International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Data Security Challenges and Solutions in Big Data Cloud Environments

Srinivas Murri*

Independent Researcher Redmond, WA, USA

Received 20 Nov 2022, Accepted 18 Dec 2022, Available online 20 Dec 2022, **Vol.12, No.6 (Nov/Dec 2022)**

Abstract

Every day, enormous amounts of data are produced at an exponential pace. Therefore, in order to get valuable insights into the modern world, analytics over data is unavoidable. Big Data (BD) makes good judgments and is strong in important applications. Big Data's meteoric rise in cloud computing has brought both new possibilities and new difficulties to the management and processing of enormous datasets. Hadoop and MapReduce are two state-of-the-art technologies used in today's world to face the heterogeneous high velocity, high variety and huge volume of data. CSPs such as AWS provide highly effective, low-cost solutions for managing and analyzing Big Data, yet security and privacy issues are still critical because of the highly sensitive nature of the information and due to the challenges brought about by intercloud migrations and high flash data rates. The present review paper aims to identify the key data security concerns in Big Data cloud computing models to analyze data privacy and integrity, availability, access control and multi-tenancy risks. It also provides an overview of kinds of security threats: encryption methods, methods of data-sharing safely, access-control mechanisms, and machine-learning-based anomaly detection that can protect from these threats. Furthermore, it emphasizes using scalable and multi-layer security architecture for the secure management of critical data through Big Data applications and maintaining their performance and capacity. Through these issues with sound security measures being addressed, corporations can explore the Big Data cloud environment for safe and reliable data processing and veracity of such environment.

Keywords: *Big Data, cloud computing environments, data security, cloud service providers, Amazon Web Services, Challenges, Solutions.*

1. Introduction

Big data refers to datasets that are too large for current software tools to handle and process in a reasonable amount of time. Military data and other unauthorized data must be securely safeguarded in an effective and scalable manner with Variety, Volume, and Velocity (Big Data). The lack of control that users have in an open environment makes data privacy and security a major worry with cloud computing. Big Data faces this issue as well [1]. Cloud computing's robust storage, computing, and distributed capabilities will enable Big Data processing, and more data will be touched by it in the world within a few years [2]. The need to investigate privacy and security concerns related to cloud computing and large data is another consideration. The privacy and security forum is a place where developers and researchers can talk about the latest findings, ideas, and experiences related to fundamental privacy and security issues and their applications in big data and cloud environments[3][4][5][6].

Figure 1 shows the cloud in action, and its ability to accommodate demands from customers all over the globe is a direct result of the network of data centers that make it possible.

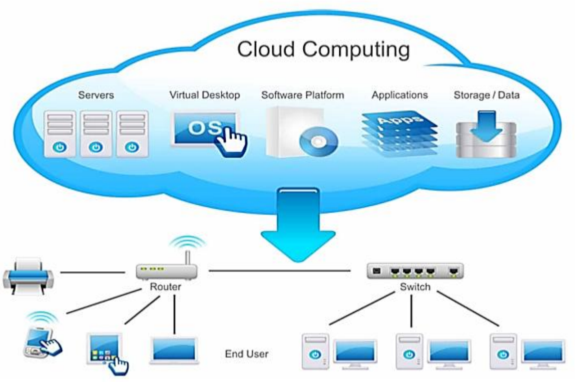

Fig. 1 Cloud computing

Cloud computing and big data, including a description, traits, and categorization of big data, as well as certain cloud computing-related debates [7]. Big data storage

*Corresponding author's ORCID ID: 0000-0000-0000-0000

DOI: https://doi.org/10.14741/ijcet/v.12.6.11

technologies, Hadoop, cloud computing, and their interaction are all covered. Data quality, data heterogeneity, privacy, availability, integrity, transformation, quality, legality, regulation, and governance are all important considerations[8][9]. Complexity has increased in handling huge data for concurrent processing, necessitating adaptation from a number of cloud-based solutions. The MapReduce program exemplifies cloud-based large-data processing[7][10].

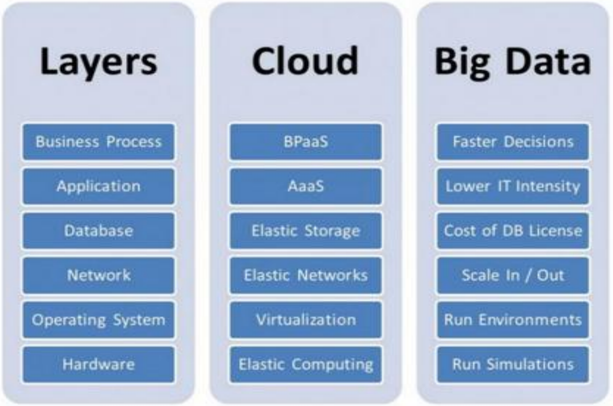

Fig.2 Big Data and Clouds [11]

Figure 2 shows the big data and cloud with layers. A cloud lets businesses manage, store, and share their Big Data in a cost-effective and user-friendly manner, as well as provide quick on-demand provisioning of server resources like CPUs. A vendor of on-demand analytics solutions supports the cloud infrastructure as a service platform, which lowers the cost of large-scale data analytics. Storage networking in a cloud is a powerful tool due to the use of a driver for high performance, and cloud computing as a whole is a great way to access resources, data, and software regardless of a system's physical location[12] [13].

This paper is structured as follows: The paper is organized as follows: Sections II and III introduce CC and Big Data, Section IV provides a list of Big Data cloud providers, Section V addresses the topic of privacy and security in the cloud, Sections VI and VII outline the problems and solutions related to data security in bigdata cloud environments, Section VIII gives a literature review, and Section IX concludes the paper.

Fundamentals of Cloud Computing

Sharing computer resources rather than relying on local servers or individual devices to manage applications is the foundation of CC. "Cloud Computing" is a style of computing in which users access resources and applications over the Internet. The term "Cloud" in this context signifies "The Internet." To disperse data processing across a large number of computers, cloud computing makes use of networks of these servers with specialized connections as an alternative to setting up individual software suites on every machine. When a user connects to the cloud via the Internet, they may access their apps in the cloud network regardless of their location [14][15]. Google Apps, including Gmail, Calendar, Docs, and Dropbox, are real-time apps that use Cloud Computing.

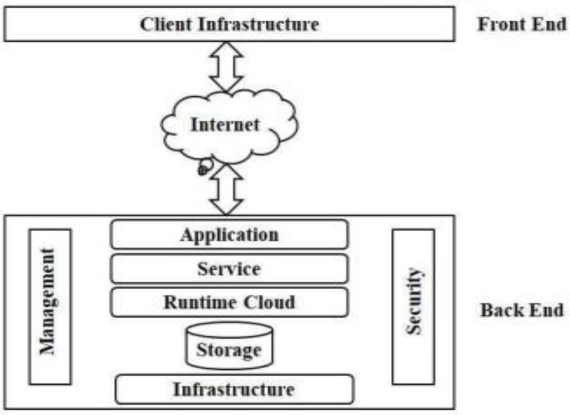

Fig.3 Cloud in architecture

Figure 3 displays the two primary parts of a cloud computing architecture: the front end and the back end. The front end includes client-side interfaces, such as web servers or mobile devices, while the back end consists of resources like servers, storage systems, and security mechanisms that manage the services[16].

Cloud Service Models[17]*:*

Infrastructure as a Service (IaaS): It offers a most adaptable cloud solution, letting customers lease storage and servers. Examples include Microsoft Azure and Amazon Web Services.

Platform as a Service (PaaS): Provides a way to build, test, and release software. GoogleAppEngine and Amazon Elastic Beanstalk are two examples.

Software as a Service (SaaS): A company offers pre-built software programs to customers via their website, usually by subscription. One example is Salesforce, while another is Microsoft Office 365.

Cloud Computing Deployment Models:

Public Cloud: The public has access to these services, while the actual infrastructure is handled by third-party suppliers. Independence, scalability, and flexibility are its features.

Private Cloud: Customised for a specific company, providing enhanced protection and control by limiting access to authorized users only.

Hybrid Cloud: Facilitates the transfer of data and applications across public and private cloud infrastructures by combining them.

Community Cloud: Shared by several organizations with common concerns, supporting collaborative functions within a specific community.

With cloud computing, you may raise or decrease the number of computers holding BigData to handle the real demand, meeting scalability requirements. Worldwide, data is expanding at an exponential pace, which poses a significant challenge for BigData due to its massive quantity. Consequently, Bigdata had difficulties in meeting the demands for increased flexibility, cost management, and fast storage development. An innovative approach to these problems has emerged with the advent of BigData on the cloud[18].

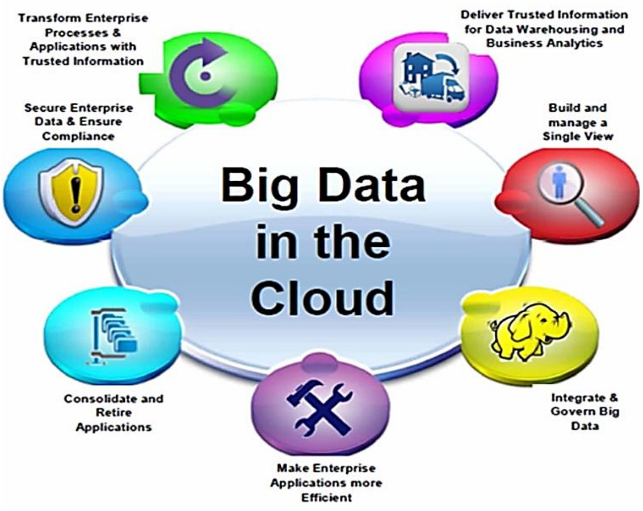
Transform Enterprise Processes & Applications with Trusted Information
Deliver Trusted Information for Data Warehousing and Business Analytics
Secure Enterprise Data & Ensure Compliance
Big Data in the Cloud
Build and manage a Single View
Consolidate and Retire Applications
Integrate & Govern Big Data
Make Enterprise Applications more Efficient

Fig.4 Adopting cloud computing technology to manage BigData[19]

Figure 4 shows how BigData is managed via cloud computing. It addresses issues encountered by application developers and offers cutting-edge solutions for managing BigData efficiently inside the Cloud. It also details new development and deployment initiatives for running data-intensive computing workloads.

Overview Of Big Data

Big data is far larger than standard data and is more challenging to handle and analyze. Big data storage necessitates scalable architecture as well as effective manipulation and storage. A vast and diverse collection of data in organized, unstructured, and semi-structured forms that is expanding quickly is referred to as "big data." The intricacy of big data makes it impossible to handle and analyze using conventional business tools; instead, it requires cutting-edge technology and sophisticated algorithms.

Characteristics of Big Data

The three Vs—volume, velocity, variety, value, and veracity—may be expanded to five Vs, as seen in Figure 5, since big data with new-generation architecture is now stored in many forms.

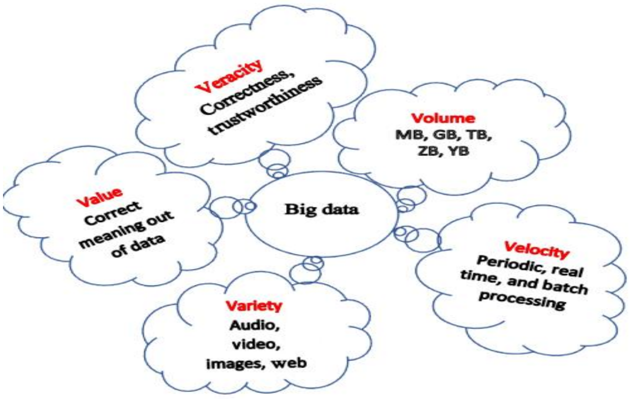
Veracity
Correctness, trustworthiness
Volume
MB, GB, TB, ZB, YB
Value
Correct meaning out of data
Big data
Velocity
Periodic, real time, and batch processing
Variety
Audio, video, images, web

Fig.5 5v's of Big data

Volume, variety, velocity, validity, value, and complexity are the six terms that define big data:

Volume: Most immediately, this book poses a threat to established IT infrastructures. By far, the majority of people picture this when they consider huge data.

Velocity: Data is expanding at a dizzying rate. The exponential growth in the processing, storage, and analysis speeds made possible by relational databases is in the millions.

Variety: Sensors, social networks, and cell phones all contribute to the proliferation of data formats. These instruments generate information in the forms of data logs, pictures, videos, audio, papers, and text. There are three more possible data types: structured, semi-structured, and unstructured.

Value: Big data is characterized by its value. This is related to the processing of data and its transformation into useful insights.

Veracity: The reliability of the data sources is crucial when working with large amounts of diverse data that are changing at a rapid pace.

Complexity: Data from many sources is connected, matched, cleaned, and transformed before being sent, allowing it to handle complexity [20].

Distributed Data in Big Data

The data is stored on many servers that the user has access to. A cloud service provider (CSP) competently and reliably delivers data to various customers. It controls the accessible data by using authentication, non-delicacy, and data recovery.

Third Party Auditor (TPA): Virtual private networks, secure socket layers, and point-to-point tunneling protocols all work together to provide security. However, several users are being accused of unauthorised access to the data. Users and cloud service providers alike may benefit from third-party authentication mechanisms in order to circumvent this kind of issue.[21] TPA monitors the transfer of data as well as the technologies and procedures used on it. The planning, implementation, and reporting phases are all necessary. At the data integrity level, security is tested.

Encryption-Based Storing of Data: The cloud storage server is encrypted using cryptographic methods to provide this level of protection. Every user that stores data on the storage cloud is given a key. Thus, only approved users may access the data kept in the cloud. They decrypt the data before retrieving it from the cloud whenever they need to access it. The user whose data was first saved in the cloud will be given a new access key [22].

Privacy-Preserving Public Auditing: Secure data is provided to consumers by using the homomorphic authenticator's approach, which guarantees the proper computation of data blocks. It is a result of merging four different algorithms. After getting access to data, a keygen algorithm creates the key. Metadata and digital signatures are both validated using the Singen algorithm.

Non-Linear Authentication: This method employs the haphazard application of a homomorphic non-linear authenticator. Digital signatures use the RSA technique for encryption and decryption of authentication data that follows. For the handshaking technique in particular, RSA takes advantage of the extensible authentication protocol. The client initiates communication with the cloud service provider by sending a request, which is then computed using a hash function.

Secure and Dependable Storage: The implementation of an error localization approach allows for the elimination of issues that have arisen with data storage on the cloud. Ensuring data security and verifying correctness are its primary functions. It employs homomorphic tokens, which validate data via erasure coding. It detects server errors and poor performance. This approach only allows for the monitoring of faults by a single server at a time, which may cause server failure in rare cases[23].

Technologies of Big Data

The following research technologies of Big Data[24]:

Hadoop: Its primary function is to group similar data nodes into clusters and to keep track of space-use statistics. In order to manage and transport data across racks, it runs on several environments. It is a framework for programming that is based on Java. Hadoop partitioned data over several servers, which facilitated the operation of various applications. Despite failures in different node clusters, the overall failure rate is lower. Hadoop is characterized by being scalable, inexpensive, adaptable, and resilient to failure. Many well-known businesses utilize Hadoop, including Google, Yahoo!, Amazon, IBM, and many more.

Hadoop Distributed File System (HDFS): It has the capacity to store vast quantities of data and is resilient to system faults. Hadoop creates clusters to spread data across computers. HDFS stores the files on the server by dividing them into blocks. On separate servers, it keeps three versions of the data. Compilation of data files that match the data node and the name node. Name node goes on to state that data node gives the data that customers query.

Map-reduce: Applications that reliably and fault-tolerantly process massive amounts of data are written using map-reduce. In order to process the data in parallel using Map tasks, it divides it into pieces. The processing data, including input and output, is saved in the file system. The unsuccessful job is also being monitored and re-executed. Two stages are used to implement the distribution of data: map and reduce. Slave nodes carry out tasks in response to instructions from the master node, which mostly involves monitoring, scheduling, and re-executing jobs[25].

Challenges in Big Data

A following research challenges of BigData[26]:

Availability: Data might be accessed by authorized individuals thanks to cloud computing. The performance of obtaining data is negatively affected since it is distributed over several clouds. Dealing with data transformation into appropriate formats is another difficulty.

Security: When customers use their credit or debit cards to access sensitive information, the security is jeopardized. In addition, every company has its own set of policies and procedures when it comes to sensitive data. Therefore, all data types need multi-level security. A privacy-protected data model is necessary to safeguard private information belonging to individuals or businesses. Hackers may save or erase data inadvertently due to insufficient security.

Scalability: The data speed and the CPU speed are not compatible. The data is not sent to the processor at the correct moment because of its huge volume. Many applications rely on parallel processing, including navigation, social networks, finance, internet search, timeliness, and thousands more. For scalability to remain intact, cloud services must reliably and promptly provide the necessary infrastructure, platform, and application resources [27].

Big Storage: Text, images, music, video, and many more formats are all possible for data. This kind of data is used by a variety of media, including mobile devices, remote sensing, aerial sensory technologies, radio frequency identification readers, and more. Large storage devices with plenty of capacity and fast input/output speeds are necessary for this kind of data. Finding specific information in large amounts of unstructured data is a major challenge. There was a delay in retrieving a large volume of data. Therefore, it becomes trickier in file systems[28].

Big Data Cloud Providers

A plethora of cloud service companies supply an assortment of big data products. Some have already achieved fame, while others are only beginning to gain recognition [29]. AWS, AT&T, Verizon/Terremark, Joyent, Rackspace, IBM, and GoGrid are just a few cloud providers that provide IaaS alternatives for big data analytics. Elastic Compute Cloud, a product of Amazon Web Services, is now one of the most well-known providers of infrastructure as a service (Amazon EC2). Amazon never intended to become a major player in the infrastructure services market. Amazon Web Services has lately begun to challenge its dominance in the cloud computing industry. Google Cloud Platform and the other clouds listed before are becoming more significant options. Two examples of such solutions include Open Stack, an open-source initiative with strong support from the business community, and Amazon Web Services, which refers to either Amazon's own offering or solutions provided by other firms that are compatible with its API[30][31]. Making a decision on a cloud platform standard so affects the accessible tools and the availability of rival suppliers offering the same technology.

The user may regulate scalability using Amazon EC2, and they can pay for resources on an hourly basis. There is significance in the usage of the word elastic in the name of Amazon's EC2. The capacity of EC2 customers to modify the infrastructure resources allotted to them in order to suit their requirements is referred to here as elasticity. Customers of its AWS portfolio may also make use of other big data services offered by Amazon. Some of them are as follows [5]:

Amazon Elastic MapReduce: Optimised for handling massive data sets. To power Elastic MapReduce, Amazon uses the Hadoop framework, which is housed on S3 and EC2. People can start using HBase today. Amazon DynamoDB is a fully managed database service that offers support for both SQL and NoSQL. Self-provisioning, transparently scalable, and easy to administer, DynamoDB is a highly available, fault-tolerant database service. Solid state discs (SSDs) are used to implement it for improved reliability and high performance.

Amazon Simple Storage Service (S3): Any quantity of data may be stored by using this web-scale service. Its architecture prioritizes speed and scalability above feature richness, making it a stripped-down data storage. You may store data in "buckets" and choose from several physical locations across the world to meet latency or regulatory requirements.

Amazon High-Performance Computing: This service offers specialized high-performance computing clusters with minimal latency that are optimized for certain activities. Amazon and other HPC companies are helping to bring HPC, which is typically utilized by academics and scientists, into the public. Amazon HPC clusters are highly adaptable and designed for certain workloads, making them easy to reconfigure for new jobs.

Amazon RedShift: An extensible MPP-based data warehousing solution, RedShift is now available in restricted preview and can handle data sets on a petabyte scale. It is compatible with several popular business intelligence products and provides a safe, dependable alternative to in-house data warehouses; it is managed by Amazon.

Big Data Privacy and Security In Cloud

One of the most talked-about technological topics in 2020 is still big data. However, amid all of the enthusiasm around Big Data's promise, the very real security and privacy issues that might impede this progress are often overlooked. The three V's of big data—Velocity, Volume, and Variety—amplify security and privacy concerns. These elements include things like expansive cloud infrastructures, a wide range of data formats and sources, the streaming nature of data collecting, and the rising number of intercloud transfers[32]. As a result, conventional security measures often fail since they are designed to protect static data on a smaller scale, rather than streaming data[17].

Modeling: making a threat model official that accounts for the majority of possible cyberattacks and data leaks.

Analysis: working with the threat model to identify practical solutions.

Implementation: implementation of the solution inside preexisting systems.

Protected information or crucial intellectual property (IP) will almost certainly be included in the massive data sets used in a Big Data deployment. Since this data is disseminated throughout the Big Data architecture as required, it is imperative that the whole data storage layer be safeguarded. Various forms of security and protection are used, including [20]:

Vormetric Encryption:

Provides comprehensive file system and volume-level protection for Big Data scenarios. Without requiring any modifications to application or system operation or management, this Big Data analytics security solution enables organizations to reap the advantages of Big Data analytics intelligence while preserving the security of their data.

Data Security Platform:

Critical data is protected by the Vormetric Data Security Platform, which stores all of your data's protections and access controls in one place. In order to protect your data from the most recent advanced persistent threats (APTs) and other security breaches, our data protection platform incorporates robust encryption, key management, granular access restrictions, and security intelligence data[33].

Encryption and Key Management:

Data encryption is essential for data breach mitigation and compliance programs. To facilitate compliance while maintaining transparency for processes, apps, and users, Vormetric offers robust, centrally controlled encryption and key management.

Fine-grained Access Controls:

Vormetric offers policy-based access controls with fine-grained control over who may access encrypted data, ensuring that only authorized processes and individuals can access encrypted data in order to fulfil stringent compliance requirements[34]. Even administrators of systems, networks, and even the cloud can only see unencrypted data if given explicit permission to do so. Despite seeing just encrypted data and not the plaintext source, administrative tasks and system updates continue to run without any issues.

Security Intelligence:

When combined with a SIEM system, the high-value security intelligence data gleaned from Vormetric logs can help detect compromised accounts and malevolent insiders, as well as patterns of access by processes and users that might indicate an advanced persistent threat (APT) attack[35].

Data Security Challenges in Big Data Cloud Environments

Big Data cloud platforms allow organizations to handle and analyze massive amounts of information effectively by providing huge processing power and storage capacity. However, in order to guarantee the availability, integrity, and confidentiality of data, these settings also present distinct security concerns. The following data security challenges in big data cloud environments are:

Data Privacy and Confidentiality

Protecting sensitive information is a critical concern in Big Data cloud environments[36]. Organizations store personal data, financial records, and other confidential information in the cloud, making it susceptible to unauthorized access. It's possible to leak data accidentally, to be attacked by insiders or to be targets of other unauthorized third parties. Using regular encryption for stored data, using communication confidentiality and imposing strict user permissions will significantly reduce these risks.

Data Integrity

It is crucial always to have accurate and credible data to ensure decisions made are correct and credible as well. The cloud environment contains information which can be modified by unprivileged users or be changed by a system during writing or reading processes. The casual paper shows how data integrity and its trustworthiness are preserved throughout its existence utilizing hash functions, digital signatures, and regularity of the integrity check [37].

Data Availability

Businesses that depend on cloud environments must provide constant data access. Some types of cyber-attacks include; failure hardware where access to data becomes unavailable and the DDoS attack, which hinders provision of services [38]. Two of the most effective ways of making data easily accessible in the occurrence of some eventuality is mirroring and shadowing and this is achieved through using redundant architectures, disaster recovery plans, and distributed storage architectures.

Access Control and Identity Management

Data control in a system with multiple users and changing needs is rather problematic. Inadequate methods of authentication may cause privilege escalation or organize unauthorized access. Strong techniques are offered by Identity-as-a-Service (IDaaS), RBAC, and MFA to guarantee that only authorized users have access to certain resources.

Secure Data Sharing

The exchange of data from one organization or multiple users, preferably through the cloud, while maintaining the privacy and security of the information is often demanded. But it leads to more serious risks of data leakage or its unauthorized redistribution. There are numerous additional ways to secure data and make sharing possible only with systems that allow for this activity: Attribute-based encryption, secure APIs, and usage auditing.

Security of Big Data Processing

Big Data applies data pipelines that are long and involve many analysis steps; thus, it is susceptible to security threats like insecure MapReduce jobs or code injection. Some ways to manage these risks include maintaining secure processing frameworks, maintaining sandboxing techniques and putting in place monitoring techniques.

Multi-Tenancy Risks

Cloud environments may accommodate many customers, which makes it challenging to keep one customer's information from leaking to another with similar infrastructure. For instance, bad tenant isolation or noisy neighbor effects are likely to threaten data integrity. The problem can be solved using VPCs, maintaining a high level of tenant isolation and utilizing the company's TEEs based on the hardware.

Security at Scale

While Big Data systems scale up, safety for big datasets and distributed structures becomes a daunting task.

Points like traditional encryption and Access control methods might not scale well. Column-wise encryption, distributed key management system, and AI-based anomaly detection are solutions that will safeguard security of the enterprise as data volumes expand limitlessly.

Data Security Solutions in Big Data Cloud Environments

Cloud environments on Big Data provoke numerous security issues connected with their complicated structure and the processing of highly sensitive information. It is revealed that enforcing strict security policies is a necessity to make and keep data available, and secure. Below are key solutions described in detail to address these challenges.

Encryption Techniques

Data privacy is provided even in cases where interception is done based on encryption, which is a standard security feature. Data which is stored in the cloud must be encrypted, and when at rest, levels of encryption like the AES are used. To bring privacy to scenarios where data is exchanged, then protocols like TLS are used. Moreover, computations may be performed on encrypted data by employing more complex approaches, such as homomorphic encryption still preserving user's privacy.

Access Control Mechanisms

If the data is sensitive there should be measures that only a few people should be allowed to access such information. RBAC and ABAC are both suitable for fine-grained user control, while RBAC keeps permissions closely connected to the employee responsibilities and ABAC – to their qualities. With regards to security, MFA employs many factors to establish a user, for instance, biometrics and passwords.

Data Masking and Tokenization

At the core of the information security measures, pertinent information is normally obscured or replaced with irrelevant credentials. Data masking is one of the greatest blessings for non-production environments to make the test data look like real data by hiding real data. Tokenisation ciphers sensitive data and generates token maps back only through a secure process and is especially beneficial for payment and private information.

Secure Data Sharing Solutions

The sharing of data is inevitable in cloud environments and must be done securely. Authenticated APIs and encrypted APIs enable users and systems to safely share data with each other. The core features of the blockchain may also be used to provide secure and unchangeable data sharing and storage while the access and alteration are monitored and documented.

Data Integrity Solutions

The issue of data integrity and authenticity as key to Big Data systems is another important concept to cover. Cryptographic hash functions are used to produce pointers for datasets to help in detecting quick tampering. Further, computational integrity and data security are built by using digital signatures that authenticate data. Due to this, immutable environments such as blockchains are used to offer a way of storing data that cannot be tampered with.

Anomaly Detection with AI and ML

AI and machine learning are being applied to make security solutions in Big Data more effective. IDS watches over the network traffic, looking for similar patterns to malicious activities, while behavioral analysis detects anomalies in user or system actions [39][40]. These techniques are vulnerable to real-time threat identification and prevention of possible attacks.

Secure Big Data Processing Frameworks

Indeed, advanced Big Data platforms such as Hadoop and Spark demand additional levels of protection. Sandboxing contains an application such that it cannot spread malware and encryption allows computation on encrypted data. Another important measure to improve cluster security and prevent unauthorized access to software and hardware components is also necessary for distributed processing systems.

Multi-Layer Security Architecture

The multilayered model of security helps to prevent multiple dangers as they are all covered. Computer security measures include firewalls and IDS that safeguard the network. Antivirus, and end-point detection and response products protect the users' machines. Disparate, approach to coding and performing vulnerability assessments ensures application security.

Literature Review

This section provides a literature review on big data security in cloud computing environments, also summary shown in Table I.

In, Jie Zhang et al. (2021) proposes a cloud-based model for data storage with an assurance of safety. The concept proposes a data storage mode based on Cache and a data security mode based on third-party authentication, which, when combined, improve data availability across the board in cloud computing. This approach takes a two-pronged approach to cloud computing security. Due to the appropriate security measures, data stored in the cloud may now be effectively protected [41].

In, Zijiao Tang et al. (2020) big data and cloud computing are two technologies that are advancing at a fast pace; both are helpful in making data storage and administration more efficient. In a large data cloud computing setting, however, there are issues with the data system's own data security. A better understanding of how to keep people's personal information safe while processing and integrating large data in a cloud computing environment is essential. Only in this way can condition-based data and information security issues be solved, and data transmission security and dependability will be improved[42].

In, Fangfang Dang et al. (2020) the elements impacting the information security of computer networks are examined using security big data, and the particular countermeasures are investigated using this data as a foundation. The difficulty of maintaining information security is rising in tandem with the number of information security challenges that individuals must deal with. The goal is to make data more stable and secure while also serving as a reference for the appropriate staff[43].

This research, Xiaojun Zhou et al. (2019) Their goals were twofold: first, to shed light on the connection between privacy, big data, and cloud computing; and second, to provide a solution to the challenge of safeguarding sensitive information in massive data sets. According to the proposed reference model, the cloud is its bedrock. An effective approach to researching the security of big data is this model, which builds the system's risk early warning and threat perception by gradually including the physical, data, interface, and application layers. Additionally, a path for future study is highlighted that will make use of blockchain technology to address issues with privacy and security in the cloud[44].

In, P. Suwansrikham et al. (2018) studies provide solutions to lessen the danger, protect personal information, and gain command. Many cloud storage providers get a portion of the massive data file. The perpetrator of an insider assault still only obtains a fragment of the file. The whole file cannot be recreated by him. Metadata is created upon file splitting. Data owners' login credentials, access paths, chunk locations, and metadata are all used to link each CSP. The notion of asymmetric security is used in this study [45].

This study, Delwar Hossain et al. (2018) offers a fresh method for protecting big data stored in the cloud, the SH-DBDS paradigm. A distributed cloud storage system will be used to upload the separated data. Unless all of the data is combined, the individual splits will be useless. An method for data splitting and joining is presented in this work. Experiments are conducted at both the local system and the AWS cloud with varying data sets ranging from 10MB to 1GB, and the results are evaluated. When assessing Big Data, both its security and performance are taken into account[46].

In, Saurabh Bahulikar et al. (2017) Data governs almost every aspect of a business, as recent technological events have shown. The article delves into the topics of data privacy, security in the cloud, and how it may be used to provide a safe foundation for virtualization and big data. Computer scientists are on the cusp of a new and exciting frontier when it comes to data science security protocols[47].

This study, Keke Gai et al. (2016) delves into these concerns and offers a fresh method for optimising file splitting and data storage on dispersed cloud servers, where operators of cloud services do not have direct access to the data. SAEDS is the name of the suggested method, and the two algorithms that back it up, Efficient Data Conflation (EDCon) and Secure Efficient Data Distributions (SED2), are rather impressive. Both efficiency and security have been tested experimentally[48].

Table 1 Summary of related work on data security in big data cloud

Reference	Purpose	Tools/Techniques	Big Data Cloud	Data Security	Challenges	Solutions
Jie Zhang et al. (2021)	To improve data availability and security in cloud computing	Cache-based data storage; third-party authentication	Yes	Yes	Data storage and security in transmission	Proposed cache-based storage and third-party authentication to ensure security and improve data availability
Zijiao Tang et al. (2020)	To enhance data security protection in big data cloud environments	Data security protection techniques	Yes	Yes	Inherent data security issues in big data and cloud environments	Developed security measures tailored to data integration and processing in big-data cloud environments
Fangfang Dang et al. (2020)	To analyse factors affecting network security and propose countermeasures	Analysis of security big data	Yes	Yes	Increasing difficulty in maintaining information security	Provided reference strategies to improve data stability and security
Xiaojun Zhou et al. (2019)	To address security and privacy protection in big	Reference model with physical, data, interface, and	Yes	Yes	Security risk warning and threat perception	Introduced a multi-layered reference model and suggested

	data cloud computing	application layers			challenges	blockchain for future cloud security and privacy issues
P. Suwansrikham et al. (2018)	To minimize risk and ensure data privacy with access control	File splitting; metadata generation; asymmetric security concepts	Yes	Yes	Insider attacks and full file reconstruction risks	Split data into chunks across multiple CSPs, with asymmetric security measures
Delwar Hossain et al. (2018)	To develop a secure and high-performance distributed big data storage (SH-DBDS) model	Data splitting and joining algorithm	Yes	Yes	Ensuring data security during storage and retrieval	Proposed SH-DBDS model with split and join mechanism evaluated on local and AWS systems
Saurabh Bahulikar et al. (2017)	To explore cloud computing's role in securing big data and virtualization infrastructures	Data privacy and secure frameworks	Yes	Yes	Privacy concerns and securing virtualization infrastructure	Discussed secure frameworks and measures to enhance big data and virtualization security
Keke Gai et al. (2016)	To propose an efficient distributed storage model for data security	SED2 Algorithm; EDCon Algorithm	Yes	Yes	Data accessibility by cloud service operators	Developed SAEDS model for secure and efficient distributed storage with novel algorithms

Conclusion

The most important things to consider while using the cloud are security and privacy. Private and secure BD has been the talk of the town in this age of Big Data. Because of the magnitude, sensitivity, and complexity of the data being processed, Big Data cloud platforms offer tremendous possibilities while also posing substantial security concerns, as discussed in this study. Since organisations embark on adopting cloud infrastructure for storing and processing mammoth data, it is vital to consider issues to do with privacy, data integrity, data availability, and access control. The evaluation underscores security requirements such as encryption of information, secure data transfer and AI-based anomaly detection and the greatest challenges include threats such as unauthorized access, data loss and data corruption. Other strategies like efficient security models and quiddities, RBAC, and MFA may decrease the risk formed by multi-tenancy and massive data flows. To protect sensitive information and intellectual property in Big Data cloud ecosystems, a multi-layered security architecture is essential, combining network, endpoint, and application security measures. In the end, companies need to implement thorough, scalable, and privacy-preserving security measures to guarantee data availability, confidentiality, and integrity, allowing for safe and effective Big Data processing in cloud settings.

References

[1] Rajesh Goyal, “The role of business analysts in information management projects,” *Int. J. Core Eng. Manag.*, vol. 6, no. 9, 2020.

[2] Y. B. Reddy, “Big Data Security in Cloud Environment,” in *Proceedings - 4th IEEE International Conference on Big Data Security on Cloud, BigDataSecurity 2018, 4th IEEE International Conference on High Performance and Smart Computing, HPSC 2018 and 3rd IEEE International Conference on Intelligent Data and Securit*, 2018. doi: 10.1109/BDS/HPSC/IDS18.2018.00033.

[3] K. Patel, “Quality Assurance In The Age Of Data Analytics: Innovations And Challenges,” *Int. J. Creat. Res. Thoughts*, vol. 9, no. 12, pp. f573–f578, 2021.

[4] B. Boddu, “The Quantum Edge: How Quantum Computing Will Transform Databases,” *Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci.*, vol. 9, no. 3, 2021, doi: https://doi.org/10.5281/zenodo.14059357.

[5] G. Galante, L. C. Erpen De Bona, A. R. Mury, B. Schulze, and R. da Rosa Righi, “An Analysis of Public Clouds Elasticity in the Execution of Scientific Applications: a Survey,” *J. Grid Comput.*, 2016, doi: 10.1007/s10723-016-9361-3.

[6] N. G. Singh, Abhinav Parashar A, “Streamlining Purchase Requisitions and Orders : A Guide to Effective Goods Receipt Management,” *J. Emerg. Technol. Innov. Res.*, vol. 8, no. 5, 2021.

[7] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, “The rise of ‘big data’ on cloud computing: Review and open research issues,” *Information Systems.* 2015. doi: 10.1016/j.is.2014.07.006.

[8] V. S. Thokala, “A Comparative Study of Data Integrity and Redundancy in Distributed Databases for Web Applications,” *Int. J. Res. Anal. Rev.*, vol. 8, no. 4, pp. 383–389, 2021.

[9] J. Thomas, “The Effect and Challenges of the Internet of Things (IoT) on the Management of Supply Chains,” *Int. J. Res. Anal. Rev.*, vol. 8, no. 3, pp. 874–878, 2021.

[10] J. Thomas and V. Vedi, “Enhancing Supply Chain Resilience Through Cloud-Based SCM and Advanced Machine Learning: A Case Study of Logistics,” *J. Emerg. Technol. Innov. Res.*, vol. 8, no. 9, 2021.

[11] E. Sayed, A. Ahmed, and R. A. Saeed, “A Survey of Big Data Cloud Computing Security,” *Int. J. Comput. Sci. Softw. Eng.*, 2014.

[12] J. Xue and J. J. Zhang, “A brief survey on the security model of cloud computing,” in *Proceedings - 9th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, DCABES 2010*, 2010. doi: 10.1109/DCABES.2010.103.

[13] M. D. H. and D. R., “An Analysis of Security Challenges in Cloud Computing,” *Int. J. Adv. Comput. Sci. Appl.*, 2013, doi: 10.14569/ijacsa.2013.040106.

[14] V. N. Inukollu, S. Arsi, and S. R. Ravuri, “Security Issues Associated with Big Data in Cloud Computing,” *Int. J. Netw. Secur. Its Appl.*, vol. 6, no. 3, pp. 45–56, 2014.

[15] R. Arora, S. Gera, and M. Saxena, “Impact of Cloud Computing Services and Application in Healthcare Sector and to provide improved quality patient care,” *IEEE Int. Conf. Cloud Comput. Emerg. Mark. (CCEM), NJ, USA, 2021*, pp. 45–47, 2021.

[16] S. A. El-Seoud, H. F. El-Sofany, M. Abdelfattah, and R. Mohamed, “Big data and cloud computing: Trends and challenges,” *Int. J. Interact. Mob. Technol.*, 2017, doi: 10.3991/ijim.v11i2.6561.

[17] R. Bishukarma, “The Role of AI in Automated Testing and Monitoring in SaaS Environments,” *Int. J. Res. Anal. Rev.*, vol. 8, no. 2, pp. 846–851, 2021.

[18] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, “Big Data computing and clouds: Trends and future directions,” *J. Parallel Distrib. Comput.*, 2015, doi: 10.1016/j.jpdc.2014.08.003.

[19] S. Ouf and M. Nasr, “Cloud Computing: The Future of Big Data Management,” *Int. J. Cloud Appl. Comput.*, 2015.

[20] Ishwarappa and J. Anuradha, “A brief introduction on big data 5Vs characteristics and hadoop technology,” *Procedia Comput. Sci.*, vol. 48, no. C, pp. 319–324, 2015, doi: 10.1016/j.procs.2015.04.188.

[21] A. Goyal, “Enhancing Engineering Project Efficiency through Cross-Functional Collaboration and IoT Integration,” *Int. J. Res. Anal. Rev.*, vol. 8, no. 4, pp. 396–402, 2021.

[22] Pranav Khare and Abhishek, “Cloud Security Challenges: Implementing Best Practices for Secure SaaS Application Development,” *Int. J. Curr. Eng. Technol.*, vol. 11, no. 06, 2021, doi: https://doi.org/10.14741/ijcet/v.11.6.11.

[23] V. V Kumar, M. Tripathi, S. K. Tyagi, S. K. Shukla, and M. K. Tiwari, “An integrated real time optimization approach (IRTO) for physical programming based redundancy allocation problem,” *3rd Int. Conf. Reliab. Saf. Eng.*, pp. 692–704, 2007.

[24] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, “Big data preprocessing: methods and prospects,” *Big Data Anal.*, 2016, doi: 10.1186/s41044-016-0014-0.

[25] R. Chandrashekar, M. Kala, and D. Mane, “Integration of Big Data in Cloud computing environments for enhanced data processing capabilities,” *Int. J. Eng. Res. Gen. Sci.*, 2015.

[26] H. Kaur, S. Goraya, and A. Prof, “Role of Big Data in Cloud Computing: A Review,” *Int. J. Eng. Res. Technol.*, vol. 8, no. 07, pp. 866–869, 2019.

[27] V. V. Kumar, A. Sahoo, and F. W. Liou, “Cyber-enabled product lifecycle management: A multi-agent framework,” in *Procedia Manufacturing*, 2019. doi: 10.1016/j.promfg.2020.01.247.

[28] V. Kumar, V. V. Kumar, N. Mishra, F. T. S. Chan, and B. Gnanasekar, “Warranty failure analysis in service supply Chain a multi-agent framework,” in *SCMIS 2010 - Proceedings of 2010 8th International Conference on Supply Chain Management and Information Systems: Logistics Systems and Engineering*, 2010.

[29] V. S. Thokala, “Integrating Machine Learning into Web Applications for Personalized Content Delivery using Python,” *Int. J. Curr. Eng. Technol.*, vol. 11, no. 6, pp. 652–660, 2021, doi: https://doi.org/10.14741/ijcet/v.11.6.9.

[30] V. S. Thokala, “Utilizing Docker Containers for Reproducible Builds and Scalable Web Application Deployments,” *Int. J. Curr. Eng. Technol.*, vol. 11, no. 6, pp. 661–668, 2021, doi: https://doi.org/10.14741/ijcet/v.11.6.10.

[31] B. Boddu, “DevOps for Database Administration: Best Practices and Case Studies,” *https://jsaer.com/download/vol-7-iss-3-2020/JSAER2020-7-3-337-342.pdf*, vol. 7, no. 3, p. 5, 2020.

[32] B. Boddu, “Data Governance and Quality in Data Warehousing and Business Intelligence,” *IJFMR*, vol. 3, no. 6, p. 9, 2021.

[33] N. Richardson, R. Pydipalli, S. S. Maddula, S. K. R. Anumandla, and V. K. Yarlagadda, “Role-Based Access Control in SAS Programming: Enhancing Security and Authorization,” *Int. J. Reciprocal Symmetry Theor. Phys.*, 2019.

[34] M. R. Kishore Mullangi, Vamsi Krishna Yarlagadda, Niravkumar Dhameliya, “Integrating AI and Reciprocal Symmetry in Financial Management: A Pathway to Enhanced Decision-Making,” *Int. J. Reciprocal Symmetry Theor. Phys.*, vol. 5, no. 1, pp. 42–52, 2018.

[35] V. K. Yarlagadda and R. Pydipalli, “Secure Programming with SAS: Mitigating Risks and Protecting Data Integrity,” *Eng. Int.*, vol. 6, no. 2, pp. 211–222, Dec. 2018, doi: 10.18034/ei.v6i2.709.

[36] R. Arora, “Mitigating Security Risks on Privacy of Sensitive Data used in Cloud-based Mitigating Security Risks on Privacy of Sensitive Data used in Cloud-based ERP Applications,” *8th Int. Conf. “Computing Sustain. Glob. Dev.*, no. March, pp. 458–463, 2021.

[37] M. Gopalsamy, “Artificial Intelligence (AI) Based Internet-ofThings (IoT)-Botnet Attacks Identification Techniques to Enhance Cyber security,” *Int. J. Res. Anal. Rev.*, vol. 7, no. 4, pp. 414–420, 2020.

[38] M. Gopalsamy, “Advanced Cybersecurity in Cloud Via Employing AI Techniques for Effective Intrusion Detection,” *Int. J. Res. Anal. Rev.*, vol. 8, no. 1, 2021.

[39] Mani Gopalsamy, “Enhanced Cybersecurity for Network Intrusion Detection System Based Artificial Intelligence (AI) Techniques,” *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 12, no. 1, pp. 671–681, Dec. 2021, doi: 10.48175/IJARSCT-2269M.

[40] K. K. SKR Anumandla, VK Yarlagadda, SCR Vennapusa, “Unveiling the Influence of Artificial Intelligence on Resource Management and Sustainable Development: A Comprehensive Investigation,” *Int. J. Creat. Res. Thoughts*, vol. 9, no. 12, pp. f573–f578, 2020.

[41] J. Zhang, “Research on the Application of Computer Big Data Technology in Cloud Storage Security,” in *Proceedings of 2021 IEEE International Conference on Data Science and Computer Application, ICDSCA 2021*, 2021. doi: 10.1109/ICDSCA53499.2021.9650284.

[42] Z. Tang, “A Preliminary Study on Data Security Technology in Big Data Cloud Computing Environment,” in *Proceedings - 2020 International Conference on Big Data and Artificial Intelligence and Software Engineering, ICBASE 2020*, 2020. doi: 10.1109/ICBASE51474.2020.00013.

[43] F. Dang, H. Liang, S. Li, D. Li, and H. Liu, “Design and implementation of computer network information security protection based on secure big data,” in *Proceedings of 2020 IEEE 3rd International Conference of Safe Production and Informatization, IICSPI 2020*, 2020. doi: 10.1109/IICSPI51290.2020.9332352.

[44] X. Zhou, P. Lin, Z. Li, Y. Wang, W. Tan, and M. Huang, “Security of big data based on the technology of cloud computing,” in *Proceedings - 2019 4th International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2019*, 2019. doi: 10.1109/ICMCCE48743.2019.00163.

[45] P. Suwansrikham and K. She, “Asymmetric Secure Storage Scheme for Big Data on Multiple Cloud Providers,” in *Proceedings - 4th IEEE International Conference on Big Data Security on Cloud, BigDataSecurity 2018, 4th IEEE International Conference on High Performance and Smart Computing, HPSC 2018 and 3rd IEEE International Conference on Intelligent Data and Securit*, 2018. doi: 10.1109/BDS/HPSC/IDS18.2018.00036.

[46] M. D. Hossain and M. Abdullah Adnan, “Secured and High Performance Distributed Big Data Storage in Cloud Systems,” in *2018 3rd International Conference on Computer and Communication Systems, ICCCS 2018*, 2018. doi: 10.1109/CCOMS.2018.8463340.

[47] S. Bahulikar, “Security measures for the big data, virtualization and the cloud infrastructure,” in *India International Conference on Information Processing, IICIP 2016 - Proceedings*, 2017. doi: 10.1109/IICIP.2016.7975336.

[48] K. Gai, M. Qiu, and H. Zhao, “Security-Aware Efficient Mass Distributed Storage Approach for Cloud Systems in Big Data,” in *Proceedings - 2nd IEEE International Conference on Big Data Security on Cloud, IEEE BigDataSecurity 2016, 2nd IEEE International Conference on High Performance and Smart Computing, IEEE HPSC 2016 and IEEE International Conference on Intelligent Data and S*, 2016. doi: 10.1109/BigDataSecurity-HPSC-IDS.2016.68.