RESEARCH ARTICLE OPEN ACCESS

Heuristic Based Job Scheduling Approach to Process Large Data Packets

Anjali Nema [1], Deepak Gour [2]
Department of Computer Science & Engineering
Technocrats Institute of Technology
Bhopal - India

ABSTRACT
Cloud computing and processing big data with the cloud environment is an important aspect. It deal with the packet processing between multiple users. Component such as virtual machine, data center and other processing unit help in data process which need a proper input optimization technique. The proper utilization of resources is always required. Thus job schedule approaches help in building the architecture beyond to process input request. Previously given approach make use of resource status, last used resource history and other similar mechanism to process the upcoming request. The lacking in these approaches are somewhere heavy burden on particular resource. In this paper an Hybrid advance Heuristic based approach for the resource allocation job scheduling approach is proposed. The proposed mechanism is experimented using Java framework and result computation efficiency shows the effectiveness of given algorithm.

Keywords:- Map reduce, Medical Data processing, Phase level scheduling, task level scheduling.

I. INTRODUCTION

A technique named Map Reduce which is a high performance computing paradigm scenario for large-scale data (Big data) processing in clusters and data centers, and it has become very popular in recent years. An open source implementation of Map Reduce is known as Hadoop. In order to work with large data processing by the companies such as Google, Yahoo!, Bing and other search engine they use technique for data processing. There have been many techniques and file distribution technique which already been given in the name of data processing. Also in the field several key challenges for the utilization and performance improvement of a Hadoop cluster are still there. Map Reduce: In order to process the data in distributed system there are few steps and logical entities which are invented by the professionals. Google first introduced a popular parallel model known as Map Reduce. Map Reduce is designed to handle and generate large scale data sets in distributed environment. A convenient way to parallelize data analysis process is also provided by Map Reduce. It has some advantages such as conveniences, robustness, and scalability. All the map tasks (and reduce tasks) are executed in a fully parallel way in a Map Reduce system. Therefore, through the use of the Map Reduce model, high- level parallelism can be achieved for data processing. The basic idea of MapReduce is to split the large input data set into many small pieces and assign small tasks to multiple devices in a distributed environment. The process of MapReduce includes two major parts, Map function and Reduce function. First of all, the input files will be automatically split and copied to different computing nodes. Later on, the inputs in key-value pair format will be sent to Map function. The input pairs will be processed by map function and it will also generate intermediate key-value pairs which will be inputs for Reduce function. The inputs which have the same key are combined and the final result will be generated by reduce function. The final result will be written into the distributed file system. Only the Map and Reduce functions are needed to be implemented by users.

II. RELATED WORK/SURVEY

The health care industry is generating large amounts of data. While most of the data is in hard copy format, the latest trend is to move towards digitization of this massive volume of data. Reports say that in U.S. alone the healthcare data reached upto 150 exabytes in 2011[2,10].
As different tasks may have different resource requirements, this resource allocation scheme may lead to resource contention or poor resource utilization. Also Run-time resource consumption varies from task-to-task, which makes it difficult to effectively utilize available resources.

Task Level Scheduling: A task level scheduling in which jobs are allocated irrespective of any system configuration using which the working is according to its equal distribution technique where the performance may vary in case the over

work is allotted to less configure system and less work is allotted to heavy duty system, thus task level technique perform less efficiency in terms of processing big data[4]. Working of Task Level Scheduling:

The master node runs a Job-tracker that is responsible for scheduling tasks on slave nodes. Then each slave node runs a Task-tracker that is responsible for launching and allocating resource for each task. For this, task-tracker launches a JVM that executes the corresponding map or reduce task. This Hadoop MapReduce adopts a slot-based resource allocation scheme.[3] The scheduler assigns tasks to each machine based on number of available slots on that machine.

Phase Level Scheduling[2] : A Phase level scheduling take first a knowledge about the system availability and its configuration , based on the notification regarding configuration it give job allocation to the available machine in distributed system , thus is provide an optimize technique to process data in available resources. Such that the technique gets a full utilization of the resources and get maximum efficiency using this concept.

Advantage of Phase level scheduling:

•Removes the limitations of task-based Map Reduce resource scheduling algorithm

•Minimization of job completion time

•Maximization of resource utilization

•Hadoop cluster[6] is shared by multiple users with multiple jobs

•Assigning resource capacity to jobs such that all jobs get an equal share of resource capacity

•Guarantee minimum shares to resource pools or jobs.

•Maintain a job queue, sorted according to fairness.

PRISM[1] : A technique which uses for the big data processing it uses a phase based scheduling for the Map Reduce contains three components : first ,Phase-based scheduler at master node which is the scheduler which having an acknowledgement of all the available resources and to which resource need to be access and assign the particular incoming job. Second, Local node manager is a component which is going to provide the job getting done by each CPU. Third, Job progress monitor which is going to take care of the job progress part and going to provide the status to the master node.

III. EXISTING ALGORITHM

The existing paper given a PRISM technique for processing the data in map reduce manner which involve the following process: Prism mechanism:

Node manager sends heartbeat message to scheduler where the purpose of this process is to get a knowledge about the available resources and the data which is already being getting process by the machine such that the new job allocation probability can monitor over the machine using this step. After getting a process request Scheduler replies a task scheduling request based on the available task for the execution. After finishing a phase execution, task asks permission to start new phase execution to node manager. Node manager forwards the request to scheduler by regular heartbeat message.

Decision to start a new phase is made based on:

Job's phase-level resource requirement

Jobs current progress information

If allowed, the node manager grants permission to task progress

After finishing the task, the status received by the node manager is forwarded to scheduler for the message as its availability with the data processing.

IV. PROBLEM DEFINITION

The problem associate with the current prism technique for map reduce and various drawbacks are associated with the technique are : Increases the complexity of the system as its required a number of process to allocate the actual task and a monitoring level may need time to have an extra step in this Map reduce. Require major change in MapReduce implementation which is might be difficult to implement and process every time in different resources and application. Each task is divided into phases with different resource consumption characteristics.

V. PROPOSED SOLUTION

A further steps need to be taken as investigate the phase level processing of big data where the straggler problem is must be included such that the actual possible implementation can be performed.

A Hybrid approach following the heuristic behavior is proposed in the current paper. The approach help in

scheduling each resource properly according to its requirement and based on current real time utilization. The technique help in allocating the resources in real time. A Re-distribution of resources is also provided which help preventing from node failure.

The experimental setup is performed using the Java framework and packet processing with the different mechanism is executed. A packet distribution and redistribution is performed. Further the next section shows the comparison analysis between the traditional job scheduling approach and proposed hybrid technique of job scheduling.

VI. RESULT ANALYSIS

This section includes the analysis of result on the basis of several parameters which we have used. The parameter for the computation comparison such as computation time, Energy consumption of each mechanism and Overall throughput is calculated. The Computed result are shown below in table 1.

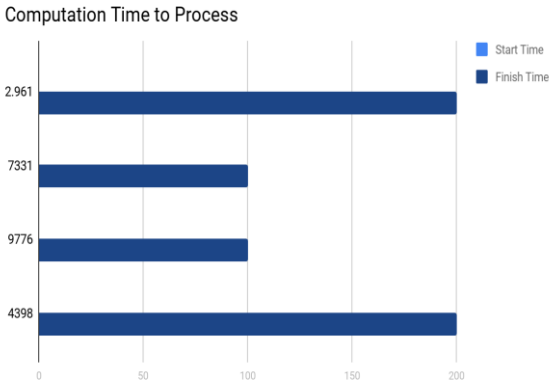
Table 1. Computation Parameters

Parameters	Resource Aware Approach	Hybrid Heuristic Behavior Approach
Computation Time (in ms)	10078	6687
Computation Cost to Process	2.0	1.0
Energy Consumption	0.9842	0.49214
Throughput	49120	12232

In the table 1 above, the computation analysis is shows between the existing approach and proposed Hybrid heuristic based mechanism of job scheduling analysis..

Table 2. Results of Various parameters.

Response Time	Start Time	Finish Time
2.961	0.0	200
7331.0	0.0	100
9776.0	0.0	100
4398.0	0.0	200



So, according to the calculations we have done in finding out the computation time can be seen in the table 1. The given above result analysis shows the efficiency of proposed approach while considering large data transmission.

VI. CONCLUSION

Job scheduling is the mechanism of resource allotment and sharing as per the requirement to take best utilization of component. There are mechanism which optimize the use of complete architecture and provide the overall efficiency. In the existing approach techniques such as resource aware, credit based , history usage based allocation for the request is performed. In this paper a novel hybrid approach using Heuristic behavior is used. The proposed technique take use of equal utilization of resource behavior of heuristic optimization. The proposed technique compute the efficiency while executing with different input file and processing then with different number of resources. The proposed technique implemented using Open source Java platform and proven the efficiency over the existing techniques for resource job allocation technique. Future work include the experimenting the technique with real time dataset. Also an application using open source social media platform and their resource allocation can be performed.

REFERENCES

[1] Qi Zhang, Student Member, IEEE, Mohamed Faten Zhani,” PRISM: Fine-Grained Resource-Aware Scheduling for MapReduce”, IEEE TRANSACTIONS

ON CLOUD COMPUTING, VOL. 3, NO. 2, APRIL/JUNE 2015 2168-7161 _ 2014 IEEE.

[2] S.hanjiang Tang, Bu-Sung Lee, Bingsheng He, "DynamicMR: A Dynamic Slot Allocation Optimization Framework for MapReduce Clusters", IEEE Transactions, 2013

[3] Divyakant Agrawal, UC Santa Barbara, Philip Bernstein, Microsoft Elisa Bertino, Purdue Univ. " Big Data White pdf", from Nov 2011 to Feb 2012.

[4] Dipali Salunkhe, Devendra Bahirat, Neha V. Koushik, Deepali Javale," Study of Hadoop Features for Large Scale Data" (IJITEE) ISSN: 2278-3075, Volume-4 Issue-6, November 2014.

[5] Dimitrios Markonis, Roger Schaer, Ivan Eggel, Henning M¨uller, Adrien Depeursinge," Using MapReduce for Large–scale Medical Image Analysis".

[6] K. P. Andriole, J. M. Wolfe, and R. Khorasani. Optimizing analysis, visualization and navigation of large image data sets: One 5000–section CT scan can ruin your whole day. Radiology, 259(2):346–362, May 2011.

[7] Seyyed Mojtaba Banaei1, Hossein Kardan Moghaddam," Hadoop and Its Role in Modern Image Processing", Open Journal of Marine Science, 2014, 4, 239-245.

[8] Dean, J. and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM, 51, 107-114.

[9] Bakshi, K. (2012) Considerations for Big Data: Architecture and Approach. Aerospace Conference-Big Sky, MT, 3-10 March 2012

[10] Apache Hadoop Available at http://hadoop.apache.org

[11] Apache HDFS Available at http://hadoop.apache.org/hdfs

[12] Apache HBase. Available at http://hbase.apache.org

[13]Map Reduce Simplified Data Processing on Large Clusters Available at http://labs.google.com/papers/mapreduceosdi04.pdf