International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.01, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Big Data and Hadoop

Shalmali Jadhav[1], Revati Pungavkar[2]

[1,2] *KIT's College Of Engineering,Shivaji University, Kolhapur(India)*

ABSTRACT

In this world of information the term BIG DATA has emerged with new opportunities and challenges to deal with the massive amount of data. The term 'Big Data' describes innovative techniques and technologies to capture, store, distribute, manage and analyze petabyte- or larger-sized datasets with high-velocity and different structures.Data is generated from various different sources and can arrive in the system at various rates. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. This paper presents an overview on Big Data, Advantages and its scope for the future research. Big Data present opportunities as well as challenges. This paper gives an introduction to Hadoop and its components.

Keywords: Big data, petabyte, Hadoop, Distributed processing, clusters, algorithms.

I.INTRODUCTION

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. at us with a great Velocity is normally refer to as Big Data. Big data can be structured, unstructured or semi-structured, which is not processed by the conventional data management methods.

Hadoop is open source software used to process the Big Data. It is very popular used by organizations/researchers to analyze the Big Data. Hadoop is influenced by Google's architecture, Google File System and MapReduce. It is a Programming framework used to support the processing of large data sets in a distributed computing environment.

II.BIG DATA

A] Big Data Definition:

Data that has extra-large Volume, comes from Variety of sources, Variety of formats and comes at us with a great Velocity is normally refer to as Big Data. Big data can be structured, unstructured or semi-structured,

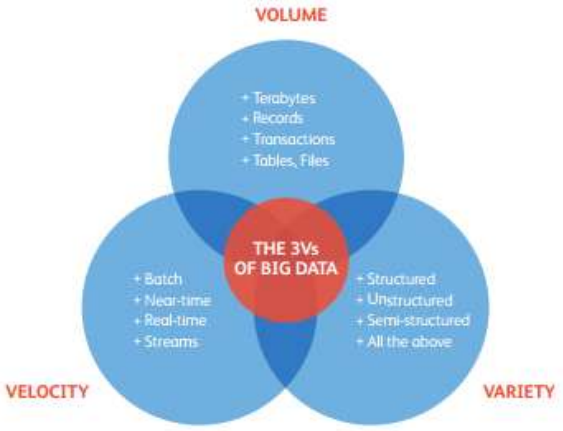
which is not processed by the conventional data management methods.Data can be generated on web in various forms like texts, images or videos or social media posts. In order to process these large amount of data in an inexpensive and efficient way, parallelism is used. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data.

B] 3 Vs of Big Data

Volume of data: Volume refers to amount of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes.

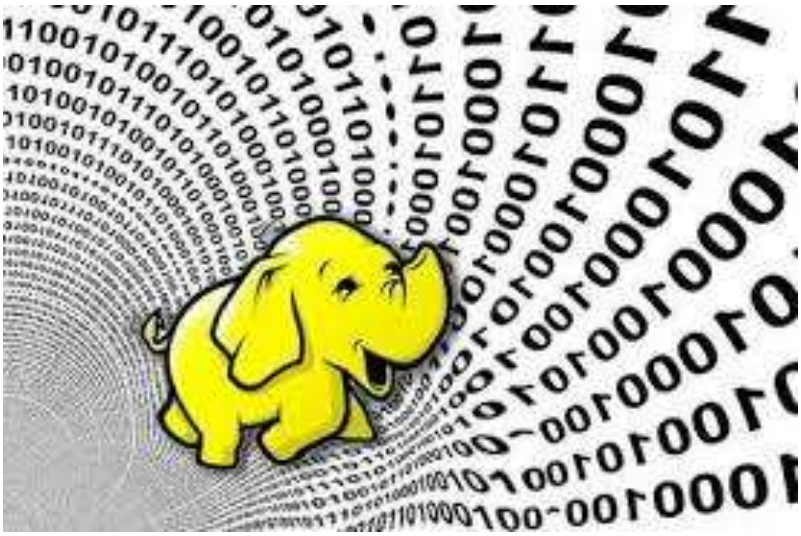
Variety of data: Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio, video, XML etc.

Velocity of data:Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.



III.HADOOP

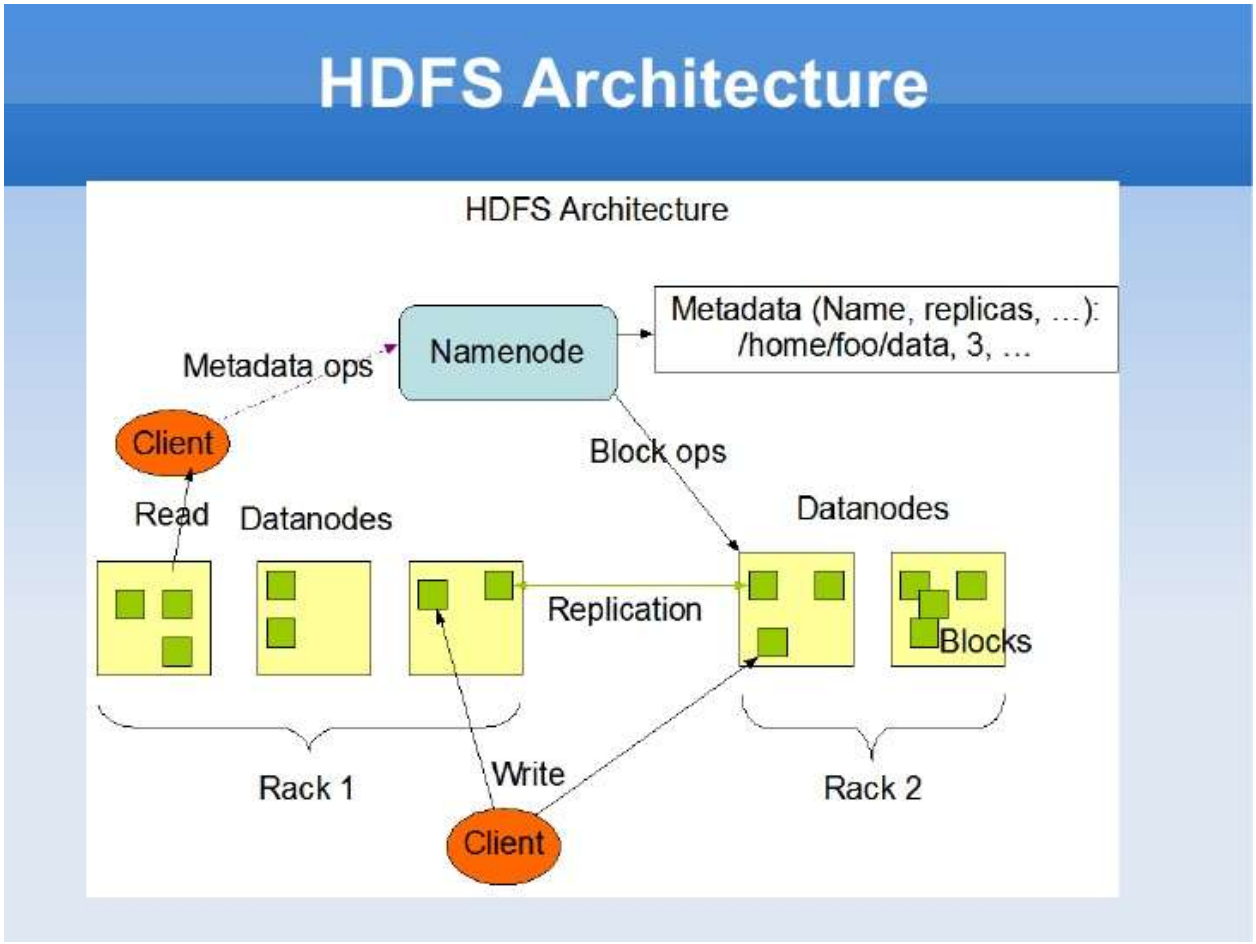
The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware.HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them.HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers.



A. Hadoop Framework

Hadoop consists of two main components:

1) *Storage*: The Hadoop Distributed File System (HDFS): It is a distributed file system which provides fault tolerance and designed to run on commodity hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS can store data across thousands of servers. HDFS has master/slave architecture. Files added to HDFS are split into fixed-size blocks. Block size is configurable, but defaults to 64 megabytes.
2) *Processing*: MapReduce [4]: It is a programming model introduced by Google in 2004 for easily writing applications which processes large amount of data in parallel on large clusters of hardware in fault tolerant manner. This operates on huge data set, splits the problem and data sets and run it in parallel.



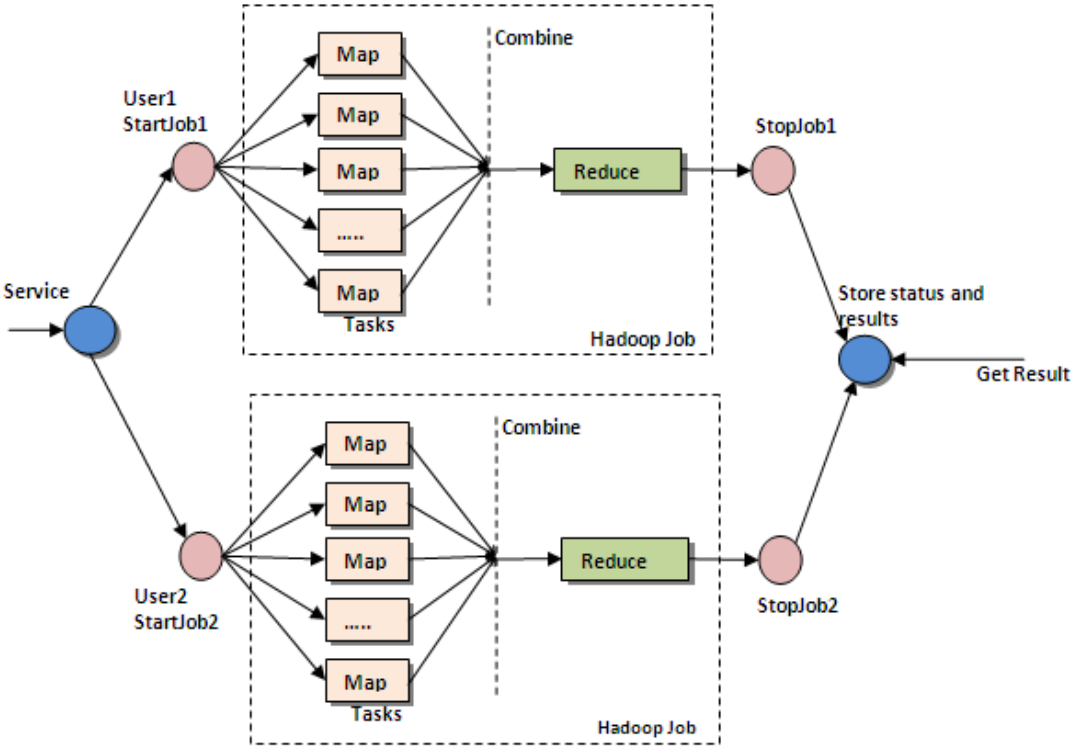
B. MapReduce Architecture

The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel.For example, a very large dataset can be reduced into a smaller subset where analytics can be applied. There are two functions in MapReduce as follows:

map – the function takes key/value pairs as input and generates an intermediate set of key/value pairs

reduce – the function which merges all the intermediate values associated with the same intermediate key.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.01, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354



IV.COMPONENTS OF HADOOP

The Table 1, Comparison among Components of Hadoop, gives details of different Hadoop Components which have been used now days. HBase, Hive, MongoDB, Redis, Cassandra and Drizzle are the different components. Comparison among these components is done on the basis of Concurrency, Durability, Replication Method, Database Model and Consistency Concepts used in the components.

- HBase – It is a widecolumn store based on Apache Hadoop and on concepts of Big Table.
- Hive – It is a Data Warehouse Software for Querying and Managing Large Distributed Datasets, built on Hadoop.
- MongoDB – It is one of the most popular Document Stores.
- Redis - In-memory Database with configurable options performance vs. persistency.
- Cassandra - Wide-column store based on ideas of BigTable and DynamoDB.
- Drizzle – MySQL fork with a pluggable micro-kernel and with an emphasis of performance over compatibility.

V.CHALLENGES AND OPPORTUNITIES

Big data comes with a lot of opportunity to deal in health, education, earth, and businesses but to deal with the data having large volume using traditional models becomes very difficult.

A. Challenges with Big Data:

1) *Heterogeneity and Incompleteness*:Heterogeneity is the big challenge in data Analysis and analysts need to cope with it. Consider an example of patient in Hospital. We will make each record for each medical test. And we will also make a record for hospital stay. This will be different for all patients. This design is not well structured. So managing with the Heterogeneous and incomplete is required. A good data analysis should be applied to this.

2*) Scale*:As the name says Big Data is having large size of data sets.Hard disks are used to store the Data. They are slower I/O performance. But now Hard Disks are replaced by the solid state drives and other technologies. These are not in slower rate like Hard disks, so new storage system should be designed.

3) *Timeliness*:Another challenge with size is speed. If the data sets are large in size, longer the time it will take toanalyze it.

4) *Privacy*:Privacy of data is another big problem with big data.For example, in social media we cannot get the private posts of users for sentiment analysis.

5) *Human Collaborations*:In spite of the advanced computational models, there are many patterns that a computer cannot detect.We need technological model to cope with this.

B. Opportunities to Big Data:

Now this is Data Revolution time. Big Data is giving so many opportunities to business organizations to grow their business to higher profit level. Not only in technology but big data is playing an important role in every field like health, economics, banking, and corporates as well as in government.

1) *Technology*: Almost every top organization like Facebook, IBM, yahoo have adopted Big Data and are investing on big data. From these stats we can say that there are a lot of opportunities on internet, social media.
2) *Government*:Big data can be used to handle the problems faced by the government.Big data analysis played an important role of BJP winning the elections in 2014 and Indian government is applying big data analysis in Indian electorate.
3) *Healthcare*:According to IBM Big data for Healthcare, 80% of medical data is unstructured. Healthcare organizations are adapting big data technology to get the complete information about a patient.
4) *Science and Research*: Big Data is the latest topic for research. There are so many papers being published on big data.
5) *Media*:Media is using big data for the promotions and selling of products by targeting the interest of the user on internet.

VI.APPLICATIONS IN DATA MINING

Big Data is very useful for Business Organizations as well as to the researchers to observe the data patterns in big data sets. Extracting useful information from large amount of big data is called as Data Mining.40 Zettabytes of data will be created by 2020 which is 300 times from 2005. be created by 2020 which is 300 times from 2005 [3]. To analyze this data to get useful information for security, health, education etc., we need to introduce new data mining system which is effective. There are many Data mining techniques which can be used with big data, some of them are:

A. *Classification Analysis*: It is a systematic process for obtaining important information about data and metadata. Classification can also be used to cluster the data.

B. *Cluster Analysis*: It is the process to identify data sets that are similar to each other. This is done to get the similarities and differences within the data. For example, clusters of customers having similar preferences can be targeted on social medial [6].

C. *Evolution Analysis*:It is also called as genetic data mining mainly used to mine data from DNA sequences. But can be used in Banking, to predict the Stock exchange by previous years' time series Data [7].

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.

VII.CONCLUSION

We have entered an era of Big Data. The paper describes the concept of Big Data along with 3 Vs, Volume, Velocity and variety of Big Data.An overview to big data challenges is given and various opportunities and applications of big data has been discussed. This paper describes the Hadoop Framework and its components HDFS and Map reduce.This paper also focuses on current researches in Data Mining. The paper describes Hadoop which is an open source software used for processing of Big Data.

REFERENCES

[1] 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.

[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] MrigankMridul, AkashdeepKhajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" in International Journal of Advance Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[5] Apache Hadoop Project, http://hadoop.apache.org/, 2013.

[6] 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. ISSN09758887; pp- 35-37

[7] Mucherino A. PetraqpapajorgjiP.M.Paradalos 1998. A survey of data mining techniques alied to agriculture CRPIT.3(3): 555560.

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

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

[10] Vidyasagar S. D, A Study on "Role of Hadoop in Information Technology era", GRA - GLOBAL RESEARCH ANALYSIS, Volume : 2 | Issue : 2 | Feb 2013 • ISSN No 2277– 8160.

[11] BIG DATA: Challenges and opportunities, Infosys Lab Briefings, Vol 11 No 1, 2013.

[12] Divyakant Agrawal, Challenges and Opportunities with Big Data, A community white paper developed by leading researchers across the United States.

[13] 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.