ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https://www.svedbergopen.com/

Reserch Paper Open Acess

Big Data Analytics (BDA): Principles, Premises, and Applications in Organizational Research

Farshad Madani[1*], Seyed Vahid Reza Nooraei[2] and Mahour Mellat Parast[3]

[1]Data Science Researcher, Portland State University, USA. E-mail: farshad.madani@gmail.com
[2]Industrial Engineering Instructor, Louisiana State University, USA. E-mail: snooraie@isu.edu
[3]Research Associate Professor, Arizona State University, USA. E-mail: mahour.parast@asu.edu

Article Info

Volume 4, Issue 2, November 2024
Received : 16 June 2024
Accepted : 11 October 2024
Published : 05 November 2024
doi: 10.51483/IJDSBDA.4.2.2024.79-91

Abstract

Big Data Analytics (BDA) is getting more widespread nowadays, and it is projected more applications and technological evolution in the near future. Nonetheless, big data analytics is at the beginning of its life cycle and many industries have not started investing to extract value from their big data. There still many managers and scholars who are not familiar what big data analytics. Many authors have been trying to provide a view from big data analytics, but they shine few aspects of big data. In this paper, we try to provide some theoretical, practical, and technological facets of big data so that managers in different industries find a comprehensive perception of what big data is, what differentiates big data analytics from other analytical disciplines such as data mining and machine learning, what are current and future applications of big data, what are challenges and barriers in big data implications, and what are required technologies exclusively developed for big data analytics.

Keywords: *Big data analytics, Big data applications, Big data technologies, Data science, Machine learning, Statistics*

1. Introduction

Big data are getting more attraction, especially among small and medium size companies, over the next decade. Data is the new form of capital that it is increasingly playing a major role in the market values of S&P 500 companies (Elsten and Hill, 2017). The volume of data is globally growing (Jyoti, 2018). IDC technologies (Technology Services Organization with primary focus in IT services) has predicted that the volume of data generated worldwide will be 163 zettabytes in 2025, which is ten times bigger than data generated in 2016 which was 16.1 zettabytes (Reinsel *et al.*, 2017). Due to quick growth of big data, related technologies have obtained much attention from managerial aspects (Yaqoob *et al.*, 2016).

* *Corresponding author: Farshad Madani, Data Science Researcher, Portland State University, USA. E-mail: farshad.madani@gmail.com*

Companies started experiencing a new transition in information technology. ERP's in 1980s, CRMs in 1990s, eCommerce in 2000s were the main stream of IT changes. In recent years, big data analytics and technologies are the main stream of information technologies evolution in industries (Minelli *et al.*, 2013). Data generation and storage is shifting from personal computers to mobiles and enterprises, and cloud technologies has a major role in this transition (Reinsel *et al.*, 2017). Smart phones, tablets, cameras, sensors, social media such as Facebook and twitter, and recently internet of things (IoTs) are everywhere and generating data. On the other hand, cloud storage and computing are getting cheaper so that even small companies can start data analytics projects to take advantage of different data sources to make more intelligent insights regard to their business. Companies are experiencing another transition in information technology (IT). While there is no agreed definition of big data yet (Addo-Tenkorang and Helo, 2016), various companies from different industries anticipate that big data will have a significant effect on their operations and processes, allowing them to make more strategic data-oriented and knowledgeable decisions. Many companies have started benefiting from digital transformation to establish their competitive position. Reportedly, this type of companies has a better business performance in comparison to their rivals that they are not actively pursuing digital transformation (Jyoti, 2017). Digital disruption is a real threat that its impact varies in different industries. For example, retail industry is more than 25% under the risk of digital disruption while hospitality is 11% under the risk. Similarly, 18% of financial services, 15% of gas and oil, 18% of government, 14% of transportation, 29% of utilities and 20% of industrial equipment all are under risk (Jyoti, 2017).

In this study, we fill an important gap in the literature in big data analytics. Prior studies in big data discussed theoretical (Arunachalam *et al.*, 2018; Blackburn *et al.*, 2017; Bumblauskas *et al.*, 2017; Günther *et al.*, 2017; Mikalef *et al.*, 2017), technological (Oussous *et al.*, 2017) and practical (Addo-Tenkorang and Helo, 2016; Lee, 2017; Oussous *et al.*, 2017; Philip Chen and Zhang, 2014) aspects of big data. While these studies enrich our understanding of big data analytics, they have overlooked Big data definition, Big data applications and Big Data analytics. We aim to fill this gap in the literature by reviewing most recent published papers.

2. Prior Studies on Big Data Analytics

Papers reviewed the literature of big data can be categorized into three main groups:

1) **Theoretical Papers:** This group of papers discuss theoretically models or concepts related to big data analytics. Mikalef *et al.* (2017) proposes a theoretical framework wherein big data technologies are addressed in order to increase business value and competitive performance. Bumblauskas *et al.* (2017) provides a framework for converting big data sets to actionable knowledge, and for decreasing potential risks. A research framework was constructed by Blackburn et al to analyze the impact of big data based on its potential to inform, enable, and transform or disrupt R&D management across four dimensions: strategy, people, technology, and process integration (Blackburn *et al.*, 2017). Günther *et al.* (2017) debates how organizations take advantage of portability, and interconnectivity as the main socio-technical features of big data. Arunachalam *et al.* (2018) provide a systematic literature review about the capabilities of big data analytics in supply chain and develop a capabilities maturity model.

2) **Technological Papers:** These papers are concentrated on technological aspects of big data analytics. For example (Oussous *et al.*, 2017) did a survey on recent Big data technologies to facilitate selecting and adopting right combination of different big data technologies. They provide not only a global view of main Big data technologies but also comparisons according to different system layers.

3) **Practical Papers:** This group of papers emphasizes on aspects such as applications, challenges, and opportunities in a specific domain. Addo-Tenkorang and Helo (2016) investigate big data applications in operations and supply chain management, and analyze the trends and perspectives in this area. Lee (2017) discusses how big data analytics is a new paradigm revolutionizing the way businesses operate in many industries. Philip Chen and Zhang (2014) investigates representative Big Data applications from typical services such as finance and economics, healthcare, supply chain management, and manufacturing sector. Oussous *et al.* (2017) explores the current condition of big data concept with its related barriers, drivers, opportunities and perceptions in architecture, engineering and construction industry with an emphasis on facilities managements.

Although the three above-mentioned groups provide very beneficial information from three different perspectives regard to big data analytics, the audience of each group misses out advantage of the others. Theoretical papers provide a mental model and frameworks that explain how BDA works and creates value for organizations. Practical papers offer applied knowledge and help the audience to gain more concrete information about how BDA functions in practice. Technological papers provide more technical view regard to how implement into current business processes and organizational structure. In this paper, we provide all three theoretical, technical, and practical benefits of BDA for practitioners. More specifically, we address: 1) Big data and non-Big data definitions, 2) Big data 5Vs model and discussions, 3) relationship between big data and data science, 4) Data mining, data science and machine learning specifications and boundaries, 5) Challenges in big data implementation, 6) Big data applications and technologies, and 7) future Big data evolution

3. What is Big Data?

Big data is defined as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika *et al.*, 2011). Big data is the mix of various kinds of gritty data. The applications that are the main sources of making voluminous amounts of data, for example Internet of Things (IoT), self-quantified, multimedia, and social media data. IoT data are generated by many types of devices such as GPS devices, intelligent/smart cars, mobile computing devices, PDAs, mobile phones, intelligent clothing, alarms, window blinds, window sensors, lighting and heating fixtures, refrigerators, microwave units, and washing machines (Raguseo, 2018).

The 5V's model is used as a common framework to describe big data (Chen *et al.*, 2012), which includes Volume, Velocity, Variety, Variability and Value as explained below:

1. Volume refers to the amount of data that an organization or an individual collects and/or generates. Currently a minimum of one terabyte is the threshold of big data (Gandomi and Haider, 2015). One terabyte stores as much data as would fit on 1,500 CDs or 220 DVDs, enough to store around 16 million Facebook photographs (Gandomi and Haider, 2015). E-commerce, social media, and sensors generate high volumes of unstructured data such as audio, images, and video. New data has been added at an increasing rate as more computing devices are connected to the internet.

2. Velocity refers to the speed at which data are generated and processed. The velocity of data increases over time. Initially, companies analyzed data using batch processing systems because of the slow and expensive nature of data processing. Real time processing becomes a norm for computing applications. According to (Chen *et al.*, 2012) it is forecasted that 6.4 billion connected devices would be in use worldwide in 2016 and that the number will reach 20.8 billion by 2020. The enhanced data streaming capability of connected devices will continue to accelerate the velocity.

3. Variety refers to the number of data types. Technological advances allow organizations to generate various types of structured, semi-structured, and unstructured data. Text, photo, audio, video, clickstream data, and sensor data are examples of unstructured data, which lack the standardized structure required for efficient computing. Semi-structured data do not conform to specifications of the relational database, but can be specified to meet certain structural needs of applications. An example of semi-structured data is Extensible Business Reporting Language (XBRL) developed to exchange financial data between organizations and government agencies. Structured data is predefined and can be found in many types of traditional databases. As new analytics techniques are developed, unstructured data are generated at a much faster rate than structured data and the data type becomes less of an impediment for the analysis.

4. IBM added Veracity as a fourth dimension (Chen *et al.*, 2012), which represents the unreliability and uncertainty latent in data sources. Uncertainty and unreliability arise due to incompleteness, inaccuracy, latency, inconsistency, subjectivity, and deception in data. Managers do not trust data when veracity issues are prevalent. Customer sentiments are unreliable and uncertain due to subjectivity of human opinions. Statistical tools and techniques have been developed to deal with uncertainty and unreliability of big data with specified confidence levels or intervals.

5. Oracle introduced *Value* as an additional dimension of big data (Chen *et al.*, 2012). Firms need to understand the importance of using big data to increase revenue, decrease operational costs, and serve customers better; at the same time, they must consider the investment cost of a big data project. Data would be low value in their original form, but data analytics will transform the data into a high-value strategic asset. IT professionals need to assess the benefits and costs of collecting and/or generating big data, choose high-value data sources, and build analytics capable of providing value-added information to managers (Lee, 2017).

Figure 1 shows the main three dimensions of big data include Volume, Variety and Velocity where we added two more dimensions *value* and *veracity* to complete possible dimensions of big data. Traditional dimensions of big data include volume, variety and velocity where we believe that two more dimension like value and veracity should be added to complete Big Data dimensions. Value dimension stands on operational cost, revenue and customer service. Similarly, veracity dimension is defined based two classes include Uncertainty and Unreliability.

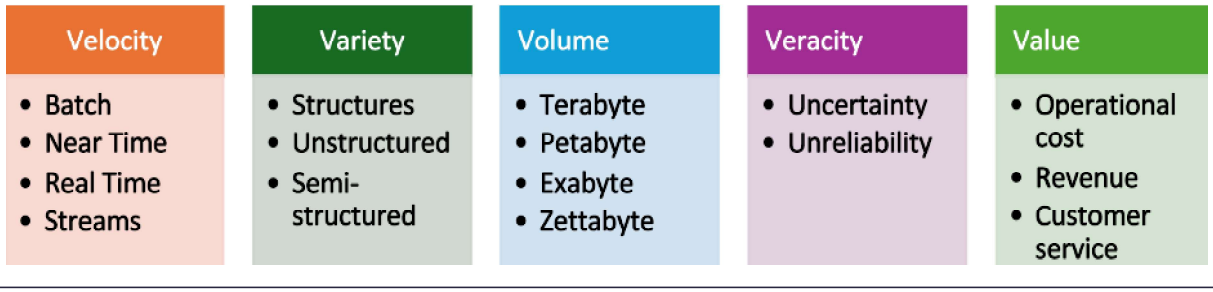

Figure 1: 5V's Model

4. What is Big Data Analytics?

Big Data Analytics (BDA) has intertwined with three major analytical terms which are Data Science, Data Mining, and Machine Learning. In order to understand carefully the boundaries of BDA, it is required to understand the domain of each of these terms.

Data science is multi-disciplinary area including computer science, math and statistics, and a knowledge domain where knowledge domain is knowledge of a specific, specialized discipline or field, in contrast to general knowledge (Gandomi and Haider, 2015). Data science definition is focused on applying data to use scientific method to businesses. Data scientists generate hypotheses, perform experiments; they also predict and build new products or interactions based on outcomes. Organizations are about to loop processes out of human and products will be updated in real time by customer reactions and other inputs (Minelli *et al.*, 2013).

Data mining is a branch of statistics. Data mining includes association rules, clustering and feature selection (Minelli *et al.*, 2013). Data mining is the process of discovering patterns in large data sets involving methods at the dealing with statistics and database systems (Minelli *et al.*, 2013). Data mining is not limited to finding solutions for the problems through specific methods. In fact, some methods such as statistical methods or machine learning techniques are applied to analyze data for rather long time. In data analysis, the statistical methods are applied for analyzing data to understand the situation we are facing. The problem specific methods for data mining also attempted to understand the meaning from the collected data (Bi and Cochran, 2014).

Machine learning is an learning algorithm from empirical data and then using those lessons to predict future outcomes of new data (Minelli *et al.*, 2013). Machine-learning methods are usually based on the application of supervised and unsupervised machine-centered techniques where learning come from machine performance (Lee, 2017). Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Use supervised learning if you have known data for the output you are trying to predict. Supervised learning uses classification and regression techniques to develop predictive models.

Classification techniques predict discrete responses—for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring. Common algorithms for performing classification include support vector machine (SVM), boosted and bagged decision trees, k-nearest neighbor, Naïve Bayes, discriminant analysis, logistic regression, and neural networks. *Regression techniques* predict continuous responses—for example, changes in temperature or fluctuations in power demand. Typical applications include electricity load forecasting and algorithmic trading. *Unsupervised learning* finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labeled responses. *Clustering* is the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data. Applications for cluster analysis include gene sequence analysis, market research, and object recognition (Alpaydin, 2010).

As some scholars have offered a definition for big data analytics, there is no significant difference between big data analytics and the three terms reviewed. Figure 2 presents the scope each of the above-mentioned

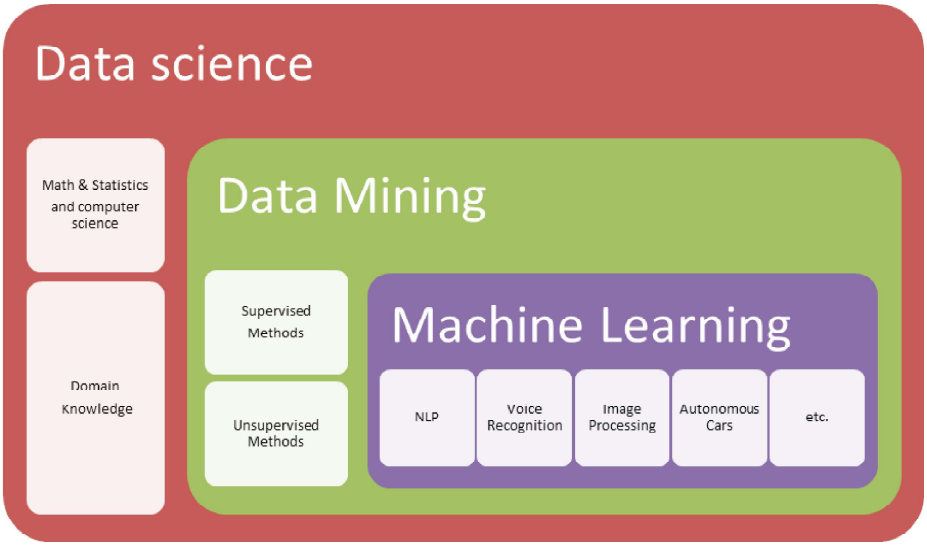

Figure 2: The Relation between Data Science, Data Mining, and Machine Learning

Table 1: Sample Definitions of Big Data Analytics

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.

terms. Big data analytics is not represented in the figure, because there is no inherent difference between big data analytics and the concepts presented. As we see in Table 1, big data analytics is a specific type of data mining or machine learning wherein data size is *big*. All common methods in descriptive, predictive, and prescriptive analytics methods are applicable in big data analytics (Minelli *et al.*, 2013). Text mining and social network analysis are widely used analytical methods in BDA on different types of data sources such as social network feeds, emails, blogs, online forums, survey responses, corporate documents, and news held by organizations. In addition, audio, images, video, and voice are other types of big data that they widely processed and analyzed in BDA (Gandomi and Haider, 2015). What distinguished big data analytics are technologies applied to store, retrieve, analyze, and visualize big data. Big data technologies are explained in the next section.

5. Big Data Analytics Applications

Big Data Analytics (BDA) generates business value (Ahmed *et al.*, 2017) through various ways such as targeted influencer marketing, better business visions, better business opportunities, autonomous decision making for real time processes, user behaviors realization, customer retention, fraud detections, and many others (Russom and Org, 2011). BDA has had substantial improvements industries such as health care, public sector administration, retailor sector, manufacturing data as well as in science including astronomy, atmospheric science, medicine, genomics, biologic, biogeochemistry and other complex and interdisciplinary scientific researches (Philip Chen and Zhang, 2014). Philip Chen and Zhang show that almost 50% of BDA applications is for the purpose of increasing operational efficiency (Philip Chen and Zhang, 2014). In addition to industrial and scientific applications of BDA, it is used for modernizing societies. More specifically, BDA is applied in developing smart buildings, smart cities, autonomous cars and smart grids. Also, there are some applications in social sciences and humanities (Rodríguez-Mazahua *et al.*, 2016).

There are many successful cases in implementing BDA to gain business values. For example, Amazon generates about 30% of its sales through analytics (Akter and Wamba, 2016). Match.com was able to have more than 50% rise in revenue two years while the company subscriber base for its core business reached 1.8 million (Akter and Wamba, 2016). By early 2002 companies started collecting IP-specific user information in the backup using cookies and server logs to know better customer requirements and need to realize new business opportunities (Bumblauskas *et al.*, 2017).

Some areas, which are more significantly affected by BDA applications, are more elaborated as below:

Finance and Economics: A case study which is related to the application of big data in finance and economics reveals that in year 2012, some financial institutions like European Hedge Fund, Global Investment Bank, Retail Banking Innovation Leader, Asia/Pacific National Bank, Expanding U.S. Property Insurer, Global European Institution, Investment Research Institution, and Community Bank applied Big Data in different areas to reach goals in price and investment policies for large portfolio trades and swaps, building merchant intelligence (Zhong *et al.*, 2016).

Healthcare: According to McKinsey reports Big Data provide more than $300 billion savings yearly in US healthcare; this saving result on reductions of 8% to national healthcare expenses. Two largest areas include Clinical operations and R&D in potential savings, with $165 billion and $108 billion in waste respectively. Big Data could assist waste reduction and inefficiency in the three areas include Clinical Operations, R&D and Public Health (Feldman *et al.*, 2012). Another application of big data in health care is a system called GNS where patients data grow exponentially and become digitized to store in Electronic Health Record and Health Information Exchange include any type of data such as test results, medical, prescriptions, and personal health devices (Minelli *et al.*, 2013).

Manufacturing: Big data application in manufacturing shows that Toyota Motor Company manages traffic through big data by totally 700,000 Toyota vehicles on traffic. According to an estimation, 16 billion dollars is made yearly by Toyota company. Similarly, Siemens through big data which is around 100,000 measurements daily in power plants provide RDS (remote diagnostic services) which is used to analyze operational behaviors (Zhong *et al.*, 2016). Big data application in supply chain denotes that United Parcel Service of North America

(UPS), assign $1 billion annually on big data to change the shipping business and cutting fuel and company costs. Similarly, FedEx has successful story in maintaining customers to track their packages in real-time through all big data (Zhong *et al.*, 2016).

Digital Marketing: Digital marketing includes applying any kind of online media channel. It is defined as existence online whether it is profitable or nonprofitable, there is no matter. And it is defined as browsing people to a website, a mobile app. Some example of digital marketing for instance Google's digital marketing which is applied in business intelligence. Similarly, ERP system is another example of digital marketing where big data are collected in a large data base. Other examples are Facebook, Google+ and Twitter (Minelli *et al.*, 2013).

6. BD Technologies

Current Big data technologies are recognized in terms of data storage (preprocessing), data processing, data visualization (post-processing) and big data analytics.

6.1. BD Storage Technologies

Big data storage usually refers to volumes that grow exponentially to terabyte or petabyte scale Although a specific volume size or capacity is not formally defined (Philip Chen and Zhang, 2014).

Largest big data practitioners like Google, Facebook, Apple, and Twitter, hyper scale computing environments are customized to run systems like Hadoop, NoSQL, and Cassandra, which have PCIe flash storage or furthermore to disk storage to cut down storage latency to a minimum amount (Chen *et al.*, 2014). One technology from Apache Hadoop is an open-source software framework written in Java for processing large-scale data sets (Dittrich and Quiané-Ruiz, 2012). The framework comprises several modules that are Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop MapReduce, which have been designed in capable of autonomously performing in software when hardware failures commonly are happened.

In the traditional data domain, a relational database perspective can cover most of the relationships and links through data that the business requires for information support. This is not the case of big data, which might not have a database, or which might use a database like NoSQL, that requires no database perspective. Because of this, big data models must be built on systems, not databases. The system components that big data models should contain are business data requirements, corporate governance and security, the physical storage used for the data, integration and open interfaces for all types of data, and the capability to handle a variety of different data models. There are commercial data modeling tools that support Hadoop, as well as big data reporting software like Tableau. When applying big data tools and methodologies, IT decision makers should include the capability to build data models for big data as one of their requirements. Many big data analytic platforms, like SQLstream and Cloudera Impala, series still use SQL in its database systems, because SQL is more capable and easier query language with high performance in stream big data real-time analytics. To store and manage unstructured data or non-relational data, NoSQL used a number of specific approaches (Zhong *et al.*, 2016). The most popular NoSQL database is Apache Cassandra. Cassandra, which was once Facebook proprietary database, was released as open source in 2008. Other NoSQL implementations include SimpleDB, Google BigTable, Apache Hadoop, MapReduce, MemcacheDB, and Voldemort. Companies that use NoSQL include Twitter, LinkedIn and NetFlix (Zhong *et al.*, 2016).

6.2. BD Processing Technologies

Distributed computing is a model where components of a software system are shared among various computers to developed efficiency and performance. Based on the narrowest of definitions, distributed computing is bounded to programs with components shared among computers within a limited geographic area. Broader definitions include shared stuffs as well as program components. In the widest sense of the term, distributed computing means that something is shared among various systems which may also be in different locations. Distributed computation technology is Hadoop ecosystem (Zhong *et al.*, 2016). Hadoop is written in Java and is a top-level Apache project that started in 2006. It emphasizes discovery from the

perspective of scalability and analysis to realize near-impossible feats. Doug Cutting developed Hadoop as a collection of open-source projects on which the Google MapReduce programming environment could be applied in a distributed system (Khan *et al.*, 2014).

The power of Hadoop platform is based on two main subcomponents: the Hadoop Distributed File System (HDFS) and the MapReduce framework (Oussous *et al.*, 2017). The Hadoop Distributed File System (HDFS) HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server (Minelli *et al.*, 2013). Apache Hadoop is one of the most well-known software platforms that support data-intensive distributed applications. It develops the computational paradigm named Map/Reduce. To store data, Hadoop relies on both its file system HDFS and a non-relational database called Apache HBase (Oussous *et al.*, 2017). Apache Hadoop platform consists of the Hadoop kernel, Map/Reduce and Hadoop distributed file system (HDFS), as well as a number of related projects, including Apache Hive, Apache HBase, and so on (Philip Chen and Zhang, 2014).

MapReduce and YARN constitute two options to carry out data processing on Hadoop. They are designed to manage job scheduling, resources and the cluster. It is worth noticing that YARN is more generic than MapReduce (Oussous *et al.*, 2017). Apache Pig is an open source framework that generates a high level scripting language called Pig Latin. It reduces MapReduce complexity by supporting parallel execution of MapReduce jobs and workflows on Hadoop (Oussous *et al.*, 2017).

Data Access Layer include Data Ingestion which are Sqoop, Flume and Chukwa, and Data Ingestion which are Sqoop, Flume and Chukwa. Data streaming include storm and spark and Storage Management include HCatalog (Oussous *et al.*, 2017). Table 2 shows the comparison of several big data platforms (Hashem *et al.*, 2015; Philip Chen and Zhang, 2014).

Table 2: Comparison of Several Big Data Cloud Platforms

	Google	Microsoft	Amazon	Cloudera
Big data storage	Google cloud services	Azure	S3	
MapReduce	AppEngine	Hadoop on Azure	Elastic MapReduce (Hadoop)	MapReduce YARN
Big data analytics	BigQuery	Hadoop on Azure	Elastic MapReduce (Hadoop)	Elastic MapReduce (Hadoop)
Relational database	Cloud SQL	SQL Azure	MySQL or Oracle	MySQL, Oracle, PostgreSQL
NoSQL database	AppEngine Datastore	Table storage	DynamoDB	Apache Accumulo
Streaming processing	Search API	Streaminsight	Nothing prepackaged	Apache Spark
Machine learning	Prediction API	HadoopþMahout	HadoopþMahout	HadoopþOryx
Data import	Network	Network	Network	Network
Data sources	A few sample datasets	Windows Azure marketplace	Public Datasets	Public Datasets
Availability	Some services in private beta	Some services in private beta	Public production	Industries

6.3. BD Visualization Tools

Visualization are ways to represent data in different formats such as tables, diagrams, and images to make data easier to understand and interpret. The high magnitude of big data hinders researcher to apply traditional visualization tools, so big data visualization is not as easy as traditional visualization (Wang *et al.*, 2015). Big data visualization goes far beyond typical corporate graphs, histograms and pie charts to more complex

representations like heat maps and fever charts, enabling decision makers to explore data sets to identify correlations or unexpected patterns.

Tableau and Microsoft Power BI are two main commercialized business intelligence tools used for big data visualization. Tableau is flexibly adaptable to different cloud technologies such as AWS, and Google cloud.

7. Big Data Implementation Challenges

Big data challenges have paid attention to the difficulties of understanding the notion of BD, decision-making of what data are generated an d collected, issues of privacy and ethical considerations relevant to mining such data (Sivarajah *et al.*, 2017). The challenges are divided into three categories. The first category is related to the characteristics of the data itself (e.g., data volume, variety, velocity, veracity and volatility). The second category is called process challenges which are related to series of how to capture data, how to integrate data, how to transform data, how to select the right model for analysis and how to provide the results and the last category of challenges are called as management challenges cover for example privacy, security, governance and ethical aspects (Sivarajah *et al.*, 2017).

Big data analytics implementation is always accompanied by barriers including outdated IT infrastructure, the inherent complexity and messiness of big data, lack of data science skills within organizations, privacy concerns, and organizational cultures. There are specific tactics for removing BD, namely: (1) the utilization of commodity hardware and specialized big data software, (2) collaboration with educational institutions, (3) installation of policies and processes that would support individual privacy, and (4) development of a clear organizational vision in relation to big data (Alharthi *et al.*, 2017).

8. Discussions

Big Data Definition: It is stated that big datasets are far more complicated, or it knows data variety is different from others because authors believe that data comes from various formats. Regarding critical elements, it is stated that the speed of data generation and data delivery are essential elements in big data since the streaming data have high-frequency in real-time decision-making and authors know specifications for big data such as high value and low veracity then they provide a new definition on big data (Sheng *et al.*, 2017). Also, it is explained that higher priority among 5V's where it is explained veracity and value, which represent the rigorousness of Big Data Analytics (BDA), are particularly important" while in prior literature there is no priority among 5V's (Nguyen *et al.*, 2018).

Scope of Big Data: For instance authors explain that while the size, scope, and scale of data are subject to limit in defining big data but it is almost impossible, they believe that the definition of big data must change completely around the analysis of the data rather than the real size of the database (i.e. large data sets or databases) because they think that still seems to be rather subjective (Bumblauskas *et al.*, 2017). One of the main problems in practice is how to know a boundary for the size and scope of the data set. They believe that data analysis based upon that information is critical to the process of efficiently big data definition and application (Bumblauskas *et al.*, 2017). It is added that the main objective of data accumulation and data analyzes is used for decision making and designing actions while creating value across all levels of the organization. Such justification is not acceptable because BDA inherently is based on 5V's. Otherwise every data analytics can be considered big data analytics (Bumblauskas *et al.*, 2017).

Limited Applications, No Clear Conclusion: Several published papers provided limited applications without a clear conclusion at the end of paper where technical knowledge in big data is missing. For instance, a literature (Mikalef *et al.*, 2017) review article defines mechanisms and processes through big data result business value to companies also authors believe that the article is designed to explain mechanisms through BDA results to competitive perform gains which is limited contribution. As a conclusion, it is not mentioned that in literatures how theoretically the driven research is missing. Similarly, there is no a clear research framework in IT business field where it is believed that framework provides a reference for the broader implementation of big data in the business context (Mikalef *et al.*, 2017). Some article proposes a maturity framework of SCA based on a four capability levels through explanation of four levels at a same time in abstract section while in conclusion, it is mentioned that SCA reveals a gap between theory and supply chain

practices, this gap is not well-defined and well discussed in article since the gap is originated from a framework of maturity of SCA as an observation (Wang and Hajli, 2016). Another article discusses a clustering to avoid duplications to gain relevant findings and identify potential gaps, use a limited number of articles published in years 2010 and 2011 while there are still articles in this area prior year 2010 (Pospiech and Felden, 2012).

Some literature shows that missing proper definitions in the area. For example, an article explains data mining as a collection of classification, regression and clustering which is almost messy definition. Similarly, the author knows machine learning as supervised and unsupervised learning, this type of definitions is inaccurate and non-total (Wang and Hajli, 2016).

Big data analytics from an evolutionary perspective is a subject to discuss. An article related to big data analytics evolution concludes that big data analytics in business does not have any changes in principles from beginning until now. For example, always more data have been best choice for making decision since data include information which is capable (Agrawal, 2014). Based on a classification in years from 1994-2015 the evolution is divided into 3 classes. The first category of evolution happened between years 1994-2004 include e-commerce, web usage mining and web structure mining. The second-class evolution happens between years 2005-2014 include social media, sentiment analysis, lexical based methods and machine learning and at last evolution beginning by year 2015 until present year include IoT applications and streaming analytics (Lee, 2017). Similar article classifies big data evolution beginning by year 1980 with ERP and CRM in 1990, the revolution continues by e-commerce in year 2000 and it is extended by Big data analytics in year 2010 (Minelli *et al.*, 2013).

9. Future BD Applications

Future application of big data is classified into three groups. The first group is about smart manufacturing for example, sensors in Ford Focus Electric car produce streams of data while the car is driven and when it is parked. Smart manufacturing provides a potential of changing how products are invented, manufactured, shipped, and sold (Rodríguez-Mazahua *et al.*, 2016). Smart city is also a new research area based on the application of IoT data. Similarly, small grid case is achieved through various connections among smart meters, sensors, control centers and other infrastructures. BDA is applied to identify at-risk transformers and to pick up abnormal behaviors of the connected devices. Grid Utilities can thus choose the best treatment or action (Panetta, 2017).

Big data technologies in a framework has different shapes, for instance a classification include data collection, data management and data utilization. For data collection, IoT makes it possible for the cloud to acquire data from many data sources including the Internet, sensors, log files, conventional Relational Database Management Systems (RDBMS), industrial networks, and tracking systems, RFIDs, WSNs, Social MIedias and Surveys. For data management, cloud computing offers reliable services by deploying cloud data centers. Some platform technologies, such as MapReduce and NoSQL, are needed to tackle big data and to retrieve relevant data effectively. For Data management, other technologies such as storage, infrastructure, Query, Hadoop, MapReduce, Data Analytics and security assurances. About data utilization, different tools are developed to analyze the retrieved data and extract knowledge to support decision-making activities. Sub systems in data utilization include SaaS, PaaS, Sales prediction, virtual manufacturing, TQM, SCM, PLM, Predictive manufacturing, Lean production, Project management, ERP I, ERP II, Enterprise alliance (Bi and Cochran, 2014). From other perspective, data technologies include data search, data sharing, and data analysis and data visualization. Since data are to be applied to make accurate decisions whenever needed, it becomes necessary that it should be available in accurate, complete and timely manner. Sharing data is more important than producing it and timely and cost-effective analytics over Big Data is now a key ingredient for success in many businesses in data analysis also visualization of big data is a complicated task for data visualization (Rodríguez-Mazahua *et al.*, 2016).

10. Conclusion

Data science is a combination of applied computer science and applied math and statistics where data may come from any field in science which is called domain knowledge. Data mining which is an application of

statistics exist in data science, so data mining is a part of data science. Furthermore, machine learning which uses learning capability of machine to predict data and find algorithms using data mining methods thus machine learning can be a part of data mining. As a result, both machine learning and data mining are some part of data science, then we defined Big Data analytics by defining its components and differences. We clarified Big Data definition by separating what it is Big Data and what it is not. Scope of big data was discussed and articles with limited contribution and ambiguous/not clear conclusions were addressed.

Big data evolution is divided into 3 categories. The first category of evolution is between years 1994-2004 include e-commerce, web usage mining and web structure mining. The second category happens between years 2005-2014 include social media, sentiment analysis, lexical based methods and machine learning and at last evolution beginning by year 2015 until present year include IoT applications and streaming analytics.

Beside three types of big data analysis include descriptive, predictive, prescriptive analysis, there is another classification for big data analysis include Data Storage, Data processing and Data visualization. This category includes present and future development of big data such as smart cities, healthcare treatments, cyber security, and disaster management.

References

Addo-Tenkorang, R. and Helo, P.T. (2016). Big Data Applications in Operations/Supply-Chain Management: A Literature Review. *Computers & Industrial Engineering*, 101, 528-543. https://doi.org/10.1016/J.CIE.2016.09.023

Agrawal, D. (2014). Analytics Based Decision Making. *Journal of Indian Business Research*, 6(4), 332-340. https://doi.org/10.1108/JIBR-09-2014-0062

Ahmed, V., Tezel, A., Aziz, Z. and Sibley, M. (2017). The Future of Big Data in Facilities Management: Opportunities and Challenges. *Facilities*, 35(13/14), 725-745. https://doi.org/10.1108/F-06-2016-0064

Akter, S. and Wamba, S.F. (2016). Big Data Analytics in E-commerce: A Systematic Review and Agenda for Future Research. *Electronic Markets*, 26(2), 173-194. https://doi.org/10.1007/s12525-016-0219-0

Alharthi, A., Krotov, V. and Bowman, M. (2017). Addressing Barriers to Big Data. *Business Horizons*, 60(3), 285-292. https://doi.org/10.1016/J.BUSHOR.2017.01.002

Alpaydin, E. (2010). *Introduction to Machine Learning*. Massachusetts Institute of Technology. Retrieved from https://books.google.com/books?hl=en&lr=&id=7f5bBAAAQBAJ&oi=fnd&pg=PR5&dq=Introduction+to+Machine+Learning&ots=C46G_iacNn&sig=lgaHslGTsOJzhLY_5a8cLgQ972k#v=onepage&q=Introduction to Machine Learning&f=false

Arunachalam, D., Kumar, N. and Kawalek, J.P. (2018). Understanding Big Data Analytics Capabilities in Supply Chain Management: Unravelling the Issues, Challenges and Implications for Practice. *Transportation Research Part E: Logistics and Transportation Review*, 114, 416-436. https://doi.org/10.1016/J.TRE.2017.04.001

Bi, Z. and Cochran, D. (2014). Big Data Analytics with Applications. *Journal of Management Analytics*, 1(4), 249-265. https://doi.org/10.1080/23270012.2014.992985

Blackburn, M., Alexander, J., Legan, J.D. and Klabjan, D. (2017). Big Data and the Future of R&D Management. *Research-Technology Management*, 60(5), 43-51. https://doi.org/10.1080/08956308.2017.1348135

Bumblauskas, D., Nold, H., Bumblauskas, P. and Igou, A. (2017). Big Data Analytics: Transforming Data to Action. *Business Process Management Journal*, 23(3), 703-720. https://doi.org/10.1108/BPMJ-03-2016-0056

Chen, Chiang and Storey. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), 1165. https://doi.org/10.2307/41703503

Dittrich, J. and Quiané-Ruiz, J.-A. (2012). Efficient Big Data Processing in Hadoop MapReduce. In *Proceedings of the VLDB Endowment*, 5, 2014-2015, VLDB Endowment. https://doi.org/10.14778/2367502.2367562

Elsten, C. and Hill, N. (2017). Intangible Asset Market Value Study?. *Journal of the Licensing Executives Society*, LII(4).

Feldman, B., Martin, E.M. and Skotnes, T. (2012). *Big Data in Healthcare Hype and Hope*. Retrieved from https://www.ghdonline.org/uploads/big-data-in-healthcare_B_Kaplan_2012.pdf

Gandomi, A. and Haider, M. (2015). Beyond the Hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management*, 35(2), 137-144. https://doi.org/10.1016/J.IJINFOMGT.2014.10.007

Ghasemaghaei, Maryam, Hassanein, Khaled and Turel, Ofir, (2015). Impacts of Big Data Analytics on Organizations: A Resource Fit Perspective. *AMCIS 2015 Proceedings*, 19. https://aisel.aisnet.org/amcis2015/BizAnalytics/GeneralPresentations/19

Günther, W.A., Rezazade Mehrizi, M.H., Huysman, M. and Feldberg, F. (2017). Debating Big Data: A Literature Review on Realizing Value from Big Data. *The Journal of Strategic Information Systems*, 26(3), 191-209. https://doi.org/10.1016/J.JSIS.2017.07.003

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A. and Ullah Khan, S. (2015). The Rise of "Big Data" on Cloud Computing: Review and Open Research Issues. *Information Systems*, 47, 98-115. https://doi.org/10.1016/J.IS.2014.07.006

Jyoti, R. (2017). *Become a Data Thriver: Realize Data-Driven Digital Transformation (DX)*. Retrieved from https://datavisionary.netapp.com/us/assets/US42988017-Become-a-Data-Thriver-Realize-Data-Driven-Transformation.pdf

Jyoti, R. (2018). *Unlock the Power of Data Capital: Accelerate Digital Transformation*. Retrieved from https://www.dellemc.com/en-us/collaterals/unauth/analyst-reports/products/storage/unlock_data_cap_accelerator_idc_white_paper.pdf

Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Ali, W.K.M., Alam, M., … Gani, A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal*, 2014, 712826. https://doi.org/10.1155/2014/712826

Kwon, O., Lee, N. and Shin, B. (2014). Data Quality Management, Data Usage Experience and Acquisition Intention of Big Data Analytics. *International Journal of Information Management*, 34(3), 387-394. https://doi.org/10.1016/j.ijinfomgt.2014.02.002

Lamba, H.S. and Dubey, S.K. (2015). Analysis of Requirements for Big Data Adoption to Maximize IT Business Value. 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), *Trends and Future Directions*, 1-6, Noida, India. doi: 10.1109/ICRITO.2015.7359268.

Lee, I. (2017). Big Data: Dimensions, Evolution, Impacts, and Challenges. *Business Horizons*, 60(3), 293-303. https://doi.org/10.1016/J.BUSHOR.2017.01.004

Loebbecke, C. and Picot, A. (2015). Reflections on Societal and Business Model Transformation Arising from Digitization and Big Data Analytics: A Research Agenda. *The Journal of Strategic Information Systems*, 24(3), 149-157. https://doi.org/10.1016/j.jsis.2015.08.002

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity | McKinsey.

Mikalef, P., Pappas, I.O., Krogstie, J. and Giannakos, M. (2017). Big Data Analytics Capabilities: A Systematic Literature Review and Research Agenda. *Information Systems and E-Business Management*, 1-32. https://doi.org/10.1007/s10257-017-0362-y

Minelli, M., Chambers, M. and Dhiraj, A. (2013). *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, John Wiley & Sons, Inc. Retrieved from https://books.google.com/books?hl=en&lr=&id=Mg3WvT8uHV4C&oi=fnd&pg=PT7&dq=big+data+big+analytics+wiley&ots=JKkMB0Ez1f&sig=D1C_0NSkcabZ3MphDbGjOYgmKew#v=onepage&q=big data big analyticswiley&f=false

Müller, O., Junglas, I., Brocke, J. vom and Debortoli, S. (2016). Utilizing Big Data Analytics for Information Systems Research: Challenges, Promises and Guidelines. *European Journal of Information Systems*, 25(4), 289-302. https://doi.org/10.1057/ejis.2016.2

Nguyen, T., Zhou, L., Spiegler, V., Ieromonachou, P. and Lin, Y. (2018). Big Data Analytics in Supply Chain Management: A State-of-the-Art Literature Review. *Computers & Operations Research*, 98, 254-264. https://doi.org/10.1016/J.COR.2017.07.004

Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A. and Belfkih, S. (2017). Big Data Technologies: A Survey. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/J.JKSUCI.2017.06.001

Panetta, K. (2017). Gartner Top 10 Strategic Technology Trends for 2018-Smarter with Gartner. Retrieved September 21, 2018, from https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/?utm_source=social&utm_campaign=sm-swg&utm_medium=social

Philip Chen, C.L. and Zhang, C.-Y. (2014). Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. *Information Sciences*, 275, 314-347. https://doi.org/10.1016/J.INS.2014.01.015

Pospiech, M. and Felden, C. (2012). Big Data—A State-of-the-Art. In *Proceedings of the Eighteenth Americas Conference on Information Systems*, Seatlle, Washington. Retrieved from https://aisel.aisnet.org/amcis2012/proceedings/DecisionSupport/22

Raguseo, E. (2018). Big Data Technologies: An Empirical Investigation on their Adoption, Benefits and Risks for Companies. *International Journal of Information Management*, 38(1), 187-195. https://doi.org/10.1016/J.IJINFOMGT.2017.07.008

Reinsel, D., Gantz, J. and Rydning, J. (2017). *Data Age 2025: The Evolution of Data to Life-Critical*.

Rodríguez-Mazahua, L., Rodríguez-Enríquez, C.-A., Sánchez-Cervantes, J.L., Cervantes, J., García-Alcaraz, J.L. and Alor-Hernández, G. (2016). A General Perspective of Big Data: Applications, Tools, Challenges and Trends. *The Journal of Supercomputing*, 72(8), 3073-3113. https://doi.org/10.1007/s11227-015-1501-1

Russom, P. and Org, T. (2011). *Big Data Analytics*. Retrieved from https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf

Sheng, J., Amankwah-Amoah, J. and Wang, X. (2017). A Multidisciplinary Perspective of Big Data in Management Research. *International Journal of Production Economics*, 191, 97-112. https://doi.org/10.1016/j.ijpe.2017.06.006

Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V. (2017). Critical Analysis of Big Data Challenges and Analytical Methods. *Journal of Business Research*, 70, 263-286. https://doi.org/10.1016/J.JBUSRES.2016.08.001

Wang, L., Wang, G. and Alexander, C.A. (2015). Big Data and Visualization: Methods, Challenges and Technology Progress. *Digital Technologies*, 1(1), 33-38. https://doi.org/10.12691/DT-1-1-7

Wang, Y. and Hajli, N. (2016). Exploring the Path to Big Data Analytics Success in Healthcare. *Journal of Business Research*, 70, 287-299.

Yaqoob, I., Targio Hashem, I.A., Gani, A., Mokhtar, S., Ahmed, E., Badrul Anuar, N. and Vasilakos, V.A. (2016). Big Data: From Beginning to Future. *International Journal of Information Management*, 36(6), 1231-1247. https://doi.org/10.1016/J.IJINFOMGT.2016.07.009

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*, 101, 572-591. https://doi.org/10.1016/J.CIE.2016.07.013

Cite this article as: Farshad Madani, Seyed Vahid Reza Nooraei and Mahour Mellat Parast (2024). Big Data Analytics (BDA): Principles, Premises, and Applications in Organizational Research. *International Journal of Data Science and Big Data Analytics*, 4(2), 79-91. doi: 10.51483/IJDSBDA.4.2.2024.79-91.