Analyzing the application areas of big data in conjunction with cloud computing

Maninderjit Kaur
Department of Computer Science and Engineering, Gulzar Group of Institutions, Khanna

Pratibha Soram
Department of Computer Science and Engineering, Gulzar Group of Institutions, Khanna

Aditi
Department of Computer Science and Engineering, Gulzar Group of Institutions, Khanna

Abstract - Due to the advancements in the current technologies, bulk data is generated on a daily basis which builds a significant challenge for users as such kind of data requires proper processing and storage. Though, there are various platforms that help to manage and analyze this enormous amount of data to extract useful insights. Cloud Computing provides a powerful environment for storing, managing and analyzing data on multiple servers. The enormous amount of data generated is termed as „big data", which represents a big challenge in cloud computing. An ample of applications in different fields exist that focus on big data concepts. This paper will summarize the application areas of big data in conjunction with cloud computing.

Keywords: Big Data; Cloud Computing; Big Data Analytics; Big Data-driven Areas; Amazon Web Services (AWS); Big Data Warehouse

I. INTRODUCTION

In the era of technical advancements, massive amounts of data which is defined as the raw collection of information is generated in every microsecond. In this era of big data, artificial intelligence, cloud computing, the Internet of Things (IoT), and many other technologies, data generated from diverse resources is particularly vital since these technologies view data as extremely crucial and their actual property or petroleum of this era.

From the previous few years, with the advancement in the big data technologies many dominant sectors of industries like manufacturing and many others have immersed in the usage of immeasurable data, these industries have redrafted their processing units and also made alterations to their business plans as well as business model.

Big Data and Big Data Analytics (Big Data tools and technologies) have recently changed the working style of many organizations and businesses, corporations, professionals, and academia get many important and new opportunities (Supriya Saker et al., 2023). Governmental and nongovernmental organizations in addition, to businesses and research institutions now regularly generate volumes of data for various specific reasons. As a result, for businesses worldwide, obtaining valuable information from these massive data sources has become essential. However, the study concludes that it is difficult to efficiently and expertly extract valuable information from BD. BDA is now unavoidably necessary to realize this. The overall benefit of BD to enhance business performance and boost market share for the majority of businesses (Isaac Kofi Nta et. al., 2022).

As the data collected from various sources is quite burden some therefore the space to repost all this massive amount of data must be vast. Despite the fact that storage costs have been declining, small and medium-sized enterprises may still be unable to afford the resources required for analyzing big data. And hence the need for cloud computing arises Cloud computing, one of the most expressive developments in modern ICT and a service for corporate applications, has offered a strong architecture for handling complex and large-scalecomputing. Integration underlies the interaction between big data and cloud computing, with the cloud serving as the product that will be represented by the big data warehouse because it cannot be created, it is kept in the storehouse. The 'relational' databases of the past are no longer enough to handle data from numerous sources. The link between big data and the cloud will now start to emerge. In this paper, there is a discussion regarding the connectivity among both terms: big data and cloud computing.

II. BIG DATA

Through technological activities, big data is created from a variety of sources. It demands sufficient processing power and advanced analytical capabilities. Big data is important since it can be used analytically to make decisions that will lead to better and faster service. It is challenging to reposit, handle, and analyze huge volume of big data with the assistance of Conventional data base systems(SupriyaSakeretal.,2023).Thenatureofdataiscomplex,requires thorough processes to uncover and transform the data into valuable insights. Big data 's methodologies and tools are used to unearth significant hidden values from vast datasets that are

diverse, intricate, and of an enormous scale.

III. BIG DATA'S5V'S

Big data refers to vast quantities of fast big data of many forms that are unable to be processed and reposited by standard computers. The fact that the problem is not just with data volume, or the "five Vs," which are the primary characteristics of big data can be used to summarize these aspects:

Volume: Volume is the most vital dimension of Big data. Hugevolume of data in number of zeta bytes is generated from many sources. (J.K. Laurila et. al., 2012)

Variety: It oversees many data kinds. The amount of structured data in databases has changed to unstructured data as more individuals use social media, smartphones, and the Internet. Numerous formats, including pictures, SMS, GPS data, audio and video clips, and more, can be used to convey data. (Parvin Ahmadi et.al.,2016).

Velocity: It illustrates the continuousness of data generation from various origins, such as Twitter and Facebook, at a given pace. A system that provides super-fast data analysis is required due to the dramatic growth in data volume and frequency(J.J. Berman, 2013).

Veracity: It shows the degree of confidence in the data's contents as well as the data's accuracy, dependability, and quality. The significant variances in the quality of the data obtained determine how accurate the analysis is. (Chen, Min, et al.).

Value: It illustrates the value of big data by showing the data's relevance following processing. This is due to the fact that the data is essentially useless by itself. Value is obtained by examining the data, information and ideas it provides. Value processing comes last, following processing of volume, velocity, diversity, contrast, validity, and visualisation. (Nabeel Zanoon et. al.,2017).

Cloud computing

Cloud computing is a quickly developing innovation that is well-known in the emerging market. Delivering dependable IaaS, hardware, and software across the Internet and to remote data centres is the promise of cloud computing. Cloud services today cover a spectrum of functions including database management, application services, storage and computation. This makes them a powerful framework, for handling large scale computingtasks.(IbrahimAbaker TargioHashemet.al.,2015). Theneed torepositandanalyse massive chunks ofdatasetshas led to the adoption of cloud computing by many enterprises. Many scientific applications for large-scaleexperiments have already been positioned in the cloud and may keep up to increase as a result of the deprivation of readily available computer resources with the growing amount of data being generated, along with the benefits of servers and reduced capital costs there is a trend towards increased reliance, on these factors.

Cloud computing characteristics

A variety of online services, including desktop application licensing, apps, and virtual server storage, can beobtained throughcloud computing, a collaborative resource system. Usingshared resources allows cloud computing to grow and become more prevalent. Cloud computing is a model that exists within distributed systems. It encompasses five characteristics that define the concept of cloud computing. NIST has highlighted key elements of the cloud are:

Self-service on demand: When needed, cloud services supply computer resources for processing and storing without requiring human intervention.

Broad network access: Networks, mobile devices, and smart gadgets are used to access Cloud computing resources. Even sensors can use cloud computing infrastructure.

Resource Bunching: Cloud platform users provide a giant array of computational resources; they are able to select the type of resource they wish to use as well as the region in which it is located, butThey are unable to determine the whereabouts of these resources.

Quick Elasticity: In order to guarantee optimal resource utilization, resources from storage devices, networks, CPUs, and applications can be scaled up or down nearly instantaneously

Measured service: Cloud systems may measure processes and resource consumption as well asmonitoring, management, and reporting in an entirely transparent way (Fonseca, N. et. al, 2015) (Zhang,Q., et.al, 2010).

Services provided by cloud computing

As stated by the encyclopedia "Cloud computing offers an shared collection of computing resources including servers, networks and services. These resources are easily accessible. Can be quickly allocated or de-allocated without administrative hassle or the need to reach out to service providers. Furthermore you only pay for what you use." The following categories can be used to group cloud computing services:

Infrastructure as a Service (IaaS): The "pay for what you need" model underpins how these services function in essence. IaaS provides high performance computing. Elastic Compute Clouds, Simple Storage Service (S3), and Amazon Web Services (AWS) are a few IaaS examples. Amazon and S3 provide online storage services. Customers can pay reasonable fees to access the world's biggest data centers. IaaS landscapeservicesarecurrentlyprovidedbyMicrosoft,HP,andGoogle.GoogleprovidesGoogleCompute

Engine for accessingIaaS services. Microsoft provides a cloud platform as well with its Windows Azure Platform. HP Cloud is a NASA and RackSpace product that is offered by HP.

Software as a Service (SaaS): Every SaaS application runs on remote cloud infrastructure with the aid of the Internet. In order to use SaaS services individuals need to have an internet connection and access, to a web browser. They connect to a machine that hosts a desktop environment equipped with all the required software packages. SaaS provides users with a variety of features and benefits compared to IaaS. (A. O"Driscoll et. al., 2013)

Platform as a Service (PaaS): PaaS furnishes users with a runtime environment, enabling them to create, access, and utilize web applications. Users, with an internet connection, can leverage PaaS, which uses the concept of pay-per-use. PaaS offers both the infrastructure (networking, storage, and services) and the platform (DBMS, business intelligence, middleware) necessary for managing the lifecycle of online applications. Notable examples of PaaS include Microsoft Azure and Google Cloud.

Data as a Service (DaaS): The data is considered as the vital part of any system. DaaS is a cloud-based service for analyzing and managing data. For these reasons, big data and the technologies that must be used with it are intimately intertwined. DaaS offers extremely effective means of processing and distributing data. SaaS (software as a service) and SaaS (storage as a service) are closelyrelated, and DaaS can be used in conjunction with either or both of these models (Nabeel Zanoon et. al.,2017).

Integration of cloud computing and big data

The electronic information society is growing at a faster rate than ever thanks to technical advancements like cloud computing. This result in the big data phenomenon and the exponential growth of big data presents challenges forthe creation of an electronic information society (Goda, K., et. al. 2012). Both big data and cloud computing relyon resources and they work together seamlessly. Big data manages the storage capabilities of the cloud infrastructure.

Big data and cloud computing work best when used together (S. Salloum, et. al. 2019). Quick expansion of big data is thought to be complex. Traditional repository system cannot match the standards for dealing with big data. The demand for data interaction across numerous remote storage locations is accelerating, and clouds are emerging as a solution to create an appropriate environment for big data. Big data issues are addressed and solutions are provided via cloud computing. As it adheres to the data policy, the cloud computing environment is flourishing to be able to take uplarge amounts of data.

The Hadoop Distributed File System (HDFS) distributes the storage of massive amounts of data. A Hadoop data storage and management system is called HDFS. The HDFS has the benefit of being inexpensive, capable of managing a cluster of thousands of nodes, and able to handle enormous amounts of unstructured data. Massive volumes of data can be reposited in a cluster using Map Reduce (MapR), a type of big data processing that operatesin a cloud environment (L. Q. Kong et. al,2020).

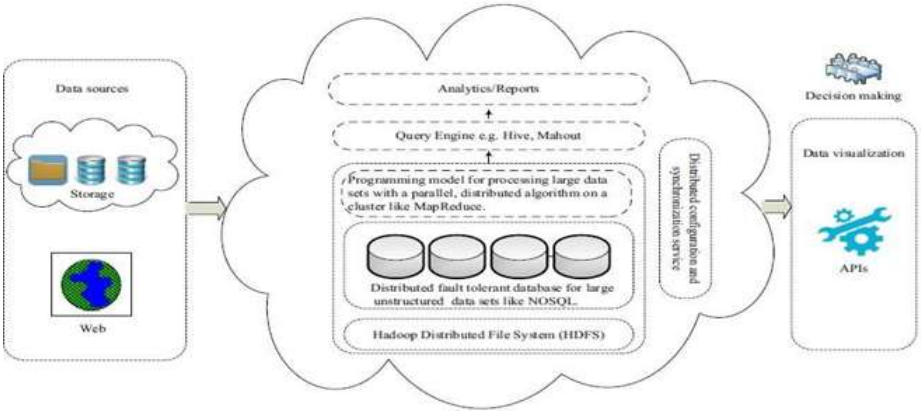

Figure1:Big data and Cloud Computing Conjunction

Virtual machine between big data and cloud computing

A virtual machine (VM) is set of programs that simulate an artificial computer environment in which the operating system (OS) and its associated applications can run on a single machine. Multiple virtual machines can be installed on a single VM. The notions of network computing, distributed systems, and parallel programming are not new because virtual technology is one of the key factors enabling the cloud. Virtualization technologies often allow one virtual machine to host many virtual machines.

Virtualization allows for the coupling of virtual metering devices into a single physical server, hence reducing the load on each one. The greatest platform for big data and conventional applications is virtualization technology. Assuming large data applications makes it easier to manage your big data infrastructure, produces results more quickly, and is more affordable. Physical or virtual infrastructure supported applications. This includes important contemporary big data, cloud, and mobile applications. Virtualization is a feature of big data systems like Hadoop that makes managing large amounts of data easier. Virtualization has various benefits over physical infrastructure. Many sources, such as multidimensional storage, web and data services, XML documents, analytical

tools, and internal and external applications, are used to generate today's virtual data. A contemporary source type that supports virtual data is data reposits (NoSQL) analytics services remains carce (MikeFergusonet.al.,2014).Virtual data is the only means of accessing and improving the heterogeneous settings used in big data initiatives. With the use of the cloud computing paradigm, customers can have a default data centre that can use data sets that weren't previously accessible by using a shared (API) for different data sets.

IV. LITERATURE REVIEW

Practitioners and researchers face challenges to stay up to date with innovations in the field of Big Data or Big data Analysis (BDA) due to the abundance articles. Therefore, earlier research made an effort to summarize many ML applications in BDA and its evolution in order to aid beginners in choosing the best ML method for BDA.
The focus of much of the research, however, was BDA on certain sectors, such as information security, healthcare, air quality, and the Internet of Things (IoT).Similar to how many are concentrated on the overview, difficulties, and strategy in BDA.

Amanpreet Kaur Sandhu, 2022 Over the last few decades, data has grown and continues to rise. Numerous sources produce data in various formats. Data variety is increasing as a result. High-speed data is generated by the mobile devices and sensor networks that are associated. Instead of requiring a specific space and expensive computer gear and software maintenance, cloud computing services are practiced to process, analyze, and reposit data. This paper investigates the association among cloud computing and big data. Furthermore, a comparison between big data and cloud services is performed. Numerous issues and worries are linked with big data, for example data privacy and security, distributed database storage, and heterogeneity/data formats.

Supriya Saker et al., 2023, reviewed articles on big data technology used in manufacturing, mining, and power sectors. Industries that have been established on three levels include the mining and manufacturing sectors.Along with this, they discussed the distribution and frequency of reviewed articles by year, as well as the big data innovations that were utilized to collect and process the enormous amounts of data from the industry. They also showed how frequently ML and data mining methods are applied to processing business data. They next discussed the progressive big data solutions that have been created to collect, organise, reposit and evaluate data pertaining to mining,manufacturing,andelectricity.Inanattempttofillthegaps,weexaminedthebigdataresearch fillsthe gaps that exist in a number of industrial sectors and offered suggestions for data-driven industry practises. For multi-dimensional big data, the quality evaluation system needs to be accurately enhanced in order to outperform later processing and storage in the general industry.

Ibrahim Abaker Targio Hashem et. al.,2015, research shows that there is a vast amount of data available now, and it is rising every day. The diversity of data kinds producedis also flourishing. The rate of data expansion and collecting has grown due to the widespread use of mobile telephones and other gadget sensors that are connected to the Internet. Companies across all industries might benefit from this data's potential to gain up-to-date business insights. For some time now, it has been able to reposit, work upon, and analyse data using cloud services; this has modified the information technology landscape and brought the promises of the on-demand service model tofruition. In this study, they examined the outgrowth of big data in cloud computing. They discussed the background ofHadoop technologyas well as MapReduce and HDFS, two ofitskeycomponents. The programme that underpins the ongoing MapReduce initiatives was given. They also discussed a few challenges associated with handling large amounts of data. The review looked at a number of factors, including volume, scalability, availability, privacy, lawful and regulatory issues, data access, governance, data integrity, data security, data transformation, and quality/heterogeneity. Additionally, the primary issues with big data in clouds were highlighted. Both academia and business must address upcoming challenges and issues. In order to ensure the sustained prosperity of data management within a cloud computing context and to collaboratively explore novel avenues, collaboration between social science academics, practitioners, and researchers is important.

Table1:Differenceamongdifferentbigdatacloudplatforms

	Google	Microsoft	Amazon	Cloudera
Big data storage	Google cloud services	Azure	S3	
MapReduce	AppEngine	Hadoop on Azure	Elastic MapReduce (Hadoop)	MapReduce YARN
Big data analytics	BigQuery	Hadoop on Azure	Elastic MapReduce (Hadoop)	Elastic MapReduce (Hadoop)
Relational database	Cloud SQL	SQL Azure	MySQL or Oracle	MySQL, Oracle, PostgreSQL
NoSQL database	AppEngine Datastore	Table storage	DynamoDB	Apache Accumulo
Streaming processing	Search API	Streaminsight	Nothing prepackaged	Apache Spark
Machine learning	Prediction API	Hadoop+Mahout	Hadoop+Mahout	Hadoop+Oryx
Data import	Network	Network	Network	Network
Data sources	A few sample datasets	Windows Azure marketplace	Public Datasets	Public Datasets
Availability	Some services in private beta	Some services in private beta	Public production	Industries

Pedro Caldeira Neves et. al. 2016 ,intend to leverage adaptive methods in this specific domain to create a method for establishing elasticity at various levels of big data systemsworking in cloud settings. The objective is to look into the mechanisms that adaptable software can employ to start scalability at various cloud stack tiers. so automatically and promptly accommodating data peaks. The paper presented gives an overview of big data in cloud environments, stressing its benefits and demonstrating how well both technologies complement one another while also outlining the difficulties the two technologies must overcome.

Nabeel Zanoon et. al.,2017 concluded that both technologies interact in ways that are complimentary to one another after studying a number of significant angles. Two elements of an integrated approach in distributed network innovations are big data and cloud computing. Because their relationship is based on the product, the storage, andthe processing as a single component, cloud service providers are encouraged to continuously develop due to the growth of big data and its requirements. The cloud is the basket and big data is the product. Cloud computing capabilities are the focus of big data. Cloud computing, on the other hand, is dependent on massive amounts and varieties of data. Based on their interactions, a model was created. Cloud computing is versatile distributed resource surroundings that uses cutting edge approaches for data processing and management while minimizing expenses. All of these particularities show how closely big data and cloud computing are related. Both are developing quickly to keep up with shifts in technology demand and usage.

V. CONCLUSION

The amount of data has expanded over the past few decades and is still growing. Numerous sources produce data in various formats. As a result, there is an increasing diversity of data. Cloud computing services are practiced to work, analyse, and reposit data without requiring a physical location or the upkeep of costly computer hardware and software. We were able to spot that no prior study on BDA platforms and modeling tools has provided a meaningful statistical analysis. It has examined cloud computing and big data from several important angles and realized that their interactions are beneficial. In the realm of distributed network technology, big data and cloud computing combine to create an integrated approach. The product, storage, and processing are the common components of their connection; the outgrowth of big data and their needs is a reason that runs cloud service providers to consistently improve. The cloud serves as the container, and big data as the product. Big data is engaged with cloud computing abilities. On the other hand, massive data sources and types are important to cloud computing. A flexible distributed resource environment is what cloud computing represents.

VI. REFERENCES

[1]. Sarker, S., Arefin, M. S., Kowsher, M., Bhuiyan, T., Dhar, P. K., & Kwon, O. J. (2022). A Comprehensive Review on Big Data for Industries: Challenges and Opportunities. *IEEE Access*.

[2]. Nti, I. K., Quarcoo, J. A., Aning, J., & Fosu, G. K. (2022). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. *Big Data Mining and Analytics*, *5*(2), 81-97.,.

[3]. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, *47*, 98-115

[4]. Zanoon, N., Al-Haj, A., & Khwaldeh, S. M. (2017). Cloud computing and big data is there a relation between the two: a study. *International Journal of Applied Engineering Research*, *12*(17), 6970-6982.

[5]. Sandhu, A. K. (2021). Big data with cloud computing: Discussions and challenges. *Big Data Mining and Analytics*, *5*(1), 32-40.

[6]. Amiri, P. A. D., & Gavgani, M. R. (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.

[7]. Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014). *Big data: related technologies, challenges and future prospects* (Vol. 100). Heidelberg: Springer..

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

[9]. Ferguson, M. (2014). Data Virtualization–Flexible Technology for the Agile Enterprise. *Intelligent Business Strategies www. sas. com*.

[10]. O'Driscoll, A., Daugelaite, J., & Sleator, R. D. (2013). 'Big data', Hadoop and cloud computing in genomics. *Journal of biomedical informatics*, *46*(5), 774-781.

[11]. Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T. M. T., Dousse, O., ... & Miettinen, M. (2012). *The mobile data challenge: Big data for mobile computing research* (No. CONF).

[12]. J.J.Berman,Introduction,in:PrinciplesofBigData,MorganKaufmann,Boston,2013,xix–xxvi(pp).

[13]. Salloum, S., Huang, J. Z., & He, Y. (2019). Random sample partition: a distributed data model for big data analysis. *IEEE Transactions on Industrial Informatics*, *15*(11), 5846-5854.

[14]. Kong, L., Liu, Z., & Wu, J. (2020). A systematic review of big data-based urban sustainability research: State-of-the-science and future directions. *Journal of Cleaner Production*, *273*, 123142.