Communication in Physical Sciences, 2024 11(4): 1030-1059

Enhancing Data Provenance, Integrity, Security, and Trustworthiness in Distributed and Federated Multi-Cloud Computing Environments

Olatunde Ayeomoni

Received:12July 2024/Accepted: 03 October 2024/Published: 26 October 2024

Abstract: *Due to the increasing trend in distributed cloud environments, strong data provenance and integrity practices are even more important than before to ensure answers to security and traceability requirements as well as compliance. The new challenges, developments, and best practices in monitoring and security of data for cloud systems are discussed in this paper. Key challenges include scalability limitations, privacy vs. transparency trade-offs, and regulatory compliance issues. To address these concerns, blockchain-based provenance tracking, AI-driven anomaly detection, cryptographic hashing, and privacy-preserving techniques such as homomorphic encryption and secure multiparty computation (SMPC) have emerged as innovative solutions. The study also examines real-world implementations in healthcare, finance, and supply chain management, demonstrating how organizations leverage provenance tracking to enhance trust, security, and operational efficiency. Additionally, the paper discusses standardization efforts such as W3C PROV and ISO 27037, which aim to improve interoperability and legal compliance. Moving forward, advancements in federated learning, decentralized identity management, and quantum-resistant cryptography will play a crucial role in enhancing provenance tracking and ensuring secure cloud ecosystems. By integrating AI-driven monitoring, blockchain scalability solutions, and adaptive compliance frameworks, organizations can build resilient, transparent, and tamper-proof data management systems in an increasingly digital world.*

Keywords*: Data Provenance, Cloud Security, Blockchain-based Integrity, Privacy-Preserving Computation and Regulatory Compliance in Cloud Computing*

Olatunde Ayeomoni
688 riddle road, Cincinnati, Ohio, 45220.
Email ayeomooe@mail.uc.edu
Orcid id 0009-0000-1890-0419

1.0 Introduction

The growing dependency on distributed cloud environments has changed how organizations store, process, and manage data (Sunyaev & Sunyaev, 2020). Cloud computing allows higher scalability, cost rectitude, and availability through which firms can operate within a very interconnected ecosystem. Even though some distributed infrastructures raise very crucial concerning-the integrity and provenance of data, organizations cannot seem to ascertain the authenticity, origin, and modifications of any data throughout its life cycle (Li & Zhang, 2021).. An absence of a structured data provenance framework gives rise to clouds easily susceptible to unauthorized alteration and cyberattacks, in addition to potential non-conformance to regulations, greatly posing risks to the partners and clients (Kommisetty, 2022).

Data provenance is simply the tracking and verification of the source of data, different modifications of data, and ultimately the flow of data across different cloud nodes (Suen *et al*., 2013). Provenance implies transparency and trustworthiness and is done by systematically keeping the track record and showing the evidence of where data originates, who changed it, and how the data changed. The very importance of provenance is supported by

industry statistics stating that IBM Security's Cost of Data Breach Report (2021) has found that more than 45% of data breaches result from a lack of visibility into data movement (Ametepe *et al*., 2021). Thus, tracking data provenance provides companies a way to reduce such risks by obtaining an auditable trail of data transformations, which is important for ensuring data integrity, regulatory compliance, and security. Besides, it is not well supported that managing provenance introduces additional complexity to distributed cloud environments because of the large number of data replicas, geographically distributed servers, and dynamic network interconnections (Zhang *et al*., 2011). Organizations should be able to have confidence that the data they use to make decisions has not been tampered with and continues to be valid, which is as important as data integrity Mechanisms for data integrity offer consistency, prevent unwanted modifications, and detect abnormalities in flow. According to Cybersecurity Ventures, data manipulation is one of the fastest growing risks, and by 2025 global losses due to cybercrime are expected to reach beyond $10.5 trillion per year (Imran *et al*., 2017). As per the risks of getting corrupted and lying data, it becomes the key requirement for implementing cryptographic hashing, blockchain-based tracking of provenance, along with real-time integrity verification techniques, both for cloud service providers and enterprises. This technology integration assuredly will earn immense trust and reliability on cloud-based applications, particularly in the healthcare, finance, and supply chain sectors, where data integrity is mission-critical (Imran & Hlavacs, 2012).

The problems with data provenance and integrity in dispersed cloud environments are due to lack of conventions frameworks, insufficient scalability and compliance limitations (Katari & Ankam, 2022).

Traditional provenance models often do not cope with how fast cloud infrastructures evolve nowadays, where real-time data interactions, AI-based processing, and hybrid cloud models intervene to muddle the tracing of provenance (Zafar *et al*., 2017). Moreover, whenever detailed data trails are furnished, privacy issues surface since provenance metadata could disclose sensitive data about users and business activities. Cloud architects are finding it hard to balance privacy and transparency when it comes to security. They looked into the provenance models based on zero-knowledge evidence, homomorphic encryption, and secure multi-party computing to enhance privacy (Imran *et al*., 2017). To establish cloud systems that are more proven, several new technologies are under development as potential solutions to these problems, such as blockchain, artificial intelligence, and decentralized identity management (Lim *et al*., 2018). Hussain & Al-Turjman (2021) opined that the blockchain technology can facilitate the prevention of tampering of origin because it brings clarity and data immutability, which is a possible tool in stopping origin manipulation. Decentralized identity models can limit access to sensitive data by authorized users and AI-based anomaly detection might be capable of detecting suspicious changes in data in real-time. Studies have shown that lucrative technology, such as cloud security systems, blockchain enabled provenance monitoring manages a minimum of 80 percent of economic fraudulent practices, even in the financial transactions, which serves as a strong example of the value of employing advanced technology (Gudala *et al*., 2022).

The work is expected to place the theoretical problems and challenges of distributed data provenance and integrity in cloud environments and technological innovation (Ametepe *et al*., 2021).

Based on several real-life case studies and applying number of provenance models and security mechanisms, the study attempts to offer best practices in deploying provenance conscious cloud infrastructures. It will also be significant in the future of paradigms of cloud

computing in terms of trust and compliance as well as data authenticity to industries, researchers, and regulators (Gudala *et al*., 2022).

Through conceptual analysis of prevailing frameworks, modern trends and directions the research contributes to the increasing literature on data provenance, cloud environment defense to provenance gaps, unlawful access and data manipulation (Dib & Rababah, 2020).

2. 0 Conceptual Foundation

2.1 Data Provenance

Data provenance is the history of a piece of data i.e. the origin of the data, the activities that make the data change its state and the path through which data passes through other systems.

It serves as a kind of record keeping that documents creation, change, and distribution of data over time (Simmhan, *et al*., 2005). In the cloud setting, provenance guarantees that operations have been performed on the data and its accountability, traceability and transparency in distributed platforms where much more data replication, processing and storage occur or are conducted all at once under different locations (Groth, 2007). Provenance gives organizations confidence in verifying the authenticity of their datasets, thus, all of which fall under data governance, compliance, and cybersecurity. Provenance metadata typically consists of timestamps, user activity logs, and transformation information, which can allow organizations to monitor and audit their data quite effectively (Trace, 2020) . Provenance is critical in building trust and reliability at a large scale while deploying distributed infrastructures. Cloud environment stores and processes large amount of data which keeps in changing at intervals making auditing and logging change slightly difficult and unauthorized modification to any of the data much recommended. Provenance mechanism helps to detect anomalies, to verify data integrity and also to do forensic investigation if security breach or corruption is identified in the data (Bettivia *et al*., 2022). There is also the need for detailed records of data processing activities to be maintained as may be required by law, for example, GDPR (General Data Protection Regulation) or HIPAA (Health Insurance Portability and Accountability Act) rules. Adoption of provenance tracking in cloud services will dramatically help to increase regulatory compliance and decrease the security risk incurred by optimizing data lifecycles (Cheney *et al*., 2009).

Therefore, the value of provenance on Cloud is much wider than simply the dimension of compliance and security; it additionally includes efficiency, and better decision-making (Singh *et al*., 2018).

Also, provenance metadata provides automatic detection of abnormal activity and real-time tracking, which, among other things, reduces the impact, and the costs of data manipulation and exposure to cyber threats. Provenance tracking has already been implemented in important industries in this context, including healthcare, financial services, supply chain management, etc. Provenance tracking is being adopted in many industries, thus achieving data attributes like quality and providing the company with the means for faster audits and better data-driven decisions. In medicine it involves the tight data provenance for the integrity of patient records without the opportunities for medical fraud. In financial systems, provenance tracking aids in discovering fraud and increasing translucence of transactions. Data provenance bolsters trust, accountability and dependability and thus forms an essential building block of modern cloud computing architectures (Mather *et al*., 2009).

2.2 Data Integrity

Data integrity is defined as the maintenance of data's accuracy, consistency, and reliability during the entire lifespan of the data. If data are altered, destroyed, or fabricated to an unknown extent, there is a trustworthiness issue. The best definition of this term would be when data are

stored, transferred, or processed in a system (Whyte, 2021). In cloud computing, where data is often accessed, modified, and shared across multiple locations, integrity plays an important role in seeing that no unauthorized access will disrupt, corrupt, or make the data inconsistent (Thokala, 2021). Protection mechanisms for data integrity-Checksum, cryptographic hash, and validation-helps to detect and protect any unauthorized data modifications. Ultimately, data integrity is the basis upon which trust can be built in digital transactions, thus averting irreversible loss of critical information while also guaranteeing safe decision-making within cloud environments (Sharma, *et al*., 2021).

Therefore, the integrity of data is paramount in cloud computing because it protects confidential data and ensures smooth operations (Aldossary and Allen, 2016). The cloud-based solutions to protect integrity are useful to businesses that are vulnerable to illicit usage, accidental data loss, and attacks of trustworthy and verifiable datasets. Cloud service providers are able to guarantee the integrity of data through redundancy, block-chain ledgers, and by detecting errors and repairing them (Kumar & Poornima, 2012). Moreover, the regulatory frameworks like GDPR, HIPAA, and ISO 27001, make it binding on an organization to have the highest possible integrity of data. Unless integrity safeguards are strongly enforced, organizations risk suffering from corruption or manipulations of data that lead to financial loss, rerouting of business confidence, and overall operational interruption (Sun *et al*., 2014).

The advantages of data integrity further compose security, organizational efficiency, trust, and regulatory compliance (Maddukuri, 2021). Organizations that focus on integrity reap benefits in terms of accurate analytics, error-free data transactions, and lower risk of perpetrated frauds. For example, in banking and finance, data integrity avails the correctness of financial transactions and fraud-proofing amendments (Aldossary & Allen, 2016). On the other side, in healthcare, upholding integrity on patient records assures safe medical decisions and adherence to legal requirements. The amalgamation of automated integrity verification, cryptography, and blockchain technology within cloud environments will yield a secure and resilient infrastructure for processing critical data. By confirming that data remain accurate, complete, and reliable, organizations will maximize the security, compliance, and trustworthiness of their digital systems (Mather *et al*., 2009).\

2.3 Distributed Cloud Environments

The characteristics of a distributed cloud environment require the cloud service to be delivered from a cross-section of multiple data centers in remote geographical locations, yet managed as a single entity (Sunyaev & Sunyaev 2020). The drawback of traditional cloud models is that it emphasizes a centralized infrastructure. In contrast, distributed cloud environments allow companies to disperse computing resources close to the end user, thereby reducing latency and enhancing availability and scalability (Greenstein & Fang 2020). The main distributed cloud model providers are Amazon Web Services (AWS), Microsoft Azure, and Google Cloud, which provide localized services under a centralized control system. When data is able to be processed locally instead of on a remote data center, this architecture is excellent for edge computing applications, and applications that demand low latency and fewer data transfer requirements, e.g.: real-time applications, regulatory compliance (Saboor *et al*., 2022).

The performance, fault tolerance and operational efficiency are the advantage of distributed cloud environments (Mather *et al*., 2009). Organizations can have more disaster recovery and reduce the amount of traffic in the networks by using computing resources. Applications such as IoT networks, autonomous systems and financial trading which require low-latency processing can gain

from this architecture (Saboor *et al*., 2022). Additionally, it makes regulatory compliance easier by enabling businesses to handle and retain sensitive data within specific geographical bounds in order to comply with laws such as the California Consumer Privacy Act and the General Data Protection Regulation.

Businesses can be able to optimize the workloads on a dynamic basis because of the extra flexibility that the distributed cloud systems provides, which allow the resources to be transparently allocated on the basis of the demand (Sarathy *et al*., 2010).

One major hindrance in use of security measures in distributed cloud environment is security. The attack surface is increased as data gets processed complex, and kept in many different places; there are more chances of data breaches, illegal access and cyberattacks (Alkadi *et al*., 2020). In that way, security systems related to communication channel, data integrity, and access control should be applied to the outsourced globally distributed cloud nodes (Mushtaq *et al*., 2017). Interoperability problems may force organizations to provide consistency, synchronization of data in numerous cloud instances, and safety. Compliance with Let's talk about the following when you are working on the global stage: The appearance of the compliance problem is observed when companies act in numerous jurisdictions where data protection laws can be different. Given these risks, there are measures that have to be employed by organizations that use end-to-end encryption, more advanced identity management techniques, and real-time threat detection in order to ensure that their distributed cloud infrastructure remains secure (Alashhab *et al*., 2022).

2.4 Theoretical Frameworks

Using security models such as provenance-based, distributed trust model, and cloud security, it may be possible to secure, monitor, and ensure data integrity (Asante *et al*., 2021). The CIA Triad Concept (Confidentiality, Integrity and Availability) is the most well known and applied model in data security and is applied to construct safe cloud infrastructures. According to Wang *et al*. (2021), there are availability: need to assure when need data and service and when data available to data and services; confidentiality local access data to unauthorized personnel use data; and integrity ensures unmodified, reliable data formed. To safeguard from a large range of cyberthreats, system failure and unauthorised access requires particular advanced encryption, authentication protocols and redusancy techniques in this model which is particularly relevant to a distributed cloud environment (Habib *et al* 2022).

One of the best theoretical frameworks for data provenance is the PROV Data Model and it had been created by World Wide Web Consortium (W3C) (Zhang *et al*., 2020). In order for traceability of the data transformations and ownership changes to be tracked, throughout cloud infrastructures, this paradigm towards data and cloud infrastructure recommends a standardization of recording, representing, and sharing the provenance information (Closa *et al*., 2017). In order to supply a sort of provenance record, the PROV model defines entities (data objects), activities (modifications) and agents (users or systems executing actions). Provenance models such as PROV enable organizations to ensure the accountability of data, auditability of process, and compliance with laws like the GDPR and HIPAA. Provenance models can also assist forensic investigations as they will help organizations track unauthorized changes and discover anomalies in data usage (Pandey & Pande, 2021).

Theory behind the provenance based on blockchain and distributed trust mechanisms: By applying these theories-the Byzantine Fault Tolerance (BFT) Model and Decentralized Trust Theory, the systems are known as blockchain models (Hermstrüwer, 2020). BFT

Model handles difficulties experienced in distributed systems when nodes turn malicious or fail in an unexpected manner. Blockchain technology applies BFT principles and therefore, it guarantees that even when a certain number of nodes fail or even are compromised, it still achieves consensus and will still keep the integrity of data (Wang *et al*., 2022). According to Decentralized Trust Theory, trust can be established in a system without relying on a central authority, which is applicable in blockchain-based provenance tracking. In the case of cloud computing, this tamper-proof data logs, automated smart contracts, and secure data exchange without requiring the need of intermediaries (Hermstrüwer, 2021).

By Game Theory and Economic Models of Trust, distribution in common distributedly achieved security and integrity of data in the cloud is also possible (Gao *et al*., 2016). Game theory models the interaction between users, cloud providers, and attackers, predicting strategy under which data will be secured based on the incentives and risks (Esposito *et al*., 2020). Reputation-based trust models, which is an economic derivative theory, is employed to give the nodes or users a rating of trust based on their past actions in a distributed cloud environment. These models are vital to security in multi-cloud and hybrid cloud systems, in which multiple stakeholders share resources and require assuring each other of reliability. To combine these theoretical frameworks, researchers and cloud providers are working together in a strategic manner to create more robust, transparent, and resilient cloud computing technology that can address provenance, data security, and distributed trust mechanisms (Kirlar *et al*., 2018).

3. 0 Data Provenance in Distributed Cloud Environments

Data provenance refers to the procedure of documenting the information and the modifications and relocations that take place between the origin of the journey to the end, until the information gets to the ultimate processing stage. It provides companies with the instruments to ensure data integrity, protection, and conformity in scattered cloud environments (Imran & Agrawal, 2022). In cloud-based systems, provenance data can be utilized to keep information of data usage, transformation, and dependency- all of which is required to be audited and held accountable. The relations among entities, activities, agents as well as their interaction with timestamps are presented in Fig. 1 to offer a conceptual view of data provenance mechanism. These links provide a means of viewing the data generation, usage, and attribution in cloud distributed systems.

3.1 Types of Data Provenance

The process through which data has been modified and relocated across the source to the place of processing is known as data provenance (Imran & Agrawal, 2022). With a good understanding of the different kinds of data provenance, organizations will be able to ensure the security, compliance, and integrity of dispersed cloud system data. There are two important kinds of data provenance, including system vs application provenance and coarse vs fine granularity (Hu *et al*., 2020). The most suitable of the above-mentioned methods will depend mostly on the degree of detailing, the processing power that is available, and the circumstances that the data will be tracked (Simmhan *et al*., 2005).

3.2 Fine-Grained vs. Coarse-Grained Provenance

Generally speaking, fine-grained provenance records changes at extremely fine levels; it is frequently interested in monitoring changes at the byte, field or attribute level (Ruan *et al*., 2021). Financial transactions, medical data, and scientific research, just to name a few examples - where the smallest modifications may have a great impact - are among those domains in which fine-grained provenance is useful (Chapman *et al*., 2020). For audits and compliance purposes, fine-grained provenance is useful because it reflects an accuracy and

responsibility of a high degree. Performance is hurt by other challenges, such as processing and storage overhead costs, at least in very large-scale cloud systems. On the other hand, coarse-grained provenance reforms data at a higher level of abstraction, focusing on the most significant processes, workflow, or dependencies instead of tracking every minor change within a transformation (Oliveira *et al*., 2018). This method is mostly applied in business process management, enterprise cloud applications, and data pipeline tracking because it does not need much space to comprehend general flow but rather minor modifications (Herschel *et al*., 2017). Thus, coarse-grained provenance is more scalable and more storage efficient; thus, it is applicable for cloud computing purposes. However, it lacks the ability to trace fine details of changes that may be a drawback for some security-sensitive applications requiring forensic detail. Many organizations currently take a hybrid approach, where fine-grained provenance is available for some critical data points and coarse-grain tracking for the workflow view (Rupprecht *et al*., 2020).

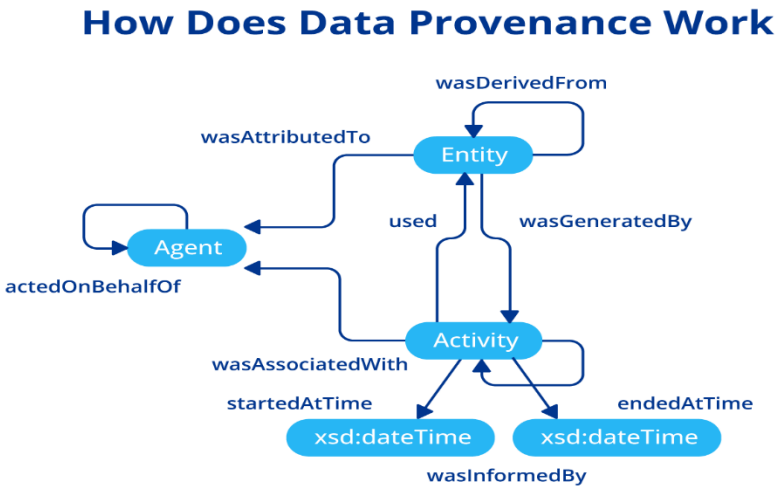

Fig. 1:Data Provenance in Distributed Cloud Environments (Astera, 2021)

3.3 System-Level vs. Application-Level Provenance

The other major distinction concerning data provenance is made between system-level and application-level provenance (Magagna *et al*., 2020). System-level provenance is automatically recorded by the underlying operating system, cloud infrastructure, or middleware for tracking the movement of data between different storage systems, networks, and computational nodes (Rupprecht *et al*., 2020). This type of provenance is hardware-independent and can be utilized to analyze system performance, detect security threats, and optimize resource allocation. Being such low-level provenance, it is usually coarse-grained and may fail to capture application-specific data changes (Muniswamy-Reddy *et al*., 2009).

On the flip side, application-level provenance is recorded in software applications, databases, or user-driven processes, capturing domain-specific transformations of data (Luczak-Rösch, 2014). For instance, application-level provenance in healthcare systems would track who modified a patient's medical record, what was changed, and why (Pinto *et al*., 2022). This form of provenance is fine-grained and caters to the unique requirements of an application, thus making it very useful for auditing, compliance, and debugging. However, it has to be custom implemented within each software system and might impose additional

computational and storage overhead if not managed well (Pinto *et al*., 2022).

3.4 Methods for Capturing Data Provenance

It has been stated that data provenance is essential for traceability, security, and compliance in a distributed cloud environment (Ametepe *et al*., 2021). And of these, the most favoured technique happens to be metadata logging, i.e. keeping the record of data that was created, modified or transferred. In such case, the metadata logs record dimensions including timestamps, user operations, system events and transformation history, thereby helping organizations ensure the life cycle of the data (Ametepe *et al*., 2021). Many cloud platforms and databases are focusing on the automation of metadata logging in order to integrate some parts of data governance, auditing, and security monitoring. While metadata logging offers a systematic and scalable means of capturing provenances, in large-scale systems, both due to both the vast amounts of data and the requirement for efficient storage/retrieval mechanisms, metadata logging quickly becomes resource-prohibitive (Imran *et al*., 2017).

Another way that data provenance can be guarded is through cryptographic hashing which can be used to ensure that data is representative and tamper-proof (Bany Taha 2015). A message hashing algorithm such as SHA-256 (Secure Hash Algorithm) generally generates a unique fingerprint of the data record that allows hand verification by the system to see if data is changed. In this sense, if a small change is made for data, the hash value will be different in magnitude, thus to warn the administrators about any unauthorized change. Cryptographic hashing algorithm is therefore applied extensively in digital signatures, file verification, and conformity audits to give assurances to the reliability of the data throughout its life cycle (Porkodi and Kesava-raja, 2021). However, this technique is change-oriented but does not record the complete history of changes, so it is the best to be used together with other provenance techniques (Li *et al*., 2022).

A state-of-the-art decentralization approach is the blockchain-based data provenance tracing (Soldatos *et al*., 2021). Immutability and tamper-proof recording of data transaction, which are the core qualities of blockchain, are perfect for ensuring maximum security in the most sensitive areas of operation, such as financial services, healthcare, and supply chain management. A data transact uniquely has a cryptographic linkage to a block position (which ultimately means any modified rewriting to the provenance record would necessitate the rewriting all future blocks, hence grossly computationally inhospitable) (Jyoti & Chauhan, 2022). In a blockchain-based provenance architecture, the precision and integrity of the data can be strengthened further, along with providing greater management of the independence of verification through smart contracts. However, despite such high security benefits, blockchain comes with scalability and performance issues as it is adopted in the cloud environment at a high volume with control over transaction processing speed and storage costs. By combining metadata logging, cryptographic hashing, and blockchain-based tracking, organizations are able to develop an encompassing, secure and verifiable framework for data provenance (Siddiqui *et al*., 2020).

4.0 Challenges in Data Provenance

Specifically, the main problem with data provenance in the dispersed cloud setting is scalability, privacy, and trust (Liang *et al*., 2017). With devices and services producing and processing vast amounts of data, and with humans continuing to seek the capability to monitor the entire lifecycle of data, data complexity continues to grow with time. Noor *et al*. (2013) observe that the incremental nature of record modifications requires provenance techniques to support the concept

of Big Data, WF-oriented operations, and real-time analytics in which data is continually updated without creating huge overheads in terms of provenance information processing and storage. A primary challenge with provenance to distributed computing can be how to provide coordinated provenance claims between multiple cloud nodes without compromising on speed. Distributed ledger management systems, indexing, and effective storage strategies are required to be efficient, since the conventional logging systems usually break in times of stress (Olawale *et al*., 2020).

Another problem is the privacy protection and the presence of legal frameworks. Without most of the specifics, some stringent laws such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) provide a guideline on how data should be used, stored, and retrieved (Mendelson, 2017). Provenance tracing also provides additional accountability and security, although security may be broken by revealing sensitive user information about data transformation.

Recording access to financial or medical information, such as but not limited to logging in, can prove to disclose confidential information such as patient history or details of transactions. To solve this issue, privacy-preserving provenance systems such as as anonymization, encryption, and access control systems should be implemented to facilitate compliance with it and perform effectively as tracing mechanisms (Torre *et al*., 2021).

In addition to the problem related to provenance manipulation and trust, cloud provenance management is a difficult topic. Since provenance records testify about the validity and reliability of data, malicious attackers might want to create their own records or seek solutions to cover illegal manipulations, security violations or fraud (Li *et al*., 2021). Conventional centralized provenance systems have the risk of insider attacks; whereby privileged administrators will attempt to conceal the traces of an incident by deleting or corrupting logs.

In the context of multi-cloud and third parties, organizations' complete trust in the external providers' ownership of accurate provenance records becomes an issue fearing incurs of mismanagement or malicious fabrication (Julakanti *et al*., 2022).

Blockchain technology has been suggested as a way to solve such problems and deliver tamper-proof provenance traceability in a decentralized manner (Westerlund *et al*., 2018). With blockchain, once provenance records have been written to the ledger, they cannot be changed without the agreement of all the nodes in the network (Shekhtman & Waisbard, 2021). This characteristic exacerbated trust and transparency and is very relevant in the field of supply chain management, financial systems and digital forensics (Batista *et al*., 2021). Nevertheless, blockchain does present challenges in scalability, as there is a wide spread ledger that requires large computational capacity and storage capacity to maintain. Continuously, research papers gravitate towards solutions, like off-chain storage and sharding and an optimized consensus algorithm, to strike the efficiency-security balance (Zhu *et al*., 2021).

Hence the data provenance problem cannot be solved without a multi-layered approach to address scalable, privacy, and trust problems at the same time. This involves combining efficient storage systems, privacy controls for compliance purposes and secure validation techniques (Kumar 2016). To accommodate both scalability and regulatory compliance, only privacy-preserving cryptographic means, AI-driven anomaly detection, and secure cloud architectures are suitable for ensuring that provenance tracking is taken full advantage of. By addressing these issues comprehensively, organizations can construct strategies that encourage trust, guarantee data integrity, and uphold compliances in dynamic digital landscapes (Mushtaq *et al*., 2022).

4.1 Ensuring Data Integrity in Distributed Cloud Environments

Finally, but not the least, data integrity is paramount to knowledge in the distributed cloud structures since it guarantees that knowledge is legal, trustful, and credible over its lifetime, between creation and destruction. The diagram in Figure 2 demonstrates that a combination of various methods, such as data validation, filtering, encryption, access controls, and audits, are used to ensure data continuity and integrity in the cloud.

Unauthorized and malicious modifications are among the threats that cloud data integrity is always vulnerable to (Zafar *et al*., 2017). An attacker that removes, adds, or changes some information in datasets is performing an operation of data tampering that can cause a fallacy which affects the effectiveness of operations and decision-making. Common vector attacks such as SQL injection, ransomware, and man in the middle (MITM) are commonly used to attack data when it's in motion or at rest. Intensive areas like healthcare, banking and law can have devastating impacts. For example, modifying financial transactions can facilitate fraud, whereas modifying medical records can result in incorrect treatments. Organizations should implement strong encryption, access controls and real-time verification procedures in order to counter these threats (Kaja *et al*., 2022).

One such worst threat is insider attacks, in which the staffs and regional administrators with special privileges may intentionally or unintentionally damage data (Saxena *et al*., 2022). In comparison to the external attackers, trusted insiders are more difficult to detect (Zhang, 2020). While carelessness can lead to accidental overwriting or data loss, employee discontentment can result in intentional manipulation of the data leading to creating disorder or releasing proprietary data for personal gain. According to a study (Silowash *et al*., 2012), RBAC, continuous monitoring, anomaly detection via artificial intelligence (AI), and zero-trust models are all necessary to identify insider threats. Another integrity problem for distributed systems is a replication complexity, where data spread on several servers or databases can become out of sync due to disconnection, synchronization failures, or delay (Tandel, 2022). Applications in banking, logistics, e-commerce, or other industries using such inflows generate risks of duplicates, outdated records and corrupted records, which may result from inconsistencies (Malik *et al*., 2016). To provide this, organizations need to implement robust data validation processes, integrity checks and conflict resolution to maintain consistency on geographically distributed systems (Gogineni, 2022).

Fig. 2: Data Integrity in Distributed Cloud Environments (Wall Street Mojo, 2020)

4.2 Techniques for Ensuring Data Integrity

Cryptographic hashing is one of the most popular techniques of ensuring data integrity through generating a uniquely fixed size hash value for any one given dataset (Hambouz *et al*., 2019). Hashing algorithms such as SHA-256 (Secure Hash Algorithm-256) and HMAC (Hash-based Message Authentication Code) work in the direction of making the digital fingerprint of the data, in which the systems can check if any nodes modified the same information (Kairaldeen *et al*, 2021). With not much modification in data, the total value of hash changes means that it is easy to detect any unauthorized change. In addition to the secret cryptographic key, HMAC has access to the extra level of protection, in other words, the generation and verification of hash values is possible only for authorized entities. Cryptographic hashing is primarily applied to file integrity monitoring, verifying databases, and secure transmission of data over computerized cloud networks (Kairaldeen *et al*., 2021).

The other important way of securing data integrity is the digital signatures that rely on Public Key Infrastructure (PKI) (Danquah & Kwabena-Adade 2020). Traditionally, digital signatures involve asymmetric encryption: Asymmetric encryption involves a private key that is employed to sign some data and a public key to verify for. This add-up to assure more data is changed between storing data and transmitting and tracing data through legally sender (Sagar Hossen *et al*, 2020). PKI, which is a larger security architecture, further provides secure key management and authentication and encryption mechanisms to withstand unauthorized modification. Such industries as finance, healthcare and government highly depend on PKI based digital certificates as protective mechanisms against sensitive transactions and regulation documents assuring data integrity and authenticity in that (Khan *et al*., 2022).

The blockchain along with the intelligent contracts offers decentralized and unchanging techniques for data integrity (Rahman *et al*., 2022). Blockchain keeps data in an indestructible and distributed ledger of transaction whereby each block is linked cryptophile to the preceding. This can prove out to be infeasible from a computation point of view when applying for wrong proof by an attempt of modifying data without changing all the blocks that follow the block containing what was changed. Hence, blockchain emerges as a good answer for trust in cloud environments, supply chains, and financial transactions (Tarafder *et al*., 2022). Smart contracts are basically self-executing pieces of code stored on the block chain, thus automation of integrity checks would be ensured by triggering actions based on the fulfillment of pre-set conditions. As such, data would be unalterable, transparent, and verifiable without reliance on centralized authorities (Ramachandran & Kantarcioglu, 2017).

Last but not least, Trusted Execution Environments (TEEs) and secure enclaves offer hardware-based security methods to prevent any potential manipulation of sensitive data (Schneider *et al*., 2022). TEEs, for instance, Intel SGX (Software Guard Extensions) and ARM TrustZone create isolated execution environments on a processor. These holistically secure critical operations from the possibility of compromise by other systems. When storing and processing data, these secure enclaves can be protective against malware, insider threats, and unauthorized access (Brandão *et al*. 2021). REEs are extensively utilized in cloud computing, financial transactions, and secure communications as they provide a trusted basis for data integrity verification and management in today's dispersed environment. By using these secure enclaves, organizations would be able to protect their cloud infrastructure against data manipulation and unauthorized changes, demonstrating that data that is trustworthy,

reliable, and immutable (Chakrabarti *et al*., 2022).

5.0 Emerging Technologies and Trends in Data Provenance and Integrity

5.1 Blockchain and Decentralized Ledgers

Blockchain and decentralised ledgers are considered to be the most cutting edge technology introduced for data authenticity and integrity (Deshpande *et al*., 2017). The blockchain offers a blockchain infrastructure for recording data provenance where data modifications are immutably recorded for auditing. Since the provenance records on a blockchain are cryptographically stored in a decentralized fashion, it eradicates the possibility of single points of failures and manipulation from within, which is usually the case for centralized systems of provenance (Madupati, 2021). Every transaction registered in a block chain has a timestamp, is signed cryptographically, and linked to the previous block, therefore forming a perpetual and verifiable history of data. This makes the blockchain an ideal tool to track the validity of data, prevent fraud, and ensure consistency on distributed cloud environments (Vagadia, 2020).

Several applied case studies are provided to show the effectiveness of blockchain-based provenance solutions in different industries (Xu *et al*., 2019). In the healthcare industry, enterprises like MedRec have built systems based on blockchain in tracking electronic medical records (EMRs) to ensure the security of data for patients, its auditability and tamper-proofing. In the context of food supply chains, the IBM Food Trust Blockchain offers companies like Walmart and Nestle the opportunity to trace the origin of food products whilst significantly reducing instances of fraud and/ or risk of contamination (Soldatos *et al*., 2021). The authors' conclusion is: For financial fraud and money laundering, the blockchain-based provenance in finance is being used; companies are tracing the source and transmission of digital assets using decentralized ledgers. There are many case studies which speak to the way in which blockchain fosters transparency, trust and accountability in provenance tracking (Dang, & Duong, 2021).

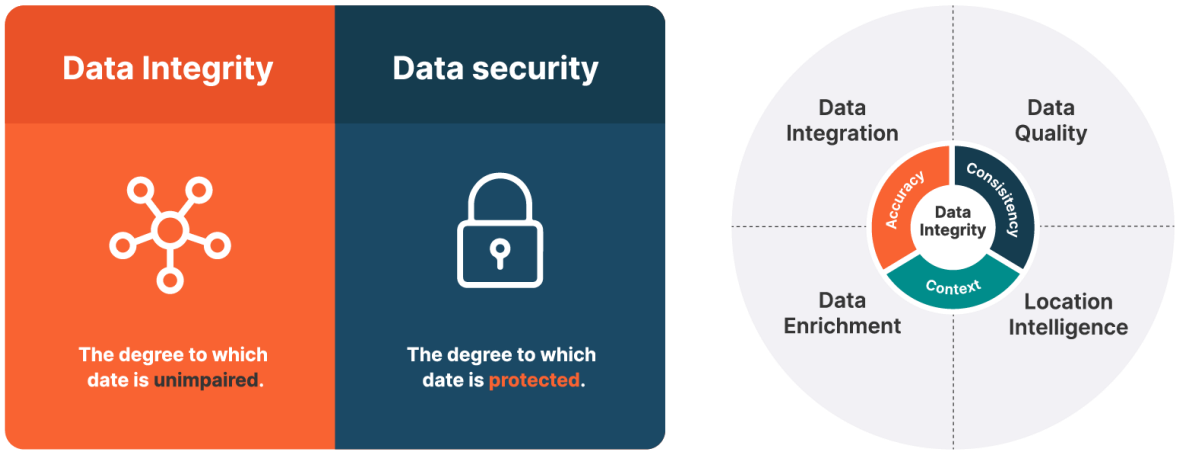

Fig. 3: Emerging Technologies and Trends in Data Provenance and Integrity (HubSpot, 2021

Some of the challenges confronting its realization include scalability, energy consumption, and regulatory implica-tions (Khan *et al*., 2021). Traditional blockchain systems such as those underpinning Bitcoin and Ethereum require great computing

resources to verify transactions, thus rendering them less practical in a high-volume cloud application. Layer 2 scaling, proof-of-stake (PoS) consensus, and hybrid blockchain models are under exploration as new solutions that would enhance efficiency while maintaining security (Madupati, 2021). Also, integrating privacy-keeping technologies (for instance, zero-knowledge proofs, homomorphic encryption) would assist in achieving equilibrium between transparency and data confidentiality while maintaining conformity with data protection regulations such as GDPR and HIPAA (Habib *et al*., 2022). The future of blockchain for data provenance and integrity will continue evolving with advancements in smart contracts, decentralized identity management, and anomaly detection supported by artificial intelligence (AI) (Jabbar *et al*., 2021). In an environment where organizations are accelerating their transition to a multi-cloud and edge computing architecture, blockchain-based provenance systems will play a significant role in supporting the integrity of distributed data systems and protecting against unauthorized modifications. With blockchain's immutable nature, cryptographic security, and decentralized trust mechanisms in place, organizations could create next-generation provenance systems to facilitate data integrity, provide verification of the data source, and satisfy auditability requirements attributed to regulations in an increasing digital world (Madupati, 2021).

5.2 AI and Machine Learning for Provenance Tracking

AI and ML are capable of benefiting the recent developments on the provenance tracking and data integrity checking tools for the distributed cloud environments (Soldatos *et al*., 2021). The traditional provenance systems typically use manual bookkeeping and rules-based tracking; this could be inefficient and rely on human beings that are prone to making mistakes. Automated provenance tracking using AI and ML does not only provide real-time pattern identification, but can also help with anomaly detection and predictive analysis, resulting in suspicious activities being detected as being out of the ordinary (Habib *et al*., 2022). In addition to the analysis of large groups of data, the use of artificial intelligence and machine learning also includes detection of inconsistent data that may indicate violation of integrity, with or without human intervention. Specifically, adding AI provenance tracking makes information management structure within organizations more transparent, secured, and compliant while reducing the risk of manipulation of data and even fraud (Habib *et al*., 2022).

One of the best use cases of AI about provenance tracking is the anomaly detection, where it detects an unusual pattern in the data flow, access log and modification history (Nedelkoski *et al*., 2019). machine learning (ML) algorithms can create a normal baseline of behaviour of data in order to detect deviations that might signal unsanctioned change, insider threat, or cyberattacks (Nedelkoski *et al*., 2019). By way of example, in financial transactions, suspicious transfers of funds or percentage modifications in data can be picked up with the help of AI-driven anomaly detection mechanisms as a fraud prevention mechanism along with compliance violations. In the healthcare industry, too, the same change will occur, where AI can be used to analyze the modification of EHR or unauthorized access to ensure that it does not breach any of the medical data systems and maintains its integrity and reliability (Zipperle *et al*., 2022).

One of the AI-based projects is automated auditing and it is a procedure where machine learning models assess provenance logs, ensure data integrity, and report violations for real-time alerts (Adelusi *et al*., 2022). Unlike typical audits that are carried out periodically and may not address the threats of security in real time, AI-based audits offer continuous monitoring

and instant determinations of threats (Lakarasu, 2022). However, the adoption of a cloud service provider has increasingly relied on the use of AI forensic tools to trace the vulnerability of data, determine malicious changes, and support the provision of regulatory compliance. AI can also optimize the tracking of provenance on blockchains as per the probed trends concerning data access and improved storage efficiency in distributed ledgers (Adelusi *et al*., 2022).

As AI and ML technologies evolve, their roles will only be magnified in ascertaining provenance integrity in multi-cloud and decentralized networks (Lakarasu, 2022). With improvements in federated learning, explainable AI (XAI), and AI-smart contracts, automated provenance verification will be further strengthened in the future, and bias, interpretability, and scalability challenges will also be addressed (Bellagarda & Abu-Mahfouz, 2022). Such organizations would be able to build emerging, self-learning provenance systems by joining AI-based anomaly detection with continuous auditing mechanisms in preventing data integrity breaches at the proactive level, thus enhancing the overall trust that would be placed in cloud environments (Ahmad *et al*., 2021).

6.0 Homomorphic Encryption and Secure Multiparty Computation

An increase in the reliance on a distributed cloud environment for sensitive data storage and processing by organizations invariably creates the challenge of ensuring privacy and integrity (Olawale *et al*., 2020). While conventional encryption techniques are used for protecting stored data and transmission, decrypting data exposes output to threats during computation process. Innovative approaches like Homomorphic Encryption (HE) and Secure Multiparty Computation (SMPC) can pave way for privacy-preserving data verification since these techniques allow computations on the encrypted data without revealing the underlying data. Such enhancements in cryptographic techniques are particularly beneficial in sectors like healthcare, finances, and government, as privacy is a big concern in such sectors (Aldossary & Allen, 2016).

Homomorphic Encryption (HE) performs computations directly on encrypted data, gives an output in encrypted form, and that output matches with the expected output generated with the same operation on plaintext data when decrypted, thus making computations without loss of confidentiality possible (Alaya *et al*., 2020). This makes it suitable to perform data processing operations with respect to confidentiality. Applications such as encrypted provenance tracing, outsourced data analytics and secure cloud computing are a good fit for it (Wood *et al*., 2020). Although there are a few classes of HE, such as fully homomorphic encryption (FHE), somewhat homomorphic encryption (SHE), and partially homomorphic encryption (PHE), the last class of HE is the most developed and allows for an infinite number of operations to be performed on the encrypted data. Nevertheless, HE has its own drawbacks such as high computational overload and low performance efficiency (Alharbi *et al*., 2020).

Another way of protecting privacy is secure multiparty computation, which allows multiple participants to collaborate to compute a function over their combined inputs, without releasing data about each individual input (Zhao *et al*., 2019). The SMPC maintains the confidentiality of inputs from each party, but permits them to know the final calculated outcome. These methods are important in federated learning, collaborative data analytics, and situations where the provenance is being monitored without disclosing raw data including when organizations share insights (Du & Atallah, 2001). SMPC enables fraud checks to be carried out between a number of banks without the disclosure of financial institution private transaction details and joint efforts may still be used to carry out medical

research utilizing patient details using encryption without violating some laws such as GDPR and HIPAA (Zhang *et al*., 2021).

With the caveat of security, homomorphic encryption and SMPC conjoined has boosted the data provenance and data integrity of the cloud settings (Zeiselmair *et al*., 2021). Apart from the prevention of insider threats without revealing the sensitive information, it can be used to validate modifications in data that may be managed for secure audits (Jayaraman & Mohammed, 2020). To optimize this technology for practical applications, there are a lot of hurdles to overcome, should it be in terms of scalability, computation efficiency, and even practical implementations. Future perspectives on privacy preserving data verification will be enhanced with the further advancements in quantum resistant cryptography, optimized cryptography schemas, AI based started cryptography optimization, hence enabling safe and high quality cloud computing. (Ghaffaripour, 2022)

6.1 Interoperability and Standardization Efforts

Interoperability and standardization in data provenance are needed for consistency, reliability, and compliance across the distributed cloud environment (Katari & Ankam, 2022). Due to the growing adoption of multi-cloud architectures and decentralized data processing together with blockchain-based provenance systems, organizations require standardized frameworks (Cheney *et al*., 2009). The absence of common standards will lead to provenance records being incompatible, fragmented, or unproven from different systems, which makes auditing, regulatory compliance, and cross-platform data integration problematic. Several international bodies and research initiatives have been developing provenance standards to enable seamless data tracking, validation, and exchange between different cloud infrastructures (Herschel *et al*., 2017).

The most famous framework have PROV under W3C from World Wide Web Consortium (W3C), which is known as Provenance Data Model (Wittner *et al*., 2022). The PROV model provides an organized path of representing and exchanging provenance data that identifies prov entities as data objects, activity being modification, or the agents involving the users or systems for changes (Closa *et al*., 2017). The application of PROV-O - an ontology for provenance representation - will be sufficient in such cases for ensuring compatibility across a multitude of cloud platforms, databases and analytics systems. Another important standard is the ISO 27037 which outlines guidelines on how to collect digital evidence and track forensic provenance. This standard is the most widely accepted standard in cybersecurity, digital forensics and legal compliance to ensure provences records can be innocently lawfully admissible and verifiable into inquiries (Missier *et al*., 2013).

Ongoing research is still trying to put more emphasis on the aspects of provenance standardization and interoperability, especially with the advent of AI-driven analytics, decentralized identity management and blockchain-based tracking systems (Shinde, 2022). The research work to be carried out is to be able to automatically validate provenance protocols computing cross-platorm metadata schemas and provenance ontologies machine-readable for vastly improved real-time data traceability and verification (Adelusi *et al*., 2022). Likewise, Open Provenance Model (OPM) and ISO/IEC 29100 Privacy Framework dream for global standards on provenance with respect to the privacy, security and data protection (Lewis *et al*., 2021).

Future developments on semantic web approaches, AI-driven provenance tracking and safe and secure multi-cloud provenance transfer protocols in future will improve interoperability and standardization more (Adelusi *et al*., 2022). It is then hoped that the

organizations running under the standardized provenance models would have a higher level of transparency, even while maintaining regulatory compliance, which would increase trust for the distributed cloud environment (Lakarasu, 2022). Adhering to global provens and interoperability standards would guarantee that companies have smooth governance of their data, thereby reducing the fragmentation risk and enhancing the borderless data sharing in an increasingly digital connected ecosystem (Mohna *et al*., 2022).

7.0 Case Studies and Practical Implementations

7.1 Real-world Use Cases:

Provenance and integrity techniques are needed for real-world applications to satisfy data needs, especially in the case of requiring security, traceability, and compliance (Adedeji, 2020). The healthcare, finance, and supply chain management are today moving towards adopting some technologies designed with some mechanisms such as blockchain, AI-driven anomaly detection, and cryptographic verification, which will ensure proving that data are tamperproof, auditable, and trust-certified (Singh *et al*., 2018). However, these technologies will provide solutions that will address more urgent matters such as operational efficiency, fraud identification, and compliance with rules. They are, therefore, the most useful for improving business processes and protecting sensitive data (Pasquier *et al*., 2018). Accordingly, one of the best examples of provenance and integrity applications is health data security, where the confidentiality and fidelity of electronic health records (EHRs) are very important (Margheri *et al*., 2020). Healthcare institutions are usually dealing with information about patients that are available in multiple systems, thus allowing unauthorized modifications, cyberattacks, and compliance breaches. A classic example of the latter is MedRec, an MIT-developed system whose main logic is to provide patients a secure, immutable storage type for their records using blockchain technology (Zarour *et al*., 2021). All changes to medical data are cryptographically recorded and verifiable according to MedRec. Thus, allowing a patient and his health providers to track those changes made under a record without any breach on HIPAA. It is possible to mention similar blockchain applications used in hospitals and pharmaceutical companies, like counterfeit drugs prevention and the assurance of interoperability of data and easy medical research procedures (Pandey *et al*., 2020).

The financial sector also attaches great importance to data integrity or provenance tracking to avoid fraud, support a transaction, or effectively meet the demands of Anti-Money Laundering (AML) (Elumilade *et al*., 2021). Banks use such blockchain-based ledger systems, and AI-driven anomaly detection systems to track records of transactions and alert suspicious activity at that point in time. In regard to views, JPMorgan chase also incorporates blockchain through cross-border payments and fraud detection to prevent such fraudulent fund transfers or manipulation of data (Edge and Sampaio, 2009). Similar to Mastercard, it has Provenance Verification System which is a system that verifies the history of transactions to prevent its fraud, using AI-based verifications and cryptography. By incorporating blockchain and AI, it will complete a network of secure, traceable, and tampered-proof trail of transactions that will enhance the confidence in a globally accepted banking system (Hasan *et al*., 2009).

The other relevant field of practical use of data provenance is that of supply chain management, where relevant businesses need to be aware of how to facilitate tracing authenticity of goods, traceability, and integrity in their transit within global logistical systems (Madanagopal *et al*., 2019). With provenance tracking, based on blockchains, the company may trace the provenance of the products to their origin, to whom they have been sold, and to what condition they were at each step of the

supply chain. A notable case is an IBM Food Trust, the blockchain-based system that was implemented by these corporations as Walmart, Nestle, and Unilever to operate their food supply chains with an aim of minimizing contamination risks (Wittner *et al*., 2022). This immediate visibility is offered by IBM Food Trust since every transaction can be tracked as a part of an immutable ledger, thus allowing businesses to identify fraudulent suppliers instantly, damaged goods, or fake products Brandín & Abrishami, 2021).

All these effective applications of the data provenance and data integrity solutions in healthcare, finance, and currently with supply chain applications show that the need to ensure secure data tracking is growing across industries (Clauson *et al*., 2018). Naturally, such case studies have demonstrated the potential of blockchain, AI, and cryptographic methods. As with any other technology, one will have always some problematic aspects on the one hand, including the problem of scalability, the regulatory compliance issues, etc. (Yaqoob *et al*., 2022). The future development of quantum-proof encryption, decentralized identity, and mechanisms based on AI to perform provenance tracking, will definitely improve the adoption of secure data provenance solutions in critical areas (Yaqoob *et al*., 2022).

It is what the organizations will benefit by the exploitation of blockchain, AI, and cryptography to all the security, forecasting, and other functionalities of the mentioned technologies as they can increase their trust in data and minimize fraud risks and simplify compliance (Jamil *et al*., 2019). As such, only the enviable future will see security, auditability, and fictitiousness of data being addressed as more industries use provenance-aware systems (Yaqoob *et al*., 2022).

8.0 Lessons Learned and Best Practices

Through the experience of painting, other lessons could be gleaned from the challenges, strengths and best practices of data security in distributed systems from the implementation of data provenance and integrity techniques in other sectors (Hasan *et al*., 2007). One of these was that Data provenance should not be an afterthought but built into system architecture from the onset. Indeed, attempting to add provenance tracking features to existing systems has been problematic for many organizations, which amplifies security risk and degrades efficiency (Hu *et al*., 2020). Organizations that employed blockchain-provenance, AI-based anomaly detections and cryptographic validations strongly asserted the importance of cross-platform interoperability, efficient data governance protocols, and pre-planning for provenance tracking to function effectively (Ikegwu *et al*., 2022).

Provisioning, additionally, has lessons to learn from transparency against privacy in the case of dissemination. Even though detailed records of provenance account for more accountability, they are also likely to hold sensitive business or private information and create risks of non-compliance under regulations like GDPR or HIPAA (Cobbe *et al*.,2020). Privacy-preserving mechanisms like homomorphic encryption, secure multiparty computation (SMPC), and zero-knowledge proofs should be incorporated in organizations to maintain data secrecy while signifying approval for verifiability (Cobbe *et al*., 2020). The incorporation of techniques based on RBAC and user authentication mechanisms into provenance systems has been proven by industry leaders to mitigate risks while offering appropriate access permission privilege to authorized parties regarding important provenance data (Alansari, 2020).

Scalability is one of the issues faced by organizations in their pursuit of provenance solutions mostly in real-time, highly transaction-intensive operational environments such as financial transactions and supply chain logistics (Alam & Roy, 2022). But, using the hybrid model with on-chain and off-chain storage has been beneficial in demonstrating

how we can increase efficiency (Malik *et al*., 2018). For less important metadata, yet less sensitive, audit logs that are critical to contain on-chain audit logs, should instead utilize security, but controlled cloud storage, rather than having the provenance data completely and/or partially stored in the blockchain networks by capturing all the internal nodes. It still protects the provenance record's integrity, but decreases the computational overhead, as well as storage costs (Yang & Xu, 2016).

In conclusion, practitioners will readily assert that effectively and proactively tracking provenance necessitates automation and AI-based monitoring (Yang *et al*., 2016). Rule-based monitoring and manual audits are not sufficient to navigate our ever-changing large-scale data context. The detection, deterrence, and reaction to fraud, data breaches, and unauthorized data changes will greatly improve for companies that employ machine learning to automate anomaly detection, compliance reporting, and real-time verification of integrity (Odetunde *et al*., 2022). In the future, enterprises should continue looking into broadly interpreted frontiers to help improve and strengthen data provenance and integrity across Cloud and Distributed Systems, such as new AI-powered tools, decentralized trust frameworks, and adaptive security models (Litke *et al*., 2019).

9.0 Open Challenges and Future Research Directions

Although the improvement of data integrity and provenance is truly unbelievable, numerous challenges to be overcome remain, especially regarding the ability to store and retrieve data in large amounts (Wang *et al*., 2015). This is because it is not computationally efficient when monitoring changes and access of large volumes of data generated by enterprises in a multi-cloud and hybrid environment (Hu *et al*., 2020). A variety of scaling challenges are generally associated with blockchain-provenance systems to process transaction latency and storage cost, notwithstanding its provenance immutability, such that future studies should look at lightweight provenance models, sharding strategies, and AI-driven data-compression strategies. Also, the federation of learning and edge computing development can assist in spreading provenance loads better, and this approach will decrease the load on centralized cloud systems (Alam & Roy, 2022). Another serious concern of provenance tracing is that it is difficult to strike a balance between privacy and transparency. The granular security and compliance aspects of businesses require extensive provenance records that can be audited, and the need to maintain such massive logs puts the company at risk which is contrary to several laws, such as the GDPR and HIPAA (Tan *et al*., 2018). To allow organizations to verify the authenticity of the underlying data without sharing confidential material, future studies need to cover privacy-preserving provenance methods like the differential privacy, zero-knowledge proofs (ZKP) and secure multiparty computation (SMPC). It will be very important to find a reasonable balance between traceability and confidentiality to develop next-generation provenance models that support user privacy interests and regulatory needs (Mihai *et al*., 2022). AWS, Azure, and GCP are proprietary cloud platforms of different levels of openness and interoperability, which makes the integration of provenance tracking with these leading cloud service providers challenging (Galiveeti *et al*., 2021). Although providers offer logging tools such as AWS CloudTrail, Azure Monitor, and Google Cloud Audit Logs, these do not often interoperate across platforms, creating a challenge for enterprises trying to maintain unified provenance records across multi-cloud environments (Quadri, 2017). Future research should look toward standardized provenance APIs, interoperability frameworks, and decentralized identity management, thus enabling easy integration across all available cloud platforms. In addition, smart contract-

based automation can offer an extra layer of multi-cloud data traceability; thus, guaranteeing the consistency and verifiability of provenance records across different providers (Zou *et al*., 2021).

The last, but by no means least, lingering hurdles imposed by regulatory and legal considerations pertaining to the implementation of provenance, as they set different compliance requirements for data tracking, for retention, and for auditing from country to country and from industry to industry (May, 2005). Regulations which include GDPR, CCPA, and ISO 27001 bind organizations to keeping accurate, clear, and verifiable provenance records while recognizing user rights for deleting and amending it (Kunz *et al*., 2020). Going forward, research should study adaptive compliance models that allow organizations align their retention of provenance policies with local legal requirements. Subsequently, machine-readable compliance frameworks and AI-powered tools for regulatory auditing might be developed to automate policy enforcement and guarantee that progeny tracking fulfills the conformity of shifting global standards (Katari, & Ankam, 2022).

10 Conclusion

This paper has made an emphasis on the importance of tracing the origin of data, modifications to the same, and the volume of data that has been transferred down the line concerning security, compliance and trust concerns. It also claims data provenance and integrity to be a critical concern, particularly to distributed cloud environments, where the only hope appears to be scalability, privacy, and trust concerns. The participants appear to be reassured by the fact that blockchain, equipped with the mechanisms such as hash cryptography, AI powered anomaly detection prototype, homomorphic encryption, and secure multiparty computation (SMPC), can pass the test of privacy preservation. W3C PROV and ISO 27037 are other standardization-oriented solutions that offer guidelines that eliminate the necessity to enhance compliance and interoperability in the cloud-based provenance systems. It is hoped that this research activity will uphold the scientific and governance principles of integration security models, authenticity of data, and creation of innovative provenance methods. The paper will also in the light of emerging technologies, give some of the recommended practices toward the development of the reliable provenance-aware systems in the variety of theoretical, practical, and technological fields. The biggest assumption that the paper points out as being the most critical in privacy-oriented provenance tracking under the jurisdiction of both the GDPR and PHI on compliance aspects of the privacy regulation is that transparency implies concealment. The heated discussions have covered several topics related to the decentralized provenance models and cloud integration, with new possibilities of studying multicluster traceability and interoperability. The hybrid solutions, an off-chain and on-chain storage combination to maintain security and efficiency and introduce better data provenance and integrity procedures, are likely to assist organizations. The most commonly used implementation in the future is the federated learning technique as well as AI-based surveillance and multiple adaptive cryptography methods that will enable scaling and tracking with privacy. The provision of additional features needed to support cross-platform provenance systems would ensure an easy integrative interface with service providers of different clouds (AWS, Azure, and GCP) without losing certain semblance of regulatory frameworks and compliance.

The creation of international provenance standards that can consider the ethical and legal

issues related to the governance and security of data then comes into the picture for policy makers. The development of AI-driven regulatory auditing tools, scaling of Blockchain, and development of technologies to enhance real-time verification of provenance should be the prime areas for future study. Distributed cloud systems will keep changing industries. Building trust, lowering fraud, and maintaining security for important data to be used across industries all rely on secure, unalterable and tamper-proof traceback. Technology advancement brought with some platforms where firms, researchers, and the global community may subsume to build an immense amount of clarity all around the digital ecosystem, an ecosystem from which the stakeholders can identify and pinpoint the positions, which should be fortified with complete truth. Moreover, this frame of argument in technology will address both being able to sustain accuracy, accountability, robustness, and support against all of the known cyber threats starting to multiply.

11.0 References

Adedeji, P. A. (2020). *Hybrid renewable energy-based facility location: a Geographical Information System (GIS) integrated multi-criteria decision-making (MCDM) approach*. University of Johannesburg (South Africa).

Adelusi, B. S., Ojika, F. U., & Uzoka, A. C. (2022). Advances in Data Lineage, Auditing, and Governance in Distributed Cloud Data Ecosystems.

Ahmad, W., Rasool, A., Javed, A. R., Baker, T., & Jalil, Z. (2021). Cyber security in iot-based cloud computing: A comprehensive survey. *Electronics*, *11*(1), 16.

Alam, K., & Roy, B. (2022). Challenges of provenance in scientific workflow management systems. In *2022 IEEE/ACM Workshop on Workflows in Support of Large-Scale Science (WORKS)* (pp. 10-18). IEEE.

Alansari, S. (2020). *A blockchain-based approach for secure, transparent and accountable personal data sharing* (Doctoral dissertation, University of Southampton).

Alashhab, Z. R., Anbar, M., Singh, M. M., Hasbullah, I. H., Jain, P., & Al-Amiedy, T. A. (2022). Distributed denial of service attacks against cloud computing environment: survey, issues, challenges and coherent taxonomy. *Applied Sciences*, *12*(23), 12441.

Alaya, B., Laouamer, L., & Msilini, N. (2020). Homomorphic encryption systems statement: Trends and challenges. *Computer Science Review*, *36*, 100235.

Aldossary, S., & Allen, W. (2016). Data security, privacy, availability and integrity in cloud computing: issues and current solutions. *International Journal of Advanced Computer Science and Applications*, *7*(4).

Aldossary, S., & Allen, W. (2016). Data security, privacy, availability and integrity in cloud computing: issues and current solutions. *International Journal of Advanced Computer Science and Applications*, *7*(4).

Aldossary, S., & Allen, W. (2016). Data security, privacy, availability and integrity in cloud computing: issues and current solutions. *International Journal of Advanced Computer Science and Applications*, *7*(4).

Alharbi, A., Zamzami, H., & Samkri, E. (2020). Survey on homomorphic encryption and address of new trend. *International Journal of Advanced Computer Science and Applications*, *11*(7).

Alkadi, O., Moustafa, N., & Turnbull, B. (2020). A review of intrusion detection and blockchain applications in the cloud: approaches, challenges and solutions. *IEEE Access*, *8*, 104893-104917.

Ametepe, W., Wang, C., Ocansey, S. K., Li, X., & Hussain, F. (2021). Data provenance collection and security in a distributed environment: a survey. *International Journal of Computers and Applications*, *43*(1), 11-25.

Ametepe, W., Wang, C., Ocansey, S. K., Li, X., & Hussain, F. (2021). Data provenance collection and security in a distributed environment: a survey. *International Journal of Computers and Applications*, *43*(1), 11-25.

Asante, M., Epiphaniou, G., Maple, C., Al-Khateeb, H., Bottarelli, M., & Ghafoor, K. Z. (2021). Distributed ledger technologies in supply chain security management: A comprehensive survey. *IEEE Transactions on Engineering Management*, *70*(2), 713-739.

Astera. (2021). *Data provenance in distributed cloud environments* [Diagram]. Retrieved from https://www.astera.com/wp-content/uploads/2021/05/Untitled-3-2.png

Bany Taha, M. M. M. (2015). *Tamper-Evident Data Provenance* (Doctoral dissertation, University of Waikato).

Batista, D., Kim, H., Lemieux, V. L., Stancic, H., & Unnithan, C. (2021). Blockchains and provenance: How a technical system for tracing origins, ownership and authenticity can transform social trust. *Building decentralized trust: multidisciplinary perspectives on the design of blockchains and distributed ledgers*, 111-128.

Bellagarda, J. S., & Abu-Mahfouz, A. M. (2022). An updated survey on the convergence of distributed ledger technology and artificial intelligence: Current state, major challenges and future direction. *IEEe Access*, *10*, 50774-50793.

Bettivia, R., Cheng, Y. Y., & Gryk, M. R. (2022). *Documenting the future: navigating provenance metadata standards*. Cham: Springer.

Brandão, A., Resende, J. S., & Martins, R. (2021). Hardening cryptographic operations through the use of secure enclaves. *Computers & Security*, *108*, 102327.

Brandín, R., & Abrishami, S. (2021). Information traceability platforms for asset data lifecycle: blockchain-based technologies. *Smart and Sustainable Built Environment*, *10*(3), 364-386.

Chakrabarti, S., Knauth, T., Kuvaiskii, D., Steiner, M., & Vij, M. (2020). Trusted execution environment with intel sgx. In *Responsible Genomic Data Sharing* (pp. 161-190). Academic Press.

Chapman, A., Missier, P., Simonelli, G., & Torlone, R. (2020). Capturing and querying fine-grained provenance of preprocessing pipelines in data science. *Proceedings of the VLDB Endowment*, *14*(4), 507-520.

Cheney, J., Chong, S., Foster, N., Seltzer, M., & Vansummeren, S. (2009, October). Provenance: a future history. In *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications* (pp. 957-964).

Cheney, J., Chong, S., Foster, N., Seltzer, M., & Vansummeren, S. (2009, October). Provenance: a future history. In *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications* (pp. 957-964).

Clauson, K. A., Breeden, E. A., Davidson, C., & Mackey, T. K. (2018). Leveraging Blockchain Technology to Enhance Supply Chain Management in Healthcare:: An exploration of challenges and opportunities in the health supply chain. *Blockchain in healthcare today*.

Closa, G., Masó, J., Proß, B., & Pons, X. (2017). W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed

environment. *Computers, Environment and Urban Systems*, *64*, 103-117.

Closa, G., Masó, J., Proß, B., & Pons, X. (2017). W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment. *Computers, Environment and Urban Systems*, *64*, 103-117.

Cobbe, J., Norval, C., & Singh, J. (2020). What lies beneath: Transparency in online service supply chains. *Journal of Cyber Policy*, *5*(1), 65-93.

Cobbe, J., Norval, C., & Singh, J. (2020). What lies beneath: Transparency in online service supply chains. *Journal of Cyber Policy*, *5*(1), 65-93.

Dang, T. K., & Duong, T. A. (2021). An effective and elastic blockchain-based provenance preserving solution for the open data. *International Journal of Web Information Systems*, *17*(5), 480-515.

Danquah, P., & Kwabena-Adade, H. (2020). Public key infrastructure: an enhanced validation framework. *Journal of Information Security*, *11*(4), 241-260.

Deshpande, A., Stewart, K., Lepetit, L., & Gunashekar, S. (2017). Distributed Ledger Technologies/Blockchain: Challenges, opportunities and the prospects for standards. *Overview report The British Standards Institution (BSI)*, *40*(40), 1-34.

Dib, O., & Rababah, B. (2020). Decentralized identity systems: Architecture, challenges, solutions and future directions. *Annals of Emerging Technologies in Computing (AETiC)*, *4*(5), 19-40.

Du, W., & Atallah, M. J. (2001, September). Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 workshop on New security paradigms* (pp. 13-22).

Edge, M. E., & Sampaio, P. R. F. (2009). A survey of signature based methods for financial fraud detection. *computers & security*, *28*(6), 381-394.

Elumilade, O. O., Ogundeji, I. A., Achumie, G. O., Omokhoa, H. E., & Omowole, B. M. (2021). Enhancing fraud detection and forensic auditing through data-driven techniques for financial integrity and security. *Journal of Advanced Education and Sciences*, *1*(2), 55-63.

Esposito, C., Tamburis, O., Su, X., & Choi, C. (2020). Robust decentralised trust management for the internet of things by using game theory. *Information Processing & Management*, *57*(6), 102308.

Galiveeti, S., Tawalbeh, L. A., Tawalbeh, M., & El-Latif, A. A. A. (2021). Cybersecurity analysis: Investigating the data integrity and privacy in AWS and Azure cloud platforms. In *Artificial intelligence and blockchain for future cybersecurity applications* (pp. 329-360). Cham: Springer International Publishing.

Gao, L., Yan, Z., & Yang, L. T. (2016). Game theoretical analysis on acceptance of a cloud data access control system based on reputation. *IEEE Transactions on Cloud Computing*, *8*(4), 1003-1017.

Ghaffaripour, S. (2022). *Novel Solutions for Privacy, Security and Trust in Modern Medical Data Management Systems* (Doctoral dissertation, Toronto Metropolitan University).

Gogineni, A (2022). Consistency Models for Cross-Cluster Data Synchronization in Large-Scale Multi-Tenant Architectures. *IJSAT-International Journal on Science and Technology*, *16*(1).

Greenstein, S. M., & Fang, T. P. (2020). *Where the Cloud Rests: The Economic Geography of Data Centers*. Harvard Business School.

Groth, P. (2007). *The origin of data: Enabling the determination of provenance in multi-institutional scientific systems through the documentation of processes* (Doctoral dissertation, University of Southampton).

Gudala, L., Reddy, A. K., Sadhu, A. K. R., & Venkataramanan, S. (2022). Leveraging biometric authentication and blockchain

technology for enhanced security in identity and access management systems. *Journal of Artificial Intelligence Research*, *2*(2), 21-50.

Gudala, L., Reddy, A. K., Sadhu, A. K. R., & Venkataramanan, S. (2022). Leveraging biometric authentication and blockchain technology for enhanced security in identity and access management systems. *Journal of Artificial Intelligence Research*, *2*(2), 21-50.

Habib, G., Sharma, S., Ibrahim, S., Ahmad, I., Qureshi, S., & Ishfaq, M. (2022). Blockchain technology: benefits, challenges, applications, and integration of blockchain technology with cloud computing. *Future Internet*, *14*(11), 341.

Hambouz, A., Shaheen, Y., Manna, A., Al-Fayoumi, M., & Tedmori, S. (2019, October). Achieving data integrity and confidentiality using image steganography and hashing techniques. In *2019 2nd international conference on new trends in computing sciences (ICTCS)* (pp. 1-6). IEEE.

Hasan, R., Sion, R., & Winslett, M. (2007, October). Introducing secure provenance: problems and challenges. In *Proceedings of the 2007 ACM workshop on Storage security and survivability* (pp. 13-18).

Hasan, R., Sion, R., & Winslett, M. (2009). Preventing history forgery with secure provenance. *ACM Transactions on Storage (TOS)*, *5*(4), 1-43.

Hermstrüwer, Y. (2020). The limits of blockchain democracy: A transatlantic perspective on blockchain voting systems. *TTLF Working Papers*, (49).

Hermstrüwer, Y. (2021). Blockchain and public administration. In *Blockchain and Public Law* (pp. 105-122). Edward Elgar Publishing.

Herschel, M., Diestelkämper, R., & Ben Lahmar, H. (2017). A survey on provenance: What for? What form? What from?. *The VLDB Journal*, *26*(6), 881-906.

Herschel, M., Diestelkämper, R., & Ben Lahmar, H. (2017). A survey on provenance: What for? What form? What from?. *The VLDB Journal*, *26*(6), 881-906.

Honar Pajooh, H., Rashid, M. A., Alam, F., & Demidenko, S. (2021). IoT Big Data provenance scheme using blockchain on Hadoop ecosystem. *Journal of Big Data*, *8*, 1-26.

Hu, R., Yan, Z., Ding, W., & Yang, L. T. (2020). A survey on data provenance in IoT. *World Wide Web*, *23*, 1441-1463.

HubSpot. (2021). *Emerging technologies and trends in data provenance and integrity* [Diagram]. Retrieved from https://4326216.fs1.hubspotusercontentna1.net/hubfs/4326216/Pillars%2520of%2520Data%2520Integrity.png

Hussain, A. A., & Al-Turjman, F. (2021). Artificial intelligence and blockchain: A review. *Transactions on emerging telecommunications technologies*, *32*(9), e4268.

Ikegwu, A. C., Nweke, H. F., Anikwe, C. V., Alo, U. R., & Okonkwo, O. R. (2022). Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions. *Cluster Computing*, *25*(5), 3343-3387.

Imran, A., & Agrawal, R. (2022). Data provenance. In *Encyclopedia of big data* (pp. 317-321). Cham: Springer International Publishing.

Imran, M., & Hlavacs, H. (2012). Provenance in the cloud: Why and how?. *CLOUD COMPUTING*, 106-112.

Imran, M., Hlavacs, H., Haq, I. U., Jan, B., Khan, F. A., & Ahmad, A. (2017). Provenance based data integrity checking and verification in cloud environments. *PloS one*, *12*(5), e0177576.

Imran, M., Hlavacs, H., Haq, I. U., Jan, B., Khan, F. A., & Ahmad, A. (2017). Provenance based data integrity checking and verification in cloud environments. *PloS one*, *12*(5), e0177576.

Jabbar, S., Lloyd, H., Hammoudeh, M., Adebisi, B., & Raza, U. (2021). Blockchain-enabled supply chain: analysis, challenges, and future directions. *Multimedia systems*, *27*, 787-806.

Jamil, F., Hang, L., Kim, K., & Kim, D. (2019). A novel medical blockchain model for drug supply chain integrity management in a smart hospital. *Electronics*, *8*(5), 505. Jamil, F., Hang, L., Kim, K., & Kim, D. (2019). A novel medical blockchain model for drug supply chain integrity management in a smart hospital. *Electronics*, *8*(5), 505.

Jayaraman, I., & Mohammed, M. (2020). Secure privacy conserving provable data possession (SPC-PDP) framework. *Information Systems and e-Business Management*, *18*(3), 351-377.

Julakanti, S. R., Sattiraju, N. S. K., & Julakanti, R. (2022). Multi-cloud security: strategies for managing hybrid environments. *NeuroQuantology*, *20*(11), 10063-10074.

Jyoti, A., & Chauhan, R. K. (2022). A blockchain and smart contract-based data provenance collection and storing in cloud environment. *Wireless Networks*, *28*(4), 1541-1562.

Kairaldeen, A. R., Abdullah, N. F., Abu-Samah, A., & Nordin, R. (2021). Data integrity time optimization of a blockchain IoT smart home network using different consensus and hash algorithms. *Wireless Communications and Mobile Computing*, *2*(1), 4401809.

Kaja, D. V. S., Fatima, Y., & Mailewa, A. B. (2022). Data integrity attacks in cloud computing: A review of identifying and protecting techniques. *Journal homepage: www. ijrpr. com ISSN*, *2582*, 7421.

Katari, A., & Ankam, M. (2022). Data Governance in Multi-Cloud Environments for Financial Services: Challenges and Solutions. *Educational Research (IJMCER)*, *4*(1), 339-353.

Khan, D., Jung, L. T., & Hashmani, M. A. (2021). Systematic literature review of challenges in blockchain scalability. *Applied Sciences*, *11*(20), 9372.

Khan, S., Luo, F., Zhang, Z., Rahim, M. A., Ahmad, M., & Wu, K. (2022). Survey on issues and recent advances in vehicular public-key infrastructure (VPKI). *IEEE Communications Surveys & Tutorials*, *24*(3), 1574-1601.

Kırlar, B. B., Ergün, S., Alparslan Gök, S. Z., & Weber, G. W. (2018). A game-theoretical and cryptographical approach to crypto-cloud computing and its economical and financial aspects. *Annals of Operations Research*, *260*(1), 217-231.

Kommisetty, P. D. N. K. (2022). Leading the Future: Big Data Solutions, Cloud Migration, and AI-Driven Decision-Making in Modern Enterprises. *Educational Administration: Theory and Practice*, *28*(03), 352-364.

Kumar, T. V. (2016). Layered App Security Architecture for Protecting Sensitive Data.

Kumar, V., & Poornima, G. (2012). Ensuring data integrity in cloud computing. *Journal of Computer Applications*, *5*(4), 513-520.

Kunz, I., Casola, V., Schneider, A., Banse, C., & Schütte, J. (2020, October). Towards tracking data flows in cloud architectures. In *2020 IEEE 13th International Conference on Cloud Computing (CLOUD)* (pp. 445-452). IEEE.

Lakarasu, P. (2022). End-to-end Cloud-scale Data Platforms for Real-time AI Insights. *Available at SSRN 5267338*.

Lakarasu, P. (2022). End-to-end Cloud-scale Data Platforms for Real-time AI Insights. *Available at SSRN 5267338*.

Lewis, D., Filip, D., & Pandit, H. J. (2021). An ontology for standardising trustworthy AI. *Factoring Ethics in Technology, Policy Making, Regulation and AI*, *65*.

Li, L., & Zhang, J. (2021). Research and analysis of an enterprise E-commerce marketing system under the big data environment. *Journal of Organizational and End User Computing (JOEUC)*, *33*(6), 1-19.

Li, W., Wu, J., Cao, J., Chen, N., Zhang, Q., & Buyya, R. (2021). Blockchain-based trust management in cloud computing systems: a taxonomy, review and future directions. *Journal of Cloud Computing*, *10*(1), 35.

Li, X., Wei, L., Wang, L., Ma, Y., Zhang, C., & Sohail, M. (2022). A blockchain-based privacy-preserving authentication system for ensuring multimedia content integrity. *International journal of intelligent systems*, *37*(5), 3050-3071.

Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., & Njilla, L. (2017, May). Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)* (pp. 468-477). IEEE.

Lim, S. Y., Fotsing, P. T., Almasri, A., Musa, O., Kiah, M. L. M., Ang, T. F., & Ismail, R. (2018). Blockchain technology the identity management and authentication service disruptor: a survey. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(4-2), 1735-1745.

Litke, A., Anagnostopoulos, D., & Varvarigou, T. (2019). Blockchains for supply chain management: Architectural elements and challenges towards a global scale deployment. *Logistics*, *3*(1), 5.

Luczak-Rösch, M. (2014). *Usage-dependent maintenance of structured Web data sets* (Doctoral dissertation).

Madanagopal, K., Ragan, E. D., & Benjamin, P. (2019). Analytic provenance in practice: The role of provenance in real-world visualization and data analysis environments. *IEEE Computer Graphics and Applications*, *39*(6), 30-45.

Maddukuri, N. (2021). Trust in the cloud: Ensuring data integrity and auditability in BPM systems. *International Journal of Information Technology and Management Information Systems*, *12*(1), 144-160.

Madupati, B. (2021). Blockchain in Day-to-Day Life: Transformative Applications and Implementation. *Available at SSRN 5118207*.

Magagna, B., Goldfarb, D., Martin, P., Atkinson, M., Koulouzis, S., & Zhao, Z. (2020). Data provenance. In *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges* (pp. 208-225). Cham: Springer International Publishing.

Malik, S. U. R., Khan, S. U., Ewen, S. J., Tziritas, N., Kolodziej, J., Zomaya, A. Y., ... & Li, H. (2016). Performance analysis of data intensive cloud systems based on data management and replication: a survey. *Distributed and Parallel Databases*, *34*(2), 179-215.

Margheri, A., Masi, M., Miladi, A., Sassone, V., & Rosenzweig, J. (2020). Decentralised provenance for healthcare data. *International Journal of Medical Informatics*, *141*, 104197.

Mather, T., Kumaraswamy, S., & Latif, S. (2009). *Cloud security and privacy: an enterprise perspective on risks and compliance*. " O'Reilly Media, Inc.".

Mather, T., Kumaraswamy, S., & Latif, S. (2009). *Cloud security and privacy: an enterprise perspective on risks and compliance*. " O'Reilly Media, Inc.".

May, P. J. (2005). Regulatory implementation: Examining barriers from regulatory processes. *Cityscape*, 209-232.

Mendelson, D. (2017). Legal protections for personal health information in the age of Big Data–a proposal for regulatory

framework. *Ethics, Medicine and Public Health*, *3*(1), 37-55.

Mihai, S., Yaqoob, M., Hung, D. V., Davis, W., Towakel, P., Raza, M., ... & Nguyen, H. X. (2022). Digital twins: A survey on enabling technologies, challenges, trends and future prospects. *IEEE Communications Surveys & Tutorials*, *24*(4), 2255-2291.

Missier, P., Belhajjame, K., & Cheney, J. (2013, March). The W3C PROV family of specifications for modelling provenance metadata. In *Proceedings of the 16th international conference on extending database technology* (pp. 773-776).

Mohna, H. A., Barua, T., Mohiuddin, M., & Rahman, M. M. (2022). AI-ready data engineering pipelines: a review of medallion architecture and cloud-based integration models. *American Journal of Scholarly Research and Innovation*, *1*(01), 319-350.

Muniswamy-Reddy, K. K., Braun, U., Holland, D., Macko, P., Maclean, D., Margo, D., ... & Smogor, R. (2009). Layering in provenance systems.

Mushtaq, M. F., Akram, U., Khan, I., Khan, S. N., Shahzad, A., & Ullah, A. (2017). Cloud computing environment and security challenges: A review. *International Journal of Advanced Computer Science and Applications*, *8*(10).

Mushtaq, M. S., Mushtaq, M. Y., Iqbal, M. W., & Hussain, S. A. (2022). Security, integrity, and privacy of cloud computing and big data. In *Security and privacy trends in cloud computing and big data* (pp. 19-51). CRC Press.

Nedelkoski, S., Cardoso, J., & Kao, O. (2019, July). Anomaly detection from system tracing data using multimodal deep learning. In *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)* (pp. 179-186). IEEE.

Nedelkoski, S., Cardoso, J., & Kao, O. (2019, July). Anomaly detection from system tracing data using multimodal deep learning. In *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)* (pp. 179-186). IEEE.

Noor, T. H., Sheng, Q. Z., Zeadally, S., & Yu, J. (2013). Trust management of services in cloud environments: Obstacles and solutions. *ACM Computing Surveys (CSUR)*, *46*(1), 1-30.

Odetunde, A., Adekunle, B. I., & Ogeawuchi, J. C. (2022). Using Predictive Analytics and Automation Tools for Real-Time Regulatory Reporting and Compliance Monitoring. *Int. J. Multidiscip. Res. Growth Eval*, *3*(2), 650-661.

Olawale, A., Ajoke, O., & Adeusi, C. (2020). Quality assessment and monitoring of networks using passive.

Oliveira, W., Oliveira, D. D., & Braganholo, V. (2018). Provenance analytics for workflow-based computational experiments: A survey. *ACM Computing Surveys (CSUR)*, *51*(3), 1-25.

Pandey, A. K., Khan, A. I., Abushark, Y. B., Alam, M. M., Agrawal, A., Kumar, R., & Khan, R. A. (2020). Key issues in healthcare data integrity: Analysis and recommendations. *IEEE Access*, *8*, 40612-40628.

Pandey, R., & Pande, M. (2021). Provenance data models and assertions: a demonstrative approach. *Semantic IoT: Theory and Applications: Interoperability, Provenance and Beyond*, 103-129.

Pasquier, T., Singh, J., Powles, J., Eyers, D., Seltzer, M., & Bacon, J. (2018). Data provenance to audit compliance with privacy policy in the Internet of Things. *Personal and Ubiquitous Computing*, *22*(2), 333-344.

Pinto, A., Cardinale, Y., Dongo, I., & Ticona-Herrera, R. (2022). An ontology for modeling cultural heritage knowledge in urban tourism. *IEEE Access*, *10*, 61820-61842.

Porkodi, S., & Kesavaraja, D. (2021). Secure data provenance in Internet of Things using

hybrid attribute based crypt technique. *Wireless Personal Communications*, *118*(4), 2821-2842.

Quadri, S. (2017). *Cloud computing: migrating to the cloud, Amazon Web Services and Google Cloud Platform* (Master's thesis, S. Quadri).

Rahman, M. S., Chamikara, M. A. P., Khalil, I., & Bouras, A. (2022). Blockchain-of-blockchains: An interoperable blockchain platform for ensuring IoT data integrity in smart city. *Journal of Industrial Information Integration*, *30*, 100408.

Ramachandran, A., & Kantarcioglu, D. M. (2017). Using blockchain and smart contracts for secure data provenance management. *arXiv preprint arXiv:1709.10000*.

Ruan, P., Dinh, T. T. A., Lin, Q., Zhang, M., Chen, G., & Ooi, B. C. (2021). LineageChain: a fine-grained, secure and efficient data provenance system for blockchains. *The VLDB Journal*, *30*, 3-24.

Rupprecht, L., Davis, J. C., Arnold, C., Gur, Y., & Bhagwat, D. (2020). Improving reproducibility of data science pipelines through transparent provenance capture. *Proceedings of the VLDB Endowment*, *13*(12), 3354-3368.

Saboor, A., Mahmood, A. K., Omar, A. H., Hassan, M. F., Shah, S. N. M., & Ahmadian, A. (2022). Enabling rank-based distribution of microservices among containers for green cloud computing environment. *Peer-to-Peer Networking and Applications*, *15*(1), 77-91.

Sagar Hossen, M., Tabassum, T., Ashiqul Islam, M., Karim, R., Rumi, L. S., & Kobita, A. A. (2020). Digital signature authentication using asymmetric key cryptography with different byte number. In *Evolutionary Computing and Mobile Sustainable Networks: Proceedings of ICECMSN 2020* (pp. 845-851). Singapore: Springer Singapore.

Sarathy, V., Narayan, P., & Mikkilineni, R. (2010, June). Next generation cloud computing architecture: Enabling real-time dynamism for shared distributed physical infrastructure. In *2010 19th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises* (pp. 48-53). IEEE.

Saxena, N., Hayes, E., Bertino, E., Ojo, P., Choo, K. K. R., & Burnap, P. (2020). Impact and key challenges of insider threats on organizations and critical businesses. *Electronics*, *9*(9), 1460.

Schneider, M., Masti, R. J., Shinde, S., Capkun, S., & Perez, R. (2022). Sok: Hardware-supported trusted execution environments. *arXiv preprint arXiv:2205.12742*.

Sharma, K., Shingatgeri, V. M., & Pal, S. (2021). Role of data digitization on data integrity. *Quality assurance implementation in research labs*, 221-245.

Shekhtman, L., & Waisbard, E. (2021). Engravechain: A blockchain-based tamper-proof distributed log system. *Future Internet*, *13*(6), 143.

Shinde, Y. A. A (2022). Comprehensive Survey on Enhancing Blockchain Data Security through the Integration of IoT and AI. *Journal Of Technical Education*, 167.

Siddiqui, M. S., Syed, T. A., Nadeem, A., Nawaz, W., & Albouq, S. S. (2020). BlockTrack-L: A lightweight blockchain-based provenance message tracking in IoT. *International Journal of Advanced Computer Science and Applications*, *11*(4).

Silowash, G. J., Cappelli, D. M., Moore, A. P., Trzeciak, R. F., Shimeall, T., & Flynn, L. (2012). Common sense guide to mitigating insider threats.

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance techniques. *Computer Science Department, Indiana University, Bloomington IN*, *47405*, 69.

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance techniques. *Computer Science Department, Indiana University, Bloomington IN*, *47405*, 69.

Singh, J., Cobbe, J., & Norval, C. (2018). Decision provenance: Harnessing data flow for accountable systems. *IEEE Access*, *7*, 6562-6574.

Singh, J., Cobbe, J., & Norval, C. (2018). Decision provenance: Harnessing data flow for accountable systems. *IEEE Access*, *7*, 6562-6574.

Soldatos, J., Despotopoulou, A., Kefalakis, N., & Ipektsidis, B. (2021). Blockchain based data provenance for trusted artificial intelligence. *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI*

Suen, C. H., Ko, R. K., Tan, Y. S., Jagadpramana, P., & Lee, B. S. (2013, July). S2logger: End-to-end data tracking mechanism for cloud data provenance. In *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 594-602). IEEE.

Sun, Y., Zhang, J., Xiong, Y., & Zhu, G. (2014). Data security and privacy in cloud computing. *International Journal of Distributed Sensor Networks*, *10*(7), 190903.

Sunyaev, A., & Sunyaev, A. (2020). Cloud computing. *Internet computing: Principles of distributed systems and emerging internet-based technologies*, 195-236.

Tabrizchi, H., & Kuchaki Rafsanjani, M. (2020). A survey on security challenges in cloud computing: issues, threats, and solutions. *The journal of supercomputing*, *76*(12), 9493-9532.

Tan, C. B., Hijazi, M. H. A., Lim, Y., & Gani, A. (2018). A survey on proof of retrievability for cloud data integrity and availability: Cloud storage state-of-the-art, issues, solutions and future trends. *Journal of Network and Computer Applications*, *110*, 75-86.

Tandel, T. (2022). *A study of modern cluster-based high availability database solutions* (Master's thesis, OsloMet-storbyuniversitetet).

Tarafder, M. T. R., Mohiuddin, A. B., Ahmed, N., Shihab, M. A., & Kabir, M. F. (2022). Block chain-Based Solutions for Improved Cloud Data Integrity and Security. *BULLET: Journal Multidisiplin Ilmu*, *1*(04), 736-748.

Thokala, V. S. (2021). A Comparative Study of Data Integrity and Redundancy in Distributed Databases for Web Applications. *Int. J. Res. Anal. Rev*, *8*(4), 383-389.

Torre, D., Alferez, M., Soltana, G., Sabetzadeh, M., & Briand, L. (2021). Modeling data protection and privacy: application and experience with GDPR. *Software and Systems Modeling*, *20*, 2071-2087.

Trace, C. B. (2020). Maintaining records in context? Disrupting the theory and practice of archival classification and arrangement. *the american archivist*, *83*(2), 322-372.

Upadhyay, D., Gaikwad, N., Zaman, M., & Sampalli, S. (2022). Investigating the avalanche effect of various cryptographically secure hash functions and hash-based applications. *IEEE Access*, *10*, 112472-112486.

Vagadia, B. (2020). Data integrity, control and tokenization. In *Digital Disruption: Implications and opportunities for Economies, Society, Policy Makers and Business Leaders* (pp. 107-176). Cham: Springer International Publishing.

Wall Street Mojo. (2020). *Ensuring data integrity in distributed cloud environments* [Diagram]. Retrieved from https://www.wallstreetmojo.com/wp-content/uploads/2020/05/Data-Integrity-1-768x337.png

Wang, J., Crawl, D., Purawat, S., Nguyen, M., & Altintas, I. (2015, October). Big data provenance: Challenges, state of the art and opportunities. In *2015 IEEE international conference on big data (Big Data)* (pp. 2509-2516). IEEE.

Wang, X., Duan, S., Clavin, J., & Zhang, H. (2022). Bft in blockchains: From protocols to use cases. *ACM Computing Surveys (CSUR)*, *54*(10s), 1-37.

Wang, Y., Su, Z., Ni, J., Zhang, N., & Shen, X. (2021). Blockchain-empowered space-air-ground integrated networks: Opportunities, challenges, and solutions. *IEEE Communications Surveys & Tutorials*, *24*(1), 160-209.

Westerlund, M., Neovius, M., & Pulkkis, G. (2018). Providing tamper-resistant audit trails with distributed ledger based solutions for forensics of IoT systems using cloud resources. *International Journal on Advances in Security*, *11*(3).

Whyte, S. T. (2021). *Reliable data collection: A tool for data integrity in Nigeria* (Doctoral dissertation, Walden University).

Wittner, R., Mascia, C., Gallo, M., Frexia, F., Müller, H., Plass, M., ... & Holub, P. (2022). Lightweight distributed provenance model for complex real–world environments. *Scientific Data*, *9*(1), 503.

Wittner, R., Mascia, C., Gallo, M., Frexia, F., Müller, H., Plass, M., ... & Holub, P. (2022). Lightweight distributed provenance model for complex real–world environments. *Scientific Data*, *9*(1), 503.

Wood, A., Najarian, K., & Kahrobaei, D. (2020). Homomorphic encryption for machine learning in medicine and bioinformatics. *ACM Computing Surveys (CSUR)*, *53*(4), 1-35.

Xu, X., Lu, Q., Liu, Y., Zhu, L., Yao, H., & Vasilakos, A. V. (2019). Designing blockchain-based applications a case study for imported product traceability. *Future Generation Computer Systems*, *92*, 399-406.

Yang, R., & Xu, J. (2016, March). Computing at massive scale: Scalability and dependability challenges. In *2016 IEEE symposium on service-oriented system engineering (SOSE)* (pp. 386-397). IEEE.

Yaqoob, I., Salah, K., Jayaraman, R., & Al-Hammadi, Y. (2022). Blockchain for healthcare data management: opportunities, challenges, and future recommendations. *Neural Computing and Applications*, 1-16.

Yaqoob, I., Salah, K., Jayaraman, R., & Al-Hammadi, Y. (2022). Blockchain for healthcare data management: opportunities, challenges, and future recommendations. *Neural Computing and Applications*, *34*(14), 11475-11490.

Yaqoob, I., Salah, K., Jayaraman, R., & Al-Hammadi, Y. (2022). Blockchain for healthcare data management: opportunities, challenges, and future recommendations. *Neural Computing and Applications*, *34*(14), 11475-11490.

Zafar, F., Khan, A., Malik, S. U. R., Ahmed, M., Anjum, A., Khan, M. I., ... & Jamil, F. (2017). A survey of cloud computing data integrity schemes: Design challenges, taxonomy and future trends. *Computers & Security*, *65*, 29-49.

Zafar, F., Khan, A., Malik, S. U. R., Ahmed, M., Anjum, A., Khan, M. I., ... & Jamil, F. (2017). A survey of cloud computing data integrity schemes: Design challenges, taxonomy and future trends. *Computers & Security*, *65*, 29-49.

Zarour, M., Alenezi, M., Ansari, M. T. J., Pandey, A. K., Ahmad, M., Agrawal, A., ... & Khan, R. A. (2021). Ensuring data integrity of healthcare information in the era of digital health. *Healthcare technology letters*, *8*(3), 66-77.

Zeiselmair, A., Steinkopf, B., Gallersdörfer, U., Bogensperger, A., & Matthes, F. (2021). Analysis and application of

verifiable computation techniques in blockchain systems for the energy sector. *Frontiers in Blockchain*, *4*, 725322.

Zhang, M., Jiang, L., Zhao, J., Yue, P., & Zhang, X. (2020). Coupling OGC WPS and W3C PROV for provenance-aware geoprocessing workflows. *Computers & Geosciences*, *138*, 104419.

Zhang, O. Q., Kirchberg, M., Ko, R. K., & Lee, B. S. (2011, November). How to track your data: The case for cloud computing provenance. In *2011 IEEE Third International Conference on Cloud Computing Technology and Science* (pp. 446-453). IEEE.

Zhang, Q., Xin, C., & Wu, H. (2021). Privacy-preserving deep learning based on multiparty secure computation: A survey. *IEEE Internet of Things Journal*, *8*(13), 10412-10429.

Zhang, Y. (2020). Mitigating Insider Threats in Enterprise Storage Systems: A Security Framework for Data Integrity and Access Control. *International Journal of Trend in Scientific Research and Development*, *4*(4), 1878-1890.

Zhao, C., Zhao, S., Zhao, M., Chen, Z., Gao, C. Z., Li, H., & Tan, Y. A. (2019). Secure multi-party computation: theory, practice and applications. *Information Sciences*, *476*, 357-372.

Zhu, P., Hu, J., Li, X., & Zhu, Q. (2021). Using blockchain technology to enhance the traceability of original achievements. *IEEE Transactions on Engineering Management*, *70*(5), 1693-1707.

Zipperle, M., Gottwalt, F., Chang, E., & Dillon, T. (2022). Provenance-based intrusion detection systems: A survey. *ACM Computing Surveys*, *55*(7), 1-36.

Zou, J., He, D., Zeadally, S., Kumar, N., Wang, H., & Choo, K. R. (2021). Integrated blockchain and cloud computing systems: A systematic survey, solutions, and challenges. *ACM Computing Surveys (CSUR)*, *54*(8), 1-36.

Conflict of interest

All declare there is no conflict of interest

Ethical Consideration

Not applicable

Availability of Data

Data shall be made available upon request.

Funding

None

Authors Contribution

The entire work was carried out by the author.