Efficient Analysis of Big Data by using Hadoop in Cloud Computing by Map Reducing

Ashish A. Patokar
Assistant Professor [Adhoc Basis]
Dept. of Computer Science & IT,
Shri Shivaji College, Akola
ashishpatokar@gmail.com

Dr. V. M. Patil
Head & Associate Professor
Dept. of Computer Science & IT,
Shri Shivaji College, Akola
vinmpatil21@yahoo.co.in

Abstract- Nowadays, a excellence volume of data from individualize resources such as sensory devices,social media networks and information serving devices are induce. This category of large data is called as big data and about 80% of the data is now in unstructured formats. Hadoop is an open source platform that expand computing of big data. Hadoop is made up of two components such as Hadoop distributed file system (HDFS) andMap Reduce engine. HDFS is made up of geographically distributed Data Node and access of these Data Node service is known as Name Node. Map Reduce is a minimization technique which constructs use of sorting, shuffling and reducing. This paper introduces the Big Data technology, framework of Hadoop, architecture of hadoop and its efficient analysis.

Keywords- *Hadoop, Cloud, Big Data, HDFS, Map Reduce*

*****

I. INTRODUCTION

Big Data analysis is the way of examining massive data sets containing a combo of data type's i.e. big data to uncover buried patterns, market, trends, and customer's choice and other effective business information. Map Reduce is abroadly used for efficient analysis of big data.

Nowadays, Cloud has become an avoid lessneed for majority of IT organizations. Application of cloud such as big data storage, retrieval and portability. Big data is a data that is combination of datasets whose size, complexity make them hard to be carry, arranged and analysis sized by traditional technologies. Big data is a heterogeneous blend of data such as structured and unstructured data. Structured data such as in rows and columns, XLS's, tables and unstructured data like images, PDF formats, manuals, media like video, audio and graphics etc[1,2].

When a organizations required to store large amount of data and analysis they will adopt two options. The first option that is either purchases a big machine with more RAM, higher disk space. The second option contacting some database vendors for solution and these two options have their own quaere. The first option has some limit to how big machine you can buy and the second option implies scale horizontally so the hadoop provides expandable storage and distributed computing efficiency [3,4]. Hadoop is the open source platform for structuring big data and solves the problem useful for analytical and operational purpose [5].

Hadoop is the combination of two components. These are the Hadoop Distributed File System (HDFS) and the Map Reduce engine. Hadoop is use in many of the world's biggest online media companies like Yahoo, Fox interactive media and LinkedIn Twitter. Many hadoop services include HTTP interfaces [6]. The hadoop platform uses proxy IP address and database of role to perform authentication and authorization [7,9]. This paper basically focuses on efficient analysis of big data processing.

II. LITERATURE SURVEY

M. Dhavapriya et al .(2016) in a Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table [1] have focused on various methods for catering to the problems in hand through map reduce framework over hadoop distributed file system (HDFS) Map reduce is minimization technique makes use of file indexing with mapping, sorting, snuffing and finally reducing map reduce techniques is implemented for big data analysis using HDFS author also focus on big data opportunities and challenges etc.

Harshawardhan S. Bhosale et al. (2016) A Review Paper on Big Data and Hadoop [2] have focused in Have describes the concepts of big data along with vs., volume, velocity and variety of big data and focuses on big data processing problem and also map reduce and HDFS architecture.

Dr. Rakesh Rathi et al. (2014) in a Big Data and Hadoop [3] have describes the term big data its explosive nature increasing rapidly in today's scenario. Also focus on hadoop and Map Reduce architecture , Big data

charcterstics. Hadoop is the solution of all the problems arises due to massive amount of data that includes audio, text and images etc.

Twinkle Antony et al. (2014)in Addressing Big data with hadoop [4] have focused on hadoop is an open source platform provides distributed computing of big data. Also describes the overall execution of hadoop. Hadoop is made of two components i.e HDFS and Map Reduce.

Vinod Sharma et al. (2014) in The Evolution of Big Data Security through Hadoop Incremental Security Model [5] have introduces the Big data technology along with its important in the modern world. Hadoop, Map Reduce and No SQL are the major big data technology and also focuss on light on other challenges and issues and the concept of big data and the existing projects.

Bhawana Gupta et al. (2014) in a Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data [6] Have proposed the use of the big data Analytics for analyzing the enterprise data and focus on framework based on hadoop for dealing the targeted attacks using big data security analytics.

Poonam S. Patil et al. (2014) in a Survey Paper on Big Data Processing and Hadoop Components [7] Have briefly introduce the map reduce frame work based on Hadoop and the current state of the art in map reduce Algorithms for big data analysis. Also focused on Hadoop components.

Suman Arora et al. (2014) in a Survey Paper on Scheduling in Hadoop [8] HAve dissucssed many techniques for making efficient scheduler for the map reduce can speed up system or data retrieval technique likequincy, Asynchronous processing Job awareness, Delay scheduling etc. For making the scheduler effective processing.

Mehak et al. (2014) in a Improving Data Storage security in cloud using Hadoop [9] Have focuses on protecting cloud data through encryption and to effectively utilize the resources of the cloud. Also discussed encryption techniques, hadoop map reduce paradigm.

Silky Kalra et al. (2014) in a A Review on HADOOP MAPREDUCE-A Job Aware Scheduling Technology [10] Have includes detailed view of various important components of Hadoop, Job awareness, Scheduling Algorithms For map reduce framework, various DDoS attack and defense methods.

Harshitha R et al. (2014) in A Survey on Scheduling Techniques in Hadoop [11] Have discussed Maps reduce Frame Work , Overview of difference scheduling techniques with their Applications. Different scheduling Techniques for enhance the data locality, efficiency, fairness and performance are discussed.

Vidyasagar S.D.et al. (2013) in a study on A Study on "Role of Hadoop in Information Technologyera" [12] Have foccused on Assumptions and Goals, Hadoops components, working process of Hadoop architecture, and requirments of hadoop.

Vishal S. Patil et al. (2013) in a Hadoop skeleton & fault tolerance in hadoop clusters [13] Have focused on framework of Hadoop along with how Fault tolerance is archived by means of data duplication and also discussed about the architectural framework of hadoop and also some of the strategies to overcome the faults tolerance in HDFS includes check point and recovery.

S. Chandra Mouliswaran et al. (2012) in a study on replica management and high availability in hadoop distributed file system (HDFS) [14]
Have Focused on how the replicas are managed in HDFS for providing high availability of data under extreme computational requirement and also focus on possible failure affect the Hadoop Cluster.

R.Thangaselvi et al. (2015) in a An efficient Map reduce scheduling algorithm in hadoop [15] Have proposed a method to improve efficiency of map reduce scheduling Algorithms and now it works better than existing. Map reduce Scheduling Algorithms by taking less amount of computation and gives high accuracy.

III. BIG DATA ANALYSIS IN HADOOP

Big Data analysis allow to a large variety of use cases reach across multiple industries. Numerous data today is not natively in systematic format. Data analysis, retrieval, organization and modeling are the essential challenges. Analysis of data is a process of auditing, converting and cleaning and designing data with the intent of detecting useful meaning information and decision support system making and it has multiple facts and approaches under a variety names in science and social domains. Industrialization use
Amazon Web Services: eg. Amazon is a name in feeding web hosting and product services. Amazon notice to offer a essential framework for customers to use without supplying much in the way of customer foundation. Amazon supports products like Hadoop, Hive and Pig sanctioning you to build your own solution on their platform and create your big data stack and storage.

 ISSN: 2321-8169
378 - 381

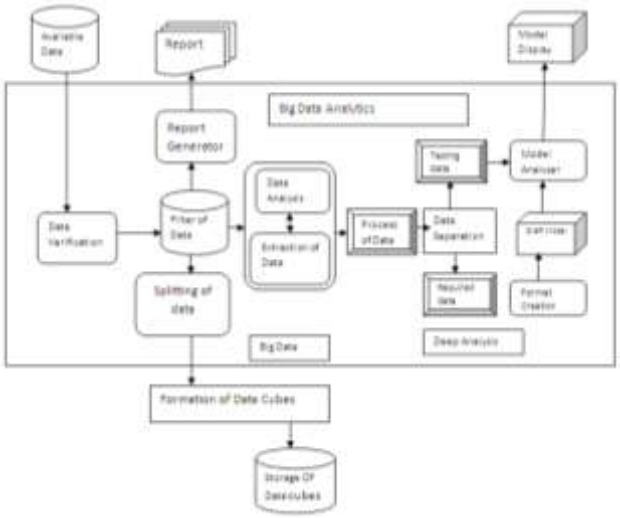

Figure 1. Big Data Analysis Tools

The Big Data analytics and detail data analysis process at different stages of batch process shown in figure. Big Data processing part is done using hadoop / PIG technology with classical ETL logic implementation. The Map Reduce model that hadoop provides can linearly scale to the processing by adding machines to the hadoop cluster. Cloud computingresources (Amazon, EMR) is common approach to the platform to perform this kind of data. The deep analysis is done in R, SPSS, and SAS using a much smaller amount of carefully sampled data that fits into a single system capacity. The detail data analysis is part usually involve data, data virtualization, data preparation model, learning model evaluation and analysis.

IV. BIG DATA ANALYSIS IN EFFECTIVE WAY BY MAP REDUCE

Hadoop map reduce is a software framework for easily writing applications, Process vast amount of data in parallel process on Large clusters of commodity hardware in Fault Tolerant manner. Map reduce splits the input data set into indecent chunks which processed by the map tasks in a completely parallel manner, and the Frame work sorts the out-put of map, which are then input to reduce tasks. And both input and out-put are stored in a file system. The map reduce frame work helps developers divide a query into steps, divide dataset into chunks and run those step pattern separate hosts[3].

The map reduce model consist of two functions, map ( ) and reduce ( ) [1].

Map Reduce libraries written in numerous programming languages, with distinct levels of optimization. The name Map Reduce originally point out to the proprietary Google technology.

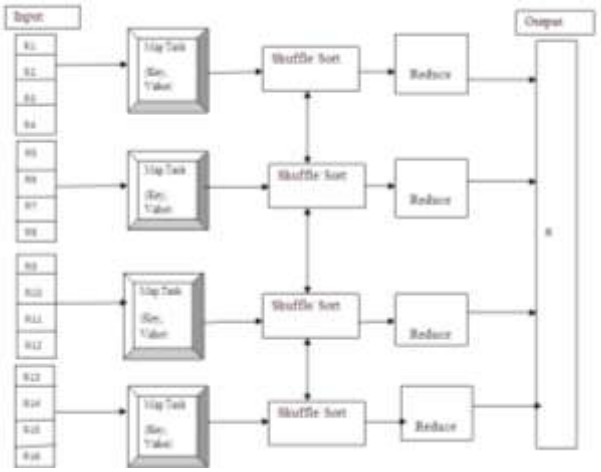

Figure 2. Map-Reduce Working

A) Map Reduce Components

- Name Node- the Name Node bearing by clients of the HDFS to locate information within the file system and feed updates for data they have added, deleted and manipulated.
- Data Node- Data Node serves two functions. It contains a section of data in HDFS and acts as compute platform for running jobs and other resort the local data within HDFS.
- Job Tracker- Job Tracker schedules jobs and tracks the assign job to task tracker.
- Task Tracker- Tracks the tasks and reports to the job tracker.

B) Map Reduce Process

Map Reduce boldness the general instability problems promote in homegrown distributed systems. The Map Reduce splits into multiple tasks i.e. Mapper and Reducer. Map Reduce has a master and Slaves. The master is recorded in "Masters" configuration file and slaves are recorded in "Slaves" and they perceive about each other.

- *Mapper*

  The Mapper maps the input key/ values pairs to a set of intermediate key/ values pairs. For example the sorter i.e. (the boy asking to the old man how old are you) only concerned about sorting people into accurate groups (in case age). In Map Reduce to take the boy is known as MAPPER.

- *Reducer*
  Reducer reduces set of interpose values which portion a key to a smaller set of values. The Reducer has three phases i.e. Shuffle, Sort and Secondary sort.

a) *Practitioner*- Practitioner allows you to distribute how outputs from map stage are sent to reducers. The key is used to obtain the partition, ideally by a hash function and hash Practitioner is the default Practitioner.

b) *Reporter*-Reporter is a skill for Map Reduce function to report progress. Mapper and Reducer utilization can use reporter to report progress.

c) *Output Collector:* Output Collector facility supplied the Map Reduce framework to collect data output by the mapper or reducer.

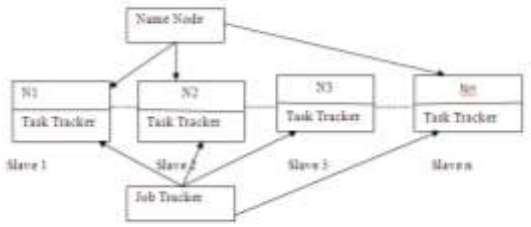

Figure 3. Map Reduce Working through Master/ Slave

V. CONCLUSION

Hadoop is an open platform to process extensive among of big data. Hadoop supply distributed storage known Hadoop distributed file system (HDFS) and expand through computing through a programming model called Map reduces. Map Reduce dividing the whole program and executed separately. Big data analysis tools as map reduce and HDFS.Ensure to help chamber/organizations better group their client and the market place, for better business decisions. To process the wide amount of data accessible drives science progress, innovation and to search new ways to some problems, which are considered impossible in the past.

REFERENCES

[1] M. Dhavapriya, N. Yasodha, " Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table", International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 1, Jan - Feb 2016,pp 5-14.

[2] Harshawardhan S. Bhosale, Devendra P. Gadekar , "A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153, pp 1-7.

[3] Dr. Rakesh Rathi,Sandhya Lohiya, "Big Data and Hadoop", International Journal of Advanced Research inComputer Science & Technology (IJARCST 2014), Vol. 2, Issue 2, Ver. 2 (April - June 2014), pp 214-217.

[4] Twinkle Antony, Shaiju Paul, "Addressing big data with hadoop", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.2, February- 2014, pp 459-462.

[5] Vinod Sharma, N.K. Joshi, "The Evolution of Big Data Security through Hadoop Incremental Security Model", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 5, May 2015,pp 3489-3493.

[6] Bhawna Gupta, Dr. Kiran Jyoti, " Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data" International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, pp 3867-3870.

[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,pp 585-590.

[8] Suman Arora, Dr.Madhu Goel, " Survey Paper on Scheduling in Hadoop", Volume 4, Issue 5, May 2014, pp 812-815.

[9] Mehak, Gagandeep, " Improving Data Storage Security in Cloud using Hadoop", Journal of Engineering Research and Applications, ISSN : 2248-9622, Vol. 4, Issue 9(Version 3), September 2014, pp.133-138.

[10] Silky kalra, Anil Lamba," A review on HAdoop MAp Reduce- A job Aware Scheduling Technology", International Journal of computational Engineering Research(IJCER), ISSN (e): 2250 – 3005, Vol, 04 , Issue, 5 , May – 2014, pp 36-40.

[11] Harshitha R, Rekha G S, Dr. H S Guruprasad, "A Survey on Scheduling Techniques in Hadoop" , IJEDR , Volume 3, Issue 1, 2014, pp 248-254.

[12] Vidyasagar S. D, "A Study on Role of Hadoop in Information Technology era",Global research analysis, Volume 2, Issue 2, Feb. 2013, pp 100-101.- 8160

[13] Vishal S Patil, Pravin D. Soni, "Hadoop skeleton & fault tolerance in hadoop clusters", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 2, February 2013, pp 247-250.

[14] S. Chandra Mouliswaran and Shyam Sathyan,"Study on replica management and high availability in hadoop distributed file system (HDFS)", Journal of Science ,Vol 2 , Issue 2 ,2012 , pp 65-70.

[15] R.Thangaselvi, S.Ananthbabu, R.Aruna," An efficient Mapreduce scheduling algorithm in hadoop", International Journal of Engineering Research & Science (IJOER), Vol-1, Issue-9, December- 2015, pp 102-108.

[16] Apache Hadoop: http:// Hadoop.apache.org.

[17] Hadoop Tutorial: http: // developer.yahoo.com

[18] http:// data.us/ RONLcAAA