ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Proxy Re-Encryption (Pre) Based Deduplication Scheme on Encrypted Big Data for Cloud

A.Shyamala Devi[1], Dr. V. Sai Shanmuga Raja[2]

M.E. Computer Science and Engineering, Shanmuganathan Engineering College, Arasampatti, Pudukkottai, India.

Assistant Professor, Department of Computer Science and Engineering, Shanmuganathan Engineering College, Arasampatti, Pudukkottai, India.

ABSTRACT: In recent years Cloud computing is the most important and required service to offer various support to its client and different assets over the Internet. The most vital and prevalent cloud Service is information stockpiling. Keeping in mind the end goal to save the security of information holders, information are frequently put away in cloud in a scrambled frame. In any case, encoded information present new difficulties for cloud information deduplication, which gets to be distinctly vital for Big Data stockpiling and handling in cloud. Conventional deduplication plans can't take a shot at scrambled information. Existing arrangements of scrambled information deduplication experience the ill effects of security shortcoming. They can't adaptable bolster information get to control and renouncement. Hence, few of them can be promptly sent by and by. In this paper, we propose a plan to deduplicate encoded information put away in cloud in light of proprietorship test and intermediary re-encryption. It coordinates cloud information deduplication with get to control. We assess its execution in view of broad investigation and PC recreations. The outcomes demonstrate the unrivaled proficiency and adequacy of the plan for potential commonsense arrangement, particularly for enormous information deduplication in Cloud Storage.

KEYWORDS: Big Data, Cloud Computing, Access Control, Proxy Re-Encryption, Deduplication.

I. INTRODUCTION

Cloud computing offers another method for Information Technology benefits by improving different assets (e.g., capacity, figuring) and giving them to clients in view of their requests. Cloud computing gives a major asset pool by connecting system assets together. It has alluring properties, for example, adaptability, flexibility, adaptation to internal failure, and pay-per-utilize. In this way, it has turned into a promising administration stage.

The most essential and well known cloud administration is information stockpiling administration. Cloud clients transfer individual or classified information to the server farm of a Cloud Service Provider (CSP) and permit it to keep up these information. Since interruptions and assaults towards touchy information at CSP are not avoidable, it is reasonable to expect that CSP can't be completely trusted by cloud clients. Also, the loss of control over their very own information prompts to big data security dangers, particularly information protection spillages. Because of the quick improvement of information mining and different examination innovations, the security issue gets to be distinctly genuine. Subsequently, a great practice is to just outsource encoded information to the cloud with a specific end goal to guarantee information security and client protection. Be that as it may, the same or diverse clients may transfer copied information in scrambled frame to CSP, particularly for situations where information are shared among numerous clients. Despite the fact that Cloud storage space is immense, information duplication extraordinarily squanders organize assets, devours a great deal of vitality, and entangles information administration. The advancement of various administrations additionally makes it pressing to convey productive asset administration instruments. Thus, deduplication gets to be distinctly basic for big data stockpiling and handling in the cloud.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Deduplication has demonstrated to accomplish high cost reserve funds, e.g., diminishing up to 90-95 percent stockpiling requirements for reinforcement applications and up to 68 percent in standard record frameworks. Clearly, the reserve funds, which can be passed back straightforwardly or in a roundabout way to cloud clients, are critical to the financial aspects of cloud business. Step by step instructions to oversee scrambled information stockpiling with deduplication in a proficient way is a reasonable issue. Be that as it may, current modern deduplication arrangements can't deal with encoded information. Existing answers for deduplication experience the ill effects of animal drive assaults. They can't adaptably bolster information get to control and renouncement in the meantime. Most existing arrangements can't guarantee unwavering quality, security and protection with sound execution.

Fig.1 System Architecture Design

II. EXISTING SYSTEM

In tha past systems, it is hard to allow data holders to manage deduplication due to a number of reasons. First, data holders may not be always online or available for such a management, which could cause storage delay. Second, deduplication could become too complicated in terms of communications and computations to involve data holders into deduplication process. Third, it may intrude the privacy of data holders in the process of discovering duplicated data. Forth, a data holder may have no idea how to issue data access rights or deduplication keys to a user in some situations when it does not know other data holders due to data super-distribution. Therefore, CSP cannot cooperate with data holders on data storage deduplication in many situations.

III. PROPOSED SYSTEM

In the proposed approach, a scheme based on data ownership challenge and Proxy Re-Encryption (PRE) is implemented to manage encrypted data storage with deduplication. We aim to solve the issue of deduplication in the situation where the data holder is not available or difficult to get involved. Meanwhile, the performance of data deduplication in our scheme is not influenced by the size of data, thus applicable for big data. Specifically, the contributions of this paper can be summarized as below: We motivate to save cloud storage and preserve the privacy of data holders by proposing a scheme to manage encrypted data storage with deduplication. Our scheme can flexibly support data sharing with deduplication even when the data holder is offline, and it does not intrude the privacy of data holders.

- We propose an effective approach to verify data ownership and check duplicate storage with secure challenge and big data support.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

- We integrate cloud data deduplication with data access control in a simple way, thus reconciling data deduplication and encryption.
- We prove the security and assess the performance of the proposed scheme through analysis and simulation. The results show its efficiency, effectiveness and applicability.

IV. ENCRYPTED DATA DEDUPLICATION

Cloud storage specialist organizations, for example, Dropbox, Google Drive, Mozyf, and others perform deduplication to spare space by just putting away one duplicate of every document transferred. Be that as it may, if customers routinely encode their information, stockpiling reserve funds by deduplication are completely lost. This is on account of the Encrypted information are spared as various substance by applying distinctive encryption keys. Existing modern arrangements flop in encoded information deduplication. For instance, DeDu is a proficient deduplication framework, however it can't deal with encoded information. Accommodating deduplication and customer side encryption is a dynamic research subject. Message-Locked Encryption plans to take care of this issue.

The most noticeable appearance of MLE is Convergent Encryption (CE), presented by Douceur et al. furthermore, others. CE was utilized inside a wide assortment of business and research stockpiling administration frameworks. Giving M a chance to be a document's information, a customer first figures a key K HðMÞ by applying a cryptographic hash work H to M, and afterward registers ciphertext C EðK;MÞ by means of a deterministic symmetric encryption plot. A moment customer B encrypting a similar record M will deliver a similar C, empowering deduplication. Be that as it may, CE is liable to a natural security confinement, to be specific, defenselessness to disconnected savage constrain lexicon assaults. Realizing that the objective information M hidden the objective ciphertext C is drawn from a lexicon S ¼ fM1; . . .;Mng of size n, an assailant can recoup M in the ideal opportunity for n ¼ jSj disconnected encryptions: for every i ¼ 1; . . . ; n, it basically CEencrypts Mi to get a ciphertext indicated as Ci and returns Mi with the end goal that C ¼ Ci.

This works since CE is deterministic and keyless. The security of CE is just conceivable when the objective information is drawn from a space too substantial to deplete. Another issue of CE is that it is not adaptable to bolster information get to control by information holders, particularly for information renouncement prepare, since it is incomprehensible for information holders to create the same new key for information re-encryption.

A picture deduplication plot embraces two servers to accomplish irrefutability of deduplication. The CE-based plan depicted in joins record substance and client benefit to acquire a document token with token unforgeability. Notwithstanding, both plans specifically encode information with a CE key, accordingly experience the ill effects of the issue as portrayed previously. To oppose the assault of control of information identifier, Meye et al. proposed to embrace two servers for intra-client deduplication and interdeduplication.

The ciphertext C of CE is further encoded with a client key and exchanged to the servers. Be that as it may, it doesn't manage information sharing after deduplication among various clients. ClouDedup additionally intends to adapt to the natural security exposures of CE, yet it can't settle the issue brought about by information cancellation. An information holder that expels the information from the cloud can in any case get to similar information since despite everything it knows the information encryption key if the information is not totally expelled from the cloud. Bellare et al. proposed DupLESS that gives secure deduplicated stockpiling to oppose beast constrain assaults. In Dup-LESS, a gathering of associated customers (e.g., organization workers) encode their information with the guide of a Key Server (KS) that is separate from a Storage Service (SS).

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

TABLE 1
System Notation

Key	Description
$(pk_i,\ sk_i)$	The public key and secret key of user u_i for PRE
DEK_i	The symmetric key of u_i
$H()$	The hash function
CT	The ciphertext
CK	The cipherkey
M	The user data
KG	The key generation algorithm of PRE
RG	The re-encryption key generation algorithm of PRE
E	The encryption algorithm of PRE
R	The re-encryption algorithm of PRE
D	The decryption algorithm of PRE
$Encrypt(DEK, M)$	The symmetric encryption function on M with DEK
$Decrypt(DEK, CT)$	The symmetric decryption function on CT with DEK
$(V_i,\ s_i)$	The key pair of user u_i in Eillptic Curve Cryptography (ECC) for duplication check
P	The base point in ECC
$x = H(H(M) * P)$	The token used for data duplication check

Customers verify themselves to the KS, yet don't release any data about their information to it. For whatever length of time that the KS stays out of reach to assailants, high security can be guaranteed. Clearly, Dup-LESS can't control information access of other information clients adaptable. Then again, a strategy based deduplication intermediary plan was proposed however it didn't consider copied information administration (e.g., erasure and proprietor administration) and did not assess plot execution. As expressed, dependability, security and protection ought to be taken into contemplations when planning a deduplication framework.

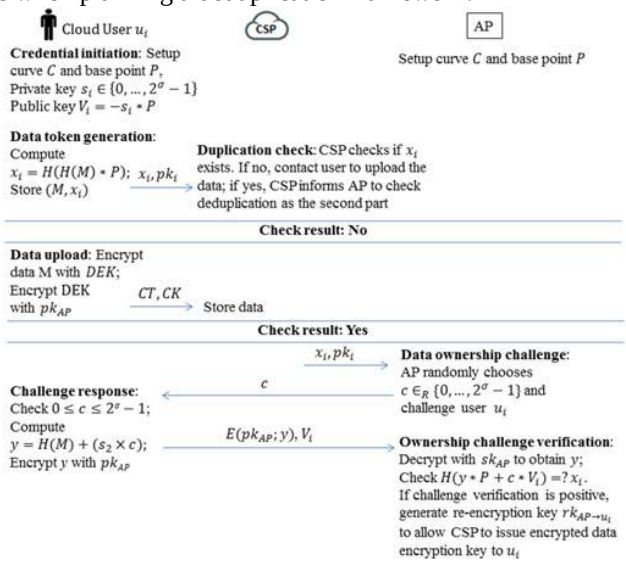

Fig.2. Data Ownership Verification.

The strict dormancy necessities of essential stockpiling lead to the attention on disconnected deduplication frameworks. Late reviews proposed procedures to enhance reestablish execution. Fu et al. proposed History-Aware

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Rewriting (HAR) calculation to precisely distinguish and rework divided lumps, which enhanced the reestablish execution. Kaczmarczyk et al. concentrated on between adaptation duplication and proposed Context-Based Rewriting (CBR) to enhance the reestablish execution for most recent reinforcements by moving discontinuity to more established reinforcements.

Another work even proposed to relinquish deduplication to diminish the piece discontinuity by holder topping. In our past work, we proposed utilizing PRE for cloud information deduplication. This plan applies the hash code of information to check proprietorship with mark confirmation, which is shockingly unreliable if HðMÞ is unveiled to a malignant client. In this paper, we propose another proprietorship confirmation way to deal with enhance our past work and intend to bolster huge information deduplication in a proficient way.

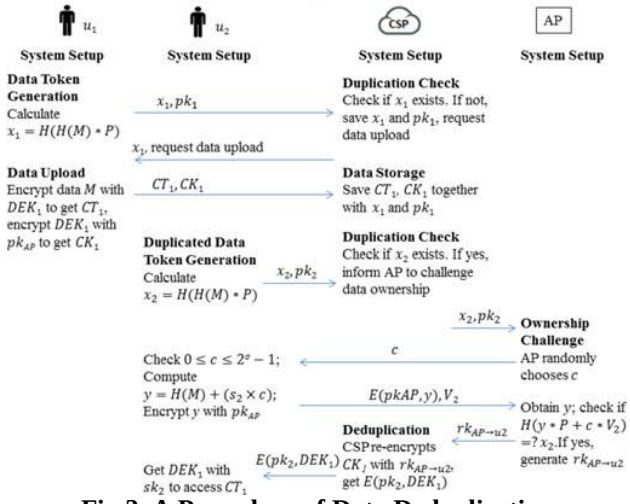

Fig.3. A Procedure of Data Deduplication

Algorithm 1: Grant Access to Duplicated Data

Input: pkj, PolicyðuiÞ, PolicyðAPÞ

- CSP requests AP to challenge ownership and grant access to duplicated data for uj by providing pkj.
- After ensuring data ownership through challenge, AP checks PolicyðAPÞ and issues CSP rkAP!ui ¼ RG ðpkAP ; skAP ; pkjÞ if the check is positive.
- CSP transforms EðpkAP ;DEKiÞ into EðpkjDEKiÞ if PolicyðuiÞ authorizes uj to share the same data M encrypted by DEKi: RðrkAP!ui ;EðpkAP ;DEKÞÞ ¼ EðpkjDEKiÞ.

Note: rkAP!ui calculation can be skipped if it has been executed already and the value of (pkj; skj) and (pkAP ; skAP) remain unchanged.

- Data holder uj obtains DEKi by decrypting EðpkjDEKiÞ with skj: DEKi ¼ DðskjEðpkjDEKiÞÞ, and then it can access dataM at CSP.

V. LITERATURE SURVEY

In the year of 2013 the authors "M. Bellare, S. Keelveedhi, and T. Ristenpart" described into their paper titled "DupLESS: Server aided encryption for deduplicated storage" such as Cloud storage service providers such as Dropbox, Mozy, and others perform deduplication to save space by only storing one copy of each file uploaded.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

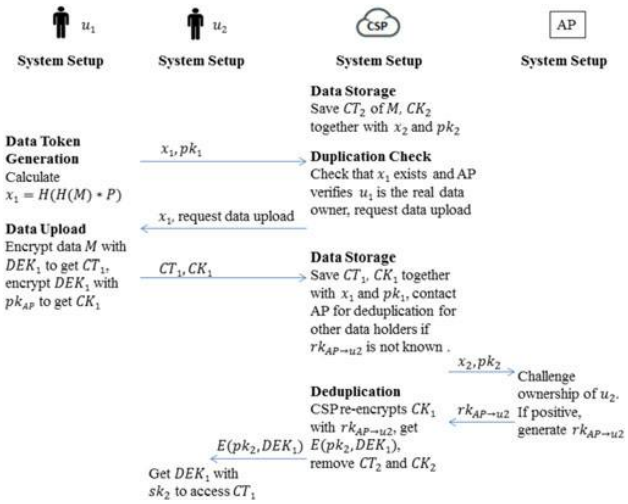

Fig.4. A Procedure of Data Owner Management

Should clients conventionally encrypt their files, however, savings are lost. Message-locked encryption (the most prominent manifestation of which is convergent encryption) resolves this tension. However it is inherently subject to brute-force attacks that can recover files falling into a known set. We propose an architecture that provides secure deduplicated storage resisting brute-force attacks, and realize it in a system called DupLESS. In DupLESS, clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol.

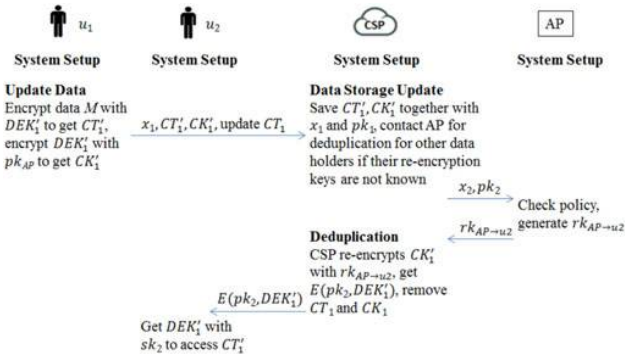

Fig.5. A Procedure of Encrypted Data Update

It enables clients to store encrypted data with an existing service, have the service perform deduplication on their behalf, and yet achieves strong confidentiality guarantees. We show that encryption for deduplicated storage can achieve performance and space savings close to that of using the storage service with plaintext data.

In the year of 2012 the authors J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer described into their paper titled "Reclaiming space from duplicate files in a serverless distributed file system" such as the Farsite distributed file system provides availability by replicating each file onto multiple desktop computers. Since this replication consumes significant storage space, it is important to reclaim used space where possible. Measurement of over 500 desktop file systems shows that nearly half of all consumed space is occupied by duplicate files. We present a mechanism to reclaim space from this incidental duplication to make it available for controlled file replication.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

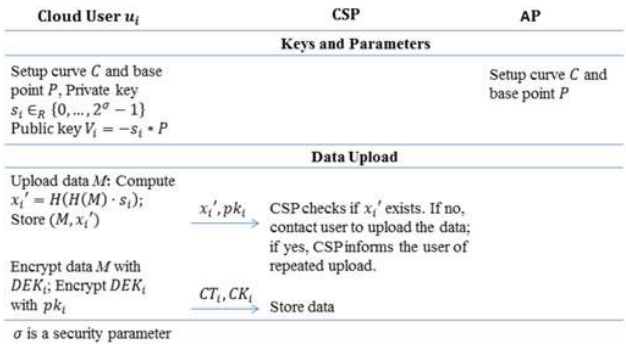

Fig.6. A Procedure of Valid Replicated Data Upload

Our mechanism includes 1) convergent encryption, which enables duplicate files to coalesced into the space of a single file, even if the files are encrypted with different users' keys, and 2) SALAD, a Self- Arranging, Lossy, Associative Database for aggregating file content and location information in a decentralized, scalable, fault-tolerant manner. Large-scale simulation experiments show that the duplicate-file coalescing system is scalable, highly effective, and fault-tolerant.

In the year of 2012 the author G. Wallace, et al described into his paper titled "Characteristics of backup workloads in production systems" such as Data-protection class workloads, including backup and long-term retention of data, have seen a strong industry shift from tape-based platforms to disk-based systems. But the latter are traditionally designed to serve as primary storage and there has been little published analysis of the characteristics of backup workloads as they relate to the design of disk-based systems. In this paper, we present a comprehensive characterization of backup workloads by analyzing statistics and content metadata collected from a large set of EMC Data Domain backup systems in production use.

This analysis is both broad (encompassing statistics from over 10,000 systems) and deep (using detailed metadata traces from several production systems storing almost 700TB of backup data). We compare these systems to a detailed study of Microsoft primary storage systems [22], showing that backup storage differs significantly from their primary storage workload in the amount of data churn and capacity requirements as well as the amount of redundancy within the data.

TABLE 2
Computation Complexity of System Entities by Comparing with [33]

Entity	Algorithm	Computations [our scheme]	Computations [33]	Complexity
Data Owner	Setup	1*PointMulti + 1* Exp	1*ModInv + 1*Exp	$\mathcal{O}(1)$
	Data upload	2 *Exp + 1*PointMulti	3*Exp	
CSP	Re-encryption	1 *Pair	1*Pair	$\mathcal{O}(n)$
Data Holder	System Setup	1*PointMulti + 1*Exp	1*ModInv + 1* Exp	$\mathcal{O}(1)$
	Challenge Response	2*Exp + 1*PointMulti	–	
	Data upload	–	3*Exp	
	Decryption for access	1*Exp	1*Exp	
AP	System Setup	1*Exp	1*Exp	$\mathcal{O}(n)$
	Ownership challenge and re-encryption key generation	2*Exp + 2*Point Multi	1*Exp	

These properties bring unique challenges and opportunities when designing a disk-based file system for backup workloads, which we explore in more detail using the metadata traces. In particular, the need to handle high churn while leveraging high data redundancy is considered by looking at deduplication unit size and caching efficiency.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

VI. CONCLUSION AND FUTURE SCOPE

Managing encrypted data with deduplication is important and significant in practice for achieving a successful cloud storage service, especailly for big data storage. In this system, we proposed a practical scheme to manage the encrypted big data in cloud with deduplication based on ownership challenge and PRE. Our scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption. Extensive performance analysis and test showed that our scheme is secure and efficient under the described security model and very suitable for big data deduplication. The results of our computer simulations further showed the practicability of our scheme. Futurework includes optimizing our design and implementation for practical deployment and studying verifiable computation to ensure that CSP behaves as expected in deduplication management.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Server aided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Secur., 2013, pp. 179–194.
[2] Dropbox, A file-storage and sharing service. (2016). [Online]. Available: http://www.dropbox.com
[3] Google Drive. (2016). [Online]. Available: http://drive.google.com
[4] Mozy, Mozy: A File-storage and Sharing Service. (2016). [Online]. Available: http://mozy.com/
[5] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system," in Proc. IEEE Int. Conf. Distrib. Comput. Syst., 2002, pp. 617–624, doi:10.1109/ICDCS.2002.1022312.
[6] G. Wallace, et al., "Characteristics of backup workloads in production systems," in Proc. USENIX Conf. File Storage Technol., 2012, pp. 1–16.
[7] Z. O. Wilcox, "Convergent encryption reconsidered," 2011. [Online]. Available: http://www.mailarchive.com/cryptography@metzdowd.com/msg08949.html
[8] G. Ateniese, K. Fu, M. Green, and S. Hohenberger, "Improved proxy re-encryption schemes with applications to secure distributed storage," ACM Trans. Inform. Syst. Secur., vol. 9, no. 1, pp. 1–30, 2006, doi:10.1145/1127345.1127346.
[9] Opendedup. (2016). [Online]. Available: http://opendedup.org/
[10] D. T. Meyer and W. J Bolosky, "A study of practical deduplication," ACM Trans. Storage, vol. 7, no. 4, pp. 1–20, 2012, doi:10.1145/2078861.2078864.
[11] J. Pettitt, "Hash of plaintext as key?" (2016). [Online]. Available: http://cypherpunks.venona.com/date/1996/02/msg02013.html
[12] The Freenet Project, Freenet. (2016). [Online]. Available: https://freenetproject.org/
[13] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. Cryptology—EUROCRYPT, 2013, pp. 296–312, doi:10.1007/978-3-642-38348-9_18.
[14] D. Perttula, B. Warner, and Z. Wilcox-O'Hearn, "Attacks on convergent encryption." (2016). [Online]. Available: http://bit.ly/yQxyvl
[15] C. Y. Liu, X. J. Liu, and L. Wan, "Policy-based deduplication in secure cloud storage," in Proc. Trustworthy Comput. Serv., 2013, pp. 250–262, doi:10.1007/978-3-642-35795-4_32.