Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.7 No.S1, 2018, pp. 16-19

Improving Security on Cloud Based Deduplication System

B. Rasina Begum[1] and P. Chithra[2]
[1]Research Scholar, Department of Computer Science and Engineering,
Mohamed Sathak Engineering College, Kilakarai, India
[2]Assistant Professor, Department of Computer Science and Engineering,
Thiagarajar College of Engineering, Madurai, India
E-Mail: rasinayousuf@gmail.com, pccse@tce.edu

***Abstract* - Cloud computing provides a scalable platform for large amount of data and processes that work on various applications and services by means of on-demand service. The storage services offered by clouds have become a new profit growth by providing a comparable cheaper, scalable, location-independent platform for managing users' data. The client uses the cloud storage and enjoys the high end applications and services from a shared group of configurable computing resources using cloud services. It reduces the difficulty of local data storage and maintenance. But it gives severe security issues toward users' outsourced data. Data Redundancy promotes the data reliability in Cloud Storage. At the same time, it increases storage space, Bandwidth and Security threats due to some server vulnerability. Data Deduplication helps to improve storage utilization. Backup is also less which means less Hardware and Backup media. But it has lots of security issues. Data reliability is a very risky issue in a Deduplication storage system because there is single copy for each file stored in the server which is shared by all the data owners. If such a shared file/chunk was missing, large amount of data becomes unreachable. The main aim of this work is to implement Deduplication System without sacrificing Security in cloud storage. It combines both Deduplication and convergent key cryptography with reduced overhead.**

Keywords: **Convergent Key Cryptography, Proof of Ownership, Deduplication, Cloud Security**

I. INTRODUCTION

Storing huge amount of data in Cloud is the recent techniques in IT arena. There are more advantages of Cloud storage especially files accessibility. Data stored in the cloud can be taken anywhere and at any time. At the same time, Cloud storage has limited bandwidth. If the Internet connection is slow or interrupted, accessing and sharing files are problematic. Cloud storage is scarce resource. It should be utilized efficiently. Else it will cost more. Data redundancy leads to wastage of space. Different enterprise has placed their data in Cloud. Many data are redundant in nature. For Example, Consider email system contains 50 instances of the same one megabyte (MB) file attachment. If it is backed up, all 50 instances are saved in cloud storage, requiring 50 MB storage space. In data Deduplication, only one copy of the attachment is stored; each further same instances refer back to the one saved copy. In this example, 50 MB storage could be reduced to only 1 MB. Different applications have different levels of data redundancy. Many Backup applications in general take advantage from de-duplication owing to the repeated full backups of an existing file system.

Deduplication helps to improve storage utilization and reduces the bandwidth problem for transmitting large sized redundant data. Instead of keeping the multiple copies of the same data, it stores the single instance. There are many benefits of Deduplication. Business get more benefits from data Deduplication include: Reduced backup costs; reduced costs for business continuity / Disaster Recovery; Increased storage efficiency; and network efficiency. There is much duplication in the data outsourced to the cloud.

There are two types of Deduplication namely File-Level Deduplication and Block Level Deduplication [1]. File-level Deduplication is a one which finds redundancies between different files and removes these redundancies to reduce storage, and Block level Deduplication, which finds and removes redundancies between data blocks. The file can be divided into block with fixed size or variable size. Fixed size blocks simplify the computations of block boundaries, while using variable-size blocks, it increases the complexity.

In the following fig.1, 64KB data is segmented into 8 chunks each size is 8KB. After segmentation, some blocks are identified and those blocks will be stored only once.

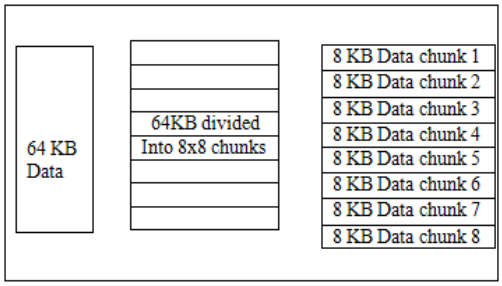

Fig. 1Data Segmentation

Block level Deduplication [7] is depicted in the following fig.2. File 1 consists of set of blocks differentiated in terms of colors. File 2 also consists of set of blocks differentiated in the same way. Here some blocks are identical. Using Deduplication concept, same blocks are stored once to avoid redundancy.

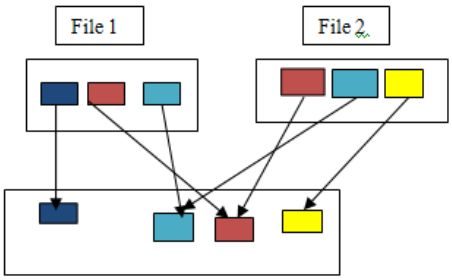

Fig. 2 Block Level Deduplication

II. PROPERTIES OF SECURE DATA

The Following Are The Properties Hold By Secure Data.

1. *Verifiability:*It assures that the content of data has not been modified by unauthorized manner and available as it is. All entries must have the sufficient information for its verification.
2. *Confidentiality:*Cloud data usually hold sensitive information. It is necessary to prevent sensitive information from illegal people. It is vital that the authorized people can get it: Access must be limited to those authorized to view the data.
3. *Tamper Resistance:*A Cloud data should be tamper resistant in such a way that no valid entries can be introduced by anyone other than the creator. In addition once those entries are created, they cannot be manipulated without detection.
4. *Correctness:* Once the data is stored in cloud storage, it will not be edited or modified by the attackers.
5. Privacy: Cloud data should not be casually traceable or linkableto other sources during transmission and in storage.

III. HASH BASED DEDUPLICATION FOR DATA CHUNKS

As per the report of International Data Corporation, the amount of data in the world is expected to attain 40 trillion gigabytes in 2020. Many data will be redundant data and waste the storage space. This proposed work will identify the data redundancy and keeping only one copy which is shared by all owners. Many storage service providers are unwilling to apply encryption over the data because identical data copies of different users using different keys will lead to different cipher texts. So hash based Deduplication helps to solve this problem.
The steps needed for eliminating redundancy between chunks are as follows.

1. Slice data into chunks [5][2]. Chunk may be fixed or variable. Ex. Data is sliced into chunks A, B, C, D, E
2. Generate Hash for all chunks and save the values. A_h, B_h, C_h, D_h, E_h
3. Do the same for next data and generate its hash value.
4. Check it with the previous hash value and store only the new hash and data to eliminate the redundancy.

It is depicted in the following fig.3. Here sample data is “Hello World”. Next calculate the hash value of the data “Hello World”. This is the encryption key used to encrypt the data. Two diverse users will automatically use the same encryption key for the same kind of data and thus generate the same kind of Cipher text. At this point, the cloud storage provider can still perform Deduplication since he can detect duplicate data.

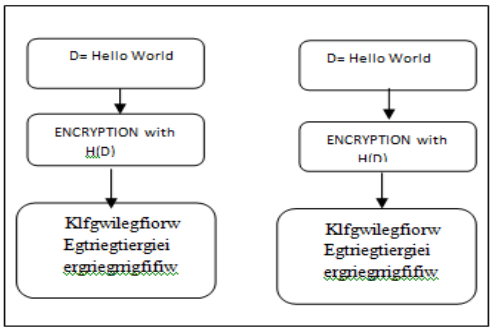

Fig. 3 Hash based Deduplication

IV. PROPOSED WORK

Confidentiality of data and Authorization of duplicate check are discussed here. Assume that data are sensitive and needed to be fully securedagainst both public cloud and private cloud. Under this assumption, two types of attackers are considered, one is, an attacker aims to extract private information from both public cloud and private cloud, and internal attacker aims to obtain more information on the file from the public cloud and duplicate-check token information from the private cloud.

It is essential that data will be encrypted in Deduplication system before sending data to the storage cloud to maintain the confidentiality of data. The data are encrypted with the conventional encryption scheme to guarantee the security of data. The following algorithms and methods are proposed for encryption and check the ownership of data.

A. Convergent Key Cryptography

In Traditional encryption, Users use different keys to encrypt the data. As a result, the identical data produces different cipher text. This process makes the Deduplication as impossible one. There are conflict solutions between Deduplication and Encryption. Convergent encryption[3] solves this problem by generating encryption key using the hash of the data segment. This technique gets the key from the original text. Common method is to compute key as the hash of the original text.

Ex., If the given message is X, then the key is K=H(X).
Here H is a cryptographic hash function.
For encryption, this key K is used.
Convergent key cryptography [4] provides data confidentiality in Deduplication System. A user receives a convergent key from original data and encrypts the

information with the convergent key. A Convergent Encryption will be outlined with three important operations:

1. KeyGen(M) – It gives Convergent Key for encryption.
2. Encrypt(K,M) -> It takes original message and key as input and generates Cipher text C.
3. Decrypt(K,C) ->this operation provides the original text M using Convergent Key and Cipher text.

B. Proof of Ownership (PoW)

It provides users to prove their ownership of data to the storage server [1] [6]. PoW is implemented as an interactive Algorithm PoW run by a user and storage server. The Server derives a short tag value ϕ(D) from aData copy D. To prove the ownership of the data copy D, the user needs to

1. Compute and send ϕ′ to the Server.
2. User present proof to the server that it owns Din an interactive way with respect to ϕ′. The PoW is successful ifϕ′ = ϕ(D) and the proof is correct.

C. Uploading File: To upload a file F, the user first performs the file-level deduplication by sending ϕ(D) to the servers. If a duplicate is found, the user will perform the file-level deduplication, otherwise, the user performs the block-level deduplication as follows. User firstly divides Dinto a set of fragments {Aj}(where j= 1,2, · · ·). For each fragment Aj, the user will perform a block-level duplicate check by computing ϕ(Aj)= TagGen(Aj), where the data processing and duplicate check of block-level deduplication is the same as that of file-level deduplication if the file Dis replaced with block Aj.

V. PERFORMANCE ANALYSIS OF DIFFERENT ALGORITHM

Here some algorithms are considered and evaluate in terms of its economy. The following fig. 4 shows the performance analysis for the protection algorithms. It's clear that AES is a lot of economical than other algorithms. The time needed to finish the method of cryptography is smaller amount as compared to different Algorithms. Thus we select AES Algorithm for encrypting the Original data.

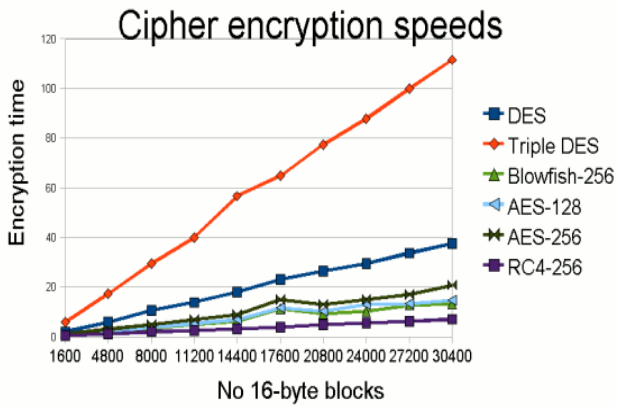

Fig.4 Performance Analysis

VI. IMPLEMENTATION

We implement Deduplication system with robust security. The algorithms were run by the storage server and the clients both are executed on a machine with Linux OS, 2.70GHz Intel(R) Core(TM) i7-4600UCPUand 8GB RAM. It has three separate programs. A Client program is used to model the data users to carry out the file upload/download process. A Private Server program is used to model the private cloud which manages the private keys and handles the tag computation. A Storage Server program is used to model the S-CSP which stores and deduplicate the files. The followings are the function calls used in our system.

1. *File Tag (File):* Computes SHA-1 hash of the File as File Tag.
2. *Dup Check Request (Token)*: Requests the Storage Server for Duplicate Check of the files.
3. *File Encryption (File)*: It encrypts the File with Convergent Encryption using 256-bit AES algorithm, where the convergent key is from SHA-256 Hashing of the file;
4. *File Upload Request (FileID, File, Token)*: It uploads the File Data to the Storage Server if the file is Unique.
5. *File Storage (FileID, File, Token)*: It stores the File on Disk.

VII. FURTHER DISCUSSION

The proposed scheme has some additional advantages.

1. *Flexibility:* The proposed scheme is highly legible to support access control on encrypted data with deduplication. One data owner can legibly update DEK. The new key can be easily issued to other data holders or eligible data users by CSP with a low cost, especially when CSP has issued the reencryption key already.
2. *Low Cost of Storage:* This can clearly save the storage space of CSP since it only stores one copy of the same data that is shared by data owner and data holders. Storing deduplication records hold some storage. But comparing with the big volume of duplicated data, this storage cost can be disregarded.
3. *Big Data Support:* The proposed scheme can efficiently carry out big data deduplication.

VIII. CONCLUSION AND FUTURE WORK

Cloud computing does not store the data on the user's computer, but in the cloud. So the data has to be encrypted first and then uploaded to the cloud for process. It is very important for the data security of the cloud computing platform, because the data is invisible to the third party and can be processed by the cloud itself. This work proposes the Deduplication system with robust security to enhance storage utilization and achieving the confidentiality of the users' outsourced data without an overhead. The current framework can be applied to only text file formats. In future, it can be extended for performing Deduplication with

tight security in other file formats such as audio, image and video.

REFERENCES

[1] Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang and Yang Xiang, Mohammad Mehedi Hassan and Abdul hameed Alelaiwi, "Secure Distributed Deduplication Systems with Improved Reliability", *IEEE Transactions on Computers*, Vol. 64, pp. 3569-3579, December 2015

[2] Tin-Yu Wu, Jeng-Shyang Pan, and Chia-Fan Lin, "Improving Accessing Efficiency of Cloud Storage Using De-Duplication and Feedback Schemes", *IEEE Systems Journal*, Vol.8, pp. 2018-218, March 2014

[3] Jin Li, Xiaofeng Chen, FatosXhafa, and Leonard Barolli, "Secure Deduplication Storage Systems Supporting Keyword Search", *Journal of Computer and System Sciences*, Vol. 81, pp. 1532-1541.

[4] Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou, "Secure Deduplication with Efficient and Reliable Convergent Key Management", *IEEE Transactions On Parallel And Distributed Systems*, Vol. 25, pp.1615-1625, June 2014

[5] Deepu S Bhaskar, B. Shylaja, "Performance Comparison of Deduplication Techniques For Storage In Cloud Computing Environment", *Asian Journal of Computer Science And Information Technology*, Vol.4, pp. 42-46, 2014

[6] A.N. Naveen, and V. Ravi, "Client Side Deduplication Scheme for Secured Data Storage in Cloud Environments", *International Journal of Engineering Research & Technology*, Vol.4, May 2015.

[7] Rongmao Chen, Yi Mu, and FuchunGuo, "BL-MLE:Block-Level Message-Locked Encryption for Secure Large File Deduplication", *IEEE Transactions on Information Forensics and Security*,Vol. 10, pp. 2643-2652, August 2015.