A PAPER ON HADOOP FILE SYSTEM (HDFS), BIG DATA AND MAP REDUCE

R.Deepa[1], S.Vaishnavi[2]
Department of Information technology, PSG College of Arts and Science, Coimbatore
Department of Information technology, PSG College of Arts and Science, Coimbatore

ABSTRACT

The word "Big Data" explains featuring methods and skills to record, store, distribute, be in change analyze petabyte or bigger sized datasets with a great velocity and various structures. Data is produced from many different sources and can place in the system at various measures. In order to process the large amount of data in a budget and maximum way connecting that is used. Big data is data whose 1) Scale 2) diversity 3) complexity requires new architecture methods 4) algorithms and analytics in change at extract value and hidden knowledge from it. Hadoop is an open source software project which provokes the divide up processing of big data sets across clusters of useful nodes. It has designed and developed to scale up from a one server to thousands of machines device with extremely high degree blame of ability. Hadoop is quickly emerged as establishment of huge data for processing tasks and planning of scales and process number of volumes of sensors data which includes from web of things sensor.

Keywords:* Big Data, *Hadoop,*Map Reduce,*HDFS.

1. INTRODUCTION

A: Definition: Big Data

Elaborates the data sets or combination of data sets whose volume, variability and velocity makes them tough to record, merge , processed by conventional technologies and tools such as relational databases and statics surfaces or packages of visualizations within the time period to make them needful. While the size is used to control whether a particular data set is considered big data which is not securely explained and continuous to change over time, most analysts recently refers terabytes to data sets from 30 - 50 terabytes to multiple petabytes as big data. Big Data Concerns are

i)Data privacy is the Huge Information we now produce contains a ton of data about our own lives, quite a bit of which we have a privilege to keep private. Progressively, we are requested to strike a harmony between the measure of individual information we uncover, and the comfort that Enormous Information controlled applications and administrations offer.

ii)Data security is even we decide we are happy for someone to have our data for a particular purpose. It is framed into three layers including infrastructure layer, computing layer, and application layer from top to bottom.

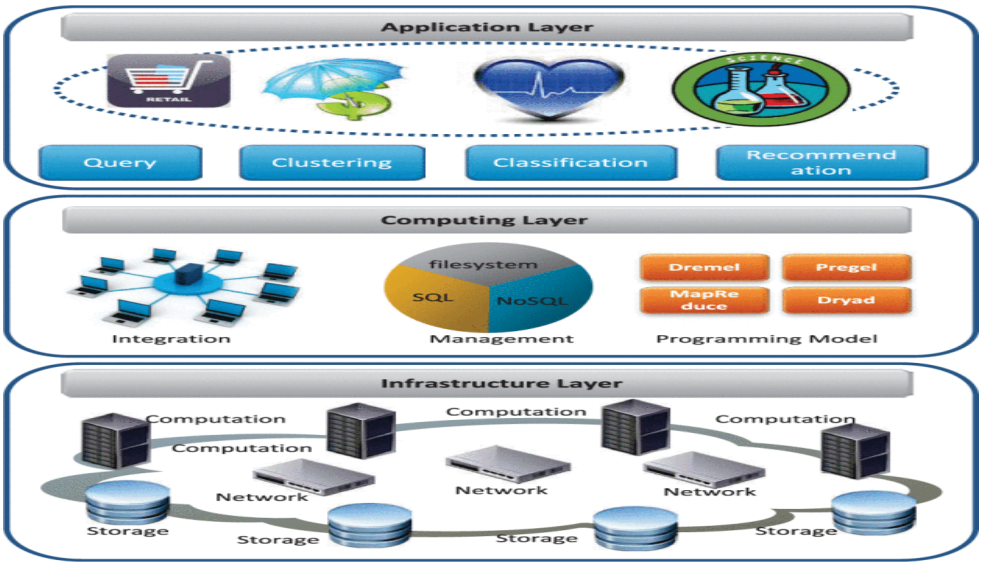

Fig1: Layered Architecture of Big Data System

b. Three V's of big data

Volume of Data: Volume which refers to enormous amount of data. It used to be representatives made information. Since information is produced by machines, systems and human cooperation on frameworks like online networking the volume of information to be dissected is huge. The volume of data is store in storehouse for undertaking in natural development.

Variety of Data: Variety refers the numerous sources and kinds of information both organized and unstructured. We used to store information from sources like spreadsheets and databases. Presently information comes as bases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, photographs, recordings, and so forth. This assortment of unstructured information makes issues for capacity, mining and dissecting information.

Velocity of Data: velocity refers to speed of processing for sensitive period of time process such a catching false and large data which must be used it streams into your enterprise in order to extend its value. The stream of data is gigantic and persistent.

C. Big Data Processing Occurrence with Problem

i) Dissimilarity and Incompleteness

When human grasp information it is good deal of dissimilarity is compact ably tolerated. Machine analysis algorithms accepts homogeneous data must be carefully structured as first step data analysis, computer system have to work toughly if it can be store many items that has size and structure is identically.

ii) Scale

Big data is about size and managing huge and quick increasing volume of data has been demanding issue of various decades. This challenge was reduced by processors by getting quicker, following law of Moore's to provide us with resources needed to cope with increasing the data volume. These are basic shift under process now. Data volume has faster scaling and computes resources and statics with CPU seeds.

iii) Timeliness

The huge data set is processed, and it analysis longer. The plan of device which deals with size also results in a system that can process in given size velocity is the context of Big data.

iv) Privacy

The privacy of data is another is another large concern and increases in the context of Big data. For the purpose of records there are strict governing laws and cannot be done. Maintaining and managing isolated which is effectively on both technical and sociological which must be labeled connection from both approaches to realize the big data. The data privacy is another large concern and increases context of big data.

V) Human Association

Note the refinements center around three specific human practices: 1 .sharing learning 2learning 3. Building accord Although, the enormous advances made in analyzing the computational which remains various patterns were humans can easily detect but algorithm of computer have a tough time finding analytics for big data will not be all computational, relatively it is been designed explicitly to have a human in the loop the recent sub field of visual analytics is attempting last with modeling and analysis phase in pipeline. A big data analysis system must support the income from multiple human experts and shared results with exploration. When hikes all team together in one room. The data machine has to acquire distributed expert income and support of their collaboration.

2. SOLUTION FOR BIG DATA PROCESSING

Hadoop is a Structure for Programming the uses to support the processing huge data sets in distributed computing environment. Hadoop was developed by Google's

MapReduce that is a software framework were an application break down into many parts. Hadoop kernel and MapReduce are consisted by the current Apache Hadoop Ecosystem. HDFS consist of many types of parts like 1) Apache Hive 2) Base and 3) Zookeeper. HDFS and MapReduce are explained in the following.

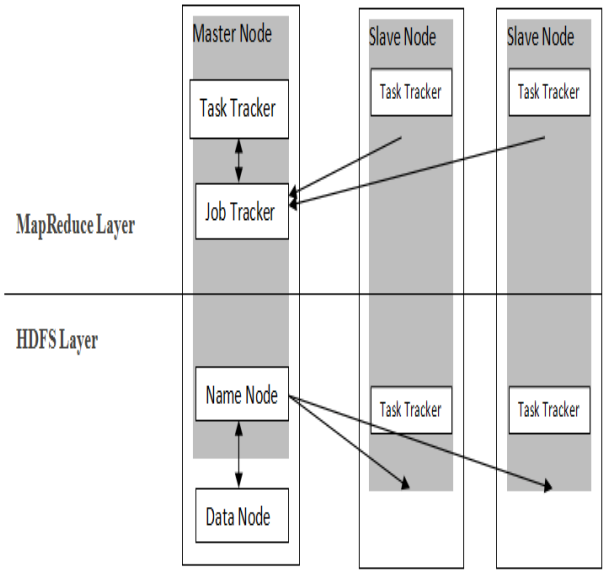

Figure 2. Hadoop Architecture

A. HDFS Architecture (Hadoop File System)

HDFS is the globe most reliable storage system. HDFS stores very large files running on cluster of commodity hardware. It works based on the principle of storing less number of larger files rather than huge number of smaller files. HDFS stores data reliably even when hardware HDFS includes in incrementally scale up and survive failure of significance parts of storage and it creates cluster of machines and it works cordially among them. Hadoop continuous works without losing data or interrupting during it works. By shifting work on the balance machines in the cluster. HDFS manages the storage on cluster by breaking the incoming files into pieces which are known as blocks and stores each the blocks redundantly access the servers. HDFS stores three full copies of file copying each part of 3 different servers.

HDFS Architecture Goals

I) Failure of Hardware:

Hardware failure is number of extension there exist huge number of components which are very predictable to hardware failure there are few components which is non-functional always so core architectural goal of HDFS is quick and automatic fault detection recovery. Empowers high throughput information get to. A record once created and written it shuts the need not be changed aside from

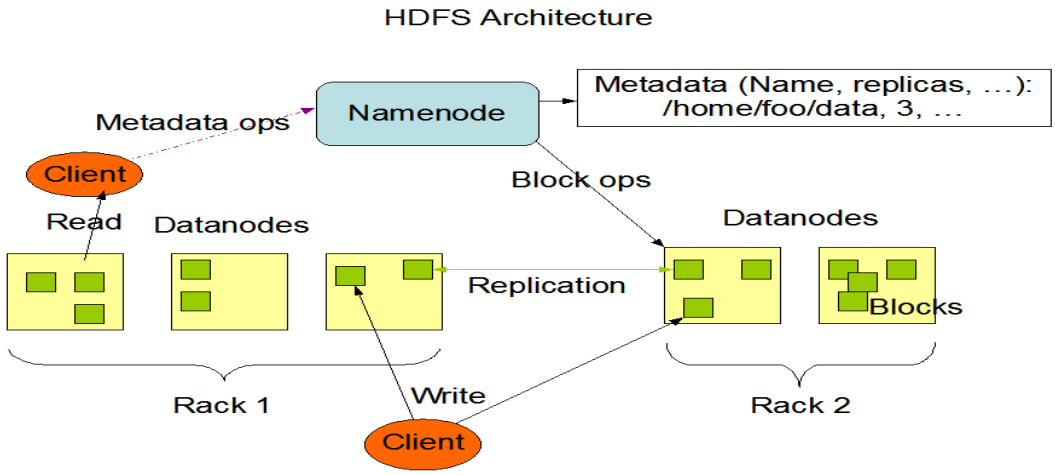

Figure 3: HDFS Architecture

II) Streaming Data Access

HDFS applications need to access for streaming of their datasets .HDFS is designed for batch processing rather than interactive use by users. Where the force is higher than data less access latency.

III) Large Datasets

HDFS always works with large datasets. It ranges from gigabytes to terabytes.

IV) Simply Coherence Model

A record once created and written it shuts the need not be changed aside from attaches. This presumption disentangles information coherency issues and empowers high throughput information get to. The simple coherency of data issues and empowers high throughput of data. A MapReduce Based Application (or) Web crawler fits well in the model.

V) Moving calculation is less expensive than Moving Data

In the event that an application does the computation close to the data it works on, it is significantly more productive than done far of. This reality ends up more grounded while managing vast data collection. The fundamental favorable position is that this expands the general throughput of the framework. It likewise limits arrange blockage. The supposition is that it is smarter to draw calculation nearer to data as opposed to moving data to calculation.

VI) Portability crosswise over heterogeneous Hardware and Software and platforms

HDFS is planned with convenient property to be compact from one stage to other. This empowers across the board reception of HDFS. It is the best stage while managing an expansive arrangement of data. Moving calculation is less expensive than moving data.

Assignment of NameNode

- Oversee document framework namespace.
- Manages customer's entrance to records.
- It additionally executes record framework execution, for example, naming, closing, opening files/directories.
- All DataNodes sends a Pulse and piece answer to the NameNode in the Hadoop bunch. It guarantees that the DataNodes are alive. A square report
- All DataNodes sends a Pulse and piece answer to the NameNode in the Hadoop bunch. It guarantees that the DataNodes are alive. A square report
  contains a rundown of all block on a datanode.
- NameNode is additionally in charge of dealing with the Replication Factor of the considerable number of blocks.

Task of DataNode

- Block replica creation, deletion, and replication according to the instruction of Namenode.
- DataNode manages data storage of the system.
- iii)DataNodes send heartbeat to the NameNode to report the health of HDFS.
- By default, this frequency is set to 3seconds.

B. Map Reduce Architecture

The processing support in the Hadoop ecosystem is MapReduce framework . the framework allows the operation specification to apply the huge data set to divide the problem and data and run it parallel. For Example: A very large dataset can reduce into smaller subset, applying the analytics. MapReduce programs are composed in a specific style affected by functional programming builds, specifical expressions for preparing arrangements of information. Here in MapReduce, we get contributions from a rundown and it changes over it into yield which is again a rundown. It is the core of Hadoop. Hadoop is so much effective and productive because of MapReduce as here parallel handling is finished.

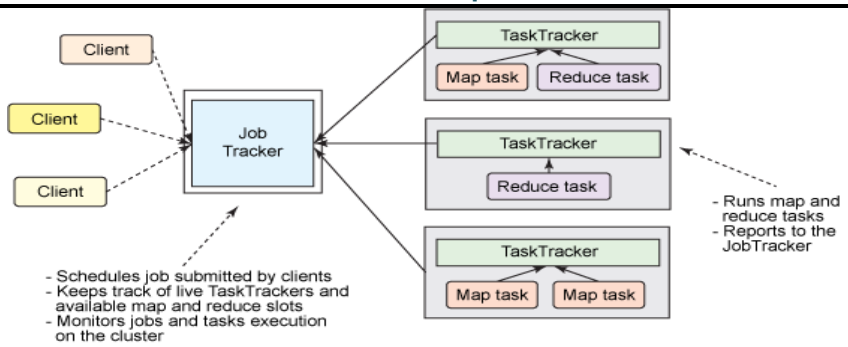

Figure 4: MapReduce Architecture

Apache MapReduce Terminologies

- How about we now comprehend distinctive phrasings and ideas of MapReduce, what is Map and Reduce, what is an job, task, task attempt and so forth.
- MapReduce is the information handling segment of Hadoop. MapReduce programs change arrangements of info data components into arrangements of yield data components. A MapReduce program will do this twice, utilizing two distinctive rundown preparing expressions.

How MapReduce Works

Input data given to **mapper** is processed through user defined function written at mapper. All the required complex business logic is implemented at the mapper level so that heavy processing is done by the mapper in parallel as the number of mappers is much more than the number of **reducers**. Mapper creates a yield which is middle of the road information and this yield goes as contribution to reducer. This transitional outcome is then handled by client characterized work composed at reducer and last yield is produced. More often than not, in reducer light handling is finished. This last yield is put away in HDFS and replication is done obviously.

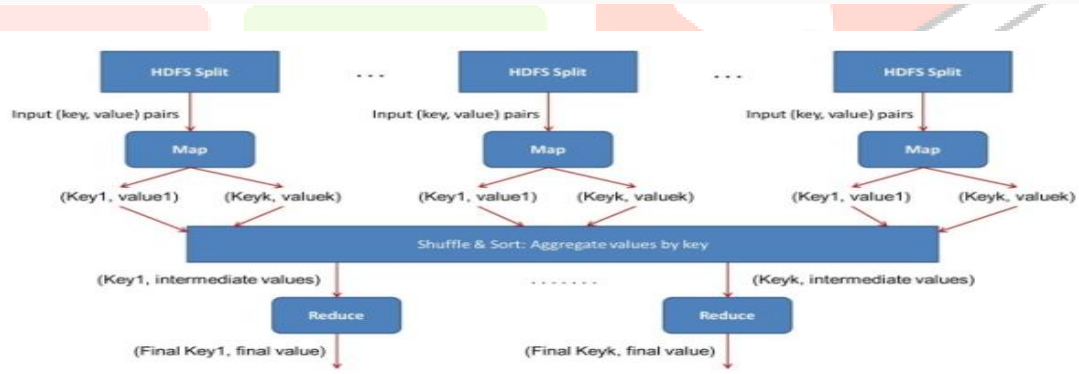

Figure 5: MapReduce Working

3. LITERATURE REVIEW

E. Madhusudhan Reddy & S.Vikram Phaneendra

Illustrated that in olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as “big data”. In this they told that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity consisting of name node, data node, edge node, HDFS to handle big data systems. Hadoop architecture handle large data sets, scalable algorithm does log management application of big data can be found out in financial, retail industry, health-care, mobility, insurance. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc [1]

Kiran kumara Reddi & Dnvsl Indira

Upgraded us with the learning that Enormous Information is mix of organized , semi-organized ,unstructured homogenous and heterogeneous information .The creator proposed to utilize pleasant model to deal with exchange of

colossal measure of information over the system .Under this model, these exchanges are consigned to low request periods where there is sufficient ,sit still transfer speed accessible . This data transmission would then be able to be repurposed for enormous data transmission without affecting different clients in framework. The Decent model uses a store – and forward approach by using organizing servers. The model is ready to suit contrasts in time zones and varieties in transfer speed. They proposed that new calculations are required to exchange enormous information and to understand issues like security, big data and to solve issues like security, compression, routing algorithms [2].

TABLE 1: Components of Hadoop comparison

Hadoop Component	Purpose	HPCC Equivalent	Notes
HDFS	Distributed file system to store files for Hadoop	None	HPCC uses native filesystem to store files
Name node	Keep track of all files stored in HDFS including all the blocks allocated to each file	Thor master node	The DFU is responsible for tracking file parts across nodes
Data node	Sub node that stores Hadoop files	Thor slave nodes	Like Hadoop namenode, Thor can store data in both the master and slavenodes
Job tracker	Scheduling job runs and managing resources	Dali	
Task tracker	Run subtasks assigned to the sub node		Dali monitors task completion on each Thor sub node
Hive	Provides DW structure to HDFS files and SQL-like declarative access to DW	Roxie + Thor	Thor is used to perform data warehousing functions like aggregations and create keyed B+ Tree indexes. Roxie is used to provide fast keyed access to aggregated data
Pig/Sqoop	Provide easy declarative language constructs to perform jobs on Hadoop	ECL	ECL is a declarative SQL-like language

CONCLUSION

We have stated and explained the events of Big Data. Thus the paper describes about the concept of Big Data. It points out of processing (operational) problems. The challenges occurred technically must be statement foe effective and speed processing of Big Data. This challenge includes the issue of scale, dissimilarity and incompleteness, timeless, privacy, human association and also the goals of architecture. It focus on MapReduce Architecture and its working, terminologies etc. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system.These technical challenges are among huge types of domain application and therefore its non-cost effective to state in the context of single domain alone. Hence, this paper designates that Hadoop is an open-source software utilized for an procedure of Big data.

REFERENCES

[1]Vikram Phaneeendra & 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).

[**2]Kiran kumara Reddi & Dnvsl Indira** "Different Technique to Transfer Big Data : survey" IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}.

[3]Balaji Palanisamy, Member, IEEE, Aameek Singh, Member, IEEE Ling Liu, Senior Member, IEEE" Cost-effective Resource Provisioning for MapReduce in a Cloud"gartner report 2010, 25.

[4]Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" Volume 4, Issue 5, May 2014" 27.

[5]Dittrich JA QuianeRuiz Efficient big data processing in Hadoop MapReduce–Proceeding of VLDB Endowment 2012.

[6]M.Maurya,S.Mahajan Performance analysis of MapReduce programs on Hadoop Clusters 2012.

[7]Sharma Y. ; Kumar S. and Pai R.M; FormalVerification of OAuth 2.0 Using Alloy Framework .International Conference on Communication Systems and Network Technologies in 2011.

[8]Ke Liu and Beijing Univ OAuth BasedAuthentication and Authorization in Open Telco API .IEEE International Conference on Communication Systems and Network Technologies in 2012.

[9]Kevin T. Smith Big Data Security: The Evolution of Hadoops Security ModelPosted on Aug 14, 2013.

[10]Priya P. Sharma and Chandrakant P. Navdeti Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution in 2014.

[11]Seonyoung Park and Youngseok Lee, Secure Hadoopwith Encrypted HDFS, Springer-Verlag Berlin Heidelberg in 2013.

[12]Dean J., Ghemawat S.: MapReduce: Simplified DataProcessing on Large Cluster, In:OSDI (2004).

[13]Ghemawat S., Gobioff H., Leung, S.: The Google FileSystem. In: ACM Symposium onOperating SystemsPrinciples (October 2003).

[14]OMalley O., Zhang K., Radia S., Marti R., Harrell C.: Hadoop Security Design,Technical Report (October2009) .

[15]White T.: Hadoop: The Definitive Guide, 1st edn.OReilly Media (2009).

[16]Guanghui Xu, Feng Xu, Hongu Ma. "Deploying and searching Hadoop in virtual machines" International Conference on Automation and Logistics, China, August 2012 IEEE Conferences.

[17]MakhoNgazimbi,PhD, "Data Clustering Using Map-Reduce" ,Boise State University ,March 2009.

[18] F. N. Afrati and J. D. Ullman. Optimizing Joins in a Map-Reduce Environment. In EDBT, pages 99– 110, 2010.

[19]Hemant Hingave & Rasika Ingle, "An approach for MapReduce based Log analysis using Hadoop", IEEE SPONSORED 2'ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEMS (ICECS '2015).

[20]Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

[**21]Vasiliki Kalavri & Vladimir Vlassov**) "MapReduce: Limitations, Optimizations and Open Issues", 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communication(2013).