IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

BIG DATA AND CLOUD COMPUTING

[1] Rohan S Abraham, [2]Abhijith Suresh, [3] Gracen K Shaji, [4]Jeena Ann
[1] Student, [2] Student, [3] Student, [4] Student
[1]PG Department of Computer Applications and AI,
[1]Saintgits College of Applied Sciences, Kottayam, Kerala, India

Abstract: In today's world, Big Data is an important area that is used in decision making and it processes large volumes of data to address some query or pattern. These are the data generated by various sources that are difficult to analyze and processed by traditional technologies. Cloud computing, a rapidly growing technology, integrates with Big Data and provides a strong infrastructure for storage, processing, and network services for large amounts of data. With Big Data in Cloud, organizations can easily access and analyze large amounts of data. In addition, big data in cloud offer increased data security, automatic backup, and disaster recovery, making them an attractive option for businesses.

IndexTerms **- Big Data, Cloud Computing, Hadoop, HDFS, Map Reduce.**

1.INTRODUCTION

Big Data refers to a huge amount of compound data that can be either unstructured or structured, generated from various number of sources such as social media, e-commerce, and others. This datas can be analyzed to identify various patterns, correlations, and other insights which offer valuable pieces of information for certain needs. Traditional relational databases are not sufficient to analyze and process data from multiple-sources, such as managing data related to records of bank transactions, sports, social gaming, etc. So, to handle these kinds of complex data, the cloud is used. And all these difficulties and barriers are much reduced as a result of integrating Big Data within cloud environment [1].

Cloud computing provides an affordable infrastructure to store, manage and process Big Data. Companies can lease computing resources from various cloud providers, eliminating the need for expensive hardware and software investments. Cloud computing also provides flexible and on-demand access to resources, making it easier for certain organizations to store, analyze, and process huge amounts of data. Through the virtualization process integration of big data with cloud is achieved. Virtualization denotes the usage and sharing of resources independent of the underlying hardware.

Big Data in the cloud is revolutionizing the way businesses operate, providing a secure, scalable, and cost-effective platform for data storage and processing. This enables organizations to extract valuable insights from their data and make decisions. With the growth of Big Data, cloud computing is driving innovation in various industries.

2.BIG DATA

Big Data refers to vast and complex sets of information generated from various sources like social media, sensors, mobile devices, and from three primary sources: machine data, social data, and transactional data, in a very short duration of time. The large volume and diverse nature of this data make it difficult to process and analyze using traditional data management tools. Big Data analytics uses innovative methods to gain valuable insights and knowledge from the data, allowing organizations to make decisions and enhance their growth. Through deep analysis and efficient processing by various Data Analytics methods, valuable information can be extracted from big data [2]. The key characteristics of Big Data are Volume, Veracity, Variety, Value, and Velocity.

2.1. Characteristics

The main characteristics of Big Data are: (Figure 1)

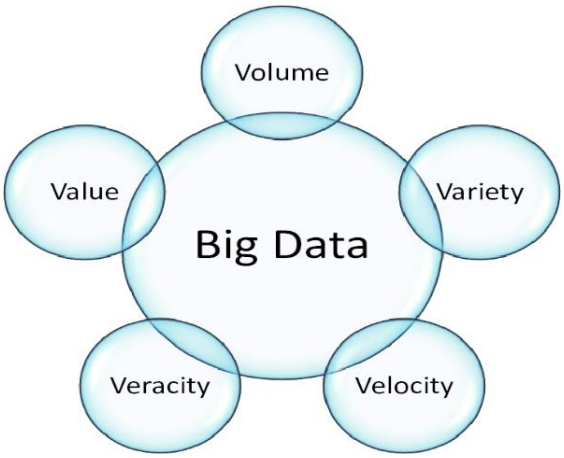

Figure 1: characteristics of big data

- Volume (Size of the Data)
- Variety (Data is Presented in Different Data Formats)
- Velocity (Speed Where Data is Produced)
- Veracity (Data Quality)
- Value (Knowledge gained from the data)

2.1.1.Volume :

The enormous quantity of data produced per second, ranges from terabytes to zettabytes. In other words ,Volume refers to the magnitude of data. Big data sizes are reported in multiple terabytes and petabytes [3]. 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 [4].

2.1.2.Variety :

It refers to various kinds of data gathered from different sources. In other words, it represents the multiplication of the types of data managed by an information system. This multiplication leads to a complexity of links and link types between these data. The variety also relates to the possible uses associated with raw data [5].

2.1.3.Velocity :

It represents the speed of data creation and the urgency in which it needs to be analyzed. This characteristic of big data puts demands on the data storage and management systems to be able to handle high-speed processing and retrieval.

2.1.4.Veracity :

It is a measure of the trustworthiness of data and its ability to provide reliable and meaningful insights. Veracity is a complex issue and can be impacted by various factors, such as data sources, data collection methods, and data processing techniques.

2.1.5.Value :

Value can be considered a critical aspect of big data as it is the end goal of collecting, processing, and analyzing big data. Value refers to the insights and benefits that organizations can derive from big data.

2.2.Advantages

Big data has several advantages, some of them are:

- Decision making: Helps in quicker and better decision making.
- Increased operational efficiency: By analyzing big data, organizations can improve their operations.
- Customer insights: Big data can be used to gather information about customer behavior and preferences, allowing organizations to make their products and services to customer needs.
- Fraud detection: Big data can be used to identify fraud and financial crime, enabling organizations to prevent fraud and reduce losses.
- Cost savings: By optimizing operations and reducing inefficiencies, big data can help organizations save money.

2.3.Challenges

- Storing and managing large amounts of data is difficult and expensive, it requires new technologies and infrastructure.
- In the case of unstructured data, ensuring the quality, accuracy, and completeness can be difficult.
- Protecting sensitive and confidential information contained in big data is a major concern, and organizations need to implement security measures to prevent unauthorized access and breaches.
- With the increasing amount of personal data being collected, there is a growing concern about privacy and the need to properly manage and protect personal information.
- Analyzing big data can be complex and time-consuming, requiring specialized skills and tools.
- Integrating big data into existing systems and processes can be challenging, and requires new technologies.

- The field of big data is still evolving, and there is a shortage of personnel with the necessary skills to effectively manage big data.

3.CLOUD COMPUTING

Cloud computing refers to a model that provides access to computing resources and services through the internet, using a pay-as-you-use approach. Instead of having to invest in hardware and software, users can access a variety of resources from different third-party vendors using a web browser or mobile app. These resources, including data and applications, are stored on remote servers that can be accessed from anywhere in the world with an internet connection.

The advantages of cloud computing over traditional computing include lower costs, scalability, and greater flexibility. Cloud computing can also boost operational efficiency and improve productivity. Cloud service providers have begun to integrate frameworks for parallel data processing in their services to help users access cloud resources and deploy their programs [6]. Additionally, it can enhance data security by storing information on remote servers, reducing the risk of loss or theft. However, there are some drawbacks to cloud computing, such as security concerns and the potential for vendor lock-in. Organizations must evaluate these potential risks against the benefits and choose a vendor that meets their requirements.

Overall, cloud computing is a transformative technology that provides organizations with access to a wide range of computing resources and services, making it easier and more cost-effective to operate their businesses. Careful consideration of both the benefits and risks is necessary when making decisions about using cloud computing.

3.1.Advantages

The advantages of cloud computing include:

- Lower costs: Cloud computing eliminates the requirement of investments in hardware, software, and the expenses related to maintenance and upgrades.
- Flexibility: Organizations can easily scale their use of cloud services to meet changing demands.
- Convenient access: Cloud resources can be accessed from any location with an internet connection, enabling remote work.
- Improved teamwork: Teams can work together more productively by sharing resources in real-time through the cloud.
- Enhanced security measures: Cloud providers usually invest in security, making data stored in the cloud more secure than on local servers.
- Increased efficiency: Cloud computing can simplify processes, minimize downtime, and automate routine tasks, leading to improved efficiency.
- Disaster recovery: Remote storage of data in the cloud decreases the risk of loss in case of a disaster.
- Environmental advantages: By reducing the need for physical servers, cloud computing can help decrease energy consumption and carbon emissions.

3.2. Cloud Service Models

There are three models of service in cloud which are software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS), Figure 2 : shows the three service models.

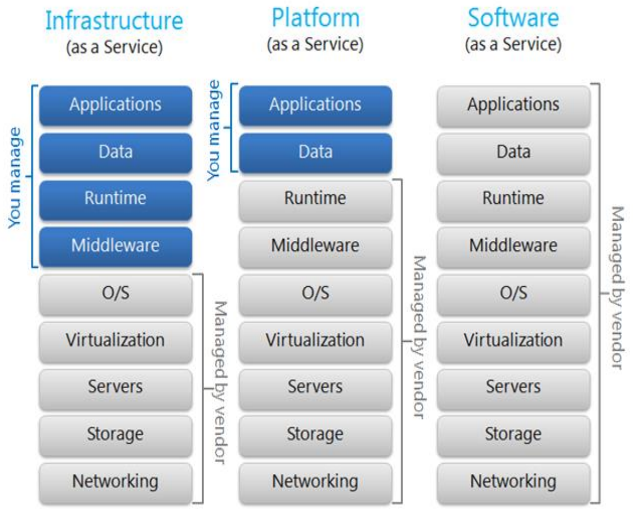

Figure 2 : cloud service models

3.2.1. *Software as a Service (SaaS)* :

In this model a third-party provider hosts software applications and makes them available to customers over a high-speed internet connection. This eliminates the need for customers to have their own installations and reduces the costs and difficulties associated with traditional software deployment and upkeep.

3.2.2. Platform as a Service (PaaS)

This model provides platform-level components such as project management tools and scalable, flexible runtime environments for software deployment. Users can configure and install the necessary software on the cloud and have access to the framework required for building, deploying, testing, and managing software resources. In essence, PaaS delivers the necessary infrastructure for software development and delivery.

3.2.3 Infrastructure as a Service (IaaS)

IaaS is the most comprehensive and flexible type of cloud service that offers a completely virtualized computing infrastructure managed over the internet. The IaaS provider handles the physical aspect of the infrastructure such as servers and data storage, while customers have full control to customize the virtualized resources according to their requirements.

3.3. Types of Cloud

There are four cloud deployment models are there, Private Cloud, Public Cloud, Hybrid Cloud, and Community Cloud.

3.3.1. Public Cloud

A public cloud is a deployment type of cloud computing where resources, are made accessible to the general public via the internet by a third-party service provider. This provider manages the infrastructure and provides technical support, while the customer can scale their usage as needed and pay only for what they use. Public clouds are known for their scalability and cost-effectiveness, eliminating the need for investment in hardware or maintenance.

3.3.2. Private Cloud

Private cloud is private to a specific organization/customer. A private cloud resources and services are used exclusively by one business or organization. Unlike public cloud, a private cloud resources are not shared by multiple organizations. All the hardware infrastructure and software are solely dedicated to one organization. The private cloud is physically located on premise that is at our organizations on-site data center or it can also be hosted by a third party service provider.

3.3.3. Hybrid Cloud

It is a combination of both private cloud and public cloud. It allows cost-effective way for businesses to increase compute capacity on demand and better flexibility in terms of data transfer. 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.

3.3.4. Community Cloud

It is a cloud service that provides services to a community of users or organizations with shared interests or concerns. In other words, here the cloud resources are shared by an organization which is of common interest for every participant which is being part of a community, whose needs are similar. Organizations using this cloud service have shared missions, security requirements, and policies.

4. INTEGRATION OF BIG DATA IN CLOUD

The cloud offers a cost-effective and efficient way to process and analyze Big Data in real-time. This environment has greatly improved Big Data analysis, leading to better results and decision-making. The cloud allows for analyzing and processing Big Data by breaking down large amounts of information into smaller pieces and processing each piece separately on different servers. With remote multi-servers and dynamic parallel resource allocation, the cloud can effectively handle large quantities of data. Integration with cloud make big data resources more monitored, productive, compliant and simpler [7].

Big Data requires a cluster of servers to manage the processing of large amounts of data with various formats. Cloud computing offers this service in an affordable way, using a cluster of servers, storage and networking resources that can be scaled up or down as needed. By using cloud computing, a single server can serve multiple customers for accessing and updating their data without the need for individual applications. The cost-effectiveness of cloud computing makes it an ideal solution for managing Big Data.

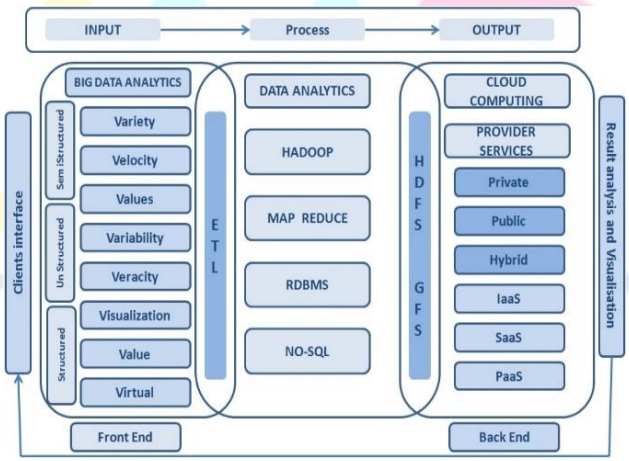

Figure 3 : Relation between cloud and big data

Cloud service providers, such as Google Cloud Platform, Amazon Web Services, Microsoft Azure, IBM, Oracle, Salesforce, etc., offer important features for Big Data handling and processing, including scalability. Additionally, they provide enhanced data security and privacy. The cloud platform offers a safe and scalable private cloud solution for the storage of Big Data and its computations. Access to the data is granted through a key and can only be retrieved with the proper 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 [8]. Big Data integration with Cloud Computing can have challenges, including:

- Data Collection: Getting large amounts of data into the cloud can be difficult as data size grows rapidly.

- Storage: Storing large amounts of growing data requires a scalable and distributed data storage system in the cloud.
- Analysis: Analyzing big data can reveal valuable information but existing tools and techniques need to be improved to handle large datasets.
- Security and Privacy: Protecting personal information and sensitive data is a major concern when accessing and analyzing big data in the cloud. Strong encryption techniques are needed.

5.BIG DATA MANAGEMENT TOOLS IN CLOUD

5.1. Hadoop

Hadoop is an open-source software framework that is used for distributed storage and processing of large data sets. It is written in Java and includes higher-level languages for writing queries and data analysis pipelines. Hadoop is popular among organizations for managing and analyzing unstructured logs and events, with its two most utilized components being Hadoop Distributed File System (HDFS) and MapReduce. It has become a popular choice for many organizations to analyze big data and make informed decisions.

5.2. HDFS

Hadoop Distributed File System (HDFS) is a distributed file system that is designed to store large amounts of data across multiple nodes in a cluster. Thereby it maintains reliability by duplicating data across multiple hosts combining parallel processing technique [9]. HDFS provides a reliable and scalable storage solution for big data applications and is an integral part of the Apache Hadoop framework. It works by breaking data into smaller blocks and storing multiple copies of these blocks across different nodes, ensuring that data is available even in case of node failure.

5.3. MapReduce

The MapReduce system is a key component of the Hadoop framework that is used for processing large datasets on a cluster. It uses a parallel algorithm to divide the work among various independent nodes. The MapReduce programming paradigm involves two stages. The Map () method consists of acquiring, filtering & categorizing datasets. The Reduce() method, which performs a summary operation on the data.The MapReduce system manages the distribution of the data across servers, coordinates all communications, and handles parallel data transfers.

5.4. NoSQL

NoSQL (Not Only SQL) provides systematic way to replicate and store 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 popular NoSQL databases include MongoDB, Cassandra, and CouchDB.

6.CONCLUSION

This paper presents how cloud computing helps in analyzing, processing, and storing big data. Big data and the 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.

However, there are some challenges regarding integration with cloud such as scalability, availability, and problems with bandwidth for data transfer, the solutions for this 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] Gupta, R., Gupta, H. and Mohania, M. (2012) "Cloud computing and big data analytics: What is new from databases perspective?," Big Data Analytics, pp. 42–61. Available at: https://doi.org/10.1007/978-3-642-35542-4_5.

[2] D.P. Acharjya, Kauser Ahmed P (2016) "A survey on Big Data Analytics: Challenges, open research issues and Tools," International Journal of Advanced Computer Science and Applications, 7(2). Available at: https://doi.org/10.14569/ijacsa.2016.070267.

[3] Gandomi, A. and Haider, M. (2015) "Beyond the hype: Big Data Concepts, methods, and analytics," International Journal of Information Management, 35(2), pp. 137–144. Available at: https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

[4] Elgendy, N. and Elragal, A. (2014) "Big Data Analytics: A literature review paper," Advances in Data Mining. Applications and Theoretical Aspects, pp. 214–227. Available at: https://doi.org/10.1007/978-3-319-08976-8_16.

[5] Riahi, Y. and Riahi, S. (2018) "Big Data and big data analytics: Concepts, types and technologies," International Journal of Research and Engineering, 5(9), pp. 524–528. Available at: https://doi.org/10.21276/ijre.2018.5.9.5.

[6] Warneke, D. and Kao, O. (2009) "Nephele," Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers [Preprint]. Available at: https://doi.org/10.1145/1646468.1646476.

[7] Bautista Villalpando, L.E., April, A. and Abran, A. (2014) "Performance analysis model for big data applications in cloud computing," Journal of Cloud Computing, 3(1). Available at: https://doi.org/10.1186/s13677-014-0019-z.

[8] Agrawal, D., Das, S. and El Abbadi, A. (2011) "Big Data and cloud computing," *Proceedings of the 14th International Conference on Extending Database Technology* [Preprint]. Available at: https://doi.org/10.1145/1951365.1951432.

[9] Kala Karun, A. and Chitharanjan, K. (2013) "A review on Hadoop — HDFS infrastructure extensions," *2013 IEEE CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGIES* [Preprint]. Available at: https://doi.org/10.1109/cict.2013.6558077.