Information Sciences 456 (2018) 159–173

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

DedupDUM: Secure and scalable data deduplication with dynamic user management

Haoran Yuan[a], Xiaofeng Chen[a,*], Tao Jiang[a], Xiaoyu Zhang[a], Zheng Yan[a], Yang Xiang[a,b]

[a] State Key Laboratory of Integrated Service Networks (ISN), Xidian University, Xi'an, P.R. China
[b] Digital Research & Innovation Capability Platform, Swinburne University of Technology, Hawthorn, Australia

ARTICLE INFO

Article history:
Received 23 January 2018
Revised 5 May 2018
Accepted 7 May 2018
Available online 8 May 2018

Keywords:
Data deduplication
Random convergent encryption
Dynamic user management
Access control
User joining

ABSTRACT

Data deduplication on cloud enables the cloud servers to store a cope of data and eliminate redundant one so that a goal to save storage space and network bandwidth is realized. Recently, many research works which are concerning to the privacy-preserving problem of dynamic ownership management in the secure data deduplication setting are published. However, to our knowledge, the existing schemes are not efficient when the cloud user joining and revocation frequently go on, especially in the absence of a trusted third party in practical cloud storage systems. In this paper, we propose a secure and scalable data deduplication scheme with dynamic user management, which updates dynamic group users in a secure way and restricts the unauthorized cloud users from the sensitive data owned by valid users. To further mitigate the communication overhead, the pre-verified accessing control technology is adopted, which prevents the unauthorized cloud users from downloading data. In other words, our present scheme also ensures that only the valid cloud users are able to download and decrypt the ciphertext from the cloud server. All this reduces the communication overhead in our scheme implementation.

1. Introduction

The adventure of cloud storage attracts a souring number of users and enterprises to store their sensitive data or database to the remote cloud server [1,7–10,18,28,29,38,42–44]. In the year 2015, Center for Democracy & Technology (CDT) issued a report that the size of the total data generated would surpass 7.9 zettabytes (ZB) [34]. Similarly, Internet Data Center (IDC) also claimed that the digital universe would grow double in size every two years, and 40 trillion gigabytes is expected in 2020 (more than 5200 gigabytes for every man, woman, and child) [17]. The fact that the data volume is increasingly growing posts a threat for the functionality of cloud server, it is essential to equip with substantial disk space and bandwidth. On the other hand, the cloud server may store abundant data, for instance, some movies or music files are stored by different cloud users, which wastes a large amount of storage space and backup space. To overcome this problem, deduplication methods over unencrypted data are presented and have attracted much attention in the past years [31]. Nowadays, deduplication techniques have been widely used in the cloud server, e.g., referred to Wuala [40], Mozy [32], Dropbox [13], and Google Drive [12]. It is seen that 90 percent of business applications disk storage space and bandwidth are saved [14].

* Corresponding author.
E-mail address: xfchen@xidian.edu.cn (X. Chen).

https://doi.org/10.1016/j.ins.2018.05.024
0020-0255/

Despite plenty of benefits, some new security challenges are also in the appearance, especially, the security of the users' data. Since the cloud server is assumed to be honest-but-curious, it may as well try to infer and analyze the outsourced data. In this situation, to protect the privacy of their sensitive data, the cloud users have to encrypt their data before outsourcing to the cloud server. But, in general different users encrypt the same data with themselves' encryption keys , which leads identical data output different ciphertexts, as well as deduplication unachievable.

Convergent encryption (CE) provides a feasible solution to protect the privacy of data and realize deduplication [11]. The main trick is as follows. It encrypts and decrypts sensitive files with convergent keys through computing the hash value of the file. After encrypting the files, the cloud user only keeps the encryption key and outsources the ciphertext to the cloud server for saving storage. Since the hash value of the same file is deterministic and same, the same file derives the same convergent key, then the same file is determinately encrypted to the same ciphertext. This allows the cloud server to perform deduplication. Unfortunately, convergent encryption scheme has the tag consistency problem. To crack it, Bellare et al. [3] formalized the notion of message-locked encryption and a randomized convergent encryption (RCE) scheme was proposed. However, this scheme also witnesses some security flaws in ownership revocation. If cloud users who keep their encryption keys revoked, they are able to decrypt the corresponding data. In the problem of dynamic ownership management, it has become a research focus that how to restrict unauthorized cloud users from the sensitive data [33]. Recently, Hur et al. presented a secure data deduplication scheme focusing on dynamic ownership management in cloud storage (SD-DOM) [20]. This scheme considered the problems of key updating and dynamic ownership management problems. But, some limitations are seen. In the stage of key generation, the cloud server fixes the maximum number of cloud users beforehand and sets a binary KEK (key-encrypting key) tree for the universe of cloud users. Therefore, when a few more cloud users join the cloud and the size further enlarges over the limited number this scheme cannot be adaptive. For detailed analysis of this scheme one is suggested in Section 3.

Although some existing schemes involved to data deduplication are able ro support both cloud user revocation and new cloud user joining, they required a fully trusted third party or cloud users for abundant computing power [20,39,41]. In addition, the cloud server cannot distinguish whether the cloud user is authorized or not before the data is downloaded. In other words, the cloud server cannot confirm that whether the cloud user belongs to the group or not and possesses decryption ability. Besides, the malicious attackers are able to download ciphertexts even though the decryption is restricted, which causes enormous communication cost for the cloud server. To the best of our knowledge, these schemes cannot perfectly support cloud user joining and verify cloud users' identity before the data downloading is completed, especially the fully trusted third party is absent in practical cloud storage systems. In this paper, we attempt to solve those above problems.

Our Contribution. In this paper, we further revisit the problem of dynamic user revocation and new cloud user joining over encrypted data deduplication. Our contributions are three folds:

- We propose a data deduplication scheme with dynamic user management which supports cloud user revocation and new cloud user joining by exploiting the re-encryption techniques.
- Compared with the existing schemes [20,39,41], our scheme does not require a fully trusted third party and the cloud user for excessive computational overhead.
- Our scheme uses the access control technique to verify the validity of the cloud users before they download data. Only when the cloud users are in group, will the cloud server send ciphertext to the cloud user. Therefore, the abundant communication cost will be reduced.

1.1. Related work

Convergent encryption (CE) provides the first clever solution for deduplication over encrypted data [11]. In this scheme, the data owner obtains the encryption key by hashing the data. Then, the user uses the encryption key to encrypt data and gets ciphertext. The same data is always deterministically encrypted with same secret key, realizing the data deduplication over encrypted data. However, the CE scheme is vulnerable to the tag consistency problem. To solve this problem, Bellare et al. proposed the randomized convergent encryption (RCE) scheme [3], which is an implementation of message-locked encryption and uses the additional tag checking mechanism to guarantee the integrity of users' data. Bellare et al. [2] proposed the DupLESS scheme with secure deduplicated storage to resist brute-force attacks. Jin et al. [22] proposed an anonymous data deduplication scheme based on proxy re-encryption algorithm. To share the convergent key among the different cloud servers, Li et al. [24] realized the convergent key management by employing security Ramp secret sharing scheme [5]. Based on the predicate encryption, Shin and Kim [35] constructed an equality predicate encryption scheme that realized the deduplication over encrypted data. However, this approach is only efficient in the case of the single user deduplication. Chen et al. [6] designed a block-level message-locked encryption (BL-MLE) scheme for secure large file deduplication [37]. Wen et al. [39] proposed a session-key-based convergent key management scheme (SKC) and a convergent key sharing scheme (CKS), both of which can securely and dynamically update in data deduplication. In the SKC scheme, it is difficult to change the session key and replace the encrypted convergent key. In the CKS scheme, abundant computing power is required. Li et al. [26] augmented the CAONT [27] and designed a rekeying-aware encrypted deduplication scheme which realized the secure and lightweight rekeying. They further extended rekeying-aware encrypted deduplication scheme with dynamic access control by integrating the primitives of CP-ABE [4] and key regression [16]. Based on the ownership challenge and proxy