ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 11, March 2016

Deduplication And Data Security With Duplicate Authorization Check

N.SUGUNA

Department of Computer Science and Technology

Prof Mr. V.R.MANI

Department of Computer Science and Technology

ABSTRACT—Cloud computing useful for sharing information among the users with unlimited storage space. Data deduplication is method for eliminating duplicate copies of data, to reduce storage space and upload bandwidth in cloud.To achieve secure deduplication in cloud, convergent encryption has been used. Convergent encryption generates identical ciphertext from same plaintext files. The duplicate files in cloud computing has to be removed from storage without the provider have access to the encryption keys.The problem is there if no efficient and reliable key management.To achieve this Dekey , a new construction in which users no need to manage any keys on their own but they securely distribute the convergent key shares across multiple servers. Encryption mechanisms not preserve the data in the cloud from unauthorized access from both inside attacker as well as outside attacker.To minimize the attacks and also to provide security in cloud computing, user behavior profiling and decoy information technology called as Fog Computing has to be implemented.

Keywords:- Deduplication, Convergent encryption key management, Dekey, User behaviour profiling, Decoy Technology.

1.INTRODUCTION

Data deduplication is technique to ignore duplicate copies of data, and it has been used to reduce storage space and bandwidth of upload .In secure deduplication convergent encryption has been adopted, a critical issue of making convergent encryption is to manage a huge number of convergent keys..By user‟s point of view, data outsourcing increases security and privacy concerns. We should trust third-party cloud providers to properly invoke confidentiality, integrity checking, and access control mechanisms contradictory to any insider and outsider attacks. However, deduplication, while enlightning storage and bandwidth efficiency, is adaptable with Convergent key management. Specifically, traditional encryption is responsible for distinct users to encrypt their data with their own keys. Many proposals have been made to secure remote data in the Cloud by applying encryption and standard access controls. It is fair to say all of the standard approaches have been demonstrated has to fail for a variety of reasons, including insider attacks, mis-configured services, faulty implementations, and the creative construction of effective and experienced attacks not envisioned by the implementers of security procedures. Constructing a trustworthy cloud computing environment is not enough, because accidents continue to happen, if it happened the information gets lost, there is no way to get it back. . The concept is that we can decrease the damage of stolen data if we decrease the value of that stolen information to the attacker. To achieve a „preventive‟ disinformation attack has to used. Secure deduplication services can be implemented by using these features such as:

1.1 User Behaviour Profiling: User profiling is a well known Technique to monitor data access in the cloud and detect abnormal data access patterns.This can be applied here to how much a user accesses their information in the Cloud. Such „normal user‟ behavior can be used to check whether abnormal access to a user‟s information is happening. The method of user behavior profiling is commonly used in fraud detection applications. That profiles would naturally include volumetric information, how many documents are typically read and how often.

1.2 Decoys: Decoy used to monitor data access in the cloud and detect abnormal data access patterns. We start a disinformation attack by responding huge amounts of decoy information to the attacker. This preserve against the misuse of the user‟s original data.Start disinformation attacks against cruel

insiders, preventing them from making the original sensitive users data from fake worthless decoy technology has been used. Its mainly used to serve two purposes:
(1) Authorize data access , when irregular information access is detected, and
(2) Confusing the attacker with bogus information

The decoys, then, serve two purposes:
Validates whether data access is authorized
.Confusing the attacker with bogus information.

Outsider attacker protection has to be done by increase the insider attacker with secure deduplication. Convergent encryption provides a data confidentiality in deduplication. It encrypts/decrypts a content with a convergent key, which has to be derived by computing the hash value of the data copy itself [8]. After generating the key data has to be encrypt and then users has to retain the keys and send the cipher text to the cloud. Encryption is deterministic but isame data copies will generate the same convergent key and the same cipher text. It allows the cloud to perform the deduplication on the cipher texts. The cipher text scan is decrypted by the respective data owners with their convergent keys.The new construction is Dekey in which users no need to maintain any keys on their own but instead securely share **out** the convergent key across multiple servers for insider attacker. Dekey uses the Ramp secret sharing scheme and explain that Dekey has limited overhead in realistic environments.

2.LITERATURE REVIEW

Data deduplication is important for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space to users. They defined the notions used in based paper, review some secure primitives used secure deduplication. Symmetric Encryption, Convergent Encryption, Proofs of Ownership (PoWs), Ramp Secret Sharing, Secure Deduplication.

2.1Cryptographic Cloud Storage

Symmetric encryption is encryption technique that is used to provide security to the users information. In traditional encryption different users they encrypt their data with their own keys. This type of encryption where the same key is need to encrypt and decrypt the message. This differs from asymmetric (or public-key) encryption, which uses one key to encrypt a message and another one to decrypt the message. Symmetric encryption is the oldest and best-known technique to secure the data. The secret key is a number, a word, or just a string of random letters, is applied to the text of a message to modify the content in a particular way. This might be as simple as shifting each letter by a number of places in the alphabet. Since both sender and recipient know the secret key, they can encrypt and decrypt all messages that use this key.The same data content of different users produce different ciphertexts so it is impossible for deduplication .

2.2Secure and Efficient Access to Outsourced Data

Convergent encryption permits cloud storage services to deduplicate data, lacking the service having access to the encryption keys that is used to protect customer files.This presents better privacy than traditional cloud storage. Typically, when cloud services encrypt data, they use their own encryption key.The encryption key is derived from the file itself with convergent encryption. Essentially, it produces identical ciphertext from identical plaintext files. Convergent encryption allows cloud storage providers to store large amounts of data at low prices, while offering better privacy than traditional cloud storage.The drawback is Privacy concerns have been raised with cloud storage services deduplicating data via convergent encryption, because deduplication can be used to “discover” which users are storing a file, if the attacker also has a copy of the file.

2.3Proofs of Ownership in Remote Storage Systems

The idea of proof of ownership (PoW) is to solve the problem of using a small hash value as a proxy for the entire file in client-side deduplication , the could use the storage service as a content distribution network. This proof mechanism in PoW presents a solution to protect the security in client-side deduplication. Like this way, a client can prove to the server that it really has the file. Dekey supports client-side deduplication with PoW to allow users to prove their ownership of data copies to the storage server. Particularly, PoW is implemented as an common algorithm (denoted by PoW) run by the prover (i.e., user) and a verifier (i.e., storage server). The verifier has to derive a short value from a data copy M.In order to prove the ownership of the data copy M, the prover wants to send short value and run a proof algorithm with the verifier. It is passed if and only if short value and the proof is correct. The notations of PoW_F and PoW_B to denote PoW for a file F and block B, respectively. Especially, the notation of $PoW_{F;j}$ will be used to denote a PoW protocol with respect to $T_j(F)$_ $TagGen_{CE}(F; j)$.

2.4 Secure De-duplication with Efficient and Reliable Convergent Key Management

Dekey is modeled to efficiently and reliably preserve convergent keys. The idea is to permit deduplication in convergent keys and distribute the convergent keys around multiple KM-CSPs.In place of encrypting the convergent keys on a per-user basis, Dekey structures secret shares on the original convergent keys (that are in plain) and allocates the shares across multiple KM-CSPs. If many users share the same block, they can access the same corresponding convergent key. This considerably reduces the storage overhead for convergent keys. Apart from this, it provides fault tolerance and permits the conver-gent keys to remain accessible even if any subset of KM-CSPs fails.

2.5 Authorized Duplicate check

In 2013 Jin Li,et.al explain for deduplication to protect the confidentiality of sensitive data while handling deduplication, the convergent encryption technique has been presented to encrypt the data before outsourcing. To protect data security, this paper presents the first attempt to functionally address the problem of authorized data deduplication. Apart from normal deduplication systems, the differential privileges of users are then considered in duplicate check besides the data itself. They also provide several new deduplication constructions supporting authorized duplicate check in a hybrid cloud environment. Security analysis explains that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, they implement a prototype of our proposed authorized duplicate check scheme and plan test bed experiments using our prototype. They show that their proposed authorized duplicate check scheme obtains minimal overhead compared to normal operations. In 2014 I.Sudha et.al proposed a completely distinct approach to secure the cloud with the decoy information technology and is called as “Fog Computing” . They use this technology to activate disinformation attacks against malicious insiders, which helps to prevent and distinguish the real perceptive customer data from fake worthless data. The Decoy Information Technology is used for validating whether data access is authorized even when abnormal information access is detected. It helps in confusing the attacker with bogus information. In 2014 Jin Li, et.al explain propose Dekey, an efficient and reliable convergent key management scheme for secure deduplication. Dekey supplies deduplication among convergent keys and distributes convergent key shares across multiple key servers, while reserving semantic security of convergent keys and confidentiality of outsourced data. They present Dekey using the Ramp secret sharing scheme and explains that it incurs small encoding/decoding overhead related to the network transmission overhead in the regular upload/download operations.

3.EXISTING SYSTEM

3.1 Introduction

Dekey is modeled to efficiently and reliably preserve convergent keys. The idea is to permit deduplication in convergent keys and distribute the convergent keys around multiple KM-CSPs.In place of encrypting the convergent keys on a per-user basis, Dekey structures secret shares on the original convergent keys (that are in plain) and allocates the shares across multiple KM-CSPs. If many users share the same block, they can access the same corresponding convergent key. This considerably reduces the storage overhead for convergent keys. Apart from this, it provides fault tolerance and permits the conver-gent keys to remain accessible even if any subset of KM-CSPs fails.

3.2 System Model

The system setup phase in Dekey is related to that in the baseline approach, but it contains an additional step for realizing the key storage in KM-CSPs. In Dekey, we take that the number of KM-CSPs is n.

S1:Input security parameter 1^{λ}, the user prepares a convergent encryption scheme, and two PoW protocols POW_F and POW_B for the file and block ownership proof.

S2: The S-CSP initializes both the rapid storage system and the file storage system and place them to be ⊥.

S3: Each KM-CSP uses a rapid storage system for block tags and a lightweight storage system for holding convergent key shares, and sets them to be ⊥.

File Upload

In order to upload file F , the user and the S-CSP carry out both file-level and block-level deduplications. The file-level dedu-plication operation is same as that in the baseline approach. More exactly, the user sends the file tag T(F) to the S-CSP for the file duplicate check. If there is a file duplicate is found, the use has to run the PoW protocol POW_F with the S-CSP in order to prove the file ownership. It avoids the block-level duplicate check and moves to the key distribution stage. If there is no duplicate exists, then block-level deduplication will be performed

File Download

The user wants to download the file F means,the user first downloads the encrypted blocks $\{C_i\}$ from the S-CSP .They has to decrypt those encrypted blocks by recovering the convergent keys. Espcially, the user sends all the pointers for F to k rather than n KM-CSPs and fetches the corresponding shares K_{ij} for each block B_i. After collecting all the shares, the user continues to recreate the convergent key K_i = Recover($\{K_{ij}\}$) for B_i. Finally, the encrypted blocks $\{C_i\}$ can be decrypted with $\{K_i\}$ to get the original file F .

4.PROBLEM DESCRIPTION

Dekey is structured to efficiently and reliably maintain convergent keys. Its concept is to permits deduplication in convergent keys and distribute the convergent keys around multiple KM-CSPs. Christo Ananth et al. [7] proposed a secure hash message authentication code. A secure hash message authentication code to avoid certificate revocation list checking is proposed for vehicular ad hoc networks (VANETs). The group signature scheme is widely used in VANETs for secure communication, the existing systems based on group signature scheme provides verification delay in certificate revocation list checking. In order to overcome this delay this paper uses a Hash message authentication code (HMAC). It is used to avoid time consuming CRL checking and it also ensures the integrity of messages. The Hash message authentication code and digital signature algorithm are used to make it more secure . In this scheme the group private keys are distributed by the roadside units (RSUs) and it also manages the vehicles in a localized manner. Finally, cooperative message authentication is used among entities, in which each vehicle only needs to verify a small number of messages, thus greatly alleviating the authentication burden. Correct breach detection and threat removal method must be there to stop the data from going into dark hands. Traditional methods have unsuccessful in providing up to the mark security to the clients, primarily those wreaked by an insider to the service provider (cloud).

A different technology of security called the fog computing, this security mechanism uses a very simple method for threat detection and data misuse, but there have been some flaws which they were not able to count up-to. It comprises a unique model for data secure in the cloud using offensive decoy technology for flooding the intruder with false data.

5.PROPOSED SYSTEM

5.1 Introduction

In Cloud Computing the Fog Computing is an new technique.The fog computing also provides data, compute, storage, and also application services to end-users as like colud. The main change is that fog gives proximity to its end users through dense geographical distribution and also it supports mobility. Access points or set-up boxes are used as end devices to host services at the network. That end devices are also called as edge network. The Fog computing expands the Quality of service and also decrease the latency. According to Cisco, due to its wide geographical distribution the Fog computing is well suited for real time analytics and big data. .Fog computing provides-Low latency and location awareness, it has Wide-spread geographical distribution, supports Mobility. The main task of fog is to deliver data and place it closer to the user who is positioned at a location which at the edge of the network.

Several methods are introduced to secure cloud data by encryption and standard access control but it is found those methods are not full proof due to variety of reasons. Customers are not only requires reliable cloud environment but also a healthy security for data and applications. The recovering of stolen or lost data is not possible. So we must have knowledge to deal with such incidences. When we lower the value of stolen data by providing decoy documents then it is possible to limit the harm of the system.

5.2 Methodology

System Parameters- The experiments are conducted using 1 GHz processor with RAM 256MB or higher and hard disk 10 GB or higher.

Experiment Factors-In order to evaluate the performance of CUSUM algorithm on the basis of time, average fluctuation and true positive

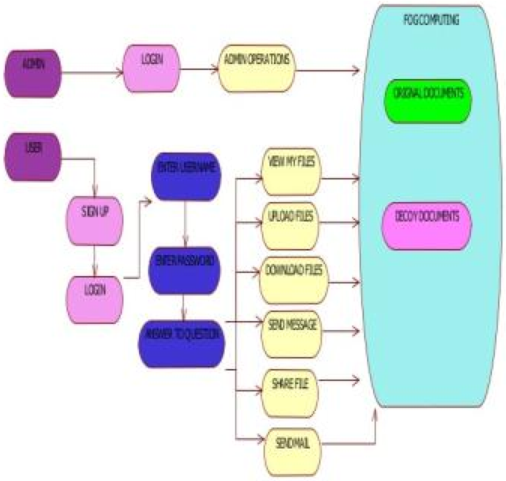

Fig 5. 1 System Architecture

5.3 Modules

5.3.1 File Upload

To upload file F , the user and the S-CSP perform both file-level and block-level deduplications. The file-level deduplication operation is identical to that in the baseline approach. More precisely, the user sends the file tag T(F) to the S-CSP for the file duplicate check. If a file duplicate is found, the user will run the user behavior profiling with the S-CSP to prove the file ownership. It skips the block-level duplicate check and jumps to the key distribution stage. If no duplicate exists, then block-level deduplication will be performed. After both file-level and block-level duplicate checks if no file duplicate is found then he can upload the file.

5.3.2.File Download

To download file F the user will run the usr behavior profiling to prove the file ownership. After verficaition the user can download the file.if he is corrsct owner he can download the correct file otherwise duplicate file has to be downloaded.

5.3.3 User Behaviour Profiling

Legitimate Users of the Cloud system are acquainted with the documents and information on the Cloud system they have stored. The search for documents is to the point and limited. A masquerade gets access to the victim's system illegitimately, is unlikely to be acquainted with the structure and contents of the file system. Their search is not to the point and widespread. The user search behaviour is profiled and developed based on this key assumption, Cusum algorithm is used which calculates the average fluctuation and thus the user behaviour is noted when changed.

Cusum algorithm

For applying cusum on **N** no of observations

Let, initial average av**1 -> N =0;**

```
Sump=Sum till previous observations =0;
For loop n=1 -> N
sump=sump (previous)+Current(n)
av(n)=sump/N
end for loop
```

Now av is the cumulative summation averages and difference in two consecutive averages gives the fluctuation**.**

5.3.4 Decoy Technology

The file system is combined together with traps, these traps are uploaded to the system with Cloud service provider. Traps can contain confidential documents like credit card details, tax returns, bank statements. These documents are places in highly egregious places. A masquerader who is not acquainted with the system and who has an ill intent may is likely to click on these false documents. Thereby the system can be notified of masquerade activity.

The hash code of all the legitimate and decoy documents upload on the system is calculated. The hash code of every document downloaded is matched with the hash code of the decoy document. If a match is found then the document is deemed to be a decoy document and an alert is generated. An insider attacker would not be able to escape detection if they access a decoy document. The hash code is based on keyed-Hash Message Authentication Code (HMAC).

HMAC code: HMAC that is keyed hashed message authentication code which is used for calculating a message authentication code. It involves a cryptographic hash function along with a secret key. We are calculating the HMAC code of the document by using the MD5 Algorithm.MD5 processes a document of variable length into a fixed length output of 128 bits.

1. Variable length to fixed length output.
2. Input n-bit blocks
3. Input divided into 512 bit blocks

4. Padding is done
5. Buffer initialization
6. Output 128 bit

In our decoy technique, we have enhanced the security of the by inserting a pseudo-generator which further jumbles the code and changes the position of each element in the code so that the attacker can never the get the original information if ever he tries.

The advantages of placing decoys in a file system are threefold:

- The detection of masquerade activity
- The confusion of the attacker and the additional costs incurred to identify the real information from bogus information.

The combination of the two techniques: The combination of user behaviour profiling with decoy technology provides a strong evidence of illegal access and helps improve accuracy of detection. Only user one technique can produce false positive results.

6.ACKNOWLEDGEMENT

I owe my deep gratitude to my respected guide Prof. Mr. V.R.MANI who gives me the valuable guidelines with a touch of inspiration and motivation to progress my way through quite substantial barrier between early problem statement and something that resembled a fine work.

7.CONCLUSION

The basic idea in this paper is to limit the damage of stolen data if we decrease the value of that stolen information to the attacker. To succed this a preventive disinformation attack has to be used. We present that secure deduplication services can be present with additional security features insider attacker on Deduplication and outsider attacker by using the detection of masquerade activity. The confusion of the attacker and the additional costs incurred to distinguish real from bogus information, and the deterrence effect which, although hard to measure, plays a significant role in preventing imitate activity by risk-averse attackers. We posit that the combination of these security features will provide unparalleled levels of security for the deduplication. Through this research we have concluded that decoy technology and fog computing together can provide security to real world problems like insider data theft attacks. In future we can extend the working of algorithm, by calculating the accuracy with other attributes such as performance evaluation of the security mechanism. The concept of fog computing is very vast other than security of data we can extend this research for network security through fog computing and also localising the user data a secure geographical locations.

REFERENCES

[1] Secure De-duplication with Efficient and Reliable Convergent Key Management" Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou in *IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 6, June 2014*

[2] P. Anderson and L. Zhang, ''Fast and Secure Laptop Backups with Encrypted De-Duplication,'' in Proc. *USENIX LISA, 2010, pp. 1-8.*

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart, ''Message-Locked Encryption and Secure De-duplication,'' in Proc. *IACR Cryptology ePrint Archive, 2012, pp. 296-3122012:631.*

[4] 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 Proc. *ICDCS, 2002, pp. 617-624.*

[5] J. Gantz and D. Reinsel, The Digital Universe in 2020: Big Data,Bigger Digital Shadows, Biggest Growth in the Far East, Dec.2012[Online]Available: *http://www.emc.com/collateral/analystreports/idc-the-digital-universe-in-2020.pdf*

[6] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg,''Proofs of Ownership in Remote Storage Systems,'' in Proc. *ACM Conf. Comput. Commun. Security, Y. Chen, G. Danezis, and V. Shmatikov, Eds., 2011, pp. 491-500.*

[7] Christo Ananth, M.Danya Priyadharshini, "A Secure Hash Message Authentication Code to avoid Certificate Revocation list Checking in Vehicular Adhoc networks", International Journal of Applied Engineering Research (IJAER), Volume 10, Special Issue 2, 2015,(1250-1254)

[8] M. Mulazzani, S. Schrittwieser, M. Leithner, M. Huber, and E. Weippl, ''Dark Clouds on the Horizon: Using Cloud Storage as Attack Vector and Online Slack Space,'' in Proc. *USENIX Security, 2011, p. 5.*

[9] F. Rocha and M. Correia(2011) , "Lucy in the sky without diamonds: Stealing confidential data in the cloud," in Proceedings of the 2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops, ser. DSNW '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 129–134.

[10] B. Grobauer, T. Walloschek, E. Stocker (2011) " Understanding Cloud Computing Vulnerabilities, Security & Privacy", IEEE (Volume:9 , Issue: 2),

March-April, 2011, Page(s) 50-57

[11] Fog Computing: Mitigating Insider Data Theft Attacks in the Cloud Position Paper Salvatore J. Stolfo Computer Science Department Columbia University New York , NY, USA Email: sal@cs.columbia.edu

[12] Jiang Zhu, D.S. Chan, M.S. Prabhu, P. Natarajan, Hao Hu, F. Bonomi (2013) "Improving Web Sites Performance Using Edge Servers in Fog Computing Architecture,"IEEE 7th International Symposium on System Engineering (SOSE), pp. 320-323, March 2013.

[13] M. Ben-Salem and S. J. Stolfo (2011) "Modeling user search-behavior for masquerade detection," Proceedings of the 14th International Symposium on Recent Advances in Intrusion Detection. Heidelberg: Springer, September 2011, pp. 1–20.

[14] Jay Heiser, Mark Nicolett, Assessing the Security Risks of Cloud Computing, 03 June, 2008

[15] Ki-Woong Park, Sung Kyu Park, Jaesun Han, Kyu Ho Park, (2010) "Towards Mutually Verifiable Billing Transactions in the Cloud Computing Environment," IEEE 3rd International Conference on Cloud Computing (CLOUD), Page(s) 139-147, July 2010

[16] J. Montelibano, A. Moore, Insider Threat Security Reference Architecture (2012) 45th Hawaii International Conference on System Science (HICSS), Page(s) 2412 - 2421, 4-7 January 2012

[17] Nahla Shatnawi, Q.A., Wail Mardini (2011). "Detection of Insiders Misuse in database Systems" proceedings of the international Multi Conference of Engineers and computer Science 2011, Hong Kong Vol. I, IMECS 2011, March 16 - 18, 2011.

[18] Marten van Dijk, Ari Juels (2010) "On the Impossibility of Cryptography Alone for Privacy-Preserving Cloud Computing" RSA Security Brief, March 2010