ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Comparative Analysis of Mapreduce Scheduling Algorithms

[1]Bijoy Joseph, [2]Mani Bushan D'Sozua

Department of M.Sc Software Technology, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India[1]

Asst. Prof, Department of MCA, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India[2]

ABSTRACT: MapReduce is a programming model to deal and process large data sets. The idea behind MapReduce is to partition the data sets into multiple smaller data sets and send it to commodity hardware for processing in parallel. The process of dividing the data sets and sending it to different compute clusters is called **Mapping**. The business intelligence drawn at these clusters is grouped together and then the final output is generated. This process is called **Reduce**.

There are many scheduling algorithms available today. But the problem with such scheduling algorithms are that they can't predict beforehand what is going to happen with the commodity hardware in the cluster. There are many other factors as well such as locality and synchronization overhead. Moreover, the data which are given to the compute clusters should be distributed evenly among them for processing.

The algorithm should be good enough to decide what has to be done in case of cluster failures and various other problems. My paper reviews various MapReduce Scheduling Algorithms. A comparative analysis of these algorithms will be done based on various scheduling issues.

KEYWORDS: Map; Reduce; Resource Scheduling; Big Data

1. INTRODUCTION

Big Data and MapReduce

Big Data are the data sets, both structured and unstructured, which are very large and complex. The volume of data is so large that traditional DBMSs can't handle and manipulate such large data sets. The storage and processing requirements of such data sets demands very high computing capability. MapReduce is one of the programming models to deal with big data. Wherein the data sets are divided and is fed to compute clusters for processing and the business intelligence is drawn.

Hadoop software platform is a very popular implementation of MapReduce programming model. Hadoop is a software ecosystem where the data sets are stored in the HDFS (Hadoop File System) and the processing part : MapReduce. These are the two different components of the Hadoop Software Framework. The entire framework is written in Java.

The HDFS – distributed file system of Hadoop provides us to store data sets on different locations. This also facilitates in very rapid processing of the data sets.

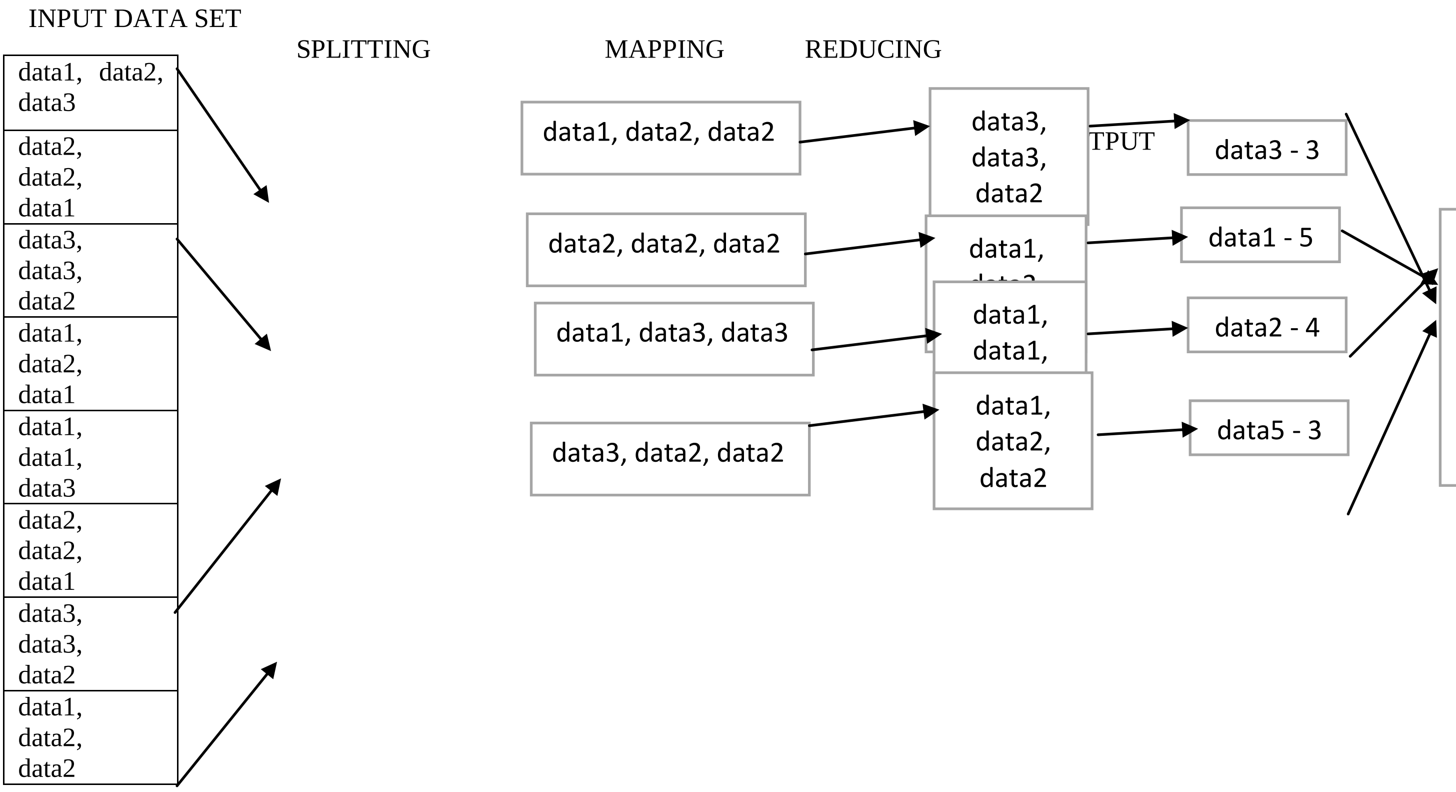

Fig: 1 Overview of a MapReduce job

II. MAPREDUCE SCHEDULING AND RELATED ISSUES

There are many scheduling algorithms available today. But the problem with such scheduling algorithms are that they can’t predict beforehand what is going to happen with the commodity hardware in the cluster. There are many other factors as well such as locality and synchronization overhead. Moreover, the data which are given to the compute clusters should be distributed evenly among them for processing.

The algorithm should be good enough to decide what has to be done in case of cluster failures and various other problems. My paper reviews various MapReduce Scheduling Algorithms. A comparative analysis of these algorithms will be done based on various scheduling issues.

III. MAPREDUCE SCHEDULING ALGORITHMS

Challenges of MapReduce

Job Scheduling: The basic concept on which MapReduce works is assigning the job (data sets) to heterogeneous clusters of commodity hardware. Since we are dealing with commodity hardware, with not a very good build quality and which does not guarantee a robust processor, the task of generating the desired result can’t be guaranteed. Some computers in the cluster may crash is the job is not distributed equally. If some computers in the cluster is busy on a task and it has not yet finished then the algorithm should be able to identify free computers or the task should be assigned to a different cluster.

Energy Efficiency of the clusters: A cluster may contain a large pool of computers ranging from a few hundreds to thousands. The energy consumption is directly proportional to the number of computer in the cluster and the overall compute clusters. There is a need to look at the energy efficiency of the clusters.

Scheduling Algorithms: FIFO Scheduler and Fair Scheduler – a comparision

1.1 FIFO Scheduler: In this scheduling technique is the hardware in the cluster is assigned free slots. When a tasked has to be mapped, the tasks are then assigned to each nodes in the cluster first come first serve. If a situation arrives when a task takes too long to complete then that task is pushed to the next nodes which are free and the next task in the node is assigned to them. All jobs should complete its execution in a timely manner.

1.2 Fair Scheduler: In this scheduling methodology, the free slots in the cluster are assigned a fair share of memory. It means that the tasks which are assigned to them should not exceed the amount of memory which has been assigned to them to complete itself. The user may assign the tasks to these slots for Map. If a process takes too long to complete over a certain period of time. To those processes are killed to free up the memory so that the next tasks in the queue take their place.

IV. RESULT

Table 1: Analysis of FIFO and Fair Scheduling algorithm: advantages and disadvantages

Scheduling Algorithms	Description	Advantages	Disadvantages
FIFO	Jobs are scheduled in first-in-first-out manner. Based on their priorities.	1. Cost for scheduling the jobs on the cluster is less. 2. When it is not running, it automatically dislocates nodes from the cluster. 3. There is less sharing of nodes therefore greater safety. 4. It is simple to implement.	1. Designed only for single type of job. 2. Can't handle heavy load 3. Job priority is not given any importance.
Fair Scheduler	The jobs are assigned equal resources for execution. If a job takes longer to execute then that job is terminated.	1. Works well with large and small clusters. 2. It is less complex 3. Fast response time for small jobs mixed with large jobs	1. It does not considers the job weight. 2. 3. Does not consider availability of resources.

V. CONCLUSION

In this paper, we have discussed about MapReduce scheduling issues. Then, we explained two popular scheduling algorithms in this field namely FIFO and Fair Schedule. We then Analyzed these algorithm based on implementation, advantages and disadvantages. We can see that most of these schedulers discussed in this paper, addresses one or more problem. And choice of this scheduler for a particular job is up to the user. From Analysis table it is evident that fair Scheduler is comparatively better than FIFO in many cases.

REFERENCES

[1] A. N Nandakumar and Y. Nandita, " A Survey on Data Mining Algorithms on Apache Hadoop Platform", International Journal of Emerging Technology and Advanced Engineering, Vol. 4, NO. 1, January 2014, pp.563-566.

[2] Z. Tang, L. Jiang, J. Zhou, K. Li, and K. Li, "A self-adaptive scheduling algorithm for reduce start time ", Future Generation Computer Systems, 2014.

[3] K. Morton, M. Balazinska and D. Grossman, "Paratimer: a progress indicator for MapReduce DAGs", In Proceedings of the 2010 international conference on Management of data, 2010, pp.507–518.

[4] Lu, Wei, et al. "Efficient processing of k nearest neighbor joins using MapReduce ", Proceedings of the VLDB Endowment, Vol. 5, NO. 10, 2012, pp. 1016-1027.

[5] J. Dean and S. Ghemawat, "Mapreduce: simplied data processing on large clusters", in OSDI 2004: Proceedings of 6th Symposium on Operating System Design and Implemention,
(New York), ACM Press, 2004, pp. 137–150.

[6] Pranoti K. Bone and A.M.Wade, "Survey on Hadoop MapReduce Scheduling Algorithms", *Grenze Int. J. of Engineering and Technology, Vol. 1, No. 1, January 2015*

[7] Seyed Reza Pakize, "Comprehensive View of Hadoop MapReduce Scheduling Algorithms" International Journal of Computer Networks and Communications Security, VOL. 2, NO. 9, SEPTEMBER 2014, 308–317, ISSN 2308-9830

[8] Rakesh Varma,"Survey on MapReduce and Scheduling Algorithms in Hadoop", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438