ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Implementation of deduplication scheme for Encrypted Data in Cloud

Aishwarya M B , Ashritha A, Chetana C H, Mamatha S, Snigdha Sen

UG Student, Computer Science & Engineering, Global Academy of Technology, Bengaluru, India

UG Student, Computer Science & Engineering, Global Academy of Technology, Bengaluru, India

UG Student, Computer Science & Engineering, Global Academy of Technology, Bengaluru, India

UG Student, Computer Science & Engineering, Global Academy of Technology, Bengaluru, India

Assistant Professor, Computer Science & Engineering, Global Academy of Technology, Bengaluru, India

ABSTRACT: Cloud computing is the delivery of computing services—servers, storage, databases, networking, software, analytics and more—over the Internet. The most important and useful cloud service is data storage. As data need to be stored in cloud, to preserve privacy of data holders, data are often stored in an encrypted form. However, encrypted data can cause data deduplication, which eliminates redundant data segments from the backup and reduces the size of the backup data. It not only reduces the storage space requirements, but also reduces the data that is transferred over the network resulting in faster and efficient data protection operations. Generally Traditional deduplication schemes suffer from security weakness. In this paper, we try to implement a deduplication scheme for encrypted data stored in cloud.

KEYWORDS: Cloud computing, encrypted data, HDFS, Deduplication

I. INTRODUCTION

cloud storage is a service model in which data is maintained, managed, backed up remotely and made available to users over a network. data centre of a cloud service provider (csp) is used when cloud users want to upload personal or confidential data. but intrusions and attacks are inevitable. moreover due to the rapid development of data mining and other analysis technologies, the privacy issue becomes serious. therefore there is a requirement to encrypt data before it gets stored in cloud. but sometimes it is obvious that the same or different users may upload duplicated data in encrypted form to csp[1], especially for scenarios where data are shared among many users. although cloud storage space is huge, still data duplication greatly wastes network resources, consumes a lot of energy, and complicates data management.therefore deduplication becomes critical for big data storage and processing in the cloud. deduplication has proved to achieve high cost savings, e.g., reducing up to 90-95% storage[2] needs for backup applications and up to 68% in standard file systems

II. RELATED WORK

In paper[3] authors proposed the concept of Designated-Verifier Provable Data Possession (DV-PDP). Here they have given DV-PDP system model and formal DV-PDP security model. In DV-PDP, data owners can designate a verifier to verify data integrity of his data. The verifier is stateless and independent from CSP, which solves the problem that the verifier can be controlled by the malicious CSP. In our design, we propose to use ECC-based homomorphism authenticator to design PDP scheme, which does not compute expensive bilinear and consume small amount of calculation and Communications. This scheme is very suitable for mobile clouds. The advantages of this scheme are through security analysis and performance analysis. This scheme is secure and highly efficient. But sometimes the private information is not needed. Private PDP is necessary in some cases.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Checking data possession in networked[4] information systems such as those related to critical infrastructures (power facilities, airports, data vaults, defense systems, and so forth) are described . Here author presented a new remote data possession checking protocol. Using a remote data possession checking protocol, the data vault customer might be able to periodically verify that the data vault provider is storing a current and complete copy or backup of the critical files. Remote data possession checking is an important component of intrusion detection systems (IDS) used to detect server corruption.. In a real deployment of a verification system using our protocol or any remote integrity checking protocol, the time to access files should be taken into account when evaluating performance. However, this is unlikely to have substantial impact, as a state-of-the-art hard drive technology allows as much as 1 Mbyte to be read in as little time as a few nanoseconds. The price paid is that the proof of possession obtained in this way is probabilistic, that is, requirement R2 above is not met with 100 percent probability.
Paper [5] described storage outsourcing is a rising trend which prompts a number of interesting security issues, many of which have been extensively investigated in the past. Prior work has addressed this problem using either public key cryptography or requiring the client to outsource its data in encrypted form. Here they constructed a highly efficient and provably secure PDP technique based entirely on symmetric key cryptography, while not requiring any bulk encryption while not requiring any bulk encryption. Also, in contrast with its predecessors, our PDP technique allows outsourcing of dynamic data, i.e, it efficiently supports operations, such as block modification, deletion and append. They developed and presented a step-by-step design of a very light-weight and provably secure PDP scheme. It surpasses prior work on several counts, including storage, bandwidth and computation overheads as well as the support for dynamic operations. However, since it is based upon symmetric key cryptography, it is unsuitable for public (third-party) verification. A natural solution to this would be a hybrid scheme combining

Paper[6] discussed about solving open problem with a novel scheme based on techniques including polynomial-based authentication tags and homomorphism linear authenticators. This design allows de-duplication of both files and their corresponding authentication tags. Data integrity auditing and storage de-duplication are achieved simultaneously. This scheme is also characterized by constant real time communication and computational cost on the user side. Public auditing and batch auditing are both supported. Hence, our proposed scheme out performs existing POR and PDP schemes while providing the additional functionality of de-duplication. We prove the security of our proposed scheme based on the Computational Diffie-Hellman problem, the Static Diffie-Hellman problem and the t-Strong Diffie-Hellman problem. Numerical analysis and experimental results on Amazon AWS show that our scheme is efficient and scalable.This polynomial based authentication tag can also be used as an independent solution for other related applications.

III. PROPOSED METHODOLOGY

Here we try to implement a deduplication scheme on encrypted data at CSP by applying PRE to issue keys to different authorized data holders based on data ownership challenge. It is applicable in scenarios where data holders are not available for deduplication control. Data duplication occurs at the time when data holder tries to store the same data that has been stored already at CSP. In case that is updated by a data owner with and the new encrypted raw data is provided to CSP to replace old storage for the reason of achieving better security, CSP issues the new reencrypted data to all data holders with the support of AP. This scheme provides a secure approach to protect and deduplicate the data stored in cloud by concealing plaintext from both CSP and AP. Fig 1 describes overall system architecture.

- File selection
- Hash Key Generation
- Encryption
- File Uploading
- Deduplication
- File Downloading
- Decryption

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

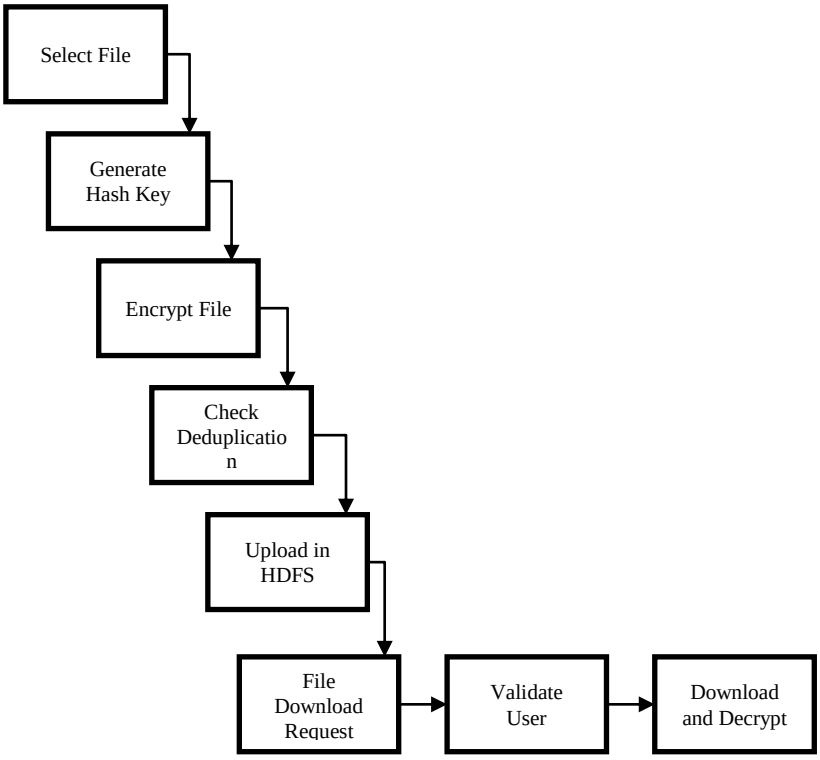

Fig 1. Overall System Architecture

File Selection: For data maintenance and computation Cloud Clients have large data files to be stored and rely on the cloud. They can be either individual consumers or commercial organizations; the resources are virtualized by Cloud Servers according to the requirements of clients and expose them as storage pools. The cloud clients may buy or lease storage capacity from cloud servers, and store their individual data in these bought or rented spaces for future utilization. Fig 2 describes file selection steps

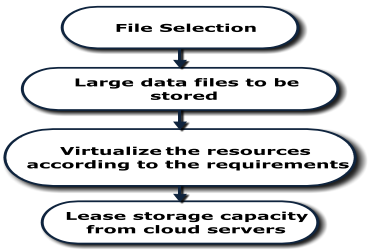

Fig 2. File selection steps

Hash Key Generation: By sending hash value of the file Hash (F) the client runs the deduplication test to the cloud server. If there is a duplicate, the cloud client performs Proof of Ownership protocol with the cloud server. If it is passed, the user is authorized to access this stored file without uploading the file. The auditor does almost the same thing as that in SecCoud. Firstly, he computes the hash of ciphertext {ctBij} and sends it to the cloud storage server for duplicate check. The auditor performs a PoW, if there is a duplicate stored in the cloud server, and the details are described in the Proof of Ownership protocol. It has been shown in Fig 3

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

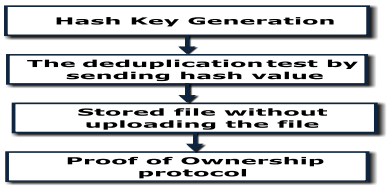

Fig 3. Hash key generation

Encryption: On convergent encryption, the system make a modification such that the convergent key of file is generated and controlled by a secret "seed", such that any adversary could not directly derive the convergent key from the content of file and the dictionary attack is prevented. In deduplication Convergent encryption provides data confidentiality. From the data content and encrypts the data copy with the convergent key a user (or data owner) derives a 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 holds, i.e., if two data copies are the same, then their tags are the same. Fig 4 shows encryption process

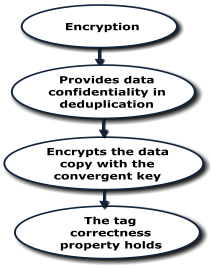

Fig 4. Encryption

File Uploading: This protocol including three phases. Client performs the duplicate check with the cloud server to confirm if such a file is stored in cloud storage or not before uploading a file in phase 1. If there is a duplicate, another protocol called Proof of Ownership will be run between the client and the cloud storage server. In phase 2 client uploads files to the auditor, and receives a receipt from auditor. In phase 3 auditor helps generate a set of tags for the uploading file, and send them along with this file to cloud server.

Deduplication: The design goal of this work is secure deduplication. Because, it requires that the cloud server is able to reduce the storage space by keeping only one copy of the same file. This objective is distinguished from previous work in that we propose a method for allowing both deduplication over files and tags regarding to secure deduplication. Similarly, we can also define a game between challenger and adversary for secure deduplication below. Notice that the game for secure deduplication captures the intuition of allowing the malicious client to claim it has a challenge file F through colluding with all the other clients not owning this file.

File Downloading:It should not represent a major additional cost to traditional cloud storage that the computational overhead for providing integrity auditing and secure deduplication, nor should they alter the way either uploading or downloading operation.

Decryption: The decryption steps shown in Fig 5 takes the convergent key ckFandciphertextctF as input and outputs the plain file F;

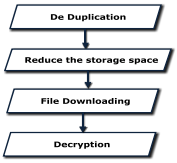

Fig 5. Decryption

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

IV. PSEUDO CODE

Hash key generation algorithm
Step 1: create an instance of messagedigest using sha-256
Step 2: using messagedigest instance from step 1 generate hash key
Step 3: read the file and convert it into bytes
Step 4: apply hash key on bytes
Step 5: convert hash value to hexadecimal string
Step 6: save the hash value and display
Deduplication check algorithm
Step 1: connect to client pc
Step 2: retrieve the hash key
Step 3: compare the hash key with the hash key stored in auditor's repository
If hash key matched then
Duplicate occurred
Else
No duplication
Step 4: Send file status of duplication to the client

Tag Generation Algorithm
Step 1: Create an instance of SecretKeySpec by passing the encryption standard which is used to generate the tag
Step 2: Convert the tag to string format
Step 3: Display the tag on the text field

CAPTCHA Generation Algorithm
Step 1: Initialize a string that contains uppercase and lowercase alphabets and numbers
Step 2: Generate a random string of length 5 characters
Step 3: Display the code on the label

File Encryption Algorithm
Step 1: Apply AES algorithm to encrypt the file contents
Step 2: Display the encrypted file contents on the text area

File Decryption Algorithm
Step 1: Apply AES algorithm to decrypt the encrypted file
Step 2: Display the original file contents on the text area

V. RESULTS

Following are few snapshots from our implementation. Fig 6 describes when user receives the status of the file duplication from the auditor. Here in this snapshot it shows no duplication for the uploaded file is found.

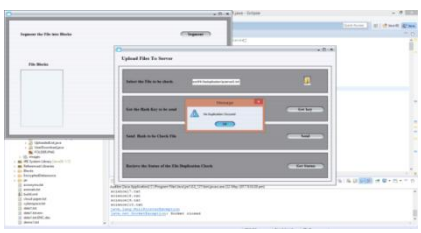

Fig 6. File duplication status page

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Fig7 describes about the snapshot when a new user wants to upload/download file he has to register first.

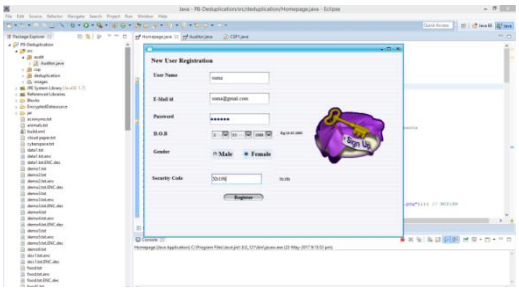

Fig 7. New user registration

Fig8 discusses about file got uploaded successfully.

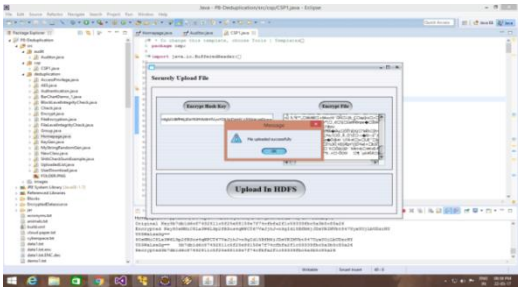

Fig 8. User uploading the file on cloud

VI. CONCLUSION AND FUTURE WORK

Future work includes optimizing this design and implementation for practical deployment and studying verifiable computation to ensure that CSP behaves as expected in deduplication management. Deduplication with encrypted data is significant in practice for achieving a successful cloud storage service, especially for big data storage. In this paper, we tried to impelement a practical scheme to manage the encrypted big data in cloud with deduplication. This scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption. Extensive performance analysis and test showed that our scheme is secure and efficient under the described security model and very suitable for big data deduplication

REFERENCES

[1] Zheng Yan,Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng,"Deduplication on Encrypted Big Data in Cloud",IEEE TRANSACTIONS,2016

[2] D.T. Meyer and W.J Bolosky, "A Study of Practical Deduplication,"ACM Transactions on Storage, 7(4), pp. 1-20, 2012,doi:10.1145/2078861.2078864

[3] 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.

[4] J.Yuan and S. Yu, "Secure and constant cost public cloud storage auditing with de-duplication," in IEEE Conference on Communications and Network Security (CNS), 2013, pp. 145–153.

[5] 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.

[6] S. Keelveedhi, M. Bellare, and T. Ristenpart, "Dupless: Server-aided encryption for de-duplicated storage," in Proceedings of the22Nd USENIX Conference on Security, ser. SEC'13. Washington,D.C.: USENIX Association, 2013, pp. 179–194.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

[7] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson ,and D. Song, "Provable data possession at untrusted stores," in Proceedings of the 14th ACM Conference on Computer and Communications Security, ser. CCS '07. New York, NY, USA: ACM, 2007, pp. 598–609.
[8] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner,Z. Peterson, and D. Song, "Remote data checking using provable data possession, "ACM Trans. Inf. Syst. Secur., vol. 14, no. 1, pp. 12:1–12:34,2011.
[9] G. Ateniese, R. Di Pietro, L. V. Mancini, and G. Tsudik,"Scalable and efficient provable data possession," in Proceedings of the 4thInternational Conference on Security and Privacy in Communication Netowrks, ser. Secure Comm '08. New York, NY, USA: ACM, 2008,pp. 9:1–9:10.
[10] C. Erway, A. K¨upc¸¨u, C. Papamanthou, and R. Tamassia, "Dynamic provable data possession," in Proceedings of the 16th ACM Conference on Computer and Communications Security, ser. CCS '09. New York ,NY, USA: ACM, 2009, pp. 213–222.
[11]F. Seb´e, J. Domingo-Ferrer, A. Martinez-Balleste, Y. Deswarte, andJ.-J. Quisquater,"Efficient remote data possession checking in critical information infrastructures," IEEE Trans. on Knowl. and Data Eng.,vol. 20, no. 8, pp. 1034–1038, 2008.
[12] H. Wang,"Proxy provable data possession in public clouds," IEEE Transactions on Services Computing, vol. 6, no. 4, pp. 551–559, 2013.
[13]Y. Zhu, H. Hu, G.-J. Ahn, and M. Yu, "Cooperative provable data possession for integrity verification in multi-cloud storage, "IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 12, pp. 2231–2244, 2012.
[14] H. Shacham and B. Waters, "Compact proofs of retrievability," in Proceedings of the 14th International Conference on the Theory and Application of Cryptology and Information Security: Advances in Cryptology, ser. ASIACRYPT '08.Springer Berlin Heidelberg, 2008, pp.90–107.

BIOGRAPHY

AISHWARYA M B is a final year UG student of Computer science and engineering department, Global academy of Technology. Her interest includes cloud computing,IoT and Big data

ASHRITHA A is a final year UG student of Computer science and engineering department, Global academy of Technology. Her interest includes cloud computing, netwoking and IoT

Chetana C H is a final year UG student of Computer science and engineering department, Global academy of Technology. Her interest includes cloud computing,IoT and Big data

Mamatha S is a final year UG student of Computer science and engineering department, Global academy of Technology. Her interest includes cloud computing,IoT and Big data.

Snigdha Sen is working as a Assistant Professor in Computer science and engineering department, Global academy of Technology. Her research interest includes cloud computing, IoT and ethical hacking