Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-018-0185-1

ORIGINAL RESEARCH

A study of big data and its challenges

Sheikh Mohammad Idrees[1] · M. Afshar Alam[1] · Parul Agarwal[1]

Received: 3 June 2017 / Accepted: 13 April 2018

© Bharati Vidyapeeth's Institute of Computer Applications and Management 2018

Abstract Big data is an emerging torrent. We are held up in a Lake of data and its intensity is continuously increasing. With the fast growth of promising applications like social media, web, mobile services, and other applications across various organizations, there is a rapid growth of data. Thus arises the notion of "Big Data". Data analysis, querying, storage, retrieval organization and modeling are the fundamental challenges associated with it. These challenges are posed due to the fact that big data is complex in nature. In this paper, we address above issues and many more to realize the bottlenecks. But we believe that appropriate research in big data will lead to a new wave of advances that will revolutionize the market and the future analysis platforms, services as well as products and will tackle all the challenges.

Keywords Big data · V's · Big data challenges · Characteristics

1 Introduction

The data revolution has just begun. The term 'Big Data' is relatively new. It has started gaining momentum a decade ago which as a result has led to tremendous increase of data set size. It has started revolutionising commerce, science, medicine, finance and everyday life. Data creation is taking place at an extraordinary rate due to advances more or less in every field of science [1]. IoT, now also referred as the Internet of Everything connects new devices and new sources that generate mounds of data with every passing second [2]. More than 90% of the data in the world has been produced in the past 2 years only [3]. As per stats [4–6], every day we create 2.5 quintillion [a million raised to the power of five (10^{30})] bytes of data. However, the word 'Big' in big data does not only imply to size. Had it been referring to size only, the way out would have been quite effortless or simple. Instead big data is a broad term that describes the enormous volume of data sets, so large and complex that it becomes difficult to handle, process, analyse, manage, store and retrieve the data sets in a specified time frame using traditional methods. Thus big data may be defined as data that is '**so huge**', '**so fast**' and '**so tough**' for present tools to process. Here 'so big' states that organisations must be able to deal with data on a peta byte level, 'so fast' indicates that the data needs to be processed very quickly, similarly 'so hard' indicates that data can not fit in the existing set of tools for processing and managing. Basically we can say that the big data differs from the traditional data in three ways—the amount of data (size), the rate of data creation and transmission (velocity) and the heterogeneity of data—structured, unstructured or semi-structured (variety).

2 Big data: 'How big? Why big?'

Data considered to be the most valuable asset in the twenty-first century (structured or unstructured, public or private, inside or outside the organisations) is growing at a rapid pace nearly 50% per month and expected to be totalling an amount of 800% in the next 5 years [6]. By 2020 data is supposed to be nearly equal to 35 zettabytes. Harnessing such immense data and extracting value from

✉ Parul Agarwal
pagarwal@jamiahamdard.ac.in

[1] Department of Computer Science and Engineering, Jamia Hamdard, New Delhi 110062, India

Published online: 24 April 2018

data is the top challenge and priority for the business and IT sector. Text messages, YouTube videos, Facebook Posts, twitter tweets, blog podcasts, emails, photos, videos and music, geographic map locations, flight sensor data and others 85% of it is unstructured and poses tough challenges for business to manage it or extract values from it and so needs to be focused. As per stats the number of text messages that are send today is greater than the entire population of the world. Consider that facebook in itself has more than 800 million active users who contribute to nearly 250 million photos that are uploaded to facebook daily and about 293,000 status updates per min [8]. The next is the twitter where 97,000 tweets are posted in every second [9]. Almost 400 tweets per min contain a YouTube link to some video. At an average viewers spent almost 2.9 billion h (3,35,000 years) on YouTube in a month and this figure is still rising. Video are being uploaded to YouTube at the rate of 90 h per min. In 1 min we have about 2.4 million Google searches [8, 9]. As per calculation almost 294 billion emails are sent everyday that is one email in almost 0.00000015th of a second [7]. Almost 20.8 million whatsapp messages are being sent every minute [9].There are almost 6 billion mobile phones and among those almost 45% are smart phones which are used to place calls, send text messages, send map locations and surf internet. There are over 200 million blogs online, out of which 34% post opinions about products or brands and 90% of customers trust these opinions made by people online. Most organisations take business decisions based on their structured data only which is equal to just 15% of the total data. As per the latest analysis, in general 60 GB of data is generated with every passing second. Big data due to its potential has found its applications like weather forecasting, crop harvest prediction, and many others [10].

2.1 Big data characteristics

Big data possesses certain characteristics which are mentioned below under the following V's as discussed in [11]:

1) *Volume* Volume depicts the vast ocean of data. It simply refers to the quantity of data. With increase in so many data sources, diverse and dense, there is a rapid rise in the level of data with every passing second.
2) *Velocity* Velocity is the rate at which the data is being generated. Velocity is the measure of how fast data is created, processed and transmitted.
3) *Variety* Variety specifies the spectrum of different data types that form the data. The data may be structured, semi-structured or unstructured.
4) *Veracity* Considering the vast amount and variety of data, most of the data is supposed to be inconsistent. This inconsistent data can create a lot of trouble to the organisations that are dependent on data. As such data veracity takes care of the analysis of the data accuracy.
5) *Value* Data is the most precious raw material in the present time. However, having such a vast volume of data available easily, it is of no use if we are not able to transform this data into value. Organisations with latest IT infrastructure systems have started working on big data and are able to generate useful values from data.

To understand the big data characteristics we can take an example of the population the human beings on earth. This data is expanding with every second, thus specifying the velocity and volume. The different types (based on colour, sex etc.) will in turn represent variety.

3 Issues associated with big data

Big data has presently changed the definition of knowledge. The analysis of big data has put forward many new issues and challenges because of its nature and complexity. These issues are briefly written down as below:

3.1 Data storage issues

From the past, all the data in the world has been continuously increasing at a rapid pace. Hence the storage of this vast growing volume of data has been an issue of concern as we were having limited storage capacity in our systems. However, with the advancements in technology this issue has been tackled. Today we are having memories up to the 8 Tb capacity (or even more) as compared to 1 GB in the past. Furthermore, data is being produced continuously by everybody and not just by professionals only. So the need of the hour is to think beyond the present day storage mediums, as every time we have formulated a new storage medium, the magnitude of data has burst out.

3.2 Data issues

Data is the most precious raw material, also considered as raw oil of the present time. However, handling of huge data has encountered many issues [12]. These are put down as below:

3.2.1 Data heterogeneity

Big data has a vast spectrum of different types of data ranging from text documents, audio, video and sensor data to simulation data. The days have vanished, when the data centres had only to deal with traditional data [13], (Documents, financial transactions, stock records and the personal files). Today, data centres have to handle structured, unstructured and semi structured data which includes photographs, Audio and Video, text documents, 3D models, geographical map locations, sensor data and the simulation data.

3.2.2 Issues due to high volume

Volume of data refers to the magnitude or size of data. With the fast advances in every field of science the volume of data in every field has increased. Considering big data, the size of data is relatively so high that is becomes impossible to process data using the traditional computing solutions like relational database tools. The data may range from even Petabyte to Exabyte scale. Data volume is going to provide a great opportunity to organisations to understand people better at the present time and in future as well [13].

3.2.3 Issues due to high velocity

Big data velocity is the rate at which data is flowing in and out of the organisations. It may be defined as the pace of creation of data or how fast data is generated and processed so as to meet the growing demands of present times. With the rising demand of e-commerce and other areas data generation rate has increased tremendously. In addition the users increasingly demand data which is streamed to them in real time basis. This rate of data generation is beyond human imagination which in itself is a major challenge to tackle.

3.2.4 Addressing sensitivity and data quality

The sensitive data offers great challenge to all the organisations. The sensitive data includes one's personal data or the confidential data of an organisation. Almost all the countries have legislative laws to preserve the sensitivity of data. This sensitive data is subjected to encryption technique in most of the companies to preserve its sensitivity. Data quality aspect of data is outstandingly important. Quality aspect of data involves accuracy, completeness, relevance, consistency in data, reliability, appropriate and accessibility. Delivering all this is posing a challenge in itself.

3.3 Processing issues

Data processing refers to general operations or manipulations performed on input data sets so as to extract the requisite information. Big data processing tools like MapReduce [14] have to deal with data that range from terabyte to Petabyte scale. In addition to this, there are tools like HPCC [15] which provide a high performance computing platform for handling such huge data. Processing of such large scale data creates many hurdles, challenges and opportunities that need to be explored. These issues are as [13]:

3.3.1 Algorithms used

Algorithms form the back bone of any computational processing system. Algorithms specify the computational way for achieving the desired results. The efficiency of a processing system mainly depends upon the complexity of the algorithms used. Despite the fact that there have been great advances in computational algorithms, the algorithms that can deal such huge data volume and velocity are not up to the mark, as such need to be addressed.

3.3.2 Handling timeliness and data scalability

Considering the data sets so large and complex, timeliness of data poses new challenges. Handling of such massive data has been a tough test to face from the last one decade. It is much easier to get data into the system than to put data out of the system. In the past this issue was dealt by making processors faster [16]. However, at present the data level is alarmingly raising as compared with the processor speeds. Data is being generated by every one of us. Moreover, the people are always busy in exploring vital data meant for them. The rate at which users demand data is considerably very high. As such matching the pace of delivering such large demanded data is a tough challenge.

3.4 Data management issues

The management plays a vital role to help in dealing big data and is more or less the most difficult obstacle to be dealt. The aim of effective data management is to help organisations to generate big benefits from big data. It includes policy based approaches to help the flow of data in a system/organisation. However, main issues faced are presented in [17]:

3.4.1 Data ownership

Data ownership [16, 18], has been the most tough issue since the time of developments in internet and cloud. Data is being generated at an enormous tempo by all of us. The instant when some data is placed over the internet, it resides there forever. As such who owns this data becomes a big question. People are voluntarily giving away their data for free to large companies like Facebook, Google, Twitter etc. without thinking about what happens to their data thereafter. Although these big companies allow us to delete our data, when it becomes obsolete for us but there are always possibilities that this data may get compromised. In the past many cases have came into limelight where the user's data has been compromised by the companies like Uber, Facebook [19], etc.

3.4.2 *Costs for infrastructure setup*

Since big data is the new forthcoming technology at the present time, as such it requires extra cost in its initial set up. Big data has to deal with high volume of data so it requires there lated infrastructure, sufficient transmission and storage capabilities to cope up. Moreover, there has to be a very high speed connectivity links to get data into and out of the system.

3.4.3 *Resource optimisation*

Optimization aims at making efficient use of the available resources so as to achieve the established goals. It includes resources like time, data, human manpower, monetary resources and physical resources. Being a new technology, the number of skilled people to handle big data is precisely low. Other resources also need expert monitoring as such create new opportunities.

3.4.4 *Governance and legal aspect*

As per stats the big data technology and services have grown nearly to 17 Billion in 2015 [13, 20]. Big data is making governments faster, smarter and more accountable. The Digital India initiative recently by the Government of India is a big data move. Big data is an upcoming area, as such the governments do not have the effective governance policies and procedures, thus need to have a look over it. Moreover, all countries in the world are having legal laws that every organisation needs to obey. These legislations enable us to preserve the various issues related to data sharing, ownership and others. These laws are new, and need some amendments at times, as with times more issues get raised [21].

3.5 Security and privacy issues

One of the critical things related with big data is the security of such vast volume of data. The rising security and privacy fears are mainly due to widespread data accessibility. Some key security and privacy issues are:

3.5.1 *Leakage of sensitive data and personal information*

Data leakage in clouds poses a foremost concern [22–24], while considering the huge amount of data that big data holds. The adequate man power, skill and infrastructure that could endow us with the security and privacy of data are lacking. Moreover, the present day data protection techniques are more susceptible to get easily breached. The activities of data hackers are becoming more terrifying due to the large volume of publically accessible available data.

Password controlled access and cryptographic approaches are being used since past for the security of data. However, these both are inadequate, as password can be at many times easily predicted or monitored and stolen. Similarly to imply encryption technique with such a large volume and variety of data is all together a new challenge in itself.

3.5.2 *Supervision over our data*

Almost all the people who are using internet are now worried about their data. A sense of being under surveillance or observation is prevailing among people as many business agencies are closely monitoring the data that people share over internet. Each move of the people (likes, dislikes and even moods) are being watched. There are high chances of using such vital data about a person against him. As an example, consider the cell phone location data of the person may help in finding the work addresses/home addresses of a person which the person might wish to hide. Similarly the frequent travel patterns may depict ones personal habits.

3.6 Forecasting issues

Forecasting is the method used to predict the personality of one thing based on the portrayal of the other related thing. Forecasting is based on the past and present data and examination of trends. The ability of big data to improve the organisational performance cannot be questioned. However, forecasting faces various issues as discussed in [25]:

3.6.1 *Lack of forecasting tools*

Considering big data in view, the traditional forecasting tools cannot hold the vast volume of size, speed and complexity of the data. Data mining tools may ease this problem to a certain extent but still are not up to mark.

3.6.2 Lack of expert staff

At present the world is facing shortage of the adequate capable expert staff that can help in dealing big data. As a new technology the world is facing a significant scarcity of analytical professionals and data scientists who can evaluate the data in the big data analytics industry. As per the stats [26] the United States alone is facing deficit of around 140,000–190,000 expert analytical professionals.

3.6.3 Impurities in data

Big data is having considerably a high level of noise [25]. Presence of noise poses a threat to the actual signal data. With raising level of noise in data, there is a threat to the forecasting big data process, as the noise is going to misrepresent the forecast results.

3.6.4 Lacking efficient algorithms for big data forecasts

The techniques used in big data must be able to convert the unstructured data into structured one and wisely capture the real time changes in data. Data mining algorithms are often not able to handle such huge vast volume of data [25]. These methods are designed relatively for low level data and thus cannot forecast results efficiently.

4 Conclusion

This paper gives a basic idea about the big data and also tries to identify the various challenges faced by this new technology. The data present in this world is immense and it continues to grow up with every passing second. It is clear from the research based on the big data that by 2020, the data present in this world will nearly get doubled [27]. As such there is a need to tackle the issues that we are going to face because of this technology in the near future. All these issues are briefly highlighted in this paper. In addition, big data analysis is the frontier that is going to boost the business economy in the near future. As such the business organisations need to improve tactics of data analytics in order to get an edge over other contenders. The V's model which characteristics the big data are actually the challenges we need to address and tackle. So to reap the benefits of "Big Data", we need to transform the challenges into strengths.

References

1. Khan I, Naqvi SK, Alam M, Rizvi SNA (2017) An efficient framework for real-time tweet classification. Int J Inf Technol 9(2):215–221
2. Kumar S, Raza Z (2017) Using clustering approaches for response time aware job scheduling model for internet of things (IoT). Int J Inf Technol 9(2):177–195
3. Adam K, Hammad I, Fakhreldin MA, Jasni MZ, Mazlina AM (2015) Big data analysis and storage. Proceedings of the 2015 international conference on operations excellence and service engineering, Orlando, Florida, USA, 10–11 Sept 2015, pp 648–659
4. Cloud Security Alliance (2016) Top ten big data security and privacy challenges. United States, Nov 2012. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten_v1.pdf. Accessed 29 Feb 2016
5. Bansal S, Rana DA (2014) Transitioning from relational databases to big data. Int J Adv Res Comput Sci Softw Eng 4(1):626–630
6. Mundial FE, World Economic Forum (2016) Big data, big impact: new possibilities for international development, Switzerland. http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf. Accessed 10 Mar 2016
7. The big data explosion—a big problem or big opportunity? https://www.youtube.com/watch?v=9UMNvlhgoZ0. Accessed 01 Sept 2016
8. Abraham Y (2015) What happens in an internet minute? How to capitalize on the big data explosion? http://www.excelacom.com/resources/blog/what-happens-in-an-internet-minute-how-to-capitalize-on-the-big-data-explosion. Accessed 07 May 2015
9. Kelly L (2016) Update: what happens in one internet minute? http://www.excelacom.com/resources/blog/2016-update-what-happens-in-one-internet-minute. Accessed 29 Feb 2016
10. Biswas SS, Agarwal P (2017) Big data in climate change. Glob J Res Anal 6(7):412–413 **(ISSN No 2277-8160)**
11. Big Data (2016) https://en.wikipedia.org/w/index.php?title=Big_data&oldid=704736013. Accessed 15 Feb 2016
12. Hashem IA, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of, big data, on cloud computing: review and open research issues. Inf Syst 47:98–115
13. Vesset D, Woo B, Morris HD, Villars RL, Little G, Bozman JS, Borovick L, Olofson CW, Feldman S, Conway S, Eastwood M (2012) Worldwide big data technology and services 2012–2015 forecast. IDC report. pp 233485
14. Noh H, Min JK (2013) An efficient data access method exploiting quadtrees on mapreduce frameworks. In international conference on database systems for advanced applications. Springer Berlin Heidelberg, 22 Apr 2013, pp 86–100
15. Agarwal P, Biswas SS (2017) Big data on cloud: a review. Int J Adv Res Comput Sci 8(2):49–51 **(ISSN: 0976-5697)**
16. Agrawal D, Bernstein P, Bertino E et al (2012) Challenges and opportunities with big data: a white paper prepared for the computing community consortium committee of the computing research association. http://cra.org/ccc/resources/ccc-led-whitepapers/
17. Kaisler S, Armour F, Espinosa JA, Money W (2013) Big data: issues and challenges moving forward. In system sciences (HICSS), 2013 46th Hawaii international conference on 2013 Jan 7, IEEE, pp 995–1004
18. Big Data Preliminary Report (2014) ISO/IEC JTC 1 information technology. https://www.iso.org/files/live/sites/isoorg/files/developing_standards/docs/en/big_data_report-jtc1.pdf. Accessed 05 Feb 2016
19. Musen K, Deng A, Alarcon T, Gil Y (2016) Data science in the news: advances and challenges for the era of big data, 24 Aug 2015. ftp://isi.edu/isi-pubs/tr-702.pdf. Accessed 12 Feb 2016
20. Géczy P (2014) Big data characteristics. Macrotheme Rev 3(6):94–104
21. Nasser T, Tariq RS (2015) Big data challenges. J Comput Eng Inf Technol 4:3. https://doi.org/10.4172/2324,9307(2)

22. Schmitt C, Shoffner M, Owen P, Wang X, Lamm B, Mostafa J, Barker M, Krishnamurthy A, Wilhelmsen K, Ahalt S, Fecho K (2013) Security and privacy in the era of big data: the SMW, a technological solution to the challenge of data leakage. RENCI White Paper Series. Text, vol 1(2). RENCI, University of North Carolina at Chapel Hill. https://doi.org/10.7921/G0WD3XHT
23. Raja J, Ramakrishnan M (2016) A comprehensive study on big data security and integrity over cloud storage. Indian J Sci Technol 9(40):1–6
24. Kishore N, Sharma S (2016) Secured data migration from enterprise to cloud storage-analytical survey. BVICAM's Int J Inf Technol 8(1):965–968
25. Hassani H, Silva ES (2015) Forecasting with big data: a review. Ann Data Sci 2(1):5–19
26. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C et al (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, Washington
27. The future of big data analytics—global market and technologies forecast—2015–2020. http://www.prnewswire.com/news-releases/the-future-of-big-data-analytics—global-market-and-technologies-forecast—2015-2020-275637471.html. Accessed 10 Mar 2016