IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 1, January 2016

A Cumulative Study to Prevent Duplication Of Data In Cloud

Budhavant Ashwini[1], Korkhele Sandhya[2], Gangarde Ashwini[3], Kothawale Shital[4]

Sinhgad Academy of Engineering Kondhwa(bk), Pune, India[1,2,3,4]

Abstract: Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. As the numbers of files are increasing the condition of storage node can't be managed. Because of high volume of files, it results in wasted hardware resources, increased control complexity of data center and less efficient storage system. Many systems are existed in the market regarding secure authorized deduplication like symmetric algorithm, Farsite distributed file system, digital fingerprint and message locked encryption. So to overcome this above flaws our proposed system put forwards an idea of secure authorized deduplication by using data hash key, bloom filter, subset vector creation and reverse circle cipher.

Keywords: Data hash key, bloom filter, subset vector creation, reverse circle cipher.

I. INTRODUCTION

Deduplication is a technique for eliminating duplicate copies of data through a de-duplication scanning process, which improves the system performance and decreases the bandwidth occupied by data transmission. Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts/decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since the encryption operation is deterministic and is derived from the data content, identical data copies will generate same convergent key and hence the same cipher text. To prevent unauthorized access, a secure proof of ownership protocol is also needed to provide the proof that the user indeed owns the same file when a duplicate is found. After the proof,subsequent users with the same file will be provided a pointer from the server without needing to upload the same file. A user can download the encrypted file with the pointer from the server, which can only be decrypted by the corresponding data owners with their convergent keys. Thus, convergent encryption allows the cloud to perform deduplication on the cipher texts and the proof of ownership prevents the unauthorized user to access the file. However, previous deduplication systems cannot support differential authorization duplicate check, which is important in many applications. In such an authorized deduplication system, each user is issued a set of privileges during system initialization each file uploaded to cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files. To handle the concept of a hybrid cloud approach for secure authorized deduplication there are many methodologies are supporting like symmetric algorithm, Farsite distributed file system, convergent encryption, SALAD, Message lock encryption,digital fingerprint.

1. Farsite distributed file system:
This system provides availability by replicating each file onto multiple desktop computers. It presents a mechanism to reclaim space from this incidental duplication to make it available for controlled file replication.

This mechanism includes convergent encryption which enables duplicate files to coalesced into the space of signal file, even if the files are encrypted with different users key, and SALAD (Self Arranging, Lossy, Associative Database) for aggregating file content and location information in decentralized, scalable, fault tolerant manner. Large scale Simulation experiments show that the duplicate file coalescing system is scalable, highly effective, fault tolerant.

2. Symmetric Algorithm:
In Symmetric Algorithm system same key is used for both encryption and decryption process. for better authentication and confidentiality and security in cloud computing it provide new duplication constructions using secure algorithm it supporting for authorized duplicate check in efficient cloud architecture, in which the duplicate check tokens of files are generated with symmetric keys. Basically it is used for the huge amount of data. By symmetric keys for attacker it is easy to break the keys.

3. Message-Locked Encryption:
In Message locked encryption system the key under which encryption and decryption are performed is itself derived from the message. It provides a way to achieve secure deduplication (space efficient secure outsourced storage), a goal currently targeted by numerous cloud storageproviders.

It provides definition both for privacy and for a form of integrity that we call tag consistency. Based on this foundation, it makes both practical and theoretical contributions. On the practical side it provides ROM security analyses of a natural family of MLE schemes that

include deployed scheme. On the theoretical side the challenge is standard model solution, and we make connections with deterministic encryption, hash functions secured on correlated input and sample-then-extract paradigm to deliver scheme under different assumptions and for different classes of message sources.

4. Digital Fingerprint:

Through hash algorithm, hash functions can generate an exclusive fixed-sized digital fingerprint for each data chunk. In order to transform the variable-length data into fixed length data, hash function scatter and remix data through mathematical function to produce a fixed value shorter than new data. This calculated hash value, the fingerprint or the signature of the data is usually expressed by strings of random characters and numbers.

A digital fingerprint is essential feature of data chunk. The optimal state is such that each data chunk has its unique fingerprint, and different data chunks have different have different fingerprints. This paper is been classified in different section like section 2 is dedicated for related work, section 3 for conclusion.

II. RELATED WORK

System that provides security and reliability by storing encrypted replicas of each file on multiple desktop machines. To free space for storing this replica the system coalesces incidentally duplicated files, such as shared documents among work groups or multiple users copies of common application programs. But this system requires solution for problem of enabling the identification and coalescing of identical files when this files are encrypted with the keys of different users. [2] And also identifying in decentralized, scalable, fault tolerant manner files that have identical content. [3] Relocating the replicas of files with identical content to a common set of storage machines.

[4] Coalescing the identical files to reclaim storage space while maintaining the semantics of separate files. [5] Symmetric encryption algorithm narrates that the same key is used for both encryption and decryption process. As the same key is used for both encryption and decryption of the message so for attacker it is easy to break the security if attacker finds the any of the key. [6] Message Locked Encryption technique describes a way to achieve secure deduplication (space efficient secure outsourced storage), a goal currently targeted by numerous cloud storage providers. It provides definition both for privacy and for a form of integrity that we call tag consistency. But unfortunately this system has a limited user because of proof of ownership. [7] Digital fingerprint is used for deduplication in which the data is divided into the number of chunks and through hash algorithms, hash function can generate an exclusive fixed size digital fingerprint for each data chunk. In order to transform the variable length data into fixed length data, hash functions scatter and remix the data through mathematical functions to produce a fixed size value shorter than the raw data, but as the data is divided into the number of chunks it loses the data integrity. To overcome some major disadvantages in secure authorized deduplication this paper tries to propose a method using some best ideas which includes data hash key, bloom filter, subset vector creation and reverse circle cipher. Here in this paper many of the ideas are been analyzed by different authors in section 2 to narrish our idea of project.

REFERENCES

[1]. ” Reclaiming Space From duplicate file in a Server less Distributed File System”, John R.Douceur, Atul Adhya, William J.Boloskey, Microsoft Research, July 2002.[2]”DHCP options And BOOTP Vendor Extensions”, Alexander and R.Droms,RFC 2132,Mar 1997.
[2]. ”Server less Network File System”, Anderson, M.Dahlin.J.Neefe, D.Roselli, R.Wang, ACM.pp.109-126, Dec 1995.
[3]. ” A Low Bandwidth Network File System”, A.Muthitacharoen, B.Chen, D.mazieres, 18th SOSP, ACM, oct 2001
[4]. ”an Efficient Secure Authorized Data deduplication Approach in Cloud computing”M.parabrahma Rao, Gunthati Prathap, vol.2, June 2015
[5]. ”Message-Locked encryption and Secure Deduplication”, Mihir Bllare, Sriram Keelveedh, March 2013
[6]. ”Improving accessing efficiency of Cloud Storage using Deduplication and Feedback Schemes”,Tin-yu wu,Jeng-shyamg pan,Chia-Fan Lin,IEEE SYSTEMS JOURNAL,VOL.8,March 2014