International Journal of Advanced and Innovative Research (2278-844)/

Volume 6 Issue 8

A SURVEY ON DATA SCHEDULING AND ITS ANALYSIS ON BIG DATA

Mukesh Rawat , Prof. Angad Singh

Research Scholar, Associate professor
Department of IT,
NRI College of Science and Technology, India
Mukeshrawat1121@gmail.com

Abstract: Now-a-days, large amount of data and its processing through a series of component via complex combination is derived. Data processing, its scheduling is a mechanism of processing large amount of data with appropriate component via scheduling it with given use. Data processing usage make virtual machine, data center and other component utilization for processing of real time or large input dataset. In this paper a survey of different scheduling mechanism, survey of different technique which take participate in multiple scheduling of data algorithms is performed. There are various past approach such as round robin, heuristic based approach, matrix based computation and other well defined peer structure architecture is defined. Paper also present a comparison of previous technique which help in exhibiting previous scheduling approach, their limitation and further work which can be done with their solution.

Keywords: Data Mining, Apache Hadoop, Cloud Computing, Quality of Service (QoS),Map Reduce.

1. Introduction

Large data processing and its executing through the multiple virtual machine platforms is a utilized structure today. In this architecture various authors have presented their own framework and communication components [1]. Data processing helps in working with large dataset, pruning, classification and using them as per the requirement [3]. Virtual machine is a component which helps in data processing as per the requirement and demand. Data utilization, data immigration, data processing,

Producing it into chunk and further deployment
Is an important aspect deals with the process.

Virtualization is concept which helps in using single machine with parallel processing and its major functionality. Data process keeps it secure and energy efficient for the user for service utilization. In the past work author have worked towards data delivery in efficient time frame and efficient throughput outcome [4].

Hadoop is a large data processing platform which use a mechanism of Map reduce. Map leads to break a task in multiple concurrent task. Further there is mapper class which help in doing this mapping concept [17]. A reduce mechanism help in re-producing the output of all concurrent process into the single output system. Map Reduce is a mechanism which supported by Hadoop library and API which is described below.

1. **Hadoop Distributed File System (HDFS)**– It is used for distributed storage of large amount of data with high throughput access to data on clusters [16].

2. **Hadoop MapReduce** – A software framework for distributed processing of data on clusters. Hadoop MapReduce programming model consist of two data processing functions. These are:
   a. **Map() function** : which performs the filtering part.
   b. **Reduce() function** : which performs the summary of the output generated by the Map() function.

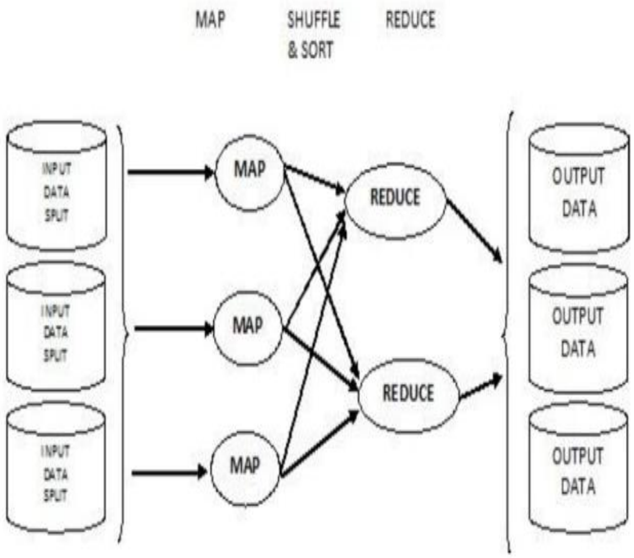

Figure 1. The MapReduce Architecture [16]

2. Related Work

There has been a big quantity of previous work in energy efficient computing systems. Comprehensively reviewing the prevailing literature closely associated with our work.

Abhijeet Desai et al [12] designed a new architecture called Advanced Control Distributed Processing Architecture (ACDPA) by combining the abstraction property of Software Defined Networking (SDN) and the distributed processing power Hadoop. Software Defined Networking (SDN) is an approach where abstraction is used to simplify the network in two layers: 1) Used for controlling the traffic. 2) Used for forwarding the traffic. SDN uses openflow protocol which is an open protocol for controlling and configuring the swites in the network. The system provides abstraction for the control of network traffic and also provide quick processing of big data i.e. network traffic. The control is done using SDN controller opendaylight [19] and the processing is done using Hadoop. The system accepts large amount of data from SDN data plane through wireshark and gives it to Hadoop for processing. The result from Hadoop is feedback to the SDN controller which controls the Quality of Service (QoS). The figure 2 shows overview of Advanced Control Distributed Processing Architecture (ACDPA).

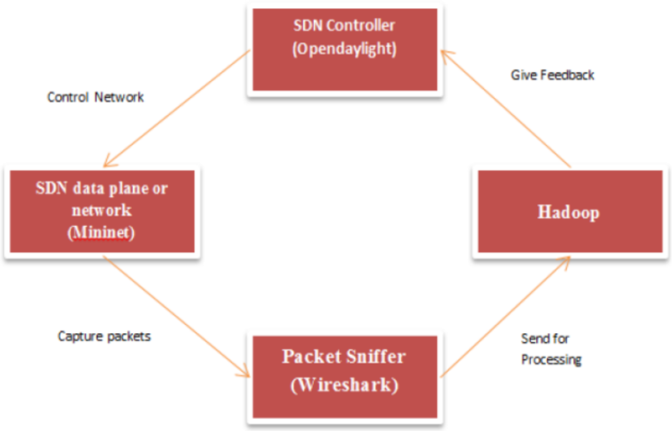
SDN Controller (Opendaylight)
Control Network
Give Feedback
SDN data plane or network (Mininet)
Hadoop
Capture packets
Send for Processing
Packet Sniffer (Wireshark)

Figure 2. Overview of Advanced Control Distributed Processing Architecture (ACDPA) [12].

Ninikrishna T et al [13] proposed an architecture which provides better security to the cloud and also fulfill Quality of Service (QoS) requirements, using Software Defined Networking (SDN) and Hadoop. Kerberos is also used to enhance the security & provide authentication and Single Sign On (SSO). Their contribution includes: 1) Set the QoS requirements for group of users of the cloud using the concept of SDN. Using SDN the segregation of the flows thus controlling the QoS can be accomplished. 2) Increases the security of the cloud by employing Kerberos and the SDN in tandem with Hadoop. The security aspect will be integrated and easier to detect & react to threats due to the programmability of the network. Long term attacks can be detected by using Hadoop analysis.

Rasha Osman et al [14] introduced a measurement-driven methodologyto quantify the impact of replication on DataBase-as-a-Service (DDaaS) environments. The methodology builds upon an analytical model representing the database cluster configuration combined with an environment model to represent the transient replication stages. The main contributions are: 1) Exploit fluid modeling techniques to approximate response time percentiles for replicated relational DBaaS platforms. 2) Methodology evaluate the performance of a relational database cluster hosted on DBaaS platforms. 3) Methodology evaluated under variable workloads and dynamic cluster re-configurations. 4) Evaluated the impact of replication overhead and time on performance stability and its effect on DB cluster performance.

Z. Zhao et al [15] proposed a cluster content caching structure in Cloud Radio Access Networks (C-RANs), which takes full advantage of centralized signal processing and distributed caching. The structure proposed improves the Quality of Service (QoS) and reduce the power consumption in real-time services. In particular, redundant traffic on the backhaul can be reduced because the cluster content cache provides a part of required content objects for Remote Radio Heads (RRHs) connected to a common edge cloud. The tractable expressions are derived for both energy efficiency and effective capacity performance, which shows that the proposed structure can improve QoS guarantees with a low power cost of local storage. The joint design of RRU allocation and RRH association has been studied to further improve the performance of cluster content caching.

Omer Ranaet al [23] measures and characterize energy consumption for high throughput workloads of a number of virtual machines running the hadoop system over an OpenNebula Cloud. The main focus of this work is to highlight how the power consumption : 1) can be related to the number of Virtual Machines (VMs) & the associated workload generated on a physical server. 2) can be monitored and understood & sub-sequently exposed to the user. In this, the approach considers two types of workloads: 1) Virtual machines deployed on the server. 2) Data analysis algorithm executed over the virtual machine.

Xiaobo Fan et al [4] presented how power usage varies over time and as the number of machines increases from individual racks to cluster of up to five thousand servers. Their key findings and contributions are: 1) Power capping using dynamic power management can enable additional machines to be hosted but is more useful as a safety mechanism to prevent overload situations. 2) Observed time intervals when large groups of machines are operating near peak power levels suggesting that power management and power gaps techniques might be more easily exploited at the datacenter- level. 3) CPU voltage/frequency scaling, a technique targeted at energy management, has the potential to be moderately effective at reducing peak power consumption once large groups of machines are considered. 4) Evaluated the benefits of building systems that are power efficient across the activity range.

3. Conclusion

In this paper a survey is performed on big data analytics using cloud based on Apache Hadoop and its importance in business. The objective of survey is comparison of power efficient techniques. Here quality of service techniques which provides low power consumption are studied. Also identified, technologies used for big data processing are mainly Hadoop and MapReduce. Hadoop MapReduce a powerful programming model is used for analyzing large set of data with parallelization, fault tolerance and other features are it is elastic, scalable, and efficient.

REFERENCES

[1] Rakesh Rathi, Sandhya Lohiya, “Big Data and Hadoop,” International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), vol. 2, pp. 214-217, April-June 2014.

[2] S. Ghemawat, J. Dean,“MapReduce: Simplified Data Processing on Large Clusters,” in the Communications of the ACM, vol. 51, pp. 107–113, January 2008.

[3] Energy star enterprise storage specification. United States Environmental Protection Agency, http://www.energystar.gov/index.cfm?c=new specs. enterprise storage, under development, April 2009.

[4] X. Fan, W. D. Weber, L. A. Barroso, “Power provisioning for a warehouse-sized computer,” In ISCA ’07: Proceedings of the 34th annual international symposium on Computer architecture, New York, USA, ACM, pages 13–23, 2007.

[5] Cloud Computing on Wikipedia, en.wikipedia.org/wiki/cloudcomputing.

[6]NIST Definition of Cloud Computing V15, csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-V15.doc.

[7] “Green Grid data center efficiency metrics: PUE and DCIE,” White Paper, The Green Grid, December 2008.

[8] M. Dhavapriya, N. Yasodha, “Big Data Analytics : Challenges and Solutions using Hadoop, MapReduce and Big Table,” International Journals of Computer Science Trends & Technology (IJCST), issue 1, vol. 4, pp. 5-14, Jan-Feb 2016.

[9] Chi-Sheng Su, Tseng-Chang Yen, “An SDN based Cloud Computing architecture and its Mathematical Model,” Information Science of Electronic & Electrical Engineering (ISEEE), pp. 1728-1731, 2014.

[10] Ashish A. Patokar, V.M. Patil, “Efficient Analysis of Big Data by using Hadoop in Cloud Computing by Map Reducing,” in National Conference on Innovative Trends in Science and Engineering (NC-ITSE), issue-7, vol. 4, pp. 378-381, 2016.

[11] https://www.oreilly.com/ideas/quality-of-service-for-hadoop-its-about-time.

[12] A. Desai, Nagegowda K S, “Advanced Control Distributed Processing Architecture (ACDPA) using SDN

and Hadoop for identifying the flow characteristics and setting the quality of service(QoS) in the network,"*2015 IEEE International Advance Computing Conference (IACC)*, Bangalore, 2015, pp. 784-788.

[13] A. Desai, Nagegowda K S, Ninikrishna T, "Secure and QoS aware architecture for cloud using software defined networks and Hadoop," 2015 International Conference on Computing and Network Communications (CoCoNet), Trivandrum, 2015, pp. 369-373.

[14] R. Osman, J. F. Pérez, G. Casale, "Quantifying the Impact of Replication on the Quality-of-Service in Cloud Databases," 2016 IEEE International Conference on Software Quality, Reliability and Security (QRS), Vienna, 2016, pp. 286-297.

[15] Z. Zhao, M. Peng, Z. Ding, W. Wang, H. V. Poor, "Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 5, pp. 1207-1221, May 2016.

[16] Hadoop. http://hadoop.apache.org

[17] Hadoop Power-By Page. http://wiki.apache.org/hadoop/PoweredBy.

[18] Fan Yang, Qing Liao, Jing ming Zhao, "An Improved parallel k-means clustering algorithm with Map Reduce," in 15th IEEE International Conference of Communication Technology (ICCT), pp. 764-768, 2013.

[19] http://www.opendaylight.org/

[20] Gridmix. HADOOP-HOME/src/benchmarks/gridmix in all recent Hadoop distributions.

[21] Hadoop Distributed File System, http://hadoop.apache.org/hdfs.