ABSTRACT

Data de duplication 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 de duplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this project makes the first attempt to formally address the problem of authorized data de duplication. Different from traditional de duplication systems, the differential privileges of users are further considered induplicate check besides the data itself. We also present several new de duplication 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, the proposed work implements a prototype of our proposed authorized duplicate check scheme and conduct test bed experiments using our prototype. The proposed work shows that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

FIGURE NO.	NAME OF THE FIGURE	PAGE NO.

LIST OF ABBREVIATIONS

ACRONYM		EXPANSION
SSP	-	Storage Service Provider
CDC	-	Content defined chunking
DFD	-	Data flow diagram

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Cloud computing provides seemingly unlimited “virtualized” resources to users as services across the whole Internet, while hiding platform and implementation details. Today’s cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified *privileges*, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever-increasing volume of data. To make data management scalable in cloud computing, de duplication has been a well-known technique and has attracted more and more attention recently.

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Deduplication can takeplace at either the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.Although data deduplication brings a lot of benefits, security and privacy concerns arise as users’ sensitive data are susceptible to both insider and outsider attacks. Traditional encryption, while providing data confidentiality, is incompatible with data deduplication.

Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identicaldata copies of different users will lead to different cipher texts,

making deduplication impossible. 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 ciphertext to the cloud. Since the encryption operation isDeterministic and is derived from the data content, identical data copies will generate the 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 accessthe 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 the 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. Before submitting his duplicate check request for some file, the user needs to take this file and his own privileges as inputs. The user is able to find a duplicate for this file if and only if there is a copy of this file and a matched privilege stored in cloud. For example, in a company,many different privileges will be assigned to employees.

In order to save cost and efficiently management, thedata will be moved to the storage server provider (SSP) in the public cloud with specified privileges andthe deduplication technique will be applied to store only one copy of the same file. Because of privacy consideration, some files will be encrypted and allowed theduplicate check by employees with specified privileges to realize the access control. Traditional de duplication systems based on convergent encryption, although providing confidentiality

to some extent; do not supportthe duplicate check with differential privileges. In other words, no differential privileges have been considered in the deduplication based on convergent encryption technique. It seems to be contradicted if we want to realize both deduplication and differential authorization duplicate check at the same time.

1.2 ORGANIZATION OF THE THESIS

Chapter 1 deals with an introduction to the project where the existing system is been discussed. It also gives an overview of how de duplication technique has been implemented with the high security.

In Chapter 2, a detailed description of the literature survey of the papers which are referred during the course of the project was summarized.

Chapter 3 gives a brief explanation on the aim and scope of the project. Here proposed system has been compared with the existing system. The issues in the existing system and the advantages of the proposed system are also discussed.

Chapter 4 deals with the methods and algorithms used. The hardware and software requirements are provided along with the system design, architecture and flow of overall project.

Chapter 5 deals with the system implementation of the project.

In chapter 6, the Results and Discussion along with the screenshots of each module has been depicted.

Chapter 7 deals with the summary and conclusion of the project. It also includes the future scope of the project.

CHAPTER 2

LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

The major part of the project development sector considers and fully survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations

[1] In this paper, they proposed an architecture that provides secure deduplication storage resisting brute force attacks, and realize it in a system called dupLESS . It enables clients encrypted data with an existing service. The encryption for deduplicated storage can achieve performance and space saving close to that of using the storage service with plaintext data.

[12] There is a mechanism to reclaim space from incidental duplication to make it available for controlled file replication.This mechanism convergent encryption, which

enable duplicate files to be coalesced into the space file, even if the files are encrypted with different users keys.

[15] It is a baseline approach in which each user holds an independent master key for encrypting the convergent keys and outsourcing them to the cloud. However, such a baseline key management scheme generates an enormous number of keys with the increasing number of users and requires users to dedicatedly protect the master key.

[17] In this project , they construct a private deduplication protocol based on the standard cryptographic assumptions is then presented and analyzed. They show that the private data deduplication protocol is probably secure assuming that the underlying hash function is collision-resilient, the discrete logarithm is hard and theerasure coding algorithm can erasure up to many fractions of the bits.

[21] In this paper, they design an encryption scheme that guaranteessemantic security for unpopular data and provides weaker security andbetter storage and bandwidth benefits for popular data. This way, datade duplication can be effective for popular data, whilst semantically secureencryption protects unpopular content. We show that our schemeis secure under the Symmetric External Decisional Diffie-Hellman Assumption.