Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 02, February -2018

Job Scheduling In Big Data Using Cuckoo Optimization Technique

D. S. Dayana [1], D. Godwin Immanuel [2]

[1]*Department of Computer Applications, SRM Institute of Science and Technology, Chennai*
[2]*Department of Electrical and Electronics Engineering, Sathyabama Institute of Science and Technology, Chennai*

Abstract — T*o examine large volume of data and to extract hidden data, applications of big data analytics is used. To schedule the job and data, cuckoo search scheduling algorithm is used in this proposed study. The cuckoo search algorithm permits providers and consumers of resources to perform the decisions for scheduling on their own. So depending on their requirement, providers and consumers achieve enough amounts of data. The objective of this paper is to minimize the overall turnaround timing and execution cost and to maximize the utilization of the resources. To achieve this objective, Cuckoo Search Algorithm (CSA)is designed depending on Confidence Time Gap (CTG). Hadoop is a software framework that stores huge volume of data in a cluster and allows to process data from all nodes. Map Reduce is a application framework used to process huge volume of data in clusters. The efficiency of Big Data Analytics is improved by implementing job scheduling using Cuckoo Search Algorithm. This algorithm is more efficient and convenient than the available resource brokers implementing various data-based job scheduling algorithms.*

Keywords- *Confidence Time Gap (CTG); turnaround time; resources; scheduling*

I. INTRODUCTION

To extract the hidden data, applications of big data analytics is used that process huge amount of data. Hadoop is a software framework that stores huge volume of data and this allows the application framework, Map Reduce to process the huge volume of data[5]. To process the data jobs are not separate and the configuration of cluster is also individual. The performance can be measured by the execution of the job, job characteristics and its clusters. When there is a need to execute several jobs, the outcome result for scheduling will be large and when we select manually, it become inefficient. The demand for data increases every year various applications. In 2000, the applications related to scientific produced hundred terabytes of data and in the year 2018, it reaches to several billion terabytes of data every year [1]. So, the processing power and the space required to store the data is not sufficient as represented by the Moore's Law [2].

High-capacity resources such as supercomputers, high-bandwidth networks and mass storage systems are required for the analysis of high-energy physics, molecular modeling and earth sciences datasets as there will be wide distributions over a wide geographic area [4]. With the help of big data analytics technology, the data that is distributed and the heterogeneous processing, resources and storage can be merged to obtain the objective. Big Data is used to store, distribute, manage and analyze larger-sized datasets with high-velocity and different structures. Most of the big data analytics technologies are job-based or resource based technologies. The data needed for processing is included in jobs and data is placed in nodes. For scheduling, the job and the data is allocated. Large-scale issue occurs when jobs and data are scheduled. So a combined scheduling technique is needed for the allocation of job and data. As huge volume of data is used, the turn-around time of the entire system gets increased.

The job scheduling problem based on job and data is focused in the proposed work. To provide the needed resources to nodes, the providers and consumers are given permission to take decision for scheduling on their own. This paper focus on the Cuckoo Search Algorithm (CSA) with heuristic methods. The computational difficult problems can be solved by cuckoo search algorithm.

II. PERFORMANCE OF HADOOP

Hadoop's performance prediction. J. Berlinska et al. [6] propose a mathematical model of MapReduce, and analyzeMapReduce distributed computations as a divisible load scheduling problem, but they do not consider the system constraints. Zaharia et al. [7] proposed prediction model forsub-tasks of Hadoop job, rather than the entire job. Lijie Xu et al. [8] extracted characteristic values related to hadoop performance and utilized machine learning methods to find the optimal value, without building performance models. Jungkyu Han et al. [9] proposed a Hadoop performance prediction model. The scheduling algorithm is used for the efficient utilization of the resources in system. It uses CloudSim as a simulator to evaluate the algorithm performance. The evaluation is simplified due to the limitations in CloudSim[12]. The incorporation of both data location and processing resources characteristics into the job scheduling decision have also been tried [10]. The speed of the computing resources is also balanced and the data transfers delay is insignificant in real life. It is also difficult for users to indicate the lifetime of their job [11].

@IJAERD-2018, All rights Reserved

III. PROPOSED METHODOLOGY

Job scheduling in big data analytics is a optimization problem. The proposed algorithm performance evaluated by Job scheduling in compound equipment system. Here an issue arises for unfitting of job scheduling in a common description system. In this proposed method, the issues in data and job scheduling issues can be solved in the big data analytics environment.

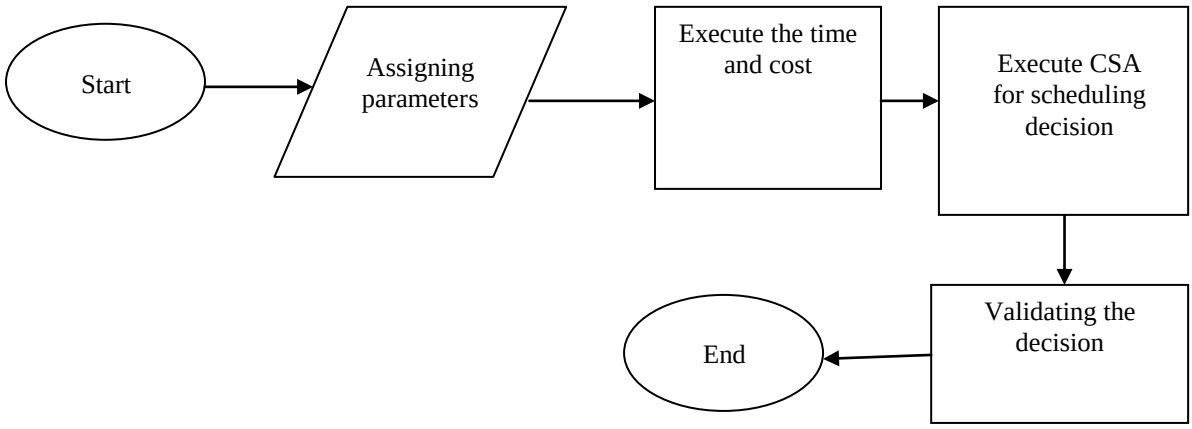

Figure 1. Flow diagram

Cuckoo search framework is used to integrate the service of job and data for each provider of resource. In the service and reservation stage, the time to complete the job is added for each job. The availability of data is done by the rank expression in scheduling. The meta scheduler allow users to interact with resource managers [13].

IV. ALGORITHM

The proposed Cuckoo search algorithm for job scheduling is shown below

Step 1: consumer node submits a job to all the provider nodes

Step 2: Acknowledgement of job is received, each provider checks whether it will be able to meet the requirement.
If yes, the price of job is send to provider in a bid;
If no, job is ignored by the provider

Step 3: From all the bids received, the consumer node selects the provider node which is charging the least cost and to that node the job is send.

Step 4: Calculate the unit price
Adjust all the provider nodes

Step 5: Resources and data is selected

Step 6: Sort in descending order considering value weight ratio and a queue is formed.

Step 7: Initialization
Randomly generate cuckoo nests
For each node calculate the fitness
Set the counter for generation. Set parameter for mutation

V. PROPOSED METHODOLOGY

The performance is evaluated using the resource files and Amazon EC2 test platform is used. The goal for the proposed methodology is to select the best scheme for resource allocation using cuckoo search algorithm. The execution time and cost is also measured using the estimation module. Then it generates a 2-D array for time and cost. So we can identify various jobs that has different time and cost on the same cluster. Also to finish its execution, the same jobs running will take different cost and time. For implementation, cuckoo search algorithm is used.

@IJAERD-2018, All rights Reserved

In huge-scale parallel and distributed processing environments, Amazon EC2 test platform is used as a toolkit for resource modeling and simulation of application in job scheduling. The parallel and cluster processing systems is used and it consumes huge volume of processing power. The performance of the job scheduling is evaluated and the proposed algorithm is working efficiently. The algorithm's performance is evaluated on various loads by set of job information with various data types submitted at various time intervals. Experiment results are compared with the previous scheduling algorithm and the CSF as shown in Table 1.

Table 1: Parameters and attributes used for simulation

Parameter	Values
Number of clusters	180
Number of nodes	30
Bandwidth capacity	200
Processing capacity	512 to 1031
Total providers	300
Total users	750
Cost of users (units)	100-1000
Cost of resources(units)	100-2000
Number of files	100-15000
Size of a file	100

Experiment results in terms of cost with Cuckoo search algorithm (CSA) and the existing Cuckoo search framework is shown in Table 2.

Table 2: Performance in terms of cost

Jobs	Resources Used	Adaptive Qos	Economic Adaptive Qos	CSA	Improvement
50	14	502	452	356	12.1
100	34	771	672	564	13.4
150	84	1196	992	897	17.2
1000	104	1486	1162	956	22.6
2000	134	1672	1212	1013	27.8

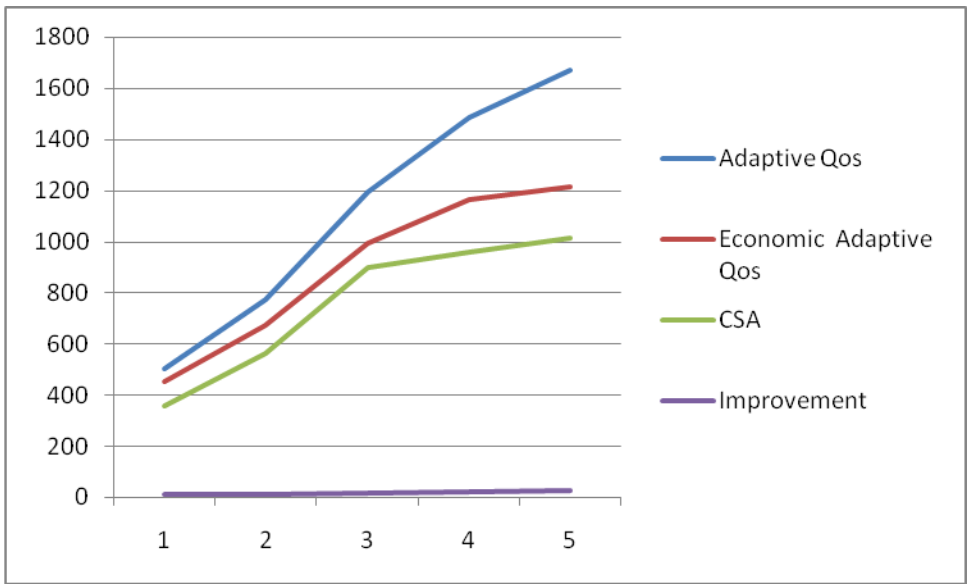

Figure 2: Performance in terms of cost

@IJAERD-2018, All rights Reserved

The results is compared with the performance of the proposed CSA is having less cost as shown in Fig.2. When the number of jobs gets increased, there is increase in the usage of system resources. The cost also gets increased when compared with existing methods.

VI. CONCLUSION

In big data analytics application processing, making decision for job and data scheduling is a big issue. The characteristics such as job, data and cluster are considered for computation of data. In this paper, cuckoo search algorithm is used to minimize the turn-around time and execution cost and to maximize the utilization of the resources. The whole system is evaluated and the proposed methodology is feasible. This study can be used for future work on job scheduling in big data analytics.

REFERENCES

[1] M. Mitchell,An Introduction to Genetic Algorithms, MIT Press, 1998.

[2] K. Deb, “An introduction to genetic algorithms”, Sadhana,vol.24,no.4-5,pp.293□315,1999.

[3] Linderoth, L., and Wright, S.J., 2003, “Decomposition algorithms for stochastic programming on a computational grid”, Computational Optimization and Applications, 24(2-3), pp.207-250.

[4] Paniagua, C., Xhafa, F., Caballé, S., and Daradoumis T., 2005, “A parallel grid-based implementation for real time processing of event log data in collaborative applications”, In: Parallel and Distributed Processing Techniques, Las Vegas, USA, pp. 1177-1183.

[5] Gray, J., and Shenoy, P., 2000, “Rules of thumb in data engineering,” in Proceedings of the IEEE International Conference on Data Engineering, San Diego, CA, February.

[6] J. Berlinska, M. Drozdowski. “Scheduling divisible MapReduce computations”, Journal of Parallel and Distributed Computing, 71(3):450-459, 2011.

[7] M. Zaharia, A. Konwinski, A.D. Joseph, R. Katz, and I.Stoica.“Improving MapReduce performance inheterogeneousenvironments” , Proc. of OSDI, 2008.

[8] Lijie Xu, □MapReduce Framework Optimization via Performance Modeling, Proc. of Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012.

[9] Jungkyu Han, Masakuni Ishii and Hiroyuki Makino. “A Hadoop Performance Model for Multi-Rack Clusters”, the 5th International Conference on Computer Science and Information Technology (CSIT), 2013.

[10] Venugopal, S., Buyya, R., and Winton, L., 2004, “A Grid Service Broker for Scheduling Distributed Data-Oriented Applications on Global Grids”, Grid Computing and Distributed Systems Laboratory Technical Report,GRIDS-TR-2004-I (University of Melbourne, Australia).

[11] Tang, M., Lee, B-S., Tang, X., and Yeo, C-K., 2006, “The impact of data replication on job scheduling performance in the Data Grid Future Generation Computer Systems”, 22, 254–68.

[12] Qinghua Lu, Shanshan Li, Weishan Zhang.,2015 “Genetic Algorithm based Job Scheduling for Big Data Analytics” 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things,33-38.

[13] G. Kalpana, Dr. D. I. George Amalarethinam., 2015 “ Cuckoo Search Based Community Workflow Scheduler Framework For Grid Scheduling With Resource Constraints” International Journal of Applied Engineering Research 10, 8981-9001.

@IJAERD-2018, All rights Reserved