Journal of Computer Science

Review

Big Data Characteristics, Architecture, Technologies and Applications

[1]Wisam A. Qader, [1]Musa M. Ameen and [2]Bilal Ismael Ahmed

[1]Department of Computer Engineering, Faculty of Engineering, Tishk International University, Erbil, Iraq
[2]Department of Information Technology, Faculty of Science, Tishk International University, Erbil, Iraq

Article history
Received: 04-06-2020
Revised: 28-04-2020
Accepted: 20-06-2020

Corresponding Author:
Wisam A. Qader
Department of Computer Engineering, Faculty of Engineering, Tishk International University, Erbil, Iraq
Email: wisam.softeng@gmail.com

Abstract: Big Data is a vast volume of data that is not easy to be stored or processed with conventional approaches within a limited period. Therefore, to manage and extract value from it, a new architecture, method and analysis are needed. Big Data poses many challenges and problems and it has different properties such as volume, velocity, variety and veracity. The goal of Big Data is not only to collect, save and organize huge volumes of data, but it is also used to evaluate, extract and visualize useful information for further processes. Big Data is a modern worldwide novel technology that has the potential to provide great benefits to business and organizations of different fields around the world and it will be more desirable in the next few years. This work describes the importance of Big Data, various challenges it faces in adapting to today's modern era, characteristics and architecture of Big Data, technologies used in Big Data and applications created using Big Data. The paper also explains MapReduce and Hadoop Distributed File System as two important models of Big Data.

Keywords: Big Data, Hadoop Distributed File System, MapReduce

Introduction

The size of data is rapidly increasing in the globe at a very high speed. The source of the Big Data is generated from audio, video, text, mobile phones, images, e-mails, health records, sensor machines, social networks, scientific data, businesses, websites, applications, etc. That huge volume of data contains two types of data: Structured and unstructured. As data in the industry grow, those data can no longer be moved with old and traditional tools and traditional databases alone cannot solve every aspect of Big Data. Therefore, more robust algorithms, new machine learning technologies and approaches need to be designed and applied for that purpose (Song and Zhu, 2016).

It can be said that Big Data is very beneficial in various areas, both economically and nationally such as Financial Services, Health Care and Medicine Services, Education, Banking, Location Information Services, Telecommunications, Media, etc. By the end of 2003 humans created 5 Exabytes (10^{18} bytes) of data. But today, the same size of data will be created in a couple of days. In 2012, the data size reached 2.72 zettabytes (1021 bytes) as it is mentioned in (Singh and Singh, 2012). The amount of data is estimated to increase by 40% each year, reaching about 16.8 zettabytes of data by 2017. If the average storage of a PC can store 500 Gigabytes (109 Bytes), this requires approximately 33 billion computers to save all of the data of the globe and this amount is about 23 zettabytes in 2019. Previously, it had been taken around 10 years to decode a human genome. But due to rapid advancements in technology, now it is possible to have the same result within a week. Google alone has about one million servers worldwide by 2012 and the number rises to 2.5 million by 2016 (Gerhardt *et al*., 2012).

According to IBM as discussed in (Center, 2012), nearly 2.5 Exabytes of data are generated every 24 h. Statistics show that approximately 90% of the current data on the globe has come from the last two years. This amount is doubling every two years. For those purposes, there should be an ecosystem to drive such a huge amount of data which is called Big Data (GreyCampus, 2019).

This paper describes and reviews the importance of Big Data, various challenges it faces in adapting to today's modern era, characteristics and architecture of Big Data, technologies used in Big Data and applications created using Big Data. The manuscript also explains MapReduce and Hadoop Distributed File System as two important models of Big Data.

Science Publications

This paper is divided into the sections as follows: Section 1 introduces big data. Section 2 describes the characteristics of Big Data. Section 3 discusses Big Data architecture. Section 4 describes the importance of Big Data. Section 5 describes sources of Big Data. Section 6 describes the technologies used in Big Data. Big Data applications and future are discussed in sections 7 and 8, respectively. Finally, this paper concludes with section 9.

Big Data Characteristics

Volume

Is the size of the data that is stored and processed. Over 3 billion people using the internet, about 40 billion Internet-connected devices, 30 billion daily Google searches, about 70,000 h of video upload on YouTube, 1.52 billion People are logged into Facebook daily. Connections with 125 million friends, uploading 350 million photos every 24 h and 2.7 billion comments are put on Facebook only. Google processes about 20 petabytes of data every day, so it translates into 66 different languages and supports a lot of services. More than One billion tweets are posted every three days on Twitter. Every day more than 570 new websites are being created (Madden, 2012). It has been reported that, in 2018, the number of IoT devices will have more than tripled since 2012 and there will be 50 billion devices that will work on the Internet (Burhan *et al*., 2018). Figure 1 shows some statistics about the number of connected devices worldwide in billions from 2012 to 2020.

Velocity

Is the speed at which the data are processed. This concept explains the speed of processing of the data coming from various sources to drive the data in almost near the real time. This characteristic is limited to the speed of the received data as well as the speed of the data flow. For example, since the data from the sensor devices are always moving to database storages, that amount is not as small as to be processed easily. Therefore, the traditional systems have insufficient capability to perform analysis on that huge amount of data (Madden, 2012).

Variety

Is processing different kinds of data. In addition to the traditional data, the generated data can be retrieved from both active and passive devices, including web pages, logs, files, e-mails, documents and even the data from the sensor devices (Madden, 2012). All those data are entirely not one type and contain raw, structured, unstructured and semi-structured data, which are hard to be processed with the existing conventional and traditional systems.

Veracity

It refers to the certainty, noise and abnormality of the data. There are many challenges in data analysis, veracity is one of the assets among those problems when values like volume and speed are compared Statista, 2019.

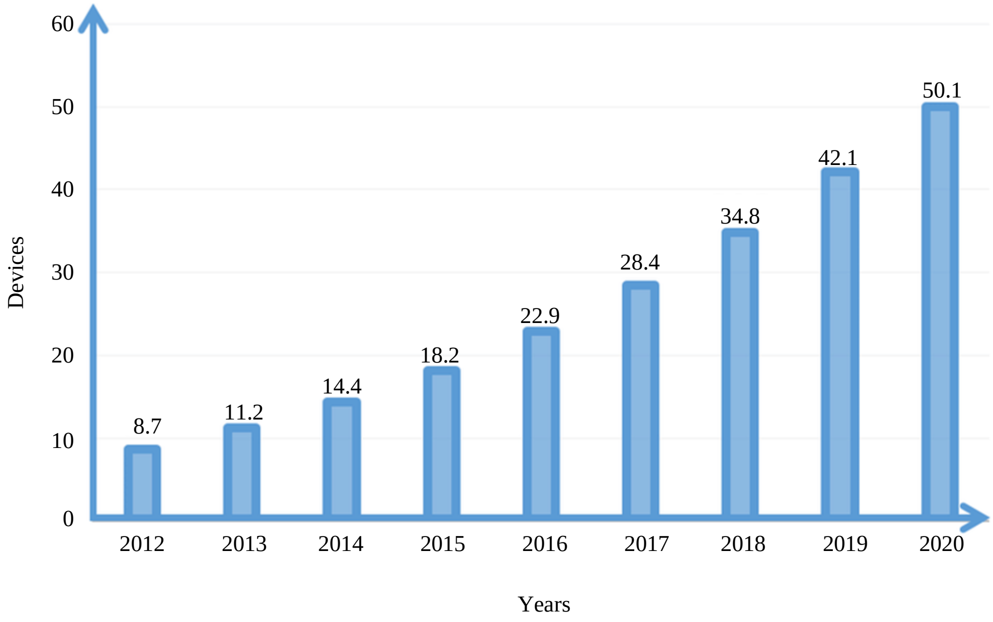

Fig. 1: Number of connected devices worldwide (Burhan *et al*., 2018)

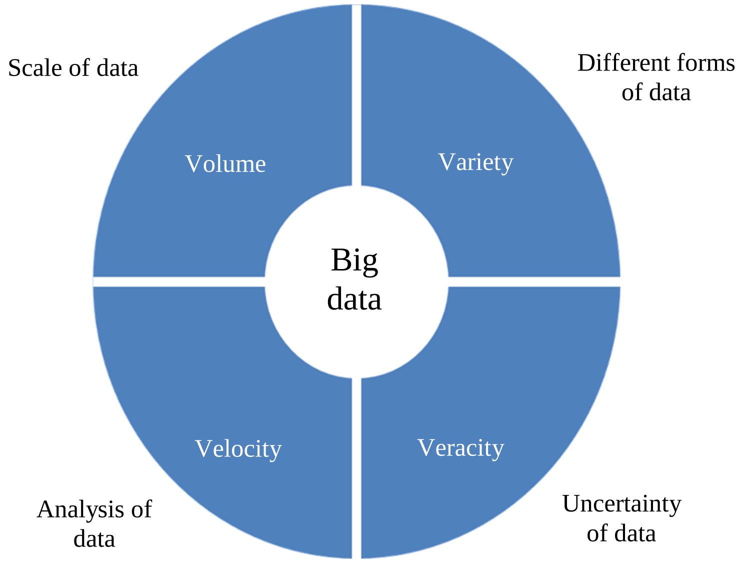

Fig. 2: 4Vs of Big Data (Sagiroglu and Sinanc, 2013)

Figure 2 contains all the 4Vs of Big Data and most sources discuss that the Big Data has four main features 4Vs as explained earlier, but some sources speculate the Big Data with 5, 6, 7Vs. Which are, validity, volatility and value.

Big Data Architecture

The first thing that comes to mind about Big Data is MapReduce and Distributed File System (DFS). Each enterprise derives and uses a different model because Big Data has not only one architecture. The first MapReduce was developed by Google in 2002, Hadoop was developed by Yahoo in 2006 and Hive by Facebook in early 2008. Hive, Impala, Spark, Cassandra, Pig and HBase all use the MapReduce model. So, here we will explain MapReduce and HDFS (Patel and Gandhi, 2018).

Hadoop Distributed File System (HDFS):

Is a file-based system which is written in Java. It has many advantages over its counterparts including a reliable and scalable storage of data. HDFS is designed to reach large-scale clusters such as commodity servers. HDFS demonstrates scalability of constructing a single cluster of up to 200 petabytes of storage and 4,500 servers and supports about one billion files and blocks. HDFS can operate under different systematic and physical situations. Data storage and data processing can be distributed on a high number of servers.

MapReduce

Is a programming paradigm used on clusters for processing high volumes of data using Parallel Distributed Algorithms (PDA). MapReduce is largely separated into two processes:

Map

In this part, the amount of workload is broken into smaller sub-workloads and sets the duties to the Mapper, then computes all unit blocks of data.

Reduce

In this part, the outputs are evaluated and combined to generate a final product. Figure 3 shows the architectures of HDFS and MapReduce.

Data is loaded into HDFS by splitting them into blocks and setting them to the data nodes in the cluster. Then, the blocks are duplicated due to availability in events like a failure. After that, the outputs are saved in HDFS and duplicated according to the settings. Finally, the clients read the results from the HDFS.

Importance of Big Data

Big Data is useful for different types of organizations, institutions, companies and even governments. This is because it can solve many questions that they did not know about before using the capability of Big Data. In other words, it is a point of reference for them to deal with and process huge amounts of data and visualize them according to their needs as mentioned in (Song and Zhu, 2016). At the same time, organizations can identify problems in a more understandable form by using Big Data. With the advancements of Big Data, businesses can move much faster in developing their strategy and become more efficient. It also helps them to eliminate problem areas as well by studying the data they have.

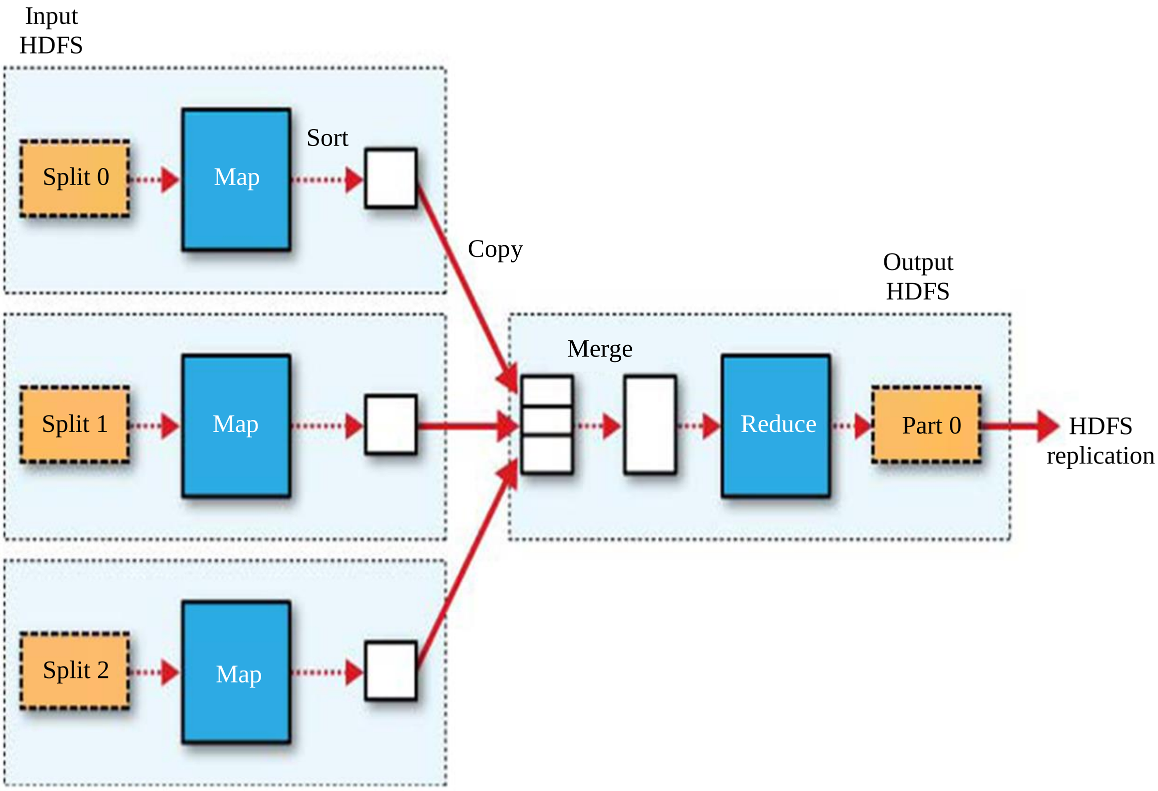

Fig. 3: HDFS and MapReduce architectures (Bakshi, 2012)

Big Data enables:

- Reducing costs
- Saving time
- Developing new products
- Making smart decisions

If Big Data is computed and evaluated in an effective and well-organized way, companies will have a better understanding of their trade, clients, products and their competitors in the marketplace. Of course, that can result in effective improvements, increase sales, reduce expenses, improve products and services (Pence, 2014).

Netflix can be mentioned as a good example where Big Data solutions helped them to develop their business and get viable gains for them. The Customer experience is very important for business owners and companies. Netflix got a better customer experience by using Big Data, to ensure that the customers stay watching its programs. Netflix is continuously analyzing shifts in:

- Program viewership
- Content customers are consuming
- The colors of the promotional visuals of its programs
- Devices that its clients are watching its programs on

For most of the media agencies, gaming and technology companies, analyzing big sets of data are a way to keep their customers, improve advertising, studying the geographical distribution of their users to serve relevant contents, showing contents according to day and night times in different countries (Bakshi, 2012). By using Big Data, Netflix has the capability to deliver the right content when subscribers want to watch.

Sources of Big Data

In this part some sources of Big Data are mentioned that generate data continuously (Sagiroglu and Sinanc, 2013):

- Social Media and Social Networks: We are all generating and using data through social networks and from the social networks we are building a world of data
- Sensor Technology: Measuring all kinds of data
- Networks: Collecting and publishing data from anywhere you are through Cables or Wi-Fi technologies
- Astronomy and Satellites: Collecting all sorts of data nearby or remotely and sharing them
- Medical: Capturing data from all diagnostic instruments and tools

- Mobile Devices: Reading and writing data and creating new datasets through apps
- Microphones: Broadcasting purposes
- Readers and Scanners: Reading objects and data
- Programs or Software: Used for a variety of tasks. It is not limited to only particular functions
- Cameras: Tracking objects for personal, commercial and security purposes

In the past, data generated from users and devices were very few, but now the statistics show a very different model, in which all devices and people are producing and consuming data. The ways of generating data cannot be counted easily, so the data are no longer applicable and has changed its form into Big Data (Sagiroglu and Sinanc, 2013).

Technologies of Big Data

Most of the companies can get benefits from generating useful information which can help them to manage business solutions and become more productive (Sagiroglu and Sinanc, 2013).

Below are some of the technologies and tools for handling and managing Big Data:

Storing:	Simple Storage Service (S3), Hadoop Distributed File System (HDFS)
Processing:	Apache Hadoop, Hive, Pig, Jaql, Apache HBASE, Cassandra, etc
NoSQL databases:	Key-value, document and graph databases
Servers:	Google App Engine (GAE), Microsoft Azure, App Engine, Elastic Cloud 2 (EC2), etc

Sagiroglu and Sinanc (2013; Assunção *et al*., 2015; Chen and Zhang, 2014) describes some of the techniques of Big Data as below:

A. Parallel Computing

It is Processing data simultaneously on multiple machines, each uses its own OS, memory, computing speed and processing on various parts of the data. Therefore, parallel computation helps to decrease time of analyzing and processing Big Data.

B. Distributed File System

There are several central servers that have files which can be accessed with appropriate authentication privileges from several remote clients on the network. The location of the files is tracked continuously using mapping scheme and HDFS. The files that are retrieved from the clients on the servers appear as regular files on the machine of the client and the user and it can be manipulated or modified in the same way it is saved on the machine locally. After the user finishes processing and modifications on the file, the file is sent back to the server through the network. The server will save the modified file for later use.

C. Apache Hadoop

Apache Hadoop is a software project that has been open sourced from the beginning and it has features of distributed processing of large data and datasets on commodity servers' clusters. Apache Hadoop is scalable; in other words, it can be expanded from one server up to several thousands of servers or machines. The clusters' application layer is able to detect and troubleshoot the failures instead of relying on high-end hardware.

D. Data Intensive Computing

It is a set of parallel computing applications that handle Big Data using parallel data computing approach. The core of this task is based on the principles of data and programs, or algorithms that have been used to perform the computation. Distributed and parallel systems of interconnected standalone computers are used for processing and analyzing Big Data. Both distributed and parallel systems work together as a single integrated computing resource.

E. Batch Processing Tools

Apache Hadoop is among the most popular and powerful batch process-based tools for Big Data operations and computations. It has a robust platform and infrastructure for other particular Big Data tools and applications. Various systems are developed and built on Hadoop platform and they are used in different fields such as machine learning and data mining for the purpose of business and commercial applications.

Applications of Big Data

As the fields related to Big Data are continuously increasing, it plays an important role and affects directly on the strategy of marketing in the huge businesses. So, the companies that miss this opportunity will also miss the future innovations and productivity. All the technologies and tools associated with Big Data can be used to boost and increase production efficiency of the company and it can be a useful tool to generate and develop new data-driven services and products. As a result, Big Data applications are opening a new era in all fields. Some of the examples of Big Data usages and applications for numerous industries are mentioned below (GreyCampus, 2019).

Telecommunications:

- Revenue assurance and price optimization

- Customer personality prevention
- Campaign management and customer loyalty
- Call detail record analysis
- Network performance and optimization
- Mobile user location analysis
- Homeland Security

Financial services and Frauds:

- Trading Analytics
- Compliance and regulatory reporting
- Risk management and analysis
- Fraud detection and security analysis
- High speed arbitrage trading
- Trade surveillance
- Abnormal trading pattern analysis
- Multi-channel sales
- Cross-channel analytics
- Recommendation engines using predictive analytics

Retail and Consumer:

- Merchandizing and market basket analysis
- Campaign management and customer loyalty programs
- Supply chain management and analytics
- Event or behavior-based targeting
- Market and consumer segmentation

Web and Digital Media:

- Large-scale clickstream analytics
- Ad targeting, analysis, forecasting and optimization
- Abuse and click-fraud prevention
- Social graph analysis and profile segmentation
- Campaign management and loyalty programs

Health and Life Sciences:

- Smart Healthcare
- Clinical trials data analysis
- Disease pattern analysis
- Campaign and sales program optimization
- Patient care quality and program analysis
- Medical device and pharmacy supply chain management
- Drug discovery and development analysis

Future of Big Data

Nowadays, we are getting and generating more data and information than in the past. We can use these data in many aspects. Quantitative change can be converted to qualitative change. Having more data enables new things that were previously impossible. So, in short; it's no longer only raw data. We have more new data, better data and more different data (Assunção *et al*., 2015).

We are still facing the fact that we have limitations on what we can get from the available data and what we can do with that data. However, most of our beliefs and assumptions about the difficulty of data processing need to be thoroughly reviewed. An incredible tremor is going to happen as the Big Data overflows through society, politics, business and industrial sectors. People shape their tools and their tools shape them as well. There are positive sides of Big Data and it also brings more advantages and chances to society. During the industrial revolution, this was certainly true, it was a destructive period of shifts, but ultimately it led to a better life (Cukier, 2019).

A new economic, political and philosophical change and movement occurred after the industrial revolution. The impact of robots, Big Data, computers and the Internet on the social and political movements and the effect that these technologies have on the economy cannot be predicted and making any assumption is not logical. Currently, all the discussions on income disparity and occupation movement seems to be headed toward that direction (Cukier, 2019).

Privacy is the second main policy that was a problem even in the past and for "small data", but it became a bigger challenge in the Big Data century. The nature of protecting personal information and data will change if potential breaches and threats of privacy occur 1,000 times per second rather than every day or every hour. It also varies when the act of gathering and collecting information occurs not so clearly and positively, as byproducts of other services and passively. It is difficult to accept the fact that how the classical privacy protection law works in the world, or even how the privacy-compromised people take action or even recognize the situation (Cukier, 2019).

It becomes terrible as the foundation of the personal information protection law that's available today in the world is the principle guaranteed by the privacy guidelines of the Organization for Economic Cooperation and development (OECP) and the enterprise needs to destroy the data once its main purpose is achieved. However, the important point of Big Data is that you do not know all the useful ways that you might be using today, so you need to store data permanently (Assunção *et al*., 2015).

As a result, regulations governing Big Data need regulators that understand they can be much the same, even more. Indeed, today's rules have less impact on securing and privacy protection. In short, advancing the existing mediocre policy a little more is meaningless. Instead, Big Data companies are seeking new, better and of course different regulations.

Big Data will affect and modify businesses, self-pace education, security, etc., which is all together can modify and directly affect the society. It is obvious that the advantages will exceed the shortcomings, but it is almost a hope and not a complete fact. The world of Big Data is still in its early stages and as a society we are not ready to deal with all the data that has been collected right now. For now, we cannot predict the future. No wonder today's technologies and trends will continually surprise us just as ancient people with abacus dreaming of smartphones and we cannot easily understand the advantages of this little device, but when he understands the first that comes to mind is that "What sufficient is that!". More it will not be more: It will be different. (Michael and Miller, 2013).

Challenges of Big Data

Big Data can produce very useful information, but there are also new challenges concerning the amount of data to store, the cost, the safety of data and the maintenance period. For example, Governmental agencies, industries, businesses and companies are continuously and increasingly dependent on video and graphic data for monitoring and criminal investigation. Security cameras are located throughout most of the governmental places and public areas. In (Assunção *et al*., 2015) multiple cameras available on police cars for tracking and recording traffic jams and congestion and dash cams for handling complaints. Most of the organizations are running tests and making experiments with video cameras that are currently wearable to store incident moments and collect evidence directly from crime scenes for later use in the courts. Because, many of these devices and machines can quickly generate huge amounts of data, it can take a long time to save and take time to process, so the operator will either execute continuously or capture only distinctive images or scenes that are considered (Michael and Miller, 2013).

Big Data can also bring new challenges and ethical difficulties. Companies use Big Data to understand and learn more about employees, boost productivity and develop innovative business processes however, these developments are costly. Keeping track of all employee actions and continually measuring performance relative to industry benchmarks brings about a level of monitoring that can destroy the human spirit. Such surveillance may bring great benefits for the company, but it will not always be the best profit for those who make up the company (Villars *et al*., 2011).

In addition, as large multimedia data sets become more common, the boundary between private and public spaces becomes ambiguous. The new online apps will quickly incorporate recent devices and wearable devices in the form of digital watches or glasses, as well as allowing users to upload videos via social media networking, as well as to enable the continuous audio-visual capture. People will essentially be cameras. This publicly available data will dwarf things generated by today's CCTV cameras (Chen and Zhang, 2014).

Conclusion

As discussed above, it may be predicted that by 2025 Big Data will be used to recreate, eliminate and digitize 80% of the business process. Therefore, you cannot miss the impacts of Big Data on different domains of businesses. In order to see and realize the power of Big Data, organizations need to determine and identify how to use their company's data to build analytical and business reports. Looking at all the latest major market analysis and researches, it can be inferred that the service market is expected to rise about 35% by 2021 compared to 2017. As Peter Sondergard, Senior Vice President of Gartner Research stated that," Information is the oil of the 21st century and analytics is the combustion engine".

Acknowledgment

We are highly thankful for all the sources we have gotten data and instructions even a statement.

Author's Contributions

Wisam A. Qader: Proposing the idea, collecting the resources and writing the article.

Musa M. Ameen: Collecting part of the resources, writing and reviewing the article.

Bilal Ismael Ahmed: Writing part of the article, improving and final reviewing of the article.

Ethics

This study is self-contained and the authors confirm that they have read and approved this document and there are no any ethical issues involved.

References

Assunção, M.D., R.N. Calheiros, S. Bianchi, M.A. Netto and R. Buyya, 2015. Big data computing and clouds: Trends and future directions. J. Parallel Distributed Comput., 79: 3-15. DOI: 10.1016/j.jpdc.2014.08.003

Bakshi, K., 2012. Considerations for big data: Architecture and approach. Proceedings of the Aerospace Conference, Mar. 3-10, IEEE Xplore Press, Big Sky, MT, USA. DOI: 10.1109/AERO.2012.6187357

Burhan, M., R.A. Rehman, B. Khan and B.S. Kim, 2018. IoT elements, layered architectures and security issues: A comprehensive survey. Sensors, 18: 1-37. DOI: 10.3390/s18092796

Center, I.I., 2012. Peer research: Big Data analytics. Intel's IT Manager Survey on How Organizations Are Using Big Data.

Chen, C.P. and C.Y. Zhang, 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inform. Sci., 275: 314-347. DOI: 10.1016/j.ins.2014.01.015

Cukier, 2019. Big Data and the future of business.

Gerhardt, B., K. Griffin and R. Klemann, 2012. Unlocking value in the fragmented world of big data analytics. Cisco Internet Bus. Solutions Group.

GreyCampus, 2019. Applications of Big Data.

Madden, S., 2012. From databases to big data. IEEE Int. Comput., 16: 4-6. DOI: 10.1109/MIC.2012.50

Michael, K. and K.W. Miller, 2013. Big data: New opportunities and new challenges [guest editors' introduction]. Computer, 46: 22-24. DOI: 10.1109/MC.2013.196

Patel, H.B. and S. Gandhi, 2018. A review on big data analytics in healthcare using machine learning approaches. Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, May 11-12, IEEE Xplore Press, Tirunelveli, India, pp: 84-90. DOI: 10.1109/ICOEI.2018.8553788

Pence, H.E., 2014. What is big data and why is it important? J. Educ. Technol. Syst., 43: 159-171. DOI: 10.2190/ET.43.2.d

Sagiroglu, S. and D. Sinanc, 2013. Big data: A review. Proceedings of the International Conference on Collaboration Technologies and Systems, May 20-24, IEEE Xplore Press, San Diego, CA, USA, pp: 42-47. DOI: 10.1109/CTS.2013.6567202

Singh, S. and N. Singh, 2012. Big data analytics. Proceedings of the International Conference on Communication Information Computing Technology, Oct. 19-20, IEEE Xplore Press, Mumbai, India. DOI: 10.1109/ICCICT.2012.6398180

Song, I.Y. and Y. Zhu, 2016. Big data and data science: What should we teach? Exp. Syst., 33: 364-373. DOI: 10.1111/exsy.12130

Villars, R.L., C.W. Olofson and M. Eastwood, 2011. Big data: What it is and why you should care. White Paper IDC, 14: 1-14.