Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Cloud Computing Technology Algorithms Capabilities in Managing and Processing Big Data in Business Organizations: MapReduce, Hadoop, Parallel Programming

Munisf Sokiyna

*Corresponding author, PhD Candidate, Department of Management Information System, Cyprus International University, Cyprus/Nicosia. E-mail: munsif.sokiyna@gmail.com

Musbah J. Aqel

Assistant Professor, Department of Management Information System, Cyprus International University, Cyprus/Nicosia. E-mail: maqel@ciu.edu.tr

Omar A. Naqshbandi

PhD Candidate, Department of Management Information System, Cyprus International University, Cyprus/Nicosia. E-mail: eng.omar@hmu.edu.iq

Abstract

The objective of this study is to verify the importance of the capabilities of cloud computing services in managing and analyzing big data in business organizations because the rapid development in the use of information technology in general and network technology in particular, has led to the trend of many organizations to make their applications available for use via electronic platforms hosted by various Companies on their servers or so-called cloud computing that have become an excellent opportunity to provide services efficiently and at low cost, but managing big data presents a definite challenge in the cloud space beginning with the processes of extracting, processing data, storing data and analyze it. Through this study, we dealt with the concept of cloud computing and its capabilities in business organizations. We also interpreted the notion of big data and its distinct characteristics and sources. Finally, the relationship between cloud computing with big data was also explained (extraction, storage, analysis).

Keywords: Cloud Computing, Data, Data Warehouse, Big Data, Artificial Intelligence (AI), Business Organizations.

DOI: 10.22059/jitm.2020.76298

Introduction

The use of social media and various search engines has increased recently. Like Facebook, Twitter, YouTube, Google, WhatsApp, Myspace, and others. As OurWorldData.com web-site mentioned that Facebook, which is one of the most important social networking sites in this period, has 2.4 billion users, and YouTube and WhatsApp have more than a billion users, as shown in figure 1(Booth, 2019b, 2019a). Figure 1 shows that there have been recent social networking sites Like TikTok that was released in the month of 9 of 2016, where this site in the mid of 2018 has half a billion subscribers, which is equivalent to that TikTok got 20 million per month during that period (BOOTH, 2019b, 2019a). So, we are talking about huge numbers of people who use these websites. As Bell (Bell, 2008) described in his study, these sites constitute a virtual world like the one we live in, but digitally. This encouraged the owners of these sites to take advantage of these huge numbers of people to develop their businesses and make accurate decisions using artificial intelligence (AI) (Noraziah, Fakherldin, Adam, & Majid, 2017). These sites produce very large quantities of data within seconds and parts of a second as well, which is called Big Data that is analyzed to use it in making decisions for various fields such as industries, marketing, and businesses in various fields. Because of these huge amounts of data, many challenges have emerged accompanying the analysis and interpretation of this data, the most important of which is the high costs of processing, storing and accessing this data.

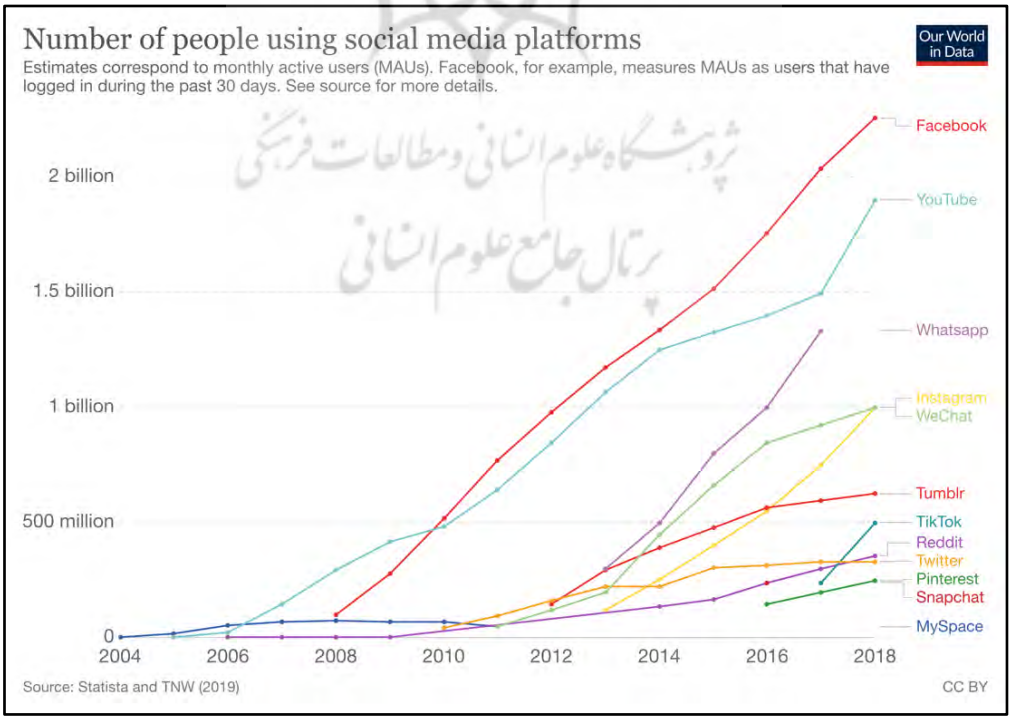
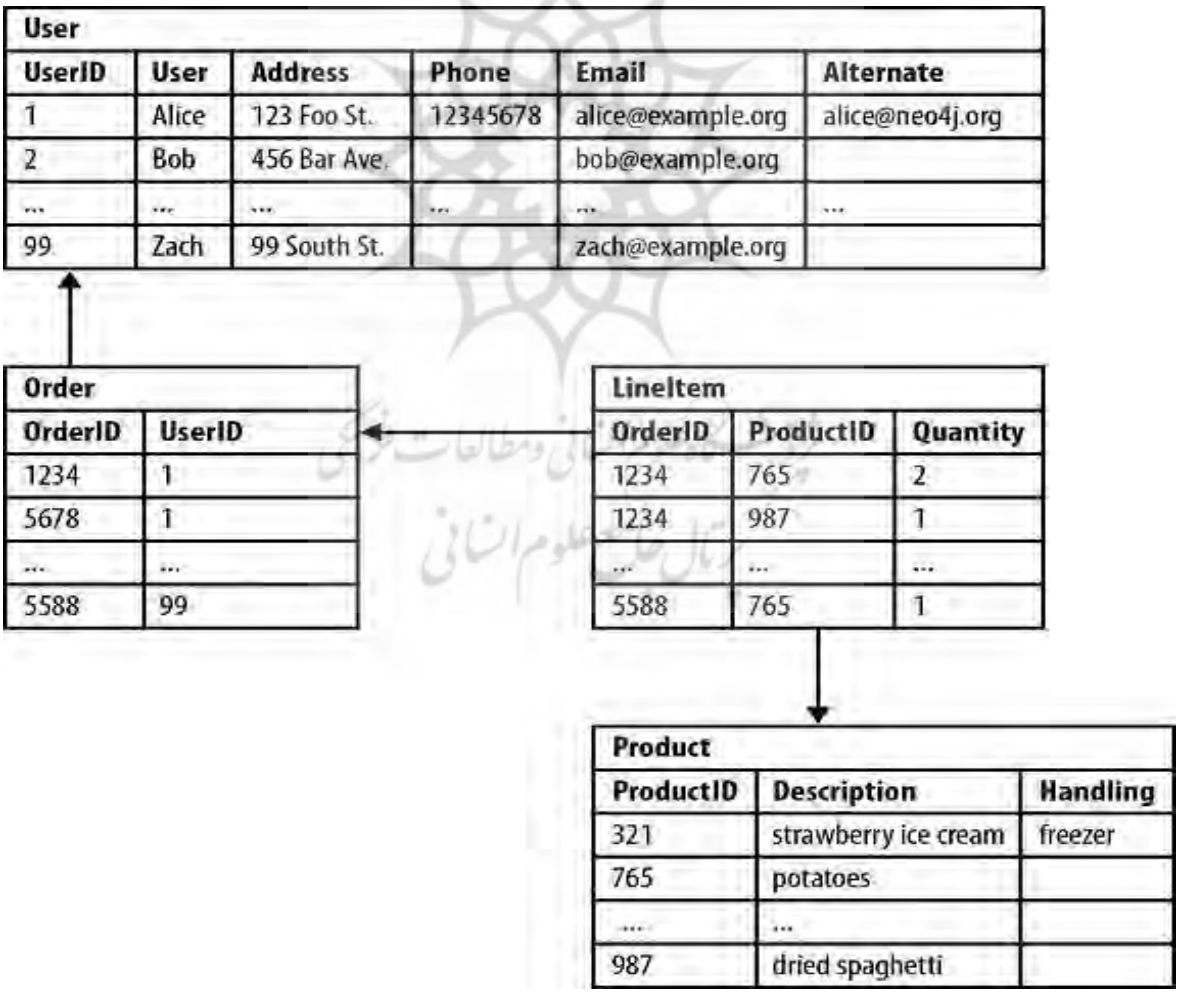


Figure 1. A number of people using social media platforms (BOOTH, 2019b, 2019a)

Therefore, significant resources must be allocated to support the analyses of this data. (Experiment, 2011; Kalil, 2012; Noraziah et al., 2017). With the rapid growth and development of social networking applications and other applications such as bioinformatics analysis and semantic web analysis, the amounts of data that have to be analyzed have increased. Based on this, Cloud computing techniques have been used, which are among the technologies that have a strong structure for cloud computing, especially large-scale computing. Cloud computing has provided many solutions for dealing with big data like multi-media, large-scale and high- dimensional data(Ji, Li, Qiu, Awada, & Li, 2012). From the big data sources except for the social media and other sites, the big data comes from smartphones, the sensors in the smart devices in addition to the audio and video recordings. These data fall into three main classifications: structured, semi-structured, and unstructured data as shown in the subsequent figures 2,3, and 4 (Noraziah et al., 2017; Yang, Huang, Li, Liu, & Hu, 2017).



User

UserID	User	Address	Phone	Email	Alternate
1	Alice	123 Foo St.	12345678	alice@example.org	alice@neo4j.org
2	Bob	456 Bar Ave.		bob@example.org	
...	...	...	...	...	...
99	Zach	99 South St.		zach@example.org	

Order

OrderID	UserID
1234	1
5678	1
...	...
5588	99

LineItem

OrderID	ProductID	Quantity
1234	765	2
1234	987	1
...	...	...
5588	765	1

Product

ProductID	Description	Handling
321	strawberry ice cream	freezer
765	potatoes	
....	...	
987	dried spaghetti	

Figure 2. Structured Data (Pickell, 2018)

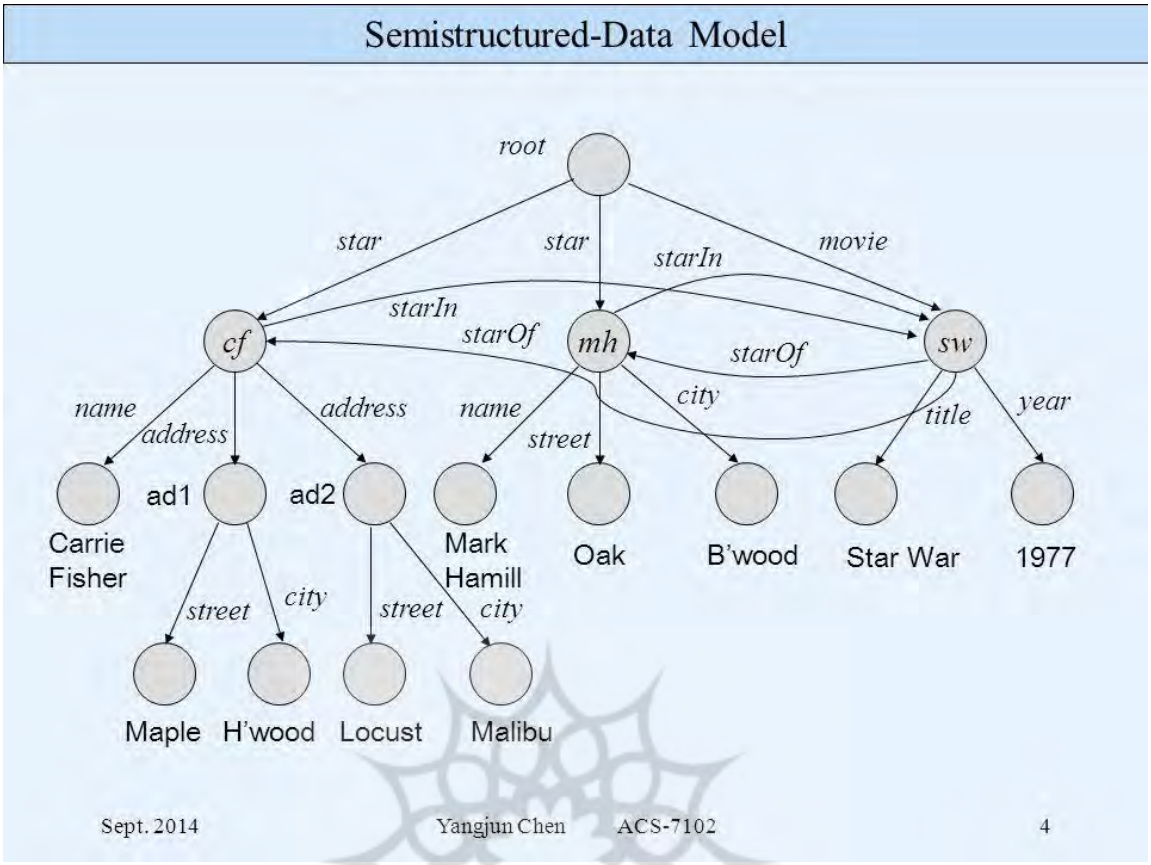


Figure 3. Semi-structured data(Yangjun, 2014)

Figure 4. Unstructured data (Gomes, 2016)

Big Data SSSS Characteristics

Human understanding and the rapid development of information technology have helped to transfer data from the traditional stage to the more accelerated methods described by volume, velocity, variety, veracity, and value represented by 5V's (Marr, 2015; Yang et al., 2017). To begin with, first V alludes to the data volume which is the most related to big data due to its importance in determining the amount of data to be analyzed because sometimes the data reach very large proportions and sometimes they are not understood in the form of logs and this requires large spaces of storage and also provide the necessary capabilities to analyze and process this data(Slagter, Hsu, & Chung, 2015).

The second V stands for velocity that alludes to the speed of transferring data from its sources such as Facebook or various sensors to supercomputers to analyze, use, and process them in decision- making processes at a fast rate to face any potential problems before they occur or worsen(Gewirtz, 2018; IMMERMAN, 2017). Because speed is not only limited to obtaining data from its sources very quickly, but also using it and processing it quickly, and this speed is required because there is some data that its value can be eroded over time (Yang et al., 2017; Zikopoulos, Eaton, & others, 2011). As for variety, as we mentioned, there are various sources from which data is produced such as sensors, social media, different packet, and other sources. Because of this, the data structure is different from each other according to the source that came from it, as the coming data do not only constitute the regular tables of databases consisting of rows and columns, as many of the data are in an unorganized form and difficult to be available in a ready manner for processing and integration with applications(Mao et al., 2015).

The fourth V refer to the Veracity of the data, and here the Veracity is represented by, among other things, the reliability and quality of big data and the extent to which they can be used for the purpose of solving a specific problem(dependability). Because there is a lot of big data that carry abnormalities or inconsistencies and sometimes duplication, therefore it is necessary to remove these matters to obtain big data of high quality suitable for processing.(Mayer-Schönberger & Cukier, 2013; Naimi & Westreich, 2014). And last but not least, the fifth V that all previous V's work for it, which represents the value of big data. It must be ensured that the data collected will benefit the company or not(Jain, 2016). In simple words from a logical perspective, there is no real benefit to the data itself, but the benefit lies in converting it into valuable information that helps companies make certain decisions to develop the business or to produce a specific product and others. This point is very important because the infrastructure used to analyze and process data is very costly, it is important to ensure that the decisions made are based on accurate data and lead to measurable results (Manyika et al., 2011; Mao et al., 2015). Figure 5 shows big data 5V's.

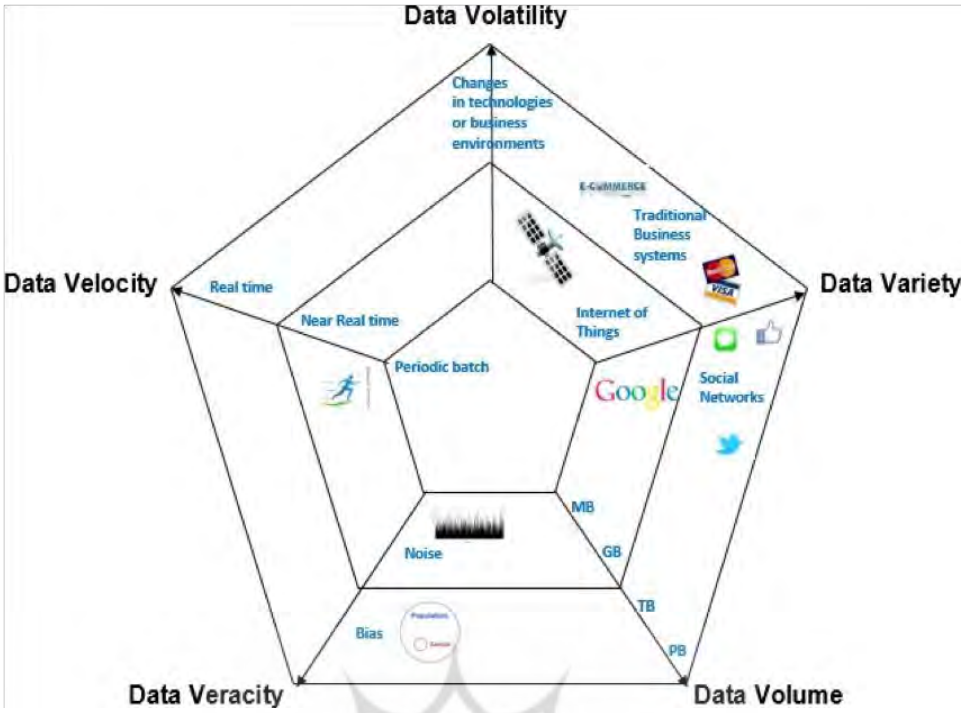


Figure 5. Big Data 5 V's (Hammer, Kostroch, & Quiros, 2017)

Big Data and Business

The growth of Big Data, particularly its adoption by organizations and industry, increases the meaning/content of Big Data (Yang et al., 2017). The first volume-based definition currently incorporates the data itself, related technologies and skills to help produce, gather, store, manage, process, analyze, exhibit and use data, just as the information and knowledge inferred (Ramapriyan, 2015). From a business point of view, the Big Data period was imagined as the subsequent step for development and innovations, rivalry, and profitability. As in the McKinsey report given its possibility to drive business incomes and make new opportunities (Manyika et al., 2011). For example, explore how big data can adapt by using a four-stage strategy, (Dasgupta, 2013).

Monsanto's corporation, a climate specialist, used big geographic data to help farmers around the world be careful about agricultural matters and did so by analyzing multi-layered climate behavior to adapt to climate change (Sykuta, 2016). (Manyika et al., 2011) expected in his study that the big data will increase the income of the current companies by nearly above 50% and encourage new business in the coming decades to invest in the use and analysis of big data because it will create new opportunities, especially in different areas of daily life. In brief, the Big Data field introduces extraordinary chances and changes in the digital life field in daily routine transactions(Mayer-Schönberger & Cukier, 2013), including

customized medication (Alyass, Turcotte, & Meyre, 2015), modified item suggestions and travel alternatives. The previous barely any years have seen this change from theoretical to reality through a large group of new technological techniques (e.g. Uber transportations).

Cloud Computing and Handling the Big Data

Big data processing has been mentioned in many previous studies but focused on stream-based and distributed processing. So that the cloud computing appeared earlier than the big data emergence, where the cloud computing is considered a new model of computation that works to present the computations as an extra tool after basics needed like water, air, electricity, gas and telephone communications with the highlights of rapid elasticity, resources pooling, on-demand, broad network access, and self-service(Mell & Grance, 2011; Pettey & van der Meulen, 2012; Zikopoulos et al., 2011). Cloud computing used big data for the purpose of analyzing it and providing proposed solutions to various digital problems in various fields such as social sciences, business, industrial and other fields(Yang et al., 2017; Naveed Q.N, 2019). There are, in fact, many of the characteristics that characterize cloud computing as a whole, and among these characteristics, there are those that greatly support the challenges facing big data and the table 1 show the characteristics of cloud computing and how it supports perspectives of big data:

1) The feature of pooling resources is one of the most important advantages of cloud computing, as it allows us to allocate different resources and share them with many clients and serve consumers with a multi-tenant model in addition to high dynamism through the speed of recalling data according to the user's request;

2) Elasticity is defined as the extent to which the system adapts to changes in the work environment so that it works to meet the different demands (On-demand)and match them with the available resources at any time and that is by forecasting and speed of monitoring(velocity)(Experiment, 2011; Yang et al., 2017);

3) variety of big data coming from different sources are handled with elasticity, resources pooling, and self-service advantages (Noraziah et al., 2017);

4) The self-service process determines the best matching services to reach the veracity of the Big Data (Ji et al., 2012; Noraziah et al., 2017).

5) the value represented as accurate forecasting, justifiable cost and customer satisfaction with on-demand service, elasticity features of cloud computing(Yang et al., 2017). Figure 6 illustrates the concept of cloud bases big data storage.

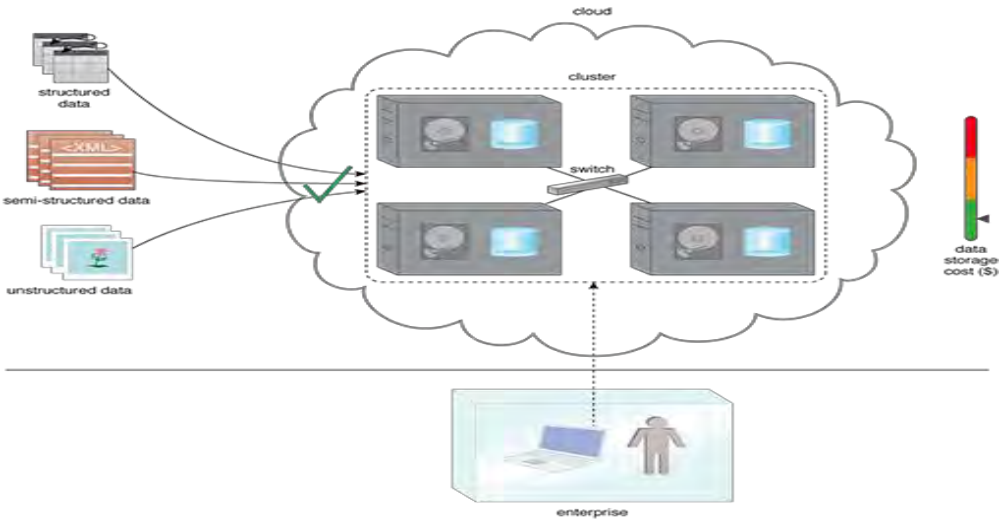


Figure 6. Cloud-Based Big Data storage(Erl & Khattak, n.d.)

Cloud computing has provided modern capabilities to carry out big data processing and analysis within organizations and various businesses. It has become using new technologies to carry out the extraction, storage, analysis as well as mining operations. Some of the most common problems that cloud computing addresses are social media analysis, encryption and decryption, simulation, pattern recognition, and data mining. To address these problems, for example, the popular Twitter site, which produces more than 4 petabytes per day, equivalent to 12 terabytes per day(Matthews, 2018). Cloud computing enables Twitter to use more than 8,000 central processing units and to use Peta and zeta bytes to perform big data processing and analysis(Bagheri & Shaltooki, 2015).

Table 1. Tackling Big Data challenges with cloud computing(Yang et al., 2017)

BigData/Cloud Computing	Rapid Elasticity	Resource Pooling	On Demand	Broad Network Access	Self service
Volume		√			√
Velocity	√		√		
Variety	√	√		√	
Veracity				√	√
Value	√		√		√

Cloud Computing Analysis Technologies

Cloud computing has turned into a new model of computing. Many frameworks and algorithms have become closely associated with cloud computing such as MapReduce, Hadoop, and Big Table, which has helped them become a powerful model for analyzing and

processing massive amounts of big data across distributed IT systems(Purcell, 2014; Voruganti, 2014). So that these technologies depend on two main principles which are the elasticity that expresses the ability of these applications to use computerized resources when needed and upon request. Another principle is the scalability, which is the ability to respond in different circumstances, such as different amounts of data and different sources(Purcell, 2014; Voruganti, 2014). Many researchers and companies defined cloud computing, however, there was a lot of ado in addition to being widely misused to represent anything online(Hadoopa, 2019). Therefore, the National Institute of Standards and Technology has defined cloud computing as" a pay-per-use model to enable conveniently, on-demand network access to a shared pool of configurable computing resources such as applications and services, storage, networks and servers that can be provided and issued quickly with least management effort or service provider interaction". Therefore, cloud computing services and capabilities have helped provide scalable and affordable solutions to meet the challenges of big data(Mell & Grance, 2011). There are some concepts associated with cloud computing that are not fairly new such as parallel programming, distributed systems, and grid computing(Kobielus, 2012). But one of the most important concepts and enablers of cloud computing is a virtualization technology that has helped cloud business models. Where this technology enables any physical machine to host multiple virtual devices (VMs) and this is an investment in capital and the total benefit of the devices. Among the most famous of these virtual applications are VMware, VMware Workstation, Oracle Virtual Box, and Fusion, which enables the user to run another operating system on the same device (Windows, Linux, Mac, and Unix) in addition to the applications associated with different virtual devices, all installed on one device (Hadoopa, 2019). Hadoop is an open-source software consider as a top-level Apache project written in Java programming language(Gu et al., 2014). Hadoop consists of two parts: the first part is the programming paradigm MapReduce and the second part the Hadoop Distributed File System (HDFS). It enables applications to work with a lot of standalone computers and process massive amounts of data, estimated in a petabyte(Bagheri & Shaltooki, 2015). Some of the major tasks of MapReduce include machine learning, social media content analysis, mining, and data mining. HDFS works to store large files across distributed and multiple machines that achieves reliability by replicating data across multiple hosts.(Kobielus, 2012). There are also many IT companies such as IBM, Yahoo and Intel that have used HDFs as a technology for storing big data. Also, regular users use public cloud computing services including Google Drive Mega Drive, Dropbox iCloud and overcome the limited capacity storage on personal computers (Gu et al., 2014). Parallel programming (e.g. MapReduce) is one of the most effective factors with big data for widespread use(Alvaro et al., 2010; Dobre & Xhafa, 2014). It is a process of using multiple resources to solve a specific problem so that. It divides the problem into a smaller series of steps, also provides instructions and implements solutions simultaneously(Alvaro et al., 2010). Figure 7 shows the usage of cloud computing in big data.

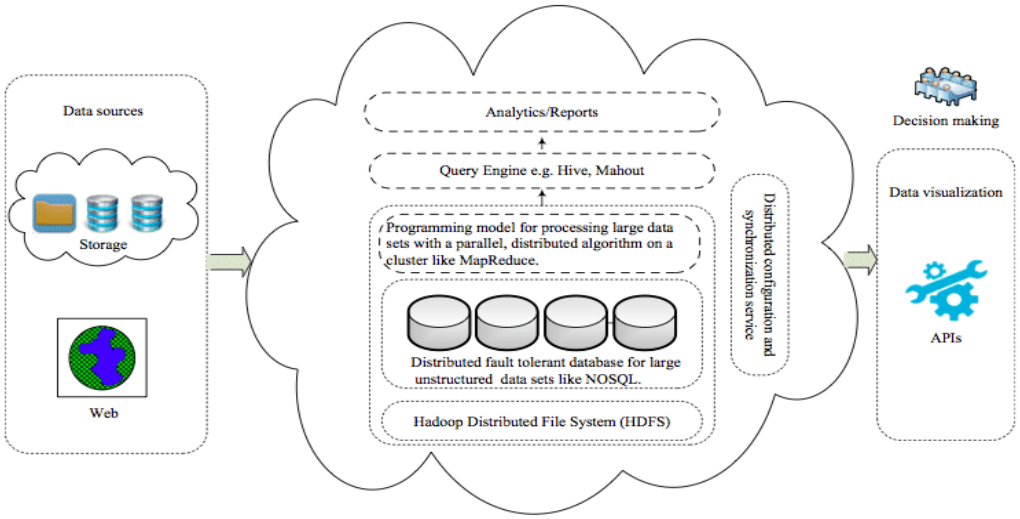


Figure 7. Usage of cloud computing in Big Data (Khan, Shakil, & Alam, 2018)

Discussion

Big data these days is very important, especially in routine life, as there is the production of these data every second or more precisely every part of the second around the world with different sources such as social networking sites, sensors, different industries, radio frequencies (RFID), and many other sources in conjunction with the rapid development of artificial intelligence technologies such as machine learning, deep learning, and neuronal networks, this undoubtedly required an increase in computer capabilities, especially in storage and processing capabilities as well (Ji et al., 2012). Therefore, this matter requires understanding and analyzing this data, then employing it, as does work. Therefore, there are basic aspects of big data that we must take into consideration to ensure the full benefit of big data, including methods of storage and management of this data as it represents a major challenge in this field as the current data management systems and techniques are not able to achieve big data, needs such as storage speed compared to data velocity so we need a hierarchical storage structure in addition to that the previous computer algorithms are unable to store the huge amounts of big data that come from different sources for several reasons including the heterogeneity of this data in addition to the distortions for that Big data system is another problem in big data management so the services and capabilities of cloud computing came to solve some of these problems that were difficult to address in the past. Among the most important things that also constitute a big challenge for big data is the analysis of big data and its processing, specifically during the processes of querying and processing big data, as speed is an important factor in these operations and usually these processes take a long time to complete(Zhou, Lu, Li, & Du, 2012). In these cases, the index is the best option for such problems as indicators in the big data between modern pre-processing technology and the appropriate index of big data are a desirable solution in these cases. For

example, if a lot of parallel data is in the application then traditional algorithms are ineffective with this case, it is necessary to use the capabilities of cloud computing with a low-cost model that enables many computers to be used at a short cost. Finally, an important issue related to big data is the security of this data. This is why cloud computing services enabled the use of big data applications on the Internet, as this enabled companies to reduce the costs of IT infrastructure to a large extent(Ji et al., 2012). However, security and privacy have a major impact on storage operations and big data processing. This is due to the great use of third parties and the infrastructure for hosting important data and due to the large increase in the volume of data and applications, this has presented many challenges represented in protecting the security and monitoring data, in contrast to traditional security methods as data security revolves around extracting and processing big data without revealing this sensitive information for users.

Conclusion

Lately, the massive amount of big data generated every day and hourly from various sources has increased. Where this research paper focused on a systematic survey on the management and processing of big data through the services and capabilities provided by the cloud computing represented by the management of big data and methods of storage, and the processes of analysis and processing in addition to its security. As the velocity of big data growth increases dramatically due to the widespread use of portable and non-portable smart devices in addition to internet-connected sensors. This matter formed positive and negative aspects, as it enabled many companies and businesses to develop their business and keep pace with it which has become significantly digitized and at the same time posed many challenges such as the rise of information technology infrastructures cost, especially those responsible for analyzing and processing large data storage. Where cloud computing applications are among the most important technologies that have helped greatly to address these challenges in addition to the parallel programming framework and Hadoop and MapReduce algorithms. These technologies helped extract, process, store, transmit and recall big data in real-time. There is no doubt that this data will continue to increase rapidly and significantly and at the same time will create new challenges accompanying this data. Cooperation between applied science scientists and computer science, in particular, is necessary in order to ensure the exploitation and success of cloud computing services and applications and to explore weaknesses that can be developed.

References

Alvaro, P., Condie, T., Conway, N., Elmeleegy, K., Hellerstein, J. M., & Sears, R. (2010). Boom analytics: exploring data-centric, declarative programming for the cloud. In *Proceedings of the 5th European conference on Computer systems* (pp. 223–236).

Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics*, *8*(1), 33.

Bagheri, H., & Shaltooki, A. A. (2015). Big Data: challenges, opportunities and Cloud based solutions. *International Journal of Electrical and Computer Engineering*, *5*(2), 340.

Bell, M. W. (2008). Toward a definition of "virtual worlds." *Journal For Virtual Worlds Research*, *1*(1).

BOOTH, C. (2019a). Number of people using social media platforms. Retrieved from https://www.statista.com/topics/1164/social-networks/

BOOTH, C. (2019b). The most popular social media networks each year, gloriously animated. Retrieved from https://thenextweb.com/tech/2019/06/11/most-popular-social-media-networks-year-animated

Dasgupta, A. (2013). Big data: The future is in analytics. *Geospatial World*, *3*(9), 28–36.

Dobre, C., & Xhafa, F. (2014). Parallel programming paradigms and frameworks in big data era. *International Journal of Parallel Programming*, *42*(5), 710–738.

Erl, T., & Khattak, W. (n.d.). i Buhler, P.(2016). Big Data Fundamentals: Concepts, Drivers & Techniques. Prentice Hall.

Experiment, T. Dz. (2011). The DØ Experiment . Retrieved from https://www-d0.fnal.gov

Gewirtz, D. (2018). Volume, velocity, and variety: Understanding the three V's of big data | ZDNet. Retrieved from https://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/

Gomes, P. (2016). Log analysis | Loggly. Retrieved from https://www.loggly.com/product/log-analysis/

Gu, R., Yang, X., Yan, J., Sun, Y., Wang, B., Yuan, C., & Huang, Y. (2014). SHadoop: Improving MapReduce performance by optimizing job execution mechanism in Hadoop clusters. *Journal of Parallel and Distributed Computing*, *74*(3), 2166–2179.

Hadoopa, A. (2019). Apache Hadoop. Retrieved from http://hadoop.apache.org//

Hammer, C., Kostroch, M. D. C., & Quiros, M. G. (2017). *Big Data: Potential, Challenges and Statistical Implications*. International Monetary Fund.

IMMERMAN, G. (2017). What is big data velocity? Retrieved from https://www.machinemetrics.com/blog/what-is-big-data-velocity

Jain, A. (2016). The 5 V's of big data - Watson Health Perspectives. Retrieved from https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/

Ji, C., Li, Y., Qiu, W., Awada, U., & Li, K. (2012). Big data processing in cloud computing environments. *Proceedings of the 2012 International Symposium on Pervasive Systems, Algorithms, and Networks, I-SPAN 2012*, 17–23. https://doi.org/10.1109/I-SPAN.2012.9

Kalil, T. (2012). Big Data is a Big Deal | whitehouse.gov. Retrieved January 30, 2020, from https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal

Khan, S., Shakil, K. A., & Alam, M. (2018). Cloud-based big data analytics—a survey of current research and future directions. In *Big Data Analytics* (pp. 595–604). Springer.

Kobielus, J. G. (2012). The Forrester Wave™: Enterprise Hadoop Solutions, Q1 2012. *Forrester Research*.

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

Mao, R., Xu, H., Wu, W., Li, J., Li, Y., & Lu, M. (2015). Overcoming the challenge of variety: big data abstraction, the next evolution of data management for AAL communication systems. *IEEE Communications Magazine*, *53*(1), 42–47.

Marr, B. (2015). *Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance*. John Wiley & Sons.

Matthews, K. (2018). Here's How Much Big Data Companies Make On The Internet - Big Data Showcase. Retrieved from https://bigdatashowcase.com/how-much-big-data-companies-make-on-internet/

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. Special Publication 800-145. *Gaithersburg: National Institute of Standards and Technology*.

Naimi, A. I., & Westreich, D. J. (2014). Big data: A revolution that will transform how we live, work, and think. Oxford University Press.

Noraziah, A., Fakherldin, M. A. I., Adam, K., & Majid, M. A. (2017). Big Data Processing in Cloud Computing Environments. *Advanced Science Letters*, *23*(11), 11092–11095.

Pettey, C., & van der Meulen, R. (2012). Gartner's 2012 Hype cycle for emerging technologies identifies" Tipping point" technologies that will unlock long-awaited technology scenarios. *Hype Cycle Special Report. P1-4*.

Pickell, D. (2018). Structured vs Unstructured Data – What's the Difference? Retrieved from https://learn.g2.com/structured-vs-unstructured-data

Purcell, B. M. (2014). Big data using cloud computing. *Journal of Technology Research*, *5*, 1.

Ramapriyan, H. K. (2015). The Role and Evolution of NASA's Earth Science Data Systems.

Slagter, K., Hsu, C.-H., & Chung, Y.-C. (2015). An adaptive and memory efficient sampling mechanism for partitioning in MapReduce. *International Journal of Parallel Programming*, *43*(3), 489–507.

Sykuta, M. E. (2016). Big data in agriculture: property rights, privacy and competition in ag data services. *International Food and Agribusiness Management Review*, *19*(1030-2016–83141), 57–74.

Voruganti, S. (2014). Map Reduce a Programming Model for Cloud Computing Based On Hadoop Ecosystem. *International Journal of Computer Science and Information Technologies*, *5(3)*.

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*, *10*(1), 13–53. https://doi.org/10.1080/17538947.2016.1239771

Yangjun, C. (2014). Semistructured-Data Model Sept. 2014Yangjun Chen ACS Semistructured-Data Model Semistructured data XML Document type definitions XML schema. - ppt download. Retrieved from https://slideplayer.com/slide/4950204/

Zhou, X., Lu, J., Li, C., & Du, X. (2012). Big data challenge in the management perspective. *Communications of the CCF*, *8*, 16–20.

Zikopoulos, P., Eaton, C., & others. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

Bibliographic information of this paper for citing:

Sokiyna, M., Aqel, M.J., & Naqshbandi, O.A. (2020). Cloud Computing Technology Algorithms Capabilities in Managing and Processing Big Data in Business Organizations: MapReduce, Hadoop, Parallel Programming. *Journal of Information Technology Management*, 12(3), 100-113.