International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Big Data: Challenges and Opportunities

P. Bastin Thiyagaraj[1], K. Gayathri [2], Dr. A. Aloysius[3]

[1]Department of Information Technology, St. Joseph's College (Autonomous), Trichy, TamilNadu, India
[2]Department of Information Technology, St. Joseph's College (Autonomous), Trichy, TamilNadu, India
[3]Department of Computer Science, St. Joseph's College (Autonomous), Trichy, TamilNadu, India

ABSTRACT

We are living in an information age and there is enormous amount of data that is flowing between systems, internet, telephones, and other media. The development of Internet, Internet of Things, and Cloud computing lead to growth of data in almost every industry and business area. Big data has rapidly developed into the hot topic, attracts extensive attention from academia, industry, and governments around the world. The data is being collected and stored at unprecedented rates. In this paper, the concept of big data is introduced briefly, including its definition, features, significances, opportunities and value. We describe the grand challenges (namely, data complexity, computational complexity, and system complexity), as well as possible solutions to address these challenges.

Keywords: Big Data, Data Complexity, Computational Complexity, System Complexity.

I. INTRODUCTION

'Big Data' is a **data** but with a **huge size**. 'Big Data' is a term used to describe collection of data that is huge in size and yet growing exponentially with time. A data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently. According to the Definition of Gartner "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"[9]. For example, Face book generates over 500 terabytes of data everyday—including uploaded photos, likes, and users' posts [10]. According to IDC's "Digital Universe" forecasts, 40 ZB of data will be generated by 2020[11].

Big Data

Big data could be found in three forms [12]:

- **Structured**: Data can be stored in database SQL in table with rows and columns.
- **Semi-structured**: Semi-structured data (CSV but XML and JSON) is information that does not reside in a relational database but that does have some organizational properties that make it easier to analyse.
- **Unstructured**: It often includes text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations

Attributes of Big Data

There are six V's include in Attributes of Big data.

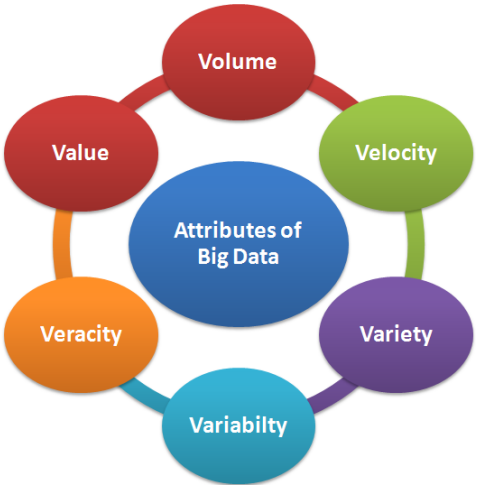

Figure 1. *Attributes of Big data*

(i)Volume – The name 'Big Data' itself is related to a size which is enormous. Size of data plays very crucial role in determining value out of data. A particular data can actually be considered as a Big Data or not and it dependent upon volume of data. Hence, **'Volume'** is one of the characteristic which needs to consider while dealing with 'Big Data'.

(ii)Variety – Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. It is also considered as the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

(iii)Velocity – The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows from the sources like business processes, application logs, networks and social media sites, sensors, MOBILE devices, etc.

(iv)Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively [1].

(v)Veracity – includes two aspects, **Data consistency** – defined by their statistical reliability and **Data trustworthiness** – data origin, collection and processing methods including trusted infrastructure, and facility [13].

(vi) Value- Value starts and ends with the business use case. The business must define the analytic application of the data and its potential associated value to the business.

II. SIGNIFIANCES OF BIG DATA

There are number of significances that are transforming our live style in Big data [2].

- National development
- Industrial upgrades
- Scientific research
- Emerging interdisciplinary research

National development: Big data, companies will upgrade and transform to the mode of Analysis as a Service (AaaS), thereby changing the ecology of the IT and other industries. In this context, the global giants of the IT industry (such as IBM, Google, Microsoft, and Oracle) have already begun their technical development planning in the big data era. At the national level, the capacity of accumulating, processing, and utilizing vast amounts of data will become a new landmark of a country's strength.

Industrial upgrades: Big data is a big problem faced by many industries, and it brings grand challenges to these industries. Many big companies, including Google, Microsoft, Amazon, Facebook, Alibaba, Baidu, Tencent, and other IT giants, are working on cloud computing technologies and cloud-based computing services.

Scientific Research: the development of unprecedented data the scientific community has to re-examine its methodology of scientific research and has triggered a revolution in scientific thinking and methods [15]. Scientific research has been about data, and as data in research continues to grow exponentially. scientific research organization in

Germany that is dedicated to providing scientists worldwide faster access to insights into the atomic structure of novel semiconductors, catalysts, biological cells and other samples, making optimal data management in this high-volume environment extremely critical.

Emerging interdisciplinary research: it deals about the Data science has been gradually coming into place. It includes information science, mathematics, social science, network science, system science, psychology, and economics [14]. It employs various techniques and theories from many fields, including signal processing, probability theory, machine learning, statistical learning, computer programming, data engineering, pattern recognition, visualization, uncertainty modelling, data warehousing, and high performance computing

III. CHALLENGES OF BIG DATA

There are many challenges in big data ranging from the design of processing systems at the lower layer to analysis means at the higher layer, as well as a in scientific research. Among these challenges, some are caused by the characteristics of big data, some, by its current analysis models and methods, and some, by the limitations of current data processing systems.

Data complexity

The big data has provided unprecedented large-scale samples when dealing with computational problems and face more complex data objects. The inherent complexity of big data (including complex types, complex structures) makes its perception, representation, understanding and computation far more challenging and results in sharp increases in the computational complexity when compared to traditional computing models based on total data.

Computational complexity

The key features of big data, multi-sources, huge volume, and fast-changing, make it difficult for traditional computing methods machine learning, information retrieval, and data mining are effectively support the processing, analysis and computation of big data. These methods cannot simply rely on past statistics, analysis tools, and iterative algorithms used in traditional approaches for handling small amounts of data.

System complexity

Big data processing systems suitable for handling a diversity of data types and applications are. For data of huge volume, complex structure, and sparse value, its processing is confronted by high computational complexity, long duty cycle, and real-time requirements.

The design of system architectures, computing frameworks, processing modes, and benchmarks for highly energy-efficient big data processing platforms is the key issue to be addressed in system complexity. Solving these problems can lay the principles for designing, implementing, testing, and optimizing big data processing systems.

IV. OPPORTUNITIES

There are opportunities in doing research in storage of data, retrieval of images, audios videos, life cycle and visualization of high dimensional data.

Storage and Retrieval of Images, Audios and Videos: Multidimensional data should be integrated with analytics over big data hence array-based in-memory representation models can be explored. With the proliferation of smart phones Images, Audios and Videos are being generated at an unremarkable pace. However, storage, retrieval and processing of these unstructured data require immense research in each dimension [5].

Data Life Cycle

Research activities might use only part of the life cycle for instance, a project involving meta-analysis might focus on the Discover, Integrate, and Analyze steps, while a project focused on primary data

collection and analysis might bypass the Discover and Integrate steps.

Plan: description of the data that will be compiled, and how the data will be
Collect: observations are made either by hand or with sensors or other instruments
Assure: the quality of the data are assured through checks and inspections
Describe: data are accurately described using the appropriate metadata standards
Preserve: data are submitted to an appropriate long-term archive (i.e. data center)
Discover: potentially useful data are located and obtained, along with the relevant information about the data (metadata)
Integrate: data sources are combined to form one homogeneous set of data that can be readily analysed
Analyse: data are analysed
Further, some scientists may create new data in the process of discovering, integrating, analysing, and synthesizing existing data [6].

Big Data Computations

Big data is ubiquitous. The good news is that it provides great opportunities for the data analyst. The assumptions are required with smaller data sets and let the data speak for itself. The size of data sets is now increasing much more rapidly than the speed of single core, of RAM, and hard drives. Many tools can't handle this well, when data is too large, the software stops working. High-Performance Computing is CPU-centric, typically focusing on using many cores to perform lots of processing on small amounts of data [7].

Visualization of High-Dimensional Data

The wide availability, increasing size, and complexity lead to new challenges and opportunities for their effective visualization. For example, genomic microarrays in biology spectrometry data in air quality research simulation parameters in nuclear safety engineering and chemical compositions in combustion simulations [8] can all be mapped to high-dimensional spaces for exploration.

V. CONCLUSION

We have entered an era of Big Data. Through the analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. We must support and encourage fundamental research towards addressing these challenges if we are to achieve the promised benefits of Big Data.

VI. REFERENCES

[1]. Prof. https://www.guru99.com/what-is-big-data.html
[2]. V. Mayer-Schonberger, K.Cukier, Big Data: A Revolution That Will Tran sform How We Live, Work, and Think, Houghton Mfflin Harcourt, 2013
[3]. T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation, 2009.
[4]. T.Kalil,"Big data is a big deal", available at: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal, 2012.
[5]. Jin, X., Wah, B. W., Cheng, X., Wang, Y.,. Significance and challenges of big data research. Big Data Research 2 (2), 59-64 , 2015.
[6]. https://www.dataone.org/data-life cycle
[7]. https://docs.microsoft.com/en us/machine-learning-server/r/tutorial-large-data-tips
[8]. http://ieeexplore.ieee.org/document/7784854/references
[9]. Amir Gandomi and Murtaza Haider "Beyond the hype: Big data Concepts, Methods and analytics", International Journal of Information Management (IJIM) ELSEVIER, 2015, pp: 137-144.

[10]. Provost, F., & Fawcett, T. (2013).Data science and its relationship to big data and data driven decision making.Big Data, 1(1), 51–59.

[11]. Cai, L and Zhu, Y 2015 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002.

[12]. https://jeremyronk.wordpress.com/2014/09/01/structured-semi-structured-and-unstructured-data/

[13]. Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure," IEEE, pp. 48–55, 2013

[14]. M. Loukides, What Is Data Science?, O'Reilly Media, Inc., 2011

[15]. Datascience,http://en.wikipedia.org/wiki/Data_science, 2014

AUTHOR'S PROFILE

P. BASTIN THIYAGARAJ is working as an Assistant Professor in the Department of Information Technology, St.Joseph'scollege(Autonomous), Tiruchirappalli, TamilNadu, India. I am having 7 years of experience in teaching and 2 years in research.

K. Gayathri is studying II MSc Computer science in the Department of Information Technology St.Joseph'scollege (Autonomous), Tiruchirappalli, TamilNadu, India.

Dr. A. ALOYSIUS is working as an Assistant Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 16 years of experience in teaching and research. He has published many research articles in the National / International conferences and journals. He has acted as a chairperson for many national and international conferences. Currently, eight candidates are pursuing Doctor of Philosophy Programmed under his guidance.