International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 05
April 2017

An Performance Analysis of Map Reduce Using Big Data and Hadoop

D.Saidulu[1], D. Ajay Kumar[2], S.P. Raghu Ram Bhattar[3], S.Raja Vamshi [4], P.Lalitha Venkat Sai[5]
[1]Associate Professor, Department of CSE, Guru Nanak Institutions, Ibrahimpatnam, Hyderabad, India
[2,3,4,5] B.Tech Students, Department of Computer Science & Engineering, Guru Nanak Institutions, Hyderabad, India

Abstract: *Big data is used to define an enormous volume of both structured and unstructured data that is so large that it's problematic to process using traditional database and software techniques. Big Data is regarded as by the dimensions volume, variety, and velocity, while there are some well-established methods for big data processing such as Hadoop which uses the map-reduce paradigm. Using MapReduce programming paradigm the big data is processed.*

Index Terms– Big Data, Parameters, Challenges, Opportunities, Hadoop

I. INTRODUCTION

Big data method genuinely a huge information; its miles a collection of massive datasets that cannot be handled using antique computing strategies. Big statistics is not best containing statistics, it additionally incorporates various equipment, strategies and frameworks. Data that has more-large Volume comes from Variety of assets, Variety of codecs and is derived at us with a fine Velocity is normally known as Big Data. Big data may be based, unstructured or semi-structured. Big information hold the information generate with the aid of diverse gadget and packages like Black container Data which is part of helicopter.

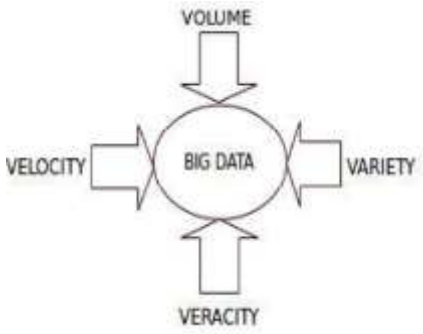

Figure 1 Four V's of BIG DATA

It catches sound of flight team, recording of Earphones and megaphones. Social media records Twitter is likewise part of social media which keep data and the view posted by way of millions of human beings. Stock exchange data which hold the statistics approximately the purchase and retail choices made on a share of different groups made by the customer. Search engine information which recruit lot of statistics from various databases. As the data is bigger from different sources in different form, it is represent by the 4 Vs.

Map Reduce Using Big Data and Hadoop

Velocity: Velocity is the speed at which data is developing and processed. For example social media posts.

Variety: Variety is one more important characteristic of big data. It refers to the type of data. Data may be in different styles such as Text, numerical, images, audio, video data. On twitter 400 million tweets are sent per day and there are 200 million active users on it.

Veracity: Veracity means anxiety or accuracy of data. Data is uncertain due to the inconsistency and in completeness.

Hadoop is an open source project hosted 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:
1. File System (The Hadoop File System)
2. Programming Paradigm (Map Reduce)

The other subprojects permit correlative services or they're rising on the center to upload better-level abstractions. There exist several predicaments in coping with storage of large amount of records. Though the storage capacities of the drives have

improved vastly but the charge of analyzing data from them hasn't shown that vast development. The reading process takes massive quantity of time and the method of writing is additionally slower. The time may be decreased through analyzing from more than one disks right now. Only the use of 100th of a disk can also appear wasteful. But if there are one hundred datasets, every of which is one terabyte and providing shared access to them is likewise a solution. There exist many issues also with the use of several pieces of hardware because it increment the possibilities of failure. This may be deflecting by using Replication i.e. growing redundant copies of the same facts at awesome devices so that in case of failure the replica of the facts is viable. The important hassle is of combinative the data being study from exceptional machines. There are so many methods are capable in allotted computing to deal with this problem but nonetheless its miles pretty difficult. All the complication mentioned is without difficulty controlled by Hadoop. The trouble of failure is directed through the Hadoop Distributed File System and problem of mixing data is directed by Map lessen programming Paradigm. Map Reduce essentially diminish the trouble of disk reads and writes by giving a programming version dealing in computation with keys and values. There are few additives of Hadoop which might be as:

Hadoop Distributed File System

Hadoop develop with a distributed File System called HDFS, HDFS stands for Hadoop Distributed File System. The Hadoop Distributed File System is a versatile, clustered way to handling files in a big data environment. HDFS is not the final terminal for files.

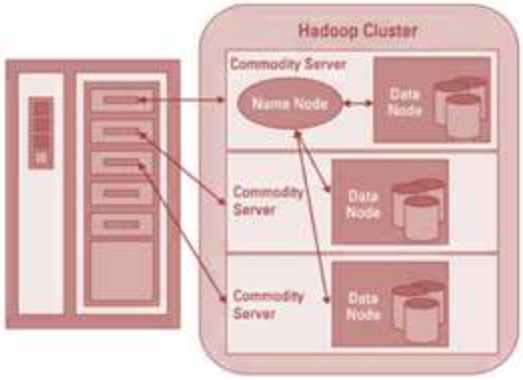

Figure 2 HDFS Architecture

It is a kind of data service that offers a different set of capabilities required when data volumes and velocity are high. Because the data is written once and then read many times. HDFS is a good choice for supporting big data analysis. HDFS works by cracking large files into small parts called blocks. The blocks are stored on data nodes, and it is the responsibility of the NameNode to notice what blocks on which data nodes make up the complete file. The Name Node also performs as a "traffic cop," handling all access to the files. The entire collection of all the files in the cluster is sometimes referred to as the file system namespace.

Map Reduce

Hadoop Map Reduce is an implementation of the algorithm advanced and managed by the Apache Hadoop project. It is supportive to deliberate about this application as a Map Reduce engine, because that is absolutely how it works. We deliver data (fuel); the engine converts the information into production rapidly. Hadoop Map Reduce consists of many stages, each with a meaningful set of operations helping to get the answers you need from big data. The development starts with a user request to run a Map Reduce program and go on until the results are written back to the HDFS.

II. RELATED WORK

Harshawardhan S. Bhosale1, Prof. Devendra Gadekar, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, (10-15 October, 2014), a

review on Big Data and Hadoop the paper describes the concept of Big Data along with 3Vs, Volume, Velocity and variety of Big Data. The paper also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. The paper describes Hadoop which is an open source software used for processing of Big Data.[3]

Shilpa, Manjeet Kaur, LPU, Phagwara, India, a review on Big Data and Methodology (5-10 October, 2013) illustrated that there are various challenges and issues regarding big data. There must support and encourage fundamental research towards these technical issues if we want to achieve the benefits of big data. Big-data analysis fundamentally transforms operational, financial and commercial problems in aviation that were previously unsolvable within economic and human capital constraints using discrete data sets and on premises hardware. By centralizing data acquisition and consolidation in the cloud, and by using cloud based virtualization infrastructure to mine data sets efficiently, big-data methods offer new insight into existing data sets. [5]

Garlasu, D.; Sandulescu, V.; Halcu, I. ; Neculoiu, G. (17-19 Jan. 2013)," A Big Data implementation based on Grid Computing", Grid Computing offered the advantage about the storage capabilities and the processing power and the Hadoop technology is used for the implementation purpose. Grid Computing provides the concept of distributed computing. The benefit of Grid computing center is the high storage capability and the high processing power. Grid Computing makes the big contributions among the scientific research, help the scientists to analyze and store the large and complex data.

Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) "Shared disk big data analytics with Apache Hadoop" Big data analytics define the analysis of large amount of data to get the useful information and uncover the hidden patterns. Big data analytics refers to the Mapreduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of implementation of Google's Mapreduce Model. In this the performance of SF-CFS is compared with the HDFS using the SWIM by the facebook job traces .SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns. [2]

Aditya B. Patel, Manashvi Birla, Ushma Nair (6-8 Dec. 2012) "Addressing Big Data Problem Using Hadoop and Map Reduce" reports the experimental work on the Big data problems. It describe the optimal solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and Map Reduce programming framework for parallel processing to process large data sets.[4]

Real Time Literature Review about the Big data According to 2013, facebook has 1.11 billion people active accounts from which 751 million using facebook from a mobile. Another example is flicker having feature of Unlimited photo uploads (50MB per photo), Unlimited video uploads (90 seconds max, 500MB per video), the ability to show HD Video, Unlimited storage, Unlimited bandwidth. Flickr had a total of 87 million registered members and more than 3.5 million new images uploaded daily.[22]

III. SYSTEM DESIGN

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the big data [20]. There are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies.

A. Hadoop

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 05
April 2017

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper [17].

MapReduce is a programming framework for distributed computing which is created by the Google in which divide and conquer method is used to break the large complex data into small units and process them. MapReduce have two stages which are [18]:

- **Map ():-** The master node takes the input, divide into smaller subparts and distribute into worker nodes. A worker node further do this again that leads to the multi-level tree structure. The worker node process the m=smaller problem and passes the answer back to the master Node.
- **Reduce ():-** The, Master node collects the answers from all the sub problems and combines them together to form the output.

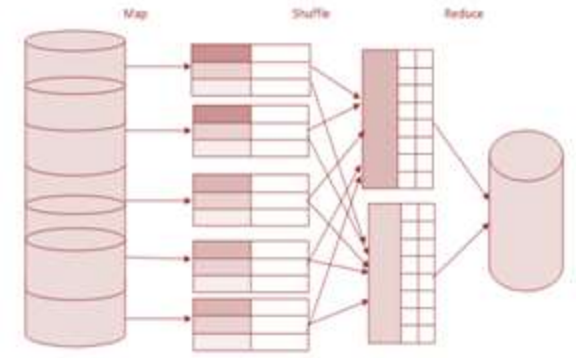

Figure 3 Functional HDFS Architecture

B. HDFS

HDFS is a block-structured distributed file system that holds the large amount of Big Data. In the HDFS the data is stored in blocks that are known as chunks. HDFS is client-server architecture comprises of NameNode and many DataNodes. The name node stores the metadata for the NameNode.NameNodes keeps track of the state of the DataNodes. NameNode is also responsible for the file system operations etc [5]. When Name Node fails the Hadoop doesn't support automatic recovery, but the configuration of secondary nod is possible.

HDFS is based on the principle of "Moving Computation is cheaper than Moving Data". Other Components of Hadoop [6]:

HBase: it is open source, Non-relational, distributed database system written in Java. It runs on the top of HDFS. It can serve as the input and output for the MapReduce.

Pig: Pig is high-level platform where the MapReduce programs are created which is used with Hadoop. It is a high level data processing system where the data sets are analyzed that occurs in high level language.

Hive: it is Data warehousing application that provides the SQL interface and relational model. Hive infrastructure is built on the top of Hadoop that help in providing summarization, query and analysis.

Sqoop: Sqoop is a command-line interface platform that is used for transferring data between relational databases and Hadoop.

Avro: it is a data serialization system and data exchange service. It is basically used in Apache Hadoop. These services can be used together as well as independently.

Oozie: Oozie is a java based web-application that runs in a java servlet. Oozie uses the database to store definition of Workflow that is a collection of actions. It manages the Hadoop jobs.

Chukwa: Chukwa is a data collection and analysis framework which is used to process and analyze the large amount logs. It is built on the top of the HDFS and MapReduce framework.

Flume: it is high level architecture which focused on streaming of data from multiple sources.

Zookeeper: it is a centralized service that provides distributed synchronization and providing group services and maintains the configuration information etc.

C. HPCC

HPCC is a open source computing platform and provide the services for management of big data workflow. HPCC' data model is defined by the user. HPCC system is designed to manage the most complex and data-intensive analytical problems. HPCC system is a single platform, a single

architecture and a single programming language used for the data processing. HPCC system is based on Enterprise control language that is declarative, on-procedural programming language

HPCC system was built to analyze the large volume data for the purpose of solving complex problem.

The main components of HPCC are:

- HPCC data refinery: massively parallel ETL engine.
- HPCC data delivery: Massively structured query engine
- Enterprise Control Language distributes the workload between the nodes

IV. CONCLUSION

As of Hadoop by means of its distributed file system & programming framework grounded on concept of mapped reduction is a dominant tool to control huge data sets. In this paper, we deliberated map reduce over a hadoop cluster by using streaming libraries of hadoop. Big-data analysis basically converts operational, financial and commercial problems in flight that were formerly unsolvable within economic and human capital constraints using discrete data sets and on-premises hardware. By centralizing data acquisition and merging in the cloud, and by using cloud based virtualization infrastructure to mine data sets efficiently, big-data methods offer new vision into prevailing data sets.

REFERENCES

[1]. S.Vikram Phaneendra & 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] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"

[3] Harshawardhan S. Bhosale1, Prof. Devendra Gadekar, JSPM's Imperial College of Engineering & Research, Wagholi, Pune,a review on Big Data

[4] Aditya B. Patel, Manashvi Birla, Ushma Nair,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"

[5] Shilpa, Manjeet Kaur, LPU, Phagwara, India, a review on Big Data and Methodology

[6]. Sagiroglu, S.; Sinanc, D., (20-24 May 2013),"Big Data: A Review"

[7]. Grosso, P.; de Laat, C.; Membrey, P.,(20-24 May 2013)," Addressing big data issues in Scientific Data Infrastructure"

[8]. Kogge, P.M.,(20-24 May,2013), "Big data, deep data, and the effect of system architectures on performance"

[9]. Szczuka, Marcin,(24-28 June,2013)," How deep data becomes big data"

[10]. Zhu, X.; Wu, G.; Ding, W.,(26 June,2013)," Data Mining with Big Data"

[11]. Zhang, Du, (16-18 July, 2013)," Inconsistencies in big data"

[12]. Tien, J.M.(17-19 July,2013)," Big Data: Unleashing information"

[13]. Katal, A Wazid, M.; Goudar, R.H., (Aug, 2013)," Big data: Issues, challenges, tools and Good practices"

[14]. Zhang, Xiaoxue Xu, Feng, (2-4 Sep. 2013)," Survey of Research on Big Data Storage"

[15]. http://dashburst.com /infographic /big-data-volume-variety-velocity/

[16]. http://www-01.ibm.com /software /in /data /bigdata/

[17]. http://searchcloudcomputing.techtarget.com /definition/Hadoop

[18]. K. Bakshi, "Considerations for Big Data: Architecture and Approach", Aerospace Conference IEEE, Big Sky Montana, March 2012

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 05
April 2017

[19].how-much-data-is-on-the-internet-and-generated-online-every-minute/

[20]. Addressing big data problem using Hadoop and Map Reduce

[21]. A Real Time Approach with Big Data-A review

[22] A Real Time Approach with Big Data-A review