International Journal of Engineering and Advanced Technology Studies
Vol.5, No.2, pp.30-38, September 2017
Published by European Centre for Research Training and Development UK (www.eajournals.org)

BIG DATA: TECHNOLOGY, OPPORTUNITIES AND CHALLENGES

Imran Khan[1] and Shaikh Abdul Hannan[2]

[1]Milliya Arts, Science College, Beed, Maharastra (India)
[2]Al-Baha University, Al-Baha, Saudi Arabia.

ABSTRACT: *Dealing with large amount of data and running analytics on those data is becoming challenging with rapidly increase in various types of data. Big data is the technology which deals with such large amount of data analytics. It covers wide range of application areas from managing data of social networking sites to the large amount of data on ecommerce portals for decision making. In this paper an attempt is made to present a review of State of Art technology in Big Data, its importance, major benefits and challenging in this domain.*

KEYWORDS: Big Data, Remote Sensing, Medical Image Processing, Hadoop, HDFS, Security, MapReduce, Hive.

INTRODUCTION

Application of Information and Communication Technology in almost every domain of life is producing large amount of structured and unstructured data. Relational Data Base management Systems are suitable to deal with structured data, it fulfils the requirement of real time applications handling structured dataset. But growing amount of unstructured data switched requirement from terabytes to petabytes due to social networking sites, sensor network devices etc. A wide range data types like audio, images, video etc. are produced and shared by means of email or social networking sites like Facebook and twitter. On social networking sites user produces record of events of life and post updates in form of large files. Relation Database Management is not suitable to store, manage and analyse such large amount of unstructured data.

In last decade internet has changed the mode of sharing data, requirement of data storage increases every day due to high definition data capturing devices. The traffic of data on internet is increasing rapidly, in every minute 347222 tweets, Apple users download 51000 Apps, 2 million searches using Google, 350 GB data processing on Facebook, 300 hours of latest videos uploads on YouTube, Skype users sends 110040 calls, over 100 million emails transaction and around 570 new websites creation demands of technology to handling large amount of data termed as big data [1, 2]. During 2012, 2.5 quintillion bytes of data were created every day [3]. IBM indicated that everyday 2.5 Exabyte data is created and it tough to analyse it. Estimation for generated data was 5 Exabyte and 2.7 Petabytes in 2003 and 2012 respectively and by 2020 it will increase by 5 times [4]. IT Companies are also developing big data management systems, as shown in Table I [5]. Most widely used big data is Apache Hadoop; it is composed of Hadoop Common, Hadoop Distributed File System - HDFS, MapReduce and YARN modules. This framework supports scalable, reliable and distributed computing. It is specially designed so that, it can scale up from a server to thousands of terminals, it not only offers local storage and computing but also allows distributed processing for large data sets.

ISSN 2053-5783(Print), ISSN 2053-5791(online)

Table I: Big Data Management Systems [5]

Company	System
IBM	Apache Hadoop, InfoSphere
Cloudera	CDH, Cloudera Standard, Cloudera Enterprise
Oracle	Oracle Big Data Appliance
Google	BigTable
Yahoo!	Shepra
Amazon	SimpleDB
Microsoft	Dryad
Facebook	Apache Cassandra
Hypertable	HyperTable
ASF	Apache CouchDB

Data generation

Data generation is a continuous process in era of information and communication technology. The huge amount of data has been generated every day in Enterprise, Medical diagnosis and research centres, Government and private offices, space research organisations, educational institutions. Individual end users of technology are also producing a massive data using gadgets on social networking sites. Rise in E-Learning technology demands the huge storage to produce and share e-contents in a robust Learning Management Systems.

Advances in remote sensing require the massive data storage technologies to handle large size Multi spectral and Hyper Spectral Images produced using satellite. Every year many Satellites have been launched by various countries to capture Digital Images of earth for weather forecasting, Geographical Information System, Disaster Management, Town Planning, Navigation etc., and it captures large size hyper spectral images. Hyper spectral images of earth, moon etc. are captured continuously at a particular interval by satellite. A single hyper spectral image needs gigabytes of storage to store information collected in many bands of electromagnetic spectrum. It initiated a demand to develop a technology to handle big data in Remote Sensing domain.

Advancements in Medical Imaging technology make possible to capture high definition Medical Images for diagnosis. Large amount of MRI, CT scan, and Endoscopy like medical images have been captured very frequently at diagnosis centre and will surely increase in future. It demands to develop the technology to handle big data in Medical Image Diagnosis and research.

Digitisation of official documents, historical documents, books, maps using advanced high quality scanners are also demands for massive storage devices. Handling big digitized documents will need a big data technology to manage the lifecycle of these files. Availability of these documents on single click can be made possible using big data technology.

Big data technology

The definition of big data varies in literature and researchers still mention different opinions while defining big data. Min Chen et. al. [6] in a survey on big data reveals that, presently, although the significance of big data has been recognized, people still shows different

ISSN 2053-5783(Print), ISSN 2053-5791(online)

opinions in its definition. As a whole, big data means "the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time". Since different research scholars, technological enterprises, technical practitioners, data analysts, and have various definitions of big data.

Apache Hadoop defined the big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope." With this definition the big data can be summed up as a datasets to which classic database software could not acquire, store, and managed and need special tool. It gives two senses, first is, in big data technology the volumes of datasets is large and may grow over time with advances in technologies generally ranged from TB to PB. Furthermore only volume of dataset is not the criterion; big data cannot be managed and processed using general database software [7].

The data centres supports to store massive amount of data. The big data mechanism not only requires massive data storage capacity but also high processing capacity and effective data communication capacity of network. Now many advanced data centres have been developed where new algorithms are developed for resolving dilemmas of large storage, robust network and big data related computing technologies.

Five V's of Big Data

Three main characteristics also termed as three V's of big data are Volume, Velocity and Variety [8, 9, 10, and 11]. With these aspects many researchers [12, 13] also added Veracity and Value as features of big data; it is illustrated in figure 1.

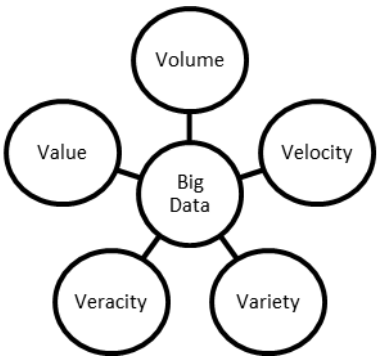

Figure1 - Five V's of Big Data

Volume:

Availability of high definition data acquisition devices is increasing the size of data day to day. The data shared by means of social networking websites and sensor networks is crossing the required space from petabytes to Zeta-bytes. Using big data tools large volume data can be managed effectively.

Variety:

Data can be stored in various formats; Variety refers to different categories of data structured, unstructured, semi-structured, mixed and raw data figure 2 illustrates it. Structured data is designed using Schema and Data models. It consist data of row and column form with relationship between them. Unlike the unstructured data doesn't designed using pre-defied data model. The semi-structured data is based on Lack strict pre-defined data models and

ISSN 2053-5783(Print), ISSN 2053-5791(online)

mixed data type is combination of various types together. Big data is able to handle unstructured data in effective manner.

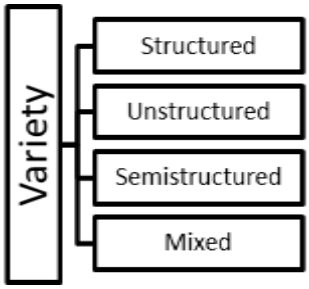

Figure 2 – Variety: a feature of big data

Velocity

This is a concept which indicates the speed at which the data generated and become historical. Big data is able to handle the incoming and outgoing data rapidly. Velocity refers to the speed at which required data is generated and produced as per requirement. The streaming of data must be at ideal speed and should deal in a timely manner. In many systems velocity of the data is important when it work on real time application; big data also promises the proper velocity of data. Speed of arrival and processing of data can categorise into [8] Batch, Near-time, Real-time and Streams. Here Batch means arrival and processing at time interval. Near-time means at small time intervals. Real time is based on continuous input, process and output, and finally streams indicate the data flows.

Variability

It terms the amount of variance adopted in summaries kept within the data bank and defined how these are closely clustered or spread out within the data set.

Value

It refers to offer value added services to the customer in enterprises and e-commerce to improve the customer relationship. To achieve this, study and observation on all customers' attitudes and market trends are needed to be analysed. Users can also query the data store to find business trends and accordingly they can change their strategies. By means of making big data open to all an entrepreneur or e-commerce company creates transparency on functional analysis.

Hadoop

It is an open source framework of the Apache Foundation, which offers solution for handling big data, processing and analysing it. Hadoop uses simple programming models for processing on large dataset across the clusters of computers. Hadoop framework is developed by Doug Cutting in 2005, and its framework is written in java. Two major components of Hadoop are Hadoop Distributed File System (HDFS) and Map Reduce framework.

ISSN 2053-5783(Print), ISSN 2053-5791(online)

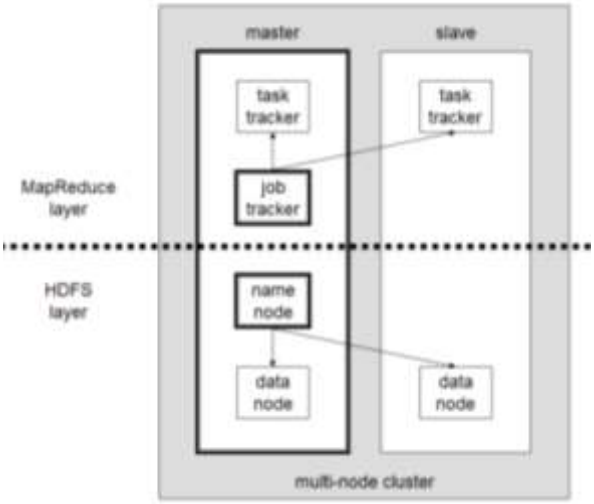

Figure 3 - Hadoop Architecture [16]

HDFS is inspired by (GFS) Google File System of Google, which offers data storage technology with features like scalable, efficient and replica based dataset storage on different nodes and forms clusters of storage; figure 3 shows architecture of Hadoop. In distributed file system the client/server based systems facilitates clients to access data stored server, process on it at fast speed as if the data were on their own computers.

Hadoop is beyond RDBMS and can handle structured, unstructured and semi structured data effectively using commodity hardware [14, 15].

Hadoop technology replicates the data across various computers, so that if a system goes down, then the data is processed on any of the replicated computers. Distributed file system reduces the risk of node failure and make possible to continue data processing even in case of failure of any node. In Hadoop application is also broken into fragments or blocks to avoid failure. Apache Hadoop consists of many components but key components are HDFS and Map Reduce. Where HDFS provides reliable storage and Map Reduce assist in analysis of data.

Hadoop Distributed File System (HDFS)

HDFS is a block-structured distributed file system and it is capable to holds large amount of data. In the HDFS the data is stored in blocks that are known as chunks. HDFS has client-server based architecture encompasses NameNode along with many DataNodes, Figure 4 – shows the architecture of HDFS.

ISSN 2053-5783(Print), ISSN 2053-5791(online)

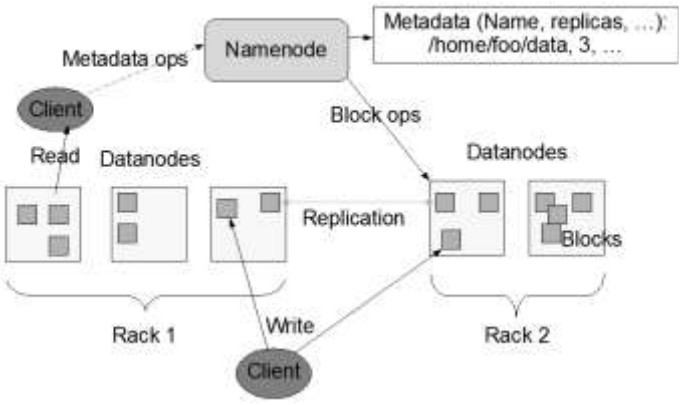

Figure 4 - HDFS Architecture [15]

The NameNode are used to store the metadata of the NameNode. Also the state of DataNodes is tracked by Name Nodes, it also see the operations of file system. In case of Node fail configuration of SecondaryNode can be done [15]

Map Reduce

It is a programming framework created by Google; it is based on 'divide and conquer' methodology for breaking complex big data into small units for doing parallel processing on big data problem. It can be further split in two stages:

Map Step: For robust system the data of master node is chopped up in many smaller sub-problem and is controlled by JobTracker node, later the result is stores in local file system, from here a reduce access it.

Reduce Step: In this step the input data is analyses and merges from map steps. To parallelize the Mas data process, multiple reduced tasks are executed parallel on worker nodes and JobTracker controls it.

Other important components of Hadoop system are as follows.

HBase

HBase is an open source database system, it is distributed, Non-relational database implemented using Java, which runs above the HDFS layer. It serves input and output for Map Reduce.

Oozie

It is a web-application and runs in java servlet. It uses the database to collect information related with workflow which is basically collection of the actions. It assists by managing Hadoop jobs in systematic manner.

Sqoop

Sqoop gives a platform to convert data from Relational databases to Hadoop and vice versa, basically it is an application based on command line interface.

ISSN 2053-5783(Print), ISSN 2053-5791(online)

Avro

It is used for data serialization and data exchange by offering such services and functionality in Apache Hadoop. According to data records, these services can be used independently as well as together.

Chukwa

It is a framework built on the upper layer of HDFS and Map Reduce framework. This framework is used for collecting data and performs analysis for processing and analysing massive amount of logs.

Pig Tool

This software analyses large datasets which consists high level language platform like SQL to express data analysis and evaluating programs. Pig Tool comprises of a compiler which produce the sequences of Map-Reduce programs. [17].

Zookeeper

Zookeeper is centralization based service, which offers groups services, distributed synchronization and maintenance for records and configuration information [18].

Apache Hive

Hive is an application on top of Hadoop developed as data warehouse framework, it offers relational model and SQL interface to write quires for processing and to analyse big data in HDFS. It assists by providing summarization, analysis and queries.

Mahout

This is a data mining and machine-learning library provide respective functionalities. It can be grouped as: collective filtering, categorization, clustering, and mining of parallel frequent patterns. The Mahout libraries can be executed by MapReduce and can also execute in a distributed mode [19].

Challenges with Big Data Processing

With all the features and functionalities which required for dealing with big data, this technology there could be some challenges and problems in big data processing. Heterogeneity and incompleteness of data, scale, complexity, Timeliness, Privacy, security, unauthentic access and Human Collaboration are few of these challenges. Jaseena et. al. presented a detail study on issues, challenges and solution in this regard [20].

Big data consist of data in different patterns and heterogeneous mixture, such data is highly dynamic and doesn't follow specific format, it may be in the form of documents, images, pdf as an email attachments. This data me be a medical images records, voice mails, audio, graphics, video, etc. It a major challenge to transform this data into structured format to analyse it later. This data is also increasing at large scale. As large data will take more time to analyse and its challenging if in a situation an instant result is needed. Unauthorized release, modification and denial of information and resources are security violation and challenging for researchers to secure the data.

ISSN 2053-5783(Print), ISSN 2053-5791(online)

CONCLUSION

Today large size of high quality digital data is produced in almost every sector of life. Social networking, emails, enterprises, Medical Image Processing, Remote Sensing, Satellite Image Processing in multi-spectral and Hyper-spectral Images and Bioinformatics are producing large amount of data with every minute. Dealing with such large amount of data and doing its analysis for decision making is becoming challenging for researchers. Big data technology came out with useful tools to handle such a huge dataset and perform analysis on it. But still many challenges and problems are to be resolve in big data. This paper focused on big data technology and an attempt is made to show importance, major benefits and challenging in this domain.

REFERENCES

[1] Viktor Mayer-Schonberger, Kenneth Cukier "Big data: a revolution that will transform how we live, work, and think", (2013) Book, ISBN 978-0-544-00269-2
[2] "How Much Data The Internet Generates In Just One Minute", Available online at, http://www.iflscience.com/technology/amount-data-internet-generates-every-minute-crazy/
[3] International Journal of Advanced Research in Computer Science and Software Engineering 3(10), October - 2013, pp. 991-995 available at http://www-01.ibm.com/software /in/data/bigdata/Shilpa
[4] http://www.iosrjen.org/Papers/vol2_issue 8%20(part-1)/K0287882.pdf
[5] Hua (Julia) Fang, Zhaoyang Zhang, Chanpaul Jin Wang, "A survey of big data research", Quantitative Health Sciences Publications and Presentations, available at http://escholarship.umassmed.edu/qhs_pp/1153
[6] Min Chen · Shiwen Mao · Yunhao Liu, "Big Data: A Survey", Springer Science + Business Media New York 2014, pp 171-209
[7] Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) "Big data: the next frontier for innovation", competition, and productivity. McKinsey Global Institute
[8] Marcos D.A, Rodrigo N.C, Silvia B, Marco A.S. Netto, Rajkumar Buyyab, "Big Data computing and clouds: Trends and future directions", Journal of Parallel Distrib. Comput. 79–80 (2015) 3–15.
[9] R.Devankuruchi "Analysis of Big Data Over the Years" International Journal of Scientific and Research Publications, Volume 4, Issue 1, January 2014 1 ISSN 2250-3153
[10] Manisha saini, Pooja Taneja, Pinki Sethi "Big Data Analytics: Insights and Innovations" International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X
[11] Katarina Grolinger, Michael Hayes, Wilson A. Higashino, Alexandra L'Heureux, David S. Allison "Challenges for MapReduce in Big Data"
[12] Steve Sonaka "Big Data and the Ag sector: More than lots of numbers" International food and agribusiness review, volume 17 issue 1, 2014
[13] Alejandro Zarate Santovena "Big Data: Evolution, components, challenges and opportunities" pp 27-33.

ISSN 2053-5783(Print), ISSN 2053-5791(online)

[14] Jayshree Dwivedi, Abhigyan Tiwary, "Big-Data-Analytics-An-Overview", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 5, ISSUE 07, JULY 2016, pp 76-81.
[15] D. Rajasekar, C. Dhanamani, S. K. Sandhya, "A Survey on Big Data Concepts and Tools", International Journal of Emerging Technology and Advanced Engineering Website: ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 2, February 2015, available at www.ijetae.com
[16] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, "A Review paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1, ISSN 2250-3153, pp 1-7
[17] K.Arun, L.Jabasheela, "Big Data Review, Classification and Analysis Survey", International Journal of Innovative Research in Information Security (IJIRIS) ISSN: 2349-7017(O), Volume 1 Issue 3 (September 2014) ISSN: 2349-7009(P) pp 17-23
[18] Sabia and Love Arora, "Technologies to Handle Big Data A Survey", Proc. Of ICCCS-2014 Conference Proceeding, available online at www.sbsstc.ac.in/icccs2014/Papers/Paper2.pdf
[19] Varsha B.Bobade, "Survey on Big data and Hadoop", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 03 Issue: 01 | Jan-2016 www.irjet.net p-ISSN: 2395-0072, pp 861-863.
[20] Jaseena K.U, Julie M. David, "Big Data Mining Issues and Challenges", Computer Science & Information Technology (CS & IT), pp 131-140 .

ISSN 2053-5783(Print), ISSN 2053-5791(online)