PSN COLLEGE OF ENGINEERING & TECHNOLOGY

(An Autonomous Institution)
Melathediyoor, Tirunelveli - 627 152.

International Virtual Conference on Computational Engineering (IVCCE- 2022)

Certificate of Appreciation

This is to certify that Ansaba R B of Saintgits College of Applied Sciences, Kottayam, Kerala,India has presented a paper entitled on Big Data in Cloud Computing Environment in the International Virtual Conference on Computational Engineering (IVCCE - 2022) on 18/03/2022 to 19/03/2022.

Convener
Dr.M.Vignesh kumar
Associate Professor

Executive Director
Dr.P.Selva kumar
Professor

Principal
Dr.V.Manikandan
Professor

INTERNATIONAL JOURNAL OF
RECENT TRENDS IN MULTIDISCIPLINARY RESEARCH
May- June 2022, Vol 2(03), 01-05.
ISSN No: 2583-0368

Big Data in Cloud Computing Environment

Ansaba R B [1], Ambily Merlin Kuruvilla [2]
[1] *BCA student, Saintgits College of Applied Sciences, Pathamuttom, Kottayam, Kerala, India*
[2] *Assistant Professor & HOD, Department of Computer Applications, Saintgits College of Applied Sciences, Kottayam, Kerala, India*

Article Type: Research

OPEN ACCESS
Article Citation:
Ansaba R B[1],Ambily Merlin Kuruvilla [2], "Big Data in Cloud Computing Environment", International Journal of Recent Trends In Multidisciplinary Research, March-April 2022, Vol 2(03), 01-05.

Accepted date: May 12, 2022

Published date : May 14, 2022

Published by 5th Dimension Research Publication.

Abstract
In today's world, Big Data is an important area that is used in decision making and it processes huge volumes of data to address some query or pattern. Data is analysed through a set of algorithms, which differs depending upon the type of data, business's aim behind the analysis, and also other factors. But bigdata possess many challenges in terms of storing and processing data. Hence cloud computing which is another emerging technology is integrated with big data which provides better infrastructure for processing, storage for enormous data, and networking services.

Key Words: Big Data; Cloud Computing; Hadoop; Hdfs; Map Reduce

1. Introduction
Single Cloud computing is a powerful model and infrastructure that is distributed across the internet which process, manage and store the data. Cloud computing offers services for enterprise applications which centralizes both data storage and perform huge scale complex computing. It can reduce maintenance cost, provide less infrastructure and accelerate automation. [1]
Cloud services enables big data to analyse, manage and process the stored data in a more efficient manner. Through virtualization process integration of big data with cloud is the being achieved. Virtualization denotes the usage and sharing of resources independent of underlying hardware. Microsoft's Cloud Hadoop includes Azure Marketplace which comprise MapR and Azure Data Lake, which comprise Data Lake Store, Azure HDInsight, Data Lake Analytics as Azure cloud services. AWS includes versions of Hadoop, Spark, and Presto which operate on the data stored in Amazon Glacier and S3. Google's managed Hadoop include Cloud Dataproc and Spark cluster which uses GCP cloud services such as Big Query and Bigtable. [2]
Cloud platform provide rich productivity suites for database, data warehouse, collaboration, business intelligence, OLAP, and development tools. Big Data processing has many challenges relating with Data collection, analysis, sharing, research and visualization. Each of these processes need different techniques, infrastructure, and highly skilled professionals. Also, it cannot be done easily with traditional programs because of resource restrictions such as computing power and time, hence we need advanced algorithms and vast databases. And all these difficulties and barriers are much reduced as a result of integrating Big Data within cloud environment. [3]
Big data represents huge amounts of complex data which can be either unstructured or structured generated by multiple sources. The traditional relational databases are not sufficient to process and analyse data from multiple-sources, such as managing data related with record of transactions, customer behaviour, mobile phone and GPS navigation, etc. So, to deal with these kinds of complex data, cloud is employed, which serve as the storehouse where the processed outcome/data will be stored. Cloud computing approach is efficient because of having advanced technologies to handle the vast amount of data. This paper discusses an overall view of cloud computing and big data, their features, Relation and integration of big data & cloud, some **big data management tools in cloud**.

2. Cloud Computing
Cloud computing is a type of service-oriented computing where software and hardware are delivered as a service over the internet. Cloud is a combination of distributed and centralized system which includes virtualized servers, operating systems, applications, etc that are dynamically supplied. It provides services relating to storage, processing and sharing of data through visualized resources over the networks. Cloud platform is completely virtual to its users and require less effort from user to operate and manage its services. Important features associated with cloud includes scalability, on-demand delivery of resources, easy accessibility, cost-effective, flexibility and reliability. [4]

It has another major feature, Pay-as-you-use which means that users have to pay only for what they need at any given time. Advantages of cloud computing include:

- Data security
- Virtualized resources
- Easy and agile development
- Less maintenance cost
- Scalable data storage.
- Services in the pay-per-use model.

Cloud Service Models:

Service-oriented architecture of cloud supports "everything as a service" and hence offers their services as different models which are: [5]

Fig 1: cloud service models

- ***Platform as a Service (PaaS)****:*

In this service model, platform level elements such as project management environments, scalable and elastic runtime environment are provided. User can configure and install required software on the cloud. In short, PaaS provides the framework needed to build, deploy, test and manage software resources.

- ***Software as a Service (SaaS):***

It is a software distribution model where cloud consumers on internet can retrieve software applications and databases that are hosted by cloud service provider. If user does not have specific software or associated compatible hardware not installed on local computer, he/she can access directly from cloud. [6]

- ***Infrastructure as a Service (IaaS):***

Computing resources are provided to consumers by the IaaS in the form of infrastructure like virtual machines, servers, operating systems, network, hardware resources and storage on demand across internet . It provides completely virtualized computing infrastructure and provides an environment to deploy and run infrastructure including hardware and software in cloud environment.

Types Of Cloud:

Before transferring a business system into cloud, there is a need to consider many factors. There are four contrasting types of Clouds and three among them are basic types and hence mostly used. [7]

- ***Public Cloud:***

This type of cloud is available to the general public. General uses of public clouds include file-sharing, online office applications, application development and testing and web-based email. Public cloud infrastructure services are provided over the internet and hence open for everyone. Through public cloud, customers and users can easily access systems and shared

resources with low cost and high efficiency. Some examples of public cloud are Microsoft Azure, Amazon Elastic Compute Cloud(EC2), Google Cloud , Alibaba Cloud, Oracle Cloud Fast Connect .

- ***Private Cloud:***

It is also called "internal cloud" or personal data center computing.

It is deployed on a private network and are meant for the unique use of a particular company. This model provides highest level of security and data privacy as it permits only authorized users. They are more expensive than public cloud. Through this model, it is only able to access systems and services within an institution or an organization.

- ***Hybrid Clouds:***

It combines and integrates both private cloud and public cloud. Hence it allows cost-effective way for businesses to increase compute capacity on demand and better flexibility in terms of data transfer. Users or customers can develop and deploy applications using public cloud and at the same time offers higher degree of security through private cloud rather than using only a public cloud.

3. Big data

Big data refers to large amounts of data or compound datasets produced by various sources like sensors, mobile devices, social media and from three primary sources: machine data, social data and transactional data, in a very short duration of time. Such data are too large, fast growing and are difficult or impossible to process using traditional methods or conventional tools and techniques. Through deep analysis and efficient processing by various Data Analytics methods, valuable information can be extracted from big data. [8]

Characteristics: Big data is characterized mainly by five Vs which are: Volume, Veracity, Variety, Value and Velocity.

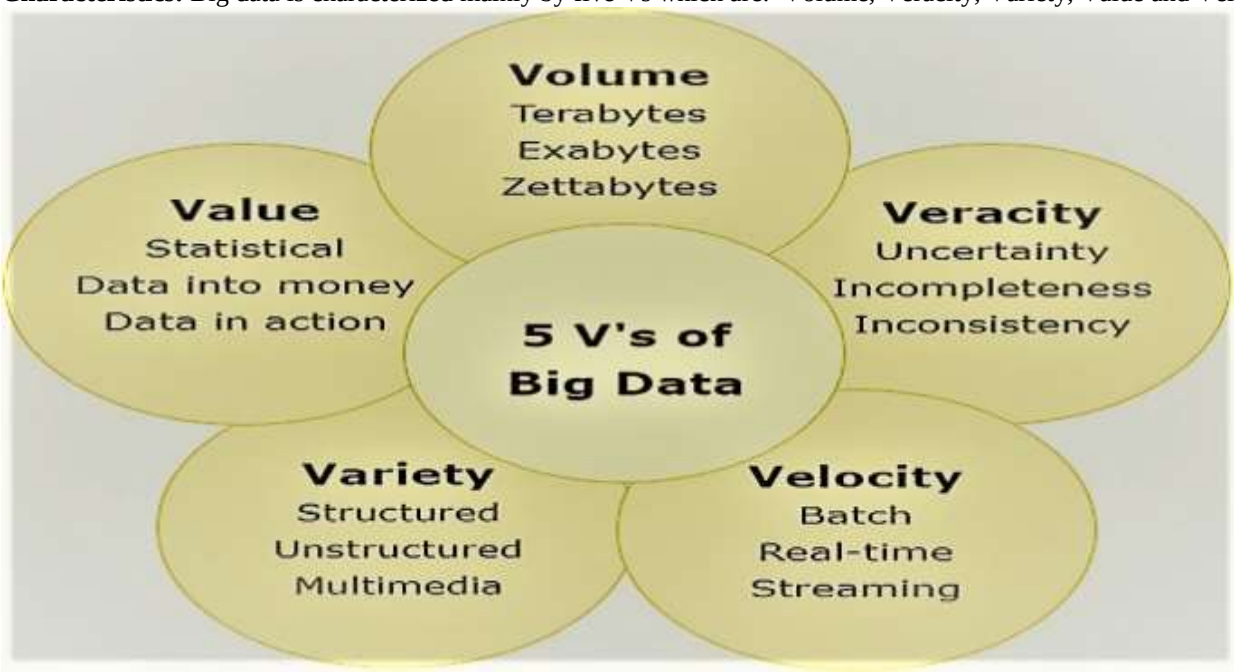

Fig 2: characteristics of big data

Volume:

It denotes incredible amount of data which are generated and stored in Gigabytes (GB), Zettabytes (ZB), and Yottabytes (YB). In coming years, the volume will rise significantly as data is being created every second from various sources like social media platforms, smart (IoT) devices, networks, machines and so on. [9]

Variety:

It refers to various kinds of data gathered from different sources. Data generated can be of different formats which can be structured, unstructured, semi structured or a mix of all these three. It can include different forms of data such as financials, logs files, social media updates, images, videos, text messages, audio, etc.

Veracity:

It denotes "compliance with truth or fact" and refers to overall quality and reliability of the data source. Low veracity can negatively affect the accuracy of the results.

Value:

It denotes the final value obtained after processing of data and produced during analysis which helps in decision making. To obtain value firstly mine data which refers to the process of conversion of raw data into useful data. Next on this retrieved data, analysis is done.

Velocity:

It denotes the speed at which data is being created, generated, collected and analysed. Velocity also associated to how fast big data is going to be processed.

Examples of data generated with high velocity include Facebook posts, data from sensors and mobile devices.

Big Data Advantages:

- Real-time monitoring of product price optimization, business and market
- Greater innovations and lifesaving applications in the healthcare industry and public health with availability of record of patients.
- Real time communication regarding customer requests, their queries and problems.
- Helps in quicker and better decision making

Challenges of Big Data:

- Difficult to Manage large volumes of data as there is always a lot of raw data to store and analyse.
- Lack of workers with adequate big data skills and talent.
- There is a chance to make wrong decisions due to unevenness of data quality and it is difficult to determine which source of data is correct.
- Poor data scalability, reliability and runtime quality issues.[10]

4. Integration of Bigdata in Cloud

Cloud platform provides one of the best environments for efficient bigdata processing and real time analysis in a cost-effective way. It has greatly improved Big Data analysis, resulting in better findings and hence decision making. Cloud environment provide services to analyse and process bigdata by breaking huge volume of information into smaller units and each of them can be processed independently in different servers. Through remote multi servers and dynamic parallel resource allocation, it is possible to handle massive amount of data accordingly in cloud environment. Integration with cloud make big data resources more monitored, productive, compliant and simpler. [11]

Cloud providers like Google Cloud Platform, Amazon Web Services, Microsoft Azure, IBM, Oracle, Salesforce, etc. provide important factor: scalability which is required for bigdata handling and processing.Another important factor is the data security and privacy which cloud platform offers. It provides more scalable and elastic Private Cloud Solution thereby a safe environment to keep big data and its computation. To store data on the cloud, a key is given to all its users and data can be accessed only by using that key.Cloud error localization is a technique which is used to identify and monitor error in big data storage and also handles bad performance of server. [12]

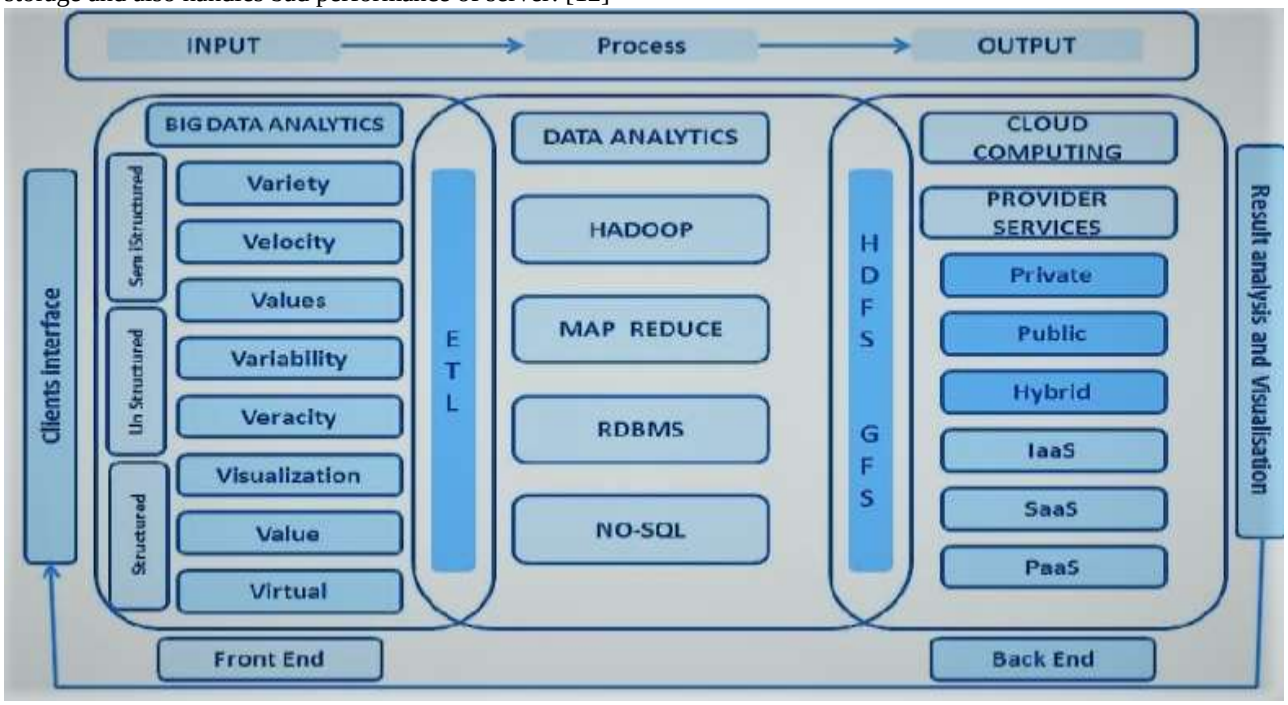

Fig 3: Relation between cloud and big data

Big data management tools in cloud:

• **Hadoop**:

Hadoop is a part of Apache project and it is a freely available java-based programming framework. Hadoop enables processing of large sets of data on a cluster of servers and applications consisting terabytes of data. So, even if some node fails, Hadoop supports with rapid transfer rates. Hadoop consists of higher-level declarative languages for big data analysis pipelines and query writing. Hadoop mainly composed of HDFS and MapReduce.

• **HDFS**

HDFS is a file system used to store or span all the nodes in a Hadoop cluster for data storage. Thereby it improves reliability and support security. HDFS usually splits files into blocks which in turn is stored on the server. Thereby it maintains reliability by duplicating data across multiple hosts combining parallel processing technique. [13]

- **MapReduce**

This is a framework which helps in writing applications that process and generates large datasets on a cluster with parallel or distributed algorithm. At first, breaking Big Data into small subunits takes place which in turn are analysed and processed by Map jobs in parallel. Map () method consists of acquiring, filtering & categorizing datasets. Reduce () method consists of final result generation and locating associated summaries. [14]

- **NoSQL**

NoSQL (Not Only SQL) systems provides systematic way to store and replicate data, giving out retrieval and appending operations from the data. These databases are not bound by the confines of a fixed schema model instead each are deployed as a cluster of nodes. Examples of NoSQL systems include Amazon DynamoDB, Azure Cosmos DB, MongoDB, Cassandra, CouchDB, and HBase.

5. Conclusion

This paper presented how cloud computing helps in analysing, storing and processing big data. Big data and cloud together comprise an integrated model of distributed network technology.Cloud supports big data in terms of security of data, encryption, data integrity, data transformation, data heterogeneity, data quality and others.
Even though there are challenges regarding integration with cloud such as scalability, availability and problems with bandwidth for data transfer, Solutions are constantly being developed by cloud providers for the efficient use of big data on cloud. So, the integration and application of big data in cloud will have a huge impact and continue to grow in the following years.

References

[1] Beri, R. & Behal, V. (2015). Cloud Computing: A Survey on Cloud Computing. Int. J. Comput. Appl., Vol. 111, pp. 19–22.
[2] Sheetal Singh, Vipin Kumar Rathi, Bhawna Chaudhary, "Big Data and Cloud Computing: Challeges and Opportunities",International Journal of Innovations in Engineering and technology, Vol. 5(4), August 2015.
[3] Gupta, H. & Mohania, M. (2012). Cloud computing and big data analytics: What is new from databases perspective? in Big Data Analytics. BDA 2012. Lecture Notes in Computer Science, Vol. 7678, pp. 42–61 (Springer Berlin Heidelberg).
[4] Fonseca, N., & Boutaba, R. (2015). Cloud services, networking, and management. John Wiley & Sons.
[5] K. Kaur, "A Review of Cloud Computing Service Models", International Journal of Computer Applications, Vol.140, No.7, pp.15-18, 2016.
[6] J. Srinivas, K.Venkata Subba Reddy and Dr. A. Moiz Qyser, "Cloud Computing Basics", International Journal of Advanced Research in Computer and Communication Engineering, Vol.1(5), 2012.
[7] Venters, W., Whitley, E.A.: A Critical Review of Cloud Computing: Researching Desires and Realities. J. Inf. Technol. 27, 179–197 (2012).
[8] D.P. Acharjya, Kauser Ahmed P, " A Survey on Big Data Analytics: Challenges, Open Research Issue and Tools", International Journal of Advanced Computer Science and Applications, Vol. 7(2), 2016
[9] N. Elgendy and A. Elragal, "Big Data Analytics: A Literature Review Paper," in Advances in Data Mining.Applications and Theoretical Aspects, 2014, pp. 214–227.
[10] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.
[11] Bautista Villalpando, L. E.; April, A. & Abran, A. (2014). Performance analysis model for big data applications in cloud computing, Vol. 3, pp. 1–20
[12] Agrawal, Divyakant & Das, Sudipto & Abbadi, Amr. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM International Conference Proceeding Series. 530-533. 10.1145/1951365.1951432
[13] K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.
[14] L. Zhao, Z. Zhou," Cloud computing model for big data processing and performance optimization of multimedia communication", Computer Communications (2020)