DE-DUPLICATION OF DATA IN CLOUD

A.Revathi[1], Department of Computer Applications,
Mr. E. Ranjith[2], MCA, M.Phil., (Ph.D), Assistant Professor,
Krishnasamy College of Engineering and Technology, Cuddalore

ABSTRACT

Data deduplication is one of the techniques which used to solve the repetition of data.The deduplication techniques are generally used in the cloud server for reducing the space of the server. To prevent the unauthorized use of data accessing and create duplicate data on cloud the encryption technique to encrypt the data before stored on cloud server. Thus to overcome the security threats, this project proposes multiple cloud storage. Thus the common forms of data storage such as files and databases of a specific user is split and stored in the various cloud storages.

Keywords: De-duplication,Cloudcomputing.

INTRODUCTION

Cloud computing is one of the emerging technology, which helped severalorganizations to save money and time adding convenience to the end users. Thus the scopeof cloudstorage is vast because the organizations can virtually store their data's withoutbothering the entire mechanism.Cloud computing technology provides services such as Software as a service (SaaS), Platform as a service (PaaS), Infrastructure as a service (IaaS) to the user on demand. Data is growing at an alarming growth. So data deduplication is a fast growing technology because it reduces storage needs by storing only a single copy of the data. Encryption and deduplication are conflicting technologies and are challenging to implement. Also, deduplication offers the benefit of reduced replication, since only unique data is present in cloud. Cloud Computing provides key advantage to the end userslike cost savings, Able to access the data irrespective of location, performance and security.

OBJECTIVES

- Easy and fast to compute the hash value for any block
- Small changes in a block data should completely change the hash value so broadly that the new hash value looks uncorrelated with the previous hash value.
- Infeasible to find two dissimilar messages with the same hash value Because the hash value of the file should be unique as we know the properties of hash function.
- When performing the secure data deduplication, CSPcan check the ownership of the authorized user. Weexploit the dynamic count filters (DCF) to achieve thedata updating and improve the retrieval efficiency ofownership verifying.
- Security analysis shows that our proposed system issecure under the proposed security model, and performanceevaluation demonstrates the effectiveness andefficiency of our proposed system.

SYSTEM MODEL

The system model of SRRS consists of three entities: Users(U), a Cloud Service Provider (CSP) and a ManagementCenter (MC), A user sends a requestto MC, and encrypts the original file, then,outsources theciphertext to CSP.Users who belong to different role groups owning thecorresponding role keys. According to the role keys andaccess control policies, the user can upload or download thespecific files from CSP. The creator of a file is unique, and isalso a special user.A CSP is responsible for data storage, management andverification. CSP stores and manages the uploaded files fromusers. In terms of verification, CSP establishes a challengelist for each file to verify the user's ownership to prevent theunauthorized user's access.A MC is a trusted third party, which is responsible for userauthorization and role key management.

Our proposed system is secure under the standard model.Specifically, the probability of an adversary runninga successful secure protocol should be negligibleunder the secure parameter; the privacy informationof users can not be acknowledged by the CSP in theprocess of performing data deduplication; the adversarycan obtain a minimum amount Smin of the privacyinformation from the authorized user to run successfulsecurity protocol. Meanwhile, the system supports dynamicupdating and revoking of privileges to achieveflexible access control.

- The bandwidth, the server memory space, the clientstorage space should be efficient and economical. Inthe process of performing authorized deduplication, thebytes of file exchanged between the user

ADVERSARY MODEL

In our proposed system, we assume that: the MC istrusted by all entities involved in our system, and will notcompromised by an adversary; the CSP is "honest-butcurious(HbC)", which performs our proposed protocol honestly,but is curious about the user's privacy information;the communication channel between MC and the user issecure in our system;we define a secure hash functionresisted against the collision attack and exploit the standardsymmetric encryption algorithm.

DESIGN GOAL

We consider two aspects of system security and performanceefficiency to construct our system, therefore, the design goalincludes security and performance requirement, which isdescribed as follows:

ROLE AUTHORIZED TREE

A novel authorization structure named role authorizedtree (RAT) based on a B+ tree, MC organizes a role group using a RAT to manage the rolekey and achieve the user's authorization.

METHODOLOGY

The data de-duplication technique is used to store single instance of redundant data and eliminates the duplicate datain datacenter. It is used to decrease the size of datacenter and reduce the replication of data that were duplicated on cloud.The de-duplication process helps to remove any block or file that are not unique and store in smaller group of blocks .

The basic steps for data de-duplication process are.

- The files are converted into small segments.
- Then new and existing data are checked for redundancy
- Metadata are updated and segments are compressed.
- Duplicate data are deleted and check the data integrity.

Data-deduplication levels

Deduplication strategy can be categorized into two main strategies as follow, differentiated by the type of basic dataUnits.

File-level deduplication

A file is a data unit when examining the data of duplication, and it typically uses the hash value of the file as itsidentifier. If two or more files have the same hash value, they are assumed to have the same contents and only one ofthese files will be stored.

Advantage:

1. If any change is made in a file it makes to save the whole file again in file level deduplication.

2. In file level deduplication indexes are small, and so it takes less time for computational when it identifiesthe duplicate copies

Block-level deduplication

This strategy segments a file into several fixed-sized blocks or variable-sized blocks, and computes hash valuefor each block for examining the duplication blocks.

Advantage:

1.Block level deduplication can eliminate or delete the small redundant chunk of data when compared to whole file.

2. Each and every file system can use same deduplication algorithm in block level deduplication.

SYSTEM ARCHITECTURE

The main components of the proposed system are:

a) **Data owner**: the interface used by the client to use the cloud storage service.

b) **Cloud**: The server of the cloud service provider (CSP) where operations such as deduplication check using hashing is carried out and the data is stored.

The end user after using the credentials logs intothe interface on the CSP webpage. The data is uploadedinto the cloud from the client machine by the end user ofthe CSP and once the data is encrypted and then the data issent to the CSP server and deduplication is performed.Then, the data owners can view or edit the data but thedata customers can only request permission to view thedata through the privilege grant module and the randomaccesskey provided to them by the CSP, once theirrequest to view the document is accepted by the dataowner.

Data flow

The end user interacts with the front end or the website of the cloud storage service provider. During, the upload process a small application is downloaded on the client machine to encrypt the data with AES 128-bit encryption using a 15 bit private key. Once the data is

encrypted the data is broken down into smaller chunks depending on the chunk size for the various data types, specified in the backend of the system. The hash values are the calculated for the data chunks using MD5 hashing technique. If the data chunk which is being uploaded has the same hash value as the data which is already exists in the cloud, then the status of the data is uploaded as duplicate in the index table which is present in CSP database and the location of the existing file is mentioned in the index table, whereas if the data chunk is not a duplicate then the status is set as original and the file upload function is run to upload the new data into the cloud, and the location of the new data is updated on the cloud along with a file id for reference in the table. The index table is only visible to the database administrator in case a manual operation must be carried out on the files on the database.

Security analysis:

The proposed schemes implement a secure approach to data deduplication. CSP doesn't have access to the plain text any time as the file is getting encrypted from the Data Owner side. The CSP computes Hash value of cipher text.CSP and Users while functioning, without collusion guarantees that data is never compromised at the cloud storage. The data is being encrypted before reaching to CSP. Thus, CSP won't be able to know the actual data in 'm'. This is ensured as user won't share the actual data with the CSP.CSP storage will have only the data that is being passed on to CSP. This ensures that only (m) data is getting stored. In case CSP Storage is compromised, the attacker won't be able to get actual data as only Data Owner will be able to read it again using the CSP interfaceprovided in his machine.

CONCLUSION

Cloud computing has reached a maturity that leads it into a productive phase. This means that most of the main issues with cloud computing have been addressed to a degree that clouds have become interesting for full commercial exploitation. This however does not mean that all the problems listed above have actually been solved, only that the according risks can be tolerated to a certain degree. Managing encrypted data with deduplication is significant in practice for running a secure, dependable, and green cloud storage service, especially for big data processes. To secure the confidentiality of sensitive data during deduplication, the convergent encryption technique is used to encrypt the data before outsourcing.

SURVEY

1. M. Bellare, S. Keelveedhi, and T. Ristenpart, ―Message-Locked Encryption and Secure Deduplication,‖ Advances in Cryptology (EUROCRYPT 13), LNCS 7881, 2013, pp. 296–312.

2. J. Li et al., ―A Hybrid Cloud Approach for Secure Authorized Deduplication,‖ IEEE Trans. Parallel Distributed Systems, vol. 26, no. 5, 2015, pp.1206–1216.

3. Z. Wan, J. Liu, and R.H. Deng, ―HASBE: A Hierarchica Attribute-Based Solution for Flexible and Scalable Access Control in Cloud Computing,‖ IEEE Trans. Information Forensics and Security, vol. 7, no. 2, 2012, pp. 743–754.

4. M. Fu et al., ―Accelerating Restore and Garbage Collection in Deduplication-Based Backup Systems via Exploiting Historical Information,‖ Proc. Usenix Ann. Technical Conf., 2014, pp. 181–192.

5. M. Lillibridge, K. Eshghi, and D. Bhagwat, ―Improving Restore Speed for Backup Systems That Use Inline Chunk-Based Deduplication,‖ Proc. 11th Usenix Conf. File and Storage Technologies, 2013, pp. 183–198.

6. Z. Yan and M.J. Wang, ―Protect Pervasive Social Networking Based on Two Dimensional Trust Levels,‖ IEEE Systems J., Sept. 2014, pp. 1–12; doi: 10.1109/JSYST.2014.2347259.

7. Iuon –Chang Lin, Po-chingChien ,‖Data Deduplication Scheme for Cloud Storage‖ International Journal of Computer and Control(IJ3C),Vol1,No.2(2012).

8. Z. Yan, W. Ding, and H. Zhu, ―Manage Encrypted Data Storage with Deduplication in Cloud,‖ Proc. Int'l Conf. Algorithms and Architectures for Parallel Processing (ICA3PP), 2015, pp. 547–561.

9. Priyadharsini, Dhamodran, Kavitha,‖A Survey On Deduplication In Cloud Computing‖, IJCSMC, Vol. 3, Issue. 11,November 2014, pg.149 – 155.SURVEY ARTICLE