JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Big Data and Library Science: An Overview

Rajesh Achra

(MCA, MLIS, NET)

Junior Judicial Assistant

Rajasthan High Court, Jodhpur (Rajasthan)

Abstract: Generally speaking, big data refers to data sets that are too large or complex to be processed with traditional application software. Larger data sets offer more statistical power, while complex data may lead to more false discoveries. A number of valuable questions can be answered by Big Data about patterns, trends, and associations in user behavior. Large data sets can improve the quality of library services by helping libraries clearly understand changing user needs, and accordingly reshaping & restructuring their services & procedures. The digital library data resources can be used as big data by applying innovative techniques and introducing important digital changes. Big data provides insights into resource utilization, decision-making, and library user needs. This paper reviews the concept of the Big data in the field of library Science.

Keywords: Big Data, Library Science, Information Science.

I. INTRODUCTION

Since the advent of new technologies, devices, and communication means such as social networking sites, the amount of data produced by mankind is growing rapidly. Up until 2003, 5 billion gigabytes of data were produced by mankind. This amount of data can fill an entire football field if it is piled up on disks. In 2011, it was created every two days, and in 2013, it was created every ten minutes. This rate continues to grow exponentially. A big data collection is a set of large data sets that traditional computing techniques cannot handle. Instead of being a single technique or tool, it has evolved into a comprehensive subject that involves a variety of tools, techniques, and frameworks. In addition to providing more accurate analysis, big data technologies can help businesses make more concrete decisions, resulting in greater operational efficiencies, lower costs, and reduced business risks. Big data requires an infrastructure capable of managing and processing huge volumes of structured and unstructured data in real time and protecting privacy and security. [1]

The types of Big Data are as follows:

- **A structured approach:** In structured data, any data that can be stored, accessed, and processed in a fixed format is referred to as structured. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kinds of data (where the format is well known in advance) and also obtaining value from them. We are, however, anticipating problems when this data grows to a huge extent, typically exceeding several zettabytes in size. [1]

- **Unstructured:** Any data with an unidentified format or structure is classified as unstructured. Its sheer size makes it difficult to process, and many challenges arise when trying to extract value from it. Examples of unstructured data include heterogeneous sources which contain a mix of text files, images, and videos. Nowadays, organizations possess a great deal of data but they lack the tools to exploit its potential due to it being in its raw or unstructured form. [2]

- **Semi-structured:** In a semi-structured data set, we can observe that the data is structured but not defined with a table definition as in relational databases. An example of semi-structured data is data in an XML file. [2]

The following characteristics describe big data:

- **Volume –** The name Big Data itself implies an enormous size. The size of data plays an important role in determining the value that can be derived from it. As well, whether a particular data is considered Big Data or not depends on its volume. Therefore, volume is one factor that needs to be considered when dealing with Big Data solutions. [3]

- **Variety –** The next aspect of Big Data is its variety. A variety of sources and data types, both structured and unstructured, is what constitutes variety. Earlier, most applications used spreadsheets and databases as their only sources of data. In recent years, data such as emails, photos, videos, monitoring devices, PDFs, audio files, etc., have also been considered in analysis applications. This variety of unstructured data poses certain challenges when it comes to storage, mining, and analyzing. [3]

- **Velocity -** The term 'velocity' refers to how fast data is generated and processed to meet the demands. The flow of data is massive and continuous from sources such as business processes, application logs, networks, social media sites, sensors, and mobile devices.

- **Variability -** This refers to the inconsistency of the data, which can hinder the ability to manage and handle it effectively at times. [4]

The advantages of big data processing:

- **Multiple benefits can be gained from processing Big Data in DBMS, including:** When making decisions, businesses can use outside intelligence, social data from search engines and sites like Facebook and Twitter can help organizations fine-tune their business strategies. [4]

- **Improved customer service:** The traditional customer feedback system is being replaced by new systems that use Big Data and natural language processing to read and evaluate consumer feedback. Identification of any risks associated with the product or service at an early stage. [4]

- **Better operational efficiency:** Before identifying what data should be moved to the data warehouse, big data technologies can be used to create a staging area or landing zone for new data. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data. [5]

II. BIG DATA IN LIBRARY SCIENCE

Libraries are in a unique position to leverage Big Data for their managerial purposes, as well as provide library services that focus on data analytics. Several libraries have already started offering research data services such as data management planning, collection, curation and archiving. This extent of the services vaires depending on the user's need and the current situation of the library. Suggestions have been made regarding services that could be provided by libraries following the life cycle put forth by DataONE. While there are some stages in which they can play a vital role, librarians should carefully consider which services they would offer and how they would resource them. [5]

Planning: In the case of planning an organizational level research project, libraries can provide a variety of services that send researchers on the right path. Librarians can help to refine research strategies by aiding in how data is searched (e.g., using appropriate search terms) for optimal efficiency. Additionally, their database design and development experience are crucial for organization and mining of Big Data (DDS, 2021). Furthermore, librarians are highly capable of helping construct Data Management Plans (DMPs), which researchers benefit from greatly. Whilst working with Big Data projects may have some unique considerations, it doesn't stray too far from ordinary DMPs. [6]

Collection: Libraries can still aid researchers in collecting Big Data, as they may already possess skills related to common Big Data sources, and these can be part of the library collection. Furthermore, working with Big Data calls for a great deal of computing power. Here libraries can partner with other units such as high-performance computing, enabling them to access technical infrastructure needed to process this data. Lastly, libraries can ensure that all data collected is properly organized and documented for later archiving. [6]

Ensure, analyze, and integrate: Once the required data has been collected, Quality assurance (QA) is essential to verify its validity for the purpose of research. Following which, the data is integrated, organised and processed with relevant tools and techniques to make it easier to comprehend and interpret the results. This helps researchers explore solutions and answers to the research problem which initiated their work in the first place. Generally, the responsibility of QA, analysis and integration of information lies with the researcher. Nevertheless, if librarians are supplying data as part of their collection, they might need to perform quality assurance on it and provide suitable documentation that help enable further QA from other researchers. [7]

Provide a description: Understandably, creating the proper documentation and metadata to accompany any data file is a challenge many researchers face. However, this is an area in which libraries are well equipped to help. Services should be offered to assist with the documentation of Big Data and provisioning of appropriate metadata. [7]

Preserve: In addition to data descriptions, libraries can also provide archiving and preservation services. Many libraries have institutional repositories and some have dedicated data repositories.

Discover: Apart from some of the more common Big Data sources, such as social media platforms like Twitter, finding useful data can be challenging. The role of librarians in research data support is to assist researchers in finding relevant and useful datasets. In addition to locating the datasets, librarians as informaticists should also be familiar with techniques for efficiently acquiring, storing, and processing the data. [8]

III. PRACTICAL APPROACH OF BIG DATA IN LIBRARIES

Twitter Archive at the Library of Congress: The partnership between Twitter, the popular short message service, and the Library of Congress, the largest library in the world, made waves when first announced in 2010. The goal was to archive and store every tweet ever posted. However, it turned out to be more difficult than anticipated as daily tweets had increased from 55 million to over 500 million in a few short years. Researchers have been eagerly awaiting access to this wealth of data ever since but there is currently no indication if or when this will be made possible by the Library of Congress. Nevertheless, it is still remarkable to witness how an institution that has stood for over 200 years can cooperate with a mere four year-old start-up thanks to big data. [9]

A metadatabase for Australian geophysical data is being created: 2013 Business Information Survey reveals that information professionals have not been strongly involved in the popularly talked-about topic of big data. Only a few of those polled indicated that they worked on such projects, with one particular example being an Australian effort to build a metadatabase. It called for librarians to effectively manage vast and complicated geophysical data, consisting of petabytes in multiple formats, plus raw and processed data requiring various licensing conditions. During this project, information experts collaborated with geophysicists, IT pros and database developers while performing tasks such as creating and consulting on necessary metadata fields, generating controlled vocabularies and search parameters, exploring additional functionalities, testing the database with feedback, importing relevant metadata and educating users before launching the new system. This case study underlines the major impact that librarians can make in regards to big data applications through providing skillful expertise related to metadata. [9]

(Harvard University Library) Big Data Applications for Books: Harvard University Library, the largest university library system across the world, released its metadata on 12 million items, including books, videos, audio recordings, images, manuscripts and maps to name a few in April 2012. While copyright restrictions do not allow these materials to be available online in full text version, their metadata is an incredibly valuable asset which can be seen as “big data for books” according to Co-Director of Harvard Library Lab David Weinberger. It was already done by the University of Michigan in November 2010 [16]. Based on the amount of data, however, neither of these cases can truly qualify as big data. For instance, the amount of data published by Harvard Library totals around 4 GB. Many data analysis projects in libraries bandy around the term "big data" even though the amount of data is negligible. This has prompted fresh criticism of the term. [10]

The Jisc and HESA Library Data Labs project: JISC and HESA have come together to create Heidi Plus, a web-based platform which provides non-profit organisations and universities in Britain with an effective system of data analysis. Released on 30 November 2015, the programme was built to grant decision makers convenience and speed when receiving information whilst also reducing time and financial costs. JISC and HESA have now launched the Library Data Labs Project, which utilises the Heidi Plus platform to enable libraries to review massive amounts of data. Unlike its prior uses, this project has a specific library-focused objective: over a three-month period, five inter-institutional teams comprising 23 different university libraries as well as a JISC team started visualizing data in response to questions put forth by the respective libraries. Sources of information they assessed included SCONUL, Ulrich, Dewey, Altmetrics, H-Index, IMD and more. [11]

IV. CONCLUSION

Academic libraries have a large selection of primary and secondary data with academic content. To maximize the benefit they offer customers, this content can be augmented with freely available online data. For example, such data may be used to catalogue library holdings and analyse customer behaviour as well as their media usage. Thus, knowledge of Big Data applications is essential in order to create value that otherwise could not be achieved.

REFERENCES

1. M. Humbel (March 2017), "Die Umsetzung von Open Data an Wissenschaftlichen Bibliotheken der Schweiz. Eine qualitative Untersuchung," in Churer Schriften zur Informationswissenschaft, Schrift 86, (available online: https://www.htwchur.ch/fileadmin/htw_chur/angewandte_zuk unftstechnologien/SII/churer_schriften/CSI86-Umsetzung_Open_Data_an_wissenschaftlichen_Bibliotheken _der_Schweiz.pdf site visited on 03.10.2018).

2. M. Humbel (March 2017), "Die Umsetzung von Open Data an Wissenschaftlichen Bibliotheken der Schweiz. Eine qualitative Untersuchung (Interview mit M. Ehrismann, M. Hotea and R. D. Wanger)," in Churer Schriften zur Informationswissenschaft, Schrift 86, (available online: https://www.htwchur.ch/fileadmin/htw_chur/angewandte_zuk unftstechnologien/SII/churer_schriften/CSI86-Umsetzung_Open_Data_an_wissenschaftlichen_Bibliotheken _der_Schweiz.pdf site visited on 03.10.2018), 74ff. and 135.

3. T. Aaron (November 2016), "5 reasons why library analytics is on the rise (Blogbeitrag)," in Musings about librarianship, (online available: https://musingsaboutlibrarianship.blogspot.com/2016/11/5-reasons-why-library-analytics-is-on.html site visited on 06.11.2018).

4. A. McGill (August 2016), "Can Twitter Fit Inside the Library of Congress?" in The Atlantic, (available online: https://www.theatlantic.com/technology/archive/2016/08/can-twitter-fit-inside-the-library-of-congress/494339/ site visited on 06.11.2018).

5. A. Foster (April 2013), "Add value or die. The fate of corporate information services. The Business Information Survey 2013" in BIR, (available online: http://journals.sagepub.com/doi/full/10.1177/02663821134842 22 site visited on 06.11.2018).

6. Alia (2016), “Website of the National 2016 Conference, 29 August-2 September 2016 Adelaide: Engage Create Lead” (available online: https://library.alia.org.au/big-data-small-library-0 site visited on 25.10.2018)

7. V. Johnson (2016), "Big Data, Small Library," Deakin (available online: https://library.alia.org.au/file/424/download?token=rx_SWTG Z site visited on 06.11.2018).

8. . Humbel (March 2017), "Die Umsetzung von Open Data an Wissenschaftlichen Bibliotheken der Schweiz. Eine qualitative Untersuchung," in Churer Schriften zur Informationswissenschaft, Schrift 86, (available online: https://www.htwchur.ch/fileadmin/htw_chur/angewandte_zuk unftstechnologien/SII/churer_schriften/CSI86-Umsetzung_Open_Data_an_wissenschaftlichen_Bibliotheken _der_Schweiz.pdf site visited on 03.10.2018).

9. Tableau Software (2018): "Tableau Website", (available online: https://www.tableau.com/ site visited on 06.11.2018).

10. C. O'Maley Voliva (April 2015), "Data Visualization for Public Libraries," in Public Libraries Online, (available online: http://publiclibrariesonline.org/2015/04/data-visualization-for-public-libraries/ site visited on 06.11.2018).

11. Tableau (2018), "Brooklyn Public Library saves time, money and headcount with Tableau", (available online: https://www.tableau.com/solutions/customer/brooklyn-public-library-saves-time-money-and-headcount-tableau site visited on 06.11.2018).