BIG DATA TECHNOLOGY: APPLICATION IN LIBRARIES AND SOCIETY

Satish Kumar,
Information Scientist,
ARIES, Nainital (UK)
Email – sklisc@gmail.com

ABSTRACT

Big Data is the digital data that is rapidly increasing on three fronts of volume, velocity and variety. Analysis of big data reveals patterns, trends and associations that can be used by businesses, academic institutions and libraries. Big Data has the potential to revolutionize not just research, but also education. Libraries have amassed an enormous amount of machine-readable data about library collections, both physical and electronic. Librarians are well skilled to discuss with researchers the value of data management and sharing strategies, and to increase their awareness of metadata standards and practices and institutional repositories. This paper introduces the concept of big data. It also discusses in detail, big data applications in libraries especially data mining, data curation and research data management.

Keywords: *Big Data, Data Mining, Data Curation, Research Data Management.*

INTRODUCTION

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes (10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data.

The complex nature of big data is primarily driven by the unstructured nature of much of the data that is generated by modern technologies, such as that from web logs, radio frequency Id (RFID), sensors embedded in devices, machinery, vehicles, Internet searches, social networks such as Facebook, portable computers, smart phones and other cell phones, GPS devices, and call center records. In most cases, in order to effectively utilize big data, it must be combined with structured data (typically from a relational database) from a more conventional business application, such as Enterprise Resource Planning (ERP) or Customer Relationship Management (CRM).

Similar to the complexity, or variability, aspect of big data, its rate of growth, or velocity aspect, is largely due to the ubiquitous nature of modern on-line, real-time data capture devices, systems, and networks. It is expected that the rate of growth of

big data will continue to increase for the foreseeable future.

Specific new big data technologies and tools have been and continue to be developed. Much of the new big data technology relies heavily on massively parallel processing (MPP) databases, which can concurrently distribute the processing of very large sets of data across many servers.
As another example, specific database query tools have been developed for working with the massive amounts of unstructured data that are being generated in big data environments.

WHAT IS BIG DATA?

"Big Data are high---volume, high---velocity, and/or high---variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (**Gartner 2012**).

Complicated (intelligent) analysis of data may make a small data "appear" to be "big"

Bottom line: Any data that exceeds our current capability of processing can be regarded as "big"

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reductions and reduced risk.

Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, combat crime and so on.

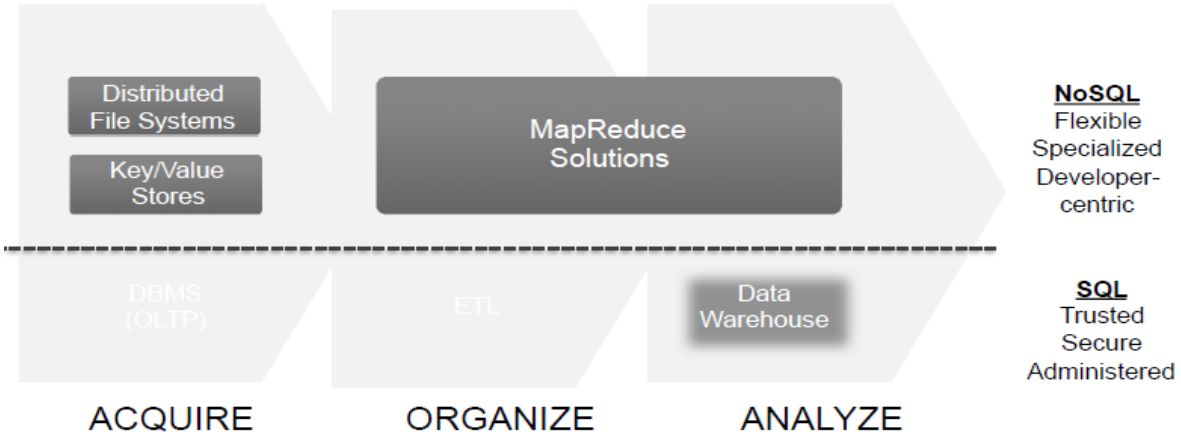

Fig. 1 Divided Big Data Spectrum

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

WHY BIG DATA?

Key enablers for the growth of "Big Data" are:

- ***Increase of data storage capacities*** – Data storage has grown significantly, shifting markedly from analog to digital after 2000.
- ***Increase of processing power*** – Computation capacity has also risen sharply.
- ***Availability of data*** – Organizations in all sectors have at least 100 terabytes of stored data, Data generation will increase due to social networks and internet of things (IoT).

WHY IS BIG DATA IMPORTANT?

When big data is effectively and efficiently captured, processed, and analyzed, companies are able to gain a more complete understanding of their business, customers, products, competitors, etc. which can lead to efficiency improvements, increased sales, lower costs, better customer service, and/or improved products and services.

CHARACTERISTICS OF BIG DATA

Big Data is not just about the size of data but also includes data variety and data velocity. Together, these three attributes form the three V's of Big Data:

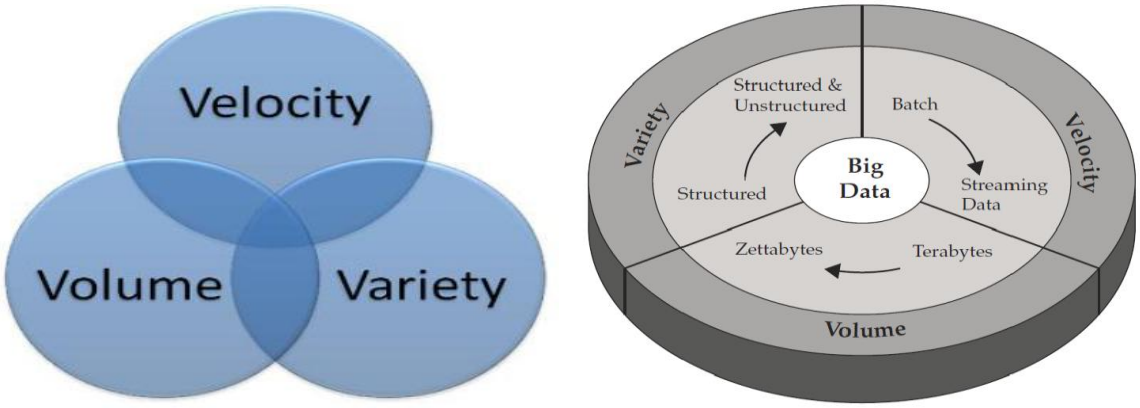

Fig. 3 Characterization of Big Data – Volume, Velocity, Variety (V3)

- **Volume**

  The quantity of generated data is important in this context. The size of the data determines the value and potential of the data under consideration, and whether it can actually be considered big data or not. The name 'big data' itself contains a term related to size, and hence the characteristic.

- **Variety**

  This helps people who analyze the data and are associated with it effectively use the data to their advantage and thus uphold the importance of the big data.

- **Velocity**

  'Velocity' in this context means how fast the data is generated and processed to meet the demands and the challenges that lie in the path of growth and development.

- **Variability**

  This refers to inconsistency the data can show at times—which hampers the process of handling and managing the data effectively.

- **Veracity**

  The quality of captured data can vary greatly. Accurate analysis depends on the veracity of source data.

- **Complexity**

Data management can be very complex, especially when large volumes of data come from multiple sources. Data must be linked, connected, and correlated so users can grasp the information the data is supposed to convey.

Data can come from a variety of sources (typically both internal and external to an organisation) and in a variety of types. With the explosion of sensors, smart devices as well as social networking, data in an enterprise has become complex because it includes not only structured traditional relational data, but also semi-structured and unstructured data.

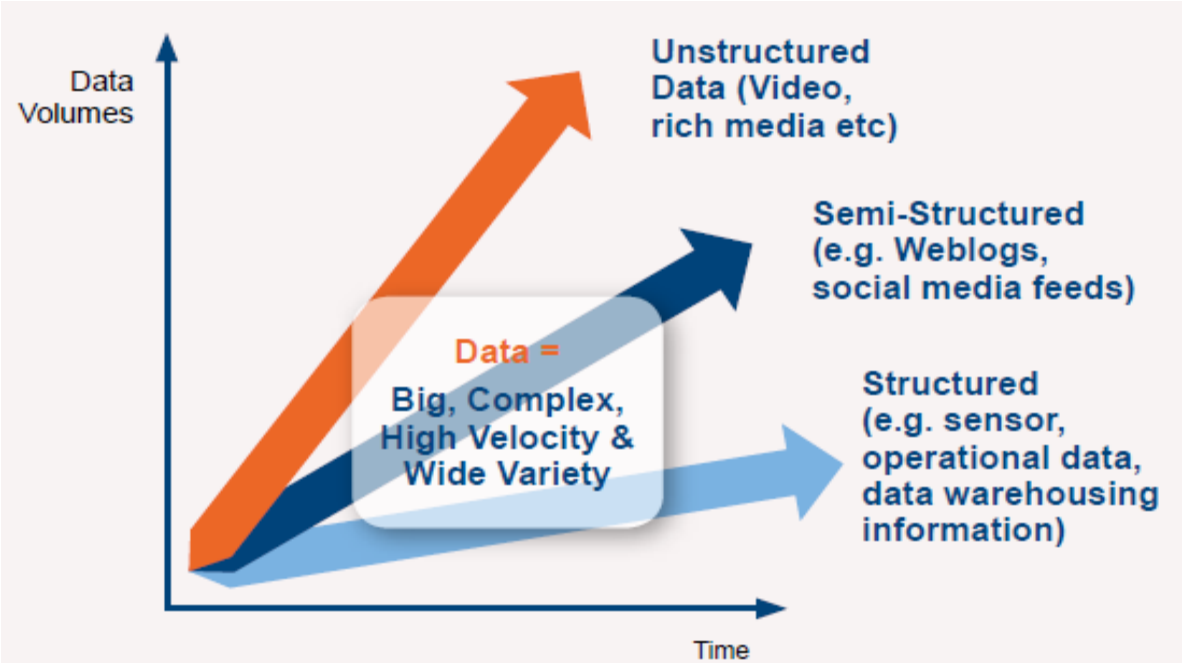

Fig. 4 Characterization of Big Data – Structured, Semi-Structured, Unstructured

- **Structured data:** This type describes data which is grouped into a relational scheme (e.g., rows and columns within a standard database). The data configuration and consistency allows it to respond to simple queries to arrive at usable information, based on an organisation's parameters and operational needs.
- **Semi-structured data:** This is a form of structured data that does not conform to an explicit and fixed schema. The data is inherently self-describing and contains tags or other markers to enforce hierarchies of records and fields within the data. Examples include weblogs and social media feeds.
- **Unstructured data:** This type of data consists of formats which cannot easily be indexed into relational tables for analysis or querying. Examples include images, audio and video files.

The velocity of data in terms of the frequency of its generation and delivery is also a characteristic of big data. Conventional understanding of velocity typically considers how quickly the data arrives and is stored, and how quickly it can be retrieved. In the context of Big Data, velocity should also be applied to data in motion: the speed at which the data is flowing. The various information streams and the increase in sensor network deployment have led to a constant flow of data at a pace that has made it impossible for traditional systems to handle.

Handling the three Vs helps organisations extract the value of Big Data. The value comes in turning the three Vs into the three **Is:**

1. **Informed intuition:** predicting likely future occurrences and what course of actions is more likely to be successful.

2. **Intelligence:** looking at what is happening now in real time (or close to real time) and determining the action to take

3. **Insight:** reviewing what has happened and determining the action to take.

APPLICATION OF BIG DATA IN INDIA

The use and adoption of Big Data within governmental processes is beneficial and allows efficiencies in terms of cost, productivity, and innovation. That said, this process does not come without its flaws. Data analysis often requires multiple parts of government (central and local) to work in collaboration and create new and innovative processes to deliver the desired outcome. Below are the thought leading examples within the Indian Government Big Data space.

- Big data analysis was, in parts, responsible for the BJP and its allies to win a highly successful Indian General Election 2014.
- The Indian Government utilises numerous techniques to ascertain how the Indian electorate is responding to government action, as well as ideas for policy augmentation.

APPLICATION OF BIG DATA IN LIBRARIES

Automation of library circulation systems and development of Worldcat by OCLC can be called as the early applications of big data in libraries. Big data affects libraries both directly and indirectly. Direct effect is in the use of big data tools to analyze big data sets of libraries. Indirect effect is through the library users who are increasingly using big data in their research.

In order to deal with big data and data mining, there is no need for librarians to acquire new skills. They need only to refocus on new issues and requirements. Librarians who have practiced cataloguing and metadata creation have the ability of conceptualizing relationships among data. They can therefore, be the best persons to advise the researchers about the data management mechanisms to be adopted from the beginning of a research project so as to make it easier to collect, organize and preserve the data that will be generated.

BIG DATA CURATION

Librarians can make big data sets more accessible, visible and useful by creating metadata schemes and taxonomies and designing standard retrieval methods. Traditionally librarians have always worked with information sources, which are the finished products. Now they should also understand the value of the 'rawdata'. With the help of new big data analysis tools, one can look at data in different ways. Information Visualization Tools enable mining the raw data for new information different than the purpose for which it was used originally.

RESEARCH DATA MANAGEMENT

Research data is the data that is produced as a result of any research activity. Such data is very valuable but it is also diverse and heterogeneous in nature and can be huge in a university setup where active research is going on. This research data is a pool of raw data which can be mined by other researchers according to their needs.

This will save their precious time that would have been spent in data collection. But, such an activity needs proper planning for infrastructure and policy framework. Researchers are normally focused on

their own domains. It is very difficult to imagine how the data collected by them can be useful for others. But a facet of the data that may be useless to the researcher who collected it might be valuable for other researchers.

The librarians have to perform the task of data curation in such cases so as to make the data more valuable and available for other future uses, especially because predicting the future uses can be very difficult. What to save and in which format, are some of the crucial decisions to be taken.

Data archiving is the library skill that can be reoriented for big data management. For researchers, data management is nothing but data storage. It can be stored locally on their computer disks or in the cloud. But archiving actually means arranging, describing, documenting and preserving the data so as to enhance it and make it easily retrievable. This is very important because storage or preservation of data is meaningless without later access. If research data is archived properly, then it would make the research reproducible and also facilitate new research.

Such initiatives are being undertaken the world over. Their goal is to build large networks and repositories to provide support for effective data management and access. But in order to do this, it is necessary to find out the needs and habits of researchers, to develop powerful data analysis tools or adapt some from the available open source software.

Hathi Trust (2015) is one such co-operative effort of more than 100 university libraries. It is administered by Michigan and Indiana universities and is a collaborative repository of digital content from research libraries. It is "committed to the long-term curation and availability of the cultural record". Its textual data corpus can be used by researchers for text mining according to their needs.

The Digital Preservation Network (DPN, 2015) is the newest and one of the most ambitious project with a vision to preserve the complete scholarly record for future generations. It replicates multiple copies of various digital repositories in diverse nodes in order to protect from "the risk of catastrophic loss due to technology, organizational or natural disasters".

For proper functioning of such initiatives for Data mining, Data curation or Research data management, a smart policy framework is a necessity because here, personal research data is being used in new ways giving rise to many information sharing issues. Policy makers therefore, should create proper guidelines for promotion of both – an information sharing environment and researcher privacy. Guidelines should also define the extent and nature of use that can be considered fair dealing. Whether the researchers will be able to opt out or control their research data should also be specified clearly. Funding issues for maintenance of the data repository should also be a major concern.

Majority of the initiatives in big data management for preservation and access are headed by the librarians. But university officials are also involved in these projects. Preservation of big data and the cultural heritage should be a joint responsibility of libraries and universities or research institutions. This is because big data management involves social and economic issues in addition to the technology issues.

CONCLUSION

Big data are extremely large data sets that can be analyzed only with the help of specialized software. This gives rise to complex visualizations regarding patterns and trends in human behavior and interactions. Big data is expanding continually on three fronts of volume, velocity and variety. Consideration of other characteristics like veracity, variability, visualization and value is also very important for developing a big data program. Businesses were the first to adopt big data analysis followed by the education sector. Big data is finding applications in data mining, data curation and research data management. Libraries are at the fore front of such initiatives with administrative and

technological support from their universities or institutions.

It's important to remember that the primary value from big data comes not from the data in its raw form, but from the processing and analysis of it and the insights, products, and services that emerge from analysis. The sweeping changes in big data technologies and management approaches need to be accompanied by similarly dramatic shifts in how data supports decisions and product/service innovation. There is little doubt that analytics can transform organizations, and the firms that lead the charge will seize the most value.

Like many new information technologies, big data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings. Like traditional analytics, it can also support internal business decisions. The technologies and concepts behind big data allow organizations to achieve a variety of objectives.

REFERENCES:

1. Why is Big data important? March 2012. www.navint.com
2. The Economist. Building with big data. [Online] Available at: http://www.economist.com/node/18741392 [Accessed on 9th July 2015].
3. Edd Dumbill. What is big data? [Online] Available at: http://radar.oreilly.com/2012/01/what-is-big-data.html [Accessed on 9th July 2015].
4. Peter Buneman. Semistructured Data. [Online] Available at: http://homepages.inf.ed.ac.uk/opb/papers/PODS1997a.pdf [Accessed on 9th July 2015].
5. IDC. IDC's Worldwide Big Data Taxonomy, 2011. [Online] Available at: http://www.idc.com/getdoc.jsp?containerId=231099 [Accessed on 9th July 2015].
6. IDC. Worldwide Big Data Technology and Services 2012-2015 Forecast. [Online] Available at: http://www.idc.com/getdoc.jsp?containerId=233485 [Accessed on 9th July 2015].
7. Erik Brynjolfsson, et al. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance. [Online] Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486 [Accessed on 9th July 2015].
8. Cisco. The Internet of Things: How the Next Evolution of the Internet is Changing Everything. [Online] Available at: http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf [Accessed on 9th July 2015].
9. John Mahoney, Hung LeHong. The Internet of Things is Coming. [Online] Available at: http://www.gartner.com/id=1799626 [Accessed on 9th July 2015].
10. Arnold, K. E., and Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In S. Buckingham Shum, D. Gašević, & R. Ferguson (Eds.), Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK 2012) (pp. 267-270). New York: ACM.
11. Diebold, Francis X. (2012). A personal perspective on the origin(s) and development of 'Big Data': the phenomenon, the term, and the discipline. Second version. PIER Working Paper No. 13-003. Retrieved from http://dx.doi.org/10.2139/ssrn.2202843 on 9.7.2015.
12. Digital Preservation Network. (2015). Retrieved from http://www.dpn.org/ on 13.7.2015.

13. GSV Capital (2014). Market commentary. Retrieved from http://gsvcap.com/market/commentary/san-francisco-37/ on 9.7.2015.

14. Hathi Trust Digital Library. (2015). Partnership Community. Retrieved from https://www.hathitrust.org/community on 13.7.2015.

15. Laney, Doug. (2001). 3D data management: controlling data volume, variety and velocity. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf on 22.07.2015.

16. Rijmenam, Mark van. (2013). Why the 3 Vs are not sufficient to describe big data. Retrieved from https://datafloq.com/read/3vs-sufficient-describe-big-data/166 on 22.07.2015.

17. Soares, Louis. (2011). The 'Personalization' of Higher Education: Using Technology to Enhance the College Experience. Retrieved from https://www.americanprogress.org/issues/labor/report/2011/10/04/10484/thepersonalization-of-higher-education/ on 10.7.2015.