International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 4, April 2015

Performance analysis of Big Data using Cloud

[1] Er.Zainab Mirza, [2]Mohammed Abbas Khute,[3] Saifuddin Shaikh,[4] Atul Rai
[1]H.O.D, Dept of Information Technology, M.H. Saboo Siddik college of Engineering, Mumbai
[2][3][4] B.E Student, Dept of Information technology, M.H. Saboo Siddik College of Engineering, Mumbai
[1]mirza_zainab@yahoo.com, [2]abbaskhute4@gmail.com, [3]shaikhsaifuddin67@gmail.com, [4]atul93rai@gmail.com

Abstract: **In last few years there has been drastic increase in accumulation of data. The accumulation is so huge that it is beyond the reach of classical computing resources. In this paper we are going to do performance analysis and their comparison using classic computing resources i.e. pc and java and Evolved computing model i.e. cloud computing and Hadoop.**

Keywords: **AWS, Cloud Computing, Big data, Hadoop, Java**

I. INTRODUCTION

In last few decades the computing power evolved from standalone computer to a widely distributed high performance server on premises and at different operational location [1].

At the same time there is continuous and drastic reduction in price of storage devices [2]. And thirdly the emergence of World Wide Web. These three factors results in drastic increase in accumulation of data. The data are of two type user generated and machine generated [3]. The data comes from user account from different web servers such as social media, emails, etc. and from machine such as log files, error dump, GPS etc.

The data is so huge that it is beyond classical computing resources. There was a need for HPC[High Performance Computing] at both hardware and software level which is found now a day's with cloud computing and Big data processing tool such as Hadoop [4].

Using cloud computing and Hadoop we are going to process Datasets with variation in its size and compare its performance with classical computing OBJECTIVE

In this paper we are going to analyze the performance of Big data by doing word count operation. The performance of word count [5] is measured using two different computing model i.e. Pc with Java and cloud with Hadoop. We have chosen Amazon web services (AWS) as cloud provider.

II. Concepts

Cloud Computing

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly *provisioned and released with minimal management effort or service provider interaction. " [6]*

There are three service model of cloud they are as follows:

Software as a Service (SaaS)

The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure2. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings. [7]

Platform as a Service (PaaS)

The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider.3 The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment. [8]

Infrastructure as a Service (IaaS)

The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls). [9]

Big data

Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. [10]

a distributed environment. It's model is based on master and slave in which parallel processing is done among slave

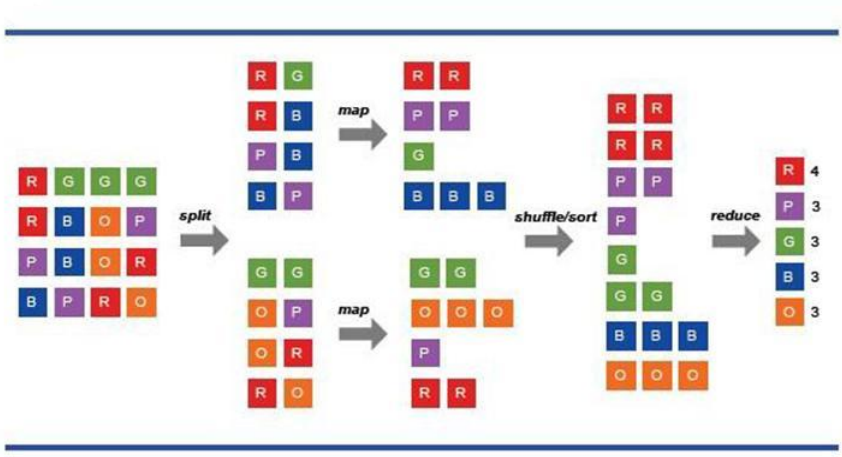


Fig 1 [5] MapReduce Example

Big data relation with cloud computing

Since the Big data is having high volume and velocity, the computing resource must be highly scalable. Cloud computing model provides on demand, highly scalable resource pool so it is the best option for processing big data. Cloud also provide pay as you go characteristic which made it a cost effective model for big data processing. [11]

III. TOOLS

Java

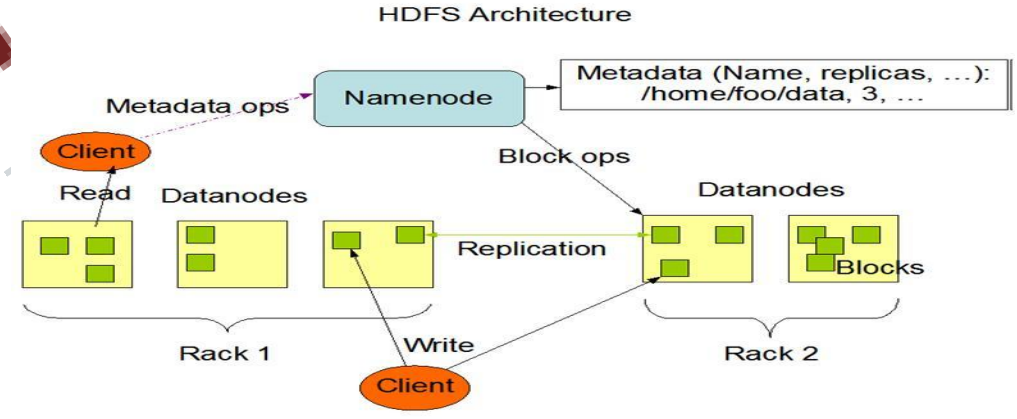
Java is one of the high levels programming language which came into existence in 1991. It is well known for its computing and portability (i.e. write once and run anywhere) feature. It has huge ecosystem with rich in variety of models and API .[12]

Hadoop

Hadoop is a open source tool owned by Apache foundation. It is specially architected and design to process Big data in Hadoop consist of HDFS and resource manager. HDFS stands for Hadoop Distributed File System. HDFS consist of name node, secondary name node, data node. The name node is responsible for file operations open, close, execute. It does not store data in HDFS but does mapping of files in HDFS. Secondary name node keeps image of name node which is used to recover if any failure occurs. Data node stores the data in HDFS(Refer Fig 2). [13]

Resource manager consist of job tracer, task tracker and map reduce. Job tracker keeps track and schedule map reduce jobs. The task tracker gets information from job tracker and it is responsible to spawn process in JVM. Map reduce is parallel programming approach which is used to process big data by analysis, mapping and reducing it (Refer Fig 1). [15]

Amazon EMR



Amazon EMR [Elastic MapReduce] is platform as a service (PaaS) which gives runtime environment for Hadoop programs. Amazon EMR is the combination of Amazon EC2 (Elastic Compute Cloud) and Hadoop distribution. Amazon EC2 is IaaS provided by Amazon .It gives compute power. [16]

Amazon S3

Amazon S3(Simple Storage Service) is a cloud storage Which can be accessed through internet. It used for storing and retrieving of data. For performing word count operation the word count application and dataset is stored in Amazon S3. [17]

IV. PERFORMANCE ANALYSIS

We have done performance analysis in three steps they are as follows:

Evaluation of Cluster size

First the performance of the data on different cluster size is measured. The cluster size on which best performance is found is used for measuring performance of different size of data set. The performance is measured in cluster size 2, 4, 8, 16.The best performance for data size 9.6 GB is obtained with cluster size 16.hence it is used. (Refer Fig 3)

Performance analysis using java

The performance is evaluated using Java and Pc. The following are the specifications.

Java: jdk1.7.0u67

Pc: Processor: Intel(R) Core™ i3-3110M CPU @ 2.40GHz.

RAM: 2Gb.

System type: 64-bit Operating System, x64-based processor.

Operating System: Ubuntu 14.04

Hard drive: 20 GB.

Files with different size are given for word count starting from 0.8 GB text file and gradually increased by 0.8 GB up to 9.6 GB. The graph is plotted with x-axis representing file size and y-axis time (in sec). (Refer Fig 4)

Performance analysis using Amazon EMR

Amazon EMR consists of Amazon EC2 and Hadoop distribution the specifications are as follows:

Hadoop: Hadoop 2.4

Amazon EC2:[18]

Instance type: General purpose.

Instance name: m3.xlarge

Processor: Intel Xeon E5-2670 v2

RAM: 15 GB.

Hard drive: 40 x 2 GB.

Java 7 (java-1.7.0-openjdk)

Amazon Linux AMI 2014.09.[19]

The file on which the word count is to be performed is uploaded on Amazon S3.S3 stands for Simple Storage Service. The word count application which is in .jar format is also uploaded on s3.

The performance is measured using cluster size 16. In this same data set is given for performance which is increased gradually by 0.8 GB. (Refer Fig 5)

V. DISCUSSION

The first case we have seen that as the cluster size increases the performance improves i.e. the time taken to perform word count on 9.6 GB file reduces gradually.

Table 1 Performance comparisons

Sr. no	File Size (in GB)	Java (Time in Seconds)	Amazon EMR (Time in Seconds)	Performance Improvement
1	0.8	56.6	113	-
2	1.6	116.9	132	-
3	2.4	156.9	134	14.6 %
4	3.2	209	145	30.62 %
5	4.0	261	161	38.31 %
6	4.8	302	167	44.7 %
7	5.6	390	202	48.2 %
8	6.4	444.4	221	50.27 %
9	7.2	495.7	225	54.61 %
10	8.0	542	231	57.38 %
11	8.8	610	260	57.38 %
12	9.6	659	277	57.96 %

Fig 3 Performance on Multiple Clusters

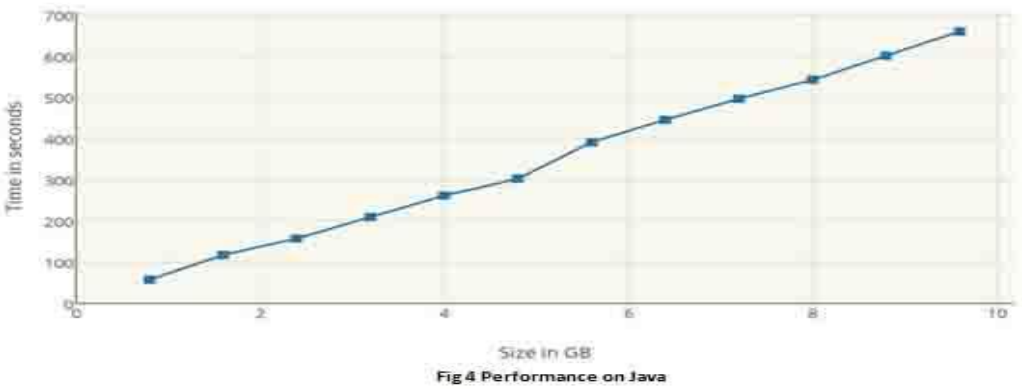


Fig 4 Performance on Java

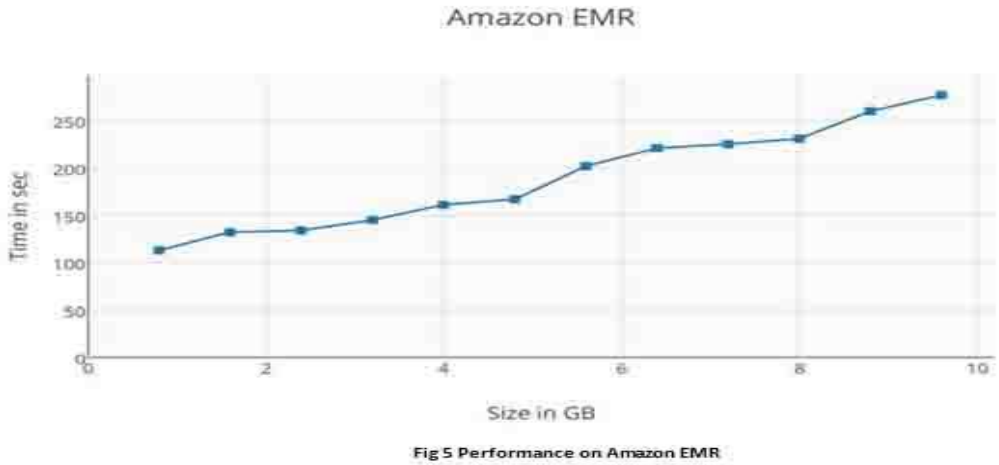


Fig 5 Performance on Amazon EMR

After comparing the time taken by java and Amazon EMR we deduce the follows

i .For the smaller file size i.e.(0.8&1.6). The performance of java is better than Hadoop.

ii. As the file increases the performance of Amazon EMR gradually increases. For 9.6 GB file the performance improved by 58%.

VI. CONCLUSION

We have done performance analysis successfully.

For smaller size the performance of java is better than Hadoop because it works on parallel processing model which takes time for provisioning job and distributing work load. For big data the performance of Hadoop in cloud is better than java.

VII. REFERENCES

[1] http://en.wikipedia.org/wiki/Distributed_computing
[2] http://en.wikipedia.org/wiki/Memory_storage_density
[3] http://en.wikipedia.org/wiki/Machine-generated_data
[4] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, "Big Data Processing in Cloud Computing Environments".
[5] Nandan Mirajkar, Sandeep Bhujbal , Aaradhana Deshmukh, "Perform wordcount Map-Reduce Job in Single Node Apache Hadoop cluster and compress data using Lempel-Ziv-Oberhumer (LZO) algorithm"
[6] http://csrc.nist.gov/publications/nistpubs/800-146/sp800-146.pdf
[7] Ibid
[8] Ibid
[9] Ibid
[10] http://www.gartner.com/it-glossary/big-data
[11] "Cloud Computing: Concepts, Technology & Architecture" (The Prentice Hall Service Technology) by Thomas Erl, Ricardo Puttini and Zaigham Mahmood (May 20, 2013)
[12] "Java The Complete Reference", 8th Edition by Herbert Schildt,8 Aug 2011
[13] https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
[14] Ibid
[15] http://en.wikipedia.org/wiki/Apache_Hadoop
[16] http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-what-is-emr.html
[17] http://docs.aws.amazon.com/AmazonS3/latest/gsg/GetStartedWithS3.html
[18] http://aws.amazon.com/ec2/instance-types/
[19] http://aws.amazon.com/amazon-linux-ami/2014.09-release-notes/