ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Study on Cloud Computing with Hadoop

Ashwini Satkar, Ashwini Patil

Asst. Professor, Dept of Computer Science, Dr. D.Y. Patil A.C.S. College, Pimpri, India

ABSTRACT: All companies need IT infrastructure to run their business in today's fast paced world. Be it a small company having 10-20 people or a big multinational organization having presence all over the world. Most of the companies have their in house IT department to fulfil daily computing needs. There are various departments looking after each device's like servers, networking equipment's. Managing them is becoming difficult task because of their complex nature. Hence cloud computing is ray of hope to solve companies computing needs. Managing daily growing data is also getting difficult. With evolution of computing devices, lots of data is generated on daily basis. Managing that data is cumbersome task. To manage that big data which arises from today's growing industry we can use the concept of cloud computing. But changes in the data patterns and applications, distribution of data of the clusters we use Hadoop platform.

KEYWORDS: Big data, Cloud computing, Hadoop, Map reduce and HDFS

I. INTRODUCTION

Big Data: Big Data is a large volume of data which arrives from multiple resources. It can be structured or unstructured data. It is very fast, too big & doesn't fit in normal databases.

Challenges with big data

- Understands the need of data
- Finding meaningful data very quickly that is maintain the speed
- We have to maintain the quality of data
- Display meaningful results

To manage that big data using traditional techniques like relational database requires more time, more money as well as more cost. Analysis of large data sets along with cloud computing requires platform like Hadoop which collects data from distributed clusters & manage it with multiple resources.

II. WHY CLOUDCOMPUTING?

Cloud Computing means accessing all your data, applications, daily files without worrying about where it is stored and without knowing physical location of your servers. User can only access front end without any idea of backend servers and infrastructure. It enables companies to consume compute resources as a utility like electricity rather than having to build and maintain computing infrastructures in-house. Cloud computing promises several attractive benefits for businesses and end users such as pay per use, self-provisioning as per demand and elasticity. Cloud computing service

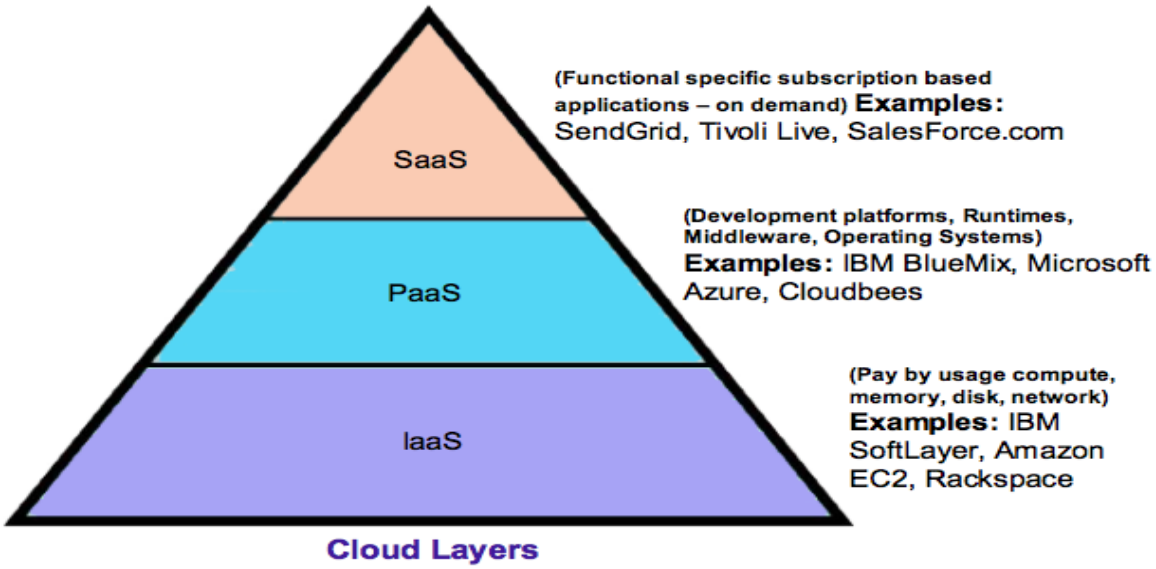
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

can be private, public or hybrid. Cloud computing uses following 3 layers which are as follows:



Cloud Layers

Cloud computing handles large amount of data but changes in the data patterns and applications has made way for the new type of storage called key value storage which are now being widely used by various enterprises. This type of open source platform is called as Hadoop

Structure of Cloud Computing

III. HADOOP

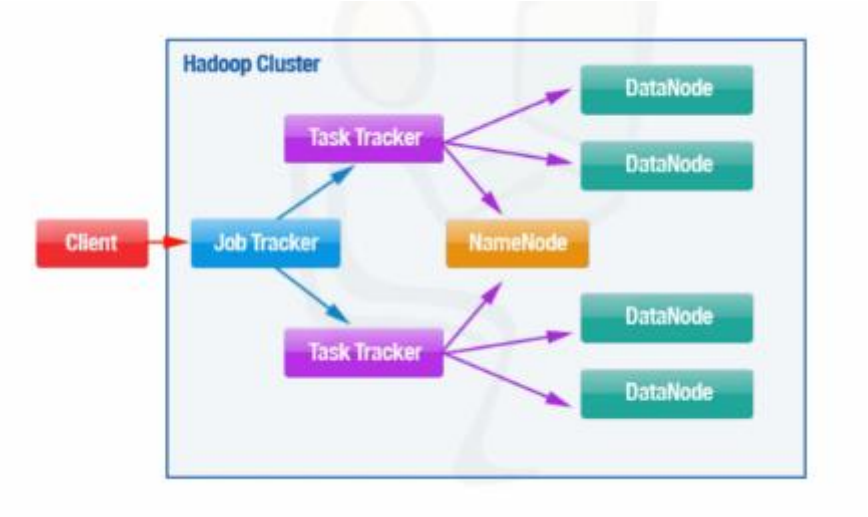
Hadoop is a framework that allows distributed processing of large data sets across the clusters of computers using a programming language. It is open source library & application programs written in JAVA language. Hadoop implements HDFS (Hadoop distributed File system).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015



General Structure of Hadoop

Components of Hadoop are

- **Job Tracker:** Responsible for scheduling the task to the slave nodes
- **Task Tracker:** Performs the task (Map and reduce functions) on the data assigned to it by Master Node
- **Name Node:** Obtains the data from client machine divides it in to chunks
- **Data Node:** It is a slave node. Responsible to store the chunk of that assigned to it by the Name Node.

A. Map Reduce Framework

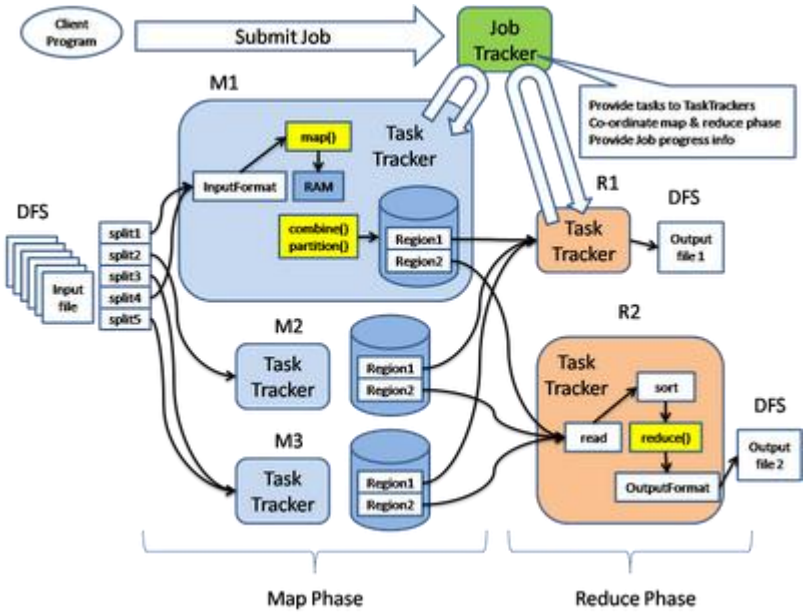
- Hadoop is an open-source Cloud computing environment that implements the MapReduce framework in Java.
- Hadoop makes processing very easy and generate large data sets on the cloud. It works on the data stored in HDFS and act as resources scheduler.
- We can divide the work into smaller chunks using MapReduce framework & concurrently process multiple chunks. We can then combine the results to obtain the final result.
- MapReduce enables one to use the massive parallelism provided by the cloud and provides a simple interface to a very complex and distributed computing infrastructure.
- If we model our problem as a MapReduce problem then we can take advantage of the Cloud computing environment provided by Hadoop.
- The MapReduce framework consists of a single master **Job Tracker** and one slave **Task Tracker** per cluster-node. The master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master.
- Both map and reduce functions can run in parallel & runtime allow to be reduced to several optimizations.
- Map-Reduce is fault resiliency which allows the application developer to focus on the important algorithmic aspects of problem & ignores the issues like data distribution, synchronization, parallel execution, fault tolerance and monitoring.
- It is used by Apache Hadoop so we avoid paying expensive software licenses
- It is flexible to modify source code to meet the evolving needs and take advantage of leading-edge innovations coming from the worldwide Hadoop community.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

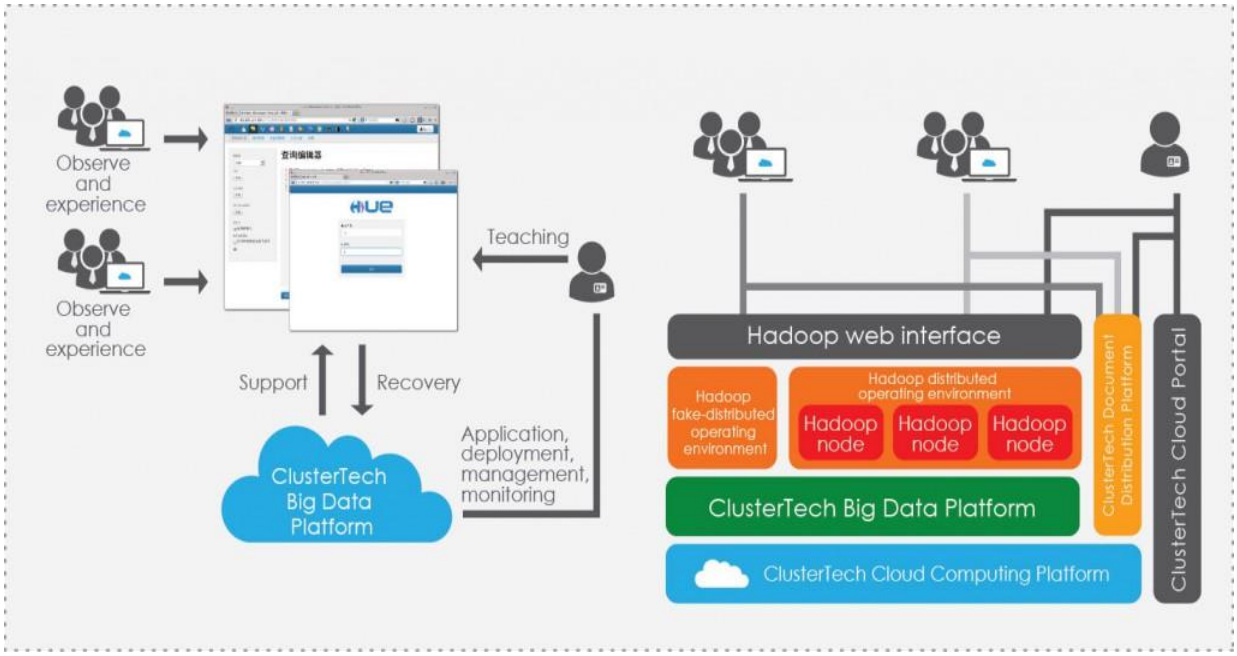
Vol. 3, Issue 12, December 2015



Working of Map Reduce

B. HDFS

Hadoop enables the development of reliable, scalable, efficient, economical and distributed computing using very simple Java interfaces. Hadoop includes a distributed file system, HDFS and a system for provisioning virtual Hadoop clusters over a large physical cluster called Hadoop On Demand (HOD).



Cloud Computing With Hadoop

In this we add the new layer as Hadoop web interface which interact with ClusterTech Big Data platform. It uses several Hadoopnode at every cluster to use distributed computing

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

IV. KEY BENEFITS

- Hadoop lowers the cost of innovation in cloud computing
- Procuring large scale resources quickly
- Handle large workload efficiently
- Handle variable resource requirements
- Running Hadoop clusters in the same cloud environment

V. HADOOP IN YAHOO!

Database for Search Assist™ is built using Hadoop.
3 years of log-data
20-steps of map-reduce

	Before Hadoop	After Hadoop
Time	26 days	20 minutes
Language	C++	Python
Development Time	2-3 weeks	2-3 days

VI. CLOUD COMPUTING & HADOOP

Cloud Computing	Hadoop
It is concept or methodology	Based on methodology of Distributed computing
On demand service	Not On demand service
Not Open source software	Open source apache software project
e.gFacebook, LinkedIn, MySpace, Twitter ,Hotmail ,Google docs	e.g. Amazon/A9, AOL, Facebook, Fox interactive media, Google , IBM, New York Times

VII. CONCLUSION

Data storage is an important element of cloud computing. In this paper we discussed use of HDFS and Map/Reduce in Hadoop framework. Hadoop provides an economically scalable solution for storing and processing large amount of structured and unstructured data over long period of time. Hadoop architecture provides a lot of flexibility. It is very easy to add information, adapt to changing needs of the cloud computing. It can answers queries which are very difficult to solve using RDBMS. Hadoop is inexpensive as compared to the other data store solution available in the market making it very attractive. It also uses commodity hardware, open-source software.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

REFERENCES

1. http://www.ijarcsse.com/docs/papers/Volume_5/1_January2015/V5I1-0335.pdf
2. http://www.ijcsit.com/docs/Volume%206/vol6issue03/ijcsit2015060323.pdf
3. Apache Hadoop. http://hadoop.apache.org/
4. www.cloudera.com
5. https://en.wikipedia.org/wiki/
6. http://wiki.apache.org/hadoop/PoweredBy
7. http://developer.yahoo.com/hadoop/

BIOGRAPHY

Mrs. Ashwini Satkar is a Assistant Professor in the Computer science Department, College of Dr. D. Y. Patil Arts, Commerce And Science, Pimpri, Pune. She received Master of Computer Science (M.C.S) degree in 2009 from SPPU,India. Her research interests are Computers, Networking.

Mrs. Ashwini Patil is a Assistant Professor in the Computer science Department, College of Dr. D. Y. PatilArts, Commerce And Science, Pimpri, Pune. She received Master of Computer Application (M.C.A) degree in 2009 from SPPU, India. Her research interests are Networking, Operating Systems.