Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 87 (2016) 38 – 43

Procedia Computer Science

Fourth International Conference on Recent Trends in Computer Science & Engineering

Chennai, Tamil Nadu, India

A Study on Deduplication Techniques over Encrypted Data

Akhila K[a]*,Amal Ganesh[a],Sunitha C[a]

[a]*Department of CSE,Vidya Academy Of Science and Technology,Thrissur 680501,India*

Abstract

In the current digital world, data is of prime importance for individuals as well as for organizations. As the amount of data being generated increases exponentially with time, duplicate data contents being stored cannot be tolerated. Thus, employing storage optimization techniques is an essential requirement to large storage areas like cloud storage. Deduplication is a one such storage optimization technique that avoids storing duplicate copies of data. Currently, to ensure security, data stored in cloud as well as other large storage areas are in an encrypted format and one problem with that is, we cannot apply deduplication technique over such an encrypted data. Thus, performing deduplication securely over the encrypted data in cloud appears to be a challenging task. Various methods that address this challenge are studied in this paper.

Keywords : Deduplication,Cloud storage,Convergent Encryption;

1. Introduction

With numerous benefits of cloud storage such as cost savings, accessibility, scalability etc., users around the world tend to shift their invaluable data to cloud storage. As the data generation rates are increasing, it is a tedious task for cloud storage providers to provide efficient storage. Cloud storage providers uses different techniques to improve storage efficiency and one of leading technique employed by them is deduplication, which claims to be saving 90 to 95% of storage [1],[2]. Data Deduplication technique evolved as an simple storage optimization technique in secondary then widely adapted in primary storage as well as larger storage areas like cloud storage area. Now, data deduplication is widely used by various cloud storage providers like Dropbox [3],Amazon S3 [4], Google Drive [5],etc. Data once deployed to cloud servers, its beyond the security premises of the data owner, thus most of them prefer to outsource their in an encrypted format. Data encryption by data owners eliminates cloud service providers chance of deduplicating it since encryption and deduplication techniques have conflicting strategies, i.e., data encryption with a key converts data into an unidentifiable format called cipher text thus

* Corresponding author. Tel.: +91 953 949 4768;
E-mail address:akhilak777@gmail.com

Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016
doi:10.1016/j.procs.2016.05.123

encrypting, even the same data, with different keys may result in different cipher texts , making deduplication less feasible .However, performing encryption is essential to make data secure, at the same time, performing deduplication is essential for achieving optimized storage. Therefore, deduplication and encryption need to work in hand to hand to ensure secure and optimized storage. Various techniques and approaches used for deduplication over encrypted data are studied in this paper.

2. Background

2.1. Deduplication

Deduplication is basically a compression technique for removing redundant data. Fig 1 explains the deduplication process before storing data onto memory. Deduplication can be categorized as file level deduplication and block level deduplication based on granularity. File level deduplication takes into account the entire file, thus even small update or append makes the file different from previous version of it and thereby reducing deduplication ratio. Where as in case of block level deduplication data blocks are considered for deduplication. Deduplication can further categorized based on location of deduplication i.e., as client side deduplication and as source side deduplication. Performing deduplication at client side is ensures bandwidth saving since only hash value of file is sent to server, if duplicate is existing [6] , [7].Deduplication is widely is used various applications like backup, metadata management, primary storage, etc. for storage optimization [8].

2.2. Convergent Encryption

Convergent encryption [2], is an encryption approach that support deduplication. With convergent encryption, encryption key is generated out of hash of plain text. Thus applying these technique identical plaintexts would produce same cipher text, and this helps in performing deduplication further.

2.3. Proof Of Ownership

Deduplication works by computing cryptographic hash function on to data and using this hash value to determine similar data. Once a duplicate copy is found then new data is not uploaded but pointer to file ownership is updated thus saving storage and bandwidth. When it comes to client side deduplication, hash values of data are computed at client and send for duplicate check. An attacker, who gains access to hash value of a data which not authorized to him/her, may claim deduplication of file and thereby gaining access to the file. To defend such an attack, a Proof Of Ownership (PoW) has been proposed in[10], and various works like[10],[11], etc adapted this method. PoW works as an interactive algorithm between two parties - a prover and verifier to prove the ownership of the file. Verifier computes a short value of data M whereas, a prover need to compute short value of M and send it to verifier for claiming ownership of M [9],[10].