

Big Data Issues from Data Challenges Perspective

G. Sandra

Sree Narayana Guru College of Advanced Studies, Cherthala, Alappuzha, Kerala, India

Abstract

Today Big data is the source of much valuable information. It is extensively used by different agencies like business organizations, governments, academicians, policy makers, etc., for knowledge discovery. Though it is beneficial to many, but all are facing crucial challenges also from it. These challenges are originating from the characteristics of big data and they can be studied from different angles; such as data challenges, process challenges and management challenges. The main objective of this paper is to review the big data challenges from the characteristics of data itself by exploring its sources and identify potential research issues developing from it.

Keywords

Big data, Data challenges, Datasets, Processing, Characteristics.

I. Introduction.

Big data is the buzzword of the modern digital world [1]. The popularity of ICT has resulted in an unprecedented growth in the data generation [2]. Scholars have calculated that its growth rate has already exceeded 2.5 exabyte and expected to be 44 fold growth in 2020 compared to what it was in 2009 [3]. Different agencies such as online networking sites, sensors, market researches, scientific data, academic researches, etc., have contributed much to this massive explosion of data, otherwise call Big Data. This rapid rise in big data has promoted competition, productivity, innovation and growth of different sectors of various economies of the world.

This plethora of data generation is both a boon as well peril to different beneficiaries of it, such as corporate world, government agencies, medias and policy makers. It is the treasure house to provide critical intelligence for deriving insights and formulating policy decisions [4]. Its beneficiaries are belonging to domains like engineering, politics, science, insurance, defense, healthcare, marketing agencies, etc. Apart from its benefits, big data stakeholders are faced many challenges also, which are broadly classified into; (i) the voluminous datasets both from size and complexity, which is called data challenges, (ii) the skill to process these data, otherwise called process challenges and (iii) challenges related to database management system[5].

This study presents a comprehensive review of the big data challenges that are generating from its characteristics. It concentrated on the discussion of data challenges only, which is rated as the most complicated one. Hence the above mentioned other challenges remain beyond the scope of this study. The rest of this study is organized as follows. Section II explains the concept 'Big Data'. Section III deals with the dimensions of big data. Section IV discusses the major research issues. Section V concludes the study.

II. Meaning of the concept 'Big Data'

The concept 'Big data' has an uncertain origin. It is believed that the term has originated by mid- 1990's [6]. But it gained its present popularity only from 2011. The IBM and similar organizations have played a leading promotional role to have current hype in its prominence. Different scholars have attempted to define the

concept 'Big data' in their own way by emphasizing different dimensions of big data. Many of them have differed in their understanding about big data, where a few concentrated on what it is and others on what it does. There is no universal definition of big data. Most of the researchers used the characteristics of big data to define the concept. In this attempt majority of them concentrated on the particular aspect of big data, i.e., size of data. The data scientist, John Rauser defined big data as "Any amount of data that's too big to be handled by one computer". To Gartner, "Big data is the high volume, high velocity and high variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [7]. IBM has of the view that "Every day, we create 2.5 quintillion bytes of data. This data comes from everywhere: sensors used to gather climate information, post to social media sites, digital pictures and videos, purchase, transaction records and cell phone GPS signals, to name a few. This data is big data". "The next frontier for innovation, competition and productivity" is also termed as big data (McKinsey & Co).

A review of the above definitions shows that the data generation at greater speed with high volume and variety can be termed as big data. The annual rapid rate of growth of data is expected to persist at a speed of 50 to 60% [8]. This rate of growth of data is a critical issue to computational science that needs serious thoughts about finding out the sources of the problems and the challenges arising from it. This study deals with it in the next section.

III. Dimensions of Big Data

The exponential growth of big data over the years has changed the outlook of the researchers with regard its dimension. Starting it with 3Vs by Gartner, now it is more than 14 Vs and a C. Gartner used to understand big data with its three characteristics, namely, Volume, Velocity and Variety. They are very high in nature. The Statistical Analysis System (SAS) has made an addition to the above by proposing Variability and Complexity [9]. Value was added as another prominent characteristic by Oracle [10]. In 2014, nine more Vs were added to the characteristics of big data and made the total to 14 Vs and they are Veracity, Volatility, Validity, Viscosity, Venue, Virality, Visualization, Vagueness and Vocabulary. These characteristics are at present the focal point of research to scholars. In order to theorizing types of data challenges, Uthayasankar Sivarajah, et.al have reviewed 227 articles published on big data challenges and found that 39.64% of articles are data challenge related to volume, 25.9% of the researchers have considered Variety as the detrimental challenge, for Veracity, it is 19.4%, for Value 13.2%, for Velocity, 7.9%, for Visualization, 2.6% and for Variability 1.8% articles have considered it as the biggest challenge [11]. To get more vision about these characteristics, a brief discussion of them is given below.

A. Volume

In big data, its volume refers to the amount of data generated which is very high. Its size is multiplying from terabyte to petabyte. The advancement of technology and communication has led to the multitudinous growth of data to the tune of zettabytes. Since the

size of the data collected is so huge, its storage is a great challenge to the enterprises and also it is very costly. The modern database management tools are inefficient to solve the above problems. The data scientists have proposed solutions like Phase Change Memory (PCM), SSD (Solid State Drive) to solve above problems. Methods like cloud computing technology, distributed big data storage, where the data collected is stored in their own original location itself [12], are widely accepted to store high volume of big data generated. But all these methods have their limitations.

B. Variety

Big data is generated by human and machine efforts in the form of text, picture, videos, e-mails, tweets, audio recording, etc. [13]. They are collected from internal and external sources. It can be classified as structured, semi-structured and unstructured. The structured digital contents are easy to visualize, process, store and model. Structured data are rigid in nature and effortless to be managed in databases. Unlike the structured data, the semi-structured data is not rigid in nature. It contains meta-model such as markers and tags for identification. For unstructured data, there is no predefined format. It may be in the form of articles, e-mails, image files, books, videos, etc. This type of data is difficult to process. So to arrive at critical decisions and knowledge discovery, the real challenges the organizations are facing may be how to comprehend such raw data and manage it [14].

C. Velocity

In big data, the velocity refers to the rate of generation of data. Due to the expansion of social networks and emergence of new technologies, high rate of influx of non-homogeneous data is generated continuously [15]. It is characterized by non-aligned data structures, mutually exclusive data format and erratic data semantics. It helps the corporate and other organizations to get the awareness about the growth of big data. To utilize the big data for knowledge discovery, one should give importance not only to how much information collected, but also how fast it can be used for getting insights, which will influence the decision making.

D. Value

The value of the datasets refers to the quality of the data and its usefulness. It really governs the level of business profits. The big data has relatively low value density. It shows that the data collected in the original format can give only low value compared to its volume. By processing high volume of data high value can be obtained. So value is depending on the scientific analysis of accurate data. Every business decision of the corporate sector is related on the insights derived from data analysis. This means that one has to get the most value out of the datasets collected. This will depend on the ability of the organization to extract timely information in real time.

E. Veracity

The veracity dimension of the big data explains the quality and accuracy or truthfulness of the data collected. Usually it is very low in big data. The collected raw data may not be directly used for knowledge extraction. Data collected may be good or bad. Hence its abnormalities, bias, inconsistencies and duplication must be removed to make the data collected utmost pure [16]. The veracity of data is also influenced by the selection and implementation of processing techniques. The analysis of datasets in the right way makes the outcome more reliable. It should be analyzed in a timely manner, if not the derivation of insights may not be relevant.

F. Variability

The inconsistencies in the velocity of big data are referred as variability [17]. Data generated from myriad resources, such as social media, sensors, cell phones, GPS signals, digital pictures, etc, are complex in nature. There is also variation in the flow rates of data and it is constantly changing, which affects the meaning of data while processing at different times. This property has become a great challenge to data scientists. Hence suitable algorithms need to be developed, especially for performing sentiment analysis, to understand the meaning of the words and the context in which it is used [18].

G. Visualization

Presenting the collected data in different visual formats is essential in big data in order to make it readable. Usually graphs and pictures are used to ease the complexity of the big data. The big data tool, 'Tableau' is the best example of visualization technique to transform the complex datasets. But many of such big data applications have only poor performance which poses great challenge to big data analytics.

H. Volatility

Volatility refers to the life duration of datasets. i.e., how long it will remain valid and can be stored [19]. In standard datasets, the data is stored for a fixed data period according to its importance and need, for which rules and regulations are framed in advance. How long the data needs to be stored is depending on many factors such as the recurring use of data, storage capacity, the value of data, its relevance, etc. A proper understanding of the volatility is depending on the awareness of the volume, velocity and variety of big data.

I. Viscosity

The datasets of big data are collected from different sources, where resistance to data flow is seen in many ways. Viscosity measures this resistance. Only by using appropriate processing techniques, this resistance can be analyzed to derive insights from the data. Since big data is complex in nature, its management is very difficult. The degree of correlation and interdependencies in big data structures are dealt by this complexity and a small change in its structure can have great impact on the system's behavior [20].

J. Virality

In data collection, for knowledge extraction, data has to be disseminated as early as possible between beneficiaries to materialize the objectives. Virality character of big data refers to how speedily the information is getting transferred between people to people networks. It is the spread speed at which the data is broadcasted or spread by a user and received by different users.

K. Validity

Data quality has different dimensions and one among them is validity. It is essential for the evaluation of data quality. It refers to the level of need for data. It is being affected by the compatibility, and correctness of big data. The validity of data may have greater similarity with data veracity. But they are not one and the same. Sometimes the data may not have any veracity problem, but may not be valid, if it lacks legibility and not understandable. The validity property of big data is used to find out the hidden relationship of big data. Similarly a dataset may be valid for a particular application, but cannot be used for another purpose.

L. Venue

The data collection is taking place in different places and arrangements. It may be through social media, government agencies, medical fields, sensors, cell phones, customer workstations, in the cloud, log files, etc the data collection is made possible. These data are generated for different purposes. Multiple platforms, databases and formats are used for data generation.

M. Vocabulary

For solving challenges of big data, data scientists have developed new concepts, theories, definitions, and techniques. For example, Apache Hadoop, MapReduce, Metadata, NoSQL, Apache Hive, Oracle NoSQL, Hadoop Distributed File System etc. It is nothing but data terminology like data structure, data model and computing architecture.

Having discussed how the characteristics of big data became the source of big data challenges, the following section reviews the major research issues developing from them.

IV. Data Challenges and Research Issues

The data challenges emanate from the characteristic of big data. The growth of big data from all sides of the digital world is associated with high degree of complexity. To solve it, consistent efforts are required for developing appropriate applications suitable to the different characteristics of big data. Hence it is imperative to know the most important research issues developing from the characteristics of big data. So a brief review of them is given below.

A. Scalability and Storage Issue

In big data, the scalability is the attribute of a system to accommodate the growing quantum of data. To analyze the speedily increasing big data and its high volume, scalable data platforms are absolutely essential. The techniques used for data analysis can be considered as scalable, only when it has the capacity to accommodate the growing quantum of data for analysis without making change in its design or code refraction. In a big dataset with millions of data, it is very difficult to change the technique used for analysis in accordance with the change in the volume of data. Similarly the storage of the ever growing big data is also a great data challenge. The storage issue is very crucial in big data management. Due to high volume and high velocity of big data, the organizations are facing the problem of finding out sufficient infrastructure for storage of data for query and data analysis. To solve this issue, different types of data storage techniques are employed, such as Direct Attached Storage System (DAS), Network Attached Storage system (NAS) and Storage Area Network (SAN). In DAS, various hard disks are directly attached to servers to store data. But it has low scalability. NAS is a network oriented system. It is equipped with mechanisms like special data storage system, disk array, storage software, etc. It has an expandable storage capacity. But the data in this system is transmitted in form of files and thus faces the problem of security and privacy. The SAS is designed with a scalable and bandwidth intensive network to store data. But it also confronts the problem of security.

B. Uncertainty Issue

The data generated in the various domains of big data is inherently uncertain because of inconsistencies, noise and incompleteness in data collection. Every characteristic of big data has its own role in the generation of uncertainty issue. It creeps into the system

while collecting, organizing and analyzing the data. If training dataset is biased, any sort of advanced techniques of analysis will not be good [21].

C. Dimensionality Issue

The dimensionality of big data refers to how many attributes of it to be considered for dataset analysis. The increase in attributes of dataset generates the problem of dimensionality. When it increases, the volume of the space also increases. Hence the available data become sparse. This is a problem to any sort of analysis for statistical significance. When the available data is sparse and dissimilar, the techniques used for big data analysis become inefficient.

D. Accuracy Issue

Accuracy in the analytical efforts of big data is a crucial factor. It is essential for getting desired results from the processing of datasets. Though the big data is resource rich, but it can be less useful if the data analysis is inaccurate. Data scientists use sampling techniques to analyze big data. Since the whole data is not considering for the analysis, the result arrived at may be inaccurate. Data analysis may be relatively accurate in the case of structured data bases, but it will be affected adversely with every increase in the data volume and variety [22].

E. Complexity

Among the big data categories, it is less complex to process the structured data. It has a rigid format. They are automatically generated without user interaction. Techniques like SQL are used to process the structured data. The data processing is too hard in the case of unstructured data because they have no definite format. To solve its complexities, data scientists are generally depending on data processing platforms like Hadoop, MapReduce, etc. To validate all items in big data is also difficult. Hence new approaches are essential to solve the complex situation.

F. Security Issue

Ensuring security always a major big data challenge. Most of the organizations allow others to access user profiles without the permission of its owners. Here the individual's right to privacy is curtailed. Now-a-days the big data leakage is the order of the day. It threatens privacy. The deluge of data also raises great privacy concern. Most of the organizations while collecting personal and sensitive data do not resort to security measures to ensure data privacy. It is desirable to design efficient and privacy preserving techniques for big data sharing and processing to ensure the confidentiality of stored data.

V. Conclusion

Big data challenges can be grouped into three categories: the data challenges, process challenges and management challenges. This study surveys data challenges related to the characteristics of data and research issues developing from it. The nature of the data challenges of big data reveals that it is not an easy task to solve them and hence the research issues. So we need advanced techniques in accordance with the nature of data challenges and research issues.

References

- [1] IDC, 'Digital Universe in 2020'.
- [2] Chen M, Mao M, Liu Y, "Big Data: A survey", Mobile Networks and Applications, Vol. 19, No. 2, pp. 171-209, 2014.

- [3] Mc.Afee A, Brynjolfsson E; "Big data: The management revolution Harvard Business Review, 90(10), pp. 60-68, 2012.
- [4] Lund S, Manyika J, Nyquist S, Mendonca L, Ramaswamy S, "Game changes: five opportunities for US growth and renewal", Mc Kinsey Global Institute Report, 2013.
- [5] Jiang H, Chen Y, Qiaoz, Weng T.H, Li K.C; "Scaling up map Reduce based big data processing on multi-GPU systems", Cluster Computing, 18(1), pp. 369-383. 2015.
- [6] Diebold F.X, "A personal perspective on the origin and development of big data: The phenomenon, the term and the discipline", Social Science Research Network, 2012.
- [7] Beyer M.A, Laney D, "Importance of big data: A definition", Stamford C.T: Gartner, 2012.
- [8] Chen D, Safran M, Peng Z, "From big data to big data mining : challenges, issues and opportunities", Database systems for Advanced Applications, Springer, pp. 1-15, 2013.
- [9] Mark Troester, "Big data meets big data analytics", [Online] Available: <http://www.sas.com/resources/WR46345.pdf>.
- [10] Oracle, "Information management and big data: A reference Architecture, [Online] Available: <http://www.oracle.com/infomgmt-big-data>.
- [11] Uthaya sankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody, "Critical analysis of big data and analytical methods", Journal of Business Research, Vol.70, pp. 263-286, 2017.
- [12] Rongxing Lu, Huizhu, Ximeng Liu, Joseph K. Liu, Jun shao, "Towards efficient and privacy preserving computing in big data era", IEEE Network, 2014.
- [13] Chen J, Chen Y, Du X, Li C, Lu J Zhaos, Zjou X, " Big data challenges: A Data Management Perspective", Frontiers of Computer Science, 7(20), pp. 157-164, 2013.
- [14] Labrinidis A, Jagadis H.V, " Challenges and Opportunities with Big Data", Proceedings of the VLDB Endowment, 5(12), pp. 2032-2033, 2012.
- [15] Chen M, Nao S, Liu Y, "Big data: A survey", Mobile Network Applications, 19(2), pp.171-209, 2014.
- [16] Mills S, Irakliotis L, Lucas S, Rappa M, Carlson T, Perlowitz B, "Demystifying big data: A Practical guide to transforming the business of government", Tech American Foundation, Washington, 2012.
- [17] Gandomi A, Haoder M, "Beyond the hype: Big data concepts, Methods and Analytics", International Journal of Information Management, 35(2), pp. 137-144, 2015.
- [18] Zhan X, Hu Y, Xie K, Zhang W, Su L, Liu M, "An evolutionary trend reversion model for stock trading rule discovery", Knowledge-Based Systems, (79), pp. 27-35, 2015.
- [19] Nawsher Khan, Ibrar Yaquooob, Ibrahim Abaker, Targio Hasem, "Big Data: Survey, Technologies, opportunities and challenges", The scientific world Journal, Vol. 7, 2014.
- [20] Stephen Kaisler, Frank Armour, Alberto Espinosa J, William Money, "Big Data: Issues and challenges moving forward", International conference, IEEE, pp. 995-1004, 2013.
- [21] Saidulu D, Sasikala R, "Machine learning and Statistical Approaches for Big data: Issues and Challenges and Research directions", International journal of Applied engineering Research, 12(21), pp. 1691-99, 2017.
- [22] Tole A. A, "Big Data Challenges", Database System Journal, 4(3), pp. 31-40, 2013.