NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 6, ISSUE 7, July-2019

DIAGNOSING VIRTUALIZED HADOOP PERFORMANCE FROM BENCHMARK RESULTS: AN EXPLORATORY STUDY

Sudhir Allam,
Sr. Data Scientist, Department of Information Technology, USA

ABSTRACT

The importance of virtualization technologies in Hadoop is explored in this article. It looked at Hadoop as a new and common platform for businesses to use to improve business performance based on broad data sets. Hadoop gains from virtualization technologies in a variety of ways, including increased resource availability and cluster stability. Customers are also requesting virtual services including CPU, RAM, disks (etc.) from service companies (e.g. Amazon) and paying "pay as you go."[1] These advantages, though, are meaningless to consumers if unreasonable output loss occurs when moving from a real to a virtual platform. According to existing research on virtualized Hadoop performance, inappropriate network and storage settings for open-source virtual implementation result in significant performance degradation. However, due to the complexities of hardware and applications, including virtualization setups and implementation scales, performance tuning remains an extremely difficult practice to implement [1]. To bridge the virtualized Hadoop implementation gap, this paper recommends a performance diagnostic approach that incorporates statistical research from several levels, as well as a heuristic performance diagnostic tool that tests the reliability and accuracy of virtualized Hadoop by tracking employee traces from common big data benchmarks. Users will easily detect the bottleneck using the insights given by this tool, validate the assessment using performance resources generated by the guest OS and hypervisor, and keep optimizing performance for virtualized Hadoop by running this tool several times. Virtualization systems, in general, are used by supervisors to maximize resource use while lowering operational costs. Virtualization systems are divided into two classes. The first is about heavy virtualization, which is focused on the principle of virtual machines (VM) [2]. Each virtual machine (VM) replicates hardware and runs its operating system (OS) that is entirely independent of the host OS. The next one is for light virtualization, which is focused on container management. Although maintaining isolation, the containers share the host OS kernel [2]. This paper looks at the efficiency of Hadoop software which utilizes virtualization technologies.

KEYWORD: Artificial Intelligence, Hadoop, Hadoop virtualized systems, virtual machines

INTRODUCTION

With new technologies expanding, businesses are increasingly collecting data. The proliferation of Hadoop technology, in particular, would intensify this data explosion. For Big Data applications, Apache Hadoop [3] has emerged as the most popular framework. Hadoop is a well-known platform for analyzing unstructured data quickly and cost-effectively. The Hadoop business, which was worth $1.5 billion in 2012, is expected to double to $50 billion by 2020 [3]. Companies will now install Hadoop clusters on physical servers, in private clouds, or the public cloud [4]. It is not yet clear what model of implementation would prevail during this time of development, but the protection and granular control provided by private clouds contribute to this model dominating for medium and large companies. Cloud suppliers have quickly embraced this opportunity as part of their offerings (IaaS, PaaS, and SaaS) [4]. Amazon, for example, was one of the first to deliver Hadoop-as-a-service through its Elastic MapReduce (EMR) [4] web service. The key benefits of these cloud platforms are fast automatic implementation and cost-effective Hadoop cluster management, which is achieved by a pay-per-use model. All these capabilities are made possible through virtualization technologies, a fundamental building block of both public and private cloud infrastructure [5]. The advantages of virtualization, on the other hand, come at the cost of increased output overhead. The problems of virtualized Hadoop clusters include not only storing massive data collections, but also data transfer throughout processing. According to other studies contrasting the output of a virtualized Hadoop cluster to a real one, virtualization overhead varies between and 10% based on the application nature. This paper explored Hadoop's virtualized technologies to understand it increases performance.

RESEARCH PROBLEM

The main problem that this explorative study aims at solving is to understand how virtualized technologies are improving the performance of Hadoop. We can now gather more data (and different types of data) than we ever have before. It may be our generation's most important intangible commodity. Traditional management frameworks, such as hierarchical relational databases, may be overwhelmed by the sheer size ("big data") and the need for scalable, low latency processing [6]. As a consequence, modern methods for storing and mining vast amounts of unstructured data have become accessible. Hadoop as-a-service is now possible thanks to these streamlined provisioning and management tools. Some systems enable administrators to distribute pre-configured models to customers, allowing them to tailor the environment to their requirements. More advanced cloud infrastructure solutions simplify Hadoop implementation and management, allowing businesses to have Hadoop clusters without requiring consumers to handle any configuration files [7]. The power to scale changing a physical cluster — taking or replacing physical nodes — calls for the whole machine to be restructured. Load balancing (making sure that each user node stores about the same volume of data) is a critical challenge when scaling and managing a cluster. Distributed resource configurations are used in certain hypervisors, such as vSphere Enterprise Edition, which can conduct automated load balancing [8]. Cluster managers simply have to be deployed on new nodes to scale a distributed structure. Once the cluster scheduler learns of the data packet, it can immediately consume the available resources and start preparing activities for it.

LITERATURE REVIEW

VIRTUALIZING HADOOP

As physical Hadoop clusters increased in complexity, developers began to wonder whether they could virtualize them. As Hadoop matured, development efforts switched to virtualization, much like other business (and Java-based) programs. One may simplify their Hadoop cluster into even fewer physical servers because every VM is given separate computational and storage services from the host systems[9]. Virtual machine licenses supported or enterprise-level applications have an upfront cost, although this can be compensated over time by the cluster's lower running costs. Virtualization is the process of transforming something into something Hadoop pioneered the technology needed to operate Hadoop in the cloud, paving the way for major vendors to sell Hadoop as a web service. Amazon Web Services was the first to start beta evaluating their Elastic Map Reduce system in 2009 [9]. While the emphasis of this analysis isn't on public cloud implementation, it's important to note that it can be valuable for ad hoc or resource sharing, particularly if the data is already in the cloud [9,10]. An organization might find that creating its cloud infrastructure is more cost-efficient for a secure, live cluster. Controlled businesses can often prefer the privacy and protection of a private hosting network. Project Serengeti, an open-sourced development and implementation tool for virtual private applications based on vSphere. Other products, such as OpenStack's Project Sahara on KVM (formerly known as Project Savanna), have also been published in the last two years [10]. Even though these applications operate on vendor-specific virtual machines, they accept the vast majority of Hadoop deployments. Additionally, they may handle managing frameworks (such as Hive and Pig) that are usually designed on top of a Hadoop cluster to meet analytical requirements [10].

BENCHMARKS HADOOP

Many benchmarks have been built using three of Cloudera's sample applications: Pi, TestDFSIO, and TeraSort. Table 1 lists the estimated amount of simultaneous maps and reduces cluster-wide assignments with different benchmarks [11]. The same for native for all virtual setups.

Table 1. Simultaneous Tasks in a virtualized system

BENCHMARK	MAP		REDUCE	
	HT on	HT off	HT on	HT off
Pi	168	84	1	1
TestDFSIO-write	140	140	1	1
TestDFSIO-read	140	140	1	1
TeraGen	280	280	0	0
TeraSort	280	280	70	70
TeraValidate	70	70	1	1

Pi

Pi is a primarily computational program using a Monte Carlo approach to approximate pi. It's almost "unbelievably complementary": all specific solutions are separate and the single reduced task collects very little data from the multiple nodes. There is a negligible amount of network traffic and storage I/O [12]. All published findings calculate 1.68 trillion observations 13]. They cover 168 total evaluated in comparison. On both homogeneous and heterogeneous instances, the task-based approach per Hadoop node is proportional to the number of CPUs or vCPUs. Both map tasks start simultaneously and end when the last map function is full. For good results, all map tasks must operate at the same pace and finish at around the same time. A sluggish map task may have a big impact on the work period.

TESTDFSIO

TestDFSIO is a storage performance test divided into two components: TestDFSIO-write writes a total of 1000020MB (roughly 1TB) to HDFS, and TestDFSIO-read passes it back in. Due to the replication factor, compose tests do twice as much I/O as reading tests and produce significant networking traffic. A total of 140 map tasks were considered to be near to ideal for the HT disabled event, while 70 map tasks were insufficient, and 210 performed close to 140 [13]. For the HT-enabled experiments, the same amount of tasks were used as additional CPU capabilities were not required to support this I/O-dominated benchmark. The number of map activities per work node was equivalent to the number of usable disks [13].

TERASORT

TeraSort filters a lot of 100-byte data. It does significant I/O computing, networking, and storage and is also known to serve actual Hadoop workflows. It is divided into three sections: generation, classification, and verification [14]. TeraGen generates the data in a similar way to TestDFSIO-write, with the exception that it requires a large amount of processing to generate the random data. The map tasks write directly to HDFS with no reduction process. TeraSort performs the real sorting and writes HDFS-sorted data in many partitioned directories [14]. The program itself overrides the specified replication factor, thus writing just one copy. The theory is that the user should still re-run the program if a data disk is missing, but input data requires backup since it can not be quickly retrieved. TeraValidate checks all the details to ensure it's in place. The map tasks do this differently with each file and the single reduced task checks which each file's final record arrives until the next file's first records. Two sizes often appearing in the published Hadoop tests have been shown to

generate the data: 10 milliards records (1TB) and 35 milliards documentation (3.5TB). TeraGen should use an equivalent of 280 processing elements, 280 concurrent (several thousand in all) applied to calculate, 70 reduced TeraSort workload, and 70 TeraValidate map deliverables.

ANALYSIS

PI

On both virtual instances, Pi is 4-10% quicker than the equivalent native cases. This is an unpredictable outcome as pure CPU programs usually display a 1-5% decrease in virtual output. A small amount of mapping processes not performing at the same pace (a Linux and ESXi scheduler function) is expressed in how long the task is reduced. Scheduler variations will account for 1-2 percent of output differences. The number of the period of all map assignments is also seen in another Hadoop statistics. In simulated instances, this is up to 9% more. In both instances, this ensures that each map role runs faster in a VM along with 100% CPU usage. Research is underway to explain this conduct further. HT enables the time spent for the native case to be reduced by 12% and for simulated cases by 15 to 18 percent [15].

TESTDFSIO

TestDFSIO is a large storage performance stress test. The findings indicate that the virtual is 7–10% slower than the respective native situations in three out of four 1-VM situations (start writing/read, HT disabled/activated). The virtual system is 4-13 percent quicker in the 4th 1-VM scenario and all multi-VM instances. It is anticipated that performance would be solely constrained by hardware (here the Storage Controllers) and that the difference between software systems would be much lower in such testing (sequential output, slightly loaded processor). For the 1-VM scenario, the average flow is lower not just because the top flow is lower but because the flow is much more acoustic in time [15]. At 20 seconds increments, the data displayed was obtained in esxtop. At finer intervals, the output simply decreases for a couple of seconds to zero. These falls are synced through all of the cluster's nodes, meaning that they are linked to the app and not to its site-based. The output with two VMs per host is far more reliable and much more reliable with multiple VMs. The original case has noise similar to the case of a single VM. Why further VMs boost the performance behavior is not known. One explanation is that it's impossible to schedule I/O to ten disks for a single Hadoop node and that it can do a great job for fewer disks. While the performance for four VMs is significantly better than two VMs over the long term of the evaluation, the former still ends 2 percent earlier. The heterogeneity of the case of 4-VM makes it hard for all tasks to complete approximately at once with the case of 2-VM [15]. The test would compose several original files about the number of disks it has in each node by changing the number of tasks per node. But replicas are uniformly spaced overall knots and this proportionality is destroyed. Heterogeneous Hadoop nodes can be of benefit, but the compromises are hard to consider. HT has no clear output impact, which is supposed to occur with a low CPU workload.

VIRTUALIZATION BENEFITS OF HADOOP

Benchmarking methods show that a computer cluster's efficiency is equivalent to a real cluster. Built-ins workflows lower the initial setup sophistication and implementation time • Quick reading of output and simple to use control capabilities are provided by streamlined control consoles. For easy scaling, nodes may be quickly inserted and deleted. Besides, private cloud implementation provides cost-effective setup and service, added benefit by improving management, growing hardware use, and offering configurational stability to increase the efficiency of a cluster [16].

THE PERFORMANCE OF THE COMPETITION

Although a hypervisor needs a certain amount of computing resources, initial performance-oriented questions regarding virtual Hadoop. To control the VM's host, the virtuality layer includes certain CPU, memory, and other re-sources7 though the effect is based on the hypervisor's characteristics used. But the efficiency of VMs has considerably improved over the last 5 to 10 years (especially for Java-based applications). Virtual Hadoop clusters compete for a physical device by utilizing best practices 8,9. Increasing the amount of VMs per host will also contribute to improved results. Several independent reviews show (up to 13 percent). The benchmarking tasks that are available on Hadoop (such as Linux VServer, OpenVZ, and LXC), such as

WordCount or TeraSuite may even provide almost natively performed.10 [17]. These findings usually overweigh efficiency issues with the various other advantages that private cloud implementation provides.

FAST DEPLOYMENT

Hadoop developers need to navigate a complex setup and initialization protocol to launch a cluster. Clusters can consist of tens to hundreds of nodes – each node has to be installed independently in a physical deployment. An administrator can accelerate initial imaging by clones VM nodes with a virtual machines cluster. It is easy to copy VMs to extend a cluster size and delete issue nodes from the backup picture and restore them. Some Hadoop of Ferings, such as BDE, will automatically mount and connect to the net. Containers provide benefits in place of VMs when it requires hours to supply plain metal and minutes to supply VMs, but only seconds to supply containers. Installation and activation may also be automatic, like BDE. Enhanced maintenance and tracking For the cluster to satisfy 24/7 accessibility requirements, a Hadoop cluster must be closely managed and a range of control resources exist to track it. Both of them are supported by the Hadoop distribution (e.g. Cloudera Manager and Pivotal's Command Centre), while others are free software (e.g. Apache Ambari) or industrial (e.g. To optimize the management tool and lifecycle, virtualization software customers now use hyperviewers management interfaces (for example vCenter and XenCenter) and implement virtualized Hadoops as a new supervised workload**.**

FUTURE IN THE UNITED STATES

VMware, the pioneer, has been able to show phenomenal benefits in US organizations, by the convergence of several physical servers on one virtual computer or VM. Virtualization of servers allowed businesses to improve their services and played a key role in the Hadoop industry. Once cloud virtualization was restricted to the creation and research of fields, it finally appeared as an integral environment on servers. Although a growing number of businesses began to move to cloud computing and services infrastructure (IaaS), the virtualization of servers was still important to organizations. While businesses dreamed of centralizing their Hadoop operations into cloud-based systems, they saw Virtualization as a modern business paradigm with an economic benefit and an alternative to ancient practices. Virtualization combined with Hadoop guarantees efficient use of resources and reduces costs by the ability to run several running programs and systems on the same host. Virtualization services provide a wide global distribution potential across different industries.

ECONOMIC BENEFITS TO THE UNITED STATES

About 60% of businesses in the U.S anticipate a faster rate of transition in the digital transformation between 2021 and 2025. The network feature and desktop virtualization fields have become the most widely implemented virtualization technologies, rendering them the main business categories of international virtualization technology. In the short to medium term, however, development is projected to be driven primarily by software-defined storage and network functionality virtualization industries. According to a major market research company, service provider SDN and NFV acquisitions would grow at a CAGR from around 45 percent between 2017 and 2020, resulting in revenue of around USD 22 billion. The technology virtualization market is projected to expand at a CAGR of 21.1 percent over the projected timeframe, from USD 1.58 billion in 2017 to USD 4.12 billion in 2022, with datasets as a key driver. The technology virtualization market in Asia-Pacific is projected to expand at the fastest rate, due to this rapid expansion of domestic companies, the dramatic increase in data production across industry verticals, and improved infrastructure creation. Although the appetite for industrial automation, data virtualization technologies, increased emphasis on minimizing technology spending, the need for business analytics, and real-time data access are projected to propel the United States to the top of the market in terms of market size.

CONCLUSION

This study provides insight into the significance of virtualized technology and its effects on Hadoop results. The findings suggest that virtualizing Hadoop can be beneficial even just for efficiency reasons. For a single VM per host, an average figure in time elapsed over the default setup is 4% across all benchmarks. This is a small price to compensate for all of the other benefits that virtualization has. On average, operating two or four small VMs on each two-socket processor outperforms the corresponding native configuration; in certain

instances, output was up to 14 percent faster. Much greater savings may be achieved by using an optimized CPU consumption measure, which is more applicable in certain settings. The findings discussed here are only a small part of a larger project centered on Hadoop virtualization. Alternative storage methods, ease of usage, bigger arrays, higher-level implementations, and a variety of other subjects will be discussed in future articles. The priorities of a well-established architecture are the same if Hadoop is implemented in a given system, a private or public cloud, or both. It will be impossible to prescribe an optimal design for a Hadoop cluster due to the vast range of empirical demands and resource requirements. Planning a private server cluster, on the other hand, would not necessitate a steep learning curve. Many businesses may make use of tools they already have (such as virtualization licenses, cluster administrators, DAS, or SAN/NAS storage). Furthermore, many of the best practices currently in operation in an IT department (such as eliminating VM contention and maximizing I/O bandwidth) can be applied to customizing a high-performance Hadoop cluster. A private cloud provides unique advantages to Hadoop managers and consumers, including comparable (or even better) efficiency. Developers may use built-in software to configure their application components without requiring IT, thanks to rapid implementation and built-in workflows that simplify initial setup sophistication. Monitoring and analyzing output is made simpler with management systems, and functionality including high efficiency and fault tolerance reduce downtime. In the coming years, there will be a greater need for low-latency data processing services, as business programs begin to migrate from physical systems to cloud storage infrastructures.

REFERENCES

1) N. Cao, Z. Wu, H. Liu and Q. Zhang, "Improving downloading performance in hadoop distributed file system", Journal of Computer Applications, vol. 30, no. 8, pp. 2060-2065, 2010.
2) P. Gupta, P. Kumar and G. Gopal, "Sentiment Analysis on Hadoop with Hadoop Streaming", International Journal of Computer Applications, vol. 121, no. 11, pp. 4-8, 2015.
3) G. Hua, C. Hung and C. Tang, "Hadoop-MCC: Efficient Multiple Compound Comparison Algorithm Using Hadoop", Combinatorial Chemistry & High Throughput Screening, vol. 21, no. 2, pp. 84-92, 2018.
4) M. Gomes Xavier, M. Veiga Neves, and C. Fonticielha de Rose. A performance comparison of container-based virtualization systems for MapReduce clusters. In Parallel, Distributed and Network-Based Processing (PDP), 2014 22nd Euromicro International Conference on, Feb 2014.
5) J. Issa, "Performance characterization and analysis for Hadoop K-means iteration", Journal of Cloud Computing, vol. 5, no. 1, 2016.
6) Y. Kwon, C. Kim, S. Maeng and J. Huh, "Virtualizing performance asymmetric multi-core systems", ACM SIGARCH Computer Architecture News, vol. 39, no. 3, pp. 45-56, 2011.
7) R. Montella, G. Giunta and G. Laccetti, "Virtualizing high-end GPGPUs on ARM clusters for the next generation of high-performance cloud computing", Cluster Computing, vol. 17, no. 1, pp. 139-152, 2014.
8) S. Park, S. Kim and Y. Ha, "Scalable visualization for DBpedia ontology analysis using Hadoop", Software: Practice and Experience, vol. 45, no. 8, pp. 1103-1114, 2015.
9) Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. Madden, and M. Stonebraker. A comparison of approaches to large-scale data analysis. SIGMOD '09, New York, NY, USA, 2009.
10) R. Peinl and F. Holzschuher. The docker ecosystem needs consolidation. In CLOSER 2015, May 2015.
11) S. Huang, J. Huang, J. Dai, T. Xie, and B. Huang. The HI bench benchmark suite: Characterization of the MapReduce-based data analysis. In Data Engineering Workshops (ICDEW), 2010, March 2010.
12) Jlassi, P. Martineau, and V. Tkindt. Offline scheduling of map and reduce tasks on hadoop systems. In CLOSER 2015, May 2015.
13) P. Paul and D. Veeraiah, "Multi-Layered Security Model for Hadoop Environment", International Journal of Handheld Computing Research, vol. 8, no. 4, pp. 58-71, 2017.
14) J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. Commun. ACM, 51(1), Jan. 2008.
15) J. PJacob and A. Basu, "Performance Analysis of Hadoop Map Reduce on Eucalyptus Private Cloud", International Journal of Computer Applications, vol. 79, no. 17, pp. 10-13, 2013.

16) Y. Samadi, M. Zbakh and C. Tadonki, "Performance comparison between Hadoop and Spark frameworks using HiBench benchmarks", Concurrency and Computation: Practice and Experience, vol. 30, no. 12, p. e4367, 2017.
17) P. Vasconcelos and G. de Araújo Freitas, "Evaluating Virtualization for Hadoop MapReduce on an OpenNebula Cloud", International Journal of Multimedia and Image Processing, vol. 4, no. 34, pp. 234-244, 2014.
18) S. Xie, "Investigation on Fast Response Performance of Dam Deformation Monitoring System with Wireless Sensor and Virtualizing Technique", International Journal of Multimedia and Ubiquitous Engineering, vol. 11, no. 7, pp. 193-204, 2016.