Journal of Artificial Intelligence Practice (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2025.080403
ISSN 2371-8412 Vol. 8 Num. 4

Cloud Computing Environment: Research on Big Data Security and Privacy Protection Strategies

Shenpeiyilin

Amazon Web Servicees Inc, Arlington County, Virginia, USA

Keywords: Cloud Computing, Big Data, Data Security, Privacy Protection, Strategy

Abstract: Against the backdrop of rapid cloud computing development, big data has become an important resource across research, education, government and enterprise sectors, characterized by large scale, diverse types and high sensitivity. However, the openness and sharing features of cloud environments also bring severe challenges to data security and privacy protection. This paper first analyzes cloud computing architectures and the characteristics of big data, and describes the main security risks that arise throughout the data lifecycle (collection, transmission, storage and use), while summarizing common threat types and privacy leakage pathways. On this basis, it discusses key technical measures such as encryption, access control, differential privacy and federated learning, and proposes a protection strategy that integrates a multi-layered security defense with a compliance-oriented governance framework. Case studies are used to validate the feasibility and practical effectiveness of the proposed strategy in preventing data breaches and improving privacy protection. The results show that building a systematic, scalable security and privacy protection system not only effectively ensures the security and trustworthiness of big data in cloud environments, but also provides strong support for future intelligent and compliant data applications.

1. Introduction

With rapid advances in information technology, cloud computing has been widely adopted across research, education, government and industry, and its distributed storage and elastic computing capabilities have substantially unlocked the value of big data. At the same time, cloud platforms' multi-tenant architecture, resource sharing and cross-regional deployments amplify data security and privacy challenges. Big data often contains highly sensitive information—personal records, research outcomes, business secrets and even strategic assets—so leaks, tampering or misuse can cause serious individual, economic and social harm. Although prior work has proposed many technical solutions (encryption, access control, differential privacy, federated learning, anonymization, etc.), research remains fragmented and rarely addresses protection across the full data lifecycle. The convergence of AI, IoT and blockchain further complicates threat surfaces and use cases. This paper therefore reviews cloud architectures and big-data characteristics, analyzes lifecycle-level risks, evaluates key technical measures and governance practices, and proposes a combined technical–compliance strategy validated by case studies to support secure, compliant big-data use in cloud environments.

2. Cloud Computing Environment and Big Data Characteristics

2.1 Cloud Computing Architecture and Service Models

As an on-demand computing paradigm that provides compute, storage and network resources, cloud computing typically features a three-tier architecture: the infrastructure layer, the platform layer and the application layer. The infrastructure layer primarily consists of large-scale distributed compute nodes and storage resources, offering elastic compute and massive data storage capabilities to upper layers. The platform layer supplies a unified runtime environment and middleware to support developers in building, testing and deploying cloud applications. The application layer serves end users by delivering diverse services and access interfaces. Through layered architecture, cloud computing enables dynamic resource scheduling and efficient utilization, meeting the high compute and storage demands of big data processing[1].From a service perspective, cloud offerings are commonly categorized as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). IaaS uses virtualization to provide users with basic resources (compute, storage and networking), allowing them to install operating systems and application environments autonomously. PaaS builds upon IaaS by offering an integrated development and runtime environment that simplifies application construction and management. SaaS delivers complete application systems to users on a ready-to-use basis, freeing users from concerns about underlying resources and platform details. With technological evolution, new models such as Data as a Service (DaaS) and AI as a Service (AIaaS) have emerged, further extending cloud capabilities for big data processing and intelligent applications. Different service models imply differences in resource control, management complexity and the allocation of security responsibilities. Generally, IaaS grants users greater flexibility but also requires them to bear more responsibilities for system security management; PaaS provides convenience but places some security and privacy obligations on service providers; SaaS lowers the usage threshold but raises particular concerns regarding data security and regulatory compliance. For big data applications, appropriately selecting and combining service models at different layers is key to ensuring both system performance and security[2].

2.2 Big Data Characteristics and Security/Privacy Challenges

Big data is commonly characterized by the “4V”s: large Volume, diverse Variety, high Velocity and low Value density. With the proliferation of IoT, mobile internet and smart terminals, data sources have become more extensive, encompassing structured business databases, semi-structured logs, and unstructured text, images and video[3]. These characteristics place stringent performance and scalability demands on systems for storage, transmission and analysis, while concurrently complicating data security and privacy protection. First, during data collection and transmission, the dispersed sources and heterogeneous collection methods make unauthorized capture and tampering more likely. If transmission channels lack robust encryption and authentication, sensitive information may be intercepted or maliciously altered in transit. Second, in storage and management, cloud platforms often centralize large volumes of data from different organizations and individuals, increasing the difficulty of preventing unauthorized internal access and ensuring tenant isolation, while creating a larger attack surface for adversaries. In addition, big data’s distributed nature may result in storage across multiple regions and data centers, raising cross-border data flow and compliance governance issues. In the analysis and application phase, privacy protection challenges become particularly acute[4]. The value of big data is often realized by deeply mining sensitive information such as user behavior, consumption patterns and health records; however, such mining can easily lead to privacy breaches. For example, data mining

algorithms can infer users' identity traits or sensitive attributes—even if data has been anonymized, re-identification may be possible through linkage with other data sources. Moreover, as AI and machine learning become widespread, model training typically requires access to large volumes of raw data, introducing additional risks of privacy leakage and data misuse. Overall, the scale, complexity and sensitivity of big data, combined with the openness and shareability of cloud environments, make data security and privacy protection cross-layered, multidimensional and high-risk issues. Therefore, designing and implementing effective big data security and privacy protection strategies in cloud environments is essential to ensure trustworthy and sustainable data utilization[5].

3. Big Data Security Issues in Cloud Computing Environments

3.1 Security Risks across the Data Lifecycle

In cloud environments, big data normally traverses multiple stages from generation to final destruction: collection, transmission, storage, processing and use, sharing and exchange, and disposal. The complexity and openness of these lifecycle stages mean that security risks are present throughout; a vulnerability at any stage may lead to severe security incidents. During data collection, risks stem from the authenticity of sources and the legality of collection processes. Sensors, mobile devices or third-party systems may be tampered with to produce false or poisoned data, undermining the reliability of subsequent analyses. Unauthorized data collection can also infringe personal privacy or violate legal requirements. In the transmission stage, data exchanged over public networks is susceptible to eavesdropping, tampering or man-in-the-middle attacks. Without secure transmission protocols and end-to-end encryption, sensitive information can be illicitly captured or its integrity compromised during transit[6].At the storage stage, the multi-tenant nature of cloud infrastructures means that different users' data may coexist on the same physical hardware. If access controls and isolation mechanisms are inadequate, risks include unauthorized access, malicious insider disclosure or misuse by the cloud provider. Distributed storage architectures may also be targeted by attackers; compromise of even a subset of nodes can lead to large-scale data exposure. During data processing and use, common risks include application vulnerabilities, code injection and operator errors. Big data analytics often involve aggregating and linking sensitive information; absent appropriate privacy protections, such aggregation may enable re-identification attacks that expose user privacy. In data sharing and exchange phases, increasingly frequent inter-organizational data interactions, cross-domain access and cross-border transfers complicate compliance and governance. Without unified access policies and compliance frameworks, data exchanged with external parties may be leaked, misused or transferred to untrusted third parties. In the disposal phase, failure to conduct secure erasure or physical destruction can leave residual data recoverable, resulting in privacy breaches and intellectual property loss. In summary, security hazards exist at every stage of the big data lifecycle, and cloud virtualization and distribution features exacerbate these risks. Effective protection therefore requires stage-specific defense measures and the construction of a full-lifecycle security assurance system[7].

3.2 Common Security Threats and Attack Techniques

In cloud environments, big data faces diverse and evolving threats. These include traditional network and application-layer attacks as well as novel methods targeting virtualization, containerization and distributed storage. Actors range from external cybercriminal groups, APT teams and extortionists to malicious or negligent insiders and third-party supply-chain attackers. Whether motivated by financial gain, political objectives or disruption, such attackers ultimately

threaten data confidentiality, integrity and availability, causing legal, financial and reputational harm to organizations. At the network and transport layers, eavesdropping, man-in-the-middle (MITM) attacks and distributed denial-of-service (DDoS) remain prevalent. Unencrypted or weakly encrypted links on public networks are vulnerable to interception, leading to data leakage[8]. DDoS attacks exhaust bandwidth or compute resources, disrupting cloud services and interrupting data processing pipelines. At the application layer, attacks such as SQL injection, cross-site scripting (XSS), cross-site request forgery (CSRF) and API abuse are especially dangerous in cloud-native services that heavily rely on REST/GraphQL and API gateways; attackers can exploit these vectors to exfiltrate data or escalate privileges. Cloud-specific virtualization and multi-tenant features introduce new attack surfaces. Virtual machine or container escape exploits, and vulnerabilities in hypervisors, container runtimes or host kernels, can break isolation and enable cross-tenant data access. Side-channel attacks that exploit shared physical resources (e.g., CPU caches, time-shared I/O) can infer sensitive information about co-located tenants. Metadata services (e.g., IMDS) and cloud provider management interfaces, if subjected to SSRF (server-side request forgery) or credential leakage, can be used for lateral movement and to steal temporary credentials, enabling broader access and compromise. Misconfiguration and privilege misuse are frequent causes of cloud data breaches[9]. Publicly readable object storage buckets (e.g., improperly configured S3 buckets), overly permissive IAM policies, hard-coded or leaked API keys/credentials, and unprotected service endpoints expose data to exploitation. In addition, compromised or misconfigured automation scripts and CI/CD pipelines can allow attackers to inject malicious code, backdoors or tampered build artifacts into the supply chain, which may then be deployed to production and lead to widespread data and model contamination. Attacks against data analytics and machine learning are increasingly important. Data poisoning injects malicious samples into training sets to cause erroneous behavior or plant backdoors. Model inversion and membership inference attacks can reconstruct sensitive attributes or determine whether a particular record contributed to training, thereby exposing individual privacy. Model-stealing attacks, by repeatedly querying a target model to reconstruct an approximate replica, undermine the model's commercial value and can facilitate evasion of security controls. Insider threats and social engineering cannot be ignored. Legitimate users or former employees may abuse access privileges, store sensitive data locally without encryption, or fall victim to social engineering that reveals credentials—often a significant source of data breaches. To address human-factor risks, technical controls must be coupled with organizational policies, audits and behavioral monitoring. Finally, attacks often follow composite, chained patterns: an attacker might obtain a low-privilege credential through phishing or brute force, exploit SSRF or API vulnerabilities to move laterally, and then leverage misconfiguration or cross-tenant flaws to exfiltrate data for ransom or sale on illicit markets. The distributed, multi-provider and third-party-dependent nature of cloud environments further amplifies such cascading effects. Therefore, identifying common threats should go beyond single techniques to analyze attack paths and adversary capability levels systematically, thereby informing the design of effective defensive strategies[10].

4. Big Data Privacy Protection Issues in Cloud Computing Environments

4.1 Primary Pathways and Manifestations of User Privacy Leakage

User privacy breaches typically follow a path from endpoints to the cloud and then to analysis and sharing; each stage carries its own technical risks and can be triggered by external attacks or internal errors. At the user and data-collection stage, endpoint devices, mobile applications, or embedded third-party SDKs that perform excessive collection, misuse permissions, or contain malicious code can expose sensitive information to untrusted parties at the outset. In addition, social

engineering and phishing attacks that deceive users or operations personnel into revealing credentials or installing harmful apps are important routes for initial data leakage.

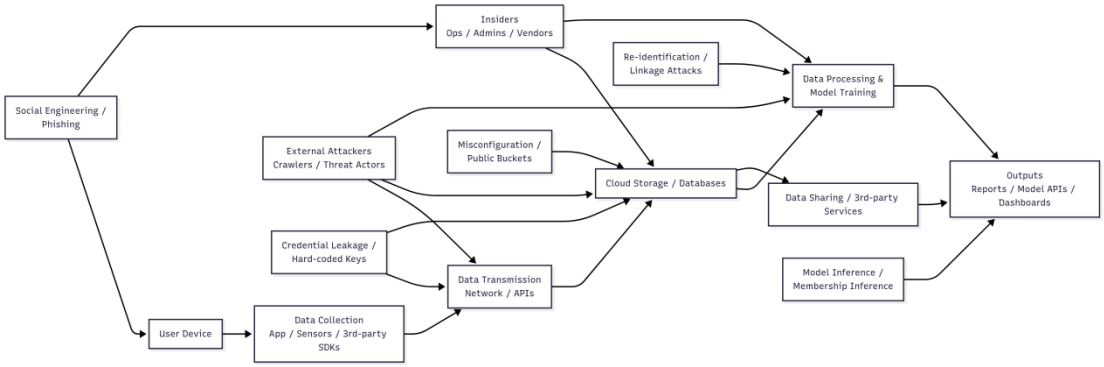

Figure 1 Primary privacy leakage pathways

As shown in Figure 1, during the transmission stage, if end-to-end encryption is not used or weak/incorrect configurations are in place (for example, outdated TLS versions or bypassed certificate verification), data in transit can be eavesdropped or tampered with; if API keys or short-lived credentials are leaked, attackers can exfiltrate large volumes of sensitive data via legitimate interfaces. Once data enters cloud storage, the multi-tenant environment and distributed storage introduce new risk surfaces: improper access-control policies, public object storage (for example, buckets without access restrictions), overly permissive IAM policies, or leaked operational credentials can directly result in large-scale data scraping by outsiders or misuse by insiders. Privacy risks in the data-processing and analysis stage are more covert and inference-oriented. When data are aggregated, cleaned and used to train machine-learning models, even if raw data have been de-identified or anonymized, attackers may still achieve "re-identification" through model inversion, membership inference, or linkage attacks that combine auxiliary data, recovering individuals' sensitive information from model outputs or statistical results. Moreover, poisoning of training data can introduce backdoors at the model level, causing leakage of sensitive attributes or manipulated predictions. In data-sharing and third-party collaboration, when data are exported, shared via APIs, or handed to outsourced partners, the risk of third-party misuse or resale increases significantly if strict contractual constraints, access auditing and least-privilege principles are not enforced. Cross-border transfers add legal and compliance complexity: jurisdictions define personal data differently and impose varying regulatory requirements, and compliance failures can result in substantial legal liabilities and fines. Manifestations of privacy breaches include, but are not limited to, personal identifiable information (PII) exposure leading to fraud or harassment; re-identification of anonymized datasets through linkage attacks; inference of sensitive attributes such as health status, preferences, or income from trained models; unauthorized access to corporate secrets or business data and their sale on illicit markets; and regulatory penalties and reputational damage caused by compliance violations (for example, unlawful cross-border transfers or failure to honor data-subject rights). It is noteworthy that many privacy incidents are not caused by a single technical fault but by cascading effects arising from misconfiguration, credential leakage, third-party dependencies and human factors (social engineering or operational mistakes).In summary, identifying and visualizing these primary pathways facilitates a "by-path" protection approach when designing privacy strategies: Organizations should minimize collection and enforce consent at the user side; strengthen encryption and key management at the transport layer; enforce strict access control and isolation at the storage layer; apply techniques such as differential privacy and federated learning at the processing layer to reduce inference risk; and constrain third parties through contracts, audits, and data-sanitization policies at the sharing layer. At the same time,

technical controls must be combined with organizational governance (operational processes, least-privilege practices, auditing and staff security awareness training) to reduce privacy incidents caused by human factors.

4.2 Privacy Protection Requirements and Representative Case Analyses

In cloud environments, privacy protection must address technical, governance and contractual dimensions. Core requirements include data minimization and purpose limitation; clear user notice and controllable rights; end-to-end encryption and encryption at rest; key and credential management; fine-grained access control and role-based authorization; and auditable accountability mechanisms. Typical risks stem from public exposure due to object-storage misconfiguration, API abuse caused by hard-coded or leaked credentials, excessive collection by third-party SDKs or instrumentation, and inference and membership-inference attacks against machine-learning models. Cross-border data flows also present major legal and compliance challenges. Targeted countermeasures include embedding security checks and policy enforcement in CI/CD pipelines, enforcing encryption and access logging for sensitive resources, implementing third-party component onboarding and runtime traffic monitoring, designing model services with differential privacy or federated learning, and applying query-rate limits and output-threshold controls. At the contractual level, data-processing agreements should clearly define responsibilities and audit rights and establish legal bases and technical isolation schemes for cross-border transfers. Meanwhile, organizations should institutionalize data classification and Data Protection Impact Assessment (DPIA) processes, strengthen incident response and forensics, and conduct regular penetration testing and compliance self-assessments. Balancing technical controls and governance processes, performing continuous risk assessments, conducting staff security training, and supervising third-party security are key to reducing privacy incidents, improving governance capabilities and ensuring compliance. Finally, privacy protection should be integrated across the full product and data lifecycle to form a sustainable closed-loop governance model, with governance effectiveness reported periodically.

5. Key Technologies for Big Data Security and Privacy Protection

Ensuring big-data security and privacy in cloud environments relies on a complementary, collaborative set of key technologies. First, end-to-end encryption and robust key management are foundational: use TLS and strong authentication for transport, apply server-side or client-side encryption for storage, and pair these with centralized key-lifecycle management and hardware security modules (HSMs/TEEs) to mitigate credential misuse and side-channel risks. Second, fine-grained access control and least-privilege policies—combining RBAC, ABAC and attribute-based authorization—enable precise control over data access in multi-tenant settings and mitigate long-term credential exposure via identity federation and temporary credentials. Third, privacy-preserving algorithms and paradigms are crucial during analysis and model training: differential privacy injects provable noise into statistical releases and query results to resist re-identification; federated learning and secure multi-party computation (SMC) enable collaborative training without centralizing raw data; and in some scenarios homomorphic encryption can provide end-to-end protection for specific computations. Fourth, data masking, anonymization and classification techniques, together with data-lifecycle management, create a pre-sharing and pre-audit defense layer; traceable data labeling and metadata management support compliance reviews. Finally, continuous monitoring and auditing, behavior-based anomaly detection, and embedding security checks into CI/CD pipelines combine technical controls with operational processes to form a “protect-detect-respond-audit” closed loop. Integrating these technologies in a

layered, risk-oriented manner and supplementing them with sound governance and contractual mechanisms is the key path to making cloud-hosted big data both usable and controllable.

6. Security and Privacy Protection Strategies in Cloud Computing Environments

Achieving security and privacy for big data in cloud environments requires a systematic strategy that combines technical measures, governance, and contractual controls. First, organizations/enterprises should establish a layered defense architecture. At the endpoint and transport layers, they must enforce data minimization, end-to-end encryption, and robust key management. At the storage and compute layers, they need to mandate encryption, implement fine-grained access control (RBAC/ABAC) and tenant isolation, and introduce sensitivity labels and lifecycle policies in the data catalog to enable automated governance. At the analytics and model layers, they should adopt techniques such as differential privacy, federated learning, and model access rate limiting to mitigate inference and membership-inference risks. Second, organizations/enterprises should embed security assurance into development and operations workflows. They can achieve this by integrating static analysis, configuration auditing, and policy enforcement into CI/CD pipelines. Additionally, they need to complement these measures with real-time monitoring, behavior-based anomaly detection, and auditable logging to form a protect–detect–respond–recover closed loop. On the governance side, organizations/enterprises must establish data classification and Data Protection Impact Assessment (DPIA) mechanisms. They should conduct periodic risk assessments and compliance self-evaluations, and mitigate human-factor risks through least-privilege principles, change-approval controls, and continuous security training. Externally, they need to regulate the supply chain through contractual clauses, service-level agreements (SLAs), and third-party security assessments. Meanwhile, they must define both the legal basis and technical isolation measures for cross-border data transfers. Finally, organizations/enterprises should implement incident response and digital forensics procedures. They need to conduct regular exercises and simulations, and publish transparent governance reports to ensure enforceability, accountability, and continuous improvement. This approach helps reconcile business innovation with the imperative to minimize privacy breaches and regulatory risk.

7. Case Studies and Empirical Evaluation

This chapter validates the feasibility and effectiveness of the proposed strategies through two representative cases: (1) reproduction and mitigation testing of a sensitive-data exposure caused by object-storage misconfiguration, and (2) risk assessment of membership inference and model-inversion attacks against publicly accessible model services. For the first case, we constructed a multi-tenant testbed with sample data, simulated erroneous permission policies, and introduced configuration checks and automatic rollback in the CI/CD pipeline. Evaluation metrics included the number of exposed records, mean time from vulnerability discovery to remediation, and counts of unauthorized access attempts. Results show that after applying the mitigation strategy, exposed record counts and unauthorized accesses decreased significantly, and mean remediation time was reduced. For the second case, we executed membership-inference and model-inversion experiments against a public API and compared attack success rates and model utility loss before and after applying differential-privacy noise, query-rate limits, and output-threshold controls. The experiments indicate that differential privacy can substantially reduce membership-inference success while preserving acceptable model performance. Together, the two empirical studies demonstrate that strategies that balance technical controls and process governance can effectively lower privacy risks, though limitations remain—such as the utility–privacy tradeoff when tuning

differential-privacy parameters and the difficulty of fully controlling third-party compliance. Accordingly, we recommend deploying protections according to business risk tiers, maintaining continuous monitoring and periodic drills, and extending empirical evaluation to cross-border and multi-vendor scenarios in future work.

8. Conclusion

This paper analyzes security and privacy issues for big data in cloud computing environments, outlining lifecycle risks, common attack vectors, and key protective technologies (end-to-end encryption, key management, fine-grained access control, differential privacy, federated learning, etc.). It proposes a systematic strategy that combines layered defenses, CI/CD security integration, governance, and contractual constraints. Case studies and empirical evaluations show that measures such as configuration auditing and differential privacy can significantly reduce leakage and inference risks, yet sustained governance is required to address challenges such as the privacy–utility tradeoff, third-party compliance, and cross-border regulation. We recommend that practitioners deploy technical and process controls according to business risk levels, strengthen continuous monitoring and incident-response exercises, and integrate privacy protection early in product and data design. These measures will provide sustainable support for trustworthy big-data use in the cloud era.

References

[1] Alabdulatif, Abdullah, Navod Neranjan Thilakarathne, and Kassim Kalinaki. "A novel cloud enabled access control model for preserving the security and privacy of medical big data." Electronics 12.12 (2023): 2646.

[2] Valivarthi, Dharma Teja. "Optimizing cloud computing environments for big data processing." International Journal of Engineering & Science Research 14.2 (2024): 1756-1775.

[3] Layode, Oluwabunmi, et al. "Data privacy and security challenges in environmental research: Approaches to safeguarding sensitive information." International Journal of Applied Research in Social Sciences 6.6 (2024): 1193-1214.

[4] Amaithi Rajan, Arun, and Vetriselvi V. "Systematic survey: secure and privacy-preserving big data analytics in cloud." Journal of Computer Information Systems 64.1 (2024): 136-156.

[5] Pawar, Ankush Balaram, Shashikant U. Ghumbre, and Rashmi M. Jogdand. "Privacy preserving model-based authentication and data security in cloud computing." International Journal of Pervasive Computing and Communications 19.2 (2023): 173-190.

[6] Banerjee, Somnath. "Challenges and Solutions for Data Management in Cloud-Based Environments." International Journal of Advanced Research in Science, Communication and Technology (2023): 370-378.

[7] Yanamala, Anil Kumar Yadav. "Emerging challenges in cloud computing security: A comprehensive review." International Journal of Advanced Engineering Technologies and Innovations 1.4 (2024): 448-479.

[8] Mohammed, Shameer, et al. "A new lightweight data security system for data security in the cloud computing." Measurement: Sensors 29 (2023): 100856.

[9] Tahseen, Asma, Sangyam Rohith Shailaja, and Yagnasri Ashwini. "Extraction for Big Data Cyber Security Analytics." Advances in Computational Intelligence and Informatics: Proceedings of ICACII 2023 993 (2024): 365.

[10] Ang'udi, Janet Julia. "Security challenges in cloud computing: A comprehensive analysis." World Journal of Advanced Engineering Technology and Sciences 10.2 (2023): 155-181.