International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 08 | Aug -2017 www.irjet.net p-ISSN: 2395-0072

BIG DATA WITH HADOOP – FOR DATA MANAGEMENT, PROCESSING AND STORING

Revathi.V[1], Rakshitha.K.R[2], Sruthi.K[3], Guruprasaath.S[4]

[1]*Assistant Professor, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Coimbatore, India*
[2,3,4] *III BCA 'A', Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Coimbatore, India*

Abstract - *Big data is a term that describes a large amount of data that are generated from every digital and social media exchange. Its size and rate of growth make it more complex to maintain for this purpose Hadoop technology can be used. Hadoop is an open source software project that enables the distributed processing of big data sets across clusters of commodity servers. It is planned to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Along with Hadoop, some techniques are also used to handle the massive amount of data. Some commonly used techniques are Map Reduce, HDFS, Apache Hive etc.*

Key Words: **Big data, Hadoop, HDFS, Map Reduce, Hadoop Components**

1. INTRODUCTION

Companies across the globe have been using data for a long time to help them make better decisions in order to enhance their performances. It is the inaugural decade of the 21st century that actually showcased a rapid shift in the availability of data and its applicability for improving the overall effectiveness of the organization. This change that was to revolutionize the role of data brought into advent the concept that became popular as Big Data. Big data is a large amount of data that floods from the daily uses of human life like telephone, mobiles, traveling, shopping, and computer, organization which evokes complication for storing, processing or accessing data using a traditional database. This is in all likelihood one of the important reasons why the concept of Big data was first embraced by online firms like Google, eBay, Facebook, Linkedin, etc., as these organizations were constructed with the foundation concept of using rapidly changing data. They did not plausibly face the challenge of integrating the raw and unstructured data with the already available ones. Since HADOOP has emerged as a popular tool for BIG DATA to analyze these data using map-reduce to get the desired output with the help of programs. It is a freely available platform to perform these type of performance. The design purpose of it is to perform millions of data in a single executive server. [1]

2. CHARACTERISTICS OF DATA

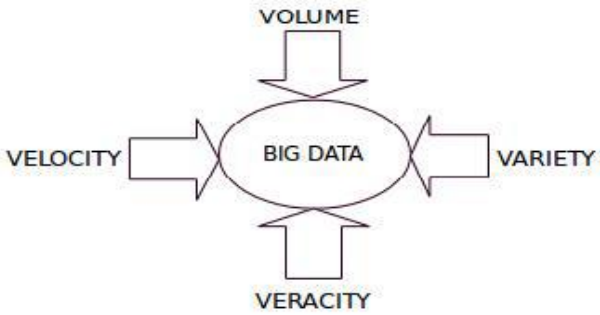

Fig -1: Four V's of Big Data

- **Volume**

Volume refers to the amount of data. The size of the data is represented as a volume. The size of the information is represented in terabytes and petabytes.

- **Variety**

Variety forms the data too big. The files arrive in various formats and of any type, it may be structured or unstructured, such as text, audio, videos, log files and more.

- **Velocity**

Velocity refers to the speed of information processing. The data adds up at high speed. Sometimes 1 minute is also late, so big data are time sensitive.

- **Value**

The possible value of Big data is huge. Value is the main root for big data because it is important for businesses, IT infrastructure system to store a lot of values in the database.

- **Veracity**

Veracity refers to noise, biases, and abnormality when we are handling with high volume, velocity, and a variety of data. All data not going to be 100% correct, there will be dirty data. [2]

3. TECHNOLOGY USED IN BIG DATA

For the determination of processing a large amount of data, the big data require exceptional technologies. The various techniques and technologies have been introduced for manipulating, examining, and visualizing the big data. On that point are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies.

3.1 HADOOP

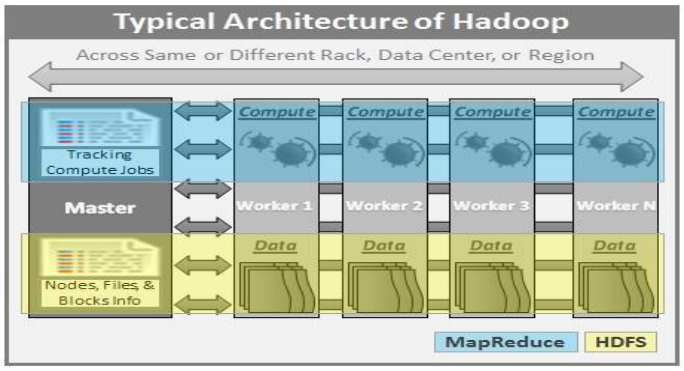

Fig -2: Hadoop architecture

Hadoop is an open source project which is developed for maintaining big data. This is implemented by Apache Software Foundation. It consists of many small sub projects which belong to the category of infrastructure for distributed computing. Hadoop mainly consists of:

- File System (The Hadoop File System [HDFS]) - Storage part
- Programming Paradigm (Map Reduce) - Processing part

The other sub projects allow correlative services or they are built along the core to add higher-level abstractions. There exist several problems in dealing with storage of vast amounts of data. Though the memory capacities of the drives have expanded massively, the rate of reading data from them hasn't shown that considerable progress. The reading process takes a vast amount of time and the process of writing is also slower. The time can be decreased by taking from multiple disks at once. Just using one hundredth of a disk may seem wasteful. Only if there are one hundred datasets, each of which is one terabyte and providing shared access to them is also a solution. There also exist many problems with using several pieces of hardware as it increases the chances of failure. Replication method that is creating redundant copies of similar data on the different devices so that if any failure occurs the copy of the data will be available. The primary problem is combining the data being read from different machines. There are so many methods are able in distributed computing to handle this problem, but even so, it is quite challenging. All the complication discussed is well managed by Hadoop. Hadoop Distributed File System identifies the failure problem and the Map reduce programming paradigm identifies the problem of combining data. Map Reduce diminishes the complication of disk records and writes by giving a programming model dealing in computation with keys and values. [3]

4. HDFS (Hadoop Distributed File System)

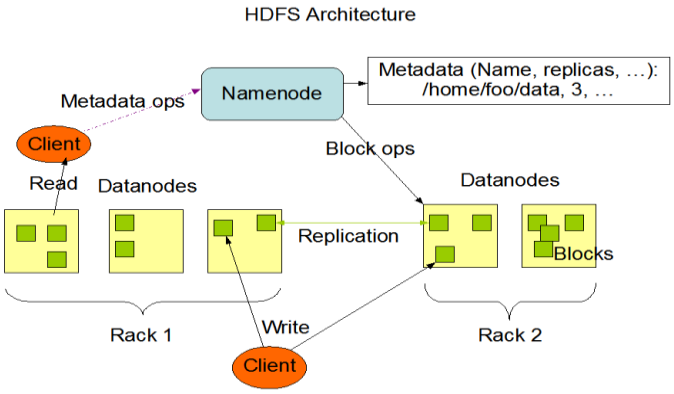

Fig -3: HDFS system

HDFS is a file system which is constructed of storing a very large number of files by using streaming data access rule by running clusters on commodity hardware. HDFS manages storage on the cluster by breaking incoming files into pieces called blocks and string each block redundantly across the pocket billiards on the server. HDFS stores, three copies of each file by copying each piece to three different hosts. The size of each block 64MB. HDFS is divided into three modes which are Name node, Data Node, HDFS Clients/Edge Node.

- **Name node**

It is a centrally placed node, which holds information about the Hadoop file system. The principal task of name node is that it records all the metadata & attributes and specific locations of files & data blocks in the data nodes. Name node acts as the master node as it stores all the information around the system and provides information which is newly added, modified and removed from data nodes.

- **Data node**

It functions as a slave node. Hadoop environment may comprise more than one data node based on capacity and performance. A data node performs two main tasks storing a block in HDFS and acts as the political program for running jobs.

- **HDFS Clients/Edge node**

HDFS Clients sometimes also know as Edge node. It works as a linker between name node and data nodes. Hadoop cluster at that point is only one client, but there are also many depending upon performance needs. [4]

5. MAP REDUCE

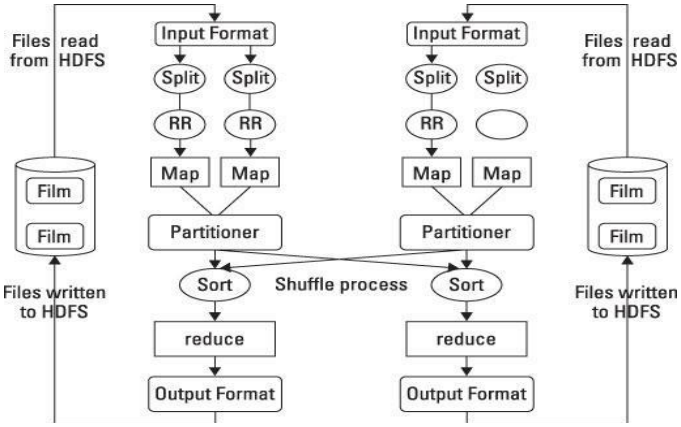

Fig -4: Map Reduce

Hadoop Map Reduce is an implementation of the algorithm developed and managed by the Apache Hadoop project. Map Reduce (map + reduce) Is a theoretical explanation for writing applications that treat large amounts of structured and unstructured data stored in the Hadoop Distributed File System (HDFS). Map reduce is an instrument which is very crucial to the analysis of the data which are structured or unstructured. In which the data are splits in the minuscule parts of the cluster to take the original form of the data. Map reduce is 100 times quicker than the Spark and 1000 times quicker than the Drill, now Drill is also working platform.
For EX: CPU is the mental capacity of computer like that Map reduction is the heart of the Hadoop. [5]

6. BIG DATA MAPPING USING MAP REDUCE TECHNIQUE

➢ Get the big data ready

When a request is made by the client to run Map Reduce program, the initial step is to search and understand about the input file. The file pattern is a random type so it is necessary to change data that can be processed by the data. This process of input process and Record reader. Input Format chooses how the file is going to be smashed into smaller pieces for processing using a function called Input Split. It then assigns a Record Reader to transform the raw data for processing by the map. Dissimilar types of Record Readers are supplied with Hadoop, offering a wide variety of conversion options.

➢ Let the Big Data Map Begin

Client's information is now in a mode tolerable to map. For all input pairs, a specific instance of a map is called to swear out the data for each input pair. Map and reduce the need to work together to maintain the data, the program should collect output information from the separate mappers and pass it to the reducers. This action can be performed by an Output Collector. A Reporter function also present information collected from map tasks. This entire project is being done on multiple nodes in the Hadoop cluster simultaneously. Later all the map tasks are complete, the common results are collected in the partition and a shuffling occurs, sorting the output for excellent processing by reducing.

➢ Reduce and combine for big data

For each output pair, Reduce is called to do its task. In a similar pattern to map, reduce collect its output while all the mappings are processed. Reduce can't begin as far as all the mapping is made out. The output of reducing is also a key and a value. Hadoop provides an Output Format feature, and it performs very much like Input Format. Output Format takes the key-value pair and formulates the output for writing to HDFS. The last task is performed to write the data to HDFS. Record Writer is used for writing and also to perform similarly as Record Reader except in reverse. It uses up the Output Format data and writes it to HDFS. [6]

7. HADOOP ECOSYSTEM

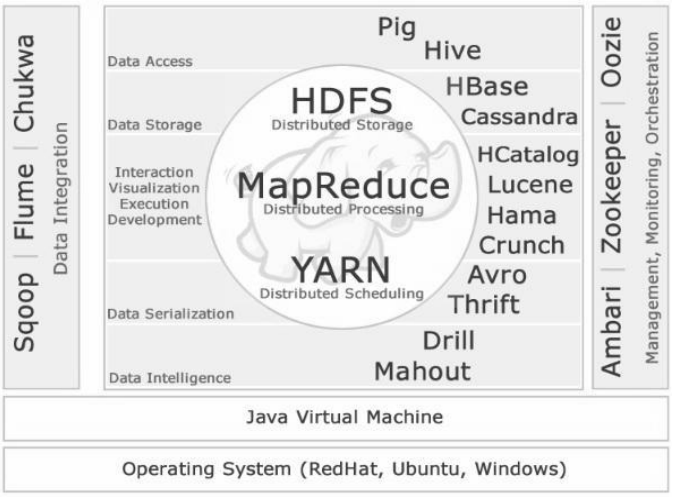

Fig -5: Components of Hadoop

7.1 DATA MANAGEMENT

✓ Ambari

It allows installation of services, so that Pig, Hive, scoop, HBase can be picked and put in. It will work across all the nodes in the cluster and also can manage the services from one centralized location like starting up, stopping, reconfiguring.

✓ Zookeeper

Zookeeper is a distributed coordination and setting up service for Hadoop cluster It is a centralized service that

provides distributed synchronization and providing group services and maintains the configuration information. With Hadoop, this will be useful to track if a particular client is down and plan necessary communication protocol around node failure.

- ✓ **Oozie**

It is a workflow library that permits us to play and to connect lots of those essential projects for instances, Pig, Hive, and Sqoop.

7.2 DATA INTEGRATION

- ✓ **Sqoop**

Sqoop is a command-line interface platform which can be employed to transfer the data from relational database environments like Oracle, MySQL, and PostgreSQL into Hadoop environment.

- ✓ **Flume and Chukwa**

Various applications, operating systems, web-services generate plenty of log data, for processing these logs a framework can be developed by using Flume and Chukwa tools. It is a way to push real time data, information right into Hadoop and also to execute real time analysis.

7.3 DATA ACCESS

- ✓ **Pig**

Pig is a platform for processing and maintaining data with easy techniques. By just writing half dozen lines of code pig can process terabytes of data.

- ✓ **Hive**

Hive is Data warehousing application that provides the SQL interface to process and analyze the big data stored in HDFS. Hive infrastructure is made along the top of Hadoop that help in providing summarization, query, and analysis.

7.4 DATA STORAGE

- ✓ **HBase**

HBase is a distributed column oriented database where as HDFS is a file system. Simply it is built on top of HDFS system. It is completely written in Java, which is a non-relational, distributed database system. It can suffice as the input and output for the MapReduce. For example, read and write operations affect all rows but just a small subset of all columns.

- ✓ **Cassandra**

Apache Cassandra is a no SQL database for managing a heavy quantity of data across many commodity servers with high availability and no single instance of failure.

7.5 INTERACTION, VISUALIZATION, EXECUTION, DEVELOPMENT

- ✓ **Hcatalog**

It is known as metadata table and storage management scheme. It is a way for tools like a Pig, Hive for interoperable also to hold a consistent view of data across those tools.

- ✓ **Lucene**

Lucent is there for full text searching an API loaded with algorithms to answer things like standard full-text searching, wild card searching, phrase searching, range searching kind of stuff.

- ✓ **Hama**

Hama is there for BSP (Book Synchronous Processing). To turn with a large amount of scientific data home is used.

- ✓ **Crunch**

Crunch is there for writing and testing and running map reduces pipeline. It basically gives full control to overall four phrases, which is going like 1. Map, 2. Reduce, 3. Shuffle, and 4. Combine. It is there to assist in joining and aggregation that is very hard to do with low-level map reduce, so Crunch is there to make you little easier inside map reduce pipeline.

7.6 DATA SERIALIZATION

- ✓ **Avro**

Avro is a data serialization format which brings data interoperability among multiple components of Apache Hadoop. Most of the components in Hadoop started supporting the Avro data format. It starts with the basic premise of data produced by component should be readily consumed by another component Avro has following features Rich data types, Fast and compact serialization, Support many programming languages like Java, Python.

- ✓ **Thrift**

It is more specific for creating flexible schemes that work with Hadoop data. It is specifically because it is intended for cross-language compatibility, so we build an application with Hadoop data in Java and if we want to use the same object in an application that you built on Ruby, Python or C++.

7.7 DATA INTELLIGENCE

✓ **Drill**

The drill is actually an incubator project and is designed to perform interactive analysis of nested data.

✓ **Mahout**

Mahout is a library for gaining knowledge about machines and data mining. It is divided into four different groups: collective filtering, categorization, clustering, and mining of parallel frequent patterns. The Mahout library belongs to the subset that can be performed in a distributed mode and can be executed by MapReduce. [7]

8. CHALLENGES

- **Heterogeneity and Incompleteness**

If we need to analyze the data, it should be structured, but when we deal with the Big Data, data may be structured or unstructured as well. Heterogeneity is the big challenge in data Analysis and analysts need to cope with it. Consider an example of a patient in the Hospital. We will arrive at each disc for each medical test. And we will likewise create a record for a hospital stay. This will be different for all patients. This figure is not well structured. Hence it goes with the team you manage heterogeneous and incomplete is required. A secure data analysis should be applied to this.

- **Scale**

As the name says Big Data is having a large size of data sets. Managing with large data sets has been a large problem for teens. Earlier, this difficulty was solved by the processors getting faster, but now data volumes are becoming huge and processors are static. The World is being given towards the Cloud technology, due to this shifted data is generated at a very high rate. This high rate of increasing data is becoming a challenging problem with the data analysts. 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 applied sciences. These are not at a slower rate like Hard disks, so the new storage scheme should be designed.

- **Timeliness**

Another challenge with the size is speed. If the data sets are heavy in size, longer the time it will take to analyze it. Any system which deals effectively with the size is likely to execute well in term of speed. There are cases when we need the analysis results directly. For example, If there is any fraud transaction, It should be dissected before the transaction is completed. And then some new system should be designed to meet this challenge in data analysis.

- **Privacy**

Privacy of data is another big problem with large data. In some nations, there are strict laws regarding the data privacy, for example, in the USA there are strict laws for health records, but for others, it is less forceful. For example, in social media, we cannot fix the private posts of users for sentiment analysis.

- **Human Collaborations**

In malice of the advanced computational models, there are many patterns that a computer cannot detect. A novel method of harnessing human ingenuity to solve a problem is crowdsourcing. The best example is Wikipedia. We are relying on the information given by the strangers, however, most of the time they are correct. But there may be other people who provide false information. For handling this kind of complexity we require a technology model to cope with this. As mankind, we can look the review of the book and find that some are positive and some are negative and come up with a decision whether to buy or not. We need systems to be that intelligent to determine. [8]

9. OPPORTUNITIES

- **Technology**

Nearly every top organization like Facebook, IBM, Yahoo has adopted Big Data and are investing in big data. Facebook handles 50 Billion photos of users. Every month Google handles 100 billion searches. From these stats, we can state that there are a lot of opportunities on internet, social media.

- **Government**

Big data can be applied to handle the problems faced by the government. Obama government announced a big data research and development initiative in 2012. Big data analysis played an important role of BJP winning the elections in 2014 and Indian government is putting on big data analysis in Indian elections.

- **Healthcare**

Many health care organizations are adopting big data techniques for deriving each and every information about patients. To ameliorate the health care and low down the cost big data analysis is required and certain technology should be adapted.

- **Science and Research**

Big data is the latest topic of inquiry. Many researchers are solving with big data. There are so many papers being printed on big data. NASA center for climate simulation stores 32 petabytes of observations.

- **Media**

Media is using big data for the promotions and selling of products by targeting the interest of the user on the net. For instance, social media posts, data analysts get the number of posts and then analyze the interest of the user. It can as well be done by getting the positive or negative reviews on the social media. [9]

10. CONCLUSION

Big Data (Hadoop) is in huge demand in the marketplace nowadays. As there a huge amount of data is lying in the industry, but there is no tool to wield it and Hadoop can implement on low-cost hardware and can be used by a large set of the audience on a large number of datasets. In Hadoop map reduce is the most important component in Hadoop. Many techniques are used to make the efficient plan for the map reduce so that it can speed up the system or data retrieval Technique like Quincy, Asynchronous Processing, Speculative Execution, Job Awareness, Delay Scheduling, Copy Compute Splitting had made the schedule effective for faster processing. [10]

REFERENCES

[1] Bijesh Dhyani, Anurag Barthwal, "Big Data Analytics using Hadoop", International Journal of Computer Applications (0975 – 8887) Volume 108 – No 12, December 2014

[2] Prof. Devendra P. Gadekar, Harshawardhan S. Bhosale, "A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications [IJSRP], Volume 4, Issue 10, October 2014 1 ISSN 2250-3153

[3] Ms. Gurpreet Kaur, Ms. Manpreet Kaur, "REVIEW PAPER ON BIG DATA USING HADOOP", International Journal of Computer Engineering & Technology (IJCET), Volume 6, Issue 12, Dec 2015, pp. 65-71, Article ID: IJCET_06_12_008 ISSN Print: 0976-6367 and ISSN Online: 0976–6375

[4] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chandler, "The Hadoop Distributed File System", October 2010, 978-1-4244-7153, IEEE

[5] Dr. Sanjay Srivastava, Swami Singh, Viplav Mandal, "The Big Data analytics with Hadoop", International Journal of Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue III, March 2016, ISSN: 2321-9653

[6] Varsha B. Bobade, "Survey Paper on Big Data and Hadoop", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 03, Issue: 01, Jan-2016

[7] Poonam S. Patil, Rajesh. N. Phursule, "Survey Paper on Big Data Processing and Hadoop Components", International Journal of Science and Research (IJSR), Volume 3 Issue 10, October 2014, ISSN (Online): 2319-7064

[8] Rahul Beakta, "Big Data And Hadoop: A Review Paper", RIEECE -2015, Volume 2, Spl. Issue 2 (2015), e-ISSN: 1694-2329, p-ISSN: 1694-2345

[9] Ivanilton Polatoa, Reginaldo Reb, Alfredo Goldman, Fabio Kona, "A Comprehensive View of Hadoop Research - A Systematic Literature Review", Journal of Network and Computer Applications, Volume 46, PP 1-25, November 2014

[10] Dr. Madhu Goel, Suman Arora, "Survey Paper on Scheduling in Hadoop", International Journal of Advanced Research in Computer Science and Software Engineering [IJARCSSE], Volume 4, Issue 5, May 2014, ISSN: 2277 128X