ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Secured Authorized Deduplication in Cloud Storage

Mitwa Parsana[1], Sweta Ghadi[2], Madhavi Bhatt[3], John Kenny[4]

B.E. Student, Dept. of CSE, Universal College of Engineering, Mumbai, India[1,2&3]

Assistant Professor, Dept. of CSE, Universal College of Engineering, Mumbai, India[4]

ABSTRACT: Cloud computing involves deploying groups of remote servers and software networks that allow centralized data storage and online access to computer services or resources. With an increasing demand of cloud storage, effective methods need to be placed to reduce hardware costs, meet the bandwidth requirements and to increase storage efficiency. Deduplication is the process of storing duplicate data in a single instance and just replicating a reference pointer. This process is intended and used to store more data logically and provide more free space physically. Many systems are stick to a limited number of file formats as they address the logical structure of the file in a certain format. This problem is solved by ingenious approach of hashing data in its binary form itself. Using this approach has an advantage of including all kinds of file formats because the proposed system has no relevance with the actual structure of the file in terms of a certain format. The data here is considered as a general entity and is divided into fixed size blocks over which hash is computed. This hash is matched whenever a new file arrives. This approach yields better results over existing approaches for deduplication.

KEYWORDS: cloud computing, cloud storage, deduplication, hashing

I. INTRODUCTION

Nowadays cloud service providers offer high availability storage and parallel computing at comparatively lower costs. As clouds are gaining prevalence, an overtly expanding amount of data is being stored on the cloud and shared with users of specific privileges, which in turn define the access specifications of the data that is stored. One of the most critical challenges in cloud storage services is the overhead of managing the increasing volume of data. To make the management scalable and easy in cloud computing, a technique called data deduplication has been used for a long time and has been gaining its due in recent times.

Data deduplication in its core is a specialized data compression technique for the elimination of duplicate copies of repeated data in storage. The aforementioned technique is used to improve the storage utilization and is also to be applied to network based data transfers to reduce the number of bytes to be transferred from one network to another. The approach states that instead of storing multiple copies with the same content, the deduplication scheme eliminates the redundant data by keeping only one physical copy and by using referential pointer to other redundant data to that copy.

II. LITERATURE SURVEY

In [3] paper, author addressed an important security concern in cross-user client-side deduplication of encrypted files in the cloud storage: confidentiality of users' sensitive files against proof protocol of their client-side deduplication scheme UH-CSD and the network transfer time of files without encryption both outside adversaries and the honest-butcurious cloud storage server in the bounded leakage model.

In [4] there are two categories of data deduplication strategy, and extend the fault-tolerant digital signature scheme proposed by Zhang on examining redundancy of blocks to achieve the data deduplication. The proposed scheme in this

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

paper not only reduces the cloud storage capacity, but also improves the speed of data deduplication. Furthermore, the signature is computed for every uploaded file for verifying the integrity of files.

In [5] it stores only the single copy of the duplicate data. Client-side deduplication tries to identify deduplication chance already at the client and save the bandwidth of uploading copies of existing files to the server. To overcome the attacks Shai Halevi1, Danny Harnik, Benny Pinkas, and Alexandra Shulman-Peleg proposes the Proof of ownership which lets a client efficiently prove to a server that that the client keep a file, rather than just some short information about it present solutions based on Merkle trees and specific encodings, and analyse their security.

In [6] introduce two models for secure deduplicated storage: authenticated and anonymous. In both the authenticated and anonymous models, a map is created for each file that describes how to reconstruct a file from chunks. This file is itself encrypted using a unique key. In the authenticated model, sharing of this key is managed through the use of asymmetric key pairs. In the anonymous model, storage is immutable, and file sharing is conducted by sharing the map key offline and creating a map reference for each authorized user. Convergent encryption does leak information that a particular ciphertext, and thus plaintext, already exists. So an adversary can get more potential knowledge by constructing a particular data chunk and checking whether it has already existed on the storage. What's worse, the encryption key is determined if a typical plaintext is given, i.e. the mapping from plaintext to ciphertext is determinate but not random, which is less secure according to semantic security principal.

III. MOTIVATION

With emerging market for cloud storage, variety of systems have been developed to provide secure storage, but the matter of the fact is none of the traditional encryption techniques are suitable for deduplication purposes. As a candidate solution, divergent encryption techniques have proven to achieve great deal of confidentiality in the deduplication process, but in whole the process suffers from the threat of dictionary attacks. Yet another approach of providing proxy re-encryption has been proposed but it's actually performance in real time environment is something that remains a questionable. Currently an algorithm which deterministically generates key without disclosing any plaintext or related information is in works. Keys are generated via a key server hence the secret key distribution centre hack results in the compromise of all the files on the cloud. This algorithm was primarily used by 4shared.com before being thrashed down by an attacker who exploited this mechanism to his benefit.

IV. PROBLEM STATEMENT

Cloud storage has been a very popular way for storing data over the internet. As the growth of the users has been exponentially rising every year so is the size of the data that is uploaded from the users. Hence deduplication becomes de facto measure for cloud service providers. The idea is to store unique copy of duplicate data which in turn greatly reduces their storage and data transfer cost. But on the flip side deduplication comes with bottlenecks in security and privacy of data. So, to get the best of both worlds we need a secure deduplication policy which would maximize deduplication and also meet the standards of security and privacy.

V. PROPOSED SYSTEM

In proposed system, block level deduplication is implemented. As the name implies, that file is being fragmented into fixed size blocks or chunks using fixed-length block approach. If the block is determined to be redundant then each block is assigned a unique identifier by using a simple checksum-based approach (MD5/SHA etc). The hash algorithm then generates a unique hash to that particular chunk and that particular unique hash will be compared with central index. If the hash is already stored in index, then it delineate that data is processed before only. Accordingly only a reference pointer is saved to the previously stored block. If the generated hash is new and does not exist in index, then that block is unique. The unique block is stored and the unique hash is updated in the index.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

VI. SYSTEM MODEL

There are three main entity of this system called user, private cloud and cloud service provider are shown in Fig. 1.

A. USER:

A user is an entity that wants to outsource data storage to the cloud and access the data later. In the authorized deduplication system, each user is issued a set of rights in the start of the system. If the User wants to upload Files to the cloud and the file is already exist then the user does not upload files.

B. PRIVATE CLOUD

Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating user's secure usage of cloud service. Specifically, since the computing resources at data user side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user with an execution environment and infrastructure working as an interface between user and the public cloud. The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users.

The public cloud does not have access to the user credentials. If compromised token generation keys can be changed regardless of the user credentials. Public cloud does not have access to decrypted files. Deduplication only happens on privileged files.

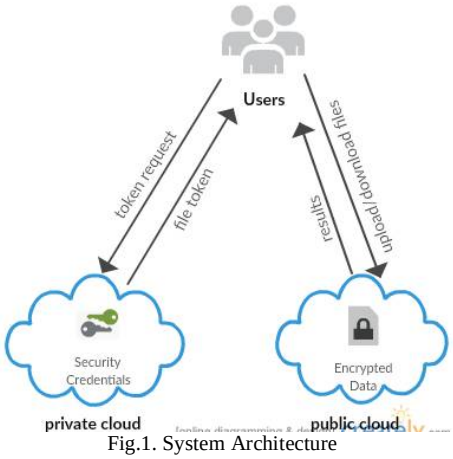

Fig.1. System Architecture

C. CLOUD SERVICE PROVIDER

The CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the CSP eliminates the storage of redundant data via deduplication and keeps only unique data. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users.

VII. ALGORITHM DETAILS

Block of data is encrypted under a key derived by hashing the data itself. The security that is implemented in this system is using SHA algorithm and AES algorithm.

A. SECURE HASH ALOGIRTHM-1

The SHA-1 algorithm belongs to a set of cryptographic hash functions which are similar to the MD(message digest) family of hash functions. But the core difference between the SHA-1 and the MD family is the more frequent use of input bits during the course of the execution in the SHA-1 algorithm than in the MD counterparts. This fact results in the SHA-1 being more secured compared to MD4 or MD5 but at the expense of slower execution.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

B. ADVANCE ENCRYPTION STANDARD

AES allows for three different key lengths: 128, 192, or 256 bits. Encryption consists of 10 rounds of processing for 128-bit keys, 12 rounds for 192-bit keys, and 14 rounds for 256-bit keys. Except for the last round in each case, all other rounds are identical. Each round of processing includes one single-byte based substitution step, a row-wise permutation step, a column-wise mixing step, and the addition of the round key. The order in which these four steps are executed is different for encryption and decryption.

C. BINARY COMMON SUBSEQUENCE

The BCSS is a modified version of the longest common subsequence algorithm that is used to operate on binary data. The binary common subsequence was introduced by Norman et el in 1999 advanced summit for algorithms. Leading edge towards millennium in multiple systems, the algorithm found its place in data duplication detection. Many of the plagiarism checker systems check out the data on basis of the BCSS. The proposed system use the algorithm to pad the patches of data to provide a much granular deduplication.

VIII. EXPERIMENTAL RESULT

The final results of the designed system are given below. From those results we get the detailed information to sign in page, file uploading and file downloading and Deduplication information. The output images given as below.

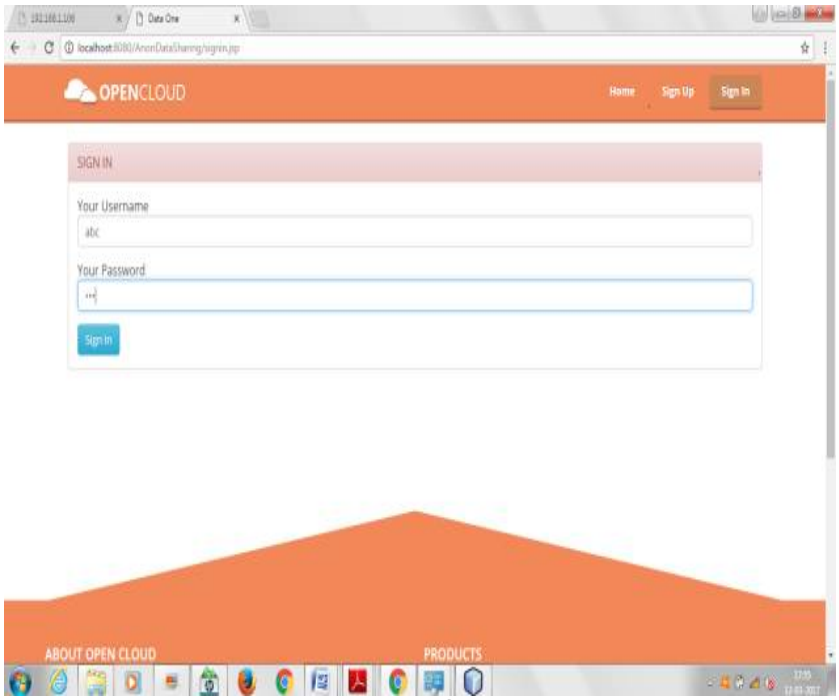

Fig. 2. SignIn Page

In Fig. 2, shows the signin page where user can login into their account. Users can only login to this page only if they had registered to this system with correct identity information.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

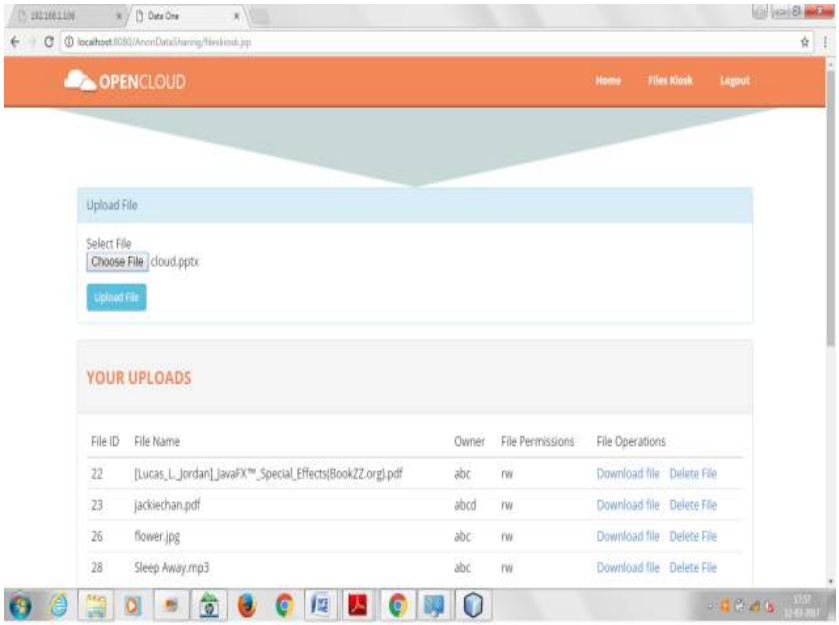

Fig. 3. File Uploading

In Fig. 3, shows the profile of user where user can upload any type of file namely text, pdf, images, videos etc. Also user can download and delete the files. Also the user can see the operation that can be performed on the files like read-write or only read and can see the owner of the file whose designation is lower level than that user.

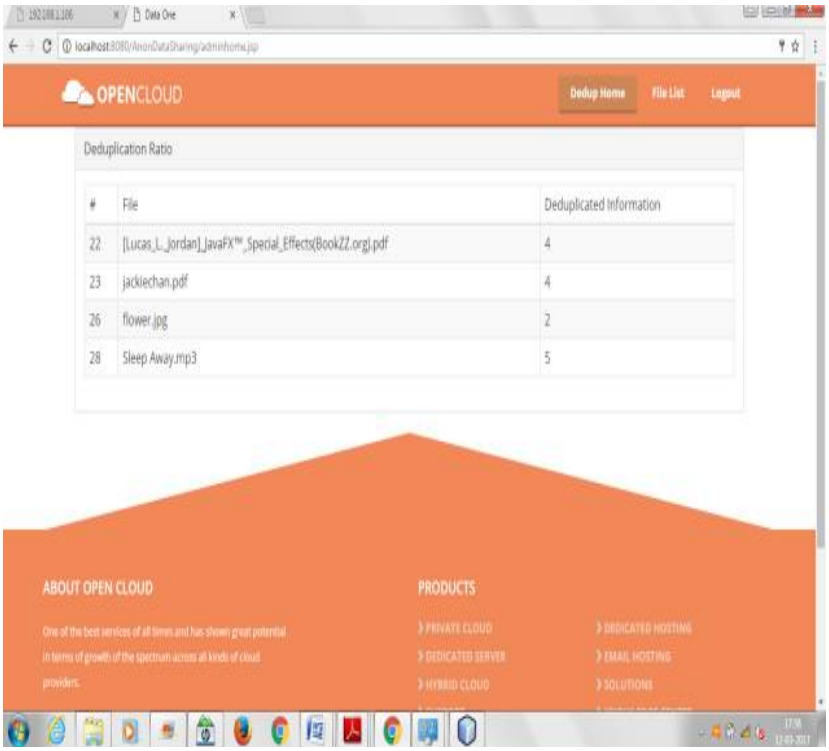

Fig. 4. Deduplication Information

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

In Fig. 4, shows the admin page where admin can see the files uploaded by the user and duplication information. Deduplication information contains the number of blocks generated by the files. Also it has file list module in it which shows the files uploaded by every user with their details.

IX. CONCLUSION AND FUTURE WORK

Nowadays most of users use cloud to store data. Increasing amount of data in cloud is a major concern. The system includes methods that are used to achieve cost effective storage and effective bandwidth usage by deduplication. The system ensures that user is an authenticated person or not. Only such a person can perform deduplication check and store data in cloud. However, data deduplication is the most crucial element for improving efficiency of the cloud system. This tchnique will play a major role in the cloud based services for storing backup data by both medium and large enterprises. As part of future work, will work on finding possible optimizations in terms of bandwidth, storage space and computation. Cloud secure deduplication may be extended with more security features such as proofs of retrievability, data integrity checking and search over encrypted data.

REFERENCES

1. Golthi Tharunn, Gowtham Kommineni, Sarpella Sasank Varma, Akash Singh Verma, 'Data deduplication in Cloud Storage', International Journal of Advanced Engineering and Global Technology, Vol-03, Issue-08, August 2015.
2. Meghana Vijay Kakde, Prof. N.B.Kadu, 'Survey Paper on Deduplicating Data and Secure Auditing in Cloud', International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016
3. J. Xu, E.-C. Chang, and J. Zhou, 'Weak leakage resilient client-side deduplication of encrypted data in cloud storage'. Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, Pages 195-206 ,ACM,2013.
4. W. K. Ng, Y. Wen, and H. Zhu. 'Private data deduplication protocols in cloud storage', In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012.
5. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. 'Proofs of ownership in remote storage systems', In Y. Chen, G. Danezis, and V. Shmatikov editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.
6. Mark W. Storer Kevin Greenan Darrell D. E. Long Ethan L. Miller, 'Secure Data Deduplication', Proceedings of the 4th ACM international workshop on Storage security and survivability Pages 1-10 ,ACM, 2008.

BIOGRAPHY

Mitwa Parsana is a B.E. Student in the Computer Science and Engineering Department, Universal College of Engineering, Mumbai University. Her interests are Cloud Computing and System Security.

Sweta Ghadi is a B.E. Student in the Computer Science and Engineering Department, Universal College of Engineering, Mumbai University. Her interests are Cloud Computing, HCI and RDBMS.

Madhavi Bhatt is a B.E. Student in the Computer Science and Engineering Department, Universal College of Engineering, Mumbai University. Her interests are Cloud Computing and Software Testing.

John Kenny is a Assistant Professor in the Computer Science and Engineering Department, Universal College of Engineering, Mumbai University. He received his M Tech. from GITAM Institute of Technology, Visakhapatnam. His interests are Machine Learning and Data Mining.