

Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data

Suhail Sami Owais

Dept. of Computer Science
Applied Science Private University
Amman, Jordan

Nada Sael Hussein

Master of Computer Science
Amman, Jordan

Abstract—There is an exponential growth in the amount of data from different fields around the world, and this is known as Big Data. It needs more data management, analysis, and accessibility. This leads to an increase in the number of systems around the world that will manage and manipulate the data in different places at any time. Big Data is a systematically analysed data that depends on the existence of complex processes, devices, and resources. Data are no longer stored in traditional database storage types or on such forms like database, which only on structured data are limited, but surpassed them to the unstructured or semi-structured data. Thus, Big Data has several characteristics and specific properties proportionate with the size of the data with the enormous and rapid development in all business and areas of life. In this work, we study the relationship between the characteristics of Big Data and extract some categories from them. From this, we conclude that there are five categories, and these categories are related to each other.

Keywords—Big Data; Characteristics; Categories; Management; Analysis; Anywhere and Anytime

I. INTRODUCTION

Big Data refers to technologies and initiatives that involve data that is too diverse, rapidly changing or massive for conventional technologies, skills and infrastructure to address efficiently. In other words, the volume, velocity, or variety of data is too great [1]. Big Data requires new technologies with a spatial architecture so that it becomes possible to extract value from it by capturing and analysis process[2]. Due to such large size of data (rising up information yields an increase in the amount of data), it becomes very difficult to perform effectively and to analyse using the existing or the known traditional techniques [3].

Big Data due to its various properties like volume, velocity, variety, variability, value, and complexity puts forth many challenges [4]. Since Big Data is an upcoming technology in the market which can provide many benefits to the business organizations, it becomes necessary that various challenges and issues associated in investigating and adapting to this technology are brought into light.

The main data categories are: the Traditional Database category and the Big Data.

A. Traditional Database:

A database is a set of relevant data by data, which means known facts that can be recorded and have implicit meaning. For example, names; telephone numbers, and addresses of the people you know. You have recorded these data in an indexed

address book or stored them on a hard drive, using a personal computer and software such as Microsoft Access or Excel. The collection of the related data with an implicit meaning is known as a database [5].

B. Big Data:

Big Data is the term used to describe huge amount of structured and unstructured data which is large. It was very difficult to process the data in the Big Data using the traditional databases and software technologies [6]. Also Big Data is companies who had to query loosely structured very large distributed data [7].

This research consists of five sections. The first one is an introduction about Big Data and its category. The second section discusses how Big Data becomes growth and from where it comes to be called as Big Data. It also demonstrates how the authors categorize the characteristics of Big Data into five categories and discuss them. The third section explains what are the benefits of Big Data. The fourth section demonstrates that Big Data can be found anywhere, anytime, and in anyplace. Finally, the last section offers conclusion and the future work.

II. BIG DATA

An exponential growth in the amount of data from all different types of data becomes enormous data known as Big Data. Thus, Big Data is a massive set of structured and unstructured and semi-structured data [8]. It is hard to manage the data in the Big Data using any of the traditional applications [9]. Big Data consists of very large datasets that will be processed by any traditional database system tools [6].

Collecting data in Big Data is done not only from the traditional corporate database only, but includes other sources as well:

1) *Data collected from Sensors*: Sensors are becoming an increasingly important component of the information stored and processed by many businesses. For this a lot of data were collected from different fields of sensors. This includes data from Fixed-Sensors such as home automation; traffic sensors; traffic-webcam sensors; scientific sensors; security-monitoring videos or images and from weather-pollution sensors as well as data from Mobile-Sensors such as data from GPS Sensors; mobile phone location, satellite images

2) *Data collected from Machines*: Where there are a lot of data collected from machines in which it is collected from

sensor data, it is known as complex data. For example: Video from security cameras, Recording voice from microphone, Mail to Mail Log Files, Satellite Imaging and Bio-Informatics.

3) *Data collected from Human:* Several data were collected from humans. This included data from human enterprise contents and from external sources and from Documents, E-mail, Web Logs and social networks like Facebook; LinkedIn; Twitter; Instagram; Flickr, and Picasa.

4) *Data collected from Business Process:* Data were collected from industry, production, refining, distribution, and from Marketing. For example, data produced by public agencies like that from medical records, data produced by businesses such as commercial transactions, and from banking records and from E-commerce and credit cards.

Big Data has many characteristics or properties mentioned by nV's characteristics [8]. Set of V's characteristics of the Big Data were collected from different researchers' publications to have Nine V's characteristics (9V's characteristics). These 9V's characteristics are: (Veracity, Variety, Velocity, Volume, Validity, Variability, Volatility, Visualization and Value).

We categorize these characteristics of the Big Data into five categories CPIVW. They are: Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data.

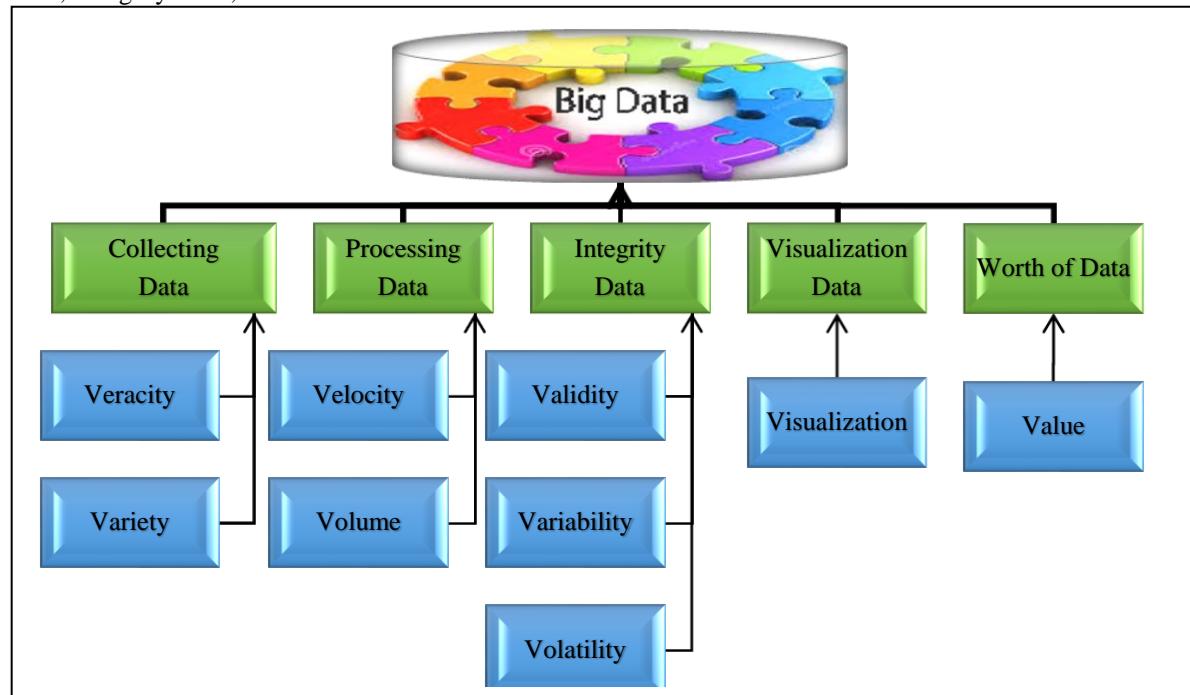


Fig. 1. Five Categories CPIVW of the Big Data with their 9 V's Characteristics

- Veracity: Big Data veracity refers to the biases, noise, and abnormality in data. The data that is stored, and mined meaningful to the problem that is being analysed, in addition the developers ask the question "Is the data that is being stored, and mined meaningful to the problem being analyzed or not?" [6]. The veracity not only talks about the quality of data, but when the users begin using the Big Data, they also become truly engaged and are more willing to invest in efforts to clean up data ideally at the source [10].
- Variety: Structured, semi-structured, and unstructured data besides text and more data types have emerged, such as record, log, audio, and hybrid data [11]. Currently, data comes in all types of formats such as emails, video, audio, transactions, and images. In other words, the variety is structure, semi-structure, and unstructured data in same place [12].

2) Processing Data:

Figure 1 demonstrates the Big Data CPIVW categories and their characteristics. The CPIVW five categories will be discussed in the next section.

2.1 Five Categories CPIVW of the Big Data and their 9 V's Characteristics:

Because of the relationship between some of the characteristics, they are all classified into different categories. The characteristics of the Big Data they clustered into groups as categories with respect to the relationship between them. Thus, we extract five categories CPIVW (Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data) from these 9V's characteristics of the Big Data.

The extracted five categories CPIVW (Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data) are discussed below:

1) Collecting Data:

Data is collected from different resources and different types to have Big Data as a massive set of structured data and unstructured data and semi-structured data. For this veracity and variety characteristics for the Big Data have such relationship between them. So, they grouped together to form collected data category.

Processing Data comes from grouping the two main V's characteristic of Big Data, velocity and volume, because there exists such a relationship between them. The processing data category talks about speed of processes on the data fit with size of Big Data. The data in processing data category passes through many of the processes and is processed on demand. Here is a description of the two Vs, velocity and volume:

- Velocity: The created information at faster pace than before, in which the different channels of Big Data increase the output content. This property means how fast the data is to be produced and processed to meet the demand [6].
- Volume: It is predicted that the data volume worldwide will reach 40 ZB by 2020 [11]. Storing different type of data from social network is possible in Big Data storage devices. For this, the amount of data is known as volume of data, where the amount of data continues to explode. Thus, to improve the company's for archiving, and tiered data importance strategies to accommodate the new volumes? There are many factors which contribute to increasing the volume streaming data and data collected from sensors and other resources. The large volume of Big Data is the primary goal of consumers to optimize future results [6]. From this, the main goal from the large volume of Big Data is to make it useful for users and consumers and optimize future results. Thus, such a programming model (MapReduce) that associated implementation for processing and generating large data sets [13].

3) *Integrity Data:*

Validity, variability, and volatility are the three Vs grouped together to categorize the data integrity. Data integrity refers to the accuracy and consistency of data stored in Big Data, and talks about the truth of data and guarantees the data as correct and not manipulated. Integrity data also ensures the quality of the data in the Big Data.

- Validity: As such, Big Data veracity is a matter of validity, meaning that the data is correct and accurate for the intended use. Clearly valid data is the key for making the right decisions [6]. Data validation is one that certifies uncorrupted transmission of data.
- Variability: Along with the velocity, the data flows may be highly inconsistent with periodic peaks, daily, seasonal, and event-triggered peak data loads can be challenging to manage, especially with unstructured data involved [14].
- Volatility: When we talk about volatility of Big Data, we can easily recall the retention policy of structured data that we implement every day in our businesses [15]. Once retention period expires, we can easily destroy it. For example: an online ecommerce company may not want to keep a one year customer purchase history. Because after one year and default warranty on their product expires so there is no possibility of these data restore ever [15].

4) *Visualization Data:*

Visualization data refers to a way to explore and understand your data, in the same way that the human brain processes information. Visualization Data category contains only the visualization characteristic for Big Data, thus the data is easy to read and analyse from complex graphs. Visualizing the data in the Big Data leads the people to understand the meaning of different data values faster when they are displayed in charts and graphs rather than reading about them from reports.

- Visualization: It is the hard part of Big Data which makes all that huge amount of data comprehensible and easy to understand and read. With the right analysis and visualizations, raw data can be put to use, otherwise raw it remains essentially useless [16]. Visualization means complex graphs that can include several variables of data while still remaining understandable and readable [17].

5) *Worth of Data*

Big Data characteristic value is the worth of data category. Worth of data talks about the cost and management of data in Big Data. Small data can have more value than a corresponding Big Data collection, and the Big Data has a high economic value. The collected data comes with many noises. While collecting data, it may contain some noises, and thus it must be filtered from any noise to help the user to analyse and take decision.

- Value: It has a low-value density as a result of extracting value from massive data. Useful data needs to be extracted from any data type and from a huge amount of data [8]. For this we must look for true value of data, in which data value must exceed its cost or management. Thus, attention must be paid to the investment of storage for data. Storage may be cost effective and relatively cheaper at the time of purchase but such underinvestment may damage highly valuable data [18]. For example storing clinical trial data for new drug on cheap and unreliable storage may save money today but can put data on risk tomorrow [19].

2.2 *Big Data CPIVW Categories Hierarchy*

After CPIVW has been clarified, the 9V's characteristics of Big Data were grouped in clusters to get the five categories CPIVW. And from this we present them in a hierarchy model as shown in Figure 2. Here it is shown that the five categories CPIVW have some dependency to get the Big Data in such manner.

III. SOME BENEFITS FROM THE BIG DATA

Distributing data around the world via Internet and the increasing amount of data tends to become vast and huge data. Thus, a lot of data can be created on a single day itself, and from where these data's collected form where. It comes from around the world and it may be more than 2.5 quintillion byte from different resources. This yields to begin most of the companies with different sizes to gain the benefits of data technology and analytics. While if your company was not up to speed so far? For this, at hand there are several ways that Big Data can help with several benefits your business. But, may or may not there is/are such obstacles that can thwart the Big Data plans.

In Information technology (IT), the executives continually evaluate the technology trends that will impact their business. Some simply deploy technology to promote the goals stipulated in business plans. Others take on the role of chief innovation, creativity and renewal officer and introduce different models of using existing data to generate new revenue and gain insight into who clients are and what they want.

Any IT organization considering a Big Data initiative should consider these five major points [20], which will bring clarity as well as revenue and benefits to a company.

Several benefits from the Big Data that may help the business are discussed as follows:

- Management Data in Big Data using an intelligent tools will be better which enable many use cases. While data may be founded in different formats that it has been successfully mined for specific purposes.
- Internet benefits. The benefit use of cloud storage (storing of data online in the cloud or in "the Internet"), and the benefit use from cloud services providers (servers, storage and applications through the Internet). Both benefits help the business from storage and the cloud computing power. When files are stored in the cloud, they can be accessed at anytime from anywhere via Internet and accessed with remote backups of data.
- Visualizing data from Big Data using business intelligence software is presented in a simple way to read and analyse. This software must be able to provide the processing engines that allow the end users to query and manipulate information quickly evenly in real time in some cases.
- Capabilities will evolve the Business data analysis methods for data from the Big Data. In which a structured data; unstructured data and semi structured data are grouping as with different types of data to form the Big Data that forms the data in such business. Some examples are text file and audio and video files. Thus, an intelligent tool needs to be used for recognizing a specific pattern based on such criteria. A natural language processing tool can prove vital to text mining, sentiment analysis, and entity recognition efforts.

IV. BIG DATA ANYWHERE-ANYTIME-ANYPLACE

Big Data yields to explore such technology that will permit and control the enormous data founded around the world. These data can be accessed via Internet from anywhere at any time and from any platform. One of the most important technologies were founded is called Cloud Computing. Big Data does not refer only to the huge amount of data, but also it refers to the computational term in a large number of fields either economic; business; medical; researches and any other fields. Because of rising in the amount of data in the Big Data which yields to increase the demands for accessing the data resources from anywhere at any time from anyplace.

Several challenges must be noticed here instead of using just single computer and accessing the data from only one storage device. Thus a software's applications must be founded

and abled to execute on anywhere over the networked devices. Via internet where each device in the networked devices may join or leave the shared data scope at anytime from anywhere at any place. Thus, for processing data in the Big Data the founded software may use concurrency to take advantage of multiple foundations.

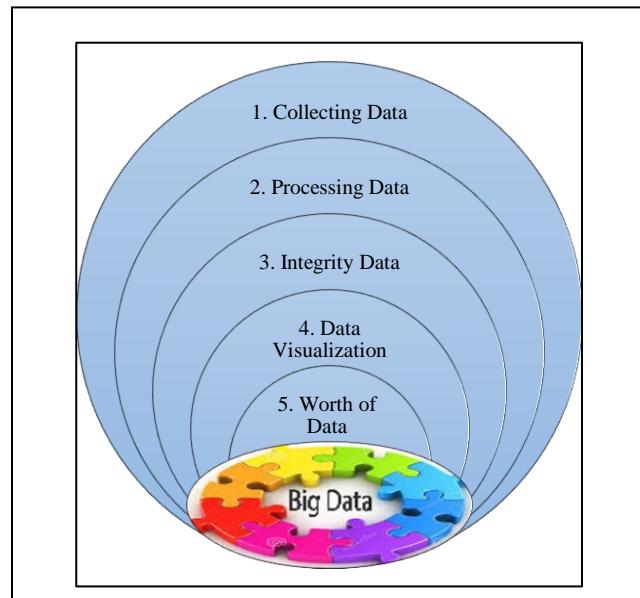


Fig. 2. Five Categories CPIVW Hierarchy for the Big Data

V. CONCLUSION AND FUTURE WORK

With the increasing amount of data, it has become difficult to be handled using such traditional tools, as that used for data bases. Big Data appears to coincide with the tremendous development in all areas of the life that led to the aggravation of the size and diversity of data sources such as data collected from Sensors; Machines; Human; and from Business Process. Data from different sources cause Big Data to be a massive set of structured and unstructured and semi-structured data. For this, it becomes too hard to manage the data in the Big Data by using a traditional applications.

Big Data has 9V's characteristics (Veracity, Variety, Velocity, Validity, Variability, Volatility, Visualization and Value). The 9V's characteristics were studied and taken into consideration when any organization need to move from traditional use of systems to use data in the Big Data.

The relationships between Big Data characteristics are studied and we extract five categories by assortment the 9V's characteristics of the Big Data, also the five categories known as CPIVW and they are (Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data).

Last but not least, it must be said that Big Data needs to be an environment fit with the size and data types. This environment is a cloud computing environment in which Big Data emerged and the type we will look at the next paper to the mechanism of Big Data work and how they are managed and analyzed within the cloud computing environment and what types of that fit better than others with large data and taking

into consideration several aspects of things, and we will look to in a timely manner.

ACKNOWLEDGMENT

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research project (Grant No. DRGS-2015-2016-61).

REFERENCES

- [1] Big Data, mongoDB White Paper, available at: <https://www.mongodb.com>, February 2015.
- [2] K. Li, H. Jiang, L. T. Yang, A. Cuzzocrea, "Big Data: Algorithms, Analytics, and Applications", 1st. edition, Taylor & Francis website, pp. 91-104, 2015.
- [3] S. Gupta, M. S.Chaudhari, "Big Data Issues and Challenges: Data analysis, storing, processing, issues, challenges and future scope", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 3, Issue 2, pp. 62-67, February 2015.
- [4] L. V. Satyanarayana, "A Survey on Challenges and Advantages in Big Data", International Journal of Computer Science And Technology, vol.6, Issue 2, April-June 2015.
- [5] R. Elmasri, S. B. Navathe, "Fundamentals of database systems," 6th. edition, Adeson-Wesley publishes, pp. 3-24, New York, 2011.
- [6] Vangie Beal, "Big Data", Available at: http://www.webopedia.com/TERM/B/big_data.html, February 2016.
- [7] A. Katal, M. Wazid, and R. H. Goudar, "Big data: issues, challenges, tools and good practices," in Proceedings of the 6th International Conference on Contemporary Computing (IC3 '13), pp. 404–409, IEEE, 2013.
- [8] Nour E. Oweis, Suhail S. Owais, Waseem George, Mona G. Suliman, and Václav Snášel. A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses). Intelligent Data Analysis and Applications, Springer Cham Heidelberg New York Dordrecht London. Volume 370, 2015.
- [9] A. Syed, K. Gillela, and C. Venugopalnd, "The Future Revolution on Big Data", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, Issue 6, June 2013.
- [10] Patricia Saporito, "2 More Big Data V's — Value And Veracity," Digitalist Magazine by SAP, Available at: <http://www.digitalistmag.com/big-data/2-more-big-data-vs-value-and-veracity-01242817>, 2014.
- [11] Full lifecycle management of Big Data White Paper, available at: www.huawei.com, 2013.
- [12] Infographics, "The four V's of Big Data," IBM Big Data & Analytics Hub, Available at: <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>, February 2016.
- [13] J. Dean, S.Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communication of ACM-50th anniversary issue: 1958-2008, vol.51, no. 1 , pp. 107-113, Jan 2008.
- [14] R. Toshniwal, K.G. Dastidar, and A. Nath, "Big Data Security Issues and Challenges," International Journal of Innovative Research in Advanced Engineering (IJIRAE), vol. 2, issue 2, February 2015.
- [15] M.A. Khan, M.Fahim Uddin, and N. Gupta, "Seven V's of Big Data," Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1), Bridgeport, Connecticut, USA, pp.3-5, April 2014.
- [16] V'S of Big Data, Available at: <http://www.godatify.com/category/tech-blog/big-data/>
- [17] Mark van Rijmenam, "Why The 3V's Are Not Sufficient To Describe Big Data,", Datafloq , Available at: <https://datafloq.com/read/3vs-sufficient-describe-big-data/166>.
- [18] V.S.Thiyagarajan, K.Venkatachalapathy, "Isolating Values from Big Data with the Help of Four V's", International Journal of Research in Engineering and Technology, vol.4, issue 1, January 2014.
- [19] J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, "Big data for Dummies," John Wiley & Sons, Inc. publishes, Hoboken, New Jersey, 2013.
- [20] Reda Chouffani, "5 Reasons to Move to Big Data (and 1 Reason Why It Won't Be Easy)," Serving Chief Information Officers (CIOs), Available at: <http://www.cio.com/article/2385690/big-data/5-reasons-to-move-to-big-data--and-1-reason-why-it-won-t-be-easy-.html>, 2013.