A brief review on Hadoop architecture and its issues

Er. Prachi Jain, Er. Alisha Gupta

Department of Computer Science and Engineering, Haryana Engineering College, Jagadhri

E mail: prachijain1992@gmail.com

Abstract— With enormous data present all over the world, the need of managing the data has also risen. Hadoop is used to maintain and process such large amount of data. Hadoop is an Apache framework which is used to store and process large amount of data. The data is stored in a distributed environment in Hadoop. Hence, Hadoop consists of hadoop distributed file system which is used to store this large amount of data. The data present in such a large scale and complex structure is termed as big data. It is very difficult to process big data using conventional computing techniques. Therefore different methodology has been opted to process big data. Map reduce is one of them. Map reduce is used for large scale processing. The data is processed by breaking down it into various jobs which are fed as input to map tasks and reducer processes the data came as output from the mapper. Various scheduling algorithms are proposed to schedule these jobs. This paper covers all the aspects related to Hadoop and big data. A study related to various work done in this field is also covered in this paper.

Keywords— Hadoop, Map reduce, Big data, Hadoop distributed file system, job tracker, task tracker, AES-MR

INTRODUCTION

Due to rapid development of internet applications, the demand of computing power has risen manifolds. Many new technologies like grid computing, cloud computing, distributed computing or parallel computing have emerged to provide enormous computing power. [1] Due to invention of cloud computing, more and more applications are now deployed in the cloud environment enabling people to have access to data at very less rate. As a result volume of data has also increased which has led to the invention of big data.

1. Big data

Big data is basically a terminology that is used for very massive data sets that have a large variation along with complex structure. These are the characteristics that usually add difficulties like storing the data, analyzing it and further applying procedures after which results are to be extracted. [2] Big Data is related to data that surpasses the usual storage, processing power, and computing capacity of traditional databases and data analysis techniques. Moreover, to process such a large amount of data, Big Data requires a large set of tools and methods which can be applied to analyze and extract patterns from large-scale data. Need of big data has risen because storage capabilities, computational processing power and availability of large volumes of data has increased. Big data can be characterized by three Vs: Volume, Variety and Velocity. Here volume means that there is large amount of data present which can be in the order of terabytes or petabytes. With variety we understand that the data comes from varied sources like text, audio, video, images etc. Velocity defines how the data is kept in motion and how the analysis of streaming data is done. [3]

2 Big Data technologies

To handle such a large amount of data one common technique is load balancing i.e. to redistribute the data in multiple servers so that load on single server is reduced. HADOOP can be used to handle big data.

2.1 HADOOP

HADOOP is basically a framework provided by Apache which is used to run applications on systems which includes thousands of nodes and data is in the order of terabytes. It handles large amount of data by distributing it among the nodes. [4] It also helps system to work properly even when a node in the network fails. As a result risk of catastrophic system failure is reduced. Apache HADOOP consists of the HADOOP kernel, HADOOP distributed file system (HDFS) and map reduce paradigm. [5]

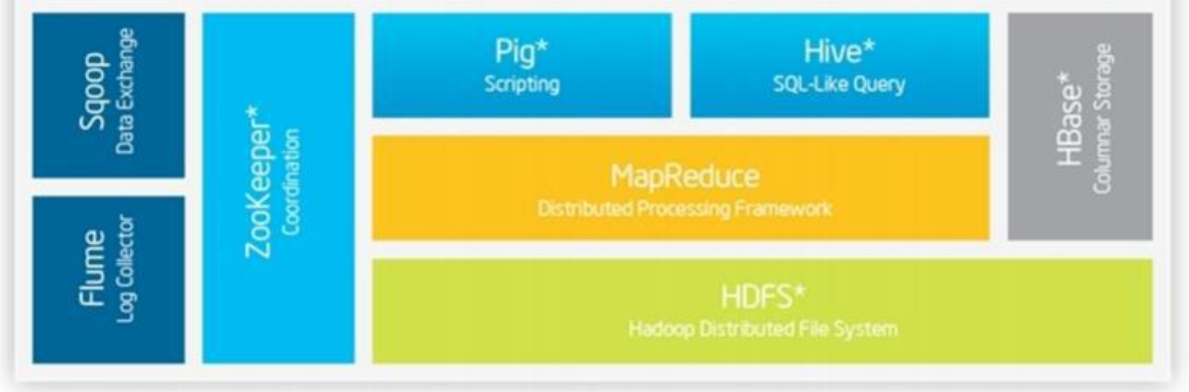

Figure 1: HADOOP Architecture

2.2 Hadoop distributed file system

There is a fault-tolerant storage system present in HADOOP which is called HADOOP Distributed File System, or HDFS. [6] With HDFS we can store huge amounts of information along with scaling up incrementally and surviving the system failure without threat of losing data.

HDFS contains three major components:

- Name node
- Data Node
- Secondary Name node

Name node, also called the master node is the one in which the information about the name system along with the information of the data blocks is present. Whereas data nodes are the nodes where actual storage is done and data can be held and moreover upon request it can be read and write as well. Secondary Name nodes are helper of master nodes. Whenever name node performs any action, a checkpoint is created and that check point saved on secondary name node. Hence, if the master node gets failed, we can restart the node and by using secondary name nodes we can retrieve checkpoints back. Therefore, secondary name nodes act as a saviour during system failure. [5]

2.3 Map reduce

Map reduce is a paradigm used for parallel processing of data using two functions: map and reduce. [7] It provides scheduling, parallelization, replication and failover. Using map and reduce phase map reduce basically encodes data for faster processing. Input to Map tasks is the fixed sized blocks which is obtained by partitioning the input data and then feeding into map tasks in parallel. Collection of key- value pair tuples are generated as intermediate output. These tuples are then sent to different reduce nodes on the basis of key values. Reduce task is used to perform following three steps:

- Copy – The output of map node is copied on reduce node
- Sort – Sorting is done on the collected map output on the basis of key values
- Reduce - Reduce function applied to the data.

Basically there is one master i.e. job tracker and many workers i.e. task trackers present in map reduce architecture. Duty of job tracker is to receive job from user, feed it into map task which breaks it down and then reduce task reduces it. After that job is assigned to task tracker. Progress of task tracker is also monitored by job tracker. Whenever all the work is completed job tracker informs user about the completion. To perform map and reduce task each task tracker is allotted a fixed number of map and reduce task slots. [8]

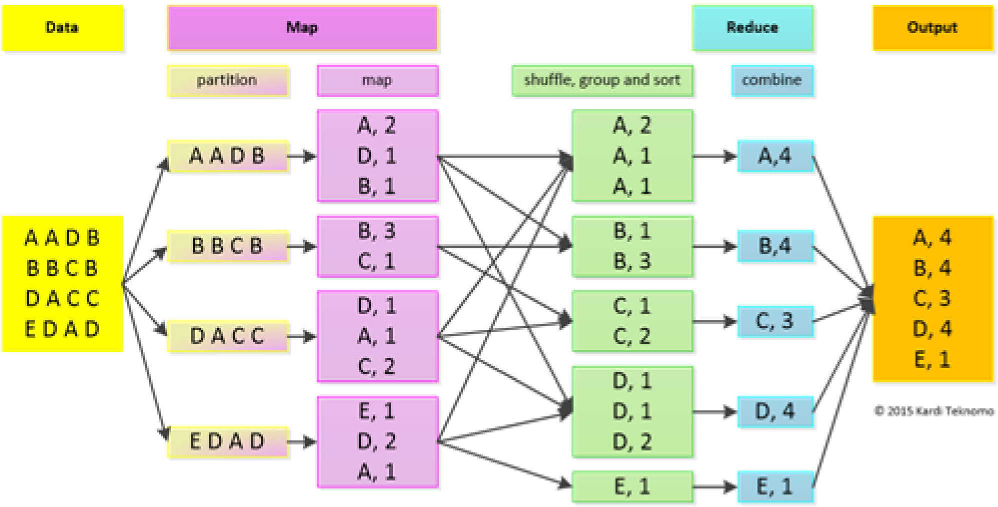

Figure 2: Map reduce

2.4 Scheduling in cloud

There are various algorithms proposed for scheduling of applications in cloud environment. Basically, scheduling can be defined as method to select and decide the task which is most appropriate to execute. It is also defined as allocation of machines to tasks so that makespan of workflow is minimized. [7] Algorithms like data aware scheduling algorithms, first come first serve, round robin, minimum completion time, heterogeneous earliest finish time etc can be used for scheduling workflows in a HADOOP environment. FIFO scheduling algorithm is the default algorithm provided by Hadoop architecture. In FIFO scheduling, jobs are executed in first come first serve order. A FIFO queue is maintained by FIFO scheduler which keeps multiple tasks in it. [4]

2.5 Encryption of big data

As a large amount of data is stored in HDFS, security of data is an important issue in HADOOP. Various algorithms have been proposed to secure data in HDFS. Kerberos is used to attain authorization and authentication in HDFS. Kerberos is a protocol which can be used to authenticate the users who are requesting access to data. However with the invention of new and advanced hacking tools, hackers can now break the security given by Kerberos, hence leaving data in an unsecured state. To overcome this issue, a new algorithm called AES- MR has been proposed which uses AES along with mapper reducer function in parallel. [6]

LITERATURE SURVEY

The work done for the implementation of Hadoop architecture is as follows:

1. Big data: B. Saraladevi et al. has covered the various issues involved in big data in their paper. Big data is a term used to store large amount of data collected from various sources. The various issues discussed in the paper are management, security, storage and processing issues. The author has used various approaches for securing data in HADOOP distributed file system. These approaches are based on Kerberos, Bull eye algorithm and name node. Using Kerberos only authorised users can access HDFS. In bull eye algorithm, how security from node to node can be obtained is explained. The third approach replicates name node so that problem arising due to server crashes can be reduced. These all algorithms are implemented in base layer of HADOOP.

2. HADOOP: Anam Alam in 2014 has provided insight into HADOOP architecture and it's uses. HADOOP has been used in a large number of areas like data intensive applications, social media, data analytics etc. It can be used for opinion mining in terms of social media. The author has worked on the major shortcoming of HADOOP i.e. incremental computations. To overcome this problem, author has used caching. [5]

3. HADOOP DISTRIBUTED FILE SYSTEM: HADOOP has two components: HADOOP distributed file system (HDFS) and Map reduce. HDFS is used to store data and map reduce is used to process the data. Nusrat Sharmin Islam et al. in their paper have proposed various data access strategies which can be used to access HDFS efficiently. They have considered the heterogeneity of data. According to the author's data access strategies, the read performance of HDFS is improved by up to 33% as compared to the default locality aware data access. Moreover, in their study execution times of HADOOP and Spark sort is also reduced by upto 32% and 17% respectively. [10]

4. MAP REDUCE: Map reduce is a framework used to process large chunks of data. Hirotaka Ogawa et al. in their paper have come up with the limitations of existing map reduce framework and have come up with a new map reduce framework called SSS. SSS is based on distributed key- value store. The authors have taken two benchmarks: word count and iterative composite benchmark. According to the study done by the authors, results show that SSS is 1- 10 times faster than conventional HADOOP. [11]

5. SCHEDULING IN HADOOP: Kamal Kc et al. has proposed deadline constraint scheduler in their paper. The conventional scheduler used in HADOOP is FIFO. The authors have tried to design a scheduler which schedules job according to the deadlines specified by users. The study ensures that only those jobs are scheduled whose deadlines can be met. The results show that whenever deadlines for different jobs are different, then scheduler assigns different number of tasks to task tracker due to which whether deadline is met or not is ensured. [9]

6. ENCRYPTION IN HADOOP: Encryption is a process to encode the data in such a way that it is not understandable. While large amount of data is stored in HDFS, security of data has become a primary concern. Viplove Kadre et al. have proposed a new encryption method called AES- MR which works in parallel to encrypt large data sets. The author have combined AES algorithm with mapper function so that encryption is done in a faster manner. The results show that encryption is not only done in a faster manner but security at HDFS level is also improved. [6]

OUTLINE OF HADOOP ARCHITECTURE AND ISSUES

TECHNIQUE	AUTHOR & YEAR	BASED ON	FINDINGS	FUTURE SCOPE
Big Data	B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvanc in 2015	Data security in big data	Three approaches implemented in base layer for securing data in HDFS: Based on Kerberos Based on bull eye algorithm Based on name node replication	These approaches can be implemented in other layers of HDFS
Hadoop	Anam Alam, Jamil Ahmed in 2014	Shortcoming in Hadoop: incremental computation	Caching can be done at three levels: Job base, Task base and hardware base	Caching can be extended at other levels as well
Hadoop distributed file system	Nusrat Sharmin Islam, Md. Wasi-ur-Rahman, Xiaoyi Lu, Dhabaleswar K. (DK) Panda in 2016	Efficient data access strategies for Hadoop and Spark with heterogeneous storage	Read performance of HDFS improved by up to 33% Execution time of HDFS reduced by up to 32% Execution time of Spark reduced by up to 17%	Evaluate impact of data access strategies on different middleware. Dynamic switching
Map Reduce	Hirotaka Ogawa, Hidemoto Nakada, Ryousei Takano, Tomohiro Kudoh in 2010	Map reduce framework	SSS and packed SSS performs 1-10 times faster than Hadoop	Provide fault tolerance Provide higher level programming tool Provide more comprehensive benchmarks
Scheduling in Hadoop	Kamal Kc, Kemafor Anyanwu in 2010	User deadline constraint scheduler	When jobs have different deadlines, scheduler assigns	Map/reduce task runtime estimation

			different number of tasks to task tracker to ensure that deadlines are met	Filter ratio estimation Data distribution Multiple map reduce cycle support
Encryption in Hadoop	Viplove Kadre, Sushil Chaturvedi in November, 2015	Encryption using AES- MR	Data security at HDFS level is enhanced Less time due to parallel processing	Increase number of workers Chunk sizes can be distributed.

CONCLUSION

Currently, the amount of data present in the world has risen so much that need for high computing power has emerged. Hadoop comes as a rescue in such scenario which has enabled us to store and process this large amount of data. Various scholars and researchers have worked on different aspects of Hadoop and big data. Work has been done on different scheduling algorithms to schedule the jobs in Hadoop. Kamal Kc and Kemafor Anyanwu have given user deadline constraint scheduler which performs scheduling on the basis of deadlines specified by users. There is a need to encrypt the data stored in Hadoop. Till date, only Kerberos is used to provide authentication and authorization. Therefore, it is possible to formulate effective encryption algorithms which can be used to encrypt the data to make it more secure.

REFERENCES:

[1] Hong Mao, Shengqiu Hu , Zhenzhong Zhang , Limin Xiao, Li Ruan " A load- driven task scheduler with adaptive DSC for map reduce" 2011 IEEE/ACM International Conference on Green Computing and Communications

[2] Priyank Jain, Manasi Gyanchandani, Nilay Khare "Big Data privacy- a technological perspective and review" Journal of big data, 2016

[3] Tripti Mehta, Neha Mangla "A Survey Paper on Big Data Analytics using Map Reduce and Hive on Hadoop Framework", International journal of recent advances in engineering and technology, vol. 4, Issue 2, February 2016

[4] Sutariya Kapil B, Sowmya Kamath S "Resource Aware Scheduling in Hadoop for heterogeneous Workloads based on Load Estimation", ICCCNT, 2013

[5] Anam Alam, Jamil Ahmed "Hadoop architecture and its issues", International Conference on Computational Science and Computational Intelligence, 2014

[6] Viplove Kadre, Sushil Chaturvedi "AES – MR: A Novel Encryption Scheme for securing Data in HDFS Environment using MapReduce", International journal of Computer applications, vol. 129- No. 12, November 2015

[7] Divya M, Annappa B., "Workload Characteristics and Resource Aware Hadoop Scheduler", IEEE 2nd International Conference on Recent Trends in Information Systems, 2015

[8] Kamal Kc, Kemafor Anyanwu, "Scheduling Hadoop jobs to meet deadlines", 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010

[9] B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvanc, "Big Data and Hadoop-A study in security perspective", 2nd International Symposium on Big Data and Cloud Computing, 2015

[10] Nusrat Sharmin Islam, Md. Wasi-ur-Rahman, Xiaoyi Lu, Dhabaleswar K. (DK) Panda, "Efficient Data Access Strategies for Hadoop and Spark on HPC Cluster with Heterogeneous Storage", IEEE International Conference on Big Data, 2016

[11] Hirotaka Ogawa, Hidemoto Nakada, Ryousei Takano, Tomohiro Kudoh, "SSS: An Implementation of Key-value Store based MapReduce Framework", 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010

[12] Xicheng Dong, Ying Wang, Huaming Liao, "Scheduling Mixed Real-time and Non-real-time Applications in MapReduce Environment", 2011 IEEE 17th International Conference on Parallel and Distributed Systems, 2011