International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1604-1609
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0228

A Data Sharing Protocol to Minimize Security and Privacy Risks of Cloud Storage in Big Data Era

S Tabassum[1], *K Nithys Sai Swaroop*[2], *K Mahendra Babu*[3], *N Naresh Babu*[4,] *S Mohammed Anas*[5]

[1]*Assistant professor, Dept. of CSE, Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh, India.*

[2,3,4,5]*4th B-Tech, Dept. of CSE, Annamacharya Institute of Technology and Science, Rajampet, Andhra Pradesh, India.*

Emails: *shaiktabassum618@gmail.com*[1], *nithyasai640@gmail.com*[2], *babumahendra604@gmail.com*[3], *nareshbabu231468@gmail.com*[4], *mohammedanasshaik03@gmail.com*[5]

Abstract

The big data cloud system uses the storage of cloud service providers to distribute data to legitimate users. Compared to traditional solutions, cloud providers store shared data in large data centers outside the trusted area of the data owner, which can provide data protection. Secret Group Key (SSGK) to protect public data and communications from unauthorized access. Unlike previous work, a group key is used to encrypt shared data, and a secret exchange scheme is used to distribute group data. Comprehensive security and performance analysis shows that our protocol significantly reduces security and data protection risks through cloud data exchange and saves about 12% of disk space.

Keywords: *Big Data,Cloud Computing,Data security,Encryption,Secret Group Key,Key Distribution,Secure Data Exchange,Cryptographic Techniques.*

1. Introduction

Data is an organization's most vital asset, forming the foundation for decision-making. It aids in curing diseases, boosting revenue, improving efficiency, and enhancing performance. Storage, analysis, and sharing are essential for optimizing operations. However, the rapid growth of data makes local storage challenging due to limited resources. To address this, most businesses have shifted to cloud services, benefiting from scalability, reliability, disaster recovery, and cost efficiency. Cloud computing offers vast storage and computational power, enabling seamless access across platforms, enhancing productivity, collaboration, and project management. With its expansion, nearly all businesses are expected to adopt cloud solutions. Despite its advantages, cloud computing faces security threats. Enterprises use cloud storage to ease local data management but risk losing control over sensitive information. Data sharing in an open environment exposes cloud servers to attacks, and even cloud providers may misuse data. Shared data among business partners, employees, and customers can be exploited, leading to confidentiality breaches, financial losses, and reputational damage. Organizations must implement strong security solutions. Several models ensure cloud data protection, focusing on leakage prevention and leaker detection. Strategies include cryptography, access control, differential privacy with machine learning, watermarking, and probabilistic techniques. Leakage prevention ensures secure sharing, while leaker detection identifies culprits. By 2021, 90% of organizational workloads had moved to the cloud. The industry was projected to grow at 14.6% annually, reaching $300 billion by 2022. With 75 billion IoT devices expected by 2025, cloud services remain integral. While cloud computing cuts costs and enhances storage flexibility, data confidentiality remains a concern, as users cannot fully trust Cloud Service Providers (CSPs). Data owners fear losing control, leading to unauthorized access. In 2021 alone, 22 billion data records were exposed, with a 5% increase expected in 2022. The global cost of a data breach hit $4.24 million, the highest in 17 years. COVID-19 further increased breach costs by $1.07 million due to remote work. Addressing data leakage is critical, requiring prevention and detection mechanisms. Though solutions exist, a systematic study is needed to determine their effectiveness.This work reviews key techniques for secure cloud data sharing, analyzing their mechanisms, strengths, and applications. A comparative analysis evaluates optimal techniques for different scenarios. Sections II–VI discuss cryptography, access control, differential privacy, watermarking, and probabilistic methods, detailing their function, research contributions, and applications. Section VII compares these techniques, while Section VIII concludes with findings and future directions. (Figure 1,2) [1-4]

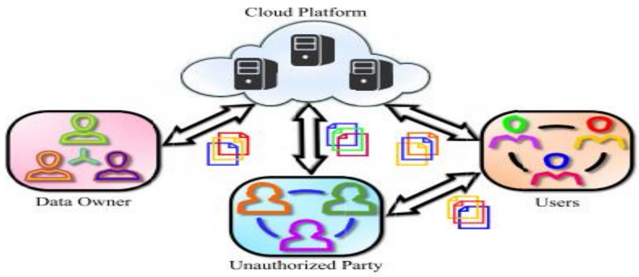

Figure 1 **Block Diagram of Sharing Environment**

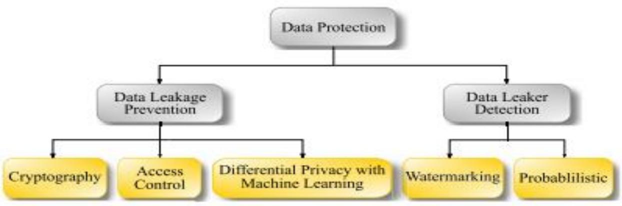

Figure 2 **Major Classification of Data Protection Techniques**

2. Cryptography Based Models

The set of entities to be encrypted is EτE_{\tau}Eτ, while SKSKSK, PBKPBKPBK, and PVKPVKPVK represent secret, public, and private keys, respectively. Symmetric cryptography maps encryption as e:Eτ×SK→Eτ∗ e : E_{\tau} \times SK \to E_{\tau}^*e:Eτ×SK→Eτ∗ and decryption as d:Eτ∗×SK→Eτd : E_{\tau}^* \times SK \to E_{\tau}d:Eτ∗×SK→Eτ such that d(e(Eτ,SK))=Eτd(e(E_{\tau}, SK)) = E_{\tau}d(e(Eτ,SK))=Eτ. Asymmetric cryptography maps encryption as e:Eτ×PBK→Eτ∗ e : E_{\tau} \times PBK \to E_{\tau}^*e:Eτ×PBK→Eτ∗ and decryption as d:Eτ∗×PVK→Eτd : E_{\tau}^* \times PVK \to E_{\tau}d:Eτ∗×PVK→Eτ such that d(e(Eτ,PBK))=Eτd(e(E_{\tau}, PBK)) = E_{\tau}d(e(Eτ,PBK))=Eτ. Here, Eτ∗E_{\tau}^*Eτ∗ represents encrypted documents. (Figure 3)

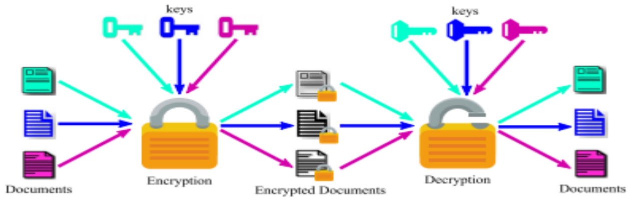

Figure 3 **Birds-Eye View of Cryptography Based Models**

The symmetric cryptography technique consists of three functions: the key generator Kgen(CG)Kgen(CG)Kgen(CG) generates key SKSKSK based on the security factor SFSFSF. The encryption function e(Eτ,SK)e(E_{\tau}, SK)e(Eτ,SK) transforms EτE_{\tau}Eτ into Eτ∗E_{\tau}^*Eτ∗, and the decryption function d(Eτ∗,SK)d(E_{\tau}^*, SK)d(Eτ∗,SK) retrieves EτE_{\tau}Eτ. Similarly, asymmetric cryptography generates public and private keys PBK,PVK=Kgen(CG)PBK, PVK = Kgen(CG)PBK,PVK=Kgen(CG), encrypts EτE_{\tau}Eτ using PBKPBKPBK, and decrypts Eτ∗E_{\tau}^*Eτ∗ using PVKPVKPVK. A cryptographic framework encrypts documents D={D1,D2,...,Dn}D = \{D_1, D_2, ..., D_n\}D={D1,D2,...,Dn} with keys K={K1,K2,...,Kn}K = \{K_1, K_2, ..., K_n\}K={K1,K2,...,Kn}, generating encrypted documents DED_EDE, which are decrypted by authorized users. Kao et al. proposed uCloud, using RSA to encrypt data via public keys while storing private keys on mobile devices. Al-Haj et al. developed two cryptographic algorithms ensuring confidentiality, integrity, and authenticity using symmetric encryption and elliptic curve digital signatures. Liang et al. introduced a Ciphertext-Policy Attribute-Based Proxy Re-Encryption Scheme, reducing communication costs. Wang et al. proposed FH-CP-ABE to secure hierarchical data, mitigating plaintext attacks but increasing computation costs.Liu et al. designed a fair data access control scheme with key reconstruction, generating fake keys for security but lacking an efficient authentication mechanism. Another CP-ABE scheme by Liu et al. reduced user-end decryption costs but had privacy limitations. Li et al. proposed LDSS for mobile cloud computing, offloading computation to proxy servers. Zaghloul et al. developed P-MOD, integrating privilege-based access control into ABE, outperforming CP-ABE and FH-CP-ABE in hierarchical encryption. Li et al. introduced an LSSS-based CP-ABE to improve policy efficiency and reduce costs. Zhang et al. proposed HP-CP-ABE with authority verification, ensuring constant-sized private keys while reducing storage and transmission costs. However, it supports only the AND policy, limiting security flexibility.

3. Access Control Based Models

The Access Control Mechanism (ACM) regulates data exposure based on type, user privileges, and permissions. An Access Control Policy (ACP) defines data distribution as (D,U,G)(D, U, G)(D,U,G), where DDD is data, UUU is users, and GGG sets access rules. Effective ACM requires predefined user privileges and secrecy levels. Fig. 4 illustrates a model where users U1,U2,U3U_1, U_2, U_3U1,U2,U3 request documents, receiving only authorized subsets. Nabeel and Bertino's scheme uses attribute-based encryption, minimizing owner overhead while ensuring confidentiality. A secure data-sharing method [27] prevents revoked users from accessing data, even in collusion with untrusted clouds. TMACS [23], a CP-ABE scheme, distributes attribute management for better security. A hierarchical access system [17] with CP-ABE reduces encryption, decryption, and storage overhead. Ali et al. [55] introduced DaSCE for cloud security, combining key management, access control, and assured deletion. Almutairi et al. [56] proposed Role-Based Access Control (RBAC) to limit data exposure in multi-tenant environments. Xu et al. [26] presented a fine-grained access model for dynamic groups, allowing policy enforcement, credential updates, and computation by untrusted CSPs. The TAFC method [57] integrates Timed-Release Encryption (TRE) with CP-ABE, enabling time-based access control. [5-8]

4. Differential Privacy with Machine Learning Based Models

A mechanism MN:D→Range(MN)MN: D \to \text{Range} (MN)MN:D→Range(MN) satisfies ϵ\epsilonϵ-differential privacy if the probability of obtaining an output OPOPOP remains bounded for any pair Di,Di′D_i, D'_iDi,Di′ differing in one record. Differential privacy in machine learning protects data by embedding statistical noise, ensuring privacy while allowing classification into categories {A, B, C, D}. Entities EτE_\tauEτ apply differential privacy using generated noise NGNGNG. The technique involves three functions: (1) Noise generation (Ngen(DP)N_{\text{gen}}(DP)Ngen(DP)), (2) Noise embedding (8e∗8^*_e8e∗), and (3) Noise extraction (9d∗9^*_d9d∗), ensuring noise application and retrieval for privacy protection. (Figure 4) [7-9]

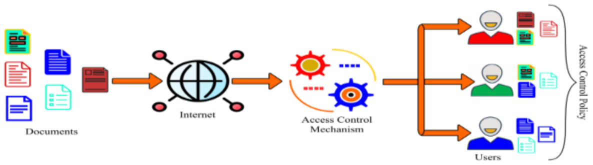

Figure 4 Schematic Representation of Access Control Based Models

Yonetani et al. introduced DPHE for secure computation but supported only one operation at a time. Hesamifard et al.'s CryptoDL enabled deep learning over encrypted data but lacked multi-key protection. Li et al.'s POCC framework used homomorphic encryption for cloud classification but assumed trusted storage servers, which is impractical. Li et al. also proposed a classifier delegation scheme using Naive Bayes and hyperplane decision-based classifiers but required frequent user interactions. PMLM used public-key encryption with ϵ\epsilonϵ-differential privacy but had high computational costs. Gao et al. developed a privacy-preserving Naive Bayes classification method but failed in truth discovery.Ma et al. introduced PDLM for deep learning over encrypted data using stochastic gradient descent (SGD), improving storage efficiency but suffering from high computational costs and low classification accuracy.

5. Water Marking Based Models

Watermarking embeds identifiable marks into data to ensure ownership and prevent unauthorized modifications. It involves key generation for security, embedding the watermark, and detecting it accurately. A robust system ensures imperceptibility, effectiveness, and resistance to attacks, verifying extracted watermarks using a similarity function. Watermarking applies to various data types, including text, images, audio, video, and relational data. (Figure 5) [10]

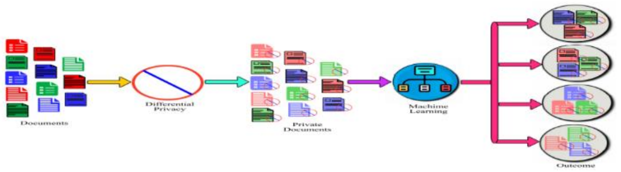

Figure 5 Standard Model for Differential Privacy with Machine Learning

Advanced techniques enhance security. Fingerprinting embeds multi-bit marks resistant to attacks, while optimization-based methods use pattern search and genetic algorithms for resilience. Mobile agent-based approaches automate detection, and security models integrate watermarking with access control, like Kumar et al.'s cloud-based Bell-La Padula model. (Figure 6) [11-12]

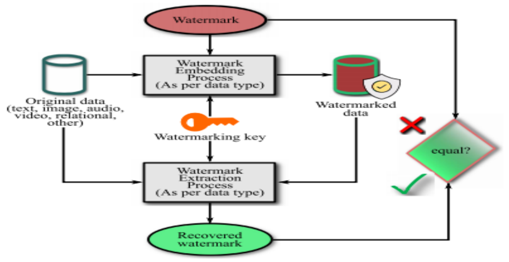
Watermark
Watermark Embedding Process (As per data type)
Original data (text, image, audio, video, relational, other)
Watermarked data
Watermarking key
equal?
Watermark Extraction Process (As per data type)
Recovered watermark

Figure 6 Key Components of Watermarking Based Models

Specialized methods address specific needs. Curvelet transforms embed watermarks in ECG signals, the LIME framework secures data sharing with watermarking and encryption, and Active Bundles protect data through access control. Image watermarking uses Hidden Markov Models for secure transactions, while cloud integrity auditing ensures secure storage.Further advancements include GAHSW for robust database watermarking, digital text watermarking for copyright protection, and JWEC for medical image security. Peng et al. introduced reversible watermarking for encrypted vector graphics, enhancing robustness. These methods demonstrate significant progress in digital watermarking for security and integrity. [13-14]

6. Probability Based Models

The probability technique assesses whether an agent UjU_jUj is responsible for a leaked dataset LLL by analyzing data overlap and the likelihood of guessing objects. Agents possessing parts of LLL may be suspected, but they can argue that the data came from other sources, such as another company or public records. The larger and rarer LLL, the harder it is to deny responsibility. If an object in LLL was exclusive to U1U_1U1, they become more suspect. Probability estimates help determine accountability, e.g., if 90% of emails are found online, the discovery probability is 0.9, whereas for bank accounts, it may be 0.2. (Figure 7) [15-16]

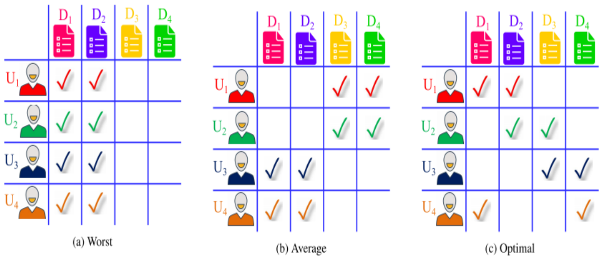

Figure 7 Distribution Strategy (a) W∗ *j* = W∗ *k* (b) minimize P *j*6=*k* |W∗ *j* ∩ W∗ *k* | (c) minimizeP *j* |W 1 ∗ *j* | P *k*6=*j* |W∗ *j* ∩ W∗ *k* |.

Data distribution strategies help identify a guilty user MUMUMU by minimizing overlap between datasets assigned to agents. The optimal strategy ensures minimal data sharing. Papadimitriou and Garcia-Molina’s agent guilt model evaluates if leaked data originated from an agent or an external source. Harel et al. introduced misuseability weights to assess data sensitivity and insider risks. Kumar et al. proposed allocation strategies to prevent unauthorized data use and enhance guilty party identification. Fan et al. developed a file distribution model minimizing overlap, improving leak detection. TM-Score extends misuseability weight evaluation for textual data leaks. Sodagudi and Kurra introduced a method for detecting malicious attackers in MANETs. Guevara et al. designed an anomaly-based detection algorithm analyzing user behavior, achieving high accuracy but being time-consuming. Ezhilchelvan and Mitrani studied VM co-residency risks in public clouds, showing promising results but requiring real-world validation. [16]

7. Comparative and Comprehensive Analysis

Among five techniques—Cryptography (CG), Access Control (AC), Differential Privacy (DP), Watermarking (WM), and Probability (PB)—only CG and DP ensure both privacy (P) and security (S). CG, AC, and DP prevent leaks (L), while WM and PB detect leakers (D). No single technique provides both. CG, AC, and DP ensure confidentiality (C), integrity (I), and accessibility (A), while WM and PB provide only integrity. CG and DP excel in data protection (DR) but lack leaker detection. For data usability (DU), utility (U) is low in CG and AC, moderate in DP, and high in WM and PB. Sharing (X) is supported by all. In cloud environments (CE), CG and AC apply to private clouds, all five to public clouds, and CG, AC, and PB to hybrid clouds. CG, DP, and WM involve data transformation, increasing costs, while AC and PB avoid it but still have overheads. CG ensures privacy and security but risks key compromise and lacks leaker identification. AC controls disclosure but cannot detect leakers. DP preserves privacy and utility but is ineffective in identifying culprits. WM detects leakers but fails if the watermark is removed. PB provides strong leaker estimation but lacks privacy and prevention. CG excels in privacy and security, AC in controlled access, DP in balancing privacy and utility, while WM and PB are best for leaker detection. No single technique is sufficient; integration is needed for full data protection. [17]

8. Conclusion and Future Work

Data protection in cloud computing and information security is a challenging task. Numerous efforts address this challenge, but a comprehensive study of existing solutions is lacking. This paper provides an in-depth analysis of key techniques for secure data sharing in cloud environments, highlighting their functionality, research gaps, and future directions. A thorough comparison of these techniques is conducted, assessing their relevance in different contexts. It is observed that no single technique can fully secure data against all involved entities. A robust solution requires integrating multiple techniques to ensure complete security in shared environments. The insights presented in this analysis serve as a milestone for researchers and emerging applications requiring secure data storage and sharing. [18]

References

[1]. K. Singh and I. Gupta, ‘‘Online information leaker identification scheme for secure data sharing,’’ Multimedia Tools Appl., vol. 79, no. 41, pp. 31165–31182, Nov. 2020.

[2]. E. Zaghloul, K. Zhou, and J. Ren, ‘‘P-MOD: Secure privilege-based multilevel organizational data-sharing in cloud computing,’’ IEEE Trans. Big Data, vol. 6, no. 4, pp. 804–815, Dec. 2020.

[3]. Gupta and A. K. Singh, ‘‘GUIM-SMD: Guilty user identification model using summation matrix-based distribution,’’ IET Inf. Secur., vol. 14, no. 6, pp. 773–782, Nov. 2020.

[4]. W. Shen, J. Qin, J. Yu, R. Hao, and J. Hu, ‘‘Enabling identity-based integrity auditing and data sharing with sensitive information hiding for secure cloud storage,’’ IEEE Trans. Inf. Forensics Security, vol. 14, no. 2, pp. 331–346, Feb. 2019.

[5]. Gupta and A. K. Singh, ‘‘An integrated approach for data leaker detection in cloud environment,’’ J. Inf. Sci. Eng., vol. 36, no. 5, pp. 993–1005, Sep. 2020.

[6]. R. Li, C. Shen, H. He, X. Gu, Z. Xu, and C.-Z. Xu, ‘‘A lightweight secure data sharing scheme for mobile cloud computing,’’ IEEE

Trans. Cloud

[7]. Comput., vol. 6, no. 2, pp. 344–357, Apr. 2018. Gupta, N. Singh, and A. K. Singh, ‘‘Layer-based privacy and security architecture for cloud data sharing,’’ J. Commun. Softw. Syst., vol. 15, no. 2, pp. 173–185, Apr. 2019.

[8]. J. Li, S. Wang, Y. Li, H. Wang, H. Wang, H. Wang, J. Chen, and Z. You, ‘‘An efficient attribute-based encryption scheme with policy update and file update in cloud computing,’’ IEEE Trans. Ind. Informat., vol. 15, no. 12, pp. 6500–6509, Dec. 2019.

[9]. Suisse. (2017). 2018 Data Center Market Drivers: Enablers Boosting Enterprise Cloud Growth. Accessed: May 19, 2019. [Online]. Available:https://cloudscene.com/news/2017/12/2018-data-center-predictions/

[10]. Gupta and A. K. Singh, ‘‘A framework for malicious agent detection in cloud computing environment,’’ Int. J. Adv. Sci. Technol., vol. 135, pp. 49–62, Feb. 2020.

[11]. Y. Li, Y. Yu, G. Min, W. Susilo, J. Ni, and K.-R. Choo, ‘‘Fuzzy identity-based data integrity auditing for reliable cloud storage systems,’’ IEEE Trans. Dependable Secure Comput., vol. 16, no. 1, pp. 72–83, Jan./Feb. 2019.

[12]. Gupta and A. K. Singh, ‘‘A probabilistic approach for guilty agent detection using bigraph after distribution of sample data,’’ Proc. Comput. Sci., vol. 125, pp. 662–668, Jan. 2018.

[13]. L. Zhang, Y. Cui, and Y. Mu, ‘‘Improving security and privacy attribute based data sharing in cloud computing,’’ IEEE Syst. J., vol. 14, no. 1, pp. 387–397, Mar. 2020.

[14]. Gupta and A. K. Singh, ‘‘Dynamic threshold based information leaker identification scheme,’’ Inf. Process. Lett., vol. 147, pp. 69–73, Jul. 2019.

[15]. S. Wang, J. Zhou, J. K. Liu, J. Yu, J. Chen, and W. Xie, ‘‘An effi- cient file hierarchy attribute-based encryption scheme in cloud computing,’’ IEEE Trans. Inf. Forensics Security, vol. 11, no. 6, pp. 1265–1277, Jun. 2016. Gupta and A. K. Singh, ‘‘SELI: Statistical evaluation based leaker identification stochastic scheme for secure data sharing,’’ IET Commun., vol. 14, no. 20, pp. 3607–3618, Dec. 2020.

[16]. W. Teng, G. Yang, Y. Xiang, T. Zhang, and D. Wang, ‘‘Attribute-based access control with constant-size ciphertext in cloud computing,’’ IEEE Trans. Cloud Comput., vol. 5, no. 4, pp. 617–627, Oct./Dec. 2017.

[17]. Gupta and A. K. Singh, ‘‘A probability based model for data leakage detection using bigraph,’’ in Proc. 7th Int. Conf. Commun. Netw. Secur. (ICCNS). New York, NY, USA: Assoc. Comput. Machinery, 2017, pp. 1–5.

[18]. L. Columbus. (Jan. 2018). 83% of Enterprise Workloads Will Be in the Cloud by 2020. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2018/01/07/83-ofenterprise-workloads-will-be-in-the-cloud-by-2020/#50d375286261