Research Journal of Computer and Information Technology Sciences _____E-ISSN 2320 – 6527
Vol. **4(9),** 22-24, September **(2016)** *Res. J. Computer and IT Sci.*

Mini Review Paper

Statistical Survey on Big Data Analytics

Kauleshwar Prasad* and Arpana Rawal
Department of Computer Science & Engineering, B.I.T Durg, Durg, India
kauleshwarprasad@gmail.com

Available online at: www.isca.in, www.isca.me
Received 24th Feburary 2016, revised 30th August 2016, accepted 5th September 2016

Abstract

This paper offers a broader definition of big data along with its importance and scopes. Various characteristics of big data: Volume, Velocity and Variety (V3) are discussed. Big Data solutions can be considered as ideal for analyzing not only raw structured data, but semi structured and unstructured data from a wide variety of sources. Due to the rapid evolution and adoption of big data by industry various researches are going on in this field. This paper basically describes the various analytics of big data like Text Analytics, Audio Analytics, Video Analytics, Social Media Analytics and Predictive Analytics.

Keywords: Volume, Velocity, Variety, Text Analytics, Audio Analytics, Video Analytics, Social Media Analytics and Predictive Analytics.

Introduction

Before 1945, US statistics reveal that American university libraries doubled in size, every sixteen years. If this growth rate exists then it has been estimated that in 2040 it will become 2PB. In 1997 M. COX and D. Ellsworth mentioned in their article that data sets are increasing in such a way that they do not fit in main memory, local disk, even remote disk. It was the first article in which the term big data were used[1]. The origin of 'Big Data' is uncertain but it is a large concept and is everywhere today. Diebold (2012) argues that the term "big data probably originated in lunch-table conversations at Silicon Graphics Inc. (SGI) in the mid-1990s, in which John Mashey figured prominently". Figure-1 shows that the average monthly frequency distribution of documents contains the term "Big Data" is increasing year by year and from 2011 it shows fast growth, Amir[2]. In April 2012, an online survey of 154 C – suite global executive was done by Harris Interactive on behalf of SAP in which he got confused definition of Big Data. As shown in Figure-2, the definitions of Big Data vary with application domains. 18% of the definitions described Big Data as explosion of new data sources (social media, mobile device, and machine-gene rated devices), 19% told that it is a requirement to store and archive data for regulatory and compliance, 24% gave their view as new technologies designed to address the volume, variety, and velocity challenges of Big Data, 28% told that it is a massive growth of transaction data, including data from customers and the supply chain, Amir[2].

Big Data Size: A Statistical View

Big Data is a term denoting large and complex data sets procured over huge timelines. In contrast to traditional database storage / processing technologies, Big Data is differentiated in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety).

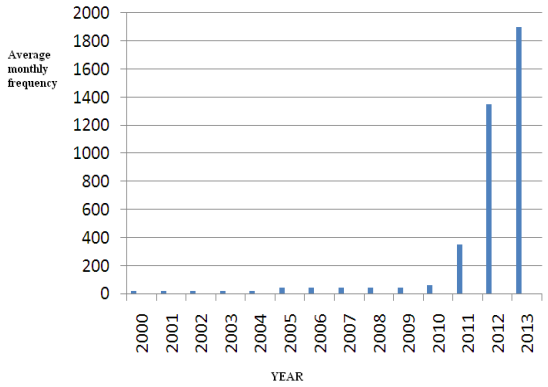

Figure-1
Frequency distribution of documents containing the term "Big Data" in ProQuest Research Library

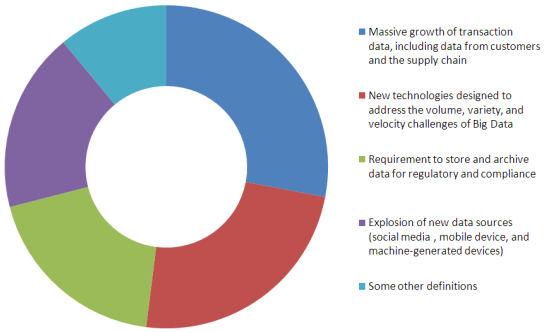

Figure-2
Big data definitions based on an online survey of 154 global executives in April 2012

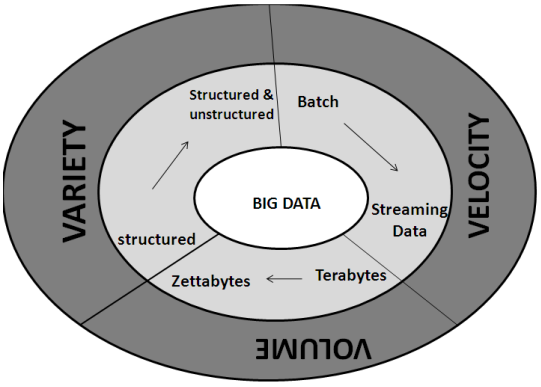

Figure-3
Characterization of Big Data

Volume: It refers to magnitude of data. Nowadays it is seen that amount of data produced from machine is much larger than non traditional approach. For example, students' attendance in all higher educational institutions of the Nation with an average student intake of 500 per annum on a specific working day can generate, on an average, Terabytes or Pita bytes of data. In domains like social sites, we find data in the video, text, music and large images format. Big Data Analytics have gone up to the extent of handling data from units of terabytes to petabytes, zeta bytes.

Velocity: It represents the speed in which data get updated. In earlier days, it was believed that the data of yesterday is recent. But nowadays the data of 1 second is old. People used to update their status second by second through social media. This high velocity data represent Big Data.

Variety: In earlier days data formats were well defined by a data schema and change slowly. But nowadays due to advancement in technology like new services added, new sensors deployed or new marketing campaigns executed new data types are needed to capture the resultant information.

Big Data Analytics on Move

The process of extracting insights from big data is Big Data analytics. It has been broken down into five stages. These five stages have been grouped into two sub process: Data Management and Analytics as shown in Figure-4.

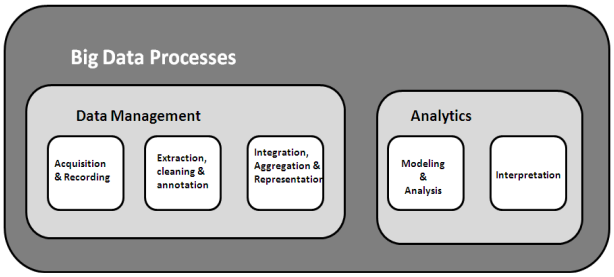

Figure-4
Processes for extracting insights from Big Data

Data Management contains processes such as acquisition and store data, extraction, cleaning and annotation and to prepare and retrieve it for analysis. Analytics contains modeling and analysis and then interpretation of data. In the following sub section, analytical techniques for both structured and unstructured data is described.

Text Analytics: The term text analytics describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, quantitative / qualitative research, or business-oriented investigations[3]. Large volumes of human generated texts can be easily converted into meaningful summaries so that evidence based decision making can be implemented. For example, in stock market information from financial news can be easily extracted or fro series of question-answering sessions during informal surveys / formal interviews, biographies of targeted celebrities and statesmen can be generated. Such tasks include subtasks like Information retrieval, Named entity recognition, co reference, Relationship fact and event extraction, sentiment analysis and Quantitative text analysis.

Audio Analytics: Audio Analytics is the process of extracting and analyzing the information from unstructured audio data. It is also called Speech Analytics. Nowadays primary application of audio analytics is in the field of healthcare and call centers. In call centers thousands of calls are attended from customers daily and these calls are analyzed and used to provide feedback to agents in real time. In healthcare, audio analytics is used for the treatment and diagnosis of certain medical conditions that affect the patient's communication patterns and even infant's cries, Amir[2].

Video Analytics: Video Analytics are also called Video Content Analysis (VCA). Nowadays it plays very important role in the field of surveillance, automated security, entertainment and healthcare etc.

Social Media Analytics: Social media is a set of variety of online platforms that allow users to create and exchange content. It has been categorized into following types: social networks (e.g., Facebook and Linkedin), blogs (e.g., Blogger and Wordpress), microblogs (e.g., Twitter and Tumblr), social news (e.g., Digg and Reddit), social bookmarking (e.g., Delicious and Stumble Upon), media sharing (e.g., Instagram and YouTube), wikis (e.g., Wikipedia and Wikihow), question-and-answer sites (e.g., Yahoo! Answers and Ask.com) and review sites (e.g., Yelp, TripAdvisor). Social media analytics is defined as the analysis of unstructured and structured data from social media channels, Barbier[7].

Predictive Analytics: As the name tells, it is based on prediction. Prediction can be done in various fields like share market, failure of jet engines based on the stream of data from several thousand sensors, customer's next moves based on what

they buy, when they buy and even what they say on social media. In this way we can say that predictive analytics refers to the technique that predict future outcomes based on historical and current data, Amir [2].

Future Research Directions

Many Business sectors like health, nutrition, education, telecommunication, marketing, sports, marketing and business management that manages big data world are heading forward with Big Data Analytics to be polished for incorporating business intelligence in future. With the ever-growing dynamic real-time organizational databases available at short analytical time spans, the need of the hour is to narrow up the analytical gaps of all the interconnected business domains across the globe. Assuming that all the global transaction frontiers adopt Big Data Analytics, then only the lifecycles of Big Data and resource / energy optimization processes can be deployed for maintaining ecological balance between mankind as well as science and technology.

Conclusion

The main objective of this paper is to review, describe and reveal on big data. This paper focuses on definitions and dimensions of big data. We reviewed analytics techniques such as text, audio, video, social media and predictive analytics. At last we show future aspects in which big data can be used.

References

1. Gu Jifa and Zhang Lingling (2014). Data, DIKW, Big data and Data science. 2nd International Conference on Information Technology and Quantitative Management, ITQM, *Procedia Computer Science,* 31, 814-821.

2. Gandomi Amir and Haider Murtaza (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35, 137-144.

3. Sabia Sheetal Kalra (2014). Applications of Big Data: Current Status and Future Scope. *International Journal on Advanced Computer Theory and Engineering*, 25-29.

4. Isaac Triguero, Daniel Peralta, Jaume Bacardit, Salvador García and Francisco Herrera (2014). MRPR: A Map Reduce solution for prototype reduction in big data classification. Neurocomputing, 150, Part A, 331-345.

5. Jacques Bughin (2016). Big data, Big bang?. *Journal of Big Data*, 3(2), 2-14.

6. Nevenka Dimitrova, Hong-Jiang Zhang Behzad Shahraray, Ibrahim Sezan, Thomas Huang and Avideh Zakhor (2002). Applications of Video Content Analysis and Retrieval. IEEE Multimedia.

7. Barbier G. and Liu H. (2011). Data mining in social media. C.C. Aggarwal (Ed.), Social network data analytics, United States: Springer.

8. Chris Eaton, Dirk Deroos, Thomas Deutsch, George Lapis, Paul C. Zikopoulos (2012). Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data. Mc Graw Hill, Ney Work, India.

9. Aggarwal Charu C. (2011). An Introduction to Social Network Data Analytics. Springer, 1-15.

10. Hua Fang et. al. (2015). A Survey of Big Data Research. *IEEE Netw.*, 29(5), 6-9.