REVIEW PAPER ON HADOOP CONFIGURATION AND IMPLEMENTATION IN VIRTUAL CLOUD ENVIRONMENT

[1]Nasrullah, [2] Ms. Akanksha Bana
[1] M.Tech (SE), School of Engineering & Technology, Noida International University,
Plot 1, Yamuna Expy, Sector 17A, Uttar Pradesh 203201
[2]Assistant Professor, School of Engineering & Technology, Noida International University,
Plot 1, Yamuna Expy, Sector 17A, Uttar Pradesh 203201

Abstract: Big Data is the collection of large datasets which cannot be handled by traditional computing techniques, to store and process large dataset within a specific time limit, therefore Apache foundation developed hadoop. Which is an open source framework and properly used in virtualized cloud environment. It is still unclear what the correct configuration must be. To solve this issues, there are different organizations which implemented Hadoop core components HDFS and MapReduce for storage and processing large datasets. According to their own perspectives. Our literature review is mainly focused on hadoop configuration and implementation in virtualized cloud environment.

Index Terms-Big data, Hadoop, Hadoop Ecosystem, HDFS, MapReduce, Hadoop Configurations

1. INTRODUCTION

The size of data is growing in the rapid rate, and it is the biggest concern nowadays for organizations and researchers. Now a days, data and web data are increasing in terms of 10 to 100 petabytes per day which cannot be handled by the traditional database management systems. The researcher’s main focus how to handle this huge amount of data, and how to store and process vast amount of data within a specific time limit. Therefore Hadoop is a basic distributed system architecture which provide complete scalability and reliability. Hadoop is developed by Apache software foundation. The crucial parts of Hadoop include (HDFS) Hadoop Distributed File System, (MapReduce) is a data processing model and (Hbase) is a distributed column oriented database these 3 parts of the hadoop is open source implementation according with the given three cores techniques Hadoop (HDFS) Hadoop distributed file system is used to provide three times distributed pieces in different systems is connected to network the main advantages of hadoop configuration and implementation in virtualized cloud environment is that if one system goes down the two other system replicas are available and if we look to the structure its quite cheap, robust and fully fault tolerated. Hadoop is free open source framework for virtualized cloud computing environment it is also implemented by Google MapReduce framework. It is a popular framework which process and generate large data on cloud. And the apache hadoop software libraries is a framework which allow as for the processing a large dataset in distributed environment across the clusters computers using simple programing models this model is design to scale up and down from single servers to thousands of machines. And due to the outstanding advantages of Hadoop there are lots of applications based on Hadoop. Especially in the field of Internet Yahoo implemented hadoop clusters for web searching. Facebook implements Hadoop clusters for data analysis and machine learning. Taobao Hadoop system is used to store and handle the data of e-commerce transactions and data recovery for amazon Searching and there or some other organizations who implements hadoop clusters for their own perspectives.

2. Big data

Big data is a large dataset which cannot be handle by the traditional computing. Big data is not merely a data, but today it becomes a complete subject in computer science and Engineering fields, which involves various producers, tools and techniques, frameworks and materials. It is so big and large size that the traditional or conventional data processing applications are inadequate in handling these techniques.

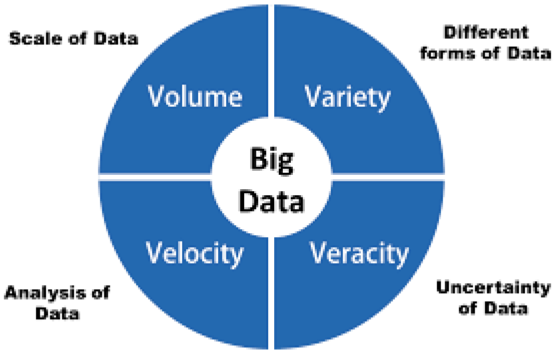

Fig 1-Four V's of big data

2.1 Variation and V's of big data are

1. **Volume:** Apache Hadoop provides system to self –jactitation and significant information sets to address such as a huge volume of information.
2. **Velocity:** Apache Hadoop handles increasing rates of data is generated, providing approaching from expansive information framework.
3. **Variety:** Hadoop bolster complex occupations to handle any mixture of structure, unstructured and semi structure information.
4. **Variability**: Apache Hadoop handles data which is complex according to different variation in size and form.

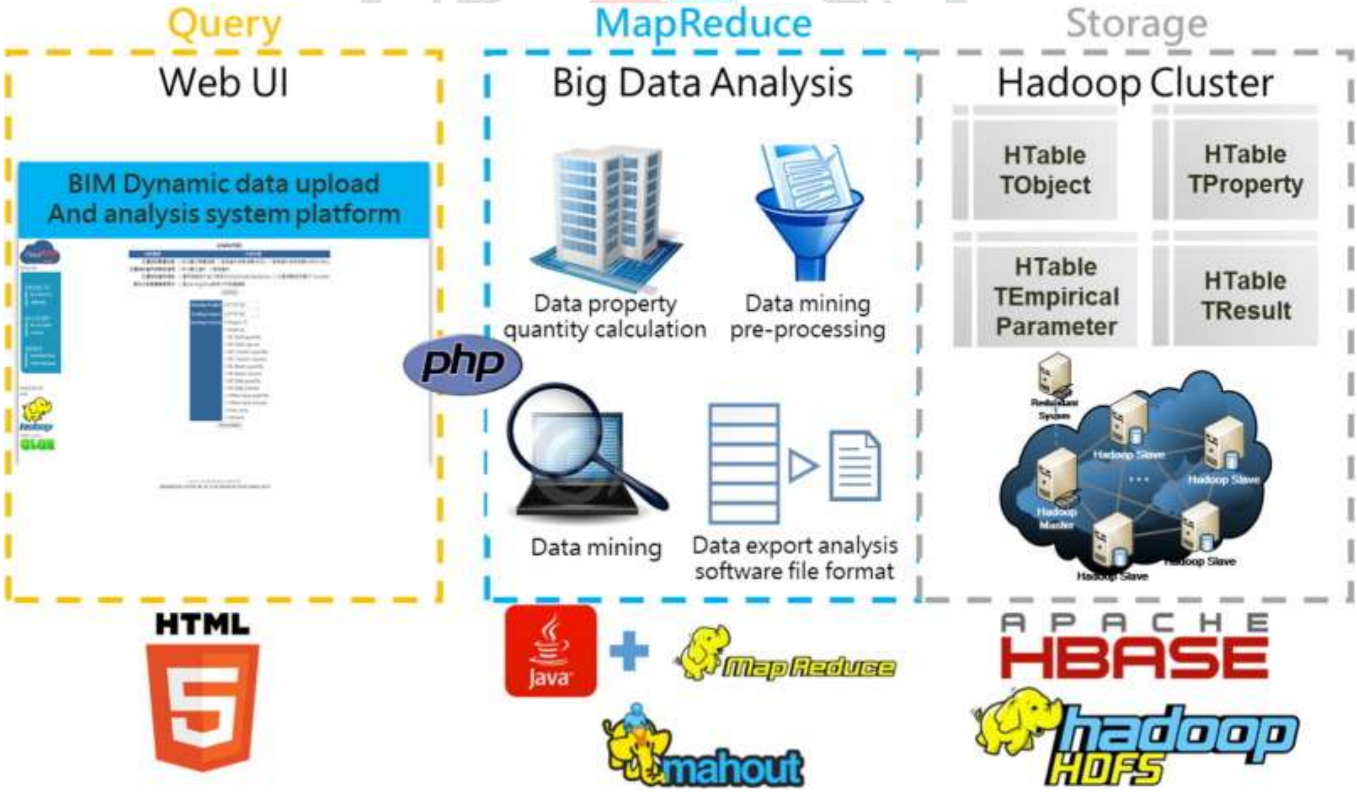

Figure 2-Architecture of Big Data System using HDFS and MapReduce

3. Hadoop Solution for Big Data Storage and Processing

The Apache foundation is used to introduce the completed project of Hadoop in Jan-2008. After Jan-2008, Hadoop is being used in many streams like government services, financial services, advertising, telecom, and many search engines. Like Google, Microsoft Azure, amazon ES3 etc. Hadoop is an open source framework for scalable, reliable, in the field of distributed computing. It is designed to connect from one single server up to 100th machines in which, each offers the local computation storage and processing. The Hadoop is a framework written in Java, The attributes of the big data can be define into four v's." i.e. Velocity, Volume, Varity and Variability. And the key attributes of Hadoop is redundant trustworthiness and reliability, that is if failure occurs in one machine, the replica will be available automatically and immediately the operator without need to do anything with data it is extremely powerful in terms of data storage and processing access and is preliminary batch processing centric and makes it easier to distributed applications using MapReduce software paradigm. Furthermore it can run one or more commodity hardware which removed the total cost and buying the expensive hardware.

3.1 Hadoop Ecosystem

The hadoop Ecosystem is used to top down structure in first version release of Hadoop 1.x in this system only two units which actually works in the environments. Ex. HDFS and MapReduce. The MapReduce process the data and result automatically get stored in the HDFS

Ecosystem of Hadoop has a top down approach. In earlier release of Hadoop 1.x there are only two units which actually works i.e. MapReduce and HDFS .MapReduce process the data and result automatically get stored in the HDFS. In the second version of Hadoop 2.x there are some new different compounds added to the ecosystem i.e. YARN (Yet another Resource Negotiator). The YARN is same like MapReduce but the way and structure of processing is little bit different. YARN processes the data in container. The containers are the logical units which consist of resource and task itself (here resource is Data Node). With the presence of YARN the deadlock situation which usually use to appear in 1.x are minimized in 2.x .In Hadoop ecosystem every data breakdown into the data chunks also known as the data blocks

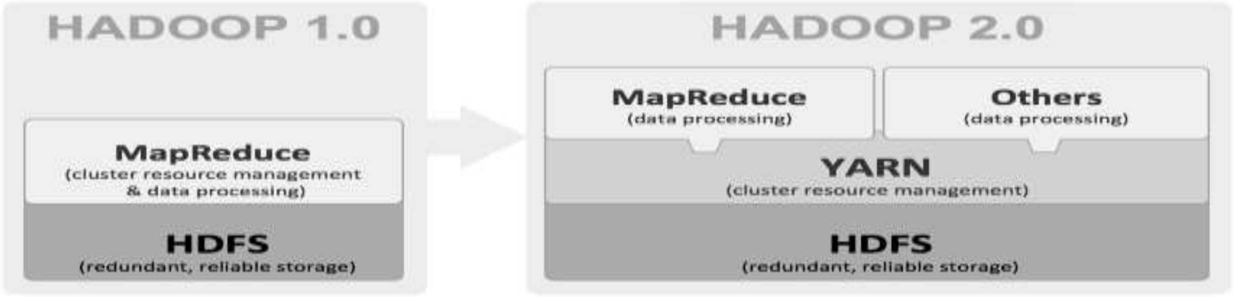
HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)

Figure 3- Hadoop1.0 and Hadoop 2.0 comparing with complete ecosystem

3.2 Hadoop has two major components: HDFS and MapReduce.

3.2.1 HDFS and its Architecture

The brain of Hadoop, and the bluestocking of Hadoop is popularly known as human brain stores data, same way (HDFS) store data in Hadoop. It accumulate the data in multiple nodes in a distributed manner. The main idea to have a distributed file system where data get stored in Datanode. HDFS has a golden rule "Read many times write once "A file can be read many times on a HDFS file system but it can only be written once. There is no need of physical data requirement. To configure this there is a need of multiple machine cluster, which is connected over LAN. These machines can be a commodity hardware also for creating a cluster. What HDFS do it get the instances or the collection of the Metadata of actual data if the User or Master gives an instruction from the Master Node, HDFS knows where the data resides in real, and it starts processing on that machine where the real data resides In persistence the master node will be free from the burden and can do rest of the things of cluster. After processing, the final output is stored over the HDFS. This is how the HDFS works in Hadoop. In addition, the cost of migration of data is saved for the industries who generates Petabytes of data quarterly basis

3.2.2. HDFS Architecture

HDFS was originally built as infrastructure for the Apache Nutch web search engine project. NameNode and Data Nodes HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, the Apache Hadoop Core project Cluster: it is a set of machines connected through LAN, MAN, WAN, VPC, Cloud etc.

NameNode: stores the metadata (like name, path and so on) and helps the client to commute with the Hadoop infrastructure.
Data Node: Data Node stores the actual data and with the help of processing unit it actually processes the data.

Block: The Related Applications that are compatible in the environment of HDFS are those deal with large data sets. These applications write their data only once but they read it one or more times and require these reads to be satisfied at streaming speeds. HDFS supports write-once-read-many semantics on files. Each typical block size is used by HDFS is 64 MB and 128 MB. Thus, an HDFS file is chopped up into 64 MB and 128MB chunks, and if possible, each chunk will reside on a different Data Node.
Data Replications: blocks of file are replicated for fault tolerance
Replica Replacement: A rack-ware replica replacement policy

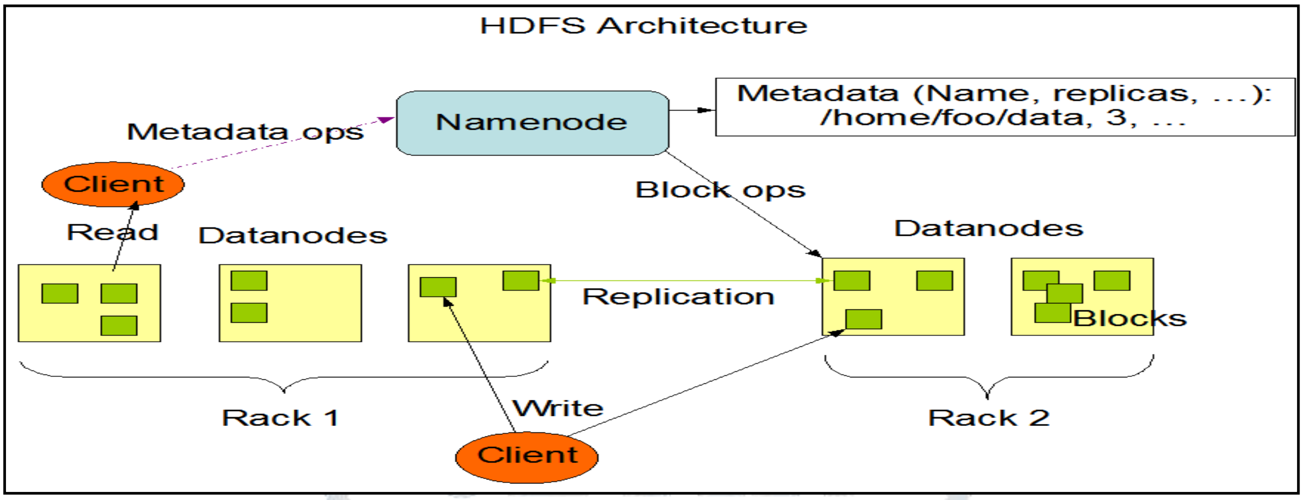

Figure 4 - HDFS files Architecture

3.2.3 MapReduce

MapReduce is a computing paradigm for processing data that resides on hundreds of computers, which has been popularized recently by Google, Yahoo, IBM, Amazon and others. This paradigm is extraordinarily powerful, but it does not provide a general solution to what many are calling "big data" so while it works particularly well known for some problems, some are more challenging. MapReduce is not a tool and it is more of a framework we have to fit our own complete solution in the framework to map and reduce, which in some situations might be challenging. It is not a feature, but rather a constraint. MapReduce framework has the total ability to process your data with distributed computing environment MapReduce is a programming model and a distributed computing framework to process extremely large data. It can be used to write auto scalable distributed applications in a cloud environment. In the MapReduce model, programmers have to reduce an algorithm into iterations of map and reduce functions known from Lisp and other functional programming languages. Writing an algorithm only consisting of these two functions can be a complicated task, but MapReduce framework is able to automatically scale and parallelize such algorithms. The framework takes care of partitioning the input data, scheduling, synchronizing and handling failures, allowing the programmers to focus more on developing the algorithms and less on the background tasks. Making it easier for programmers without extensive experience with parallel programming to write applications for large distributed systems. Hadoop MapReduce framework is written in Java programming language and is designed for applications processing vast amounts of data in parallel, in a cloud network. Because clouds often consist of commodity hardware, the MapReduce framework must provide reliable and fault tolerant execution of distributed applications. A running Map Reduce job consists of various phases such as: Map->Sort->Shuffle->Reduce**.**

3.2.4 Command –line Interface:

HDFS has many interfaces of interacting but the command line is the way, which is more convenient and most widespread way of interacting with the kernel of the Linux system by researchers and developers.
FS Shell: The File System Shell is the interactive way to communicate with the file system in Hadoop. It accesses the HDFS with some Linux based 'root' commands which help to do particular task over the HDFS. Few of the commands are as follows:
Hadoop fess –less /: to access the HDFS and list all the files/ directory present over the HDFS.
Hadoop fess –less /user: to get into the user directory over the HDFS.
Hadoop fess –cat /user/myfile.txt: to read the content of file (myfile.txt) present in the user directory over HDFS.
Hadoop fess –put /file source /final destination over the HDFS: this command is used to put any local file to HDFS
Hadoop fess -chow [-R] [OWNER] [: [GROUP]] URI [URI]: Change the owner of files. With -R, make the change recursively through the directory structure. The user must be a super-user.
Hadoop fess -cap URI [URI …] <dust>: Copy files from source to destination. This command allows multiple sources as well in which case the destination must be a directory. Example:

- hadoop fess -cap /user/hadoop/file1 /user/hadoop/file2
- hadoop-fess-cap/user/hadoop/file1 /user/hadoop/file2/user/hadoop/dir. :

Hadoop fess -midair <paths>Takes path Uri's as argument and creates directories. The behavior is much like UNIX midair –p .creating parent directories along the path.

Example-1:

- hadoop fess -midair /user/hadoop/dir1 /user/hadoop/dir2
- hadoop fess -midair hdfs://nn1.example.com/user/hadoop/dir hdfs://nn2.example.com/user/hadoop/dir

Hadoop fess -mv URI [URI …] <dust>: Moves files from source to destination. This command allows multiple sources as well in which case the destination needs to be a directory. Moving files across file systems is not permitted.

Example-2:

- hadoop fess -mv /user/hadoop/file1 /user/hadoop/file2
- hadoop fess -mv hdfs://nn.example.com/file1

hdfs://nn.example.com/file2 hdfs://nn.example.com/file3
hdfs://nn.example.com/dir1

Hadoop fess -put <locals> ... <dust>: Copy single sec, or multiple sacs from local file system to the destination file system. Also reads input from stein and writes to destination file system.

- hadoop fess -put local file /user/hadoop/hadoopfile
- hadoop fs -put localfile1 localfile2 /user/hadoop/hadoopdir
- hadoop fs -put localfile hdfs://nn.example.com/hadoop/hadoopfile
- hadoop fs -put - hdfs://nn.example.com/hadoop/hadoopfile

Reads the input from stdin.

4. Literature Review

Sunita, B. Aher et.al. Is mentioned that to implement MapReduce techniques of hadoop in the cloud environment and presented the complete steps to run program on Hadoop they execute the dataset consisting different combination subject of Computer Science and Engineering courses in Hadoop with the total explained result. Using MapReduce techniques in the cloud environment and running hadoop clusters is efficient solution to this business analytics problems. In the cloud large number of computer is connected through internet so how the hadoop make sense in the cloud are discussed.

Nikhil.Gupta. Hadoop will be the most required technology for the virtualized cloud. And this prove is given by the by the hadoop clusters offered by different virtualized cloud venders in many business and some of the most important steps is required making arrangement in a cloud distributed environment, single node hadoop cluster, backed by HDFS the most important steps for Hadoop configurations are

Step 1- First Install Ubuntu and configure java on it, then add hadoop dedicated system user

Step 2- Configure SHH and connect to the server, disable IPV6 and then installed hadoop

Step 3- After the SHH configuration make the directory where hadoop will stores files

Step 4- Start node cluster, copy Local data to HDFS and Run Map reduce job

Huang,Lu.et.al Hadoop is an open source implementation in virtualized cloud environment says that there are some existing problems in hadoop like single point of failure NameNode, HDFS small files problems, some bottlenecks issues and dispatcher is the pluggable module in hadoop according to user application requirement it can design their own dispatcher with the three task dispatcher are discussed. Hadoop performance and optimization can be improved by the MapReduce from the following aspect avoid unnecessary reduce task, Reused the writable task full in external file and add a combiner job too.

Jinto, Thomas et.al the work which is down by hadoop implementation in virtual machine is used to reduce the resource initialization for this reason we provide the solution and algorithm which gives an efficient resource initialization and utilization. This algorithm allow the cloud service provider to reduce the cost and total reduce the distribution efficiency to first response time for the significant MapReduce workload.

Dharmik, H. the security works never end, hadoop is a secure system there are different procedure and guidelines to make hadoop more secure for cloud using hadoop to provide more flexibility authentication and authorization to improving data production working on the Kerberos protocols authorization where the description of the TGT plays complete roles and securely implement hadoop in the virtualized cloud environment

E.Rabi, P. Says that in the cloud environment well-known data processing engine for big data is hadoop to process huge number of datasets, hadoop MapReduce consist of all data processing functions, components and application of hadoop which discussed. And provided overview for implementation in cloud environment. The complete execution process and programming model in MapReduce the job consist map and reduce function how these function called there are some steps take actions in the places.

Madhavi, V. et.al the author is try to identify the performance issued and configuration in HDFS on heterogynous number of clusters and also completely suggested for data placement mechanism in HDFS this mechanism is provide how to distribute fragments of an input file to heterogeneous node which base on the computing capacities for this manner we also provide security for the data which is need for processing.

Srihari R et.al mentioned that hadoop framework allow different types of users which can run hadoop clusters across the number of machines and also manage them there experimental evaluation V-Hadoop allows the elastics which can utilized the of the underline physical infrastructures and provide complete management time and cost efficient to all users who implement hadoop in virtualized cloud environment.

5. Conclusion

This paper provide an overview of Big data, Hadoop, MapReduce and command line hadoop configuration and 4 V's of big data is discussed. The different issues and challenges of Hadoop are introduce here and then for them what are the solutions which are proposed in different papers. The data has introduce whole new attribute for storing, processing and analysis. For providing the new opportunity and solutions of the real word problems and how to implement generic configuration of Hadoop from small to huge organizations. This paper also discus Hadoop components with different tools and techniques and how to use it in virtualized cloud environment. HDFS and MapReduce are the core components of Hadoop we have configured this two core components by using command line interface. Many organization provides their own hadoop distribution and configuration in the virtual cloud infrastructures and virtualization is enable on demand clusters Hadoop is also adapted the wide areas form social life, engineering, science and has change the way of thinking and solving real world problems.

References

[1] A. Sunita.B, Ather, and Anita,R. Kulkarni "Hadoop MapReduce: A programming Model for large Data Processing", American Journal of computer Science and Engineering Survey, AJCS, pp. 001-010, 2014

[2] B. Nikhil, G. "Cloud computing Techniques for Big Data and Hadoop Implementation", International Journal of Engineering Research & Technology (IJERT), vol. 3,(4), April-2014

[3] C. Huang Lu, Hu, Ting-T, and Hen Hai, S. "Research on Hadoop Cloud Computing Model and its Application", Third International Conference on Networking and Distributed Computing, IEEE, 2012.

[4] D. Jinto.T, Pawan.M,"Efficient Resource Utilization in Hadoop on Virtual Machine", International Journal of computer science and Mobile computing, vol.4, 2, pg. 965-569, April- 2015

[5] E. Dharmik, H. "A Survey on Data Security System for Cloud Using Hadoop", International Journal of Innovative Research in Computer and Communication Engineering, vol.4, Issue 11, 2016.

[6] F. Rabi. P. Panday. "Big Data Processing with Hadoop-Map Reduce in Cloud System", International Journal of Cloud Computing and Services Science.IJCST, vol., 2, 2016

[7] G. Madhavi, V, and DR, Shrinvivas, D. "Critical Study of Hadoop Implementation and Performance Issues", Computer society of India, pp. 03-10, 2013.

[8] H. H.Shihari, R. Bryan, J. Khuzaimam, D. "V-Hadoop Virtualized Hadoop Using Containers", Duke Computer Science Colloquium, pp. 03-08, 2018.