Virtual Hadoop: The Study and Implementation of Hadoop in Virtual Environment using CloudStack KVM

[1]Arun S Devadiga, [2]Shalini P.R, [3]Aditya Kumar Sinha

[1]PG Scholar, [2]Assistant Professor, [3]Principal Technical Officer
[1]Computer Science and Engineering,
[1] NMAM Institute of Technology, Nitte, Karnataka, India
[1] Centre for Development of Advanced Computing (CDAC), Pune
[1]arundevadiga1@gmail.com, [2]shalini.pr.2007@gmail.com, [3]saditya@gmail.com

***Abstract*—The paper focuses on using Hadoop tool in virtual environment using CloudStack KVM for solving big data related problems. Hadoop is an apache tool which is used to process a huge amount of data concurrently. Since, Hadoop is an open source application; it has been used throughout the industry. Using Hadoop in virtual environment provides a way for parallel computing, and helps in deployment and management of applications for distributed computing. MapReduce component of Hadoop is used here for large-scale parallel applications and via virtualization we can improve the existing computing resources, which is essential in cloud computing field. By deploying virtual machine management of Hadoop we can have effective management of resource for large number of node in terms of configuration, deployment and resource utilization. Currently, there are many open source solutions for building cloud environment. One among them is CloudStack, which is an open source cloud platform that allows building all kind of cloud environment including private, public and hybrid cloud. KVM virtual machine provides the virtual environment. Hence, this article explains the work involved in integrating the Hadoop, CloudStack and KVM. This integration will result in virtual Hadoop which will allow user to process huge amount of data concurrently in virtual environment, with efficient use of resources.**

***Index Terms*— KVM, Virtualization, Hadoop, MapReduce, Distributed Environment, CloudStack.**

I. INTRODUCTION

Big data [1] is a buzzword, which is used to describe the huge amount of either structured or unstructured data. The data that is so huge that it's difficult to process using the traditional software and the database techniques [2]. According to the National data authority, the digital data is doubled every two years. These massive amounts of data are due to the public records, online transactions, digital media, social networking, blogs, emails, and trading, scientific experiments and so on. Hence, this large amount of data presents a significant challenge on efficiently analyzing, storing, querying and utilizing the data on many computing industry. Many of them have a misconception about the problem which is caused by the big data; they think the problem is only due to its size. The difficulty in handling this huge amount of data is due to its three V attributes (i.e., Volume, variety and velocity) [3]. Hence, to capture and store these data a lot of work has been proposed [2] [4] to solve the disadvantage of traditional database systems.

Hadoop [5], which is an Apache project providing reliable and distributed framework to store and analyze the huge amount of data. To provide high storage and high performance in satisfying the applications processing demands, there has been a development of hundreds or thousands of commodity computers connected using a local area network. With parallel processing of data in all these computers also called as nodes form a data center of its own. MapReduce [6] were the programming model used by these large data center (i.e., Hadoop) to implement and process the large dataset. Hadoop also uses HDFS [5] for reliable, distributed storage of data with the fault tolerant advantages. If any of the node fails due to system or process failure during the MapReduce process, then system returns to another node and carry forwards with the processing Hadoop in a physical cluster gives a parallel, scalable, higher efficiency (i.e., 1TB data can be sorted in 62 seconds [5]).

Nowadays, there are lot of an open source solutions for providing various cloud environments that include public, private or hybrid clouds such as Eucalyptus [7], OpenNebula [8], and CloudStack [9]. CloudStack is an open source cloud platform that allows building all kind of cloud environment including private, public and hybrid cloud. So, Virtualization is instantiated in CloudStack to provide the virtual environment. Virtualization is a representation of the computer logically using the software [10]. In a physical computer, it had a single operating system running one or more applications. In virtualization computer, software which runs on a single physical computer will abstract the physical resources and will maps or shares it to different virtual computers. Hence several operating systems can be run inside the single physical computer. The effect on one virtual computer will not affect the other virtual computer. Main advantage of virtualization is the effective use of hardware. That is, by sharing resources to each virtual machine, there has been complete utilization of the hardware. Billions of dollars have been invested on the research on controlling heat dissipation in data center. The only way is to use the less number of servers, hence virtualization on server allows less physical hardware and less dissipation of heat. Nowadays, there has been a significant increase in the cost of the hardware, hence virtualization allows fewer physical hardware and hence reduced cost. Some of the parameters which adds up to the cost saving are easier maintenance and lesser electricity. Redeployment and backups are made easier in virtualization by using a snapshot mechanism. Hence there is a faster disaster recovery in virtual environment.

In this article, it explains the work involved in the integration between the CloudStack, Hadoop and KVM which results in virtual Hadoop. Hence, one can imagine services to the users to efficiently launch their huge amount of data to the system with less managing and deploying work. So has to produce faster data processing capacity, reliability, lesser power dissipation and complete usage of the hardware, the above mentioned advantages of Hadoop, Virtualization and CloudStack are integrated.

II. RECENT WORKS

There have been a lot of researches going on in the field of virtualization. Some of the recent works have been discussed here. Today, virtualization is getting more and more popular in a cloud environment, one best example is the Amazon elastic MapReduce [11] [12]. A. Iordache et al [12] in their paper proposes a cloud based MapReduce, which is offered by the Amazon web services called as elastic MapReduce (EMR). This EMR allows user to sign up to the Amazon web service and after getting sign in, user can submit their MapReduce jobs using EMR API which is developed by some programming models like python or java. These MapReduce jobs are then sent to the Hadoop cluster, which consist of three virtual machines (VM) [11]. That is, Unique master VM, which acts as HDFS and schedules MapReduce task over other VMs, Multicore VMs which produces storage for HDFS and computes all the MapReduce tasks. Finally a multitask VMs which don't store any data but executes MapReduce task. In the thesis paper [9], they present a novel method of designing and implementing Resilin. Like EMR, the Resilin acts as a mediator between the Infrastructure as a Service (IaaS) and the client, hence acts as an EMR API and performs distributed MapReduce computation. The virtualization combined with Hadoop is also bioinformatics application by A. Matsunaga et al [13]. In this work, they discusses about integrating machine virtualization, Hadoop and network virtualization to deploy BLAST [14]. The validation is carried out by deploying two virtual cluster based on Xen [15]. The evaluation of the BLAST application in physical and virtual machine is carried out. Finally, giving a brief insights about the result saying that BLAST in virtual environment is better than the blast in the physical machine. Since, virtualization is easy to install and can be economically used, hence it has been an emerging part of cloud computing. Y. Geng et al [16] builds a model for data allocation in a virtual environment. Since, the CPU core as been increased, the virtual instance that can be created is also increased. Henceforth, there will be increase in I/O interference in virtual cloud causing a serious problem in the efficiency of the system. In this strategy, the file blocks are stored across the machines and replicas in different machine. Also HDFS will be aware of the virtual machine location. Hence work load can be balanced and I/O interference can be reduced, since localities of the virtual machines are aware. Since the overwhelming popularity in the field of cloud computing made the researchers [17], to implement MapReduce in virtual environment to provide higher efficiency and the stability. Y. Yang et al [18] discusses about the impact of virtual machine on Hadoop. The paper also focuses on effect of different virtualization technologies like OpenVZ, KVM, and Xen [15] on MapReduce environment. Also in research paper by J. Li et al [19] discusses the performance impact of three hypervisors (i.e., Xen, KVM and the commercial hypervisor).

Hadoop for a computation and storage uses single server to thousands of machines. Even though, Hadoop in a physical cluster provides a good performance for processing huge amount of data. But, the huge amount of data launch to system with less managing and deploying work is necessary. And also efficient resource utilization and power saving is also necessary Hence there has been a lot of research and study [17-19] to improve the performance of the Hadoop system. In our paper, we develop a virtual Hadoop to improve the reliability, easier managing and also to improve the power saving metrics. The Hadoop clusters will be deployed in Virtual machines like Xen and KVM. Next, the research on the impact of these virtual machines on Hadoop clusters and also comparing the Hadoop in virtual and physical machine is carried out.

III. PROPOSED SYSTEM

The core framework of the virtual cloud is similar to the reference [20] [13]. The proposed model mainly focuses on deploying Hadoop in CloudStack KVM, which results in virtual Hadoop. Later, Various MapReduce programs and datasets are given as the input to the virtual Hadoop created. The performance of the Hadoop in virtual environment is compared with the Hadoop in Physical environment. Its shows that Hadoop in virtual environment produces better performance than the Hadoop in physical environment. Fig .1 shows the architecture of the Deploying Hadoop in Virtual Environment using CloudStack KVM.

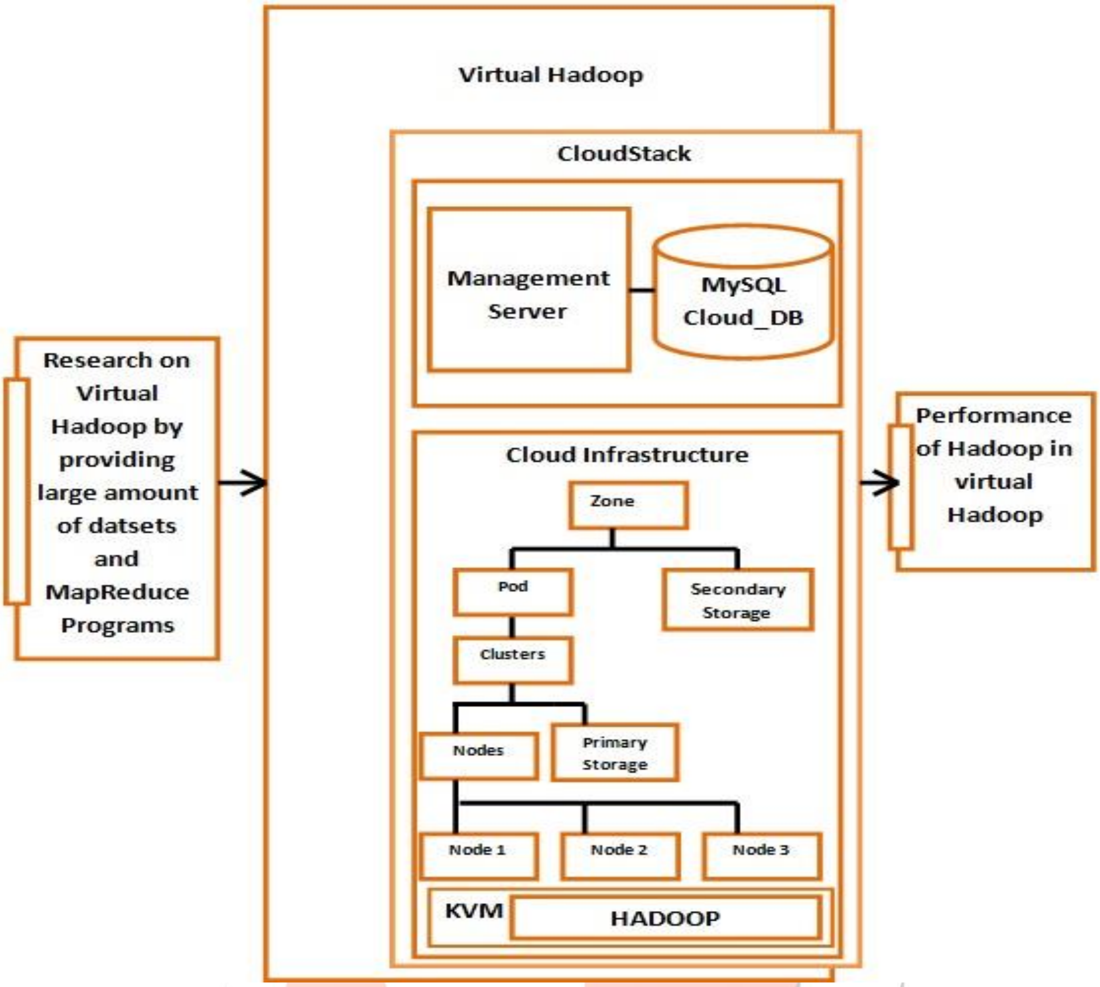

Fig 1. Architecture of the Deploying Hadoop in Virtual Environment Using CloudStack KVM

Deploying Hadoop in virtual machine using CloudStack KVM consists of following two steps:

Cloud and Virtual environment using CloudStack KVM

CloudStack is open source software used to provide cloud environment include public, private and hybrid cloud, which was developed by Citrix [21]. The CloudStack architecture is shown in the Fig. 2. CloudStack components are Management Server (MS), Availability Zone (AZ), Compute Nodes (CN), Clusters and POD. Zone is a collection of multiple pods which acts like a single data center. An Availability Zone can be defined as a single datacenter with many pods and secondary storage. Pod is similar to the rack of hardware with several clusters. Compute nodes are the hypervisor nodes where virtual machines are executed. The CloudStack supports Xen, KVM, VMware and oracle Virtual Machines. A cluster is a collection of hypervisor enabled host and a primary storage system. A Host is the Compute Node (CN) included in the cluster. Now, setting up the CloudStack in our proposed system allows a demand cloud infrastructure, where user can use the virtual service on pay policy.

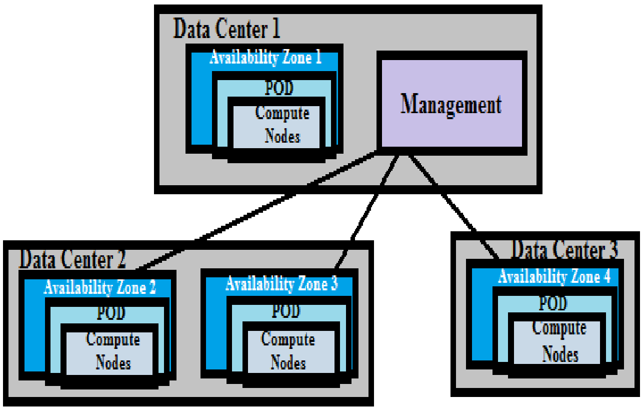

Fig 2. CloudStack architecture

The CloudStack KVM deployment consists of two major steps:

1) Installation of Management server: CloudStack is not only able to exist or occur without any conflict with the existing infrastructural resources, but it's easy to install, configure and manage by any users. The CloudStack uses management server to manage all the resources available and controls all the allocation of hypervisors or virtual machine to hosts. The management server provides an API or User Interface for users to manage cloud Infrastructure, assigning IP addresses and storages to the VMs. Before installing CloudStack, some of the system related configurations are done as shown in Table I including IP address, gateway, and DNS and Host name. To host the management server, OS must prepare which is as follows:

1. Set up the root password for the OS and login to your OS as the root.
2. Assign static IP addresses as per the Network connection
3. Set up SElinux permissive by default for access control or security policies and make sure machine can reach the Internet.
4. For time synchronization set the Network Time Protocol (NTP) and to point it to NTP servers edit NTP configuration file. Finally, restart the NTP client.

Table 1. Configuration Table

Host Name		Cloud Manager/Node
Hardware	CPU	Intel i5-2450M, 2.50GHZ
	RAM	6G DDR3
	Hard Disk	750G
Network	IP	10.11.32.22
	DNS	10.11.32.7
	Gateway	10.11.32.10
Software	OS	Ubuntu-10.04 LTS; RHEL 5.4-5.x 64-bit or 6.2+ 64-bit; CentOS 5.4-5.x 64-bit
	Software	CloudStack Management Server, MySQL
Storage	Primary Storage	NFS
	Secondary Storage	NFS

After setting up the system, management server installation can be performed as follows:

1. Download the CloudStack management server from the http://sourceforge.net/projects/CloudStack/files/CloudStack Acton/ and install all the CloudStack packages.
   # tar xzf CloudStack-VERSION-N-OSVERSION.tar.gz
   # cd CloudStack-VERSION-N-OSVERSION # ./install.sh
   Then choose "M" to install the Management Server software.
2. Install and configure MySQL database. Finally, restart the MySQL services and then invoke MySQL as the root user.
3. Setup the database
4. Create two directories for primary and secondary storage for using Management server as the NFS server and also create the agent.

2) Installation of Kernel based Virtual Machine (KVM): KVM [22] [23] is a full virtualization technology developed for linux on x86 hardware platform. The core of its virtualization is build in with loadable kernel module, kvm.ko and it is one of the virtualization machine monitor (VMM). KVM was developed by Qumranet in Israel. KVM also consist of hardware assisted virtualization that is Intel VT and AMD-V and with little paravirtualization is in progress that is in the form of device driver. The KVM virtualizes traditional linux kernel with guest mode, while this guest mode as its own kernel and user mode and executes all the guest OS codes [23]. Since KVM is full virtualization, this makes it simpler. KVM as somewhat different architecture then that of Xen that is it resides in the Host OS (i.e., Linux) and set of system calls (ioctl-s) is provided by the KVM to create a Virtual machine from the userspace [24]. Every virtual machines created by the KVM is treated as the ordinary process by the host OS. KVM is just a kernel level extension, not the complete tool. The actual tool in the userspace is QEMU emulator, just for the sake of simplicity it is called as KVM. The KVM installation consists of following steps:

1. Prepare the system Virtual Machine(VM) template
2. Install KVM in the host
3. Install CloudStack agent and NTP
4. NTP must be edited to confirm that all hosts in a pod have same time.

3) CloudStack Infrastructure configuration:

After the deployment and running of the CloudStack management server, CloudStack infrastructure must be configured. This can be done by entering the web console address of CloudStack in a browser (i.e., http://10.11.32.22:8080/client). User must log into using the username and password which was created during the installation. Configure CloudStack infrastructure by adding zones, pods, clusters, hosts, primary storage and the secondary storage. The fig.3 shows the console where it shows the number of zones, pods, hosts, clusters and memory added for deploying Hadoop in it. The left side of the panel shows the basic control column and the above shows the logged user status.

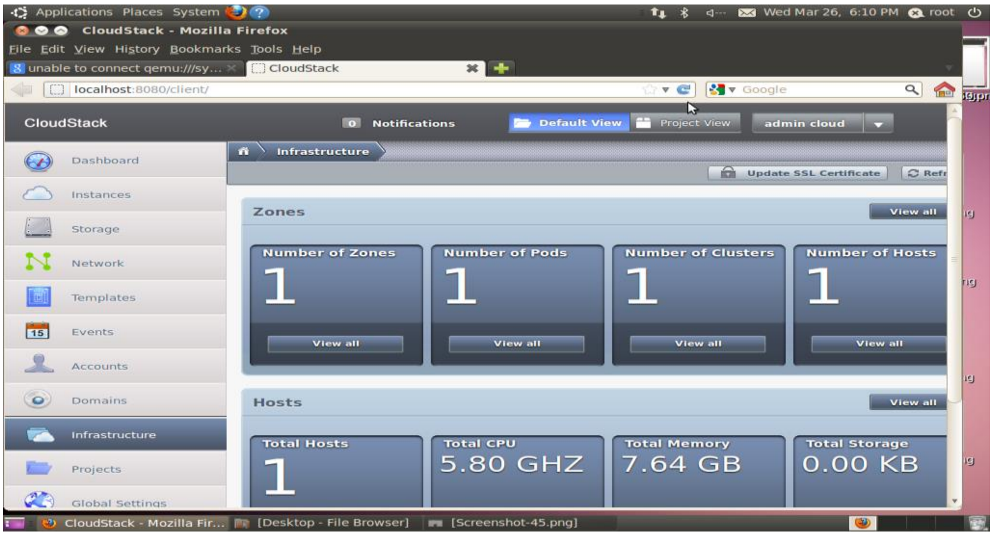

Fig 3. CloudStack web console

After configuring the CloudStack infrastructure, the virtual machine needs be created. User must use ISO or template file to create an OS instance in a virtual machine created inside CloudStack. Once the Virtual OS is installed, the Hadoop must be deployed in it to result in Virtual Hadoop.

Deployment of Hadoop in CloudStack KVM to produce virtual Hadoop

Hadoop [5] is a distributed, scalable tool which is used to solve the Big data Problem. Today, the world has been digitalized there has been a huge amount of data (Facebook itself generates 25TB of data daily [5]). If the data was just huge then it would not have been the problem. The problem is the three V attributes of the big data: Volume, Velocity and Variety. The data today is coming in a higher speed. For Example, CERN atomic experiments generate data at 40TB per second [5]. The big data also comes in different variety, that is data can be image, video format etc. Hence, this huge amount data cannot be processed or stored using traditional database (i.e. RDBMS) because the traditional database works on structured data, but the data today is structured or unstructured. Hence, to resolve this problem Apache foundation has developed a tool called HADOOP. Hadoop uses two of its ecosystem that is HDFS to store the data in a distributed and reliable fashion. MapReduce to process the huge amount of data parallelly and efficiently.

The HDFS [25] is a distributed, scalable, reliable file storage system, where it uses master/slave technology to store the huge amount of data. It uses Namenode (Master node), to store all the metadata, namespace of the data to be stored and many datanodes (slave node) to store the data with periodically reporting the status to the namenode. The user can read/write directly from the datanode with the permission of the namenode.

The MapReduce [6] [26] also uses a master/slave technology to process the huge amount of data parallelly. The MapReduce use Job tracker (Master) and Task tracker (Slave) to solve the problem of analyzing a data. The MapReduce uses Map and reduce functions to structure the unstructured data. Firstly, map function maps the unique key to the value and sorts the value. Later, reduce function removes the duplication and gives index to it (Fig .4).

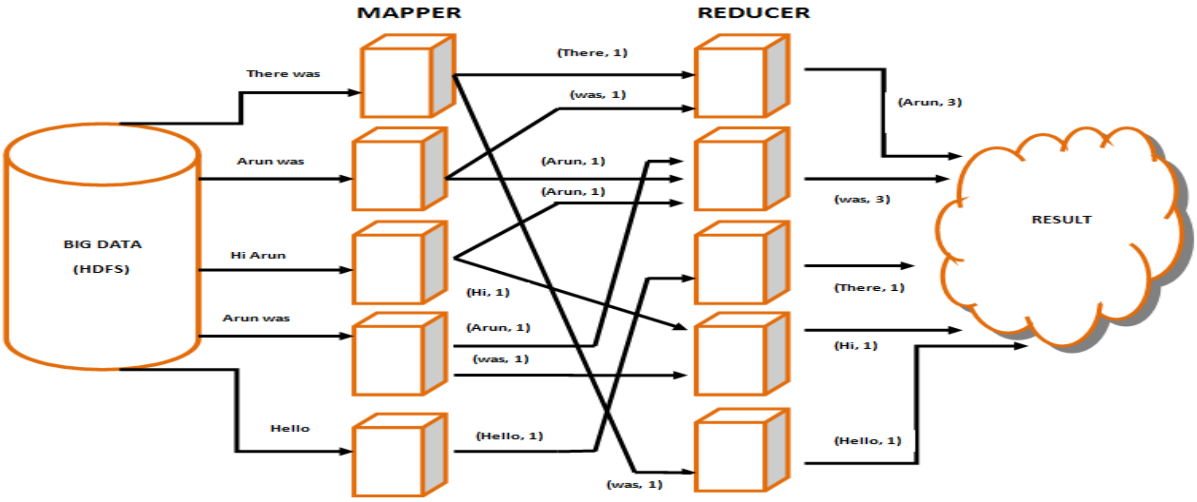

Fig 4. MapReduce Architecture

Hence to deploy Hadoop in CloudStack KVM following steps need to be followed:

1. Before installing Hadoop Java JDK needs to be installed, since Hadoop uses Java as the programming model. So the latest version of java can be downloaded from *http://www.oracle.com/* [27], *selected jdk-7u4-linux-i586.tar.gz.*
2. Untar the Java file and set all the environment variables.
3. Download Hadoop from the *http://archive.apache.org/* [28] and unpack and install Hadoop
   $ sudo apt-get install openssh-server
4. Configure Hadoop by editing the configuration file hdfs-site.xml, mapred-site.xml and core-site.xml. Also install Hadoop eclipse plug-in [29] and change the eclipse java perspective to MapReduce environment.

IV. VIRTUAL HADOOP

The basic MapReduce consist of one namenode and many datanode. Namenode is master of data nodes and responsible for managing the datanodes. Job tracker is the master of the task tracker and looks for task management. Now, the Hadoop is virtualized using any of the VMM like Xen, KVM or OpenVZ. There will be many instances of the virtual machines running in the single machine. Each instance may consist of its own datanode and the task tracker to carry out the processing of huge amount of data. Figure 5 shows the virtual Hadoop architecture. The instances in the single physical machine will be sharing the same CPU, Memory, I/O access and whatever physical resources provided by the physical machine.

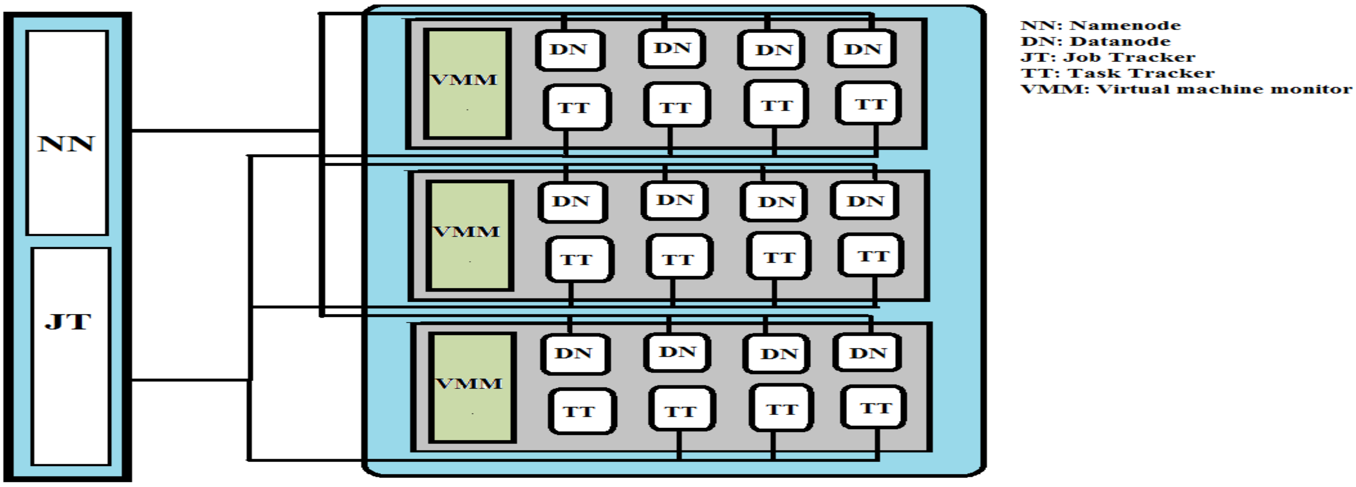

Fig 5. Hadoop Virtualization architecture

For example, consider two cases Hadoop in physical environment (i.e., many machines) and one more case that is virtual environment. In physical environment, some of the resources will be underutilized when running small map-reduce jobs and there will be many machines running, which increases cost for power usage. The physical environments also tend to have problem when managing these huge amount of data. In virtual environment, since single machine is utilized by the multiple instances, the resources will be fully utilized. The problem which occurs in virtual machine is it takes slightly longer response time for a small query. Because in virtual machine the datanode and task tracker need to wait for the resources, since many instances are trying to access the same resources at same time. Unlike, in physical machine there will be single instance in single machine, there is no problem for resources.

To solve these problem many algorithms are been developed like LATE [30], File block allocation algorithm [20] etc. Hence these algorithms will overcome the problem of low response rate in virtual environment. Still Hadoop in virtual environment has higher efficiency in managing resources, power saving and provides reliable Hadoop. These above mentioned parameters holds good when Hadoop is run on several physical machines. Because there is the problem of managing the physical machines and power overhead. The virtual environment solves this problem by making Hadoop instances to run in single physical machine, with full utilization of resources.

V. PERFORMANCE ANALYSIS

For the experimental purpose, the virtual Hadoop is deployed in the system with Ubuntu 10.04 LTS operating system. The implementation is performed in standalone computer of 6GB RAM, 2.50GHZ CPU. The Hadoop MapReduce programs are implemented using Java 7 JDK. NFS server acts as a CloudStack primary storage with Ubuntu 10.04 LTS X86. The performance analysis of virtual Hadoop based on the Execution time of the large data set scaling from 100 MB, 200MB, 300MB and 400MB from Wikipedia database [31]. Wikipedia provides the dump data from any Wikimedia foundation project [32]. Next process is to feed these data into the Hadoop cluster. These Microsoft excel files are uploaded into HDFS and MapReduce programs are used to process these dataset. The MapReduce java programs used here maps the rows one by one in key value format and reduce the dataset row by row.

The performance analysis of the virtual Hadoop with the variation in the dataset scaling increasingly from 100MB, 200MB, 300MB and 400MB. The following figure 6 shows the increase in the execution time with the increase in the dataset.

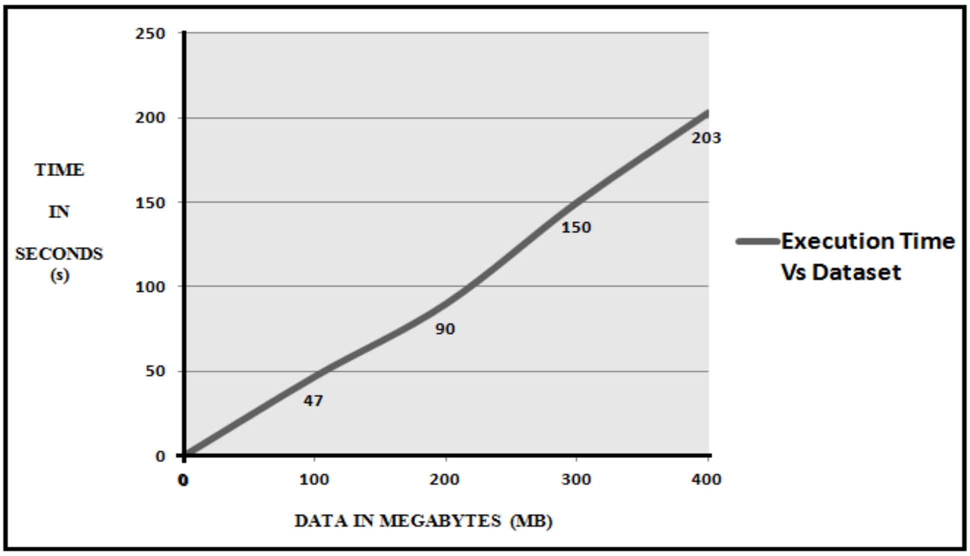

Fig 6. Execution Time Vs Dataset for Virtual Hadoop

Results above shown in fig 6 briefs the relationship between the dataset and the execution time when the dataset scalling from 100MB, 200MB, 300MB and 400MB fed to the virtual Hadoop. This result shows the behavior of the MapReduce application with increase in the size of the dataset in virtual Hadoop. Virtual Hadoop works linearly with the increase in the input size for MapReduce application, hence having major impact on execution size with the increase in the datasize.

The datasets scaling from 100 MB to 400 MB are fed to the Hadoop in physical clusters and the result shown in fig. 7 briefs that the behaviour of MapReduce application in physical cluster has slight lower execution time but almost equal than the MapReduce application in the virtual cluster. Eventhough, the execution time is almost similar between the virtual and physical Hadoop, the major advantage in virtual Hadoop occurs due to the CloudStack KVM virtualization. That is, the virtual Hadoop using CloudStack KVM allows complete utilization of the resources available. The physical Hadoop needs to maintain clusters of commodity computers but the virtualization alows CloudStack KVM based Virtual Hadoop to maintain the clusters of computers in a single or less number of computers. Since , the execution time of virtual Hadoop is sligtly more but almost similar than the physical Hadoop, the Virtual Hadoop using CloudStack KVM produces more advantages.

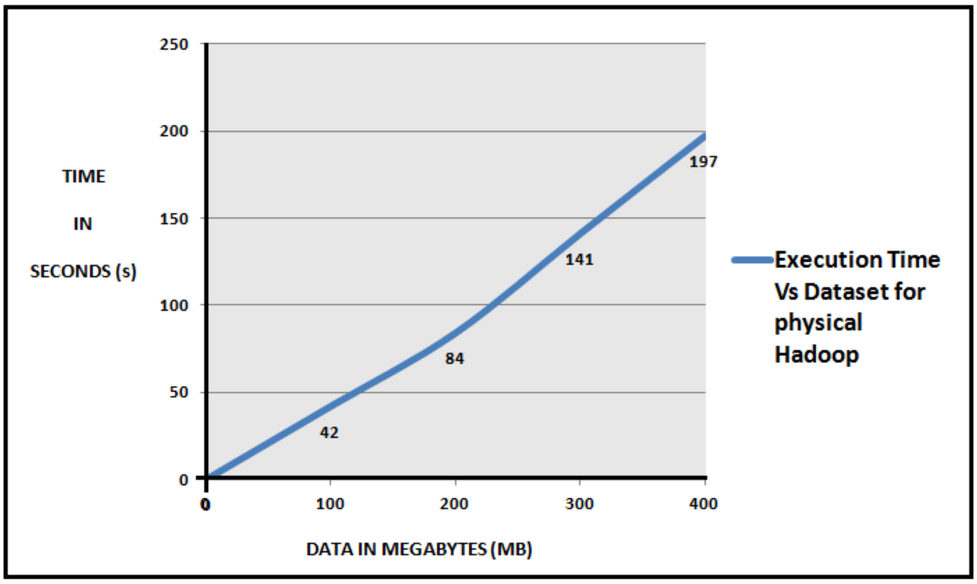

Fig 7. Execution Time Vs Dataset for Physical Hadoop

VI. Conclusions

This discusses about the deployment of virtual Hadoop using CloudStack KVM. Hadoop is an apache tool which is used to process a huge amount of data concurrently. Since, Hadoop is an open source application; it has been used throughout the industry. Using Hadoop in virtual environment provides a way for parallel computing, and helps in deployment and management of applications for distributed computing. MapReduce component of Hadoop is used here for large-scale parallel applications and via virtualization we can improve the existing computing resources, which is essential in cloud computing field. The deployment of Hadoop in virtual environment allows user to process the large amount of data without using the large amount physical commodity clusters. The paper discusses about the method to deploy and configuring CloudStack, KVM and Hadoop to produce the Virtual Hadoop.

The result shows that the virtual Hadoop has slightly higher but almost similar execution time to execute the MapReduce program than the Physical Hadoop. But the advantages are that the management is easier, fully utilizing the computing resources, make Hadoop more reliable and save power. Hence this advantage proves that virtual Hadoop using CloudStack as higher efficiency compared to the Physical Hadoop.

REFERENCES

[1] S. Sagiroglu, D. Sinanc, “Big Data: A Review”, *International Conference on Collaboration Technologies and Systems (CTS)*, 2013, pp.42-47

[2] Jefrey Dean and Sanjay Ghemawat, "MapReduce: simplified data processing on Large clusters. Commun. ACM", January 2008, pp.107-113

[3] A. Katal, M. Wazid, R H Goudar “Big Data: Issues, Challenges, Tools and Good Practices”, *Sixth International Conference on Contemporary Computing (IC3)*, 2013, pp. 404 – 409

[4] F. Chang, J. Dean, S.Ghemawat, W. C. Hsieh, Deborah A. Wallach, M. Burrows, T. Chandra, A. Fikes, R E. Gruber, “Bigtable: A distributed storage system for structured data”, in *proceedings of the 7th conference on usenix symposium on operating systems design and implementation - volume 7*, 2006, pages 205-218.

[5] Tom white, "Hadoop: The Definitive guide", Yahoo Press, 2010..

[6] Apache Hadoop, http://Hadoop.Apache.org .

[7] Eucalyptus. http://open.eucalyptus.com

[8] OpenNebula. http://opennebula.org

[9] CloudStack. http://www.cloud.com

[10] "Virtualization in education", IBM, October 2007. Retrieved 6 July 2010.

[11] Amazon Elastic MapReduce (Amazon EMR), http://aws.amazon.com/elasticMapReduce/.

[12] A. Iordache, C. Morin, N. Parlavantzas, E. FelleR, P. Riteau, ”Resilin: Elastic MapReduce over Multiple Clouds *Cluster”, Cloud and Grid Computing (CCGrid), 13th IEEE/ACM International Symposium on Digital Object Identifier: 10.1109/CCGrid.2013.48, 2013*

[13] A. Matsunaga, M. Tsugawa, J. Fortes, "CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications", *Fourth IEEE International Conference on eScience*, 2008

[14] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman., “Basic Local Alignment Search Tool”, *Journal of Molecular Biology, 1990, v. 215(3), pp.403-410,doi:10.1006/jmbi.1990.9999.*

[15] J. Fischbach, D. Hendricks, and J. Triplett, Xentop. Xen builtin Utility, 2005.

[16] Y. Geng, S. Chen, Y. Wu, R. Wu, G. Yang, W. Zheng, "Location-aware MapReduce in Virtual Cloud", *International Conference on Parallel Processing*, 2011

[17] C. Ning, W. Zhong-hai, L. Hong-zhi, and Z. Qi-xun, “Improving downloading performance in Hadoop distributed file system”, *Journal of computer applications*, vol. 30, 2010.

[18] Y. Yang, X. Long, X. Dou, C. Wen, "Impacts of Virtualization Technologies on Hadoop", In *Third International Conference on Intelligent System Design and Engineering Applications*, 2013

[19] J. Li, Q. Wang, D. Jayasinghe, J. Park, T. Zhu, C. Pu, "Performance Overhead Among Three Hypervisors: An Experimental Study using Hadoop Benchmarks", In *IEEE International Congress on Big Data*, 2013

[20] G. Xu, F. Xu, H. Ma "Deploying and Researching Hadoop in Virtual Machines", *Proceeding of the IEEE International Conference on Automation and Logistics Zhengzhou, China*, August 2012

[21] F. Gomez-Folgar, A. Garcia-Loureiro , T. F. Pena, R. Valin, "Performance of the CloudStack KVM Pod primary storage under NFS version 3", *10th IEEE International Symposium on Parallel and Distributed Processing with Applications*, 2011

[22] Kernel Based Virtual Machine (KVM), www.linux-kvm.org.

[23] J. Che, Y. Yu, C. Shi, W. Lin, "A Synthetical Performance Evaluation of OpenVZ, Xen and KVM", *IEEE Asia-Pacific Services Computing Conference*, 2010

[24] D. Petrovi´c and A. Schiper, "Implementing Virtual Machine Replication: A Case Study using Xen and KVM", *26th IEEE International Conference on Advanced Information Networking and Applications*, 2012.

[25] Hadoop distributed File system, www.Hadoop.Apache.org/hdfs

[26] MapReduce, www.Hadoop.Apache.org/MapReduce

[27] JAVA 1.6, http:// www.oracle.com/technetwork/java/javase/downloads/java-archive-downloads-javase6-419409.html.

[28] Hadoop 0.20.1, http://archive.apache.org/dist/Hadoop/core/Hadoop-0.20.1/.

[29] Hadoop Eclipse Plugin, http://code.google.com/p/Hadoop-eclipse-plugin/downloads/detail?name=Hadoop-0.20.1-eclipse-plugin.jar&can=2&q=.

[30] Matei Zaharia, Andrew Konwinski, Anthony D. Joseph, Randy H. Katz, Ion Stoica, “Improving MapReduce Performance in Heterogeneous Environments**”** *Technical Report No. UCB/EECS-2008-99*, *http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-99.html*, 2008.

[31] Wikipedia: Database Download, http://en.m.wikipedia.org/wiki/Wikipedia:Database_download/

[32] Wikipedia Dump data, http://dumps.wikimedia.org