A Study on Big Data Processing using HADOOP

Prajitha.P[1] Dr.V.Kathiresan[2]
[1]MCA Student [2]Head of Department
[1,2]Department of Computer Applications
[1,2]Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India - 641 049.

Abstract— The fast expansion of Internet and world wide website has led to vast amounts of information available online. This information needs to be processed, analyzed, and associated to complete correct in sequence with the reason of store, handle, access, and procedure free amount of data presented online and the data that is created in structured and unstructured type, Data logical computing is required which satisfy have to be search, analyze, mine, and visualize the huge quantity of data and information. at present a variety of data concentrated technologies equivalent, map reduce are used which uses computing applications which require huge volumes of data and assign mostly of their processing time to input and output and utilization of data. We will study previously available Data thorough technologies with Hadoop data complete to provide high look that be supposed to be fault flexible more than hardware failures, reduce communications errors, and software bugs and execute a variety of data concentrated analysis level.

Key words: Big data difficulty, Hadoop, cluster, Map Reduce

I. INTRODUCTION

Big Data pass on near vast amount of data that is normally in peta bytes. Big Data is an exhortation that is associated with volume of data that cannot be process in traditional environments. Hadoop is open source software that makes utilize of Map Reduce for spread processing. It has involved worldwide concentration for processing big data in the real world. Big data dispensation should have certain individuality such as volume, velocity and variety. Big data can transform market in business, government and other aspects.[1] It has bang on the society Big data handing out has certain phase that consist of data acquirement and recording, information extraction and cleaning, data representation, aggregation and integration, data modeling, analysis and query processing, and interpretation. It is a great scope of information, new kinds of data and analysis, real time information, data entry from new technologies.[7]The Apache Hadoop project consists of the HDFS and Hadoop Map Reduce in addition to other modules. The software is model to collect the processing power of clustered computing while organization failures at node level. [8]The Map Reduce software framework which was initially introduce by Google in 2004 is a encoding Model, which at the present adopted by Apache Hadoop it consists of splitting the large chunks of data and Map &Reduce phases.

II. BIG DATA

Big data is the new word that contains huge and complex dataset. It is complicated to handle these dataset not including new technology. The McKinney Global Institute published a statement on big data that describe the different business opportunities in big data [6]. Paulo Boldi, One of the authors says "Big Data does not require big technology, it need big brainpower". It consisted two types in Big Data they are:

A. Structured Data:

The data can be without difficulty analyze. It is an arithmetical form, figures, and transaction data

B. Unstructured Data:

The data contain complex information such as Email attachments, Images comments on social networking sites. [5] These data cannot be easily analyzed.

C. Three V's in Big Data Managing:

- Volume of data: Volume refers to quantity of data. [4] It is stored in endeavor repositories have developed from megabytes and gigabytes to peta bytes.
- Variety of data: Data variety exploded from structured and legacy data stored in endeavor repositories to unstructured, semi structured, audio, video, XML and so on.
- Velocity of data: It defines has motion of data. Data created rapidly, process and analyze[3].

III. PROCESSING BIG DATA WITH HADOOP

Hadoop that can processing data on large clusters. It extends Map Reduce and also provides different capability to it including caching mechanism. As real time application require to process huge amount of data inside of data mining and data analysis. [4] In Hadoop the major program model is called Map Reduce which is proper for processing big data. Hadoop is a spread file system that supports giving out huge amount of data in terabytes or more in spread environments such as cloud computing. [5] A Map Reduce is already a scalable and efficient program model that is enhanced further. Map reducing program model is previously individual used by many companies to process huge data. They consist of Yahoo, Face book, Google and so on. Hadoop is an open source Map Reduce framework that has been improved and presented as Hadoop. Hadoop includes loop aware task scheduler, and caching mechanism besides other common requirements as there in Hadoop[2].

IV. PROBLEM WITH BIG DATA PROCESSING

Big Data has approach since be living in the worlds that uses the strapping use of rising data technology. At present live huge amount of data, the different challenge are features with reference to the managing of such general data .The

Challenges consist of the unstructured data, real time analytics, error acceptance, processing and storage of the data. The bulk of the data is increasing day by day with the exponential development of the organizations. The different operations are use in support of the data dealing

All rights reserved by *www.ijsrd.com*

out that includes the tagging, highlighting, searching, indexing and so on. Data is generating as of the various sources in the shape of structured as well as unstructured form [8]. Big data sizes differ since a few dozen terabytes to various peta bytes of data. The processing and study of huge amounts of data or produce the important information's in the difficult task.[7] A Big data is the most recent technology that know how to be useful for the big business organization, therefore it is required that different issue and challenge related among this technology must carry out into radiance. The two main problems about big data are the storage faculty and the processing of the data.

V. HADOOP SOLUTION FOR BIG DATA PROCESSING

Hadoop is a encoding skeleton used to maintain the handing out of large data set in a dispersed computing environment.[9] Hadoop was developed by Google's Map Reduce that is a software framework wherever an application split into different part. [4] The Current Apache Hadoop environment consist of the Hadoop Kernel, Map Reduce ,HDFS and statistics of different mechanism like Apache Hive, Base and Zookeeper, HDFS and Map Reduce are explain in next points.

A. Hadoop Distributed File System:

Hadoop include fault-tolerant storage space system is called the Hadoop Distributed File System. [4] HDFS is intelligent store vast amount of information, level up incrementally and endure the failure of important parts of the storage space communications without lose data.[8] HDFS manage storage space on the cluster by breaking external store into pieces know as blocks and storing each of the blocks redundantly crossways the pool of servers.

B. Map Reduce:

The processing maintain in the Hadoop environment is the Map Reduce skeleton. The skeleton allows the design of process to be applying to a vast data set. For example, it is very huge dataset can be cheap into a smaller subset anywhere analytics can be apply. [6]In Hadoop these kind of operation are written as Map Reduce job in Java. It is a higher level language like Hive and Pig that build writing these programs easier. [7] The output of these jobs can be written back to moreover HDFS or located in a traditional data warehouse. There are two functions in Map Reduce as follow.

C. Map Function:

The map function takes a key, value pair and output a list of in-between values with the key. [8] The map function is written in such a method that various map function can be executed at one time.

D. Reduce Function:

The reduce function then take the output of the map functions and do various development on them, generally combine value to produce the preferred result in a output file.

VI. CONCLUSION

This paper describes the model of Big Data alongside through 3 Vs, Volume, Velocity and variety of Big Data. It is focus on Big Data processing problems. This technological challenge must be concentrate on for capable and fast processing of Big Data. And Its built-in processing big data with hadoop, problem with big data processing, hadoop solution for big data processing, HDFC, MapReduce, Map function, Reduce function. These technological challenges are regular crossways a huge range of application domains. This paper describes Hadoop which is an open source software use for dealing out of Big Data.

REFERENCES

[1] Aditya B. Patel, Manashvi Birla, Ushma Nair, (6-8 Dec. 2012),“ Big Data Problem Using Hadoop and Map Reduce”.
[2] Jens DittrichJorgeArnulfoQuian´eRuiz (2013).Big Data Processing in Hadoop Map Reduce. USA: Proceedings of the VLDB Endowment. P1-2.
[3] Sagiroglu, S.; Sinanc, D., (20-24 May 2013),”Big Data: A Review” .
[4] Ankita S. Tiwarkhede1, Prof. Vinit Kakde2 A Review Paper on Big Data Analytics International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438
[5] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N “ Big Data using Apache Hadoop and Map Reduce” Issue 5, May 2014” .
[6] Harshawardhan S. Bhosale1, Prof. Devendra P. Gadekar2A Review Paper on Big Data and Hadoop International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.
[7] Dr. Shoban Babu Sriramoju Associate Professor, Dept. of CS, Varadha Reddy Engineering College, Warangal, India A Review on Processing Big Data International Journal of Innovative Research in Computer and Communication Engineering Vol. 2, Issue 1, January 2014.
[8] Prof. R.A.Fadnavis Dept. Of Information Technology and Samrudhi Tabhane Dept. Of Information Technology Yeshwantrao Chavan College Of Engineering Nagpur, India R.A.Fadnavis et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015.
[9] M.Dhavapriya, N. Yasodha Department of CS NGM College, Pollachi Tamil Nadu - India 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.
[10] Getaneh Berie Tarekegn PG, Department of Computer Science, College of Computing and Informatics, Assosa University, Assosa, Ethiopia big data: security issues, challenges and future scope International Journal of Computer Engineering & Technology (IJCET) Volume 7, Issue 4, July–Aug 2016.

All rights reserved by *www.ijsrd.com*