The Correlation of Cloud Computing and Big data: Review & Benefits

[1]Payal Patel, [2]Riddhi Patel

[1]M.E Scholar, [2]Professor
[1]Department of Computer Engineering,
[1]LDRP Institute of Technology and Research, Gandhinagar, India

Abstract: The phrase "big data" was developed in answer to the fast expansion of global data as a technology capable of storing and analysing vast and diverse volumes of data, offering deep insights about consumers and experiments to both businesses and research. 4V's of big data – volume, velocity, variety, and veracity – enables traditional data warehouses to tackle data management and analytics Cloud computing appears to be an ideal platform for storing large amounts of data.. On the other hand, working with big data in the cloud has its own challenges, including balancing two opposing design principles. This paper includes a brief overview of big data and relationship between the cloud and big data. It also introduces the challenges of delivering big data.

Keywords - Big data, Cloud Computing, Database, Hadoop, HDFS, MapReduce.

I. INTRODUCTION

There has been an increasing demand to store and process more and more data, in domains such as finance, science, and government. Systems that support big data, and host them using cloud computing, have been developed and used successfully. Whereas big data is responsible for storing and processing data, the cloud provides a reliable, fault tolerant, available and scalable environment so that big data systems can perform by Hashem et al. (2014) [1]. Big data, and in particular big data analytics, are viewed by both business and scientific areas as a way to correlate data, find patterns and predict new trends. Therefore there is a huge interest in leveraging these two technologies, as they can provide businesses with a competitive advantage, and science with ways to aggregate and summarize data from experiments such as those performed at the Large Hadron Collider (LHC).

Cloud computing is a metaphor used by Technology or IT Services companies for the delivery of computing requirements as a service to a heterogeneous community of end-recipients [2]. Also it is a technology based on Internet system that provides remote data centers to manage information services and applications. Cloud computing allows individual users and companies to administer files, information and applications without installing specific software on their computers just having Internet access [3]. 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 obstacles and challenges that this relationship may encounter [23].

II. BIG DATA: CONCEPT AND DEFINITION

"Big Data" refers to large and difficult data to store, manage, and analyse using traditional databases. Scalable architecture is required for efficient storage, manipulation, and analysis. This huge amount of data originates from a variety of sources: Smartphones and social media postings; sensors such as traffic lights and utility meters; point-of-sale terminals; consumer wearables like fitness trackers. Different methods are used to extract hidden values from this diverse and complicated data and turn it into meaningful insights, better decision making, and a competitive edge. The data set-at-rest qualities of volume and variety of data from many domains or categories, as well as the data-in-motion features of velocity, or rate of flow, and variability, demand a new architecture to achieve efficiency. Figure 1 shows all the properties of big data.

a. Volume

The enormous quantity of data created each second from many sources such as social media, mobile phones, vehicles, credit cards, M2M sensors, photos, and videos, allowing users to data-mine the hidden information and patterns found in them.

b. Velocity

The rate, at which data is created, transmitted, gathered, and analysed. Data created at an ever-increasing rate must be analysed, and the speed of transmission and access to the data must stay instantaneous to enable real-time access to various applications that rely on this data.

c. Variety

Data is created in many formats, either structured or unstructured. Structured data, such as a person's name, phone number, address, financial information, and so on, maybe organized in a database's columns. This data is quite simple to enter, save, query, and analyse. Unstructured data, which accounts for 80 % of today's world data, is harder to trace and retrieve than structured data.

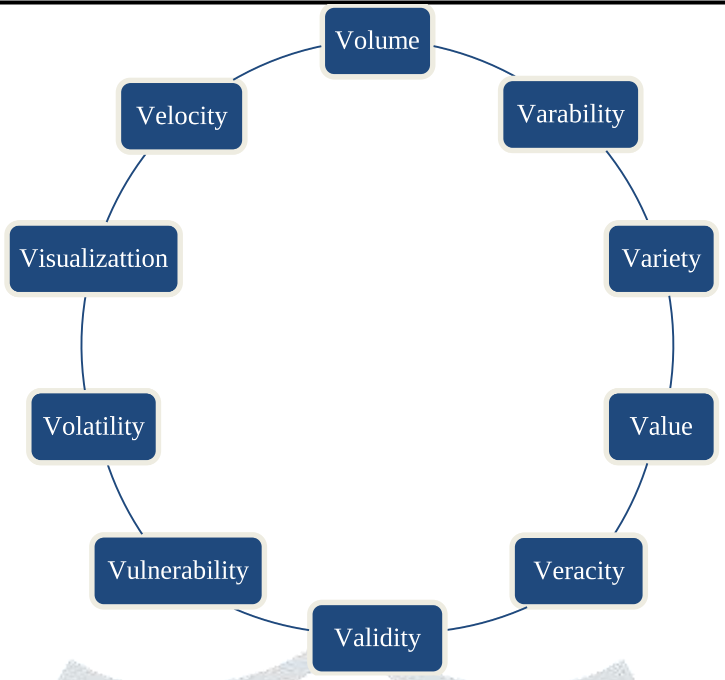

Fig. 1 V's of Big Data [4]

d. Variability

It refers to the high degree of irregularity in data flow and its volatility during peak periods. The unpredictability is caused by a plethora of data dimensions originating from a variety of diverse data kinds and sources. Variability may also refer to the irregular rate with which large data is absorbed into data storage.

e. Value

The hidden value revealed from data for decision-making is referred to as value. Big data may provide significant value in a variety of ways, including better understanding your consumers, targeting them appropriately, streamlining processes, and enhancing machine or business performance.

f. Veracity

It refers to the quality and dependability of the data source. Its significance is determined by the context and the meaning it brings to the analysis. Knowledge of the validity of the data, in turn, aids in a better understanding of the risks involved with data-driven analysis and business choices.

g. Validity

The correctness of the data gathered for its intended application. To guarantee uniform data quality, standard definitions, and metadata, proper data governance processes must be implemented.

h. Vulnerability

This refers to the security of the data that has been gathered and kept.

i. Volatility:

The period of time that data must be stored historically before it is deemed irrelevant to the current research.

j. Visualization

Refers to the process of making data intelligible to nontechnical stakeholders and decision-makers. The development of sophisticated graphs that turn data into information, information into insight, insight into knowledge, and knowledge into an advantage for decision making is known as visualization [4].

2.1 Big Data Formats

Big data is generated by a variety of sources, including sensors and free texts such as social media, unstructured data, metadata, and other geographical data gathered from weblogs, GPS, medical devices, and so on [5]. Because big data is acquired from many sources, it comes in a variety of formats, including:

1. Structured data

This is data that has been arranged into tables or databases and is ready to be processed.

2. Unstructured data

This is the majority of data; it is the data that individuals produce daily in the form of words, photos, videos, communications, log records, click-streams, and so on.

3. Semi-structured or multi-structured data

It is considered structured data, but unlike XML documents or JSON, it is not intended in tables or databases [6].

III. Cloud Computing

It is a word used to describe on-demand computing resources and systems that may deliver a variety of integrated computer services without being constrained by local resources to simplify user access. Data storage, backup, and self-synchronization are examples of these resources, as are software processing and scheduling activities [7]. Cloud computing is a shared resource system that may provide a range of online services, including virtual server storage, applications, and desktop programme licensing. Cloud computing is able to expand and offer volume by using shared resources [8].

3.1 Cloud Computing Characteristics

One of the distributed systems that reflects a complex model is cloud computing. The National Institute of Standards and Technology (NIST) have highlighted essential elements of the cloud by condensing the idea of cloud computing into five characteristics, which are as follows:

1. *On-demand self-service:*
   Cloud services deliver computer resources such as storage and processing when and where they are needed, with no human interaction.
2. *Broad network access:*
   Cloud computing resources are available via the network, and mobile and smart devices, as well as sensors, may access cloud computing resources
3. *Resource Pooling:*
   Cloud platform users share a broad array of computing resources; the user may choose the type of resources and the geographical region they desire, but they cannot choose the actual physical placement of these resources.
4. *Rapid Elasticity:*
   Storage media, network, processing units, and applications resources are constantly accessible and may be raised or lowered nearly instantly, allowing for high scalability to assure optimal resource use.
5. *Measurable service:*
   Cloud systems may measure processes and resource usage, as well as surveillance, control, and reporting [9][10][11].

3.2 Cloud Computing Services

Although there is a wide range of cloud computing services accessible, most of them fit into one of the following categories:

1. Software as a Service (SaaS)

The most widely utilized business choice in cloud services and represents the largest cloud market. SaaS allows consumers to access apps through the internet. Third-party suppliers are responsible for maintaining SaaS-based applications. Because most SaaS apps operate directly in the browser, the client doesn't have to download or install any software. Applications, runtime, data, middleware, OS, virtualization, servers, storage, and networking are all managed by the SaaS provider, making it easier for businesses to streamline their maintenance and support.

2. *Platform as a Service (PaaS)*

A term used to describe how software is delivered as a Platform as a Service (PaaS) is a business model that allows developers to access hardware and software resources via the internet to create bespoke applications. PaaS enables rapid, easy, and cost-effective application creation, testing, and deployment. This paradigm enables businesses to design and develop applications that are incorporated into PaaS software components, while enterprise operations or third-party providers manage the operating system, virtualization, servers, storage, networking, and the PaaS software itself. Because these apps are cloud-based, they are scalable and highly available.

3. *Infrastructure as a Service (IaaS)*

Through virtualization technology, the Infrastructure as a Service cloud computing paradigm offers companies a self-servicing platform for accessing, monitoring, and controlling distant data center infrastructures such as compute, storage, and networking services.

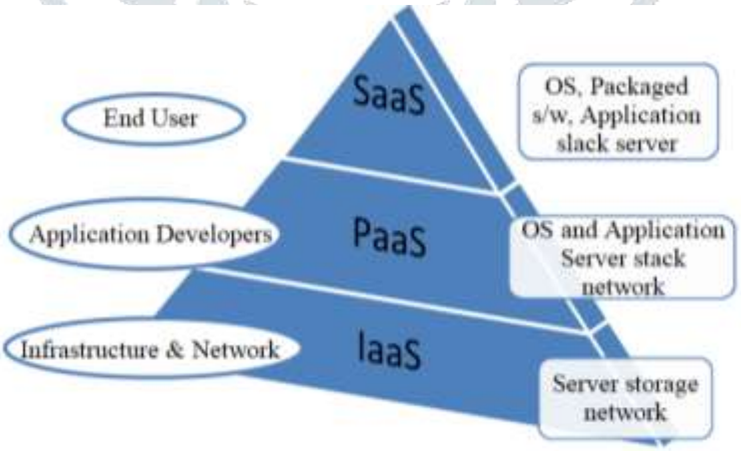

Fig. 2 Cloud Computing Services [14]

Users of IaaS are in charge of managing applications, data, runtime, middleware, and the operating system, while providers are in charge of virtualization, servers, hard drives, storage, and networking. IaaS delivers the same capabilities as data centers without the need to physically maintain them [4]. Figure 2 depicts the many cloud computing services available

IV. The Connection Between Cloud Computing and Big Data

Big data and cloud computing are merged. Big data enables practitioners to perform distributed queries over massive data sets using commodity computing and provide results in a timely way. Hadoop, a family of distributed data processing technologies, relies on cloud computing as the underlying engine. A distributed fault-tolerant database is handled in a cluster using a programming paradigm for large datasets using a parallel distributed algorithm. The primary goal of data visualization is to provide a visual representation of analytical results in the form of various graphs for decision-making. Big data, which is based on cloud computing, makes use of distributed storage technologies rather than local storage linked to a computer.

Cloud computing and big data are mutually beneficial. The fast growth of big data is viewed as a concern. Traditional storage can't keep up with the demands of big data, much alone the necessity for data to be shared across several distributed storage sites. Clouds are evolving and providing solutions for the appropriate environment of big data [13], whereas traditional storage cannot meet the requirements for dealing with big data, let alone the need for data exchange between various distributed storage locations. Big data challenges are solved via cloud computing [14]. The cloud computing environment is growing to be able to absorb large amounts of data because it adheres to the data splitting policy, which means that data is stored in many locations or availability areas. Resource pooling is utilized to give flexibility on-demand in cloud computing environments designed for general-purpose applications. As a result, big data appears to be a good fit for the cloud computing environment [15].

One of the most essential components of big data is information privacy and security. Maintaining privacy and security in the cloud is a major problem since data is hosted and processed on third-party services and infrastructure. The greater the amount, diversity, and validity of big data, the greater the risk of privacy and security. In many respects, mobile health has changed the way health care is delivered. Through different health care mobile applications, users have been able to maintain their lifestyle, health and fitness, drug reference, and diagnosis. Because everything is sent through the mobile internet, data privacy might be compromised if the network is not adequately protected or exposed to outsiders. Various new laws, privacy rules, regulations, protections, industry standards, and service level agreements must be maintained between providers and customers in order to give users with trust and ensure that their data is not exposed [16].

V. Big Data in the Cloud

Scalability, fault tolerance, and availability are all important for storing and processing large amounts of data. Through hardware virtualization, cloud computing provides all of these benefits. Since a result, big data and cloud computing are complementary ideas, as the cloud makes large data more accessible, scalable, and fault-tolerant. Big data is viewed as a profitable business prospect by the business world. As a result, a number of new businesses, including Cloudera, Horton works, Teradata, and others, have begun to focus on providing Big Data as a Service (BDaaS) or DataBase as a Service (DBaaS). Consumers may also consume big data on demand through companies like Google, IBM, Amazon, and Microsoft. Next, we'll look at two case studies: Nokia and RedBus, both of which demonstrate how big data may be used successfully.

a. Nokia

One of the first businesses to recognise the value of big data in cloud settings was Nokia . Several years ago, the firm employed different DBMSs to meet the needs of each application. Realizing the benefits of combining data into a single application, the firm chose to transition to Hadoop-based systems, combining data within the same domain and utilising analytics algorithms to get adequate insight into its clients. Hadoop's cost per terabyte of storage was lower than a typical RDBMS since it runs on commodity hardware proposed by Cloudera, 2012 [17].

Nokia collects enormous amounts of data from mobile phones on a petabyte scale in order to better understand their user interactions and improve the user experience with their phones. The company deployed Hadoop data warehouse as an infrastructure to store this daily petabyte of unstructured data acquired from mobile phones in use, services, log files, and other sources in order to derive business decision strategies and obtain a holistic perspective of user engagement.

b. Redbus

RedBus is India's leading firm specialising in online bus ticketing and hotel reservations. This business needed to build a sophisticated data analysis tool in order to get insight into their bus booking service [18]. Its databases have the potential to grow to 2 terabytes in size. The app would have to be able to analyse booking and inventory data from hundreds of bus companies that serve millions of people. This massive data is then sent to Big Query, which runs complex queries and provides answers to various analytical queries in seconds, such as how many times a customer searched for a destination and found only a few bus services available, and any technical issues that may arise during booking and alert the appropriate team.

c. Tweet Mining in cloud

To collect and analyse tweets, Noordhuis et al. [19] employed cloud computing. All computations were done using Amazon's cloud infrastructure. The crawling of tweets was followed by the use of a page ranking algorithm. Google uses PageRank to determine the relevance of a web page. In general, if the author of a Web page A links to another Web page B, the author of Page A is likely to consider Page B significant. The number of in-edges defines the significance of a certain page in this way. As a result, the greater the number of inbound links to a page, the more important it is. The similar concept applies to tweets, where a twitter user following another tweet is considered relevant instead of a website page. To compute PageRank, we must crawl the Twitter social network. Nearly 50 million nodes and 1.8 billion edges were scanned, accounting for almost two-thirds of Twitter's projected customer base.

VI. Challenges of Big Data

The bandwidth and latency [20] are two elements and problems that impact the timely processing of large data. In the link between big data and cloud computing, there are various problems that may be described.

a. Data Storage

Traditional storage of big data is troublesome because hard drives fail frequently, data protection measures are ineffective, and the pace of big data necessitates storage systems that can expand fast, which is difficult to do with traditional storage. Cloud storage services provide nearly limitless capacity with high fault tolerance, making them attractive options for addressing big data storage issues.

b. Variety of Data

The expansion of nearly infinite sources of data causes big data to organically expand, increase, and vary. The diverse nature of big data is a result of this development. In general, data from many sources, of various forms and formats, is highly interconnected. They are inconsistent and have conflicting shapes. Data can be stored in an organised, semi-structured, or unstructured format. Data can be stored in a structured, semi - structure, or unstructured fashion. Semi-structured data formats are only minimally suited for today's database systems, whereas structured data formats are ideal. Because unstructured data has a complicated format that is difficult to express in rows and columns, it is unsuitable.

c. Data Transfer

Data is collected, inputted, processed, and outputted at various stages. Because big data transport is difficult, data compression techniques must be used to lessen the volume, which is a barrier for transfer speed. Cloud computing reduces costs by distributing storage resources and transferring data over high-speed lines, while virtual resources and resource consumption at the user's request reduce expenses.

d. Privacy and data ownership

Because the cloud is an open environment, the user's ability to monitor it is restricted. Big data poses a significant privacy and security problem. In the real world, big data and cloud computing interact. Cloud computing is expected to access about 40% of global data by 2020, according to (IDC) projections. To enable massive data processing, cloud computing provides powerful storage, computation, and dissemination capabilities. As a result, there is a high need for research on data privacy and security issues in both cloud computing and big data.

VII. Technologies of Big data

Hadoop

Hadoop is a tool for forming data node clusters and storing data in a space-efficient manner. It can manage and transport data between racks in a variety of environments. Hadoop distributes data across several servers, making it easier to execute various applications. Despite the failure of diverse node clusters, it has a reduced rate of system failure. Hadoop Framework is a framework used by Google, Yahoo, Amazon, IBM, and other well-known organizations. Hadoop is a tool that allows data-intensive applications to run on top of one another. Task trackers, job trackers, data engine, and fetch manager are all part of the Hadoop architecture. Task trackers are used to keep track of the tasks that need to be completed. Even when a large nodes fail, this strategy reduces the chance of the entire system failing. Hadoop allows for a scalable, cost-effective, adaptable, and fault-tolerant computing solution [21].

The data engine provides data processing information. The data is fetched by the fetch manager during the execution of a given job. Hadoop is divided into two primary projects– Hadoop Distributed File System (HDFS) and MapReduce [22].

a. Hadoop Distributed File System (HDFS) —

It can withstand system failures while also storing massive amounts of data. The data is fetched by the fetch manager during the execution of a given job. The Hadoop framework is used in a variety of applications. It keeps three copies of the data on three separate servers. Both the data node and the name node have their own set of data files. The name node is in charge of accessing all sorts of files, whereas the data node communicates with itself to execute file system operations.

b. Map- Reduce —

Map-reduce is a programming model for handling huge amounts of data in a reliable and fault-tolerant way. It divides the data into pieces that are handled in parallel by Map tasks. During the processing, the input and output data is saved in a file system. It also keeps track of the unsuccessful job and re-runs it.

VIII. Conclusion

The big data paradigm has resulted in the discovery of hidden information from the data as a result of innovation and competitiveness driven by improvements in cloud computing. In this study, we present an overview of big data in cloud environments, highlighting its benefits and demonstrating how effectively the two technologies work together, as well as the problems that both confront.

References

[1] Hashem, I.A.T. et al., 2014. The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47, pp: 98–115.

[2] Pedro Caldeira Neves, Bradley Schmerl, Jorge Bernardino and Javier Cámara. Big Data in Cloud Computing: features and issues. September, 2016

[3] Mansor Zauir, Mohamad M. Al Rahhal, Abdullah Al-Faifi, Alaaeldin M. Hafez, Hassan Abdalla. Survey Of Data Mining Usage In Cloud Computing : https://www.researchgate.net/publication/312211163

[4] Manoj Muniswamaiah, Tilak Agerwala and Charles Tappert. Big Data In Cloud Computing Review And Opportunities. International Journal of Computer Science & Information Technology (IJCSIT) Vol 11(4), August 2019

[5] Nabeel Zanoon , Abdullah Al-Haj and Sufian M Khwaldeh. Cloud Computing and Big Data is there a Relation between the Two: A Study. International Journal of Applied Engineering Research, Volume 12, Number 17 (2017), pp: 6970-6982

[6] Liebowitz, J. (Ed.). (2014). Bursting the big data bubble: The case for intuition-based decision making. CRC Press.

[7] Calheiros, Rodrigo N., et al. "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms." Software: Practice and experience 41.1 (2011), pp: 23-50.

[8] R.Subhulakshmi, S.Suryagandhi, R.Mathubala, P.Sumathi, An evaluation on Cloud Computing Research Challenges and Its Novel Tools, International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST) Volume 2, Special Issue 19, October 2016.

[9] Fonseca, N., & Boutaba, R. (2015). Cloud services, networking, and management. John Wiley & Sons.

[10] https://www.ibm.com/blogs/cloudcomputing/2014/01/cloud-computing-definedcharacteristics-service-levels/

[11] Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. Journal of internet services and applications, 1(1), pp: 7-18.

[12] https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/

[13] TIAN, Wenhong Dr; ZHAO, Yong Dr. Optimized cloud resource management and scheduling: theories and practices. Morgan Kaufmann, 2014.

[14] S. Hemalatha, M.S. Kokila and S. Krithika. Cloud Computing in Big Data Analytics. International Journal of Engineering and Management Research Volume-6, Issue-1, January-February-2016, pp: 381-383

[15] Wei-Dong Zhu, Manav Gupta, Ven Kumar, Sujatha Perepa, Arvind Sathi, Craig Statchuk, Building Big Data and Analytics Solutions in the Cloud, IBM Redbooks - 2014.

[16] Mohaiminul Islam and Shamim Reza. The Rise of Big Data and Cloud Computing. Internet of Things and Cloud Computing 2019; 7(2), pp: 45-53

[17] Cloudera, 2012. Case Study Nokia: Using big data to Bridge the Virtual & Physical Worlds.

[18] Kumar, P., 2006. Travel Agency Masters big data with Google bigQuery

[19] Sakr, S. & Gaber, M.M., 2014. Large Scale and big data: Processing and Management Auerbach, ed.

[20] Parvin Ahmadi Doval Amiri and Mina Rahbari Gavgani, 2016. A Review on Relationship and Challenges of Cloud Computing and Big Data: Methods of Analysis and Data Transfer. Asian Journal of Information Technology, 15: 2516-2525

[21] Venkatesh H, Shrivatsa D Perur and Nivedita Jalihal. A Study on Use of Big Data in Cloud Computing Environment, International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 6 (3) , 2015, 2076-2078

[22] Balu Srinivasulu and Andemariam Mebrahtu. Concepts and Technologies of Big Data Management and Hadoop File System. International Journal of Computer Trends and Technology (IJCTT) –Volume 44(2), February 2017 ISSN: 2231-2803, Page 80

[23] Prof. R. N. Yeotikar. Study of Relation between Big Data & Cloud Computing: Big Data Challenges & Issues. International Journal of Creative Research Thoughts (IJCRT), www.ijcrt.org http://doi.one/10.1727/IJCRT.17197 IJCRTICGT077 581

JETIR