METADATA USABILITY FOR BIG DATA

Dr. Manish Kumar Singh,

Information Scientist,
Central Library, B.H.U., Varanasi

ABSTRACT

Metadata use is an essential component in information retrieval systems. It is equally important element in traditional database management systems or homogeneous group of datasets. Big Data can have any of these two forms. Big data can be characterized by its three characteristics i.e. volume, velocity and variety. Metadata usage depends on its generation, relevance, distribution, storage, etc. This dependency is more complicated with bigger Big Data, as volume and variety increases. This paper analyses the usability of metadata in the perspective of Big Data. Usability is identified for variation in volume, velocity and variety.

Keywords: *Metadata usage, Big Data metadata.*

INTRODUCTION

The term 'Big Data' refers to the universe of data residing in multiple heterogeneous datasets, together containing enormous amount of data along with access mechanism for the user or other application. This definition covers the three important characteristics of the Big Data, i.e. volume, velocity and variety. The term 'volume' refers to the enormous size of the universe of datasets. The term 'velocity' refers to the needed speed at which the distributed data must be accessible to a query maker so as to give a real time experience. The third term 'variety' refers to the heterogeneous nature of different datasets. Datasets may be structured database or unstructured file or a data stream. Even the structured database may follow different database models.

Several digital initiatives, like digital marketing, digital research and the social web have created a situation in which information seekers are now gaining access to huge quantities of data on an unprecedented scale, leading to new research challenges in data curation and processing. Funding initiatives like Digging into Data (Digging Into Data, 2012) are now explicitly encouraging researchers to engage in studies that lead to greater understanding, use, and applications of Big Data. Big Data resource consolidates many different information sources without pre-supposing the questions that might be asked and allows the business, or its Data Scientist proxies, to explore the data and discover valuable insights (Rowlands, 2013). Effective harnessing, maintenance and use the ever-growing volume of structured data and unstructured content creates competitive advantages by helping a business make better, faster decisions.

Metadata is tightly interwoven into our data, whether it is structured dataset or unstructured one. Metadata allows us to do sophisticated processing/analysis against structured data. With structured data, we can ask queries. There are no attributes in unstructured data that allow us to make anything but the most basic of queries. Because of lack of attribution, we cannot do sophisticated analysis and processing of data that is unstructured. To assure that the users get the real time experience while using Big Data, despite the huge volume, high velocity and enormous variety of Big Data, it is essential that the usage analysis of metadata be done to improve upon the quality of the metadata. This task is obviously quite different from the metadata usage in traditional database management systems. Rest of the paper covers the literature review followed by metadata usage analysis.

LITERATURE REVIEW

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, accuracy, completeness, query answering capabilities, ownership, and access and privacy policies. Dataspaces are not a data integration approach; rather, they are more of a data co-existence approach.

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 ignore 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.

Rowlands (2013) states that Big Data means that a lot of the metadata we have long cherished might go away and that new types of metadata are going to need to be managed. The metamodel is going to change massively. And there are going to be disconnects between the metadata we use now, and "Big" metadata.

METADATA CREATION AND MAINTENANCE

The metadata creation and maintenance tasks need to be tuned to the requirements of Big Data. These are described below (Vemuganti, 2013). **Metadata Discovery:** In the case of Big Data, the metadata needs to be discovered from various datasets, as all the datasets are not expected to have their metadata with them. As the data is sourced from multiple sources and all these sources have different formats, some structured and some others unstructured, the discovered metadata needs to be harmonized. The process of metadata discovery needs to be formalized. **Metadata Collection:** To enable the search over the universe of Big Data, a metadata collection mechanism should be implemented. A robust collection mechanism should address the volume, velocity and variety characteristics of the Big Data. A technology or a process for metadata collection needs to be in place. **Metadata Storage:** The collected metadata needs to be stored in database structures conforming to a model for enterprise metadata storage. If existing models are not fit for the requirements for the enterprise then suitable custom models can be developed. **Metadata Distribution:** For enabling the search over the universe of Big Data, the metadata will need to be distributed to consuming applications. A formal distribution structure should be put in place to enable this distribution. Along with distribution, creation of application interfaces is also required.

METADATA USABILITY ANALYSIS

Finding the sought data with minimal expense of time and effort is a key challenge when collecting data from a Big Data repository, as these metadata based data systems can carry out searches across both structured data applications and unstructured content repositories. Incorporating meaningful metadata attributes into structured data and unstructured content makes information assets more actionable.

We can classify metadata into three categories according to their usage. The **first** is business metadata, or data that conforms to transaction regulations and are tailored for business use. This is a more common based on intuitive relations and organized so that users can search for particular sets of data more easily. The **second** type is database metadata, the set of labels referring to the datasets, and the structure/organization/schema of data. This is also used for security purposes, such as keeping track of when the users last accessed the data and possibly even for what reason. The **third** type of metadata is application metadata, an elaborate type of data that explains what other metadata means and who has access to it.

A usage test was conducted on a single source of Big Data using different types of metadata. The usage of data, measured in successful retrieval and quantified by variance (~resilience) in defined deviation in three characteristics, is tabulated for each of the types of metadata.

Table 1: Scaled variance in defined deviation for successful retrieval using metadata

	Volume	Velocity	Variety
Business metadata	8.5	8.7	8.4
Database metadata	6.9	8.1	4.3
Application metadata	5.1	7.1	3.2

CONCLUSION

We have to use Big Data to ensure that the tremendous wealth of information provided by it can made to use to maximum extent with usable metadata. Metadata usage analysis is essential to fine tune the metadata performance in retrieval of the relevant information, as the performance of queries depends largely on metadata. After performing usage analysis on three categories of metadata by varying volume, velocity and variety of the big data on a single big data source, it was revealed that database metadata has the least resilience to the bigger big data source and have relatively stable performance in heterogeneous data source. Business data showed least resilience to the variations.

REFERENCES

- 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).
- 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).
- Digging into Data. (2012). Avalable at http://www.diggingintodata.org/Default.aspx (Accessed on 25-01-2015).
- Franklin, Michael. Halevy, Alon and Maier, David (2005). From databases to dataspaces: a new abstraction for information management. SIGMOD Rec. 34(4):27—33.
- Furlough, Mike (2012). Research Libraries and “Big Data”. CENDI/NFAIS Workshop. Washington, DC.
- 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).
- Lesk, Michael (2013). Curators of the Future. New Technology of Library and Information Service. 29(3): 1-7.
- 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).
- Routzahn, Robert (2013). Shine a Light on Big Data. Available at: http://www.ibmdatamag.com (Accessed on 25-01-2015).
- Rowlands, Ian (2013). Big Data Changes the Metadata Point of View. Available at: http://www.dataversity.net/big-data-changes-the-metadata-point-of-view/ (Accessed on 25-01-2015).
- 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).
- Siwacch, Gautam and Esmailpour, Amir (2014). Encrypted Search & Cluster Formation in Big Data. IN ASEE 2014 Zone I Conference, University of Bridgeport, Bridgpeort.
- 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.
- 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).

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