INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

ADAPTIVE HANDLING OF 3V'S OF BIG DATA TO IMPROVE EFFICIENCY USING HETEROGENEOUS CLUSTERS

Radhakrishnan R[1], Karthik S[2]

[1]M.E. CSE Krishnasamy College of Engineering and Technology, S.Kumarapuram, Cuddalore, Tamil Nadu – 607109, radhakrishnan4me@gmail.com
[2] Associate Professor & HOD (Department of CSE), Krishnasamy College of Engineering and Technology, S.Kumarapuram, Cuddalore Tamil Nadu - 607109, karthiks1087@gmail.com

Abstract: - Big data is the trending technology that caters to handle scalable data. Volume, Variety and Velocity are the 3V's of big data. Volume refers to the size of the data, variety refers to the types of data and velocity refers to the speed data transfer. The scheduling algorithm co-ordinates the tasks and executes it in the clusters. The existing scheduling algorithm does not efficiently use the heterogeneous cluster resources. The objective of this paper is to propose an adaptive scheduling algorithm to handle the 3V's efficiently. For this we propose the heterogeneous adaptable computing method that handles the data with the combination of CPU-GPU execution along with heterogeneous distributed file system. This type of adaptive scheduling is estimated to be efficient compared to the existing scheduling in the Hadoop as it explores the possibility of utilizing the resources available in the heterogeneous cluster, further it also makes it easier to add heterogeneous hardware to have scalability.

Keywords*:* Big Data, Hadoop, Scheduling algorithm

1. Introduction

Big data – a large volume of data is being processed in many areas like e-commerce, health care, e-governance, education, scientific research, weather monitoring, etc. In recent years big data has become an active and interesting research area. Most of the corporates and enterprises are adopting to the changing technological advances and have started to use big data.The large volume of data is managed using distributed systems, clusters and cloud. Big data is often characterized by volume, velocity and variety known as 3V's of big data. Volume is the amount of data, with the different forms other than text like images, videos, and audio which obviously leads to exponential growth of terabytes to zettabytes of data. Velocity is the speed of data movement. This is an important factor for the live services. Variety is the multiple formats which has to be processed ranging from various office formats to multimedia and other custom application formats.

Big data is gaining a lot of attention as it has a lot of scope to work on it and the applications of big data are the need of the hour with the fast pace of internet penetration. Big data is using in wide areas from scientific application to user data analytics. It can handle massive amount of data and are scalable to the expanding requirements. The handling of data can be classified as the handling of the 3V's of big data. The volume, variety and velocity. Hadoop is one of the frameworks which are used to implement the big data. It can handle a large amount of data and has its own scheduling algorithm. It is good and is designed for the homogeneous clusters. But it is not adaptable and is inefficient to handle the large amount of data using the heterogeneous clusters.

To handle the volume, variety and velocity of the big data in an efficient way we propose the adaptive handling algorithm which uses the CPU-GPU combination along with the heterogeneous file systems [1] which will increase the efficiency by utilizing the hardware in an appropriate way. Normally the CPU computing is done and the GPU computing is the recent trend that is being exploited to make the executions of the parallel processing much faster. GPU has a many cores which can carry many parallel tasks and execute it in less time. But not all the processes are suitable to be efficiently executed using GPU. So the combination of CPU-GPU will yield very good results. The problem here is to allocate the suitable task to the suitable computing methodology. This problem is addressed in this paper. A lot of work has been done previously by many researchers in the GPU computing. We use the right task allocation using proper classification of the task that can be scheduled to the right hardware.

2. Related Works:

The computing using GPUs to clouds is done by first addressing the performance requirements with the use of multi-layer parallelism , second by addressing the elasticity by online provisioning and allocation of cloud-based resources , third by addressing the predictability using performance envelope and fourth by characterizing the interaction between the execution engine architecture with other layers[2]. The hybrid GPU/CPU execution is efficient to perform massive parallel computations that are commonly used in the cryptanalysis and cryptography [3]. Mars framework is an implementation on the Hadoop platform that helps to utilize the GPU cores. This also helps in integrating the Phoenix to perform co-processing between the GPU and the CPU [4].

The big data volume handling using the heterogeneous distributed file systems is a three step process where the data nodes of different file types are formed first then the file size is analysed and then the storage of the data is made based on the suitable file system using the analysed result.[1].

The advances in the scheduling process of big data is made through many scheduling algorithms. A simple task scheduling algorithm uses the weighted round-robin method which improved the efficiency to a certain extend [7]. The bandwidth aware scheduling process addressed the task allocation using the software defined network which can provide data locality in an optimized way[8]. The adaptive task scheduling algorithm adjusts the workload in the dynamic environment in the heterogeneous clusters where the task trackers can adapt. ATSDWA obtains tasks with respect to the computing ability and are self-regulative [9].

3. Proposed Work

The objective of this paper is to handle the 3V's of big data in an efficient way. For this I propose an adaptive scheduling algorithm AH3V. First indexing of the volume, velocity and variety of streaming data is made. Priority based on the pattern of 3V's are made using the indexed data. Based on this pattern and priority, the streaming data is administered which improves the efficiency for vast amount of scalable streaming data. This is also a secure way of scheduling as it does not log and depend on the client details. The implementation of the experimental setup is made using the Hadoop and YARN based framework. Further the future possible enhancements are outlined.

4. Architecture

The Hadoop architecture has the job tracker and task tracker which is used for scheduling. The job tracker manages the jobs and decides to accept or reject the job that is incoming to the server. The task tracker manages the tasks by proper management and communication between the master node and the slave nodes. Task tracker identifies the right slave node to be used for the task to be processed. We modify the architecture by introducing the data handler, monitoring, task coordinator, AH3V server and AH3V client. The mars framework [4] is used to handle the processes that are to be executed using the GPU.

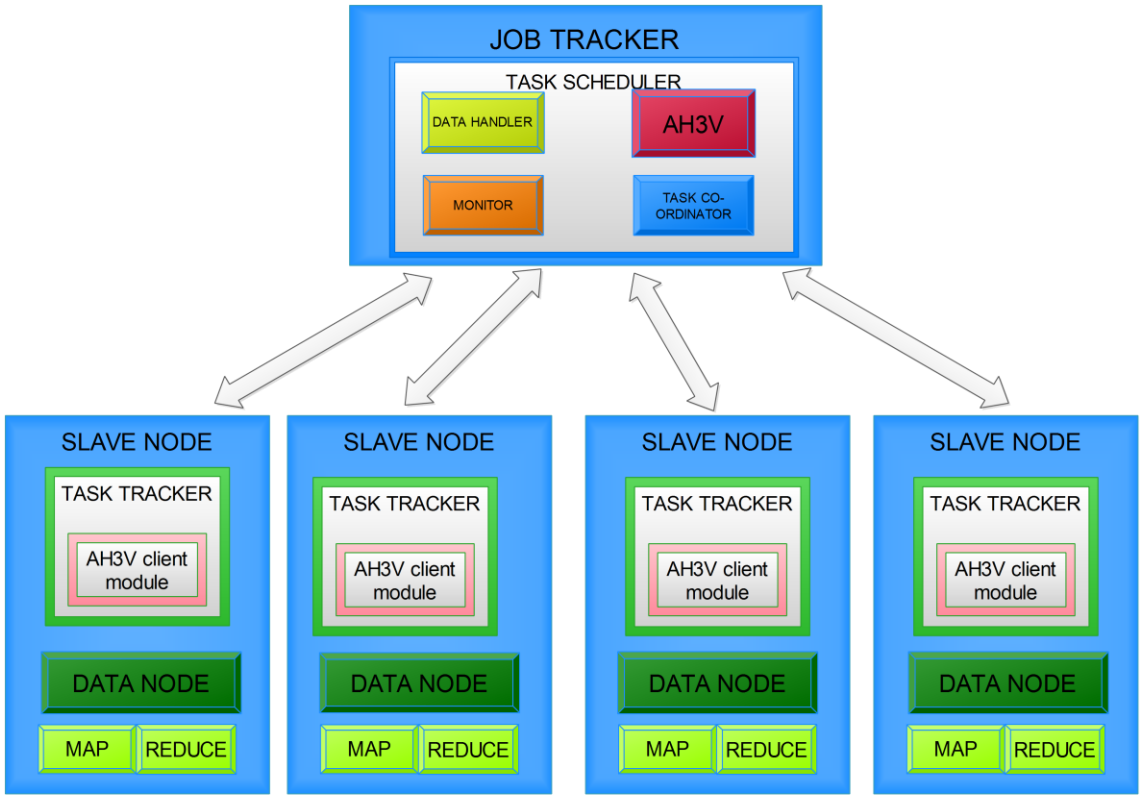

Figure 1: Architecture Diagram

5. Data Flow

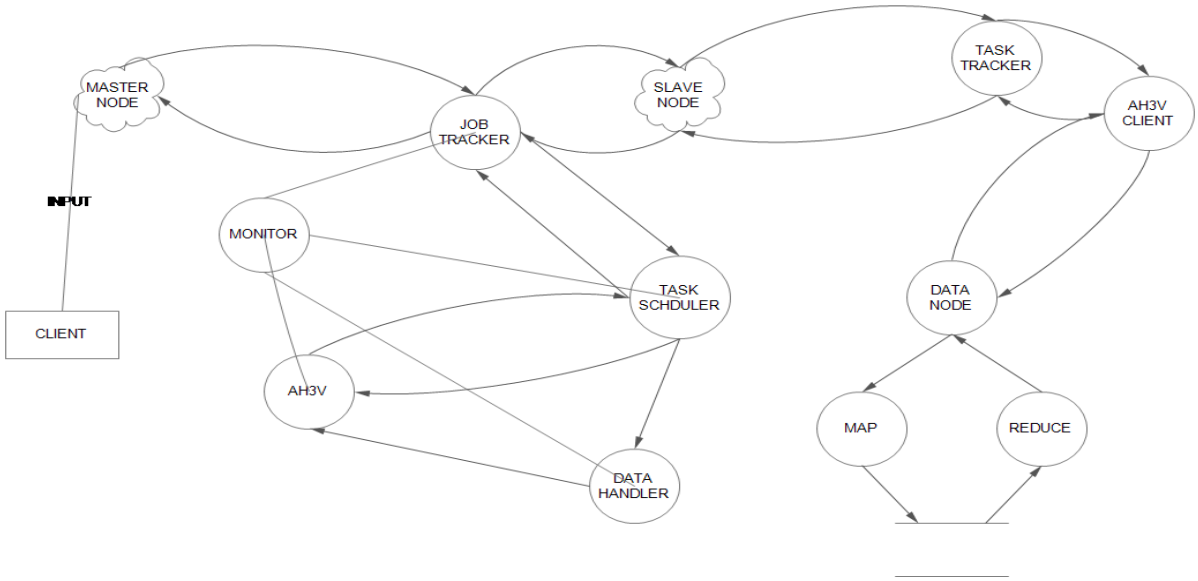

Figure 2: Data Flow Diagram

The data flow starts form the incoming of data from the client. This is received by the master node and sends it to the job tracker. Job tracker with the help of data handler and task scheduler executes the AH3V server module. The job tracker communicates with the slave node where the AH3V client module in the task tracker receives the task to be done and executes it in the data node. After which the map and the reduce processes takes place to complete the process executions.

6. Modules

6.1. DFS Integrator

This is the starting phase where the distributed file system is integrated. This is a little bit of complex work and the tools of the Hadoop framework are used in to make the integration of the different file system. The process involved in this module can be summarized as below

- Distributed File System Integrator
- Configuration of Hadoop framework
- Making of DFS file format
- Integration of Hadoop framework with DFS

6.2. Data Handler

The data handler handles the data that the system receives from various sources and does the configuration works and the process involved in this module can be summarized as below

- Formation of Different data nodes
- Data node configuration
- Name node configuration

6.3. AH3V Server

In the AH3V Server module the volume that is received from the different sources are organized and the algorithm core part is worked on in this module. Incoming data is classified based on file size and frequency of access as below

- Small file size with high frequent access
- Small file size with less frequency access
- Small file size with unknown frequency of access
- Large file size with high frequency access
- Large file size with less frequency access
- Large file size unknown frequency of access

The classified file size are then allocated the right node with the distributed file system based on the following comparison

Table 1: Distributed file system comparisons

	HDFS		**Ceph**		**GlusterFS**		**Lustre**	
Input/Output	**I**	**O**	**I**	**O**	**I**	**O**	**I**	**O**
1 X 20GB	407s	401s	419s	382s	341s	403s	374s	415s
1000 X 1MB	72s	17s	76s	21s	59s	18s	66s	5s

For the variety handling the classification of the following is done

- Modeling and rendering color correction and grain management composting
- Finishing and effects editing encoding and digital distribution
- On-air graphics on-set
- Simulation
- Other normal processing and usual sequential execution

After this classification the normal and sequential execution processes are sent to the CPU based execution cluster. The processes which could be massively parallelized are sent to the GPU based execution cluster.

6.4. AH3V Client

The AH3V client resides in the task tracker of the data nodes. It receives the tasks to be executed. It uses the right scheduling algorithms based on the type of cluster it has. The CPU cluster utilizes the usual sequential algorithm and the GPU cluster utilizes the mars framework to execute the task it has received. It also sends the status of the execution to the monitoring and the co-ordinating module to keep the processes updated.

6.5. Task co-ordinator

The task co-ordinator acts as the intermediate between all the processes and makes a record of all the processes that are done. It makes the communication between different modules. It ensures that the same task are not assigned to the different nodes.

6.6. Monitoring

Monitoring module monitors the health of different nodes and gives an alert if any node has technical issues. It has the classification algorithm and verifies the allocation done by the task co-ordinator. It also records the status of all the task in different nodes by logging the jobs done by different nodes which is later used by the AH3V server module to mine the past data, identify the suitable cluster for the jobs and adapts to the future job scheduling in the heterogeneous cluster environment.

7. Conclusion and Future Work

We described the ways and means of achieving the efficiency of the scheduling algorithm for the 3V's of big data using the Hadoop framework. The proposed approach is efficient than the existing system which does not adapt during the run time for the large amount of data. The use of the proposed algorithm make the system usable for the different environments where the unexpected amount of data , unexpected types of data and the unexpected streams of sources comes from random user base. Future work is to improve the cost efficiency where the cost of implementation in the large data centers are not considered here. This will also extends the efficiency improvement of the other V's of big data like value, virtue and velocity.

References

1. Radhakrishnan R , Karthik S. "Efficient Handling of Big Data Volume Using Heterogeneous Distributed File Systems". International Journal of Computer Trends and Technology (IJCTT) V15 (4):151-154, Sep 2014. ISSN:2231-2803. www.ijcttjournal.org. Published by Seventh Sense Research Group.
2. Varbanescu, Ana Lucia, and Alexandru Iosup. "On Many-Task Big Data Processing: from GPUs to Clouds." *MTAGS Workshop, held in conjunction with ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC)}*. ACM}.
3. Niewiadomska-Szynkiewicz, Ewa, et al. "A hybrid CPU/GPU cluster for encryption and decryption of large amounts of data." *Journal of Telecommunications and Information Technology* (2012): 32-39.
4. He, Bingsheng, et al. "Mars: a MapReduce framework on graphics processors."*Proceedings of the 17th international conference on Parallel architectures and compilation techniques*. ACM, 2008.
5. Ciznicki, Milosz, Krzysztof Kurowski, and Jan Węglarz. "Evaluation of selected resource allocation and scheduling methods in heterogeneous many-core processors and graphics processing units." *Foundations of Computing and Decision Sciences* 39.4 (2014): 233-248.
6. Wang, Zhenzhao, et al. "SepStore: Data Storage Accelerator for Distributed File Systems by Separating Small Files from Large Files." *Internet of Vehicles–Technologies and Services*. Springer International Publishing, 2014. 272-281.
7. Wang, Dan, Jilan Chen, and Wenbing Zhao. "A Task Scheduling Algorithm for Hadoop Platform." *Journal of Computers* 8.4 (2013): 929-936.
8. Qin, Peng, et al. "Bandwidth-Aware Scheduling with SDN in Hadoop: A New Trend for Big Data." *arXiv preprint arXiv:1403.2800* (2014).
9. Xu, Xiaolong, Lingling Cao, and Xinheng Wang. "Adaptive Task Scheduling Strategy Based on Dynamic Workload Adjustment for Heterogeneous Hadoop Clusters."

Author Biography

Radhakrishnan R has received his B.E. (CSE) degree in THE YEAR 2012. At present he is pursuing M.E. (CSE) in Krishnasamy College of Engineering and Technology, Cuddalore, Tamil Nadu, India. He has published one international journal article. His research interests lies in the areas of BIG DATA, Data Mining, Cloud Computing and Distributed Computing.

Karthik S completed his B.E. (CSE) degree in the year 2005, M. Tech (CSE) degree in the year 2007, MBA (HRM) in the year 2008, M. Phil (CSE) degree in the year 2009. Currently he is pursuing Ph.D. in the area of BIG DATA. Currently he is working as a HOD/ Associate professor in Computer Science and Engineering at Krishnasamy College of Engineering & Technology, Cuddalore, Tamil Nadu, India. His research interests lies in the areas of BIG DATA, DBMS, Data Mining, Data warehousing, Cryptography & Network Security, and Cloud Computing. He has published 3 International Journals and 4 research papers in National/ International conferences. Also he is life member of Indian Society of Technical Education of India (ISTE). He attended many workshops & National seminars in various technologies and also attended Faculty development Programme.