RESEARCH ARTICLE OPEN ACCESS

Analysis of Dynamic Workflow scheduling Algorithm for big data application

Roopa G Yeklaspur[1], Dr.Yerriswamy.T[2]
1(Department of CSE KLE Institute of Technology Hubballi)
2 (Department of CSE KLE Institute of Technology Hubballi)

Abstract:
In the era of big data, even though we have large infrastructure, storage data varies in size, formats, variety, volume and several platforms such as hadoop, cloud since we have problem associated with an application how to process the data which is varying in size and format. Data varying in application and resources available during run time is called dynamic workflow. Using large infrastructure and huge amount of resources for the analysis of data is time consuming and waste of resources, it's better to use scheduling algorithm to analyse the given data set, for efficient execution of data set without time consuming and evaluate which scheduling algorithm is best and suitable for the given data set. We evaluate with different data set understand which is the most suitable algorithm for analysis of data being efficient execution of data set and store the data after analysis

Keywords — **Dynamic data set, workflow scheduling algorithm, HDFS.**

I. INTRODUCTION

Efficiency is important factor for the analysis of data, progress in the efficiency and improvement within given time span and effort for the output of data analysis in the execution environment and storage in the large infrastructure known as Hadoop Distributed File System(HDFS). The given application consists of sub application which changes data in terms of time these data are analysed using scheduling algorithms such as MinMin, Minimum Completion Time(MCT), First Come First Serve(FCFS), where these data i.e Big Data does not remain static changes dynamically how we access the application which in turn consists of sub application varies both in speed and volume to process and store these large data in a executional environment need is HDFS In Hadoop distributed file system different types of file and huge data file can be stored in hadoop environment. HDFS Stores huge data on clusters and provides inexpensive and reliable large amount of data. HDFS provides large storage and capacity for large data analysis. How to process large data in an efficient way as data changes dynamically during the execution of application it is very difficult for the user to understand and process the data in Big Data.

Due to change and variation in data to process large data is big task. In HDFS computing environment it provides storage, processing and analyzing big data,efficient execution of data can be managed by the scheduling algorithm and resources. Scheduling algorithm manages the data workflow management system in an efficient way within time constraints. The scheduling algorithm FCFS, MinMin, Distributed Heterogeneous Earliest Finish Time(DHEFT), MaxMin which performs the best for the given data set in cloud environment. The resources and time if it is preallocated and the given data set length is not known than there is huge wastage of resources and time. Performance analysis work should be done on the given data set using scheduling algorithm. In this paper we are analyzing the data set using scheduling algorithm and which algorithm is suitable and do the fast analysis for the given data set. Different algorithm performance is different for different data set and data set file length may vary. To check the performance of the these algorithm data set file should be known in priori so it is easy for the analysis and check which algorithm give better result. Among all these algorithm DHEFT performs best in all the scenario. The scheduling of task is done only one task at a time, to show the performance of all scheduling algorithm. Virtual machine can execute only one task at a time.

.

II. SYSTEM MODEL

A. *Workflow Model*

Given the workflow model, the large data set is divided in to sub task, each sub task

is interdependent until all sub task are completed no other new data set is taken, the large data set is divided into sub task called as 11,12,13, the sequence of execution is important in the workflow model, each sub task completes its execution in interdependent way until previous subtask completes its execution.

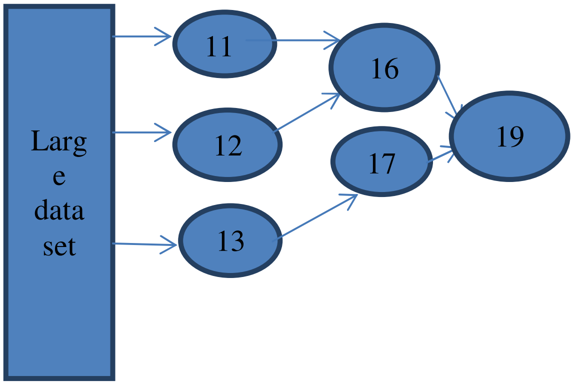

Figure.1. Data workflow

B. Cloud Server Model

Cloud server is the main module which provides dedicated resources to each user to perform all processing activities. Virtual machines are createdand registered user can only access the server. It is the main server which receives file for uploading to virtual machine and this file is divided in to number of tasks. Virtual machine executes the files uploaded and executes one task at a time and remaining tasks are shared between virtual machines.

C. *Scheduling Algorithm*

Scheduling are often implemented so they keep all computer resources busy. Scheduling assigns task to virtual machine and share all computer resources effectively.

During execution phase initially we need to decide which physical machine is most suitable and appropriate to assign the task and providing the most suitable resource the task is mapped to the physical resource for execution, after providing resources schedule the task to the appropriated working virtual machine.

There are various scheduling algorithm these algorithm are used for the given large data set to evaluate how these algorithm are most suitable for the large data set to analyze the performance of algorithm on the large data set. The algorithm are MinMin Algorithm, Data Aware Scheduling Algorithm, MaxMin Scheduling Algorithm, FCFS Scheduling Algorithm, MCT Algorithm, DHEFT Algorithm. To achieve the best results, performance and resource utilization of user provided datasets the jobs are scheduled according to the arrival of tasks.

- FCFS Scheduling Algorithm : In this algorithm the tasks are kept in ready queue, the tasks are selected first come first serve basis, not according to minimum task completion time, the tasks are assigned to available resources in the arrival order. If longest job is selected first task the smallest task has to wait for longer time.

- MinMin Scheduling Algorithm : The tasks are sorted in ascending order with minimum completion time and allocates resources to the fastest job, this process repeats until all the jobs are scheduled.

- MaxMin Scheduling Algorithm : This algorithm is reverse of MinMin algorithm and selects the maximum completion time and allocates the resources to the task.

- Data Aware Scheduling Algorithm : In this algorithm the data is transferred and stored into vacant virtual machine to its closest resources among the pool of virtual machine to schedule and execute the dataset.

- Distributed heterogeneous Earliest Finish Time : is an algorithm it enables us to calculate the average bandwidth between each pair of virtual machines.

- Minimum Completion Time : is an algorithm which assign the task to the available resources which completes its execution with minimum completion time.

III. EXPERIMENTAL SET UP

We present experimental set up in Clouderaenvironment :

A. *Experiment Environment*

Main aim of the experiment is to check to the performance of scheduling algorithm on big data. In this paper we have used Cloudera for experimental set up which includes eclipse,virtualmachine. Eclipse is an Integrated development environment.It contains a base workspace and primary usedfor developing java appilcations. In this application we have created 3 virtual machines, User Registration, Cloud server for the computation of Dynamic data set. These computation of tasks is assigned to the virtual machines by the server and data set is divided in to number of tasks and these tasks are assigned to all virtual machines at the same time to execute.

B. *Workflow*

In this paper we use the dynamic data set to check performance analysis of the different size data set and same size data set using scheduling algorithm and compare the performance results of scheduling algorithm for different size data set.

C. *Performance Metric*

The metric we used to perform of the scheduling algorithm on each of the different size dataset. To minimize the execution time of the task we evaluate the performance.

D. *Scenarios*

To evaluate the algorithms we show the execution analysis. During the execution phase two things to be considered that is delay and task length. Delay can occur due system fault or memory required for the application might not be enough and with the increase in task length this will occur because of each time change in input dataset size this change is due to big data that is speed and volume. Following are the results.

IV RESULTS AND GRAPH

A. *Plotted Graph*

In this results will compare the results for different dynamic data sets on various scheduling algorithm. Figure 2. Shows the analysis of data set using scheduling algorithm we have considered initially 12kb of data set the plotted graph shows the result analysis of x-axis is the scheduling algorithm and y-axis varies with time. Among the six algorithm which performs best is the question? FCFS has the worst performance among all the six algorithms. This is because how each of them schedules the task of the given data set is different from each other. The makespan is sum of data processing time, transfer time of data from storage to execution time, waiting time in the queue and computation time[1].

Communication cost is same for all algorithm but differs in computation capacity and waiting time in queue. DHEFT performs best among all six algorithm its selects the less computation task and also selects the machine which performs the task in less time without waiting for the resources. MinMin algorithm selects the task with minimum completion time task to the machine. MaxMinselects the task with maximum completion task to the available machine here completion of time takes larger time, the first available machine is assigned the which is having the maximum computation task. Other algorithm performs with the availability of machines depending on the computational task. DHEFT algorithm performs better than all other algorithm as per the analysis of plotted graph the grey color signifies the DHEFT algorithm which takes less time for the computation of given data set which had resources available with computation capacity and less waiting time. As per the plotted graph we can compare the timing taken to analysis for the given data set of 12kb by various scheduling algorithm.

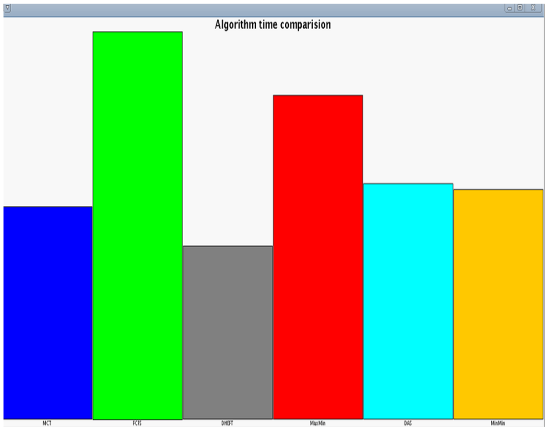

Figure.2. Algorithm time comparision for 12kb of data set.

Colour indication MCT-blue, FCFS-green, DHEFT-grey, MaxMin-red, DAS-blue, MinMin-yellow.

Next we consider figure 3, the dataset of 25 kb, here also MinMin selects the minimum task size and schedules the dataset to the available machine,the performance remains the same for 25kb of dataset even though it selected minimum task size the performance remains same because of the increase in data size computational capacity requirement is more, DHEFT performance is best in this dataset, MaxMin algorithm performance in this dataset is worst among all the other algorithm because longest task is selected for computation and delay might incur during processing of task waiting for the resources, the remaining algorithm performance is not good as DHEFT.

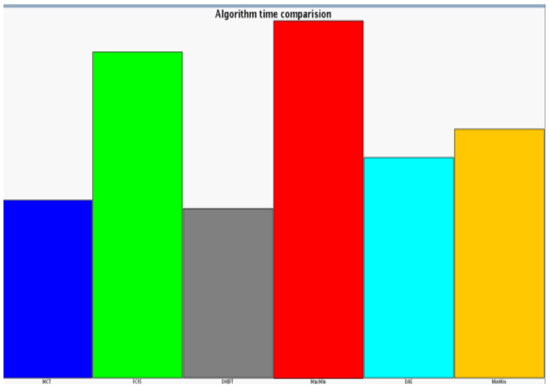

Figure.3. Algorithm time comparision for 22kb of data set.

Next 55kb of dataset figure 3 shows the plotted graph of the result, in this given data set of 55kb DHEFT performance is better ,DAS performance is better than other algorithm the task is assigned to the closest machine where data transformation is reduced of the whole work span and MinMin algorithm selecting minimum task but takes computational time of all the sorted tasks to be completed in order by checking the availability of resources.MaxMin performance is always vice versa of MinMin algorithm, remaining algorithm performance is worst such as FCFS.

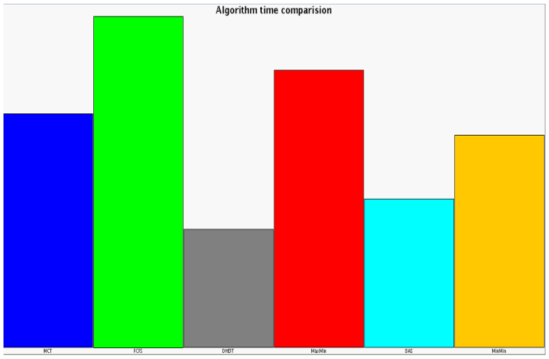

Figure.4. Algorithm time comparision for 55kb of data set.

V.CONCLUSION

The performance of scheduling algorithm on various data set of big data including the FCFS, MCT, MinMin, MaxMin, DAS, DHEFT algorithm in order to improve efficiency of big data. Different data set performance of algorithm will vary in different way with dynamic dataset.Hence dynamic big data resulted its performance on varying data set size and the availability of resources and computation capability. Hence dynamic big data after the data analysis the data is stored in encrypted form in HDFS. Dynamism associated with big data can have huge influence on scheduling algorithm with different size of data set. For future work other scheduling algorithm can be applied to the dynamic data and size of the input data size can also be enhanced.

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

REFERENCES

1. *Chaochao Zhou, Saurabh Kumar Garg,"Performance Analysis of Scheduling Algorithms for Dynamic Workflow Applications",School of Engineering and ICT,University of Tasmania Hobart, Australia..*
2. *National Institute of Standards and Technology, The NIST Definition of Cloud Computing, NIST, US.Department of Commerce, Special Publicaton, 2011.*
3. *H. Topcuoglu,"Performance-Effective and Low-Complexity Task Scheduling for Heterogeneous Computing", IEEE Transactions on Parallel andDistributed Systems, vol. 13, no. 3, 2012, pp. 280-294.*
4. *S. Abrishami, M. Naghibzadeh, and D.H.J. Epema,"Deadline-constrained Workflow Scheduling Algorithms for Infrastructure as a Service Clouds ", Future Generation Computer Systems, vol. 29, 2013, pp. 158-169.*
5. *L.K Arya and A.Verma "Workflow Scheduling Algorithms in Cloud Environment- A Survey" Proceddings of 2014 RAECS unjab*
6. *M.A. Rodriguez and R. Buyya, "Deadline Based Resource Provisioning and Scheduling Algorithm for Scientific Workflows on Clouds", IEEE Transactions on Cloud Computing, vol. 2, no. 2, 2014, pp. 222-235.*
7. *J. Yu and R. Buyya, "A Taxonomy of Workflow Management Systems for GridComputing," journel of Grid Computing, Springer, pp. 171-200,*
8. *J. Yu and R. Buyya, "A taxonomy of scientific workflow systems for gridcomputing," SIGMOD Record, Vol. 34, no. 3, pp. 44-49, 2005.*
9. *G. Singh et al. "Workflow task clustering for best effort systems with Pegasus," InProceedings of the 15th ACM Mardi Gras conference: Fromlightweight mash-upsto lambda grids: Understanding the spectrum of distributed computingrequirements, applications, tools, infrastructures,interoperability, and theincremental adoption of key capabilities. ACM, 2008.*
10. *W. Chen, R. Silva, E. Deelman, and R. Sakellariou, "Balanced Task Clustering inScientific Workflows," In proceedings IEEE 9th International Conference oneScience, pp. 188-195. 2013.*
11. *L. Meyer et al. "An Opportunistic Algorithm for Scheduling Workflows on Grids", High Performance Computing for Computational Science, 2007,*