International Journal of Technology, Leadership and Sciences (IJTLS)
ISSN: 3107-9083 www.ijtls.com Paper ID: TLS-26/V2/I2-020
DOI: https://doi.org/10.5281/zenodo.19127082

SECURE CLOUD DATA DEDUPLICATION FOR INTEGRITY ASSURANCE AND STORAGE OPTIMIZATION

[#1]**G. LAKSHMI,** ***Associate Professor,***
[#2]**SUNAINA SADAF,** ***B.Tech Student,***
[#3]**MOHAMMED SANIA,** ***B.Tech Student,***
[#4]**BIYYALA SATWIKA,** ***B.Tech Student,***
[#5]**EDLA ADARSH,** ***B.Tech Student,***
Department of AIML,
TRINITY COLLEGE OF ENGINEERING AND TECHNOLOGY, PEDDAPALLY, TG.

ABSTRACT: Duplicate data elimination, which is frequently referred to as deduplication, is one method of significantly reducing storage expenses. The necessity of data deduplication has grown in tandem with the popularity of cloud computing and data storage. Keeping a copy of copied content as a backup is one way that cloud providers might minimize expenses. Cloud computing has completely changed the way services are provided by allowing for the integration of multiple resources across the Internet. The most important and widely used cloud service is data storage. In order to protect the privacy of data owners, encryption is a basic element of most cloud storage options. Data encryption makes cloud data deduplication more difficult, which is a major problem because data processing and storage rely on deduplication. Some out-of-date techniques can't crack encrypted files. Current techniques of decrypting encrypted data have security flaws that make them ineffective for revocation and flexible data access control, such as being vulnerable to brute force assaults. With this limitation, the amount of data that can be used for practical purposes with any level of assurance is limited. Following the instructions in this article will decrypt your encrypted cloud data and remove any duplicates. The login test cases and deduplication process flow diagram of the completed solution were inspired by data deduplication in cloud storage. An AES-based encryption scheme is used for data security, while RAS is employed for hash creation.

Keywords: Deduplication, Encryption Technique, Decryption Technique.

International Journal of Technology, Leadership and Sciences (IJTLS)
ISSN: 3107-9083 www.ijtls.com Paper ID: TLS-26/V2/I2-020
DOI: https://doi.org/10.5281/zenodo.19127082

1. INTRODUCTION

Providers of cloud computing services should naturally work to eradicate data duplication, given that their customers are often offered almost unlimited storage capacity. Extraneous data is not compressed using this method. Users may compress data on a regular basis.

Data duplication is the practice of keeping the same information twice. Cloud services do not employ deduplication technology, therefore customers may get an error message if they try to upload duplicate files to the cloud. Backups can yield space reductions of 90% to 95%, whilst conventional file systems may only reach a deduplication storage efficiency of 68%.Data security, privacy, portability, and total ownership cost are all profoundly affected by encryption. Unfortunately, encryption prevents deduplication from being implemented. When two pieces of encrypted data look the same, it's hard to tell them apart. On the other side, deduplication finds and keeps copies separately. If consumers opt for the widely used encryption method, cloud storage providers will make deduplication useless. Unencrypted cloud storage leaves customer data susceptible to unauthorized access. Check that the area is secure.

Data replication is a common method for recovering keys in convergent encryption. This plan is an attempt to meet two separate needs. It would appear that convergent encryption is a good way to eliminate unnecessary data and safeguard privacy; nevertheless, there are some known drawbacks to it. The main topics covered in this essay are data compression and cloud storage. Among the many ways that IT services are offered, cloud computing is one of them. A certain amount of storage space and processing power is allotted to each user.

The use of interconnected computers in a network, a large pool of resources can be created through cloud computing. It has many benefits, including being cost-effective, scalable, and fault-tolerant. The provision of services can now be built upon it. Among the most important and often used cloud services is data storage. Because user data is securely stored in the provider's data center, cloud users may rest confident.

Data saved by customers in the cloud could be accessible to unauthorized individuals due to service providers' perhaps insufficient security procedures. Customers run the danger of having their privacy invaded and their data exposed when they give up control of their information. Naturally, worries about the safety of private data have surfaced in response to the rapid development of data mining and analytical technologies. Customers' privacy and security are compromised when data is encrypted before being transmitted to the cloud.

International Journal of Technology, Leadership and Sciences (IJTLS)
ISSN: 3107-9083 www.ijtls.com Paper ID: TLS-26/V2/I2-020
DOI: https://doi.org/10.5281/zenodo.19127082

The Cloud Service Provider may still come across copies of customer data even after it has encrypted it. More and more people getting involved makes this a more realistic possibility. An unorganized data management system, more network congestion, and wasted energy are all possible outcomes of cloud data duplication, which is a major concern in and of itself. The addition of supplemental services will highlight the importance of efficient resource management.

Deduplication makes it harder to store encrypted data efficiently. Still, processing encrypted data is beyond the capabilities of standard industrial de-duplication methods. It is currently possible to launch brute force assaults on the deduplication mechanism. Access limitations or retractions cannot be applied to data that is constantly changing. At this time, we do not have a solution that guarantees secrecy, security, or long-term profitability. For a number of reasons, getting data owners to authorize deduplication is infamously difficult. Since data users might not be ready to implement this management technique or might not be online at the same time, we apply temporary storage delays as a precaution.

Data owners could be reluctant to chip in if deduplication is going to be a lengthy process that involves a lot of talking and numbers. Individuals' privacy could be at risk in the search for duplicate data. As a result of data hyper-distribution, data proprietors might not be able to supply clients with data or compression keys. As a result, data deduplication initiatives cannot be coordinated between data proprietors and CSPs. This method successfully removes duplicate records from cloud storage, as shown by the results.

2. RELATED WORK

Deduplication can only work if the Document Object Model (DOM) is present. The use of convergent encryption solved the problem of duplicate data. The key K = H(F) represents the hash code of a specific data item F. Bellare developed DupLESS, an encryption method, which encrypts data F using Key K and makes it decryptable for anyone with both elements. It flattens data kept on servers. Due to the need for a separate key server for each key, DupLESS's block-level deduplication computations are slow. To encrypt data on the client side, Liu came up with an impractical solution that required data owner identification and copy destruction. For use in hybrid cloud settings, Cui created a deduplication solution that works with access control lists (ACLs). In order to facilitate backup identification, a quick hash has just been created. Because many data segments could have the same short hashes, collisions and offline brute force attacks are possible.

International Journal of Technology, Leadership and Sciences (IJTLS)
ISSN: 3107-9083 www.ijtls.com Paper ID: TLS-26/V2/I2-020
DOI: https://doi.org/10.5281/zenodo.19127082

The complex technologies eventually put simple plans into action. Due to its ineffectiveness with regularly replicated content, this strategy is best suited for individuals who rely on cloud storage. These strategies won't work if consumers' views on data ownership change over time.

B. A dynamic control system that combines AP and Pub-CSP, re-encrypting intermediates with a third party that is either reliable or semi-reliable, can handle content duplication and unexpected changes in ownership. In order to distribute convergent keys, Wen came up with a mechanism. Data consumers do not have the processing power to decrypt and encrypt data, or to get convergence keys from private shares.

Hur showcased techniques to drastically cut down on redundant data generated by the many data consumers in the cloud. Afterwards, Yan adjusted his approach to incorporate several methods for storing data. Unfortunately, the owner cannot guarantee availability or constant access. The access point can automatically allow access even when the owner isn't around. Integrating group keys, attribute-based access control, and trusted entities, Premkamal improved upon earlier systems. Security is put at risk by these tactics, even though they work for dynamic ownership management. Many proxy re-encryption solutions rely on third parties that data owners don't know, therefore they may be hesitant to give access to an unknown entity.

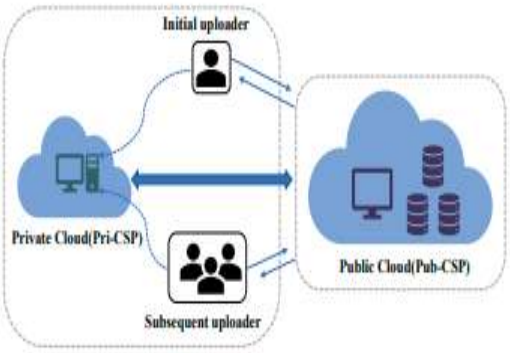


Fig. 1. Architecture of a data deduplication system.

Along with taking into account Connecting to a proxy service or another third party allows a malevolent individual to access files A and B. Here is the main goal of the research that has been suggested.

3. SYSTEM MODEL

Here we shall talk about the data deduplication system's operation and the characteristics of a threat actor. This research just uses file-level deduplication as its metric.

International Journal of Technology, Leadership and Sciences (IJTLS)
ISSN: 3107-9083 www.ijtls.com Paper ID: TLS-26/V2/I2-020
DOI: https://doi.org/10.5281/zenodo.19127082

Hybrid Architecture for Secure Deduplication

The data compression system's three main parts are shown in Figure 1..

- **Data users (DU).** Maintaining access to Pub-CSP data is a primary focus for the group. A public key, a secret signature key, and a proxy re-encryption (PRE) key (sku, pku) are all distributed to each user of the deduplication system. When a file is initially published, the user is regarded as its creator (u1,A). When a file FA is present, they take on the responsibility of holder (ui,A).
- **Public Cloud (Pub-CSP).** operation that oversees a platform for collaborative cloud storage. Our research is based on Pub-CSP's accessibility and suitable capacity.
- **Private Cloud (Pri-CSP).** Building an execution environment and infrastructure for data consumers can solve the problems of Pub-CSP's present volatility and DU's restricted processing capabilities. Data owners, hash values, and re-encrypted keys are all kept in Pri-CSP's database. Data is compiled into this set of documents. By definition, this system is based on the assumption that PriCSP is utilized by a single major organization. The company's ownership, management, and overall direction are all handled by the corporation itself. As a result, Pri-CSP deserves everyone's trust. Following the rules of the system is important to Pub-CSP, even though it is interested in raw user data. We think that the competing financial interests of Pub-CSP and Pri-CSP will ensure that they will never work together.

Security Requirements

There are additional security measures that must be followed in order to restore control, prevent collusion, validate ownership, and ensure data privacy and consistency.

- **Data privacy**. A key component of data privacy is safeguarding sensitive information, particularly that stored on the Pub-CSP server, from unauthorized access.
- **Data consistency.** The tags are immune to harm. Those with access to the information can detect when the ciphertext has been altered.
- **Ownership verification**. The marks are completely safe. Modifications to the ciphertext can be detected by those who have access to the data.
- **Ownership revocation** Possession proof. Information whose ownership cannot be verified, including ciphertext and related messages, should not be shared with anyone.

- Retaining something. The name of a rightful owner will be removed from the register and they will be denied access if they request the deletion or modification of their data from the cloud.
- **Collusion resistance**. No one should have access to the raw data without the necessary legal custody, regardless of whether they are collaborating with the Pub-CSP or other illegal data consumers.

4. PERFORMANCE EVALUATION

Consider the current state of affairs. Our data deduplication method is compared and contrasted with DedupDUM, RCE, CE, and randomized convergent encryption in this study. The main characteristics of the system are the ability to verify ownership, maintain consistent tag integrity, manage titles in real-time, encrypt data to eliminate duplicates, and regulate access. Each option significantly improves secrecy thanks to its encrypted storage. A tag consistency attack, however, might put Scheme CE at risk. In order to maintain data integrity and guarantee that the IDs in the received data are consistent, DU can use a variety of strategies. Using the public key of the DU, DedupDUM generates a group key that can be used for dynamic ownership management. Making new users and removing old ones is now much easier. Having just an identity, a tag, and some fake ciphertext does not guarantee that the user has full access to the original file.

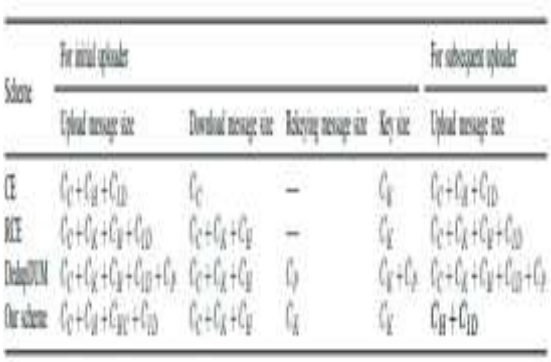
The possibility that the attacker is working with an unscrupulous cloud server is also not considered. To dynamically track who owns data F, our technique differs from past approaches by keeping two separate inventories—one at the Pri-CSP and the other at the Pub-CSP. Since our method uses pkui to generate the re-encrypted key REKui, adding and removing cloud users is a breeze. Find out that DU does not have legal ownership of file F, and deduplication will be performed.

The result is a dramatic drop in transportation costs. For the B section of the test, Table 1 shows how much each of the four routes of communication will cost. Cloud user ID (CC), hash code (CID), hash code group (CH), and encrypted data volume (CHC) are the acronyms for these concepts. While F and CK denote a key's capacity, Cp denotes the size of a public key. When sending F's initial data, all three methods—CE, RCE, and DedupDUM—create upload messages that are identical in size. Before deduplication, our method finds the rightful owner of DU0 by expanding the hash code set HC(F). Our technique stands out from the rest

International Journal of Technology, Leadership and Sciences (IJTLS)
ISSN: 3107-9083 www.ijtls.com Paper ID: TLS-26/V2/I2-020
DOI: https://doi.org/10.5281/zenodo.19127082

because, as shown in Table 1, it uploads H(F) only to prepare the way for the upcoming upload of F, before any authentication for access or ownership takes place.

Table 1 Communication Overhead.



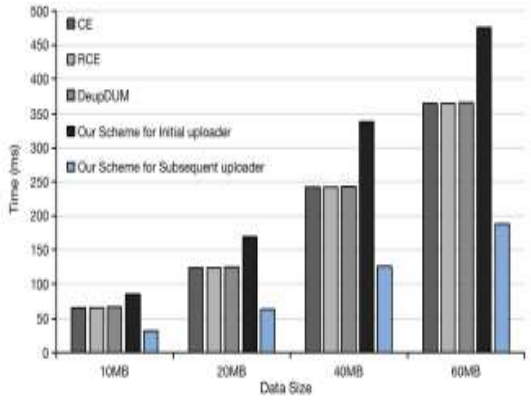


Fig. 2. Computation time for upload.

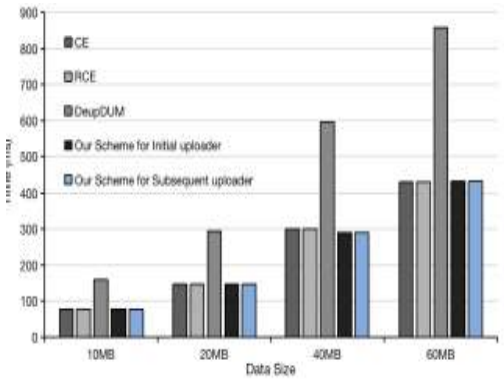


Fig. 3. Computation time for download.

Our method improves the re-encryption key in conjunction with DedupDUM. However, CE and RCE do not account for significant changes when determining the dimensions of the rekeying message. Although the group key controls ownership revocation, once identified, the data F that determines the encryption key K in DedupDUM does not change. Owners can

International Journal of Technology, Leadership and Sciences (IJTLS)
ISSN: 3107-9083 www.ijtls.com Paper ID: TLS-26/V2/I2-020
DOI: https://doi.org/10.5281/zenodo.19127082

leave the system vulnerable by using Pri-CSP to decrypt data before rekeying. As long as ownership could be revoked and we were willing to put in the extra effort, we found that using Pri-CSP instead of DU made our approach more secure. assessment of the results. Here, we highlight the ways in which we differ from other writers' approaches and compare and contrast them. We used the most recent versions of the crypto library (1.4.1) and the umbral library (0.3.0) to test all cryptographic operations to guarantee fairness. The AES algorithm generates the 128-bit key required for encryption and decoding. The amount of disk space could be anywhere from 10 MB to 60 MB. The test machine's 3.1 GHz Intel(R) Core(TM) i5-7300HQ CPU and 16.0 GB of RAM should be more than sufficient. On one of the file-sharing sites we examined, we observed an upload processing time. The DedupDUM system's calculations are similar to those made using the CE and RCE methodologies. The procedures for encrypting data F with AES are depicted in Figure 2, which includes calculating the hash code and hash code set, confirming and certifying the signature, decrypting the re-encrypted key, and ultimately encrypting data F. Comparing this method to other solutions, there is just a modest improvement. The advantages of our approach for concurrent file uploads are obvious. Unlike earlier systems, ours just needs to re-upload H(F) before access and title are verified. By using our approach, the exorbitant cost of communication can be considerably reduced.

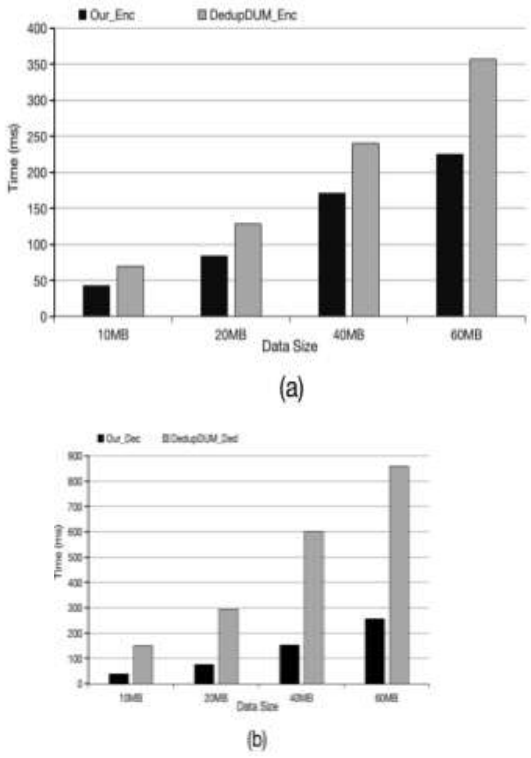


Fig. 4. Computation time for (a) encryption and (b) decryption.

International Journal of Technology, Leadership and Sciences (IJTLS)
ISSN: 3107-9083 www.ijtls.com Paper ID: TLS-26/V2/I2-020
DOI: https://doi.org/10.5281/zenodo.19127082

Figure 4 displays the amount of time needed to encrypt and decrypt data. Transmission takes no more than 0.373 seconds, while duplicate deletion takes 0.251 seconds. The suggested deduplication method can significantly cut down on upload times.

5. SECURITY ANALYSIS

Our method's security is assessed based on how well it protects data ownership, confidentiality, consistency, revocation, and resistance to collusion.

A. Data Privacy

No one should have access to raw data, including the trustworthy but inquisitive Pub-CSP. Consequently, there are usually two kinds of risks connected to Pub-CSP: users being misinformed. Only the re-encrypted key for the authorized DU will be kept by Pub-CSP in the case of an attack. Only once it has been encrypted with PRE using Pri-CSP can this key be decrypted using the DU's private key. If Pub-CSP relies on Pri-CSP and authorized users to earn revenue, it will be unable to decode plaintext using the cipher key. By using (Fid, u1, id) to request data F, an unauthorized user u2 receives the response "Fid, CT, H(F), (u1, id, REKu1)" from Pub-CSP. This indicates that while u1 is legal, u2 and id do not pass the ownership verification. Since REKu1 can only be decrypted with user u1's private key, user u2 is unable to decode CT and access F's contents. As a result, neither the interested but well-intentioned Pub-CSP nor the malevolent actors who shouldn't have access to the data are informed.

B. Data Consistency

Upon decryption, data owners may find that data deduplication activities are susceptible to damaging tag consistency assaults. If FB equals FA, user u2 can generate a fictitious ciphertext CTB and submit it to Pub-CSP as a CTA, assuming that both u3 and u2 have access to the same data FA. After the real person has finished

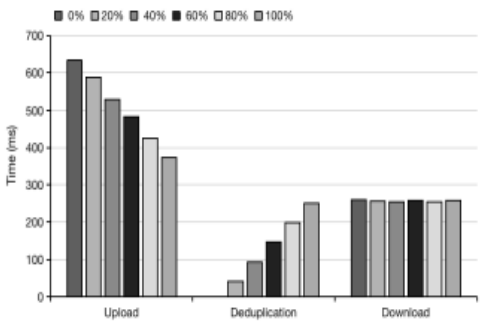


Figure5 displays the amount of time required to compute each operation using a duplicate ratio.

International Journal of Technology, Leadership and Sciences (IJTLS)
ISSN: 3107-9083 www.ijtls.com Paper ID: TLS-26/V2/I2-020
DOI: https://doi.org/10.5281/zenodo.19127082

As it expands, U3 plans to upload FA to facilitate the transmission of H(FA) to Pub-CSP for duplicate verification. PriCSP has been requested by Pub-CSP to eliminate all existing instances of H(FA). Pub-CSP designates users u3 (CTB, REKu3) following data deduplication. When H(Decrypt(De(sku3, REKu3, CTB)) = H(FA), user u3 verifies this. In the event that u3 is not stopped, the data will leak to Pub-CSP. Thus, our approach guarantees the accuracy of all data.

C. Data Ownership Verification

By randomly selecting a hash code from the dataset, namely between 10.5% and 14.3% of F, our method determines the data's true owner. Because the function H is not invertible and Fx is chosen at random, it is challenging to calculate H(Fx) in situations where the original plaintext is unavailable.

D. Ownership Revocation

Data F should only be accessible by authorized users. Ownership of u0 is revoked by naming Pri-CSP as its owner. Before re-uploading the plaintext to Pub-CSP, owner u0 encrypts it using the new symmetric key. When the original owner u1 of the data revokes ownership or other holders choose to remove or modify their data, the owner u0 updates the re-encrypted keys of the remaining users. When an access check fails, the most recent ciphertext cannot be decrypted by the original data owner using the original cipher-key. ***E.***

Collusion Resistance

This essay will examine the risks of utilizing Pub-CSP in coordinated attacks as well as how to use Pri-CSP's dependability to identify the perpetrators. Once an impostor, u1, obtains the plaintext of data F through a dishonest Pub-CSP, the Pub-CSP instructs Pri-CSP to use the phony data to deduplicate u1's data. Prior to providing the re-encrypted key REKu1, PriCSP policies require that u1 contain the data F. The cipher key will remain with Pub-CSP and the ownership verification will fail because u1 does not have access to the plaintext. Second, because each key is unique to its owner, criminals will still be unable to decipher cipher-keys even if they collaborate. Collaboration is not possible with our method.

6. CONCLUSION

Utilizing deduplication in a hybrid cloud architecture, we devised a method for efficiently and securely processing encrypted data by approaching the issue as one of ownership. Storage management is handled by Pub-CSP, dynamic ownership and deduplication are handled by Pri-CSP, and Pri-CSP serves as a proxy. Furthermore, our methodology

demonstrates that the original, unencrypted data can only be accessed by the legitimate owner, whereas encrypted data can only be accessed by authorized entities. Our security research, comparison with prior research, and implementation-based performance evaluation have confirmed that our approach is secure, efficient, and impervious to collusion and duplication assaults.

REFERENCES

[1] Ibrahim, Abaker, Targio, Hashem, Ibrar, Yaqoob, Nor, Badrul, Anuar, and Salimah, "The rise of "big data" on cloud computing: Review and open research issues," Information Systems, vol. 47, no. Jan., pp. 98– 115, 2015.

[2] F. M. Awaysheh, M. N. Aladwan, S. Alawadi, J. C. Cabaleiro, and T. F. Pena, "Security by design for big data frameworks over cloud computing," IEEE Transactions on Engineering Management, vol. PP, no. 99, 2021.

[3] Duan and Qiang, "Cloud service performance evaluation: status, challenges, and opportunities-a survey from the system modeling perspective," Digital Communications & Networks, pp. 101–111, 2016.

[4] Z. Yan, L. Zhang, W. Ding, and Q. Zheng, "Heterogeneous data storage management with deduplication in cloud computing," IEEE Transactions on Big Data, pp. 1–1, 2017.

[5] S. Quinlan and S. Dorward, "Venti: A new approach to archival storage," proc.usenix conf.on file & storage tech, 2002.

[6] G. R. Blakley and C. Meadows, "Security of ramp schemes," in Advances in Cryptology, Crypto 84, Santa Barbara, California, Usa, August, 1984.

[7] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Server-aided encryption for deduplicated storage," in Usenix Conference on Security, 2013.

[8] X. Jin, L. Wei, M. Yu, N. Yu, and J. Sun, "Anonymous deduplication of encrypted data with proof of ownership in cloud storage," IEEE/CIC International Conference on Communications in China, 2013.

[9] J. Li, X. Chen, M. Li, J. Li, P. P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management," IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, 2014.

[10] J. Li, Y. K. Li, X. Chen, P. Lee, and W. Lou, "A hybrid cloud approach for secure authorized deduplication," Parallel & Distributed Systems IEEE Transactions on, vol. 26, no. 5, pp. 1206–1216, 2015.