Chapter 16
Secure Deduplication with Encrypted Data for Cloud Storage

Pasquale Puzio
SecludIT, France & EURECOM, France

Melek Önen
EURECOM, France

Refik Molva
EURECOM, France

Sergio Loureiro
SecludIT, France

ABSTRACT

With the continuous increase of the number of users and the size of their data, data deduplication becomes a necessity for cloud storage providers. By storing a unique copy of duplicate data, cloud providers greatly reduce their storage and data transfer costs. The advantages of deduplication unfortunately come with a high cost in terms of new security and privacy challenges. In this chapter we propose ClouDedup, a secure and efficient storage service which assures block-level deduplication and data confidentiality at the same time. Although ClouDedup is based on convergent encryption, it remains secure thanks to the definition of a component that implements an additional encryption operation. Furthermore, as the requirement for deduplication at block-level raises an issue with respect to key management, we suggest to include a new component in order to implement the key management for each block together with the actual deduplication operation. In this chapter we show how we have implemented the proposed architecture, the challenges we have met and our solutions to these challenges.

CASE DESCRIPTION

Introduction

With the potentially infinite storage space offered by cloud providers, users tend to use as much space as they can and vendors constantly look for techniques aimed to minimize redundant data and maximize space savings. A technique which has been widely adopted is cross-user deduplication. The simple idea behind deduplication is to store duplicate data (either files or blocks) only once. Therefore, if a user wants to upload a file (block) which is already stored, the cloud provider will add the user to the owner list of that file (block) only. Deduplication has proved to achieve high

DOI: 10.4018/978-1-4666-8210-8.ch016

space and cost savings and many cloud storage providers are currently adopting it. Deduplication can reduce storage needs by up to 90-95% for backup applications (Opendedup, 2014) and up to 68% in standard file systems (Meyer, 2012).

Along with low ownership costs and flexibility, users require the protection of their data and confidentiality guarantees through encryption.

Unfortunately, deduplication and encryption are two conflicting technologies. While the aim of deduplication is to detect identical data segments and store them only once, the result of encryption is to make two identical data segments indistinguishable after being encrypted. This means that if data are encrypted by users in a standard way, the cloud storage provider cannot apply deduplication since two identical data segments will be different after encryption. On the other hand, if data are not encrypted by users, confidentiality cannot be guaranteed and data are not protected against curious cloud storage providers.

A technique which has been proposed to meet these two conflicting requirements is convergent encryption (Douceur, Adya, Bolosky, Simon, & Theimer, 2002) whereby the encryption key is usually the result of the hash of the data segment. Although convergent encryption seems to be a good candidate to achieve confidentiality and deduplication at the same time, it unfortunately suffers from various well-known weaknesses (Bellare, Keelveedhi, & Ristenpart, 2013)(Perttula, 2008) including dictionary attacks: an attacker who is able to guess or predict a file, can easily derive the potential encryption key and verify whether the file is already stored at the cloud storage provider or not.

We cope with the inherent security exposures of convergent encryption and propose ClouDedup, which preserves the combined advantages of deduplication and convergent encryption. The security of ClouDedup relies on its new architecture whereby in addition to the basic storage provider, a metadata manager and an additional gateway are defined: the gateway adds an additional encryption layer to prevent well-known attacks against convergent encryption and thus protect the confidentiality of the data; on the other hand, the metadata manager is responsible for the key management task since block-level deduplication requires the memorization of a huge number of keys: we define an efficient key management mechanism to avoid users to store one key per block.

To summarize the key features of ClouDedup:

- ClouDedup assures block-level deduplication and data confidentiality while coping with weaknesses raised by convergent encryption. Block-level deduplication renders the system more flexible and efficient;
- ClouDedup preserves confidentiality and privacy even against potentially malicious cloud storage providers thanks to an additional layer of encryption;
- ClouDedup offers an efficient key management solution through the metadata manager;
- The new architecture defines several different components and a single component cannot compromise the whole system without colluding with other components;
- ClouDedup works transparently with existing cloud storage providers. As a consequence, ClouDedup is fully compatible with standard storage APIs and any cloud storage provider can be easily integrated in our architecture.

BACKGROUND

Deduplication

According to the data granularity, deduplication (for a practical example see Figure 1) strategies can be classified into two main categories: file-level deduplication (Wilcox-O'Hearn & Warner, 2008) and block-level deduplication (Cox, Murray,