A STUDY ON CLOUD COMPUTING, BIG DATA AND HADOOP

N. Karthick1[1], Dr. R. Rajkumar[2]

ABSTRACT

Big data bring new and exciting opportunities to companies who utilize it. The study is used to produce the details about Cloud Computing, Big Data and Hadoop. In day to day life people are handling a large volume of data in their business environment. This study is used to show how to handle the large volume of data for the progress of business with the help of current technology.

Keywords: Big data, Cloud Computing, Current technology

I. INTRODUCTION

Cloud computing

Cloud computing is used to share, host and offer services over the Internet. It is also referred to as utility computing that involves a large number of computers connected through a communication network. Cloud computing provides service to people based on their needs. Utility computing refers to the packaging of computing resources. Distributed computing uses a distributed system to solve computational problems. In parallel computing many calculations are carried out simultaneously. In parallel computing large task are divided into sub groups and then it will start work on it [1][2].

[1]Assistant Professor, Department of MCA & SS

[2]Assistant Professor, Department of ECS, VLB Janakiammal College of Arts and Science, Coimbatore

Cloud is classified into three types:(shown in Figure1)

- Private
- Public
- Hybrid

Private cloud will be operated by own organizations. Only people authorized by them can access the cloud service. The cloud services are provided via cloud computing technology. The organizations have to pay an amount based on their need.

Public cloud represents cloud computing services that anyone can access but cannot own. In this model the service control is not with the user.

Hybrid cloud is a combination of both private and public clouds. The usage depends upon the user's interest.

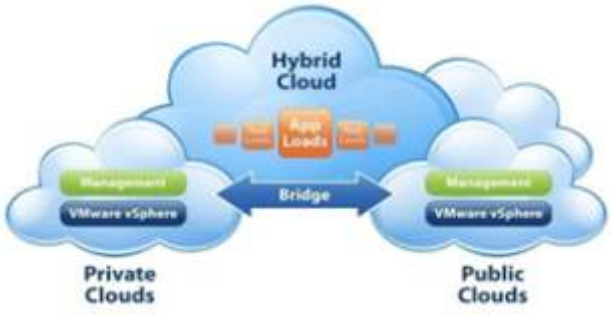

Figure 1: Types of Cloud

Clients form a very important computing component. In cloud computing the client represents the device that the end user uses to interact and image their information

in the cloud [3]. Clients are classified into three categories:

a) Mobile
b) Thick
c) Thin

- Mobile represents the mobile phone devices; it includes blackberry, smart phone, windows phone, etc.
- Thin client does not have internal hard drivers. It is just used to project all the information.
- A thick client represents the normal computer devices used in cloud environment.

I. Big Data

- Big data is normal data, But Big in size. Big data aim to solve all problems in a better way.[5][6]

Big data structure can be categorized into three types:

- Structured
- Unstructured
- Semi structured

People can accomplish the following business-related tasks by using Big data analytics:

- Determination of root causes, issues & failure
- Detecting fraudulent behavior before it affects an organization.
- Recalculating entire risk portfolios in minutes.

Figure 2: Big data Process

3 V's are involved in Big data process[10]

- Volume
- Velocity
- Variety
- Volume represents data quantity, Velocity data speed and Variety data types.
- Storage Area Network (SAN) can provide massive storage that can yield Infinite data. It has multiple application servers and programs run on each application server. But all date reside within one SAN. Each server gets the data from SAN before an execution. After the execution the result will be stored in SAN.[7][9]

Figure 3 represents the work process of SAN.

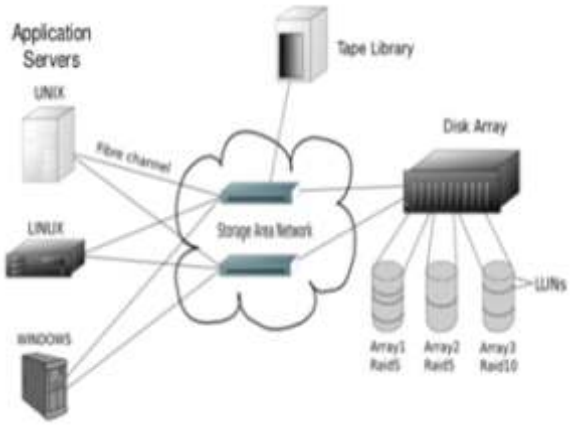

Figure 3: SAN Work Process

- Every system has some draw backs; so, the following things are considered major issues in SAN:
- Data synchronization is required during exchange.
- Partial failures are also difficult to handle.
- It needs huge bandwidth.
- Much amount of time needs to be spent for transferring the data.

- Following figure represents the traditional approach of SAN

I. Hadoop

In 2000 there was a critical challenge to handle the big data. In 2004 Google released two operation part [4][8]:

- Google File System (GFS)
- Map Reduce programming model

Hadoop was inspired by Google. Hadoop is the name of a toy elephant named Doug in a TV program. Hadoop is an open source software frame work, used for managing big data. Hadoop is not a tool but a framework of tools.

The most important part of Hadoop is

- Hadoop distributed file system (HDFS)
- Map Reduce programming model
- The following two figures represent the HDFS & Map reduce model classification:

The Hadoop architecture is classified into two parts:

- Master node
- Slave node

Figure 4 represents the part involved in Hadoop architecture.

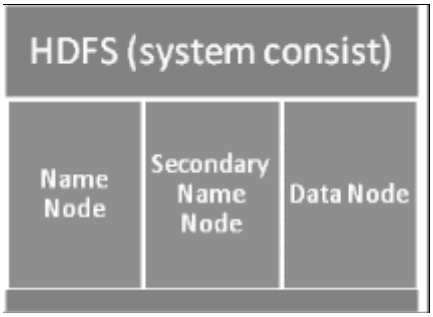

Figure 4: Hadoop Architecture Parts

- Name node is responsible for the distribution of the data throughout the Hadoop cluster. Cluster is the representation of a group of elements.
- Secondary node is otherwise known as backup node. It contains all the snapshot about the name node.
- Job tracker consists of slave nodes to perform a task and it will schedule the task to slave nodes.
- Task tracker is used to perform logic operation known Map & Reduce on Data.

Characteristics of Hadoop

- Scalable
- Cost effective
- Flexible to use
- Fault tolerant
- Reliable
- Easy to use

Advantages of Hadoop :

- Consists of Master & Slave nodes to perform tasks.
- It can handle all types of node failures.

Hadoop node classification is shown in Figure 5

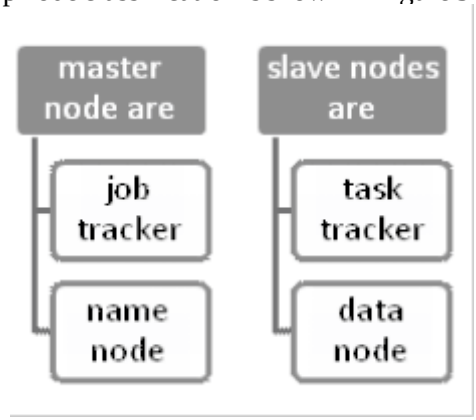

Figure 5: Hadoop Nodes

Limitations of Hadoop :

- Hadoop is not a platform to solve all kinds of problems.
- If data are too small Hadoop will not be able to perform a process.
- Not suitable if large tasks cannot be divided into sub programs.
- To perform Real time and stream-based processing Hadoop is not suitable.

CONCLUSION

This paper clearly gives the basic operations involved in cloud, big data &Hadoop. By understanding the terminology clearly a user can get awareness to handle a large volume of data with the help of cloud and its component in their business environment.

REFERENCES

[1] Anirban Kundu and etal, Introducing New Services in Cloud Computing Environment, International Journal of Digital Content Technology and its Applications, Vol.4(5), 2010, pp. 143-152

[2] RajkumarBuyya and et al, Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems, Future Generation Computer Systems, Vol. 25(6), 2009,pp. 599-616

[3] Dhinesh Babu and Venkata Krishna, Honey bee behavior inspired load balancing of tasks in cloud computing environments, Applied Soft Computing, Vol.15(5), 2013, pp.2292-2303

[4] Suman Arora and Dr. Madhu Goel, Survey Paper on Scheduling in Hadoop, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4(5), 2014, pp. 812- 815.

[5] Catteddu D. (2010) Cloud Computing: Benefits, Risks and Recommendations for Information Security. In: Serrão C., Aguilera Díaz V., Cerullo F. (eds) Web Application Security. IBWAS 2009. Communications in Computer and Information Science, vol 72. Springer, Berlin, Heidelberg.

[6] Open Cloud Computing Interface, http://www.occi-wg.org/doku.php

[7] Rajkumar Buyya, Chee Shin Yeo, "Cloud Computing and Emerging IT Platforms: Vision Hype and Reality for Delivering Computing as the 5th Utility", Future Generation Computer Systems, pp. 599-616, 2009.

[8] Eduardo Correia, Cloud Computing, Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics, 10.4018/978-1-5225-7598-6.ch009, (109-116)

[9] Deniz Tuncalp, Management of Privacy and Security in Cloud Computing, Web-Based Services,10.4018/978-1-4666-9466-8.ch070, (1585-1610), (2016).

[10] Ahmed Mehrez, Cloud Systems in Supply Chains, (2015).