Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

The Big Data analytics with Hadoop: Review

Swamil Singh[1], Viplav Mandal[2], Dr. Sanjay Srivastava[3]
[1,2] Student, Computer Science Department, MGM's College of Engineering And Technology, Noida
[3]Professor, Computer Science Department, MGM's College of Engineering And Technology, Noida

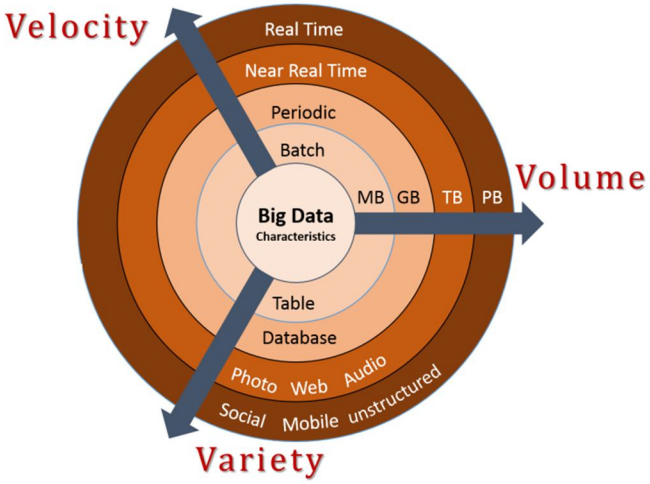
Abstract: ***Big data is huge amount of data. It is form of structure and unstructured, structure data have sql data and unstructured data is images, videos and social media data etc, today 80% data is unstructured and 20% data is structure. Big data is become from public places, industry, organization, and business function. Big data is represent by volume, velocity, variety and veracity. Hadoop is a method that processing the big data. It is open source that is use for storing and managing big data. Hadoop is also based on HDFS file system (Hadoop Distributed File System). Hadoop is cluster based which consist of data nodes and name nodes. Hadoop is using programming model that name is MapReduce.***
Keywords: Big data, Hadoop, HDFS, Map Reduce, big data concept.

I. INTRODUCTION

A. *Big Data*

Big data is a collection of large amount data that is a form of structure and unstructured. Big data has emerged with new opportunities to deal with huge amount of data. Data can be generated from web in various form like images, videos and text. Big data cannot be handled by old technology. Big data is not only containing data. It also contains the tools, techniques and frameworks. Big data consist various parameter that is volume (size), velocity (time sensitive), veracity (consistency), variety (structured and semi structured) and the ratio (20% to 80%). Big data is a term used to describe the exponential growth and availability of data. All these data have becomes in form of petabyte, zettabyte, and yottabyte. Hadoop is able to handle the variety and velocity while the system need velocity. Big data can also transform economy to businesses and other aspects. It has its high impact on society. Big data has processing certain phases like data acquisition, data representation, cleaning, aggregation, data modeling and query processing. Challenges of big data is processing size, speed and time, accuracy and performance.

B. *Characteristics of Big Data*



1) *Volume:* volume means large amount of data that is generated in every second. Now days data is increasing from petabytes to exabyte, these data is generated by machine , human intraction and networks like social media the volume of data to be analyzed is massive. 45 zettabytes data will be created by 2020 which is 300 times in 2005-06, data volume measures the amount of data available in an organization. Sometimes the same data is re-evaluated with multiple anagles and even thoug the orignal data is the same the new found intelligence creates explosion of the data. The big volume indeed represent Big Data.
2) *Velocity:* velocity define the motion of data and speed of processing of data. The importance of data's velocity the incresing rate at which data flows into an organization. Data velocity management is more than a bandwidth. Here good example of data

velocity is social media posts.

3) *Variety:* Data variety is define in form of struture and unstructerd data where structure data is 20%(text, message etc) and unstrutured data is 80% (video, audio, images,social media etc.). data variety is also important characterstics of big data.

II. LITERATURE REVIEW

Garlasu, D.; Sandulescu, V.; Halcu, I. ; Neculoiu, G. (17-19 Jan. 2013),"A Big Data implementation based on Grid Computing", Grid Computing – This paper present the advantage about the storage capabilities and the processing power and the Hadoop technology is used for the implementation purpose. Grid Computing provides the concept of distributed computing. The benefit of Grid computing center is the high storage capability and the high processing power. Grid Computing makes the big contributions among the scientific research, help the scientists to analyze and store the large and complex data.

Aditya B. Patel, Manashvi Birla, Ushma Nair (*6-8 Dec. 2012)* "Addressing Big Data Problem Using Hadoop and Map Reduce"- This paper present the experimental work on the Big data problems. It describe the optimal solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and Map Reduce programming framework for parallel processing to process large data sets.

Richa Gupta, Sunny Gupta, Anuradha Singhal, "Big Data: Overview", - This paper present an overview on big data, its importance in our live and some technologies to handle big data. This paper also states how Big Data can be applied to self-organizing websites which can be extended to the field of advertising in companies.

Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future" - The paper presents a broad overview of the topic big data mining, its current status, controversy, and forecast to the future. This paper also covers various interesting and state-of-the-art topics on Big Data mining.

Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) "Shared disk big data analytics with Apache Hadoop"- This paper present the Big data analytics define the analysis of large amount of data to get the useful information and uncover the hidden patterns. Big data analytics refers to the Mapreduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of implementation of Google's Mapreduce Model [2]. In this the performance of SF-CFS is compared with the HDFS using the SWIM by the facebook job traces .SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns.

Jimmy Lin "**MapReduce Is Good Enough**?" – This paper present used Hadoop which is currently the large –scale data analysis "hammer" of choice, but there exists classes of algorithms that aren't "nails" in the sense that they are not particularly amenable to the Map Reduce programming model . He focuses on the simple solution to find alternative non-iterative algorithms that solves the same problem. The standard Map Reduce is well known and described in many places .Each iteration of the page rank corresponds to the Map Reduce job. The author suggested iterative graph, gradient descent & EM iteration which is typically implemented as Hadoop job with driven set up iteration &Check for convergences. The author suggests that if all you have is a hammer, throw away everything that's not a nail.

Sagiroglu, S.; Sinanc, D. (20-24 May 2013),"Big Data: A Review" - This paper present the big data content, its scope, methods, samples, advantages and challenges of Data. The critical issue about the Big data is the privacy and security. Big data samples describe the review about the atmosphere, biological science and research. Life sciences etc. By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big data easy to get the information from the complicated data sets [6]. The overall Evaluation describe that the data is increasing and becoming complex. The challenge is not only to collect and manage the data also how to extract the useful information from that collected data. According to the Intel IT Center, there are many challenges related to Big Data which are data growth, data infrastructure, data variety, data visualization, data velocity.

Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis: Challenges and Solutions" - This paper presents various methods for handling the problems of big data analysis through Map

Reduce framework over Hadoop Distributed File System (HDFS). Map Reduce techniques have been studied in this paper which is implemented for Big Data analysis using HDFS.

III. CHALLENGES OF BIG DATA

A. Privacy

Data privacy is also a large problem of big data. In some countries very rigorous laws related to privacy for example in USA there

are rigorous laws for health record, but for other it is very forceful.

B. Heterogeneity and incompleteness

It is big challenge in data analysis. Because when data is analysis then data is structure and when we talk about big data then data may be form of structure and unstructured. Consider an example of patient in hospital. We will create each record for each medical test. And we also create a record for hospital stay. This will be dissimilar for all patients. This design is not good structured. So managing with the Heterogeneous and incomplete is required. A well data analysis should be applied to this.

C. Scale

As we know from its name big data have a large datasets. Conventional software tools are not sufficient for managing expand of volumes of data, organization and data analysis. Mostly technologies are based on cloud, so due to cloud technology data is produced in high rate. Due to this data is becoming a provocation issue to data analysis.

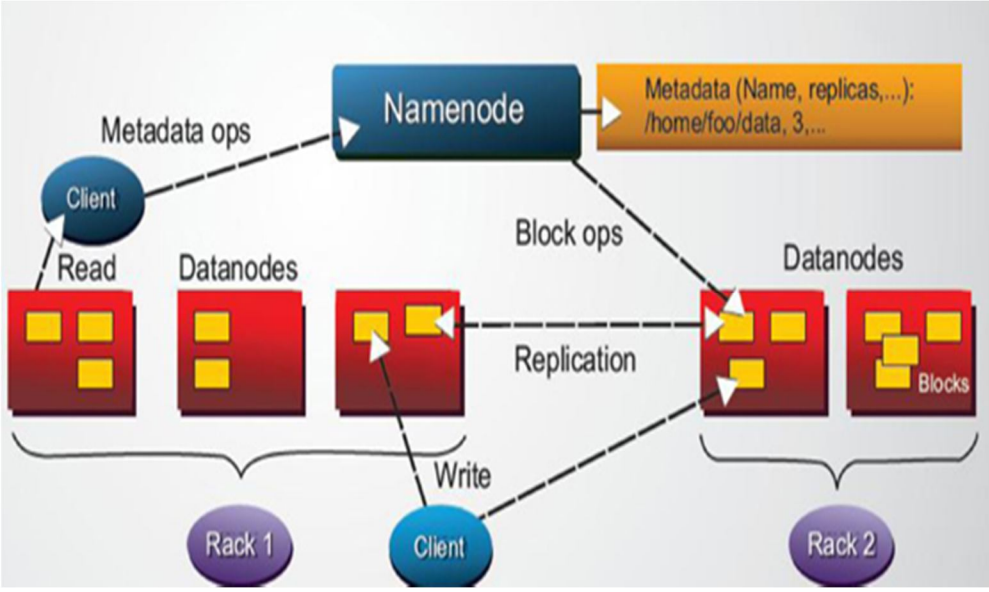
IV. TECHNOLOGIES OF BIG DATA

A. Hadoop

Hadoop is a product of apache foundation and it is also an open source product. Hadoop is a very helpful software for big data. It is a very famous for researcher to analyze the big data. Hadoop permit writing applications that fast process huge amounts of data in simillar on huge clusters of compute nodes. Hadoop was developed by Google's map reduce and it break down the application into various parts. And it is also a programming framework. Hadoop is based on mapreduce, HDFS and its components is a Hive, Hbase, Flume, Sqoop. Hadoop cluster is a used Hadoop Distributed File System to manage the data. It is better for storing terabyte and petabyte data on cluster. Hadoop is highly fault tolerant and high capacity distributed file system. Hadoop is divides every file into small fixed size.

B. Hadoop components

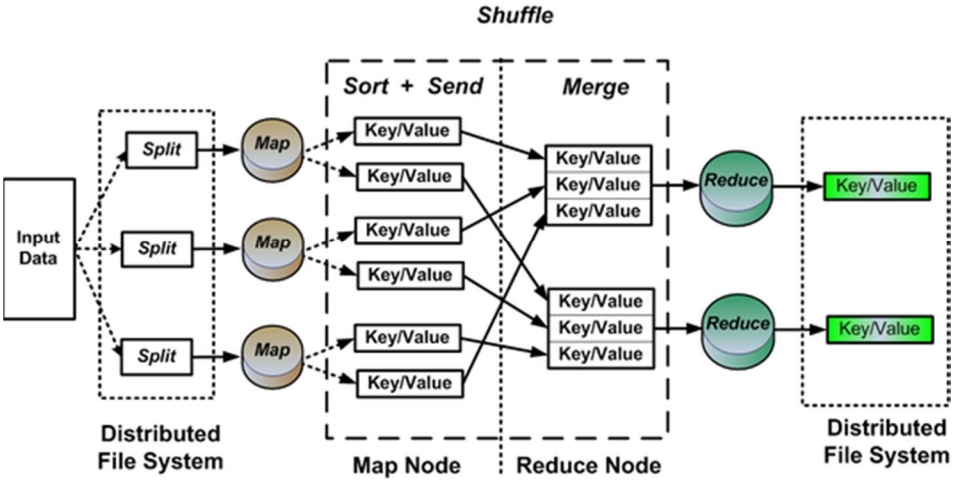
1) HDFS



Hadoop Distributed File System is a high fault tolerant distributed file system and it is storing data on a cluster in a form of fixed blocks. Here NameNode is called the master node which manage the blocks present in datanode. HDFS is biuld using Java programming language. It keeps a record of how the files in HDFS are divided into blocks, in which nodes these blocks are stored and by and large the NameNode manages cluster configuration. The NameNode is also responsible to take care of the replication factor of all the blocks. If there is a change in the replication factor of any of the blocks, the NameNode will record this in the EditLog. Datanodes execute the low-level read and write appeal from the file system's clients. Datanode send a report time to time on all blocks that is present in cluster to the NameNode. If any node have break, it process have never damaged when cluster operating operating the cluster, by shifting work to the remaining machines in the cluster.

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

2) Map Reduce



Map reduce is programming model and developed by google in 2004, it is divide the work into map and reduce, Shuffle is working between map and reduce for merge the key value, Mapreduce is also based on hadoop distributed file system which is store and manage the data. MapReduce is a simple programming model for processing huge data sets in parallel. The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node. The MapReduce framework operates exclusively on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job.

Input and Output activity of a MapReduce work –

(input) <k1, v1> -> **map** -> <k2, v2> -> **combine** -> <k2, v2> -> **reduce** -> <k3, v3> (output)

Mapreduce is also called the heart of hadoop. Mapreduce concept is very simple for who have also well known about the clustered scale-out data processing solutions.

V. CONCLUSION

In this review paper Big data is a huge amount of data which is in form of structure and unstructured. Technology is used in big data is hadoop which apache software and it is open source product, hadoop use a distributed file system name is Hadoop distributed File System which is manage and storing the huge amount of data in cluster, this paper is also describe on challenges and technologies of big data. Hadoop is plays a good role in big data.

VI. ACKNOWLEDGMENT

I am highly obliged to Dr. Sanjay Srivastava for his guidance, support and keen supervision. And I also thank to friends for supporting me.

REFERENCES

[1] Garlasu, D.; Sandulescu, V.; Halcu, I. ; Neculoiu, G. (17-19 Jan. 2013),"A Big Data implementation based on Grid Computing", Grid Computing.
[2] Aditya B. Patel, Manashvi Birla, Ushma Nair (6-8 Dec. 2012) "Addressing Big Data Problem Using Hadoop and Map Reduce".
[3] Richa Gupta, Sunny Gupta, Anuradha Singhal, "Big Data: Overview".
[4] Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future".
[5] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) "Shared disk big data analytics with Apache Hadoop".
[6] Jimmy Lin "MapReduce Is Good Enough?"
[7] Sagiroglu, S.; Sinanc, D. (20-24 May 2013),"Big Data: A Review".
[8] Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis: Challenges and Solutions.
[9] Sumit Kumari "A Review Paper on Big Data and Hadoop".
[10] Rahul Beakta "Big Data and Hadoop: A Review Paper".
[11] Raja Appuswamy_, Christos Gkantsidis, Dushyanth Narayanan,Orion Hodson, and Antony Rowstron "Scale-up vs Scale-out for Hadoop: Time to rethink?".
[12] Dr. Shoban Babu Sriramoju "A Review on Processing Big Data".
[13] Jaseena K.U.1 and Julie M. David "ISSUES, CHALLENGES, AND SOLUTIONS:BIG DATA MINING".
[14] Silky Kalra, Anil lamba "A Review on HADOOP MAPREDUCE-A Job Aware Scheduling Technology".
[15] Ms. Gurpreet Kaur and Ms. Manpreet Kaur "REVIEW PAPER ON BIG DATA USING HADOOP".
[16] Dr. Siddaraju, Sowmya C L, Rashmi K, Rahul M "Efficient Analysis of Big Data Using Map Reduce Framework".
[17] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop".
[18] K.Arun and Dr.L.Jabasheela "Big Data: Review, Classification and Analysis Survey".

[19] Chen He, Derek Weitzel, David Swanson, Ying Lu "HOG: Distributed Hadoop MapReduce on the Grid".
[20] Ivanilton Polatoa,b,⇤, Reginaldo R´eb, Alfredo Goldmana, Fabio Kona "A Comprehensive View of Hadoop Research - A Systematic Literature Review".
[21] Jeffrey Shafer, Scott Rixner, and Alan L. Cox "The Hadoop Distributed Filesystem:Balancing Portability and Performance".
[22] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler "The Hadoop Distributed File System".
[23] Shilpa and Manjit Kaur "BIG Data and Methodology-A review".
[24] Suman Arora and Dr. Madhu Goel "Survey Paper on Scheduling in Hadoop"
[25] http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/
[26] http://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/
[27] http://www.forbes.com/sites/oreillymedia/2012/01/19/volume-velocity-variety-what-you-need-to-know-about-big-data/2/#66f12a727c1d
[28] http://www.edureka.co/blog/apache-hadoop-hdfs-architecture/
[29] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
[30] https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/
[31] http://blog.cloudera.com/blog/2012/10/analyzing-twitter-data-with-hadoop-part-2-gathering-data-with-flume/