The 10 Vs, Issues and Challenges of Big Data

Conference Paper · March 2018

DOI: 10.1145/3206157.3206166

6 authors, including:

Nawsher Khan
Abdul Wali Khan University Mardan
35 PUBLICATIONS **669** CITATIONS
SEE PROFILE

Mohammed Alsaqer
King Khalid University
5 PUBLICATIONS **87** CITATIONS
SEE PROFILE

Solmaz Salehian
Oakland University
13 PUBLICATIONS **74** CITATIONS
SEE PROFILE

The 10 Vs, Issues and Challenges of Big Data

Nawsher Khan [1, 2], *Mohammed Alsaqer* [1]
[1] College of Computer Science, King Khalid University Abha, Saudi Arabia
[2] Department of Computer Science, Abdul Wali Khan University, Pakistan
nawsherkhan@gmail.com, msalsaqr@kku.edu.sa

Habib Shah [1], *Gran Badsha* [1], *Aftab Ahmad Abbasi* [1]
[1] College of Computer Science, King Khalid University Abha, Saudi Arabia
hurrahman@kku.edu.sa, gran16178kku@gmail.com, aftab_a_abbasi@hotmail.com

Soulmaz Salehian[3]
[3] Department of Computer Science & Engineering, Oakland University, Rochester, MI 48309-4401, United States
ssalehian@oakland.edu

***Abstract* — In this emerging computing and digital globe, information and Knowledge are created and then collected with a rapid approach by wide range of applications through scientific computing and commercial workloads. Over 3.8 billion people out of 7.6 billion population of the world are connected to the internet. Out of 13.4 billion devices, 8.06 billion devices have a mobile connection. In 2020, 38.5 billion devices will be connected and globally internet traffic will be 92 times greater than it was in 2005. The use of such devices and internet not only increase the data volume but the velocity of market brings in fast-track and accelerates as information is transferred and shared with light speed on optic fiber and wireless networks. This fast generation of huge data creates numerous challenges. The existing approaches addressing issues such as, Volume, Variety, Velocity and Value in big data research perspective. The objectives of the paper are to investigate and analyze the current status of Big Data and furthermore a comprehensive overview of various aspects has discussed, and additionally has been described all 10 Vs' (Issues) of Big Data.**

CCS CONCEPTS

- Data → Big Data → Computer Systems → Database Management System → Data Issues & Challenges.

Keywords— Big Data, Data Management.

I. INTRODUCTION

Over 3.8 billion people (50%) out of 7.6 billion population of the world (at the end of 2017) are connected to the internet. Out of 13.4 billion devices, 8.06 billion have a mobile connection. In 2020, 38.5 billion devices will be connected to Internet and globally traffic will be 92 times greater than it was in 2005. IDC predicts that data generation will be 44 times more in 2020 than this generation was in 2009. It is evident with the introduction of big data, cloud computing, and Internet-of-Things (IoT).

Growth of the data generation has shown the significant upward trend rapidly. In 2010, our globe created over 1ZB of data and the data generation was rose to 7ZB by 2014 [1]. International Data Corporation (IDC), a technology research firm, estimated that data increase is doubling every two years. Every second more data over the Internet creates than was stored in the entire Internet just 20 years ago [2].

The human beings has always intended to gather data. They have been moving from textual data to richer data including images, videos, and interactive maps as well as associated metadata such as geo-location information and time/date stamps [3]. The rapid rise in technology consequently has led to data overflow. Big Data refers to this overwhelming data flow which demands more sophisticated computational platforms and specific data storages. Big Data is not only about the data itself, but it also involves challenges and capabilities associated with processing, analyzing and storing the large data sets to enhance decision making and to optimize discovery and processing in timely and cost-effective manner. Big Data todays has been becoming critically important for business success and it has introduced a new era of data-driven insights to businesses. Bain & Company [4] surveyed at more than 400 companies about their data, analytics capabilities and their decision making speed and effectiveness. They found out that only 4% of companies are really good at analytics and the group that puts into play the right people, tools, data and intentional focus. These companies are already using analytics insights to improve either the way that they operate or their products and services. The results of this analytics insights show a significant differences for these companies and they are twice as likely to be among the top performers within their industries ; Three times more likely to execute decisions as intended ; Five times more likely to be faster in decision making.

Although Big Data provides many opportunities, it brings many challenges and issues and demands many requirements. In this paper, a wide range of these challenges and opportunities have been described. The rest of the paper is organized as; Section II has drawn a road map for Big Data. Section III has explored the current status of Big Data. Section IV has focused on the theme of this paper and has described the 10 Vs' characteristics, issues and challenges of Big Data.

II. BIG DATA ROAD MAP

There is debate about what the age of Big Data is and when the first time is emerged. Barnes in his article [5] argued that the past for Big Data is potent and proponents ignore it at their peril. He also mentioned Big Data's problematic assumptions and practices first criticized 40 years ago by opponents of geography's quantitative revolution and it did not first emerge in 2008. He believes that Big Data has been made possible because of the particular combination of different elements, each of which has its own history that comes together at this present time. But precisely because these elements in their earlier incarnation have a history and issues can still remain even in the new form. Diebold in [6] investigated that the non-academic origin of Big Data was introduced by John Mashey and others at Silicon Graphics (SGI) in the mid-1990s. But, the first academic reference was by Weiss and Indurkhya in 1998 in computer science and Diebold in 2000 in statistics/econometrics. Furthermore, Chen et al [7] pointed out that the concept of Big Data for the first time was announced in June 2011, by EMC/IDC, published a research report titled Extracting Values from Chaos [8]. Teradata system [9] was presented as the first commercial parallel database in 1979. In late 1990s, the benefits of parallel database were widely recognized in the database field [7]. However, with the advent of Big Data many other problems such as resource management and data analysis arose. Google in 2004, proposed MapReduce [10] as a distributed parallel computing framework planned to process large amounts of data, and in 2003 presented Google File System (GFS) [11] to address the need of Google's data storage system Apache Software Foundation developed Hadoop [12] which was inspired by Google MapReduce. Apache Hadoop was released in 2011 initially and provides MapReduce implementation and a distributed file system called HDFS [13]. Explosion of data and advancing trends in technology open doors to new approaches and techniques to understanding data, making decisions and make business more intelligent.

III. CURRENT STATE OF BIG DATA

A remarkable amount of web-based, highly detailed and contextualized, sensor and mobile data generating and arriving at a Terabyte, Petabyte and even Exabyte scale (The Economist 2010), using by business and organizations for new research, discovery and insights. Although, a large population in emerging globe has no access to internet and digital devices, means they are not yet part of the generation of Big Data. Although, sooner or later, Big Data management issues and challenges will get worse, if not considered seriously today. Big Data has become the vital part for everyone's life and Big Data has hidden solutions for many serious real life problems.

Big Data plays an significant role to understand human and machines as both are the data generating agents. It is identified and recognized as one of the present and future research factors. Big Data is generated every time continuously from all corners of the globe. It is a sharpness to make intelligence and value out of it. Consequently, it is important to understand the characteristics of Big Data before leveraging it in various context. To analyze and discuss various Vs explored by various authors in academic paper will open doors towards understanding the data and finding true value of Big Data. Laney characterized and categorized the concept of Big Data by, volume, velocity and variety, known as 3Vs [26]. Another 4th V to describe Big Data called veracity added by IBM [33]. [27], [32] and more have discussed 'value' as another 5th V of Big Data. [10] and more have added 'variability' as another 6th V of Big data. Our previous paper has described more two 7th & 8th Vs', 'validity' and 'volatility' to define Big Data [35]. Hence, several more academic research papers are reviewed to know and understand the views of different researchers about the characteristics of Big Data.

However, Big Data is still in its infancy stage and has not been reviewed in general. This study comprehensively surveys and classifies the various attributes including its Volume, Velocity, Variety, Value, Variability, Veracity, Validity, Volatility, Viability, and Viscosity of Big Data with more detail

IV. CHARACTERISTICS, ISSUES AND CHALLENGES

High-volume, high-velocity and high-variety of Big Data have revolutionized many aspects of traditional systems for storing, processing and analyzing data and created many new challenges. These 3Vs (volume, velocity and variety) have been specified as main dimensions and characters of Big Data which make it different from traditional data. In this section, discussion focuses on characters of Big Data, as originating data play pivotal roles in capturing, storing and analyzing of data. Figure 1 has described the Characteristics, Issues and Challenges of Big Data and next section has discussed each one briefly.

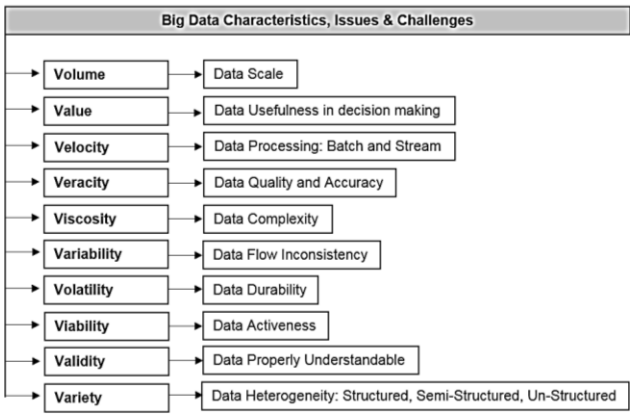

Fig. 1. Big Data Characteristics, Issues and Challenges

A. Volume

Data volume refers to the large data set generated by science and education, business and human interaction records. Data volume plays significant role in storage and processing [14]. However, the storage capacity introduces less challenge in comparison with processing, due to rapidly developing storage technologies on one hand and decreasing the storage price on the other hand. So, cost-effective storage solutions and Cloud technologies provide the opportunities for organizations to store data. Data volume, however, has a fundamental influence on data processing, data management and decision making. Because the size of data has been growing faster in comparison with the computational power of processing system [15], [16].

B. Value

Data value refers to data usefulness in decision making and is one of the most significant factors in Big Data, because it has direct impact on business profits. For instance, McKinsey [20] estimated a range of health-care initiatives and asserted the potential impact could account for 300 billion to 450 billion in reduced health-care spending, or 12 to 17 percent of the 2.6 trillion baseline in US health-care costs if the early successes were scaled up to create system-wide impact. Also, [21] pointed out that organizations agreed the use of business information and analytics differentiates them in industry as top performers and low performers. Thus, the value lies in the meticulous analysis of accurate data.

C. Velocity

Data velocity refers to the speed at which data is generated. There are two ways that data can be imported: batch data, where the dataset is loading all at once, and streaming data where data flow is continuously imported and processed while it is being generated. Stream processing is key for choosing Big Data analysis platform, because real time process most often is time sensitive and demands faster and near-instant analytics results. For example, Hadoop performs well for processing archive data in a batch model, while Apache Spark is more efficient for real time and interactive job analysis [17].

Sometimes 5 seconds is too late. For time-sensitive processes like catching fraud etc. big data must be used as it streams into the enterprise in order to increase its value. Five million trade events and actions are scrutinized to identify potential fraud daily. Five hundred million daily call detail records are being analyzed in real-time to predict customer churn quickly. Velocity has the categories of Batch, Near Time & Real Time.

D. Veracity

Data veracity focuses on data quality and accuracy and defines how data can be trusted when important decision needs to be made regarding the collected data. Data is categorized in good, bad, or undefined data which could be because of data inconsistency, incompleteness, ambiguity, latency, deception, and approximations [23]. Regarding the data sensitivity, organization should apply efficient strategies in order to protect data and comply with the regulatory requirements.

E. Viscosity

Complex Data management can be very complex for huge data sets especially when they come from different sources, because linking, matching and transformation are significant activities [24]. Complexity of Big Data deals with the degree of correlation and interdependencies in big data structures such that small changes can have significant effect in system's behavior or might not apply any changes at all [18].

F. Variability

Data variability refers to inconsistent data flow, and can be shown at times. This property has become challenging by increasingly usage of digital media, which is the main root of peak in data loads [22].

G. Volatility

For the understanding of Volatility, we need to understand volume, variety, and velocity of Big Data. Big Data volatility refers to the life duration, that for how long time data is valid and for how long time it should be stored [35]. In this domain, for real time data, we need to determine at what point and when is data no longer relevant and applicable to the current analysis. For some sources, the data should always be present there, but for others, may be this is not the case. Therefore, this is need to understand the requirements, availability and lifetime of data. In a data standard setting, data are kept for decades to build an understanding of importance of data. In Big Data production environment, where data is massively growing every second, we need to establish rules and regulation for data availability for smooth running of enterprises work processes.

Due to the volume, velocity, and variety of big data, this is very necessary to understand volatility. For some specific sources, the same data will always be there; but for others, this is not will be the case. Understanding what data is out there and for how long can help us to define retention requirements and policies for big data. Being a professional, big data help us to recognize better ways to

design and deliver our products and services. This is possible only when big data is integrated into the operating processes of organizations and companies.

H. Viability

The meaning of Viability is that the Big Data should have the capability to be live and active forever, and able for developing, and to produce more data when need. But we need more to analyze huge data sets in real time. We need carefully to select the attributes and factors that are most likely to predict outcomes that matter most to businesses. Using Big Data, We are collecting multidimensional data, not simply collecting a large number of records, but spans a broadening array of variables. The secret is uncovering the latent, hidden relationships among these variables.

By mining a sample of the data and executes some simple statistical tests and calculations to conclude if there is a statistically major and significant correlation between the chosen variable and customer churn. If so, we have to establish the viability of those variables and will want to expand our scope and further investigation resources into generating, collecting and refining that data source. Then we can repeat this process of confirming the viability of key variables until our model describes a high level of predictability.

I. Validity

Validity of data might be having same ideas with veracity of data, but, actually they don't have same concepts and theories. When the status of data changes from exploratory to actionable, data should be valid. A data set may not have any veracity problems but may not be also valid if not properly understandable. This property of Big Data is essential to find the presence of hidden relationships among elements within huge Big Data generation sources. Similarly, a set of data might be validated for a specific applications or usage and will be invalid for another applications or usage. The validity of Big Data producing sources and consequent analysis must be exact correct if the results of analysis are going to be used for the process of decision making.

J. Variety

Data variety refers to the degree of data organization. Unstructured data lacks sufficient degree of organization, while structure organization has a high degree [15]. Structured database encompasses a unified data format and can be managed with database tools easily. However, unstructured and semi-structured data are more difficult to analyze and make a decision. Relational DataBase Management System (RDBMS) has been used by traditional data analysis systems. This traditional RDBMSs have required expensive hardware and only apply to structured data [7]. Thus, semi-structured and unstructured data demand more advanced infrastructure for processing, and traditional management tools are not able to handle the high-volume and heterogeneity of Big Data. Big data platforms should handle data from various sources, such as e-mail, mobile devices, distributed sensors, web pages which are in different forms of text, image, audio, video and multimedia data. Data variety might empower a system in some ways, but it also add more complexity and it is probably one of the obstacle for effectively using huge data set and making a decision [18]. Different data formats generally need more requirements for allocation and processing. Volume, velocity and variety have been considered as the main dimensions of Big Data, However, recently more characters have been added to this category. For instance, IBM and Microsoft have introduced veracity or variability [19] and [16] has added value as another dimension of Big Data.

Data variety is a measure of the richness and fruitfulness of the data representation which are in text, images, video, audio, etc. forms. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl.

IV. Big Data Processing Taxonomy

Although different attributes of Big Data provides many new opportunities, it introduces many new challenges and needs new requirements and technologies. Rapidly increasing heterogeneous data sets demand advances on data storage and Big Data Characteristics Volume Velocity Variety Value Variability Veracity Complexity Data Scale Data processing: batch and stream data heterogeneity: structured, semi-structured, unstructured Data usefulness in decision making Data flow inconsistency Data quality and accuracy Degree of correlations and interdependencies.

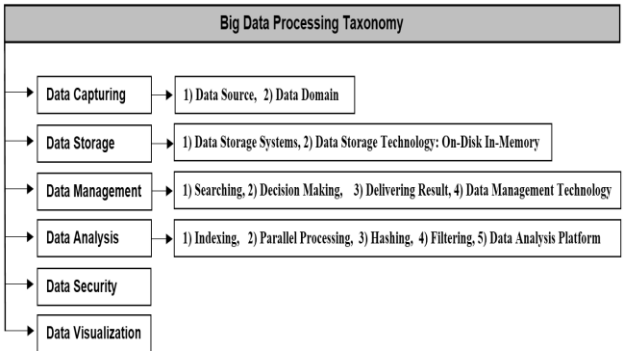

Fig. 2. Big Data Characteristics, Issues and Challenges

Mining technologies that make it possible to preserve huge generated data directly or indirectly by users and then analyze it to yield useful new insights [3]. These associated changes with Big Data reflect six main categories including Data capturing, data storage, data management, data processing, data security and Visualization, which will be discussed as potential opportunities, technologies and challenges in the next coming paper.

CONCLUTION

Growth of the data generation by more than 3.8 billion people (50%) out of 7.6 billion population of the world (at the end of 2017) is becoming hot and critical issue and importance. According to the IDC prediction, Data production will be 44 times larger in 2020 than it was in 2009. Although Big Data provides many opportunities, it brings many challenges and issues and demands many requirements. In this paper, a wide range of these challenges and opportunities have been described. This paper has explored the current status of Big Data and specially focused on the theme of this paper and has described the 10 Vs characteristics, issues and challenges of Big Data. In future work author will explore more with the possible solutions of theses discussed properties and challenges.

REFERENCES

[1]. Richard L Villars, Carl W Olofson, and Matthew Eastwood. Big data: What it is and why you should care. White Paper, IDC, 14, 2011.

[2]. Andrew McAfee, Erik Brynjolfsson, et al. Big data: the management revolution. Harvard business review, 90(10):60–68, 2012.

[3]. Katina Michael and Keith W Miller. Big data: New opportunities and new challenges [guest editors' introduction]. Computer, 46(6):22–24, 2013.

[4]. R Wegener and V Sinha. The value of big data: How analytics differentiates winners. bain & company, 2013.

[5]. Trevor J Barnes. Big data, little history. Dialogues in Human Geography, 3(3):297–302, 2013.

[6]. Diebold Francis. On the origin (s) and development of the term'big data'. Technical report, PIER Working paper, 2012.

[7]. Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. Mobile Networks and Applications, 19(2):171–209, 2014.

[8]. [8] John Gantz and David Reinsel. Extracting value from chaos. IDC iview, 1142(2011):1–12, 2011.

[9]. Teradata, 1979.

[10]. Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. Communications of the ACM, 51(1):107– 113, 2008.

[11]. Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In ACM SIGOPS operating systems review, volume 37, pages 29–43. ACM, 2003.

[12]. Hadoop, 2011.

[13]. Dhruba Borthakur et al. Hdfs architecture guide. Hadoop Apache Project, 53, 2008.

[14]. Nada Elgendy and Ahmed Elragal. Big data analytics: a literature review paper. In Industrial Conference on Data Mining, pages 214–227. Springer, 2014.

[15]. Peter Geczy. Big data characteristics. The Macrotheme Review, 3(6):94– 104, 2014.

[16]. CL Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. Information Sciences, 275:314–347, 2014.

[17]. Soulmaz Salehian and Yonghong Yan. Comparison of spark resource managers and distributed file systems. In Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloudSocialCom-SustainCom), 2016 IEEE International Conferences on, pages 567–572. IEEE, 2016.

[18]. Stephen Kaisler, Frank Armour, J Alberto Espinosa, and William Money. Big data: Issues and challenges moving forward. In System Sciences (HICSS), 2013 46th Hawaii International Conference on, pages 995–1004. IEEE, 2013.

[19]. Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Abdullah Gani, Salimah Mokhtar, Ejaz Ahmed, Nor Badrul Anuar, and Athanasios V Vasilakos. Big data: From beginning to future. International Journal of Information Management, 36(6):1231–1247, 2016.

[20]. Basel Kayyali, David Knott, and Steve Van Kuiken. The big-data revolution in us health care: Accelerating value and innovation. Mc Kinsey & Company, 2(8):1–13, 2013.

[21]. Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S Hopkins, and Nina Kruschwitz. Big data, analytics and the path from insights to value. MIT sloan management review, 52(2):21, 2011.

[22]. Avita Katal, Mohammad Wazid, and RH Goudar. Big data: issues, challenges, tools and good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on, pages 404–409. IEEE, 2013.

[23]. S Mills, S Lucas, L Irakliotis, M Rappa, T Carlson, and B Perlowitz. Demystifying big data: a practical guide to transforming the business of government. TechAmerica Foundation, Washington, 2012.

[24]. Abdullah Gani, Aisha Siddiqa, Shahaboddin Shamshirband, and Fariza Hanum. A survey on indexing techniques for big data: Taxonomy and performance evaluation. Knowledge and Information Systems, 46(2):241–284, 2016.

[25]. Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data applications in the government sector. Commun. ACM, 57(3), 78-85. doi:10.1145/2500873

[26]. Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6, 70.

[27]. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144. doi:https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[28]. Gani, A., Siddiqa, A., Shamshirband, S., & Hanum, F. (2016). A survey on indexing techniques for big data: taxonomy and performance evaluation. Knowledge6and Information Systems, 46(2), 241-284. doi:10.1007/s10115-015-0830-y

[29]. Shah, N., Irani, Z., & Sharif, A. M. (2017). Big data in an HR context: Exploring organizational change readiness, employee attitudes and behaviors. Journal of Business Research, 70, 366-378. doi:https://doi.org/10.1016/j.jbusres.2016.08.010

[30]. American Institute of Physics (AIP). 2010. College Park, MD, (http://www.aip.org/fyi/2010/)

[31]. Mervis, J.: Agencies Rally to Tackle Big Data. Science 336, 22 (2012) Importance of Big Data.

[32]. Addressing Big Data Issues in the Scientific Data Infrastructure. Yuri Demchenko, SNE Group, University of Amsterdam. TNC2013, 3 June 2013.https://tnc2013.terena.org/includes/tnc2013/documents/bigdata-nren.pdf

[33]. Schroeck, M. et al., 2012. Analytics: The real-world use of big data.

[34]. 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.

[35]. Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, et al., "Big Data: Survey, Technologies, Opportunities, and Challenges," The Scientific World Journal, vol. 2014, Article ID 712826, 18 pages, 2014. doi:10.1155/2014/712826