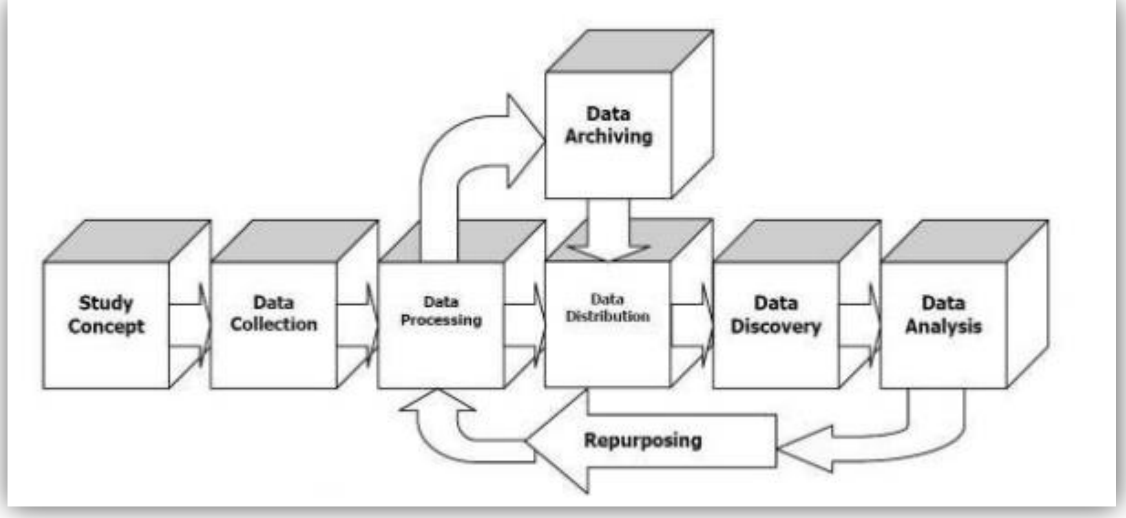


Figure4. DDI data lifecycle

At the DDI Lifecycle model, the processing of the data is done as described here.

Concept: at this stage, a global vision of the data is being defined, and all the various concepts and definitions are being organized in a hierarchical structure.

Collection: at this stage, all relevant data are being gathered including descriptive information regarding the interested survey itself.

Processing: at the stage of processing, various tasks are taken place, such as control of data, clean-up of data, data evaluations and data coding.

Archiving: this step, allows data archiving by clustering obsolete and unused data.

Distribution: all the relevant data are being distributed to the various systems.

Discovery: at this stage, the data re able to be described by metadata in the course of discovery.

Analysis: the examination and determination of data takes place at this step.

Repurposing: this step is a completely new conceptual framework, since the re-definition of data needs to be done. Specifically, the relationship between the data conceived during the design process and the possibility of defining both primary and secondary data sources in the collection phase [18].

We need to notice that this data lifecycle model uses ontologies and archiving of data which reduces effort in the business that uses it. Furthermore, the re-use of data offer benefits in various departments.

Lifecycle for Big Data

In 2014 in the report of [20] a new lifecycle model for Big Data was defined. Big data, due to the amount of data and their unstructured nature, are more complex to analyse and to bring them in a form that could be usable. According to this report, new scientific discovery methods need to be implemented in order to improve the collection, processing and usage of the data. In this model we see for the first time a filtering and an enrichment phase which were added after the stage of collection of data. The filter and enrichment have as an aim, to reduce the mass of data initially collected. The characteristic of this phase remains on the storage phase, where the data are retained during all the stages of the lifecycle. This enables the use of the data and the reformation throughout the whole period.

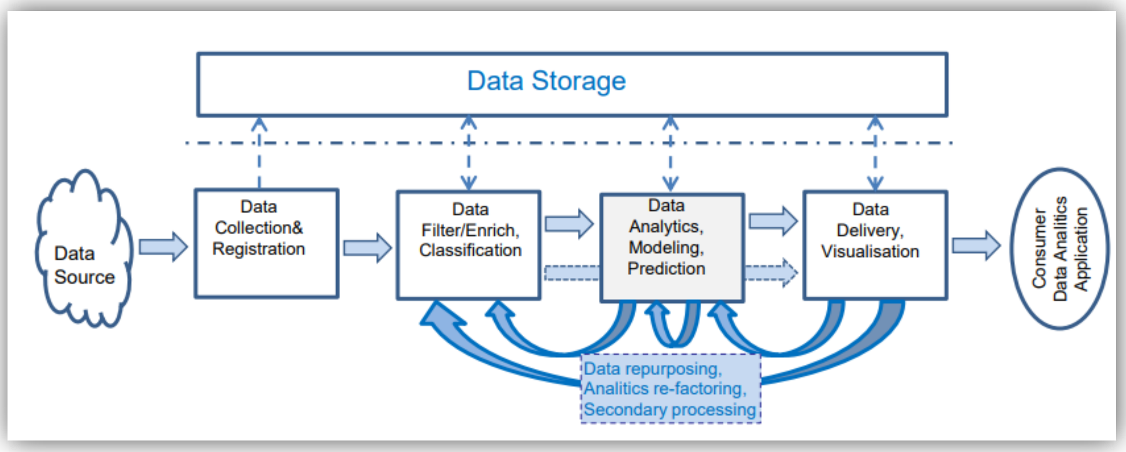

Data Storage
Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer Data Analitics Application
Data repurposing, Analitics re-factoring, Secondary processing

Figure5. Big Data lifecycle model

Data integrity, control and accountability are three characteristics that must be supported through the complete lifecycle. Verifying the reliability of the data is one of the most important components of this lifecycle. As we see also in the picture above, storage takes place during the whole lifecycle. This data lifecycle, adapted to Big Data is not very different from other traditional lifecycles. The new stage of data Filter/Enrich Classification after the data collections seems really interesting. Although, it should be investigated further more whether or not it is a compatible model for big data, since the data storage step does not facilitates the high volume of big data.

Hindawi Lifecycle

According to [21] , during each stage of the data lifecycle, the management of Big Data is the most demanding issue. The data lifecycle that was presented is using technologies and terminologies of Big Data. Existing practices are analyzed but with a different scientific approach. This lifecycle consists of the following stages: collection, filtering & classification, data analysis, storing, sharing & publishing, and data retrieval & discovery. The figure was also presented in the same report.

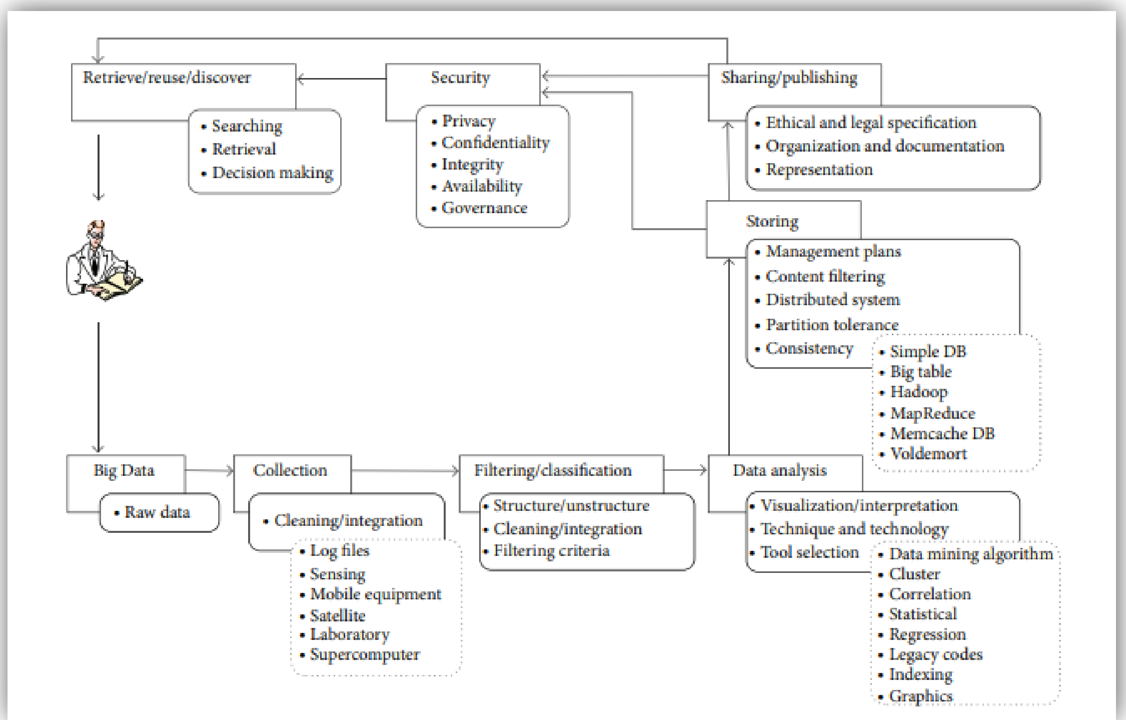


Figure6. Proposed data life cycle using the technologies and terminologies of Big Data

Big Data, the lifecycle starts with that raw data collected by researchers and organizations. At this stage, the data increase their value through input from individual program officers, or scientific research projects. All data are being transformed from their initial state and are stored in a value-added state.

Collection, data collection is traditionally the first stage of any data lifecycle. In data collection, special techniques are utilized to retrieve raw data from various data sources and environments. Large amount of data are created in forms of log files, data from sensors, from mobile equipment, from satellite data, laboratory or even supercomputers.

Filtering & Classification, the data collected until now are in a very high volume but not that structured. At this phase, the classification of structure and unstructured data is enabled, as well as filtering with specific criteria.

Data Analysis is a challenging task due to the complexity of the data, although it is one of the most crucial steps since it gives the opportunity to handle abundant information that could affect the business. Data analysis aims to understand the relationships between features and to develop methods data mining that could predict observations. Techniques that can analyze such large amount of data are some of the following: data mining, visualization, statistical analysis, and machine learning. Data mining are preferred for big data, since can automatically discover useful patterns in large datasets.

Storing/Sharing/Publishing, at this stage data are collected and analyzed for storing, sharing and publishing, to the interested audience, e.g. to governments, researchers, industries, communities etc. Big data due to their extensive datasets, they need to be stored and managed in a reliable and accessible manner. As far as the sharing/publishing is concerned it is a real challenge to develop a large scale distributed system for storage, processing and analysis.

Security, this stage of the lifecycle describes the security of data and also it clarifies the roles in the data management in order to protect legitimate privacy, confidentiality and intellectual property.

Retrieve/Reuse/Discover, data retrieval ensures the data quality, value addition and preservation by reusing existing data. The reusability of published data must be guaranteed and specified in scientific communities. European Commission supports Open Access to scientific data from publicly funded projects.

In this lifecycle we recognize that due to the fact that filtering precedes the analysis, the large amount of data flow is being reduced and controlled.

2.2.2 Big Data Lifecycle Stages

In the section above we discussed about various data lifecycle models not only for big data but also for normal data and for different principles. In our report below we will now concentrate on the lifecycle phases that concern only Big Data. Afterwards on our analysis we will choose the most developing and promising phases in order to find out which kind of tools and technologies are being used in order to perform the specific phase.

As we have already mentioned, Big Data is different from traditional data. Their characteristics such as volume, velocity**,** variety but also the complexity of the datasets in the data ecosystem, makes it challenging to develop one universal approach as the Big Data lifecycle. This field is still evolving and each industry according to their needs is developing a different lifecycle which covers as much as possible their final goal. Of course approaches from traditional data management are also being transferred and used for Big Data use cases as well, but its solution needs to be enhanced. The phases below have been proposed from [22] and are the one that we will endorse for our analysis later on as well.

- **Business Case Evaluation,** a well-defined business case is the first stage of a Big Data lifecycle. The business case should be created assessed and approved from the stakeholders in order to start the data analysis.
- **Data Identification** follows and is the step where the datasets required are being identified for further analysis according of course to the project, the business case selected.
- **Data Acquisition and Filtering** is the phase where the data are being gathered from all the sources that where identified in the previous step and a filtering is applied, meaning the valuable data is being separated from corrupt or invaluable datasets.
- **Data Extraction** is the fourth stage of the Big Data lifecycle. During the data extraction, diverse data are being extracted and transformed into a format that the Big Data solution can use for the purpose of the data analysis. This step is needed because some of the data which are identified as and input for the analysis may come in incompatible formats.
- **Data Validation and Cleansing** is one of the most interesting stages. Invalid data can fabricate wrong analysis results. Since data input in Big Data analysis can be unstructured without validity, therefore stage 5 is crucial. At this stage complex validation rules are being established and any invalid data is being cleansed.

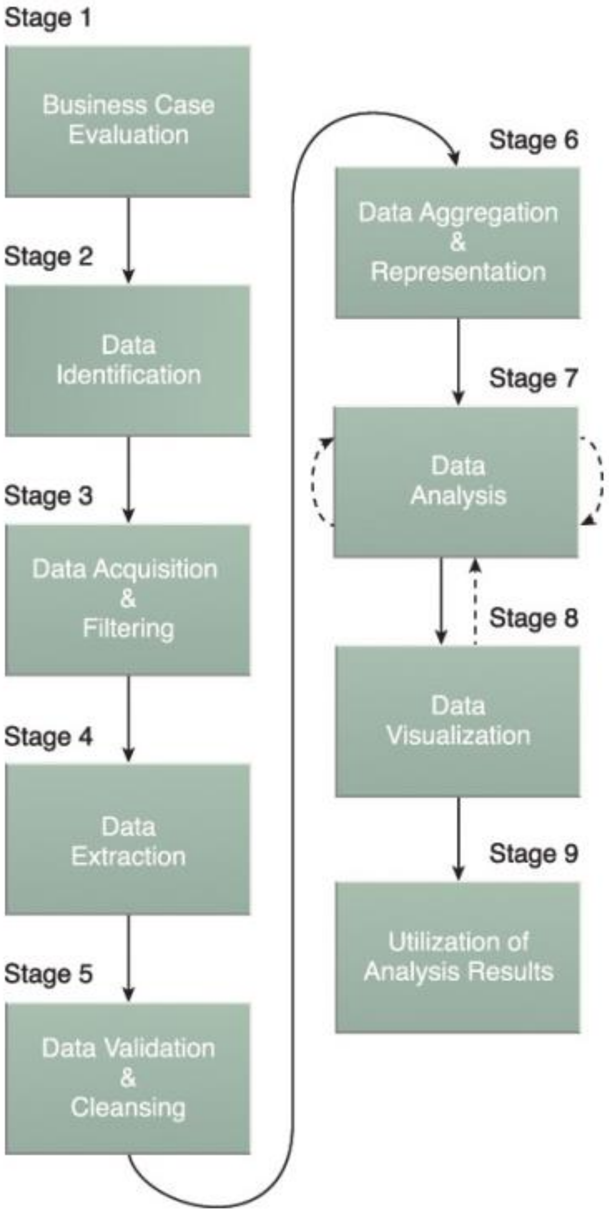


Figure7.Big Data Lifecycle Stages

- **Data Aggregation and Representation,** is a step dedicated to integrate multiple and various datasets in a harmonization and homogeneous view. The challenge at this stage is that data may appear in two or three datasets, or they need to be adhered with other fields in order to make sense. This aggregation takes place here.
- **Data Analysis,** after many various steps and a lot of data processing, we have brought the data into a form that an actual analysis can be carried out. Generally, the analysis task involves more than one type of analytics

according the type of the required result this step can be simultaneously simple but very challenging. Sometimes combination of data mining and complex statistical analysis techniques may be required in order to discover patterns or relationships between variables.

- **Data Visualization** is the phase where data visualization techniques and tools are being used to communicate and visualize the analysis results for effective interpretation by the business users.
- **Utilization of analysis,** is the stage where is defined how and where can the processed analysis being furthered used. Processed analysis could be for example dashboards that support business for decision making or specific reports etc.

From the above lifecycle steps, we will select the most interesting and innovative one and will describe tools and technologies for them in the next chapters.

In the section bellow we will continue the literature review regarding tools and technologies for big data the historical evolvement.

2.3 Big Data Technologies

As explained until now, Big Data is a term that refers to combinations of data with a large size, with complex characteristics and with an uncontrollable rate of growth. All these characteristics make Big Data challenging to capture, manage, process or analyze. These tasks are not possible to be performed by using conventional technologies and tool such as statistics or visualization packages. In order to analyze data in such high volume and in so many different data formats, big data analytics is usually required. The analysis is being performed by using specialized software tools and applications for predictive analytics, data mining and of course data optimization. Some of the tools and will concern us on the following section. We will try to provide a review on various tools that have been studied and their development.

We have to mention that according to the Big Data lifecycle step that we are examining various tools and technologies exist.

2.3.1 Big Data Challenges

For us to understand the development of Big Data tools and technologies we need to understand first of all the challenges, to identify, classify, analyze and utilize Big Data. According to the work of [23] of course Big Data generate opportunities for various sectors and businesses. Some of the major challenges in Big Data analysis include data inconsistence and data incompetency, scalability and data security. The first step for data analysis as already mentioned is that data must be transformed to a well-constructed form. In order to improve the quality of the data and the analysis results, some data processing techniques need to be performed. Such techniques are data integration, transformation, reduction which could be applied in order to deal with the data sets that come from heterogeneous sources.

Let's see the challenges according to the different sub-processes as described by [24]. We will start by describing the challenges that analysts face regarding *data capture and storage*. Big Data has changes the way of capturing and storing data, including different storage devices, architectures and mechanism. The accessibility of the data is one of the most crucial point. Data should be easily accessible and at the right format for further analysis. Also in regards to the storage, that existing storage architecture has severe limitations when it comes to large scale systems. Optimizing data access is a way to improve the performance of data analysis. If we deep dive in to the challenges of *data analysis* challenges we see that the first that comes up is the Big Data volume. So the biggest and most important challenge is scalability when we deal with Big Data. The last few years researchers have been focused on developing analysis algorithms to deal with the increasing and changeable data. In aspect of Big Data analytical techniques include not only hardware architecture adaptation but also software architecture adaptation. *Data visualization* is another area in which Big Data face severe challenges. The main objective of data visualization is to present the results into a more intuitive and effective way by using usually graphs. For Big Data applications it is particularly difficult to perform data visualization due to the high volume of the data. Big Data visualization tools are still under development and years after years even more have been invented and more technique are being used some of them we will analyze later on.

2.3.2 Big Data Techniques

Big Data techniques for analyzing the large volume of data need to be driven by specified applications in order to process efficiently the data into low run times. Those applications are for example; machine learning applications, data mining applications or statistical techniques which are able to explore patterns for the high volume of data. Let's explore below some of the most usually common data analysis techniques distinguished by the various disciplines that exists.

Optimization Methods have been used during the last years to solve quantitative problems in a lot of fields such as engineering and economics [25]. Global optimization problems have been addressed with various computational strategies such as genetic algorithm which can become highly efficient. We identify however the challenge that these computational solutions have very high complexity in memory and time consumption. Real-time optimization may also be necessary for Big Data applications. A research on real-time optimization has been performed by [26]. Another approach for data optimization is data reduction. For data reduction, there are two main factors described by [27] , in order for data to be reduces, they need to be dimensionally reduced and then compressed. For the researchers to be able to extract useful information from high dimensional data a proposed methodology is the Dynamic Quantum Clustering (DQC), which is based on quantum mechanics techniques. With this technique, hidden structures of data could be exposed and the interpretation of information is done in a significant. Way. Another method for dimensional data reduction is the Feature Hashing (FH) method. The FH method does not preserve the data quality. In order to move to the second approach which is the data compression, the idea of clustering is being used. Data compression method can achieve online clustering by analyzing correlating similarities of data.

Another data analysis technique that also uses data clustering and classification is Data Mining.

Data Mining, involves methods from machine learning and statistics and its enables the extraction of valuable information from datasets including also regression. Big Data mining is generally more challenging than the traditional one, but it can be used to cope with huge workloads. Existing methods for clustering can be enhanced also for Big Data; some of them involve the hierarchical clustering, K-Mean and so on. An application of data mining for Big Data is bioinformatics which is developed from traditional biology to approaches that combine integrative database analysis. According to [28] , the data mining process for Big Data is divided in six steps. The first four which include, data cleansing, integration, transformation and selection is known as the data preprocessing technique and the next three steps, involving the data mining engine, pattern evaluation and knowledge presentation are known as the mining process. The Big Data Mining process require new developed data mining engines and techniques since different type of data are being handled and they come from various sources so higher speed for mining is necessary.

Machine Learning is a technique which is worth mentioning, since it has been an evolvement of artificial intelligence. With this technique, algorithms are designed that allow computers to generate behaviors based on empirical data. The most impressive characteristic of machine learning is the discovery knowledge and the ability for machines to make intelligent decisions automatically. In regards to Big Data, machine learning algorithms need to be upgraded and adapted in order to deal with both supervised and unsupervised learning. Some frameworks that are used for Big Data machine learning are the Map/Reduce method, DryadLINQ and IBM parallel machine learning toolbox. A technique which has a wide range of application coverage and in used in artificial intelligence is the Artificial Neural Network (ANN). ANN are computing systems designed to simulate the way human brain analyzes processes and information they use layers and nodes. The more nodes in neural network exist, the higher accuracy they can provide. The most utilized networks for artificial intelligence are based on statistical estimations and classification optimization. The combination of deep machine learning and parallel training implementation techniques provide potential ways in processing Big Data.

Visualization Approaches are the techniques which are used in order to display data in a graphical, more intuitive way in order to interpret and translate the data into knowledge. For Big Data visualization to work, the data size needs to be significantly reduced before the actual data reading. Also the way of how Big Data are being visualized is critical and often affects the final result. Visualization methods have the aim to understand the relationship between attributes with respect to the context, to find out data which are far away from center and to understand the underlying structure of the hidden patterns in the data. All these could be visible to the researches by using graphical representations in a form much more compatible and user friendly.

Big Data Visualization is still on of the most challenging area and needs to be developed.

2.3.3 Big Data Tools and Technologies

Until now we described some of the Big Data techniques that have been developed and used, lets us explore now some Big Data Technologies with their specifications and characteristics.

Big Data technologies can be divided in two main categories:

1. Operational Big Data technologies, which cover the generated amount of data on a daily basis, such as online transactions, social media or any data form specific companies used for the analysis with the support of dig data software. These data act as raw data to feed the analytical big data technologies. Some cases of operational big data technologies are, online trading, online ticket booking or even purchasing form online shops.

2. Analytical Big Data Technologies refer to the advance adaptation of big data technologies. It includes the real investigation of high volume data that is crucial for business decisions. Some examples in this category are stock marketing, weather forecasting or even medical-health records.

Some of the most upcoming Big Data Technologies that influence the IT industries and the market in the recent time are the following. [29]

Artificial Intelligence (AI)

We discussed a little bit for artificial intelligence on the paragraph above, but we will look into depth what is its contribution in Big Data. Artificial Intelligence is a wide-band computer science that designs smart machines which are able to accomplish various tasks. In order to design those machines human intelligence is also mainly required. Some examples in our nowadays life of artificial intelligence machines are, self-driving, face detection devices, digital assistants, chabots, e-payments and so on. Artificial Intelligence enables the likelihood to achieving a goal by intellectualizing and making decisions. Big data and artificial intelligence have reciprocal relationship, since, AI depend heavily in the success of big data and the other way around, the more data are used for the machine learning models, the better the machines become. With higher amount of data and better quality, better insights into business problems can be given and therefore the better the problems can be solved.

Furthermore, using data from different sources and in different formats, artificial intelligence can increase knowledge and as a result, it enables more accurate predictions and better analysis.

NoSQL Databases

This big data technology refers to separate databases technologies that are designing modern applications. It represents a non-relational database that implements methods for retrieval of data. Those databases are widely used for big data analytics. The NoSQL databases store unstructured data, such as big data but nevertheless they provide higher performance, flexibility and adaptability for huge amount of data and for various datatypes. Examples of these databases are MongoDB, Redis and Cassandra. Particularly, NoSQL databases have the following main characteristics. Flexible schema, unlike the relational model here that differences in attributes or structures is supported. Often the databases have no restriction or even understanding on the schemas, it is completely left up to the application to maintain the consistency in the schema. NoSQL databases follow *the BASE* approach. The BASE approach covers the basically availability, soft state and eventual consistency instead of the ACID(atomicity, consistency, isolation, durability). We see that NoSQL databases are more compatible with Big Data than the relational one. Different categories of NoSQL include document databases (e.g. MongoDB), key-value databases (e.g. Redis) and graph databases (Neo4j).

Data Lakes

Data lakes refer to a repository of data in all formats both structured and unstructured at any scale. In the process of data gathering and initializing, big data can be stored at their initial form, without transformation into structured data. Data lake technology enables the various types of data analytics from the raw data un processed, such as dashboards, data visualization to big data transformation, real-time analytics and of course machine learning. Data lakes are being used from organizations and firms that have the need to explore new types of analytics such as machine learning, or have sources of log files from various inputs such as social media and streams. Data lakes support business to respond faster to opportunities, to sustain productivity and take better decisions. Let’s see some of the advantages if data lakes [30]. Data Lakes are user friendly. Instead of having various storage locations, in data lakes there is a central storage location for all data and all applications. This enables capacity, security and logically higher performance. Furthermore, data lakes are more flexible. By using native mechanisms, data lakes allow standard-based access to data. This enables companies to separate IT components so that they could be used separately the one form the other. Another characteristic of data lakes is the compatibility. They work with many applications, tools and technologies and have not many restrictions. All Operating Systems are supported, such as Window, Mac, Linux, Unix or Hadoop. Last but not least, data lakes provide high efficiency, since they require les storage in datacenters and there is no need for data duplications or parallel systems. We can conclude that data lakes are a compatible technology for Big Data.

Blockchain

Blockchain is a technology that is still under development, since it has been invented only the past few years. Many of the colossal companies such as AWS, IBM, and Microsoft have started introducing solutions in blockchain technology. Blockchain is briefly a Distributed Ledger Technology that makes history of any data unchanged and transparent through the use of decentralization and cryptographic hashing. Blockchain provides a highly secure ecosystem and is mostlz used for various applications of big data in industries of banking, finance, and insurance, healthcare and so on. Using blockchain technology for Big Data analytics process, two main demands of Big Data are being fulfilled. Big Data generated from blockchain technologies is secure, as it needs to comply with the network architecture. And the blockchain Big Data are valuable. The big data is structured, complete and are the perfect source for further analysis [31].

Hadoop Ecosystem

The Hadoop Ecosystem is composed of a platform that supports finding solution for problems in Big Data. The ecosystem includes a variety of components and services such as ingesting, storing, analyzing and maintaining. Hadoop ecosystem includes Apache projects but also commercial tools and solutions. Hadoop Ecosystem has mainly two components HDFS (Hadoop Distributed File System) and MapReduce (Data processing using programming). HDFS is inspired by Google’s file system while MapReduce engine enables the processing of nodes [32]. MapReduce is a processing framework, which works based on the divide and conquers approach with the aim to minimize the completion time of a task. Hadoop was originally designed to overcome the scaling issue without interruption using three main processes, storage, processing and management of resource. Nowadays, Hadoop is mainly used for batch analysis of large sets of data. The key features offered by Hadoop and are advantages for Big Data are the following [33]. *Accessibility;* Hadoop runs on large clusters of off-the-shelf machines and therefore it can provide easy access to various systems. There are no barriers of distance. R*obust;* Hadoop can easily overcome problems and inefficiency on the system since it runs on off-the-self hardware which are inexpensive, widely available and interchangeable. *Scalable;* Hadoop can handle more flexible larger datasets, since it can simply add more nodes to the clusters. *Simple;* Hadoop ecosystem provides the availability to the programmers to write quick and efficient programs in parallel without having any restriction in the language. *Cost Effective;* Hadoop uses off-the-self hardware so no expensive servers or systems are required.

The working environment in Hadoop provides various analysis tools, data warehousing, data querying and data mining tools were also machine learning algorithms are included. All the above characteristics of Hadoop make it a compatible candidate for Big Data analysis.

Apache Spark

Apache Spark is a unified analytics engine for large-scale data processing. With already build in features for SQL, machine learning and graph processing; Apache Spark is one of the faster generators for Big Data transformation. It supports many languages of big data, such as Python, R, Scala and Java. The characteristics of Apache spark is speed, it can run workloads 100 times faster than a similar engine such as Hadoop. Apache Spark can achieve high performance using query optimizer and a physical execution engine. It is easy to use, since Apache Spark offers over 80 operators that make it easy to build parallel apps and can be used interactively from various shells. Furthermore, Spark owns a stack of libraries including SQL and DataFrames for machine learning, GraphX and Spark Streaming, where all of these libraries could be combined in the same applications. And another important advantage of Apache Spark is that it can run everywhere. Sparks can run on Hadoop, Apache Mesos, Kubernetes, it can stand alone or in the cloud. It can access diverse data sources. Apache was designed with the main objective to decrease the big data processing time. This is succeeded by decreasing the waiting time between interrogating and program execution timing. The Spark might be used within Hadoop mainly for storage and processing purposes. It is much faster that the MapReduce component that Hadoop uses.

Predictive Analytics

A part of Big Data analytics, attempts to predict future behavior by analyzing historical facts. Predictive analytics use a variety of statistical techniques such as data mining, predictive modeling and machine learning that help accomplish its goal. With the tools and models of predictive analytics, companies investigate history and current data in order to discover patterns, trends and behaviors that occur at a particular time. An example would be to explore the relationships among various trending parameters. Those models are designed to assess the risk on specific set of possibilities. Using Big Data, predictive analytics can now be applied in much wider processes in a company [34]. More means and mechanisms are available to collect and measure results of the company's decisions. Furthermore, predictive analytics offer various possibilities since it is now possible to link big data with costumer's metrics to performance of internal supporting functions.

Prescriptive Analytics

Prescriptive Analytics is one of the final stages of business analytics and involves both descriptive and predictive analytics. Prescriptive analytics make use of machine learning, to help enterprises to take decisions based on a computer program's predictions. Prescriptive analytics has many benefits that enterprises can take advantage of with Big Data, such as improved use of resources and increased insight into patters and habits of customers. Furthermore prescriptive analytics is also useful for front line workers, since it supports analysis in to a detailed and personalized way. Moreover, prescriptive analytics enable the analysis on high scale data such as industrial data so that companies can scale internal decision processes. Prescriptive analytics uses mathematical programming, heuristic search and stimulation modeling to identify the optimal actions.

According to [23] the advanced technologies for developing Big Data science has the purpose to invent more scientific methods of managing, analyzing, visualizing and exploring information from heterogeneous and large scale datasets. The effort to understand and process Big Data has as a final target the economic benefit and the social evolution. Some more complex and ongoing technologies to utilize Big Data are Granular computing, which indicates a computational theory where granules e.g. classes, clusters, subsets, groups and intervals are effectively used to build and efficient computational model for complex applications with huge amount of data and information. Another technology is Bio-Inspired computing, where the way a human brain works can indicate ways to interact with Big Data. Bio-inspired computing, captures and processes sensory data received every moment of a day in a efficient way. Moreover, an upcoming Big Data technology is Quantum computing. Quantum computing combines ideas from classical information theory, computer science and quantum physics [35]. A quantum computer has a much larger memory in comparison to the traditional computers, which can process at the same time a big set of inputs. Last a really interesting technology which we will analyse a little bit more is Cloud Computing.

Cloud Computing

Cloud Computing [36] is a new style of computing in which dynamically scalable and often virtualized resources are provided as services on the internet. Cloud computing offers high availability and easy scalability therefore is an appropriate technology for Big Data. In simple words, cloud computing is the use of virtual computers. This technology can deliver applications and services over the internet, but also can be extended also to infrastructure as a service such as AmazonEC2 or platform as a service, such as Google App Engine and Microsoft Azure. Cloud Computing is a highly feasible technology and many researches has tried to apply it to Bog Data problems. Until now both, distributed MapReduce and cloud computing are being combined to get an effective scalable computing. Cloud Computing offers CloudView, a framework for storage, processing and analysis of high volume data in the cloud computing environment. CloudView is formulated by using the Map-Reduce model and it is performed in real-time. The advantages that cloud computing offer to Big Data are flexibility, the possibility to transfer and share easily large dataset. We need to mention although that cloud computing has as a negative point the low time and cost efficiency since it is challenging to upload or download big data in the cloud.

There are many more big data techniques but we concentrated only in some of them. We will explore those techniques also later in the next chapters where we will try to do taxonomy of Big Data tools and technologies.

In the next subsection we will identify some of the lately developed tools to harness Big Data.

2.4 Big Data Tools for major Big Data areas

In this subsection we will present taxonomy of Big Data tools and platforms which were performed by [37] separated into the major areas of data processing which are Big Data Infrastructure, Big Data Storage, Big Data Computing and Retrieval and Big Data Service Management. We will explore the various tools and techniques that were developed for these particular areas.

Big Data Infrastructure

Before getting into detail on to Big Data tools it is really importang to understand how big data clusters work. We have mentioned in our report many times clusters and clustering in the various steps of data lifecycle processes. The concept of Big Data Clusters has two main classifications. It includes the Cluster Configuration and Topology, which deals with the model of how Big Data are separated into various machines/nodes with regard to the service each one hosts. The second class is Cluster Deployment, which handles the deployment of those nodes into hardware infrastructure.

Cluster Configuration and Topology; a Big Data Cluster is separated into the Data Nodes and the Management Nodes. The Data Nodes are used in order to store the data in a easily distributed way and to process them for transformation and access. While the Management Nodes are used as a frontage for the enterprises applications with the aim to execute the various use cases. Both data and management nodes are connected over network. There two main types of networks used for Bog Data Clusters. The first is the data network, which is used from all nodes in order to interconnect and is used for data ingestion, data processing and data access. The second network it the management network, which is used for the management of all nodes. In order to succeed its task, various tools such as web interfaces are used.

Cluster Deployment; one of the most efficient cluster deployment architecture that is discovered is the Shared Nothing Architecture (SNA). As described until now, Big Data used cases need accessing and storing of an enormous amount of data and that accessing storing task is a time consuming process. The SNA architecture ensures the availability of enough space to store and access data in parallel for many applications. Furthermore, big data clusters can be deployed also in various other ways such as in virtualized environments or in Appliance modus too.

Big Data Storage

Big data storage is one of the most crucial and important step for data processing. The large volume of data and the different data formats need to be stored appropriately so that it can later on be efficiently retrieved and computed. Data Storage involves key concepts such as data model, partitioning, replication, compression, format, indexing and persistence.

Big Data technologies reinforce various types of data models to represent the data for access and exploitation. Some of them are the NoSQL Databases, which we already covered before, the Graph data model, where data is presented as nodes and links. At the Graph model, problems such as finding relationships between nodes, connectivity or computing distances can be explored. Moving on to the data partitioning task, every big data technology need to follow some approach to partition the data across the various data nodes. Multiple approaches exist for data partitioning. Some of them are range partitioning, where the range of an attribute is the key to partition the nodes. Hash partitioning has as a key characteristic the hush function which is applied to the attributes of the nodes. List partitioning in which each partition is specified as a list. By moving on Data Replication increases data availability and provides the same data in multiple copies. This has as an advantage the security of the data since even if there is data loss due to hardware break out, the same data can be found in many data nodes. Data compression refers to the ability to reduce the volume of data. Some of the tools that are being used in order to compress the data are Gzip, LZO, Snappy and bzip2, which has the highest data compression ratio in comparison to the other three. Ongoing to the data formats, at this point the different storage formats that are used to store data and process data are being includes. Some of the storage formats that exist today are: Delimited Text Files that are the files that consist data in a humanly readable format such as CSV and TSV files. Optimized Row Columnar Files or ORC which covers a columnar storage format, this storage format is popularly used by Hive. Another storage format is Sequence Files, which are consisting of binary value pairs. There are three types of sequence files formats, uncompressed value records, only compressed value records and value records separated and compressed in blocks. This type of storage format is being used for data that are difficult to split such as XML and JSON. Continuing with the data indexing, this big data technology is being used in order to identify particular record in files by identifying the block of data it belongs before, and it also helps to identify the exact location of the data record within a specific block. Data persistence, covers two approaches, the first is the storage of the data in local disk of each data node and the second handles the set of distributed file systems (APIs) which can be implemented by an external vendors ensuring all requirements are followed.

Big Data Computing and Retrieval

After the data processing is done and the data are stored persistent and available the stage of data computing and retrieval takes its place. Big Data computing and retrieval can happen in three ways. By processing high volume data in rest, by processing high volume of continuously streaming data and by accessing data randomly for reading/writing in a high volume data in rest. Let's explain in detail the important components below.

Distributed Processing Engine is one of the fundamental abstraction methods for big data computing and retrieval. The distributed processing engine addresses the need for ingesting, processing, filtering, querying, modeling exporting and archiving high volume data in to the big data infrastructures.

Directed Acyclic Graph (DAG) Based Distributed Processing Engine is an approach where each task gets classified into a various set of tasks. This approach works by partitioning the input data into various data nodes. In each of these nodes, a specific task, e.g. Task1 gets executed and the outputs from the nodes become the input to the next nodes which solve the second or third task this process goes on until the end of the tasks. At this approach the tasks are executed in parallel by shipping the necessary functions to the data nodes required.

Multi-Level Serving Tree (MLST) Based Distributed Processing, is an approach based on the multi-level serving tree popularized by Google Dremel. This approach uses the concept of serving tree with multiple levels to execute a job. As soon as a server receives an incoming query from a client it rewrites the query into subqueries based on metadata information. The intermediate servers perform a parallel aggregation where alla the parallel results of the query are being assembled back in the root server.

Long Running Shard Processes Based Distributed Processing Engines is used for providing high throughput random read/write application and search. This is a special approach where, instead of many levels there are only two used. The master processes and the slave processes. Once the appropriate pairing is identified, the master processes connect directly to the client and interaction happens between them.

There are more distributed processing engines, which will be not analyzed at this stage of the report.

Application Component is the connecting element between the distributed processing engine and the client query. It used one or more engine to fulfil a request coming from the client. The application component manages the requests, their security and connection pool etc. these components are long running processes in order to be able to handle the incoming requests. In some cases they can also be scheduled to run periodically.

Data Access Interfaces

One of the widely known tools for data access even in big data technologies is SQL. Apart from SQL based interfaces, though, big data technologies support the read/write in APIs also in various other programming languages such as Java, Python, C++, Scala etc. Furthermore big data technologies such as NoSQL Databases, Stream Event Processing Technologies and Search technologies can be also used.

Big Data Service Management

This concept covers the models related to the management of services that are running on big data clusters. Some of the most important components at this stage are resource management, high availability management and monitoring.

Resource Management verifies that computing resources of all nodes and the network is available to handle all clients' requests.

High Availability needs to be ensured in order for all the requests to be handled even when a node running breaks through the task or if the process dies. The master process should always be checked if it is live and running.

Monitor needs to happen at multiple levels in big data technologies. Services, requests, software resources and hardware resources need to be monitored throughout the whole process. In some cases specific interfaces exist for this reason.

2.4.1 Big Data tools based on batch and stream processing

Big data Tools can be divided into three main subcategories, the one that are based on batch processing, the one based on stream processing and the tools based on interactive analysis [23].

Batch Processing Tools

Some of the most used tools are the following:

Apache Hadoop and Map/Reduce – Apache Hadoop is a software platform that support distributed applications. The application consists of the Hadoop kernel, Map/Reduce, Hadoop Distributed file Systems (HDFS) and many other projects such as Apache Hive and Apache HBase. Map/Reduce, is a model for processing large volume of datasets and was firstly presented by Google and developed by Yahoo. With the addition of this model, Hadoop became a powerful software framework for writing applications which process in parallel high volume data on large clusters. As discussed also before, the Map/Reduce framework works with a "master" which tracks and collect all the tasks and distributes them to the "slaves"-data nodes that implement the tasks as required.

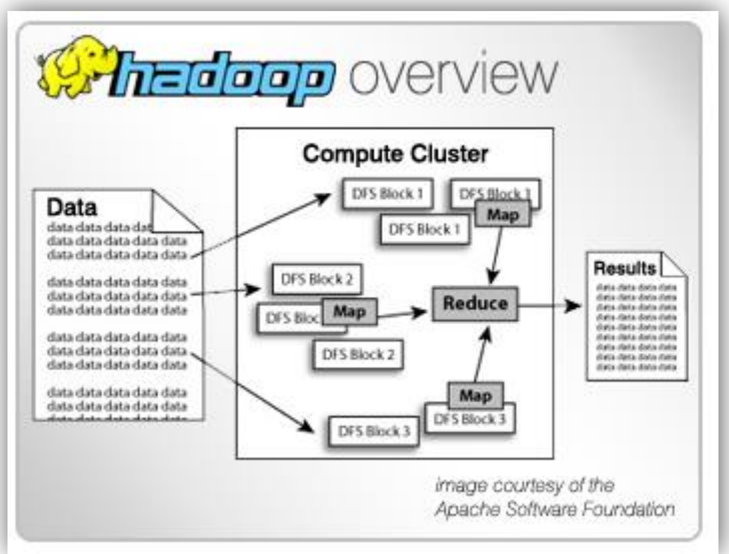


image courtesy of the Apache Software Foundation

Figure8.Apache Hadoop

Dryad **-** Is another programming model that is able to implement and distribute parallel programs. Its advantage is that it handles scalability easily from small clusters to larger clusters. This model is based on dataflow graph processing. Dryad provides a large number of functionalities such as generating the job graph, scheduling processes, handling failure on clusters collecting performance metrics, visualizing the results and more. As a model it is more complex and powerful and many applications e.g. Dryad LINQ have built based on it.

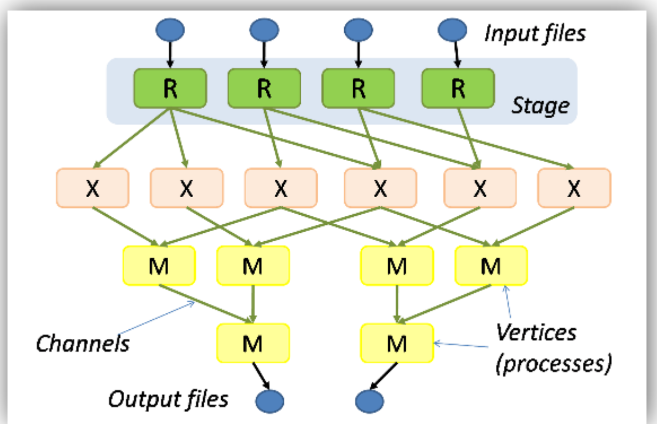


Figure9.Dryad model

Apache mahout – Provides machine learning techniques for high volume and intelligent data analysis applications. Companies such as Google, Amazon, and Yahoo! have implemented machine learning algorithms into their projects. Mahout's main component include, clustering, classification, pattern mining regression and reduction.

Jaspersoft BI suit **-** Is an open source software that is able to generate reports form database columns. This business intelligent platform is a highly scalable big data analytic option. It has the capability to visualize data fast on popular storage platforms such as MongoDB, Cassandra, Redis, Riak and CouchDB. One of the best characteristics of Jaspersoft BI Suit is that it can explore Big Data without even extracting it, transforming or loading it. HTML5 is used for the data visualization.

Tableau **–** Consists of three main components to process large scale data sets, Tableau Desktop - a visualization tool, Tableau Server – a business intelligence system that provides web-based analytics and Tableau Public- creates interactive visuals.

Karmasphere studio and analyst **-** Karmasphere is another Hadoop-based Big Data platform for business data analysis. It offers a new approach for self-service access and analytics to Big-Data in an efficient and collaborative wat. It enables the users to process their Big Data applications into a user friendly environment.

Talend Open Studio **–** is open source software, where users can perform their data analysis visually in graphical environment. It is developed form Hadoop and includes HDFS, Pig, HCatalog, HBase, Sqoop or Hive. Talend Open Studio is a user friendly and easy application to use; the interested party can easily drag and drop varieties of an icon into a canvas. The only negative is that it does not offer enough detail in order to deeply understand problems or mechanisms.

Stream Processing Tools

Stream processing tools are mostly real-time and high performance engines. For stream data applications such as processing log-files, machine to machine (M2M), telematics, all there require real-time response for processing the high volume of data. Stream Big Data is characterized by high volume, high velocity and complicated data types. Examples of real-time big data platoforms are SQL stream, Strom and StreamCloud which we will also analyze later on in more detail. To understand a little bit more the Real- Time processing concept, it means that ongoing data are being processed, this requires of course small response time and no latency. Bellow we will see some of the most used stream processing systems/platforms/tools.

Storm- is an open source distributed computation system for processing streaming data. It is especially designed for real-time processing in comparison to Hadoop which is a batch processing platform. Storm as system it guarantees that all data will be processed and generally it is easy to set up and operate. It provides many applications such as real-time analytics, interactive operation system and online machine learning and so on. Storm works also with clusters, but in order to implement real-time computation different topologies per task should be created.

S4- is a general purpose scalable computing platform for processing continuous streams of data. It was released by Yahoo and later became and Apache Incubator project. S4 allows programmers to develop applications. The processes are robust, decentralized, scalable and extensive. The S4 job is a modular job, where large scale stream data could be easily and dynamically processed.

SQLstream, s-Server – is another Big Data platform for processing large scale streaming data in real time. S-Server focuses on automatic operations and on intelligent of streaming data. This platform works fast, as it uses in-memory processing “NoDatabase” technology. It has good performances in real-time data collection, transformation and sharing and is mostly used for Big Data analytics and management.

Splunk- is a real-time intelligent platform for utilizing information from machine generated data. Splunk offers a combination of cloud technologies and Big Data technologies for analysis and monitor, which can be presented in various ways such as graphs and reports in via web interface.

Apache Kafka - works as a tool to manage streaming and operational data by using in-memory analytical techniques. Kafka has various advantages and therefore is used form many companies worldwide. Some of the main characteristics are high-throughput, support for distributed processing and support for parallel data load.

SAP Hana- is another analytics platform that aims to provide real time analysis on business processes by using in-memory analytics. Some of the main Big Data real time analytics of SAP Hana are, operational reporting, data warehousing, and predictive analysis.

Interactive Analysis Tools

Big Data tools based in interactive analysis was created form the urge to process and analyze interactively the data. Interactive analysis brings the data in a interactive environment, where users are able to perform their own analysis of information. Two of those tools are below.

Google's Dremel **-** is able to run aggregation queries oven limitless row tables in seconds by combining multi-level execution trees and columnar data layout. It was proposed by Google and is scalable for processing Big Data.

Apache drill – is another distributed system for interactive data analysis, but its characteristic is that it is more flexible and can support different data formats, data sources and query languages.

Every Big Data platform has it specific use and functionality, other are for analysis, other for machine learning and other for stream data processing. The amount of big data tools and technologies available is big and it is impossible to cover all of them. In this report the more relevant where presented.

2.5 Taxonomy for (big) data tools and technologies

In this section, we will research and report existing taxonomy papers for (big) data tools and technologies. We will try to discover the criteria of the taxonomy that other researchers have used and define what would be useful for our work.

2.6 Literature review method

The literature review process followed in this chapter is described below. First, the criteria for the choice of the references and the papers studied were defined. All papers are dependent on the useful information they contain, the period of interest and the number of citations, as well as the origin of the paper for example, if it is a journal, a book section, a conference or another source. Then, the next step was to identify the concept and to find related databases and sources. The search procedure of this thesis included the use of range of relevant sources such as IEEE Xplore, Emerald, Google Scholar, Springers and Elsevier. The resulting papers were then filtered based on year, abstract, content, citations etc.

The literature review seeks to provide a description and evaluation of the current state of big data tools and technologies. It gives an overview of the explored sources and the already existing research papers on the specific topic.

At last, the literature review and all the sources were evaluated, combined and reported here.

3. Research methodology

We will conduct a qualitative research for the detailed analysis of the key (big) data tools throughout their lifecycle. In this section, we will define the research problem, research goals, and research questions and describe qualitative research approach.

First of all, we will define what qualitative research is and why we chose to work on this paper with this specific methodology. According to [38] qualitative research is a term that covers a wide range of techniques. It is an approach that provides you with the ability to investigate and examine people's experiences and work by using various set of research methods. Some of these research methods are interviews, focus group discussions, observation, content analysis or visual methods. Qualitative research, nevertheless is not only restricted to the application of these methods. In order to be able to perform a qualitative analysis it is important to identify the issues and to understand the meanings and interpretations. The qualitative researchers have to study things by attempting to make sense of or interpret other results on a specific topic. For a researcher to conduct qualitative research, both using the research methods and incarnating the concepts and the assumptions are equally important. The purpose of qualitative research is to seek contextualized understanding of phenomena, identify processes and understand the context of people's studies.

Let us analyze below in depth the qualitative research methods and the steps that we followed in order to perform our analysis. And why we chose to use qualitative research.

In order to start with the research and with the writing of the paper it is meaningful to fully understand the scope and the definitions of the topic of research and analysis. After comprehending the terms the first phase for the qualitative research approach is the research design.

3.1 Research Design

Before starting our research analysis, first of the research needs to be carefully designed. The research design includes the overall strategy that will be developed in order to integrate the various components of the study and the information found into a coherent way. Of course the whole research strategy should be focused on the way how the research problem will be effectively addressed. This constitutes the blueprint for the collection and the analysis of the relevant information. By developing the research design, the researcher chooses the research methods and information sources that are best appropriate for the research question and the research problem. At the stage of research design, the control and optimization of the research methodology takes place. At this point it is important to define the clear strategy of your research in order to address effectively the questions.

The research design consists of four major steps.

1. Information Selection: at this step you define the most appropriate information that would be useful for our analysis, the source and the media of selection. In order to complete these step, it is necessary to consider the following. What are the sources of information that address directly the research topic? What is the scope of investigation and what has the history determined to be the most useful info to select. Furthermore, the procedures to obtain this information should as well be defined at this initial step. And of course the comparison on how the information where selected in the past provides us with an overview of how it would also be suitable to select various information from various sources.

2. Information processing: involves the step where all the knowledge obtained from previous researches will be processed in a way that will serve our study. What we can avoid and what we should conclude in order to answer the main question of the analysis. The main objective of this stage is to utilize the language in order to represent and interpret concepts.

3. Instrument creation: this step covers the final outcome that we want to create by analyzing the information gathered. In our case is the taxonomy of the (big) data tools that we want to present at the end. In order to have the final conclusion we will use also the personal experience on the various tools existing.

4. Application validation: At this stage we measure how the research is being validated and at which extend it could be applied and used from next researchers. At this stage it is also important to test if the information or the taxonomy made is valid and reusable.

After considering our main topic and taking into account all the components we choose to use qualitative research. We derived information from online sources, books, journals from valid web sited and editors that have high citations.

Below we will explain more about qualitative research.

3.2 Why Qualitative research

Qualitative research relies primarily on words, images and some kind of description as opposed to numbers in quantitative studies. If we try to distinct the qualitative research versus the quantitative research we see that with the qualitative research we seek understanding, while when we do quantitative research we are trying to generalize, we are trying to get for an objective more generalized information. If we look in detail another type of classification we see that Qualitative research methodology is broader, where the whole picture of a topic is investigated and it has an exploratory character. On the other side, the quantitative method has a narrow point of view, it is focused on specific topic and it has a conclusive character.

In general, qualitative research is considered more flexible in comparison to the quantitative research.

3.3 Qualitative Research Approach and Methods

When we talk about the qualitative research approaches we mean the way how the research is being conducted. There are two main approaches that are used in the qualitative research. The Deductive research aims in testing the theory. The theory is being deducted from the data. Deduction moves from more general to more specific information. The other approach is the inductive research and it is the opposite. From more specific information we generalize and explore the general relationship.

When we talk about qualitative research methods, we can group the research method into two main classifications. The first one is the population based qualitative research method where we are working with surveys, polls, interviews, focus groups and observations. In this classification human beings are the objects from where we get our data for analysis, they are the main source of information.

While the second type of classification is called the desktop research approach. The main source of information here are documents and two types of research methods are the case study research and the record keeping. In this research method class all information is derived from online researches, literature reviews and case study researches from papers already published.

Let us describe the qualitative research methods classified bellow by describing them in a few words below:

Document reviews: identify patterns or communication; describe characteristics of organizations or processes.

Observations: learn about behaviors and interactions in natural settings, study cultural aspects of a specific context.

Interview: explore individual experiences and perceptions in rich detail.

Focus groups: generate unique insights into shared experiences and social norms.

Usability Testing: inform or design decision and usability issues and find solution for them.

Case study research: research strategy that investigate an interaction within its real-life context.

Intervention research: systematic study of purposive change strategies.

Ethnographic research: observe and interact with participants and their real-life environment.

Sketching study: observational sketching generates a different form of visual data which has considerable potential.

There are many more research methods but we chose to describe only sthe most common one.

The qualitative research methodology that will be used in this paper and for our study belongs in the second main classification which consists of the detailed analysis of documents and existing papers and at the end a hand-on experience on the big data tools that will be compared. Below we will analyze in depth how document analysis is used as a Qualitative Research method. It can be spontaneous with qualitative data and it is easy to be used for analysis. However we need to recognize that there is a risk of time consumption, since it takes more time to perform an analysis. Those are the main reasons we also concluded for the qualitative research. Continuing, we recognize that qualitative research is focused on the meaning of experience. It has as philosophical roots constructivism and interpretivism, while the main goals of the research are seeking understanding, describing and discovering. Qualitative research is designed to be flexible, evolving and emergent and of course for the data selection the researcher himself is the instrument for obtaining information either from documents or from other people. There are four main goals of qualitative research which also coincides with our objective and specifically, the goals are the following.

To test, weather the theory that we found a documented is applicable to the study domain. To explain maybe why something happens the reason for its existence. Another goal of qualitative research is to provide detailed description of a specific topic. There can be cases where you have a domain where the literature already existing does not describe as detailed and thorough. So a qualitative research can be used in order to generate more detailed and descriptive documents. Last but not least, a goal of Qualitative research is to explore certain phenomena or certain behavior. When no previous knowledge is available, then exploration can be used for qualitative research. These goals can be embedded into the research design when qualitative research is being performed. About the research design we discussed on a previous section.

Continuing on with the main goals of qualitative research we should definitely mention that qualitative research could be used in order to identify and characterize, patterns of behaviors, individual perceptions, and important factors or inform predictions about relationships on specific domains.

These are a general idea on the goals of qualitative research methodology that can furthermore be specified at the research design step, where it is important to define the goal in order to adapt the research on it.

3.4 Document Analysis as a Qualitative Research Method

Document Analysis is the procedure of reviewing and evaluating documents [39]. Documents could be either at a printed form or at an electronic form (computer based, internet documents). As in any other qualitative research method, all documents need to be analyzed, examined and interpreted in order to invoke a meaning and to obtain understanding on specific topics. The procedure to analyze documents involves finding, selecting, interpreting and combining all information contained in the documents in order to address the research question. We chose for our research methodology the document analysis due to its many advantages that it has to offer and due to the fact that it is the most appropriate methodology for our research topic. Below we will be presented some of the advantages of document analysis.

Analyzing documents is less time consuming and therefore a more efficient research method. It requires selecting the data and information from existing sources instead of collecting them. Furthermore the high availability of many documents is a benefit for the research strategy. With the use of internet many documents are in the public domain and easily accessible. That makes document analysis a really attractive option for qualitative researchers. We have to mention that this kind of analysis is also cost-effective. Document analysis is less costly that other research methods and is mostly used in the cases where the collection of information is not feasible. As mentioned before, the information included in a document are already gathered and reported what is needed as an extra step is the evaluation of the content of the information and the interpretation of the context. Going on, documents are not human beings that could be affected by the research process. They do not have any "uncooperative" behavior and are easily to handle in relation to the method of observation for example where the researcher needs to be aware in order to interpret reactions and needs to evaluate possible influence to the research. This offers stability to the analysis, while documents are suitable for repeated reviews. Another critical point in relation to other research methodologies is the accuracy that documents offer. Names, references and details are accurately reported which can be really helpful on the research process. Last but not least, documents provide a wide range of coverage not only regarding a specific topic but also in time. The historical evolvement of analysis and researches is more approachable.

Of course, by using document analysis we need also be really careful on the selection of the documents for our analysis since each research has a specific objective that might not coincides with our objective. Also sometimes it is hard to retrieve specific documents so the difficulties are still there.

In our study apart from the document analysis that will be the main research methodology, also the personal experience will be an important factor in order to conduct the final findings and conclusion section that will be presented at the next section.

3.5 Qualitative Research Validation

Last topic in this section will be presented below, a topic that is really important when we conduct qualitative research methodology and it is called validation. Validation is a very important key step, as by default qualitative research produces subjective results however the role of the researcher is, to avoid the bias and to assure replicability. Generally the researchers are strongly influencing the researching approach and it is very difficult to make it objective. This is happening because during a qualitative research, we are not focused on the large sampling size and we are not looking to generalization. We are looking mainly to understand to interpret and observe relations and specific study topics. And since this work is done by the researchers it is subjective. This does not mean that the work is being undermined or that it is less important, but it needs to have a very important serious step to justify to the readers of the study that even if it is subjective and we as researchers we deliberate specific choices but we were not biased.

Let us see how we avoid bias and how we achieve replicability. A first question that needs to be posed in order to validate our research in qualitative research method is whether our study results could be generalized to a wider population. This has to do with generalization and is categorized under external validity. Continuing on we need to inspect if with our study we delivered the results that were intended. This is related to the intention of the research versus the reality and the final outcome and is referred to the internal validity. With all those approaches we look at four major concepts. When we look on the internal validation we are controlling the credibility and the confirmability of our research. On the other side when we talk about the external validity we are looking at the dependability and transferability of our study.

Therefore it is always crucial to look as researchers on these validation approaches both internal and external and to support that the outcome is not biased and it can be replicated and used also from further researchers.

To conclude, in order to perform successfully qualitative research we need to focus on the main components of the qualitative research. The research design needs to be specifically constructed in such a way that we have a clear understanding on the sources from where we collect our data, how we process our data and how we validated. Of course in order to come to these points we already need to have defined the research questions and the research problem in order to be focused during our study.

On the following section we will go in depth into the different kind of tools and we will try to do taxonomy of them focused on some specific characteristics which will be analyzed in depth.

4. Analysis and Findings

In this section, we will present our research results and analysis about the main big data tools and technologies that are best fit with phases of (big) data lifecycle. We will concentrate on the phases only that are more crucial in the complete lifecycle of the data. Also, we will taxonomy the tools and technologies divided in open source tools or commercial. Our approach would be also to get some first experience if possible with the tools in order to simplify the process for selecting the criteria.

4.1 Big data use cases and selection of lifecycle stages

In order to define which data lifecycle stages are the most important for our study, we need first of all to identify the most popular use cases where (big) data are being used. The most popular one are the following.

Internet of Things, these are numerous ways in which analytics can be applied to internet of things, for example, sensors can be used to select data that could be analyzed to achieve various insights from a variety of sources. Big data allow the gathering of rich insight to businesses. Furthermore big data popular use cases are related Information Security and Data Warehouse Optimizations. Big Data tools are being used to remove some of the challenges of the data warehouses. Even the health care industries including governments are looking for patterns that can ensure and optimize the public health care. The big challenge of big data is storing and processing the data at a specified timestamp. The traditional technologies are not sufficient, so Hadoop Technology and various data tools have emerged to solve the challenges faced in a big data environment. So there are many data tools that support users in saving money effort and time. Let us see below some of the most common big data tools divided into categories:

1) Data Storage Management some example here for big data tools are MongoDB, Cassandra, neo4j, HBASE(APACHE) which are no sequel databases. Furthermore, talend, Hadoop, Microsoft HD Insight and Apache Zookeeper are all popular for data storage management.

2) Data Cleaning, the next broad category is data cleaning. Data need to be cleaned up and well structured. Examples of such tools which helps into finding and reshaping the data into usable datasets are Microsoft Excel and open Refine.

3) Data Mining is a process of discovering insights into a database. Some of the popular tools for data mining are Teradata and Rapidminer.

4) Data Visualization tools are a useful way of conveying complex data insights into a pictorial way that is easy to understand. For example, Tableau and IBM Watson Analytics and Plotly are some common tools.

5) Data Reporting, for data reporting PowerBI tools are used.

6) Data Ingestion is the process of getting data into Hadoop form which can be done by using Sqoop, Flume or Storm.

7) Data Analysis requires asking questions and finding the answers to data. The popular tools used for data analysis are Hive, Pig, MapReduce and Spark.

8) Data Acquisition is also used for acquiring the data and sqoop, Flume and Storm are popular at this step too.

All these various big data tools have many advantages, which we will summarize bellow. Big Data Tools, provide the analysts with advanced analytics algorithms and models. They help the user to run on big data platforms such as Hadoop or any high-performance analytics systems. The tools help the user not only to work with structured data but also unstructured or semi-structured data coming from multiple choices. Furthermore, it is quite easy to analyze and visualize the data in a form that helps in conveying the complex data insights in an easier way in order for the users to understand better the data. Last but not least, data tools help to integrate other technologies very easily.

For our analysis we will concentrate on big data tools that are used for the following lifecycle steps:

- ➔ Data Analysis
- ➔ Data Visualization and
- ➔ Data Storage Management

We will concentrate our study on these stages since they are the most promising ones and more and more tools are being developed for them. We will try to stay concentrated on data driven governments. Also, we will taxonomy the tools and technologies divided in open source tools or commercial. And a clear selection of criteria for the taxonomy will be defined.

Let us start bellow by exploring the various big data tools and technologies per lifecycle stage.

4.2 Big data tools for data analysis

Big Data Tools and big data techniques for analysis is important step throughout the complete lifecycle of the big data process. Government and public sector, social networking and internet are some of the areas where data analytics tools are widely used and emerge the need for development [40]. Below, we will identify and try to taxonomy trends in the use of big data tools and analytic techniques.

Both public and private sectors experience the last years more and more problems with processing big data. This happens due to the fact that the size of population is huge and it grows even more, and in this population different age groups are being formed and each of them have their own need. As we have seen also from projects of the European Union which refer to the eGovenrment Action Plan 2016-2020 it is more and more important for the public sector and the governments to offer end to end digital public services to all citizens. Govenrments are trying to handle and process big data for the citizen's benefit and to build a modern era such as Smart City applications. These Smart City applications use electronic data collection sensors and devices connected over the internet to monitor all the information about the residents and to provide optimizations for their lives.

Let us see bellow some of the tools in detail and their differences.

Apache Hive

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Apache enables the users by providing them with an interface SQL-like to query various data stored in databases and file systems that integrate with Hadoop. Apache Hive is and Open-Source software.

Features: Apache Hive is mainly used for effective data aggregation method, adhoc querying and analysis of huge volumes of data [41]. With Hive different storage types can be used such as text, HBase, Optimised Row Columnar format and RCFiles. Furthermore the storage of the metadata is performed with a relational database management system. This reduces significantly the time to perform semantic checks during query execution. Another characteristic is that Hive has built-in user-defined functions to handle dates, strings and other data mining tools. It supports also the extension of these functions to manipulate user cases which are not supported by the already built-in functions. Going on, it includes SQl-like queries (HiveQL) which are converted into MapReduce or Spark jobs. Last, Hive is operating on compressed data stored into Hadoop ecosystem using algorithms to compress lossless data.

Architecture: The architecture of Hive is depicted below according to [42] **.**

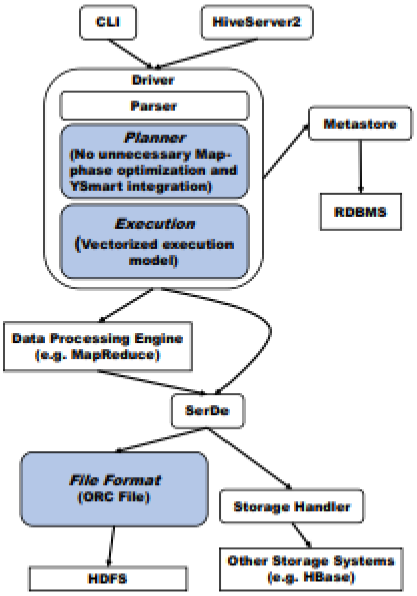


Figure10.Apache Hive Architecture

Hive provides two interfaces to users to submit their statements. These interfase are Command Line Interface (CLI) and HiveServer2 as shown in the picture above. Via these two inputs, are the users' statements submitted to the driver. The driver first parses the statement and then passes the Abstract Syntax Tree (AST) corresponding to this statement to the Planner. The Planner later on analyses the different type of statements. During this process important and needed metadata are being retrieved from the Metastore. Queries used for data retrieval and data processing are analyzed by the Query Planner. Generally, Hive translates the queries to executable forms for the data processing engine which in this case it is for example Hadoop MapReduce. After all jobs have finished, the Driver will give back to the user who has submitted the statement the results of the query. In case we have data stored in different storage systems such as HBase, then a corresponding Storage Handler is needed.

Apache Pig

Apache Pig is another platform which runs in Hadoop clusters and is designed as well to analyze and process large datasets. In comparison to Hive, Pig uses a specific language which is called Pig latin and is similar to SQL. The positive is that this language does not require as much code in order to analyze data. Apache Pig is as well open source software for data analysis.

Features: Apache Pig provides a rich set of operators like filters and join to perform several operations. It is easy to learn write and read. Furthermore it is extendable, so the user can make its own user-defined functions and processes. Generally, fewer line of codes are needed. Pig supports a multi query approach which significantly reduces the development time. The data structure is multivalued nested and richer. Going on, Pig can handle the analysis of both structured and unstructured data. Pig is mostly used from programmers and researchers for data analysis. Last pig does not support a web interface in comparison to Hive, so it is more difficult to be used from the end user that has the query.

Architecture: Apache Pig has the following major components: Parser, optimizer, compiler, execution engine and execution mode. Parser manipulated all statements or commands written in Pig Latin. At this first stage, several checks are being performed on the syntax and type and it gives as an outcome a Directed Acyclic Graph which represent all the logical operators of the scripts as nodes and data flow edges. The Optimizer, performs the optimization of the activities on the output such as split, merge, transform and so on. It removes all necessary data. At the compiler, the output that the optimizer generates is transformed to MapReduce jobs. The execution engine is responsible to produce the desired results. At the last, MapReduce step, the programmer executes the Pig Latin statement on the data already stores in the HDFS-Hadoop Distributed File System. The picture for the architecture is depicted below.

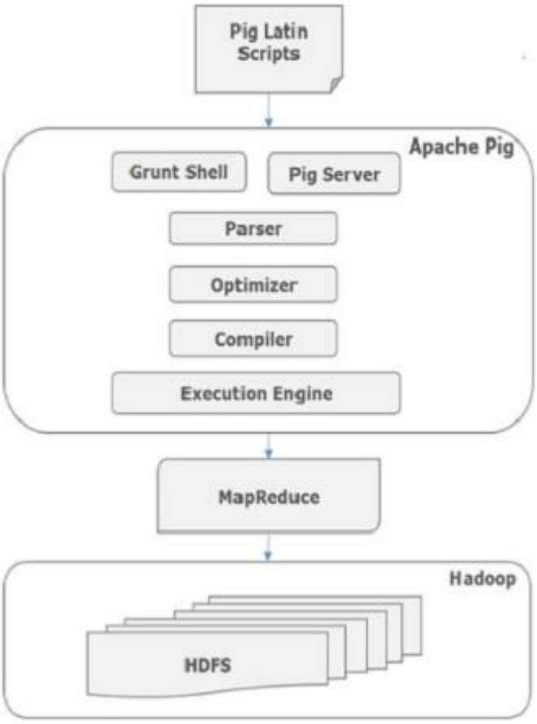


Figure11. Apache Pig Architecture

Splunk

As we have already mentioned in a previous chapter, Splunk is a real-time intelligent platform which offers a combination of cloud technologies and Big Data technologies for analysis and monitor of big data. Splunk is commercial software, meaning non open-source which its aim is to convert and communicate between machine data an artificial intelligence which makes it suitable for big data analysis.

Features: Buy using Splunk software, it is easy to universally collect either indexing or machine data from virtually any source. It provides to the end user the possibility by using powerful search processing language to search and analyze real-time and historical data in depth. It has powerful reporting and analysis capabilities and the analysis could be customized according to the users' needs. The Splunk software is resilient and scalable on any hardware. Furthermore, it is a flexible platform for big data apps and has also the ability to be integrated with Hadoop for reliable and bi-directional interoperability.

Architecture: There are three different stages in the Splunk data pipeline. The first one is the Input stage, where Splunk software consumes the raw data from the sources and transforms it into blocks. The second stage is the storage stage, where parsing and indexing are taking place. In the parsing phase, Splunk software examines, analyzes and transforms the data to extract only the relevant information. This is also known as event processing. At indexing phase, Splunk software writes the parsed events to the index on the disk. It writes both compressed raw data and the corresponding index file. This has as benefit that is can be accessible during the searching stage.

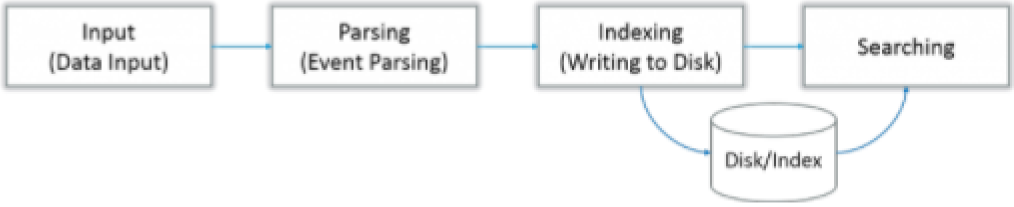


Figure12. Splunk Data Pipeline

Last, at the searching stage is controlled the user accesses, views and uses the indexing data. The search function manages also the search process.

After explaining the structure above we will see below how the complete architecture looks like.

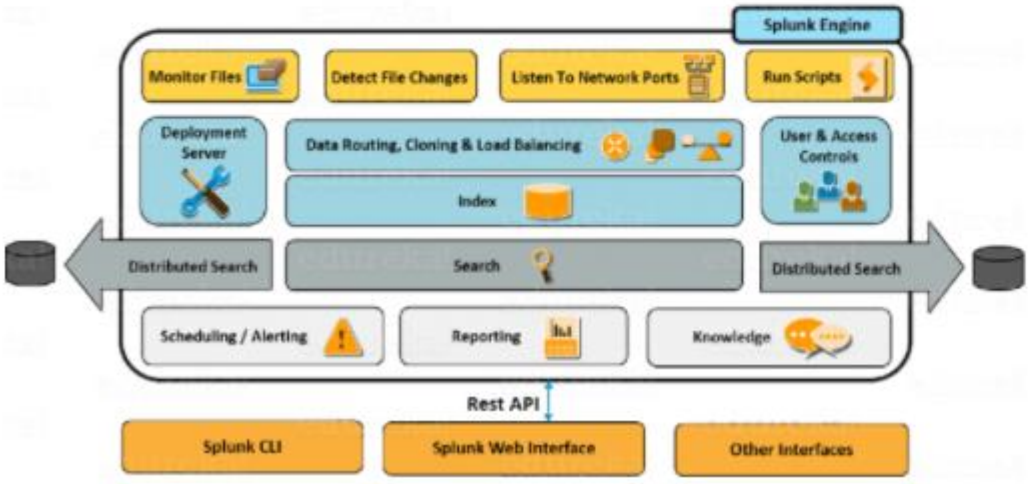


Figure13. Splunk Architecture

Data can be retrieved from various network ports by running scripts for automating data forwarding. It is possible to monitor the files and detect the real time changes in the data. As mentioned above, when data are received they are stored in the Index and from there each user is able schedule specific searches for data analysis or create alerts. Splunk uses a web interface for the end user to perform big data analysis and structured and unstructured data could be analyzed. As we already said, Splunk is not open-source software so each version according to the price paid can offer various features as well [43].

Of course there are many more tools for data analysis we selected indicatively three from a deep look into their features and architectures. There are many software and platforms that are provides free to the users, open source and are really powerful. Of course by selecting a commercial tool there is a higher option of customization.

Let's go on now whit the other two important steps for big data lifecycles and let us explore the tools for these steps.

4.3 Big data tools for data visualization

It is hard to think of a professional industry that does not benefit from making big data more understandable. Big Data visualization relies on powerful computer systems to ingest raw data and process it to generate graphical representations that allow humans to interpret and understand a high amount of data into seconds. Visualization is important and preferred to be used with big datasets since it offers the ability to comprehend to identify easy trends, patters and outliers even in the most complex data types and structures. Even the European Union has set the importance of big data visualization since more and more governments are using tools offered by varioys companies in order to analyze and help the decision making. Project where taken place not only in the Action Plan 2016-2020 but also to the Horizon 2020 Program. An example of using visualization techniques is project PoliVisu [44] where big data and visualization tools where used to tackle urban planning and mobility. Data visualization has a positive impact on the daily business and some of the most important benefits are the following. By visualizing the data, an overall view of the performance can be presented. Analysis takes place in seconds. It helps the involved parties to make data-driven decisions. It can analyze the customer behaviors and help identify trends and patterns. Furthermore by visualizing the data it is easier to share insights across organization and deliver clear messages to the audience.

We will analyze depth some of the data visualization tools and try to understand in detail their features, how they work and the differences.

Tableau

Tableau is a interactive data visualization software, which started as a desktop application but as the tool has developed it is being used even more and more in enterprises and for big data visualization. Tableau is being used by various types of organizations both in the private and the public sector. Some organizations that use Tableau are Federal Governments, Government intelligence organizations, multinational financial institutions and so on. The mission of Tableau is to help people see and understand their data. Tableau is a commercial tool which needs to be purchased. Below we will analyze some of the most distinct features of Tableau and will see how it works.

Features: Tableau is one of the leading software of the big data visualization market, enabling various types of graphs creation, charts, dashboards, stories, maps and other elements without programming. Furthermore it includes various not only descriptive but also inferential statistics with analytical charts generation. Tableau is possible to be integration with other tools as well, tools for data analysis such as Excel, SQL, SAP, Amazon, and others. By using Tableau it enables to collaborate with various users and share securely big data from various data sources either they are on premise or on –cloud. It offers connectivity to both live data sources but also data extraction from external data sources as in-memory data. Last, Tableau supports different kind of data connectors such as Presto, Google Analytics, Cloudera, Hadoop, Salesforce and so on which enables the user to retrieve data from everywhere.

Architecture: The architecture of Tableau is highly scalable which serves, mobile clients, web clients and desktop installed software. Let's take a deeper look. According to [45] the architecture of the server of Tableau is separated in various layers.

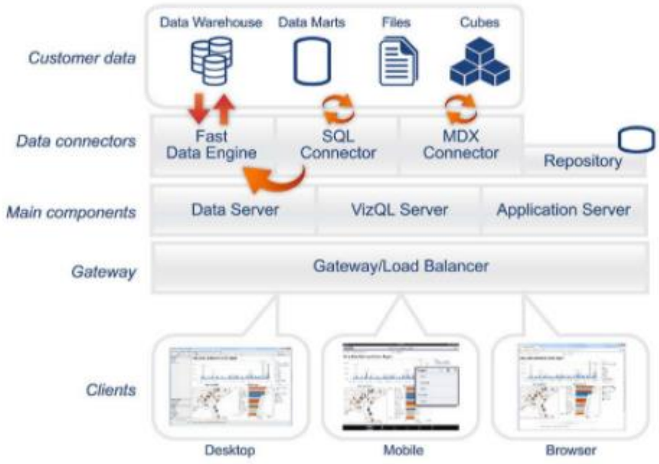


Figure14. Tableau Architecture

The first layer involves Data Layer, in the data layer each enterprise or customer can provide their data which can be in various formats. No specific architecture of format is required. Later on, The Data Connectors, ensure that data will be either live connected or in-memory connected according to the enterprise's needs. The three main Components of the architecture are: Data Server, where Tableau allows the customers to store the data sources and it also manages the metadata, VizQL Server, which sends queries directly to the data source returning the result that need to be presented and last but not least, Application Server, which handles all permissions that the user will receive for the software. Last through the Gateway the requests with the procedures are being shared and the clients can visualize and work on the interactive dashboards created.

QlikView

QlikView is a business intelligence tool which is used for converting raw data into knowledge. QlikView offers data visualization in a meaningful and innovative way. It is a commercial tool which can bring the whole business into control. QlikView is simple and easy software which supports the consolidation, analysis, search and visualization of various data sources. QlikView is a solution that has been used in the public sector and mostly into delivering federal government agency services effectively, transparently and efficiently. Qlik helps governments to deliver the citizens the transparency they demand. It enables to improve public health by exploring the healthcare services and to succeed cost efficiency. Last it has been used in order to improve that state and local government services and to save budget. An example where QlikView was used is in Swedish municipalities, since Qlik is also a Swedish company. Below we will analyze some of the features of QlikView.

Features: QlikView uses in-memory data model and it provides to the enterprises the ability to handle huge data sets instantly and accurately. It provides full filtering data, meaning it is possible to search across data both directly and indirectly. QlikView is a business discovery platform with fast and powerful visualization capabilities. Furthermore, it enables the automated data integration. The usage of dynamic applications such as dashboards and analyses from various forms of data presentation is possible as well. QlikView can easily convert data into graphical analytics and can be used to consistent reporting. It provides flexibility and ensures for the enterprises easy implementation and high scalability. It is considered a low cost –quick return on investment software since it does not need a long period to be implemented and running.

Architecture: The QlikView architecture according to [46] reflects the separation of roles. The front end is where end users interact with the data that they are allowed to view by using QlikView.

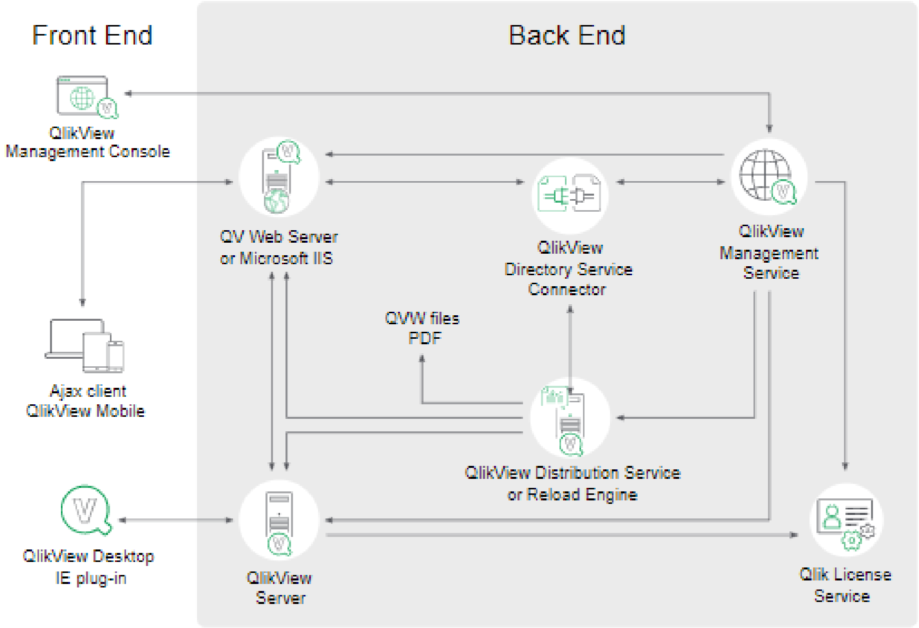


Figure15. QlikView Architecture

The back end is where all source documents are being stored for processing. These source files contain scripts to extract data from various data sources, for example from data warehouses, Microsoft Excel files, SAP, Salesforce and so on. The back end uses the infrastructure resources for clustering. What is important to mention is that the QlikView Server – Client communication requires three primary processes which must be able to communicate with each other in a consistent and secure manner. This interaction may involve various machines and multiple network connections. We will not go any further in detail at this stage.

Plotly

Let us explore now an open-source tool for data visualization. Plotly provides online graphing, analytics and statistics tools as well as scientific graphic libraries for Python, R and JavaScript. It is an on-line tool which can be used by everyone. Below we will describe some of the features and benefits of Plotly.

Feature: Plotly is a user friendly visualization software that offer highly advanced visualization tools. There is no need for any special training or knowledge for the end user to fully utilize the tool and its features. It gives also the ability to the user to fully customize functions since it includes an open development process. Furthermore, Plotly consists of high-powered tools for analysis and data scientists can freely work with languages familiar to Python or R. It provides an improved productivity, since it can speed up work and avoid bottlenecks and delays through the use of centralized dashboards. Another positive feature is that it enables team collaboration and file sharing. Going on, Plotly reduces drastically the costs on enterprises that use it as data tool since with its features it can replace a full team with developers and experts. Scalability is also crucial since Plotly can work not only with individual researchers but also startups and bigger enterprises. Last, it has an open Application Programming Interface that can be used to fully customize any user's experience. It can easily integrate third parta apps and work with existing workflow structures [47].

Architecture: The Ploty architecture is consisted of the following modules.

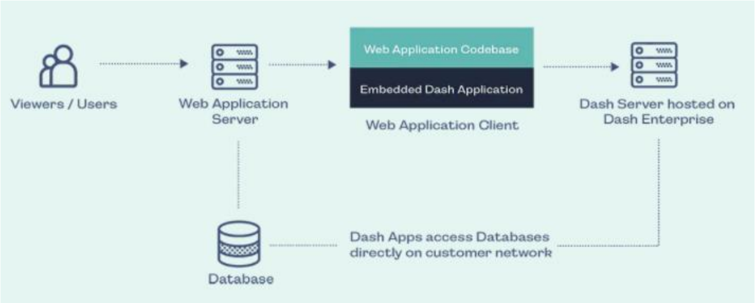

Viewers / Users
Web Application Server
Web Application Codebase
Embedded Dash Application
Web Application Client
Dash Server hosted on Dash Enterprise
Database
Dash Apps access Databases directly on customer network

Figure16. Plotly Architecture

Graph objects, that contains the objects or templates used to visualize. Plotly Express, which is a high level API of the Plotly and it is much easier to draw charts. The subplots, which contain the helper functions for layouts of the multi-plot figures and the Figure Factories, which provides many special types of figures such as annotated Heatmaps, Dendrograms, Gantt charts and so on. Last, I/O(input/output) is the module which involves the low level interface for displaying, reading and writing figures for static image, html and so on.

Below we will explore some of the big data tools for data storage.

4.4 Big data tools for data storage

After big data analysis, all data need to be stored in order to be ready for usage and also available for any re-use in the future. Big data storage is concerned with storing and managing data in a way that is scalable and would satisfy the needs of each application that require access to the data. The big data storage system should ideally allow storage of virtually unlimited amount of data and handle the data with flexibility and efficiency. The read and write in the data storage system should be possible and of course due to the nature of big data, various data modes and both structured and unstructured data should be supported [48]. It is obvious that these entire requirements are not fully satisfied, but over the years many new storage systems have tried to at least partly address these challenges.

It is visible on a cross-sector level the need to move towards data-driven economies and the necessity for data platforms are addressed even from the open data initiatives of the European Union. Technology vendors support the move to data governments and technologies as we can see from the continuously developing tools and technologies.

Some of the tools and technologies are reported below.

MongoDB

MongoDB is an open-source general purpose, document base, distributed database built for modern application developers and for the cloud era [49]. MongoDB is a distributed NoSQL data manager of a documental sort. This means that it is a non-relational database. In MongoDB, data is stored as documents. These documents are stored in JSON (JavaScript Object Notation) format. These documents support embedded fields, so related data and lists of data can be stored with the document instead of an external table. MongoDB as data manager system has the following characteristics [50]

Features: MongoDB as a database has flexible storage. This happens since it is sustained by JSON and does not need to define prior any schemes. Moreover, it provides the possibility to create multiple indexes on any attribute, which facilitates it is use since no MapReduce or any other process is needed. MongoDB has a high capability for growth, replication and scalability by just increasing the number other machines. Furthermore, it provides an independent file storage support, for any size, based on a storage specification which is implemented by all supported drivers. Finally, MongoDB is an application which is suitable for storing high volume of data since it has no complex joins and the structure of any object is clearly defined.

Architecture: MongoDB architecture consists of the following major components. Database, which is the ‘’physical’’ container for the data. Each database has its own set of files on the file system with multiple databases existing on a single MongoDB Server.

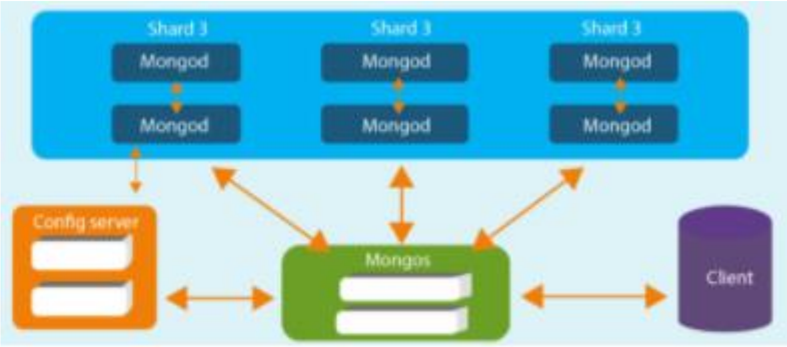


Figure17. MongoDB Architecture

The second component is Collection. With collection in MongoDB a group of database documents is being meant. The entire collection, for example a table exists within a single database. There are no schemas and various documents can have various fields. The only common thing that the documents in a collection have, is that they serve the same purpose for delivering the same end result. The last component of MongoDB architecture is the Document itself. As document, a set of key-values is being described. Each document is associated with dynamic schemas. The benefit of having dynamic schemas is that there is no need for all the documents to have the same structure or fields. Also various types of data are also supported. That is the reason why MongoDB database is also suitable for big data since those data are unstructured, come from various sources and have various types.

Apache Cassandra

Apache Cassandra is another free and open-source distributed database. It is a NoSQL database management system that is designed to handle large amount of data across many severs offering high availability and consistency. Cassandra is preferred for big data storage. Generally, NoSQL tools such as Cassandra but also other tools that we studied above have the characteristic that are no relational, are open source, cluster friendly, schemaless and work on the twenty-first-century web [51]. NoSQL tools are mainly used for large scale data and easy development, Cassandra as a storage system implements “no single points of failure”, by using redundant nodes and data. The main objective of Cassandra is to overcome the challenges of global scale applications and to offer a new design of database model as the systems already in place are struggling to keep up with the new requirements. The aim is to offer full multi-master database replication, global data availability of low latency, flexible schema, online load balancing and cluster growth and more.

Features: Some of the most important features that Cassandra has to offer as a database are the following. Cassandra is highly scalable; as a result, it enables the addition of more hardware which consequently enables attachment of more customers and more data as per requirement. As already mentioned, Cassandra has not a single point of failure and it is constantly and consistently available for business-critical applications that no failure is allowed. Furthermore, Cassandra is linearly scalable. It maintains always a quick response time since one can easily increase the number of nodes in the cluster. Cassandra is fault tolerant since each node in the database has a copy of the same data. So even if one node can no longer serve the other could still serve the requirement. What should also be mentioned is that Cassandra enables flexible data storage. It supports all kind of data formats, structured, semi-structured and unstructured and gives you the possibility to transform the data structures according to the user's need. Continuing, this database offers easy data distribution, since the replication of data via various data centers is possible. Last, Cassandra was designed to run on commodity hardware, it performs very fast writes and it can store high amount of data without decreasing the read efficiency.

Architecture: Unlike legacy systems, based on master-slave architectures, Cassandra implements ring-type architecture where its nodes are logically distributed like a ring. Data are automatically distributed across all the nodes and are replicated for redundancy. All data is kept in memory. Hash values of the keys are used to distribute the data among nodes in the cluster. As hash value is defined the number that maps a given key to a numeric value. Cassandra architecture supports multiple data centers where the data could be replicated as well. Copies of data could be kept in both data centers for remote backup for example. The benefits of this kind of architecture are that it enables transparent distribution of data to nodes. The location of the data can be determined in the cluster based on the data. Furthermore, any node can accept any request; as a result all requests are being fulfilled with the higher consistency.

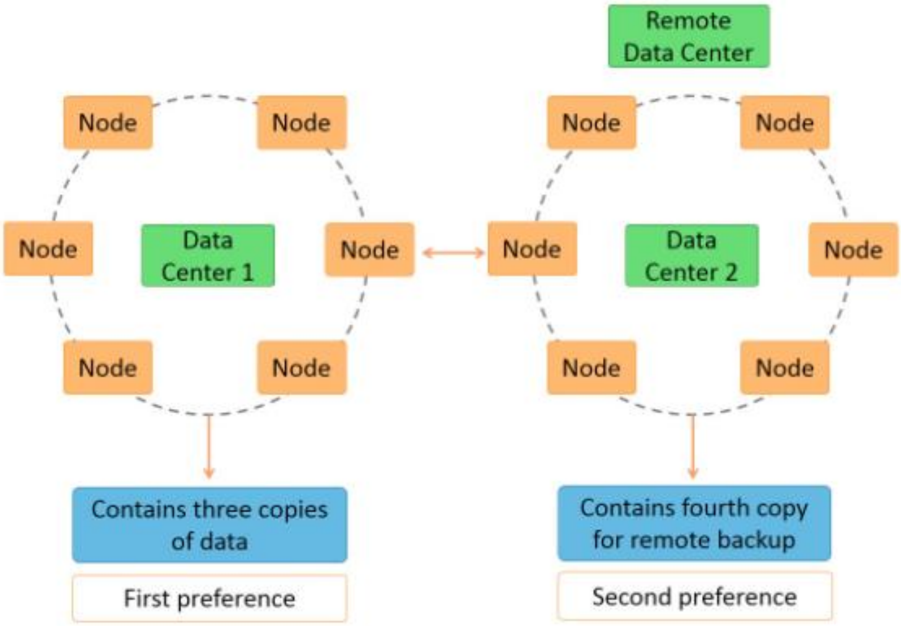


Figure18. Apache Cassandra Architecture

Another characteristic of Cassandra's architecture is that it supports network topology with multiple data centers, multiple racks and nodes. The read and write processes ensure fast read and fast write of data. Inter-node communication is possible and node, disk, rack or data-node failures can be handled.

Summarizing, Cassandra is brilliant choice as big data storage system, because it can easily deal with velocity, variety and complexity issues and can handle massive datasets. It provides a homogeneous environment regardless of how the data look like or in which format. And last is gives the opportunity to the users to customize and set up the environment according to their needs.

Cloudera

Cloudera's Enterprise Data Hub is a big data platform powered by Apache Hadoop at the core. Cloudera provides a scalable, flexible and secure environment for handling big data interactively and at real-time. Cloudera is a US based software company for data warehousing, data engineering and machine learning. Cloudera is widely recognized as the leading Apache Hadoop software and service provider in the big data landscape [52].

Features: Cloudera offers the best data platform constructed on Hadoop. It has designed data platforms which are fast, easy and secure by offering the possibility to solve even the most complex business issues and challenges when it comes to data. By using Cloudera it is feasible for the enterprises or the users to build an enterprise data hub and leverage the power of data by exposing its hidden values. Moreover, the users could add security governance and management functions that are required to create an enterprise-grade foundation for data. Furthermore, Cloudera as a data warehouse, for handling data storage has high performance in both on-premises deployments and as a cloud service. It provides an open architecture, where even if the data are stored it can be accesses by many users and more tools this provides more value with lower cost. Cloudera pricing is available on a by quote basis, it is not open source and has various versions according to the users' needs.

Architecture: Let us see in detail some of the characteristics of the architecture and how Cloudera works. Cloudera Manager Architecture consists of a Server, Management Service, several relational databases, Manager Agents and Filesystem-based runtime state storage.

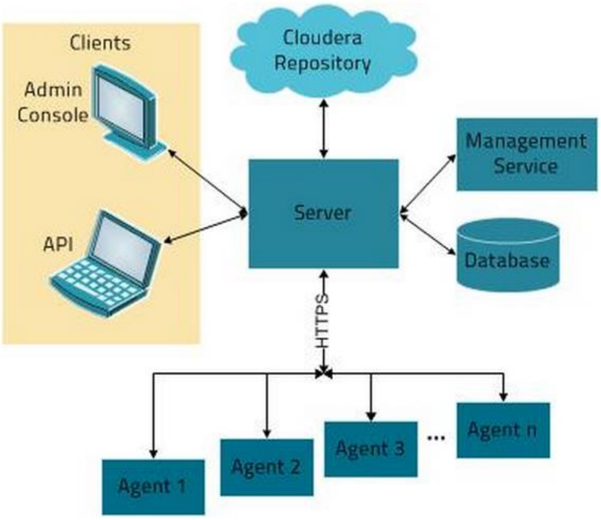


Figure19. Cloudera Architecture

The Server and some of the Service roles use a relational database to store their operational data. Some of other services such as host monitor use the filesystem to store their data. The Cloudera Manager Agent software includes an agent and a supervisor process. The agent handles the communication with the Manager and with the roles of Cloudera Management Service. The supervisor process handles the local

Cloudera-deployed process lifecycle and handles failed processes [53]. What Cloudera has to offer as database for data storage is real time data management; it collects stores and organizes massive unstructured data in real time.

Generally all the tools that we described above can be used for almost all the stages of the big data lifecycle. These are tools that have many various features but we chose to classify them into this categorization according to their strengths. On the next section we will try to make a comparison matrix for all these tools taking into consideration some of their attributes.

4.5 Taxonomy and comparison of the tools

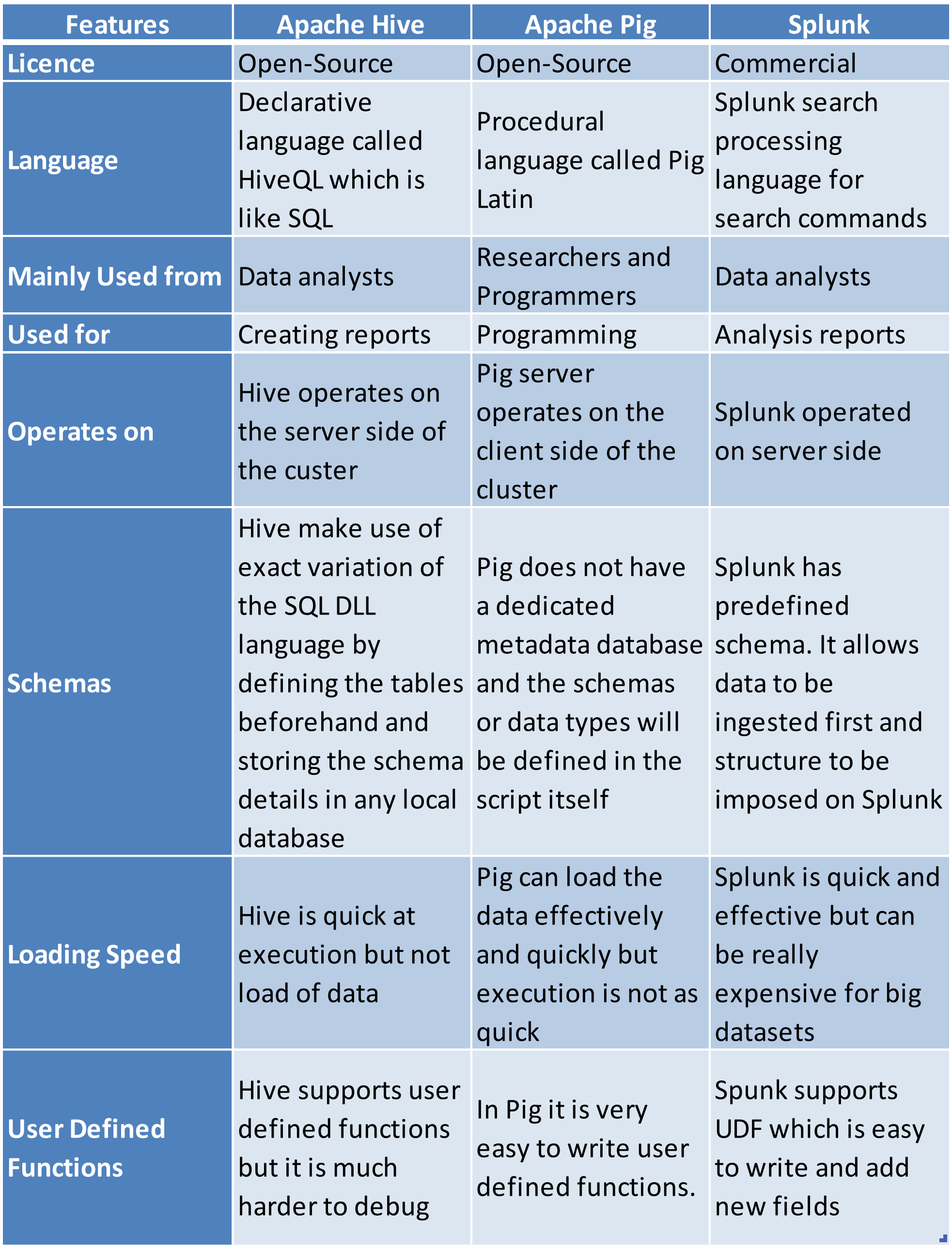
4.5.1 Comparison of analysis tools

Below is a comparison table for the paradigms/tools we mentioned above according to their key features and the facilities they support. In order to compare the tools we took into consideration some of the features that are the most common and most easy to be explained. The limitations of Apache Hive are that it does not offer real-time queries and low level update also as seen in the table below, the load of data can have high latency. Furthermore update and delete operations are not supported in Hive and it is not designed for online transitional process.

Focusing on Pig, we see that although it is easy to write user defined functions, the errors that are produced are not easy identified and not easy to be corrected. Furthermore, Pig is still in the development, even if it has been around for some time. Another huge disadvantage of Pig is that as it does not need a specific schema, sometimes the data structure can be transformed to "raw" data type which leads to propagation for other steps of the data processing. Finally, in Pig may occur latency at execution since commands other that dump and store a result are not being executed? This increases the time of debugging and resolving the issue.

Continuing with our last tool which is Splunk, and is the only commercial tool, we see that the first disadvantage is that it can be really expensive, especially for large data volumes. At the analysis sometimes, although it offers a real-time architecture and endless capabilities, it is difficult to understand the searches. Specifically, regular expressions and search syntax may be tricky. The various dashboards provided by

Splunk are functional but not as effective as expected. Last, it's a tool with a complex architecture, therefore training and lots of time to learn this tool is required.

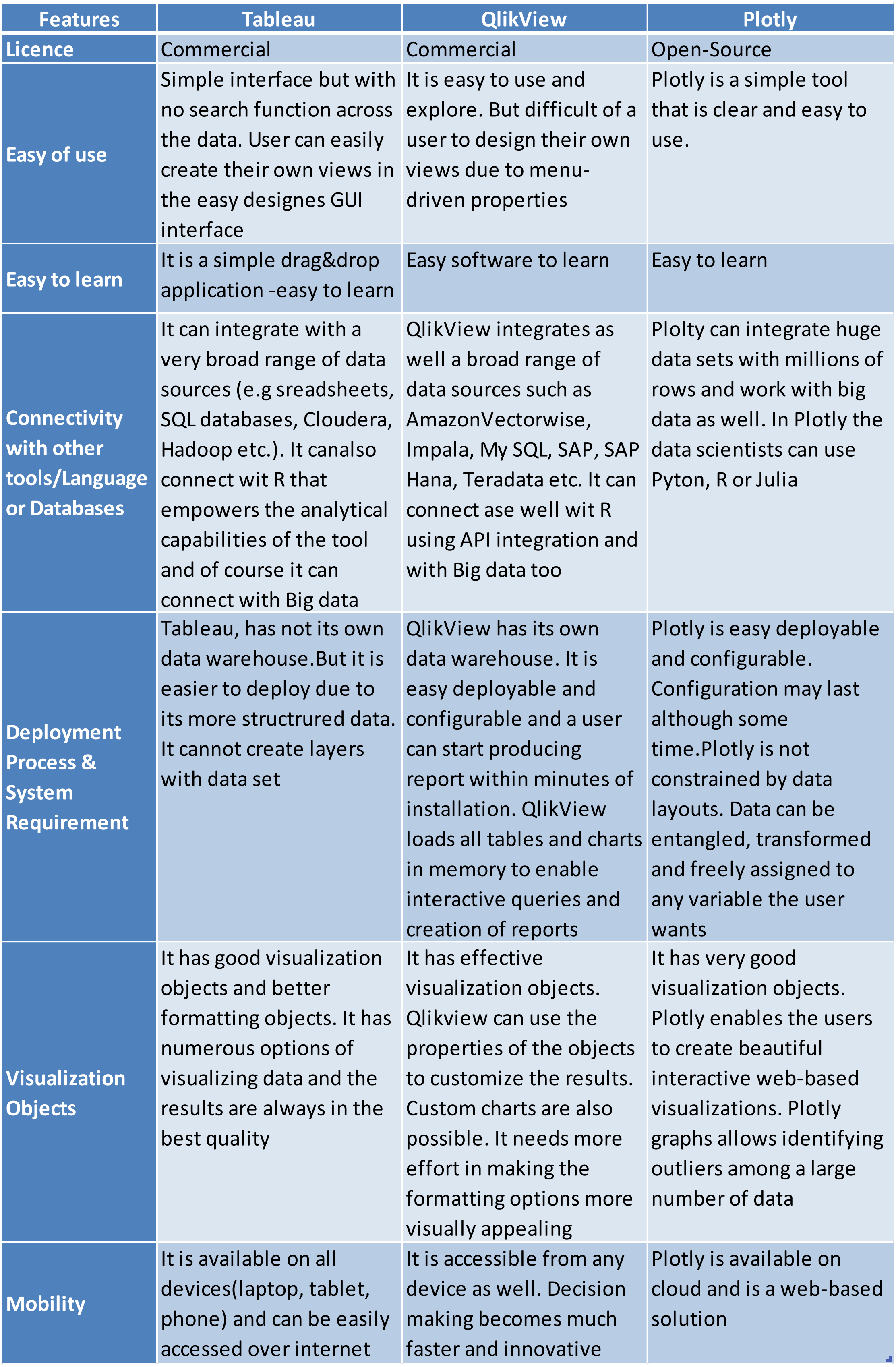


Features	Apache Hive	Apache Pig	Splunk
Licence	Open-Source	Open-Source	Commercial
Language	Declarative language called HiveQL which is like SQL	Procedural language called Pig Latin	Splunk search processing language for search commands
Mainly Used from	Data analysts	Researchers and Programmers	Data analysts
Used for	Creating reports	Programming	Analysis reports
Operates on	Hive operates on the server side of the custer	Pig server operates on the client side of the cluster	Splunk operated on server side
Schemas	Hive make use of exact variation of the SQL DLL language by defining the tables beforehand and storing the schema details in any local database	Pig does not have a dedicated metadata database and the schemas or data types will be defined in the script itself	Splunk has predefined schema. It allows data to be ingested first and structure to be imposed on Splunk
Loading Speed	Hive is quick at execution but not load of data	Pig can load the data effectively and quickly but execution is not as quick	Splunk is quick and effective but can be really expensive for big datasets
User Defined Functions	Hive supports user defined functions but it is much harder to debug	In Pig it is very easy to write user defined functions.	Spunk supports UDF which is easy to write and add new fields

Figure20. Big Data Analysis Tools _ Comparison

4.5.2 Comparison of the data visualization tools

Moving on to the next set of tools that we analyzed and looked in depth we will compare them as well taking into consideration the most critical features and the most easy to be interpreted by the end users. Below we have the comparison table.



Features	Tableau	QlikView	Plotly
Licence	Commercial	Commercial	Open-Source
Easy of use	Simple interface but with no search function across the data. User can easily create their own views in the easy designes GUI interface	It is easy to use and explore. But difficult of a user to design their own views due to menu-driven properties	Plotly is a simple tool that is clear and easy to use.
Easy to learn	It is a simple drag&drop application -easy to learn	Easy software to learn	Easy to learn
Connectivity with other tools/Language or Databases	It can integrate with a very broad range of data sources (e.g sreadsheets, SQL databases, Cloudera, Hadoop etc.). It canalso connect wit R that empowers the analytical capabilities of the tool and of course it can connect with Big data	QlikView integrates as well a broad range of data sources such as AmazonVectorwise, Impala, My SQL, SAP, SAP Hana, Teradata etc. It can connect ase well wit R using API integration and with Big data too	Plolty can integrate huge data sets with millions of rows and work with big data as well. In Plotly the data scientists can use Pyton, R or Julia
Deployment Process & System Requirement	Tableau, has not its own data warehouse.But it is easier to deploy due to its more structrured data. It cannot create layers with data set	QlikView has its own data warehouse. It is easy deployable and configurable and a user can start producing report within minutes of installation. QlikView loads all tables and charts in memory to enable interactive queries and creation of reports	Plotly is easy deployable and configurable. Configuration may last although some time.Plotly is not constrained by data layouts. Data can be entangled, transformed and freely assigned to any variable the user wants
Visualization Objects	It has good visualization objects and better formatting objects. It has numerous options of visualizing data and the results are always in the best quality	It has effective visualization objects. Qlikview can use the properties of the objects to customize the results. Custom charts are also possible. It needs more effort in making the formatting options more visually appealing	It has very good visualization objects. Plotly enables the users to create beautiful interactive web-based visualizations. Plotly graphs allows identifying outliers among a large number of data
Mobility	It is available on all devices(laptop, tablet, phone) and can be easily accessed over internet	It is accessible from any device as well. Decision making becomes much faster and innovative	Plotly is available on cloud and is a web-based solution

Figure21. Big Data Visualization Tools _ Comparison

Diving into their characteristics, we see that in regards to Tableau, although it is software with remarkable visualization capabilities and really high performance, it can be a really expensive tool especially for small to medium companies. Furthermore it lacks functionality required for a full business intelligence tool, such as large-scale reporting and static layouts. Also, it has only limited capacity for result sharing. Last but not least, Tableau is strictly a visualization tool. That means that Tableau allows you to do very basic preprocessing. Usage of another tool such as Power BI or even Excel is required in order to preprocess the data prior to loading.

Going on with ClikView, although it is a tool that has many benefits to offer, such as data sharing, low maintenance, it is a self-service BI tool and is really quick in data delivery it has its own disadvantages as well. QlikView proves to be inefficient at time in real-time data analysis. Furthermore although it can load heavy data, the computer's RAM sets a limit to it. Another limitation of QlikView is its difficult application development, since it demands technical expertise and good knowledge of SQL. Last, it can become quite expensive for small and medium enterprises, since it requires a lot of extra purchases. Concluding, QlikView software although its disadvantages it manages to maintain a spot in the lot as one of the best data discovery and visualization tool.

Continuing with Plotly, although it is platform for users that just want to make simple visualizations and do not know much or any code, it has its limitations as well. It can be really challenging to visualize complex data or make complex, multi-field charts using the Plotly's online tool. Therefore in order to visualize more complex data, knowledge of Python is required. Furthermore, by using Plotly sometimes it is challenging to connect graphs to the same underlying dataset; this may cause some limitations on data handling. Last, although Plotly is an open-source tool, it also has a version that can be purchased. We see that in the free version it lacks a little bit in functionality.

As we saw above all tools have their advantages and disadvantages, each company or data scientist can choose their preferable tool according their need. Most of the bigger companies usually choose more than one tool, so that full coverage of data analysis and visualization is covered.

4.5.3 Comparison of the data storage tools

Continuing we will perform the comparison of the third set of tools that we reported above, the one that are mostly used for data storage. Once again we will perform a comparison to those features that are most critical for these specific tools.

Comparing in depth the tools, we recognize that MongoDB has many aspects that are entirely favorable but there are still areas that the database does not perform as well as other databases. Some of the disadvantages of MongoDB are the lack of support on transactions. There are some applications that need transactions in order to update multiple documents, in MongoDB there is a potential for data corruption. Furthermore, some of the downsides of the database could lead to duplicate data sets which are difficult to handle and could lead as well to corrupted data. Last, enjoying MongoDB's quick speeds and high performance is only available with the right indexes. With out of order composite indexes the database will operate really slowly.

Continuing with Cassandra and its limitations, what needs to be mentioned is that Cassandra does not provide relational data properties. Furthermore it does not support aggregates. Making excessive requests and reading more data slows down that actual transaction, resulting in latency issues that are also a disadvantage for the database. Finally, to store huge amount of data in Cassandra, JVM (java virtual machine) is required to manage memory which itself is a language so the automatic memory management is not done by the application but by a language in Cassandra.

Finally, focusing on Cloudera, as main disadvantage we see that it is not fully an open-source tool. Some of components are privately owned and need to be purchased. Furthermore the custom features of open source software tools supported only be Cloudera can be really challenging. Concluding, there are no hard limits in regards to the data storage, bur gradual performance degradation will occur by increasing components such as the number of databases, the number of tables, files and users etc.

Features	MongoDB	Cassandra	Cloudera
Licence	Open-Source	Open-Source	Commercial
Primary database model	Object and document-riented store of data	Wide column store	is a multi-model and supports key value, wide column and relational or users can provide their own model
Data scheme	Schema-free, the database can input documents of different structures	Schema-free, but it is a little bit more stationary database not as flexible	Schema-free. There is no need to pre-define a schema.
Data availability	During the a failover time in case of recovery, the database is not responding to requests, this happens because MongoDB has a sigle master directing multiple slave nodes	Cassandra utilizes multiple master inside a cluster. This ensures the high availability of data all the time	Cloudera support high availability implicitly because it comprises distributed processes.
Scalability	In MongoDB only that master node can write and accept input. Since it has only one master node, this database is limited in terms of writing scalability.	Having multiple master nodes, is increases Cassandra's capabilities. This allows the database to coordinate numerous writes at the same time with better speed and higher scalability	Cloudera database provides unparalleled scale and flexibilty for applications. This enables to bring together and process data of all types and from multiple sources
Supported Progamming Languages	MonogDB supports a wide range of programming languages such as C, C++,PHP, Java, Python, R etc.	Cassandra supports as well a wide range of languages but significantly less than MongoDB. Some of them are Perl, PHP, Ruby and Scala	Cloudera supports various programming languages such as Apache Groovy, C, C++, Java, Python etc.
Aggregation	MongoDB has a built-in aggregation framework. This allows the database to retrieve data by utilising an ELT(extract, load and transform) pipeline to transform the documents into aggregated results	Cassandra has no aggregation framework and requires external tool like Hadoop, Spark and others	Cloudera supports aggregatiion by having the YARN Log Aggragation features.

Figure22. Big Data Storage Tools _ Comparison

4.6 Synopsis

We recognize that the volume of the various big data tools for all the data lifecycle stages is very high. The tools cover a wide range of the enterprises and data driven government's needs. In order to specify which tool is appropriate for which case some factors need to be considered of course. Since big data is one of the most valuable resources in an organization it is crucial to select the tools wisely. The first big decision that needs to be made is whether the hosting of the big data tool will take place in the organizations' own data center or if a cloud based solution will be used. Cloud-based big data applications are popular for several reasons such as scalability and ease of management. However cloud is not always the best option, since organizations with high compliance and security requirements such as governments need to keep the sensitive data on premise. Going on, it needs to be considered whether a commercial or an open-source tool will be selected. However the open source solutions are plenty enterprises sometimes find it difficult to get them up and running and configured for their need. Last, it needs to be decided if batch or streaming big data applications will be used. The earliest big data solutions such as Hadoop process batch data only, but organizations increasingly prefer the real-time data analysis. In some cases organizations select also data processing architectures that can handle both real-time and batch data. Clearly before choosing the right application the total overview of the enterprise/organization needs to be clear and understood. Knowing what the enterprise wants to accomplish is key for the selections, meaning the goal needs to be clarified as much as possible. Furthermore, it is always a good way to start with a small project as pilot before moving on to implementing the tool for the whole company. It is important though, that even for the smaller project, a holistic approach will be followed. This way it will be easier to extend the application and the end-to-end solution created by the architects for the full company. Last, the culture of the organization needs to become data-driven so that it will support the change and will integrate the tools for analysis.

5. Discussion and Conclusion

From the whole report and the documents that we reviewed, we conclude that big data paly a really important role in the technological evolution and specifically in the decision making processes of any data-driven government and company. As we saw in our research big data have attracted huge attention of academic and researchers. We used for our article 53 sources that helped us perform the taxonomy of the big data tools. To utilize at the most these sources we kept the search not too specific but in a range that it would return the wanted results. As discussed in the literature review chapter we selected our sources by some criteria such as citations and if they are academic journals and from specific databases. From these big data papers we extracted the data to provide an overview on current frames on big data technologies and tools. In addition we provided an overview of the technologies for big data platforms such as big data tools for analysis and storage and saw big data analytics techniques. Last we saw in details some of the top tools for big data processing going into depth on their architecture, the way of working their advantages and disadvantages. It needs to be mentioned that we came across many challenges. There were not many reports and documentation regarding the tools and techniques that data-driven governments use. It was nearly impossible to find another report with taxonomy of tools and even though there are millions articles for big data, in the European Union websites it was hard to find a concentrated report on this topic. Let us see in detail what we understood for big data from our research and experience. Firstly, the main function of big data systems and tools is to manage various data types coming from various sources. For example, a smart city system manages data from mobile phones, sensors and many other devices so that public safety, effective traffic control and general security can be managed and monitored. Health and medical care systems obtain sensors data from patients in offering them novel therapies. Road traffic monitoring systems maintain data from surveillance cameras in order to detect road accidents. Governments gather the citizens' data from their electronic Identification Cards in order to provide them flexibility in state matters and ensure their integration in case of moving inside the European Union for example. As we mentioned several times in our report big data are unstructured and can be found in different formats. In order to use them either the data need to be structured before entering the database or tools and technologies can be used such as Hadoop, Spark that support also unstructured data. All the tools that we analyzed support big data processing and are applicable on big data applications.

Secondly we saw that in developing the characteristics of new hardware and tools the traditional algorithms and data processing techniques were enhanced. Companies that developed the new tools were trying to overcome all the challenges and offer a better solution for big data.

Thirdly, of course big data is still a chaotic world. To many end users and governments the capabilities of the big data tools are still not known. The easiest way to fill this gap is by using visualization tools that show to the end user what big data contains and how they could help them to spot bottlenecks and anomalies into a process.

Lastly, from our literature research we spotted that big data applications are still limited in crucial areas such as e-commerce and finance. Since big data are still in the phase of development the tools and techniques existing are not completely solve real big data problems. However, the effort to develop and invent solutions is clear and visible. While the needs are increasing the development of big data tolls will increase rapidly as well. The more the tools are the more the needs will increase. These are relationally connected and therefore the speed of development is really high. We recognize that the growing modernized public areas require the development of efficient applications and infrastructures. The increase of data analytics capabilities, mobile cloud computing and data mining technologies are able to increase subsequently the research works and applications in order to support the areas. Furthermore, the growth of the number of devices connected to the IoT (internet of things) and the increase of data consumption reflect the connection between big data and IoT. The union of IoT and big data has created several opportunities for developing more applications in order to decrease the problems that exist.

It is expected that the next years more and more governments and companies’ even users will start using big data technology, as they will see more benefit from the big data application even at the lowest level of the daily life. Therefore more and bigger investments are coming from both government and private sector in order to benefit from the advantages of big data. Big data applications are important, since as we analyzed in the previous sections, superior storage techniques are being used, more beneficial computer architectures are being created, more efficient data-intensive techniques are being developed such as cloud computing, social computing etc. It is really an added value the evolution of big data for both parties' public and private, the management of human resources and capital investment but also the innovative and creative developments of big data application only support and help the daily life.

Recommendations

The European countries and all the data-driven companies should continue focusing on big data applications and support those kinds of initiatives since the decision making will be more efficient and effective. All the countries move on through the digital transformation era and therefore projects on big data applications will only bring benefits. This way the data-driven governments will be able to deliver to their citizens easier, faster and in a more economical way by having everything digitalized. Apart from that, we would recommend a creation of one single source of information for all the available big data tools, techniques and initiatives which were developed and are used from the data-driven governments. This will enable the interested parties not only public but also private to be aware of what is accessible and already available for usage. As we saw from our research all the tools are many and it is impossible to cover them all so a next research that would make sense would be to try to gather all the tools ever developed for big data for the various data lifecycle stages. We see that some tools are overlapping other and some are really providing a different characteristic.

Closing, the world of big data is really interesting and the insights that the interested party will get are valuable. The techniques and tools are so many and different and it is one of the best discovery of the new digital are.

Bibliography

[1] José María Cavanillas, Edward Curry, Wolfgang Wahlster, "New Horizons for a Data-Driven Economy," in *ARoadmap for Usage and Exploitation of Big Data in Europe*, Luxemburg, Springer, 2015, pp. 29-38.

[2] D. Laney, "3D data management: Controlling data volume, velocity, and variety.," META Group., 2001.

[3] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H., "Big data: The next frontier for innovation, competition, and productivity," in *McKinsey Global Institute*, 2011, p. 156.

[4] M. Loukides, "What is data science," O'Reily Radar, 2010.

[5] A. Jacobs, "The pathologies of big data.," in *Communications of the ACM*, 2009, pp. 36-44.

[6] IDC, "IDC's worldwide big data taxonomy.," 2011.

[7] Wikipedia, "Big data," Wikipedia article. http://en.wikipedia.org/wiki/Big_data.

[8] M. 2.0., "Big data definition – Mike 2.0.," 2014.

[9] NESSI, "Big data: A new world of opportunities. NESSI White Paper.," 2012.

[10] O'Ria´in, S., Curry, E., & Harth, A., "XBRL and open data for global financial ecosystems: A linked data approach," *International Journal of Accounting Information Systems,* pp. 141-162, 2012.

[11] Abhinandan Banik, Samir Kumar Bandyopadhyay, "Big Data- A Review on Analysing 3Vs," *Journal of Scientific and Engineering Research,* 2016.

[12] Gang-Hoon Kim, Silv ana Trimi, Ji-Hyong Chung, "Big-Data Applications in the Government Sectror," Communications of the ACM (Association for computing machinery), United States, 2014.

[13] M. EL ARASS , I. TIKITO , SOUISSI N., "Data Lifecycle analysis: towards intelligent cycle," IEEE, Morocco, 2017.

[14] J. B. J. Reynolds, "In the context of the Convention on Bilogical Diversity," in *World Conservation Monitoring Centre*, 1996.

[15] J. Rowley, "The wisdom hierarchy: representations of the DIKW hierarchy," in *The wisdom hierarchy: representations of the DIKW hierarchy*, U.K, Journal of Information Science,, 2006, pp. 164-180.

[16] S. Allard, "DataONE: Facilitating eScience through Collaboration," *Jurnal of eScience Librarianship.*

[17] L. Lin, T. Liu, J. Hu, and J. Zhang,, "A privacy-aware cloud service selection method toward data life-cycle," in *20th IEEE International Conference on*, 2014.

[18] X. Ma, P. Fox, E. Rozell, P. West, and S. Zednik, "Ontology dynamics in a data lifecycle: challenges and recommendations from a geoscience perspective," 2014, pp. 407-412.

[19] A. Gregory, "The Data Documentation Initiative (DDI): An introduction for Nationsl Statistical Institutes," Open Data Foundation, 2011.

[20] Y. Demchenko, "Defining Architecture Components of the Big Data Ecosystem," in *2nd BDDAC2014 Symposium, CTS2014 Conference*, Mineapolis, USA, 2014.

[21] Nawsher Khan,1,2 Ibrar Yaqoob,1 Ibrahim Abaker Targio Hashem,1, "Big Data: Survey, Technologies, Opportunities, and Challenges," *The Scientific World Journal,* p. 18, 2014.

[22] Thomas Erl, Paul Buhler, Wajid Khattak, "Big Data Analytics Lifecycle," in *Big Data Fundamentals: Concepts, Drivers & Techniques*, The Pearson education service technology series from Thomas Erl , 2016, p. 11.

[23] C.L. Philip Chen ⥣, Chun-Yang Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," 2014, p. 34.

[24] James P. Ahrens, Bruce Hendrickson, Gabrielle Long, Steve Miller, Robert Ross, Dean Williams, "Data-intensive science in the us doe: case studies and," *Comput. Sci. Eng. 13,* pp. 14-24, 2011.

[25] Muhammad Sahimi, Hossein Hamzehpour, "Efficient Computational Strategies for Solving Global Optimization Problems," *Computing in Science & Engineering , IEEE,* pp. 74-83, 2010.

[26] Gayathri Seenumani, Jing Sun, Huei Peng, "Real-time power management of integrated power systems in all electric ships leveraging multi time scale property," *IEEE Trans. Syst. Technol.,* vol. 20, no. 1, pp. 232-240, 2012.

[27] Zhibin Guan∗, Tongkai Ji∗‡, Xu Qian∗, Yan Ma∗ and Xuehai Hong†, "A Survey on Big Data Pre-Processing," IEEE, 2017.

[28] Kamlesh Kumar Pandey, Diwakar shukla2 , "Challenges of Big Data to Big Data Mining with their Processing Framework," IEEE, 2018.

[29] N. Tyagi, "Analytics Steps," 11 May 2020. [Online]. Available: https://www.analyticssteps.com/blogs/top-10-big-data-technologies-2020.

[30] *Data lakes: The bedrock of Big Data Processing,* https://blog.unbelievable-machine.com/en/data-lakes-the-bedrock-of-big-data-processing.

[31] *Blockchain and Big Data,* https://itsvit.com/blog/blockchain-big-data-match-made-heavens/.

[32] Meenakshi Saroha, Aditya Sharma, "Big Data and Hadoop Ecosystem: A Review," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT),* India, 2019.

[33] A. Gupta, "Big Data Analysis Using Computational Intelligence and Hadoop: A Study," IEEE.

[34] S. Early, "Big Data and Predictive Analytics: What's New?," *IT Professional,* vol. 16, no. 1, pp. 13-15, 2014.

[35] A. Steane, "Quantum computing," UK, 1998.

[36] Borko Furht, Armando Escalante,, Handbook of Cloud Computing, Springer, 2011.

[37] S. Mazumder, "Big Data Tools and Platforms," in *Big Data Concepts, Theories, and Applications,* Springer, 2016, pp. 29-128.

[38] Monique Hennink, Inge Hutter, Ajay Bailey, Qualitative Research Methods,

SAGE, 1st Edition 2010 , 2nd Edition 2020.

[39] G. Bowen, "Document Analysis as a Qualitative Research Method," *Qualitative Research Journal,* p. 15, 2009.

[40] Azlinah Mohamed;Maryam Khanian Najafabadi ;Yap Bee Wah;Ezzatul Akmal Kamaru Zaman;Ruhaila Maskat, "The state of the art and taxonomy of big data analytics: view from new big data framework," Springer, 2019.

[41] Aditya Bhardwaj, Vanraj,Ankit Kumar,Yogendra Narayan, Pawan Kumar, "Big Data Emerging Technologies: A CaseStudy with Analyzing Twitter Data using Apache Hive," IEEE, 2015.

[42] Yin Huai,Ashutosh Chauhan,Alan Gates, Gunther Hagleitner,Eric N. Hanson,Owen O'Malley,Jitendra Pandey,Yuan Yuan,Rubao Lee,Xiaodong Zhang, "Major Technical Advancements in Apache Hive".

[43] Murali Maghanagopal, Merritte Stidston, Shibani Singh, Sandeep Togrikar, Francisco Casas, Jason Beyer, Brian Wooden, "High-Performance data analytics with Splunk on Intel Hardware".

[44] [Online]. Available: http://www.polivisu.eu/vision.

[45] [Online]. Available: https://intellipaat.com/blog/tutorial/tableau-tutorial/tableau-architecture/#:~:text=Tableau%20has%20a%20highly%20scalable,shared%20views%20on%20Tableau%20Server..

[46] [Online]. Available: https://help.qlik.com/en-US/qlikview/April2020/Subsystems/Server/Content/QV_Server/QlikView-Server/QVSRM_FunctionalArchitechture.htm.

[47] [Online]. Available: https://comparecamp.com/plotly-review-pricing-pros-cons-features/#:~:text=The%20main%20benefits%20of%20Plotly,%2C%20scalability%2C%20and%20total%20customization.&text=Plotly%20is%20a%20user%2Dfriendly,utilize%20all%20tools%20and%20features..

[48] Martin Strohbach,Jörg Daubert,Herman Ravkin,Mario Lischka, "Big Data Storage," in *New Horizons for a data-driven Economy,* Springer, pp. 119-

141.

[49] [Online]. Available: https://www.mongodb.com/.

[50] Paula Catalina Jaraba Navas,esid Camilo Guacaneme Parra,José Ignacio Rodríguez Molano, "Big Data Tools: Haddop, MongoDB and Weka," in *4 Tools for Data Analytics*, Springer.

[51] Raul Estrada, Isaac Ruiz, "Storage:Apache Cassandra," in *Big Data Smack*, Mexico, Springer, 2016.

[52] Aisling O'Driscoll,Jurate Daugelaite, Roy D.Sleator, "Big data Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics.*

[53] [Online]. Available: https://docs.cloudera.com/documentation/enterprise/latest/topics/admin_cm_ha_deploy_arch.html.