ISSN: 2454-4248
469 – 473

Literature Survey of Big Data

Lalit Gandhi,
Research Scholar, U.I.E.T,
M.D.UniversityRohtak, Haryana INDIA

Abstract: Traditional DBMS software packages are inadequate to handle voluminous data sets called as Big Data. Big data alludes to datasets that are big, as well as high in assortment and speed. Because of the quick development of such data, arrangements should be considered and given with a specific end goal to deal with. In this paper we have done literature survey of Big data, its characteristics, structure and architecture is discussed in detail. Then we have highlighted some of applications of Big data.

*****

I. Introduction

Traditional DBMS software packages are inadequate to handle voluminous data sets called as Big Data. Data capturing, storage and analysis are the main challenges among Big Data. Searching, sharing, transfer, visualization, querying, updating, information privacy is tedious in Big Data. Earlier three concepts – velocity, variety and volume were related with big data. Lately, “big data” referred use of predictive, user behavior analytics or other advanced data analytics methods that extract data. Scientists, medical practitioners, government agencies find it hard to handle large data. Data sets grow rapidly and internet of things devices are not capable enough to deal with them. IDC reported that volume of data is growing exponentially and by 2025 there will be 163 zettabytes of data. RDBMS and related packages cannot handle big data. ”Big Data” definition depends upon user and their tools for handling voluminous data e.g. some organizations my consider ten gigabytes of data as big whereas for others terabytes of data size becomes significant consideration.

II. BIG Data

Big data term is in use since 1990. Big data is data set with sizes outside the capability of commonly used software tools. This is hard for commonly used DBMS software to encapsulate, curate, achieve, and treat data within manageable time. Big data contains unstructured, semi-structured and structured data. Importance of unstructured data is high. Big data "size" is varying since evaluation of the term. “Size” definition started from few gigabytes and now leading to terabytes to zetabytes. Big data need tools and techniques to fetch, control and manipulate.

Big data is the information assetwith high volume, velocity and variety, which need specific skill and logical approaches for its transformation. V*eracity*, is added by some organizations to describe reality. In another form “Big data” is where parallel computing tools are needed to handle data. There exist considerable difference between "big data" and "Business Intelligence".

- Business Intelligence uses descriptive statistics whereas Big Data uses inductive statistics.
- Business Intelligence utilizes data with high information density for measuringthings, detect trends, etc. in contrast with “Big Data” that uses data with low information density for depicting behaviors.

III. Different Forms of Big data:

a) Structured Data:

Structured data means data in the organized format. When we can map the collected data ware house in relational database format. RDBMS structure is enforced on current big data, so we know mapping of columns and how they are associated with tables and corresponding table space. Complete data is represented in the form of Relations or classes, Attributes, Schemas. Complete data is organized and follow the same order, all of them have same format and description associated with. Complete data follow the format and constrained to length defined.

b) Semi Structured Data:

Semi structured data, definition is partially clear and often thedistinguishing linksare not clear. There is no fixed schema defined for the semi structured data, tags are added with data, to associate the same with organizational structure. This helps to analyze and organize the data. The concept can be thought of as XML instead of HTML. Data can be available in the form of database system and file system. Data structure is partially organized and not completely. Grouping of similar data is done, but may not have same attributes for mapping.

c) Unstructured Data:

This data is not organized and formats cannot be indexed easily. Indexing in terms of structured data and semi structured (partially) data is done as mapping is done using relational database. Audio, Video and image files are associated with this data format Data is of any type, and no structure is defined, no sequence, no constraints, no regulations and randomness is spread across data.

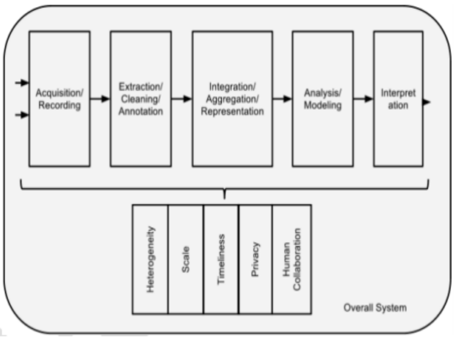
Big data is an important concept as it helps the organizations to improve their key performance indexes, it is not only with quantity of data but quality of data is also considered. Collection of data is done from multiple sources and integrated to analyze. It helps in reduction of time and cost. Analysis of data can guide for market trends and new product developments along with strategy development &decision making. Big data is also associated with analytics of business related tasks and a lot of tasks can be done using big data analytics like – Root Cause analysis of failures, issues and defects along with risk analysis. It can also help to detect and analyze fraud.

IV. Big Data Characteristics:

Big data concept can be defined through 3V model represents volume, velocity and variety of information.

Based on its properties the definition of big data modified by Gartner as high volume, high velocity and high variety of information that need new forms of processing required for decision making and optimization. Three-V model can be extended to Four-V and Five-V by adding value added and veracity. We can without much of a stretch express that these models, give a clear and every single acknowledged definition identified with what all is joined in a big data based application, arrangement, issue and system.
Volume: The amount of created and put away data. The extent of the data decides the esteem and potential knowledge and whether it can be viewed as big data or not.
Variety: The sort and nature of the data. This helps individuals who investigate it to viably utilize the subsequent understanding. Big data draws from content, pictures, sound, video; in addition to it completes missing pieces through data combination.
Velocity: In this specific situation, the speed at which the data is created and handled to address the demands and difficulties that lie in the way of development and improvement. Big data is regularly accessible continuously.
Variability: Irregularity of the data set can hamper procedures to handle and oversee it.
Veracity: The data nature of collected data can change significantly, influencing the precise investigation.

V. Pipeline Architecture of Big Data:



Data Acquisition:
There are multiple data sources for data collection and recording. Since huge data is collected so it is possible that some part of the data may be of no use. Different sorting and filtering techniques can be used to select the useful data. Sorting and filtering techniques must be of such types that no important information is left. Other important concept is creation of metadata for the fitment of recorded data. Recording the data about the information toward the beginisnot helpful since in the pipeline, it will continue changing and translated in various ways while being brought through the information investigation stages.

Extraction and Cleaning:
Mere collection of data does not resolve the purpose. Data is collected and it is not in the format that can be analyzed. To perform the same we need to process the data and extract the useful information. Pulling out the best useful information that can be used for decision-making and getting insight from the same. Refining the data again and again to get the useful information in continuous manner is among the technical challenges of the big data. Data cleaning is another main key area that considers number of constraints to clean data.

Data Integration, Aggregation, and Representation:
Once data is collected from different sources and useful information is extracted, cleaning is done, then it becomes necessary to integrate the data collected from different sources. Data is integrated for the purpose of usefulness. Aggregation of the different chunks is done to get the data to be used further for meaningful purpose. Then data is represented in the useful format for further processing. Here also the techniques required are of main concern as integration of different formats and representation of the same is not easy.

Data Modeling, Analysis and Query Process:
Big data is generally unique in relation to conventional database administration frameworks and thus the strategies for questioning and mining in Big Data are not the same as

customary measurable systems, which will be distinctive on Big and little examples. When we discuss Big Data, we frequently connect it with being dynamic, between connections, conniving and uproarious rather than the way toward mining, which will require spotless, reliable, coordinated, efficient, accessible data, which can be gotten to through mining interfaces utilizing decisive inquiries, adaptable mining calculations and figuring situations for big data. When we are discussing the already specified subjects, we can without a lot of an extend express that the data mining itself can be used to help with trust rest of data too upgrading nature of the same, understanding the related semantics of the data and give shrewd and clear addressing limits.

Interpretation:
We should comprehend that analysis is of limited regard, if customer cannot comprehend the same with reference to Big Data. Notwithstanding whether examination is done, each one of the reports and outlines are made, in spite of all that one needs to sit and translate the same. The illustration cannot happen, while sitting alone in a work area territory or vacuum, as the individual looking at the reports and outlines needs to manage the doubts that were used while making the examination and following the methods.

VI. Techniques and Technologies:

Following are the techniques and technologies utilized for Big Data.

Techniques:

Techniques for analyzing data, such as A/B testing, machine learning and processing are used. Big data technologies, like business intelligence, cloud computing and databases are used. Visualization uses charts, graphs and other displays of the data. Some of the techniques described as:

Data mining: Data-driven decision-making and its described as “searching or ‘digging into’ data file to extract the required information to understand better.

Cluster analysis: This is type of data mining, splitting the larger groups into smaller groups of similar objects where the properties of objects known in advance.

Crowd sourcing: This technique is used for data collection rather than analyzing.

Machine learning: This is evaluation of computers based on the empirical data and machine learning techniques are used to interpret from the existing database.

Text analytics: Major part of data is collected in text form. Transforming text from unstructured data to meaningful form is text analytics.

Association rule learning: “Association Rules” for larger databases to find the association between different facts and data.

Technologies:

There are number of software products and technologies that are available to facilitate big data analytics.

EDW: EDWs are the Enterprise data warehouses that are used for data analysis.

Visualization products: Big question with big data is how to represent the results. Many representation products fulfill this requirement, and can represent the data. Representation can also help in getting the information as a result.

MapReduce:This technique processes big data which uses distributed computing. As its name suggests two main jobs are achieved using the same, firstly Map and then Reduce. Map takes an arrangement of data and proselytes the same into another arrangement of data. Singular components are broken into tuples. Reduce undertaking which takes contribution from Mapped yield data and diminish the same into little and granular rows of relational data called tuples. Hadoop: It is an open-source framework which can store and process big data in a distributed environments. Hadoop is an apache based software framework, which is resulted from MapReduce and Big Table. Hadoop can utilize clusters of computers with simple programming models.

NoSQL databases: NoSQL database, valuable for extensive arrangements of appropriated data. NoSQL is particularly valuable when a venture needs to get to and examine colossal measures of unstructured data or data that is put away remotely on various virtual servers in the cloud. The most famous NoSQL database is Apache Cassandra. Other NoSQL incorporate SimpleDB, Google BigTable, Apache Hadoop, MapReduce, MemcacheDB, and Voldemort.

VII. Applications of Big Data:

1. Government: The utilization of big data inside administrative procedures builds proficiency as far as cost, profitability and advancement. Numerous administration associations can work by and large to convey the coveted yield.
2. International development: Big data innovation can make vital commitments yet in addition exhibit remarkable difficulties to International improvement. Progressions in big data examination offer practical chances to enhance basic leadership in basic advancement regions, for example, social security, work, financial efficiency, wrongdoing, security, and catastrophic event and asset administration.
3. Manufacturing: Big data gives a foundation to straightforwardness in manufacturing industry, which is the capacity to unwind vulnerabilities, for example, conflicting part execution and accessibility. Prescient

assembling as an appropriate approach toward close to zero downtime and straightforwardness requires tremendous measure of data and propelled forecast instruments for a methodical procedure of data into helpful data.

4. Healthcare: Big data investigation has helped human health care enhance by giving customized drug and prescriptive examination, clinical hazard mediation and prescient investigation, waste and care inconstancy diminishment, mechanized outside and inside detailing of patient data, institutionalized medicinal terms and patient registries and divided point arrangements. A few regions of change are more optimistic than implemented.
5. Media: The Media business gives off an impression of being moving far from the conventional approach of utilizing particular media conditions, for example, daily papers, magazines, or TV programs and rather takes advantage of customers with innovations that compass focused on individuals at ideal circumstances in ideal areas. A definitive point is to serve or pass on, a message or substance that is (factually) in accordance with the customer's outlook.
6. Internet of Things: Big data and the IoT work in conjunction. Data extricated from IoT gadgets gives a mapping of gadget interconnectivity. Such mappings have been utilized by the media business, organizations and governments to all the more precisely focus on their crowd and increment media effectiveness. IoT is additionally progressively received as a methods for social occasion tactile data, and this tangible data has been utilized as a part of restorative and assembling settings.
7. Information technology: Big data has come to unmistakable quality inside Business Operations as an instrument to enable representatives to work all the more productively and streamline the accumulation and circulation of Information Technology (IT). The utilization of big data to determine IT and data accumulation issues inside an endeavor is called IT Operations Analytics (ITOA).[89] By applying big data standards into the ideas of machine knowledge and profound figuring, IT divisions can foresee potential issues and move to give arrangements before the issues even happen.

VIII. Conclusion:

This paper presented the basic understanding of big data, its characteristics and architecture. We have also looked in to the application areas. Also different tools and techniques that are required for big data analysis and presentations are discussed. There is a lot of area in the research of big data that we can work upon as future era is of big data.

IX. References:

[1] Johnston, Casey (6 April 2012). "Google Trends reveals clues about the mentality of richer nations". ArsTechnica. Retrieved 9 April 2012.

[2] Tobias Preis (24 May 2012). "Supplementary Information: The Future Orientation Index is available for download" (PDF). Retrieved 24 May 2012.

[3] Philip Ball (26 April 2013). "Counting Google searches predicts market movements". Nature. Retrieved 9 August 2013.

[4] Tobias Preis, Helen Susannah Moat and H. Eugene Stanley (2013). "Quantifying Trading Behavior in Financial Markets Using Google Trends". Scientific Reports.3: 1684. doi:10.1038/srep01684. PMC 3635219. PMID 23619126 .

[5] Nick Bilton (26 April 2013). "Google Search Terms Can Predict Stock Market, Study Finds". New York Times. Retrieved 9 August 2013.

[6] Christopher Matthews (26 April 2013). "Trouble With Your Investment Portfolio? Google It!". TIME Magazine. Retrieved 9 August 2013.

[7] Philip Ball (26 April 2013). "Counting Google searches predicts market movements". Nature. Retrieved 9 August 2013.

[8] Bernhard Warner (25 April 2013). "'Big Data' Researchers Turn to Google to Beat the Markets". Bloomberg Businessweek. Retrieved 9 August 2013.

[9] Hamish McRae (28 April 2013). "Hamish McRae: Need a valuable handle on investor sentiment? Google it". The Independent. London. Retrieved 9 August 2013.

[10] Richard Waters (25 April 2013). "Google search proves to be new word in stock market prediction". Financial Times. Retrieved 9 August 2013.

[11] David Leinweber (26 April 2013). "Big Data Gets Bigger: Now Google Trends Can Predict The Market". Forbes. Retrieved 9 August 2013.

[12] Jason Palmer (25 April 2013). "Google searches predict market moves". BBC. Retrieved 9 August 2013.

[13] E. Sejdić, "Adapt current tools for use with big data," Nature,vol. vol. 507, no. 7492, pp. 306, Mar. 2014.

[14] Stanford. "MMDS. Workshop on Algorithms for Modern Massive Data Sets".

[15] DeepanPalguna; Vikas Joshi; VenkatesanChakaravarthy; Ravi Kothari & L. V. Subramaniam (2015). Analysis of Sampling Algorithms for Twitter. International Joint Conference on Artificial Intelligence.

[16] Kimble, C.; Milolidakis, G. (2015). "Big Data and Business Intelligence: Debunking the Myths". Global Business and Organizational Excellence. 35 (1): 23–34. doi:10.1002/joe.21642.

[17] Chris Anderson (23 June 2008). "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". WIRED.

[18] Graham M. (9 March 2012). "Big data and the end of theory?". The Guardian. London.

[19] "Good Data Won't Guarantee Good Decisions. Harvard Business Review". Shah, Shvetank; Horne, Andrew; Capellá, Jaime;. HBR.org. Retrieved 8 September 2012.

[20] Big Data requires Big Visions for Big Change., Hilbert, M. (2014). London: TEDxUCL, x=independently organized TED talks

[21] Alemany Oliver, Mathieu; Vayre, Jean-Sebastien (2015). "Big Data and the Future of Knowledge Production in Marketing Research: Ethics, Digital Traces, and Abductive Reasoning". Journal of Marketing Analytics. 3 (1): 5–13. doi:10.1057/jma.2015.1.

[22] Jonathan Rauch (1 April 2002). "Seeing Around Corners". The Atlantic.

[23] Epstein, J. M., & Axtell, R. L. (1996). Growing Artificial Societies: Social Science from the Bottom Up. A Bradford Book.

[24] "Delort P., Big data in Biosciences, Big Data Paris, 2012"(PDF). Bigdataparis.com. Retrieved 8 October 2017.

[25] "Next-generation genomics: an integrative approach" (PDF). nature. July 2010. Retrieved 18 October 2016.

[26] "BIG DATA IN BIOSCIENCES". ResearchGate. October 2015. Retrieved 18 October 2016.

[27] "Big data: are we making a big mistake?". Financial Times. 28 March 2014. Retrieved 20 October 2016.

[28] Ohm, Paul. "Don't Build a Database of Ruin". Harvard Business Review.

[29] Darwin Bond-Graham, Iron Cagebook – The Logical End of Facebook's Patents, Counterpunch.org, 2013.12.03

[30] Darwin Bond-Graham, Inside the Tech industry's Startup Conference, Counterpunch.org, 2013.09.11

[31] Al-Rodhan, Nayef (16 September 2014). "The Social Contract 2.0: Big Data and the Need to Guarantee Privacy and Civil Liberties – Harvard International Review". Harvard International Review. Retrieved 3 April 2017.

[32] Barocas, Solon; Nissenbaum, Helen; Lane, Julia; Stodden, Victoria; Bender, Stefan; Nissenbaum, Helen (June 2014). Big Data's End Run around Anonymity and Consent. Cambridge University Press. pp. 44–75. doi:10.1017/cbo9781107590205.004. ISBN 9781107067356.

[33] Lugmayr, Artur; Stockleben, Bjoern; Scheib, Christoph; Mailaparampil, Mathew; Mesia, Noora; Ranta, Hannu; Lab, Emmi (2016-06-01). "A COMPREHENSIVE SURVEY ON BIG-DATA RESEARCH AND ITS IMPLICATIONS – WHAT IS REALLY 'NEW' IN BIG DATA? -IT'S COGNITIVE BIG DATA!".

[34] danahboyd (29 April 2010). "Privacy and Publicity in the Context of Big Data". WWW 2010 conference. Retrieved 18 April 2011.

[35] Jones, MB; Schildhauer, MP; Reichman, OJ; Bowers, S (2006). "The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere" (PDF). Annual Review of Ecology, Evolution, and Systematics. 37 (1): 519–544. doi:10.1146/annurev.ecolsys.37.091305.110031.

[36] Boyd, D.; Crawford, K. (2012). "Critical Questions for Big Data". Information, Communication & Society. 15 (5): 662–679. doi:10.1080/1369118X.2012.678878.

[37] Failure to Launch: From Big Data to Big Decisions, Forte Wares.

[38] Gregory Piatetsky (12 August 2014). "Interview: Michael Berthold, KNIME Founder, on Research, Creativity, Big Data, and Privacy, Part 2". KDnuggets. Retrieved 13 August 2014.

[39] Pelt, Mason. ""Big Data" is an over used buzzword and this Twitter bot proves it". siliconangle.com. SiliconANGLE. Retrieved 4 November 2015.

[40] Harford, Tim (28 March 2014). "Big data: are we making a big mistake?". Financial Times. Financial Times. Retrieved 7 April 2014.

[41] Ioannidis, J. P. A. (2005). "Why Most Published Research Findings Are False". PLoS Medicine. 2 (8): e124. doi:10.1371/journal.pmed.0020124. PMC 1182327. PMID 16060722.

[42] Lohr, Steve; Singer, Natasha (10 November 2016). "How Data Failed Us in Calling an Election". The New York Times. ISSN 0362-4331. Retrieved 27 November 2016.

[43] Markman, Jon. "Big Data And The 2016 Election". Forbes. Retrieved 27 November 2016.