International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Big Data with Cloud Computing: Discussions and Challenges

Pallavi Sunil shewale, Paurnima Kawale, Arunadevi Khaple
Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra, India

Article Info
Volume9, Issue 2
Page Number: 470-478

Publication Issue
March-April-2022

Article History
Accepted :03March2022
Published :10March2022

ABSTRACT

With the recent advancements in computer technologies, the amount of data available is increasing day by day. However, excessive amounts of data create great challenges for users. Meanwhile, cloud computing services provide a powerful environment to store large volumes of data. They eliminate various requirements, such as dedicated space and maintenance of expensive computer hardware and software. Handling big data is a time-consuming task that requires large computational clusters to ensure successful data storage and processing. [1] Security and Privacy of information is the biggest challenge to cloud computing. Security and privacy issues can be overcome by employing encryption, security hardware and security applications.[2]

In this work, the definition, classification, and characteristics of big data are discussed, along with various cloud services, such as Microsoft Azure, Google Cloud, Amazon Web Services, International Business Machine cloud, Hortonworks, and MapR. A comparative analysis of various cloud-based big data frameworks is also performed.

Various research challenges are defined in terms of distributed database storage, data security, heterogeneity, and data visualization.[1]

Key words: big data; data analysis; cloud computing; Hadoop

I. INTRODUCTION

With recent technological advancements, the amount of data available is increasing day by day. For example, sensor networks and social networking sites generate overwhelming flows of data. In other words, big data are produced from multiple sources in different formats at very high speeds [1] At present, big data represent an important research area. Big data are rapidly produced and are thus difficult to store, process, or manage using traditional software. Big data technologies are tools that are capable of storing meaningful information in different types of formats. For the purpose of meeting users' requirements and analyzing and storing complex data, a number of analytical frameworks have been made available to aid users in analyzing complex structured and unstructured data [3]. Several programs, models, technologies, hardware, and software have been proposed and designed to access the information from big data. The main objective of these technologies is to store reliable and accurate results for big data [4] In

addition, big data require state-of-the-art technology to efficiently store and process large amounts of data within a limited run time.

Three different types of big data platforms are interactive analysis tools, stream processing tools, and batch processing tools[4]. Interactive analysis tools are used to process data in interactive environments and interact with real-time data. Apache Drill and Google's Dremel are the frameworks for storing real-time data.

Stream processing tools are used to store information in continuous flow. The main platforms for storing streaming information are S4 and Strom. Hadoop infrastructure is utilized to store information in batches.

Big data techniques are involved in various disciplines, such as signal processing, statistics, visualization, social network analysis, neural networks, and data mining. Mohajer et al. designed an interactive gradient algorithm that receives controlled messages from neighboring nodes. The proposed method uses a self-optimization framework for big data.

II. DEFINITIONS OF BIG DATA

Big data refers to data sets that are too large or complex to be dealt with by traditional data processing application software. Data with many fields (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data analysis challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: volume, variety, and velocity.[The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and value [5].

2.1 Characteristics of big data

Big data are characterized by three Vs: volume, velocity, and variety. These characteristics were introduced by Gartner to define the various challenges in big data [1]

With new-generation architecture, data are now stored in different types of formats; hence, the three Vs may be extended to five Vs, namely, volume, velocity, variety, value, and veracity[1]

(1) Volume: Data are generated by multiple sources (sensors, social networks, smartphones, etc.) and are continuously expanding. The Internet produces global data in large increments. In 2012, approximately 2.5 exabytes (EB) of data were produced every day. According to the report of International Data Cooperation, the volume of data in 2013 doubled, reaching 4.4 zettabytes (ZB). In 2020, the volume of data reached 40 ZB. Table 2 shows the names of the units of data that can be measured in bytes[14].

(2) Velocity: Data are exponentially growing at high speeds. Millions of connected devices are added on a daily basis, thereby leading to increases in not only volume but also velocity[15, 16]. One relevant example is YouTube, which generates big data at high speeds[17, 18].

Table 3 presents the number of users in India who had used social media networks by February 2021. Figure 1

Table 1 Units of data.

Name of unit	Equals	Size in bytes
Bit	1 or 0	1/8
Nibble	4 bits	1/2
Byte	8 bits	1
Kilobyte (KB)	1024 bytes	210
Megabyte (MB)	1024 KB	220
Gigabyte (GB)	1024 MB	230
Terabyte (TB)	1024 GB	240
Petabyte (PB)	1024 TB	250
Exabyte (EB)	1024 PB	260

Zettabyte (ZB)	1024 EB	270
Yottabyte (YB)	1024 ZB	280

Table 2 Users in India as of February 2021.

Application name	Count	Application name	Count
WhatsApp	53 Crore	Instagram	21 Crore
YouTube	44.8 Crore	Twitter	1.75 Crore
Facebook	41 Crore		

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many multinational companies to process the data and business of many organizations. The data flow would exceed 150 exabytes per day before replication.

There are five v's of Big Data that explains [7] shows the five Vs of big data.

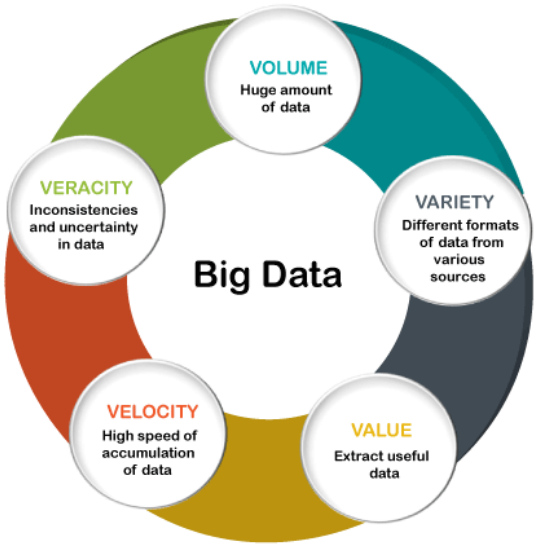

Fig 1: Five Vs of big data.

(3) Variety: Data are generated in multiple formats via social networks, smartphones, or sensors. These tools produce data in the form of data logs, images, videos, audio, documents, and text. Data may also be structured, semistructured, and unstructured[1]. Collected data can be unstructured, semi-structured or structured in nature. Unstructured data is data that is unorganized and comes in different files or formats. Typically, unstructured data is not a good fit for a mainstream relational database because it doesn't fit into conventional data models. Semi-structured data is data that has not been organized into a specialized repository but has associated information, such as metadata. This makes it easier to process than unstructured data. Structured data, meanwhile, is data that has been organized into a formatted repository. This means the data is made more addressable for effective data processing and analysis.[8]

(4) Value: Value is an important characteristic of big data. It relates to how data can be dealt with and converted into meaningful information[1]. The last V in the 5 V's of big data is value. This refers to the value that big data can provide, and it relates directly to what organizations can do with that collected data. Being able to pull value from big data is a requirement, as the value of big data increases significantly depending on the insights that can be gained from them.[8]

(5) Veracity: Veracity refers to the quality, correctness, and trustworthiness of data. Therefore, maintaining veracity in data is mandatory[1,2]. For example, data in huge amounts create confusion, whereas small amounts of data can convey incomplete or half information.[1]. Data can sometimes become messy and difficult to use. A large amount of data can cause more confusion than insights if it's incomplete. For example, concerning the medical field, if data about what drugs a patient is taking is incomplete, then the patient's life may be endangered.[8]

2.2 Types of big data

Data are produced at unprecedented rates from various sources, such as financial, government, health, and social networks. Such rapid growth of data can be attributed to smart devices, the Internet of Things, etc. In the last decades, companies have failed to store data efficiently and for long periods[1,2]. This drawback relates to traditional technologies that lack adequate storage capacity and are costly. Meanwhile, big data require new storage methods backed by

powerful technologies[7]. Big data can be classified into several

III. CLOUD COMPUTING

Cloud computing offers a cost-efficient and scalable solution to store big data. According to the National Institute for Standards and Technology, "Cloud Computing is based on pay-per-use services for enabling convenient, on-demand network access to a shared pool of configurable computing resources such as servers, networks, and services that can be rapidly provisioned and released with minimal management effort or service provider interaction". Cloud computing services can be[3] categories. Figure 2 depicts the classification of big data.

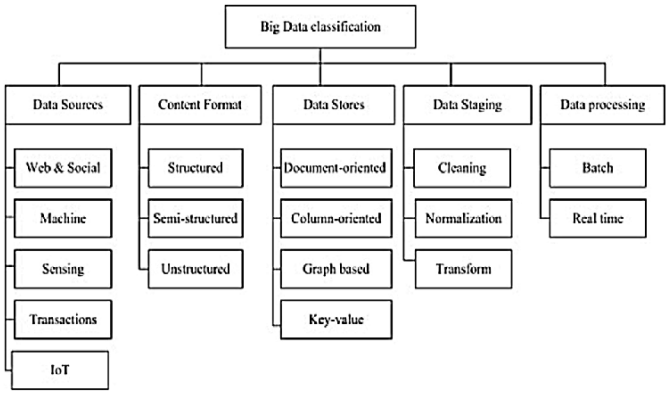

Fig 2: Types of big data.

Table 4 Types of big data.

Type	Category	explanation
	Social Media	Social media represents an important aspect of big data. Facebook, Twitter, emails, and microblogs are social media sources that generate massive amounts of data daily[27].
	Machine Genarated data	Software and hardware, such as medical devices, computers, and other types of machines that generate data without human interferences.
Data Source	Sensing	Various types of sensing devices that generate data and convert them into signals
	Transaction	Financial, business, and work data generate time-based dimensions that define data.
	IOT	Tablets, smartphones, and digital camera devices are connected over the Internet and thus generate huge amounts of data and information.
Content format source	Structured-data	Structured-data are in a consistent order with a well-defined format. The advantage of structured-data is that they are easy to maintain, access, and store on computers. Structured-data are stored in the form of rows and columns; an example is a DataBase Management System (DBMS)
	Semi-structured data	Semi-structured data can be considered as another form of structured-data. It inherits a few properties of structured-data that do not represent the data in database models. An example is Common Separated Value (CSV) files

	Unstructured data	Unstructured data do not follow the formal structure rules of data models. Images, videos, text messages, and social media posts are examples of unstructured data.
Data store sources	Key value stores	Key value stores are used to store and access data in key/value pairs. They are basically designed to store massive data and manage heavy loads. Apache HBase, Apache Cassandra, Redis, and Riak are examples of key value store databases
	Graph stores	Graph stores are used to analyze data on the basis of the relationships between nodes, edges, and properties. Neo4j is an example of a graph store.
	Column family stores	Column family stores keep data and information within a column of a table at the same location on a disk in the same way a row store keeps row data together. Google Bigtable is an example of column family stores.
	Document-oriented stores	Document-oriented stores offer complex data forms in multiple formats, such as XML, JSON, text, string, array, or binary forms. CouchDB and MongoDB are examples of document-oriented stores
Data Staging	Cleaning	Cleaning is a process in which noisy data, outliers, and missing values are removed.
	transformation	In data transformation, data are transformed in an appropriate format for analysis.
	Normalization	Normalization is a process used to reduce redundancies from data
Data processing	Batch data processing	MapReduce-based systems are used to process data in the form of batches. Apache Hadoop, Apache Mahout, Skytree Server, and Dryad are examples of batch processing.
	Real-time data processing	Streaming systems, such as S4, are based on distributed frameworks that allow users to design applications for processing continuous unbounded streams of data

IV. CLOUD COMPUTING SERVICES

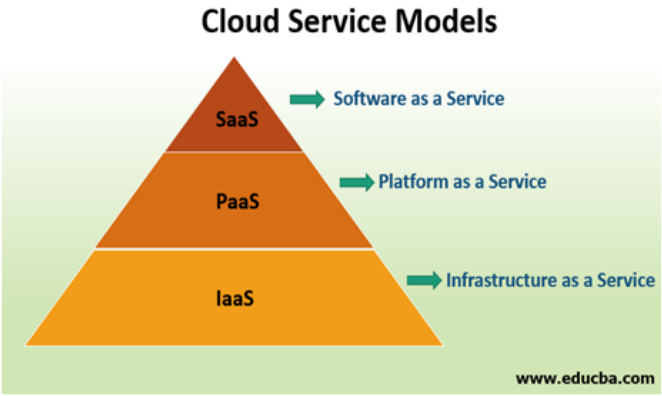

Fig 3: CLOUD COMPUTING SERVICES

classified into the following three categories[1]:

(1) Infrastructure as a Service (IaaS): These services are basically based on the principle of "pay for what you need". It provides high-performance computing to customers. Amazon Web Services (AWS),Elastic Compute Cloud, and Simple Storage Services (S3) are examples of IaaS. AWS and S3 provide online storage services. At nominal charges, customers can easily access the world's largest data centers. At present, three companies provide IaaS landscape services: Google, Microsoft, and HP. Google provides Google Compute Engine to access IaaS services. Microsoft also provides a cloud platform through its Window Azure Platform. HP offers HP Cloud, which is designed by NASA and Rack Space.

(2) Software as a Service (SaaS): With the help of the Internet, all applications are run on remote cloud infrastructure in SaaS. To access SaaS services, users need an Internet connection and a web browser, such as Google Chrome or Internet Explorer[40]. Users connect to a desktop environment via a virtual machine, in which all software programs are installed. SaaS provides more facilities to users than IaaS.

(3) Platform as a Service (PaaS): It provides a runtime environment to users. It allows users to create, test, and run web applications. Users can easily access PaaS on the basis of the pay-per-use mode using an Internet connection. PaaS provides the infrastructure (networking, storage, and services) and platform (DBMS, business intelligence, middleware) for running a web application life cycle. Examples of PaaS include Microsoft Azure and Google Cloud[41].

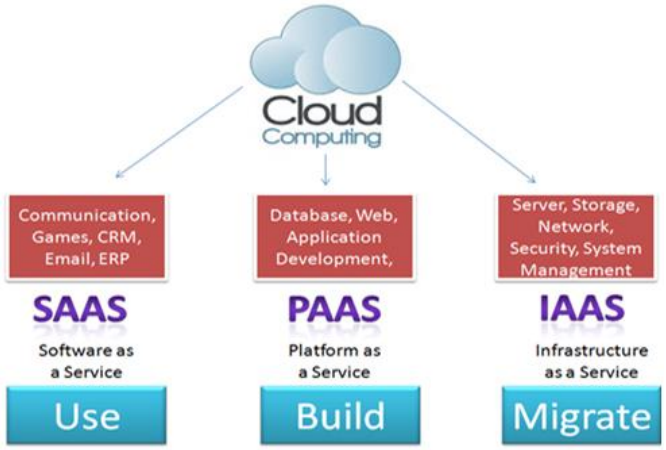

Fig 4 : Cloud computing services.

The cloud computing environment has two important aspects: the frontend and the backend. From the frontend side, users access cloud services through an Internet connection; at the backend, all cloud services are runFigure 3 shows the various types of cloud computing services[42].

Big data and cloud computing are closely associated. With technological changes, big data models Provide distributed processing, parallel technologies, large storage capacity, and real-time analysis of heterogeneous databases.

Data security and privacy are also considered in big data models. Big data require large amounts of storage space and thus entail the use of cloud computing. Cloud computing offers scalability and cost savings[43]. Moreover, it provides massive amounts of storage capacity and processing power. Cloud computing works on different types of technologies, such as distributed storage and virtualization, and processes data for different types of tasks. It accesses distributed queries over multiple datasets and gives responses in a timely types of tasks. It accesses distributed queries over multiple datasets and gives responses in a timely example of big data processing in a cloud environment that allows the storage of massive amounts of data in a cluster[9].

In other words, MapR is an efficient and cost effective model for processing big data. The MapR framework comprises the map and reduce functions for handling big data.

Cloud computing also plays an important role in distributed system environments by facilitating storage, boosting computing power, and aiding network communication. Big data technologies store data in cloud clusters rather than in local storage file systems. Several companies provide big data cloud platforms.

Moreover, various cloud computing platforms are available to store big data. Table 5 shows a comparative analysis of big data cloud frameworks for storing massive amounts of data[10]. Cloud services such as Microsoft Azure, Google Cloud, AWS, IBM, Hortonworks, and MapR are compared on the basis of various parameters.

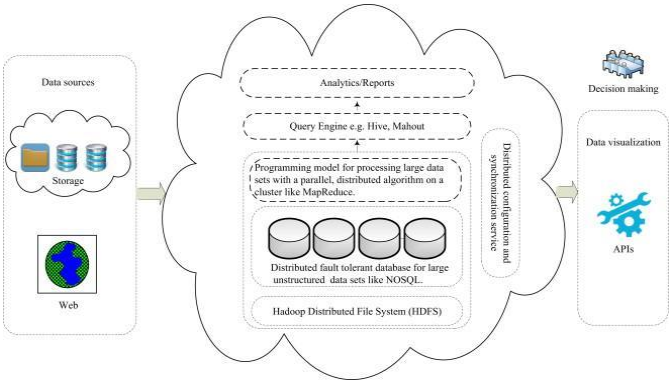

Fig 5: Big data and cloud computing.

V. Research Issues in Big Data

As data are growing at exponential rates, a number of issues and problems emerge during the processing and storage of big data. Few tools are available to resolve these issues and problems in a cloud environment.

Technologies, such as PigLatin, Dryad, MongoDB, Cassandra, and MapR, are not able to resolve these issues in big data processing. Even with the help of Hadoop and MapR, users cannot execute queries on databases, and they have low-level infrastructures for data processing and management. Some issues and problems in big data are summarized as follows[11]:

(1) Distributed database storage system:

Numerous technologies are used to store and retrieve huge amounts of data. Cloud computing is an important aspect of big data. Big data are generated by multiple devices on a daily basis. At present, the main issue in distributed frameworks is the storage of data in a straightforward manner and the processing and migration of data between distributed servers.

(2) Data security: Security threats are an important issue in a cloud computing environment. Cloud computing has been transformed with modern information and communication technologies, and several types of unresolved security threats exist in big data. Data security threats are magnified by the variety, velocity, and volume of big data. Meanwhile, various issues and threats, such as the availability of data, confidentiality, real-time monitoring, identity and access authorization control, integrity, and privacy, exist in big data when used with cloud computing frameworks.

Therefore, data security must be measured once data are outsourced to cloud service providers[11].

(3) Heterogeneity: Big data are heterogeneous in nature because data are gathered from multiple devices in different formats, such as images, videos, audio, and text. Before loading data into a warehouse, they need to be transformed and cleaned, and the processes present challenges in big data[12]. Combining all unstructured data and reconciling them for use in report creation are incredibly difficult to achieve in real-time.

(4) Data processing and cleaning: Data storage and acquisition require preprocessing and cleaning, which involves data merging, data filtering, data consistency, and data optimization. Thus, processing and cleaning data are difficult because of the wide variety of data sources[13]. Moreover, data sources may contain noise and errors, or they may be incomplete. The challenge is how to clean large amounts of data and how to determine whether such data are reliable.

(5) Data visualization: Data visualization is a technique to represent complex data in a graphical form for clear understanding. If the data are structured, then they can be easily represented in the traditional graphical way. If the data are unstructured or semistructured, then they are difficult to visualize with high diversity in realtime.
heterogeneity/data formats.

VI. CONCLUSION

Cloud computing has transformed the way businesses around the world do business in a way that many people are unaware of. Understanding the difference among various types of cloud computing and identifying which one is best suited for a growing business is tremendously important. This paper provides the knowledge of the introduction to cloud computing, its concepts, models and services. The paper also discusses the comparison of all cloud computing deployment models in table form. These clouds are compared against supported platforms, supported languages, storage capacity, services, and products. Fig. 3 shows Public cloud is the most popular general deployment option, with a usage share of over 61%. Traditional on-premises deployment, with just under half (49%) of shared use, ranks second. Hybrid cloud, which combines public cloud services with on-premises private cloud infrastructure, ranks third, with approximately 39% usage. The study encouraged respondents to choose from several of the five cloud deployment options. It shows a tenth (9%) selected all five, and almost a fifth (19%) selected four out of five. Among them two-thirds (64%) selected at least two cloud deployment options. The upshot is that while the public cloud is by far the most popular choice, most of the organizations surveyed employ a mix of cloud types. Interestingly, multi-cloud or the use of multiple cloud computing and storage services in a single homogeneous network architecture had the fewest users (24% of respondents). [1]

VII. REFERENCES

[1]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9663258 BIG DATA MINING AND ANALYTICS ISSN 2096-0654 03/06 pp 32 – 40 Volume 5, Numbe r 1, Mar ch 2022 DOI: 10.26599/BDMA.2021.9020016 C The author(s) 2022. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). Big Data with Cloud Computing: Discussions and Challenges

[2]. https://www.tutorialspoint.com/cloud_computing/cloud_computing_challenges.htm

[3]. J. H. Yu and Z. M. Zhou, Components and development in big data system: A survey, J. Electr. Sci. Technol., vol. 17, no. 1, pp. 51–72, 2019.

[4]. S. Kumar and K. K. Mohbey, A review on big data based parallel and distributed approaches of pattern mining, J. King Saud Univ. – Comput. Inform. Sci., doi:

[5]. https://en.wikipedia.org/wiki/Big_data

[6]. F. Ridzuan and W. M. N. Wan Zainon, A review on data cleansing methods for big data, Procedia Comput. Sci., vol.

[7]. https://www.javatpoint.com/big-data-characteristics

[8]. https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data

[9]. L. Q. Kong, Z. F. Liu, and J. G. Wu, A systematic review of big data-based urban sustainability research: state-of-thescience and future directions, J. Clean. Prod., vol. 273, p. 123142, 2020.

[10]. P. P¨a¨akk¨onen and D. Pakkala, Reference architecture and classification of technologies, products and services for big data systems, Big Data Res., vol. 2, no. 4, pp. 166–186, 2015.

[11]. M. Wook, N. A. Hasbullah, N. M. Zainudin, Z. Z. A. Jabar, S. Ramli, N. A. M. Razali, and N. M. M. Yusop, Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling, J. Big Data, vol. 8, no. 1, pp. 1–15, 2021.

[12]. S. M. Shamsuddin and S. Hasan, Data science vs. big data @ UTM big data centre, in Proc. of 2015 IEEE Int. Conf. Science in Information Technology, Yogyakarta, Indonesia, 2015, pp. 1–4.

[13]. T. Y. Yang and Y. Zhao, Application of cloud computing in biomedicine big data analysis cloud computing in big data, in Proc. of the 2017 Int. Conf. Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, India, 2017, pp. 1–3.