International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume 2, Issue 11 (November '2016)

SURVEY ON BIG DATA TOOLS AND TECHNIQUES

K. Syed Mohamed Bukari,
Assistant Professor,
P.G. Department of Computer Science,
The New College,
Chennai, India.

Abstract: Big Data is the term used to identify the datasets that are large in size and complexity. Big Data are now swiftly expanding in domains. Big Data includes e-mail messages, snaps, business deals, surveillance video recordings, posts of social Medias, mobile phone GPS signals and RFID readers, microphones, cameras, sensors and activity logs. That's why, the data that exceeds the processing capacity of conventional database systems. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity. The Big Data is a challenge of the most exciting opportunities for the next era. Any Big Data is for decision making. This study paper includes the information about Introduction about Big Data, Challenges in Big Data, Tools and Techniques to handle Big Data, and its Applications.

Keywords: Big Data, Challenges, Datasets, Heterogeneous, Hadoop, MapReduce, HDFS, Pig, HBase, Hive, Avro, Oozie, Chukwa, ZooKeeper

I. INTRODUCTION

Big Data means Business Intelligence Data. It means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. In the beginning, Internet is used for only the purpose of sending e-mails messages, browsing queries. But recently internets are used for more purpose. So the internet generates very huge amount of data at every seconds and it will increased day by day. So such type of data is difficult to process because it contains the billons records of million people. Information that includes the web sale, social networks, online transactions, e-mails, audios, videos, images, sensor data, science research, health records, search queries, click streams, posts, cloud computing, mobile phones and their applications. Under the explosive increase of global data, the term of big data is mainly used to describe enormous datasets. Compared with traditional datasets, big data typically includes masses of unstructured data that need more real-time analysis. In addition, big data also brings about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and also incurs new challenges, e.g., how to effectively organize and manage such datasets. The rampant growth of connected technologies creates even more acceleration in Big Data trends – it is a "DATA TSUNAMI"

Umbrella of Big Data:

Big data involves the data produced by different devices and applications. Given below are some of the areas that come under the umbrella of Big Data.

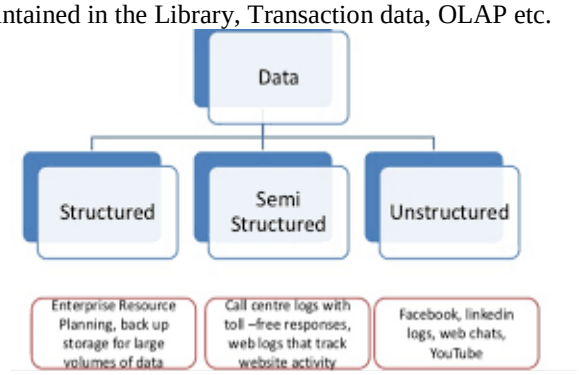
- **Black Box Data**: It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data**: Social media such as Facebook, WhatsApp and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data**: The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Transport Data**: Transport data includes model, capacity, distance and availability of a vehicle.
- **Power Grid Data**: The power grid data holds information consumed by a particular node with respect to a base station.
- **Search Engine Data**: Search engines retrieve lots of data from different databases.

There are three types of data

Unstructured Data: Any Web page data is unstructured. Example: Clustered data (text file, video file, audio file etc.)

Semi Structured Data: Something is mandatory and others are optional. For example, sending E – Mail to address is mandatory and attaching files and subject are optional.XML files are comes under this category of data.

Structured Data: It is well organized. Examples: Data maintained in the Library, Transaction data, OLAP etc.



II. CHALLENGES IN BIG DATA

The need of big data generated from the large companies like YouTube, Facebook, Google, yahoo, etc. for the purpose of analysis of enormous amount of data which is in structured form or even in unstructured form. Google contains the large amount of information. So, there is the need of Big Data Analytics that is the processing of the complex and massive datasets. Data analysis is to exhibit the hidden patterns in the data. This data is different from structured data in terms of five parameters: Variety, Volume, Value, Veracity and Velocity (5V's) are the challenges of big data management are:

1. **Volume:** Data is ever-growing day by day of all types ever Kilo Byte, Mega Byte, Tera Byte, Peta Byte, Exa Byte, Zetta Byte, Yotta Byte of information. The data results into massive files. Excessive volume of information is main issues of storage. This main issue is resolved by reducing storage value. Data volumes are expected to grow more than 50 times by 2020.

2. **Velocity:** The data comes at high speed. Sometimes one minute is too late, so big data is time sensitive. Most organisations data velocity is main challenge. The credit card transactions and social media messages done in millisecond and data generated by this putting in to databases.

3. **Variety:** Data sources are extremely heterogeneous. The files comes in various formats and of any type, it may be unstructured or structured such as text, audio, log files, videos and more. The varieties are endless, and the data enters the network without having been quantified or qualified in any way.

.

4. **Veracity:** The increase in the range of values typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data. Big data and analytics technologies work with these types of data.

5. **Value:** It is a most important V in big data. Value is main buzz for big data because it is important for IT infrastructure system, businesses to store large amount of values in database.

III. TOOLS AND TECHNIQUES

Hadoop: Hadoop is an open source project of the Apache Foundation that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is a specially designed file system with commodity hardware for processing large data sets. But not recommended for small data sets. It is a framework written in java originally developed by Doug Cutting who named it after his son's toy elephant. Hadoop uses Google's Map Reduce and Google File System technologies as its foundation. It is optimised to handle massive quantities of data which could be structured, unstructured or semi structured, using commodity hardware, that is, relatively inexpensive computers.

HDFS:

HDFS is a block-structured distributed file system soring very large files that holds the large amount of Big Data. In the HDFS, the data is stored in blocks that are known as chunks. HDFS stores three copies of each file by copying each block to three different servers. Size of each block is 64MB. HDFS architecture is broadly divided into following three nodes which are Name Node, Data Node and Secondary Node.

1. **Name Node:** It is centrally placed node, which contains information about Hadoop file system. The main task of name node is that it records all the metadata & attributes and specific locations of files & data blocks in the data nodes. Name node acts as the master node as it stores all the information about the system .and provides information which is newly added, modified and removed from data nodes.

2. **Data Node:** It works as slave node. Hadoop environment may contain more than one data nodes based on capacity and performance. A data node performs two main tasks storing a block in HDFS and acts as the platform for running jobs.

3. **Secondary Node:** When Name Node fails the Hadoop doesn't support automatic recovery, but the configuration of secondary node is possible.

MapReduce:

MapReduce is a programming framework for distributed computing which was created by Google using the divide and conquer method to break down complex big data problems into small units of work and process them in parallel. MapReduce can be divided into two stages:

a) Map Step: The master node data is chopped up into many smaller sub problems. A worker node processes some subset of the smaller problems under the control of the Job Tracker node and stores the result in the local file system where a reducer is able to access it.

b) Reduce Step: This step analyses and merges input data from the map steps. There can be multiple reduce tasks to parallelize the aggregation, and these tasks are executed on the worker nodes under the control of the Job Tracker.

Components of Hadoop:

- **HDFS:** A highly fault tolerant distributed file system that is responsible for storing data on the clusters.
- **MapReduce:** A powerful parallel programming technique for distributed processing on clusters.
- **HBase:** A scalable, distributed database for random read/write access.
- **Pig:** A high level data processing system for analyzing data sets that occurs a high level language.
- **Hive:** A data warehousing application that provides a SQL like interface and relational model.
- **Sqoop:** A project for transferring data between relational databases and Hadoop.
- **Avro:** A system of data serialization.
- **Oozie:** A workflow for dependent Hadoop jobs.
- **Chukwa:** A Hadoop subproject as data accumulation system for monitoring distributed systems.
- **Flume:** A reliable and distributed streaming log collection.
- **ZooKeeper:** A centralized service for providing distributed synchronization and group services along with maintenance of the configuration information and records.

IV. APPLICATIONS

McKinsey Global Institute specified the potential of big data in five main topics:

Healthcare: clinical decision support systems, individual analytics applied for patient profile, personalized medicine, performance based pricing for personnel, analyze disease patterns, improve public health.

Public sector: creating transparency by accessible related data, discover needs, improve performance, customize actions for suitable products and services, decision making with automated systems to decrease risks, innovating new products and services.

Retail: In store behaviour analysis, variety and price optimization, product placement design, improve performance, labour inputs optimization, distribution and logistics optimization, and Web based markets.

Manufacturing: improved demand forecasting, supply chain planning, sales support, developed production operations, web search based applications.

Personal location data: smart routing, geo targeted advertising or emergency response, urban planning, new business models.

V. CONCLUSION

In this survey paper, an overview of big data's concept, challenges, tools, techniques and applications have been reviewed. The results have given away that regardless of the fact that accessible information, tools and techniques available in the literature, there are numerous focuses to be viewed as, discussed, analyzed, developed, improved and so on. Although this paper obviously has not resolved the complete subject about this substantial topic, emphatically it has provided some useful discussion and we can conclude that Hadoop is possibly one of the best solutions to maintain the Big Data.

VI. REFERENCES

[1]. Sagiroglu, Sinanc, "Big Data: A Review", 978-14673-6404-1/13 IEEE.
[2]. Sabia, Arora, "Technologies to Handle Big Data: A Survey".
[3]. Shilpa, Manjit Kaur, "Big Data and Methodology – A review", Volume 3, Issue 10, October 2013.
[4]. http://www.ibm.com/software/in/data/bigdata/ (last access - 10/11/2014).
[5]. K. Bakshi, "Considerations for Big Data: Architecture and Approach", Aerospace Conference IEEE, Big Sky Montana, March 2012.
[6]. D. Garlasu, V. Sandulescu, I. Halcu, G.Neculoiu, "A Big Data implementation based on Grid Computing", Grid Computing, 17-19 January 2013.
[7]. R.D. Schneider, "Hadoop for Dummies Special Edition", John Wiley & Sons Canada, 978-1-118-5051-8, 2012. [8] Yuri Demchenko, "The Big Data Architecture Framework (BDAF)", Outcome of the Brainstorming Session at the University of Amsterdam, 17 July 2013.
[8]. Definitions of Big Data - open tracker, http://www.pcmag.com/encyclopedia/term/62849/big-data
[9]. White paper – Big Data-as-a service, A Market & Technology Perspective-EMC Solution group
[10]. http://en.wikipedia.org/wiki/Big_data
[11]. http://hadoop.apache.org/
[12]. Rathod, Chauhan, "A Survey on Big Data Analysis Techniques" IJSRD - International Journal for Scientific Research & Development Vol. 1, Issue 9, 2013 ISSN (online): 2321-0613
[13]. http://iihtofficialblog.blogspot.com/2014/07/5-vs-of-hadoop-big- data.html
[14]. http://www.slideshare.net/shilpasoi/v3-i10-0415
[15]. http://experimentingbigdata.blogspot.com/2014_12_01_archive.html
[16]. A Survey Paper on Data Mining With Big Data RohitPitre Computer Engg. K.J.C.O.E. & M.R. Pune Vijay Kolekar Computer Engg. K.J.C.O.E. & M. R. Pune
[17]. X. Wu, C. Zhang, and S. Zhang, "Database Classification for Multi-Database Mining," Information Systems,vol. 30, no. 1, pp. 71- 88, 2005
[18]. K. Su, H. Huang, X. Wu, and S. Zhang, "A Logical Framework for Identifying Quality Knowledge from Different Data Sources," Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.
[19]. Apache Hadoop. Available at http://hadoop.apache.org
[20]. Apache HDFS. Available at http://hadoop.apache.org/hdfs
[21]. Apache Hive. Available at http://hive.apache.org
[22]. Apache HBase. Available at http://hbase.apache.org