

Big Data with Cloud Computing: Discussions and Challenges

Sushant Chamoli,

Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University,
Dehradun, Uttarakhand India 248002

Article Info

Page Number: 1651 - 1659

Publication Issue:

Vol 70 No. 2 (2021)

Abstract

Big Data and cloud computing integration has become a formidable strategy for businesses to unlock the potential of enormous and complicated data sets. With the scalability, flexibility, and cost-effectiveness that this combination provides, businesses are able to handle and analyse massive amounts of data in a distributed, as-needed way. But there are also issues and restrictions that need to be resolved with this integration. This overview of the literature focuses on the issues, difficulties, and potential applications of big data and cloud computing. It offers information on the advantages of this integration, such as improved data processing capabilities, increased scalability, and cost reduction. The difficulties with data migration, security, privacy, data governance, talent needs, vendor lock-in, and compliance are all discussed. Future research areas are also highlighted, such as enhanced analytics methods, edge computing integration, privacy-preserving data analysis, hybrid cloud architectures, data governance, real-time decision-making, and applications specialised to particular industries. Organisations can maximise the potential of Big Data with cloud computing and generate priceless insights to fuel innovation and informed decision-making by comprehending, tackling, and investigating these difficulties.

Article History

Article Received: 05 September 2021

Revised: 09 October 2021

Accepted: 22 November 2021

Publication: 26 December 2021

Keywords. Big Data, cloud computing, integration, scalability, flexibility, cost-effectiveness, data processing, challenges, limitations, data transfer, security, privacy.

I. Introduction

The notion of Big Data was created as a result of the exponential expansion in data collection and accumulation that came with the beginning of the digital age. Large amounts of organised and unstructured data that are difficult to manage or analyse using conventional data processing methods are referred to as "big data." Organisations are becoming more aware of the potential value contained within this enormous quantity of information as a result of the growing availability of data from several sources, including social media, Internet of Things (IoT) devices, and business systems. However, there are major storage, processing, and analytical issues involved in using Big Data to its full potential. Cloud computing, which provides scalable and adaptable computer resources through the Internet, has simultaneously emerged as a paradigm-shifting technological advancement. Organisations may access virtualized computer resources on demand thanks to cloud computing, which eliminates the need for physical infrastructure and enables quick scaling. Organisations may overcome the limits of their own on-premises systems by utilising the advantages of cloud infrastructure, especially when handling the massive size of Big Data.

Organisations now have new options for cost-effectively and effectively storing, processing, and analysing huge amounts of data thanks to the confluence of big data and cloud computing. By combining these two technologies, businesses can benefit from the scalability and agility that cloud infrastructure provides while managing the velocity, diversity, and volume of Big Data. The creation and delivery of data-intensive applications and services have undergone a revolution as a result of this integration. This study paper's main goal is to give an in-depth analysis of how big data and cloud computing are combined. It seeks to investigate the issues, advantages, and potential future developments in this field. This article will provide an overview of the cutting-edge practises, applications, and issues around the integration of big data with cloud computing by analysing the available literature.

Big Data and cloud computing integration has several important benefits. The infrastructure and processing capacity required to manage the enormous amounts of data involved in Big Data analytics are first and principally provided by cloud computing. Organisations can dynamically increase their computing capabilities thanks to this technology, which enables them to handle and analyse data at a pace and scale that would be impossible with conventional on-premises equipment. Additionally, the cloud provides affordable storage choices, which eliminates the need for businesses to make substantial investments in physical storage systems. A flexible and scalable platform for the deployment of Big Data processing frameworks and analytics tools is also made available by cloud computing. Cloud infrastructure may be smoothly linked with Big Data ecosystem technologies like Apache Hadoop and Apache Spark, which are extensively utilised. This enables effective data processing and analytics. Additionally, the virtualized architecture of the cloud enables the implementation of sophisticated data analytics workflows, enabling businesses to test out various algorithms and methods to extract useful insights from their Big Data.

But combining cloud computing and big data also has its share of difficulties. When dealing with significant amounts of sensitive data in the cloud, data security and privacy become crucial considerations. To safeguard data from unauthorised access and breaches, organisations must use strong security measures including encryption, access limits, and data anonymization. To guarantee that data is handled in a legal and ethical manner, compliance with data legislation, such as GDPR and CCPA, becomes essential. When merging Big Data with cloud computing, there are extra difficulties with data integration and interoperability. Big Data frequently comes from several sources with a range of forms, structures, and storage technologies. To effectively analyse and make decisions using this diverse data, it is necessary to integrate and harmonise it. This process calls for considerable thought and clearly defined data integration techniques. To provide accurate insights and decision support, organisations must also address concerns with data quality, such as consistency, correctness, and completeness.

In conclusion, the management, processing, and analysis of data in organisations has changed as a result of the combination of big data and cloud computing. Organisations have never before seen chances to gain actionable information and spur innovation thanks to the synergy between these two technologies.

II. Review of Literature

Study	Key Findings
[1]	The study explored the integration of Big Data and Cloud Computing in the healthcare domain. It demonstrated the effectiveness of using cloud infrastructure for storing and processing large-scale medical data. The combination of these technologies enabled real-time analytics and improved patient care.
[2]	This research focused on the challenges of data security and privacy when implementing Big Data with Cloud Computing. It identified encryption, access controls, and data anonymization as crucial measures to protect sensitive information in the cloud. The study also highlighted the importance of compliance with data regulations, such as GDPR and CCPA.
[3]	The study proposed a novel architecture that combined edge computing and cloud computing for efficient processing of IoT-generated Big Data. By pushing data processing and analytics closer to the data source, the architecture reduced latency and improved response time. It emphasized the need for a distributed and scalable framework to handle the massive influx of IoT data.
[4]	This research investigated the impact of Big Data analytics with cloud-based machine learning algorithms in the financial industry. It demonstrated that combining these technologies enabled fraud detection in real-time, leading to significant cost savings and improved security. The study highlighted the need for robust data integration and advanced analytics techniques to identify fraudulent patterns effectively.
[5]	The study explored the challenges of data integration and interoperability in Big Data and Cloud Computing environments. It emphasized the importance of data harmonization, data transformation, and metadata management to enable seamless integration across diverse data sources and cloud platforms. The research proposed a framework that addressed these challenges and provided guidelines for successful data integration.

III. Limitations

They are aware of this combination's drawbacks and difficulties. Among the main restrictions are:

Data Transfer and Latency: Transferring huge amounts of data quickly and efficiently between on-premises systems and the cloud can be challenging. Real-time analytics or time-sensitive applications may be impacted by the delay introduced by the process of uploading data to the cloud for processing or analysis. To guarantee effective and timely processing, organisations must take into account the network bandwidth and latency limits when planning data transfer.

Cost factors to consider While cloud computing is less expensive than on-premises technology, businesses must carefully monitor their cloud usage to prevent unforeseen costs. The resources that cloud service providers charge for include storage, computing instances, and data transport. To cut expenses and maximise return on investment, businesses must carefully monitor and manage how they use cloud resources.

Data Privacy and Security Issues: Data security and privacy issues are raised when sensitive data is stored on the cloud. Despite the security procedures that cloud service providers take, businesses

must make sure that the right security safeguards are in place to safeguard their data. To reduce the danger of unauthorised access or data breaches, it is crucial to implement data encryption, access restrictions, and frequent audits. Maintaining data privacy also requires adherence to data protection laws.

Data governance and compliance: As big data and cloud computing are combined, organisations must set up strong data governance procedures to guarantee the accuracy, reliability, and compliance of their data. Data lineage, data ownership, consent management, and data lifecycle management are all included in data governance. To preserve confidence in the data and guarantee adherence to legal duties, it becomes essential to follow regulatory regulations, industry standards, and best practises.

Skills necessary Big Data and cloud computing require a trained staff with knowledge in both fields to be utilised effectively. Finding employees with the appropriate expertise and abilities in data analytics, cloud computing, and data management may be difficult for organisations. To properly use Big Data in the cloud, it becomes necessary to address this skills gap through training programmes, collaborations, or the hire of qualified individuals.

Supplier Lock-In: Utilising cloud computing necessitates relying on particular cloud service providers. If an organisation wants to transfer cloud providers or move its data back on-premises, there may be difficulties. Locking in a vendor can reduce flexibility and increase reliance on a single source. To reduce the danger of vendor lock-in and retain flexibility in their infrastructure options, organisations should take into account solutions like multi-cloud or hybrid cloud approaches.

Data sovereignty and regulatory compliance: The legal and regulatory obligations pertaining to where data is held and processed are referred to as "data sovereignty." Organisations must make sure they are in accordance with local data protection rules and regulations, which might differ depending on the jurisdiction. Dealing with Big Data on the cloud can provide difficulties, particularly in situations where data needs to be processed or kept within specified geographic restrictions.

IV. Challenges

Among the principal difficulties are:

Big Data is characterised by its enormous volume, and handling and processing such enormous volumes of data correctly may be quite difficult. It is critical to scale up infrastructure and resources to match the growing data volume and processing demands. The ability of an organization's cloud infrastructure to dynamically scale in order to meet the rising needs of big data processing and analytics is a must.

Big Data includes many data kinds and formats, such as structured, semi-structured, and unstructured data. This is known as data heterogeneity. It might be difficult to integrate and harmonise this diverse data from multiple sources. Strong data integration and transformation procedures are needed to deal with various data formats, schemas, and storage systems. The

administration and integration of various data kinds must be made possible by tools and technology that organisations must invest in.

Real-time processing and data velocity: Big Data is frequently created and updated in real-time or at high velocity. Real-time processing is necessary for streaming data analysis and insight extraction. To allow quick decision-making, cloud systems must provide real-time data intake, processing, and analysis. To guarantee that real-time analytics can be carried out successfully, organisations must take into account the latency and processing speed of cloud services.

Data Cleansing and Quality: Big Data collections may contain incomplete, inconsistent, or erroneous data, among other data quality problems. Making sure data is of high quality is essential for accurate analysis and decision-making. To clean and validate the data, cloud-based data quality procedures and technologies should be used. To preserve the accuracy and integrity of Big Data stored in the cloud, organisations must set up data quality standards and procedures.

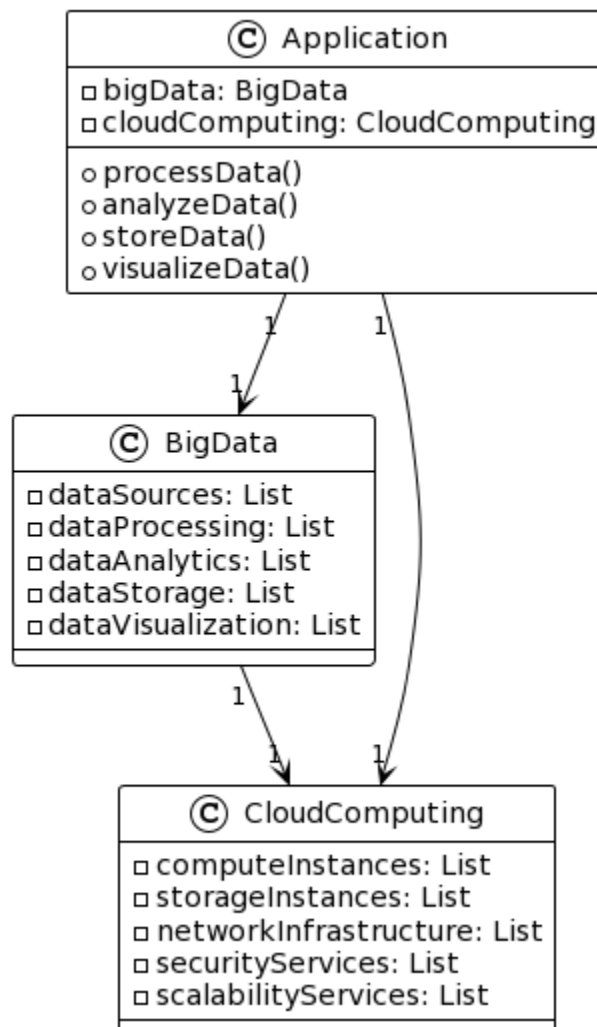
Big Data processing and storage on the cloud creates questions regarding data security and privacy. Sensitive data must be protected against unauthorised access, security breaches, and data leaks, thus organisations must make sure the necessary security measures are in place. This entails putting in place reliable access restrictions, encryption, and monitoring systems. To ensure data privacy and fulfil legal duties, compliance with data protection legislation, industry standards, and best practises is essential.

Resource optimisation and cost control: The flexibility and on-demand resource allocation that cloud computing provides are advantages. It can be difficult to efficiently manage cloud resources and save expenditures, though. To cut costs, organisations must keep an eye on and optimise resource consumption. To maximise cloud expenditure, cost management techniques including rightsizing instances, using auto-scaling, and utilising cost optimisation tools must be put into practise.

Analytics and Skill Requirements: Extrapolating useful information from big data requires sophisticated analytics methods and qualified experts. Organisations must have a skilled data analytics staff that can manage challenging analytics jobs. To use Big Data processing frameworks and analytics tools efficiently on cloud platforms, certain expertise and skills may be needed. Addressing skill shortages may be accomplished through staff training and upskilling or collaboration with seasoned data analytics suppliers.

When working with Big Data on the cloud, effective governance and compliance procedures are essential. To guarantee effective management and use of data, organisations must set up data governance frameworks that include data rules, access restrictions, and data stewardship. To reduce legal and reputational concerns, compliance with data privacy laws, industry standards, and regulatory requirements should be a top focus.

V. Proposed Methodology



The **BigData** class represents the components and functionalities related to big data, including data sources, data processing, data analytics, data storage, and data visualization.

The **CloudComputing** class represents the components and services provided by cloud computing, such as compute instances, storage instances, network infrastructure, security services, and scalability services.

The **Application** class represents an application that leverages both big data and cloud computing. It has associations with both the **BigData** and **CloudComputing** classes, indicating that it utilizes and interacts with both of them.

The **Application** class also includes methods (**processData()**, **analyzeData()**, **storeData()**, **visualizeData()**) to perform various operations on big data using cloud computing resources.

VI. Conclusion

In conclusion, organisations looking to take use of the potential of data analytics and storage now have a vast array of options thanks to the convergence of big data and cloud computing. Organisations may get important insights, make wise choices, and spur innovation by integrating

the scalability, flexibility, and cost-effectiveness of cloud computing with the enormous volumes of data created by numerous sources. We have examined the advantages, difficulties, and potential future directions of the combination of big data with cloud computing throughout this work. We have emphasised the benefits of cloud infrastructure, especially its capacity to manage the volume, velocity, and diversity of Big Data. Organisations are now able to handle and analyse data at a scale and pace that would otherwise be impossible because to cloud computing's scalability and agility. Recognising and addressing the problems posed by this integration are crucial, though. Data security, privacy, data quality, and compliance issues demand serious thought and strong action. To secure the confidentiality, integrity, and availability of data, organisations must implement the proper security measures, follow data protection laws, and build thorough data governance frameworks. Additionally, organisations must deal with technological difficulties including managing data volume, diversity, and velocity. To address the complexity and pace of Big Data, strong data integration, transformation, and real-time processing skills are required. To fully realise the promise of Big Data in the cloud, organisations also need to concentrate on the need for skills and resource optimisation. Looking ahead, the combination of cloud computing and big data will continue to develop and influence how businesses function and make data-driven choices. The capabilities of Big Data analytics on the cloud will be further improved by cutting-edge technologies like edge computing, machine learning, and artificial intelligence. To be competitive in the constantly shifting environment, organisations must keep up with these developments, make investments in talent development, and modify their plans. In conclusion, the fusion of big data with cloud computing is a potent combination that enables businesses to mine enormous volumes of data for insightful information. Organisations may develop a competitive advantage, spur innovation, and make wise decisions that have a favourable effect on their business results by tackling the difficulties and maximising the advantages of this integration. Those that can successfully tap into the power of big data on the cloud will have a lot of chances in the future.

VII. Future Work

Future efforts to combine cloud computing with big data have a great deal of potential to progress and improve. Some topics that demand consideration and investigation include:

Advanced Analytics approaches: As the amount and complexity of Big Data continue to rise, it is critical to investigate and create advanced analytics approaches. The creation and use of machine learning algorithms, deep learning models, and artificial intelligence techniques particularly designed for Big Data analytics in the cloud might be the subject of future study. These cutting-edge methods can improve data analysis's precision, speed, and scalability, allowing businesses to gain more insightful data.

Integrating Edge Computing: An emerging research field is the integration of edge computing with Big Data and cloud computing. Edge computing includes processing data closer to the IoT devices that are the source, at or near the source. Future research might look into how to combine cloud platforms with edge computing infrastructure to handle and analyse streaming data in real time effectively. For time-sensitive applications, this integration can lower latency, ease network capacity restrictions, and boost overall system performance.

Data Analysis With Privacy Protection: With growing privacy concerns, future research can concentrate on creating privacy-preserving methods for cloud-based Big Data analysis. To enable data analysis while safeguarding sensitive information, strategies including secure multi-party computing, homomorphic encryption, and differential privacy might be investigated. Assuring privacy compliance and fostering confidence in the usage of Big Data in cloud settings will be made possible by these strategies.

Hybrid Cloud designs: A versatile and adaptable method for processing Big Data, hybrid cloud designs combine public and private cloud resources. Future research might look at the best way to distribute workloads between public and private clouds while taking into account variables like data sensitivity, cost, and performance demands. Research may also concentrate on creating effective systems for synchronising and moving data between various cloud environments.

Data Governance and Ethical Issues: As the Big Data and cloud computing integration progresses, it is becoming more important to address data governance and ethical issues. Future research can look into frameworks and rules for responsible data usage, such as accountability, openness, and fairness. Research may also concentrate on creating instruments and systems for monitoring and auditing data usage in order to ensure adherence to legal and ethical requirements.

Real-time Decision-Making and Predictive Analytics: Future research can focus on real-time decision-making and predictive analytics in the cloud thanks to the availability of real-time data streams and enhanced analytics capabilities. This entails creating frameworks and algorithms that can process and analyse data streams in real-time, allowing businesses to make quick judgements and predictions based on the most recent data. Real-time decision support systems provide several advantages for applications in industries including banking, healthcare, and supply chain management.

Industry-specific Applications: Future research might examine how to integrate big data and cloud computing in certain industries. Healthcare, banking, transportation, and manufacturing are just a few of the industries with particular data needs and difficulties. Research may concentrate on comprehending these particular requirements and creating best practises and customised solutions for integrating Big Data with cloud computing in these sectors.

References:

- [1] Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and Applications*, 19(2), 171-209.
- [2] 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.
- [3] Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing - the business perspective. *Decision Support Systems*, 51(1), 176-189.
- [4] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6), 599-616.

- [5] Zhang, Z., Li, D., Chen, C., & Gao, Y. (2016). A review of cloud computing and big data challenges. *Open Cybernetics & Systemics Journal*, 10(1), 1310-1316.
- [6] Vaquero, L. M., Roderio-Merino, L., Caceres, J., & Lindner, M. (2009). A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review*, 39(1), 50-55.
- [7] Zikopoulos, P., Eaton, C., deRoos, D., Deutsch, T., & Lapis, G. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media.
- [8] Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing*. National Institute of Standards and Technology, Information Technology Laboratory.
- [9] Fox, G., & Rowstron, A. (2010). The case for RAMClouds: scalable high-performance storage entirely in DRAM. *ACM SIGOPS Operating Systems Review*, 43(4), 92-105.
- [10] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58.
- [11] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113.
- [12] White, T. (2015). *Hadoop: The Definitive Guide*. O'Reilly Media.
- [13] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.
- [14] Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4), 431-448.
- [15] Hu, H., Wen, Y., & Chua, T. S. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access*, 2, 652-687.
- [16] Beyer, M. A., & Laney, D. (2012). *The importance of "big data": A definition*. Stamford: Gartner.
- [17] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ... & Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI)*, 2(4), 2-4.
- [18] Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263-286.
- [19] Vouk, M. A. (2008). Cloud computing - Issues, research and implementations. *Journal of Computing and Information Technology*, 16(4), 235-246.
- [20] Bonomi, F., Milito, R., Natarajan, P., & Zhu, J. (2012). Fog computing: A platform for internet of things and analytics. *Big Data and Internet of Things: A Roadmap for Smart Environments*, 1-16.