International Journal of Engineering and Applied Computer Science (IJEACS)
Volume: 02, Issue: 06, ISBN: 978-0-9957075-7-3, June 2017

# Survey on Big Data Analytics

Dhruva M.S
Assistant Professor
Dept. of Computer Science & Engineering
Rajeev Institute of Technology
Hassan, India

Shashikala M.K
Assistant Professor
Dept. of Computer Science & Engineering
Rajeev Institute of Technology
Hassan, India

***Abstract*—This paper aims to highlight distinct features of Big of information. We are living on the planet with huge varieties and tremendous volume of data Information is the new money. 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 [1]. 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 [9] Big Data with respect to time [3]. 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 [6], 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 [2]. 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 [11]. Big data forward to enlarge rapidly, driven by mutation and modification in elementary technologies [15]. Conventional data administration and analysis system substantially depend on Relational database management system (RDBMS) [17].

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 [18].

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 [18].

At the point when does examination turn out to be Big Data Analytics? The size that characterizes Big Data has developed. In 1975 participants of the primary VLDB (Very substantial databases) gatherings stressed over dealing with the Millions of information focuses found in US statistics Information. Huge 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 examination can prompt productive showcasing, better nature of administrations [1]. Huge Data examination venture are promptly rising as the honorable answer for perceive business and innovation slants that are irritating conventional information administration strategies. Examination finds necessity and conceivable arrangements. With enormous information investigation, the associations are attempting to recognize leave surveys, new business actualities and patterns. This paper incorporates writing overview of Big Data examination in segment 2. Segment 3 contains foundation and information types of Big Data. Segment 4 contains Big Data investigation in detail and area 5 contains systems to break down enormous information and segment 6 finishes up the paper [8].

## 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 has implement enormous [12] 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[13].

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.

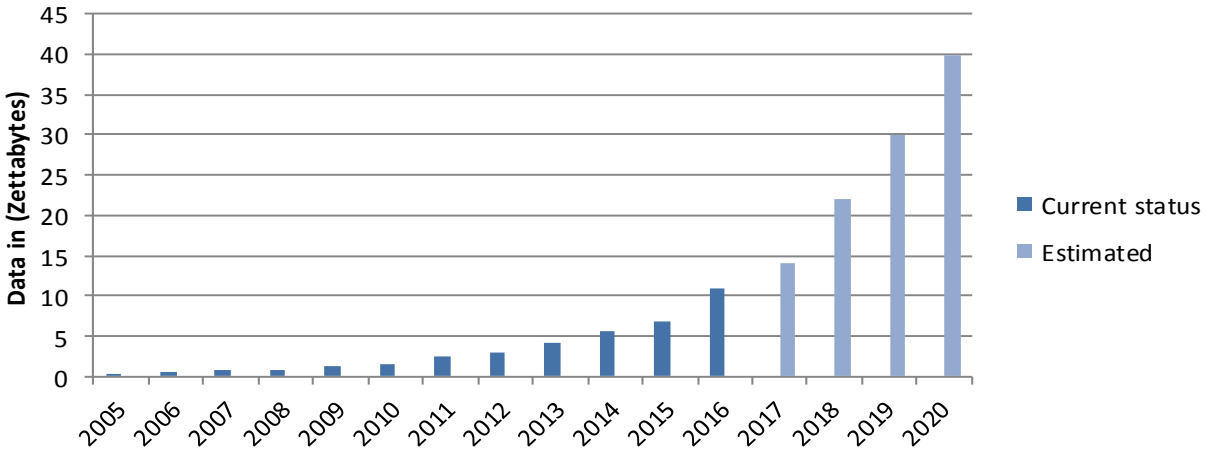



Figure 1. Data volume growth by year in zettabytes

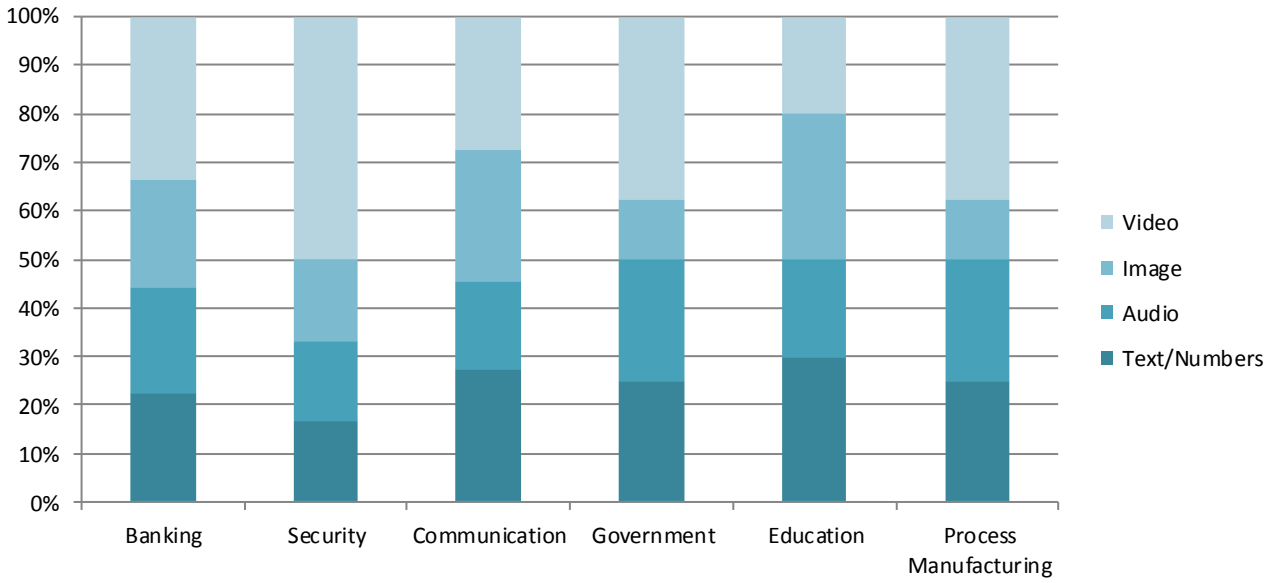



Figure 2: Variations possible in generating and growth of data

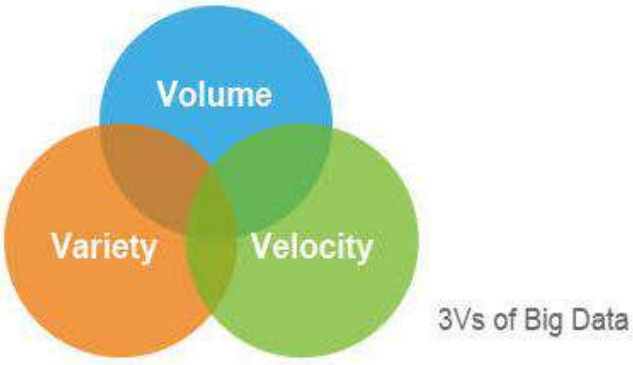
It is observed that social networking sites like Facebook have 750 million users with 350 million photo uploads per day, LinkedIn has 110 million addressee and Twitter has 320 million accounts with 500 million tweets per day. From industry, government and research community, Big Data has led to advancement in the research field that has attracted immense interest [14].

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 [17], 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 [10].

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 [1]. Audio and video nature of data is from various fields of communication and media.

## III. BIG DATA

Enormous Data is the term that contains expansive and complex datasets. It is dreary work to deal with these datasets without new innovation. The Mckinsey Global Institute (MGI) distributed a give an account of Big Data that portrays the different business openings that huge information uncover. Paulo Boldi, One of the creators says "Enormous Data does not require huge machines; it needs huge knowledge [13].




### A. *V's of Big Data*

- Volume: This would allude to the information from various sources, information being in enormous limit. It can incorporate all and any sort of information, including the information that is made from all the associated gadgets, versatile information, web and every one of the information that is being come about because of this correspondence. [19]
- Velocity: Speed not just includes the speed at which the information is exchanged, however will likewise include, information streams, formation of organized records, access to information and conveyance. The issues don't just lie with the speed of approaching information additionally to stream active information for cluster preparing. [10]
- Variety: This alludes to shifted information sorts and the same can be amassed from different sources, sources being: interpersonal organizations, cell phone, sensors in the types of recordings, pictures, sound, logs and so on. This information can be profoundly organized (information gotten from the conventional database frameworks), semi-organized (nourishes like surveys, remarks) or unstructured (snaps, sounds, pictures, recordings). [8]

### B. *Data Forms*

Data collected can be broadly classified into the following categories.

1) *Structured data*: This alludes to shifted information sorts and the same can be amassed from different sources, sources being: interpersonal organizations, cell phone, sensors in the types of recordings, pictures, sound, logs and so on [5]. This information can be profoundly organized (information gotten from the conventional database frameworks), semi-organized (nourishes like surveys, remarks) or unstructured (snaps, sounds, pictures, recordings). 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 organization of the information can be in content or numerical, however it is regular understanding that for each individual there is a one of a kind identifier as far as Age. [7]
   The whole information is composed as far as Entities (Semantic Chunks). [5]
   - Relations or Classes (Similar elements are gathered together). [7]
   - Attributes (Same portrayals for elements existing in similar gatherings)[1]
   - Schema (All Entities in the gathering have a portrayal related with it. [2]
   - All are available and take after same request.[3]
   - All of them have same organization characterized and length characterized. [5]

2) Semi-organized information: The configuration of information don't affirm an express and correct mapping, however the labels related with the information, if discovered related with authoritative structure, at that point similar information would be less demanding to sort out

and break down. A similar idea depicted here would originate before the possibility of XML.

- Data is accessible in many configurations, in the present situation, electronically
- File Systems e.g., Web information [11]
- Data Exchange Formats, e.g., Scientific information
- Data that is not totally organized, but rather
- Similar sections will be assembled and semantically sorted out
- Entities might not have same traits in the gathering

3) *Unstructured information*: Unstructured information would be in an arrangement that can't be effectively filed. Ordering is the technique for alluding social tables with the end goal of questioning or investigation. This would incorporate the record sorts that are related with sound, video and picture documents. [10]

- Data – Any sort.
- No Format and legitimate successions.

## 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 [6].

Researchers of big data analytics have found the data collected is divided into various Big Data application such as follows [17].

### 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 [10].

### 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 [9] 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 [6]. 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. [15]

### B. *Classification*

A method in which to recognize the classifications of new datasets and dole out into predefined classes for instance grouping of mushroom as consumable or toxic. It is utilized for information mining [14].

### 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 [6].

### 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 [2].

## VI. CONCLUSION

In this paper, idea of Big Data is explored. Huge information is the huge entomb related datasets and it create from different sources like web-based social networking, remarks and audits, brilliant and sensible gadgets, email connections and so on. There is many-sided quality in Big Data, for example, Velocity, Variety and Volume. These three terms are all the more trying for Big Data investigation. If writing overview indicates exponential development of information in businesses from 2005 year. There are varieties conceivable while producing and putting away information whether information is in sound, video, pictures and content. In Big Data Analytics, specialists partitioned created information into different enormous information application, for example, organized information investigation, content examination, web investigation, interactive media examination and portable examination. Many difficulties in the enormous information framework require additionally look into consideration. Investigate on common Big Data application creates benefit for business association, upgrade the adequacy of government segments among the general population.

## 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] Vishal S Patil, Pravin D. Soni , " HADOOP SKELETON & FAULT TOLERANCE IN HADOOP CLUSTERS ", International Journal of Application or Innovation in Engineering & Management (IJAIEM)Volume 2, Issue 2, February 2013 ISSN 2319 – 4847.

[4] Sanjay Rathe, " Big Data and Hadoop with components like Flume, Pig, Hive and Jaql" International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

[5] Yaxiong Zhao, Jie Wu and Cong Liu, " Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework ",TSINGHUA SCIENCE AND TECHNOLOGY ISSNl1007-0214l 05/10l lpp39-50 Volume 19, Number 1, February 2014.

[6] Parmeshwari P. Sabnis, Chaitali A.Laulkar , "SURVEY OF MAPREDUCE OPTIMIZATION METHODS ", ISSN (Print): 2319-2526, Volume -3, Issue -1, 2014.

[7] Puneet Singh Duggal ,Sanchita Paul ," Big Data Analysis Challenges and Solutions ", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

[8] Chen He,Ying Lu,David Swanson, " Matchmaking: A New MapReduce Scheduling Technique ", EECS Department, University of California, Berkeley, Tech. Rep.,April 2009.

[9] Apache HDFS. Available at http://hadoop.apache.org/hdfs [14] Apache Hive. Available at http://hive.apache.org.

[10] Apache HBase. Available at http://hbase.apache.org

[11] Apache Pig. Available at http://pig.apache.org

[12] A Review Paper on Big Data Analytics Ankita S. Tiwarkhede1, Prof. Vinit KakdeInternational Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

[13] 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.

[14] ARPN Journal of Engineering and Applied Sciences ©2006-2015 Asian Research Publishing Network (ARPN). VOL.10, NO. 8, MAY 2015 ISSN 1819-6608

[15] P. Russom, et al. Big data analytics, TDWI Best PracticesReport, Fourth Quarter.

[16] American Institute Of Physics(AIP), 2010. College Park, MD(http:// www.aip.org /fyi/2010/)

[17] http://www.oyster-ims.com/wp-content/uploads/2014/01/Global-datavolume

[18] http://www.deltapowersolutions.com/media/images/news/news-2014-big-data-3v(en)

## AUTHORS' PROFILE

**Dhruva M.S.** completed his B.E. degree and M.Tech. degree in Computer Science and Engineering from Visvesvaraya Technology University, Belgaum, India. Currently he is working as Asst. Professor in the Department of Computer Science and Engineering at Rajeev Institute of Technology, Hassan, India. His areas of interest include multimedia networks, Compiler Design and Algorithms.

**Shashikala M. K** completed her B.E. degree and M. Tech. degree in Computer Science and Engineering from Visvesvaraya Technology University, Belgaum, India Currently she is working as Asst. Professor in the Department of Computer Science and Engineering at Rajeev Institute of Technology, Hassan, India. Her areas of interest include networks, algorithms.