A Review of Literature on the Integration of Big Data and Cloud Computing

Mr. Vikram Bajaj[1*]

[1*]RNB Global University-Bikaner

***Corresponding Author:** Mr. Vikram Bajaj
[*]RNB Global University-Bikaner

Abstract:
In the foreseeable future, there will be a notable shift towards prioritizing data-centric resources in the technological landscape. It is imperative for Cloud Computing and Big Data infrastructures to enhance their security measures accordingly. Recent technological developments underscore the critical importance of data, influencing virtually all organizational operations. Cloud computing has played a pivotal role in enabling data storage, privacy protection, and the deployment of various Big Data applications. This paper aims to provide an overview of the potential and challenges associated with large data applications in cloud computing, emphasizing the need for efficient data processing and introducing conceptual models to address these demands.

Keywords: Cloud computing, Big Data.

I. INTRODUCTION

[1] Big data refers to methods for analysing, extracting information from, or generally dealing with data collections that are too massive or complicated for typical data processing application software to handle. As the term implies, big data refers to a large volume of data. 4V's can be used to explain typical data qualities. Simply described, cloud computing is the on-demand availabilityof computer system resources, primarily data storage and computational power. Cloud computing allows users to collaborate with others while accessing, using, working on, and modifying their work. Big data gives insights and information, while cloud computing allows users to work when and when they want. Characteristic analysis, storage management and cloud, big data processing, and finally obtaining insights (a piece of information) from the massive data available are all part of the analytics process. One of the most crucial things nowadays is digital security. Security is crucial in the context of big data since the data contains confidential information, secret keywords, and passwords that, if compromised, can have disastrous implications. As a result, when it comes to large data and cloud computing, security is crucial.. Security can be achieved in a variety of ways, includingNode Authentication, encryption, access control, and honeypot nodes, among others. Data storage, speed, security, processing, transmission, visualisation, architecture, integration, and quality are some of the issues that this system may face throughout implementation. Big data software packages make a broad collection of tools and options that allow an individual to map the wholedata landscape across the enterprise, allowing the individual to understand the internal dangers he or she confronts. This is seen asone of the most significant benefits, as large data ensures data security. Node Authentication, encryption, access control, honeypotnodes, and other methods can all be used to increase security. Data storage, speed, security, processing, transmission, visualisation, architecture, integration, and quality are some of the issues that may arise during the deployment of this system. The software packages for Big Data give a rich collection of tools and options that allow an individual to map the whole data landscape across the firm, allowing the individual to understand the threats he or she confronts internally. Big data keeps data safe, which is regarded one of the key

advantages.
The distributed software architecture gave rise to the cloud computing notion. The goal of cloud computing is to provide hosted services through the internet. Cloud computing in information technology has spawned a slew of new user groups and industries in recent years [1]. Cloud computing services are rendered through data centres located all over the world. Cloud computing servicesinclude Microsoft SharePoint and Google apps, to name a few.

II. BIG DATA

"Big Data" refers to data that is large, difficult to store, manage, and analyse using typical databases. For efficient storage, manipulation, and analysis, a scalable architecture is required. Smartphones and social media posts; sensors, such as traffic signals and utility metres; point-of-sale terminals; and consumer wearables, such as fit metres and electronic health records, all contributeto the vast volume of data. Various technologies are combined to uncover hidden values in this diverse, complicated data and turn it into actionable information, better decision-making, and a competitive edge. Big data has the following qualities.. Thecharacteristics of big data are.

A. Volume

III. Refers to the huge amounts of data generated every second from several sources such as social media, cell phones, autos,credit cards, M2M sensors, pictures, and videos, allowing users to data mine the hidden information and patterns in them.

A. Velocity

The rate at which information is generated, transported, collected, and analysed. Data is being generated at an ever-increasing rate,and the speed of transmission and access to the data must stay unchanged to enable for real-time access to the multiple applicationsthat rely on it.

B. Varity

This term refers to data that has been generated in a variety of formats, both structured and unstructured. Within the columns of a database, structured data such as name, phone number, address, financials, and so on can be organised. This type of information issimple to enter, store, query, and analyse. Unstructured data, which makes up 80% of today's data, is more difficult to sort and extract value from. Text messages, audio, blogs, photographs, video sequences, social media updates, log files, machine and sensordata are examples of unstructured data.

C. Veracity

This term refers to the data source's quality and reliability. Its significance is determined by the context and meaning it contributesto the study. Knowledge of the data's authenticity aids in a better understanding of the risks connected with data-driven analysis and business decisions.

IV. BIG DATA

1. Analysis Type - Whether the data will be analyzed in real time or in batches. For fraud detection, banks use real-time analysis, whereas batch production could be used for corporate strategic options.
2. Processing Methodology - Whether predictive, ad-hoc, or reporting approach is required depends on the business requirements.
3. Data Frequency - Determines how much of data is ingested and the rate of its arrival. Data could be continuous as in real-time feeds and also time series based.
4. Streams might be historical, transactional, or real-time data.
5. Data Format - Relational databases may hold structured data, such as transactions. NoSQL data warehouses can store unstructured and semi-structured data. The types of data stores that will be utilised to store and process them are determined by the formats.
6. Data Source - This identifies the source of the data, such as social media, machines, or human-generated

data. Determines
7. Data consumers - List of all users and applications which make use of the processed data [4].

V. CLOUD COMPUTING

Today internet is becoming an imperative tool in our day to day life, since the users are becoming abundant. In recent years cloud computing has emerged as a fundamental concept. The cloud makes use of computing resources both hardware and software whichis provided in the form of a service over the internet. Mobility, huge availability and cost efficiency are the factors responsible for making cloud computing popular in today's era. On the contrary it induces more threats to the safety of the organization's statisticsand information. Cloud computing gives huge computation power and storage capacity limit by means of using countless PCs together, empowering clients to send applications cost-effectively without heavy infrastructure investment. Cloud clients can decrease immense upfront investment of IT foundation and focus on their own core business. However, various potential clients are still reluctant to take advantage of cloud because of privacy and security concerns [6] [7].

The array of available cloud computing services is vast, but most fall into one of the following categories:
SaaS: As a Service (SaaS) The largest cloud market and most widely used business option in cloud services is software as a service.SaaS allows consumers to access applications via the internet. Third-party suppliers maintain SaaS applications, and the client uses a browser to access the interfaces. Because most SaaS applications run directly in the browser, the client is not required to download or install any software. Applications, runtime, data, middleware, OS, virtualization, servers, storage, and networking are all managed by the SaaS vendor, making it easy for businesses to streamline their maintenance and support.

PaaS: PaaS stands for Platform as a Service. The Platform as a Service approach makes hardware and software components availablethrough the internet, allowing developers to create unique applications. PaaS makes application development, testing, and deployment fast, easy, and cost-effective. This strategy enables businesses to design and develop applications that are incorporatedinto 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 applications are cloud-based, they are scalable and highlyavailable.
IaaS: Infrastructure as a Service Infrastructure as a Service cloud computing model provides self-servicing platform for accessing, monitoring and managing remote data center infrastructures such as compute, storage and networking services to organizations through virtualization technology. IaaS users are responsible for managing applications, data, runtime, middleware, and OS while providers still manage virtualization, servers, hard drives, storage, and networking. IaaS provides same capabilities as data centerswithout having to maintain them physically [6].

VI. LITERATURE

R. Kune et al. [4] studied big data computing model and detailed its technologies and characteristics. Later, the authors presented how a cloud computing platform would be utilized to host big data analytics. Additionally, the authors discussed an emerging big data computing platform over clouds. This study was very rich in term of defining the cloud computing model for big data, nevertheless, cloud computing QoS criteria weren't discussed nor considered during the study.

M. S. Al-Hakeem [5] proposed a new service model for big data on the cloud. The author defined a set of technologies required tointegrate big data with cloud computing in order to provide a new cloud service founded on the same basic cloud offerings (Infrastructure as a Service, Platform as a Service and Software as a Service). However, the author did not manage to discuss the various QoS aspects that would affect this new service model.

P. C. Neves et al. [6] described both cloud computing and big data systems where they focused on the issues yet to be addressed. In addition to that, the authors enumerated the issues related to big data computing model and determined the cloud computing solutions. Furthermore, the paper presents a cloud computing model that lists the cloud computing solutions advantages and disadvantages in case of a cloud service for big data.

C.X. Mavromoustakis et al. [7] paper exhibits the challenges related to big data and the opportunities that come with cloud computing model as a solution. Additionally, the authors carried out an expense analysis towards estimating the long term profits of implementing big data as a service on the cloud model. This study lacks the definition of the cloud computing QoS criteria and the SLA stipulations given that the authors considered reviewing the cost benefits of big data on the cloud.

N. Zanoon et al. [8] managed to elaborate a study on both cloud computing, big data paradigms and concluded that the relationshipbetween them is complementary. The authors developed a compatibility chart between big data and cloud computing in term of features in the interest of revealing the comprehensive relationship between big data and cloud computing.

Neelay Jagani et al. [8] have present the Big Data implementation and application in Cloud Computing. 4 V’s in big data can be applied in Cloud computing to get better performance, higher input details, better insights, reliable and secure platforms at comparatively lower costs. Different analytics, technology involved in coupling of big data with cloud computing, the challenges involved in this process, trends applications of the domain and security factors involved are discussed in this paper.

Bala M. Balachandran et al In this paper, we outline the benefits and challenges involved in deploying big data analytics through cloud computing. We argue that cloud computing can support the storage and computing requirements of big data analytics. We discuss how the consolidation of these two dominant technologies can enhance the process of big data mining enabling businessesto improve decision-making processes. We also highlight the issues and risks that should be addressed when using a so called CLaaS, cloud-based service model.

VII. CHALLENGES

In spite of all the advantages of the integration between cloud computing and big data, there are some challenges and risks thatought to be thought while deploying big data on a cloud environment.

A. Data storage

As a result of technology improvements, we are seeing an exponential increase in data. However, due to a lack of storage space, themajority of the generated data is ignored or deleted. As a result, the primary obstacles for Big Data analysis are storage mediums and quicker communication speeds. The present storage solutions lack the processing power required to handle Big Data. Standardphysical storage systems make it difficult to save data since hard disc drives fail frequently and traditional data protection techniquesare ineffective.. In addition to this, the velocity of Big Data must be such that the storage systems must be able to scale up quicklywhen required, which is actually difficult to achieve with these traditional storage systems. Due to this ever growing data, data mining tasks has increased considerably which has led to wide diversity of data. There’s a need to pay more attention for designingstorage systems and to make efficient data analysis tools that will provide guarantees on the output since the data is gathered fromdifferent sources. Moreover, machine learning algorithms can be designed for analyzing the data which will help in improving theefficiency and scalability. The unlimited storage along with high fault tolerance offered by Cloud storage services (such as: AmazonS3, Elastic Block Store) provides solutions to address Big Data storage challenges. But, it’s very expensive to host and transfer BigData on the cloud since the size of data is gigantic.

B. Data Transmission

Another challenge is how to move vast amounts of big data (let’s take for example hundreds of terabytes of data) into a public cloudin a short period of time? How will we deal with the storage, reliability, privacy, and security issues? Transferring gigantic volumesof data in different stages of data life cycle poses challenges in each of these stages. Therefore, we need to devise smart pre- processing techniques and data compression algorithms to effectively reduce the data size before transferring the data. For transferring data from local data centers to cloud platforms, we need to develop efficient algorithms which will automatically recommend the appropriate cloud service (location) based on the geotemporal principles (since data can be at different locations) to maximize the data transfer speed while the minimizing cost.

C.Computational complexities

For processing large volumes of data, we require dedicated computing resources, which we usually handle by the increasing speed of storage, network and CPU. However, the processing power and the computing resources provided by the traditional computing system is insufficient for processing the data. The virtually unlimited and on-demand processing power offered by cloud computing acts as a partial solution. However, shifting to the cloud results in some issues. First, the network bandwidth of cloud computing isvery limited which affects the efficiency of computation over large volumes of data. Second, the data is dispersed at different locations which makes it difficult to gather it for pre-processing. The essential features of cloud computing such as virtualization, pooled resources of data and high computing power makes it a difficult task to track and ensure data locality, and hampers its abilityto support data processing which involves intensive communication and exchange of data.

D. Data securities

Some security vulnerabilities arise due to the integration of Big Data and Cloud Computing. Also, the data security policies and schemes work with the structured data which is stored in conventional DBMS and aren't effective in handling highly heterogeneousand unstructured data. Therefore, we need to make effective policies for data access control and safety management so as to incorporate new data management systems and storage structures. Ensuring data confidentiality, integrity and availability becomes elemental in this cloud era since the data owners have limited control over the data and various resources. Heterogeneity is one of the most known Big Data's cloud security vulnerability. In many cases the deployment of Big Data requires it to deploy on a new cloud platform which will need new security tools to be developed as the existing security tools and practices won't work for such platforms. These security tools should include encryption, authentication, intrusion detection, access control, monitoring and event logging. Along with the security policies, while integrating Big Data to the cloud environment, consolidation plans should be taken into consideration.

E. Data privacy

Some security vulnerabilities arise due to the integration of Big Data and Cloud Computing. Also, the data security policies and schemes work with the structured data which is stored in conventional DBMS and aren't effective in handling highly heterogeneousand unstructured data. Therefore, we need to make effective policies for data access control and safety management so as to incorporate new data management systems and storage structures. Ensuring data confidentiality, integrity and availability becomeselemental in this cloud era since the data owners have limited control over the data and various resources. Heterogeneity is one of the most known Big Data's cloud security vulnerability. In many cases the deployment of Big Data requires it to deploy on a new cloud platform which will need new security tools to be developed as the existing security tools and practices won't work for such platforms. These security tools should include encryption, authentication, intrusion detection, access control, monitoring and eventlogging. Along with the security policies, while integrating Big Data to the cloud environment, consolidation plans should be taken into consideration.

VIII. CONCLUSION

Big data and cloud computing play a huge role in the current digital world. The application of Big Data in Cloud Computing seemsto have a huge potential in the coming years. While using Software as Service, typically, big data plays a pretty important role in giving insight, in cloud computing applications. Big Data when applied in cloud computing, has many applications in different fields. Some of these applications include improved analysis due to large data size, creation of an efficient infrastructure while reducing the cost in the long run and allowing better integrity and availability and security of the cloud platform, letting the businesses and platforms grow through the means of big data.In the big data era of innovation and competition driven by advancements in cloud computing has resulted in discovering hidden knowledge from the data. In this paper we have given an overview of big data applications in cloud computing and its challenges in storing, transformation, processing data and some gooddesign principles which could lead to further research.

REFERENCES

1. R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige, R. Buyya. “The anatomy of big data computing”. Software Practice and Experience, Wiley Online Library. 9 October 2015.
2. P. C. Neves, B. Schmerl, J. Bernardino, J. Camara. “Big data in cloud computing: features and issues”. In Proceedings of the 2016 International Conference on Internet of Things and Big Data. 23-25 April 2016.
3. G. Skourletopoulos, C. X. Mavromoustakis, G. Mastorakis, J. M. Batalla, C. Dobre, S. Panagiotakis, E. Pallis. “Big data and cloud computing: A survey of the state-of-the-art and research challenges”. Advances in Mobile Cloud Computing and Big Data inthe 5G Era, Studies in Big Data 22, Springer International Publishing, 2017.
4. N. Zanoon, A. Al-Haj, S. M. Khwaldeh. “Cloud computing and big data is there a relation between the two: A Study”. International Journal of Applied Engineering Research. Volume 12, 2017.
5. Bala M. Balachandran, Shivika Prasad, “Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence”, International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, Marseille, France, Procedia Computer Science 112 (2017) 1112–1122, 2017.
6. Neelay Jagani, Parthil Jagani “BIG DATA IN CLOUD COMPUTING: A LITERATURE REVIEW” International Journal of Engineering Applied Sciences and Technology, 2021 Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 185-191,2021.
7. Sivarajah, Uthayasankar, et al. "Critical analysis of Big Data challenges and analytical methods." Journal of Business Research 70 (2017): 263-286.Ji, Changqing, et al. "Big data processing in cloud computing environments." 2012 12th internationalsymposium on pervasive systems, algorithms and networks. IEEE, 2012.