RESEARCH ARTICLE OPEN ACCESS

Optimized And Secure Data Backup Solution For Cloud Using Data Deduplication

Siva Ramakrishnan S(M.Tech)[1],Vinoth Kumar P (M.E)[2]

[1](Department Of Computer Science Engineering, SRM University, Chennai)

[2](Assistant Professor ,Department Of Computer Science Engineering, SRM University, Chennai)

--************************----------------------------------

Abstract:

In the advent of cloud computing technology, more and more customers are adopting a Cloud based approach for their infrastructure requirements which is being offered as an Infrastructure as a service model (IAAS).We will be presenting a model consisting of shared and dedicated storage which is differentiated by the cost involved in the services as well as the security mechanisms involved. Users can store the files either in a dedicated storage spaces which is highly secure as it is not physically shared among any other cloud users or opt to store the files securely in a shared storage along with files from other cloud users.

The infrastructure involves both dedicated and shared storage spaces, the user has the option to mandatorily store the file only in the dedicated storage or leave it to the system to decide the services to be used. In a cloud scenario security is a major concern as multiple users may use the same infrastructure. So we present an approach where we encrypt the data with a convergent key which is obtained by hashing the data copy. The users can retain the keys and send the cipher text to cloud after the key generation and encryption. Multiple users who own the same file would generate the same hash value, convergent key and the same cipher text. Users are provided with a pointer from the server, so the user need not upload the same file again.

Keywords: - **cloud computing; Deduplication; Convergent Encryption; Shared Storage, Dedicated Storage**

--************************----------------------------------

Introduction:

Though cloud computing has greatly reduced the cost involved in infrastructure spending, one can save valuable storage spaces at the public cloud by leveraging the concepts of deduplication and convergent encryption to greatly enhance the security in a cloud based setup. Infrastructure involves a setup which consists of two modes of storage – Shared and dedicated storage. Shared storage are storage spaces constituted of RAID disks which are shared among other cloud users but partitioned in such a way that users cannot see other user files. Dedicated storage are storage spaces which are built from RAID disks which are not shared with other users.

The difference between the two services is the cost involved, for example say a cloud vendor may charge 10$ for 1 GB of dedicated storage, whereas the same vendor would charge 5 $ for 1 GB of shared storage .This is a win-win scenario for both the vendor and the user as the user wins by paying less for the non-critical data by sharing the underlying storage among other users and the vendor wins as the system cost is shared by multiple cloud users.

The user has the option to mark the file as confidential and make it mandatory to save the file in the dedicated storage. If the file is not marked as confidential the system decides where to store file in shared or dedicated storage, here we implement the concept of deduplication to check for duplicated files among multiple cloud users and store the file only once in the shared storage space.

Security has always been a concern with cloud infrastructure so we introduce encryption in Deduplication, Traditional encryption requires different users to encrypt their data with their own keys. So identical data copies from different users will result in different cipher texts, making deduplication impossible. Convergent encryption [2] has been proposed to enforce data confidentiality while making deduplication feasible.

.Preliminaries:

In this section, we first define the notations used in this paper, the notations used in this paper are listed

In TABLE 1.

Acronym Description
S-CSP Secure-cloud service provider
PoW Proof of Owner
(*pLU, sLU*) User's public and secret key pair
LF Convergent encryption key for file *F*
P1U Privilege set of a user *U*
P1F Specified privilege set of a file *F*
ϕ' *F,p1*Token of file *F* with privilege *p1*

• KeySE(1λ) ! *L* is the key generation algorithm
That generates *l* using security parameter 1λ;
• EnSE(l,*M*) ! *C* is the symmetric encryption algorithmthat takes the secret *l* and message *M* and
Then outputs the ciphertext *C*; and

• DeSE(*κ*,*C*) ! *M* is the symmetric decryption algorithm that takes the secret land ciphertext *C* and
then outputs the original message *M*.

Convergent Encryption:

Convergent encryption is achieved by deriving the convergent key from each original data copy and encrypts the data copy with the convergent key. Also a tag is derived for the data copy by the users, the tag will be used to detect the duplicate copies. We conclude that the two data are the same when their tags are the same. The user first sends the tag to the server first to check for duplicate copies and verify if identical copies are already stored on the server. The convergent key and tag are independently derived and the convergent key cannot be found from the tag.

Proof of Ownership:

Proof of ownership allows the users to prove their ownership of the copies of the data to the secure cloud service provider(S-SCP) it is an interactive algorithm which executed by a prover who is the user and the verifier the storage server in this case. ϕ(M) is a short value derived from the data copy by the user. The prover needs to send ϕ' to the verifier such that ϕ'= ϕ(M). The proof of ownership almost follows the threat model in a content distribution network where the entire file is not known to the attacker but has the accomplices who have the file.

Identification Protocol:

The two phases of identification protocol are Proof and Verify. In Proof stage a user can demonstrate his identity to a verifier by performing identification proof related to his identity. The user inputs private key sLU which is sensitive like a private key of a public key in his certificate, which the user does not like to share. The verification is done by the verifier with input of the public information pLU related to sLU. The verifier outputs accept or reject to denote proof is passed or not.

System Model:

At a high level our setting of interest is a small and medium business (SMB) consisting of group of affiliated clients say employees of a company who will use the S-CSP and store data with a deduplication technique. This technique is mostly used as data backup solutions and disaster recovery applications to reduce the storage space greatly.

There are four entities defined in our system, that is, users, and Shared storage, dedicated Storage and S-CSP in public cloud as shown in Fig 1. The Secure cloud service provider performs deduplication by checking if the contents of two files are same and stores only any one of them.

The right to access a file is defined based on a set of privileges. The definition of a privilege varies across various applications. For example, we may define role based privilege according to the job Positions (e.g., Director, Project Head, and Technical Engineer), or we may define a time-based privilege that specifies a valid time period (e.g., 2015-01-01 to 2015-01-31) within which a file be accessed.

Each privilege of the user is represented in the form of a short Message called a token. Every file is associated with file tokens, which denotes the tag with the specified privileges. User computes and sends duplicate-check tokens to the public cloud for authorized duplicate check.

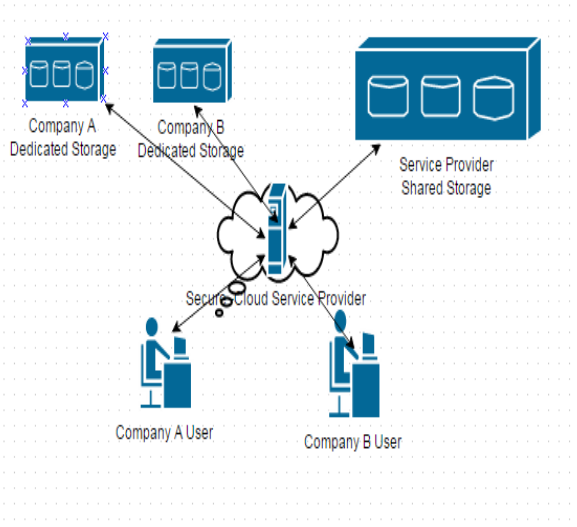

Figure 1 – System Model

Implementation:

The implementation involves a prototype of the proposed system where we model three modules as separate C++ programs.A client program is used for file upload whereas a private server program is used to model private cloud which manages the private keys and file tokens. Cryptographic opeartions of hashing and encryption are implemented by open ssllibrary[4], the inter communication between modules is based on http using GNU Linmicrohttpd and libcur

The following function calls are invoked in order to support token generation and deduplication process along with the file upload procedure.

* FileTag(ile) - SHA-1 hash of the File is computed as file tag.

- TokenReq(Tag, UserID) the private server is requested for file token generation with user ID and File tag.

- DupCheckReq(Token) duplicate check of the file is done by sending the file token received from private server.

• ShareTokenReq(Tag, {Priv.}) the private server is requested to generate share file token with file tag and target sharing privilege set.

• FileEncrypt(File) This function call encrypts the file with convergent encryption using 256 AES algorithm in cipher block chaining mode where the convergent key id SHA-256 hashing from the file.

• FileUploadReq(FileID, File, Token) This function call uploads the file data to the storage server if the file is unique and the file token stored is updated.the implementation of our private server includes request handlers for token generation and maintains key storage with hash map.

• TokenGen(Tag, UserID) - This function call loads the associated privilege keys of the user and generate the token along with HMAC-SHA-1 algorithm;

Evaluation:

Authorization steps like file token generation along with share token generation produces certain overhead, which is compared with the convergent encryption and share token generation. The evaluation is based on varying different overhead factors like 1) number of files stored 2) File Size 3) privilege set size 4) deduplication ratio. The evaluation was done with three machines of the below configuration with an Inter I5 2.66 Ghz CPU with 4 GB of RAM and installed with Ubuntu 14.04 connected over a 100 Mbps Ethernet network. The upload process is broken in to six steps namely,

* Tagging

* Token generating

* Deduplication check

*Sharetoken generation

* Encryption and

* Transfer.

Conclusion:

In this paper, the concept of convergent encryption was proposed in a public cloud infrastructure which enables the users to securely transfer the files to and fro from the Public cloud storage provider, we also had space savings by imposing Data deduplication techniques to the files which are redundantly stored by different users from different companies. The security analysis shows that our scheme is secure in terms of outsider attacks. As a proof of concept, A prototype has been implemented for proposed authorized deduplicate check scheme. The proposed model would lead to greater cost saving in a small and medium business setup as we store only confidential data in the dedicated storage space which is costlierthan the shared storage space where we store all possible redundantdata from the same company users as well as other company users. Also the implementation is secure to the extent that the users don’t have to worry about security attacks or loss of information.

Acknowledgements:

This paper is an extended version of [1] submitted on the category of Parallel and Distributed Systems, IEEE Transactions on (Volume:PP , Issue: 99)

References:

[1] A Hybrid Cloud Approach for Secure Authorized Deduplication Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou

[2] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer.Reclaiming space from duplicate files in a serverless distributed file system. In ICDCS, pages 617–624, 2002.

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013.

[4] OpenSSL Project. http://www.openssl.org/.

[5] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman.Role-based access control models. IEEE Computer, 29:38–47, Feb1996.

[6] R. D. Pietro and A. Sorniotti. Boosting efficiency and securityin proof of ownership for deduplication. In H. Y. Youm andY. Won, editors, ACM Symposium on Information, Computer andCommunications Security, pages 81–82. ACM, 2012.