Review Article

Indian Journal of Library and Information Science
Volume 11 Number 1, January - April 2017
DOI: http://dx.doi.org/10.21088/ijlis.0973.9548.11117.11

Big Data: A Bird's Eye View

Vishnu Kumar Gupta*, Pawan K. Saini**

Abstract

Big data is being produced by everybody around us round the clock. Newly emerged big data technology is the biggest invention in the field of computer science and technology during the last decade. We have to create rules and laws to prevent abuse of big data along with new developments of this technology enabling useful capabilities. For information scientists and librarians, big data is a big deal and great challenge, and also a good opportunity to play a significant role in the ever growing universe of big data.

Keywords: Big Data; Data Explosion.

Introduction

The term 'Big data' not only draws a lot of attention but also becomes a hot topic now-a-days. When checking the historical record of big data, there show to be many occasions that will help elaborate its birth. This term was popularized and coined by Roger Magoulas in 2005, when he was "the director of market research at O'Reilly Media and leads a team that builds an open source analysis infrastructure and provides analysis services, including technology trend analysis, to business decision-makers at O'Reilly and beyond [1]." In 2007, this concept earned popularity with the release of "Apache open source project Hadoop beta version [2]."

Big data as a buzzword in the field of information technology used to explain "an oceanic volume of structured and unstructured data, that is too big in size that it is very difficult to process using traditional database and software [3]." The phrase is also known as 'Enterprise big data' in the field of business enterprises. Presently, in most business enterprises, "the data is too large and runs very fast which exceeds current processing capacity. Big data has the sufficient capability to help enterprises to improve their activities and operations, which makes decisions faster and more intelligent. It originated with Web search companies who had the problem of querying very large distributed aggregations of loosely-structured data [4]." Google originated "MapReduce to support distributed computing on large data sets on computer clusters [5]." On the basis of Google's MapReduce, and Google File System (GFS) papers, "Doug Cutting developed Hadoop while he was at Yahoo, and entitled it after his son's stuffed elephant [6]."

Big data is being produced by everybody around us round the clock. Every modern digital technology and social media, and social sites generates it. Smart phones, mobile gazettes and sensor devices disseminate it very fast. It is coming from all the directions from various sources at an alarming speed, and quantity. Analytical capabilities, processing power, and optimal skills are needed to get meaningful value from big data. It is developing a culture in which technologists, CEOs, and business leaders have to join forces to realize real value of data. It deeply enables all human resources to make better decisions, which focuses on customer satisfaction, optimizing operations, preventing frauds and threats, and generating new revenue sources. But accelerating

Author's Affiliation: *Librarian, Deptt. of Bioscience and Biotechnology Library, Banasthali University (Rajasthan)-304022. **Assistant Professor, DLIS, Central University of Haryana, Mahendergarh.

Reprint's Request: Vishnu Kumar Gupta, Librarian, Deptt. of Bioscience and Biotechnology Library, Banasthali University (Rajasthan)- 304022.
E-mail- vishnu5966@gmail.com

Received on 23.11.2016, **Accepted on** 07.12.2016

demand for insights really needs some elementary new approaches and skills. These new approaches and skills are required to completely control the power of big data.

Previous Studies Related to Big Data

Many organizations and enterprises are massively looking to find actionable insights into their data. Various big data projects develop from the need to answer particular social and business-oriented questions. Any organization or enterprise can increase sales, boost efficiency, and make better operations in risk management and customer service with the help of right big data analytics.

Quin Street, which is a Webopedia parent company and founded by Doug Valenti in 1999, extensively surveyed "540 enterprise decision-makers involved in big data purchases to learn which business areas companies plan to use big data analytics to improve operations. About half of all respondents said they were applying big data analytics to improve customer retention, help with product development and gain a competitive advantage. Notably, the business area getting the most attention relates to increasing efficiencies and optimizing operations. Specifically, 62 percent of respondents said that they use big data analytics to improve speed and reduce complexity [19]."

It is a widespread belief that the proper availability and easy access to research data, in particular to 'big data' shows the apex form of knowledge and intelligence, which makes an arena of accuracy, truth, and objectivity. When big data is viewed "as an answer of too many crucial problems and questions, it is considered as an apparatus that warns privacy, shortens civil freedoms, and ushering increased state. The shifts to be expected of big data are probably more subtle than these, even though we cannot see this clearly among our current hopes and fears [20]."

Boyd and Crawford's (2012) understanding of the big data as a phenomenon, explains "acknowledging that the decisive factor is not the attribute of data, but the ability to search, aggregate, and cross-reference large data bases by virtue of the storage and processing capacity of modern information and communication tools and techniques, such as internet, WWW, computers etc. Information literate researcher scholars and information scientists must know that data is no more an exclusive issue for the science, technology and medicine, but it is present in the social sciences, humanities, arts and culture, as well [20]."

Ajana (2015) shows many hot issues in light of the application of big data in the field of immigration management and border security. "Large financial investment in the computing technologies of borders and their securitization continues to be a focal point for many governments across the globe [21]." He focused with a specific instance such technologies, i.e. 'Big Data' analytics. During the last 20 years, "the technology of big data has achieved an unusual popularity among a variety of arenas, such as business, government, scientific and research fields. While big data techniques are often extolled as the next frontier for innovation and productivity, they are also raising many ethical issues [21]."

Data sets containing so much, perhaps some kind of sensitive data, and the tools and techniques to extract and make use of this information give rise to many other chances for illegal use and unauthorized access. Much of our preservation of privacy in society depends on current inefficiencies. For instance, CCTV/video cameras always eyed people in many places, viz. Airport security lines, ATMs, urban intersections, and convenience stores. If these technological players are connected or networked together, and modern sophisticated computing technology makes it possible to correlate and analyze these data streams, the prospect for abuse becomes significant. Furthermore, cloud computing technology becomes a cost-effective tool for malicious agents, e.g., to apply massive parallelism to break a cryptosystem or to launch a zombie/botnet.

A newly developed application of big data is the addition of sensors and other micro electronic devices to engineer to order (ETO) goods such as one of a kind ships and buildings. The proper set up and operation of smart ships and smart buildings function with the help of micro electronic sensor devices. This is necessary to examine what challenges will need to be met before project businesses can achieve informational effects and transformational effects from big data technologies. A study of Fox and Do (2013) reveals "a causal mechanism and causal context for project business big data application. This type of critical realist analysis can be applied to enable better understanding of necessary causal mechanisms and causal contexts for other ICT innovations [22]."

Trottier (2014) finds that big data is always meaningful in use. "While they may be contained in databases in remote locations, big data do not exist in a social vacuum. Their impact cannot be fully understood in the context of newly assembled configurations or 'game-changing' discourses. Instead, they are only knowable in the context of

existing practices. These practices can initially be the sole remit of public discourse shaped by journalists, tech-evangelists and even academics. Yet embodied individual and institutional practices also emerge, and this may contradict or at least complicate discursive assertions. Moreover, the range of devices and practices that make up big data are engaged in a bilateral relation with these practices. They may be a platform to further reproduce relations of information exchange and power relations. Yet they may also reconfigure these relations [23]."

Using two case studies approaches, first, online community specialist groups linked to rural activities, and second, from a policy shift relating to firearm legislation in the English context, Hillyard (2014) describes how the "technologies of big data might apply to rural contexts. It considers the relative advantages and disadvantages of such 'new' innovations [24]." He provides "insight into the rural context and makes a case that such locales are not immune from the influence of the data verse. The appearance of 'big data' is not without political implications. The case of UK firearm legislation reform demonstrates the implications of policy falling short of its potential and how a social science analysis can unpack the operation of power as well as position the debate more broadly [24]."

Prescott (2014) illustrates how Nielsen Holdings, a global company, reacted to transforms in their optimal competitive industry brought about by latest developments in computing technology. This study shows "the strategic management decisions that enabled Nielsen to regain its competitive advantage. It furthermore describes the functioning of the resource-based view (RBV) of strategy, dynamic capabilities framework, and digital data genesis (DDG), in a turbulent business environment [25]."

Girtelschmid and his team (2014) propose and evaluate "a novel system architecture for Smart City applications which uses ontology reasoning and a distributed stream processing framework on the cloud [26]." Generally, automated inference and semantic modeling methodologies are applied in the field of Smart City. When applied in large scale, semantic models faced performance problems. They addressed "the problem domain by using methods from Big Data processing in combination with semantic models. The architecture is designed in a way that for the Smart City model still traditional semantic models and rule engines can be used. However, sensor data occurring at such Smart Cities are pre-processed by a Big Data streaming platform to lower the workload to be processed by the rule engine [26]."

According to Wiseman (2014), "the development of a knowledge society in the Arabian Gulf is a nested and contextualized process that relies upon the development of nation-specific knowledge economies and region-wide knowledge cultures. The role of internationally comparative education data and mass education systems in the Gulf as mechanisms for the development of knowledge economies, societies, and cultures are discussed and debated in relation to the unique contextual conditions countries operate within. The role of big data and mass education in creating expectations for achievement, accountability, and access is shown to significantly contribute to the development of knowledge societies by providing the infrastructure and capacity for sustainable change, which potentially leads to the institutionalization of knowledge acquisition, exchange, and creation in the Gulf and beyond [27]."

Rogers and Gravelle (2011) feels that "As the government's strategy for the implementation of the 'Big Society' gains momentum within an increasingly difficult financial framework. They discuss some of the major implications of this approach for partnership working in crime and disorder reduction. It considers whether the approach is a totally new one or merely an extension of previous government policy, while considering some of the advantages and disadvantages of extending the 'Big Society' ideology [28]. While considering the key problems of implementing such an approach, the authors also emphasizes the opportunities that might show themselves for enhanced community consultation in the delivery of partnership working.

What is Big Data: a Technology or a Volume?

As mentioned above, the phrase 'Big data' might be representing the very huge volume of data, which is not always true. This phrase, particularly used by technologists, business enterprises and vendors, may refer to the technology that requires managing the large amount of data. This term is originated with the origin of World Wide Web and Web search engine enterprise groups, viz. Microsoft, Yahoo, Google, etc., who required handling extremely big distributed collections of loosely structured data. Everyone leave digital files of his daily activities. E-mails are saved in corporate information systems; social media sites updates are filed; and phone conversations and chatting are saved and stored in digital formats.

According to McKinsey, big data is defined as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze [29]." Edd Dumbill of O'Reilly Media refers to it as "data that exceeds the processing capacity of

conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures [29]."

We are producing huge volume of data that are quantitatively very difficult to imagine. Several things that have changed are now able to analyze more complex varieties of data such as video recordings in CCTV, including photo images and conversation; smart phone records of conversations, etc. In the universe of big data, the following "4 Vs are characterized to define big data:

Volume– the huge amount of data generated every moment;

Velocity– the speed at which new data is generated and moves around. Credit card fraud detection is a good example where millions of transactions are checked for unusual patterns in almost real time;

Variety– the increasingly different types of data from financial data to social media feeds, from photos to sensor data, from video capture to voice recordings; and

Veracity– the messiness of the data, just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech.

So, we have a lot more data than ever before, in more complex formats, that are often fast moving and of varying quality- why would that change the world? The difference is that we now have tools that allow us to analyze vast amounts of data by breaking the task of processing very large data sets down into smaller tasks that are run in parallel using a large cluster of computers [7]."

Examples of Big Data

Some explicit examples of big data are petabytes (2^{10} terabytes), exabytes (2^{10} petabytes), zettabytes (2^{10} exabytes), and yottabytes (2^{10} zettabyte) of data containing of billions to zillions of records of a majority of people from various sources, viz. World Wide Web (WWW), marketing and sales figures, customer care centre, social media sites, telecommunication and mobile companies' data, and so on so far. This data may be considered as loosely structured data which is generally incomplete, unfinished and properly not accessible.

Big Data Explosion: How Measuring the Big Data?

According to IDC "1.8 zettabytes (which is equal to 1.8 trillion GBs) of information was generated by the World in 2011, which is enough data to fill 57.5 billion 32GB Apple iPads. That is enough iPads to build a Great iPad Wall of China twice as tall as the original. In 2012, it reached 2.8 zettabytes and IDC now forecasts that we will generate 40 zettabytes (ZB) by 2020 [8]."

As explained earlier and looking at the "sheer volume of 1.8 zettabytes of data, which is equivalent to:

- Every person in the United States of America tweeting 3 tweets per minute for 26,976 years nonstop.
- Every person in the world having more than 215 million high-resolution MRI scans per day.
- More than 200 billion HD movies (each 2 hours in length)—would take 1 person 47 million years to watch every movie 24x7.
- The amount of information needed to fill 57.5 billion 32GB Apple iPads. With that many iPads we could:
  - Create a wall of iPads, 4,005-miles long and 61-feet high extending from Anchorage, Alaska to Miami, Florida.
  - Build the Great iPad Wall of China—at twice the average height of the original.
  - Build a 20-foot high wall around South America
  - Cover 86% of Mexico City.
  - Build a mountain 25-times higher than Mt. Fuji [10].

The quantity of data is rapidly growing at a dangerous and exciting speed. In the words of Eric Schmidt, Google's Executive Chairman: "From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days…and the pace is accelerating [10]."

The main driving forces behind this never stopping and tremendous growth are driven by modern Information and Communication (ICT) and money. "New information taming technologies are driving the cost of creating, capturing, managing and storing information down to one-sixth of what it was in 2005. Additionally, since 2005 annual enterprise investments in the Digital Universe—cloud, hardware, software, services, and staff to create, manage, store and generate revenue from the information—have increased 50% to $4 trillion (USD) [11]."

It is very surprising fact that "data is exploding, but how much quantum of data is out there? Gartner predicts that enterprise data will grow 650 percent in the next five years, while IDC argues that the world's information now doubles about every year and a half. Twitter was the fastest growing social network in 2008 by 1382 % [12]." Again, according to the Gartner,

"the total number of text messages sent in 24 hours is more than 6,700,000,000, which exceeded the total population of the planet [12]."

Now, consider the following candid facts to understand the data explosion. "Every minute of every day people on this Earth create:

(i) Over 204 million email messages;

(ii) More than 2 million Google search queries;

(iii) About 48 hours of new YouTube videos;

(iv) 684,000 bits of content shared on Social Networking Site Facebook;

(v) Over 100,000 tweets; and

(vi) US $ 272,000 spent on e-commerce [13]."

Structured Data

Structured Data is data that "belongs to a fixed field within a record or file. It includes data containing in relational databases and spreadsheets. Structured data initially depends on creating a data model- a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type, such as numeric, currency, alphabetic, name, date, address; and any restrictions on the data input, such as number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M or F [14]."

Generally, structured data is managed by using a programming language, called Structured Query Language (SQL). SQL is basically programmed by International Business Machines Corporation (IBM) in the beginning of 1970s for handling, searching, querying, and answering data in relational database management systems (RDBMs).

An advantage of structured data is that it can be easily entered, stored, analyzed, accesses, and queried. Some decades ago, due to the high input cost and performance constraints of storage and processing, spreadsheets and relational databases using structured data were the only way to efficiently and effectively manage data. At that time, any data or information that could not properly fix within a compressed organized structure would have to be fitted and stored on traditional methods on papers in a traditional filing cabinet.

Unstructured Data

Big data is directly embedded with unstructured data and simply means massively huge data sets that are very hard to manage and analyze with conventional old methods, tools and techniques. Big data includes both structured and unstructured data, but IDC estimates "that 90 percent of big data is unstructured data [15]." Several methods and tools are designed to manage and analyze big data, which can easily manage unstructured data.

Unstructured data generally refers to "information that doesn't reside in a traditional row-column database. As one may expect, it is the opposite of structured data, the data stored in fields in a database. Unstructured data files often contain text and multimedia content [16]." Some examples of unstructured data are web sites, e-mails, word processing documents, videos files, photos, audio files, Power point presentations, and several other varieties of academic, social, political, and business related documents. Here it is important to know that when these kinds of records might have an own internal structure, they are yet considered 'unstructured' doe to the data they have does not fit clearly in a data set or database. Some specialists calculated that "80 to 90 percent of the data in any organization is unstructured and the amount of unstructured data in organizations is growing faster than structured databases [16]."

Semi-Structured Data

Semi-structured data is a bridge between the two. It is considered as a kind of structured data, but lacks the rigid data model structure. With semi-structured data, tags and markers are used to recognize some certain components within the data, but the data do not have a rigid structure. For instance, now word processors (software) may contain metadata presenting the name of the author, and the creation date, along with the heap of the document just being unstructured text. E-mails contain the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments. Photos and other graphics can be tagged with keywords such as the creator, date, location and keywords, making it possible to organize and locate graphics. Extended Mark-up Language (XML) and other markup languages are generally applied to handle semi-structured data.

Big Data Analytics

Big data analytics refers to "the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. Big data analytics will help organizations to better understand the information contained within the data and will also help identify the data that is most

important to the business and future business decisions. Big data analysts basically want the knowledge that comes from analyzing the data [17]."

As technology to cut down data silos and analyze data improves, business may be changed in all sorts of ways. According to *Datamation*, "today's advances in analyzing big data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook. The business cases for leveraging big data are compelling. For instance, Netflix mined its subscriber data to put the essential ingredients together for its recent hit House of Cards, and subscriber data also prompted the company to bring Arrested Development back from the dead [18]."

Another example is from the largest mobile carriers in the world. Orange, a French company launched its "Data for Development project by releasing subscriber data for customers in the Ivory Coast. The 2.5 billion records, which were made anonymous, included details on calls and text messages exchanged between 5 million users. Researchers accessed the data and sent Orange proposals for how the data could serve as the foundation for development projects to improve public health and safety. Proposed projects included one that showed how to improve public safety by tracking cell phone data to map where people went after emergencies; another showed how to use cellular data for disease containment [19]."

Conclusion

Big data is neither a fixed material nor it have enchanted power to avail latest business analytics on its own. Big data is also not a complete technology. However, it is a paradigm change in the level of thinking on how to achieve insight from data with augmented volumes and various continue changing formats.

Big data is an umbrella term used for complex data sets and traditional data processing techniques are inadequate to handle them. Some challenges to manage these data sets are capture, analysis, storage, search, curation, sharing visualization, privacy, and transfer of data. Experts and technologists of this arena have only started to see its power to gather, manage, and process data in everyone's life. In short, any discussion about big data could not be complete without describing the augmenting aspects about individual's freedom. Several aspects have been presented, viz. how credit card companies, retailers, search engine providers and mail or social media sites make use of our private data. All in all, every big thing, big idea, big business model, big organization, big society, big country, big money, big technology, big problem, and big solution is driven by big data. Undoubtedly, newly emerged big data technology is the biggest invention in the field of computer science and technology during the last decade. We have to create rules and laws to prevent abuse of big data along with new developments of this technology enabling useful capabilities.

For information scientists and librarians, big data is a big deal and great challenge, and also a good opportunity to play a significant role in the ever growing universe of big data. Information scientists and librarians are fully skilled, and aware with the service-oriented knowledge to help and support universities, businesses, nonprofit organizations, and governments.

References

1. http://www.oreilly.com/pub/au/2717
2. http://en.wikipedia.org/wiki/List_of_Apache_Software_Foundation_projects
3. http://www.webopedia.com/TERM/B/buzzword.html
4. http://wikibon.org/wiki/v/Enterprise_Big-data
5. http://en.wikipedia.org/wiki/MapReduce
6. http://en.wikipedia.org/wiki/Doug_Cutting
7. http://analyticsweek.com/big-data-the-mega-trend-that-will-impact-all-our-lives-bernard-marr-linkedin/
8. Richard L. Villars and Carl W. Olofsons, (2011), "Big data: What it is and why you should care?", IDC Report.
9. http://www.emc.com/about/news/press/2011/20110628-01.html
10. http://analyticsweek.com/big-data-the-mega-trend-that-will-impact-all-our-lives-bernard-marr-linkedin/
11. http://www.emc.com/about/news/press/2011/20110628-01.html
12. Raymond Paquet, (2010), "Gartner Webinar: technology trends you can't afford to ignore", Gartner Webinar.
13. http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html
14. http://www.webopedia.com/TERM/S/structured_data.html

15. Richard L. Villars and Carl W. Olofsons, (2011), "Big data: What it is and why you should care?", IDC Report.
16. http://www.webopedia.com/TERM/U/unstructured_data.html
17. http://www.webopedia.com/TERM/B/big_data_analytics.html
18. http://www.ultramaxit.com/big-data-analyics/
19. http://www.webopedia.com/TERM/B/big_data_analytics.html
20. D. Boyd and K. Crawford. "Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon", Information, Communication and Society, 2012; 15(5):662-679.
21. Btihaj Ajana. "Augmented borders: Big Data and the ethics of immigration control", Journal of Information, Communication and Ethics in Society, 2015; 13(1):58-78. DOI: http://dx.doi.org/10.1108/JICES-01-2014-0005.
22. Stephen Fox and Tuan Do. "Getting real about Big Data: applying critical realism to analyse Big Data hype", International Journal of Managing Projects in Business, 2013; 6(4):739-760. DOI: http://dx.doi.org/10.1108/IJMPB-08-2012-0049.
23. Daniel Trottier. Big Data Ambivalence: Visions and Risks in Practice, in Martin Hand, Sam Hillyard (ed.) Big Data? Qualitative Approaches to Digital Research (Studies in Qualitative Methodology, Emerald Group Publishing, 2014; 13:51-72. DOI: 10.1108/S1042-319220140000013004.
24. Sam Hillyard. 'Where No-One Can Hear You Scream': An Analysis of the Potential of 'Big Data' for Rural Research in the British Context, in Martin Hand , Sam Hillyard (ed.) Big Data? Qualitative Approaches to Digital Research (Studies in Qualitative Methodology) Emerald Group Publishing, 2014; 13:231-249. DOI: 10.1108/S1042-319220140000013014.
25. Michael E. Prescott. "Big data and competitive advantage at Nielsen", Management Decision, 2014; 52(3):573-601. DOI: http://dx.doi.org/10.1108/MD-09-2013-0437.
26. Sylva Girtelschmid , Matthias Steinbauer , Vikash Kumar , Anna Fensel , Gabriele Kotsis. "On the application of Big Data in future large-scale intelligent Smart City installations", International Journal of Pervasive Computing and Communications, 2014; 10(2):168-182. DOI: http://dx.doi.org/10.1108/IJPCC-03-2014-0022.
27. Alexander W. Wiseman (2014) "Strategically planning the shift to a Gulf knowledge society: The role of big data and mass education" in Alexander W. Wiseman , Naif H. Alromi , Saleh Alshumrani (ed.) Education for a Knowledge Society in Arabian Gulf Countries (International Perspectives on Education and Society) Emerald Group Publishing, 2014; 24:277-304.
28. Colin Rogers and James Gravelle. "Partnership working in the Big Society", Safer Communities, 2011; 10(2):26-31. DOI: http://dx.doi.org/10.5042/sc.2011.0183.
29. http://radar.oreilly.com/2012/01/what-is-big-data.html.