International Journal of Research in Advanced Engineering and Technology
ISSN: 2455-0876; Impact Factor: RJIF 5.44
www.engineeringresearchjournal.com
Volume 2; Issue 6; November 2016; Page No. 26-30

Big data: A study of challenges and applications

Shefali Gupta

B.E Student, Department of Computer Science& Engineering Model Institute of Engineering and Technology, Kot Bhalwal, Jammu & Kashmir, India

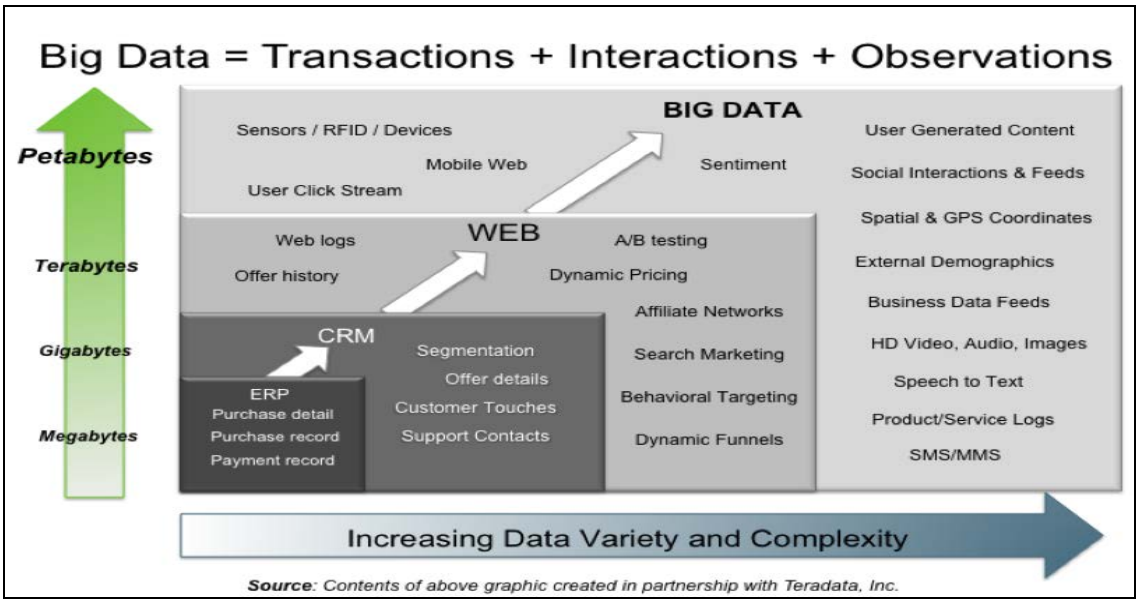
Abstract

The era of Big Data has only just begun. Big data is an emerging paradigm application to datasets whose size or complexity is beyond the ability of commonly used computer software and hardware tools. Big Data creates a radical shift in how we think about research. *"Big data…… has become prevalent in our lives" (Betsy Burton, Gartner Analyst).* Big data becomes very important driver for innovation and growth. The focus of present paper is to provide a summary existing literature to know about the background of big data along with its challenges and applications. This paper also elaborates the 'HACE Theorem' that states the characteristics of the Big Data revolution.

Keywords: big data, 3 v's, HACE theorem, big data challenges, big data applications

1. Introduction

Big data is the name used everywhere now a days in distributed paradigm on web. The era of Big Data is underway. The amount of data that is generated and stored is increasing rapidly, even exponentially. It may be doubling every two years, according to one estimate (IDC 2011) [4]. The need of Big Data comes from the big Companies like yahoo, Google, facebook etc for the purpose of analysis of big amount of data which is in unstructured form. Very large datasets commonly referred to as *big data*, have become common in the study of everything from genomes to galaxies, including, importantly, human behavior. *"Information is oil of 21st century and analytic is the combustion of engine." (Peter Sondergaard, Gartner Research).* Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. Big data is a buzzword, or catch-phrase, meaning a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity. Despite this, big data has the potential to help companies improve operations and make faster, more intelligent decisions. Revolutions in science have often been preceded by revolutions in measurement (Sinan Aral cited in Cukier, 2010) [2]. It offers a profound change at the levels of epistemology and ethics.



Big data can be characterized by 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily.

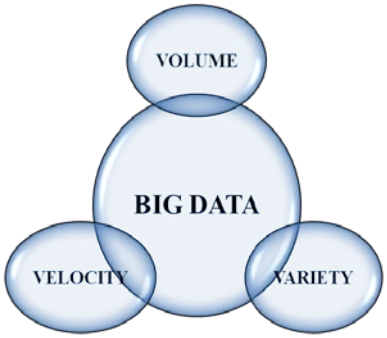


Fig 1: Big Data in 3V's.

The etymology of 'Big Data' has been traced to the mid-1990s, first used by John Mashey, retired former Chief Scientist at Silicon Graphics, to refer to handling and analysis of massive datasets (Diebold, 2012) [3]. *"The more data there is, the better my chances of finding the 'generators' for a new theory."(John Seely Brown).*

The aim of this research paper is to know about the background of big data along with its applications and challenges that exist in the Big Data, based on a comprehensive review of existing research literature. The rest of the paper is organized as follows: Section 2 provides comparison between big and small/ traditional data, while Section 3 discussed about HACE Theorem which highlights the characteristics of big data. Section 4 highlights challenges which big data faces and Section 5 provides some applications that could be significant in solving and analyzing the real world problems. Finally, Section 6 concludes the paper.

2. Comparing Small/Traditional and Big Data

Based on a review of definitions of Big Data, Kitchin (2013, 2014) [5] contends that Big Data are qualitatively different to traditional/small data along seven axes.

Table 1: shows the comparison of small and Big Data:

Seven Axes	Small Data	Big Data
Volume	Limited To Large	Very Large
Velocity	Slow, Freeze-Framed/Bundled	Fast, Continuous
Variety	Limited To Wide	Wide
Exhaustivity	Samples	Entire Population
Resolution& Indexicality	Course & Weak To Tight & Strong	Tight & Strong
Relationality	Weak To Strong	Strong
Extensionality& Scalability	Low To Middling	High

Table 2: shows comparison of Big Data and Traditional Data under the dimension of 3V's Volume, Velocity and Variety.

Characteristics	**Big Data**	**Traditional Data**
Volume	Terabyte, Petabyte, Exabyte	Gigabyte
Velocity	More rapid	Per hour, day
Variety	Structured, semi structured or unstructured	Structured

3. HACE Theorem

The theorem was proposed in IEEE paper that we have referred and it states that, Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. This characteristic of big data makes it as extreme challenge for useful knowledge discovery. For above consider the scenario we can imagine that a number of blind men are trying to size up a giant elephant, where blind people are asked to draw the picture of an of the elephant according to the part of information each collected during the process. As each persons view is limited to his region each can think that trunk of elephant as a hose, leg as a tree trunk, body as a wall and tail as a bombastic long shot. Problem can be make more complicated by assuming that size of elephant growing rapidly and the pose also changing continually. Also blind men are exchanging information and learn on their respective feelings.

"The more people you have playing with the data, the more people are going to do useful things with it". (Kim Taipale)

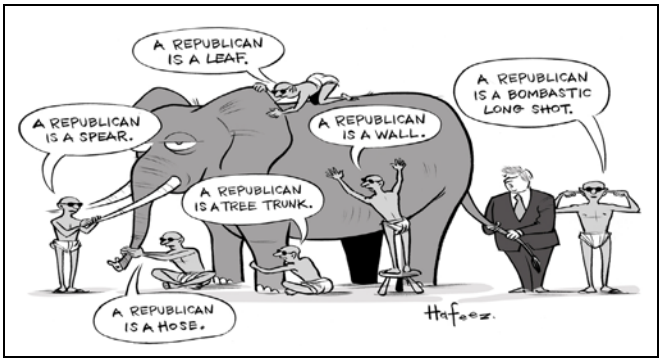


Fig 2: The blind men and the giant elephant

This information may be biased. So the results of each blind person's prediction is something different than actually what it is

- **Huge Data with Heterogeneous and Diverse Dimensionality:** Heterogeneous means from different data sources and every data collection requires a unique recording protocol. For example, each single human being in a biomedical world can be represented by using simple demographic information such as gender, age, family disease history, and so on.
- **Autonomous Sources with Distributed and Decentralized Control:** Each data source is independent to produce and gather data without having centralized control. But, the huge volumes of the data also make an application susceptible to attacks or malfunctions, if the whole system has to rely on any centralized control unit.
- **Complex and Evolving Relationships:** Complexity increases with varied sources of data collection and constantly increasing data size and its incoming speed evolving into relationships such as one-to-many, many-to-many. This gives useful patterns and acts as insight for analyzing the data.

4. Challenges in Big Data

Big data technologies are maturing to a point in which more organizations are prepared to pilot and adopt big data as a core component of the information management and analytics infrastructure. Big data, as a compendium of emerging disruptive tools and technologies, is positioned as the next great step in enabling integrated analytics in many common business scenarios.

As big data wends its inextricable way into the enterprise, information technology, practitioners and business sponsors alike will bump up against a number of challenges that must be addressed before any big data program can be successful. Some of the big data challenges are Heterogeneity and incompleteness, Timeliness, Human collaboration, Privacy and security, and Analysis.

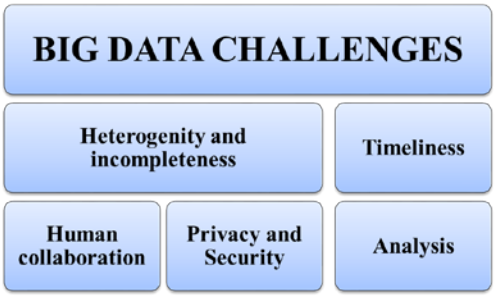


Fig 3: Big data challenges

4.1 Heterogeneity and Incompleteness

Machine analysis algorithms expect homogeneous data, and cannot understand nuance. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge.

4.2 Timeliness

There are many situations in which the result of the analysis is required immediately. Given a large data set, it is often necessary to find elements in it that meet a specified criterion. The larger the data set to be processed, the longer it will take to analyze. It is difficult to design a structure when data is growing in very high speed.

4.3 Human Collaboration

Inspite of the tremendous advances made in computational analysis there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Analytics for big data will not be all computational- rather it will be designed explicitly to have a human in the loop. A Big Data analysis system must support input from multiple human experts, and shared exploration of results.

4.4 Privacy and security

The privacy of data is another huge concern, and one that increases in the context of big data. Privacy is the most sensitive issue, legal and technological implications. In its narrow sense, privacy is defined by the *International Telecommunications Union* as the "right of individuals to control or influence what information related to them may be disclosed". For example in the healthcare industry, record of individual is very personal. But it can be available from multiple sources. So, it is difficult to maintain privacy and security.

4.5 Analysis

Working with new data sources brings about a number of analytical challenges. The relevance and severity of those challenges will vary depending on the type of decisions that the data might eventually inform. Big data is coming from various data sources. So analytics is a challenge.

5. Applications of Big Data

Big data applications solve and analyze real world problems. Internet users and machine-to-machine connections are causing the data growth. *"Big data is at the foundation of all the megatrends that are happening today, from social to mobile to cloud to gaming". (Chris Lyuch, Vertica Systems).* Real time areas are defined following in which big data is used:

5.1 Big data in healthcare

High-performance analytics are new technologies making easier to turn massive amounts of data into relevant and critical insights used to provide better care. Analytics helps to predict disease history and its trends. Unstructured data can be captured through text mining from patient records. It means information can be collected without causing additional work for clinicians. A massive amount of data collected from different sources provides the best practices for today, and will help healthcare providers identify trends so they can achieve better results to improve medical facilities all around the world.

5.2 Network Security

Big data is changing the landscape of security technologies. The tremendous role of big data can be seen in network monitoring. Big data analytics is an effective solution for processing of large scale information as security is major concern in enterprises. Fraud detection is done by using big data analytics. Phone and credit card companies have

conducted large-scale fraud detection for decades. Mainly big data tools are particularly suited to become fundamental for forensics.

5.3 Market and business

Big Data is the biggest game-changing opportunity for sales and marketing, since 20 years ago the Internet went main stream, because of the unprecedented array of insights into customer needs and behaviours. Big data reveals customers' behaviour and proven ways to elevate customer experiences. These insights ensure your business's success.

5.4 Sports

Sport, in business, an increasing volume of information is being collected and captured. Technological advances will fuel exponential growth in this area for the foreseeable future, as athletes are continuously monitored. Statistics can be analyzed and collected to better understand what are the critical factors for optimum performance and success, in all facets of elite sport. Injury prevention, competition, Preparation, and rehabilitation can all benefit by applying this approach. Used consistently this is a powerful measure of progress and performance.

5.5 Education Systems

By using big data analytics in field of education systems, remarkable results can be seen. Data on students online behaviour can provide educators with important insights, such as if the course has to be modified or not based on students reception. This modification can be done by making students answer set of online questionnaire and track the accuracy and time taken to answer those questions.

5.6 Gaming industry

The amount of data that video game players are generating on a daily basis is growing rapidly. People playing video game and generated lot of data in separate areas: game data, player data and session data. In order to improve their game development, game experience, studios are turning to commercial Hadoop distributions such as MapR to analyze, collect and process data from these massive data streams. Armed with this valuable insight from big data, video game publishers are now able to enhance game player engagement and increase player retention by analyzing gamers‟ social behaviour, activity and tracking players‟ statistics, calculating rewards, quickly generating leader boards, changing game play and mechanics and delivering virtual prizes, so as to creating meaningful gaming experiences for their customers.

5.7 Telecommunication Industry

Today big challenges for telecommunication are volume, variety and complexity. Telcos combine ETL and traditional relational databases with big data technologies on a single platform. Telcos technology parses, transforms and integrates the vast amount of data generated by location sensors, IPv6 devices, 4G networks and machine to machine monitors‟ information. Telcos parse and transforms from multiple formats and sources including unstructured mobile, media, web and machine monitor provide data. Telcos masking, managing and identifying sensitive data for regulatory compliance.

6. Conclusion

Big Data is an emerging concept that describes innovative techniques and technologies to analyze large volume of complex datasets that are exponentially generated from various sources and with various rates. Big Data is Challenging task. There are several challenges at data, model and system level. We need computing platform to handle this Big Data. In this paper, an overview of big data along with it application and challenges are discussed. It is known that big data is an emerging trend in all science and engineering domains and also a promising research area.

“It is estimated that by 2020, there could be four times more digital data than all the grains of sand on earth”.

Big Data is becoming the new Final Frontier for research related to scientific domain and for other business applications as it is having tremendous wealth of information. Thus, big data will become an excellent Opportunity in the forth coming years. The paper also discussed about the HACE Theorem. The main goal of our paper is to know about the challenges and make an analysis of various big data applications that are use in IT industries or organization to store massive amount of data. Big data has become an important area of research and will prove to be so in the future.

7. References

1. Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M *et al*. Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. 2012. http://cra.org/ccc/resources/ccc-led-whitepapers/
2. Cukier K. Data, data everywhere. The Economist. 2010. (Accessed on 20th June 2016).
3. Diebold F. A personal perspective on the origins and development of ‘big data’: The phenomenon, the term, and the discipline. 2012. Available at: http://www.ssc.upenn.edu/_fdiebold/papers/paper112/Diebold_Big_Data.pdf (accessed on 20th June 2016).
4. Gantz J, Reinsel E. Extracting Value from Chaos, IDC’s Digital Universe Study, sponsored by EMC. 2011.
5. Kitchin R. The real-time city? Big data and smart urbanism. GeoJournal. 2014; 79:1-14.
6. Kitchen R. Big data new epistemologies and paradigm shifts. Big data & Society. 2014, 1-12.
7. Kitchen R, Mc Ardle G. What makes big data, big data? Exploring the ontological characteristics of 26 data sets. Big data & Society, 2016, 1-10.
8. Sabia, Kalra S. Applications of big data: current status and future scope. International Journal on Advanced Computer Theory & Engineering. 2014; 3(5):2319-2526.
9. Sarkar D, Nath A. Big data – A pilot study on scope and challenges. International Journal of Advance Research in Computer Science and Management Studies. 2014; 2(12):9-19.
10. Sherin A, Uma S, Saranya K, Saranya-Vani M. Survey on big data mining platforms, algorithms and challenges. International Journal of Computer science and Engineering Technology. 2014; 5(9):854-862.

11. Thakur B, Mann M. Data mining for big data. International Journal of Advance Research in Computer Science and Software Engineering. 2014; 4(5):469-473.
12. Wu X, Zhu X, Wu G, Ding W. Data mining with big data. IEEE Transactions on knowledge and Data Engineering. 2014; 26(1):97-105.
13. http://en.wikipedia.org/wiki/big-data. 20th June 2016.