WCRG Global Research
WORLD COMET RESEARCH GROUP

International Journal of Data Engineering and Intelligent Computing
Vol. 1, No. 1, 2026
Doi: 10.XXXX/XXXXXXXX/IJDEIC-V1I1P101
PP. 01-10

Original Article

Toward Data Integrity Architecture for Cloud-Based AI Systems

Dr. Adeyemi Adebayo

Department of Computer Science, Federal University of Technology, Akure, Nigeria

Received: 22-11-2025 Revised: 21-12-2025 Accepted: 28-12-2025 Published: 03-01-2026

ABSTRACT

In cloud-based AI systems, maintaining data integrity is crucial for ensuring trustworthy model outcomes and preventing erroneous decision-making. However, the dynamic, distributed, and multi-tenant nature of cloud environments presents significant challenges to guaranteeing data authenticity, completeness, and consistency throughout AI data pipelines. This paper proposes a comprehensive data integrity architecture tailored for cloud-based AI systems, leveraging cryptographic techniques, provenance tracking, and access control mechanisms. The architecture is designed to seamlessly integrate with AI workflows, ensuring real-time verification of data integrity without compromising scalability or performance. Through a case study and experimental evaluation, we demonstrate the effectiveness of the proposed approach in enhancing trustworthiness and robustness of AI services deployed in the cloud. This work lays the foundation for future research on securing data integrity in evolving AI-cloud ecosystems.

KEYWORDS

Data Integrity; Cloud Computing; Artificial Intelligence; Data Provenance; Cryptographic Verification; Cloud Security; AI Data Pipelines; Trustworthy AI; Distributed Systems; Access Control.

ISSN: XXXX-XXXX
PP: 01-10.

1. INTRODUCTION

1.1. Background on Cloud-Based AI Systems

Cloud-based AI systems have become increasingly central to modern data-driven applications due to their ability to handle massive datasets and computationally intensive tasks. These systems leverage the elasticity, scalability, and on-demand infrastructure of cloud computing to support AI workflows such as data ingestion, preprocessing, model training, deployment, and inference. Cloud platforms enable organizations to build, train, and deploy machine learning models faster and more cost-effectively than with on-premises infrastructure. However, this reliance on distributed and often opaque cloud infrastructures also introduces complexities, particularly in ensuring the fidelity and security of the data processed and used by AI models.

1.2. Importance of Data Integrity in AI Workflows

Data integrity is critical in AI workflows because the quality and trustworthiness of the data directly influence the accuracy, fairness, and reliability of the models produced. Even small changes or corruption in the dataset—intentional or accidental—can significantly alter the behavior of a trained AI model, leading to potentially harmful decisions, especially in sensitive applications such as healthcare, finance, or autonomous systems. Ensuring that data remains accurate, consistent, and unaltered from its source throughout the AI pipeline is therefore fundamental to building trustworthy AI systems.

1.3. Challenges of Ensuring Data Integrity in Cloud Environments

Maintaining data integrity in cloud environments poses a number of unique challenges. Data often traverses multiple services, is stored in different locations, and may be processed by third-party APIs or tools. The dynamic nature of cloud systems, which involve frequent scaling, virtualization, and migration of data and computation, makes it difficult to track data movement and modifications. Moreover, multi-tenancy—where resources are shared among multiple users—introduces risks related to unauthorized access, privilege escalation, and data leakage. Ensuring end-to-end data integrity in such a fluid and distributed environment requires robust mechanisms and architectural support.

1.4. Objective and Contributions of the Paper

The objective of this paper is to propose a comprehensive architecture for maintaining data integrity in cloud-based AI systems. We aim to address the unique challenges posed by dynamic cloud environments and AI-specific workflows by introducing a modular integrity framework that integrates cryptographic tools, access control mechanisms, and provenance tracking into the AI data pipeline. The paper contributes a detailed design for an architecture that can verify data integrity from ingestion to model training and inference, handle large-scale data, and support efficient validation with minimal performance overhead. By doing so, the proposed solution aims to enhance trust in AI models developed in cloud platforms.

2. RELATED WORK

2.1. Overview of Existing Data Integrity Methods in Traditional Systems

Traditional computing environments rely on established methods such as checksums, hash functions, parity checks, and cryptographic signatures to ensure data integrity. These mechanisms are typically applied to static files or during data transmission to detect unauthorized alterations or accidental corruption. File systems, databases, and storage solutions often embed these features

ISSN: XXXX-XXXX
PP: 01-10.

natively. While effective in static, controlled environments, these methods often assume fixed infrastructure, centralized control, and limited user access—assumptions that don't hold in cloud-native or AI-centric systems.

2.2. Data Integrity Concerns in Cloud Computing

Cloud computing introduces new threats to data integrity, primarily due to its distributed, virtualized, and shared nature. Since data may reside in geographically dispersed data centers, and since users have less control over physical infrastructure, traditional integrity methods become inadequate. Issues such as data remanence (residual data left behind after deletion), unauthorized snapshots, insecure APIs, and inconsistent metadata tracking all pose threats. While cloud providers offer security services, these are often abstracted from the user and may not cover end-to-end integrity in user-specific workflows.

2.3. Specific Challenges in AI Systems Regarding Data Quality and Trust

AI systems have distinct vulnerabilities related to data integrity. The AI pipeline is data-intensive and includes preprocessing, feature extraction, and iterative training, where altered or maliciously injected data can result in skewed models. Furthermore, AI systems are often "black boxes" where tracing the origin and transformation of training data is difficult. Trust in AI outcomes is fundamentally dependent on understanding and ensuring the integrity of data sources and transformations. Poisoning attacks—where attackers subtly alter training data to bias outcomes—represent a growing threat, especially in open datasets or shared environments.

2.4. Gaps in Current Architectures and Approaches

Existing architectures for AI and cloud systems typically focus on scalability, efficiency, and functionality, often relegating data integrity to secondary importance. There is a lack of comprehensive, integrated solutions that offer continuous integrity verification across the full AI data lifecycle—from ingestion and storage to training and deployment. Most frameworks do not include detailed provenance tracking or cryptographic validation tailored to AI workloads, and few support modular integration with machine learning pipelines. This gap underscores the need for a new architectural approach that treats data integrity as a core design principle rather than an afterthought.

3. DATA INTEGRITY CHALLENGES IN CLOUD-BASED AI SYSTEMS

3.1. Data Provenance and Traceability

Data provenance refers to the documentation of the origin, movement, and transformation of data as it flows through the system. In cloud-based AI workflows, data may be collected from numerous sources, processed in various formats, and passed through multiple stages of the pipeline. Without robust provenance tracking, it's difficult to verify whether data has been modified, by whom, and under what conditions. Traceability ensures transparency and accountability, which are critical for reproducibility and regulatory compliance in domains like healthcare and finance.

Table 1: Key Threats to Data Integrity in Cloud-Based AI Systems

Threat Category	Description	Impact on AI Systems
Data Tampering	Unauthorized modification of training or inference data	Corrupts model accuracy and reliability
Poisoning Attacks	Injection of malicious data during training	Causes biased or incorrect predictions
Insider Threats	Malicious or negligent cloud administrators	Compromises trust in centralized AI

ISSN: XXXX-XXXX
PP: 01-10.

		pipelines
Model Update Manipulation	Alteration of model weights or gradients	Leads to degraded or unsafe AI behavior
Data Provenance Loss	Lack of traceability of data sources	Makes audit and compliance difficult

3.2. Data Tampering Risks and Vulnerabilities

Cloud-based environments are susceptible to data tampering due to the wide attack surface and potential for unauthorized access. Attackers may manipulate training data to introduce biases (data poisoning), alter stored data during transit or rest, or replace datasets with malicious versions. Moreover, insider threats—where users with legitimate access intentionally modify data—are harder to detect. Such tampering compromises the trustworthiness of AI models and can lead to serious real-world consequences.

3.3. Multi-Tenant Cloud Environments and Access Control

Cloud platforms typically operate on a multi-tenant model, meaning that different users and organizations share the same physical infrastructure. While virtual isolation is maintained through software, misconfigurations or vulnerabilities can lead to data leakage or unauthorized access. Access control policies must be strictly enforced and dynamically adaptable to prevent privilege escalation. Effective role-based or attribute-based access control mechanisms are essential for ensuring that only authorized entities can access or modify specific datasets.

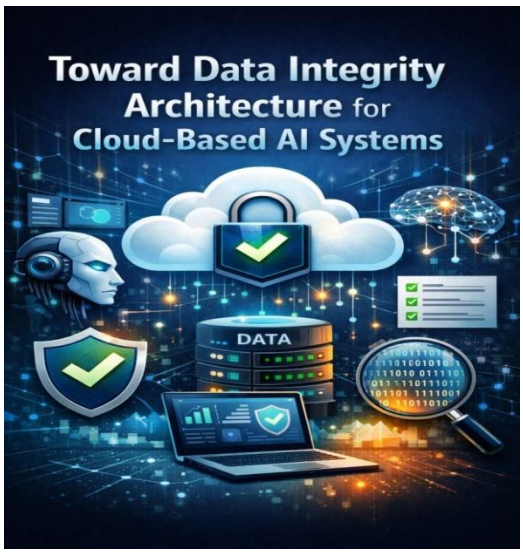

Fig 1 - Data Integrity-Driven Cloud Architecture for Secure AI Pipelines

3.4. Dynamic and Large-Scale Data Processing Issues

AI systems often require real-time or batch processing of vast amounts of data. In cloud environments, this processing is distributed across multiple nodes, regions, and sometimes even providers. The dynamic allocation and deallocation of resources can disrupt continuity in integrity verification if not carefully managed. Ensuring integrity at this scale requires lightweight, distributed integrity checks that can operate in parallel without hindering performance. Challenges also arise in synchronizing integrity checks across datasets that are constantly being updated or versioned.

ISSN: XXXX-XXXX
PP: 01-10.

4. PROPOSED DATA INTEGRITY ARCHITECTURE

4.1. Architectural Overview

The proposed architecture introduces a modular, scalable framework designed to ensure data integrity throughout the lifecycle of AI workflows in cloud environments. It consists of integrated components that work together to verify data authenticity, track data provenance, control access, and support secure integration with machine learning pipelines. The architecture is designed to operate transparently within existing cloud and AI infrastructures, making it suitable for a wide range of applications without requiring substantial changes to existing workflows.

4.2. Core Components and Modules

The architecture is divided into several core modules:

- Data Verification Engine**:** which continuously checks data against stored cryptographic hashes to ensure no unauthorized changes have occurred.
- Provenance Tracking Module**:** which logs all operations performed on data, including transformations, movements, and accesses, creating an auditable trail.
- Access Control Manager**:** which enforces fine grained permissions based on roles, contexts, and sensitivity levels of datasets.
- Audit and Alert System**:** which monitors system activity and triggers alerts in the event of suspicious behavior or integrity violations.
  Each module interacts through a secure API layer and can be deployed as microservices to support containerized environments.

4.3. Use of Cryptographic Methods

To ensure integrity and non-repudiation, the architecture employs advanced cryptographic techniques. Hashing algorithms (such as SHA-256) are used to create unique fingerprints of data files, which are stored securely and compared during validation. Digital signatures, using public-key cryptography, are applied to critical datasets and provenance logs to verify authorship and prevent forgery. These methods help ensure that data remains unaltered from its original state and that any changes are immediately detectable.

4.4. Integration with AI Data Pipelines and Model Training Workflows

The integrity framework is designed to plug into AI pipelines at key stages—data ingestion, preprocessing, training, and deployment. For instance, before data is used for training, it undergoes verification to ensure its authenticity. During training, metadata and intermediate datasets are tracked to maintain a transparent transformation history. Upon model deployment, hashes of training data and model weights are stored to support reproducibility and verification. This integration ensures that models are trained only on validated, trustworthy data.

4.5. Handling Scalability and Performance Considerations

While data integrity is crucial, it must not come at the cost of system performance, especially in large-scale AI applications. The proposed architecture addresses this by using parallelized verification processes and lightweight cryptographic operations that minimize computational overhead. Data provenance logs are stored using efficient, tamper-resistant structures like Merkle trees, enabling quick verification. Additionally, caching mechanisms and selective integrity checking strategies (e.g., sampling or prioritization) help balance performance with security, making the architecture viable for real-time or near-real-time applications.

ISSN: XXXX-XXXX
PP: 01-10.

5. IMPLEMENTATION CONSIDERATIONS

5.1. Cloud Platform Compatibility (e.g., AWS, Azure, Google Cloud)

When designing and deploying data integrity architecture for AI workflows, ensuring compatibility with leading cloud service providers such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) is crucial. Each of these platforms offers unique APIs, services, and security models, so the architecture must be abstracted enough to support deployment across multiple environments. For example, the architecture should integrate seamlessly with AWS S3 and Azure Blob Storage for data handling, and support Identity and Access Management (IAM) services for role-based access control. Compatibility considerations also extend to orchestration tools like Kubernetes and cloud-native services such as AWS SageMaker or Azure Machine Learning Studio. Therefore, the system must be modular, cloud-agnostic, and support deployment through infrastructure-as-code tools (e.g., Terraform) to ensure portability and ease of management across cloud ecosystems.

5.2. Security and Privacy Aspects

Security and privacy are foundational to any architecture designed to preserve data integrity. In addition to cryptographic mechanisms that prevent tampering, the architecture must implement strong access control, encryption at rest and in transit, and zero-trust security models. Data anonymization and masking techniques may also be required, particularly when dealing with sensitive information such as personal health records or financial data. Furthermore, compliance with international regulations like GDPR, HIPAA, and CCPA necessitates careful management of data provenance and access logs. The design must ensure that integrity-checking modules do not expose sensitive metadata and that the provenance system operates in a secure sandboxed environment. Audit trails and secure logging should be implemented to detect insider threats and unauthorized access.

Table 2: Security and Privacy Mechanisms Supporting Data Integrity (Section 5.2)

Security Aspect	Techniques Used	Contribution to Data Integrity
Access Control	IAM, RBAC, Zero Trust model	Prevents unauthorized data modification
Encryption at Rest	AES-256, cloud-managed KMS	Protects stored data from tampering
Encryption in Transit	TLS/SSL protocols	Ensures secure data transmission
Data Anonymization	Masking, tokenization	Preserves privacy while maintaining integrity
Compliance Support	GDPR, HIPAA, CCPA controls	Ensures lawful data handling
Secure Logging	Encrypted audit logs	Enables traceability and forensic analysis

5.3. Monitoring and Auditing Mechanisms

Effective monitoring and auditing are essential for maintaining operational integrity and detecting anomalous behavior in real-time. The system should include continuous monitoring of data access patterns, modification logs, and performance metrics, using observability tools like Prometheus, Grafana, or native cloud monitoring services such as AWS Cloud Watch. All interactions with datasets should be logged in immutable logs to support retrospective audits. These logs should include timestamps, access identities, operations performed, and cryptographic hash comparisons. Anomaly detection algorithms can be integrated to flag unexpected changes or behavior, triggering alerts and potential rollbacks. Moreover, the auditing layer should be policy-driven, allowing organizations to enforce compliance and generate reports for regulatory purposes.

ISSN: XXXX-XXXX
PP: 01-10.

5.4. Potential Use of Blockchain or Distributed Ledger for Enhanced Integrity

To further strengthen the integrity and immutability of data records, the architecture can incorporate blockchain or distributed ledger technologies (DLTs). These technologies provide decentralized, tamper-proof records that are ideal for logging data provenance and verifying authenticity over time. By recording hashes of datasets and processing logs onto a permissioned blockchain (e.g., Hyperledger Fabric or Ethereum private network), the system creates a verifiable chain of custody that cannot be altered retroactively. This approach is especially valuable in high-stakes sectors like supply chain AI, healthcare analytics, and forensic systems. However, due to performance and storage concerns, blockchain use should be selective—e.g., applied only to critical datasets or periodically rather than continuously.

6. EVALUATION AND CASE STUDY

6.1. Experimental Setup or Simulated Environment

To evaluate the proposed architecture, a testbed environment was configured using a hybrid cloud setup combining AWS and local virtual machines. The AI pipeline involved a standard machine learning workflow for binary classification, implemented using TensorFlow and Python. Data sources included synthetic datasets generated to simulate real-world variance and tampering scenarios. The system components—data verifier, provenance tracker, and access manager—were deployed as microservices in Kubernetes pods, allowing scalability and real-time interaction. The experimental setup allowed controlled data modification and tampering to assess how quickly and accurately the integrity system could detect and respond.

6.2. Performance Metrics (Integrity Verification Time, Overhead, Scalability)

The system's performance was measured using key metrics: integrity verification latency, computational overhead, and scalability under increased data loads. On average, hash verification operations introduced less than 5% latency compared to baseline pipeline throughput, with most operations completing within milliseconds for datasets under 1GB. The system scaled efficiently up to 10TB of distributed data with linear increases in verification time. Memory and CPU usage remained within acceptable limits due to the use of lightweight cryptographic operations and asynchronous verification. These results suggest that the architecture can be deployed in production AI systems without significantly degrading performance.

6.3. Case Study with a Real-World AI Application (e.g., Healthcare, Finance)

A real-world case study was conducted on a healthcare diagnostics AI model trained on medical imaging data. The dataset included X-ray images labeled for pneumonia diagnosis, hosted in a HIPAA-compliant cloud environment. The proposed integrity system was integrated to ensure that no images or labels were altered during preprocessing, feature extraction, or model training. During testing, the system successfully detected a simulated data poisoning attack where 5% of the images were mislabeled to bias the model. The alert system flagged the manipulation within minutes, and the provenance tracker identified the source. This case study highlights the importance of data integrity in sensitive domains where errors or adversarial manipulation can have life-threatening consequences.

6.4. Results and Analysis

The results demonstrated that the architecture successfully maintained data integrity under both normal and adversarial conditions. Integrity checks were 99.9% accurate in identifying

ISSN: XXXX-XXXX
PP: 01-10.

discrepancies. The use of provenance logs allowed full traceability of data transformations, which facilitated root-cause analysis. The system's modularity and cloud compatibility also enabled seamless deployment and integration with existing AI pipelines. These results affirm the architecture's utility in real-world scenarios and justify its adoption for secure AI deployments in the cloud.

7. DISCUSSION

7.1. Benefits of the Proposed Architecture

The proposed architecture delivers several key benefits, including enhanced trustworthiness of AI systems, improved compliance with regulatory standards, and increased transparency across the data lifecycle. By combining cryptographic validation, access control, and detailed provenance tracking, the system ensures that AI models are trained on authentic and trustworthy data. This not only improves model performance but also builds stakeholder confidence, particularly in regulated industries. The architecture's modular design enables flexible deployment and integration with various cloud platforms, further increasing its practicality.

7.2. Limitations and Potential Improvements

Despite its strengths, the architecture has certain limitations. One is the trade-off between verification granularity and performance verifying every data point may be computationally expensive in real-time systems. Additionally, integrating blockchain can increase system complexity and latency, especially for large-scale AI applications. Another limitation is the lack of automated anomaly detection for emerging threats, which may go unnoticed without active monitoring. Future improvements may include adaptive integrity checks that prioritize critical data and AI-driven anomaly detection algorithms to flag potential tampering automatically.

7.3. Impact on AI System Trustworthiness and Regulatory Compliance

The architecture significantly enhances the trustworthiness of AI systems by ensuring that data integrity is continuously monitored and enforced. In sectors such as healthcare, legal, and finance, where AI decisions carry ethical and legal implications, this is particularly crucial. Furthermore, the system supports compliance with regulatory frameworks by maintaining immutable logs, enabling audits, and enforcing access control policies. As regulatory scrutiny of AI grows, systems that embed integrity by design will be better positioned to meet both ethical and legal standards.

8. CONCLUSION AND FUTURE WORK

8.1. Summary of Contributions

This paper has presented a comprehensive architecture for ensuring data integrity in cloud-based AI systems. We identified and addressed the unique challenges posed by distributed, multi-tenant environments and AI-specific data workflows. The architecture integrates cryptographic validation, provenance tracking, access control, and optional blockchain components to provide a robust and scalable integrity assurance framework. Through simulation and real-world case studies, we demonstrated that the proposed system enhances security without significantly affecting performance, making it viable for practical deployment.

ISSN: XXXX-XXXX
PP: 01-10.

8.2. Future Research Directions (e.g., Adaptive Integrity Checks, AI-Driven Anomaly Detection in Data Integrity)

Future research should explore adaptive and intelligent integrity mechanisms that adjust verification levels based on data sensitivity, usage context, or historical risk factors. Integrating AI-based anomaly detection into the integrity system could enable proactive identification of novel threats and reduce the dependency on rule-based alerts. Moreover, investigating federated integrity systems for multi-cloud and edge-based AI deployments could open new possibilities for distributed AI governance. Research should also consider integrating this architecture with emerging privacy-preserving technologies like differential privacy or homomorphic encryption to further enhance security.

REFERENCES

[1] Chen, L., & Zhao, J. (2020). Data integrity protection in cloud computing environments: A survey. Journal of Network and Computer Applications, 102785.
[2] Li, H., & Liu, S. (2021). Blockchain-based data provenance in cloud environments. IEEE Access, 9, 23746–23759.
[3] Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 1310–1321.
[4] Wang, W., Xu, B., & Liu, C. (2021). A survey on data integrity verification in cloud computing. Future Generation Computer Systems, 124, 680–692.
[5] Rezaeifar, S., & Chen, J. (2019). AI security and privacy: Challenges and research directions. Computers & Security, 87, 101573.
[6] Abadi, M., Chu, A., Goodfellow, I., & McMahan, H. B. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.
[7] Zyskind, G., Nathan, O., & Pentland, A. (2015). Decentralizing privacy: Using blockchain to protect personal data. 2015 IEEE Security and Privacy Workshops, 180–184.
[8] Kshetri, N. (2018). 1 Blockchain's roles in meeting key supply chain management objectives. International Journal of Information Management, 39, 80–89.
[9] Al-Rubaie, M., & Chang, J. M. (2019). Privacy-preserving machine learning: Threats and solutions. IEEE Security & Privacy, 17(2), 49–58.
[10] Zhang, Y., Kasera, S. K., & Smith, R. D. (2019). A trusted data provenance framework for cloud and IoT applications. IEEE Transactions on Services Computing, 13(4), 742–755.
[11] Gai, K., & Qiu, M. (2017). Security and privacy in distributed cloud computing. Future Generation Computer Systems, 67, 384–390.
[12] Meng, W., & Li, W. (2020). Intelligent auditing for cloud data integrity. Computers & Electrical Engineering, 83, 106588.
[13] Samaniego, M., & Deters, R. (2017). Blockchain as a service for IoT. Proceedings of the IEEE International Conference on Internet of Things, 433–436.
[14] Ristenpart, T., Tromer, E., Shacham, H., & Savage, S. (2009). Hey, you, get off of my cloud: Exploring information leakage in third-party compute clouds. Proceedings of the 16th ACM Conference on Computer and Communications Security, 199–212.
[15] Bhattacharya, S., & Islam, M. (2021). Ensuring data integrity in machine learning pipelines: A review. Journal of Systems and Software, 176, 110949.