International Journal of Computer Engineering & Technology (IJCET)
Volume 8, Issue 5, Sep-Oct 2017, pp. 11–14, Article ID: IJCET_08_05_002
Available online at
http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=8&IType=5
Journal Impact Factor (2016): 9.3590(Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6367 and ISSN Online: 0976–6375

BIG DATA PARADIGM AND A SURVEY OF BIG DATA SCHEDULERS

Suja Cherukullapurath Mana
Assistant Professor, Dept. of Computer Science,
Sathyabama University, Chennai, Tamil Nadu, India

ABSTRACT

As the volume and scale of information need to be stored in a database increases, the traditional database methodologies become inadequate. As the versatility of data increases, data becomes difficult to represent using traditional database systems. Many big data systems have emerged to store the big data. Hadoop is one such system. Job scheduling is a vital consideration in all big data systems. There are several job scheduling algorithms being used in big data systems. This paper surveys and compares some of the scheduling algorithms used in big data systems.

Key word: Big data, Hadoop, job scheduling.

Cite this Article: Suja Cherukullapurath Mana, Big Data Paradigm and a Survey of Big Data Schedulers. *International Journal of Computer Engineering & Technology*, 8(5), 2017, pp. 11–14.
http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=8&IType=5

1. INTRODUCTION

The term big data represent any dataset which cannot be represented and interpreted efficiently using traditional database systems currently available. Dataset coming from banking, healthcare, weather, e commerce, Social media fields can be classified as big data [1] . Three important characteristics of big data are volume, velocity and variety [1]. Due to the exponential growth of data there are some challenges in processing these datasets [2]. Keeping the privacy and security of data, analysis of data, efficient storage and data management are some of the key challenges of handling big data. Job scheduling in big data applications is also an interesting field of research. This survey paper progresses in such a way that in the coming sections it will briefly describes the architecture of an efficient big data system and then provide a comparison study of some of the job scheduling algorithms used in big data applications.

2. ARCHITECTURE OF AN EFFICIENT BIG DATA SYSTEM

Many efficient big data processing systems are in rise, out of which Hadoop [3] is one of the efficient big data management system. The key part of Hadoop is the Map Reduce function [1]. MapReduce has two user defined functions namely Map and Reduce [4]. Another important component of Hadoop is Hadoop distributed file system [HDFS]. HDFS is capable

of storing huge data files. HDFS has two nodes namely Name node and Data node .Metadata is stored in Name node and data is stored in Data node.

As mentioned above map reduce function consist of two important modules. Map module accepts input data and produce intermediate data in the form of key value pairs. Reduce function will accept the input from Map function and produces the final result. The input to Map function will be obtained from HDFS. A typical Map-Reduce function diagram is given below.

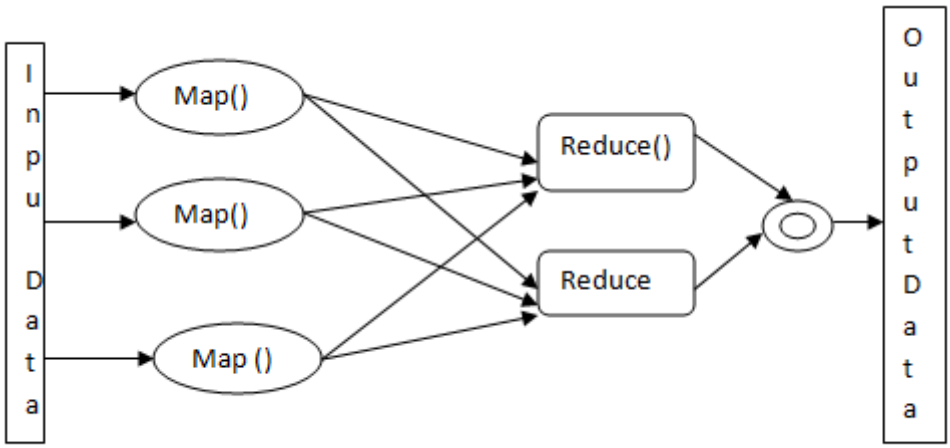

Figure 1

Input from HDFS function is passed to the Map function. Map function process this input and send the output to the Reduce function as shown in the diagram [5]

One of the key challenges in the Hadoop system is an efficient job scheduling policy within the Map Reduce architecture. An efficient scheduling policy must perform the best utilization of resources. Some of the factors affecting the scheduling policies are Locality, Fairness and Synchronization [1] , Following section discusses some of the schedulers used in big data applications . Classification of Hadoop scheduler is done based on scheduling strategy, time and resource availability [1].

3. FIRST IN FIRST OUT (FIFO) SCHEDULER

As per FIFO scheduler the jobs submitted first will get priority over the job submitted later. This kind of scheduler is mainly used when the order of execution of job is not important. When a new job arises, oldest job from the queue will be replaced by this new job. Only after finishing the current job, the resource will be allocated to other processes. It will result in long wait time for some processes.

3.1. Capacity Scheduler

In capacity scheduler several queues with defined capacities will be created. Each queue will have defined Map, Reduce slots [1]. Overall capacity of the cluster will be some of capacities of these queues. Queues will be monitored and if one queue is not using its allocated capacity the remaining capacity will be temporarily allocated to some other queues. The capacity scheduler was first developed by Yahoo [6]. One of the notable characteristic of capacity scheduler is the ability to control resource allocation based on physical machine resources [1].

3.2. Fair Scheduler

The main idea of fair scheduler is to allocate resources fairly. The fair scheduler groups jobs into pools and assign fair shares among these pools. There will be a guaranteed minimum share for each pool . Minimum pool represents the least amount of resources that any process will get [1]. Excess shares will be divided equally among pools.

3.3. Delay Scheduler

Waiting approach is used by delay scheduler to enhance locality [1]. The idea behind this category of scheduling is to wait and see if the data is present in local node itself. For example if a node request for a task and if the data for that task is not found on local node , then the scheduler will skip that task and look for subsequent jobs for the time being . But if that task has to wait for a long , then the scheduler will launch a non local task and avoid starvation.

4. DEADLINE CONSTRAIN SCHEDULER

Deadline constrain scheduler schedules job based on the deadline specified by user [7]. The job with nearest deadline will be scheduled and executed first. This type of scheduler accepts user constraints and deadlines as inputs and then checks to see if the job can be executed within the available slot. A job is determined to be schedulable if the total minimum number of operations for both Map and Reduce is less than or equal to available slots [1].

4.1. Matchmaking Scheduler

Matchmaking scheduler enhances the locality feature of Map task [1] . Scheduler will pick the local task first than the non local task. A locality marker is used to mark to ensure that each node get a substantial opportunity to grab local tasks. But one of the drawbacks of this kind of scheduler is that it may result is high response time for Map task.

5. CONCLUSIONS

This paper provides a brief overview of a big data management system. It then briefly surveys some challenges in performing an efficient job scheduling in big data system. Some schedulers like FIFIO, Capacity scheduler, fair scheduler etc are being discussed. Depending on the various characteristic requirements the best scheduling algorithm will be selected while designing a big data application. For example if we need to resolve the fairness issue fair scheduler and capacity scheduler are best suitable .If we are considering locality issue delay scheduler is the best option . Thus this paper provide a survey of some of the best suitable scheduling algorithms for big data applications.

REFERENCES

[1] Harshadkumar Prajpathi, Vipul Dabhi,Sanjay Chaudhary ”A survey of job scheduling algorithms in big data processing” 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)

[2] M. Kaur and Shilpa, “Big Data Visualization tool with Advancement of Challenges,” Int. J. of Advanced Research in Comput. Sci. and Software Eng., vol. 4, 2014.

[3] Apache Hadoop. (2014, September 10) [online]. Available: http://hadoop.apache.org.

[4] S. Bardhan and D. A. Menasce, “The Anatomy of Map Reduce Jobs, Scheduling, And Performance Challenges,” Conf. of the Comput. Meseaurement Group, San Diego,CA , 2013.

[5] https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm

[6] J. Chen, D. Wang, and W. Zhao, “A Task Scheduling Algorithm for Hadoop Platform,” J. of Comput., vol. 8, no. 4, pp. 929-936, 2013.

[7] K. Kc and K. Anyanwu. "Scheduling hadoop jobs to meet deadlines." Cloud Computing Technology and Sci. (CloudCom), 2010 IEEE 2nd Int. Conf. IEEE, pp. 388-392, 2010.

[8] Getaneh Berie Tarekegn and Yirga Yayeh Munaye, Big Data: Security Issues, Challenges and Future Scope, International Journal of Computer Engineering and Technology, 7(4), 2016, pp. 12–24.

[9] K. Prema and Dr. A.V. Sriharsha, Differential Privacy in Big Data Analytics for Haptic Applications. International Journal of Computer Engineering & Technology, 8(3), 2017, pp. 11–19.

[10] Naga Raju Hari Manikyam and Dr. S. Mohan Kumar, Methods and Techniques To Deal with Big Data Analytics and Challenges In Cloud Computing Environment. International Journal of Civil Engineering and Technology, 8(4), 2017, pp. 669-678.

[11] Ms. Gurpreet Kaur and Ms. Manpreet Kaur. Review Paper on Big Data Using Hadoop. International Journal of Computer Engineering and Technology, 6 (12), 2015, pp. 65-71.