Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Information Professionals and Big Data

Ajay Shanker Mishra
Library Information Assistant
IISER Bhopal, Madhya Pradesh, India

Abstract: - The main objective of this article to review the concept of Big Data. Firstly, a definition the term and the important characteristics of Big Data are given. Secondly, describe the management of big data. Next, the article explore relevancy of the term "Big Data" with Information Science and in general context.

Key Words: Big Data, Characteristics, Library Information Services, OPAC, RDM.

I. INTRODUCTION

The 21st century known as information age. Data or information play a vital role in socio-economic development of individual, society and nation. Nobody can survive without appropriate or relevant information. Data is important part of research and development programmes and resolve social, economic problems of nations. Many research and development progrmaames in research laboratory, institution and university form a new data.

Over the past 20 years, data has increased in a large scale in various fields. According to a report from International Data Corporation (IDC), in 2011, the overall created and copied data volume in the world was 1.8ZB (1.6 trillion gigabytes), which has increased by nearly nine times within 5 years [1]. Such figure will double at least every other 2 years in the near future.

Wurman and Bradford (1996) They assert that: There is a tsunami of data that is crashing onto the beaches of the civilized world. This is a tidal wave of unrelated, growing data formed in bits and bytes, coming in an unorganized, controlled, incoherent cacophony of foam. It is filled with the floatsam and jetsam. It is filled with the sticks and bones and shells of inanimate and animate life. The tsunami is a well of data – data produced at greater and greater speed, greater and greater amounts to store in memory, on tape, on disks, on paper, sent by streams of light faster, more and more and more.[2]

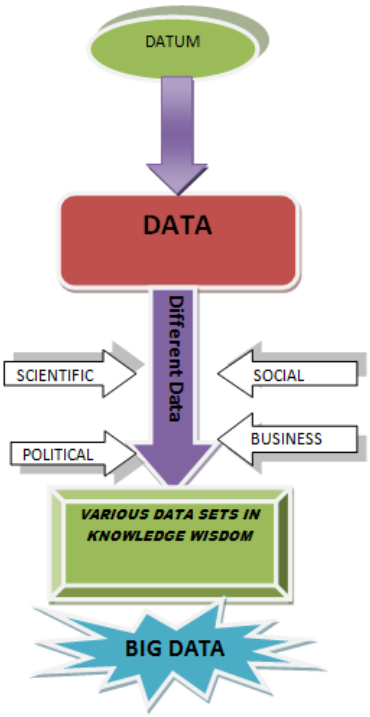

Fig.1 How Datum become Big Data?

As a result, the vast amount of data and information available for organizations for analysis is exploding [3]. This provides organizations with completely new operating possibilities, while simultaneously generating numerous new challenges. In this context the term "Big Data" has emerged and is being used more and more commonly in the knowledge wisdom.

The term of big data was coined under the explosive increase of universe data and was mainly used to describe these enormous datasets. In this paper, we introduce the concept of big data, and try to define the Big Data. In particular, paper describes its 5Vs characteristics, including Volume, Variety, Velocity, and Value. The Paper explores how the datum becomes Big Data.

II. CONCEPT OF BIG DATA

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 reduction and reduced risk.[8]

In short 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.[4] 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. [5]

III. DEFINITIONS OF BIG DATA

Big Data is also called as "Very large data, Extreme data, and Total data, etc." and the first criterion was the volume of data. Although there is no exact definition of Big Data, it sometimes refers to more data than ZB (Zeta Byte) range and also means data which require distributed parallel processing technology for the analysis of large volume of data such as Hadoop. 1 ZB is a huge amount of data which corresponds to one trillion Giga bytes [6].

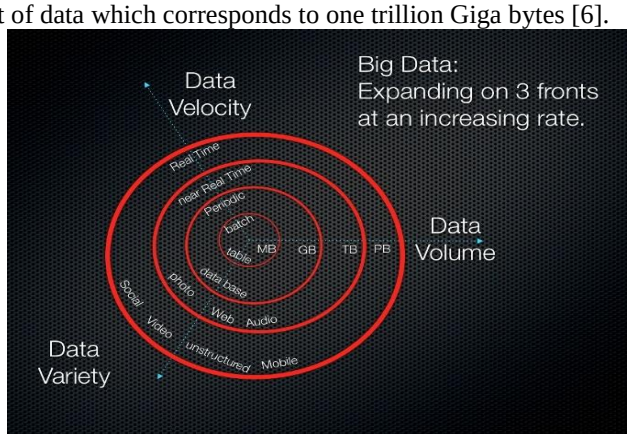

Fig.2 Various datasets and Big Data

IT persons, Institute and Information professionals define the Big Data in different ways and various definitions obtain from different literary source. Some of those are specified below:

Big Data is the amount of data beyond the ability of technology to store, manage and process efficiently (Manyika et.al, 2011).

Big Data is a term which defines the hi-tech, high speed, high-volume, complex and multivariate data to capture, store, distribute, manage and analyze the information (TechAmerica Foundation, 2014).

Big data is 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, 2014; Gürsakal, 2014).

Big Data Technologies are new generation technologies and architectures which were designed to extract value from multivariate high volume data sets efficiently by providing high speed capturing, discovering and analyzing (Gantz and Reinsel, 2011).

Hashem et. al. define Big Data by combining various definitions in literature as follows:

The cluster of methods and technologies in which new forms are integrated to unfold hidden values in diverse, complex and high volume data sets (Hashem et.al., 2015).

As the definitions suggest, there are some points to take into consideration in big data sets. The data should be complex and multiple together with its size. Therefore conventional methods have difficulty in analyzing big data sets and new methods and technologies are needed. [7]

IV. CHARACTERISTICS OF BIG DATA

Big data can be described by the following characteristics: [8] [9]

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 is the category of big data, and an essential fact that data analysts must know. 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.

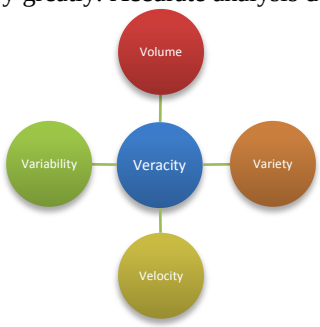

Fig.3 Characteristics of Big Data

Factory work and Cyber physical systems may have a 6C system:

Connection (sensor and networks)
Cloud (computing and data on demand)
Cyber (model and memory)
Content/context (meaning and correlation)
Community (sharing and collaboration)
Customization (personalization and value)
(https://en.wikipedia.org/w/index.php?title=Big_data&oldid=676703166)

V. CLASSIFICATION OF BIG DATA

The characteristic of big data can be understood better by dividing it into classes. These classes are Data Sources, Content Format, Data Stores, Data Staging and Data Processing (Hashem et.al, 2015).

Data Sources: Web & Social, Machine, Sensing, Transactions and IoT
Content Format: Structured, Semi-Structured and Unstructured
Data Stores: Document-oriented, Column-oriented, Graph based and Key-value
Data Staging: Cleaning, Normalization and Transform
Data Processing: Batch and Real time

VI. HOW TO PROCESS BIG DATA?

Basically, data processing is seen as the gathering, processing, management of data for producing "new" information for end users [10].In present time there are many key issues to process vast amount of data. Various literary source provide various methods to manage or process Big Data.

Karmasphere5 currently splits Big Data analysis into four steps: Acquisition or Access, Assembly or Organization, Analyze and Action or Decision. Thus, these steps are mentioned as the **"4 A's"**. The Computing Community Consortium [11] similarly to [10], divides the organization step into an Extraction/Cleaning step and an Integration step.

The paper reviews various literary sources and tries to explore Big Data process. The step wise Big Data process is illustrated below:-

First Step: Big Data Generation:- Data generation is the first step of big data. Specifically, it is large-scale, highly diverse, and complex datasets generated through longitudinal and distributed data sources. Such data sources include sensors, videos, click streams, and/or all other available data sources. At present, main sources of big data are the operation and trading information in enterprises, logistic and sensing information in the IoT, human interaction information and position information in the Internet world, and data generated in scientific research, etc. The information far surpasses the capacities of IT architectures and infrastructures of existing enterprises, while its real-time requirement also greatly stresses the existing computing capacity.[12]

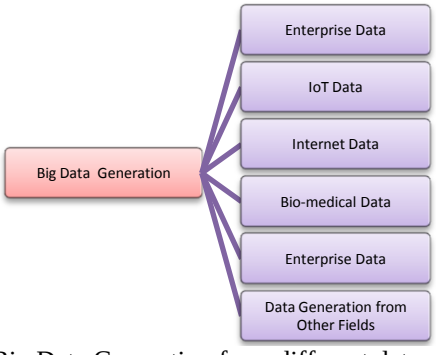

Fig.4 Big Data Generation from different data sources

Second Step: Acquisition of Big Data:- As the second phase of the big data system, big data acquisition includes data collection, data transmission, and data pre-processing. During big data acquisition, once the raw data is collected, an efficient transmission mechanism should be used to send it to a proper storage management system to support different analytical applications.

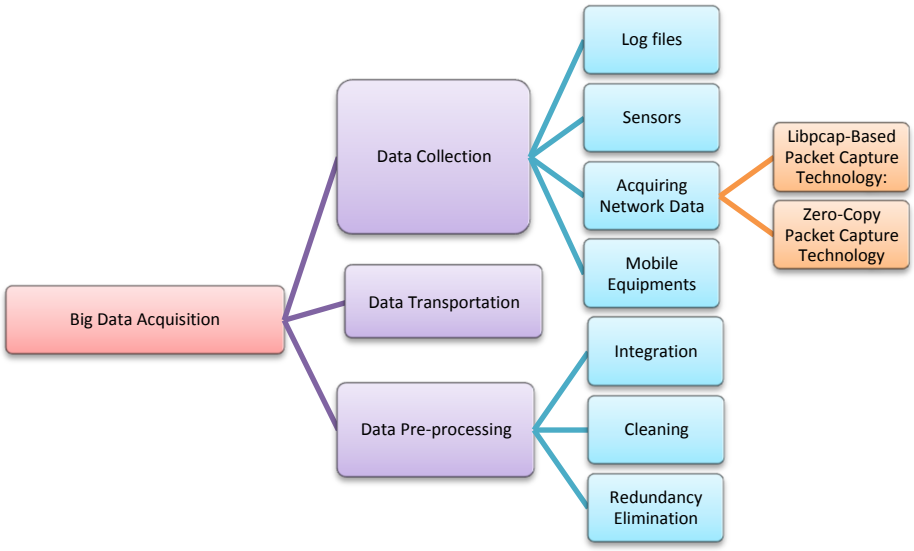

Fig.5 Big Data Acquisition Process

VII. SOME EXAMPLES OF BIG DATA IN GENERAL

Various literary sources review the example or application area of Big Data. Big data is used efficiently in numerous fields. Some of them are listed below:

(A) In Government and Administration:-

(A) In 2012, the Obama administration announced the Big Data Research and Development Initiative, to explore how big data could be used to address important problems faced by the government. [13]

(B) Big data analysis was, in parts, responsible for the BJP and its allies to win a highly successful Indian General Election 2014. [14]

(B) Retail:-

Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.[15]

(C) Retail banking:-

FICO Card Detection System protects accounts worldwide.

(D) Real estate: -

Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day. [16]

(E) Medical, Health Science and Allied science:-

*Decoding the human genome originally took 10 years to process, now it can be achieved in less than a day. the DNA sequencers have divided the sequencing cost by 10,000 in the last ten years, which is 100 times cheaper than the reduction in cost predicted by Moore's Law.[17]

*The (http://www.nasa.gov/centers/goddard/news/releases/2010/10051. html)NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.[18]

VIII. INFORMATION PROFESSIONAL AND BIG DATA

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including library information science. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

Amy Affelt, Director of Database Research Worldwide at Compass Lexicon and author of The Accidental Data Scientist, a newly-published book by Information Today, discussed some characteristics of Big Data and its possibilities for information professionals. She noted that big data is different from other data; here are some sources of it. Verification of big data and determining its value are opportunities for information professionals. Although the
data is big, the insights gained from it are even bigger. [20]

Libraries are part of a larger information and learning, knowledge sharing, ecosystem whether they operate within a community, campus or organization.[21] In this, Data explosion age Library information professional play role as a "Accidental Data Scientist".[22] Librarians have always been played a main role in information management and organization and dissemination . In the context of eResearch, information specialists such as librarians, archivists, curators and records managers may play key roles. Data librarians are professional library staff engaged in managing research data, using research data as a resource, or supporting researchers in these activities.

Clearly, data services are a hot area in academic libraries. But how is this trend playing out in libraries at teaching focused institutions. As I will illustrate below, there are rich opportunities to expand library reference and instruction services to support quantitative reasoning initiatives and data intensive undergraduate research. Data curation and management services, a major interest at research libraries, are also an emerging opportunity at liberal arts institutions as are the collection and management of field research data.

What role can information professionals play to help support and use of big data within your organization? Information professionals are surprised they already user Big Data concept.

Here we are explore some examples of Big Data application which has already been libraries and information centers are used:-

8.1-Online Databases:

IDC predicts that big data is growing at an annual rate of 60% for structured and unstructured data. Businesses need to do something with all that data, and traditionally databases have been the answer. With cloud technology, providers are rolling out more ways to host those databases in the public cloud, freeing users from dedicating their own dedicated hardware to these databases, while providing the ability to scale the databases into large capacities.[23]

Information professionals might be surprised to know they already used Big Data concept at own library. Generally we are providing access to online databases that allow users to view articles from magazines, journals, or proceedings also providing a collection of print books or e-books that users can borrow all these services use SAS (Statistical Analysis System).

8.2- Online public Access Catalogue:

Libraries have amassed an enormous amount of machine-readable data about library collections, both physical and electronic, over the last 50 years. However, this data is currently in proprietary formats understood only by the library community and is not easily reusable with other data stores or across the Web.

An Online Public Access Catalog (often abbreviated as OPAC or simply Library Catalog) is an online database of materials held by a library or group of libraries. It is a computerized library catalog available to the public. Most OPACs are accessible over the Internet to users all over the world. Users search a library catalog principally to locate books and other material physically located at a library.AWS service integrate public data sets or link and serve to his/her user community.

Public Data Sets on AWS provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications. AWS is hosting the public data sets at no charge for the community, and like all AWS services, users pay only for the compute and storage they use for their own applications. Learn more about Public Data Sets on AWS and visit the Public Data Sets forum.

Scientists, developers, and many other technologists from many different industries are taking advantage of Amazon Web Services to perform big data analytics and meeting the challenges of the increasing volume, variety, and velocity of digital information. Amazon Web Services offers a comprehensive, end-to-end portfolio of cloud computing services to help you manage big data by reducing costs, scaling to meet demand, and increasing the speed of innovation. Examples:

1-The OCLC data hub is working to become a relevant and useful source in the web of big data and cloud computing

2-Worlcat

8.3- Statistical Analysis of Usage Data:

We also use SAS software to assist our marketing efforts by identifying employees who are frequent library users, then targeting our outreach efforts to those users. By manipulating this data, we can drive usage of our resources and services.[24]

Statistical Analysis System (SAS):It's the science of collecting, exploring and presenting large amounts of data to discover underlying patterns and trends. Statistics are applied every day – in research, industry and government – to become more scientific about decisions that need to be made.[25] libraries have a variety of output measures which are largely dependent on the types of services they provide, as well as the ability of the library organization to gather these statistics.. For example Library

Resource Usage Electronic resource usage (COUNTER and related statistics);

Print usage (checkouts, renewals, in-house use counts, etc.);

Services Exhibits;

Gate counts;

Library instruction (sessions, number of students, hours);

Public events (attendance and number of events);

Reference requests/research appointments.

8.4-Research Data Management in Library and Big Data:

The management of research data is now a major challenge for research organizations. Vast quantities of born-digital data are being produced in a wide variety of forms at a rapid rate in universities and research centers. creating the so-called ‘‘volume’’,‘‘variety’’ and ‘‘velocity’’ challenges of data [26][27].

DMPonline is a web-based tool to help researchers and research support staff to produce data management and sharing plans that meet funders’ mandates. It was first demonstrated at the Jisc conference in London in 2010 and has since undergone regular updates and some major redevelopment. The tool has received international recognition, and was shortlisted along with the work of US colleagues on DMPTool for the DPC's Digital Preservation Awards at the end of 2012.[28]

“The Cambridge Big Data Strategic Research Initiative brings together researchers from across the University to address challenges presented by our access to unprecedented volumes of data. Our research spans all six Schools of the University, from the underlying fundamentals in mathematics and computer science, to applications ranging from astronomy and bioinformatics, to medicine, social science and the humanities.”

In parallel, our research addresses important issues around law, ethics and economics, in order to apply Big Data to solve challenging problems for society.

Cambridge Big Data supports collaboration and knowledge transfer in this growing field.”[29]

8.5 Social Media

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. Thanks to digital technologies, more and more human activities leave imprints whose collection, storage and aggregation can be readily automated. In particular, the use of social media results in the creation of datasets which may be obtained from platform providers or collected independently with relatively little effort as compared with traditional sociological methods.[34]

Online social media systems have created new ways for individuals to communicate, share information and interact with a wide audience (Aharony, 2012; Lin & Ranjit, 2012; Rees & Hopkins, 2009). For organizations, social media provide new avenues for communication and collaboration with their stakeholders. However, any value created for an organization through social media comes not from any particular platforms, but from how they are used (Busch, 2011; Culnan, McHugh, & Zubillaga, 2010; Dickson & Holley, 2010). While social media may be widely used by individuals and many organizations, their use in higher education generally is still relatively new (Busch, 2011;

Forkosh-Baruch & Hershkovitz, 2012). Now in these days every library has created own account on face book, twitter, link din etc. The potential value of social media for academic libraries was recognized comparatively early on, with the term‘Library 2.0’ (referring to the application of web 2.0 online

tools to library functions) being coined by Casey in 2005 (Harinarayana & Raju, 2010; Mahmood & Richardson, 2011; Nguyen, Partridge, & Edwards, 2012). Following a period of piloting (Rees & Hopkins, 2009), applications of socialmedia systems in library settings are now commonly reported (Chen, Chu, & Xu, 2012; Glazer, 2012; Mahmood & Richardson, 2011; Thornton, 2012). The growing ubiquitous presence and use of social media means that many libraries are using social media to engage their stakeholders in the online environment (Burkhardt, 2010; Collins & Quan-Haase, 2012).

IX. CONCLUSION

The emergence of big data opens number of opportunities to information professional and library information sector. In the IT era, the **“T”** (Technology) was the main concern, while technology derives the development of data. In the big data era, with the prominence of data value and the advances in

I “(Information), data will drive the progress of technologies in the future. Big data will not only change the social and economic life, but also influence everyone’s ways of living and thinking, which is just beginning. Library Information professional play with Big Data concept in our routine work in every and each sector, library professional serve to user with online database, OPAC, Blog like Facebook, Linnkdin, and Twitter.

" Librarians and information professionals have always worked with data in order to meet the information needs of their users, thus "Big Data" is not a new concept for them though it is spawning new approaches along with a language all its own.”

REFERENCES

[1] John Gantz and David Reinsel. Extracting value from chaos. IDC iView, pages 1–12, 2011.

[2] Ifijeh, Goodluck I. (2010). Information explosion and university libraries: Current trends and strategies for intervention. Chinese Librarianship: an International Electronic Journal, 30. URL: http://www.iclc.us/cliej/cl30doraswamy.pdf

[3] E. Brynjolfsson, A. McAfee, (2012), “Big data: The managementrevolution”, Harvard Business Review, pp. 60-68, October 2012.

[4] Snijders, C.; Matzat, U.; Reips, U.D.(2012). " 'Big Data': Big gaps of knowledge in the field of Internet" (http://www.ijis.net/ijis7_1/ijis7_1_editorial.html). International Journal of Internet Science 7: 1–5.

[5] Ibrahim; Targio Hashem, Abaker; Yaqoob, Ibrar; Badrul Anuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "big data" on cloud computing: Review and open research issues". Information Systems 47: 98-115.doi:10.1016/j.is.2014.07.006(https://dx.doi.org/10.1016%2Fj.is.2014.07.006).

[6] Korean President's Council on National ICT Strategies, The national development strategy in Big Data age, pp.49- 58, August 2012
[7] Özköse, Hakan, Ari,Emin Sertac and Gencer, Cevriye. (2015). Yesterday, Today and Tomorrow of Big Data. Procedia - Social and Behavioral Sciences 195 (2015) 1042 – 1050
[8] Hilbert, M. Big Data for Development: A Review of Promises and Challenges. Development Policy Review. accessible at martinhilbert.net/big data for development
[9] Hilbert, M. (2015). Digital Technology and Social Change [Open Online Course at the University of California](freely available). https://www.youtube.com/watch?v=XRVIh1h47sA&index=51&list=PLtjBSCvWCU3rNm46D3R85efM0hrzjuAIg Retrieved fromhttps://canvas.instructure.com/courses/949415
[10] http://www.gartner.com/newsroom/id/1731916.
[11] H.V. Jagadish, D. Agrawal, P. Bernstein, E. e. a. Bertino,Challenges and Opportunities with Big Data, The CommunityResearch Association, 2015.
[12] M. Chen et al(2014)., Big Data: Related Technologies, Challenges and Future Prospects, Springer Briefs in Computer Science, DOI 10.1007/978-3-319-06245-7__3.
[13] Kalil, Tom. "Big Data is a BigDeal" (http://www.whitehouse.gov/blog/2012/03/29/bigdatabigdeal). White House. Retrieved 26 September 2012.
[14] "News: Live Mint" (http://www.livemint.com/Industry/bUQo8xQ3gStSAy5II9lxoK/AreIndiancompaniesmakingenoughsenseofBigData.html). Are Indian companies making enough sense of Big Data?. Live Mint http://www.livemint.com/. 20140623.Retrieved 20141122.
[15] "Data, data everywhere" (http://www.economist.com/node/15557443). The Economist. 25 February 2010. Retrieved 9 December 2012.
[16] Wingfield, Nick (20130312)."Predicting Commutes More Accurately for Would Be Home Buyers NYTimes.com"http://bits.blogs.nytimes.com/2013/03/12/predictingcommutesmoreaccuratelyforwouldbehomebuyers/).Bits.blogs.nytimes.com. Retrieved 20130721.
[17] Delort P., OECD ICCP Technology Foresight Forum, 2012. http://www.oecd.org/sti/ieconomy/Session_3_Delort.pdf#page=6
[18] Webster, Phil. "Supercomputing the Climate: NASA's Big Data Mission" (http://www.csc.com/cscworld/publications/81769/81773supercomputing_the_climate_nasa_s_big_data_mission). CSC World. Computer Sciences Corporation.
[19] George, Gerard and Haas, Martine R., BIG DATA AND MANAGEMENT, Academy of Management Journal 2014, Vol. 57, No. 2, 321–326. http://dx.doi.org/10.5465/amj.2014.4002
[20] http://www.libconf.com/2015/04/28/data-scientist-a-new-role-for-librarians/
[21] http://internet-librarian.infotoday.com/2015/
[22] http://books.infotoday.com/books/Accidental-Data-Scientist.shtml
[23] http://www.networkworld.com/article/2162274/cloud-computing/10-of-the-most-useful-cloud-databases.html
[24] https://www.sla.org/IO/2014/MayJune2014.pdf
[25] http://www.sas.com/en_us/insights/analytics/statistical-analysis.html
[26] McAfee A, Brynjolfsson E (2012) Big data: The management revolution. Harv Bus Rev 90: 60–68.
[27] Laney D (2001) 3D data management: Controlling data volume, velocity and variety. Stamford, CT:META Group. Available: http://blogs.gartner.com/douglaney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
[28] http://library.ifla.org/958/1/119-davidson-en.pdf
[29] http://www.bigdata.cam.ac.uk/resources/research-data-management
[30] https://en.wikipedia.org/w/index.php?title=Big_data&oldid=676703166
[31] http://www.ands.org.au/guides/dmframework/dmskills-information.html
[32] https://aws.amazon.com/big-data/
[33] http://www.socialsamosa.com/wp-content/uploads/2014/11/11.jpg
[34] Tufekci, Zeynep. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other MethodologicalPitfalls. In ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014.
[35] Palmer, Stuart (2014).Characterizing University Library Use of Social Media: A Case Study of Twitter and Facebook from Australia: The Journal of Academic Librarianship 40 (2014) 611–619

ABOUT THE AUTHOR

Mr. Ajay Shanker Mishra currently working as Library Information Assistant at Indian Institute of Science Education and Research Bhopal .Prior to joining IISER Bhopal he was Librarian at Hitkarini Dental College and Hospital, Jabalpur and also served as a Guest Lecturer in Library Information Science Department Rani Durgavasti Vishvavidylya , Jabalpur. He holds a M.Phil. in Library Information Science from the APS University Rewa and also qualified UGC-NET Examination.

