Chapter 2
Secure Data Deduplication on Cloud Storage

Shivansh Mishra
Indian Institute of Technology Varanasi (BHU), India

Surjit Singh
https://orcid.org/0000-0002-2386-7729
National Institute of Technology Kurukshetra, India

ABSTRACT

Deduplication is the process of removing duplicate data by storing only one copy of the original data and replacing the others simply as a reference to the original. When data is stored on cloud storage, client-side deduplication helps in reducing storage and communications overheads both from the client as well as the server perspective. Secure deduplication is the practice by which the data stored on the cloud is secured from external influences such that the clients maintain the privacy of their data, and the server also gets to take advantage of deduplication. This is done by encrypting the data using different schemes into ciphertext, which makes sense only to the original client. The schemes created for secure deduplication on cloud storage provide a solution to the problem of duplication detection in encrypted ciphertext. This chapter provides a brief overview of secure deduplication used on cloud storage along with the issues encountered during its implementation. The chapter also includes a literature review and comparison of some deduplication techniques.

INTRODUCTION

In the modern world there have been several technologies which have inherently changed the way we interact with the digital ecosystem. Of these, cloud computing has been one of the most influential ones. The simple fact that we can store our data or off load processing to an off-site location instead of actually using our local devices has been revolutionary. This has increased the possibilities through which we can use data stored on cloud across various devices and for various purposes. This trend of off-loading storage and computing capacity to third party sources does not seem to be stopping anytime soon. Rather

DOI: 10.4018/978-1-5225-7335-7.ch002

with the advent of technologies like IoT (Internet of Things), the sheer amount of cloud storage required is only projected to increase at a mammoth pace. According to a recent report the total volume of data stored will double every two years until 2020 (Gantz & Reinsel, 2012) and more than 75% of the data produced is considered to be duplicated (Reinsel & Gantz, 2010). Hence huge savings can be achieved by identifying this duplicate data and deduplication is a possible solution to this situation.

Data deduplication is the process which looks for redundant sequences of data across different comparison windows. The first unique version of a data object is stored and the other duplicates are just referenced to the original data object rather than stored again (as shown in Figure 1). This process is completely hidden from users and applications trying to retrieve the stored data. But the process of comparing two similar data objects byte by byte is very cumbersome. Hence, for deduplication the first step is to create a data fingerprint for each object that is written to the storage device. This fingerprint should be a unique identity key to the data object. Also there is an additional requirement that generating and comparing such fingerprints should not be too difficult. When new data comes which has to be written to the device, the fingerprint of the new data are matched with the ones of data objects which have already been written to storage. All duplicate data copies except the first one are not written to the actual storage but are just referenced as pointers to the location of first unique copy. If a previously unseen data object is encountered – one whose fingerprint doesn't match any others on the storage device – the full data object is written to the storage. Sometimes hashes are used as data fingerprints. Different data structures are used to perform matching of hashes as the sheer number of data objects which are processed is very high. Hence both the hash generation as well as the hash matching schemes are optimized to provide the most efficient results. The deduplication that is performed on cloud storage has an additional criteria to be secure i.e., the cloud storage provider should not be privy to the actual plaintext that is being stored on the client. Also since spoofing attacks are very common on internet, care should be taken that the mechanism employed is immune to such attacks.

In this chapter we provide a brief overview of some important aspects of deduplication. We look at some of the advantages of deduplication and contrast them with some of its disadvantages. Some general

Figure 1. Figure shows result of deduplication on duplicate data such that total number of stored data blocks are reduced. Figure also shows the pointer based storage approach for efficient storage utilization where similar blocks are mainly referenced instead of being restored (here similar colour blocks signify similar data blocks in the same storage array).