Cataloguing to Facilitate Big Data Analytics

Manish Kumar Singh ***D K Singh***

Abstract

"Big Data" is the popular term used to denote the collection of large data sets possessed by multiple systems. The inherent characteristics of this Big Data are the difficulty in processing due to sheer scale and accessibility of data and also unmanageability through a traditional Database Management System. The size of this data set is ever increasing with increasing pace and addition of multi-exabytes per day. Apart from these, big data normally comprise of heterogeneous dataset, both structured and unstructured and also containing diverse data and file formats. It is very difficult to locate and retrieve the relevant information in real time from the universe of big data. Librarians, coming out of the walled library, can be expected to contribute in this task with their expertise in information organization and management. In this paper various challenges to the big data are identified and to address the challenges mechanisms for creating big data catalogue have been identified. Various mechanisms are discussed and compared and it is proposed to use the technique of library classification and cataloguing to catalogue the datasets in the big data thereby facilitating the information retrieval in the universe of big data.

Keywords: Big Data, Heterogeneous Datasets, Big Data Catalogue, Metadata.

1. Introduction

The term "Big Data" is used to refer to large datasets that are diverse, complex and of a massive scale. Their sizes are normally beyond the capacity of a database management tool to handle within acceptable time limit. Integration of heterogeneous data sources presents formidable logistical as well as analytical challenges. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. It is a very large distributed aggregation of structured, semi-structured and unstructured data that is often incomplete and inaccessible, but has the potential to be mined for information.

Another interesting aspect is that in recent years there has been a movement in the sciences especially to save, curate, and make accessible data from papers for other researchers who may also want to use that same data for other purposes.

The dataset acting as a component of the big data may consist of a statistical database, an unstructured dataset of collection of web pages, a natural language text, a dataset of message posts in social networking website, a live data stream, a collection of unstructured data files, etc. It is important to note that the information sought by the user may be derived from any of the above datasets or a combination of these, e.g. the message post database may contain information relevant for research in gene mutation, the structured statistical database may contain information relevant for research in ancient history, and like.

10th International CALIBER-2015
HP University and IIAS, Shimla, Himachal Pradesh, India
March 12-14, 2015

In many big data projects, there is no large data analysis happening, but the challenge is the extract, transform, load part of data pre-processing. Big data is particularly a problem in business analytics because standard tools and procedures are not designed to search and analyze massive datasets. Big data can be contrasted with small data, another evolving term that's often used to describe data whose volume and format can be easily used for self-service analytics. A commonly quoted axiom is that "big data is for machines; small data is for people."

The typical procedure of extracting the relevant information from the big data is commonly known as big data analytics. The procedure for analysis can be depicted with the help of Figure-1.

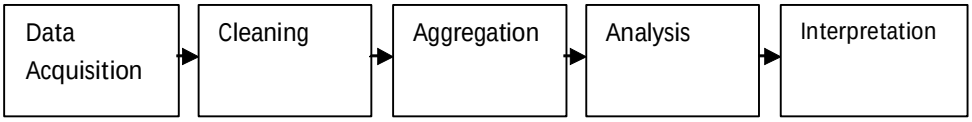

Figure 1: Steps in Big Data analysis

1.1 Library initiatives in the Big Data

The librarians have also contributed their part in the big data phenomenon. There are many collaborative initiatives launched all over the world by the library and information services organizations. Some of them are as below.

(i) DataOne, Data Concervancy and Data to Insight Center have all been funded by NSF through DataNet.

(ii) MetaArchive, a dark preservation archive making use of peer-to-peer technology, and GeoMAPP, focsed on the preservation of local/state government spatial data were both funded by Library of Congress's NDIIPP program.

(iii) The Library of Congress's NDIIPP program has given rise to the National Digital Stewardship Alliance, which includes members from academia, industry, and government, convened to work on Content, Standards and Practices, Infrastructure, Innovation, Outreach.

(iv) DataCite is an organization founded by several European national libraries and including some North American libraries to work with the publishing industry to develop the mechanisms to assign persistent, unique identifiers to datasets so that they can be cited.

(v) The Digital Preservation Network is the newest and perhaps the most ambitious. Over 75 universities have contributed at least $20k to fund an investigation of a NATIONAL preservation network.

2. Literature Review

The term “big data” was coined around eight years back and the research was started around the same time. There have been many researches in this field till now. Only few of them are in the field of creating metadata or catalogue for the big data.

Franklin et al. (2005) have proposed the idea of dataspaces and the development of DataSpace Support Platforms (DSSP), as a means of addressing the challenges of information management of the

organization's many diverse but often inter-related data sources. This paper suggest that for each participant in the dataspace, the catalog should include the schema of the source, statistics, rates of change, accu-racy, completeness, query answering capabilities, ownership, and access and privacy policies.

Siwach (2014) proposed an approach for identifying the encoding technique to advance towards an expedited search over encrypted text leading to the security enhancements in big data.

Cárdenas et al. (2013) have given the differentiators of traditional and big data and emphasized on volume, variety and velocity of the data. In the paper they investigated security from first generation 'Intrusion detection systems' to third generation 'Big Data in analytics'. Focus is on big data security and the use of cluster Infrastructures that makes it more reliable and available.

Sugimoto et al. (2012) discussed on the tools, techniques, and theories that LIS can bring to Big Data research and the role that the LIS discipline should play in this new era.

Lesk (2013) has highlighted the role of librarians by stating that the significance of analytics for libraries is that the skills needed for this work are similar to data management skills, and if, as is likely, all large libraries are doing web analytics, they are employing people who have that set of skills, and combined with librarianship, are 2/3 of the way to being scientific data curators.

Routzahn (2013) gives information about the IBM initiative of the IBM Big Data Catalog planned by IBM claimed to be designed to simplify the process that enables end users, data scientists, and other business analysts to peruse data. It is expected to ingest and store metadata from every available source, and it will classify data by such factors as origin, lineage, and potential value.

Vemuganti (2013) concluded that metadata and its management is an often ignored area in enterprises with multiple consequences. The absence of robust metadata management processes lead to erroneous results, project delays and multiple interpretations of business data entities. These are all avoidable with a good metadata management framework.

3. Challenges of Big Data Analytics

There are many challenges posed by the inherent characteristics of the big data. Various challenges posed by these characteristics are identified below.

3.1 Heterogeneity and IncomPleteness

The requirement of a data analysis algorithm is the structured homogeneous data. But, the datasets in the big data can be non-homogeneous having varied file formats and also unstructured or with varied structures. As shown in the Figure-1, the processes of cleaning and aggregation are used to convert the heterogeneous datasets into a form suitable for data analysis. Likewise, even after data cleaning and error-correction, some incompleteness and some errors in data are likely to remain.

3.2 Metadata Knowledge

The knowledge or absence of knowledge concerning metadata of the datasets is another challenge for big data analytics. The structured data stored in database management systems are expected to have their metadata in the form of data dictionary or system catalogue. But, the unstructured data in various file structures cannot be expected to have theirs. The catalogue information for such datasets needs to be created and distributed. Moreover, there must be uniformity in the metadata information. A com-

mon format catalogue, like a library catalogue, is needed for varying datasets belonging to big data.

3.3 Scale

Another very ominous challenge for big data analytics is the very large and also rapidly increasing volume of data. The query over the big data must be executed speedily over all the relevant datasets. The sub-queries need be distributed and executed in parallel on each of the relevant datasets and the result should be aggregated before presentation of results. The execution of this process over big data will not be easy to satisfy the requirement of real time presentation of results.

3.4 Timeliness

In today's highly competitive business environment companies not only have to find and analyze the relevant data needed, but it must be obtained very quickly. The big data query for information retrieval should be executed speedily in almost real time to satisfy the requirements of applications.

3.5 Result Accuracy

We can find and analyze data quickly and put it in the proper context for the audience that will be consuming the information, the value of data for decision-making purposes will be compromised if the data is not accurate or irrelevant. This is a challenge with any data analysis, but when considering the volumes of information involved in big data, it becomes even more pronounced. Data visualization will only prove to be a valuable tool if the data quality is assured. It's always best to have a pro-active method to address data quality issues so problems won't arise later.

4. 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 dataspace 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 crowdsourcing and automated metadata discovery. Another option is to follow manual classification and cataloguing by the experts.

4.1 Crowdsourced Metadata

Crowdsourcing 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. Crowdsourced metadata are generally perceived to match with the user search criteria.

4.2 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.

4.3 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 crowdsourced metadata and automated metadata discover are the technically correct subject classification of the dataset by the expert and non-repetitive nature of subject classification will avoid future efforts.

5. Conclusion

While outlining the challenges of big data analytics, it has been observed that the inherent characteristics like heterogeneous and unstructured data sets in the big data pose a serious problem in data acquisition and cleaning process required in big data analysis. It is observed that there are many challenges in the practical use of big data. It is proposed to use a catalogue for big data to address various challenges of big data use. Librarians have a great role in contributing metadata to the big data. Classification and cataloguing of big data by library experts and data scientists is quite efficient in comparison of crowdsourcing of metadata and automated metadata discovery for cataloguing procedure to catalogue the dataset metadata.

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. Piatetsky, Gregory (2014). Interview: Michael Berthold, KNIME Founder, on Research, Creativity, Big Data, and Privacy, Part 2. Available

at: http://www.kdnuggets.com/2014/08/interview-michael-berthold-knime-research-big-data-privacy-part2.html (Accessed on 13.01.2015).

4. Gartner Group (2011). Pattern Based Strategy: Getting Value From Big Data. Available at: http://www.gartner.com/it/page.jsp?id=1731916 (Accessed on 13.01.2015).

5. Cárdenas, Alvaro A. Manadhata, Pratyusa K. and Rajan, Sree (2013). Big Data Analytics for Security Intelligence. Big Data Working Group Cloud Security Alliance. Available at: https://cloudsecurityalliance.org/ download/big-data-analytics-for-security-intelligence/ (Accessed on 13.01.2015).

6. 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.

7. Furlough, Mike (2012). Research Libraries and "Big Data". CENDI/NFAIS Workshop. Washington, DC.

8. Stromberg, Joseph (2013). The vast majority of raw data from old scientific studies may now be missing. Available at: http://www.smithsonianmag.com/science-nature/the-vast-majority-of-raw-data-from-old-scientific-studies-may-now-be-missing-180948067/(Accessed on 25-01-2015).

9. Lesk, Michael (2013). Curators of the Future. New Technology of Library and Information Service. 29(3): 1-7

10. Routzahn, Robert (2013). Shine a Light on Big Data. Available at: http://www.ibmdatamag.com (Accessed on 25-01-2015).

11. Franklin, Michael. Halevy, Alon and Maier, David (2005). From databases to dataspaces: a new abstraction for information management. SIGMOD Rec. 34(4):27--33

12. Balke, Wolf-Tilo. Efficient Outsourcing for Metadata Generation. Available at: http://boemund.dagstuhl.de/mat//Files/12/12171/12171. BalkeWolfTilo.Slides.pdf (Accessed on 25-01-2015).

13. San Diego Supercomputer Center (1997). Massive Data Analysis Systems. Available at: http://www.sdsc.edu/MDAS/Reports/MDAS.Final.SciTech/ techreport-97.1/techreport.html (Accessed on 25-01-2015).

14. Vemuganti, Gautam (2013). Metadata Management in Big Data. Infosys Labs Briefings. Vol. 11 No. 1.

About Authors

Dr. Manish Kumar Singh, Information Scientist, Central Library, Banaras Hindu University, Varanasi.
Email: mks.clbhu@gmail.com

Dr. D K Singh, Dy. Librarian, Central Library, Banaras Hindu University, Varanasi.
Email: dksingh5@yahoo.com