IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

# Secure Data Deduplication

**Vijay M Hittalmani[1], Prof. Vidya I. Hadimani[2]**

Student, Department of CSE, KLE Dr. M S Sheshgiri College of Engg & Tech., Belagavi, Karnataka, India[1]

Professor, Computer Science & Engg, KLE Dr. M S Sheshgiri College of Engg & Tech., Belagavi, Karnataka, India[2]

**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. To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data deduplication. Different from traditional deduplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. We also present several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement a prototype of our proposed authorized duplicate check scheme and conduct test bed experiments using our prototype. We show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

**Keywords:** Deduplication, authorized duplicate check, confidentiality, hybrid cloud

## I. INTRODUCTION

Cloud computing helps to keep the data elastic in nature the user can use the cloud service as a pay per use facility and need not pay for any extra hardware. Even though the cost is reduced cloud services provide a good quality of service. Cloud storage can handle huge amount of data without any burden.
The operation overhead is managed by moving the appropriate type of files to the cloud. Deduplication is good but is difficult to achieve on the encryption level.

If two users are trying to upload the same data but the data is encrypted by different keys then the encrypted data is difficult to analysis for similarity. The encryption using convergent key can play an important role.

Using the same key to encrypt can produce the same hash code for the same data and it will be easy to identify the similar data. Hence the ciphertext generated using the convergent key can help to identify the duplicate data and can help to save the memory.

A. Objective of the project:
Keeping the data secured from unauthorized users and to save the memory in the cloud storage. For this, the cloud service provider (CSP) uses a technique of deduplication at encryption level. This is a facility that allows the owner to save a lot of memory by keeping only one original copy of the file and linking the replica to the original file.
\
The security is maintained by encrypting the data and two level authorizations is done. In today's generation there is need to store huge amount of data for long term use. The data can include personal information, medical records etc. To store such data there is a need of lot of resources and it is very costly.

## II. LITERATURE SURVEY

Literature survey plays a very important role in project development. Literature survey provides the required knowledge about the project and its background. It also helps in following the best practices in project development. Literature survey also helps in understanding the risk and feasibility of the project.

1) In the real world more often we tend to see the data that are two or more in database. The records which are duplicate will share the different keys that will make the duplicates matching task difficult and will result in errors. Errors will usually occur due to lack of standard formats, incomplete information or transcription errors. The duplicate detection algorithm is used which detects the duplicate records and also some of the metrics are considered that will help us to detect the similar field entry of data that is done. [1]
2) Data Deduplication is a technique that is mainly used for reducing the redundant data in the storage system which will unnecessarily use more bandwidth and network. So here some common technique is being defined which finds the hash for the particular file and with that the process of deduplication can be simplified, David Geer. [11]
3) De-duplication is the technique that is most effective most widely used but when we apply it to multiple users the cross-user deduplication tend to have some security problems. Simple mechanisms can be used to do the cross-user deduplication that reduces data from leaking and also some of the security issues are discussed with how exactly to identify the files and to encrypt them while sending is discussed, Danny Harnik, Benny Pinkas, and Alexandra Shulman- Peleg. [10].
4) Dup Less: Encryption for deduplicate storage for cloud service provider like Drop box saves spaces by storing

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

only one original copy. Dup less is used to provide secure deduplicate storage as well as storage resisting brute-force attacks. Drawback here is that Dup less only works on text files and images [2].

5) In 2013, Neal Leavitt [4] studied the different problems arriving in the deduplication in multi-tenant environment. Different authors proposed the use of the single key encryption, i.e., to generate keys from the hash code of plain text.

6) Clouded up: a technique that provides secure deduplication by encrypting the data but its limitations are only up to text files. [3]

7) In [4], the authors have proposed that the advantages of cloud computing lies in reduced storage management and increased efficiency. The cloud computing uses less hardware compared to the traditional systems. It is highly flexible and less costly.

8) In [7], the authors have proposed that Proper encryption steps to be used in the cloud. While cloud computing gives flexibility the use of good encryption technique is important. Cloud computing helps to reduce the hardware costs and makes IT companies more profitable.

9) Far site - a deduplication system that focuses very less on authorization of users. Data on the cloud is stored in normal form. In Far site deduplication is performed only on text and images and it does not support all the file formats. [5]

## III. PROPOSED SYSTEM

In the proposed system the owner of the cloud service provides his cloud service to various users who want to use it. The users can easily upload and download their respective files from the cloud. The files are stored on the cloud in the encrypted format. The files get decrypted only when the authorized user downloads the specific file, hence security is preserved. In the backend only one original copy of each file is stored and if the same file is uploaded by many users only a link is made to the original file, which saves enormous amount of memory on the cloud.

## IV. ARCHITECTURE

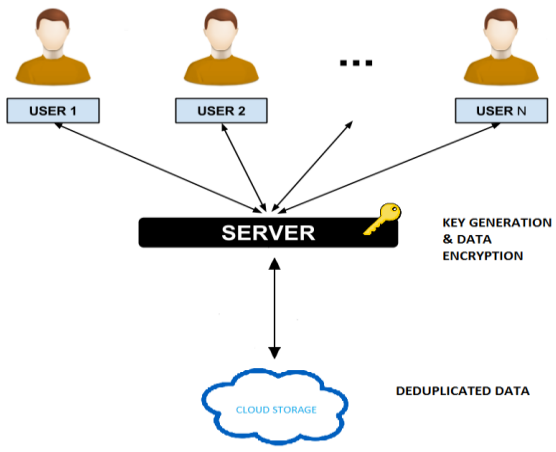


Figure 1: Architecture.

This diagram shows the deduplication at encryption level where many users who use a cloud service try to upload their files on the cloud storage. The deduplication technique is applied on the cloud storage where only one original copy of each file is stored.

The users can easily download their files, which get decrypted while downloading and can be viewed by the user in its original form.

## V. MODULES

This project consists of the following modules:

- Data Owner Registration
- Data User Registration
- Data user upload/download file

Data Owner Registration

- In this module, if a data owner has to view all the files uploaded by the users then he/she should login first.
- When the Owner is successfully logged in he/she can view all the files uploaded by the users and can check the status of the files.
- The owner can delete any file if he pleases to do so.

Data User Registration

- In this module, the user has to login to upload/download files.
- When the User is successfully logged in he/she can view all the files uploaded by him/her and can upload or download any files.

Data User Upload/Download file

- In this module, the user can view all his uploaded files and can download any of his file.
- If the user wants to delete a file then an OTP is sent to the user's mail id which should be correctly entered to delete the file.

## VI CONCLUSION

Management of data has become very common for cloud services. Cloud services prefer to focus on their core business than to manage huge amount of data getting added each day. This work studied various aspects of deduplication of data. The various needs of the cloud services such as the data management, data deduplication, and encryption were studied.

The project implemented a scenario where the cloud service can deduplicate the uploaded data. The data users are provided with the proper data. The data was prevented from unwanted exposure and unauthorized access by placing a proper access control mechanism.

Authorized data deduplication aims at data security to keep the data secured and avoid unauthorized access. Deduplication at encryption level saves a lot of memory and memory can be utilized efficiently.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

## REFERENCES

[1] Private and Hybrid Clouds, Published by the IEEE Computer Society, 2013.
[2] Dup LESS: Server-Aided Encryption for Deduplicate Storage by Mihir Bellare, Sriram Keelveedhi, Thomas Ristenpart, 2013.
[3] Clouded up: Secure Deduplication with Encrypted Data for Cloud Storage Refik Molva, Melek, 2013.
[4] Neal Leavitt, "Hybrid Clouds Move to the Forefront" Published by the IEEE Computer Society, MAY 2013.
[5] Rev De dup: A Reverse Deduplication Storage System Optimized for Reads to Latest Backups, 2013.
[6] Chuanyi Liu, Xiaojian Liu, and Lei Wan. Policy-based de-duplication in secure cloud storage. In Trustworthy Computing and Services. Springer, 2013.
[7] Yang Zhang, Yongwei Wu and Guangwen Yang, Droplet: a Distributed Solution of Data Deduplication, 2012.
[8] Private Data Deduplication Protocols in Cloud Storage by Wee Keong Ng, Yonggang Wen, 2012.
[9] Mihir Bellare, Sriram Keelveedhi, and Thomas Ristenpart. Message-locked encryption and secure deduplication. In Advances in Cryptology–EUROCRYPT 2013, Springer, 2012.
[10] Danny Harnik, Benny Pinkas, Alexandra Shulman- Peleg "Side Channels in Cloud Services Deduplication in Cloud Storage",2010.
[11] David Geer, "Reducing the Storage Burden via Data Deduplication", December 2008.

## BIOGRAPHIES

**Vijay M. Hittalmani** received B.E. degree in Computer Science & Engg. from Visvesvaraya Technological University of Belagavi in 2013. Currently pursuing M.Tech degree in Visvesvaraya Technological University, Belagavi.

**Prof. Vidya I. Hadimani** is currently working as Assistant Professor in Computer Science & Engg department, KLE Dr. M S Sheshgiri college of Engg & Tech, Belagavi. She did her Bachelors of Engg in Computer Science & Engg from Karnatak University, Dharwad in the year 2001 and did her M.Tech in Computer Science & Engg from Visvesvaraya Technological University, Belagavi in the year 2010.