World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Security issues analysis based on big data in cloud computing

Vedaprada Raghunath *

IT Director, IMR soft LLC, USA.

World Journal of Advanced Research and Reviews, 2024, 23(03), 2549–2557

Publication history: Received on 14 August 2024; revised on 24 September 2024; accepted on 26 September 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.3.2913

Abstract

This study will investigate the security problems that are linked with the Hadoop ecosystem, Big Data, Map Reduce, and cloud computing. The objective of this research is to analyze these concerns. Specifically, the primary emphasis is placed on the security concerns that are linked with big data in cloud computing. Big data applications have a lot to offer organizations, companies, and a broad range of industries, no matter how big or little. There have been issues with Hadoop and cloud computing security, therefore we also look at several possible solutions to these difficulties. The field of cloud computing security is experiencing fast growth in several areas, including computer security, network security, information security, and data privacy. Data, apps, and the infrastructure related to them are greatly helped by cloud computing's use of rules, technologies, controls, and big data tools to keep everything safe. Researchers also anticipate that cloud computing, big data, and their associated benefits will represent the most exciting new areas of study.

Keywords: Cloud Computing; Big Data; Hadoop; Map Reduce; HDFS (Hadoop Distributed File System)

1. Introduction

The rate at which data is produced has been increasing at an exponential rate over the past few years. In the early phases, many companies are seeking efficient ways to collect, store, and analyze the massive volumes of data generated by various sources, including high throughput instruments, sensors, and networked devices. To get there, big data technologies may reap the benefits of cloud computing in a large way. Virtualized resources can be assembled, connected, configured, and reconfigured on demand with the use of automated technologies. These facilitate the easy implementation of cloud services, which in turn makes it much easier for organizations to accomplish their goals.

Many people are worried about the security and privacy implications of cloud computing, which includes issues like multitenancy, trust, accountability, loss of control, and the paradigm shift that comes with using the cloud. Therefore, it is expected that cloud platforms that handle sensitive data in big quantities will put in place organizational and technical safeguards to prevent data protection breaches that could cause costly and devastating damage.

The phrase "sensitive information" is used in discussions about cloud computing to describe data that originates from a broad range of industries and academic disciplines. One common type of sensitive data stored in cloud computing systems is health-related data. Most people will naturally want their health records kept private. Consequently, there has been a rise in the need for privacy and data security measures to prevent individuals from being monitored or having their database records leaked due to the proliferation of new cloud technologies. Data Protection Directive (DPD) and Health Insurance Portability and Accountability Act (HIPAA) are two examples of protective laws from the EU and the US, respectively. Both of these pieces of legislation require the protection of confidential information when it comes to the management of personally identifiable information.

* Corresponding author: Vedaprada Raghunath

2. Cloud Computing

The phrase "cloud computing" describes an approach to application management that does not rely on individual devices or local servers but rather on the sharing of computing resources. As "The Internet" is what the term "Cloud" means when discussing cloud computing, the term "cloud computing" describes a method of delivering computer services via the Internet's interconnections. Utilizing the ever-increasing processing power to execute millions of instructions per second is a key goal of cloud computing.

To spread the processing of data across the servers, cloud computing makes use of networks that consist of a wide set of computers that are connected to one another in a particular manner. This approach requires just a single software application to be installed on each computer, as opposed to installing an entire software suite. This software serves as both a login mechanism for a web-based service and a host for all of the user's necessary apps. There is a major change in the necessary workload when employing a cloud computing system.

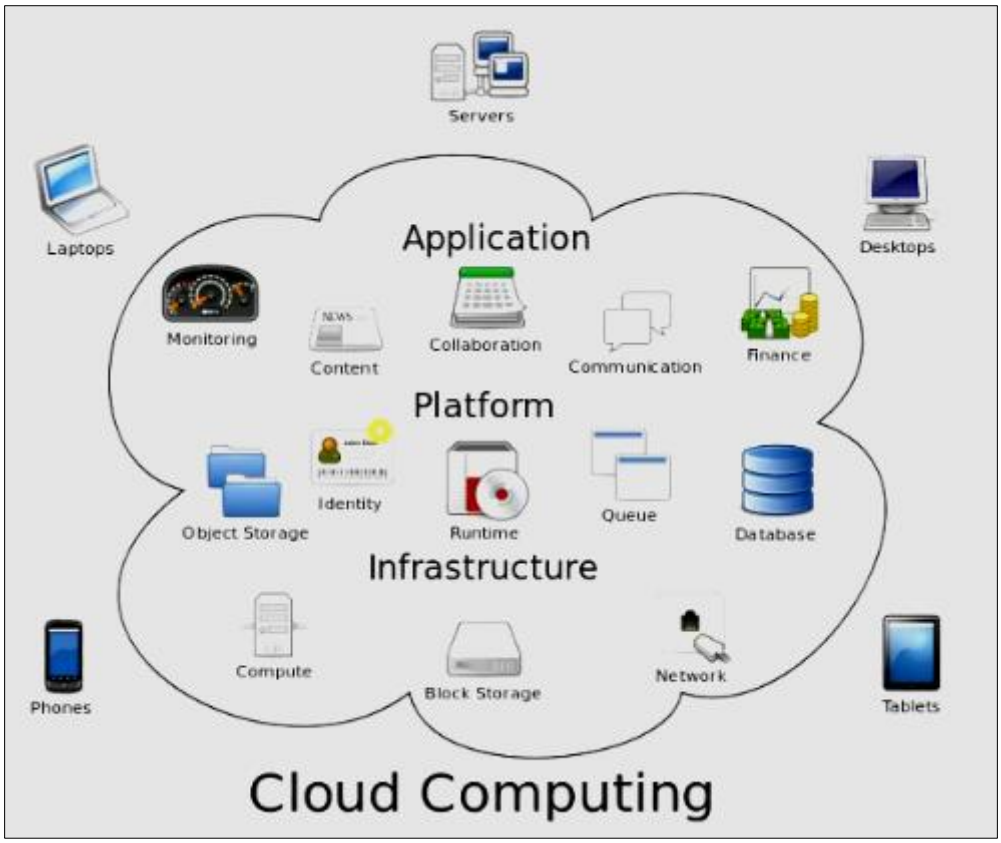

Figure 1 Cloud Computing

Local computers don't have to take the whole load when it comes to program execution anymore. There has been a recent shift toward using cloud computing shown in figure 1to cut costs associated with computing resource use. The cloud network, instead, is a collection of interconnected computers that takes care of everything. Users can save money on software and hardware due to this adjustment.

3. Literature review

3.1. Big Data and their major characteristics

The fact that cloud computing may provide consumers with a number of useful functions is something that we are already aware of [1-6]. Due to the unique nature of cloud computing, both consumers and service providers are keen to build a more secure cloud environment. Upon further inspection, it becomes apparent that cloud computing makes use of the following delivery data models [7]: Software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS) are the three main classes of cloud computing services. Cloud Computing Features (CCF) comprise,

among other things, the following: computational capability (CCF5), energy efficiency (CCF4), storage (CCF1), service (CCF2), applications (CCF3), and internet-based storage (CCF2). These days, data is being generated at an exponential rate across many industries, including manufacturing, commerce, science, and even people's private lives.

According to [8,9], if these data are handled correctly, they have the potential to provide new information about the economy, society, and environment. Furthermore, they can enable individuals to quickly adapt to new possibilities and changes.

Moreover, traditional data processing tools like database sets and data warehouses are becoming inadequate to manage the massive amounts of data that are required. Big Data (BD) describes these challenges, and it offers a new field of study for researchers to explore [10]. To be "a big thing in the field of modern technologies" is one way to characterize "Big Data" [11]. Under the BD umbrella are the five Vs: volume, velocity, variety, veracity, and value (shown in Figure 2).

There is a particular origin or "source" for each and every piece of data that is associated with the phrase "Big Data," which, depending on Variety, could give rise to a variety of data kinds. An attempt is made in this article to go over the major Big Data sources (BDSs) and the problems connected with each of them in terms of Big Data usage generally.

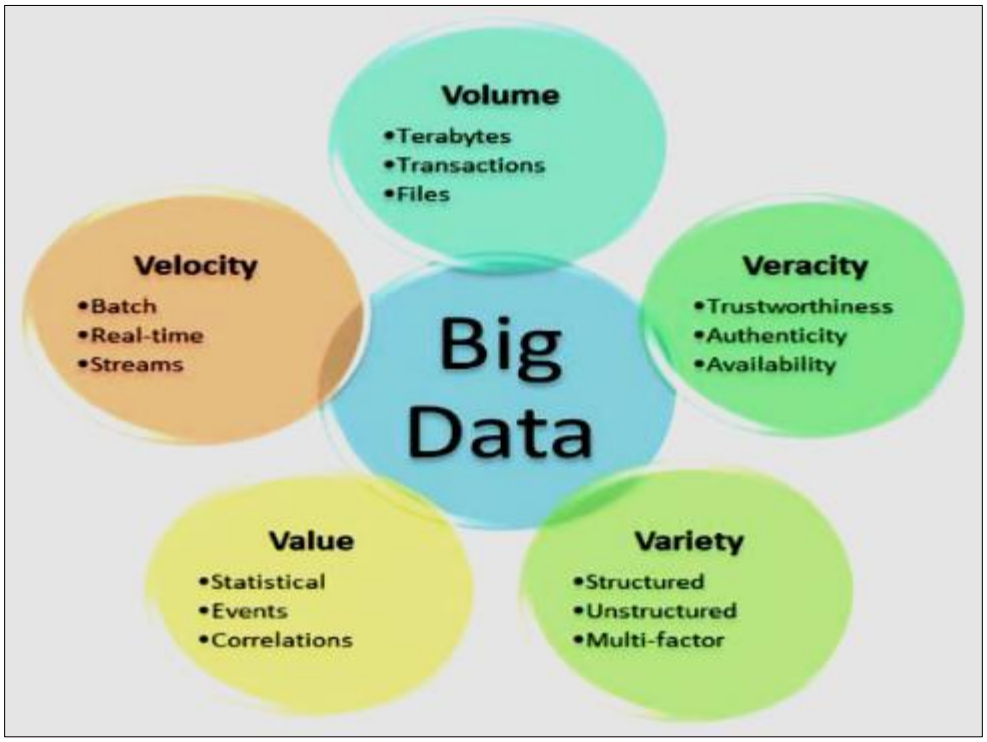

Figure 2 Big Data's 5 Vs and what makes them unique

- **BD**S1—Earth, Marine, and Space Sciences: The integration of sensor and computational simulation technologies has made it possible to present, monitor, and analyze complex earth, sea, and space systems, leading to the generation and collection of large data sets at different space-time scales every second for operational purposes. As an example, software used for earth, sea, and space observation gathers terabytes of photos every day [12], with spectrum analysis, space, and time processing gradually increasing [13-14].
- **BD**S2—Internet of Things: All the things that could link to the internet and communicate with one another are collectively known as the Internet of Things (IoT) [14,15,18]. Any and all data collected by Internet of Things (IoT) sensors can be characterized as BD because it contains both spatial and temporal information. New prospects and the rapid growth of Smart Cities could be achieved by the integration of IoT-BD in network environments and the use of technologies like Cloud Computing [15,19,21].
- **BD**S3—Social Sciences: The social sciences may undergo a paradigm shift as a result of the Big Data produced by platforms like Instagram, Twitter, and Facebook [16, 22,23].

- **BD**S4—Business: Big Data-related strategy, optimization, and competitive decision-making could benefit from corporate analytics and data. There is a lot of potentially dangerous geographical information in the data associated with the earlier cases, such as the exact time and location of a transition [17,24,25].
- **BD**S5—Industry: Industry 4.0, also known as the fourth industrial revolution, is characterized by its products and manufacturing systems that leverage technologies such as IoE-based BD to establish autonomous ad hoc networks [18,26,27,28,29].

4. Benefits of big data security

By ensuring big data is secure, businesses can use big data to its maximum capacity while reducing vulnerability, increasing confidence, and feeling innovation and growth. The main advantages of large data security will be examined in figure 3.

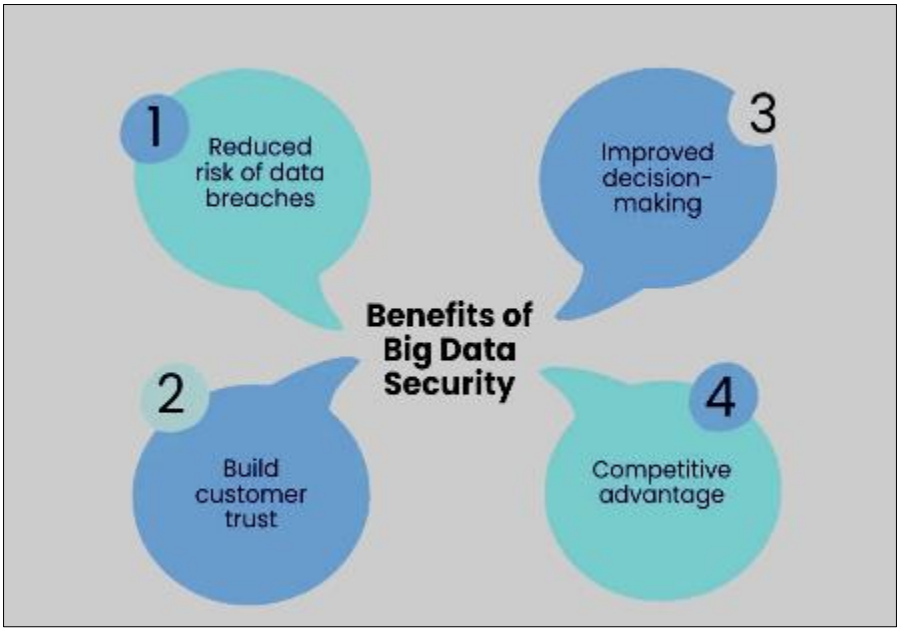

Figure 3 Benefits of Big Data security

4.1. Reduced risk of data breaches

Big data security uses multiple layers of protection to keep critical information private, legitimate, and easily accessible, reducing the likelihood of data breaches. Security measures like as encryption, role-based access control, threat detection, and real-time monitoring can significantly lessen the occurrence of data breaches. Implementing firewalls, intrusion detection systems (IDS), and intrusion prevention systems (IPS) into big data security solutions further reduces the probability of data breaches by monitoring the network and stopping any suspicious actions.

4.2. Increased customer trust

Building trust with customers in today's digital world is absolutely dependent on data security. The frequency of data breaches has made consumers wary of companies that collect and use their private information. Statista reports that just 46% of American customers have faith in their banks and other financial organizations to keep their personal information secure. Customer and company confidence in data privacy and security is low, according to these numbers. When it comes to safeguarding client data, big data security is a lifesaver. Customers are more inclined to trust and stick with a business that they perceive as caring about their personal information and data security. To validate their data security commitment and reassure consumers that their data is safe, several firms use reputable third parties to conduct security audits.

4.3. Improved decision-making

By preventing illegal access, big data security aids in keeping data accurate and intact. Only authorized individuals can access sensitive information thanks to security features including authentication, restricted access, and encryption. Finding the proper insights and trends is easier in a safe data environment, which helps stakeholders make data-driven

decisions. By analyzing large amounts of data, financial institutions may do things like better identify fraudulent activities, manage risk, and extend credit to customers with excellent credit histories. But this can only happen if the data is accurate and safe.

4.4. Competitive advantage

Businesses may gain a competitive edge with big data security by protecting their most valuable assets and empowering them to make decisions based on data. A company's ability to retain customers depends on its ability to earn their trust and loyalty through the protection of their data and privacy. Companies that take security seriously also tend to attract partners who can contribute to the expansion of those businesses. All of these things help the business expand and stay ahead of rivals that haven't gotten into big data security analytics quite yet.

4.5. Common big data security challenges

It is crucial for companies to comprehend big data security concerns in today's digital world, as attackers use complex technology and inventive approaches, making big data security a major obstacleas seen in figure 4.In order to assist you take the necessary steps to secure your data, let's examine the biggest big data security challenges.

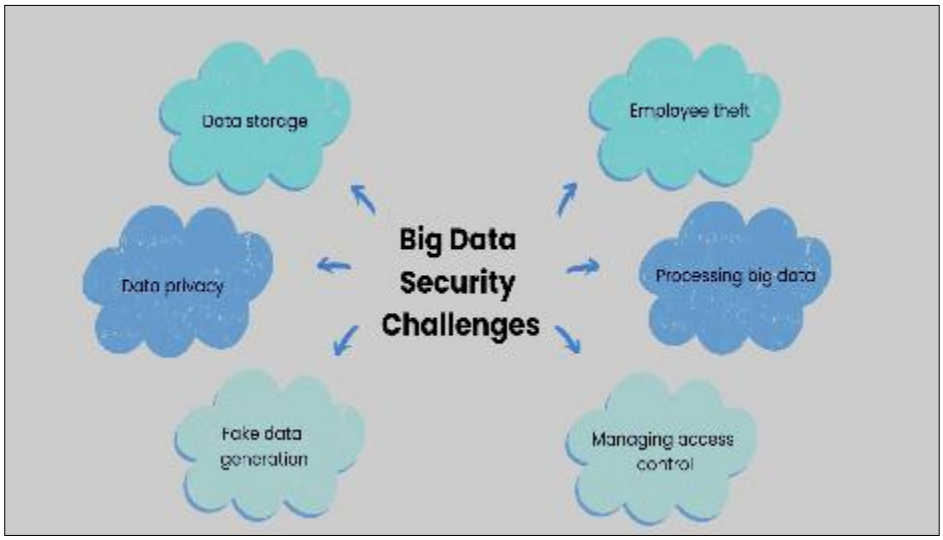

Figure 4 Big Data security challenges

4.6. Data storage

Data security is an issue with big data because of the massive amounts of data involved in processing and storing it. It is challenging to successfully integrate security measures for all data types stored by big data systems, which include unstructured, structured, and semi-structured data. Furthermore, sensitive information may exist in more than one place due to data redundancy and replication, which is typical in big data architecture and raises the possibility of illegal access.

4.7. Data privacy

Due to the nature of big data systems' frequent collection and storage of massive volumes of personally identifiable information, data privacy poses a serious threat to big data security. Because it gathers information from so many different places, both online and off, it makes it tough for companies to protect customer information. In addition, there is a higher chance of data breaches and illegal access due to the fact that big data platforms include exchanging data with third-party apps and services.

4.8. Fake data generation

The ability to create false data and then use it to trick or manipulate big data systems is another threat to big data security. Businesses may end up making poor choices due to erroneous data and insights caused by this difficulty. Criminals may, for instance, create phoney product reviews to influence the buying decisions of unsuspecting consumers. Additionally, attackers can benefit from using phony data to conceal actual data, which makes the theft of critical information easier.

4.9. Managing access control

Distributed and exceedingly complicated, big data systems store data in various physical places and on various servers. Because of this, it's not easy to set up and administer access restrictions that are compatible with all data types. Additionally, big data platforms may store and share massive amounts of data with other apps and services. There is always a greater risk of unwanted access to so large and diverse data, and managing access to it is a huge task.

4.10. Processing big data

Due to the data's exposure to numerous third-party programs and servers, processing big data—defined as sophisticated and distributed data across numerous systems—involves substantial risk. Rapid data creation and processing, sometimes in real time, characterize big data systems. This rapid speed makes it harder and harder to monitor security problems and respond rapidly.

Data processing while maintaining security measures needs meticulous planning and the application of strong security standards, especially as the volume increases.

4.11. Employee theft

Even more so for individuals engaged in big data analysis, all employees have access to the data to a certain degree. Even more concerning is the fact that some workers have first-hand knowledge of the company's security procedures, passwords, and access controls for its data systems. The ability to obtain sensitive information can be used by an employee who has access to a big data system. They can also damage the company's finances and image by manipulating data.

5. The proposed approaches

We lay out a number of safety features that would make cloud computing environments safer. Given the heterogeneity of the cloud infrastructure, we offer a range of solutions that, taken as a whole, will ensure its safety. In order to address the security issue mentioned before, the suggested solutions advocate for the utilization of various technologies and techniques. The recommendations for security are made in a way that does not hinder the scalability and efficiency of cloud systems. For a cloud environment to be secure, the following precautions should be implemented.

5.1. File Encryption

Any crucial data can be stolen by a hacker since it is stored on the machines in a cluster. Hence, it is imperative that all stored data be encrypted. It is recommended to utilize separate encryption keys on each system and to keep the key information centrally, protected by robust firewalls. This ensures that the data remains unusable for any malicious purposes, even in the event that a hacker manages to obtain it. We ensure the security of user data by storing it in an encrypted manner.

5.2. Network Encryption

It is imperative that every network communication adheres to industry standards on encryption. It is suggested that the RPC procedure calls occur over SSL so that relevant information cannot be extracted or changed even if a hacker gets to tap into network communication packets.4

5.3. Logging

It is important to keep track of all map reduction jobs that alter the data. It is also important to record the details of the users who are accountable for those tasks. In order to detect any malicious operations or data manipulation in the nodes, these logs should be examined on a regular basis.

- **Software Format and Node Maintenance**Regular formatting of the nodes running the software is required to eradicate viruses. To improve system security, it is recommended to upgrade all application software as well as Hadoop software.
- **Nodes Authentication**It is important to authenticate nodes whenever they join a cluster. A node that is known to be malicious should not be allowed to join the cluster. To distinguish between legitimate and malicious nodes, authentication methods like as Kerberos might be employed.

- **Rigorous System Testing of Map Reduce Jobs**It is recommended that developers test their map reduce jobs in a distributed environment rather than on a single machine. This will help to guarantee that the jobs are stable and resilient.
- **Honeypot Nodes**There should be honey pot nodes in the network; these nodes look normal on the outside, but they're actually traps. The hackers would be caught in these honeypots, and the appropriate measures would be taken to eradicate them.
- **Layered Framework for Assuring Cloud**The secure virtual machine, storage, data, and network monitoring layers make up a tiered architecture for cloud computing assurance [16], as seen in Figure 5. The policy, cloud monitoring, dependability, and risk analysis layers all work together to provide cross-cutting services.

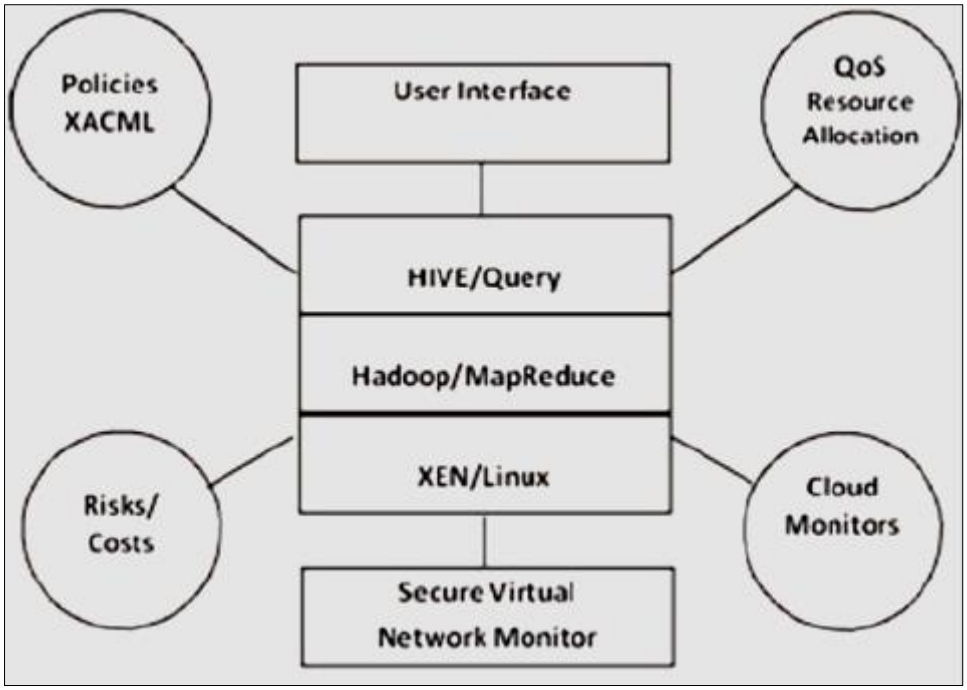

Figure 5 Layered framework for assuring cloud

- Publication of Third-Party Secure Data to the Cloud To make the most of available resources, cloud computing facilitates data storage at a remote location. Hence, this data must be safeguarded and only authorized individuals should have access to it. In essence, what this boils down to is the safe release of data needed for data outsourcing and other external publications by third parties. The computer acts as an independent publisher in a cloud setting, storing sensitive information on the server. In order to keep this data secure and guarantee its validity and completeness, the methods we've already covered are essential.
- Access Control A solid security approach for a distributed environment would be to incorporate differential privacy and required access control. When it comes to protecting their private information, data providers will have the final say. The mathematical limit on potential privacy violations will likewise be under their control. Users are able to conduct data computations in the aforementioned method with no data leaking out. We will be utilizing SELinux to ensure that no information leaks out. SELinux, short for Security-Enhanced Linux, is a kernel module that allows the Linux operating system to implement a security policy for access control. A patch to the Java Virtual Machine and an update to the Map Reduce framework will be used to implement differential privacy. The cloud service's user identification pool will be stored via its built-in applications. This means the cloud provider won't have to keep track of individual user identities across all of their apps. Cloud services will also allow for third-party authentication, in addition to the methods already mentioned. The user gaining access to the cloud and the cloud service will both have faith in the third party. An extra safeguard for the cloud service will be provided via third party authentication.

6. Conclusion

Considering the widespread adoption of cloud environments in both the business and research sectors, it is imperative that companies that operate in these cloud settings prioritize security measures. Cloud environments can be

safeguarded against complicated business operations by utilizing the ways that have been provided. Research of the CC and IoT-based BD was provided here with the intention of addressing the issues that they are now facing in terms of management and security. To be more specific, we integrated them in order to explore the qualities that are associated to them and the advantages that come with their combination. As a result, we presented the contribution that Big Data has made to CC, with the intention of bridging the scientific gap that currently exists in this area. In addition to that, this work demonstrates how CC enhanced the functionality of IoT-based BD. Lastly, we came up with a novel security model for a more sustainable environment after conducting research on the security concerns associated with the integration of CC and BD.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Chatzievangelou, D.; Bahamon, N.; Martini, S.; del Rio, J.; Riccobene, G.; Tangherlini, M.; Danovaro, R.; De Leo, F.C.; Pirenne, B.; Aguzzi, J. Integrating Diel Vertical Migrations of Bioluminescent Deep Scattering Layers Into Monitoring Programs. Front. Mar. Sci. 2021, 8, 661809. [Google Scholar] [CrossRef]

[2] Aguzzi, J.; Flögel, S.; Marini, S.; Thomsen, L.; Albiez, J.; Weiss, P.; Picardi, G.; Calisti, M.; Stefanni, S.; Mirimin, L.; et al. Developing technological synergies between deep-sea and space research. Elementa: Sci. Anthr. 2022, 10, 00064. [Google Scholar] [CrossRef]

[3] Aguzzi, J.; Costa, C.; Calisti, M.; Funari, V.; Stefanni, S.; Danovaro, R.; Gomes, H.; Vecchi, F.; Dartnell, L.; Weiss, P.; et al. Research Trends and Future Perspectives in Marine Biomimicking Robotics. Sensors 2021, 21, 3778. [Google Scholar] [CrossRef]

[4] Schnase, J.L.; Duffy, D.Q.; Tamkin, G.S.; Nadeau, D.; Thompson, J.H.; Grieg, C.M.; Mclnerney, M.A.; Webster, W.P. MERRA Analytic Services: Meeting the big Data Challenges of Climate Science Through Cloud-Enabled Climate Analytics-as-A-Service. Computers. Environ. Urban Syst. 2017 61, 198–211. [CrossRef]

[5] Kotsiou, V.; Papadopoulos, G.Z.; Chatzimisios, P.; Theoleyre, F. LDSF: Low-Latency Distributed Scheduling Function for Industrial Internet of Things. IEEE Internet Things J. 2020, 7, 8688–8699. [Google Scholar] [CrossRef]

[6] Plageras, A.P.; Psannis, K.E.; Stergiou, C.; Wang, H.; Gupta, B.B. Efficient IoT-based sensor BIG Data collection-processing and analysis in Smart Buildings. Futur. Gener. Comput. Syst. 2018, 82, 349–357. [Google Scholar] [CrossRef]

[7] Zhou, P.; Zhong, G.; Hu, M.; Li, R.; Yan, Q.; Wang, K.; Ji, S.; Wu, D. Privacy-Preserving and Residential Context-Aware Online Learning for IoT-Enabled Energy Saving With Big Data Support in Smart Home Environment. IEEE Internet Things J. 2019, 6, 7450–7468. [Google Scholar] [CrossRef]

[8] Stergiou, C.L.; Psannis, K.E.; Gupta, B.B. IoT-based Big Data secure management in the Fog over a 6G Wireless Network. IEEE Internet Things J. 2021, 8, 5164–5171

[9] Xiong, J.; Zhang, Y.; Tang, S.; Liu, X.; Yao, Z. Secure Encrypted Data With Authorized Deduplication in Cloud. IEEE Access 2019, 7, 75090–75104. [Google Scholar] [CrossRef]

[10] Gai, K.; Qiu, M.; Zhao, H. Privacy-Preserving Data Encryption Strategy for Big Data in Mobile Cloud Computing. IEEE Trans. Big Data 2018, 7, 678–688. [Google Scholar] [CrossRef] [Green Version]

[11] Mishra, S.K.; Puthal, D.; Sahoo, B.; Jena, S.K.; Obaidat, M.S. An adaptive task allocation technique for green cloud computing. J. Supercomput. 2018, 74, 370–385. [Google Scholar] [CrossRef]

[12] Chaudhary, R.; Aujla, G.S.; Kumar, N.; Rodrigues, J.J. Optimized Big Data Management across Multi-Cloud Data Centers: Software-Defined-Network-Based Analysis. IEEE Commun. Mag. 2018, 56, 118–126.

[13] Ramya Manikyam, J. Todd McDonald, William R. Mahoney, Todd R. Andel, and Samuel H. Russ. 2016.Comparing the effectiveness of commercial obfuscators against MATE attacks. In Proceedings of the 6th Workshop on Software Security, Protection, and Reverse Engineering (SSPREW'16)

[14] R. Manikyam. 2019.Program protection using software based hardware abstraction.Ph.D. Dissertation.University of South Alabama.

[15] GPB GRADXS, N RAO, Behaviour Based Credit Card Fraud Detection Design And Analysis By Using Deep Stacked Autoencoder Based Harris Grey Wolf (Hgw) Method, Scandinavian Journal of Information Systems 35 (1), 1-8.

[16] R Pulimamidi, GP Buddha, Applications of Artificial Intelligence Based Technologies in The Healthcare Industry, Tuijin Jishu/Journal of Propulsion Technology 44 (3), 4513-4519.

[17] R Pulimamidi, GP Buddha, AI-Enabled Health Systems: Transforming Personalized Medicine And Wellness, Tuijin Jishu/Journal of Propulsion Technology 44 (3), 4520-4526.

[18] GP Buddha, SP Kumar, CMR Reddy, Electronic system for authorization and use of cross-linked resource instruments, US Patent App. 17/203,879.

[19] Nadella, G. S. (2023). Validating the Overall Impact of IS on Educators in U.S. High Schools Using IS-Impact Model – A Quantitative PLS-SEM Approach, DAI-A 85/7(E), Dissertation Abstracts International, Ann Arbor, ISBN 9798381388480, 189, 2023.

[20] Gonaygunta, Hari, Factors Influencing the Adoption of Machine Learning Algorithms to Detect Cyber Threats in the Banking Industry, DAI-A 85/7(E), Dissertation Abstracts International, Ann Arbor, United States, ISBN 9798381387865, 142, 2023.

[21] Hari Gonaygunta (2023) Machine Learning Algorithms for Detection of Cyber Threats using Logistic Regression, 10.47893/ijssan.2023.1229.

[22] Hari Gonaygunta, Pawankumar Sharma, (2021) Role of AI in product management automation and effectiveness, https://doi.org/10.2139/ssrn.4637857.

[23] Sri Charan Yarlagadda, Role of Artificial Intelligence, Automation, and Machine Learning in Sustainable Plastics Packaging markets: Progress, Trends, and Directions, International Journal on Recent and Innovation Trends in Computing and Communication, Vol:11, Issue 9s, Pages: 818–828, 2023.

[24] Sri Charan Yarlagadda, The Use of Artificial Intelligence and Machine Learning in Creating a Roadmap Towards a Circular Economy for Plastics, International Journal on Recent and Innovation Trends in Computing and Communication, Vol:11, Issue 9s, Pages: 829-836, 2023.

[25] B. Nagaraj, A. Kalaivani, S. B. R, S. Akila, H. K. Sachdev, and S. K. N, "The Emerging Role of Artificial intelligence in STEM Higher Education: A Critical review," International Research Journal of Multidisciplinary Technovation, pp. 1–19, Aug. 2023, doi: 10.54392/irjmt2351.

[26] D. Sivabalaselvamani, K. Nanthini, Bharath Kumar Nagaraj, K. H. Gokul Kannan, K. Hariharan, M. Mallingeshwaran, Healthcare Monitoring and Analysis Using ThingSpeak IoT Platform: Capturing and Analyzing Sensor Data for Enhanced Patient Care, IGI Global eEditorial Discovery, Pages: 25, 2024. DOI: 10.4018/979-8-3693-1694-8.ch008.

[27] Amol Kulkarni, Amazon Athena Serverless Architecture and Troubleshooting, International Journal of Computer Trends and Technology, Vol, 71, issue, 5, pages 57-61, 2023.

[28] Amazon Redshift Performance Tuning and Optimization,International Journal of Computer Trends and Technology, vol, 71, issue, 2, pages, 40-44, 2023