A Survey on Big Data and its Charactristics

Prof. A. Angel preethi
Assistant Professor
St. Joseph's college,
Trichy -2

Dr. B. Vani
Assistant Professor
Srimad Andavar Arts and Science College
Trichy -2

Abstract:- **Every day we come across huge amounts of data that are come from internet users, social network sites, mobile phones, share markets. So these data cannot be handled using traditional data base management system. These kinds of data are called big data. Big data are data whose scale, distribution, diversity, and timeliness require the use of new procedural architectures and analytics to enable insights that disengage new sources of business value. In this paper discusses about the big data, characteristics of big data which contains 5 V's namely volume, velocity, veracity, variability and value. The characteristics of big data Hadoop, map reduce and hadoop distributed frame work are also discussed. This survey is useful for the researchers those who are going to start the research in big data.**

Key words: Big Data, Hadoop, Map Reduce, HDFS (Hadoop Distributed File System)

INTRODUCTION:

Today, data is more extremely rush into the structure of our lives than ever before. We desire to use data to solve problems, get better safety, and produce profitable affluence. The collection, storage, and analysis of data is on an rising and apparently abundant trail, fueled by increases in processing power, the crate ring costs of calculation and storage, and the growing number of sensor technologies embedded in devices of all kinds. In 2011, some estimated the amount of information created and replicated would surpass 1.8 zeta bytes, 4 in 2013; estimates reached 4 zeta bytes of data generated worldwide.

Every day we create 2.5 quintillion bytes of data and so much that 90% of the data in the world today has been created from last two years. This data comes from social media sites and networks like face book, twitter, sensor used to gather climate information, digital images, videos, scientific instruments, bank and credit card transactions, grocery stores, mobile devices, online acquisitioning, transaction reports and cell phone Global positioning system signals. For example flicker, a public picture sharing site, where in an average of 1.8 billion photos per day is received from February to march 2012 along with more than 200 hours of video every minute [1]. Every minute we upload 100 hours of video on YouTube. In addition, every minute over 200 million emails are sent, around 20 million photos are viewed almost 300,000 tweets are sent and almost 2 to 5 million queries on Google are performed. This shows that it is very difficult to process the data [3]. These are all huge volumes of data in the world. To manage, analyze, summarize, visualize, and discover knowledge from these huge amounts of data and extract value and knowledge from this data in a timely manner and a scalable fashion. This large amount of data is known as big data.

The model of generating data and consuming data has changed. In older days few companies are generating data and all others are consuming that data. But, now a day's all of us are generating and all of us are using that data. Big data are real time in nature than traditional data ware house applications. Big data produces large volumes of data sets typically stemming from more than one data source and being processed by data analyzer or processor.

Big data describes any capacious amount of prepared data, partially prepared data and unprepared data that has the possibility to be mined for information. Relational data base management systems (RDBMS) and desktop processor are often having difficulty in processing the big data. It requires amazingly equivalent software it running on tens, hundreds, or thousand's of servers. The big data may vary depends upon the software tools.

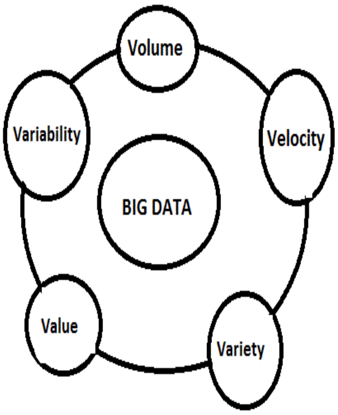
Simply we can state that big data are "large volumes of structured and unstructured data which cannot be handled by standard database management systems, Relational Database Management Systems, Object relational database management systems". The size of the Big data are moving from many dozen terabytes to many peta bytes.

Gartner et al [4] defined "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization". According to Wikimedia, "In information technology, Big data are a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools" [5]

The goal of this paper is to provide the overall view of the big data and its characteristics.

CHARACTERISTICS OF BIG DATA:

It is usually acknowledged that big data can be explained according to three V's: Velocity, Variety and Volume. In a 2001 research report, META Group analyst Doug Laney defined big data as being three-dimensional, increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Later in 2012 Gartner updated the definition of big data as high volume, high velocity, and high variety [3]. But generally we can define there are 5v's in big data.



.

Figure 1 5 V's of big data.

Volume:

Generally volume is the quantity of data. In big data the Data volume is increasing exponentially. The size of the data determines the value and latent of the data and whether it can be considered big data or not. The volume of data collected and processed is unprecedented. This explosion of data from web-enabled appliances, wearable technology, and advanced sensors to monitor everything from vital signs to energy use to a jogger's running speed will drive demand for high-performance computing and push the capabilities of even the most sophisticated data management technologies.

For example Dalvi et al [2] collected and analyze the nature and characteristics of data where it comes from many data sources. The study includes schools, hotels, libraries, automotive and restaurants and they got the result as each domain has tens and thousands of sources on the web. This is higher than the number of data sources considered in conventional data integration.

Variety:

The data has many forms like unstructured, Quasi structured, semi structured, Structured, text and multimedia. Unstructured data is data that has no inherent structures and is usually stored as different type of files.

Quasi structured is textual data with erratic data formats, can be formatted with effort, tools and time for example web click stream data that may contain some inconsistencies in data values and formats.

Semi structured is Textual data files with a discernable pattern, enabling, parsing for example XML data files that are self describing and defined by an XML schema. Structured data are containing a defined data type and format. For example transaction data and OLAP.

Figure 1 represents the five V's of big data that consists of volume, velocity, variety, variability, and value.

Li et al [5] studied a stock market domain containing 55 sources. 153 global attributes are manually matched from 333 local attributes the number of providers for each global attribute observes Zipf's law, with 13.7% of the attributes provided by at least one third of the sources and over 86% of attributes provided by fewer than 25% of the sources.

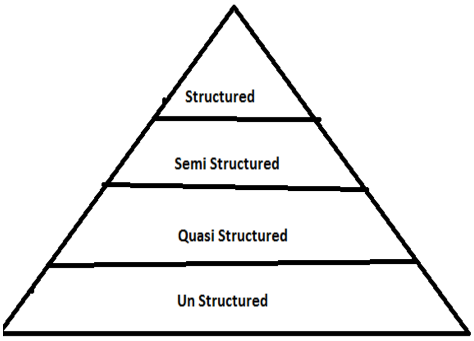


Figure 2. Structure of data

Figure 2 represents the structure of the data and how it is arranged.

Velocity:

The data in motion is called as velocity of data. Data is being collected continually to make available many of the data sources.

For example many of the data sources that provide continually changing information like stock market, gold rate, weather report, volume of share traded report these are all continually changing data sources.

Variability:

Data sources are of slightly different in the quality and coverage, accuracy and timeliness of data provided.

For example, the work by Dalvi et al. [6] showed that with strong head aggregators such as yelp.com, collecting homepage URLs for 70% restaurants that are mentioned by some websites required only 10 sources; however, collecting URLs for 90% restaurants required 1000 sources, and collecting URLs for 95% restaurants required 5000 sources. Similarly, the work by Li et al. [7] showed that even in the stock market domain, inconsistent values were provided by different sources for over 80% of the data items whose values should be fairly stable. This is consistent with the belief that "1 in 3 business leaders do not trust the information they use to make decisions."

Value:

Value is one of the important aspects of big data. Value is frequently shown as the fourth leg of the Big Data chair; Value does not discriminate Big Data from not so big data. It is equally true of both big and little data that if we are making the attempt to accumulate and examine it then it must be perceived to have value.

When making an effort to realize the concept of Big Data, the words Map Reduce and "Hadoop" cannot be avoided

Hadoop:

Hadoop is a distributed file environment. It is a free java based programming frame work. Hadoop supports the processing of large sets of data in a distributed computing environment.

It is a part of the apache project supported by the Apache software foundation. Hadoop cluster uses the master slave structure [7]. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures [13].

Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data. Hadoop has two main sub projects – Map Reduce and Hadoop Distributed File System (HDFS).

Map Reduce:

Map Reduce is also called infrastructure or frame work. Hadoop Map Reduce is a framework [8] used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. The data will be first divided into individual chunks. After that the chunks are processed by Map jobs in parallel. The map reduce framework sorted the output. This output is the input to the reduce task. The file system stored both input and the output of the job.

HADOOP DISTRIBUTED FILE SYSTEM (HDFS):

HDFS [9] is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures.

BIG DATA APPLICATIONS:

Data exploration and analysis turned into a difficult problem in many sectors in the span of big data. With large and complex data, computation becomes difficult to be handled by the traditional data processing applications which triggers the development of big data Applications [10]. Google's map reduces frame work and Apache hadoop are the software systems for big data applications [11]. The applications generates huge amount of intermediate data. There are two main applications in big data that are bio informatics and manufacturing. Bioinformatics [12] requires a large scale data analysis that uses Hadoop. The combination of sensory data and historical data constructs the big data in manufacturing.

CONCLUSION:

Every day twitter process 7 tera bytes of data and face book process 9 tera bytes of data. Not only social sites, the data which come from many data sources which are not handled by traditional data base tools is called as Big data. This data has high volume, velocity and variability. Big data are widely used in health care and Bio informatics. So this survey is useful for researchers those who are starting a career in big data.

REFERENCES:

1. F. Michel, "How Many Photos Are Uploaded to Flicker Every Day and Month?"http://www.flickr.com/photos/franckmichel/6855169886/, 2012.
2. N. N. Dalvi, A. Machanavajjhala, and B. Pang. "An analysis of structured data on the web". *PVLDB*, 5(7):680–691, 2012.
3. https://datafloq.com/read/3vs-sufficient-describe-big-data/166.
4. Douglas and Laney, "The importance of 'big data': A definition,"2008.
5. X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava. "Truth finding on the deep web: Is the problem solved? *PVLDB"*, 6(2), 2013.
6. Wie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.
7. K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.
8. F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications." *Big Data, 2013 IEEE International Conference*, Silicon Valley, CA, Oct 6-9, 2013, pp. 32 - 37.
9. Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for big-data applications using the MapReduce framework." *INFOCOM, 2013 Proceedings IEEE*, Turin, Apr 14-19, 2013, pp. 35 - 39.
10. Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." *Open Cirrus Summit (OCS), 2012 Seventh*, Beijing, Jun 19-20, 2012, pp. 48 - 52.
11. Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.
12. Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. Final Roadmap, March 2012. [online] http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf
13. Y.Demchenko, Z.Zhao, P.Grosso, A.Wibisono, C. de Laat, Addressing Big Data Challenges for Scientific Data Infrastructure. The 4th IEEE Conf. on Cloud Computing Technologies and Science (CloudCom2012), 3 - 6 December 2012, Taipei, Taiwan. ISBN: 978-1-4673-4509-5
14. E.Dumbill, What is big data? An introduction to the big data landscape. [online] http://strata.oreilly.com/2012/01/what-is-big-data.html
15. What is big data? IBM. [online] http://www.01.ibm.com/software/data/bigdata/
16. What is big data? [online] http://www.gartner.com/it-glossary/big-data
17. Roundup of Big Data Pundits' Predictions for 2013. Blog post by David Pittman. January 18, 2013. [online] http://www.ibmbigdatahub.com/blog/roundup-big-data-pundits-predictions-2013
18. Big Data prediction for 2013. Blog by Mike Gualtieri. [online] http://blogs.forrester.com/mike_gualtieri
19. The Big Data Long Tail. Blog post by Jason Bloomberg on Jan 17, 2013. [online] http://www.devx.com/blog/the-big-data-long-tail.html

20. The 3Vs that define Big Data. Posted by Diya Soubra on July 5, 2012 [online] http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data
21. Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. October 2010. [online] Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
22. Y.Demchenko, Z.Zhao, P.Grosso, A.Wibisono, C. de Laat, Addressing Big Data Challenges for Scientific Data Infrastructure. The 4th IEEE Conf. on Cloud Computing Technologies and Science (CloudCom2012), 3 - 6 December 2012, Taipei, Taiwan. ISBN: 978-1-4673-4509-5
23. Reflections on Big Data, Data Science and Related Subjects. Blog by Irving Wladawsky-Berger. [online] http://blog.irvingwb.com/blog/2013/01/reflections-on-big-data-data-science-and-related-subjects.html

B. Vani. B is working as Lecturer in Srimad Andavar Arts and Science college, Bharathidasan University, Trichy, Tamil Nadu,India. She has 16 years of experience in teaching and 4 years in research. Her area of research is wireless network security. Her research interests on Denial of Service attack on wireless network. She has published around 20 internal/ National papers. Other areas of interest include OOAD & UML, Software quality and Testing and Computer Networks.

A. Angelpreethi is working as Lecturer in the Department of Computer science,St.Joseph'scollege(Autonomous), Tiruchirappalli, tamilnadu,India. She received her master of philosophy degree from St. joseph's college (Autonomous), Tiruchirappalli. Her current area of research is Data mining, Big data, Cloud computing.