Volume 5, Issue 1, January 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Perspective on Big Data Technologies

Varuna Upadhyay
Computer Science, Pascific University
Rajasthan, India

Dr. Ashok Jain
Director Aravali Institute of Engineering
Haryana, India

Abstract- Big data is a term for massive data sets having large and complex structure with the difficulties of storing, visualizing and analyzing for further processes or results. With the prevalence of Cloud Computing today, there are so many technologies emerging on the Internet and generating huge volume of data. The irresistible service generated data become too large and complex to be effectively processed by traditional approaches. The need to process and analyze the large volumes of data has also increased. In several enterprise business and IT applications, there is a need to process petabytes of data in efficient manner daily. The increasing of big data problem in industry due to the inability of conventional database systems or process the big data sets within tolerable time limits. This paper presents an overview of big data and its architecture.

Keywords- Big data, Architecture, Hadoop, Framework.

I. INTRODUCTION

In the present internet era there is production and consumption of so much data. The increasing data is known as Big data. This is a new era where human activities and scientific pursuits will be aided by not only human and financial assets but also data assets. Big Data is a relatively new term that came from the need of big companies like Yahoo and Google to analyze big amounts of unstructured data. Its need could be identified in a number of other big enterprises as well in the other important fields. This paper is divided in different sections. Section 2 presents the related work of the technology. Section 3 consists of architecture of Big data.

II. RELATED WORK

There are so many definitions for big data such as a collection of data sets so large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications [1]. Gartner [2] big data is high volume, high variety and high velocity information assets that demand cost effective and innovative forms of information processing for enhanced insight and decision making. In the paper [3] the author explain 3Vs model. According to McKinsey [4] Big Data refers to datasets whose size are beyond the ability of typical database software tools to capture, manage, analyze and store. According to O'Reilly [5] Big data is data that exceeds the processing capacity of conventional database systems. The data is too big with moving too fast or it does not fit the structures of existing database architectures. Microsoft [6] Big data absolutely has the potential to change the way of organizations and academic institutions. IBM [7] creates everyday 2.5 quintillion bytes of data. It is so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere like posts to social media sites, pictures and videos. The collection of facts such as value and measurement is known as data. A very large data is called as big data. [8]

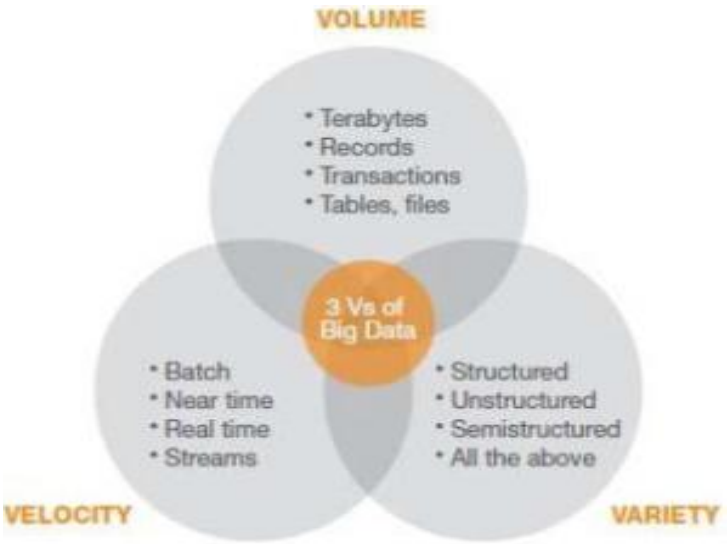

Figure 1. 3Vs Model of Big Data.

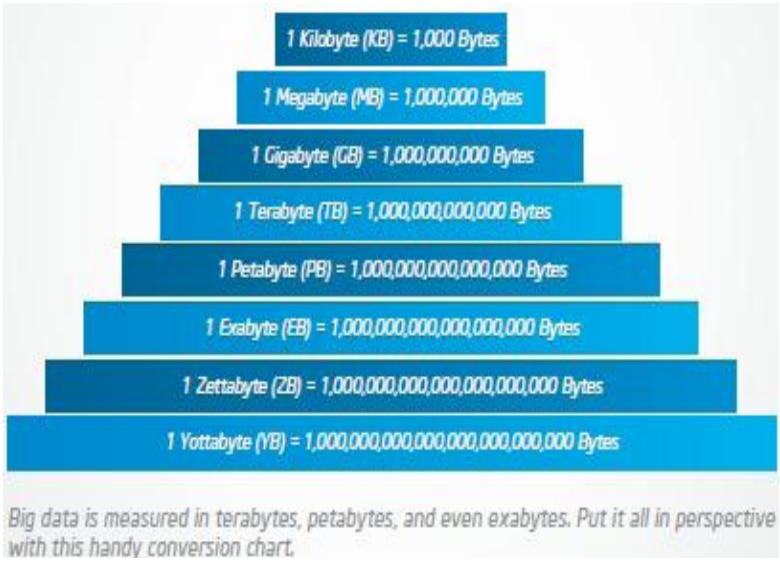

Figure 2. Mountain of Big Data.

Big data is mainly divided into three parts [9] Big Data Science, Big Data Frameworks and Big Data Infrastructure.

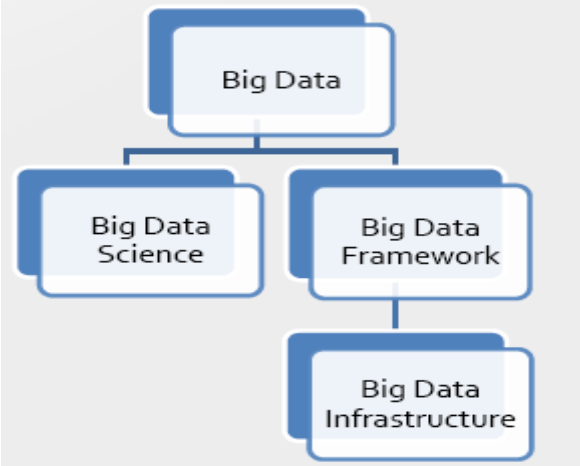

Figure 3. Big Data Parts.

A. Big Data Science

Big data science is the study of techniques covering the achievement, training and evaluation of big data. These techniques are a synthesis of both IT and mathematical approaches.

B. Big Data Frameworks

This framework is consisting of software libraries along with their associated algorithms. It will allow distributed processing and analysis of big data problems across clusters of compute units.

C. Big Data Infrastructure

Big data infrastructure is consisting of one or more big data frameworks that include management interfaces, networking, storage, back up systems and actual servers. This is used to solve specific big data problems or to serve as a general purpose analysis and processing engine.

III. BIG DATA ARCHITECTURE

This section consists of Big data Apache Hadoop MapReduce architecture.

A. Hadoop MapReduce

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data in-parallel on large clusters of commodity hardware in a reliable and fault tolerant method. The term Hadoop was derived from Google's MapReduce [11] and Google File System (GFS) [12]. A MapReduce is a job that is usually splits the input data set into independent chunks which are processed by the map tasks in a completely parallel method. In this the structure sorts the outputs of the maps and that are then input to the reduce tasks. The Hadoop software consist of MapReduce framework which is a single master JobTracker and one slave TaskTracker with thecluster node [13]. The master is responsible for scheduling the job component tasks on the slaves by monitoring them and executing the failed tasks.

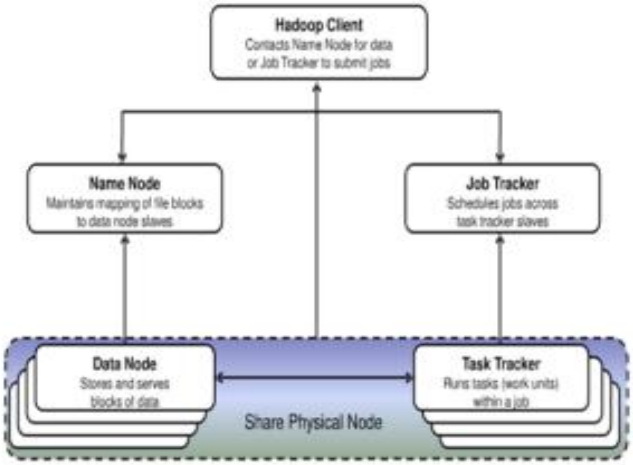

Figure 4. MapReduce JobTracker and Task Tracker

B. HDFS (Hadoop Distributed File System)

The Hadoop Distributed File System (HDFS) [14] is a distributed file system providing fault tolerance and designed to run on commodity hardware. This provides high throughput access to application data and is suitable for applications that have large data sets. It is also gives a distributed file system that can store data across thousands of servers and a means of running work across those machines. HDFS has master slave architecture.

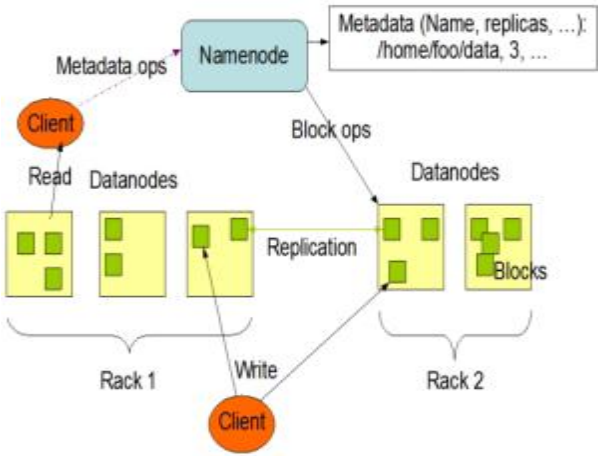

Figure 5. Hadoop Architecture

IV. CONCLUSION

Big data supports structured and unstructured data. The data which is increasing very rapidly is known as big data. Hadoop structure is a well known structure to avoid most of the upfront licensing and loading costs common to traditional relational database systems. This paper presents a better understanding of big data application with the model of Hadoop file system. Big data is a latest upcoming technology in the market which can bring huge benefits to the business organizations.

REFERANCES

[1] www.en.wikipedia.org
[2] Gartner, (2013) "IT Glossery" IEEE publication, pp.2-8.
[3] Gueyoung,J., Gnanasambandam, N., Mukherjee,T.,(2012), " Big data Analytics" Information & Computing Technology" (ICCICT), pp.19-20.
[4] Kinsey, M.,(2011) " Big dat:The next frontier for innovation, competition and productivity" pp.6-8.
[5] Reilly,O.,(2012) "An introduction to the big data landscape".
[6] Microsoft (2012) The big Bnag: How the Big data explosion is changing the world".
[7] IBM (2012)Big Data at the Speed of Business.
[8] Intel IT center(2013) Planning Guide Getting Started with Big Data.
[9] NIST Information Technology Laboratory Tackling Big Data.
[10] Xiongpai, Q., Huiju, W., Furong, L.,Baoyao, Z.,Yu, C., Cuiping, L., Hong, C., (2012) "Beyond Simple Integration of RDBMS and MapReduce – Paving theWay toward a Unified System for Big Data Analytics: Vision and Progress", Second International Conference on Cloud and Green Computing, pp2-9.
[11] Dean, J., and Ghemawat, S.,(2008) "Mapreduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, no. 1, pp. 107–113.

[12] Ghemawat,S., Gobioff,H., and Leung,S., (2003)"The google file system," ACM SIGOPS Operating Systems Review, vol. 37, pp. 29–43.

[13] Patel,A., Birla,M., Nair,U.,(2012) "Addressing Big Data Problem Using Hadoop and Map Reduce" NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING, pp.2-6.

[14] Borthakur,D. "The hadoop distributed file system: Architecture and design," (2007) Hadoop Project Website, vol. 11, pp5-8.

AUTHOR

[1] **Varuna Upadhyay** is a research scholar of Pacific Academy of Higher Education and Research University, Udaipur-Raj (India). After completing MCA from Rajasthan Vidhyapeeth University Udaipur she is persuing PhD on "Analysis the Impact of Cloud Computing in Indian Education System". She has relevant experience of teaching in the field of computer science in Nirmala College, Mumbai.

[2] **Dr. Ashok Jain**

Author is a Director of Aravali Instittute of Technical Studies, Udaipur-Raj (India). After completing MTech (IT) and MBA, he did his PhD from Mohan Lal Sukhadia University on "A Critical Evaluation of e-Governance Implementation in Rajasthan State". He is having more than 32 years of experience in the field of information technology. He is the member of "Special Interest Group on e-Governance" of Computer Society of India. His area of interest is to study and provide consultancy for successful implementation of e-Governance and e-Learning implementation in India. The author received "Rashtriya Ratana Award" in 2002 for individual outstanding performance. He is the research guide and life member of CSI, IE (India), IIIE, IIMM, ISTD and many professional bodies. He is also the active member of Internet Governance Capacity Building Program (IGCBP) whose head office is in MALTA.

