Big Data: Impact on Library and Information Service

Riya Datta (Sengupta)*, Dr. Biswajit Das@

* riyadtt30@gmail.com, MLIS, Jadavpur University

@biswajit@myself.com ,University Librarian University of Gour Banga

Abstract

Libraries play an important role at the intersections of government, universities, research institutes, and the public since they are storing and managing digital assets. The large amount of data and those data in library need to be transformed into information or knowledge which then be used by researchers or users. Librarians might need to understand how to transform, analyze, and present data in order to facilitate knowledge creation. For example, they should know how to make big datasets more useful, visible and accessible. With new and powerful analytics of big data, such as information visualization tools, researchers/users can look at data in new ways and mine it for information they intend to have. This paper discussed the characteristics of datasets in library, the research work on library big data and then summarized the applications in this field.

Keywords: Big data, Big Data Architecture, Hadoop cluster, Big Data and Libraries, Classification and Cataloguing of Big Data, Metadata

Introduction

Libraries collect a large amount of data, such as books, research articles and reports, both in physical and electronic formats. The collection was originally for researchers or public users to find necessary information they need. However, this data becomes so large and the format is so various which might affect the efficient use. Although a lot of library data has been digitalized, most of them have not been used for data mining or big data technology. On the other hand, although some work has been done in the past on how to maintain those library collections in order to efficiently and effectively use, there is no much research on using meta data to organize digital assets so that the big data and cloud computing technology could be used. The three

common characteristics of big data include high-volume, velocity and variety of information. Although a few researches have linked library data into big data, some researchers raise questions about that since there are no clear characteristics of velocity. In addition, based on the current terminology, it seems that the database management systems is enough for storing and processing library data, thus does not require big data technology such as distributed systems for analysis or processing. Therefore, it might worth to clarify this doubt. Moreover, it seems that there is no much general review works on the research for library "big data".

What is Big Data?

Big data represents the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value. 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 philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data. Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale.Big data can be described by the following characteristics:

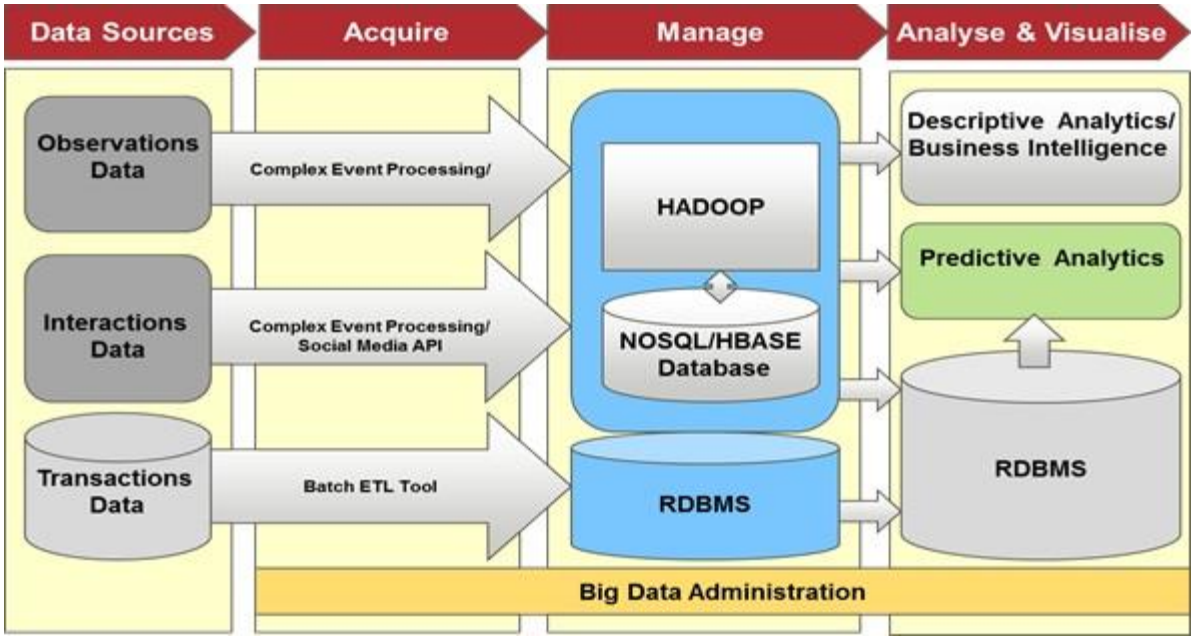
- **Volume :** The quantity of generated and stored data. The size of the data determines the value and potential insight and whether it can be considered big data or not.
- **Variety:** The type and nature of the data. This helps people who analyze it to effectively use the resulting insight. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.
- **Velocity:** In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time.
- **Variability:** Inconsistency of the data set can hamper processes to handle and manage it.
- **Veracity:** The data quality of captured data can vary greatly, affecting the accurate analysis.

Big Data Architecture

Reasoning with all data is possible only if the three functional aspects of data that are interconnected. These aspects evolved data from raw to a product by undergoing R&D. These aspects are mapped to a platform:

- **Data lake or data platform:** Raw data lands here and is stored. This platform has the capacity and cost profile to capture large volumes of data from varied sources without any loss or time lag. This is typically a Hadoop cluster.
- **Data product :** BI, analytical dashboards and relational forms of predictive analytics happen here. This platform supports high reuse and high concurrency of data access. It supports mixed workloads of varying complexity with large number of users. This is typically a parallel EDW platform.

- **Data R&D :** Deep analytics from big data happens here. This platform determines new patterns and new insights from varied sets of data through iterative data mining and application of multiple other big data analytic techniques.



Big Data and Libraries

Libraries and librarians are uniquely suited to working with big data. Libraries have a long tradition of being early technology adopters, and big data is no exception. There are several key ways for librarians to get involved. One of these is through collection development and preservation of data sets. As more users become interested in working with big data, they will need guidance and material to work with. Librarians are well-positioned to help users understand how and where to find these data sets and to preserve them for future users. Another way librarians can get involved with big data is by working within their institutions to help with research data management. Researchers need assistance with data management, especially because many funding agencies now have very specific data retention guidelines that must be followed. Librarians can help researchers, even in the planning phases of their projects, to appraise and think about the archival and preservation options for their data, as well as the potential for sharing their data.

What Librarians Actually Need to Know About Big Data

Because of its prevalence and potential impacts, librarians need to know the basics of big data and how it affects academic research. Corporate librarians need to know how companies leverage big data, how such data mining provides a competitive advantage, and how students might need to grapple with big data sets in future employment. Science librarians need to know

how big data differs from other scientific data and the impact of emerging software and hardware used for its analysis. Humanities and Social Science librarians should know that big data is becoming more commonplace in their disciplines as well, and is no longer restricted to corpus linguistics. Librarians in all disciplines, in order to facilitate the research process, will need to be aware of how big data is used and where it can be found.

Steps in Big Data analysis :

Cataloguing Big Data

The ability to find the right data is a key challenge when collecting data from a big data repository. While most suggested solutions to solve the big data problem revolve around leveraging text analytics tools, metadata-based content management platforms provide a competitive alternative to get this data under control. The catalogue contains information about all the participants in the data space and the relationships among them. In the traditional database management systems the catalogue is commonly known by the term data dictionary and system catalogue in distributed databases. This task is very easy for small data. But, in the case of big data consisting of enormous sized multiple heterogeneous datasets, maintaining a common schema is not possible.

It is of utmost importance that the dataset be catalogued and the same is distributed to the intermediary channels for making them accessible to the interested data users. Entrusted with the task of preparing a system catalogue for the big data, there are many prevalent options that can be adopted. The two common methods used for cataloguing big data are crowd sourcing and automated metadata discovery. Another option is to follow manual classification and cataloguing by the experts.

- **Crowd sourced Metadata:** Crowd sourcing is the method of opening the responsibility of metadata creation to the feeling of the data users who tag the dataset or data items in the datasets with their opinion or use of it. Crowd sourced metadata are generally perceived to match with the user search criteria.
- **Automated Metadata Discovery:** Automated metadata discover is the process of using automated metadata tools to discover the semantics of a data element in datasets. The matching used for automated metadata generation could be lexical matching, semantic matching or statistical matching. This type of metadata generation is becoming quite efficient and largely used in the statistical datasets of big data.

Classification and Cataloguing of Big Data by Librarians and Data Scientists

The above two methods of metadata generation suffers from inaccuracies leading to false results in big data analytics. Therefore, there is a need for expert human intervention in this process. Librarians are specialists in information management and organization. The data curation component of the big data problem involves information management and organization roles. Librarians must take a leading role in working with big data to avoid a situation where this emerging specialty becomes the servant only of proprietary interests. Librarians also need to embrace a role in making big datasets more useful, visible and accessible by creating taxonomies, designing metadata schemes, and systematizing retrieval methods.

This mechanism for cataloguing the big data is to use the technical expertise of librarians or data scientists to manually assign a subject class to each dataset and catalogue the schema of each dataset belonging to big data. This schema definition is then combined with the system information to obtain the system catalogue. This system catalogue is distributed physically to various cloud hosting services or access locations. Various advantages of this scheme in comparison of crowd sourced metadata and automated metadata discover are the technically correct subject classification of the data set by the expert and non-repetitive nature of subject classification will avoid future efforts.

Conclusion

Big data in library might have less challenge to study, but more challenge to engage with it due to budget and technical issues. There is also absence of big data methods training on most social science curricula. Big data can certainly help libraries make more cost-effective, innovative decisions or recommendations that users wish to have. The research data are increasing very fast, and more and more researchers wish to use collections as a whole, mining and organizing the information in novel ways. Without big data analysis, some patterns might not be easily found. The data collected when library users use the service are very helpful in improving the overall user experience, and satisfaction of library services. The ability to collect and analyze massive amounts of data will be a competitive advantage across all industries, including library. The big data currently might be suitable only for those organizations with large set of data and funding. The traditional DBMS or data analytics might be still a dominant approach. In future, the actual platforms or technologies need to be created in library to utilize this big fecility big data.

References

1. Siwacch, Gautam and Esmailpour, Amir (2014). Encrypted Search & Cluster Formation in Big Data. IN ASEE 2014 Zone I Conference, University of Bridgeport, Bridgpeort.

2. Snijders, C. Matzat, U. and Reips, U.-D. (2012). 'Big Data': Big gaps of knowledge in the field of Internet. International Journal of Internet Science 7: 1–5.

3. Gartner Group (2011). Pattern Based Strategy: Getting Value From Big Data. Available at: http:/ /www.gartner.com/it/page.jsp?id=1731916 (Accessed on 18.03.2018).

4. Cárdenas, Alvaro A. Manadhata, Pratyusa K. and Rajan, Sree (2013). Big Data Analytics for Security Intelligence. Big DataWorking Group Cloud Security Alliance. Available at: https:// cloudsecurityalliance.org/ download/big-dataanalytics-for-security-intelligence/ (Accessed on 23.01.2018).

5. Sugimoto, Cassidy R. Ding, Ying and Thelwall, Mike (2012). Library and information science in the big data era: Funding, projects, and future [a panel proposal] IN Proceedings of the American Society for Information Science and Technology. Volume 49, Issue 1, pages 1–3.