SAJET

South Asian Journal of Engineering and Technology Vol.2, No.17 (2016) 58–63

ISSN No: 2454-9614

Integrated Cloud Mechanism for Multi Architecture De-duplication

M. Nivetha, G. Sathiyabama*, A. Sumithra, S. Vijayanand

Department of Computer Science and Engineering,SVH Engineering College, Gobi, Tamilnadu, India.

*Corresponding Author: Nivetha .M

E-mail: sumithradevi95@gmail.com

Received: 12/11/2015, Revised: 14/12/2015 and Accepted: 10/03/2016

Abstract

There is huge amount of data stored in cloud; de duplication is useful and efficient technique to make data management more scalable. Data de duplication is a specialized data compression technique for removing duplicate copies of repeating data in storage. To reduce the amount of storage space and save bandwidth in cloud storage various Data compression techniques are used, Data deduplication is one of the important data compression techniques. Data de duplication removes the duplicate copies of repeating data. To encrypt the sensitive data, the convergent encryption technique has been proposed. It is first formal attempt to address the problem of secure authorized data deduplication. Data and the differential privileges of users are considered while duplicate check. That is differing from traditional deduplication systems. We propose deduplication construction supporting system which authorizes duplicate check in public multi cloud architecture. Security analysis shows that proposed system is secure in terms of definitions specified in the proposed security model. Prototype of proposed authorized duplicate check is implemented and experiments conducted using this prototype.

Index Terms - Deduplication, hybrid cloud, authorized duplicate check

Reviewed by* **ICETSET'16 *organizing committee*

1. Introduction

A model for delivering information technology services in which resources are retrieved from the internet through web-based tools and applications, rather than a direct connection to a server. Data and software packages are stored in servers. However, cloud computing structure allows access to information as long as an electronic device has access to the web. This type of system allows employees to work remotely.

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 .In computing, data deduplication is a specialized data compression technique for eliminating duplicate copies of

repeating data. De-duplication is specialize data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. 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 l de-duplication, 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.

Related and somewhat synonymous terms are **intelligent (data) compression** and **single- instance (data) storage**. **This** technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk.

Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match frequency is dependent on the chunk size), the amount of data that must be stored or transferred can be greatly reduced. It gets a high deduplication ratio which is as better as global chunk-based deduplication and very low overhead than that of global chunk-based deduplication. Shang and Li **[13]** pointed out several shortcomings of existing works and discussed the corresponding possible solutions for data deduplication for cloud systems. The challenging issues for cloud data deduplication are still the balance of trade-off between storage efficiency and performance.

This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match frequency is dependent on the chunk size), the amount of data that must be stored or transferred can be greatly reduced.

It gets a high deduplication ratio which is as better as global chunk-based deduplication and very low overhead than that of global chunk-based deduplication. Shang and Li **[13]** pointed out several shortcomings of existing works and discussed the corresponding possible solutions for data deduplication for cloud systems. The challenging issues for cloud data deduplication are still the balance of trade-off between storage efficiency and performance. The access rights of the stored data. 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 **[17]** has been a well-known technique and has attracted more and more attention recently. Data although data deduplication brings a lot of benefits, security and privacy concerns arise as users' sensitive

2. Literature Survey

In archival storage systems, there is a huge amount of duplicate data or redundant data, which occupy significant extra equipments and power consumptions, largely lowering down resources utilization (such as the network bandwidth and storage) and imposing extra burden on management as the scale increases.

So data de-duplication, the goal of which is to minimize the duplicate data in the inter level, has been receiving broad attention both in academic and industry in recent years.

In this paper, semantic data deduplication (SDD) is proposed, which makes use of the semantic information in the I/O path (such as file type, file format, application hints and system metadata) of the archival files to direct the dividing a file into semantic chunks .Which are actually the storage units in the storage devices, so as to speed up the I/O performance as well as ease the data management.

3. System Design

A new model of hybrid cloud computing architecture based on cloud bus is proposed. The system is based on local private cloud, combined with one or more type(s) of public cloud(s). The internal structures of private cloud and public cloud are the same, including infrastructure and virtualization layer, cloud platforms layer, cloud bus layer, cloud application layer, the management center and storage centers. The layer of infrastructure and virtualization is designed to incorporate the underlying hardware resources into a virtual cluster, providing a variety of virtual resources to the upper layer. The layer of cloud platform is used to run Web applications or services, and carry application-specific development and application integration through its open interfaces. The cloud bus layer, consisting of a control bus, a number of node buses and adapters, is designed to manage and monitor the various services of the cloud platform layer.

The proposed model of the architecture can accelerate the migration of the existing IT environments to cloud computing environments, reduce the investment, and make full use of IT resources. The conceptions of cloud computing are introduced, and then a hybrid cloud computing platform for smart grid is designed. After that, the distinguished characteristics of the proposed platform are explained in detail, following with the introduction of some potential power system applications. Finally, some notable state-of-the-art products that can be used to build the proposed platform are introduced. Cloud-based design and manufacturing (CBDM) refers to a service-oriented networked product development model in which service consumers are able to configure products or services and reconfigure manufacturing systems through Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), Hardware-as-a-Service (HaaS), and Software-as-a-Service (SaaS**).[1]** Adapted from the original cloud computing paradigm and introduced into the realm of computer-aided product development, Cloud-Based Design and Manufacturing is gaining significant momentum and attention from both academia and industry.

4. Proposed System Description

Traditional encryption, while providing data confidentiality, is incompatible with data deduplication.

Specifically, traditional encryption requires different users to encrypt. It is aiming at efficiently solving the problem of deduplication with differential privileges in cloud computing, we consider a **hybrid cloud architecture** consisting of **a public cloud and a private cloud**. Unlike existing data deduplication systems, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges. A new deduplication system supporting differential duplicate check is proposed under this hybrid cloud architecture where the S-CSP (Storage Cloud Service Provider) resides in the public cloud. The user is only allowed to perform the duplicate check for files marked with the corresponding privileges. Furthermore, we enhance our system in security. Specifically, we present an advanced scheme to support stronger security by encrypting the file with differential privilege keys. In this way, the users without corresponding privileges cannot perform the duplicate check. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP.

4.1 Hybrid Cloud Architecture for Data Deduplication

Their data with their own keys. Thus, identical data copies of different users will lead to different ciphertexts, making de-duplication impossible. Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible.

It encrypts/decrypts 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 generate the same convergent key the same ciphertext. To prevent unauthorized access, a secure Proof of ownership (POW) Propose another advanced deduplication system supporting authorized duplicate check. In this new deduplication system, hybrid cloud architecture is introduced to solve the problem. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server instead. In this way, the users cannot share these private keys of privileges in this proposed construction, which means that it can prevent the privilege key sharing among users in the above straightforward construction. To get a file token, the user needs to send a request to the private cloud server. To perform the duplicate check for some file, the user needs to get the file token from the private cloud server. The private cloud server will also check the user's identity before issuing the corresponding file token to the user. The authorized duplicate check for this file can be performed by the user with the public cloud before uploading this file. Based on the results of duplicate check, the user either uploads this file or runs POW (Proof of Ownership).

System Architecture

S-CSP:

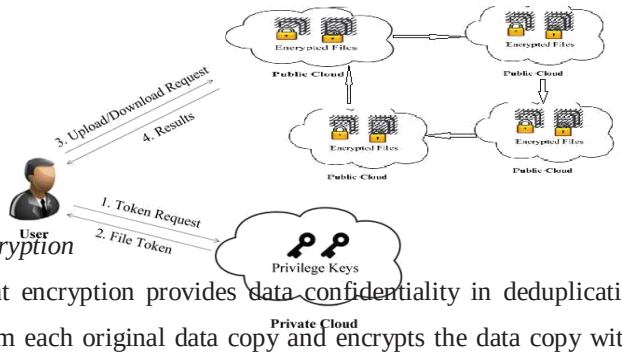
This is an entity that provides a data storage service in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data.

4.2 Authorized duplicate check:

Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate.

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



4.4 Convergent encryption

Convergent encryption provides data confidentiality in deduplication. A user (or data owner) derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also 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. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored. Note that both the convergent key and the tag are independently derived and the tag cannot be used to deduce the convergent key and compromise data confidentiality. Both the encrypted data copy and its corresponding tag will be stored on the server side.

4.5 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.

5. Conclusion

In this project, 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 test bed experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

6. Feature Enhancement

Though the above solution supports the differential privilege duplicate, it is inherently subject to brute-force attacks launched by the public cloud server which can recover files falling into a known set. More specifically, knowing that the target file space underlying a given cipher text C is drawn from a message space S = {F1 , · · · , Fn } of size n, the public cloud server can recover F after at most n off-line encryptions. That is, for each i = 1, · · · , n, it simply encrypts Fi to get a cipher text denoted by Ci . If C = Ci , it means that the underlying file is encrypted . Security is thus only possible when such a message is unpredictable. This traditional convergent encryption will be insecure for predictable file.

We design and implement a new system which could protect the security for predictable message. The main idea of our technique is that the novel encryption key generation algorithm. For simplicity, we will use the hash functions to define the tag generation function sand convergent keys in this section. In traditional convergent encryption, to support duplicate check, the key is derived from the file F by using some cryptographic hash function F = H(F). To avoid the deterministic key generation, the encryption key F for file F in our system will be generated with the aid of the private key cloud server with privilege key KP.

References

[1] OpenSSL Project, (1998). Available: http://www.openssl.org/
[2] P. Anderson and L.Zhang ,“Fast and secure laptop backups with encrypted de-duplication,” in Proc. 24th Int. Conf. Large Installation Syst. Admin., 2010, pp. 29–40.
[3] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless Server-aided encryption for de-duplicated storage,” In Proc. 22nd USENIX Conf. Sec. Symp., 2013, pp. 179–194.
[4] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Message-locked encryption and secure deduplication,” in Proc. 32nd Annu. Int.Conf. Theory Appl. Cryptographic Techn., 2013, pp. 296– 312.
[5] M. Bellare, C. Namprempre, and G. Neven, “Security proofs for identity-based identification and signature ,”. Cryptol.,vol. 22, no. 1, pp. 1–61, 2009.
[6] M. Bellare and A. Palacio, “Gqand schnor identification schemes: Proofs of security against impersonation under active and concurrent attacks,” inProc. 22nd Annu. Int.Cryptol. Conf. Adv. Cryptol.,2002, pp. 162– 177.
[7] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider,“Twinclouds: An architecture for secure cloud computing,” in Proc. Workshop Cryptography Security Clouds, 2011.