An experimental study for integrity auditing and data de-duplicating in cloud storage

Sonal Purohit *,Rohit singh thakur

Abstract— Development of cloud computing technology and use of cloud service for data storage is increasing rapidly from last few years. Cloud services make use of increasing computing power and manage a huge amount of data. It takes heavy efforts in management of data stored over cloud server. As the data is stored on cloud server, security in cloud environment is an important issue for any customer. It becomes a challenge to achieve data de-duplication with integrity auditing in cloud servers. To solve these issues two secure systems are proposed in this paper that are S-Cloud and P-Cloud. Integrity auditing and maintenance of a Map Reduce cloud achieved by using S-Cloud, by which clients can generate data tags before uploading and audit the integrity of data having been stored in cloud. In proposed work, S-Cloud reduces the computation while file uploading and auditing phases. Clients of cloud service always worry about data security and integrity, of its data so that it needs to encrypt the data before the data is reached to the cloud server and only a single copy of data is maintained over there to achieve data de- duplication which can be achieved by the use of P-Cloud.

Index Terms Cloud Storage, Data De-duplication, Integrity auditing, SHA, MD5, AES Encryption, S-Cloud, P-Cloud

——————————   ◆   ——————————

1 Introduction

Cloud computing is the service where IT resources are provided on ”pay per use” basis. Cloud computing provides dynamically scalable infrastructure for application, data and file storage, due to these features many people influenced to store data in cloud.

In recent survey it is expected to reach 40 trillion GB (gigabytes) data in 2020, it implies that cloud storage is used by large number of customer worldwide. Cloud storage is provided and managed by third party so that multiple issues are arising , such as data protection, data recovery and availability, management capabilities, regulatory and compliances restrictions, integrity auditing and data de-duplication. Cloud storage fails to deal some important emerging needs for example the abilities to achieve integrity auditing and detecting replicated files present in cloud server.

Two major problems are focused in this paper, which are integrity auditing and data de-duplication.

In traditional in-house storage data is under the control of enterprise, but in cloud storage data is transferred over the internet and stored to random server space which generates concern on the integrity auditing of client files. Storage of Cloud is vulnerable from both side [1], and there some clouds spoofs the data without notifying to others just to maintain their position. Sometimes they discard the data only to save money and space in the cloud. In today era cost and space playing vital role in infrastructure network, so that the cloud servers might even actively and deliberately discard rarely accessed data files belonging to an ordinary client. Big data and clients constrained resource capabilities, the prime challenge as how can the client efficiently perform periodical integrity verifications even without the local copy of data files.

Second problem is de-duplication of data .Cloud services are availed by large amount of data at distributed servers. Files that being stored in different servers may be duplicated which leads to memory wastage in the cloud storage [2]. Therefore the challenge is to keeping only single copy for each file and to attach a pointer for each client who owns or demands to store same file, by maintaining the fact that no two owners of same duplicate files should be aware of the ownership by another client because this reveals that some other client has the exact same file, which could be sensitive information. Even should not be aware of the duplicity of its data. As per survey by EMC [3], 75% of recent digital data is a duplicated copy, that’s why this concept of de-duplicating the data is becomes promising requirement in the cloud computing environment.

Our research is based on integrity auditing and de-duplication in cloud, as demonstration S-Cloud and P-Cloud introduces integrity auditing with maintenance of a Map Reduce cloud, which helps clients generate data tags before uploading as well as audit the data integrity have been stored in cloud with reduction in computation for tag generation. P-Cloud is used to prevent the leakage of side channel information and design a proof of ownership protocol (POW) [4] between clients and servers, which allows clients to proof that they exactly owns or demands target data.

Client always want to encrypt their files before uploading to cloud storage which is achieved by using S-Cloud and the P-Cloud is used to ensure the file confidentiality and give the assurance of de-duplication in cloud storage based on a static, hash code (i.e. short value)[4].

SECTIONS

Further this paper is organized as follows: In Section II is review then integrity auditing and de-duplication related work. In Section III, we introduce some background. Section IV and Section V respectively propose S-Cloud and P-Cloud. Section VI security analysis. Finally Section VII draws the conclusion of this paper VIII describes future related work.

II. RELATED WORK

As our work is related with the following major problems in the cloud storage system, let us discuss them one by one.

Integrity auditing- Cloud Storage relieves the client from storage, management and maintenance of large amount of data. But at the same time arise a lot of security issue as the client doesn't have control over it because of the storage at data at uncertain domain. In process of uploading and downloading, the data could deliberately be tampered, get corrupt or modified in between [5] considered a new cloud storage architecture with two independent servers for integrity auditing to reduce the computation load at client side. Recently, Li et al. [6] utilized the key-disperse paradigm to fix the issue of a significant number of convergent keys in convergent encryption. Here proposed system checks the correctness of the data [7].

Data de-duplication- Data de-duplication is a kind of data compression technique for eliminating duplicate copies of replicated data in storage. Data de-duplication technique is use to reduce the network bandwidth and also manages the storage. De-duplication eliminates redundant data by keeping only a single copy and pointing other redundant data to that copy. De-Duplication identifies duplicate files in the cloud [8]. Also the same file can be uploaded by different users [2]. Due to this there is the duplication of file, resulting in wastage the scarce storage resource. [2] De-Duplication ensures duplicate data [6] is physically stored only once, and the proof of ownership with complete transparency provided in the system. Convergent encryption [9] is a promising cryptographic primitive for ensuring data privacy in de-duplication.

Security- cloud storage avail various files with different size, it is hard to check whether which file is corrupt (tampered) or not. These challenges comes under every cloud deployment model(i.e. .private, public, hybrid),and to address these issues TPA manage audit of miscellaneous data along with timestamp details.

Infrastructure Cost- As the traditional system works the cost for cloud storage is increasing. Computation task in cloud storage is also there in traditional storage. Here proposed system is trying to provide data de-duplication with integrity auditing in minimal cost basis.

III. PRELIMINARY

Bilinear Map and Computational Assumption: Definition 1 (Bilinear Map): Let G and G_T be two cyclic multiplicative groups of large prime order p. A bilinear pairing is a map $e : G \times \rightarrow GT$ with the following properties:

- Bilinear: $e(g_1{}^a ; g_2{}^b) = e(g_1; g_2)^{ab}$ for all $g_1; g_2 \in R\ G$ and a; b $\in_R$ Z p;
- Non-degenerate: There exists $g_1; g_2 \in G$ such that $e(g_1; g_2) \neq 1$;
- Computable: There exists efficient algorithm to compute $e(g_1; g_2)$ for all $g_1; g_2 \in_R G$.

The examples of such groups can be found in super singular elliptic curves or hyper elliptic curves over finite fields, and the bilinear pairings can be derived from the Weil or Tate pairings. For more details, see [12].

Here describe the Computational Diffie-Hellman problem, the hardness of which will be the basis of the security of our proposed schemes.

Definition 2 (CDH Problem): The Computational Diffie- Hellman problem is that, given $g; g^x; g^y \in G_1$ for unknown $x; y \in Z^*{}_p$, to compute g^{xy}.

Convergent Encryption: Convergent encryption [14][15][13] provides data confidentiality in de-duplication. A user derives a convergent key from the data content and encrypts the data copy with the convergent key. In addition, the user 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 [14] holds, i.e., if two data copies are the same, then their tags are the same. Formally, a convergent encryption scheme can be defined with four primitive functions:

- Key Gen(F) : The key generation algorithm takes a file content F as input and outputs the convergent key ckF of F;
- Encrypt (ckF;F) : The encryption algorithm takes the convergent key ckF and file content F as input and outputs the cipher text ct F;
- Decrypt (ckF; ctF) : The decryption algorithm takes the convergent key ckF and cipher text ct F as input and outputs the plain file F;
- Tag Gen(F) : The tag generation algorithm takes a file content F as input and outputs the tag F of F. Notice that in this paper, we also allow Tag Gen(·) to generate the (same) tag from the corresponding cipher text as with [14][16].

Proposed System: Our research based on cloud system namely S-Cloud and P-Cloud to achieve better performance and QOS(quality of service) parameters for accessing massive data on cloud. Firstly in the plain data file it performs integrity auditing with achieving data de-duplication and then perform both integrity auditing and de-duplication on that encrypted data file**.** Proposed architecture of S-cloud and P-cloud discuss below-

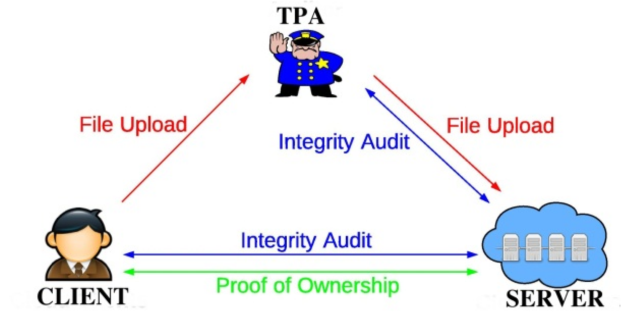
IV. S-CLOUD

S-Cloud is a system in which of integrity auditing and secure de-duplication are only present in plain files.S-Cloud system have following entities:

Client – Client have large data files to be stored and rely on the cloud storage for data maintenance and management. Either individual consumers or commercial organizations can be there**.**

Server- Clients may buy or lease storage capacity from servers, and store their individual data in these bought or rented spaces for future utilization. Server are virtualized pool of resources which client need when required**.**

Third Party Auditor (TPA) –TPA helps clients upload and audit their data maintains a Map Reduce cloud and acts like a certificate authority.Its public key is made available to the other entities in the system. The S-Cloud system supporting file-level de-duplication includes the various protocol which covers in this paper.Fig(1) have S-cloud architectural design:



Fig(1)S-Cloud Architecture

(i)**File Uploading Protocol –** Allow the client to upload the file. Protocol includes three phases

Phase 1 (client →server): client performs the duplicate check with the server before uploading a file. If there is a duplicate, third protocol runs between client and server. Otherwise, the following protocols (including phase 2 and phase 3) are run between these two.

Phase 2 (client → TPA): client uploads files to the TPA, and receives a receipt.

Phase 3 (TPA →server): TPA helps generate a set of tags for the uploading file, and send them along with this file to cloud server.

(ii)**Integrity Auditing Protocol**-It is a protocol for integrity verification and allowed to be initialized by any except the server.

In this protocol, the server is prover, while the TPA or client is verifier. Protocol includes two phases

Phase 1 (client/TPA → server): client or TPA generates a set of challenges and sends them to the server.

Phase 2 (server →client/TPA): based on the stored files and file tags, server tries to prove that it exactly owns the target file by sending the proof back to client or TPA.

(iii)**Proof of Ownership Protocol**: This protocol initialized at the server for verifying that the client owns a target file. Triggered along with file uploading protocol to prevent the leakage of side channel information. Server is verifier, and client is prover here. Includes two phases

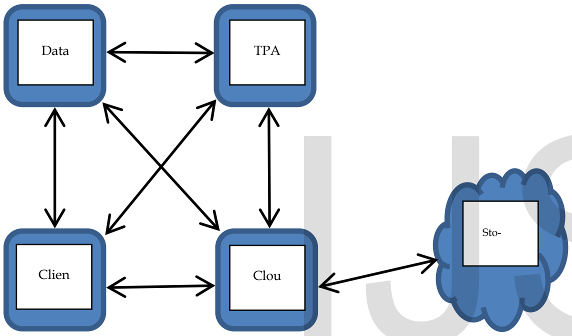
Phase 1 (server → client): server sends some challenges to the client.

Phase 2 (client →server): client responds with the proof of file ownership, and server verifies the validity of proof.

V. P-CLOUD

P-Cloud is used for maintaining integrity auditing and managing de-duplication on encrypted files. The only difference is the file uploading protocol in P-Cloud involves additional phase for communication between client servers. P-Cloud also having the same three protocols(i.e., the file uploading protocol, the integrity auditing protocol and the proof of ownership protocol) Only one difference is there between P-Cloud and S-Cloud. In P-Cloud an additional phase for communication between client and key server is involves in file uploading protocol. Client needs to communicate with the key server to get the key for encrypting the uploading file before the phase 2 in S-Cloud. Another design goal of file confidentiality is required.

.



Fig(2) P-Cloud Architecture

• **File Confidentiality**- The design goal of file confidentiality requires preventing the servers from accessing the content of files. Specially, we require that the goal of file confidentiality needs to be resistant to dictionary attack means if client have a knowledge of dictionary which include all possible data files it cannot able to recover the data file[19].

VI Security Analysis

Algorithms: Generally following algorithms used to provide data encryption

.

AES algorithm: Advanced Encryption Standard (AES) is an encryption standard (symmetric key encryption). In our work author uses the AES algorithm for the encryption purpose .P-Cloud use AES algorithm for encrypting the client data which is stored in the cloud storage. This ensures security measures [11] in this work. When client wants to upload their data files first it encrypts the file by using AES algorithm. With the help of symmetric keys, the same key can encrypt and decrypt.

MD5 algorithm: Message Digest is one way where a master fingerprint has been generated for the purpose of providing a message authentication code (hash code) [17].Hashing is done to generate a unique hash value for each data item being uploaded [10]. This hash value is used for auditing purpose to identify duplicate files [8]. The same hash value is used as a checksum to ensure data integrity [10]. MD5 is the extension of MD4 algorithm has three rounds which make it quite faster in comparison to MD4. MD5 is one way hash function that deals with security features.

Secure Hash Algorithm (SHA): SHA is also a hashing algorithm. It is used for important information like password hashing. SHA is used to create digital signatures of the important data files. When SHA algorithm is running on the data, it generates a hash value/signature (i.e short value). If the data changes in any way, the hash value/ signature will not match and thus we would able to know that the data has been corrupted, tampered or modified with.

MD5 and SHA algorithms are cryptography hash functions and used to generate hash values. They take a piece of data, compact it and create a suitably unique output called hash value/signature. The difference between lies is in what algorithm they use to create the hash value. SHA is more secure than MD5. MD5 is now broken as a way was discovered to easily generate collisions and should not be used nor trusted anymore [18] .Here in our work security is the major concern so here SHA algorithm is used.

VII CONCLUSION

Here we propose two systems S-Cloud and P-Cloud. S-Cloud introduces an auditing entity with maintenance of a Map Reduce cloud, which helps clients generate data tags. In this paper we can achieve integrity auditing and de-duplication simultaneously. We need to store only one copy of encrypted data. To deduplicate the authentication tags generated by different owners, we aggregate the tags. The integrity of deduplicated data can be correctly checked by the TPA on behalf of any owner. The major goal is to help the users to store their data on the cloud with confidentiality and security. Compared with previous work, the computation by user in S-Cloud is greatly reduced during the file uploading and auditing phases. S-Cloud uses proof-of-ownership protocol for secure data, de-duplication also prevent from side channel information leakage on internet. P-Cloud is an advanced method for S-Cloud motivated by the fact that customers always want to encrypt their data before uploading allow secure integrity auditing and data de-duplication on that encrypted data. The analysis and experimental results show that our scheme is secure and efficient.

VIII FUTURE WORK

Data de-duplication and integrity auditing have vast popularity in cloud storage and further it will be increase. Our research are gaining better results for textual data in which author has taken several text file for upload/download. S-Cloud and P-Cloud perform very well for textual data and this performance will continue for multimedia such as audio, video etc data in cloud. Another future aspect related with this paper as author has used SHA algorithm for encryption but due to its complexity it hard to understand for new comers . To overcome this drawback, modified version of convergent encryption can be used by introducing two approaches - domain separation and cryptographic tuning. This gives a better authorized de-duplication approach.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, “A view of cloud computing,” Communication of the ACM, vol. 53, no. 4, pp. 50–58, 2010.

[2] S. Keelveedhi, M. Bellare, and T. Ristenpart, “Dupless: Serveraided encryption for deduplicate storage,” in Proceedings of the 22Nd USENIX Conference on Security, ser. SEC’13. Washington, D.C.: USENIX Association, 2013, pp. 179–194. [Online].

[3] J. Yuan and S. Yu, “Secure and constant cost public cloud storage auditing with deduplication,” in IEEE Conference on Communications and Network Security (CNS), 2013, pp. 145–153.

[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, “Proofs of ownership in remote storage systems,” in Proceedings of the 18th ACM Conference on Computer and Communications Security. ACM, 2011, pp. 491–500.

[5] J. Li, X. Tan, X. Chen, and D. Wong, “An efficient proof of retrievability with public auditing in cloud computing,” in 5th International Conference on Intelligent Networking and Collaborative Systems (INCoS), 2013, pp. 93–98.reliable convergent key management,” IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 6, pp. 1615–1625, June 2014.

[6] Mehdi Sookhak a,n, HamidTalebian a, EjazAhmed a, AbdullahGani , Muhammad Khurram Khan. ”A review on remote data auditing on single cloud server”www. elsevier.com/locate/jnca

[7] R Sravan Kumar &A.Saxena “Integrity auditing and Proofs in Cloud Storage”

[8]https://view.officeapps.live.com,http://www.cs.sjsu.

[9] J. Douceur, A. Adya, W. Bolosky, P. Simon, and M. Theimer, “Reclaiming space from duplicate files in a serverless distributed file system,” in 22nd International Conference on Distributed Computing Systems, 2002, pp. 617–624.

[10] A.K.Dubey, N Namdev and S.S.Shrivastava “Cloud-user security based on RSA and MD5 algorithm for resource attestation and sharing in java environment

[11] AnthonyVelte& Robert C.Elsenpeter “Cloud Computing a Practical Approach”, McGraw-Hill, Inc. New York, NY, USA ©2010

[12] D. Boneh and M. Franklin, “Identity-based encryption from the weil pairing,” in Advances in Cryptology — CRYPTO 2001, ser. Lecture Notes in Computer Science, J. Kilian, Ed. Springer Berlin Heidelberg, 2001, vol. 2139, pp. 213–229.

[13] J. Douceur, A. Adya, W. Bolosky, P. Simon, and M. Theimer, “Reclaiming space from duplicate files in a serverless distributed file system,” in 22nd International Conference on Distributed Computing Systems, 2002, pp. 617–624.

[14] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Message-locked encryption and secure deduplication,” in Advances in Cryptology – EUROCRYPT 2013, ser. Lecture Notes in Computer Science, T. Johansson and P. Nguyen, Eds. Springer Berlin Heidelberg, 2013, vol. 7881, pp. 296–312.

[15] M. Abadi, D. Boneh, I. Mironov, A. Raghunathan, and G. Segev, “Message-locked encryption for lock-dependent messages,” in Advances in Cryptology – CRYPTO 2013, ser. Lecture Notes in Computer Science, R. Canetti and J. Garay, Eds. Springer Berlin Heidelberg, 2013, vol. 8042, pp. 374–391.

[16] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, “Secure deduplication with efficient and reliable convergent key management,” IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 6, pp. 1615–1625, June 2014.

[17] William Stallings, Cryptography and NetworkSecurity: Priciples and Practice,5th Edit ion Prent ice Hall; 5 edit ion (January 24, 2010).

[18] Piyush Gupta et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4492-4495 A Comparative Analysis of SHA and MD5 Algorithm

[19] S. Keelveedhi, M. Bellare, and T. Ristenpart, “Dupless: Serveraided encryption for deduplicated storage,” in Proceedings of the 22Nd USENIX Conference on Security, ser. SEC’13. Washington, D.C.: USENIX Association, 2013, pp. 179–194. [Online].
https://www.usenix.org/conference/usenixsecurity13/technicalsessions/presentation/bellare

.