A Survey on Big Data

Ashok Kashyap G C
Information Science and Engineering (B.E),
Rajeev Institute of Technology
(Affiliated to VTU, Belgaum) Hassan, India

Pooja B S, Assistant Professor,
Dept. of Information Science and Engineering,
Rajeev Institute of Technology
(Affiliated to VTU, Belgaum) Hassan, India

***Abstract*— This paper aims to highlight distinct features of Big Data. We are living in the world with vast varieties and tremendous volume of data. Data is the new currency. Data is being generated by sensors, digital images, activities in social media sites, forensic science, business informatics, research activities across various domains and many other internet based events as either source data or as cumulative to the existing data. This data increase from time to time exponentially from heterogeneous sources, techniques and technologies. The data is categories as "Big Data". The core theme focuses on linking Big Data with people in various ways. Big Data is enormous in Variety, Velocity and Volume. It may be in any of the form like structured, unstructured. The idea of Big Data analysis is to manage giant volume of data, obtain beneficial information, suggest casualties and support decision making. This survey provides comprehensive review and audit of Big Data analytics. Leading and evolving applications of Big Data analytics are discussed. Some of the techniques for efficient analysis of Big Data are also illustrated.**

Keywords: Big Data management, Big Data Analytics, Big Data Analyzing technique

I. INTRODUCTION

It was in early 21st century concept of Big Data came into existence and started to evolve. It was the first time when attributes like volume, structure and speed were used for the describing the nature of data. Big Data's important attribute is the volume. Data is quantified by counting the space occupied, digital transactions, statistical tables, or files but it was found more constructive to symbolize Big Data with respect to time. The very next is the variety of data. This happens, as data come from variety of sources like census, blogs, logs, streams, issue of nationalized identification cards, research data, partial structured data from business-to-business processes, satellite images. The last attribute is the velocity that refers to speed of applying the analytics and processing the data.

Big Data is vast in majority and complex data. Dissimilarity, storage and transport, privacy and security, and complexity problems with Big Data impede the progress at all stages of that can create value from data. There are various sources of Big Data, for example: Opinion polls, audio-visual files, scientific data, various database tables, email attachments etc. Big Data has great importance in fields like research, public sector services, healthcare services, web/social, manufacturing, artificial intelligence, education and cyber-physical models. Big data have priority in every sector in the global economy. It was calculated that by 2005, practically all arena in the economy will have 200 terabytes of minimum data stored per company having more than 1,000 employee. Big data forward to enlarge rapidly, driven by mutation and modification in elementary technologies. Conventional data administration and analysis system substantially depend on Relational database management system (RDBMS).

There major aspects in which RDBMS and Big Data differs are

1) RDBMS is limited to structured data, but big data supports various data processing architectures.

2) RDBMS provides an insight to a problem at the small level, big data offer better view and efficient operations on metadata and unstructured data.

When does analytics become Big Data Analytics? The size that defines Big Data has grown. In 1975 attendees of the first VLDB (Very large databases) conferences worried about handling the Millions of data points found in US census Information. Big Data Analytics is the course of classifying bulk datasets to the variety of data type i.e. indirect relations, digitized documents, consumer priorities and other useful details. The analytics can lead to efficient marketing, better quality of services. Big Data analytics project are instantaneously emerging as the dignified solution to recognize business and technology trends that are disturbing traditional data management procedures. Analytics helps to discover requirement and possible solutions. With big data analytics, the organizations are trying to identify exit polls, new business facts and trends. This paper includes literature survey of Big Data analytics in section 2. Section 3 contains background and data forms of Big Data. Section 4 contains Big Data analytics in detail and section 5 contains techniques to analyze big data and section 6 concludes the paper.

II. LITERATURE SURVEY

Over last many years, there are many researchers and scholars completed their work successfully on Big Data. Many articles have been published in the various journals and magazines (For example Forbes, Harvard Business review, Optimize, The Wall street journal).The Government of India have implement enormous techniques of Big Data to determine the feedback of Indian electorate to government plans and policies. The Obama Administration has announced that, it would invest the 200 million dollars on big data research plan in March 2012.

Reports of International Data Corporation predicts that global Data from 2005 to 2020 will grow by factor of 10. The

global data volume will grow from 0.13 Zettabyte's to 40 Zettabyte's, depicting double accumulation for every two years. IBM evaluates that everyday 2.5 quintillion bytes of data will be originated. Out of which 90% of the data in the world today is created from the last two years. It is observed that social networking sites like Facebook have 750 million users with 350 million photo

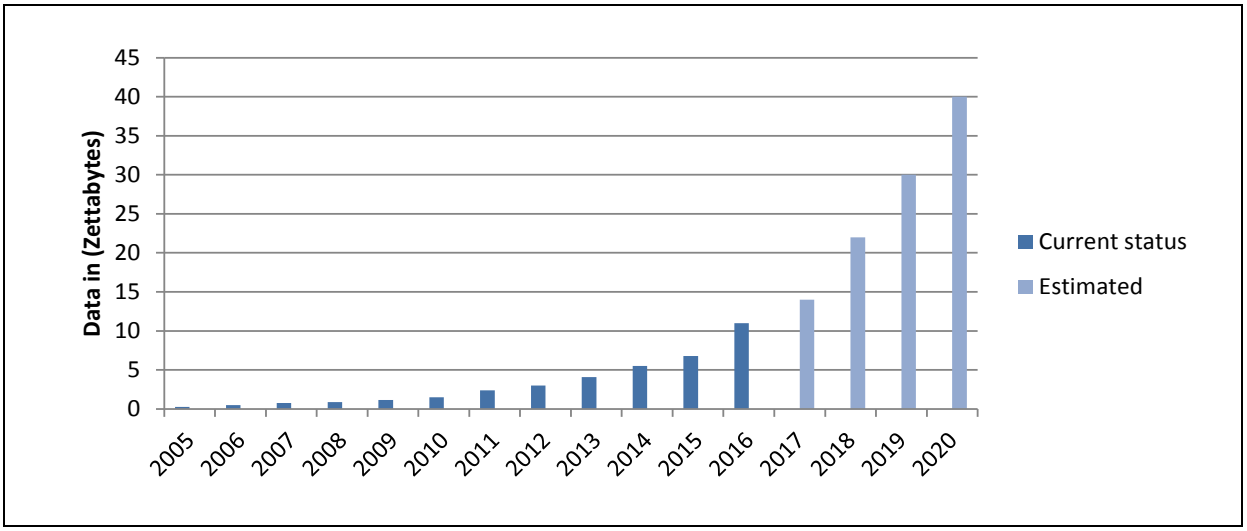


Figure 1: Data volume growth by year in zettabytes

uploads per day, LinkedIn has 110 million addressee and Twitter has 320 million accounts with 500 million tweets per day. From industry, government and research community, Big Data has led to advancement in the research field that has attracted immense interest. The major concern is coverage on both industrial reports and public media for example: The Economic Times, The Hindu, Times of India. Smart devices and mobile phones are the best way to get data from people in divergent aspect, the huge magnitude of data that mobile carrier can process to make our day to day life easier. In the Figure 1, it would present that the amount of data practically increased from the year 2005 to 2016 and the estimation, that data would increase from the 2017 to 2020. However, consider exponential growth in data from the year 2005, when enterprise system and user level data flood into data warehouse.

Figure 2 illustrates the diversity in data stored from different sectors. The type of data induced and stored are audio, video, digital images and text format and differ from one sector to another. Text/numeric data will be from the sectors that are directly related to research and development community, public zone like banking, government and health care. Audio and video nature of data is from various fields of communication and media.

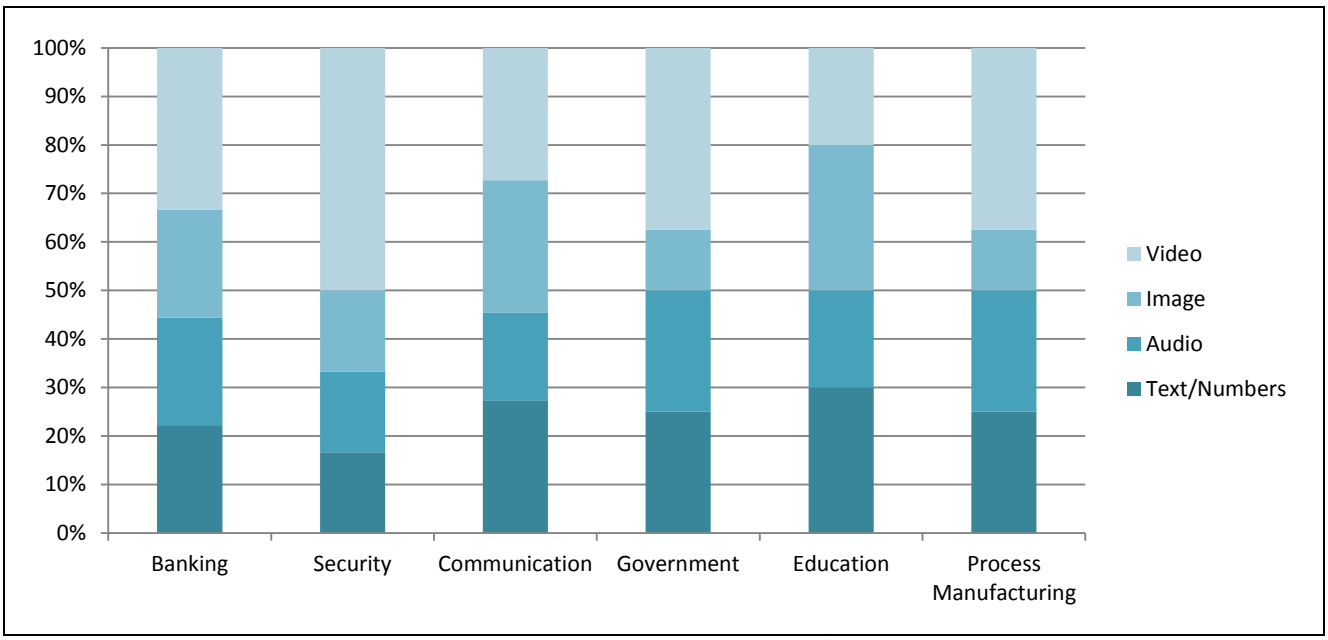
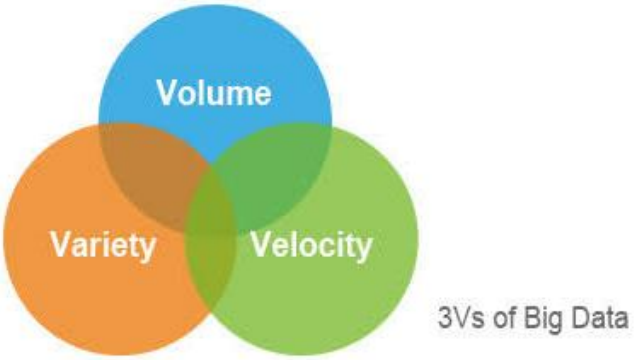


Figure 2: Variations possible in generating and growth of data

III. BIG DATA

Big Data is the term that contains large and complex datasets. It is tedious work to manage these datasets without new technology. The Mckinsey Global Institute (MGI) published a report on Big Data that describes the various business opportunities that big data reveal. Paulo Boldi, One of the authors says "Big Data does not need big machines, it needs big intelligence".



3Vs of Big Data

A. *V's of Bigdata*

- VOLUME: This would refer to the data from multiple sources, data being in huge capacity. It can include all and any kind of data, including the data that is created from all the connected devices, mobile data, internet and all the data that is being resulted from this communication.

- VELOCITY: Velocity not only involves the speed at which the data is transferred, but will also involve, data streams, creation of structured records, access to data & delivery. The issues do not only lie with the velocity of incoming data but also to stream outgoing data for batch processing.

- VARIETY: This refers to varied data types and the same can be accumulated from various sources, sources being: social networks, smartphone, sensors in the forms of videos, images, audio, logs etc. This data can be highly structured (data fetched from the traditional database systems), semi-structured (feeds like reviews, comments) or unstructured (clicks, audios, images, videos).

B. *Data forms*

Data collected can be broadly classified into the following categories

1) Structured data: Data will be placed in the current data warehouse i.e. in the relational database management system. The structure of the relational database management system will be enforced on the current data warehouse system, which is inclusive to understand the meaning associated with it. Details will be provided such that data with respect to which columns are placed where, whom are they associated with and how the columns are associated in between tables. The format of the data can be in text or numerical, but it is common understanding that for every person there is a unique identifier in terms of Age.

The entire data is organized in terms of Entities (Semantic Chunks).

- Relations or Classes (Similar entities are grouped together).
- Attributes (Same descriptions for entities existing in the same groups)
- Schema (All Entities in the group have a description associated with it.
    - All are present & follow same order.
    - All of them have same format defined and length defined.

2) *Semi-structured data:* The format of data do not confirm an explicit and exact schema, however the tags associated with the data, if found associated with organizational structure, then the same data would be easier to organize and analyze. The same concept described here would predate the idea of XML.

- Data is available in many formats, in the current scenario, electronically
    - File Systems e.g., Web data
    - Data Exchange Formats, e.g., Scientific data
- Data that is not completely structured, but
    - Similar entries will be grouped and semantically organized
    - Entities may not have same attributes in the group

*3) Unstructured data***:** Unstructured data would be in a format that cannot be easily indexed. Indexing is the method of referring relational tables for the purpose of querying or analysis. This would include the file types that are associated with audio, video and image files.

- Data – Any type.
- No Format and proper sequences.

IV. BIG DATA ANALYTICS

Big Data analytics permit enterprises for better analysis of a mix of structured, semi structured and unstructured resulted due to reviews by the customers, precious business statistics. The Mckinsey Global Institute, propagated a major research work in June 2011 on Big Data. Its overloading conclusion: Big Data is "a key basis of competition and growth". The expression Analytics (inclusive of Big Data form) is often used broadly to wrap up data-driven decision making. The term analytics classified into major subdivision: Corporate analytics and Academic research analytics. In Corporate Analytics, data is treated as the asset and major concentration is on increasing the revenue. In Academic Analytics, Researchers make use of data to test Hypothesis and form theories.

Researchers of big data analytics have found the data collected is divided into various Big Data application such as follows.

A. *Structured Analysis*

In structured analytics, data is generated from high degree of business organizations and scientific research fields. These data is organized and queried by RDBMS, Data warehousing, and various search algorithms. Data is grown by different research areas like Privacy, preserving, data mining, E-commerce.

B. *Text Analytics*

In Text analytics, text the most common way of storing the information and it includes e-mail, digital libraries, chat messages, and social media contents. Text analytics also known as Text mining, concentrate on deriving correct and effective information from massive text file. Text mining system is relay on statistical pattern learning and Natural Language Processing (NLP) with importance on the letters.

C. *WebAnalytics*

The objective of Web analytics is to fetch the information from Web Pages. Web Analytics also called Web mining.

D. *Multimedia Analytics*

Multimedia data includes animated sequences, graphic objects, and computer aided draft and drawing, audio-visual files. It has grown at a gigantic rate. Multimedia analytics refers to extract advantageous knowledge and semantics exist in multimedia data. Data types of multimedia data are printable characters, sound, volume, pixels.

E. *Mobile Analytics*

Mobile data traffic increased to 7.2 Zettabytes per month at the end of 2016. Vast collection of data and application leads to mobile analytics. Mobile analytics involves RFID (Radiofrequency identification), mobile phones, sensors etc.

V. TECHNIQUE FOR ANALYSIS OF BIG DATA

There are several techniques that can be used to process datasets. Some techniques are machine learning, A/B testing. These techniques, analyze new combination of datasets

A. *A/B Testing*

A technique in which particular or reference group is compared with a variety test of groups to determine the best performance between variants. Reference group is a constant called as control group and the test group is the variable called as the treatment group. Changes will be implemented on the objective variable, e.g., acceptance rate of products. An example application is fraud detection with suspect as reference (constant) and forensic data collected at the crime scene as a treatment group (variable). When the variable manipulated in the treatment is more than one, technique is often called "A/B/N" testing

B. *Classification*

A technique in which to identify the categories of new datasets and assign into predefined classes for example classification of mushroom as edible or poisonous. It is used for data mining.

C. *Crowdsourcing*

A technique in which collected data submitted by large group of people or community i.e. crowd. It is usually through network media such as web.

D. *Data Mining*

Method in which exact pattern of data from large existing datasets are examined to generate new information by exercising certain rules. It has applications in machine learning and artificial intelligence.

VI. CONCLUSION

In this paper, concept of Big Data is presented. Big data is the massive inter related datasets and it generate from various sources like social media, comments and reviews, smart and sensible devices, email attachments etc. There is complexity in Big Data such as Velocity, Variety and Volume. These three terms are more challenging for Big Data analytics. Provided literature survey shows exponential growth of data in industries from 2005 year. There are variations possible while generating and storing data whether data is in audio, video, images and text. In Big Data Analytics, researchers divided generated data into various big data application such as structured data analytics, text analytics, web analytics, multimedia analytics and mobile analytics. Many challenges in the big data system need further research attention. Research on typical Big Data application generates profit for commercial organization, enhance the effectiveness of government sectors among the people.

ACKNOWLEDGMENT

I would like to thank my guide and all people who encouraged and helped me to prepare this paper. Finally, I'm indebted to all websites and journal papers which I have refer to prepare this survey paper successfully.

REFERENCES

[1] Understandable Big Data: A survey Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle LE2I UMR6306, CNRS, ENSAM, Univ. Bourgogne Franche- Comté, F-21000 Dijon, France
[2] Yuri Demchenko ―The Big Data Architecture Framework (BDAF)‖ Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.
[3] A Review Paper on Big Data Analytics Ankita S. Tiwarkhede1, Prof. Vinit KakdeInternational Journal of Science and Research (IJSR) ISSN (Online): 2319-7064
[4] Survey Paper on Big Data. C. Lakshmi*, V. V. Nagendra Kumar International Journal of Advanced Research in Computer Science and Software Engineering. Volume 6, Issue 8, August 2016.
[5] ARPN Journal of Engineering and Applied Sciences ©2006-2015 Asian Research Publishing Network (ARPN). VOL. 10, NO. 8, MAY 2015 ISSN 1819-6608
[6] P. Russom, et al. Big data analytics, TDWI Best PracticesReport, Fourth Quarter.
[7] American Institute Of Physics(AIP), 2010. College Park, MD(http:// www.aip.org /fyi/2010/)
[8] http://www.oyster-ims.com/wp-content/uploads/2014/01/Global-datavolume
[9] http://www.deltapowersolutions.com/media/images/news/news-2014-big-data-3v(en)