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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Efficient Secure Authorized Deduplication in Hybrid Cloud using OAuth

Celina George

Student, Computer Science, Thejus Engineering College, Thrissur, India

Abstract: In Cloud computing involves deploying groups of remote servers and software networks that allow centralized data storage and online access to computer services or resources. Clouds can be classified as public, private or hybrid. 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. Datadeduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. This 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 and save bandwidth. To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique is used to encrypt the data before outsourcing. 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. Convergent encryption allows the cloud to perform deduplication on the ciphertexts and the proof of ownership prevents the unauthorized user to access the file. To enhance the system in security OAuth is used. OAuth (Open Authorization) is an open protocol for token-based authentication and authorization on the Internet used in hybrid cloud to enhance the security. OAuth enables the system to ensure that the user is an authenticated person or not. Only such authenticated user got the token for uploading and downloading in pubic cloud.

Keywords: OAuth, S-SCP, convergent encryption

I. INTRODUCTION

In computing, data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. A Hybrid Cloud is a combined form of private clouds and public clouds in which some critical data resides in the enterprise's private cloud while other data is stored in and accessible from a public cloud. As cloud computing becomes famous, an increasing amount of data is being stored in the cloud and used by users with specified privileges, which define the access rights of the stored data.

The critical challenge of cloud storage or cloud computing is the management of the continuously increasing volume of data.In the deduplication process, duplicate data is deleted, leaving only one copy of the data to be stored. Indexing of all data is still retained should that data ever be required. In general the data deduplication eliminates the duplicate copies of repeating data.

A. Cloud Computing

Cloud computing is a recently evolved computing terminology or metaphor based on utility and consumption of computing resources. Cloud computing involves deploying groups of remote servers and software networks that allow centralized data storage and online access to computer services or resources. Clouds can be classified as public, private or hybrid. The criticisms about it are mainly focused on its social implications. This happens when the owner of the remote servers is a person or organization other than the user, as their interests may point in different directions, for example, the user may wish that his or her information is kept private, but the owner of the remote servers may want to take advantage of it for their own business.

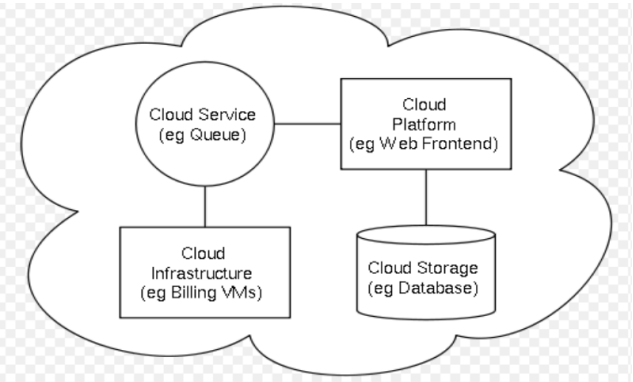

Fig.1. Architecture of cloud computing

B. Data De-Duplication

Typically, there is no one best way to implement data de-duplication across anentire an organization. Some forms of data de-duplication, such as compression, have been around for decades. Most recently, we have seen the introduction of sub-file de-duplication.These three [2] types of data de-duplication are described below.

Data Compression

Data compression works within a file to identify and remove empty space that appears as repetitive patterns. Data compression has been available for many years, but being isolated to each particular file, the benefits are limited when comparing data compression to other forms

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

of de-duplication. Data compression will not be effective in recognizing and eliminating duplicate files, but will independently compress each of the files.

Single-Instance Storage

Single-instance storage environments are able to detect and remove redundant copies of identical files. After a file is stored in a single-instance storage system than, all the otherreferences to same file, will refer to the original, single copy. Single-instance storage systems compare the content of files to determine if the incoming file is identical to an existing file in the storage system. While file-level de-duplication avoids storing files that are a duplicate of another file, many files that are considered unique by single-instance storage measurement may have a tremendous amount of redundancy within the files or between files.

Sub-file De-Duplication

Sub-file de-duplication detects redundant data within and across files as opposed to finding identical files as in SIS implementations. Using sub-file de-duplication, redundant copies of data are detected and are eliminated-even after the duplicated data exist, within separate files. Sub-file data de-duplication has tremendous benefits even where files are not identical, but have data elements that are already recognized somewhere in the organization. Fixed-length sub-file de-duplication uses an arbitrary fixed length of data to search for the duplicate data within the files. So most of the organizations widely use data depulication technology, which is also called as, single-instance storage, intelligent compression, and capacity optimized storage and data reduction.

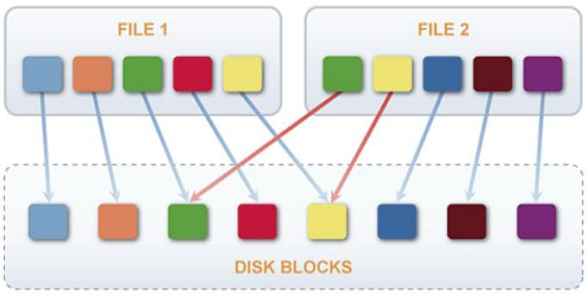

Fig 2: Example of De-duplication

B.1 Deduplication methods

One of the most common forms of data deduplication implementations works by comparing chunks of data to detect duplicates. For that to happen, each chunk of data is assigned an identification, calculated by the software, typically using cryptographic hash functions. In many implementations, the assumption is made that if the identification is identical, the data is identical, even though this cannot be true in all cases due to the pigeonhole principle; other implementations do not assume that two blocks of data with the same identifier are identical, but actually verify that data with the same identification is identical.[6] If the software either assumes that a given identification already exists in the deduplication namespace or actually verifies the identity of the two blocks of data, depending on the implementation, then it will replace that duplicate chunk with a link.

Once the data has been deduplicated, upon read back of the file, wherever a link is found, the system simply replaces that link with the referenced data chunk. The deduplication process is intended to be transparent to end users and applications.

- Chunking. Between commercial deduplication implementations, technology varies primarily in chunking method and in architecture. In some systems, chunks are defined by physical layer constraints (e.g. 4KB block size in WAFL). In some systems only complete files are compared, which is called Single Instance Storage or SIS. The most intelligent (but CPU intensive) method to chunking is generally considered to be sliding-block. In sliding block, a window is passed along the file stream to seek out more naturally occurring internal file boundaries.

- Client backup deduplication. This is the process where the deduplication hash calculations are initially created on the source (client) machines. Files that have identical hashes to files already in the target device are not sent, the target device just creates appropriate internal links to reference the duplicated data. The benefit of this is that it avoids data being unnecessarily sent across the network thereby reducing traffic load.

- Primary storage and secondary storage. By definition, primary storage systems are designed for optimal performance, rather than lowest possible cost. The design criteria for these systems is to increase performance, at the expense of other considerations. Moreover, primary storage systems are much less tolerant of any operation that can negatively impact performance. Also by definition, secondary storage systems contain primarily duplicate, or secondary copies of data. These copies of data are typically not used for actual production operations and as a result are more tolerant of some performance degradation, in exchange for increased efficiency.

To date, data deduplication has predominantly been used with secondary storage systems. The reasons for this are two-fold. First, data deduplication requires overhead to discover and remove the duplicate data. In primary storage systems, this overhead may impact performance. The second reason whydeduplication is applied to secondary data, is that secondary data tends to have more duplicate data. Backup application in particular commonly generate significant portions of duplicate data over time. Data deduplication has been deployed successfully with primary storage in some cases where the system design does not require significant overhead, or impact performance.

B.2 Benefits

- Storage-based data deduplication reduces the amount of storage needed for a given set of files. It is most

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

effective in applications where many copies of very similar or even identical data are stored on a single disk—a surprisingly common scenario. In the case of data backups, which routinely are performed to protect against data loss, most data in a given backup remain unchanged from the previous backup. Common backup systems try to exploit this by omitting (or hard linking) files that haven't changed or storing differences between files. Neither approach captures all redundancies, however. Hard-linking does not help with large files that have only changed in small ways, such as an email database; differences only find redundancies in adjacent versions of a single file (consider a section that was deleted and later added in again, or a logo image included in many documents).

- Network data deduplication is used to reduce the number of bytes that must be transferred between endpoints, which can reduce the amount of bandwidth required. See WAN optimization for more information.

- Virtual servers benefit from deduplication because it allows nominally separate system files for each virtual server to be coalesced into a single storage space. At the same time, if a given server customizes a file, deduplication will not change the files on the other servers—something that alternatives like hard links or shared disks do not offer. Backing up or making duplicate copies of virtual environments is similarly improved simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material.

II. LITERATURE REVIEW

Some existing works in this area is as follows:

Hybrid Cloud is the architecture that provides the Organization to efficiently work on both the private and public cloud architecture in combination by providing the scalability to adopt. Here some of the basic concepts and idea proposed byauthors and how best and easy to adopt this environment is explained by Neal Leavitt. [3]

An intelligent workload factoring, service for organization customers which makes the best use of the present publicCloud services including their private owned data centers. It allows the organization to work between the off-premisesand the on-premises infrastructure. The efficient core technology that is used for intelligent workload factoring is a fastredundant data element detection algorithm, that helps us factoring all the incoming requests based on the data contentand not only on volume of data, Hui Zhang, Guofei Jiang, Kenji Yoshihira, Haifeng Chen and AkhileshSaxena. [4]

The term ―Cloud‖ has many definitions one among them is to provide infrastructure as a service system where the ITinfrastructure will be deployed in the particular cloud service provider, data center as virtual machine. The growingpopularity of laas will help us to transform the organization present infrastructure into the required hybrid cloud orprivate cloud. OpenNebula Concept is being used that will provide the features that are not present in any other cloudsoftware, Borja Sotomayor ,Rubén S. Montero and Ignacio M. Llorente, Ian Foster. [5]

Data Deduplication is a technique that is mainly used for reducing the redundant data in the storage system which willunnecessarily 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. [6]

De-duplication is the technique that is most effective most widely used but when it is applied across the multiple usersthe cross-user deduplication tend to have to many serious privacy implications. Simple mechanisms can be used whichcan enable the cross-user deduplication which will reduce the risks of the data leakage and also some of the securityissues are discussed with how exactly to identify the files and to encrypt them while sending is discussed, Danny Harnik, Benny Pinkas, Alexandra Shulman- Peleg. [7]

M. Bellare [8] design a system, DupLESS that combines a CE-type scheme with the ability to obtain message-derived keys with the help of a key server (KS) shared amongst a group of clients. The clients interact with the KS by a protocol for oblivious PRFs, ensuring that the KS can cryptographically mix in secret material to the per-message keys while learning nothing about files stored by clients. These mechanisms ensure that DupLESS provides strong security against external attacks and that the security of DupLESS gracefully degrades in the face of comprised systems. Should a client be compromised, learning the plaintext underlyinganother client‟s cipher text requires mounting an online brute force attacks.

Aim of M. Bellare [9] is to formalize a new cryptographic primitive, Message-Locked Encryption (MLE), where the key under which encryption and decryption are performed is itself derived from the message. MLE provides a way to achieve secure de-duplication, a goal currently targeted by numerous cloud-storage providers. They provide definitions both for privacy and for a form of integrity that they call tag consistency. They provide ROM security analyses of a natural family of MLE schemes that includes deployed schemes. They make connections with deterministic encryption, hash functions secure on correlated inputs.

G. Neven [10] provides either security proofs or attacks for a large number of identity-based identification and signature schemes defined either explicitly or implicitly in existing literature. Underlying these is a framework that on the one hand helps explain how these schemes are derived and on the other hand enables modular security analyses, thereby helping to understand, simplify, and unify previous work. They also analyze a generic folklore construction that in particular yields identity-based identification and signature schemes without random oracles.

J. Xu [11] proposed growing need for secure cloud storage services and the attractive properties of the convergent cryptography lead us to combine them, thus, defining an innovative solution to the data outsourcing security and efficiency issues. Our solution is based on a cryptographic usage of symmetric encryption used for enciphering the

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

data file and asymmetric encryption for meta data files, due to the highest sensibility of these information towards several intrusions. In addition, thanks to the Merkle tree properties, this proposal is shown to support data de duplication, as it employs an pre-verification of data existence, in cloud servers, which is useful for saving bandwidth. Besides, our solution is also shown to be resistant to unauthorized access to data and to any data disclosure during sharing process, providing two levels of access control verification. Finally, we believe that cloud data storage security is still full of challenges and of paramount importance, and many research problems remain to be identified.

III. PROPOSED WORK

A. Overview

The main aim of proposed system is to efficiently solving the problem of deduplication with differential privileges in cloud computing and also provide an secured authorized deduplication using an OAuth protocol. We consider a hybrid cloud architecture consisting of a public cloud and a private cloud. Unlike existing data deduplication systems, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud. A new deduplication system supporting differential duplicate check is proposed under this hybrid cloud architecture where the S-CSP resides in the public cloud. The user is only allowed to perform the duplicate check for files marked with the corresponding privileges.And the system uses an OAuth protocol to support efficient authorization. OAuth acts as an intermediary on behalf of the end user, providing the service with an access token that authorizes specific account information to be shared.only the users which are authenticated in this way can undergone deduplication check.

B. Preliminaries

B.1 Symmetric encryption

Symmetric encryption uses a common secret key κ to encrypt and decrypt information. A symmetric encryption scheme consists of three primitive functions:

- $\text{KeyGen}_{SE}() \rightarrow \kappa$ is the key generation algorithm that generates κ using security parameter ;
- $\text{Enc}_{SE}(\kappa, M) \rightarrow C$ is the symmetric encryption algorithm that takes the secret κ and message M and then outputs the ciphertext C; and
- $\text{Dec}_{SE}(\kappa, C) \rightarrow M$ is the symmetric decryption algorithm that takes the secret κ and ciphertext C and then outputs the original message M.

B.2 Convergent encryption

Convergent encryption , provides data confidentiality in deduplication. A user derives a convergent key from eachoriginal data copy and encrypts the data copy with the convergent key. The user also derives a tag for the data copy, such that the tag will be used to detect duplicates. Here, we assume that the tag correctness property holds, i.e., if two data copies are the same, then their tags are the same. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored. Note that both the convergent key and the tag are independently derived, and the tag cannot be used to deduce the convergent key and compromise data confidentiality. Both the encrypted data copy and its corresponding tag will be stored on the server side.

- $\text{KeyGen}_{CE}(M) \rightarrow K$ is the key generation algorithm that maps a data copy M to a convergent key K;
- $\text{Enc}_{CE}(K, M) \rightarrow C$ is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs a ciphertext$C$;
- $\text{Dec}_{CE}(K, C) \rightarrow M$ is the decryption algorithm that takes both the ciphertext$C$ and the convergent key K as inputs and then outputs the original data copy M; and
- $\text{TagGen}(M) \rightarrow T(M)$ is the tag generation algorithm that maps the original data copy M and outputs a tag $T(M)$.

B.3 Proof of ownership

The notion of proof of ownership enables users to prove their ownership of data copies to the storage server. Specifically, PoW is implemented as an interactive algorithm run by a prover and a verifier. The verifier derives a short value $\phi(M)$ from a data copy M. To prove the ownership of the data copy M, the prover needs to send ϕ' to the verifier such that ϕ & prime;= $\phi(M)$. The formal security definition for PoW roughly follows the threat model in a content distribution network, where an attacker does not know the entire file, but has accomplices who have the file. The accomplices follow the "Bounded retrieval model", such that they can help the attacker obtain the file, subject to the constraint that they must send fewer bits than the initial min-entropy of the file to the attacker

B.4 Identification Protocol

An identification protocol can be described with two phases: Proof and Verify. In the stage of Proof, a prover/user U can demonstrate his identity to a verifier by performing some identification proof related to his identity. The verifier performs the verification with input of public informationpkU elated to skU. At the conclusion of the protocol, the verifier outputs either accept or reject to denote whether the proof is passed or not. There are many efficient identification protocols in literature, including certificate-based, identity-based identification etc.

C. System Model

C.1 OAuth

OAuth (Open Authorization) is an open protocol for token-based authentication and authorization on the Internet.OAuth protocol is used to ensure the users if they are securely authenticated or not. This protocol is used

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

above the architecture for authorized deduplication.The process for obtaining the token is called a flow.

OAuth is an authorization protocol - or in other words, a set of rules - that allows a third-party website or application to access a user's data without the user needing to share login credentials.

OAuth's open-source protocol enables users to share their data and resources stored on one site with another site under a secure authorization scheme based on a token-based authentication.

It is important to realize OAuth is a way to authenticate users to ensure a user is who he says he is. The protocol was designed to circumvent problems with the sharing of user credentials in distributed and Web 2.0 environments. With OAuth, resources stored on one website can be shared or accessed by a user once he is authenticated via OAuth. There is no need for the user to create a new account on the website and, at the same time, the website is not privy to the user's credentials.

OAuth operates much like a client/server computing model, where a primary website storing the user resources acts as a server and the website or application accessing that data is a client. The primary website establishes an OAuth interface (otherwise called an API) and secret key for the requesting website as a means of establishing a session to validate the user. Once the user requests access to the data or resources ofthe client website, he or she takes a side trip and is forwarded to the login procedure of the primary website, at which time the user provides his or her login credentials. Upon successful authentication there, an authorization token is sent from that primary website to the requesting website as an acknowledgment of authentication - allowing the user the access of data or other resources originally requested.

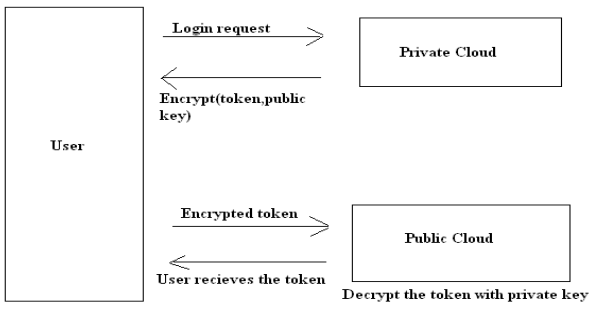

Fig.3.Token generation using OAuth

C.2 Hybrid Architecture for Secure Deduplication

At a high level, our setting of interest is an enterprise network, consisting of a group of affiliated clients who will use the S-CSP and store data with deduplication technique. In this setting, deduplication can be frequently used in these settings for data backup and disaster recovery applications while greatly reducing storage space. Such systems are widespread and are often more suitable to user file backup and synchronization applications than richer storage abstractions. There are three entities defined in our system, that is, users, private cloud and S-CSP in public cloud as shown in Fig.4 The S-CSP performs deduplication by checking if the contents of two files are the same and stores only one of them.[1]

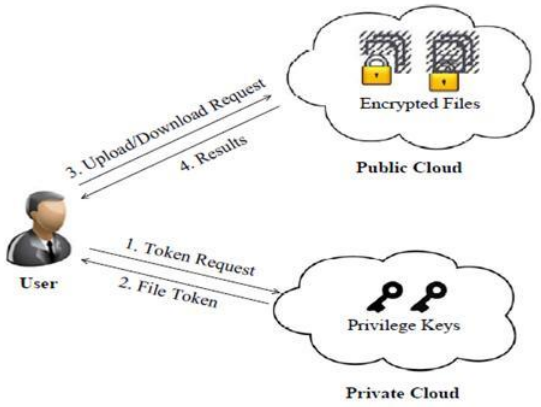

Fig.4. Architecture for Authorized Deduplication

• S-CSP

The S-CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data.

1. Start

2. Get unencrypted file tag

3. Accept the privilege based token from user

4. Validate the token and assert privilege level

5. Run deduplication check only on the privileged files

6. If the same tag is found along the privileged files, mark deduplication search as successful and grant access to the encrypted file

7. Stop

• Data Users

A user is an entity that wants to outsource data storage to the S-CSP and access the data later. 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. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system.

• 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/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud. The private keys for the privileges are

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

managed by the private cloud, who answers the file token requests from the users.

1. Start

2. Authenticate the user based on their credentials

3. Generate the unique privilege based token

4. Return the token to the user

5. Stop

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.

IV. CONCLUSION

Hybrid cloud architecture provides a lots of benefits with the use of both public and private cloud. Nowadays most of users use cloud to store data. Increasing amount of data in cloud is a major concern. In order to reduce the space and to efficiently utilize ,data deduplication is used. OAuth protocol enables the system to ensure the user is an authenticated person or not. Only such a person can perform deduplication check and store data in cloud.

REFERENCES

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, 2014, A Hybrid Cloud Approach for Secure Authorized Deduplication
[2] Pooja S Dodamani, Pradeep Nazareth, 2014, A Survey on Hybrid Cloud with De-Duplication
[3] Neal Leavitt ,2013 Hybrid Clouds Move to the Forefront.
[4] Hui Zhang, Guofei Jiang, Kenji Yoshihira, Haifeng Chen and AkhileshSaxena ,2009, A Hybrid Cloud ComputingModel
[5] Borja Sotomayor, Rubén S. Montero and Ignacio M. Llorente, Ian Foster ,2009, Virtual Infrastructure Management in Private and Hybrid Clouds.
[6] David Geer, 2008,Reducing the Storage Burden via Data Deduplication.computer.org.
[7] Danny Harnik, Benny Pinkas, Alexandra Shulman- Peleg , 2010, Side Channels in Cloud Services Deduplication in Cloud Storage.
[8] M. Bellare, S. Keelveedhi, and T. Ristenpart , 2013, Dupless: Server aided encryption for deduplicated storage.
[9] M. Bellare, S. Keelveedhi, and T. Ristenpart,2013, Message-locked encryption and secure deduplication.
[10] M. Bellare, C. Namprempre, and G. Neven,2009, Security proofs for identity-based identification and signature scheme.
[11] J. Xu, E.-C.Chang, and J. Zhou,2013,Weak leakage-resilient client-side deduplication of encrypted data in cloud storage.