A Review Paper on Big Data using Hadoop

Divya Jain
UG Student,
Department of Computer Science & Engineering
Udaipur, Rajasthan 313002, India
jain.divya744@gmail.com

Divija Ameta
UG Student ,
Department of Computer Science & Engineering
Udaipur, Rajasthan 313002, India
dameta.ameta@gmail.com

Charu Kavadia
Asst. Professor,
Department of Computer Science & Engineering
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan 313022 India
charukavadia@gmail.com

ABSTRACT

The data growth is increasing day by day due to the ever increasing data channels and the categories available to the different sources. Since every organisation has a larger availability of the data so as to have the quantity as well as quality of new information. These large volumes of data and information are known as Big Data. It mainly means to deal with the massive amount of the data available to various organisations. Big Data mainly deals with the analysis of the data and to convert the massive unstructured form of data into something structured form. There are various technology which deals with the Big Data such as Mapreduce & Hadoop. Hadoop is the open source software build on the Mapreduce. This paper mainly explains the Big Data, its uses and advantages. Along with this an introduction to Hadoop and its components is also done in this paper.

Keywords: Big Data, Hadoop, HDFS, Mapreduce and Hadoop Clustering

I. INTRODUCTION

Data is the main asset of any work, or an organisation. Data can be generated from the different sources and thus can be finding data in the system at various points. These large amount of data need to be process to have the knowledgeable part from the data know as information. These can be done using various Big Data techniques.

Big Data

Big Data is a modern technique and technology [1] to capture, manage, store, distribute, and analyze larger-sized datasets with different data structures and velocity. Big data can consist of structured, unstructured or semi-structured data depending upon the capability of the system to hold it. It actually manages the data which are complex, have variable size, deals with the large volumes which cannot be handle by the local sources and techniques. Big Data is mainly define by the four V's. These are as follows:

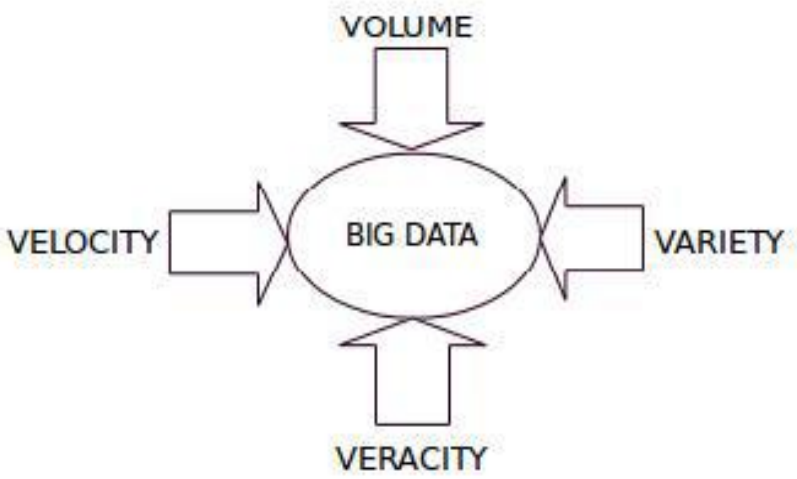

Fig.1Big Data Features

Volume means large amount of data generated in every second. Data volumes are doubling every year. Machine generated data are examples for this characteristic. Data volume is tremendously increasing from gigabytes to petabytes [2]. A research says that 40 Zettabytes of data will be generated by the end of 2020 which will be 300 times from 2005 [3]. Independent market and research studies have found that data is increasing at alarming rate is in the unstructured form which needs to be managed and can be only handle by the Big Data.
Velocity: means analysis if the streaming data which actually defines the speed at which the data is produced and developed. For time sensitive systems Big Data can be used to find out the maximise value of the enterprises.
Variety: means the type of the data. As there are many different form of data available to the system such as text, image, video, audio etc.
Veracity means accuracy or uncertainty of data. Data is uncertain due to the incompleteness and inconsistency.

II.CHALLENGES WITH BIG DATA

Any new technology has obstacles and hurdles, without handing these obstacles may lead to failure [4] of the technology implementation and not give good results.

- **Privacy and Security**: Either data is small or vast, security of data is main concerned. Major issues with security of data are:
    - Authentication of each user and team members accessing the data.
    - Proper use of encryption on data while travelling or at rest.
    - Storing of data with proper fault tolerant mechanism.
    - Restrict access based on user needs.
    - For example, in social media website some post are restricted for analysis.
- **Different types of data**: Data is heterogeneous in nature. It means data is structured, unstructured and semi structure data. To handle and analysing these data is very big challenge. Processing of structured data is easy but how to handle unstructured and semi structured data.
- **Scale**: As name implies data is very huge and data is increasing day by day in tera bytes, peta bytes etc. Major problem is where we have to store the data? In earlier years this problem is solved by vertical scaling, processors and hard disk is increased but they again become slower than horizontal scaling and cloud technology is come. Now data is stored at hard disk and cloud
- **Human Resource and Man Power:** Big data is its early stage so youth are taking interest but big data must go in research, industries and development areas.
- **Information Growth:** A large amount of data in enterprise consists of unstructured data, which is growing at an alarming rate than traditional relational information. This massive information is a threat to most well-prepared IT organizations.

- **Heterogeneity**: In Big Data, data may be structured or unstructured so as per our traditional form data needs to be in the structured form to be processed. Heterogeneity is the big challenge in analysis of the data.

III. HADOOP FRAMEWORK

To resolve above challenges and issues Hadoop is introduced. It is an open source framework to handle huge amount of data is provided by Apache Software Foundation in 2012. Its implementation is in Java and handles not only structured, semi-structured but also structured data. It uses the concept of parallel working; concurrent users can access the data without a delay. Hadoop is influenced by Google's architecture. To store huge amount of data HDFS [5, 6, 12] (Hadoop File System) is given and for processing Mapreduce technique is given. For maintain resources YARN (Yet another Resource Negotiator) technique is introduced.

Hadoop is consisting of two main components:

- Storage
- Processing

Storage: For storage Hadoop file system is given, which is based on distributed design of file system. It is specially design file system which runs on commodity of hardware. It means that Hadoop is low cost technology. Hadoop file system can hold huge amount of data and provide parallel processing.

To achieve fault tolerance of data is stored at multiple nodes or machine. It has five main services:

- Name Node
- Secondary Name Node
- Job Tracker
- Data Node
- Task Tracker

Master Services

Slave Services

HDFS Architecture: Its architecture follows the master slave relationship.

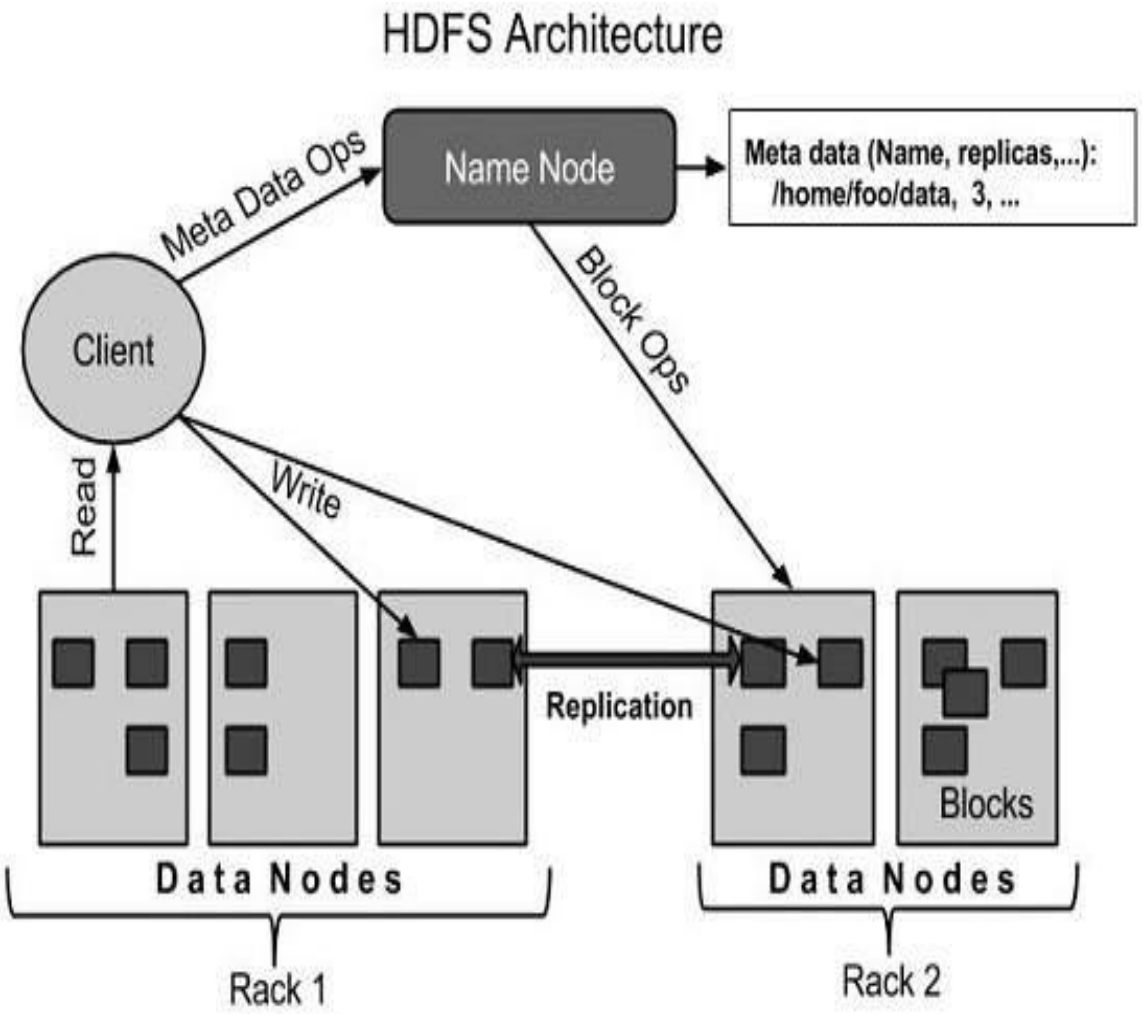

Fig.2.HDFS Architecture

Name Node: It is a master node which manages the file system name space or we can say it stores the Meta data, information data about data. It means that which data is stored on which data node.
Data Node: The nodes where the data is actually stored and perform read and write operations.
Job Tracker: When a request is send by client then contact to name node to get Meta data then it assigns the task to task tracker. It manages the entire task trackers.
Task Tracker: It performs the task and sends results back to Job Tracker.

Map Reduce: It is a processing technique which is combination of map and reduces.
Map: For mapping data files are break into size of block which is called input split. Then by using Record Reader which convert data into key, value pair. This conversion is done because mapper understands only key, value only. Then Mapper runs the program and gives output into key, value from.

Reduce: Reduce phase is combination of shuffle and reduce phase. First shuffle is run then reduce phase. Finally merge the output and stored at HDFS.

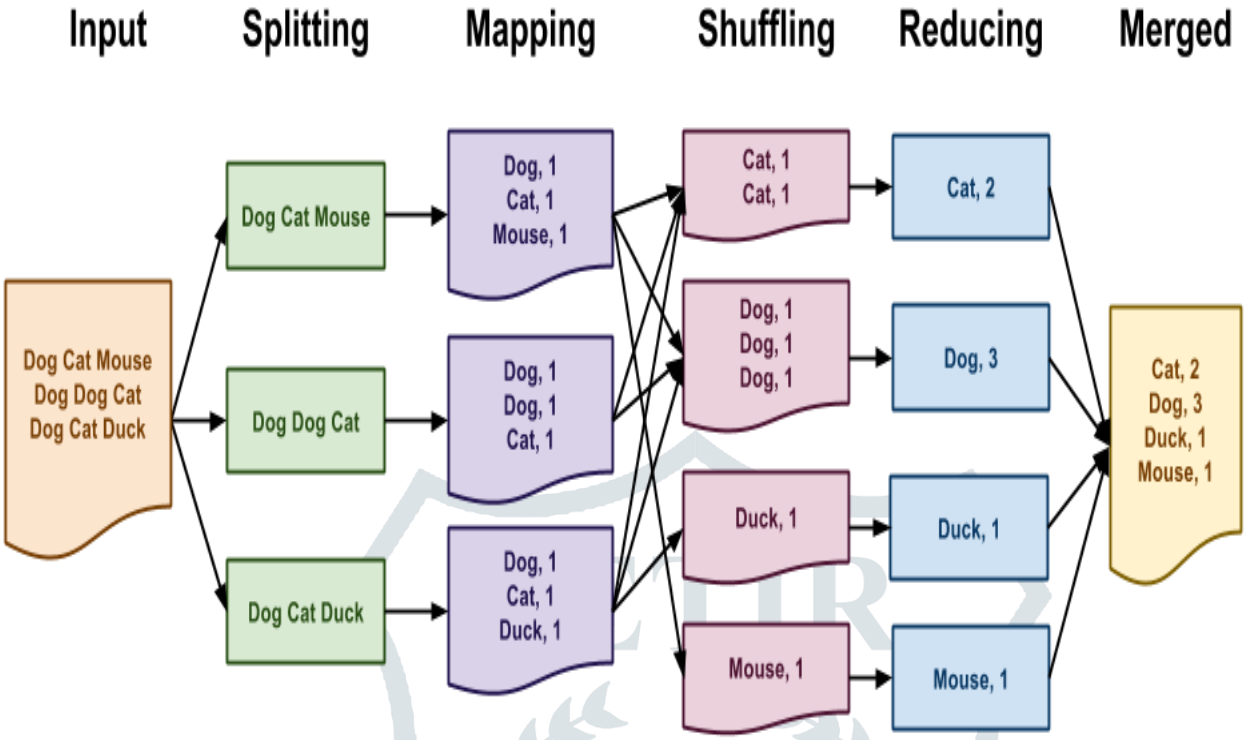

Fig.3. Map Reduce Programming Model

Here in this example, counting of words is done. First file is splited then mapper plays its role and counts the number of words. After that shuffle phase started and similar words are put into one file. Then Reducer phase counts actual number of words and finally merged output is given back to client.

IV.LITERATURE SURVEY

Harshawardhan S. Bhosale1, Prof. Devendra Gadeka [6, 12], gave a review on Big Data and Hadoop. They discussed the 3'Vs features of Big Data. These are volume, velocity and variety of data. They also discussed the problem of faster processing of data.

S. Vikram Phaneendra & E. Madhusudhan Reddy [7], illustrated the how RDBMS is failed to handle huge amount of data, which is called as 'Bidg Data'. Also explained how big data differs from traditional data in from of volume, velocity, variety, and complexity. They also gave the illustration of Hadoop Technology in various eras like health-care, finance, insurance etc. Also discussed the various issues of big data.

Real Time Literature Review [9] about the Big data is being given in this research. According to 2013, facebook a social networking site has on an around 1.11 billion people active accounts from which 751 million using facebook from a mobile. Flicker is the another example of the Big Data having feature of Unlimited photo uploads, the ability to show HD Video, Unlimited storage, unlimited video upload. Flicker had a total of 87 million registered members out of which more than 3.5 million upload images daily

Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W.[8], shared disk big data analytics with Apache Hadoop" Big data analytics is the analysis of large amount of data so as to obtain the useful information and find the hidden paths. It refers to the Mapreduce Framework developed by the Google. Hadoop an open source platform is used for the purpose of implementation Mapreduce Model.

Hadoop and Map Reduce [10] gives the experimental work on the Big data problems. The optimal solution using Hadoop cluster, Hadoop Distributed File System (HDFS) is being shown in this research work. Map Reduce is being explained to show the programming framework for parallel processing .

Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. [11], A Big Data implementation based on Grid Computing", Grid Computing is having the advantage of the storage capabilities and the processing power. It is used with the Hadoop technology for the implementation purpose. Grid Computing works on the concept of distributed computing. These provides the benefit of high storage capability and the high processing power.

Shilpa, Manjeet Kaur, LPU, Phagwara [18], a review on Big Data and Methodology , illustrated various challenges and issues in big data. They also explain the fundamental research towards these technical issues. Big-data analysis transforms data in the financial, operational, and commercial problems using discrete data sets. By cloud based virtualization infrastructure which are used to mine data sets efficiently, big-data methods new analysis to the data sets.

V. APPLICATION IN DATA MINING

Data is very important to identify hidden patterns and clubbing of similar patterns into one. These Data sets are useful for not only business organizations but also for researchers to extract useful data. Identifying useful information from raw data is called Data Mining. In this digital technology, there is huge amount of data on the internet in the form of text, images, videos, social posts and live data. These data is increasing day by day. Data may reach 300 times of 2005 [13]. To get useful information from this data mining technique is very useful in various areas like health-care, insurance, education, finance, security, banking etc

A. **Classification Analysis:**

Classification technique is used to club similar data at one point. It is a good approach to obtain information about data and Meta data

B. **Cluster Analysis:**

It is a systematic approach to identify heterogeneous and homogeneous data. For example clusters of customers having similar preferences can be targeted on Social medial [14].

C. **Evolution Analysis:**

It is genetic approach to find useful information or we can say it extract information from DNA sequences. It is used in Banking Domain, to predict Stock exchange and share market analysis[15].

D. **Outlier Analysis:**

Some observations, identifications of items are done which do not make a pattern in a Data Set. In medical and Banking problems this is used.

VI. CONCLUSION

In this review paper, we described the overview of Big data and Hadoop technology. We discussed the various problems of Big data and then we discussed about its solution (Hadoop). If we want to achieve benefits of Hadoop Technology then there must be support and research towards this technology. We presented how Hadoop framework stores the data and processed the data and how all the master and slave node are processed during map reduce phase. This paper can be focused for Data Mining where we can use clustering approach to find hidden patterns of users and clients like what he/she like, where he/she visits regularly etc.

ACKNOWLEDGEMENT

We would like to express my deep gratefulness to all the people who have supported me during this work. In particular, I offer my sincerest gratitude to my friend, **Ms. Divija Ameta**, who spared her valuable time in guiding me for my dissertation work. She has always been there to direct the way, provide insight and take part on all aspects of this dissertation work also in developing the dissertation, for supporting and motivating me in the process of the dissertation.

REFERENCES

[1]Sumit Kumari ," A Review Paper on Big Data and Hadoop"

[2] SMITHA T, V. Suresh Kumar "Application of Big Data in Data Mining" in International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 7, July 2013).

[3] IBM Big Data analytics HUB, www.ibmbigdatahub.com/infographic/four-vs-big-data

[4] Katal, A Wazid, M.; Goudar, R.H., (Aug,2013)," Big data: Issues, challenges, tools and Good practices"

[5] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop" in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.

[6] Harshawardhan S. Bhosale1, Prof. Devendra Gadekar, JSPM's Imperial College of Engineering & Research, Wagholi, Pune,a review on Big Data

[7] S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[8] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"

[9] A Real Time Approach with Big Data-A review

[10] Aditya B. Patel, Manashvi Birla, Ushma Nair,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"

[11]Garlasu, D.; Sandulescu, V; Halcu, I. ; Neculoiu, G. ;,(17-19 Jan. 2013),"A Big Dataimplementation based on Grid Computing", Grid Computing

[12]Puneet Singh Duggal, Sanchita Paul, Big Data Analysis: Challenges and Solutions in International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

[13] IBM Big Data analytics HUB, www.ibmbigdatahub.com/infographic/four-vs-big-data

[14] Smitha.T, Dr.V.Sundaram, "Classification Rules by Decision Tree for disease prediction" International journal for computer Application, (IJCA) vol 43, 8, No-8, April 2012 edition. ISSN0975- 8887; pp- 35-37

[15] Mucherino A. Petraq papajorgji P.M.Paradalos 1998. A survey of data mining techniques alied to agriculture CRPIT.3(3): 555560.

[16] Anupam Jain, Rakhi N K and Ganesh Bagler, arxiv.org/abs/1502.03815 Spices Form The Basis Of Food Pairing In Indian Cuisine.

[17] MIT Technology Review, http://www.technologyreview.com/view/535451/data-miningindian- recipes-reveals-new-food-pairing-phenomenon/

[18]Shilpa, Manjeet Kaur, LPU, Phagwara, India, a review on Big Data and Methodology