2nd International Conference on Recent Innovations in Science, Engineering and Management
JNU Convention Center, Jawaharlal Nehru University, New Delhi (ICRISEM-15)
22 November 2015 www.conferenceworld.in ISBN: 978-81-931039-9-9

BIG DATA CHALLENGES AND PERSPECTIVES

Meenakshi Sharma[1], Keshav Kishore[2]

[1]Student of Master of Technology, [2]Head of Department,

Department of Computer Science and Engineering, A P Goyal Shimla University, (India)

ABSTRACT

In this paper we discuss the various challenges of Big Data. "Big Data" tries to solve new kind of problems which arises due to information explosion i.e. data growing at very high speed and is having very large volume. It works on the area of data Processing, data Storage and data Analytics. on day to day basis , there is a production of large amount as well as complex data in every sector . Traditional data base system was not able to handle growing, larger-sized datasets with high-velocity and different structures problems. Then Big Data came into existence. Big Data volume will grow by a factor of Exabyte's, representing a double growth every years. To handle massive amount of data different technologies used by Big data and most popular technologies are Hadoop, Map Reduce, HDFS, No SQL. These technologies handle massive amount of data in Megabyte, Gigabyte, Terabyte, Petabyte, and Exabyte with optimum speed.

Keywords: Big data, 5v's, Hadoop, HDFS, Map Reduce.

I. INTRODUCTION

Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools.**[**1**]** For instance, an International Data Corporation **(**IDC) report predicts that, from 2005 to 2020, the global data volume will grow by a factor of 300, from 130 Exabyte to 40,000 Exabyte, representing a double growth every two year.[2] IBM indicates that 2.5 Exabyte data is formed every day that is extremely tough to investigate. The estimation about the generated data is that till 2003 it was represented about 5 EB of data, then until 2012 is 2.7 ZB of data and till 2015 it is expected to increase 3 times.[3] The need of Big Data comes from the big Companies like yahoo, Google, facebook etc for the purpose of analysis of big amount of data which is in unstructured form. Traditional data management and analysis systems are based on the relational database management system (RDBMS). It is evident that the traditional RDBMS could not handle the huge volume and heterogeneity of Big Data. For solutions of permanent storage and management of large-scale disordered datasets, hadoop distributed file systems and NoSQL (Not Only SQL) Databases are good choices [7]

II. BIG DATA

The National Institute of Standards and Technology (NIST) suggests that, ``*Big Data* is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant horizontal scaling for efficient processing."[2]. *Big Data* is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or

analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful.[8]

Attributive Definition: It defines the four salient features of *Big Data*, i.e., volume, variety, velocity and value.

Comparative Definition**:** It defined *Big Data* as datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

III. EXAMPLES OF BIG DATA

Many firms such as e bay, Amazon so on has established a number of large-scale data warehouses to store their business data. "It estimates that the volume of business data worldwide, across all companies, doubles every 14 months" [10]

Another example is *Flicker*, picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012 .Assuming the size of each photo is 2 megabytes (MB), this resulted in 3.6 terabytes (TB) storage every single day. As "a picture is worth a thousand words", the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters etc.[1]

IV. BIG DATA PARAMETERS

Data Volume**:** It refers to the quantity of data that is being manipulated and analyzed in order to obtain the desired results.[5] Data is ever-growing day by day of all types ever Kilobyte Megabyte ,Gigabyte, Terabyte, Petabyte , Exabyte.,[3]

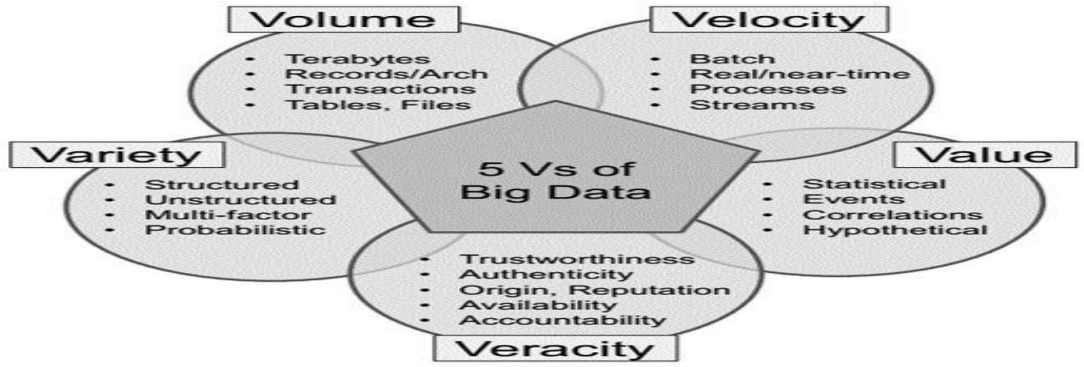

Fig. 1 Parameters of Big Data

Data Variety**:** It represents the type of data that is stored, analyzed and used. 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.[5]

Data Velocity**:** The data comes at high speed. Sometimes one minute is too late so *Big Data* is time sensitive [6]

Data Value : Which addresses the need for valuation of enterprise data?. The data we are working with is valuable for use or not.[1]

Data 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.

V. BIG DATA FRAMEWORK AND TECHNOLOGIES

Big Data differs from the traditional data and cannot be stored in a single machine. Furthermore, *Big Data* lacks the structure of traditional data. For the purpose of processing the large amount of data, the *Big data* requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the *Big Data*. E.g Hadoop, Mapreduce, HDFS**[4]**

5.1 Hadoop

In 2002, Dough Cutting develop an open source web crawler project, the Google published map reduced into 2006. Dough Cutting developed the open source, map reduced and HDFS.[9]

Hadoop is a Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop was inspired by Google's MapReduce Programming paradigm [6]

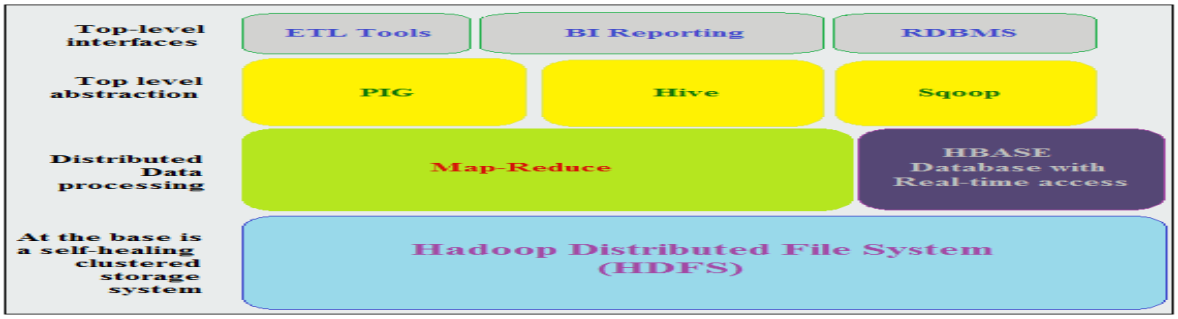

Fig. 2 Architecture of Hadoop

5.2 HDFS (Hadoop Distributed File System)

It is applied when the amount of data is too much for a single machine. The role of HDFS is to split data into smaller blocks and distribute it throughout the cluster. The name node stores the metadata for the Name Node .Name Nodes keeps track of the state of the Data Nodes.. Name Node is also responsible for the file system operations etc.[8]

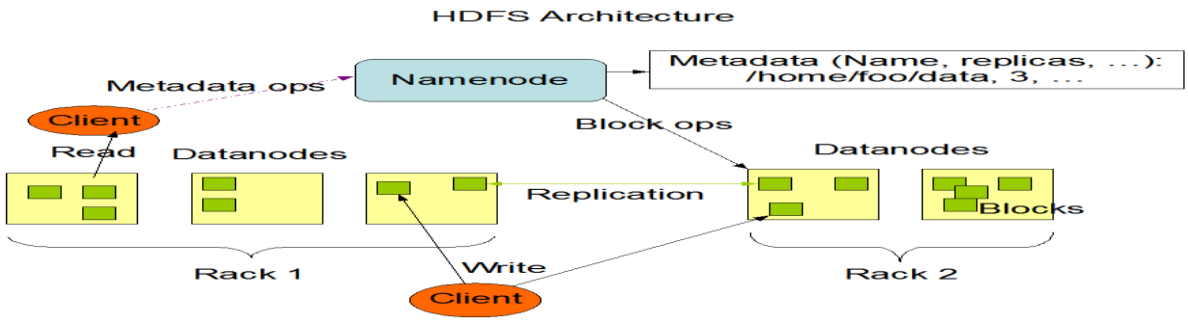

Figure 3: HDFS Architecture

5.3 Mapreduce

Map-Reduce was introduced by Google in order to process and store large datasets. [6] MapReduce have two stages which are: [4]

5.3.1 Map : The master node takes the input, divide into smaller sub parts and distribute into worker nodes. A worker node further do this again that leads to the multi-level tree structure. The worker node process the m = smaller problem and passes the answer back to the master node.

5.3.2 Reduce: The, Master node collects the answers from all the sub problems and combines them together to form the output.[3]

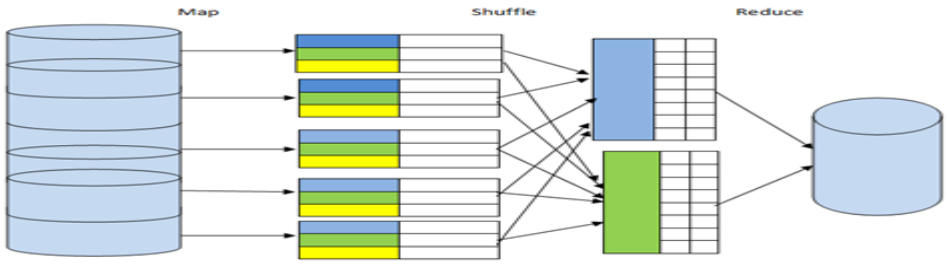

Fig. 4 Mapreduce Architecture

5.1.1 H Base: It is scalable, distributed database for random read/write access. It can serve as the input and output for the MapReduce. [3]

Pig: It is a high level data processing system where the data records are analyzed that occurs in high level language.

Hive: It is a application developed for data warehouse that provides the SQL interface as well as relational model. It helps in providing conclusion, and analysis for respective queries [6]

Sqoop: is a command-line interface application that provides platform which is used for converting data from relational databases and Hadoop or vice versa.

ETL Tool : It is the process used to extract data from multiple sources, transform it to fit your analytical needs, and load it into a data warehouse for subsequent analysis, a process known as "Extract, Transform & Load" (ETL).

BI Reporting: Business intelligence (BI) is a technology driven process including ad hoc reporting, self-service data visualization and dashboards, predictive analytics, mobile apps, and more – can help business users improve performance. It encompasses a variety of tools, methodologies that enable organizations to collect data from internal and external sources, prepare it for analysis, develop and run queries against the data, and create reports, dashboards and data visualizations to make the analytical results available to corporate decision makers as well as operational workers

VI. CONCLUSION

In this paper we analyze the Big data challenges, concepts and techniques that are suitable for big data. The techniques Hadoop which use the map reduce and HDFS paradigm is possibly the best solutions to maintain the Big Data. Hadoop is flexible and having several components and mostly used to analyze the big data. In future we will try to overcome these challenges.

REFERENCES

[1] M.K Kakhani 1, S. Kakhani2 and S.R. Biradar 3, Research Issues in Big Data Analytics ,The Federated Conference on Computer Science and Information Systems., 3, 2014, 245–249.

[2] H.HU1, Y. WEN 2 , TAT-SENG CHUA1, AND XUELONG LI 3, Toward Scalable Systems for Big Data Analytics, A Technology Tutorial, IEEE, 2 ,655-687, 2014.

[3] D. Rajasekar1, C. Dhanamani 2 , S. K. Sandhya 3, A Survey on Big Data Concepts and Tools, 5,2015.

[4] Shilpa,M. kaur, Big Data and Methodology-A review, International Journal of Advanced Research in Computer Science and Software Engineering,3(10),2013.

[5] Alexandra Adrian TOLE, Big Data Challenges, Database Systems Journal, 4(3) ,2013.

[6] Sabia1 and L. Arora 2, Technologies to Handle Big Data: A Survey, International Conference on Communication, Computing & Systems, 2014

[7] V. Shukla 1,P.K. Dubey 2, Big Data: Moving Forward with Emerging Technology and Challenge , International Journal of Advance Research in Computer Science and Software Engineering, 2(9), 2014.

[8] H S. Bhosale1, Prof. D. P. Gadekar2, A Review Paper on Big Data and Hadoop, International Journal of Scientific and Research Publications, 4(10),2014.

[9] N Gupta (Author) Ms. K.Saxena(Guide), " Cloud Computing Techniques for Big Data and Hadoop Implementation", International Journal of Engineering Research & Technology (IJERT), 3(4), 2014.

[10] MATTURDI Bardi1, ZHOU Xianwei 2 , LI Shuai 2 , LIN Fuhong2*1 , Big Data Security and Privacy A Review, China Communications Supplement (IEEE) ,No.2, 2014.