Journal of Computer and System Sciences 81 (2015) 1532–1541

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Secure deduplication storage systems supporting keyword search

Jin Li [a,b,*,1], Xiaofeng Chen [b], Fatos Xhafa [c], Leonard Barolli [d]

[a] School of Computer Science and Educational Software, Guangzhou University, Guangzhou 510006, PR China
[b] State Key Laboratory of Integrated Service Networks (ISN), Xidian University, Xi'an 710071, PR China
[c] Department of Languages and Informatics Systems, Technical University of Catalonia, Spain
[d] Department of Information and Communication Engineering, Fukuoka Institute of Technology, Japan

ARTICLE INFO

Article history:
Received 1 July 2014
Received in revised form 14 December 2014
Accepted 14 December 2014
Available online 13 January 2015

Keywords:
Data deduplication
Outsourcing
Privacy
Keyword search

ABSTRACT

Data deduplication is an attractive technology to reduce storage space for increasing vast amount of duplicated and redundant data. In a cloud storage system with data deduplication, duplicate copies of data will be eliminated and only one copy will be kept in the storage. To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. However, the issue of keyword search over encrypted data in deduplication storage system has to be addressed for efficient data utilization. This paper firstly proposes two constructions which support secure keyword search in this scenario. In these constructions, the integrity of the data can be realized by just checking the convergent key, without other traditional integrity auditing mechanisms. Then, two extensions are presented to support fuzzy keyword search and block-level deduplication. Finally, security analysis is given.

1. Introduction

Data deduplication is a technique for reducing the size of data in storage systems by detecting and eliminating redundant data and has been widely employed for data backup to minimize network and storage overhead. The system will perform duplicates for incoming data and does not record the data if it matches the existing stored data and refers other redundant data to that copy.

There have been a number of deduplication systems proposed with various deduplication strategies such as file-level or block-level deduplications depending on the size of deduplication. It can also be divided into inline deduplication system and post-process deduplication system, which performs deduplication before or after storing them respectively. Especially, with the advent of cloud storage, data deduplication techniques attract more and more attention from both academic and industrial community. It has become one of critical techniques for the management of the ever-increasing volume of data in the information society. According to the analysis report of IDC, the volume of data in the world is expected to reach 40 trillion gigabytes in 2020 [2] with very high increasing speed. Today's commercial cloud storage services, such as Dropbox,

* Corresponding author at: School of Computer Science and Educational Software, Guangzhou University, Guangzhou 510006, PR China.
E-mail address: jinli71@gmail.com (J. Li).

[1] A preliminary version of this paper has been presented at AINA2014 [1].

http://dx.doi.org/10.1016/j.jcss.2014.12.026
0022-0000/

Mozy, and Memopal, have been applying deduplication to save the network bandwidth and the storage cost with client-side deduplication.

To protect the confidentiality of outsourced data, the notion of convergent encryption [3] has been proposed. With the convergent encryption, the confidentiality of data can be achieved while realizing deduplication. In the deduplication system based on convergent encryption, the data will be encrypted/decrypted with a convergent key which is derived by computing the hash value of the content of data copy itself [3–5]. With this novel technique, identical data copies will generate the same convergent key and be encrypted into the same ciphertext, which allows the cloud to perform deduplication on the ciphertexts.

However, data encryption makes effective data utilization to be difficult given that there could be a large amount of outsourced data files. The keyword search over encrypted data in cloud computing has been proposed in recent years [6–12] and popularly used to selectively retrieve files of interest. Through this technique, both the data and keyword are encrypted such that privacy of keyword information will be also achieved besides the confidentiality of data. In a searchable encryption scheme, an index will be built by the data owner for each file or each keyword, which is then uploaded to the cloud server with the encrypted data. By integrating the trapdoors of keywords within the index information, effective keyword search can be realized while both file content and keyword privacy are well-preserved.

Although such a technique allows the cloud server to perform search on behalf of user without knowing the information of the file or the keywords, the existing searchable encryption techniques have not been considered in a deduplication storage system with additional duplicate check and deduplication storage phases.

In this paper, we show how to construct secure deduplication systems with keyword search in cloud computing. We also formalize the security model for the encrypted keyword search in deduplication systems. Both the keyword privacy and content privacy can be achieved in the constructions. In our constructions, two kinds of index for keyword search have been proposed and analyzed. The integrity of the uploaded files could be verified by the users after downloading and decryption without any other traditional integrity auditing mechanisms. Such a property is achieved based on the technique of convergent encryption. The security analysis is also given under the proposed security model.

This paper is organized as follows. In Section 2, we present the system model and security requirements of the keyword search in deduplication storage system. In Section 3, the building blocks required in this paper are presented. Our constructions are described in Section 4 and Section 5. The security and performance analysis is given in Section 6. The related work is described in Section 7. Finally, we draw conclusions in Section 8.

2. Problem formulation

2.1. System model

In this paper, we consider a cloud storage system consisting of data owner, data user and cloud storage server. The files are assumed to be encrypted by the data owner before uploading to the cloud storage server. We assume the authorization between the data owner and users is appropriately done with some authentication and key-issuing protocols. After uploading the encrypted files, authorized users are allowed to perform the keyword search on these encrypted data with the aid of cloud server. In more details, an authorized user sends a request to selectively retrieve data files of his/her interest. Upon receiving the request, the cloud storage server is responsible for pinpointing a set of files matching the searching request. All the matched file ID will be returned to the user as the searching result. Both client-side deduplication and server-side deduplication are supported in our system.

- *Data owner*. The data owner is an entity that outsources data storage to the storage server and access the data later. In a client-side data deduplication system, only the first data owner of a file needs to upload while the following data owners of the same file does not require to upload the duplicate data any more.
- *Users*. The entity of users in the deduplication systems makes registration at cloud storage server and has privileges to access some data files shared by some data owners.
- *Cloud storage server.* The cloud storage server is an entity that provides the data storage service for the users. Furthermore, the cloud storage server will also perform duplicate check before users upload their files. If there is identical content stored in cloud storage server, the users are not required to upload the file again, which can reduce the storage cost at the server side and save the upload bandwidth at user side.

Both file-level and block-level deduplications are supported in our system. To upload a file, a user first performs the file-level duplicate check. If the file is a duplicate, then all its blocks must be duplicates as well; otherwise, the user further performs the block-level duplicate check and identifies the unique blocks to be uploaded. Each data copy is associated with a tag for the duplicate check, which is computed with a hash function on the data file.

2.2. Adversary model

We consider a semi-trusted cloud storage server in our deduplication storage system. To prevent the cloud storage server from getting sensitive information, data files are encrypted before uploading by the data owners. The cloud server will