International Journal of Engineering & Technology, 7 (4.15) (2018) 16-18

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Performance Evaluation of Hadoop in Cloud for Big Data

Mohammed Fakherldin[1], Ibrahim Aaker Targio Hashem[2]*, Abdullah Alzuabi[3], Faiz Alotaibi[4]

[1]Faculty of Computer Science and Information Systems, Jazan University, Saudi Arabia
[2]School of Computing and Information Technology, Faculty of Built Environment, Engineering, Technology and Design, Taylor's University Lakeside Campus, Subang Jaya, Selangor, Malaysia
[3]Arabian Gulf University, Road 2904, Building 293, Manama 329, Bahrain
[4]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
**Corresponding author E-mail: targioabaker.targiohashem@taylors.edu.my*

Abstract

Recent trends in big data have shown that the amount of data continues to increase at an exponential rate. This trend has inspired many researchers over the past few years to explore new research direction of studies related to multiple areas in big data. Hadoop is one of the most popular platforms for big data, thus, Hadoop MapReduce is used to store data in Hadoop distributed file systems. While, cloud computing is considered an excellent candidate for storing and processing the big data. However, processing big data across multiple nodes is a challenging task. The problem is even more complex using virtualized clusters in a cloud computing to execute a large number of tasks. This paper provides a review and analysis of the impact of using physical versus cloud cluster in the processing a large amount of data. This analysis has an impact on the processing in terms of execution time and cost of using each one of them. The result indicates that the use of cloud virtual machines helped better utilize the resources of the host computer.

Keywords: *Cloud computing; Hadoop; MapReduce.*

1. Introduction

Recent trends in big data have shown that the amount of data continues to increase at an exponential rate. This trend has inspired many researchers over the past few years to explore new research direction of studies related to multiple areas in big data. Cloud computing is considered an excellent candidate for storing and processing the big data. Cloud computing "is a model for allowing ubiquitous, convenient, and on-demand network access to many configured computing resources (e.g., networks, server, storage, application, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [1]. Moreover, cloud computing has numerous promising features to address the fast growing of economies and technological obstacles. It offers an overall cost of ownership and permits companies to focus on the core business without worrying about issues such as infrastructure, flexibility, and availability of resources [2].

While existing research on Hadoop has shown great improvement, numerous performance challenging problems yet to be solved. Hadoop composed of one master node and several data nodes dependents on the size of the cluster. Users can make their choices according to the availability of the resources and the nodes.

A multidimensional resource may be considered using a series of resources such as CPU, memory, network bandwidth. For example, a slot management scheme is proposed by [3] in order to enable dynamic slot configuration in Hadoop. The idea behind slot management scheme is to improve resource utilization and reduce the completion time of multiple jobs.

This paper aims to provide analysis on the performance of Hadoop applications using physical and cloud cluster. This analysis has an impact on the processing in terms of execution time and cost of using each one of them. Normally, during the execution of the big data using Hadoop cluster, the processing time of the job on a machine is assumed to be fixed in advance. However, in reality, it demands resources to complete the job and the execution time is determined internally by many resources allocated.

The rest of the paper is structured as follows: Section 2 provides an overview of Hadoop cluster and the life cycle of the big data process. Section 3 presents the performance analysis. Finally, a conclusion is provided in Section 4.

2. Overview of HADOOP Cluster

Hadoop has offered a new alternative way to efficiently mining petabytes of unstructured information across multi-machines with lower cost commodity hardware. As an important technology, Hadoop is increasingly becoming an attention to the necessity of big data processing in recent years. Hadoop is developed using Java program and is an open source which offers an effective way to make use of the current infrastructure and cost for the distributed processing of huge amount of data using a large cluster of commodity hardware [4-5].

The mastermind behind the development of Hadoop is Doug Cutting after his son's toy elephant [6] based on Google File System GFS [7] and Google's MapReduce distributed computing environment. Hadoop comprises of two components, which are distributed file system called Hadoop Distributed File System (HDFS) and the computational layer that implements a processing paradigm called MapReduce.

Hadoop distributed file system (HDFS): HDFS is developed to provide data storage for cluster commodity hardware [8]. The concept is based on Master-slave architecture where one node act as the master and multiple nodes act as slaves storing blocks of data. The namenode manages a hierarchy of file system and directors namespace (i.e., Metadata). File systems are presented in a form of NameNode which register attributes such as access times, modification, permission and disk space quotas. The file content is split into large blocks and each block of the file is independently replicated across DataNodes for redundancy as well as periodically sends a report of all existing blocks to the NameNode.

MapReduce Data processing engine: MapReduce [9] is a simplified programming model for processing large amount of datasets pioneered by Google for data-intensive applications. MapReduce model built on top of the Google File System (GFS) [7] and adopted by open source Hadoop implementation, whose development popularized by Yahoo.

YARN: Hadoop Yarn is a framework, which provides a management solution for big data in distributed environments [10]. The main idea is to separate the resource management and job scheduling from the data processing allows Hadoop to supports various big data computing paradigms such as interactive analysis and stream processing. Moreover, Yarn offers Hadoop framework a giant flexibility in terms of job completions, which offer an effective management and monitoring of the workloads.

Table 1 illustrates the advantage and disadvantage of the Hadoop platform in terms of physical and cloud cluster deployment.

Table 1: Advantages and disadvantages of the Hadoop platform

Hadoop Platform	Advantage	Disadvantage
Physical cluster	Easy to manage Make use of low-cost community computers	Slow in terms of processing Costly Require special place
Cloud cluster	Hadoop clusters can be set up on demand. Physical infrastructure can be reused. You only pay for the CPU time you need. The cluster size can be expanded or contracted on demand.	Difficult to manage Sometime not reliable

More recently, there has been some effort to the proposed Hybrid environment in MapReduce framework. For instance, Sharma, Wood and Das introduced hierarchical scheduler for hybrid data centers with two-phase named HybridMR [11]. The algorithm comprises of multiple virtual machines and physical to make use of both paradigms. Firstly, the information acquired from HybridMR profiles to estimate virtualization overheads based on incoming MapReduce jobs to gauge can automatically guide placement between physical machines and virtual machines. Secondly, HybridMR builds run-time resource prediction models and performs dynamic resource orchestration to minimize the interference within and across collocated interactive and MapReduce applications.

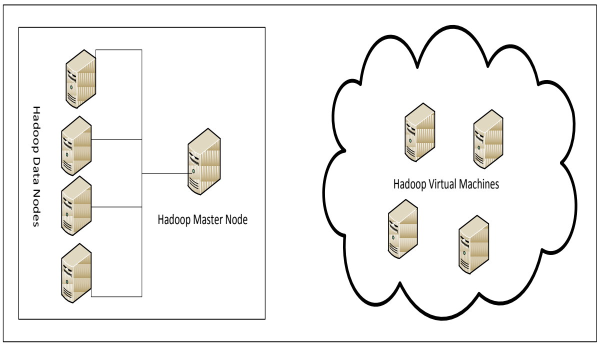

Fig. 1: Physical vs. Cloud cluster.

Figure 1 shows the difference between physical and cloud cluster. Both of them consist of multiple machines act as the data nodes and master node. The physical cluster is connected using local area network (LAN) each machine may have its own specification in terms of RAM, storage, and processor. However, cloud computing is based on virtual machines that are the computer architectures and provide the functionality of a physical computer. The performance of each virtual machine is dynamic, meaning that the resources of the machine can change on demand.

Big Data processing differs from traditional data processing primarily due to the volume, velocity and variety characteristics of the data being processed. The first step in the big data processing lifecycle (see Figure 2) is the collection of data known as data sources. This data sources can be video, audio, images, text files and database data. Big data can come from different sources such as social media, bank transactions, cloud, IoT, industries, etc. All these collected data has to be re-processed before it can be sent for analysis. There are many tools used for pre-processing data such as hive and pig. Then the result is kept in NoSQL databases for later used. Generally, Hadoop is one of the most popular platforms for big data, thus, the Hadoop MapReduce is used to store data in Hadoop distributed file systems. For processing the stored data, the MapReduce programming model is used based on two main functions namely, map and reduce. Besides MapReduce, there are many other tools which can be used on top of MapReduce for analysis such as Mahout, spark, and storm. The final output can be used for decision making by the organization. Figure 2 shows the big data processing lifecycle.

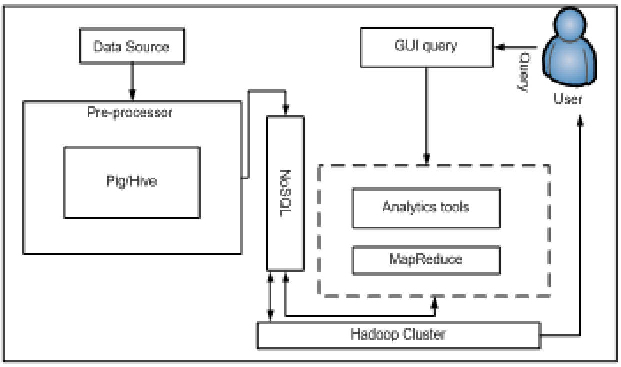

Fig. 2: Big data process lifecycle.

3. Performance Analysis

This section provides analysis of the impact of using physical versus cloud cluster at the processing a large amount of data. The reason to conduct such analysis is to identify the importance of cluster usage in terms of the cost of execution time and the utilization of the resources to complete the tasks. A physical cluster is a group of computers connected by a local area network (LAN), where Hadoop distribution is installed directly on the physical machines that are bounded by disk I/O. In contrast, cloud computing is a type of computing that relies on sharing heterogeneous computing resources to deliver computing services over the Internet in a convenient and scalable manner [12].

In the cloud, Hadoop distribution is installed on the virtual machines, where multiple "machines" does not require full physical resources at all times because the underlying infrastructure is shared [6]. Moreover, Hadoop has specifically designed for storing and processing unstructured data in a homogeneously distributed computing environment, which run on commodity hardware. The experiments were carried out in the cloud and physical cluster machines. We use five PCs as well as five Virtual Machine (VM) with the following configurations: 2.80 GHz processor, 2 GB main memory, and 1000 GB disk space. Hadoop cluster is used on Linux Ubuntu 14.04 where one machine runs a NameNode and ResourceManager, and the remaining are running DataNode and DataManager. Moreover, we use PingER data sets of different sizes varying from 500MB to 2 GB.

Table 2 provides a result of the comparison between the physical cluster and cloud cluster in terms of execution time.

Table 2: Comparison between physical and cloud cluster in terms of execution time

Data size (GB)	Physical Cluster (s)	Cloud (s)
1	46.28	42.04
2	53.44	51.67
3	69.30	50.57
4	97.57	60.39
5	124.92	90.23

As shown in Figure 3 and 4, the experimental result illustrates an increasing number of data size of the cluster, significantly increase the time necessary to run the application on both physical cluster and cloud. It shows for both physical cluster and cloud require less time to finish the jobs. Hence, running jobs on a cluster with a large number of data is the main motivation for using Hadoop to process big data. Given the relatively low commodity hardware of physical cluster, the results were fairly significant. Using five nodes of machines to run the jobs practically increase the runtime, compared with one node. Thus, increasing the number of machines could lead to increase in runtime as the cluster size increased.

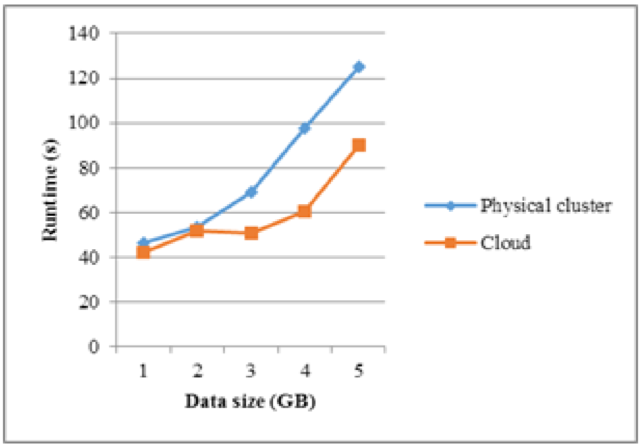

Fig. 3: Comparison between physical cluster and cloud in terms of runtime.

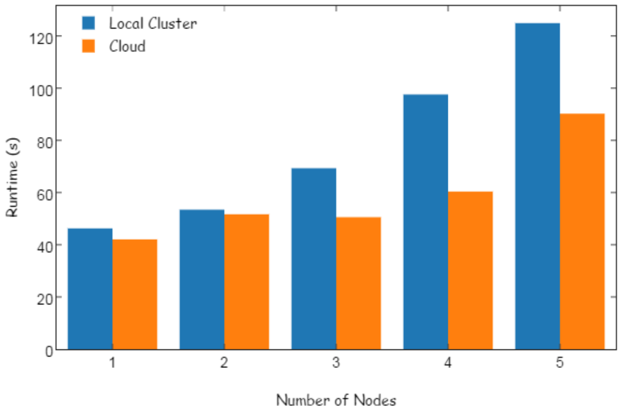

Fig. 4: Comparison between physical cluster and cloud in terms of runtime.

Nevertheless, the results indicate the unavailability of free resources. Running five cloud nodes put a considerable load on the host computer running the virtualization software and pushed the CPU utilization to 100%. This indicates that the use of cloud virtual machines helped better utilize the resources of the host computer. However, these machines require an optimal scheduling algorithm in order to reduce the overall execution time. The monetary cost to complete the entire workflow based on cloud platforms (e.g., Amazon EC2) is also an important metric as the resources are claimed on demand and will be charged as long as it is used. The monetary cost is closely related to the completion time. However, they do not always correlate due to the pricing scheme in the cloud.

4. Conclusion

Hadoop has offered a new alternative way to efficiently mining petabytes of unstructured information across multi-machines with lower cost commodity hardware. This paper aims to provide analysis on the performance of Hadoop applications using physical and cloud cluster. This analysis has an impact on the processing in terms of execution time and cost of using each one of them. The results indicate the unavailability of free resources. Running five cloud nodes put a considerable load on the host computer running the virtualization software and pushed the CPU utilization to 100%. This indicates that the use of cloud virtual machines helped better utilize the resources of the host computer. However, the continuous growth in the size of the data-centers containing Hadoop MapReduce clusters of hundreds and thousands of machines to support many users has let in a tremendous increase in the energy consumed to operate these large-scale data centers. Consequently, energy efficiency becoming a key open issue in the development of different techniques and approaches to optimize power management in Hadoop clusters [13]. Furthermore, in interactive data analysis, MapReduce workload runs in large clusters, whose size and cost make energy efficiency a critical concern on MapReduce, particularly in a cloud environment in which large equipped infrastructure is involved.

References

[1] Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf.
[2] Aceto, G., Botta, A., De Donato, W., & Pescapè, A. (2013). Cloud monitoring: A survey. Computer Networks, 57(9), 2093-2115.
[3] Yao, Y., Wang, J., Sheng, B., Tan, C., & Mi, N. (2015). Self-adjusting slot configurations for homogeneous and heterogeneous hadoop clusters. IEEE Transactions on Cloud Computing, 5(2), 344-357.
[4] Apache Hadoop Project Members. Apache Hadoop. https://hadoop.apache.org/.
[5] Zoll, Q., Zhu, Y., & Feng, D. (2010). A study of self-similarity in parallel I/O workloads. Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies, pp. 1-6.
[6] White, T. (2012). Hadoop: The definitive guide. O'Reilly Media Inc.
[7] Ghemawat, S., Gobioff, H., & Leung, S. T. (2003). The Google file system. Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, pp. 29-43.
[8] Borthakur, D. (2008). HDFS architecture guide. Hadoop Apache Project, 53, 1-13.
[9] Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
[10] Vavilapalli, V. K., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S., & Saha, B. (2013). Apache hadoop yarn: Yet another resource negotiator. Proceedings of the ACM 4th Annual Symposium on Cloud Computing, pp. 1-16.
[11] Sharma, B., Wood, T., & Das, C. R. (2013). Hybridmr: A hierarchical mapreduce scheduler for hybrid data centers. IEEE 33rd International Conference on Distributed Computing Systems, pp. 102-111.
[12] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58.
[13] Ibrahim, S., Jin, H., Lu, L., He, B., Antoniu, G., & Wu, S. (2012). Maestro: Replica-aware map scheduling for mapreduce. Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 435-442.