# Big Data Analytics: A Literature Review Perspective

## Sarah Al-Shiakhli

**LULEÅ**
**UNIVERSITY**
**OF TECHNOLOGY**

# Abstract

Big data is currently a buzzword in both academia and industry, with the term being used to describe a broad domain of concepts, ranging from extracting data from outside sources, storing and managing it, to processing such data with analytical techniques and tools.

This thesis work thus aims to provide a review of current big data analytics concepts in an attempt to highlight big data analytics' importance to decision making.

Due to the rapid increase in interest in big data and its importance to academia, industry, and society, solutions to handling data and extracting knowledge from datasets need to be developed and provided with some urgency to allow decision makers to gain valuable insights from the varied and rapidly changing data they now have access to. Many companies are using big data analytics to analyse the massive quantities of data they have, with the results influencing their decision making. Many studies have shown the benefits of using big data in various sectors, and in this thesis work, various big data analytical techniques and tools are discussed to allow analysis of the application of big data analytics in several different domains.

Sarah Al-Shiakhli

# Contents

# 1. Introduction

Big data refers to datasets which are both large in size and high in variety and velocity of data, characteristics which make it difficult for them to be handled using traditional techniques and tools (Constantiou, I.D. and Kallinikos, J., 2015). This has generated a need for research into and provision of solutions to handle and extract knowledge from such datasets. Due to the large quantities of data involved, multiple technologies and frameworks have been created in order to provide additional storage capacity and real-time analysis. Many models, programs, software, hardware, and technologies have thus been designed specifically for extracting knowledge from big data (Oussous et al., 2018), as the extensive but rapidly changing data from daily transactions, customer interactions, and social networks has the potential to provide decision makers with valuable insights (Provost and Fawcett, 2013; Elgendy and Elragal, 2014; Elgendy and Elragal, 2016).

Big data analytics have already been extensively researched in academia; however, some industrial advances and new technologies have mainly been discussed in industry papers thus far (Elgendy and Elragal, 2014; Elragal and Klischewski, 2017). The link between research in academia and industry may be best understood when summarised and reviewed critically, and as a literature review represents the foundation for any further research in information systems, it may be regarded either as a part of such research or as research itself. However, this requires more than a literature summary, as it must show the relationship between different publications and identify relationships between ideas and practice.

An effective literature review provides the reader with state-of-the-art reporting on a specific topic and also identifies any gaps in the current state of knowledge of that topic. Literature reviews have played a decisive role in scholarship, particularly where scientists are looking for the new knowledge created by explaining and combining existing knowledge processes. The literature search process used determines the quality of a literature review (Webster and Watson, 2002), and the literature review writing goal is to reconstruct available knowledge in a specific domain, offering access to subsequent literature analysis. The process should thus be described comprehensively, allowing the reader can assess the knowledge available within the relevant field in order to use the results in further research (Vom Brocke, J. et al., 2009).

This thesis aims to present a literature review of work on big data analytics, a pertinent contemporary topic which has been of importance since 2010 as one of the top technologies suggested to solve multiple academic, industrial, and societal problems. In addition, this work explains and analyses different analytic methods and tools that have been applied to big data.

Recently, the focus has been on big data in the research and industrial domains, which has been reflected in the sheer number of papers, conferences, and white papers discussing big data analytic tools, methods, and applications that have been published. In writing this literature review, the same procedure was followed as in most commonly used literature reviews in information systems, such as Vom Brocke et al. (2009). The papers were chosen based on both novelty and discussion of important topics related to big data and big data analytics in manners that serve the purpose of the research. The selected publications thus focus on big data analytics during the period 2011 to 2019. Most of the references were selected from prestigious journals or conferences, with a limited

number of white papers included; the search engines used included LTU library, Google Scholar, IEEE Xplore, Springers, ACM DL, Websco, Emerald, and Elsevier.

## 2. Research Question

In order to develop a general overview of the topic, a literature study is an appropriate way to identify the state-of-the-art in big data analytics. Big data is important because it is one of the main technologies currently used to solve industrial issues and to provide roadmaps for research and education. The question thus becomes What is the state of the art in big data analytics?

This research question is important to academia due to a lack of similar studies addressing the state of the art in big data analytics. To the best of the researcher's knowledge, no similar research has been conducted in recent years, despite big data analytics providing a basis for advancements at both technological and scientific levels (Nafus and Sherman, 2014; Elgendy and Elragal, 2014).

- A literature review on big data analytics shows what is already known and what should be known;

- It identifies research gaps in big data analytics by noting both "hot" topics that have already been studied extensively and solved problems in big data analytics, and those problems that are unsolved and research questions that remain unanswered and untouched;

- It opens the door for other researchers, better supporting the explosive increase in big data analytics;

- This research also frames valid research methodologies, goals, and research questions for such proposed study (Levy and Ellis, 2006; Cronin et al., 2008; Hart, 2018).

For industry, a literature review helps with examining areas in big data analytics that are already mature as well as identifying problems that have been solved and those that have not been solved yet. This clarity helps investors and businesses to think positively about big data (Lee et al., 2014; Chen, M. et al., 2014).

With regard to society, big data analytics help to address economic problems such as allocating funds, making strategic decisions, immigration problems, and healthcare problems such as cost pressures on hospitals, adding an extra dimension to addressing such societal problems (Chen et al., 2012).

# 3. Research Method

The research method for this work is a classic literature review, which is important because big data analytics is a vital modern topic that requires a solid research base. A literature review reconstructs the knowledge available in a specific domain to support a subsequent literature analysis. Many literature review processes are available, and three of the most common are shown in Figure 1; one of these, most commonly used in the information systems field, is followed in this work.
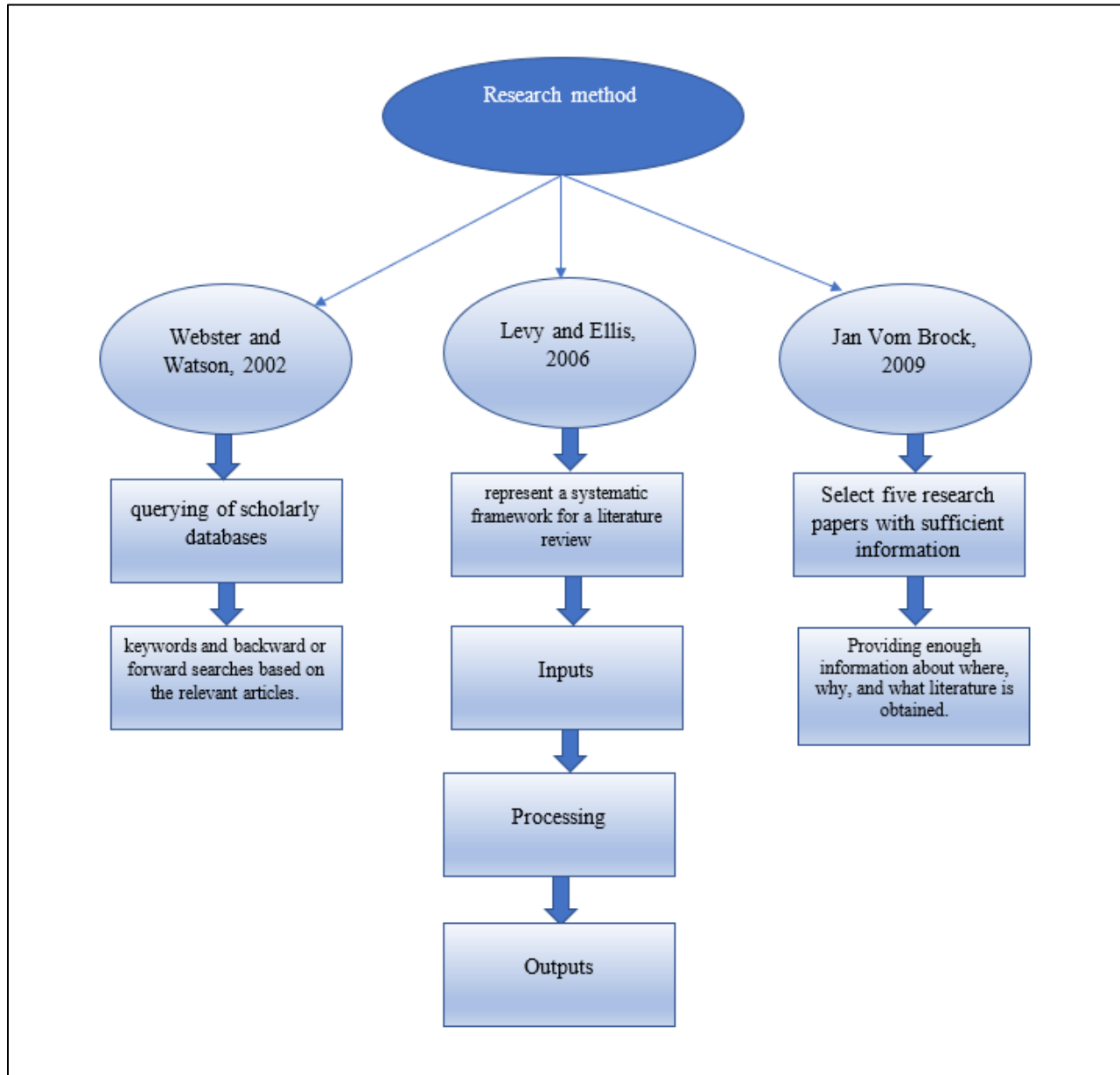


Figure 1: Approaches to writing an IS literature review.

A literature search according to Webster and Watson (2002), as shown in Figure 2, includes, the querying of scholarly databases with keywords and backward or forward searches on the basis of relevant articles discovered. This type of research is used for conducting many literature reviews and can be used to support a researcher's ideas at a given time. It includes citation searching, which allows the use of applicable articles both backwards and forwards in time. Reviewing such an article's references list to identify older articles that influenced or contributed to the author's work is called a *backward search*, while finding more recent articles that cite the article is called a *forward search*.
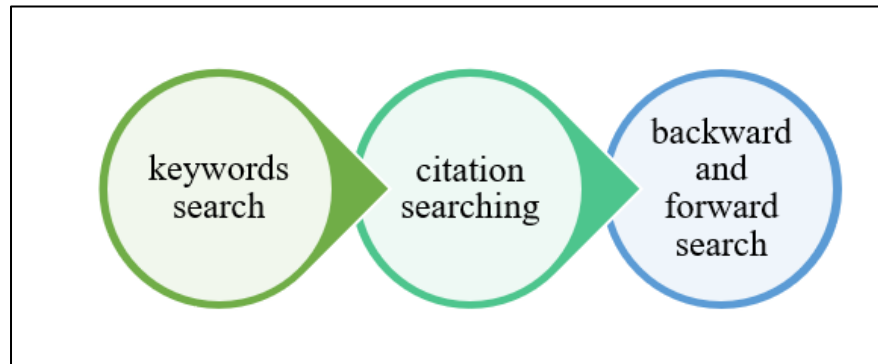


Figure 2: Research method according to Webster and Watson (2002).

However, Levy and Ellis (2006) suggest a more systematic framework for a literature review. A three-stage approach as shown in Figure 3 is suggested by the proposed framework: 1. Inputs, 2. Processing, 3. Outputs. The process should include "all sources that contain IS research publications", though this is challenging, as it is difficult and complicated to search and analyse such a vast quantity of articles (Levy and Ellis, 2006).



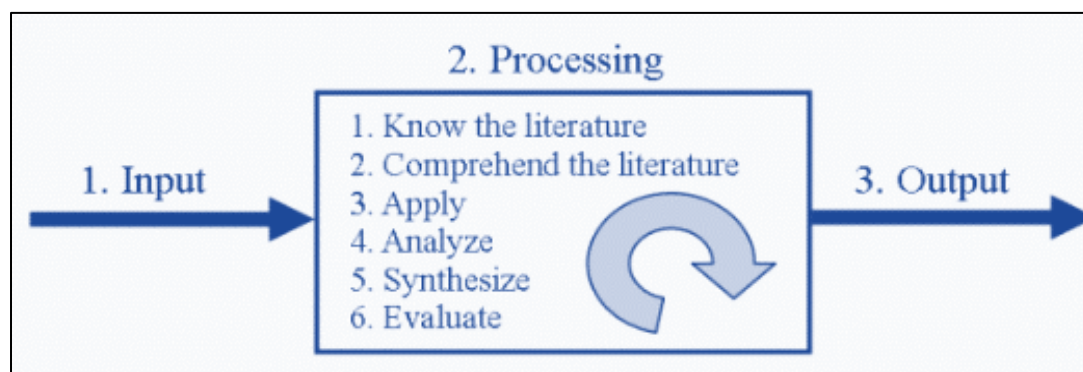Figure 3: The three stages of the effective literature review process, adopted from (Levy and Ellis, 2006).

The third research method, described by Vom Brocke et al. (2009) shows that only five research papers are required for a review as long as they contain sufficient information and are chosen for sensible reasons, and that this can be regarded as adding more value to both the authors and the

community than a review with a broad range of contribution analysis without sufficient information about where, why, and what literature was obtained. Such literature reviews are useful as any review article must document the literature search process. This method is based on the literature review analysis of results gained from ten of the most important information systems outlets based on a keyword search and a defined time period; it thus deliberately does not consider taking all available IS research papers or sources and analysing them. The processes for this are shown in Figure 4.

This research follows the procedure suggested by Vom Brocke et al. (2009) for writing a literature review as this method focuses on choosing papers for sensible reasons. The criteria for choice are dependent on the useful information that can be gained from such papers, the period of interest, and the number of citations, as well as whether the paper is from a peer-reviewed journal, conference, or other respectable source. These criteria are thus not randomly dependent on time periods or gathering all sources within all of the research field's publications.

Select a reference

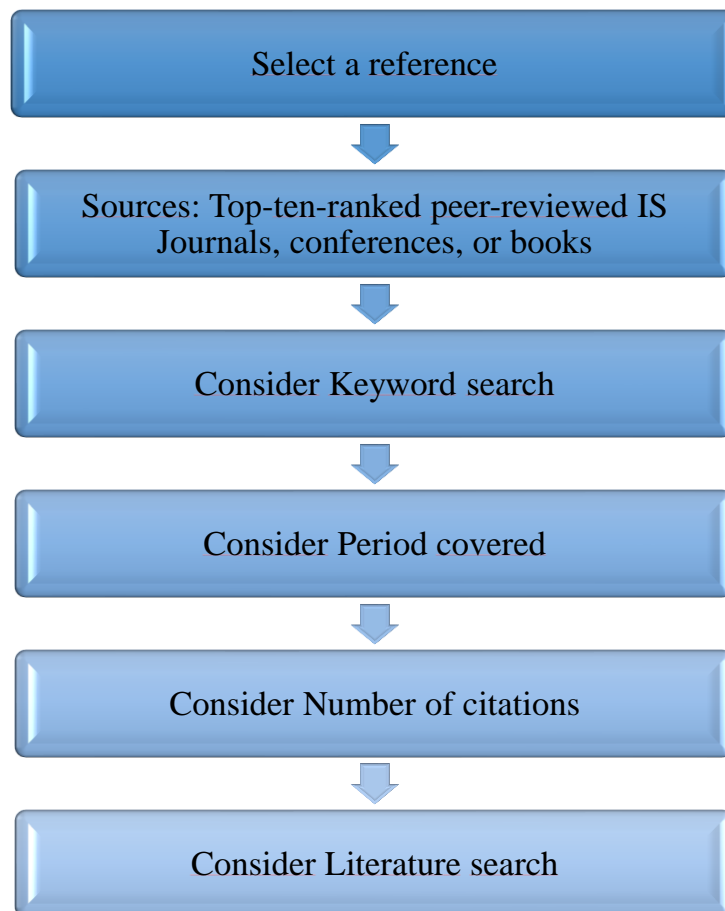Sources: Top-ten-ranked peer-reviewed IS Journals, conferences, or books

Consider Keyword search

Consider Period covered

Consider Number of citations

Consider Literature search

Figure 4: Stages of the effective search for the literature review process[1]

---

The literature review processes followed in this thesis are shown in Figure 5. They include

**Identifying the concept and review scope**

▶ Identifying the concept means determining what is needed to achieve the goal, and what work should be done to deliver the project. Such planning consists of documenting the project goals, features, tasks, and deadlines. In this research, this referred to the process of developing a literature review perspective on big data analytics.

**Finding related databases and sources**

▶ The search procedure for this thesis included the use of a range of relevant sources, such as ACM DL, IEEE Xplore, Emerald, EBSCO, WoS, LTU library, Google Scholar, Springers, and Elsevier.

▶ The resulting papers were then filtered based on year, abstract, content, citations, etc. The searches on big data analytics were filtered based on the top ten ranked peer-reviewed papers such as MIS Quarterly: Management Information Systems and Information Systems Research, with keyword searches including terms such as "Big data" and "big data analytics" for the period 2011 to 2019.

**Literature search**

▶ Analytical reading of papers refers to reading the papers chosen based on the aforementioned criteria deeply in order to understand the goals and the messages of those papers. Accordingly, the first step is to prepare the reading, reading the paper more than once and writing notes. The second is to use advanced reading techniques to re-read the paper to gain a better picture of and more insight into the paper's work as well as developing a better understanding. A final evaluative reading of the paper is then required.

**Literature analysis and synthesis**

▶ This literature review seeks to provide a description and evaluation of the current state of big data analytics. It designed to give an overview of the explored sources based on extensive searches around this topic, showing how the research covers a large study field in both academia and industry.

▶ Writing a literary analysis and synthesis for this topic thus involved generating a discussion based on several sources and showing the relationships between the sources, particularly when different ideas or focuses emerged in the research that required explanation or demonstrated new ideas or theories.

**Reviewing and combining the result**

▶ The research results from the big data analytics literature review are combined, then the work is reviewed, alongside an explanation of the methodology used and the debates arising.
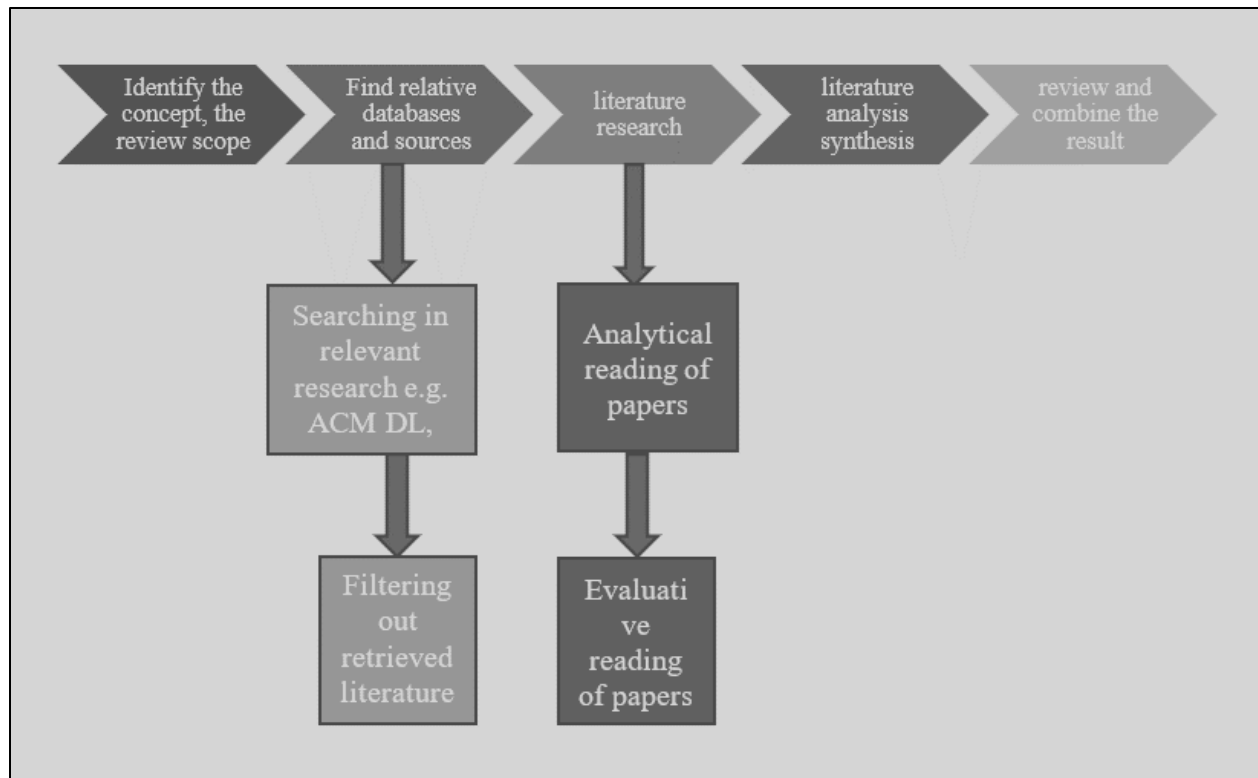
Figure 5: Literature review processes.

# 4. Scope delimitation and risks

The scope of this research will be determining the shortcomings in reviewing big data analytics, one can determine what has been defined and what is the criteria for selecting the analytics and tools for big data. The review can reveal which problems have been solved, and what else should be known. Moreover, it helps notifying the researchers about what have been presented which might open the doors for them to conduct more analytics for big data being an important topic nowadays and people directing toward this concept.

The main challenges of using big data, which need to be resolved before it can be used effectively, include security and privacy issues, data capturing issues, and challenges in data analysis and visualization to raise the positive role of big data analytics to many sectors. Storing the massive volume of data coming from different sources is another key point that needs to be addressed yet not currently resolved with the available tools. That created a need for studying and exploring new analytics method which might help in addressing some difficulties in some sectors such as in retail, banking, healthcare, etc.

Determining the possible solutions to the shortcomings with, data visualization, predictive analytics, descriptive analytics, and diagnostic analytics which are solutions to big data challenges in capturing and analysing the data. Organisations and individual use statistical models and

artificial intelligence modelling. Also, machine learning algorithms can integrate statistical and artificial intelligence methods to analyse massive amounts of data with high-performance. One solution for the storage challenge is utilizing Hadoop (Apache platform) that has the power to process highly large amounts of data. By separating the data into smaller parts then assigning some parts of the datasets to separate servers (nodes). Organizations should observe data sources, with end-to-end encryption used to prevent gaining access to the data in transit.

Companies must examine their cloud providers, as many cloud providers do not encrypt the data because of the massive amount of data convey at any given time, while encryption/decryption slows down the stream of data. Big data privacy solutions include protecting personal data privacy during gathering data such as personal interests, habits, and body properties, etc. of users who do not aware or easy to gain information from them. Also, protecting personal privacy data which might discharge during storage, transmission, and usage, even if it gained with the user permission.

Possible risks for this thesis could be in conducting the research which is how to identify a suitable subject based on the important point of finding a personal practical or professional need or a personal urge to face the research question. The risk was to confront two essential sources of confusion concerning the final success means in thesis writing. The first was the uncertainty about the understanding of the assessment criteria that will be applied to the work. The second relates to the insecurity concerning the risks that will be faced along the journey. The limitations of the study were those characteristics of design and the methodology that have been chosen which impacted the application results of the study. As the chosen method was a literature review according to Vom Brocke et al. (2009), the criteria for selecting references was not easy and a lot of references comply to the criteria of multi-dimension as aforementioned in the research method section.

## 5. What is "big data"?

Big data generally refers datasets that have grown too large for and become too difficult to work with by means of traditional tools and database management systems. It also implies datasets that have a great deal of variety and velocity, generating a need to develop possible solutions to extract value and knowledge from wide-ranging, fast-moving datasets (Elgendy, N. and Elragal, A., 2014).

According to the Oxford English Dictionary, "Big data" as a term is defined as "extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions". Arunachalam et al. (2018) argued that this definition does not give the whole picture of big data, however, as big data must be differentiated from data as being difficult to handle using traditional data analyses. Big data thus inherently requires more sophisticated techniques for handling complexity, as this is exponentially increased.

By 2011, the term big data had become quite widespread, but shows the frequency distribution of the "big data" in the ProQuest Research Library more clearly (Gandomi and Haider, 2015).
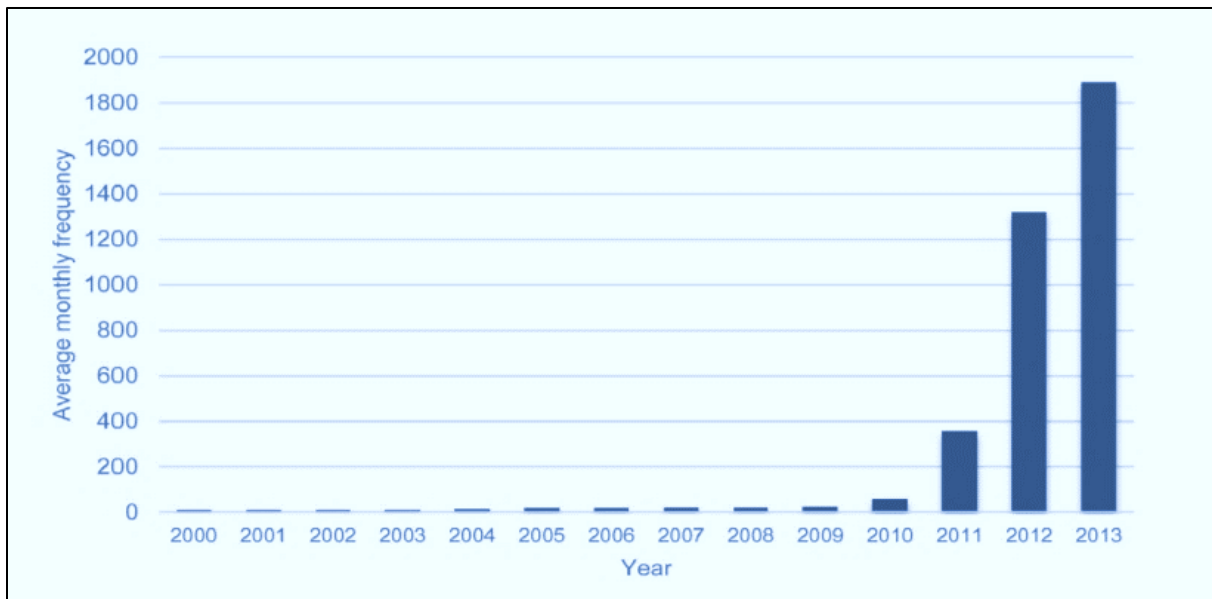
Figure 6: Frequency distribution of "big data" in the ProQuest Research Library (Gandomi and Haider, 2015).

Research by Gandomi and Haider (2015) shows that different definitions of big data are used in research and business. These big data definitions are vary depending on the understanding of the user, with some focused on the characteristics of big data in terms of volume, variety, and velocity, some focused on what it does, and others defining it dependent on their business's requirements. Figure 7 shows the different definitions of big found in an online survey of 154 C-suite global executives conducted by Harris Interactive on behalf of SAP in April 2012.

Early research work (Laney, 2001) focused on big data definition based on the 3Vs (volume, velocity, and variety). Sagiroglu and Sinanc (2013) later presented a big data research review and examined its security issues, while Lomotey et al. (2014) defined big data by 5Vs, extending the work done by Laney (2001) from 3Vs to include value, and veracity (Al-Barashdi and Al-Karousi, 2019 ).

Ren et al. (2019) thus recently developed a set of up-to-date big data definitions, as shown in Table 1. Figure 8 shows predictions of global data volume provided by International Data Corporation (IDC) (Tien, J.M., 2013). Besides the massive volume of big data, the complex structure of this new data and the difficulty in managing and protecting such data have added further issues. Since the idea of big data was raised, it has thus become one of the most popular focuses in both technical and engineering areas (Wang et al., 2016).
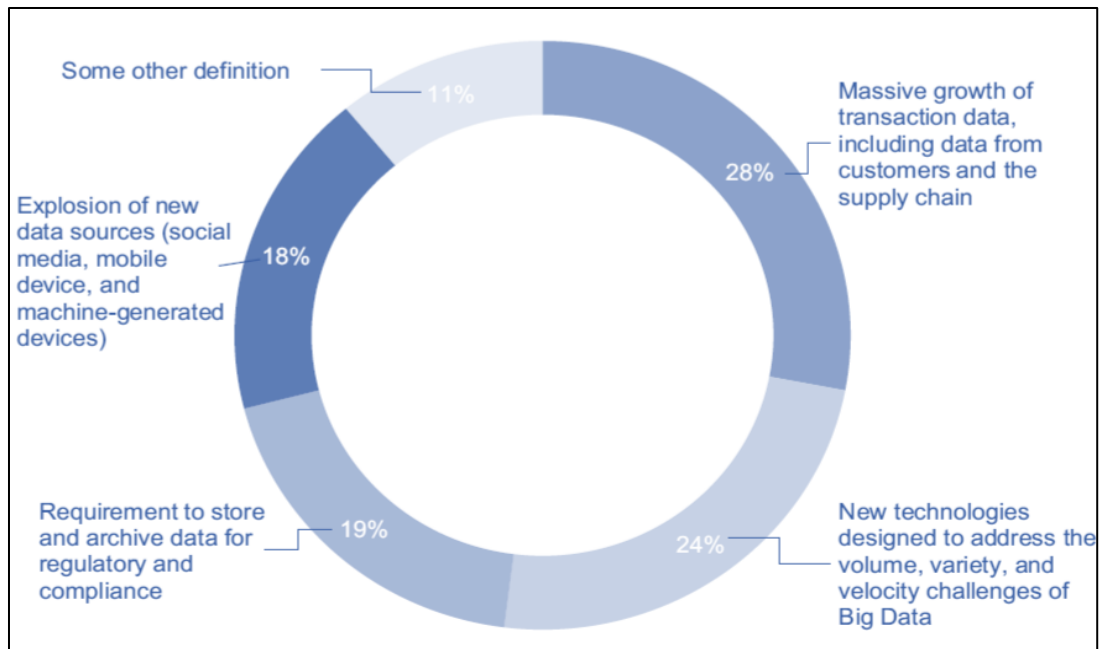
Figure 7: Definitions of big data (Online survey of 154 global executives in April 2012, Gandomi and Haider, 2015).

To realise big data's potential, the data should be gathered in a new way which enables it to be utilised for different purposes many times without recollection; this can be seen today in the many devices connected to the internet and the huge amount of data accesses even by individuals. By 2020, the predicted value of data is posited to double every 24 months (Mayer-Schonberger and Padova, 2015).



Figure 8: Global data volume predicted by IDC (Wang et al., 2016)

Table 1 shows various big data definitions or characteristics from the period 2001 to 2017.
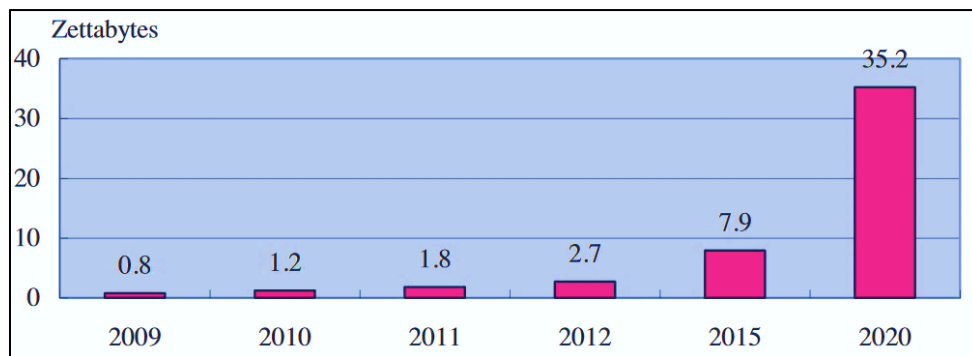
Table 1: Six representative definitions of big data adopted from (Ren et al., 2019).

| Authors/organizations | Definitions or characteristic |
|---|---|
| Laney (2001) | Characterized by 3Vs theory, namely volume, variety, and velocity. Volume: with the generation and collection of masses of data, data scale becomes increasingly big; Velocity: timeliness of big data, specifically, data collection and analysis must be rapidly and timely conducted; Variety: the various types of data, which include semi-structured and unstructured data as well as traditional structured data |
| Gantz et al. (2011) | Describes a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis |
| Manyika et al. (2011) | Refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. |
| Mashingaidze and Backhouse (2017); Daki et al. (2017) | Includes datasets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. |
| NIST (2012) | Means the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies |
| Zikopoulos et al. (2013) | Big data contains four dimensions, namely volume, variety, velocity, and veracity. Veracity: the unreliability and uncertainty inherent in some sources of data. |

Grover and Kar (2017) highlight that the number of big data articles published in reputable journals is increasing, as shown in Figure 9.
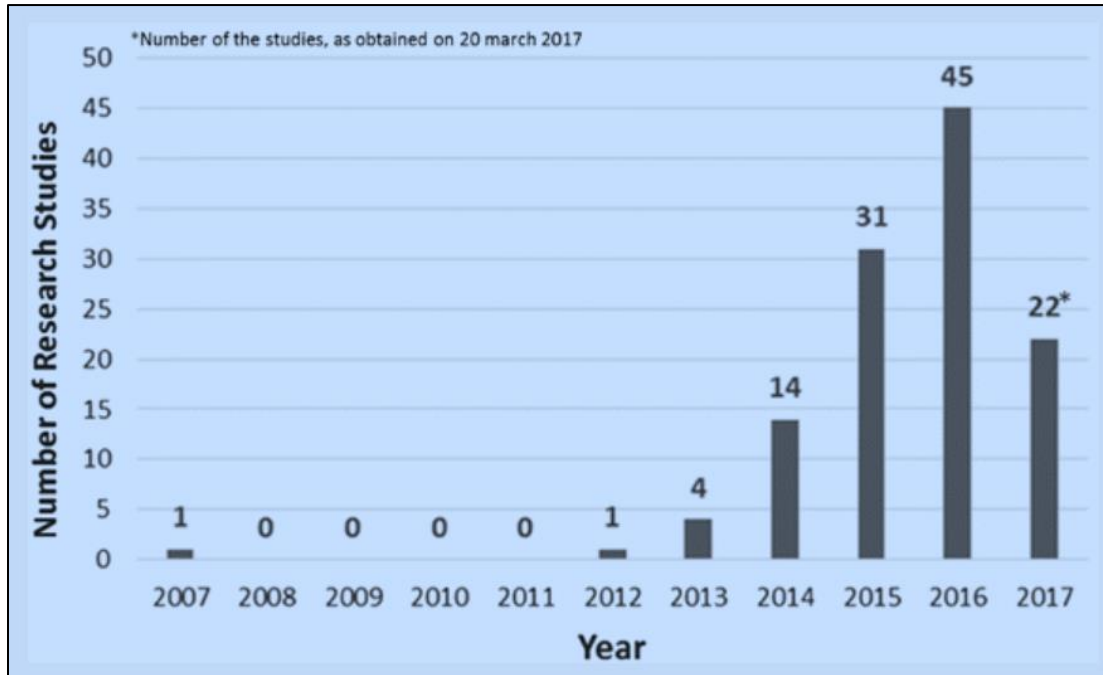


Figure 9: Yearly distribution of "big data" research studies (Grover and Kar, 2017).

Mikalef et al. (2018) also provided an overview of big data definitions in past studies, as shown in Table 2.

Table 2: Sample definitions of big data adopted from (Mikalef et al., 2018).

| Author(s) and date | Definition |
| --- | --- |
| Russom (2011) | Big data involves the data storage, management, analysis, and visualization of very large and complex datasets |
| White (2011) | Big data involves more than simply the ability to handle large volumes of data; instead, it represents a wide range of new analytical technologies and business possibilities. These new systems handle a wide variety of data, from sensor data to Web and social media data, improved analytical capabilities, operational business intelligence that improves business agility by enabling automated real-time actions and intraday decision making, faster hardware and cloud computing including on-demand software-as-a service. Supporting big data involves combining these technologies to enable new solutions that can bring significant benefits to the business |
| Beyer and Laney (2012) | Big data: 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 |
| McAfee et al. (2012) | Big data, like analytics before it, seeks to glean intelligence from data and translate that into business advantage. However, there are three key differences: Velocity, variety, volume |
| Gantz and Reinsel (2012) | Big data focuses on three main characteristics: the data itself, the analytics of the data, and presentation of the results of the analytics that allow the creation of business value in terms of new products or services |
| Boyd and Crawford (2012) | Big data: a cultural, technological, and scholarly phenomenon that rests on the interplay of (1) Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large datasets. (2) Analysis: drawing on large datasets to identify patterns in order to make economic, social, technical, and legal claims. (3) Mythology: the widespread belief that large datasets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy |
| Schroeck et al. (2012) | Big data is a combination of volume, variety, velocity and veracity that creates an opportunity for organizations to gain competitive advantage in today's digitized marketplace |
| Bharadwaj et al. (2013) | Big data refers to datasets with sizes beyond the ability of common software tools to capture, curate, manage, and process the data within a specified elapsed time |
| Kamioka and Tapanainen (2014) | Big data is large-scale data with various sources and structures that cannot be processed by conventional methods and that is intended for organizational or societal problem solving |
| Bekmamedova and Shanks (2014) | Big data involves the data storage, management, analysis, and visualization of very large and complex datasets. It focuses on new data-management techniques that supersede traditional relational systems, and are better suited to the management of large volumes of social media data |
| Davis (2014) | Big data consists of expansive collections of data (large volumes) that are updated quickly and frequently (high velocity) and that exhibit a huge range of different formats and content (wide variety) |
| Sun et al. (2015) | Big data: the data-sets from heterogeneous and autonomous resources, with diversity in dimensions, complex and dynamic relationships, by size that is beyond the capacity of conventional processes or tools to effectively capture, store, manage, analyze, and exploit them |

| Author(s) and date | Definition |
|---|---|
| Opresnik and Taisch (2015) | Big data typically refers to the following types of data: (1) traditional enterprise data, (2) machine-generated/sensor data (e.g. weblogs, smart meters, manufacturing sensors, equipment logs), and (3) social data |
| Constantiou and Kallinikos (2015) | Big data often represents miscellaneous records of the whereabouts of large and shifting online crowds. It is frequently agnostic, in the sense of being produced for generic purposes or purposes different from those sought by big data crunching. It is based on varying formats and modes of communication (e.g. text, image, and sound), raising severe problems of semiotic translation and meaning compatibility. Big data is commonly deployed to refer to large data volumes generated and made available on the Internet and the current digital media ecosystems |
| Akter et al. (2016a) | Big data is defined in terms of five 'Vs:' volume, velocity, variety, veracity, and value. 'Volume' refers to the quantities of big data, which are increasing exponentially. 'Velocity' is the speed of data collection, processing and analyzing in the real time. 'Variety' refers to the different types of data collected in big data environments. 'Veracity' represents the reliability of data sources. Finally, 'value' represents the transactional, strategic, and informational benefits of big data |
| Abbasi et al. (2016) | Big data differs from 'regular' data along four dimensions, or '4 Vs'— volume, velocity, variety, and veracity |

The abovementioned definitions are complimentary to each other at some points such as defining the big data by 5Vs in Lomotey et al. (2014). At other points, some of them are contradicting with the six representative definitions adopted from (Ren et al., 2019) that are shown in Table 1. They defined big data in term of three 'Vs' and they focused on the size of data ignoring the other dimensions.

When taken from the user understanding viewpoints, these definitions show different angles of big data used in research and business as in Gandomi and Haider (2015). The characteristics in terms of volume, variety, and velocity are the focus in some of them, whilst the function and requirements are the focus points in others such as the business requirements and how the data is stored.

However, the definition adopted in this work is the one that contains all the dimensions (i.e. the 5Vs). This is because it is regarded as being of very high density, timeliness, and different structure, format, and sources, which requires high performing processing.

# 6. Big data characteristics

Based on the various big data definitions, it is obvious that the size is the dominating characteristic despite other characteristics' importance. Laney (2001) proposed the three V's as the dimensions of challenge to data management, and the three V's constitute a common framework (Laney, 2001; Chen et al., 2012). These three dimensions are not independent of each other; if one-dimension changes, the probability of changing another dimension also increases (Gandomi and Haider, 2015).

A further two dimensions are often added to the big data characteristics, *veracity* and *variability* (Gandomi, A. and Haider, M., 2015) as shown in Figure 10. The five V's reflect the growing popularity of big data. The first V is, as always, volume, which is related to the amount of generated data (Grover and Kar, 2017). The second V is for the velocity (big data timeliness), as all data collection and analysis should be conducted in a timely manner (Chen, M., Mao, S. and Liu, Y., 2014). The third V refers to variety, as big data comes in many different formats and structures such as ERP data, emails and tweets, or audio and video (Russom, 2011; Elragal, 2014; Watson, 2014; Watson, 2019). The fourth V refers to big data's "huge value but very low density", causing critical problems in terms of extracting value from datasets (Elragal, 2014; Chen et al., 2014; Raghupathi and Raghupathi, 2014). The fifth V references veracity, and questions big data credibility where sources are external, as in most cases (Addo-Tenkorang and Helo, 2016; Grover and Kar, 2017; Al-Barashdi and Al-Karousi, 2019 ). Veracity is related to credibility, the data source's accuracy, and how suitable the data is for the proposed of use (Elragal, A, 2014).

Using big data requires the correct technical architecture, analytics, and tools to enable insights to emerge from hidden knowledge to generate value for business, and these depend on the data scale, distribution, diversity, and velocity (Russom, 2011). Big data is most easily characterised by its three main features, however: Data Volume (size), Velocity (data change rate) and Variety (data formats and types as well the data analysis types required) (Elgendy and Elragal, 2014; Schelén, Elragal, and Haddara, 2015; Chen and Guo, 2016; Elragal and Klischewski, 2017).

Streaming data is the leading edge of big data, as it can be collected in real-time from multiple websites. The addition of the final V, veracity, has been discussed by several researchers and organisations in this context. Veracity focuses on the quality of the data, which may be good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, or approximations. As most big data sources are external, they lack governance and have little homogeneity (Elragal, 2014; Elgendy and Elragal, 2014; Russom, 2011).

The important thing for modern organisations seeking competitive advantages is how to manage and extract the value from data. Big data combines technical challenges with multiple opportunities, and thus extracting business value represents both challenge and opportunity at the same time. This puts big data business perspective side-by-side with technical aspects and showing how big data adds value to organisational objectives has become a crucial aspect of research in this field. Manyika et al. (2011) clarified how big data can generate value-add for organisations by

➢ making information clear and applicable more frequently;
➢ allowing organisations to create and store transactional data in digital form, making it easier for them to gather more precise information about inventories and products;
➢ using sophisticated big data analytics to improve decision making quality;

> ➤ utilising big data to shape the next generation of products and services (Elragal, A, 2014).

Quantifying big data can be done in terms of storage size, number of records, transactions, tables, or files. Big data comes from multiple diverse sources collected for many purposes (Constantiou and Kallinikos, 2015), including IoT data, logs, clickstreams, and social media. For all of those sources to be used for analytics requires joining up unstructured data (such as texts in natural language) and semi-structured data (such as extensible mark-up language (XML), JSON or rich site summary (RSS) feeds) to a common structured data framework (Elgendy and Elragal, 2014; Elragal, 2014).
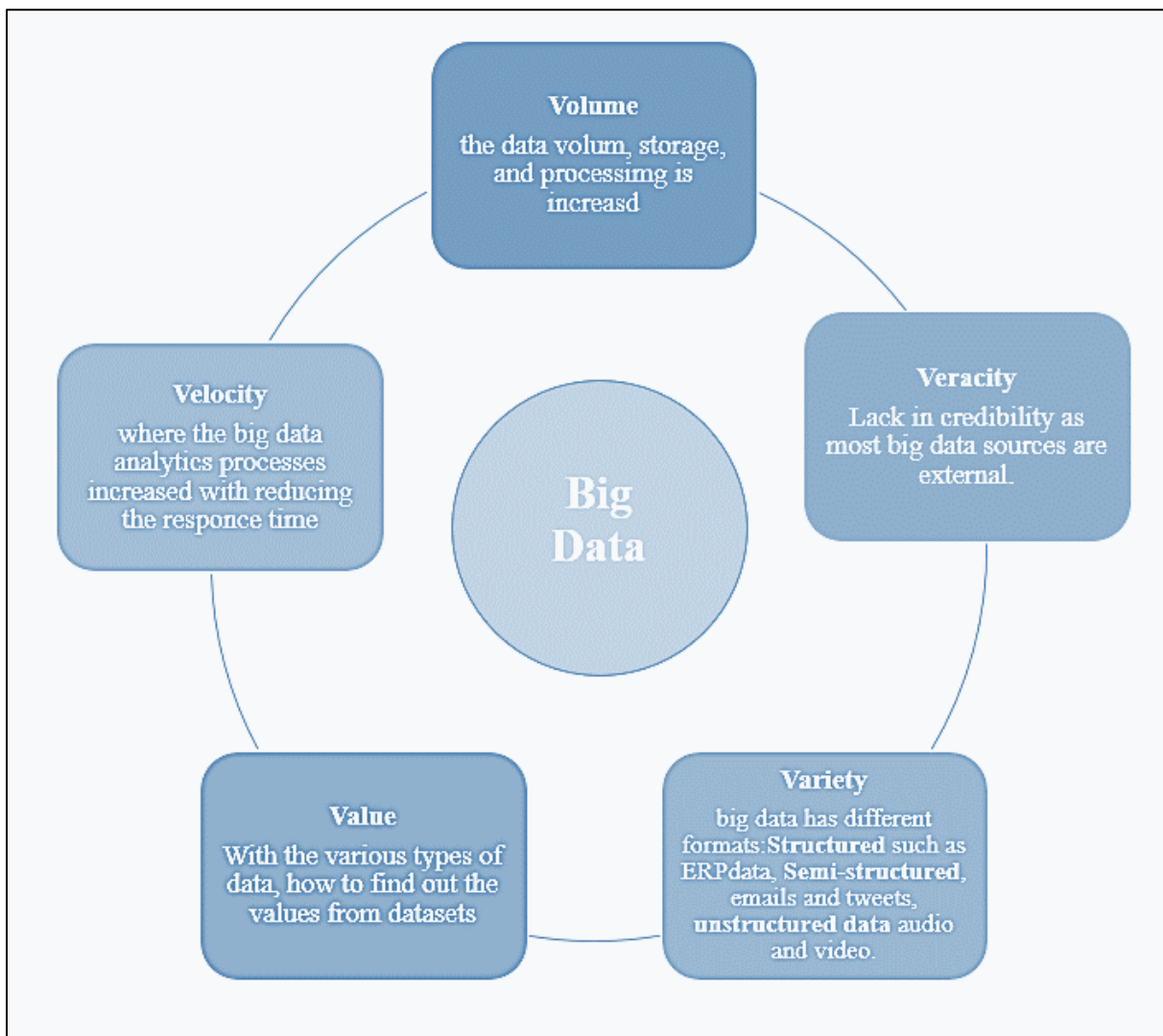
Figure 10: Big data in terms of the 5 V's.

Multi-dimensional data can be used to add historical context to big data. The variety of big data is as important as its volume, while velocity or speed can describe how difficult big data may be to

handle. Velocity may refer to data generation frequency or data delivery frequency. Depending on data inconsistency, incompleteness, ambiguity, latency, deception, and approximations, big data quality can also be characterised as undefined, good, or bad (Data, D.B., 2012).

According to Mikalef et al. (2018), various researchers focus on different aspects of big data, as shown in Table 3.

Table 3: Defining characteristics of big data, adopted from (Mikalef et al., 2018)

| Attribute | Definition |
|---|---|
| Volume | Volume represents the sheer size of the dataset due to the aggregation of a large number of variables and an even larger set of observations for each variable. (George et al. 2016) |
| Velocity | Velocity reflects the speed at which data are collected and analyzed, whether in real time or near real time from sensors, sales transactions, social media posts, and sentiment data for breaking news and social trends. (George et al. 2016) |
| Variety | Variety in big data comes from the plurality of structured and unstructured data sources such as text, videos, networks, and graphics among others. (George et al. 2016) |
| Veracity | Veracity ensures that the data used are trusted, authentic, and protected from unauthorized access and modification. (Demchenko et al. 2013) |
| Value | Value represents the extent to which big data generates economically worthy insights and/or benefits through extraction and transformation. (Wamba et al. 2015) |
| Variability | Variability concerns how insight from media constantly changes as the same information is interpreted in a different way, or new feeds from other sources help to shape a different outcome. (Seddon and Currie 2017) |
| Visualization | Visualization can be described as interpreting the patterns and trends that are present in the data. (Seddon and Currie 2017) |

3Vs: volume, velocity, variety (Chen and Zhang 2014)

4Vs: volume, velocity, variety, veracity (Zikopoulos and Eaton 2011; Schroeck et al. 2012; Abbasi et al. 2016)

5Vs: volume, velocity, variety, veracity, value (Oracle 2012; Sharda et al. 2013)

7Vs: volume, velocity, variety, veracity, value variability, visualization (Seddon and Currie 2017)

# 7. Big data analytics (BDA): tools and methods

## 7.1. Big data storage and management

The most difficult problem that needs to be solved to handle big data effectively is storage; it is not necessarily easy to deal with large quantities and varieties of data (Elgendy and Elragal, 2014; Zhong, et al., 2016; Lv, Z. et al., 2017).

There are many big data storage and analysis models. Where the large amount of data is caused by the sheer variety of users and devices, a data centre may be necessary for storing and processing the data. Establishing network infrastructure is necessary to help gather this rapidly generated data, which is then sent to the data centre before being accessed by users (Lv et al., 2017).

Research by Yi et al. (2014) identifies the components of the network that must be established, such as an original data network, the bridges used for connecting and transmitting to data centres, and at least one data centre.

Another study (H. Eszter, 2015) highlighted the issues in using big data through specific locations and showed that the users could not select data through the data network. For storage models, the most important challenge is how to deal with the sheer amount of data, as ultra-scalable solutions can block the processing of certain data sources, causing inefficiency. Building more scalable big data technology is a challenge, and any new technology must offer data gathering and distribution among nodes spread through the world (Lv et al., 2017).

Structured data storage and retrieval methods include "relational databases, data marts, and data warehouses" (Elgendy, N. and Elragal, A., 2014). Data is extracted from outside sources, then transformed to fit operational needs, and finally loaded into the database. The data is then uploaded from the operational data store to longer-term storage using Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT) tools. The data is then cleaned, transformed, and catalogued before use (Bakshi, 2012; Elgendy and Elragal, 2014).

A big data environment requires analysis skills, unlike the Enterprise Data Warehouse (EDW) traditional environment (Hartmann, T. et al., 2019).

➢ The big data environment accepts and demands all possible data sources. On the other hand, EDW approaches data sources with caution, as it is more streamlined towards supporting structured data (Elgendy and Elragal, 2014; Hartmann et al., 2019).

➢ Due to increasing number of data sources and data analyses possible, big data storage requires agile databases to give analysts the opportunity to produce and adapt to data easily and quickly (Elgendy and Elragal, 2014; Hartmann et al., 2019).

➢ A big data repository must be deep, allowing analysts to analyse the datasets deeply by using complex statistical methods (Elgendy and Elragal, 2014; Hartmann, T. et al., 2019).

Hadoop is a popular big data analytics framework. Hadoop "provides reliability, scalability, and manageability by providing an implementation for the MapReduce paradigm as well as gluing the storage and analytics together" (Elgendy, N. and Elragal, A., 2014). Hadoop includes HDFS which

is for the big data storage and MapReduce for big data analytics, and it can process extremely large amount of data by dividing the data into smaller blocks, then specifying datasets to be distributed across cluster nodes (Raghupathi and Raghupathi, 2014; Elgendy and Elragal, 2014). Hadoop incorporates several technologies: "Hive is a data warehouse implementation for Hadoop, MapReduce is a programming model in Hadoop, and Pig is a querying language for Hadoop which has similarities to the SQL language for relational databases" (Zuech et al., 2015).

First-generation technology generated the Apache Spark project in software terms (Watson, 2019), but Hadoop has a great deal more power, which offers advantages to analytics in terms of memory. It can work with both batch and real-time workloads, is easy to program with Java code, and can connect to Apache projects and other software within a closed ecosystem. Hadoop's components are shown in Figure 11 (Watson, 2019):

1. Spark SQL runs SQL-like queries on structured data.
2. Spark streaming provides real-time data processing.
3. MLib provides a machine learning library of algorithms and utilities.
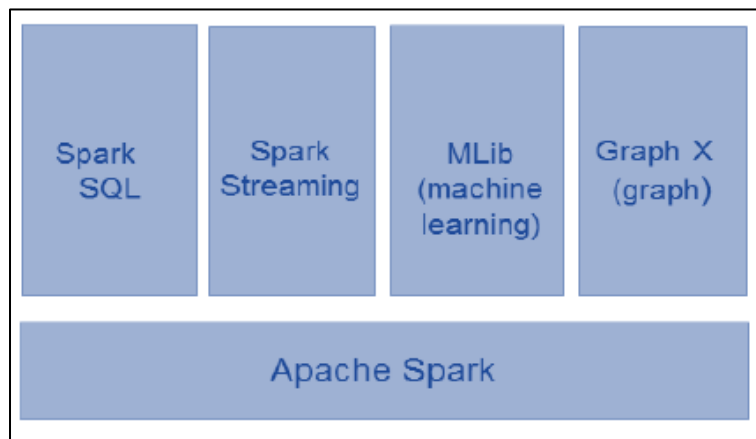4. Graph X provides application algorithms.



Figure 11: Spark Components (Watson, 2019)

## 7.2. Big data analytics processing

Analytics processing is the next issue after big data storage. According to He et al. (2011), big data analytics processing has four critical requirements:

a) *Fast data loading*: limited interference between disk and network, to speed up query execution.

b) *Fast query processing*: workloads are heavy, therefore real-time requests should be processed as quickly as possible to satisfy user requirements. The data placement structure should also have the ability process multiple queries as query volumes increase.

*c)* *Highly efficient utilization of storage space*: as user activities grow rapidly, they need scalable storage capacity and computing power. As disk space is limited, it is necessary to manage data storage during processing and address the space issues adaptively.

*d)* *Strong adaptivity to highly dynamic workload patterns*: the underlying system should be highly adaptive, as data processes have different workload patterns and the analysing of big datasets has many different applications and users, with different purposes and methods (Elgendy, N. and Elragal, A., 2014).

The work presented by García et al. (2016) shows that using big data frameworks for storing, processing, and analysing data has changed the context of knowledge discovery from data, mainly in terms of data mining processes and pre-processing, with a particular focus on the rise of data pre-processing in cloud computing. The presented solution covered various data pre-processing technique families with factors such as maximum size supported examined in terms of big data and data pre-processing throughout all of the families of methods. Moreover, various big data framework such as Hadoop, Spark, and Flink were discussed.

## 7.3.  Big data analytics

Big data growth continues apace, and many organisations are now interested in managing and analysing data. Organisations trying to benefit from big data are adopting big data analytics to facilitate faster and better decisions, as it is not easy to analyse datasets with analysis techniques and infrastructure based on traditional data management (Constantiou et al., 2015). The need for new tools and methods specialised for big data analytics is thus also growing. The emergence of big data is affecting everything from data itself to its collection and processing, and, finally, the extracted decisions. Providing big data tools and technologies can help in managing the growth of network-produced data, which is otherwise exponential, as well as in increasing the capability of organisations to scale and capture the required data to reduce database performance problems (Elgendy, N. and Elragal, A., 2014). Further big data analytics definitions are clarified in Table 4.

Opening any popular scientific or business publication today, whether online or in the physical world, generally involves running into a reference to data science, analytics, big data, or some combination of these terms (Agarwal and Dhar, 2014). Some researchers are focusing on big data definitions (Akter et al., 2016; Mikalef et al., 2018), while others analyse the tools, techniques, and procedures required for analysis (Russom, 2011), and others seek to explain big data analytics' impact on business value ( Mikalef et al., 2018).

Table 4: Sample definitions of big data analytics, adopted from (Mikalef et al., 2018)

| Authors and date | Definition |
|---|---|
| Loebbecke and Picot (2015) | Big data analytics: a means to analyze and interpret any kind of digital information. Technical and analytical advancements in BDA, which—in large part—determine the functional scope of today's digital products and services, are crucial for the development of sophisticated artificial intelligence, cognitive computing capabilities, and business intelligence |
| Kwon et al. (2014) | Big data analytics: technologies (e.g. database and data mining tools) and techniques (e.g. analytical methods) that a company can employ to analyze large-scale, complex data for various applications intended to augment firm performance in various dimensions |
| Ghasemaghaei et al. (2015) | Big data analytics, defined as tools and processes often applied to large and disperse datasets for obtaining meaningful insights, has received much attention in IS research given its capacity to improve organizational performance |
| Lamba and Dubey (2015) | Big data analytics is defined as the application of multiple analytic methods that address the diversity of big data to provide actionable descriptive, predictive, and prescriptive results |
| Müller et al. (2016) | Big data analytics: the statistical modeling of large, diverse, and dynamic datasets of user-generated content and digital traces |

People now aim to both to collect data and understand its importance and meaning for use in making decisions. The data to be analysed is large in volume and consists of various types. "Massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous" (Ma et al., 2014), are features of big data that require changes in statistical and data analysis approaches.

It is also important to understand the content of big data. The process of applying algorithms to analyse the content of big data is part of data analytics, which is used for 1) analysing sets of data information and their relationships, 2) extracting previously unknown valid patterns, and 3) for detecting important relationships between stored variables.

In this section, various big data analyses will be discussed, beginning with the data analysis techniques available and some of the common big data analytics suites, finally discussing several big data platforms and tools. Data analysis techniques can be characterised into four types, as shown in Figure 12:
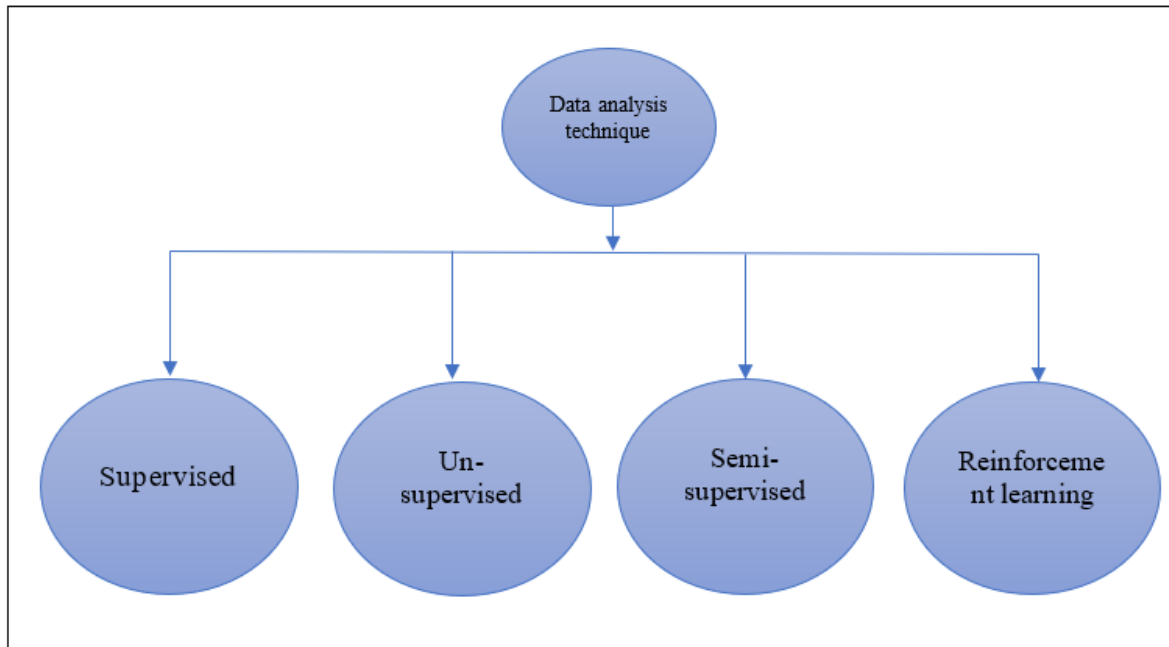
Figure 12: Data analysis techniques

### 7.1.1. Supervised techniques

A supervised technique refers to where data are trained and tested, and the training data is labelled. Labelled means that the full history of what has happened to the data is known, and thus the history for the data variables is known.

Supervised learning involves training a system based on labelled data and this requires a supervisor with the ability to expect the output from each input that can train the system according to its expectations. When the system is trained, it can give predictions within "many applications of classification and fault detection and channel coding and decoding" (Kotsiantis, et al., 2007; Cui, et al., 2019). This technique is used for approximating a function between the input and output.

The idea is for the system to learn the training dataset's classifiers (the labelled documents) then to automatically apply this classification to an unknown dataset's un-labelled documents. This learning technology thus involves learning from example (Boyd-Graber et al., 2014; Müller et al., 2016; Breed and Verster, 2019).

Regression is an example of the supervised learning algorithm, as are Linear Regression, Decision Trees (DT), Support Vector Machine (SVM), K Nearest Neighbour (K-NN), Naive Bayes Classifier (NBC), Random Forest, and neural networks (NN). However, many of these supervised techniques cannot be used with wireless networks, and as the learning techniques are dependent on the data training, the results are also restricted (Cui et al., 2019).

**Regression Analysis:** is mathematical tool used to discover correlations between several variables based on experimental or observed data. Where analysis defines the relationships between variables as non-random, such analysis may make the correlations between variables appear simpler and more regular (Lei et al., 2016), as shown in Figure 13.
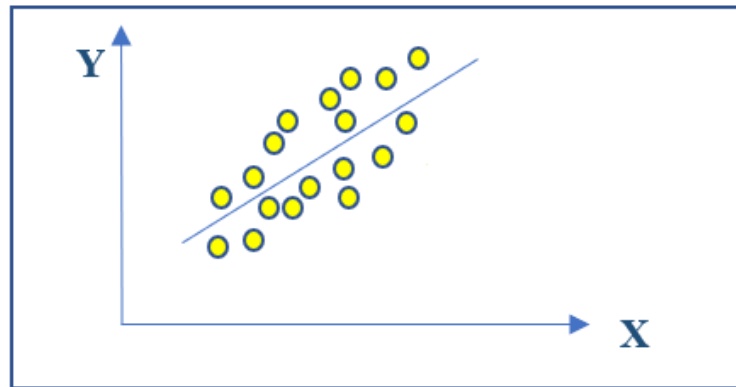


Figure 13: Regression analysis

Structured data mostly utilises predictive analytics, and this overshadows other analytics forms for 95% of big data (Gandomi and Haider, 2015). However, new statistical techniques for big data have emerged which clarify the differentiation of big data from smaller data sets. In practice, however, most statistical methods were designed for smaller datasets, in particular, samples.

Usually, scientists make predictions based on theories in the prediction domain. However, big data analytics can deliver predictions that depend on the sequence of data processing and execution. According to Kitchin (2014) and Müller et al. (2016),

- big data brings new challenge as it is generated from different system sources. The data retrieved from each source system should thus be sent to a central repository;
- the relationship between operations should be defined to allow reconstruction of datasets from multiple sources;
- the knowledge discovery process should be automated from data or datasets to make predictions;
- generating new theories is required to create and improve models. Predicted target theory generates a set of predictors; however, some theories explain the relationships between independent and dependent predictors more effectively;
- there is a shift from theory-driven to process-driven prediction based on analysing the BDA steps and identifying the challenges, theoretically informing future BDA needs throughout data acquisition, pre-processing analysis, and interpretation.

### 7.1.2. Un-supervised techniques

Here, the training data is unlabelled. Unlabelled means that the history of the data is missing, there is no history available for data variables, and the data have not been trained and tested. Thus, unsupervised techniques require separate training data (Boyd-Graber et al., 2014; Müller et al., 2016; Breed and Verster, 2019).

Unsupervised learning requires deducing functions for presenting unknown structures from unlabelled data. This technique does not require a supervisor, which means that the system must have the ability to proceed independently with training based on unlabelled data input (Cui et al., 2019).

Examples of unsupervised learning algorithms include clustering algorithms, combinatorial algorithms, A priori algorithms, Self-Organizing Maps (SOM), and applications of game theory. These techniques are used for classifying the input data into different clusters or classes based on the data distribution (Jiang et al., 2017; Cui et al., 2019).

**Cluster Analysis:** This method is based on grouping objects and classifying them depending on shared features. It is used for differentiation between objects to allow division into clusters. Thus, data which are related to each other or have the same features will be placed in a cluster or a group and unrelated data will be in other groups (Wu et al., 2018; Cui et al., 2019), as shown in Figure 14.
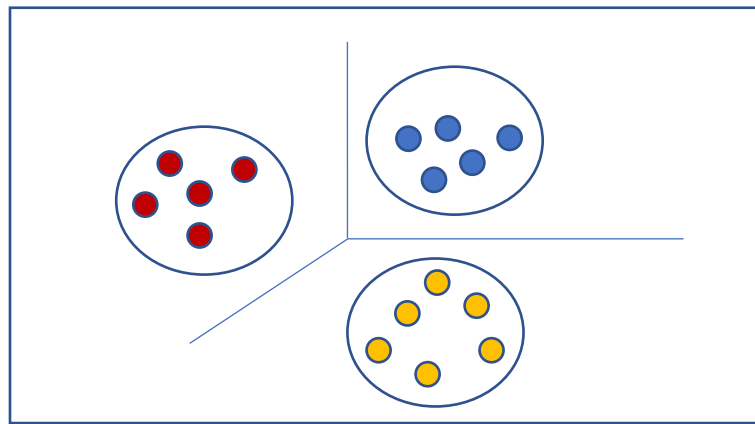


Figure 14: Cluster analysis.

### 7.1.3. Semi-supervised techniques

Where some of the data is labelled and some is unlabelled, supervised and unsupervised techniques can also be mixed. Algorithms are applied for both labelled and unlabelled data, and even with incomplete information or missing training sets, some of the dataset's classifiers can be learned.

Both supervised and unsupervised techniques focus on one aspect (target separation or independent variable distribution, respectively), and using them together may thus give better results (Breed, D.G. and Verster, T., 2019).

### 7.1.4. Reinforcement learning (RL)

Reinforcement learning involves setting and classifying real-time data changes in a way that allows the learning framework to adapt based on those changes (Wu et al., 2018; Cui et al., 2019).

The components of an RL algorithm are the agent; the environment; and the actions. The actions are taken by the algorithm based on the environment, and depending on the feedback from the environment, it determines whether the action is positive, thus using it again in future, or negative, thus discarding it. An example of reinforcement learning is Markov Chains (Markov Decision Process) (Müller et al., 2016). The difference between RL and supervised or unsupervised learning is that RL works based on the feedback which is either good or not depending on the situation and is hence dynamic, while supervised and unsupervised learning give static solutions (Cui, et al., 2019).

The RL process includes an actor which acts in the environment with its own copy of the data; the data can thus be stored in a separate replay memory and sampled by the learner to be computed within the policy parameters. The actor learners then receive the updated policy parameters (Mnih et al., 2015; Mnih et al., 2016).

The Map-Reduce framework was utilized by Li and Schuurmans (2011) for parallelising batch reinforcement learning methods with linear function approximation (Mnih et al., 2016). Applying parallelism helped speed up large matrix operations but did not assist the collection of experience or stabilise learning.

The reinforcement learning goal is to develop policies that help in decision making. An example, is Q-learning, where the algorithm has no knowledge of the data but has the ability to find out about the data in an automated way (Wu et al., 2018; Cui et al., 2019). Q-learning is one of the most popular reinforcement learning algorithms, though it learns unrealistically high action values as it includes "a maximization step of overestimated action values, which tends to prefer overestimated to underestimated values" (Hester et al., 2018). The Q-learning algorithm is thus best used for overestimating action values in specific conditions.

Recently, Q-learning has been combined with deep neural networks to produce Double Q-learning (DQN); that combination also suffers from overestimations (Mnih et al.,2015). Deep neural networks are artificial neural networks with multiple layers between the input and output layer, which help RL algorithms to provide effective performance. However, it was previously thought was that combining simple online RL algorithms with deep neural networks was unstable (Mnih et al., 2013; Mnih et al., 2015; Schulman et al., 2015; Mnih et al., 2016; Van Hasselt et al., 2016). The common idea arising from early studies was that the data sequences observed by online RL agents were not stable, and had no strong correlations to RL updates. However, data can be batched if the agent's data are stored in an experience replay memory (Schulman et al., 2015) or sampled from different time steps randomly (Mnih et al., 2013; Mnih et al. , 2016; Van Hasselt et al., 2016), and the Double Q-learning algorithm can work with large-scale function approximation (Hasselt,

H.V., 2010). Thus, a new algorithm known as Double DQN (a combination of Double Q-learning with neural networks) has been constructed which offers higher scores on several games; however, this algorithm has not displayed more accurate value estimation (Hester et al., 2018).

## 7.4.    Analytics techniques

 **Correlation Analysis:** this is an analytical method used to determine the relationships such as "correlation, correlative dependence, and mutual restriction, among observed phenomena and accordingly conducting forecast and control" (Chen, M., Mao, S. and Liu, Y., 2014), as shown in Figure 15. Positive correlation, on the left means while one variable increases so does the other. No linear correlation on the middle means there is no visible relationship between the variables. Negative correlation on the right means as one variable increases, the other decreases (Chen, M., Mao, S. and Liu, Y., 2014).
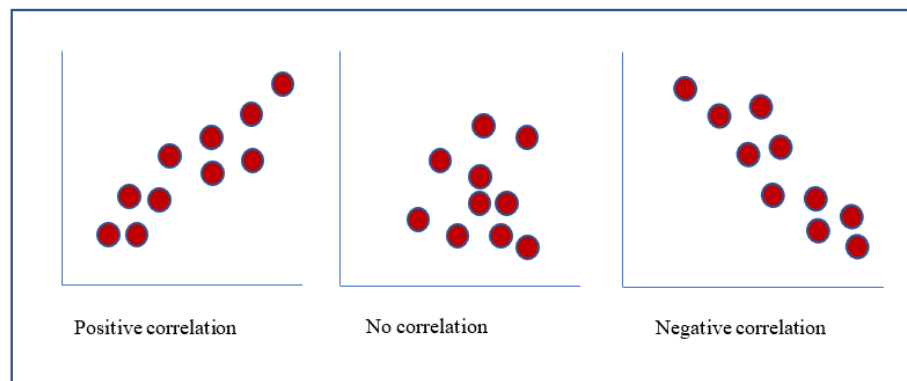


Figure 15: Correlation Analysis.

**Text Mining**: This converts the content from unstructured text to structured text in order to help uncover the meaning and the information contained.

**Factor Analysis:** This groups several related variables into a single factor, which means that fewer factors are used in analysis, which is thus simpler.

The research presented in Schelén et al. (2015) examines the state-of-the-art in big data at that time and discusses research agendas. In addition, it defines the basic technology and toolsets used. It is not easy to analyse datasets with traditional data management techniques (Constantiou and Kallinikos, 2015); therefore, new methods and tools have been developed for big data analytics, as well as for storing and managing such data. These solutions thus need to be studied in terms of handling datasets and extracting knowledge and value. In addition, the rapid changes in data volume, variety, velocity, and value require decision makers to know how to obtain valuable insights.

Traditional data analysis uses formal statistical methods to analyse data, constructing, extracting, and refining useful data, and identifying subject matter relationships in order to maximise the value of data. It can now be regarded as an analysis technique to be used for special kinds of data, though many traditional data analysis methods are still be used for big data analysis where analysts have backgrounds in statistics and computer science.

Association rules, clustering, classification, decision trees, and regression are the most common data analytics methods; however, some additional analyses have become common in terms of big data, especially in terms of social media, which relies on social networking and content sharing. Social network analysis is thus dependent on the relationships between social entities. Text mining used to analyse the contents of documents and to develop an understanding of the information therein. Sentiment analysis is then used to analyse the emotions underlying that content, and this more important form of analysis uses language processing to identify such information.

Finally, advanced data visualisation is becoming an important analysis tool, as this enables faster and better decision making (Russom, 2011; Elgendy and Elragal, 2016). Some of the more common models and analyses are explained further below, and shown in Figure 16:

- **Text analytics:**
  - ➢ **Sentiment Analysis:** This is based on understanding the subjects' emotions from their text patterns to help in organising viewpoints into good or bad, positive or negative (Mouthami et al., 2013). This analysis helps firms by alerting them where customers are dissatisfied or seeking to shift to other products, allowing preventative actions to be taken (Elgendy, N. and Elragal, A., 2014).


- **Audio analytics or speech analytics using technical approaches:**
  - ➢ **LVCSR**: large-vocabulary continuous speech recognition, indexing and searching.
  - ➢ **Phonetic-based systems:** work with sounds or phonemes (Gandomi and Haider, 2015).

- **Social media and social network analysis (SNA):** Social media depends on multiple tools and frameworks for collecting, monitoring, summarising, analysing, and visualising social media data, and SNA depends on social entities' relationships with each other to measure the knowledge linking parties, including who shares information, what information, and with whom. SNA tries to get develop network patterns, while social media tries to uncover useful patterns and user information using text mining or sentiment analysis (Elgendy and Elragal, 2014; Gandomi and Haider, 2015).

- **Data Visualisation**: This can be used even by decision makers with little knowledge about the data, as it presents the information visually prior to deep analysis. Advanced Data visualisation (ADV) offers strong potential growth to big data analytics as it allows analysis of data at several levels by taking advantage of human perceptual and reasoning abilities (Manyika et al., 2011; Russom, 2011; Elragal, and Klischewski, 2017).

- **Predictive analytics**: This is based on statistical methods such as associative rules, clustering, classification and decision trees, regression, and factor analysis (Fan et al., 2014; Bradlow et al., 2017; Breed and Verster, 2019).
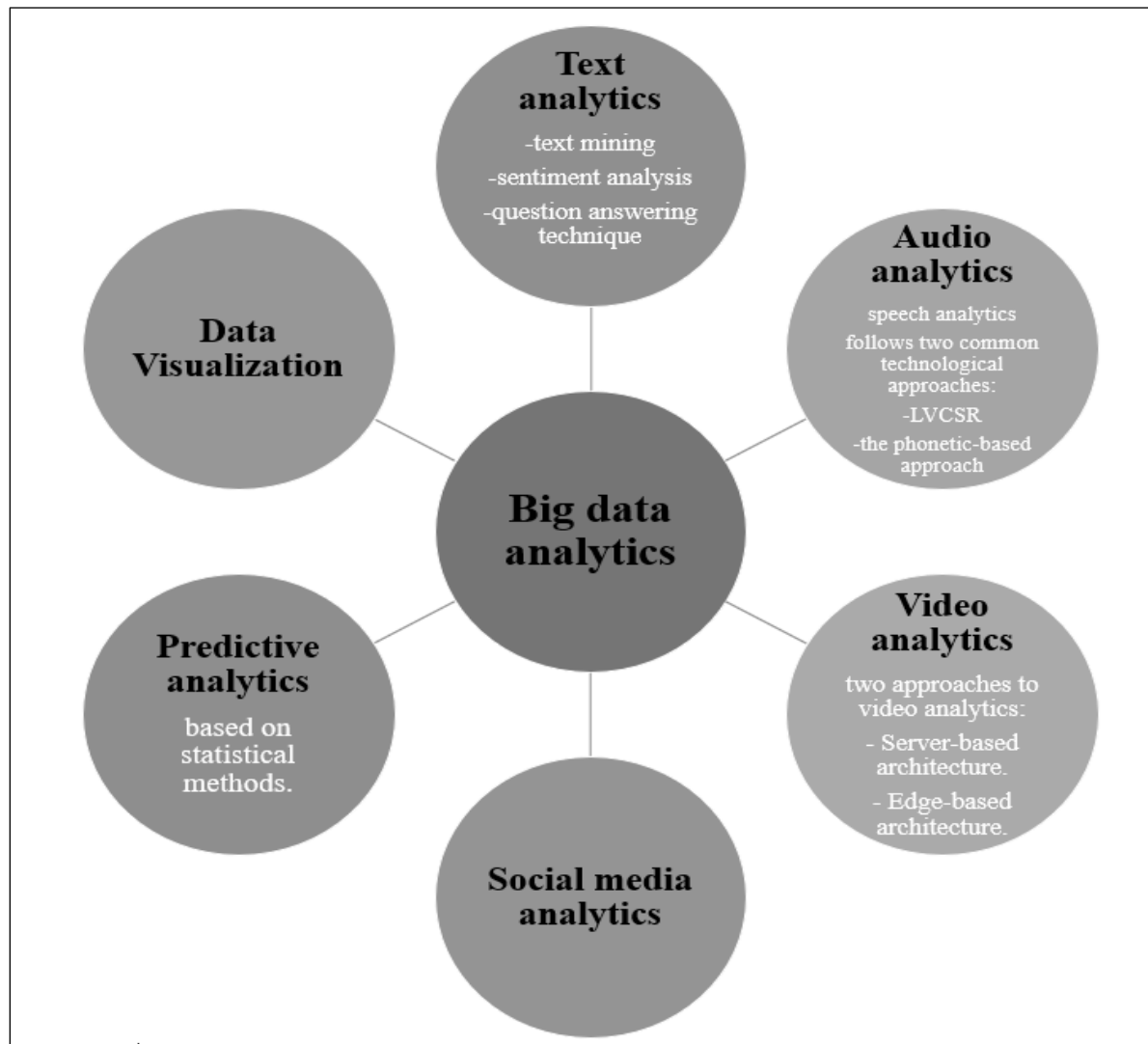


Figure 16: Common big data analytic methods.

The other types of big data analytics used for systematic review are presented by Grover and Kar (2017), and these include descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics, as shown in Figure 17.

Organisations and individual rend to use statistical models for predictive purposes, as most predictive models are built with statistical criteria. Artificial intelligence modelling is also

becoming more popular. Machine learning algorithms can combine statistical and artificial intelligence methods in order to analyse large amounts of data with high-performance (Watson, 2019).



Figure 17: Other types of big data analytics[2]

Descriptive analytics describes either what has happened or what is going to happen, while diagnostic analytics estimates the reason for something having happened, which requires techniques for discovering a problem's root causes. Predictive analytics attempts to determine the most likely future outcomes by applying statistical models (Waller and Fawcett, 2013), while prescriptive analytics explains and predicts the future and describes outcomes using tools such as

optimisation, simulation, business rules, algorithms, and machine learning (Banerjee et al., 2013; Grover and Kar, 2017).

---

[2] Based on (Grover and Kar, 2017)

The distribution of the research studies selected for systematic review across industry domains and analytic types in terms of big data analytics is shown in Figure 18.
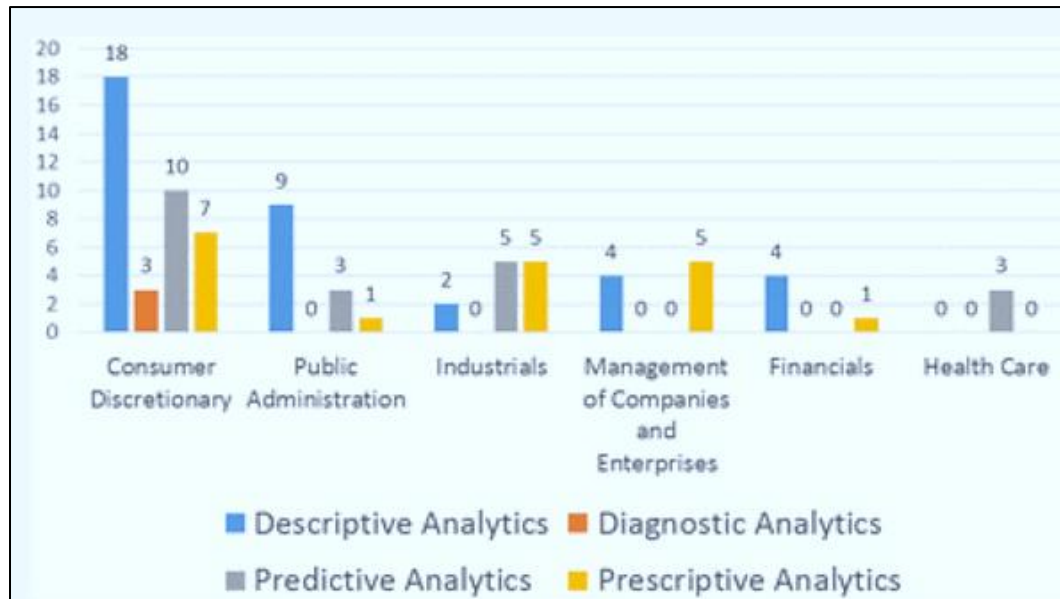


Figure 18: Distribution of research studies selected for systematic review across industry domains and analytics types, adopted from (Grover and Kar, 2017)

## 7.5. Big data platforms and tools

There are now multiple big data analytics tools and the study done by Oussous et al. (2018) showed the importance of carefully choosing the right tool for the circumstances. The choice is dependent on the "nature of datasets (i.e., volumes, streams, distribution), the complexity of analytical problems, algorithms and analytical solutions used, systems capabilities, security and privacy issues, the required performance and scalability in addition to the available budget" (ibid). Some of big data platforms and tools are shown in Figure 20.

- *Apache Mahout:* This is an open source machine learning software library that can be used for executing algorithms via MapReduce, a framework for processing large datasets (Eldawy and Mokbel, 2015). Mahout encompasses several Java libraries, ensuring efficiency of processing large datasets by allowing application of large-scale machine learning applications and algorithms. It provides an optimised algorithm in which Mahout converts machine learning tasks presented in Java into MapReduce jobs (Acharjya et al., 2016).

- *R:* This is a programming language often used for big data analysis, which offers relatively easy solutions to performing advanced analysis on large data sets via Hadoop. As compared to Mahout, in term of types and algorithms, R provides a more complete set of classification models; however, it is limited by its nature as an object-oriented programming language, which can cause problems with memory management compared to other solutions. In many

cases, use in combination with Mahout is thus recommended (Team, R.C., 2000), as R can be used to execute small data exploration while Hadoop/Jaql executes the larger operations.

- *Alteryx:* This tool offers data blending and an advanced analytics platform where analysts can merge internal business processes, third-party tools, and cloud data centres. Also, it allows data analytics utilizing some tools in a single workflow (ur Rehman et al., 2016).

- *Google Cloud Platform (GCP)* is one of the leaders among cloud Application Programming Interfaces (APIs). Despite the fact that it was established a few years ago, GCP has realized a significant growth since it suits the public cloud services that are based on massive, solid infrastructures. It gives the developer the ability to build a range of programs starting from simple websites to complex world-wide distributed applications. GCP platform contains a set of physical assets (e.g., computers and hard disk drives) and virtual resources (e.g., virtual machines, a.k.a. VMs) hosted in Google's data centres around the globe (Challita et al., 2018 ).

- *H2O* is an open source framework offering parallel processing, analytics, math, and machine learning libraries beside data pre-processing and assessment tools. Furthermore, it offers a web-based user interface that eases its use by analysts and statisticians who have limited programming backgrounds. It also provides support for Java, R, Python, and Scala (Landset et al., 2015).

- *MicroStrategy* provides an integrated big data analytics platform where the data is stored in Hadoop clusters and the users are given permission to access the desktop computer and mobile devices. This tool offers real-time visualization and interactions to implement fast decisions (ur Rehman et al., 2016).

- *RapidMiner:* is a programming-free data analysis platform. It provides the user with the ability to "design data analysis processes in a plug-and-play fashion by wiring operators". It allows importing operators for various data formats (e.g., Excel, CSV, XML). It prepares a set of operators for massive datasets with further attributes from open data sources which give an advantage of a better predictive and descriptive models (Ristoski et al., 2015).

- *Datameer:* Datameer Analytics Solution (DAS) is a business integration platform for Hadoop. It contains data source integration, "an analytics mechanism with a spreadsheet interface", designed with analytic functions and visualization to help business users in reports, charts and dashboards. Datameer can bring data from both structured such as Oracle, IBM DB2, and unstructured sources such as Twitter, Facebook, LinkedIn or e-mails (Di Martino et al., 2014).

- *Microsoft:* Microsoft platform provides predictive analytics capability called SSAS and integrated in the SQL Server. This platform offers "efficiency in Azure's cloud data source's integration and deployments as a web service" also, the simplicity of utilizing for data scientists (ur Rehman et al., 2016).

Figure 19 is adopted from (Raghupathi, W. and Raghupathi, V., 2014) and shows 1) data sources; 2) the big data states that need to be processed and transformed; 3) big data tools and platforms wherein these decisions are made depending on the inputs, tool selection, and analytical models chosen; and 4) the big data analytics applications. Figure 20 shows the big data and AI Landscape in 2018 which is adopted from (Goncharov, 2019).
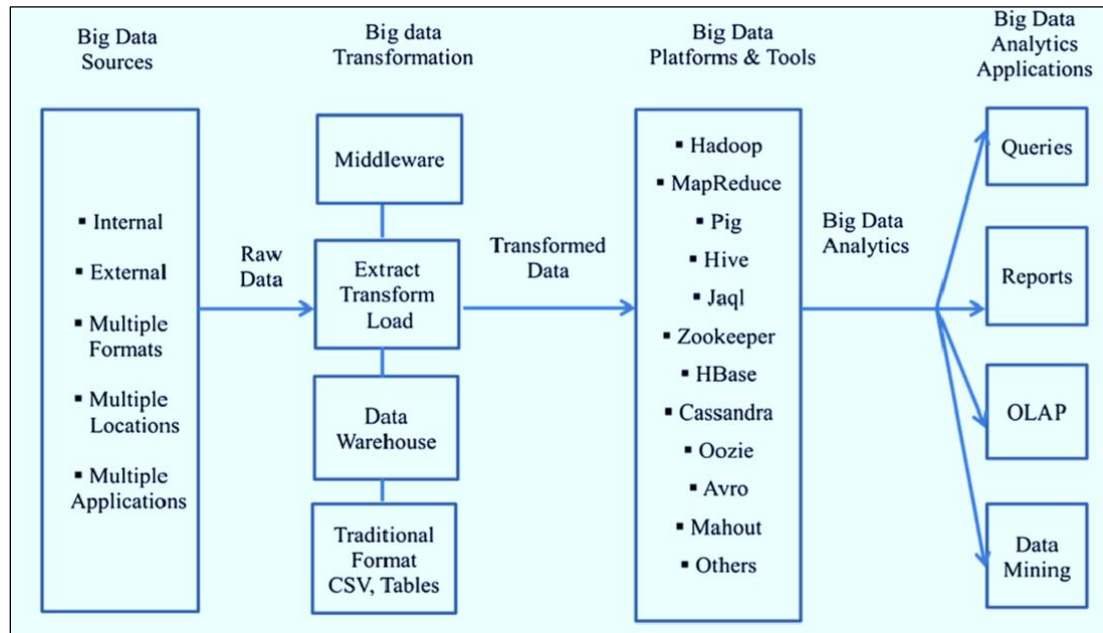


Figure 19: An applied conceptual architecture of data analytics, adopted from (Raghupathi and Raghupathi, 2014).
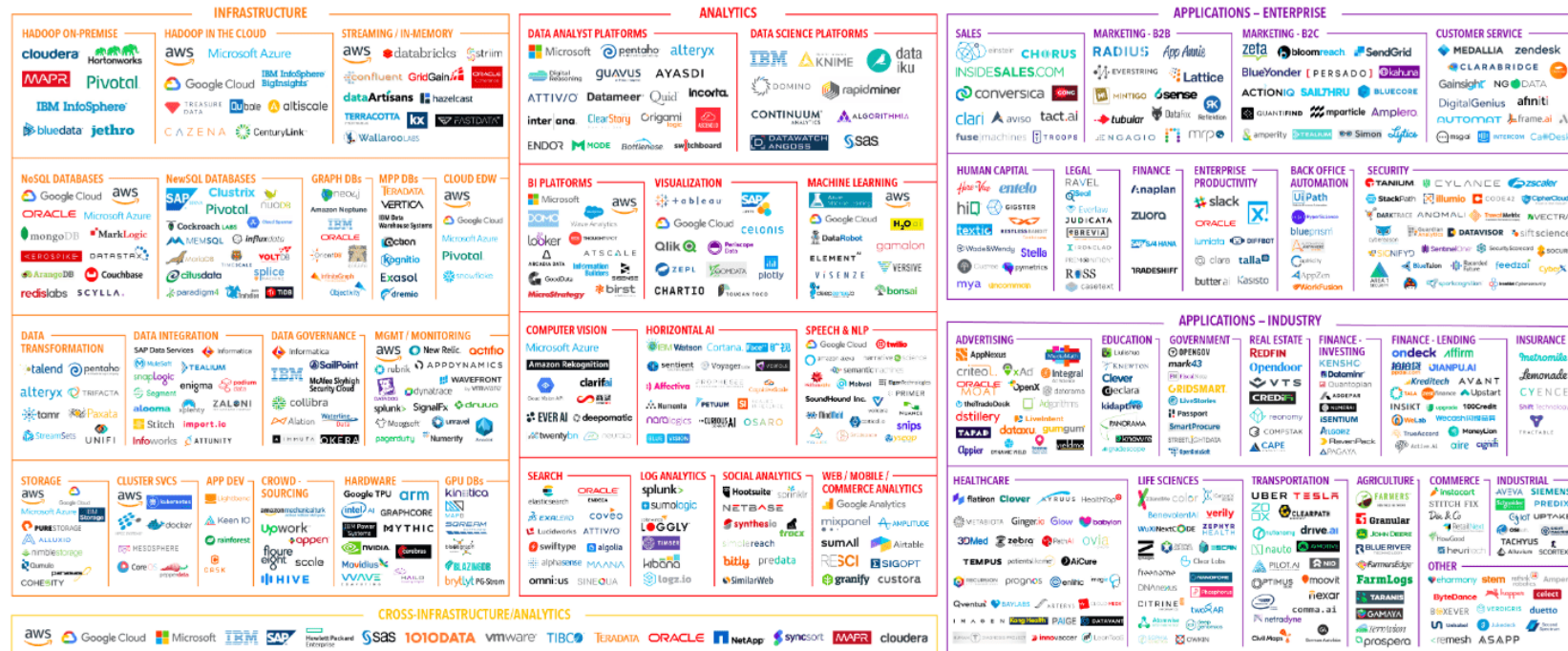
Figure 20: Big Data and AI Landscape in 2018, adopted from (Goncharov, 2019)

# 8. Big Data Analytics and Decision Making

LaValle et al. (2011) examined big data analytics capability (BDAC) and defined it as the ability to use big data in decision making. The study by Wixom et al. (2013) similarly focused on BDAC in terms of driving business value, recognising the value of BDAC in terms of strategy, data management, and human impact by conceptualising BDAC dimensions. That study showed that establishing BDAC leads to maximising business value by increasing decision speed and allowing big data usage to spread more widely through an enterprise.

Chen et al. (2012) showed that business analytics and related technologies help organisations develop better understanding of their own businesses and markets, while LaValle et al. (2011) showed that "top-performing organisations make decisions based on rigorous analysis at more than double the rate of lower performing organisations" (Sharma et al., 2014). Similarly, according to Kiron et al. (2014) BDAC is "the competence to provide business insights using data management, infrastructure (technology) and talent (personnel) capability to transform business into a competitive force".

Research by Akter et al. (2016) built a BDAC strategy based on previous studies which showed the importance of management and technology in the big data environment. This study proposed an integrated BDAC model and examined its impact. Elgendy (2013) further proposed a Big Data, Analytics, and Decisions (B-DAD) framework wherein big data analytics tools and methods are combined in the decision-making process.

In all of the models examined, the intelligence phase is the first phase of the decision-making process. In this phase

- data collected from internal and external sources are used to identify problems and opportunities;
- big data sources are clearly identified;
- further data are collected and gathered from different sources, being stored and sent to the user;
- after defining the data sources and types of the data required for the analysis, the data is processed through big data storage and management tools;
- organizing, preparing, and processing the big data is completed using either big data processing tools or a high-speed network using Extract, Transform, Load or Extract, Load, Transform (ETL/ELT) processes.

These phases are shown in detail in Figure 21.

The next phase is the design phase, in which developing and analysing the possible courses of actions is done by means of conceptualization or developing a problem representative model. In this phase, the framework divided into model planning, data analytics, and analysis. Where a data analytics model is selected, this is planned, applied, and then analysed.

The third phase of decision making is the choice phase, and in this phase, the proposed solution impact is evaluated. The final phase in decision making is the implementation phase; in this phase, the proposed solution is implemented (Elgendy, N. and Elragal, A., 2016).
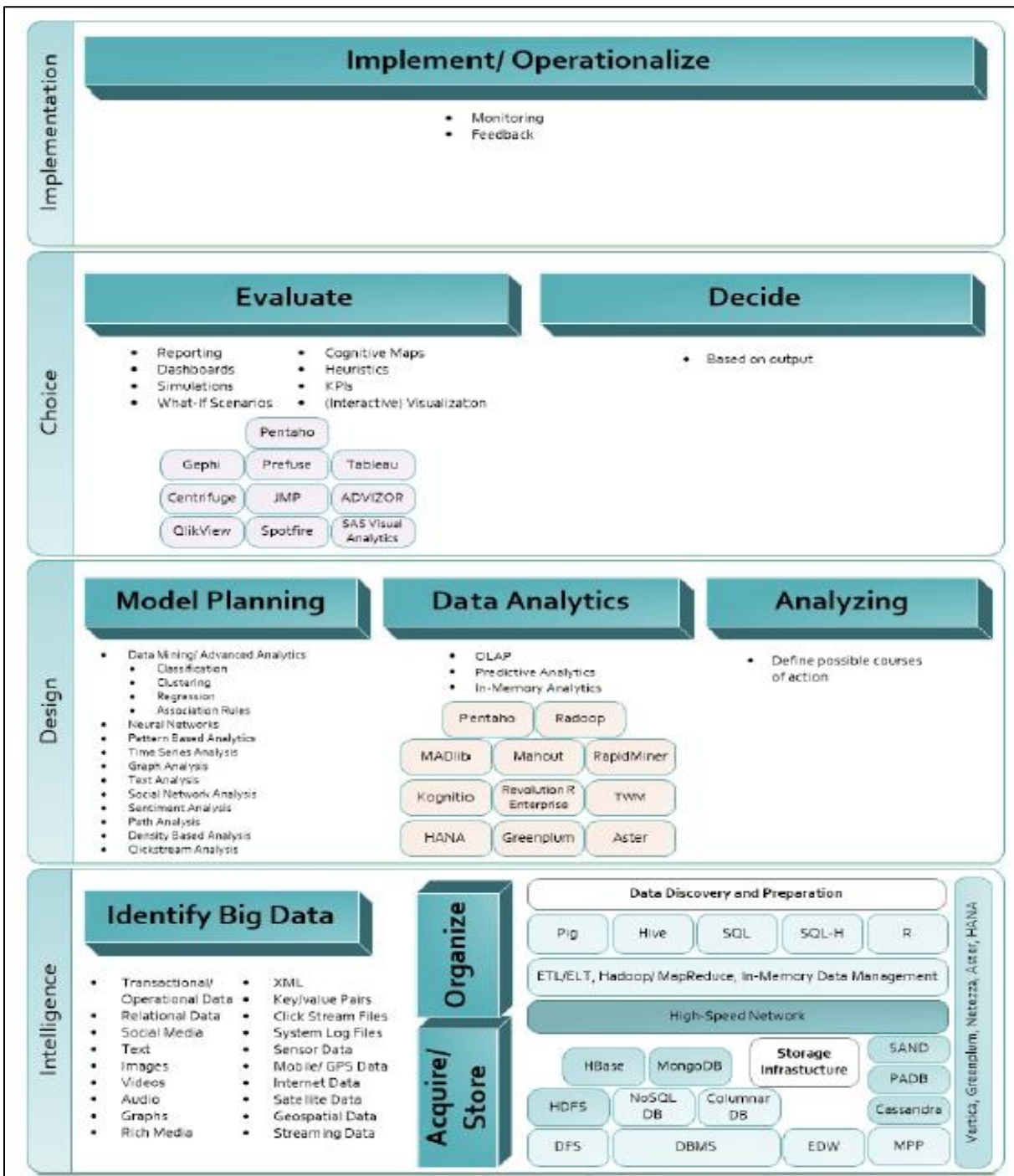
Figure 21: B-DAD framework, adopted from (Elgendy and Elragal, 2016).

Elgendy and Elragal (2016) shows the decision-making process and how big data analytics can be integrated into it. Using the methodology of design science, the B-DAD can be used to map big data tools and analytics to various decision-making phases. As a result, the added value gained by integrating big data analytics into the decision-making process can be identified (Elgendy and Elragal, 2014; Elgendy and Elragal, 2016).

Despite certain challenges, decision making is supported by advanced technologies and tools in each phase of processing and applying big data, and the use of big data now plays an important role in many decisions making and forecasting domains such as healthcare, retail, tourism, marketing, the financial sector, and transportation (Elgendy and Elragal, 2014).

Big data use requires decision support, however. The decision maker must identify the values required and focus on finding methodologies, technologies, and tools that allow them to select the best decision; this process thus relies on the assumption that the decision maker is sensible and reasonable (Wang et al., 2016).

Generally, decision making occurs at the stage of each big data procedure, including data storage, data cleaning, data analysis, data visualisation, and prediction. However, it is sometimes difficult to achieve a suitable solution for each procedure, and many technologies and techniques can be used for decision making in big data work. Some decision making requires input from many disciplines, including data mining, statistics, machine learning, visualisation, and social network analysis. Specific big data tools come in three classifications types: batch processing, stream processing, and hybrid processing tools (Wang, et al., 2016). The relationship between decision science and big data is clarified in Figure 22.
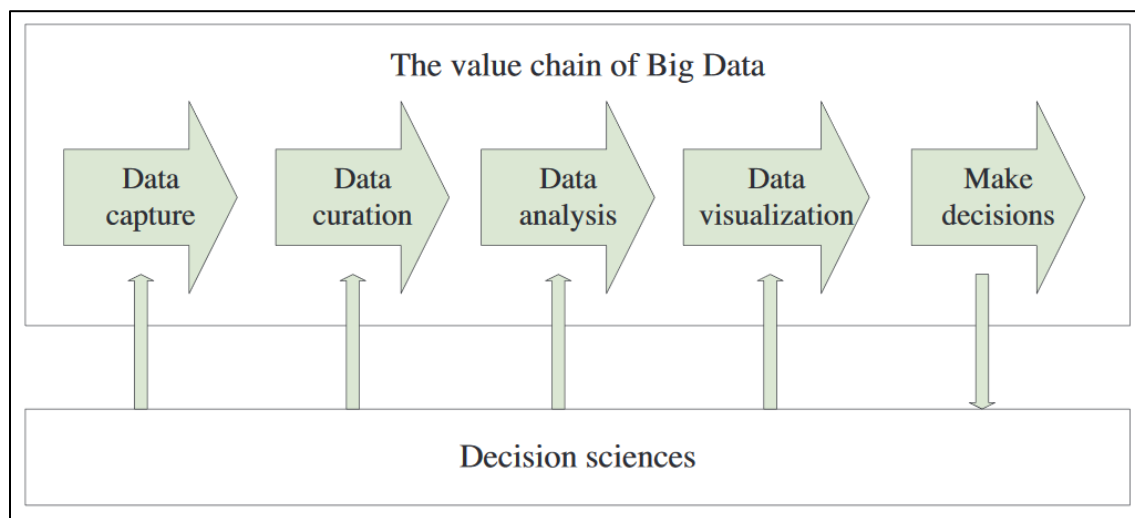


Figure 22: The relation between big data and decision sciences, adopted from (Wang et al., 2016)

# 9. Big data analytics challenges

Many studies have focused on the use of analytics techniques such as data mining, visualisation, statistical analysis, and machine learning; however, there is a need to develop new analytic approaches in order to handle big data challenges such as the time required for processing when the volume of the data is very large (Oussous et al., 2018). Oussous et al. thus presented the difficulties in applying current analytical solutions, including machine learning, deep learning, incremental approaches, and granular computing.

Chen et al. (2014) similarly addressed big data applications, opportunities, and challenges, and examined several techniques to handle big data challenges, such as cloud computing and quantum computing, to examine their efficacy. Wang.et al. (2016) presented a big data overview that included four categories: 1) concepts, big data characteristics, and processing paradigms; (2) state-of-the-art techniques for decision making in big data; (3) decision making applications of big data in social science; and (4) big data's current challenges and future directions.

The work of Ali et al. (2016) explained big data's potential and applications. It presented big data techniques and offered some background to big data analytical approaches. The study highlighted several big data technical challenges such as crowdsourcing, bias and polarization, technology usage, and scaling. New technologies and services such as cloud computing and hardware price reductions have also increased the information rates available from the Internet, representing a big challenge to the data analytics community.

**The main challenges of using big data, which need to be resolved before it can be used effectively, include**

## 9.1. Data Security issues

In public affairs, privacy, internet access disparities, and legal and security issues are key concerns, and managers and policymakers in these areas should work to overcome these limitations. Public managers and policymakers are also, however, generally working under the restrictions of a limited budget, multiple constituencies, and short time frames for extracting knowledge big data (Mergel, Rethemeyer, and Isett, 2016; Grover and Kar, 2017).

Watson (2019) presented some security issues with big data and gave some suggestions for avoiding big data security risks. The security concern inherent in big data include the fact that big data comes from many different sources, some of which may have weak security as well as a variety of formats and large volumes. Any security breaches may thus affect multiple companies and result in financial losses, and thus, appropriate actions should be taken to reduce such big data security risks.

Data sources should be monitored by organisations, with end-to-end encryption used to prevent anyone from accessing the data in transit. Companies should also check their cloud providers, as many cloud providers do not encrypt the data due to the quantity of data transferred at any given time, as encryption/decryption slows down the flow of data.

Big data is defined by the 5V's, and these charecteristics, especially the volume aspect, mean that it cannot be processed with traditional data analytic techniques. Large amounts of complex data need time for analysis. Therefore, big data faces intrusion detection challenges, as the system busy times are extended. Although many security monitoring systems have been developed to improve data security, intrusion detection is still challenging, even for isolated systems. The issues include how to store large quantities of data safely, how to maintain security, and how to track data that flows quickly from different sources. Solution to these challenges include taking a more comprehensive approach to monitoring the data that comes from different sources in order to develop better situational awareness of the threats in cyberspace. This helps minimise false alarms and maximise intrusion detection. The big data challenges for intrusion detection can also be addressed by using traditional computing storage platforms such as Hadoop, an open source distributed storage platform used for storing large amounts of data that flows quickly (Suthaharan, 2014; Zuech et al., 2015).

Suthaharan (2014) proposed using big data technologies such as Hadoop to address intrusion detection issues, and in addition, he proposed the 3Cs, Cardinality, Continuity, and Complexity, for use in developing mathematical and statistical tools. Here, Cardinality refers to the number of records, Continuity refers to the data's continuous growth over time, and Complexity refers to the data type variety (Zuech et al., 2015). Learning from the data is executed by the User Interaction and Learning System (UILS) which gives the user permissions to interact with the system and control the storage requirements. The network traffic is captured by a Network Traffic Recording System (NTRS), which stores it locally in the Hadoop Distributed File System (HDFS) or the Cloud Computing Storage System (CCSS).

Based on Hadoop technology, Cheon and Choe (2013) proposed an intrusion detection system architecture. They added additional Hadoop-based nodes to those used in analyses, varying from zero to eight replays of files; they then evaluated their efficiency. They found that the efficiency of performance was increased, and that the system spent less time processing the datasets (Zuech, et al., 2015).

Blazquez and Domenech (2018) proposed a big data architecture based on an analysis of economic and social behaviour in the digital era. This study addressed the issues raised by several economic and social topics by presenting multiple data sources and proposing a taxonomy for classifying these depending on the purpose of the agent used to generate the data.

Lan, et al. (2010) used data fusion across diverse heterogeneous sources to improve intrusion detection. As a result, they found that traditional security products such as firewalls, intrusion detection systems, and security scanners do not work together, and thus protecting networks with minimal network knowledge. The authors suggested utilising a form of data fusion known as Dempster-Shafer (D-S) evidence theory in order to better understand heterogeneous sources (Zuech et al., 2015). D-S evidence theory is a common data fusion technique used by researchers within the Intrusion Detection domain, which applies probabilistic techniques to monitor the system.

## 9.2.   Data privacy issues

Gathering data from users might lead to privacy challenges where the gathering process may cause the data context and semantics to be modified, leading to faulty and inefficient policies (Ali et al., 2016).

Lv et al. (2017) showed that one potential problem in big data is data security and privacy, as big data applications often contain sensitive information such as medical records and banking transactions which is not appropriate for normal data transmission protocols. Data security and privacy must thus be considered before the adoption of any protocol for sharing information. The challenges caused by the inclusion of sensitive information and the requirements for access control or certification are generally well known; however, secured certification mechanisms remain challenging to implement, and anonymisation approaches decrease data confidence (Wang et al., 2016).

Big data privacy contains two aspects: the first is that the personal data privacy should be protected during data gaining such as personal interests, habits, and body properties, etc. of users who do not aware or easy to gain information from them. The second aspect is that the personal privacy data might discharge during storage, transmission, and usage, even if it gained with the user permission. For example, currently, Facebook is considered as a big data company with the most social networking service SNS data. Even though, some researchers gained data from public pages of Facebook users who did not change their privacy setting through an information-gaining tool (Chen, M., Mao, S. and Liu, Y., 2014).

## 9.3.   Data storage, data capture and quality of data

Capturing and storing data is not easy, especially as data sets are increasingly growing in size and complexity. There is often not enough space to store such big data, and many sectors and fields such as the financial and medical areas are forced to delete data. Capturing and creating valuable data is only done at a high cost (Chen et al., 2014).

Oussous et al. (2018) discussed big data characteristics in terms of it being processed by many analytics tools and visualisations. The big data platforms layer and its components and technologies were explained. In term of capabilities, different technologies were compared, and big data systems categorised according to their features and the services provided to users. They showed that big data use still has many technical issues that need to be studied. They also presented big data computing systems' challenges, examining difficulties on various different levels "including data capture, storage, searching, sharing, analysis, management and visualisation". This included examining security and privacy issues. The size of big data is increasing exponentially, and this makes the current technology unable to handle such big datasets.

Modern big data challenges thus include big data management where the challenge lies in collecting, integrating, and storing data with minimal requirements (hardware and software). Big data management also requires cleaning data for reliability then aggregating data from different sources before encoding the data for security and privacy purposes (Chen et al., 2014; Najafabadi et al., 2015).

Big data cleaning challenge lies in the data's complexity: velocity, volume, and variety (Khan et al., 2014). Big data aggregation challenges are involved synchronising outside data sources and

distributed big data platforms (including applications, repositories, sensors, networks, etc.) into a cohesive system (Oussous et al., 2018). Also, in imbalanced system capacities, the challenge lies in the computer architecture and capacity, as imbalanced system capacities might affect big data application performance (Chen et al. 2014).

Furthermore, the challenge in imbalanced big data is how to classify imbalanced datasets, as "classical learning techniques are not adapted to imbalanced data sets" (Oussous et al., 2018). Big data analytics challenges lie in the complex data analysis required to understand the relationships among data features. Some data analysis requires real-time analysis, such as navigation, social networks, finance, biomedicine, astronomy, and intelligent transport systems, while other analyses require accurate result but not necessarily the same levels of speed. The challenge with big data analysis mainly arises due to the 5V's and their effects on dataset performance (Qiu et al., 2016).

One solution for the storage challenge is using Hadoop (Apache platform), which is an open-source distributed data processing platform with the power to process extremely large amounts of data. It does this by dividing the data into smaller parts then specifying some parts of the datasets to separate servers (nodes) (Raghupathi, W. and Raghupathi, V., 2014)

## 9.4.  Challenges in data analysis and visualisation

Data analysis challenges arise from data complexity, which in turn comes from the data's complex types and structures. Standard data analysis techniques face difficulties in handling such big data as it is more difficult to understand the distribution laws of big data (Wang et al., 2016).

Big data visualisation challenges come from the data's high dimensions and size. The main goal of data visualisation is to explain knowledge effectively by using diagrams; in order to transfer information easily to the user, hidden knowledge in the complex and large-scale data sets is rendered visible. For more accurate data analysis, however, abstracting information in schematic formats, including features or variables representing units of information is valuable. Nevertheless, because of the large size and high dimensions of big data, it can also be difficult to manage data visualisation in big data applications (Chen C.P et al., 2014; Wang et al., 2016).

Oussous et al. (2018) discussed the value of data mining methods in several domains. Data mining methods are significant when used for discovering patterns and extracting value hidden across massive datasets. Applying traditional data mining techniques, such as association mining, clustering, and classification, to big data is, however, inefficient and inaccurate. The volume, speed, and variability of such data makes it unsuitable for long-term storage and analysis. Several data mining methods have thus been adapted to contain detecting techniques to take the data environment into account.

Günther et al. (2017) noted that some empirical studies and some old ideas have characterised much big data value realisation; the study examined six debates identified in terms of "how organisations realize social and economic value from big data that require attention from future research". Two additional features of big data were also identified, portability and interconnectivity, and those features were utilised to show the effect of big data value realisation in organisations. At the end of the study, the authors argued that the continuous interactions between work practices, organisational models, and stakeholder interests prompted calls for empirical research on cross-level interactions and alignment results from realising big data value, as shown in Figure 23. Several suggestions for further study were also presented:

- Work on improving big data business models and innovative approaches, such as the development of a four-stage big data maturity model, allowing organisations to reach functional excellence despite the ability to develop business model transformation occurring only in the last stage;
- Relevant systems, such as Hadoop, which have the ability to work with both big data and more traditional data being identified for various cases (Ekbia et al., 2015);
- Examining the dependency on size of organisations that can adopt big data (Ekbia et al., 2015);
- Examining appropriate organisational models for creating and appropriating value from big data (Karpovsky and Galliers, 2015; Ekbia et al., 2015);
- Further investigation of two key issues: 1) controlled and open big data access when data analytics can be considered a competitive advantage, as organisations may be opposed to exchanging data with perceived competitors (Jagadish et al., 2014); and 2) minimising and countering the social risks of big data value realisation (Clarke, 2016).



Figure 23: Summary of debates related to big data value realisation, adopted from (Günther et al., 2017).

# 10.      Big data analytics applications

A recent survey (Nashua, 2017) highlighted the growth in the use of big data analytics in companies. The study examined companies' use of big data compared to the previous year in 2015, 2016, and 2017. The results indicated that over 50 percent of organisations were using big data by 2017 (Watson, 2019), as shown in Figure 24.
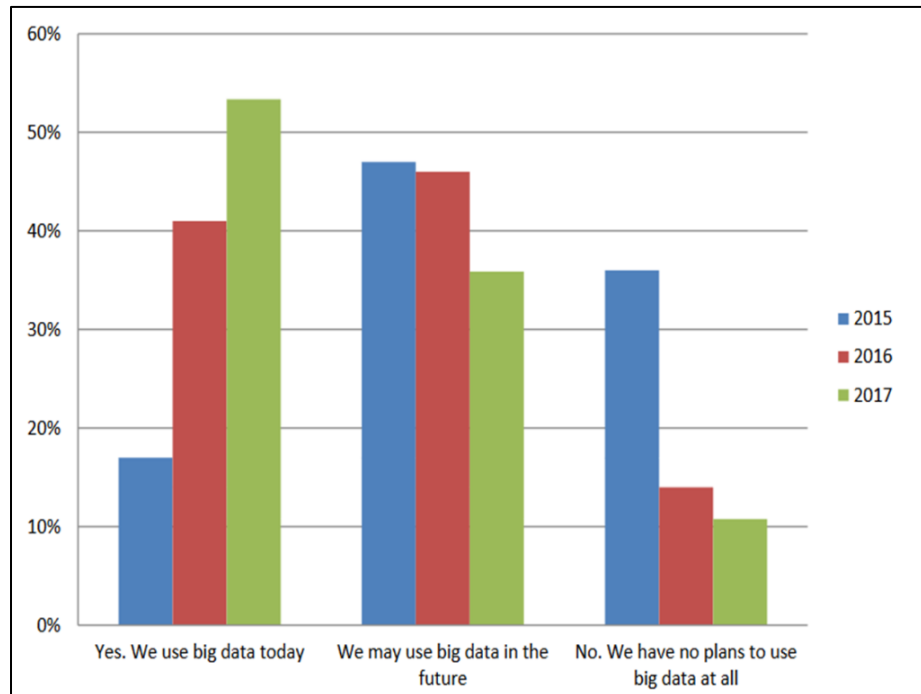


Figure 24: Adoption of Big Data from 2015 to 2017[3]

The potential key resource of many organisations' business models is thus big data where such "business models are reflections of the realized strategy" (Casadesus-Masanell et al., 2010; Günther et al., 2017). Business models represent the ability of the organisation to create and appropriate value. Organisations must rethink their use of big data in relation to their business models, using analytics to develop access to new data sources and techniques to improve efficiency and effectiveness (Woerner et al., 2015).

Grover et al. (2018) argued that to achieve strategic business value from big data, significant investment in both infrastructure and analytic technologies are required to enable skilled analysis and strategic positioning. Businesses thus need to access cutting edge tools and hire data-savvy people who understand the relevant technologies.

Watson (2014) wrote a paper about big data analytics which was published by the association for information systems (CAIS). That paper showed the advances in technologies, applications, and the impact of big data analytics at that time. In 2019, the same researcher (Watson, 2019) highlighted several important recent developments in big data analytics including

---

[3] Adopted from (Dresner Advisory Services, 2017; Nashua, 2017; Watson, 2019

➢ Continued adoption of the big data analytics,
➢ Growth in the number of big data applications,
➢ Development of the Hadoop ecosystem technology,
➢ Data lakes,
➢ Advanced analytics models, and
➢ Algorithmic transparency principles.

Big data analytics has the potential to be applied to demand forecasting, analysing potential needs based on previous work used to classify analytics techniques (Hofmann et al., 2018).

Big data analytics have also been applied in many areas, serving different sectors.

**Some big data analytics applications:**

## 10.1.    Healthcare

Research by Manyika (2011) showed that big data might help in reducing waste and improving efficiency in clinical operations, research and development, and public health by means of

- statistical tools and algorithms;
- predictive modelling to produce new drugs and devices more quickly;
- analysing records of diseases to improve epidemiology (Elgendy, N. and Elragal, A., 2014);
- allowing faster development of vaccines; and
- identifying the data relevant to provide services and prevent crises.

Raghupathi and Raghupathi (2014) described big data analytics in healthcare and identified several remaining challenges; big data analytics has the power to develop care, save lives, and minimise the costs, using the recent data explosion to extract insights in order to allow healthcare providers to make better decisions. The potential benefits gained from using big data in healthcare include, but are not limited to, discovering diseases quickly, thus making treatment easier and more effective; identifying healthcare fraud quickly in order to manage specific individuals; and improving population health.

Furthermore, Zhong et al. (2016) presented big data applications in healthcare and showed how big data can be embedded into daily life to offer the ability to examine experiences of illness and healthcare. Big data analytics thus have a large impact on the healthcare sector, reducing operational costs and improving patients' quality of life (Elgendy and Elragal, 2014; Wamba et al., 2017).

## 10.2.    Banking

Handling massive volumes of data of many different types is not easy. Big data analytics offers potential benefits to industries such as banking by allowing analysis of customer log files and the handling of customer interactions. Combining structured and unstructured data types in this way can give companies a better view of both their customers and operations (Davenport, T.H. and Dyché, J., 2013).

Analytics offers banks the ability to segment customers depending on their risk profiles, credit usage, and similar markers, offering products tailored to their needs and ability to handle money. Analytics are utilized throughout the industry, such as in retail banking operations, where every customer transaction is tracked and matched to the customer. Banks have adopted new requirements for data science (analytics) in a wise manner to avoid reduced performance, and data science (analytics) offers them a resolution to next generation business problems (Banerjee et al., 2013).

Due to the massive number of transactions and activities in financial institutions such as banks, big data development is inevitable, and this directly impacts on the management of scarce resources by individuals, groups, and organisations. Big data analytics is thus used by the financial service sector to predict client behaviours and to gain advantages based on understanding customers and employees (Zhong et al., 2016; Breed and Verster, 2019; Rana, 2019).

Big data analytics is used widely in the field: McKinsey & Company, a global management consulting company, uses big data analytics to develop its services and to improve service performance. This company uses big data analytics to analyse consumer behaviours to upgrade services and to forecast customer behaviours, to allow fraud detection, and to determine financial risk assessments (Zhong et al., 2016).

## 10.3. Retail

Big data analytics has a massive impact on retail industries, improving the customer experience and reducing fraud (Wamba et al., 2017).

The retail sector is of major importance in modern society, as almost everyone nowadays musy buy their basic needs. Predicting demand for items allows retailers to offer better services to customers (Singh et al., 2015; Lekhwar et al., 2019), and retailers can use customers' billing data to gather information for business intelligence. A Hadoop distributed file system (HDFS) tool is using to store, process, and analyse such data to allow the extraction of more information (Singh et al., 2015).

Big data analytics provides these organisations with more information on market decisions and help in segmenting customer based on their characteristics. Social media analytics can also be used to inform companies about what their customers prefer. Applying sentiment analysis to such data provides the organisation with early warnings when the customer turns to different products, allowing action to be taken by the organisation (Elgendy, N. and Elragal, A., 2014).

Organisations have used segmentation of customers for many years, but this is now assisted by complex big data techniques such as real-time micro-segmentation which offers better-targeted advertising (Manyika et al., 2011; Elgendy and Elragal, 2014). Organisations can also gain better targets for social marketing by understanding customer behaviors and predicting market sentiment trends (Russom, 2011; Elgendy and Elragal, 2014).

Retailers are thus using data analytics in order to address new challenges and find opportunities based on increases in market expectations, competition, and volatility. In many companies, additional accuracy, clarity, and insight can be provided by the adoption of data analytics techniques, and such intelligence can be extended toward industry supply chains (Hofmann et al., 2018).

## 10.4.     Telecommunications

Big data analytics can improve the quality of management in telecommunications by making use of real-time data analyses and monitoring machine logs. Predictive analytics can also be used to minimise performance variability and to prevent quality issues by providing early warning alerts (Elgendy, N. and Elragal, A., 2014).

Big data analytics platforms used in the telecommunication field face the major challenge of how to store and process big data; traditional analysis techniques are too expensive in many cases. Big data techniques such as Hadoop can help in reducing the storage costs, particularly where storage modules such as the Hadoop Distributed File System (HDFS) and computation modules such as MapReduce are included (Çelebi, 2013).

Big data analytics has the power to extract more information than traditional data analytics, which can help in improving mobile cellular networks. Such mobile cellular networks generate and carry massive amounts of data such as calls and mobile application activities that consist of both structured and unstructured data types. Traditional data analytics deals only with structured data, and thus it is almost impossible to handle that data with traditional data analytics (He et al., 2016).

# 11.     Implications of research

Big data has the power to change research and education. Improving the students' results by refining the student's performance during courses and understanding their behaviours can be done using big data analytics. Also, big data analytics gives students the ability to matching their interests to the available programs; thus, can choose the best school or educational program. On the other hand, big data helps teachers to understand the knowledge level of each student and tune the teaching technique with the most valuable effect on an individual basis. Consequently, using big data for guiding instructions in academia has a significant role in enhancing the educational services by allowing students to access online instructors and communities at low-cost content.

Technologies for controlling and analysing data are broadly available. Companies take advantage of capturing the data to support accurate and stable business experimentation that direct decision makers. It might also evaluate outputs, business models, and restoration in customer experience. Trends allow directing a revolutionary transformation in research, invention, and business marketing.  Some firms like Amazon, Google, and eBay analyse elements that control performance to determine factors that raise sales income and track the activity of users (Bughin et al., 2010).

Big data has an essential impact on financial institutions as they keep modifying their methods for segment credit card customers. Companies such as Brick and Mortar are utilizing big data to test the ability to guide customer data by collecting transactional information from millions of customers, then use the collected information in analysing new opportunities such as optimizing the most effective promotions. Other companies use data mining to gather information from social media. Southwest Airlines, Ford motor, and Pepsico analyse consumer posts on social media like Facebook and Twitter to standard the immediate influence on a movement and track the consumer opinions about their products.

Big data has an impact on many aspects of society resulting in societal benefits. On the medical system, for example, it gives benefits of saving lives, as using big data enables the doctors to decide which medication is the best to a patient. The patient does not need to wait long times, get severe reactions, or dying from using medicines that do not fit with their case.

# 12.        Conclusion and Future Research

The purpose of this study was to offer a literature review on the topic of big data analytics. This began with the presentation of a general background to the topic, including big data definitions and characteristics, followed by a review of big data analytics tools and methods.

This thesis presented data analysis techniques characterised in four sections: supervised, unsupervised, semi-supervised, and reinforcement learning. Some analytics techniques were also presented, such as clustering, correlation, regression, and factor analytics, and some big data tools and platforms such as Hadoop, Apache Mahout, and R were explained in relation to these. Big data storage, management, and analytics processing were also discussed, and some emergent advanced data analytics techniques further examined.

Various big data tools, methods, and technologies have been discussed in this research, offering readers examples of the necessary technologies, and prompting developers to come up with ideas about how to provide additional big data analytics solutions to help in decision making.

Big data analytics has been applied in various areas, serving many different sectors. Big data analytics has the potential to improve care, save lives, and reduce costs in the healthcare sector. It also benefits industries such as financial institutions by allowing analysis of customer log files to help develop a better understanding of customer needs. The retail sector has a significant impact on society and using big data analytics in this sector can again help managers to better understand people's needs, thus prompting the development of better services. Big data analytics are also used in the telecommunications sector, where they help in monitoring machine logs and addressing quality issues.

Some big data analytics challenges were discussed in this work, particularly with regard to security and privacy. Some examples of how big data analytics can be used to handle issues such as intrusion detection and big data characteristics such as size, velocity, variety, value and external sources were also given. Finally, some real-world big data analytics applications were introduced.

Big data is a significant area which offers many potential benefits and innovations. It is a remarkable domain with a promising future, if approached correctly. The difficulty with big data comes mainly from its size, which requires proper storage, management, integration, cleansing, processing, and analysis. The sheer volume, velocity, speed, and variety of data increases the difficulty of dealing with it in terms of traditional data management, creating a need to study and explore new analytics methods which might help in overcoming such difficulties to promote the positive role of big data analytics to as many sectors as possible. Future research could thus usefully focus on big data analytics challenges with regard to security and privacy issues, based

on big data's weakness in coming from many different sources; a focus on cloud providers and security breaches which affect multiple companies would also be advised.

# 13.     References

Acharjya, D.P. and Ahmed, K., 2016. A survey on big data analytics: challenges, open research issues and tools. *International Journal of Advanced Computer Science and Applications,* pp. 511-518.

Addo-Tenkorang, R. and Helo, P.T., 2016. Big data applications in operations/supply-chain management: A literature review. *Computers & Industrial Engineering journal,* Volume 101, pp. 528-543.

Agarwal, R. and Dhar, V., 2014. Big data, data science, and analytics: The opportunity and challenge for IS research. *IS research Journal.*

Akter, S., Wamba, S.F., Gunasekaran, A., Dubey, R. and Childe, S.J., 2016. How to improve firm performance using big data analytics capability and business strategy alignment?. *International Journal of Production Economics,* pp. 113-131.

Al-Barashdi, H. and Al-Karousi, R., 2019. Big Data in academic libraries: literature review and future research directions.. *Journal of Information Studies and Technology,* p. 13.

Ali, A., Qadir, J., ur Rasool, R., Sathiaseelan, A., Zwitter, A. and Crowcroft, J., 2016. Big data for development: applications and techniques.. *Big Data Analytics journal,* Volume 1, p. 2.

Arunachalam, D., Kumar, N. and Kawalek, J.P., 2018. Arunachalam, D., Kumar, N. anUnderstanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice.. *Transportation Research Part E: Logistics and Transportation Review journal,* Volume 114, pp. 416--436.

Bakshi, K., 2012. *Considerations for big data: Architecture and approach conference.* s.l., IEEE, pp. (1-7).

Banerjee, A., Bandyopadhyay, T. and Acharya, P., 2013. Data analytics: Hyped up aspirations or true potential?. *Vikalpa journal,* Volume 38, pp. 1-12.

Blazquez, D. and Domenech, J., 2018. Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change journal,* Volume 130, pp. 99--113.

Boyd-Graber, J., Mimno, D. and Newman, D., 2014. Care and feeding of topic models: Problems, diagnostics, and improvements.. *Handbook of mixed membership models and their applications Journal ,* Volume 225255.

Bradlow, E.T., Gangwar, M., Kopalle, P. and Voleti, S., 2017. The role of big data and predictive analytics in retailing. *Journal of Retailing,* pp. 79-95.

Breed, D.G. and Verster, T., 2019. An empirical investigation of alternative semi-supervised segmentation methodologies. *South African Journal of Science,* Volume 115, pp. pp.92-98.

Bughin, J., Chui, M. and Manyika, J., 2010. Clouds, big data, and smart assets: Ten tech-enabled business trends to watch.. *McKinsey quarterly,* Volume 56, pp. 75-86.

Casadesus-Masanell, R. and Ricart, J.E, 2010. From strategy to business models and onto tactics. *Long range planning journal,* Volume 43, pp. 195-215.

Çelebi, Ö.F., Zeydan, E., Kurt, Ö.F., Dedeoğlu, Ö., İieri, Ö., AykutSungur, B., Akan, A. and Ergüt, S., 2013. On use of big data for enhancing network coverage analysis. *In ICT 2013 IEEE journal,* pp. 1-5.

Challita, S., Zalila, F., Gourdin, C. and Merle, P., 2018 . *A Precise Model for Google Cloud Platform..* s.l., IEEE, pp. 177-183.

Chen G, Guo X., 2016. Big data commerce. *Inf Manag. Journal.*

Chen, C.P and Zhang, Chun-Yang, 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences Journal,* Volume 275, p. 314–347.

Chen, H., Chiang, R.H. and Storey, V.C., 2012. Business intelligence and analytics: from big data to big impact.. *MIS quarterly,* pp. 1165-1188.

Chen, M., Mao, S. and Liu, Y., 2014. Big data: A survey.. *Journal of mobile networks and applications,* Volume 19, pp. 171-209.

Chen, M., Mao, S., Zhang, Y. and Leung, V.C., 2014. *Big data: related technologies, challenges and future prospects.* s.l.:Springer.

Cheon, J. and Choe, T.Y., 2013. Distributed processing of snort alert log using hadoop.. *International Journal of Engineering and Technology IJET,* Volume 5, pp. 2685-2690.

Clarke, R., 2016. Big data, big risks. *Information Systems Journal,* Volume 26, pp. 77-90.

Constantiou, I.D. and Kallinikos, J., 2015. New games, new rules: big data and the changing context of strategy. *Journal of Information Technology,* Volume 30, pp. 44-57.

Cronin, P., Ryan, F. and Coughlan, M., 2008. Undertaking a literature review: a step-by-step approach.. *British journal of nursing,* pp. 38-43.

Cui, Q., Gong, Z., Ni, W., Hou, Y., Chen, X., Tao, X. and Zhang, P., 2019. Stochastic Online Learning for Mobile Edge Computing: Learning from Changes.. *IEEE Communications Magazine journal,* Volume 57, pp. 63-69.

Data, D.B., 2012. A Practical Guide to Transforming the Business of Government.. *TechAmerica Foundation˝ s Federal Big Data Commission Journal.*

Davenport, T.H. and Dyché, J., 2013. Big data in big companies. *International Institute for Analytics.*

Di Martino, B., Aversa, R., Cretella, G., Esposito, A. and Kołodziej, J., 2014. Big data (lost) in the cloud.. *International Journal of Big Data Intelligence,* Volume 1, pp. 3-17.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V.R., Tsou, A., Weingart, S. and Sugimoto, C.R., 2015. Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology,* Volume 66, pp. 1523-1545.

Eldawy, A. and Mokbel, M.F., 2015. *Spatialhadoop: A mapreduce framework for spatial data.* s.l., IEEE.

Elgendy, N. and Elragal, A., 2014. *Big data analytics: a literature review paper.* s.l., Springer, cham, pp. 214-227.

Elgendy, N. and Elragal, A., 2016. *Big Data Analytics in Support of the Decision Making Process.* s.l., Elsevier.

Elgendy, N., 2013. Big Data Analytics in Support of the Decision Making Process. MSc Thesis. p. 164.

Elragal, A. and Klischewski, R., 2017. Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. *Journal of Big Data,* p. 19.

Elragal, A, 2014. *ERP and big data: the inept couple.* s.l., Elsevier, pp. 242-249.

Fan, J., Han, F. and Liu, H., 2014. Challenges of big data analysis. *National science review,* Volume 1, pp. 293-314.

Gandomi, A. and Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management,* Volume 35, pp. 137-144.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M. and Herrera, F., 2016. Big data preprocessing: methods and prospects.. *Big Data Analytics Journal,* p. 9.

Goncharov, I., 2019. *Big Data and AI Landscape in 2018 ,* s.l.: Four Megatrends in Big Data in 2019 and Beyond .

Grover, P. and Kar, A.K., 2017. Big data analytics: a review on theoretical contributions and tools used in literature.. *Global Journal of Flexible Systems Management,* 18(3), pp. 203-229.

Grover, V., Chiang, R.H., Liang, T.P. and Zhang, D., 2018. Creating Strategic Business Value from Big Data Analytics: A Research Framework.. *Journal of Management Information Systems,* pp. 388-423.

Günther, W.A., Mehrizi, M.H.R., Huysman, M. and Feldberg, F., 2017. Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems,* Volume 26, pp. 191-209.

H. Eszter, 2015. Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science journal,* Volume 659, pp. 63-76.

Hart, C., 2018. *Doing a literature review: Releasing the research imagination..* s.l.:Sage.

Hartmann, T., Fouquet, F., Moawad, A., Rouvoy, R. and Le Traon, Y., 2019. GreyCat: Efficient what-if analytics for data in motion at scale.. *Information Systems journal.*

Hasselt, H.V., 2010. *Double Q-learning.* s.l., s.n., pp. 2613-2621.

He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X. and Xu, Z., 2011. *RCFile: A fast and space-efficient data placement structure in MapReduce-based warehouse systems.* s.l., IEEE, pp. 1199-12.

He, Y., Yu, F.R., Zhao, N., Yin, H., Yao, H. and Qiu, R.C., 2016. Big data analytics in mobile cellular networks. *IEEE access,* Volume 4, pp. 1985-1996.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I. and Dulac-Arnold, G., 2018. *Deep q-learning from demonstrations.* s.l., s.n.

Hofmann, E. and Rutschmann, E., 2018. Big data analytics and demand forecasting in supply chains: a conceptual analysis.. *The International Journal of Logistics Management,* Volume 29, pp. 739-766.

Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R. and Shahabi, C., 2014. Big data and its technical challenges. *Communications of the ACM,* Volume 57, pp. 86-94.

Jiang, C., Zhang, H., Ren, Y., Han, Z., Chen, K.C. and Hanzo, L., 2017. Machine learning paradigms for next-generation wireless networks.. *IEEE Wireless Communications journal,* Volume 24, pp. 98-105.

Karpovsky, A. and Galliers, R.D., 2015. Aligning in practice: from current cases to a new agenda. *Journal of Information Technology,* pp. 136-160.

Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Ali, M., Kamaleldin, W., Alam, M., Shiraz, M. and Gani, A., 2014 . Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal.*

Kiron, D., Prentice, P.K. and Ferguson, R.B., 2014. The analytics mandate. *MIT Sloan management review,* p. 1.

Kitchin, R., 2014. Big Data, new epistemologies and paradigm shifts.. *Big Data & Society Journal,* p. 2053951714528481.

Kotsiantis, Sotiris B and Zaharakis, I and Pintelas, P, 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering journal,* 160(2), pp. 3--24.

Lan, F., Chunlei, W. and Guoqing, M., 2010. *A framework for network security situation awareness based on knowledge discovery.* s.l., IEEE, pp. V1-226.

Landset, S., Khoshgoftaar, T.M., Richter, A.N. and Hasanin, T., 2015. A survey of open source tools for machine learning with big data in the Hadoop ecosystem.. *Journal of Big Data,* Volume 2, p. 24.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety.. *META group research note,* Volume 6, p. 1.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S. and Kruschwitz, N., 2011. Big data, analytics and the path from insights to value. *MIT sloan management review,* p. 21.

Lee, J., Kao, H.A. and Yang, S., 2014. Service innovation and smart analytics for industry 4.0 and big data environment.. *Procedia Cirp Journal,* pp. 3-8.

Lei, Y., Jia, F., Lin, J., Xing, S. and Ding, S.X., 2016. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data.. *IEEE Transactions on Industrial Electronics journal,* Volume 36, pp. 3137-3147.

Lekhwar, S., Yadav, S. and Singh, A., 2019. *Lekhwar, S., Yadav, S. and Singh, A., 2019. Big Data Analytics in Retail.* Singapore, Springer, pp. 469-477.

Levy, Y. and Ellis, T.J., 2006. A systems approach to conduct an effective literature review in support of information systems research.. *Informing Science Journal,* p. 9.

Li, Y. and Schuurmans, D., 2011. *MapReduce for parallel reinforcement learning.* Berlin, Heidelberg, Springer, pp. 309-320.

Lomotey, R. K., & Deters, R., 2014. *Towards knowledge discovery in Big Data.* s.l., IEEE Computer Society, p. 181–191.

Lv, Z., Song, H., Basanta-Val, P., Steed, A. and Jo, M., 2017. *IEEE Transactions on Industrial Informatics journal,* Volume 13, pp. 1891-1899.

Ma, C., Zhang, H.H. and Wang, 2014. Machine learning for big data analytics in plants.. *XTrends in plant science Journal,* pp. 798-808.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H., 2011. Big data: The next frontier for innovation, competition, and productivity..

Mayer-Schonberger, V. and Padova, Y., 2015. Regime Change: Enabling Big Data through Europe's New Data Protection Regulation.. *Colum. Sci. & Tech. L. Rev. Journal,* p. 315.

Mergel, I., Rethemeyer, R. K., & Isett, K., 2016. Big data in public affairs. *Public Administration Review Journal,* pp. 928-937.

Mikalef, P., Pappas, I.O., Krogstie, J. and Giannakos, M., 2018. Big data analytics capabilities: a systematic literature review and research agenda.. *Journal of Information Systems and e-Business Management,* Volume 3, pp. 547-578.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M., 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602.*

Mnih, Volodymyr and Badia, Adria Puigdomenech and Mirza, Mehdi and Graves, Alex and Lillicrap, Timothy and Harley, Tim and Silver, David and Kavukcuoglu, Koray, 2016. *Asynchronous methods for deep reinforcement learning.* New York, NY, US, s.n., pp. 1928--1937.

Mnih, Volodymyr and Kavukcuoglu, Koray and Silver, David and Rusu, Andrei A and Veness, Joel and Bellemare, Marc G and Graves, Alex and Riedmiller, Martin and Fidjeland, Andreas K and Ostrovski, Georg and others, 2015. Human-level control through deep reinforcement learning. *Nature journal,* Volume 518, p. 529.

Mouthami, K., Devi, K.N. and Bhaskaran, V.M., 2013. *Sentiment analysis and classification based on textual reviews..* s.l., IEEE.

Müller, O., Junglas, I., Brocke, J.V. and Debortoli, S., 2016. Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems,* pp. 289-302.

Nafus, D. and Sherman, J., 2014. Big data, big questions| this one does not go up to 11: the quantified self movement as an alternative big data practice.. *International journal of communication,* p. 11.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. and Muharemagic, E., 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data,* p. 1.

Nashua, N., 2017. *Big Data Analytics Market Study,* s.l.: Dresner Advisory Services.

Orange-Roglá, Sergio; Chalmeta, Ricardo, 2019. Framework for implementing a big data ecosystem in organizations.. *Communications of the ACM Journal,* 62(1).

Oussous, A., Benjelloun, F.Z., Lahcen, A.A. and Belfkih, S., 2018. Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences,* pp. 431-448.

Provost, F. and Fawcett, T., 2013. Data science and its relationship to big data and data-driven decision making. *Big Data Journal,* pp. 51-59.

Qiu, J., Wu, Q., Ding, G., Xu, Y. and Feng, S., 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing,* p. 67.

Raghupathi, W. and Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. *Health information science and systems,* p. 3.

Rana, S., 2019. Moving in the Realm of Big Data: Using Analytics in Management Research and Practices.. *SAGE journal,* 8(1), pp. 7-8.

Ren, S., Zhang, Y., Liu, Y., Sakao, T., Huisingh, D., Almeida, C. M. V. B, 2019. A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: A framework, challenges and future research direction.. *Journal of Cleaner Production,* Volume 210, pp. 1343-1365.

Ristoski, P., Bizer, C. and Paulheim, H., 2015. Mining the web of linked data with rapidminer.. *Web Semantics: Science, Services and Agents on the World Wide Web,* Volume 35, pp. 142-151.

Russom, P., 2011. *Big data analytics,* s.l.: TDWI best practices report, fourth quarter.

Sagiroglu, S. and Sinanc, D., 2013. *Big data: A review.* s.l., IEEE, pp. 42-47.

Schelén, O., Elragal, A. and Haddara, M., 2015. *A roadmap for big-data research and education.,* s.l.: Luleå tekniska universitet.

Schulman, J., Levine, S., Abbeel, P., Jordan, M. and Moritz, P., 2015. *Trust region policy optimization.* s.l., s.n., pp. 1889-1897.

Sharma, R., Mithas, S. and Kankanhalli, A., 2014. Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations.. *European Journal of Information Systems,* Volume 23.

Singh, J. and Singla, V., 2015. Big data: tools and technologies in big data.. *International Journal of Computer Applications.,* Volume 112.

Suthaharan, S., 2014. Big data classification: Problems and challenges in network intrusion prediction with machine learning.. *ACM SIGMETRICS Performance Evaluation Review,* Volume 4, pp. 70-73.

Team, R.C., 2000. R language definition.. *R foundation for statistical computing.*

Tien, J.M., 2013. Big data: Unleashing information.. *Journal of Systems Science and Systems Engineering,* 22(2), pp. 127-151.

ur Rehman, M.H., Chang, V., Batool, A. and Wah, T.Y., 2016. Big data reduction framework for value creation in sustainable enterprises. International Journal of Information Management, 36(6), pp.917-928.. *International Journal of Information Management,* Volume 36, pp. 917-928.

Van Hasselt, H., Guez, A. and Silver, D., 2016. *Deep reinforcement learning with double q-learning.* s.l., s.n.

Vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R. and Cleven, A., 2009. *Reconstructing the giant: On the importance of rigour in documenting the literature search process..* s.l., s.n.

Waller, M.A. and Fawcett, S.E., 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management.. *Journal of Business Logistics,* Volume 34, pp. 77-84.

Wamba, S.F., Gunasekaran, A., Akter, S., Ren, S.J.F., Dubey, R. and Childe, S.J., 2017. Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research,* pp. 356-365.

Wang, H., Xu, Z., Fujita, H. and Liu, S., 2016. Towards felicitous decision making: An overview on challenges and trends of Big Data.. *Information Sciences journal,* Volume 367, pp. 747-765.

Watson, H., 2014. Tutorial: Big data analytics: Concepts, technologies, and applications.. *Communications of the Association for Information Systems CAIS,* Volume 34, p. p.65.

Watson, H., 2019. Update Tutorial: Big Data Analytics: Concepts, Technology, and Applications.. *Communications of the Association for Information Systems,* Volume 44, p. 21.

Webster, J. and Watson, R.T., 2002. Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly,* pp. xiii-xxiii.

Wixom, B.H., Yen, B. and Relich, M., 2013. Maximizing Value from Business Analytics. *MIS Quarterly Executive.*

Woerner, S.L., Wixom, B.H., 2015. Big data: extending the business strategy toolbox.. *Journal of Information Technology,* Volume 30, pp. 60-62.

Wu, Celimuge and Yoshinaga, Tsutomu and Chen, Xianfu and Zhang, Lin and Ji, Yusheng, 2018. Cluster-based content distribution integrating LTE and IEEE 802.11 p with fuzzy logic and Q-learning. *ieee Computational intelligenCe magazine journal,* pp. 41-50.

Yi, X., Liu, F., Liu, J. and Jin, H., 2014. Building a network highway for big data: architecture and challenges. *IEEE Network journal,* Volume 28, pp. 5-13.

Zhong, R.Y., Newman, S.T., Huang, G.Q. and Lan, S., 2016. Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Computers & Industrial Engineering journal,* pp. 572-591.

Zuech, R., Khoshgoftaar, T.M. and Wald, R., 2015. Intrusion detection and big heterogeneous data: a survey.. *Journal of Big Data,* p. 3.