Cloud Computing Techniques in Big Data

Sanjay Kumar, Avnish Mishra, Manoj Kumar, Sandeep Kushwaha, C.K Raina
Adesh Institute of Technology, Gharuan, Mohali (Punjab)

Abstract—: Today we live in a modern world. In our daily life we communicate with other people using many information technologies. By using these information technologies, we produce large amount of data which required huge storage and processing. The cloud is an online mechanism on which the data is stored on virtual servers. The big data technique is used for analyzing the data as the size of data is very large. The big data processing represents a new challenge in cloud computing as the data processing techniques involves data storage acquisition and processing of data. Now the question arises what is the relationship between the cloud computing and big data. After studying this paper, you will able to find the relation among these two.

I. INTRODUCTION

Today we produce large amount of data by communication through various information technologies. This data required large amount of space to store and processing this data. We use cloud computing to store the data on various virtual servers. N

ow question arises what is cloud computing and what is big data? How cloud computing helps in processing the big data? Here is the answers of your all these questions

II. CLOUD COMPUTING

The term cloud computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide the services. These services have been known as referred to as Software as a Service (SaaS). The terms such as PaaS (Platform as a Service) and IaaS (Infrastructure as a Service) are used by sellers of the cloud (vendor) to describe their products, but we avoid these because accepted definitions for them still changing very frequently. There is brittle line between "low-level "infrastructure and a higher-level "platform ". We believe that both of these are indifferent than different, and we can consider them together. Similarly, some term such as "grid computing" from the high-performance computing community, suggests the protocols that offer shared computation and storage over long distances; however, these protocols did not lead to a software environment that grew beyond its own community. The data center hardware and software is called a cloud. When a cloud is made available in a pay-as you- go manner to the common public, we call it a public cloud; the service being sold is utility computing. We use the term private cloud to refer to internal data centers of a business or other organization, not made available to the general public, they are large enough to benefit from the advantages of cloud The cloud computing is the sum of SaaS and utility computing, but does not include medium sized data centers, even if these depend on virtualization for management. People can be the users or the providers of SaaS, or users or providers of utility computing. We focus on cloud users and cloud providers, which received less attention than SaaS users. There are some cases in which the same actor plays multiple roles

Figure 1: Cloud Computing

Types of cloud computing models:

Cloud Computing Models Cloud Providers offer services that can be grouped into three categories.

1. Software as a Service (SaaS): In SaaS model a complete application is offered to the customer, as a service on demand. A single instance of the service runs on the cloud & multiple end users are serviced. On the customers" side, there is no need for upfront investment in servers or software licenses, while for the provider, the costs are lowered, since only a single application needs to be hosted & maintained. Today SaaS is offered by companies such as Salesforce, Google, Zoho, Microsoft, etc.

2. Platform as a Service (Paas): in this model, a layer of software, or development environment is encapsulated & offered as a service, upon which other higher levels of service can be built and run. The customer has the freedom to make his own applications, which run on the provider's infrastructure. To meet manageability and scalability requirements of the applications, PaaS providers offer a predefined combination of OS and application servers, such as restricted LAMP platform (Linux, J2EE, Apache, Ruby, MySQL and PHP) etc. Google's App Engine, Force.com, etc. are some of the popular PaaS examples.

3. Infrastructure as a Service (Iaas): this model provides basic storage and computing capabilities as standardized

services over the network. Servers, storage systems, data Centre, space, networking equipment etc. are pooled and made available to handle workloads. The customer would typically deploy his own software on the infrastructure. Some common examples are Amazon, 3 Tera, Go Grid, etc.

Types of cloud:

Public Cloud The public cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services. In this type of clouds, resources are offered as a service, usually over an internet connection, for a pay-per-usage fee. Users can scale their use on demand and they do not need to purchase hardware to use that service. Public cloud providers manage the infrastructure and pool resources into the capacity required by its costumer. Public clouds are available to the general public or large organizations, and are owned by a third-party organization that offers the cloud service. A public cloud is hosted on the internet and designed to be used by any costumer with an internet connection to provide a similar range of capabilities and services. Public cloud users are typically residential users and connect to the public internet through an internet service provider's (IPS) network. Google, Amazon and Microsoft are examples of public cloud vendors who offer their services to the public. Data created and submitted by consumers are usually stored on the servers of the third party vendor. The advantages of public cloud include:

1. Data availability and continuous uptime.

2. 24/7 technical expertise.

3. on demand scalability.

4. Easy and inexpensive setup

5. No wasted resources Drawbacks of public cloud.

6. Data security

Privacy Another issue with public cloud is that you may not know where your data is stored or how it is backed up, and whether unauthorized users can get access to it. Reliability is another concern for public cloud networks. A recent two-day Amazon cloud outage, for example, left dozens of major e-commerce websites disabled or completely unavailable

Examples of public cloud include:

1. Amazon AWS

2. Google Apps

3. Salesforce.com

4. Microsoft BPOS

5. Microsoft Office 365

Private Cloud: Private clouds are built exclusively for a single enterprise. They aim to address concerns on data security and offer greater control, which is not available in a public cloud. There are two variations to a private cloud: -

On-premise Private Cloud: On-premise private clouds, also known as internal clouds are hosted within one's own data center. This model provides a more standardized process and protection, but is limited in aspects of size and scalability. IT departments would also need to incur the capital and operational costs for the physical resources. This is best suited for applications which require complete control and configurability of the infrastructure and security. –

Externally hosted Private Cloud: This type of private cloud is hosted externally with a cloud provider, where the provider facilitates an exclusive cloud environment with full guarantee of privacy. This is best suited for enterprises that don"t prefer a public cloud due to sharing of physical resources.

3 Hybrid Cloud: Hybrid Clouds are the combination of both public and private cloud models. With a Hybrid Cloud, service providers can utilize third party Cloud Providers in total or partial manner thus it increase the flexibility of computing. The Hybrid cloud environment is capable of providing on-demand, externally provisioned scale. The ability to augment a private cloud with the resources of a public cloud can be used to manage any unexpected surges in workload.

III. BIG DATA

The term big data refers to the data which is large in size. This large amount of data is comes and composed by the digital electronic computations from various resources.it required high power and high processing speed to analysing the data. The importance of the big data lies in the analytical use which provide the fast and efficient results and also provide faster and better services.

Characteristics of big data: the dig data follows five 'Vs. The five 'Vs' are as follows:

1. **Volume**: It represents the amount of data that produced by the multiple sources which show the huge data is in numbers e.g. Zettabytes. The volume of the data is most important dimension in which concerns with the big data.

2. **Variety**: It represents the different data types, with increasing the number of social network and smart phones users, the form of data has changed from structured data in databases to unstructured data and semi structured data.

4. **Veracity:** It represents the quality of the data, which demonstrate the confidence in the data content and the accuracy of the data. The quality of the data captured can differ, which affects the accuracy of analysis of the data.

5. **Value:** It represents the importance of big data, it demonstrate the importance of data after analysis of the data.

The types and nature of data:

1. **Structured data**: It is the organized data in the format or a definite structure. For example the data is in the form of tables or databases to be processed.

2. **Unstructured data:** It represents the data in a n unstructured manner it makes biggest proportion of data. For example data that generated by the people on daily basis as texts, images, records, videos, messages, log click-streams etc.

3. **Semi-structured data or multi-structured**: It represents the structured data but that is not designed in the form of tables or databases.

IV. The Model between the cloud and big data

The most common model are used for providing the big data analytics solution in cloud computing are PaaS and SaaS. The IaaS model is not used usually for high-level application that are used for data analysis but are capable to handle storage and computing requirement of the data, these Cloud computing models can accelerate the potential for the scalable analytical solutions .the cloud computing is member of the distributed computing family that provides us the resources in the form of user services such as software as a service (SaaS), infrastructure as a service (IaaS) and a platform as a service l (PaaS), but with the advantage of the big data, cloud computing models are gradually shift to big database service including (BDaaS ,AaaS) which is known as (DaaS) database as a service which means that the database services are available for applications that are deployed in an computing environment. The BDaaS is a form of service same as the software as a service or infrastructure as a service. Huge data as a service often relies on cloud storage to maintain continuous data access to the enterprise that owns the information and the provider it works with and is considered to be hosted in the cloud. Similar types of services include (SaaS) or service-based infrastructure, (IaaS) where the specific data is used as service as a options which help us to handle big data. It provides a lot of facilities for the companies today , where all of these are combined to create the ultimate solution for the companies , still DBaaS is relatively hazy term, but mostly DBaas is refers to the host of outsourced services and the functions are related to Big Data handling in the cloud computing environment. the models for the cloud based big data analytics are divided into two types of services for Cloud analytics, Analytics as a Service (AaaS), where the analytics is provided to the clients on their demand and they can pick the solutions they need for their purpose, and the Model as a Service (MaaS) where models are offered as the building blocks for analytics solutions .recently, the terms such as Analytics as a Service (AaaS) and Big Data as a Service (BDaaS) are becoming more popular. They comprise the services for the data analysis same as IaaS offers computing resources. However, these analytics services are still lacking in the well-defined contracts it may be difficult to measure the quality and the reliability of results and the input data promises on running times.

V. Relationship between cloud computing and big data

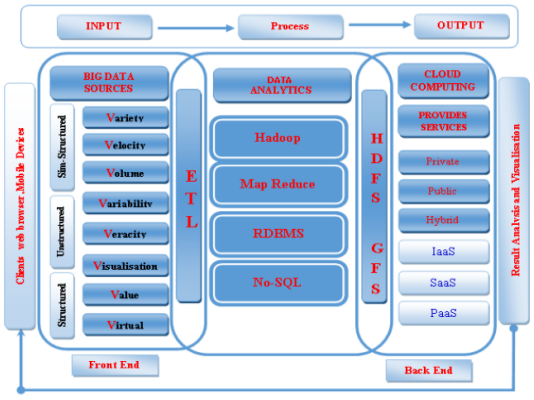
The development of the technology has led to the rapid development of the electronic information society and cloud computing is a trend in the development of the technology, this leads the phenomenon of the big data and the rapid increase in size of big data. It is a major problem that may face the development of the electronic information society.

The cloud computing and big data club together, as the big data is concerned with storage and capacity of the data in the cloud systems, cloud computing uses huge amount of the computations and storage resources. So by providing the big data application with computational capabilities, big data accelerates and stimulates the development of the cloud computing. The distributed storage technology in the computing environment helps to manage the big data. The big data cloud computing are complementary to each other. The problem associated with big data is the rapid growth of data. The Clouds are evolving and providing the solutions for the appropriate computing environment of big data while the traditional database cannot meet the requirements for dealing with the big data, in addition we need for data exchange between the various distributed storage locations. The cloud computing provides the solutions and addresses problems with the big data. The cloud computing environment is expanding to be able to digest the big amounts of data as it follows the policy of data splitting, which means, to store the data in more than one location so that availability area increases. Cloud computing environments are built for the general purpose of workloads and resource pooling which are used to provide flexibility on demand. Thus, the cloud computing seems to be well suited for the big data. Big data processing and storage require the expansion which is provided by the cloud through the use of the virtual machines that helps big data to evolve and become accessible. It is the consistent relationship between them. IBM, Google, Amazon and Microsoft are examples of the success in using big data in the cloud computing environment. In order to fit with big data the cloud computing environment must be modified to suit with the data and cloud work together. Many changes are required to be made on cloud for example CPUs to handle big data and others.

Big companies such as Google, Yahoo, Facebook, and others offer web-based services. The amount of data they collect on daily basis through the online user interactions that overwhelmed traditional IT capabilities. Thus, there is a need of the development of basic infrastructure components has to be developed. The Apache Hadoop has been introduced as a realistic benchmark in managing the big amounts of data which is unstructured in nature. Apache Hadoop is a open platform distributed software that is used for storing and processing data. By using Hadoop, a person store big amounts on tens of 10,000's of servers with effective scaling performance in terms of cost. MapReduce is based on the distribution of a data set between the multiple servers, the partial results are reassembled. The Big data are characterized into different types that require big data. ETL process deals with data diversity of

the data. The ETL represents several the functions such as extraction, transformation, and loading. These three functions are combined into a single tool to pull data from the one database and place it in another database. It helps us to convert the databases from one form to another form .The effectiveness of big data relies on data integrity. If the big data is stored at the local level, it will consume a huge amount of work to manually merge all the data to manage it. The cloud can do itself all the work for the user, by offering one site to store and manage all the commercial data. The cloud computing offers many features and benefits to big data through ease of use, low cost in resource utilization on supply, access to resources and demand, and reduces the use of solid equipment used to handle big data. The goal of both the big data and the cloud is to increase the value of a company while reducing the investment costs. The cloud reduces the cost of managing of the local software, while the big data reduces the investment charges. It seems that these two concepts club together to provide greater value to the companies. Any technological system must pass through the several main stages. The computing system follows the input, processing and output model. Input is given through input devices and then processed through the CPU. The results of the information are produced. The relationship between the data and cloud computing is as follow

Data is stored on several external and distributed storage units. But the computer system, the store the data internally or locally. Thus, the relationship between the big data and cloud computing represents the input, processing and output model. The big data is entered through devices such as the mouse, joy stick, light pen, cellular devices and other smart devices. Processing is done through the tools and techniques used by the cloud computing in the provided service, and the outputs are the results, it represents the importance of data after processing. The input and output model defines input, output and processing the tasks which are required to convert input to output. Inputs represents flow of data and raw materials. The processing step includes all tasks required for transformation of inputs. The output is data flowing from the transformation process.



VI. COMPATIBILITY BETWEEN BIG DATA AND CLOUD COMPUTING IN TERMS OF CHARACTERISTICS

Characteristics Big Data	**Concept**	**Characteristics Cloud Computing**
Velocity Visualisation	Data Rates Data Representation	•Network Bandwidth •Gigabit rates today •Broad network access •Anywhere access - public cloud •Resource pooling:
Variety Veracity	DataType Data Sources Trustworthiness Of The Data	•Cloud data management , No-SQL Databases •Anywhere Access Public Cloud •Mapreduce/Hadoop is Data Processing And Analytics Technology • SLA , QoS • ETL technology
Volume	Size Data	•Scalability- Elasticity According to Demand •Cost Pay-As-You-Go Based on Usage. Reduced cost Reduced cost •Resource Pooling: •On-Demand Self-Service
Virtual	Physical infrastructure data	• Virtual Machine (VM) Is A Software Application •Resource Pooling: Physical Infrastructure
Value	Data Analysis Results, Reports	• OLAP • OLTP

VII. COMMON POINTS BETWEEN BIG DATA AND CLOUD COMPUTING

1. Infrastructure: The cloud computing provides the infrastructure to the big data analysis. As the big data is huge in size therefore, it requires space to store and process the data. the cloud is scalable and provide high end management and security to the big data.

2. On- demand: The cloud has both user and server. The data comes from both server and user. The data size big in size and need large space to store the data. The cloud provides the on-demand resources to store and access the big data.

3. Security: The data whether it is big or small need to store, process, and security. The cloud environment provides guaranty complete confidentiality of the data as only authorized user can access the data. The identity and controls

are a provided to the data and resources according to the need of the user

4. Continuity: The cloud provides the continuity.it provide the resources even in the case of absence and fault in the component. The cloud is distributed over geographical locations so there are a high availability of error to occur this increase the fault tolerance techniques.

VIII. CONCLUSION

The cloud computing and the big data are complementary to each other .The cloud represent the container to store the products and big data represent the product. The cloud ensures the continuity of the process and ignores the faults. The cloud reduces the cost of infrastructure by virtualization. It provides the on demand resources. All these characteristics integrated a relation between the cloud and big data.

IX. REFERENCES

[1]. [Online Available] https://www.google.co.in/search?client=ucweb-b-bookmark&ei=M1y7WuOfKcvEvQT4uYWIAQ&q=cloud+computing+techniques+for+big+data&oq=cloud+computing+techniques+in+big+&gs_l=mobile-gws-serp.1.0.0i22i30j33i160l2.118.2329.3796.0.552.2887.2-1j4j2j1.1.mobile-gws-wiz-serp.0j33i22i29i30.xsA7ZcLV%2B0c

[2]. [Online Available] http://ieeexplore.ieee.org/xxsA7ZcLV%2B0c%3D

[3]. [Online Available] https://www.google.co.in/search?client=ucweb-b-bookmark&ei=M1y7WuOfKcvEvQT4uYWIAQ&q=cloud+computing+techniques+for+big+data&oq=cloud+computing+techniques+in+big+&gs_l=mobile-gws-serp.1.0.0i22i30j33i160l2.118.2329.3796.0.552.2887.2-1j4j2j1.1.mobile-gws-wiz-serp.0j33i22i29i30.xsA7ZcLV%2B0c%3D

[4]. [Online Available] https://www.google.co.in/search?client=ucweb-b-bookmark&ei=M1y7WuOfKcvEvQT4uYWIAQ&q=cloud+computing+techniques+for+big+data&oq=cloud+computing+techniques+in+big+&gs_l=mobile-gws-serp.1.0.0i22i30j33i160l2.118.2329.3796.0.552.2887.2-1j4j2j1.1.mobile-gws-wiz-serp.0j33i22i29i30.xsA7ZcLV%2B0c%3D