International Journal of Innovative Technologies
WWW.IJITECH.ORG
ISSN 2321-8665
Vol.03,Issue.02,
June-2015,
Pages:0248-0251

Hadoop Data for Big Data Processing

SHAYMAA TAHA AHMED[1], T. RAMDAS NAIK[2]
[1]PG Scholar, Dept of Computer Science, Nizam College, Osmania University, Hyderabad, TS, India.
[2]Assistant Professor, Dept of Computer Science, Nizam College, Osmania University, Hyderabad, TS, India.

Abstract: In this paper we discussed the rapid growth of Internet and WWW has led to vast amounts of information available online. This information needs to be processed, analyzed, and linked to achieve correct Information. In order to store, manage, access, and process vast amount of data available online and the data that is created in structured and unstructured form, Data intensive computing is needed which satisfies the need to search, analyze, mine, and visualize the large amount of data and information. Nowadays various data intensive technologies (parallel, map reduce) are used which uses computing applications which require large volumes of data and devote most of their processing time to I/O and manipulation of data. We will analysis already available Data intensive technologies with Hadoop data intensive to provide high performance that should be fault resilient over hardware failures, reduce communications errors, and software bugs and execute a variety of data intensive analysis benchmarks.

Keywords: Big Data Problem, Hadoop Cluster, Hadoop Distributed File System, Parallel Processing, Map Reduce.

I. INTRODUCTION

We are living in an age when an explosive amount of data is being generated every day. Data from sensors, mobile devices, social networking websites, scientific data & enterprises – all are contributing to this huge explosion in data. This sudden bombardment can be grasped by the fact that we have created a vast volume of data in the last two years. Big Data- as these large chunks of data is generally called- has become one of the hottest research trends today. Research suggests that tapping the potential of this data can benefit businesses, scientific disciplines and the public sector – contributing to their economic gains as well as development in every sphere. The need is to develop efficient systems that can exploit this potential to the maximum, keeping in mind the current challenges associated with its analysis, structure, scale, timeliness and privacy. There has been a shift in the architecture of data-processing systems today, from the centralized architecture to the distributed architecture. Enterprises face the challenge of processing these huge chunks of data, and have found that none of the existing centralized architectures can efficiently handle this huge volume of data. These are thus utilizing distributed architectures to harness this data. Several solutions to the Big Data problem have emerged which includes the Map Reduce environment championed by Google which is now available open-source in Hadoop.

Hadoop's distributed processing, Map Reduce algorithms and overall architecture are a major step towards achieving the promised benefits of Big Data. Map Reduce & Hadoop are the most widely used models used today for Big Data processing. Hadoop is an open source large-scale data processing framework that supports distributed processing of large chunks of data using simple programming models. The Apache Hadoop project consists of the HDFS and Hadoop Map Reduce in addition to other modules. The software is modelled to harvest upon the processing power of clustered computing while managing failures at node level. The Map Reduce software framework which was originally introduced by Google in 2004 is a programming model, which now adopted by Apache Hadoop, consists of splitting the large chunks of data, and „Map" & „Reduce" phases (Fig. 1). The Map Reduce framework handles task scheduling, monitoring and failures.

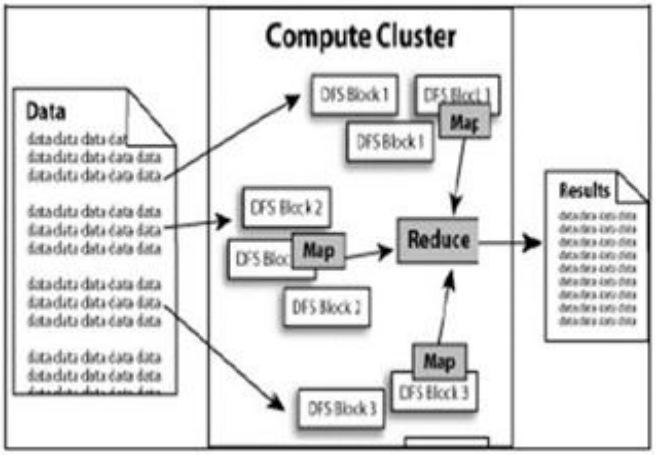

Fig. 1. Map Reduce in Hadoop.

II. PRELIMINARY

A. Literature Survey

The process of the research into complex data basically concerned with the revealing of hidden patterns. Sagiroglu, S.; Sinanc, D. (20-24 May 2013),"Big Data: A Review" describe the big data content, its scope, methods, samples, advantages and challenges of Data. The critical issue about the Big data is the privacy and security. Big data samples

describe the review about the atmosphere, biological science and research. Life sciences etc. By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big data easy to get the information from the complicated data sets [6]. The overall Evaluation describe that the data is increasing and becoming complex. The challenge is not only to collect and manage the data also how to extract the useful information from that collected data. According to the Intel IT Center, there are many challenges related to Big Data which are data growth, infrastructure, data variety, data visualization, data velocity.

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 [5]. 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 [2]. 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.

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[3]. 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.

B. Problem Definition

Big Data has come up because we are living in society that uses the intensive use of increasing data technology. As there exist large amount of data, the various challenges are faced about the management of such extensive data. The challenges include the unstructured data, real time analytics, fault tolerance, processing and storage of the data and many more. The size of the data is growing day by day with the exponential growth of the enterprises. For the purpose of decision making in an organizations, the need of Processing and analyses of large volume of data is increases. The various operations are used for the data processing that includes the culling, tagging, highlighting, searching, indexing etc. Data is generated from the many sources in the form of structured as well as unstructured form [20]. Big data sizes vary from a few dozen terabytes to many pet bytes of data. The processing and analysis of large amount of data or producing the valuable information is the challenging task. As the Big data is the latest technology that can be beneficial for the business organizations, so it is necessary that various issues and challenges associated with this technology should bring out into light. The two main problems regarding big data are the storage capacity and the processing of the data.

III. PROPOSED METHODOLOGY

We propose a map reduce technique in Hadoop for data intensive use. MapReduce programming consists of writing two functions, a map function, and a reduce function. The map function takes a key, value pair and outputs a list of intermediate values with the key. The map function is written in such a way that multiple map functions can be executed at once, so it's the part of the program that divides up tasks. The reduce function then takes the output of the map functions, and does some process on them, usually combining values, to generate the desired result in an output file. Fig.2 shows a picture representing the execution of a MapReduce job

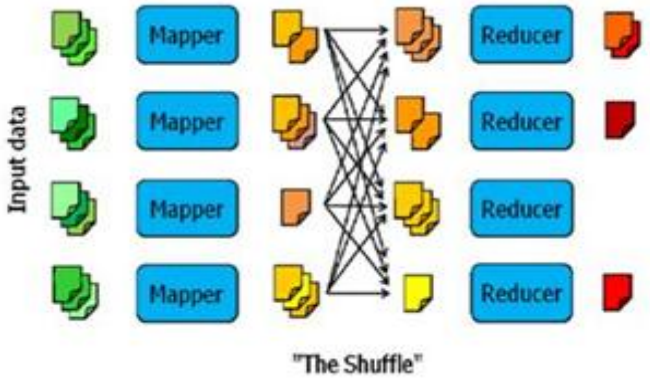

Fig. 2. MapReduce Job.

When a MapReduce program is run by Hadoop, the job is sent to a master node, the job tracker, which has multiple "slave" nodes, or task trackers that report to it and ask for new work whenever they are idle. Using this process, the job tracker divides the map tasks (and quite often the reduce tasks as well) amongst the task trackers, so that they all work in parallel. Also, the job tracker keeps track of which task trackers fail, so their tasks are redistributed to other task trackers, only causing a slight increase in execution time. Furthermore, in case of slower workers slowing down the whole cluster, any tasks still running once there are no more new tasks left are given to machines that have finished their tasks already as shown in Fig.3. Not every process nodes have a small piece of a larger file, so that when a file is accessed, the bandwidth of a large number of hard disks is able to be utilized in parallel. In this way, the performance of

Hadoop may be able to be improved by having the I/O of nodes work more concurrently, providing more throughput.

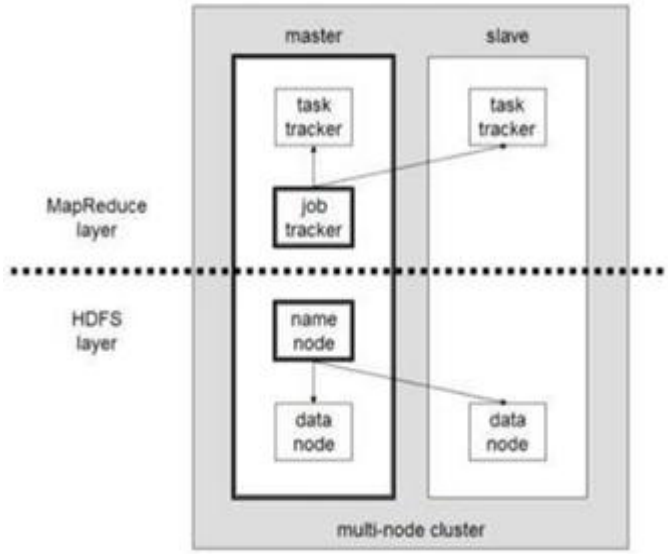

Fig. 3. HDFS cluster setup.

Our project will works in the following manner in below 7 tasks:

- The Map-Reduce library in the user program first splits the input into M pieces of typically 16 megabytes to 64 megabytes (MB) per piece. It then starts up many copies of the program on a cluster of machines.
- One of the copies of the program is special- the master copy. The rest are workers that are assigned work by the master. There are M map task and R reduce tasks to assign; the master picks idle workers and assign each one a task.
- A worker who is assigned a map task reads the contents of the contents of the corresponding input split. It parses key/value pairs out of the input data and passes each pair to the user-defined Map function. The intermediate key/value pairs produced by the Map function are buffered in memory.
- Periodically, the buffered pairs are written to local disk partitioned into R regions by the partitioning function. The location of these buffered pairs on the local disk are passed back to the master, who is responsible for forwarding these locations to the reduce workers.
- When a reduce worker is modified by the master about these locations.it uses remote procedure calls to read buffered data from the local disk of map workers. When a reduce worker has read all intermediate data, it sorts it by the intermediate keys. The sorting is needed because typically many different key map to the same reduce task.
- It passes the key and the corresponding set of intermediate values to the user's Reduce function. The output of the Reduce function is appended to the final output file for this reduce partition.
- When all map task and reduce task have been completed, the master wakes up the user program. At this point, the Map-Reduce call in the user program returns back to the user code.

IV. CONCLUSION

Hadoop with its efficient Data mining technique & programming framework based on concept of mapped reduction, is a powerful tool to manage large data sets. With its map-reduce programming paradigms, overall architecture, ecosystem, fault- tolerance techniques and distributed processing, Hadoop offers a complete infrastructure to handle Big Data. Users must leverage the benefits of Big-Data by adopting Hadoop infrastructure for data processing. However, the issues such as lack of flexible resource management, application deployment support, and multiple data source support pose a challenge to Hadoop‟s adoption. Proper skill training is also needed for achieving large scale data analysis. These challenges has been overcome so that we can tap the full potential of Hadoop data management Power.

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

[4]Yu Li ; Wenming Qiu ; Awada, U. ; Keqiu Li,,(Dec 2012),” Big Data Processing in Cloud Computing

[5]Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. ;,(17-19Jan. 2013),”A Big Data implementation based on Grid Computing”, Grid Computing

[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”K. Elissa, “Title of paper if known,” unpublished.

[10]R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[11]Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[12]M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Author's Profile:

Shaymaa Taha Ahmed, M.Sc (IS) 2nd Year Student, Nizam College, Osmania University, Hyderabad India.
Email id: memprh@gmail.com.

T. Ramdas Naik, B.Tech, M.C.A. M.Tech (Ph.D) Assistant Professor, Dept. of Computer Science, Nizam College, Osmania University, Hyderabad, India.
Email: ramdas.teja@gmail.com,