A Survey on Big Data and Cloud Computing

D. Asir Antony Gnana Singh
Department of Computer Science and Engineering,
Anna University, BIT Campus,
Tiruchirappalli, India
asirantony@gmail.com

B. Tamizhpoonguil
Department of Computer Science and Engineering,
Anna University,
BIT Campus,
Tiruchirappalli, India
thamizhpoonguil@ gmail.com

E. Jebamalar Leavline
Department of Electronics and Communication Engineering,
Anna University,
BIT Campus,
Tiruchirappalli, India
jebilee@gmail.com

Abstract**—** In the information age, analyzing and extracting the knowledge from the data is a challenging task since the data are being accumulated massively from various sources and sectors. These massively accumulated data is known as big data since it possess the characteristics such as high volume, different variety and high velocity. Processing these big data using the normal work station is quiet complex since it is saturated with the vertical scalability. Therefore, processing the big data is a challenging task. Hence, the cloud computing arrives for handling massive data for storing and analyzing them to obtain the knowledge and make decisions to improve the productivity and the services. Therefore, conducting the study on the big data and the cloud computing is important to promote the research and development activities in the field of the big data and the cloud computing. Therefore, this paper presents a survey on the big data and cloud computing.

Keywords*-Big data, Cloud Computing, Literature review on Big data, A survey on Big data and Cloud computing*

*****

I. INTRODUCTION

Big data is a collection of huge amount of data. These data may have the three characteristics namely volume, velocity, and variety. Volume denotes the huge amount of data. The data are generated rapidly day-by-day and collected from different sources in different formats. These large volume ranges from gigabytes to tera or petabytes and more. Today the data are not only generated by humans but also by machines too. Velocity denotes the speed of data transmission from various sources such as sensors, surveillance, etc. The processing speed of the application may vary according to their nature of execution. Hence, processing the data that are received in different speed is a challenging task in big data. In the recent past, the data are generated in large volume that may be in semi or unstructured form. The industry needs the structured data in any one of the formats since the industry suffers to process the unstructured data. Therefore, the industries are in the need of processing the big data to extract the knowledge and make decision and they are pushed to use big data analytics tools and techniques for the knowledge extraction process.

The cloud computing provides a platform to process the big data by providing various services such as software as service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS). In SaaS, the software is provided as the service to analyze the big data. In PaaS, the operating system is provided to setup the platform to carry out the big data analytic tasks. The infrastructure as a service (IaaS) provides the storage space to store the big data for further process. Therefore, cloud computing plays a significant role in the big data analytics to extract the knowledge from the big data. Hence, studying the big data and cloud computing is important to extract the knowledge from the big data to make decision and prediction to improve the quality of the services and production. Therefore, this paper presents an study on big data and cloud computing

The rest of the paper is organized as follows: Section II presents the literature review, Section III presents the cloud infrastructure, and Section IV presents the experimental setup, results and discussion. Section V concludes the paper.

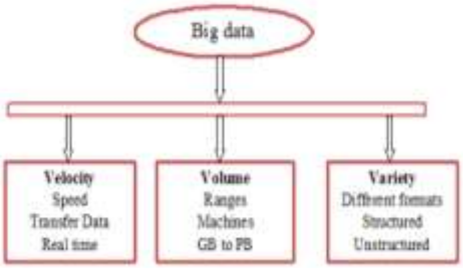

Figure 1. Big data characteristics

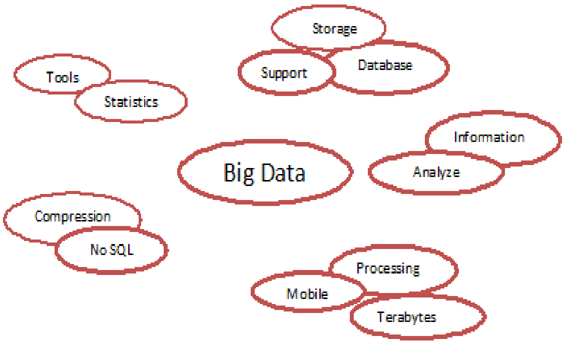

Figure 2. Representation of big data

II. LITERATURE SURVEY

This section presents the literature review on various research works carried out in the field of big data and cloud computing. Venkatesh et al presented a study on big data in cloud computing environment. In this study, they explored that the cloud computing performs and handles complex computing tasks with big data. In recent past, data are generated and massively increasing day-by-day. To handle these data, we need a large space, efficient computing hardware and software. To overcome these challenges, we need big computing power and space for computing those big data. For that purpose, the Hadoop environment is used for computations on big data and to perform the data analytics. This paper also presented different data processing techniques and the challenges faced by the developers and the database management system

(DBMS) designers. The authors discussed the big data management concepts along with the need of security in big data. This paper also proposed a encryption-based security mechanism to secure the data that are stored in the cloud environment [1].

Marcus et al discussed how to carry out big data analytics in cloud with four important areas of analytics such as data management, model development, user interaction, and business models. Large space, efforts, hardware, expensive software and time computational tasks are needed to process the big data. Cloud computing is used to overcome this problem by providing on-demand space with pay-as-you-go-option. The infrastructure provided is always to be scaled up and adapted to the appropriate system [2].

Changqing et al proposed different techniques for big data processing in system and application aspects. Social network, semantic web and bio-informatics analysis are the proof for being continuously growing data to be processed. In this paper, they processed the data processing using two aspects; big data processing mechanisms and cloud data management. The mechanisms are known as cloud computing platform, cloud architecture, cloud database, and cloud storage. Literally they used the map reduce concepts to compute the big data in cloud environment [3]. Bernice et al defined various data analysis methodologies that are enabled by the recent architectures and techniques. However, there is a biggest commitment of using hardware and processing space when using of big data. Big data is prohibitive to medium and small sized companies. Cloud computing is used for those small and medium sized companies. The map reduce techniques are used for storing and processing big data that requires a network attached storage and parallel processing system. Then cloud computing provides network access to share resources in on-demand basis. Three types of cloud computing are used PaaS, SaaS, and HaaS (hardware as a service) that reduce the hardware cost and processing cost [4].

Kalpana et al addressed a data storage security in cloud computing, and introduced an effective and flexible distribution verification mechanism. Cloud computing involves the next generation data storage which is maintained by service providers. If storing the data, ensure that the data meet the security challenges. In this paper, they proposed a data storage security in cloud and mainly focused on the quality of service (correctness of data). During the storage verification, if the data error has been detected, they simultaneously identify the misbehaving servers by using the detailed mechanisms used for data security and performance analysis in the cloud storage system [5]. Ranjan et al described a new computing technology paradigm in IT industry [6].

Gurudatt et al proposed a method to provide storage rent space to a smaller company or an individual. In general, an enterprise is always expected to store backup data whenever they need. To carry out this task, in the past days, the enterprise needs a big tape library and a network administrator in order to maintain the backups on the network which is very costly. But in the recent days, this is carried out by storage service provider (SaaS) and the storage cost is very less. So, the companies sign an agreement with the service provider to rent a particular storage space in terms of gigabyte or data transfer basis and the data is automatically transferred over the network in order to make data backup. If the company’s stored data is lost, the network admin can contact the service provider and they provide the copy of the data. In this paper, they covered many key technologies that are applied in cloud computing [7].

According to Raghavendra et al, a widespread growth of the data in many areas such as business, medical, science thus they result in data explosion. Knowledge discovery in rapidly growing data is a challenging task. Hence, a new paradigm is used for computing the large scale data, and also new techniques and different mathematical models are built for data analytics [8].

Shwetha et al presented a cloud computing methodology for IT industry. The internet is widely used all over the world. Cloud computing is employed using internet and practically it is visualized by all which is the next step or generation in the IT industry. This paper proposed a mechanism to secure the stored data and to provide the quality of service in data accessing. To provide the security, an effective and legible scheme is used for data storage correctness and to allow the trustworthy persons to access the stored data and data error localization [9].

Charlotte et al presented the emerging technologies for big data and cloud computing integration. These two have been rapidly growing in the IT industry for the past decade. Integrating big data with cloud is an important part which serves as a driving force in business as well as IT industry. In this paper, they discussed about various methodologies, solutions, facing problems, and benefits by integrating those two emerging technologies [10].

III. CLOUD INFRASTRUCTURE

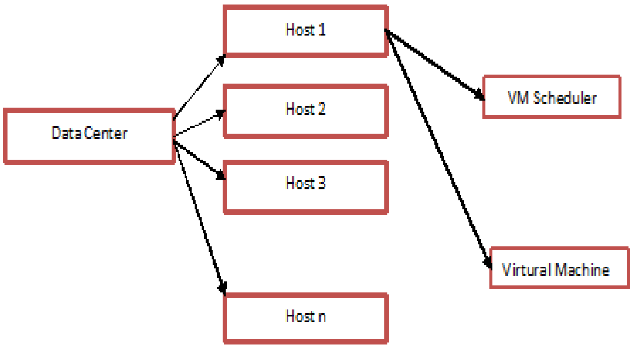

Figure 3. Cloud Infrastructure is connected with host and virtual machine

The cloud infrastructure is illustrated in Figure 3 which consists of a data centre, hosts, virtual machine scheduler and virtual machines (vm). These are interconnected with one another. The number of datacenters connected with a host varies from 1 to n and the host can manage vm scheduler and vms. The cloudlet scheduler can only divide the central process unit (CPU) resources of virtual machines among cloudlets. There are two types of scheduler. They are scheduler space shared and scheduler time shared.

1) Scheduler - space shared: Assigning the CPU specific cores to vms.
2) Scheduler - time shared: The capacity of a core can be distributed by dynamically among vms.
3) Vm scheduler calculates the processing core of a host which are allocated to vm and also determining how many processing core will be authorised to every vms.

TABLE I. LIST OF ABBREVIATIONS USED

Name	List of abbreviations
CIS	Cloud information service
VM ID	Virtual machine
MIPS	Million instructions per second
RAM	Random access memory
BW	Bandwidth
Vmm	Virtual machine monitor

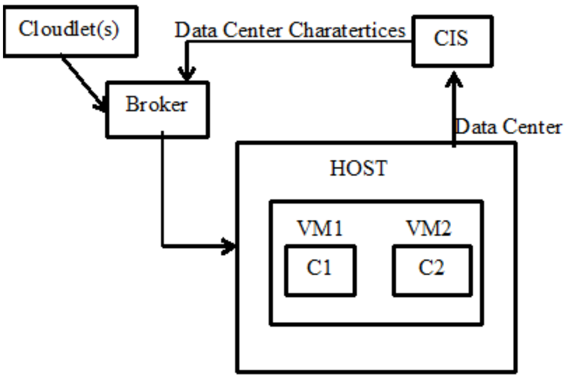

Figure 4. Basic framework of cloudsim simulation tool

IV. EXPERIMENTAL SETUP AND RESULT

The experiment is conducted to create a datacenter with one host and run one cloudlet on it. The steps that are followed to create a data center and a host and cloudlet using Cloudsim are as follows:

A. Descritpions and definitions

1. Set number of user
   - In this step, the numbers of cloud users are initialized. This value also is called as broker count.
2. Initialization of common variables
   - The variables are used to perform the tasks.
3. Cloud information service (CIS) creation
   - CIS is a kind of registry which contains resources that are available on cloud.
4. Data centre Creation
   - Data centre has some host and each host have some set of virtual machines. Data centre needs to register with the CIS as per the cloud sim infrastructure register once created and the registration process is happening in the CIS.
5. Data centre characteristics
   - Each host can have number of processing element (PE), RAM and BW
6. Virtualization
   - Cloud environment works on virtualization which actually differs from other technologies.
7. Host
   - Host can be virtualized into number of virtual machines.
8. Virtual Machine parameters
   - PE, RAM, BW
9. Data centre broker instances
   - Broker needs to submit tasks to the data center.
   - Broker is basically a data center broker class which is always responsible for submitting the tasks to the data center.
   - At the initial state, it interacts with the CIS and retrieves the resource information.
10. Cloudlets
    - The set of cloudlets submitted to broker.
    - Broker has details of the data center and directly interacts with the data centre.
    - Assign the cloudlets to some virtual machine which is running on the host.
11. Start Simulation
12. Stop simulation process
13. Print the status of the simulation

B. Steps to conduct the experiment

First step: Initialize the cloudsim package

This must be called before creating any other entities such as creating the calendar and trace flag which is not much useful in it. This is used to allocate the number of cloud users in it.

Second step: Create Datacenters

Datacenters are providing the resource providers to users in Cloudsim. Each datacenter may have some host.

Third step: Create Broker

Broker is assigning the tasks to the datacenter and acting as an intermediate between the cloudlets and the datacenter.

Fourth step: Create one virtual machine

This is the one virtual machine tasks created the virtual machine which is always under the host.

VM description

```
int vmid = 0;
int mips = 1000;
long size = 10000; // image size (MB)
int ram = 512; // vm memory (MB)
long bw = 1000;
int pesNumber = 1; // number of cpus
String vmm = "Xen"; // VMM name
```

Fifth step: Create one Cloudlet

Cloudlet is another name of the tasks which are provided by the cloud user.

Sixth step: Start the simulation process

Start the stimulation process once the tasks can be submitted to the datacenter.

Final step: Print results when simulation is completed

Start the stimulation process, once the tasks are submitted to the data center.

C. Discussions:

The experiments show that the processing time is gradually increased with respect to MIPS (Million Instructions per Second). MIPS are a RISC architecture which has simple and regular instruction set. In the MIPS, there is only one mode for addressing memory (base plus displacement) and fixed size for instructions. The below implementation is displaying the values to create a data centre with one host and run one cloudlets on it.

TABLE II. EXPRESS THE CONFIGURATION FOR VIRTUAL MACHINE DESCRIPTION

Variables	Values
Vmid	0
Mips	1000
Size	10000
Ram	512
Bw	1000
PesNumber	1
Vmm	"xen"

TABLE III. EXPRESS THE PROCESSING TIME VARIATION WITH RESPECT TO CHANGING THE MIPS

MIPS	Time
1	400000
10	40000
100	4000
1000	400

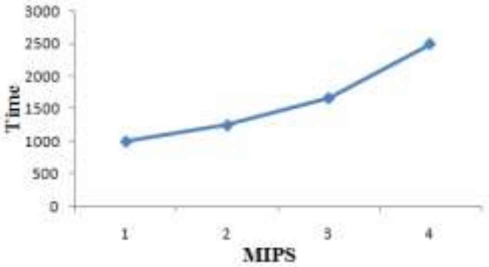

Figure 5. Expressing the variation of processing time with respect to the MIPS increased by gradually.

The below implementation is displaying the values to create a datacenter with one hosts and run two cloudlets on it. The cloudlets run in VMs with the same MIPS requirements.

TABLE IV. SHOWING THE VM DESCRIPTION CONFIGURATION FOR ONE HOST AND TWO CLOUDLETS

Variables	Values
Vm ID	0
MIPS	250
Size	10000
Ram	2048
BW	1000
PesNumber	1
VMM	"Xen"

TABLE V. SHOWING THE PROCESSING TIME IS INCREASING

MIPS	Time
100	2500
150	1666.67
200	1250
250	1000

The processing time is displayed for two cloudlets running in a host along with the same MIPS requirements. But, the two cloudlets will take the same time to complete the execution in the virtual machine using the same MIPS requirements. In the above Table, the time is showing the same time for two cloudlets even changing the MIPS values in the VM description. The below implementation displayed the values to create a datacenter with two hosts and run two cloudlets on it. The cloudlets run in VMs with different MIPS requirements. The cloudlets will take different time to complete the execution depending on the requested VM performance.

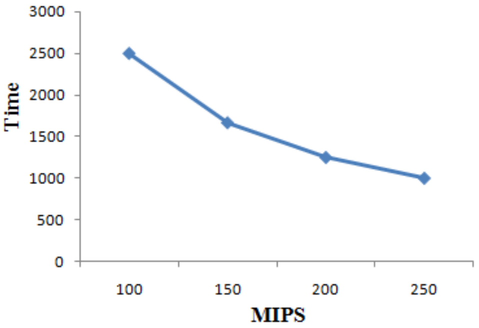

Figure 6. Plotted the diagram as per the MIPS changes in the VM description / configuration for the two cloudlets

The below implementation is for displaying the values to create a datacenter with two hosts and run two cloudlets on it. The cloudlets run in VMs with different MIPS requirements:

TABLE VI. SHOWING THE VM DESCRIPTION FOR TWO HOSTS AND TWO CLOUDLETS

Variables	Values
Vm ID	0
MIPS	250
Size	10000
Ram	2048
BW	1000
PesNumber	1
VMM	"Xen"

TABLE VII. PROCESSING TIME VARIES WITH RESPECT TO MIPS

MIPS	Time1	Time2
100	200	400
150	133.33	266.67
200	100	200
250	80	160

The processing time is displayed for the two cloudlets and two hosts are running in a host along with the same MIPS requirements. The previous experiment takes the same time for two cloudlets and one host running in the virtual machine. But in this experiment, the two cloudlets and two hosts are taking different time to complete the execution in the virtual machine using the same MIPS requirements. So the time is changed for the two cloudlets by changing the MIPS values. In the above table, according to changing the MIPS values in the VM description, it is showing the time is constantly varying based on the two cloudlets and two hosts.

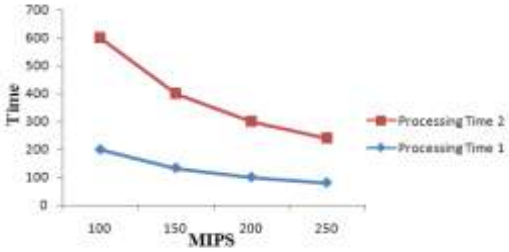

Figure 7. Plotted the diagram as per the MIPS changes in the VM Description / Configuration

From the above tables, it is observed that the execution of time gets vary with respect to the MIPS value changes.

V. CONCLUSION:

This paper presented a method for storing the data on cloud using the cloudsim package. The processing time is varying depends upon on the MIPS range values. The above implementation is displaying the values to create a data centre with host’s number increasing and run number of cloudlets on it. The result charts are displayed based on the different MIPS requirements. The processing time is increasing gradually according to the MIPS which means that it takes increase or decrease in time with respect to the instruction processed per second.

REFERENCES

[1] Venkatesh H, Shrivatsa D Perur, Nivedita Jalihal A Study on Use of Big Data in Cloud Computing Environment (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2076-2078

[2] Assunção, Marcos D., Rodrigo N. Calheiros, Silvia Bianchi, Marco AS Netto, and Rajkumar Buyya. "Big Data computing and clouds: Trends and future directions." Journal of Parallel and Distributed Computing 79 (2015): 3-15.

[3] Ji, Changqing, Yu Li, Wenming Qiu, Uchechukwu Awada, and Keqiu Li. "Big data processing in cloud computing environments." In Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on, pp. 17-23. IEEE, 2012.

[4] Purcell, Bernice M. "Big data using cloud computing." Journal of Technology Research 5 (2014): 1.

[5] Kalpana Batra, Ch. Sunitha, Sushil Kumar.: “An Effective Data Storage Security Scheme for Cloud Computing”. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 4, June 2013.

[6] Kumar, Ranjan, and G. Sahoo. "Cloud Computing Simulation Using CloudSim." arXiv preprint arXiv:1403.3253 (2014).

[7] Gurudatt Kulkarni, Ramesh Sutar, Jayant Gambhir,: “Cloud Computing-Storage as Service” International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622, Vol. 2, Issue 1,Jan-Feb 2012, pp.945-950.

[8] Raghavendra Kune1, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige and Rajkumar Buyya.: “The anatomy of big data computing” Published online 9 October 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.2374, Softw. Pract. Exper. 2016; 46:79–105.

[9] Bindu, B. Shwetha, and B. Yadaiah. "Secure data storage in cloud computing." International Journal of Research in Computer Science ISSN (2011): 2249-8257.

[10] Charlotte Castelino, Dhaval Gandhi, Harish G. Narula, Nirav H. Chokshi, :”Integration of Big Data and Cloud Computing” International Journal of Engineering Trends and Technology (IJETT) ISSN: 2231-5381– Volume 16 Number 2 – Oct 2014.

[11] http://www.cloudbus.org/cloudsim/