IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A SECURE FRAMEWORK FOR CLIENT-SIDE DEDUPLICATION IN BIGDATA APPLICATIONS

[1]C.Surya, [2]S.Manimegalai, [3]P.Meena, [4]A.Kowsalya

[1]Assistant Professor,[2]UG Student,[3]UG Student, [4]UG Student
[1]Computer Science and Engineering,
[1]SriRamakrishna College of Engineering, Perambalur, India
[2,3,4]Computer Science and Engineering,
[2,3,4]SriRamakrishna College of Engineering, Perambalur, India

Abstract: Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data deduplication. Different from traditional deduplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. We also present several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement a prototype of our proposed authorized duplicate check scheme and conduct tested experiments using our prototype. We show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

***Index Terms* – Deduplication, minimal overhead, convergent encryption, duplicate check.**

I. Introduction

Cloud computing provides seemingly unlimited “virtualized” resources to users as services across the whole Internet, while hiding platform and implementation details. One critical challenge of cloud storage services is the management of the ever-increasing volume of data. To make data management scalable in cloud computing, deduplication has been a well-known technique and has attracted more and more attention recently. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage.

Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Deduplication can take place at either the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files. Although data deduplication brings a lot of benefits, security and privacy concerns arise as users’ sensitive data are susceptible to both inside and outside attacks.

Thus, identical data copies of different users will lead to different ciphertexts, making deduplication impossible. Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts decrypts a data copy with a *convergent key*, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the ciphertext to the cloud. Since the encryption operation is deterministic and is derived from the data content, identical data copies will generate the same convergent key and hence the same cipher text. To prevent unauthorized access, a secure proof of ownership protocol is also needed to provide the proof that the user indeed owns the same file when a duplicate is found.

The user is able to find a duplicate for this file if and only if there is a copy of this file and a matched privilege stored in cloud. For example, in a company, many different privileges will be assigned to employees. In order to save cost and efficiently management, the data will be moved to the storage server provider (SCSP) in the public cloud with specified privileges and the deduplication technique will be applied to store only one copy of the same file. Because of privacy consideration, some files will be encrypted and allowed the duplicate check by employees with specified privileges to realize the access control. Traditional deduplication systems based on convergent encryption, although providing confidentiality to some extent; do not support the duplicate check with differential privileges. In other words, no differential privileges have been considered in the deduplication based on convergent encryption technique. It seems to be contradicted if we want to realize both deduplication and differential authorization duplicate check at the same time.

To prevent unauthorized access, a secure proof of ownership protocol is also needed to provide the proof that the user indeed owns the same file when a duplicate is found. After the proof, subsequent users with the same file will be provided a pointer from the server without needing to upload the same file. A user can download the encrypted file with the pointer from the server, which can only be decrypted by the corresponding data owners with their convergent keys.

II. LITERATURE SURVEY:

This section reviews the literatures on various data deduplication schemes that are carried out by different researches.
P. Anderson and L. Zhang. Fast proposed an algorithm which takes advantage of the data which is common between users to increase the speed of backups, and reduce the storage requirements. This algorithm supports client-end per-user encryption which is necessary for confidential personal data. Finally, they discussed the use of the prototype in conjunction with remote cloud storage, and present an analysis of the typical cost savings[1].

M. Bellare, S. Keelveedhi, and T. Ristenpart proposed an architecture that provides secure deduplicated storage resisting brute-force attacks, and realize it in a system called DupLESS. In DupLESS, clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol. It enables clients to store encrypted data with an existing service, have the service perform deduplication on their behalf, and yet achieves strong confidentiality guarantees.[2]

S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider approaches the user could communicate with a trusted cloud(either a private cloud or a cloud based on multiple secure hardware modules) which encrypts and verifies the data stored and operations performed in the untrusted commercial cloud. By separating the computations the trusted cloud is principally used for security-critical operations in the less time-critical setup phase, while the entrusted commercial cloud employed for processing large quantity of computation, thus realizing the secure outsourcing of cloud-computing data.[3]

J. Li, X. Chen, M. Li, J. Li, P. Lee introduced a baseline approach in which each user holds an independent master key for encrypting the convergent keys and outsourcing them to the cloud. However, such a baseline key management scheme generates an enormous number of keys with the increasing number of users and requires users to dedicatedly protect the master keys. To this end, they proposed Dekey , a new construction in which users do not need to manage any keys on their own but instead securely distribute the convergent key shares across multiple servers. Security analysis demonstrates that Dekey is secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement Dekey using the Ramp secret sharing scheme and demonstrate that Dekey incurs limited overhead in realistic environments.[4]

C. Ng and P. Lee proposed *RevDedup*, proposed a deduplication system that optimizes reads to the latest backups of virtual machine (VM) images using reverse deduplication. In contrast with conventional deduplication that removes duplicates from new data, RevDedup removes duplicates from old data, thereby shifting fragmentation to old data while keeping the layout of new data as sequential as possible. They, evaluate our RevDedup prototype using a 12-week span of real-world VM image snapshots of 160 users. Hence showed that RevDedup achieves high deduplication efficiency, high backup throughput, and high read throughput.[5]

Wen Xia ; Hong Jiang ; Dan Feng ; Yu Hua proposed SiLo, a near-exact and scalable deduplication system that effectively and complementarily exploits similarity and locality of data streams to achieve high duplicate elimination, throughput, and well balanced load at extremely low RAM overhead. The main idea behind SiLo is to expose and exploit more similarity by grouping strongly correlated small files into a segment and segmenting large files, and to leverage the locality in the data stream by grouping contiguous segments into blocks to capture similar and duplicate data missed by the probabilistic similarity detection. SiLo also employs a locality based stateless routing algorithm to parallelize and distribute data blocks to multiple backup nodes. By judiciously enhancing similarity through the exploitation of locality and vice versa, SiLo is able to significantly reduce RAM usage for index-lookup, achieve the near-exact efficiency of duplicate elimination, maintain a high deduplication throughput, and obtain load balance among backup nodes.[6]

Nesrine Kaaniche and Maryline Laurent implemented cloud-based services for large scale content storage, processing, and distribution. Security and privacy are among top concerns for the public cloud environments. Towards these security challenges, they propose and implement, on OpenStack Swift, a new client-side deduplication scheme for securely storing and sharing outsourced data via the public cloud. The originality of our proposal is twofold. First, it ensures better confidentiality towards unauthorized users. That is, every client computes a per data key to encrypt the data that he intends to store in the cloud. As such, the data access is managed by the data owner. Second, by integrating access rights in metadata file, an authorized user can decipher an encrypted file only with his private key.[7]

Chuan Lin ; Qiang Cao ; Jianzhong Huang ; Jie Yao ; Xiaoqian Li ; Changsheng Xie presented HPDV, a highly parallel deduplication cluster for VM images, which well utilizes the parallelism to achieve high throughput with minimum interference on the foreground VM services. The main idea behind HPDV is to exploit idle CPU resource of VM servers to parallelize the compute-intensive chunking and fingerprinting, and to parallelize the I/O-intensive fingerprint indexing in the deduplication servers by dividing the globally shared fingerprint index into multiple independent sub-indexes according to the operating systems of VM images. To ensure the quality of VM services, a resource-aware scheduler is proposed to dynamically adjust the number of parallel chunking and fingerprinting threads according to the CPU utilization of VM servers. Our evaluation results demonstrate that compared to a state-of-the-art deduplication system for VM images called Light, HPDV achieves up to 67% deduplication throughput improvement.[8]

III. SYSTEM ANALYSIS

3.1 Existing System

Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To protect the confidentiality of sensitive data while supporting deduplication, Cloud computing provides seemingly unlimited "virtualized" resources to users as services across the whole Internet, while hiding platform and implementation details. Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified *privileges*, which define the access rights of the stored data.

3.2 Proposed System

The convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data deduplication. Different from traditional deduplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. We also present several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement a prototype of our proposed authorized duplicate check scheme and conduct tested experiments using our prototype. We show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

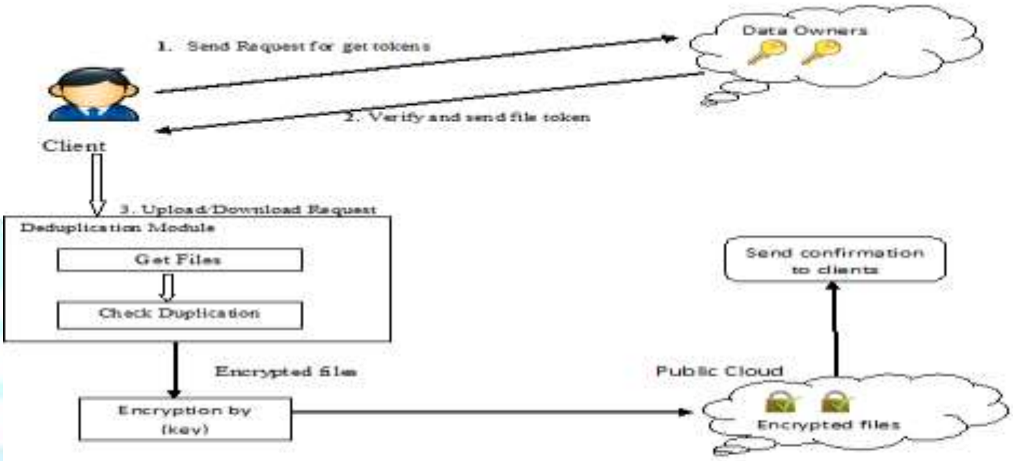
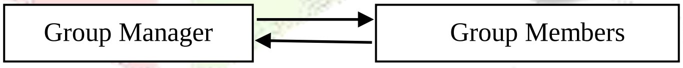


Figure 3.1: System Architecture

IV. RESEARCH METHODOLOGY

4.1Registration

For the registration of user with identity ID the group manager randomly selects a number. Then the group manager adds into the group user list which will be used in the traceability phase. After the registration, user obtains a private key which will be used for group signature generation and file decryption.



3.2 File Uploading

The canonical application is data uploading. The data property is especially useful when we expect the delegation to be efficient and flexible. The schemes enable a content provider to share her data in a confidential and selective way, with a fixed and small cipher text expansion, by distributing to each authorized user. The file upload with to the data cloud storage for distribution during upload data comparison process is done with the using the md5 hash value for checking file content to avoid de duplication attack file already present or not.

3.3 Naïve Bayes Classifier

Binary classifiers are generated for each class of event using relevant features for the class and classification algorithm .Binary classifiers are derived from the training sample by considering all classes other than the current class as other, e.g., Normal will consider two classes: normal and other. The purpose of this phase is to select different features for different classes by applying the information gain or gain ratio in order to identify relevant features for each binary classifier.

3.4 AES Encryption

To protect the client's privacy, we apply the anonymous AES in branching programs. To reduce the decryption complexity due to the use of AES, we apply recently proposed decryption outsourcing with privacy protection to shift client's pairing computation to the cloud server. The adversary launches Key Generate algorithms to query for as many private keys as he wants, which correspond to attribute sets A1, . . . ,Aq being disjoint in chargedby all authorities {Ak }, but none of these keys satisfy T0.Besides, he also conducts arbitrarily many computations using the public and secret keys that he has (belonging to compromised authorities).

IV. RESULTS AND DISCUSSION

In this paper, the notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct tested experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

REFERENCES

[1] M.Shyamala Devi, V.Vimal Khanna and A.NaveenBhalaji "EnhancedDynamic Whole File De Duplication for space optimization in privatecloud, Vol 4, Augest 2014..

[2] A. Pasqual Puzi, B. RefikMolva "Bloak level De-Duplication withEncrypted Data "OGCC Vol 1, 2014

[3] Edna Dias Canedo,Rafael Timoteo de sousa, "Trust Model For ReliableFile Exchange in Cloud Computing" Vol. 4, Feb 2012

[4] Jinji ,yankili,Patrick P.C " A hybrid cloud approach for secureauthorized de-duplication " IEEE Vol PP No : 99 2014

[5] Dinesh H.A, Agrawal V.K "Multilevel accessing technique for cloudservice "Vol 2, 2012

[6] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplicationwith efficient and reliable convergent key management. In IEEETransactions on Parallel and Distributed Systems, 2013

[7] C. Ng and P. Lee. Revdedup: A reverse deduplication storagesystem optimized for reads to latest backups. In Proc. of APSYS,Apr 2013

[8] Chuan Lin ; Qiang Cao ; Jianzhong Huang ; Jie Yao ; Xiaoqian Li ; Changsheng Xie, "HPDV:A Highly Parallel Deduplication Cluster for Virtual Machine Images"