IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 6, June 2015.
www.ijiset.com
ISSN 2348 – 7968

Secure Authorized Deduplication by Hybrid Cloud Approach

Divya V Kandalkar, Prof. R.R.Bhambare

Department of Electronics and Telecommunication, Savitribai Phule Pune University,
Sir Visveswaraya Institute of Technology, Nasik, India

Abstract

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. 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 a 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 testbed experiments using our prototype. We show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

Keywords: Deduplication, authorized duplicate check, confidentiality, hybrid cloud

1. Introduction

With the rapid development of processing and storage technologies and the success of the Internet, computing resources have become cheaper, more powerful and moreubiquitously available than ever before. This technological trend has enabled the realization of a new computing model called cloud computing, in which resources (e.g., CPU and storage) are provided as general utilities that can be leased and released by users through the Internet in an on-demand fashion. The term Deduplication (also Data Deduplication) describes a popular technique that allows cloud storage providers to signicantly decrease the amount of needed storage space. The principle of deduplication is as follows: only a single copy of each piece of data is stored. If a user wants to store data that the cloud storage provider already has stored in the past, the storage provider simply creates a link to that data instead of storing another copy. There are some variations of how deduplication may be realized:

(1) File level deduplication vs. block level deduplication. File level deduplication means that only a single copy of each file will be stored. Block level deduplication means that each file will be split up into blocks and only a single copy of each block will be stored. Identical files or blocks are detected by comparing the hash value" with a list of known files or blocks.

(2) Server-side deduplication vs. client-side deduplication. In the case of server-side deduplication, each file a user wants to store is transmitted to the cloud storage provider. For every file, the provider checks if he has to store the file or only needs to create a link to an already stored file. The user cannot detect if the cloud storage provider uses data deduplication. In the case of deduplication by the client, the client software transmits the hash value of the file to the cloud storage provider. Only if the provider is not already in possession of the file it will be transmitted. This variation of data deduplication has the effect of not only saving storage space, but also bandwidth. It is easy to detect if a cloud storage provider uses this kind of data deduplication by inspecting the log files or observing the amount of data that is transferred.

(3) Single user deduplication vs. cross user deduplication. Single user deduplication means that data deduplication is carried out separately for each user: If user A wants to store a file he has already stored in the past or in a different folder, the cloud storage provider only creates a link to that file. In the case of cross user deduplication, data deduplication is carried out across all users: If user A wants to store a file that another user B has already stored, the cloud storage provider only creates a link to that file instead of storing an additional copy. In general, data deduplication is carried out completely in the background which means that the user usually cannot choose whether data deduplication should be used or not.

ISSN 2348 – 7968

2. Literature Review

2.1 Automated Certification for Compliant Cloud-based Business Processes

A key problem in the deployment of large-scale, reliable cloud computing concerns the difficulty to certify the compliance of business processes operating in the cloud. Standard audit procedures such as SAS-70 and SAS- 117 are hard to conduct for cloud based processes. The paper proposes a novel approach to certify the compliance of business processes with regulatory requirements. The approach translates process models into their corresponding Petri net representations and checks them against requirements also expressed in this formalism. Being Based on Petri nets, the approach provides well-founded evidence on adherence and, in case of noncompliance, indicates the possible vulnerabilities. Keywords: Business process models, Cloud computing, Compliance certification, Audit, Petri nets
.

2.2 Automatic protocol blocker for privacy preserving public auditing in cloud computing

Cloud Computing has been envisioned as the next generation architecture of IT enterprise, due to its long list of unprecedented advantages in the IT history: on-demand self-service, ubiquitous network access, location independent resource pooling, rapid resource elasticity, usage-based pric-ing and transference of risk . As a disruptive technology with profound implications, Cloud Computing is transforming the very nature of how businesses use information technology. One fundamental aspect of this paradigm shifting is that data is being centralized or outsourced into the Cloud. From users' perspective, including both individuals and IT enterprises, storing data remotely into the cloud in a flexible on demand manner brings appealing benefits: relief of the burden for storage management, universal data access with independent geographical locations, and avoidance of capital expenditure on hardware, software, and personnel maintenances, etc. While these advantages of using clouds are unarguable, due to the opaqueness of the Cloud—as separate administrative entities, the internal operation details of Cloud Service Providers (CSP) may not be known by cloud users—data outsourcing is also relinquishing user's ultimate control over the fate of their data. As a result, the correctness of the data in the cloud is being put at risk due to the following reasons. First of all, although the infrastructures under the cloud are much more powerful and reliable than personal computing devices, they are still facing the broad range of both internal and external threats for data integrity. Secondly, for the benefits of their own, there do exist various motivations for cloud service providers to behave unfaithfully towards the cloud users regarding the status of their outsourced data. Examples include cloud service providers, for monetary reasons, reclaiming storage by discarding data that has not been or is rarely accessed or even hiding data loss incidents so as to maintain a reputation. In short, although outsourcing data into the cloud is economically attractive for the cost and complexity of long-term large-scale data storage, it does not offer any guarantee on data integrity and availability. This problem, if not properly addressed, may impedence the successful deployment of the cloud architecture. Recently, the notion of public auditability has been proposed in the context of ensuring remotely stored data integrity under different systems and security models. Public auditability allows an external party, in addition to the user himself, to verify the correctness of remotely stored data. However, most of these schemes do not support the privacy protection of users' data against external auditors, i.e., they may potentially reveal user data information to the auditors. , i.e., they may potentially reveal user data information to the auditors, From the perspective of protecting data privacy, the users, who own the data and rely on TPA just for the storage security of their data, do not want this auditing process introducing new vulnerabilities of unauthorized information leakage towards their data security.

2.3 Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing

Cloud computing is the long dreamed vision of computing as a utility, where users can remotely store their data into the cloud so as to enjoy the on-demand high quality applications and services from a shared pool of configurable computing resources. By data outsourcing, users can be relieved from the burden of local data storage and maintenance. Thus, enabling public auditability for cloud data storage security is of critical importance so that users can resort to an external audit party to check the integrity of outsourced data when needed. To securely introduce an effective third party auditor (TPA), the following two fundamental requirements have to be met: 1) TPA should be able to efficiently audit the cloud data storage without demanding the local copy of data, and introduce no additional on-line burden to the cloud user. Specifically, our contribution in this work can be summarized as the following three aspects:

ISSN 2348 – 7968

1) We motivate the public auditing system of data storage security in Cloud Computing and provide a privacy-preserving auditing protocol, i.e., our scheme supports an external auditor to audit user's outsourced data in the cloud without learning knowledge on the data content.
2) To the best of our knowledge, our scheme is the first to support scalable and efficient public auditing in the Cloud Computing. In particular, our scheme achieves batch auditing where multiple delegated auditing tasks from different users can be performed simultaneously by the TPA.
3) We prove the security and justify the performance of our proposed schemes through concrete experiments and comparisons with the state-of-the-art.
To enable privacy-preserving public auditing for cloud data storage under the aforementioned model, our protocol design should achieve the following security and performance guarantee:
1) Public auditability: to allow TPA to verify the correctness of the cloud data on demand without retrieving a copy of the whole data or introducing additional on-line burden to the cloud users.
2) Storage correctness: to ensure that there exists no cheating cloud server that can pass the audit from TPA without indeed storing users' data intact.
3) Privacy-preserving: to ensure that there exists no way for TPA to derive users' data content from the information collected during the auditing process.
4) Batch auditing: to enable TPA with secure and efficient auditing capability to cope with multiple auditing delegations from possibly large number of different users simultaneously.
5) Lightweight: to allow TPA to perform auditing with minimum communication and computation overhead.

3. Proposed System

There are three entities defined in our system, that is, users, private cloud and S-CSP in public cloud as shown in Fig. 1.**Hybrid cloud.** An environment comprised of two or more of the above specified cloud computing deployment models in a manner where they are bound together using technology that supports application, service or data portability, migration and interoperability.

- **Public cloud.** An environment provisioned for open use by the general public.
- **Private cloud.** An environment provisioned for exclusive use by a single organization comprising multiple (internal) consumers. Sometimes called an enterprise cloud, private clouds can be on-premise or off-premise, managed internally or managed by a third-party provider.

In this paper, we will only consider the file level deduplication for simplicity. In another word, we refer a data copy to be a whole file and file-level deduplication which eliminates the storage of any redundant files. Actually, block-level deduplication can be easily deduced from file-level deduplication, which is similar . Specifically, to upload a file, a user first performs the file-level duplicate check. If the file is a duplicate, then all its blocks must be duplicates as well; otherwise, the user further performs the block-level duplicate check and identifies the unique blocks to be uploaded. Each data copy (i.e., a file or a block) is associated with a token for the duplicate check.
• *S-CSP*. This is an entity that provides a data storage service in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users.

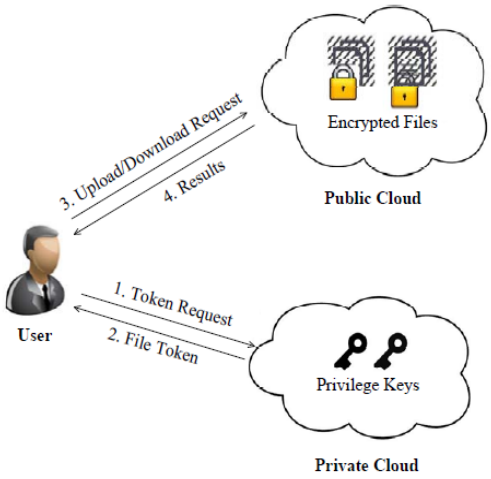

Fig 1- Block diagram of proposed system

To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data. In this paper, we assume that S-CSP is always online and has abundant storage capacity and computation power.
• *Data Users.* A user is an entity that wants to outsource data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the

ISSN 2348 – 7968

convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

• *Private Cloud.* Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating user's secure usage of cloud service. Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user
and the public cloud. The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.

This is a novel architecture for data deduplication in cloud computing, which consists of a twin clouds (i.e., the public cloud and the private cloud). Actually, this hybrid cloud setting has attracted more and more attention recently.

Under the assumption, two kinds of adversaries are considered, that is, 1)external adversaries which aim to extract secret information as much as possible from both public cloud and private cloud; 2) internal adversaries who aim to obtain more information on the file from the public cloud and duplicate-check token information from the private cloud outside of their scopes. Such adversaries
may include S-CSP, private cloud server and authorized users.

4. SECURITY ANALYSIS

Various security challenges related to these deployment models are discussed below:

• Cloning and Resource Pooling: Cloning deals with replicating or duplicating the data. According to Bernd Grobauer et al. [2], cloning leads to data leakage problems revealing the machine's authenticity. While Wayne A. Pauley [37] describes resource pooling as a service provided to the users by the provider to use various resources and share the same according to their application demand. Resource Pooling relates to the unauthorized access due to sharing through the same network. While the study on Virtual and Cloud Computing by various researches states that a Virtual Machine can easily be provisioned, they can also be inversed to previous cases, paused, easily restarted, readily cloned and migrated between two physical servers, leading to non-auditable security threats.

• Motility of Data and Data residuals: For the best use of resources, data often is moved to cloud infrastructure. As a result the enterprise would be devoid of the location where data is put on the cloud. This is true with public cloud. With this data movement, the residuals of data is left behind which may be accessed by unauthorized users. According to Rohit Bhadauria et al. [38], data-remnant causes very less security threats in private cloud but severe security issues may evolve in public cloud donations. This again may lead to data security threats like data leakage, data remnants and inconsistent data, as stated by Hassan Takabi et al. [4]. The authors have also mentioned that in order to solve the problems with data storage the optimal solution of cryptography can be thought of effectively.

• Elastic Perimeter: A cloud infrastructure, particularly comprising of private cloud, creates an elastic perimeter. Various departments and users throughout the organization allow sharing of different resources to increase facility of access but unfortunately lead to data breach problem. In private clouds, according to Krishna Subramanian [5], the resources are centralized and distributed as per demand. The resource treatment transfers resources based on the requirements of the users thus leading to problems of data loss, where any user may try to access secure data with ease. Moreover, Marios D. Dikaiakos et al. [39] states that elasticity of various cloud based resources would lead to store replicated data on untrusted hosts and this would then lead to enormous risks to data privacy.

• Shared Multi-tenant Environment: Kui Ren et al. [27] define multitenancy as one of the very vital attribute of cloud computing, which allows multiple users to run their distinct applications concurrently on the same physical infrastructure hiding user data from each other. But the shared multi-tenant character of public cloud adds security risks such as illegal access of data by other renter using the same hardware. A multi-tenant environment might also depict some resource contention issues when any tenant consumes some unequal amount of resources. This might be either due to genuine periodic requirements or any hack attack. Hsin-Yi Tsai et al. [6], has shown that multi-tenancy makes the impact of VM Hopping attack potentially larger than conventional IT environment.

• Unencrypted Data: Data encryption is a process that helps to address various external and malicious threats. Unencrypted data is vulnerable for susceptible data, as it does not provide any security mechanism. These unencrypted data can easily be accessed by unauthorized users. According to Cong Wang et al. [40], unencrypted data risks the user data leading cloud server to escape various data information to unauthorized users. For example, the famous file sharing service Dropbox was accused for using a single encryption key for all user data the company stored. These unencrypted, insecure data, as per Marjory S. Blumenthal [32], incite the malicious users to misuse the data one or the other way.

ISSN 2348 – 7968

• Authentication and Identity Management: With the help of cloud, a user is facilitated to access its private data and make it available to various services across the network. Identity management helps in authenticating the users through their credentials. But according to Rosa Sánchez et al. [35], a key issue, concerned with Identity Management (IDM), is the disadvantage of interoperability resulting from different identity tokens and identity negotiation protocols as well as the architectural pattern. While Jianyong Clien et al. [8] have mentioned that IDM leads to a problem of intrusion by unauthorized users. They even discussed that in order to serve authentication, apart from providing a password, a multi-factor authentication using smart card and fingerprint must be implemented for attaining higher level of security.

5. Results

Proof of ownership. Halevi et al. [11] proposed the notion of "proofs of ownership" (PoW) for deduplication systems, such that a client can efficiently prove to the cloud storage server that he/she owns a file without uploading the file itself. Several PoW constructions based on the Merkle-Hash Tree are proposed [11] to enable client-side deduplication, which include the bounded leakage setting. Pietro and Sorniotti [16] proposed another efficient PoW scheme by choosing the projection of a file onto some randomly selected bit-positions as the file proof. Note that all the above schemes do not consider data privacy. Recently, Ng et al. [15] extended PoW for encrypted files, but they do not address how to minimize the key management overhead.

Convergent Encryption. Convergent encryption [8] ensures data privacy in deduplication. Bellare et al. [4] formalized this primitive as message-locked encryption, and explored its application in space-efficient secure outsourced storage. Xu et al. [23] also addressed the problem and showed a secure convergent encryption for efficient encryption, without considering issues of the key-management and block-level deduplication. There are also several implementations of convergent implementations of different convergent encryption variants for secure deduplication (e.g., [2], [18], [21], [22]). It is known that some commercial cloud storage providers, such as Bitcasa, also deploy convergent encryption.

Secure Deduplication. With the advent of cloud computing, secure data deduplication has attracted much attention recently from research community. Yuan et al. [24] proposed a deduplication system in the cloud storage to reduce the storage size of the tags for integrity check. To enhance the security of deduplication and protect the data confidentiality, Bellare et al. [3] showed how to protect the data confidentiality by transforming the predicatable message into unpredicatable message. In their system, another third party called key server is introduced to generate the file tag for duplicate check. Stanek et al. [20] presented a novel encryption scheme that provides differential security for popular data and unpopular data.

6. Conclusion

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 testbed experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

References

[1] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui. "A secure cloud backup system with assured deletion and version control". In 3rd International Workshop on Security in Cloud Computing, 2011.

[2] C. Ng and P. Lee. Rev dedup: "A reverse deduplication storage system optimized for reads to latest backups". In Proc. of APSYS, Apr 2013".

[3] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. "Secure deduplication with efficient and reliable convergent key management". In IEEE Transactions on Parallel and Distributed Systems, 2013 ."

[4] J. Xu, E.-C. Chang, and J. Zhou. "Weak leakage resilient client-side deduplication of encrypted data in cloud storage". In ASIACCS, pages 195206, 2013.."

[5] J. Yuan and S. Yu. "Secure and constant cost public cloud storage auditing with deduplication." IACR Cryptology ePrint Archive, 2013:149, 2013".

[6] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan. "Sedic: privacyaware data intensive computing on hybrid clouds". In Proceedings of the 18th ACM conference on

ISSN 2348 – 7968

Computer and communications security, CCS11, pages 515526, New York, NY, USA, 2011. ACM.
[7] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: "Serveraided encryption for deduplicated storage". In USENIX Security Symposium, 2013.
[8] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message locked encryption and secure deduplication. In EUROCRYPT, 2013.
[9] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller." Secure data deduplication ". In Proc. of StorageSS, 2008.
[10] M. Bellare, C. Namprempre, and G. Neven." Security proofs for identity-based identifcation and signature schemes". J. Cryptology, 22(1):161, 2009.
29
[11] P. Anderson and L. Zhang. "Fast and secure laptop backups with encrypted deduplication". In Proc. of USENIX LISA, 2010
[12] R. D. Pietro and A. Sorniotti. "Boosting efficiency and security in proof of ownership for deduplication". In H. Y. Youm and Y. Won, editors, ACM Sym posium on Information, Computer and Communications Security, pages 8182. ACM, 2012.
[13] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman Peleg. "Proofs of ownership in remote storage systems". In Y. Chen, G. Danezis, and V. Shmatikov, edi- tors, ACM Conference on Computer and Communications Security, pages 491500. ACM, 2011
[14] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. "Twin clouds: An
architecture for secure cloud computing". In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.
[15] W. K. Ng, Y. Wen, and H. Zhu. "Private data deduplication protocols in cloud storage". In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441446. ACM, 2012.
[16] Z. Wilcox-O Hearn and B. Warner. "Tahoe: the least authority file system". In Proc. of ACM Storage's, 2008

Divya V kandalkar has completed B. E. Electronics and from Rashtrasant Tukadoji Maharaj Nagpur University and pursuing M. E. E&TC (VLSI design and Embedded System) from Sir Visveswaraya Institute of Technology, Nasik. Her field of interest are cloud computing.

Prof. R.R.Bhambare has completed B.E in Electronics from Pune University and ME (Electronics) from Dr.Babasaheb Ambedkar. Marathwada University and currently pursuing PhD from Rashtrasant Tukadoji Maharaj Nagpur University. His major fields of studies are: Embedded Systems, Advanced Communication Engineering, SDR .He is working as Associate Professor in dept of E&TC at Sir Visveswaraya Institute of Technology, Nasik.