ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. 2, Special Issue 11, March 2016

A Secure Data Deduplication Scheme for Distributed Cloud Storage

N.K. Karthika[1], J.Mannar mannan[2], Dr. M. Sundarambal[3]
PG Scholar[1], Assistant Professor[2] , Department of IT, Anna University Regional Centre,
Tamil Nadu, India
Email id: karthipr2@gmail.com

Abstract:-Present decades of computing technology move towards cloud computing. Cloud technology provides various services. Storage as a service (Saas) in the cloud provides storage place for users where one can store and retrieve data over clouds. The duplicates of the redundant files on clouds occupy large storage area. To address these issues de-duplication is needed over cloud environment. In this paper, de-duplication is based on a multiple file properties and important key words of a file. This proposed method can overcome limitations existing in the present tag based de-duplication.

Cloud computing is the provisioning of different types of computing services over the Internet. Cloud computing provides a shared pool of resources, including storage space of data, networks, processing power of computer, and specialized corporate and large scale of user applications. De-duplication of data/file, also called as the single instance storage is a method of increasing storage space by eliminating redundant data. Redundant data is replaced with a pointer to the unique data copy. The convergent encryption, a cryptographic technique has been used to encrypt the data to protect the confidentiality of the data while performing de-duplication. The implementation of RSA algorithm in the proposed system is used for secure data encryption. De-duplication of files in the distributed cloud storage is done based on four criteria's such as name, content, size and tag. We show that our proposed method is secure and the storage in the distributed cloud is increased in terms of user's authorization and de-duplication respectively when compared to normal operations.

Index terms: Distributed cloud, redundant data, De-duplication, Data security.

1 INTRODUCTION

Cloud computing is one of the type of internet based computing that provides shared processing resources and data to computers and other devices based on demand of clients. Cloud computing and storage solutions provide users and various enterprises with various kinds of capabilities to store and process their data in data centers. Cloud computing provides unlimited resources (virtualized) as services to the users on demand across the whole internet. The services are classified and are identified as Infrastructure as a service (Iaas), Platform as a service (Paas) and Software as a service (Saas). One critical challenge of the distributed cloud storage is the management of the increasing volume of data.

Distributed cloud is the application of cloud computing technologies for the interconnection of data and applications served from multiple geographic locations. Distributed cloud computing speeds communications for global services and enables more responsive communications for specific regions. Distributed file system for cloud is a file system that allows many clients to have access to the same file providing important operations such as create, delete, modify, read and write. Each file may be partitioned or divided into several parts called chunks. Each data parts or chunk is stored in remote machines. The storage of data in files in a hierarchical tree where the directories are represented by the. There are several ways to share files in a distributed cloud architecture. Users can share resources from any device, anywhere and everywhere through internet which is characterized by means of scalable and elastic resources such as physical servers that are virtualized and allocated dynamically.

Distributed access control architecture is used for multi tenant and virtualized environments. The design of this architecture is based on the principles from secure data management and software engineering. From a security management perspective, the goal of the cloud is to meet user's access control requirements. The most important security concepts in the cloud computing are identified as Confidentiality, Integrity and Availability(CIA). Confidentiality keeps private data from being disclosed and maintains privacy. Integrity refers to the fact that data is not corrupted. To ensure data availability, data recovery is allowed. Data must be coded and once when the coded data is lost, it can be recovered from fragments and these fragments are constructed during the coding phase.

Resource sharing across distributed clouds depends on the collaborative environment. Three types of collaborations such as federated, loosely coupled and ad hoc can fulfill the authorization requirements. Federated collaboration is identified by a high degree of trust and mutual dependence among collaborating clouds and supports a long term inter operation. Local policies govern interactions among multiple clouds in a loosely coupled collaborative environment. This technique is more flexible and autonomous in terms of various access policies and resource management. In ad hoc collaboration, a user is only aware of a few remote sharable services. To ensure secure inter operation via discovered resources and services in a dynamic inter operation environment where clouds can join and exit in an ad hoc manner, suitable authentication and authorization of user's mechanism need to be developed.

To make data management scalable in distributed cloud computing, de-duplication technique has been carried out. Data de-duplication is a compression technique for eliminating the repeated copies of data in the cloud storage. This technique is used to improve the storage space in the distributed cloud storage. Despite of keeping multiple data copies with the same content, de-duplication eliminates identical copies by keeping one original data copy. Data de-duplication can generally operate at the file (or) block level. For file level de-duplication, the duplicate copies of the same file are removed and for de-duplication in block level, identical blocks of files that are not identical are eliminated. De-duplication with the distributed cloud storage not only reduces the storage space requirements, but also reduces the data that is transferred over the networks which results in faster and efficient data protection operations.

Encryption and decryption of files is based on RSA algorithm. The encryption key is public and differs from the decryption key which is kept secret and this algorithm can perform bulk encryption-decryption operations at much higher speed. RSA involves a public key and a private key. The public key can be known by every users in cloud and is used for encrypting messages. The messages that are encrypted with the public key can be decrypted in a specified amount of time using the private key. Traditional encryption provides data confidentiality and is not compatible with data de-duplication. The identical data copies of different users will lead to different cipher text since this technique requires different users to encrypt their data with their own keys.

Convergence encryption called as the content hash keying, produces identical cipher text from identical plain text files. A tag is generated in order to find the duplicates. Encryption keys are generated from the chunk data, thus identical chunks will always encrypt to the same cipher text. The information, the user needs to access and decrypt the chunks is encrypted using a key known only to the user. An RSA algorithm is proposed which allows a message sender to generate a public key for encryption and the generated private key has been sent to the receiver using a secured database.

Convergence encryption called as the content hash keying, produces identical cipher text from identical plain text files. A tag is generated in order to find the duplicates. Encryption keys are generated from the chunk data, thus identical chunks will always encrypt to the same cipher text. The information, the user needs to access and decrypt the chunks is encrypted using a key known only to the user. An RSA algorithm is proposed which allows a message sender to generate a public key for encryption and the generated private key has been sent to the receiver using a secured database.

De-duplication in the distributed cloud storage is based on four criteria’s such as the Name, size, content and tag. UltraCompare checks the file sizes when searching for duplicate files and this will set a minimum and maximum file size for the detection of duplicates. Any file having a size outside of the range will be completely ignored by the Find Duplicates search. Tags are another type of file property, designed to be customized by the user. A tag cloud is a visual representation of text data, used to predict tags on websites. A tag is generated by the user to detect duplicates and is stored in the private cloud with the file. Secure data de-duplication is achieved by means of authorization. For the detection of duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored.

2 RELATED WORK

In 2002, Douceur et al. [1] studied the problem of de-duplication in multi-tenant environment. The authors proposed the use of the convergent encryption, i.e., retrieval of keys from the hash of plaintext. Then, Storer et al. [2] pointed out some security problems in cloud, and presented a novel security model for secure data de-duplication. Moreover, these two protocols is only focused on server-side de-duplication and they did not considered the data leakage settings, against unauthorized users. In order to prevent private data leakage, Halevi et al. [4] proposed the concept of Proof of Ownership (PoW), in terms of security and performances while introducing three different constructions.

In 2012 M. Bellare et.al explains a new cryptographic primitive, Message-Locked Encryption (MLE), where the key under which encryption and decryption are performed is itself retrieved from the message. MLE offers a way to achieve secure de-duplication, a goal currently targeted by numerous cloud-storage providers [5]. An MLE scheme is a symmetric encryption scheme in which the key used for both the encryption and decryption is itself retrieved from the message. Instances of this primitive are looking for widespread deployment and application for the purpose of secure de-duplication [6], but in the absence of a theoretical treatment, they have no prior indication of what these proposed methods do or do not accomplish. In 2013 Jin Li,et.al presented several new de-duplication applications that supports authorized duplicate check in a hybrid cloud architecture. Security analysis shows that the scheme is secure in terms of the definitions specified in the proposed security model. They show that their proposed authorization of duplicate check scheme achieves minimal overhead compared to normal operations [7].

In 2014 Jin Li, et.al [8] proposed Dekey, an efficient and reliable convergent key management scheme for secure authorized de-duplication. Dekey applies de-duplication among convergent keys and the convergent key that is distributed is shared across multiple key servers, while maintaining semantic security of convergent keys and confidentiality of the data that is outsourced. They implement Dekey using the Ramp secret sharing scheme and demonstrate that it achieves small encoding/decoding overhead when compared to the network transmission overhead in the regular upload/download operations. For the purpose of saving resources consumption in both network bandwidth and storage capacities, many kinds of cloud services, namely Dropbox, wuala and Memopal, apply client side de-duplication [3]. This concept reduces the storage of identical data in cloud servers and reduces the consumption of network bandwidth associated to transmitting the same contents of data several times.

PKC scheme uses one key for encryption and a different key for decryption. Modern PKC was first described using a two-key crypto system in which two parties could engage in a secure communication over a non-secure communications channel without having to share a secret key [9]. RSA is one of the first and still most common PKC implementation that is in use today for key exchange or digital signatures. Bellare et al. [10] showed how to protect the data confidentiality by transforming the predictable message into unpredictable message. In their system, another third party called key server is introduced to generate the file tag for duplicate check.

Security is traded for storage efficiency as for every file that transits from unpopular to popular status, storage space can be reclaimed. Once a file reaches the popular status, space is reclaimed for the copies uploaded so far, and normal de-duplication can take place for future copies. Standard security mechanisms (such as Proof of Ownership [4, 11]) can be applied to secure this step. Such mechanisms are not required in the case of unpopular files, given that they are protected by both encryption layers and cannot be de-duplicated. Convergent encryption (CE) based solutions [12], provides security only for unpredictable messages even in the best case, and are vulnerable to brute-force attacks. The simple approach of sharing a secret key across clients with a deterministic encryption scheme [13] fails to achieve compromise resilience.

Douceur et al. [12, 14], introduced convergent encryption, attempting to combine data confidentiality with the possibility of data de-duplication. Convergent encryption of a message consists of encrypting the plaintext using a deterministic (symmetric) encryption scheme with a key which is deterministically derived solely from the plaintext. Clearly, when two users independently attempt to encrypt the same file, they will generate the same ciphertext which can be easily de-duplicated. Unfortunately, convergent encryption does not provide semantic security as it is vulnerable to content-guessing attacks. Juels and Guajardo [15] present a protocol in which a possibly malicious device generates an RSA key in cooperation with a certificate authority. Their protocol prevents a device from generating an ill-formed keypair.

3 SYSTEM DESIGN

The cloud environment is deployed using the aneka tool which is built on xen server. De-duplication in the distributed cloud is based on different file properties and tag generation mechanism. The data to be stored in the distributed cloud is encrypted based on one of the cryptographic technique, RSA algorithm. Cryptography is a science of using mathematics to encrypt and decrypt data and it provides storage space to store the sensitive information. Using RSA algorithm, two keys are generated. Public key is used for encryption and the secret key is used for decryption.

The user stores the encrypted file in the disk memory, found in the distributed cloud. The de-duplication is performed in the storage element, disc. This de-duplication technique is based on four criteria's such as the Name, Size, Content and Tag. This technique improves the storage space in the distributed cloud, thereby increasing the bandwidth. De-duplication process can take place at either the file level or the block level. For file level de-duplication, it removes duplicate copies of the same file. De-duplication can also take place at the block level, which eliminates duplicate blocks of data that occurs in a non-identical files.

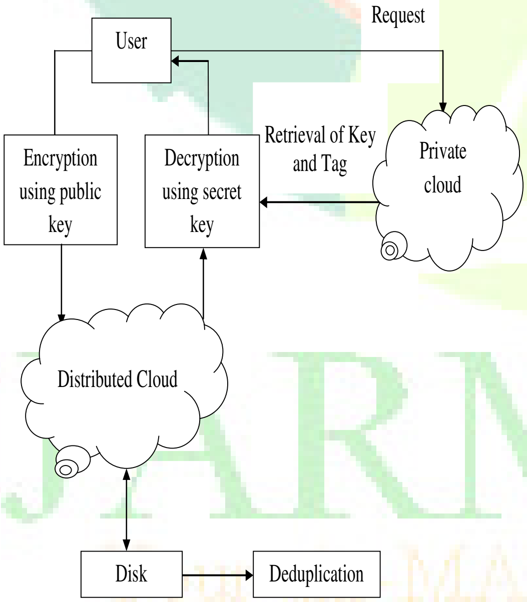


Figure 1. Architecture for secure data de-duplication

A tag is a relevant keyword or term corresponding to a piece of information such as a picture, a geographic map, or a video clip. Tags are like keywords that one can assign to their files when uploading in a distributed cloud. More than one tag can be assigned to any file. Clicking on the tag presents a list of files to which that corresponding tag is assigned. Tagging is an flexible way to group and to find the files. Tags can be sorted; the number of tags visible can be changed accordingly and recent tags can only be displayed on screen. The private key generated by the RSA algorithm along with the tag is stored in the private cloud in order to prevent the data leakage and to prevent the secret key from sharing across the different users.

Users have access to the private cloud server, a semi trusted third party which will performe de-duplicable encryption by providing file tags and the secret key for the requesting users. If a file is a duplicate, then all its data blocks must be duplicates as well; otherwise, the user identifies the unique blocks to be uploaded by performing the block level duplicate check. Each file is associated with a tag for the duplicate check. Thus the storage space in the distributed cloud is increased and thus the bandwidth is increased. The user then by using the secret key decrypts the file and the original file can be obtained. Thus, the distributed cloud provides storage space for users to store the file and are retrieved for later use. The security of files can be increased by the use of the distributed cloud when compared to the storage in other clouds.

4 IMPLEMENTATION

4.1 RSA Algorithm

The original file to be encrypted is based on RSA algorithm. The user encrypts the original file by using the public key generated. The key generation process is carried out and the encrypted and the decrypted files are generated by RSA algorithm. The public key is known to all users to encrypt the file and the private key is kept secret and is stored in the private cloud server. The user also generates the tag while uploading the file and is stored in the server along with the private key.

4.2 Encrypt / Decrypt Process

Figure 2 shows the original file is encrypted using RSA algorithm and is stored in the distributed cloud. Disc is a storage element in the cloud where the files are to be stored and the de-duplication process takes place in the disc memory to increase the storage space. Encryption is the process of converting plain text to cipher text which is not understandable to human beings in order to improve the security of a file.

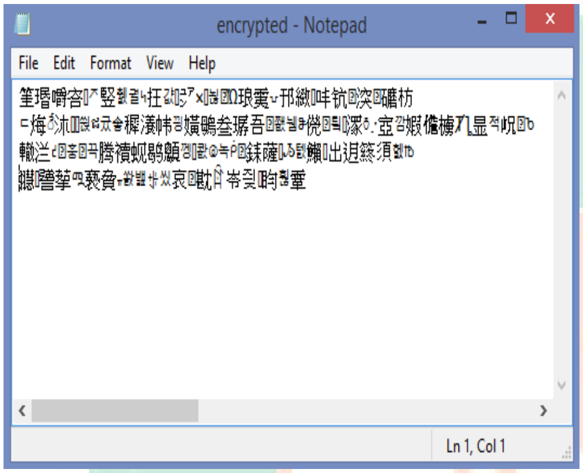
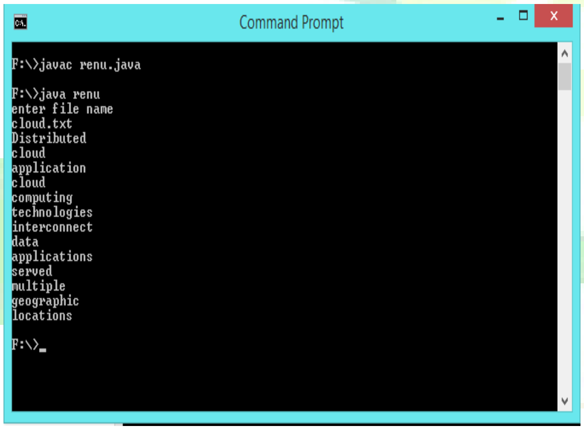


Figure 2. Encrypted File

The encrypted file in the distributed cloud is decrypted by using the secret key stored in the private cloud server. The key is obtained by submitting the request to the server.

4.3 Retrieval of File Attributes and Tags

The file attributes is retrieved to check whether another file is stored in the same storage element with the same file attributes in order to detect the duplicates in the distributed cloud.

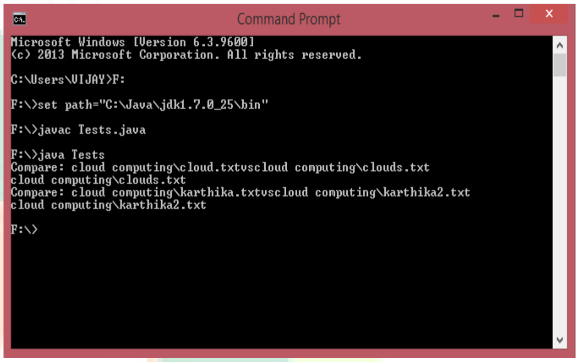


```
F:\>javac renu.java

F:\>java renu
enter file name
cloud.txt
Distributed
cloud
application
cloud
computing
technologies
interconnect
data
applications
served
multiple
geographic
locations

F:\>_
```

Figure 3. File Tags

Figure 3 shows the generation of tags from the original file. A tag is generated by the user while encrypting the file before uploading to the cloud and it is used for the detection of duplicates. There may be more than one tags assigned to the same file in order to make de-duplication possible and the storage space can be increased to a greater extent.

4.4 File De-duplication



```
Microsoft Windows [Version 6.3.9600]
(c) 2013 Microsoft Corporation. All rights reserved.

C:\Users\VIJAY>F:

F:\>set path="C:\Java\jdk1.7.0_25\bin"

F:\>javac Tests.java

F:\>java Tests
Compare: cloud computing\cloud.txtvscloud computing\clouds.txt
cloud computing\clouds.txt
Compare: cloud computing\karthika.txtvscloud computing\karthika2.txt
cloud computing\karthika2.txt

F:\>
```

Figure 4. File De-duplication

Figure 4 shows that the duplicate files in the cloud are identified and the duplicate files are eliminated to increase the bandwidth. Duplicate files are identified based on name, file size, content and the file tags.

5. CONCLUSION

In this paper, the notion of the data de-duplication was proposed to protect the data security. Several new de-duplication constructions are also presented that supports the duplicate check in the distributed cloud architecture, in which the duplicate check tags of files are generated by the user. Security analysis demonstrates that the scheme is secured by the use of private cloud which stores the secret key and the tag of a file. Data encryption and decryption is made accurate by the use of RSA algorithm. Several de-duplication scheme in the proposed work shows that this model incurs minimal overhead compared to convergent encryption.

References

[1] J. R. Douceur, A. Adya, W. J. Bolosky, D.Simon, and M. Theimer. Reclaiming space from duplicate files in a serverless distributed file system. In Proceedings of 22nd International Conference on Distributed Computing Systems (ICDCS, 2002).

[2] M. W. Storer, K. Greenan, D. D. Long, and E. L. Miller. Secure data deduplication. In Proceedings of the 4th ACM International Workshop on Storage

Security and Survivability, StorageSS ’08, pages 1–10, New York, NY, USA, 2008. ACM.

[3] D. Harnik, B. Pinkas, and A. Shulman-Peleg. Side channels in cloud services: Deduplication in cloud storage. IEEE Security And Privacy, 8(6):40–47, 2010.

[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Proceedings of the 18thACM conference on Computer and communications security, CCS’11, pages 491–500, New York, NY, USA, 2011. ACM.

[5] P. Anderson and L. Zhang, ‘‘Fast and Secure Laptop Backupswith Encrypted De-Duplication,’’ in Proc. USENIX LISA, 2010,pp. 1-8.

[6] M. Bellare, S. Keelveedhi, and T. Ristenpart, ‘‘Message-Locked Encryption and Secure Deduplication,’’ in Proc. IACR Cryptology ePrint Archive, 2012, pp. 296-3122012:631.

[7] .Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou” A Hybrid Cloud Approach for Secure Authorized Deduplication” IEEE Transactions On Parallel And Distributed System VOL:PP NO:99 YEAR 2013.

[8] Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou “Secure Deduplication with Efficient and Reliable Convergent Key Management” IEEE Transactions On Parallel And Distributed Systems, VOL. 25, NO. 6, JUNE 2014.

[9] Hellman, M. and J. Diffie, 1976. New Directions in Cryptography. IEEE transactions on Information theory, vol. IT-22, pp:644-654.

[10] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Server aided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

[11] Di Pietro, R., Sorniotti, A.: Boosting efficiency and security in proof of ownership for deduplication. In: ASIACCS '12, New York, NY, USA, ACM (2012) 81-82

[12] Storer, M., Greenan, K., Long, D., and Miller, E. Secure data deduplication. In Proceedings of the 4th ACM inter- national workshop on Storage security and survivability (2008), ACM, pp. 1–10.

[13] Rogaway, P., and Shrimpton, T. A provable-security treatment of the key-wrap problem. In EUROCRYPT 2006 (St. Petersburg, Russia,May 28 – June 1, 2006), S. Vaudenay, Ed., vol. 4004 of LNCS, Springer, Berlin, Germany, pp. 373–390.

[14] Douceur, J.R., Adya, A., Bolosky, W.J., Simon, D., Theimer, M.: Reclaiming space from duplicate files in a serverless distributed file system. In: ICDCS '02, Washington, DC, USA, IEEE Computer Society (2002) 617-632

[15] A.Juels and J. Guajardo. RSA key generation with verifiable randomness. In PKC, Feb. 2002