IJCRT Publish Paper record is available at DOI: http://doi.one/10.1727/IJCRT.17197

Study of Relation between Big Data & Cloud Computing: Big Data Challenges & Issues

Prof. R. N. Yeotikar
MCA, PGDCS, MCM
Department Of Management
Vidya Bharati Mahavidyalaya, Amravati

Abstract

The amount of data in world is growing day by day. Data is growing because of use of internet, smart phone and social network. Big data is a collection of data sets which is very large in size as well as complex. Now a days, Big data is one of the most talked topic in IT industry. It is going to play important role in future. Big data changes the way that data is managed and used. Communicating by using information technology in various ways produces big amounts of data. Such data requires processing and storage. The cloud is an online storage model where data is stored on multiple virtual servers. Big data processing represents a new challenge in computing, especially in cloud computing. Data processing involves data acquisition, storage and analysis. In this respect, there are many questions including, what is the relationship between big data and cloud computing? The answer to this question will be discussed in this paper, where the big data and cloud computing will be studied, in addition to getting acquainted with the relationship between them in terms of safety and challenges. I have suggested a term for big data, and a model that illustrates the relationship between big data and cloud computing.

Keywords: **Big Data; security; systematic mapping study, cloud, resources, 'five Vs'.**

I. INTRODUCTION

Data is the raw material for information before sorting, arranging and processing. It cannot be used in its primary form prior to processing. Information represents data after processing and analysis. The technology has been developed and used in all aspects of life, increasing the demand for storing and processing more data. As a result, several systems have been developed including cloud computing that support big data. While big data is responsible for data storage and processing, the cloud provides a reliable, accessible, and scalable environment for big data systems to function. Big data is defined as the quantity of digital data produced from different sources of technology for example, sensors, digitizers, scanners, numerical modeling, mobile phones, Internet, videos, e-mails and social networks. The data types include texts, geometries, images, videos, sounds and combinations of each. Such data can be directly or indirectly related to geospatial information.

Cloud computing refers to on-demand computer resources and systems available across the network that can provide a number of integrated computing services without local resources to facilitate user access. These resources include data storage capacity, backup and self-synchronization. Most IT Infrastructure computing consists of services that are provided and delivered through public centers and servers based on them. Here, clouds appear as individual access points for the computing needs of the consumer. They are an online storage model where data are stored on multiple virtual servers, rather than being hosted on a specific server, and are usually provided by a third party. The hosting companies, which have advanced data centers, rent spaces that are stored in a cloud to their customers in line with their needs.

The relationship between big data and the cloud computing is based on integration in that the cloud represents the storehouse and the big data represents the product that will be stored in the storehouse, since it is not possible to create storehouses without storing any product in them. The traditional databases known as 'relational' are no longer sufficient to process multiple-source data. For example, how can these traditional methods deal with data such as record of transactions, customer behavior, mobile phone and GPS navigation, and others. Here comes the role of cloud computing. At this

point, a relationship between big data and the cloud will arise. In this paper, the relationship between them will be discussed, in addition to the challenges and issues that Big Data may encounter.

II. BIG DATA

Big data comes and is composed through electronics operations from multiple sources. It requires proper processing power and high capabilities for analysis. The importance of big data lies in the analytical use which can help generate an informed decision to provide better and faster services.

The term big data is called on the huge amount of high-speed big data of different types; this data cannot be processed and stored in regular computers. The main characteristics of big data known as 'five Vs', are as follows:

1. Volume: It represents the amount of data produced from multiple sources which show the huge data in numbers by zeta bytes. The volume is most evident dimension in what concerns to big data.
2. Variety: It represents data types, with, increasing the number of Internet users everywhere, smart phones and social networks users, the familiar form of data has changed from structured data in databases to unstructured data that includes a large number of formats such as images, audio and video clips, SMS, and GPS data.
3. Velocity: It represents the speed of data frequency from different sources, that is, the speed of data production such as Twitter and Facebook. The huge increase in data volume and their frequency dictates the need for a system that ensures super-speed data analysis.
4. Veracity: It represents the quality of the data, it shows the accuracy of the data and the confidence in the data content. The quality of the data captured can vary greatly, which affects the accuracy of analysis. Although there is wide agreement on the potential value of big data, the data is almost worthless if it is not accurate.
5. Value: It represents the value of big data, i.e. it shows the importance of data after analysis. This is due to the fact that the data on its own is almost worthless. The value lies in careful analysis of the exact data, the information and ideas it provides. The value is the final stage that comes after processing volume, velocity, variety, contrast, validity and visualization .

III. CLOUD COMPUTING

It is a term that refers to on-demand computer resources and systems that can provide a number of integrated computer services without being bound by local resources to facilitate user access. These resources include data storage, backup and self-synchronization, as well as software processing and scheduling tasks. Cloud computing is a shared resource system that can offer a variety of online services such as virtual server storage, and applications and licensing for desktop applications. By leveraging common resources, cloud computing is able to achieve expansion and provide volume.

A. *Characteristics of cloud computing*

That cloud computing is one of the distributed systems that represents a sophisticated model. NIST has identified important aspects of the cloud, as it shortened the concept of cloud computing in five characteristics as follows:

1) *On-demand self-service*: Cloud services provide computer resources such as storage and processing as needed and without any human intervention.

2) Broad network access: cloud computing resources are accessible over the network , mobile and smart devices even sensors can access computing resources on the cloud.

3) Resource Pooling: Cloud platform users share a vast array of computing resources; users can determine the nature of resources and the geographic location they prefer but cannot determine the exact physical location of these resources.

4) Rapid Elasticity: Resources from storage media, network, processing units and applications are always available and can be increased or decreased in an almost instantaneous fashion, allowing for high scalability to ensure optimal use of resources.

5) Measured service: Cloud systems can measure the processes and consumption of resources as well as surveillance, control and reporting in a completely transparent manner.

B. Cloud computing service models.

Cloud computing types are classified on the basis of following models:

1) *Software as a service (SAAS)*: Cloud service providers provide various software applications to users who can use them without installing them on their computer. The user is not responsible for anything other than adjusting the settings and customizing the service as appropriate to his needs. SAAS helps big-data clients to perform data.
2) *Platform as a service (PAAS)*: Cloud service providers provide platforms, tools and other services to users, where the cloud service provider manages everything else, including the operating system and middleware., with resources that enable you to deliver everything from simple cloud-based apps to sophisticated.
3) *Infrastructure as a service (IAAS)*: Cloud service providers provide infrastructure such as storage, computing capacity, etc. is a form of cloud computing that provides virtualized computing resources over the Internet , In an IaaS model, a third-party provider hosts hardware, software, servers, storage and other infrastructure components on behalf of its users.
4) *DaaS* : It is the alternative cloud computing model, as it differs from traditional models like (SAAS, IAAS, PAAS) in providing data to users through the network, as data is considered the value of this model in conjunction with cloud computing based on solving some of the challenges in managing a huge amount of data. For these reasons, DaaS is closely related to big data whose technologies must be utilized. DaaS provides highly efficient methods of data distribution and processing. DaaS is closely related to SaaS (storage as a service) and SaaS (software as a service) which can be combined with one of these models or both of them.

IV. THE RELATIONSHIP BETWEEN THE CLOUD AND BIG DATA

Cloud computing is a trend in the development of technology, as the development of technology has led to the rapid development of electronic information society. This leads to the phenomenon of big data and the rapid increase in big data is a problem that may face the development of electronic information society. Cloud computing and big data go together, as big data is concerned with storage capacity in the cloud system, cloud computing uses huge computing and storage resources. Thus, by providing big data application with computing capability, big data stimulate and accelerate the development of cloud computing. The distributed storage technology in environmental computing helps to manage big data.

Cloud computing and big data are complementary to each other. Rapid growth in big data is regarded a problem. Clouds are evolving and providing solutions for the appropriate environment of big data while traditional storage cannot meet the requirements for dealing with big data, in addition to the need for data exchange between various distributed storage locations. Cloud computing provides solutions and addresses problems with big data. The cloud computing environment is expanding to be able to absorb big amounts of data as it follows the policy of data splitting, that is, to store data in more than one location or availability area. Cloud computing environments are built for general purpose workloads and resource pooling is used to provide flexibility on demand. Therefore, the cloud computing environment seems to be well suited for big data.

Big data processing and storage require expansion as the cloud provides expansion through virtual machines and helps big data evolve and become accessible. This is a consistent relationship between them. Google, IBM, Amazon and Microsoft are examples of the success in using big data in the cloud environment. In order for the cloud environment to fit with big data the cloud computing environment must be modified to suit data and cloud work together. Many changes are needed to be made on the cloud: CPUs to handle big data and others.

V. CHALLENGES IN BIG DATA AND CLOUD COMPUTING

The security challenges in cloud computing environments fall under several levels: the network level which includes dealing with network protocols and network security such as distributed nodes, distributed data, and communications between the nodes; authentication level where the user handles encryption / decryption techniques, authentication methods such as contract administrative rights, authentication of applications and nodes, and logging entry; the data level which is concerned with data integrity and availability as well as data protection and data distribution. Cloud computing follows the policy of shared resources, where the privacy of data is very important because it faces some challenges like integrity, authorized access, and availability of (backup / replication). Data integrity ensures that data is not corrupted or tampered with during communication. Authorized access prevents data from infiltration attacks while backups and replicas allow access to data efficiently even in case of technical error or disaster in some cloud location .

Big data face some challenges as they can be classified into groups: data sets, processing and management challenges. When dealing with big amounts of data we face challenges such as volume, variety, velocity and verification which are also known as 5V of big data . Also, in the field of computer networks the

cost of communications is a major concern compared to the cost of processing the same data, as the challenge is to reduce the cost of communications to the minimum while meeting the requirements of storage and additional data from the general cloud to handle big data . Among the factors and challenges that affect the processing of big data in a timely manner is The bandwidth and latency . Where several challenges can be summarized in the relationship between big data and cloud computing.

i) Data Storage: The storage of big data through traditional storage is problematic because hard drives often fail, data protection mechanisms are not effective, and the speed of big data requires storage systems in order to expand rapidly, which is difficult to achieve with conventional storage systems. Cloud storage services offer almost unlimited storage with a great deal of error tolerance, which offers potential solutions to address the challenges of big data storage.

ii) Variety of data: Big data naturally grow, increase and vary, which is the result of the growth of almost unlimited sources of data. This growth leads to the heterogeneous nature of big data. Generally speaking, data from multiple sources of different types and representations are highly interrelated. They have incompatible shapes and are inconsistent. A user can store data in structured, semi-structured, or unstructured formats. Structured data format is suitable for today's database systems, while semi- structured data formats are only fairly suitable. Unstructured data is inappropriate because it contains a complex format that is difficult to represent in rows and columns.

iii) Data transfer: The data goes through several stages: data collection, input, processing, and output. Big data transfer is a challenge, so data compression techniques need to be reduced to reduce the volume, where data volume is a hindrance to transfer speed. It also affects the cost, while cloud computing provides distributed storage resources and data transfer on high-speed lines, reducing costs through virtual resources and resource use at user's request.

iv) Privacy and data ownership: The cloud environment is an open environment and the user's role in monitoring is limited. Privacy and security are an important challenge for big data. Big data and cloud computing come together in practice. According to (IDC) estimates, by 2020, around 40% of global data will be accessed by cloud computing. Cloud computing provides strong storage, calculation and distribution capability to support big data processing. As such, there is a strong demand to investigate the privacy of information and security challenges in both cloud computing and big data.

VI. BIG DATA TECHNICAL ISSUES AND CHALLENGES

A. *Fault Tolerance*: With the advent of technologies like cloud computing the aim must remain such that whenever failure occurs the damage done must occur within acceptable threshold rather than the entire work requiring to be redone. Fault-tolerant computing is tedious and requires extremely complex algorithms. A foolproof, cent percent reliable fault tolerant machine or software is simply a far-fetched idea. To reduce the probability of failure to an acceptable level we can do.

B. *Divide the entire computation to be done into tasks* and assign these tasks to different nodes for computation.

C. *Keep a node as a supervising node* and look over all the other assigned nodes as to whether they are working properly or not. If a glitch occurs the particular task is restarted. There are however certain scenario where the entire computation can't be divided into separate tasks as a task can be recursive in nature and requires the output of the previous computation to find the present result. These tasks can't be restated in case of an error. Here checkpoints are applied to keep the state of the system at certain intervals of time so that computation can restart from the last checkpoint so recorded.

D. *Data Heterogeneity***:** 80% of data in today's world are unstructured data. It encompassed almost every kind of data we produce on a daily basis like social media interaction, document sharing, fax transfers, emails, messages and a lot more. Working with unstructured data is inconvenient and expensive too. Converting these to structured data is unfeasible as well.

E. *Data Quality***:** As has been mentioned earlier, storage of big data is very expensive and there is always a tiff between business leaders and IT professionals regarding the amount of data the company or the organization is storing. The quality of data is an important factor to be looked into here. There is no point in storing very large data sets that are irrelevant as better result and conclusions can't be drawn from them. Ensuring whether the amount of data is enough for a particular conclusion to be drawn or whether the data is relevant at all are further queries.

F. *Scalability***:** The challenge in scalability of big data has led to cloud computing. It is capable of aggregating multiple different workloads with different performance goals into very large clusters. This needs high level of sharing of resources that is quite expensive and brings along with it various challenges like executing various jobs so that the goal of every workload is met successfully. It also has to deal with system failures in an efficient manner as it is quite common when working with large clusters. Hard disk drives being replaced by solid state drives and phase change technology do not have the same performance between sequential and random data transfer. The kind of storage device to be used is thus a large question looming around big data storage issue.

VII. BIG DATA PROCESSING ISSUES AND CHALLENGES

Effective processing of big data requires immense parallel processing and new analytics algorithms so as to provide rapid information. Often it may be unknown how to deal with a very large and varied volume of data and whether all of it needs to be analyzed. Challenges also include finding out data points that are really of importance and how to utilize the data to extract maximum benefit from it.

VIII. BIG DATA PRIVACY AND SECURITY ISSUES AND CHALLENGES

Often in big data analysis, the personal information of people from a database or from social networking sites need to be combined with external large data sets. Thus facts about anyone which might have been confidential become open to the world. Often it leads to taking insights in people's lives of which they are unaware of. Often it happens that a more educated person having better knowledge and concepts about big data analysis takes advantage of predictive analysis over a person who is less educated than him.

IX. CONCLUSION

Big data and cloud computing have been studied from several important aspects, and we have concluded that the relationship between them is complementary. Big data and cloud computing constitute an integrated model in the world of distributed network technology. The development of big data and their requirements is a factor that motivates service providers in the cloud for continuous development.

Cloud computing represents an environment of flexible distributed resources that uses high techniques in the processing and management of data and yet reduces the cost. All these characteristics show that cloud computing has an integrated relationship with big data. Both are moving towards rapid progress to keep pace with progress in technology requirements and users.

To handle big data and to work with it and obtaining benefits from it a branch of science has come up and is evolving, called Data Science. Data Science is the branch of science that deals with discovering knowledge from huge sets of data, mostly unstructured and semi structured, by virtue of data inference and exploration. It's a revolution that's changing the world and finds application across various industries like finance, retail, healthcare, manufacturing, sports and communication. As far as security is concerned the existing technologies are promising to evolve as newer vulnerabilities to big data arise and the need for securing them increases.

ACKNOWLEDGEMENT

I would like to thank Vidya Bharati Mahavidyalaya for providing full time internet service to do this research and development, for providing library facility also. I thank with high gratitude to my beloved colleagues for providing guidance to prepare this paper.

REFERENCES

[1]. Neves, Pedro Caldeira, Bradley Schmerl, Jorge Bernardino, and Javier Cámara. *"Big Data in Cloud Computing: features and issues."*

[2]. *https://en.wikipedia.org/wiki/Cloud_computing*

[3]. McAfee, Andrew, and Erik Brynjolfsson. *"Big data: the management revolution."* Harvard business review 90.10 (2012): 60-68.

[4]. *Challenges and Security Issues in Big Data Analysis*. Reena Singh. Kunver Arif Ali. IJIRSET. Volume: 5. Issue: 1. January 2016.

[5]. *Security Issues Associated With Big Data in Cloud Computing*. K. Iswarya Assistant Professor, Department of Computer Science. Idhaya College for Women Kumbakonam India. SSRG International Journal of Computer Science and Engineering (SSRG - IJCSE) – volume:1 issue 8 October 2014