Secure Duplication Detection in Cloud using Chunk Based Technique: Survey

Bhagyashri Badgujar[1] Pallavi Kalase[2] Shyamli Bharati[3] Kajol Singh[4]
[1,2,3,4]BE Student
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Indira College of Engineering & Management, Pune, Maharashtra, India

Abstract— Cloud computing is the long dreamed vision of computing as a utility. Besides all the benefits of the cloud computing security of the stored data need to be considered while storing sensitive data on cloud. Cloud users cannot rely only on cloud service provider for security of their sensitive data stored on cloud. To achieve optimal usage of storage resources, many cloud storage providers perform de-duplication, which exploits data redundancy and avoids storing duplicated data from multiple users. System proposes a new approach to achieve more efficient Deduplication for (encrypted) large files. Our approach, named Block-Level Message-Locked Encryption (BL-MLE), can achieve file-level and block-level Deduplication, block key management, and proof of ownership simultaneously using a small set of metadata.

Key words: Third Party Authenticator, AES Algorithm, RSA Algorithm, SHA 512 Algorithms, Deduplication

I. INTRODUCTION

Uploading large files would consume extensive bandwidth; source-based Deduplication seems to be a better choice for large file outsourcing. Unlike target-based Deduplication which requires users to upload their files regardless of the potential data redundancy among those files, source-based Deduplication could save the bandwidth significantly by eliminating the retransmission of duplicated data.

Deduplication system, the user firstly sends a file identifier to the server for file redundancy checking. If the file to-be-stored is duplicated in the server, the user should convince the server that he/she indeed owns the file. Otherwise, the user uploads the identifiers/tag of all the file blocks to the server for block-level Deduplication checking. Finally, the user uploads data blocks which are not stored in the server.

Proof-of-Ownership (PoW) is necessary for source-based Deduplication. PoW is an interactive protocol between a proverb (file owner) and a verifier (data server). By executing the protocol, the proverb convinces the verifier that he/she is an owner of a file stored by the verifier. PoW protocol in which presents three schemes that differs in terms of security and performance.

In the previous encryption data privacy, is opposing the Deduplication happens file level and block level. The replicated copies of the same file is eradicated by file level Deduplication .For the block level duplication which eliminates duplicates blocks of data that occur in non-identical files.

II. LITERATURE REVIEW

"Data Deduplication is the technique which used to reduce redundancy in the storage data, there are two types of strategies are used for Deduplication purpose one is file level Deduplication, block level data Deduplication. In file level Deduplication, single instance storage is used to perform Deduplication task. In block level data Deduplication, data files are divided into blocks and these blocks are compared to either these blocks are contains same value or not. That way the task of data Deduplication is performed.

Wen Xia et al discuss the main challenges facing large-scale data reduction is how to maximally detect and eliminate redundancy at very low overheads. In this paper, we present DARE, a low-overhead Deduplication-aware resemblance detection and elimination scheme that effectively exploits existing duplicate-adjacency information for highly efficient resemblance detection in data Deduplication based backup/archiving storage systems. The main idea behind DARE is to employ a scheme, call Duplicate-Adjacency based Resemblance Detection (DupAdj), by considering any two data chunks to be similar (i.e., candidates for delta compression) if their respective adjacent data chunks are duplicate in a Deduplication system, and then further enhance the resemblance detection efficiency by an improved super-feature approach.

A three-tier cross-domain architecture, with an proficient and privacy-preserving big data Deduplication in cloud storage referred to as EPCDD achieves both privacy-preserving and data availability, and resists brute force attacks. In addition, the accountability can take into consideration to offer better privacy assurances than existing schemes [4].

The paper [5] the protocol that prevents unauthorized access by using a secure proof of ownership protocol. The protocol uses authorize de duplicate check for hybrid cloud architecture.

Data is of prime importance for individuals as well as for organizations. [11] As the amount of data being generated increases exponentially with time, duplicate data contents being stored cannot be tolerated. Thus, employing storage optimization techniques is an essential requirement to large storage areas like cloud storage. Deduplication is a one such storage optimization technique that avoids storing duplicate copies of data. Currently, to ensure security, data stored in cloud as well as other large storage areas are in an encrypted format and one problem with that is, we cannot apply Deduplication technique over such an encrypted data.

Stanek et al. introduced the concept of "data popularity" arguing that data known/owned by many users do not require as strong protection as unpopular data; based on this, presented an encryption scheme, where the initially semantically secure cipher text of a file is transparently downgraded to a convergent cipher text that allows for Deduplication as soon as the file becomes popular. In this paper we propose an enhanced version of the original scheme. Focusing on practicality, we modify the original scheme to improve its efficiency and emphasize clear functionality. The

efficiency based on popularity properties of real datasets and provides a detailed performance evaluation, including comparison to alternative schemes in real-like settings. Importantly, the new scheme moves the handling of sensitive decryption shares and popularity state information out of the cloud storage, allowing for improved security notion, simpler security proofs and easier adoption.

Bhagyashree Bhoyane et al introduced Cloud computing is the long dreamed vision of computing as a utility. Besides all the benefits of the cloud computing security of the stored data need to be considered while storing sensitive data on cloud. Cloud users cannot rely only on cloud service provider for security of their sensitive data stored on cloud.To achieve optimal usage of storage resources, many cloud storage providers perform de-duplication, which exploits data redundancy and avoids storing duplicated data from multiple users.

Akhila K et al the amount of data being generated increases exponentially with time, duplicate data contents being stored cannot be tolerated. Thus, employing storage optimization techniques is an essential requirement to large storage areas like cloud storage. Deduplication is one such storage optimization technique that avoids storing duplicate copies of data. Currently, to ensure security, data stored in cloud as well as other large storage areas are in an encrypted format and one problem with that is, we cannot apply Deduplication technique over such an encrypted data.

III. ALGORITHMS

A. AES Algorithm

AES is depending on a intended standard called as a substitution-permutation network, and is quick in both software and hardware.[8] Unlike its predecessor DES, AES does not use a Feistel network. AES is a variant of Rijndael which has a fixed block size of 128 bits, and a key size of 128, 192, or 256 bits. The multiple of 32 bits, both with a minimum of 128 and a maximum of 256 bits.

AES operates on a 4×4 column-major order matrix of bytes, termed the state, although some versions of Rijndael have a larger block size and have additional columns in the state. The key size used for an AES cipher specifies the number of repetitions of transformation rounds that convert the input, called the plaintext, into the final output, called the cipher text. The number of cycles of replication is given:

- 10 cycles of replication for 128-bit keys.
- 12 cycles of replication for 192-bit keys.
- 14 cycles of replication for 256-bit keys.

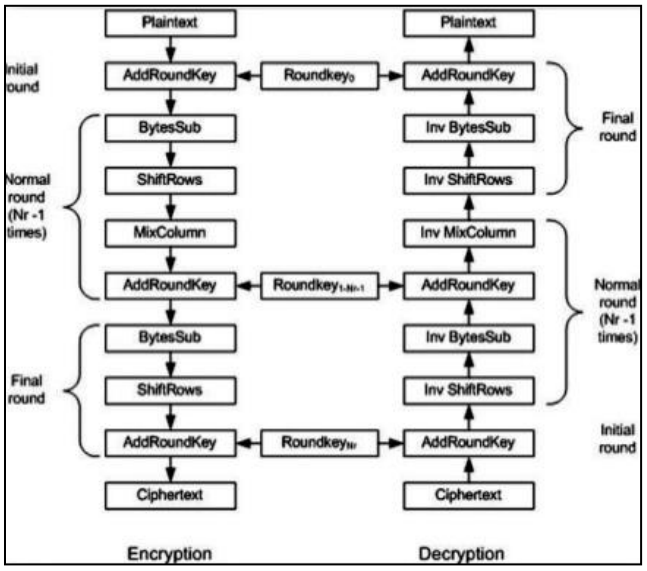

Fig. 1: AES Algorithm Flowchart

Each round consists of several processing steps, each containing four similar but different stages, including one that depends on the encryption key itself. A set of reverse rounds are applied to transform cipher text back into the original plaintext using the same encryption.

B. RSA Algorithm

Key Generation

The keys for the RSA algorithm are generated the following way:

1) Choose two distinct prime numbers p and q.
– For security purposes, the integer's p and q should be chosen at random, and should be similar in magnitude but 'differ in length by a few digits to make factoring harder. Prime integers can be efficiently found using a primarily test.
2) Compute n = pq.
– n is used as the modulus for both the public and private keys. Its length, usually expressed in bits, is the key length.
3) Compute $\varphi(n) = \varphi(p)\varphi(q) = (p - 1)(q - 1) = n - (p + q - 1)$, where φ is Euler's totient function. This value is kept private.
4) Choose an integer e such that $1 < e < \varphi(n)$ and gcd (e, $\varphi(n)$) = 1; i.e., e and $\varphi(n)$ are coprime.
5) Determine d as $d \equiv e{-}1 \pmod{\varphi(n)}$; i.e., d is the modular multiplicative inverse of e (modulo $\varphi(n)$)
– This is more clearly stated as: solve for d given $d \cdot e \equiv 1 \pmod{\varphi(n)}$
– e having a short bit-length and small Hamming weight results in more efficient encryption – most commonly 216 + 1 = 65,537. However, much smaller values of e (such as 3) have been shown to be less secure in some settings.
– e is released as the public key exponent.
– d is kept as the private key exponent.
– The public key consists of the modulus n and the public (or encryption) exponent e. The private key consists of the modulus n and the private (or decryption) exponent d, which must be kept secret. p, q, and $\varphi(n)$ must also be kept secret because they can be used to calculate d.

C. SHA Algorithm

In cryptography, SHA-1 (Secure Hash Algorithm 1) is a cryptographic hash function designed by the United States National Security Agency and is a U.S. Federal Information Processing Standard published by the United States NIST. SHA-1 produces a 160-bit (20-byte) hash value known as a message digest. A SHA-1 hash value is typically rendered as a hexadecimal number, 40 digits long.

Image Description: One iteration within the SHA-1 compression function: A, B, C, D and E are 32-bit words of the state; F is a nonlinear function that varies; n denotes a left bit rotation by n places; n varies for each operation; Wt is the expanded message word of round t; Kt is the round constant of round t; denotes addition modulo 232.

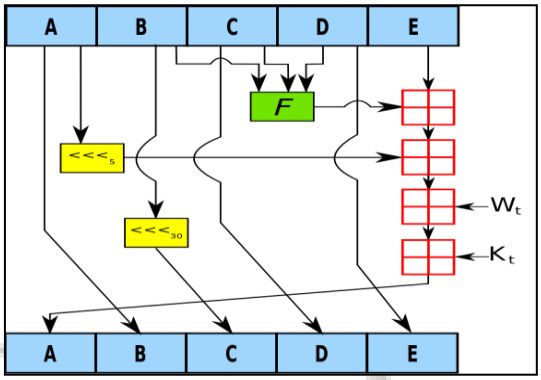

Figure: SHA Algorithm

IV. ADVANTAGES

1) Provides robust security to personal data.
2) Maintain User data Integrity to highest levels.
3) Protect privacy of user by making file inaccessible to any unauthorized personnel.
4) Multi-party approval helps in file usage control.

V. APPLICATIONS

1) Data security Application over cloud.

VI. CONCLUSION

This work survey and analyzed the concept of Deduplication which can provide more space savings than file-level Deduplication does in large file storage. This paper exploits the survey on data redundancy and avoids storing duplicated data from multiple users System focus on the Deduplication

REFERENCES

[1] Wen Xia, Member,Hong Jiang "DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads", ,IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016.
[2] Zheng Yan, Wenxiu Ding,Xixun Yu,"Deduplication on Encrypted Big Data in Cloud",IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 2, APRIL-JUNE 2016.
[3] Rongmao Chen,Yi Mu,"BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication",IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.
[4] "Xue Yang,Rongxing Lu,"Achieving Efficient and Privacy-Preserving Cross-Domain Big Data Deduplication in Cloud",IEEE Transactions on Big Data.
[5] " Mr.Vinod B Jadhav ,Prof.Vinod S Wadne "Secured Authorized De-duplication Based Hybrid Cloud Approach" International Journal of Advanced Research in Computer Science and Software Engineering – 2014.
[6] Aparna Ajit Patil, Asst. Prof. Dhanashree Kulkarni "Block Level Data Duplication on Hybrid Cloud Storage System" International Journal of Advanced Research in Computer Science and Software Engineering – 2015.
[7] Pasquale Puzio, Refik Molva, Melek O¨ nen, Sergio Loureiro, "ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage".
[8] Chunlu Wang, Jun Ni, Tao Xu, Dapeng Ju "TH_Cloudkey: Fast, Secure and lowcost backup system for using public cloud storage" IEEE 2013.
[9] Aparna Ajit Patil, Asst. Prof. Dhanashree Kulkarni "Block Level Data Duplication on Hybrid Cloud Storage System" 2015, IJARCSSE.
[10] Jan Stanek, and Lukas Kencl," Enhanced Secure Thresholded Data Deduplication Scheme for Cloud Storage". IEEE 2016.
[11] Akhila Ka ,Amal Ganesha, Sunitha Ca, "A Study on Deduplication Techniques over Encrypted Data" Elsevier 2016.

