Performing Big Data over Cloud on a Test-Bed

Vishal Dubey
Bachelor of Technology
Bharat Institute of Technology
Partapur by Pass Meerut

Saumya Gupta
Bachelor of Technology
Bharat Institute of Technology
Partapur by Pass Meerut

Sapeksh Garg
Bachelor of Technology
Bharat Institute of Technology
Partapur by Pass Meerut

ABSTRACT

Data analytics has been rapidly growing in a variety of application areas like mining business intellect for processing the huge amount of data. MapReduce programming paradigm adds itself well to these data-intensive analytics jobs, given its one of the well known ability to scale-out and force several machines to parallely process data. This paper introduces a detailed analysis of big data over cloud computing with several mapred techniques from system and application aspects. Here in this work we say that such Mapper and Reducer based analytics provide a better result over cloud platform. However, End-Users in this environment has the ability to use MapReduce applications to minimize the incurred cost, while obtaining the best performance. From the implementation point of view, we describe the key issues and challenges of big data on cloud and on local system as well. At last, the challenges come across in implementing MapReduce functions over Hadoop and the analysis of standalone, clustered and virtualized systems over our test-bed.

General Terms

Mapper, Reducer, Hadoop, Cloud, Virtualization

Keywords

Cloud computing, Big data, Hadoop, HDFS, Map Reduce, virtualization

1. INTRODUCTION

The increasing number of web serving millions of Internet users and dealing with more than petabytes of data on daily basis, the coming of cheap and resilience storage capacity resulting in a remarkable growth in data retention and High Availability (HA), and also high availability of economical resources to process that diverse data, have all taken the place of the need for large-scale data processing. Extracting facts from these large datasets, known as data analytics. Data analytics is shown to be useful in several scenarios—analytics enable web data mining (example, web search and indexing), enable extracting intelligence and in analyzing these large volumes of data demands a highly scalable solution. MapReduce is one well known approach that enables data analytics by parallely processing data. MapReduce can efficiently scale out to clusters of thousands of commodity nodes as the data and processing demands scale, and also can automatically handle failures in these huge and massive settings. In the last decades, the continuous increase of computational power has produced an awesome flow of data in a flexible and easy manner. Big data is the technology widely used for large and complex datasets for which traditional data processing applications are insufficient. **Big data** is not only becoming more available but also more understandable to computers and also the most buzz word **Cloud** is actually a powerful technology The most significant quality and feature of cloud is high availability, elasticity, flexibility and reduced cost. IaaS is a technology to implement faster applications since virtual machines often run faster than normal computer with the same hardware configuration. In our project, we had mainly focused on MapReduce functionality on host and over private cloud. Bare-Metal virtualization allows a single computer to be partitioned or divided into multiple virtual machines which may run its own OS simultaneously which provide true processing of the computer cores. Apache Hadoop [1] is a software framework that supports data-intensive distributed applications; it enables applications to work with many hundreds and thousands of nodes and petabytes of data efficiently.

2. BIG-DATA

Big data [2] is a well known term in today’s world which means data sets whose size is beyond the ability of commonly used software tools to hold and process the data within the limited time. The size of big data is continuously increasing and that data is moving on network from one location to other. At present data size varying from terabytes to petabytes of data in single diverse dataset and in future it is more likely to increase. Google, Facebook and other social site generate petabytes of data on a daily basis. Rapid Businesses are responsible for driving the growth of big data at exponential rate which is thrilling. Moreover resilient data storage, expert management and capturing values of huge size of data are enterprises very big challenges. We live in era where different and bulky data flow is a common term; it’s very tedious task to measure the total precise volume of data stored by machine on the given interval of time.

3. RELATED WORKS

Prashant Chauhan, Abdul Jhummarwala, Manoj Pandya in December, 2012 provided an overview of Hadoop. This type of computing can have both homogeneous or heterogeneous platform and hardware. The concept of cloud computing [3] and virtualization [4] has derived much momentum and has turned a more popular phrase IT sectors.

Harrison [8] Carranza and Aparicio Carranza on the topic Virtualization [9][10] in Linux a Key Component for Cloud Computing provided an overview of virtualization. For map reduce we came across Implementation of MapReduce Algorithm and Nutch Distributed File System in Nutch[11] by Dr. C. Chandrasekar and Kowsalya N. At there own pace all the organizations have started implementing these new technologies to further cut down costs through improved machine utilization, reduced administration time and also infrastructure costs at a very exponential rate. We also implement this technology to save time in various processing in our test-bed.

4. TERMS

4.1 Cloud Computing

Cloud computing is a style of computing in which dynamically scalable and often virtualize resources are provided as a service over the network. Users need not to have any prior knowledge of the technology or infrastructure in the "cloud" that supports them.

4.2. Hadoop-HDFS-MapReduce

Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computer using simple programming models which is designed to expand from single servers to clusters of machines, each offering the local compute and storage .Rather than rely on hardware to deliver high availability the library itself is designed to detect and handle failures at the application layer, so providing a high availability service on the top of clusters of computers each of which may be prone to failures. Apache Hadoop is an open source software project to enable data-intensive computing on large clusters .It includes a distributed file system (HDFS), programming feature for Mapper and Reducer, and infrastructure application for grid computing as well. We can Design framework for capturing workload statistics and replaying workload simulations to allow the assessment of framework improvements. We will deploy Hadoop cluster consist of number of several nodes and than these will be used to store data and process it in a parallel and distributed mechanism. Hadoop provides a reliable shared storage and compute system. The storage is provided by HDFS, and compute by MapReduce. Files are stored in a redundant fashion across multiple machines to ensure their durability to failure and high availability to very parallel applications. Hadoop has its own filesystem which replicates data to multiple nodes to ensure if one node holding data goes down, there are at least minimal of two other nodes from which that piece of information can be retrieved and used this is called higher availability of data. This makes Hadoop exceptional and unique, simplified programming model which allows the user to quickly write and test clustered systems, and its efficient, automatic division of data and distribution of work across machines and in turn utilizing the underlying parallelism of the CPU cores.

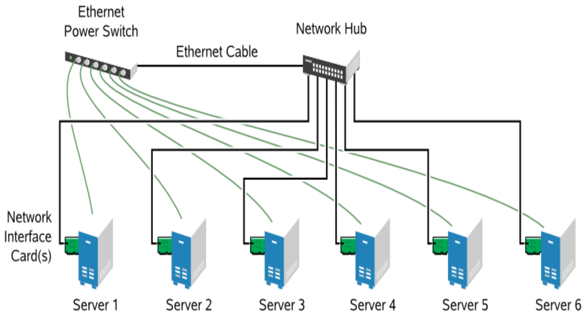

Fig. 1 MapReduce: Isolated Processes

MapReduce [5] is a programming model and an associated implementation for processing and generating diverse and gigantic and diverse datasets. Users specify a mapper function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reducer function that integrate all intermediate values associated with the same intermediate key. The MapReduce library groups together all intermediate values associated with the same intermediate key and passes them to the Reduce function. There are two types of nodes that control the job execution process: a jobtracker and a number of Tasktrackers. Management and synchronization of all the jobs run on the system by scheduling tasks to run on tasktrackers is done by Jobtracker. Tasktrackers start its child daemon and run tasks and send progress reports to the jobtracker, which also keeps a record of the overall progress made by tasktrackers of each job for each node. If any how the task fails, the tasktracker notify to jobtracker and than it can be reschedule on a different tasktracker and once the task is complete , the child daemon get terminated itself. Hadoop also divides the input to a MapReduce job into fixed-size pieces called chunks or input splits or splits. Tasktrackers is responsible for creating one map task for each split, which execute the user specific map function for each record in the split.

5. EXPERIMENT

5.1 Experimental setup

In this, we have a cluster of six computers in which one was namenode and jobtracker, one was client and other four were datanodes and tasktrackers and also five computers as Redhat Hypervisors [6] version 3.1, RHEV-Manager for the observation on Redhat Enterprise Limited version 6.4 with hard disk capacity with 500 GB and 4 GB of RAM.

5.2 Experimental result and analysis

In this section, we test the performance of parallel distributed computing by applying different map-reduce operations. We mainly consider the trend in CPU time spent with the increase in the volume of data, and we try to show the difference in CPU time between a single PC implementation and the parallel Hadoop implementation. We have considered two phases one with Hadoop technology and other without Hadoop. We have started from a basis example that was Word Count followed by Aggregate Word hist and terasort. We analyzed the trends of CPU time in different cases.

5.3 Observation on BIT Test-Bed

Observation 1: WORD COUNT

Table 1 . Host computer

System	123MB file(ms)	1 GB with 3 replicas.(ms)	1 GB with 1 replica(ms)
Standalone	47230	333400	330280
1 datanode	47470	331110	323000
2 datanode	46350	325220	321620
3 datanode	45770	323770	311270
4 datanode	43350	312520	309220

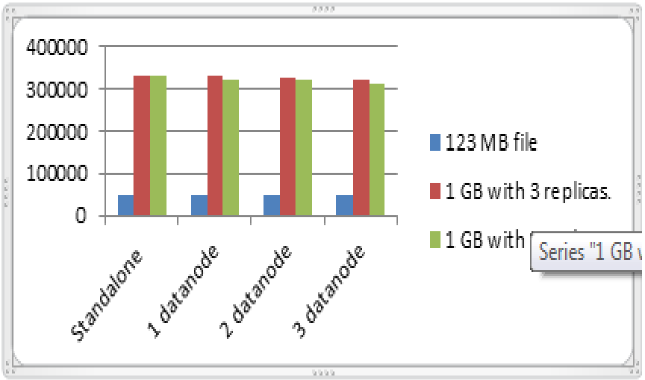

Fig. 2 Host computer

Table 2. Over our Cloud

System	123MB file (ms)	1GB file with 3 replicas(ms)	1GB with 1 replica(ms)
Standalone	44320	313210	310150
1 datanode	44070	313100	310010
2 datanode	43770	311050	309580
3 datanode	43250	309700	305010
4 datanode	42950	306870	300230

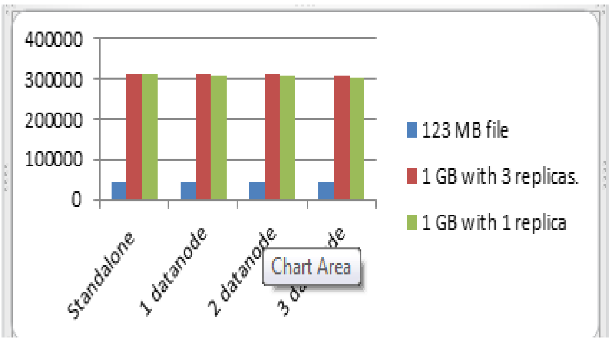

Fig. 3 Over Cloud

Observation 2: AGGREGATE WORD HIST

Table 3. Host computer

System	CPU time spent
Standalone	285660
1 datanode	285870
2 datanode	284970
3 datanode	283560
4 datanode	283150

Table 4 . Over our Cloud

System	CPU time spent(ms)
Standalone	264530
1 datanode	264420
2 datanode	264280
3 datanode	264170
4 datanode	264090

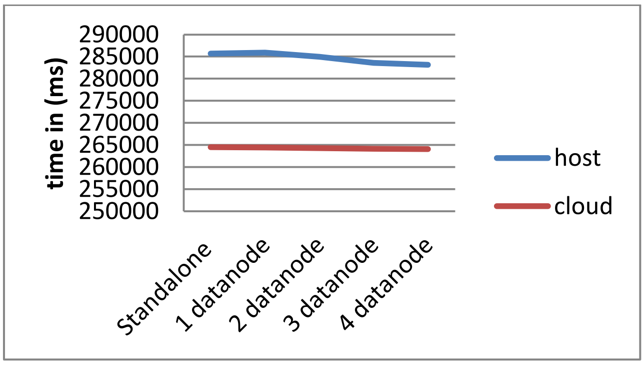

Fig.4 Host computer v/s Cloud

Observation 3: TERASORT

Table 5 . Host computer

System	CPU time spent
Standalone	5200
1 datanode	5250
2 datanode	5130
3 datanode	5090
4 datanode	4950

Table 6 . Host computer

System	CPU time spent
Standalone	4900
1 datanode	4850
2 datanode	4790
3 datanode	4680
4 datanode	4570

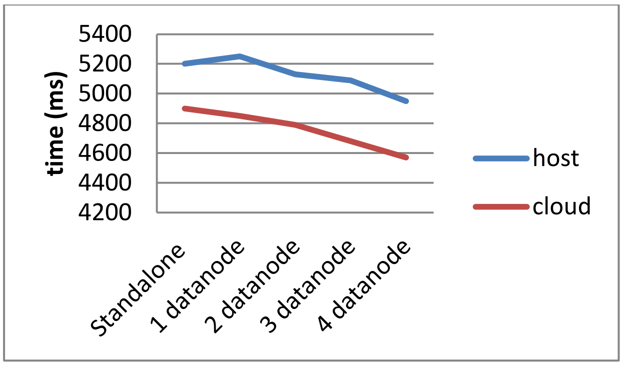

Fig. 5 Host computer v/s Cloud

5.4 Challenges Faced While Using Hadoop

- Challenge of virtualization: As the systems are old we face problem regarding the virtualization, as old system lack in hardware virtualization as the systems do not have VT-X device which is vital for hardware virtualization. Virtualization refers to the act of creating a virtual version of something including but not limited to a virtual computer hardware, Operating system and storage device or computer network resource. Hardware virtualization refers to the act of creating a virtual machine that acts like a real computer with an operating system.

- Challenge of Networking: In configuring IP address for all the nodes we have to provide the static IP address for all nodes. In resolving hostname, hostname is an alias that is assigned to an IP node to identify it as a TCP/IP host. Hostname can be up to 255 characters long. Hostname resolution means mapping a hostname to an IP address.

- Challenge of Firewall: A **firewall** is a network security system that controls the incoming and outgoing network traffic based on an applied rule set.

6. CONCLUSION

After performing several mapred operations on host and on our cloud i.e. on the Cloud Test-bed of Bharat Institute of Technology, we came to know that operations which were performed over the cloud gave better result than the host. Cloud reduces traffic on network, storage and in searching perception. A gigantic amount of data could be stored and heavy computations could be done on a standalone system with full safety and security at cheap cost but consumes large amount of time whereas the clustered system which use Hadoop is time efficient and effective and when we performed same operation over the cloud we can clearly observed that the time taken for the processing is comparatively low than the standalone and clustered computer as well as host computer thus we can conclude that the virtual system provide the best performance because of the direct interaction with the physical hardware. We have also observed that the pipelining provided the fast replication of data from one node to another which has decreased large amount of copying time and increases the efficiency of I/O processing. Pipelining [12] has drastically reduced the time and efficient processing is seen in case of clustered Hadoop environment in both the cases.

7. DISCUSSION AND FUTURE WORK

While performing the operation and several other tasks on out test-bed what we have noticed that till now we are not able to fully utilize the true power of Hadoop and moreover the Hadoop on Cloud, if fully optimized than it will be truly most demanded technology in terms of data analytics. This observation implies that a Hadoop job’s performance can potentially be improved, while incurring the same cost, since Hadoop framework is built to exploit the parallelism. This work proposes an approach to minimize cost, and maximize performance in a cloud setting. We however, say that if our resources were not virtualized than the result would be different which would be dependent on network traffics, physical hardware and with virtualized resources that had the potential for energy and time savings.

8. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable feedback on this work .This work was supported in part by Bharat Institute of Technology under the guidance of Mr. Kshitiz Saxena (Head of Computer Science and Engineering Department).

9. REFERENCES

[1] “What is Apache Hadoop,” http://hortonworks. com /hadoop/,2011-2014

[2] “Gartner IT Glossary,” http://www.gartner.com/it-glossary/big-data/,2013.

[3] “Fern Helper, Bringing big data to the enterprise ,” http://www.ibm.com/software/ in/data/ bigdata/, January 2012

[4] Cloudera. http://www.cloudera.com/.

[5] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In Proc. of OSDI, 2004.

[6] RedHat Enterprises virtualization workbook for student.

[7] Big Data Processing in Cloud Computing Environments College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

[8] Virtualization in Linux a Key Component for Cloud Computing

[9] “White Paper: Ten Things You Need to Know About Virtualization.” www.datacore.com

[10] B. Golden. Virtualization for Dummies, Wiley: Hoboken, New Jersey: 2008.

[11] Implementation of MapReduce Algorithm and Nutch Distributed File System in Nutch published by IJCA 2011

[12] Pipeline computing. From Wikipedia en.wikipedia.org/wiki/Pipeline_(computing