IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue-15, May 2019 E-ISSN: 2347-2693

A Survey on Service Oriented Scheduling for Big Data Cloud

R Srinath[1*], Arun Biradar[2]

[1]Dept.of ISE, The National Institute of Engineering, Mysuru, India
[2]Department of Computer Science and Engineering, EWIT, Bengaluru, India

DOI: https://doi.org/10.26438/ijcse/v7si15.254256 | Available online at: www.ijcseonline.org

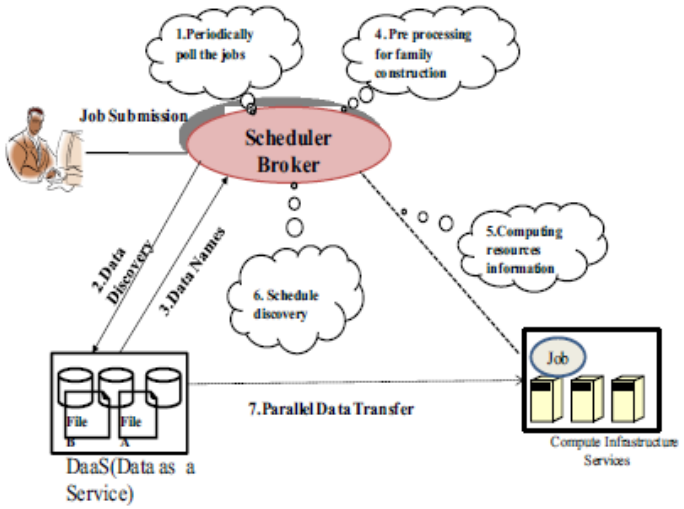
***Abstract*—** Big Data is an emerging data intensive computing technology to extract intrinsic information from large scale variety forms of rapidly growing data. Big Data Analytics is a data science paradigm, which employs several statistical and machine learning tools for effective and quick decision making. As Cloud computing technologies are coming into reality, several Cloud providers are offering large scale computing and storage facilities as services based on pay and consumption models to the end users. Due, to their service oriented delivery of Clouds, these are turning as back end infrastructure to address several big data mining problems in Big Data computing. As the convergence of Clouds and Big Data is turning into new area aka "Big Data Clouds", there is a need to address several under pinning technical elements of Big Data computing in Clouds. In this paper, we discuss Service oriented scheduling mechanisms to serve Big Data Analytics in Clouds infrastructure as services. Our main focus is on scheduling aspects, which bring out the several issues, thus meeting the constraints, and Quality of Service (QoS) parameters. We initially, bring about several challenges in scheduling the Big Data problems over Clouds infrastructure, followed by offering the service oriented analytics deliver over Clouds based SLAs and the Quality of Service while considering the dead line, budget constraints**.**

***Keywords*—**Scheduling, Big data cloud , Quality of service

I. INTRODUCTION

Big Data is an emerging data intensive computing technology to extract intrinsic information from large scale variety forms of rapidly growing data. Big Data Analytics is a data science paradigm, which employs several statistical and machine learning tools for effective and quick decision making. As Cloud computing technologies are coming into reality, several Cloud providers are offering large scale computing and storage facilities as services based on pay and consumption models to the end users. Due, to their service oriented delivery of Clouds, these are turning as back end infrastructure to address several big data mining problems in Big Data computing. As the convergence of Clouds and Big Data is turning into new area aka "Big Data Clouds", there is a need to address several under pinning technical elements of Big Data computing in Clouds.

In Clouds, the resources are offered as services to the end users. Such services are dynamically provisioned to the customers according to their need basis. As Big Data Analytics require the computing and storage resources for analysis, the same could be delivered by the respective Cloud providers, with whom the users are subscribed for carrying out the necessary computing. Hence, Clouds make as back end platform for computing on demand and payments are made on based on the resources those are consumed. However, the challenges here include, performing the application execution without violating the service Level Agreements and meeting the constraints. The major constraints are from scheduling the applications with the constraints like time and budget. The scheduling mechanisms need to be designed such that, the results could be delivered meeting the Quality of Service, without violating the Service Level Agreements thus meeting the constraints.



The system architecture is as below:

Figure 1: System Architecture

Rest of the paper is organized as follows, Section I contains the introduction of Cloud, Big data Analytics and

scheduling , Section II contain the related work of scheduling in big data cloud, Section III contain the some metrics to consider for Scheduling in big data cloud, IV concludes.

II. RELATED WORK

In this section, Previous works on scheduling in Data Grids [1][2]have been more concerned with the relationship between job assignment and data replication based on computation and data proximity. Mohammed et.al [3] discussed a Close-to-Files algorithm, searching the entire solution space for a combination of computational and storage resources for minimizing the processing time with the restriction of one dataset per job for execution.

Srikumar[4] described scheduling the distributed data intensive applications on global grids based on a set coverage approach for cost and time minimizing problems. This approach is based on the availability of both computation and data resources; however, data transfer from replicated sites and the selection of efficient computing nodes for minimizing the execution times are not addressed.

Big Data computing frameworks such as Apache Hadoop [5] is an open source implementation for MapReduce scheduling methods; the examples are Fair [6], Capacity [7], and Throughput [8].

Fair Scheduler is a pluggable group scheduler where in each group gets equal time slots for computation. Capacity Scheduler is similar to FIFO within each queue, but limiting the maximum resources per queue. Throughput Scheduler reduces overall job completion time on heterogeneous cluster by actively assigning tasks to computing nodes based on the server capabilities. Shared Scan Schedulers S3

[9] allows sharing the scan of a common file for multiple jobs arriving at different time intervals thus improving the performance of multiple jobs which are operating on a common data file.

[12] The proposed family/group scheduling model addresses the data intensive problems to minimize the turnaround time of the jobs where the computing and data resources are decoupled. The paper uses group scheduling for data-aware , big data cloud. The scheduling is implemented using genetic algorithm/

III. METRICS

Scheduling in big data cloud, the following metrics are used.

- Data proximity
- Cost
- Time
- Fair
- System Capacity
- System throughput
- FIFO
- Total compute time

IV. CONCLUSION

As the convergence of Clouds and Big Data is turning into new area aka “Big Data Clouds”, there is a need to address several under pinning technical elements of Big Data computing in Clouds one among is scheduling. The objective is to minimize the turnaround time of the jobs over the computing nodes. Determining the scheduling aspects, which bring out the several issues, thus meeting the constraints like dead line, budget constraints, and Quality of Service (QoS) parameters

ACKNOWLEDGMENT

We express thanks to Dr. RajKumar Buyya, Director, CLOUDS Lab, University of Melbourne, Australia for initiating the thought of Big Data Cloud Scheduler. We express our thanks to Dr. Raghavendra Kune, Scientist, Advanced Data Processing Research Institute (ADRIN), Dept. of Space, Hyderabad, India for providing the valuable guidance.

REFERENCES

[1] K. Ranganathan, and I. Foster, “Decoupling Computation and Data Scheduling in Distributed Data-Intensive Applications”, Proc. 11th IEEE Symposium on High Performance Distributed Computing (HPDC). Edinburgh, UK, USA, July 2002.

[2] T. Phan, K. Ranganathan, and R.Sion, “Evolving toward the perfect schedule: Co-scheduling job assignments and data replication in wide-area systems using a genetic algorithm”, Proc. 11th Workshop on Job scheduling Strategies for Parallel Processing. Cambridge MA: Springer-Verlag, Berlin, Germany, June 2005.

[3] H. Mohamed, and D. Epema, “An evaluation of the closeto-files processor and data co-allocation policy in multiclusters”, in Proc. 2004 IEEE International Conference on Cluster Computing, San Diego, CA, USA, Sept. 2004.

[4] S. Venugopal, Scheduling Distributed Applications on Global Grids, Ph.D. Thesis, University of Melbourne, Australia, July 2006.

[5] Apache Hadoop, http://hadoop.apache.org/ (15.06.2014).

[6] Fair Scheduler , http://hadoop.apache.org/docs/r1.2.1/fair_scheduler.pdf(11. 06.2014)

[7] Capacity Scheduler, http://hadoop.apache.org/docs/r1.2.1/capacity_scheduler.pdf (11.06.2014)

[8] S. Gupta, C. Fritz, R. Price, J. D. Kleer, and C. Witteveen, Throughput Scheduler: learning to schedule on heterogeneous

Hadoop clusters, Proceedings of the International Conference on Autonomic Computing, ICAC 2013, June, 2013, San Jose, CA, USA.

[9] L. Shi, X. Li , and K.L. Tan, S3: An efficient Shared Scan Scheduler On MapReduce Framework, International Conference on Parallel Processing, ICPP 2011, Taipei, Taiwan, September 2011**.**

[10] M. D. Assuncao, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, Big Data Computing and Clouds: Trends and Future Directions, Journal of Parallel and Distributed Computing, Available online 27 August 2014, DOI: 10.1016/j.jpdc.2014.08.003

[11] R. Calheiros, R. Ranjan, A. Beloglazov, C. Rose, and R. Buyya, CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms, Software: Practice and Experience, 41(1): 23-50, Wiley Press, New York, USA, January 2011S. Tamilarasan, P.K. Sharma, "*A Survey on Dynamic Resource Allocation in MIMO Heterogeneous Cognitive Radio Networks based on Priority Scheduling*", International Journal of Computer Sciences and Engineering, Vol.**5,** No.**1**,pp.**53-59**, **2017**.

[12] R. Kune, K. P. Kumar, A. Agarwal, C. R. Rao, R. Buyya, "*Genetic Algorithm based Data-aware GroupScheduling for Big Data Clouds*", *Proc. International Symposium on Big Data Computing (BDC 2014)*, pp. 96-104, December 2014.

Authors Profile

Mr. R. Srinath pursed Bachelor of Engineering from Kuvempu University, Karnatak in 1997 and Master of Technology from VTU, Belagavi in year 2002. He is currently pursuing Ph.D. and currently working as Associate Professor in Department of Information Science and Engineering, The National Institute of Engineering, Karnataka, since 2006. He is a life member of CSI & IEI. His main research work focuses on Big Data Analytics, Data Mining, and Cloud Computing. He has 19 years of teaching experience and 2 years of Research Experience.

Dr Arun Biradar. Professor & Head, Department of Computer Science & Enginnering, EWIT, Bengaluru. He is a life member of ISTE. He has published more than 50 research papers in International and National Journal. He has presented his research work in many International and National conferencs. His main area of research is Wireless Sensor Networks, Cloud Computing, Big Data Analytics, Genetic Algorithm and Artifical intelligence. He has more than 28 years of experience.

