A Review Paper on Big Data and Hadoop

Yashika Verma[1] Sumit Hooda[2]
[1]Student [2]Faculty
[1,2]Department of Computer Science & Engineering
[1,2]Gurgaon, Haryana, India

Abstract— Today, organizations in every industry are being showered with imposing quantity of new information. As there are many more data channels and categories available along with traditional sources. Therefore the rate of data growth is increasing more and more which results in a very large volume of data. These vastly larger volumes and new assets are known as Big Data. Technologies such as MapReduce & Hadoop are used to extract value from Big Data. Hadoop is well adopted, standard-based, open source software framework build on the foundation of Google's MapReduce. There are also new data storage techniques that have arisen to bolster these new architectures, including very large file system running on commodity hardware. This new data storage technology is HDFS. This file system is meant to support enormous amount of structured as well as unstructured data.

Key words: Big Data, Hadoop, HDFS, MapReduce and Hadoop Clustering

I. INTRODUCTION

A. *Big Data:*

The definition of Big Data contains three different terms. We can say it Power of 3V's of Big Data which are defined as-

1) Volume of Data:

Numerous independent market and research studies have found that data volumes are doubling every year. On top of all this extra new information, a significant percentage of organizations are also storing three or more years of historic data.

2) Variety of Data:

Studies also indicate that 80 percent of data is unstructured (such as images, audio, tweets, text messages, and so on). And until recently, the majority of enterprises have been unable to take full advantage of all this unstructured information.

3) Velocity of Data:

Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value. [17]

B. *Challenges of Big Data:*

Big data also has its own unique set of obstacles such as: [15]

1) Information Growth:

Over 80 percent of the data in the enterprise consists of unstructured data, which tends to be growing at a much faster pace than traditional relational information. This massive information threaten to swamp all but the most well-prepared IT organizations. [18]

2) Processing Power:

The customary approach of using a single, expensive, powerful computer to crunch information just doesn't scale for Big Data. As we soon see, the way to go is divide-and-conquer using commoditized hardware and software via scale-out. [6]

3) Physical Storage:

Capturing and managing all this information can consume enormous resources, outstripping all budgetary expectations.

4) Data Issues:

Lack of data mobility, proprietary formats, and interoperability obstacles can all make working with Big Data complicated.[3]

5) Cost:

Extract, transform, and load (ETL) processes for Big Data can be expensive and time consuming, particularly in the absence of specialized, well-designed software.[8]

II. HADOOP—COMPUTATIONAL & STORAGE SOLUTION

To address the above mentioned issues, the Hadoop framework is designed to provide a reliable, shared storage and analysis infrastructure to the user community. The storage portion of the Hadoop framework is provided by a distributed file system solution such as HDFS, [22] while the analysis functionality is presented by MapReduce. Several other components are part of the overall Hadoop solution suite. The MapReduce functionality is designed as a tool for deep data analysis and the transformation of very large data sets. Hadoop enables the users to explore/analyze complex data sets by utilizing customized analysis scripts/commands. [1] In other words, via the customized MapReduce routines, unstructured data sets can be distributed, analyzed, and explored across thousands of shared-nothing processing systems/clusters/nodes. Hadoop's HDFS replicates the data onto multiple nodes to safeguard the environment from any potential data-loss. [19]

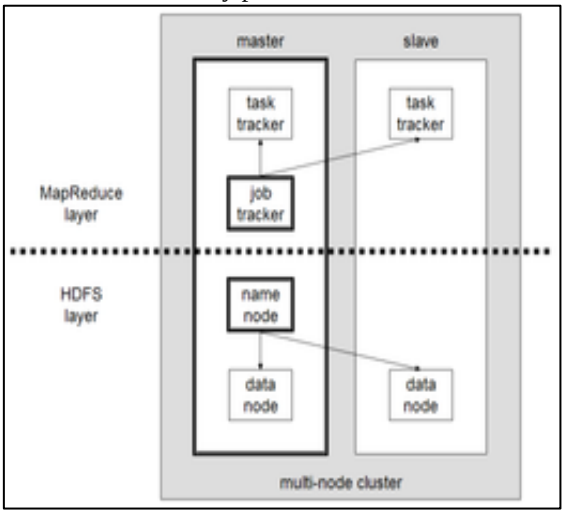

Fig. 1: Hadoop—Computational & Storage Solution

A. *HDFS:*

The Hadoop Distributed File System (HDFS) represents a distributed file system that is designed to accommodate very

large amounts of data (TB or PB) and to provide high-throughput (streaming) access to the data sets. Based on the HDFS design, the files are redundantly stored across multiple nodes to ensure high availability of the parallel applications.

1) Architecture:

An HDFS cluster encompasses two types of nodes (Name and DataNodes) that operate in a master slave relationship. In the HDFS design, the NameNode reflects the master, system namespace, maintains the file system tree as well as metadata for all the files and directories in the tree. All this information is persistently stored on a local disk via two files that are labelled the namespace image and the edit log, respectively.[7] The NameNode keeps track of all the DataNodes where the blocks for a given file are located. That information is dynamic (and hence is not persistently stored), as it is reconstructed every time the system starts up. Any client can access the file system on behalf of a user task by communicating with the NameNode and the DataNodes, respectively. The DataNodes store and retrieve blocks based on requests made by the by clients or the NameNode, and they do periodically update the NameNode with lists of the actual blocks that they are responsible for. [6]

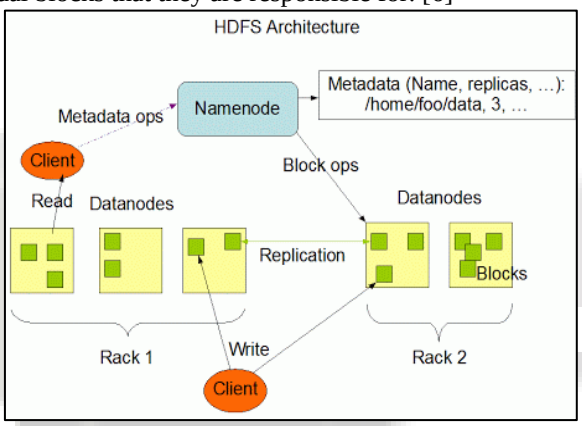

Fig. 2: Architecture

B. Map Reduce:

MapReduce is a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers. The MapReduce framework works in two main phases to process the data which are the Map phase and the Reduce phase. [20]

1) Map Reduce Design:

It takes data set as input which is divided into splits (Split 1, Split 2…. Split N).

- Map: Mapping is done on those splits. After mapping some sorting and shuffling algorithms are applied to splits.
- Reduce: Reduce phase is used to reduce the splits and store them into the centroids files on distributed cache.

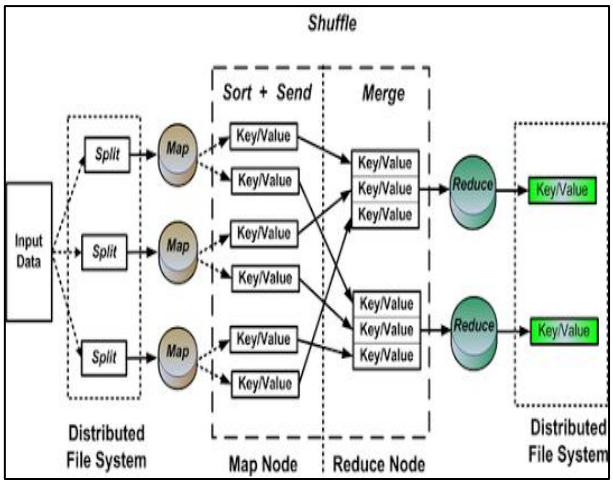

Fig. 3: Map Reduce Design

III. HADOOP CLUSTER

A hadoop cluster is a special type of a computational cluster designed specially for sorting and analyzing huge amount of unstructured data in a distributed environment. Such cluster run Hadoop's open source distributed processing software on low-cost commodity computers. [24]

A. Purpose of Clustering:

1) Hadoop clusters are known for boosting the speed of data analysis applications.
2) They are used to increase the throughput.
3) Hadoop clusters are highly resistant to failure because each piece of data is copied onto other cluster node which ensures that the data is not lost if one node fails.

IV. LITERATURE REVIEW

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 3 Vs, 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. [23]

V. CONCLUSION

Big Data is comprised of large data sets that can't be handle by traditional systems. Big data includes structured data, semi-structured and unstructured data. The data storage technique used for big data includes multiple clustered network attached storage (NAS) and object based storage. The Hadoop architecture is used to process unstructured and semi-structured using map reduce to locate all relevant data then select only the data directly answering the query. There have been extensive studies on various clustering methods and especially the k-means clustering has been given a great attention. The advent of Big Data has posed opportunities as well challenges to business.

REFERENCES

[1] Bakshi, K.,(2012)," Considerations for big data: Architecture and approach"
[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] Yu Li; Wenming Qiu; Awada, U. ; Keqiu Li,,(Dec 2012)," Big Data Processing in Cloud Computing Environments"
[7] Garlasu, D.; Sandulescu, V; Halcu, I. ; Neculoiu, G. ;,(17-19 Jan. 2013),"A Big Data implementation based on Grid Computing", Grid Computing
[8] Sagiroglu, S.; Sinanc, D. ,(20-24 May 2013),"Big Data: A Review"
[9] Grosso, P. ; de Laat, C. ; Membrey, P.,(20-24 May 2013)," Addressing big data issues in Scientific Data Infrastructure"
[10] Kogge, P.M.,(20-24 May,2013), "Big data, deep data, and the effect of system architectures on performance"
[11] Szczuka, Marcin,(24-28 June,2013)," How deep data becomes big data"
[12] Zhu, X.; Wu, G.; Ding, W.,(26 June,2013)," Data Mining with Big Data"
[13] Zhang, Du,(16-18 July,2013)," Inconsistencies in big data"
[14] Tien, J.M.(17-19 July,2013)," Big Data: Unleashing information"
[15] Katal, A Wazid, M.; Goudar, R.H., (Aug,2013)," Big data: Issues, challenges, tools and Good practices"
[16] Zhang, Xiaoxue Xu, Feng,(2-4 Sep. 2013)," Survey of Research on Big Data Storage"
[17] http://dashburst.com/infographic/big-data-volume-variety-velocity/
[18] http://www-01.ibm.com/software/in/data/bigdata/
[19] http://searchcloudcomputing.techtarget.com/definition/Hadoop
[20] K. Bakshi, "Considerations for Big Data: Architecture and Approach", Aerospace Conference IEEE, Big Sky Montana, March 2012
[21] how-much-data-is-on-the-internet-and-generated-online-every-minute/
[22] [Addressing big data problem using Hadoop and Map Reduce
[23] A Real Time Approach with Big Data-A review
[24] http://android/data/kingsoftoffice/dataclusteringmapreduce.pdf

