Deduplication in Encrypted Data: A Comprehensive Review

Neelu Verma[1], Dinesh Singh[2]

[1]M.Tech. Scholar, [2]Assistant Professor,
Department of Computer Science Engineering, DCRUST, Murthal

Abstract

Cloud computing provides several facilities to the users amongst which the most prevalent is data storage. The users are permitted to download or upload their data anywhere and anytime to the server. But retrieving data leads to several issues related to the confidentiality and privacy. The users generally store their data on the cloud in encrypted form in order to ensure security and privacy. This encrypted data presents various new challenges. Because of such issues compression technique known as data deduplication is discussed in this paper. Deduplication technology, which removes duplicate copies, is a smart solution which saves space and traffic over disk. In this paper, proposed strategies and operations for deduplication are also discussed which describe the procedure for de duplication mechanism. It also discusses about the key operations performed for data deduplication.

Keywords: *deduplication, encrypted data, hashing, cloud*

1. Introduction

Nowadays, cloud computing has become the future scope of numerous IT corporations [24].These days cloud computing is empowering its user by providing limitless and efficient space for storage and also the accessibility and availability of data everywhere and at every time [25].As the digital data is increasing at a fast frequency, the services of the cloud are gaining popularity. Most of the enterprises have decided to outsource all their records to the CSP in order to have better administration of their resources [22].As the rate at which data is increasing in the past years, there is a need for a new data compression technology known as data deduplication. This is an effective compression technique which can identify the redundancy in data and removes it and can also improve the storage space utilization and reduces the bandwidth of transmitted data. The term data deduplication basically means the technique which stores only a single copy of any data in place of storing multiple replicas of the same data and further providing simply the links of those copies [3]. Data compression technique also reduce the number of disks used in any process to minimize the energy costs [27]. So, some of the important advantages of deduplication are:

a) Reducing Input Output traffic, which results in higher Input Output throughput.

b) Saving disk space, which automatically leads to saving cost on purchasing storage devices

Besides having those advantages, deduplication still has some issues which hinder its acceptance by all the users. The major problem is its privacy of data. Even though the data is uploaded in the encrypted format to the cloud, but still there is no guarantee about the trustworthiness of the CSP [22]. Also, the data that has been encrypted by the traditional encryption methods which use a complete file as their encryption unit cannot take profit of deduplication techniques because of several issues [29]. To deal with such duplication difficulty data deduplication is a mechanism which is used to free up the storage space. In deduplication mechanism, hash code is used for deleting the duplicate data from the system. Hash code is a static length. When user uploads data for storage, its hash code is created using Hashing Algorithm such as SHA, MD5, Tiger, Whirlpool and so on and then hash code verification is done by the server from the hash index which has the record of all the hash codes. If that generated hash code matches with the other hash code present it means that data is duplicated and there is no need to store again that already stored data, then in this case data will be deleted but a pointer will be connected to the original data. If hash code does not match with the other hash codes, then that data will be uploaded to the cloud and new hash code entry will be done in hash index.

The deduplication becomes inconsistent with the conventional encryption method. All the various clients may encrypt the same file into various cipher texts, making deduplication a challenge. So, convergent encryption use hash code as a key for encryption. Due to this, the similar file would give the similar cipher text. But this is also not efficient enough as any attacker may get the complete file from the server by just knowing the hash of the file. Therefore, to solve this issue, the concept of Proof of Ownership came over which the owner prove itself to the server that he is the actual owner of the file [24]. The data deduplication strategies have been categorized as follows:

Block-level de-duplication can segment the files in blocks and also stores one copy of every block only. The system may either use blocks of fixed size or of variable size.

File-level de-duplication is another popular service type in which only one copy of every file is being stored. [28][16]

Deduplication can generally occur on two sides as explained below.

Server Side Data deduplication: It occurs on the site of the server. This deduplication takes place under CAP and CSP. Whenever any user upload a file to the cloud, the server verifies that whether the same data is already present or not. If the file is not present in the cloud, then that file will get uploaded else the server rejects that file. So, here the server does the deduplication after the reception of file.

Client Side Data deduplication: It uses the concept of hashing or tag generation using hashing algorithms. It compares the calculated hash code and sends hash code before uploading the file into the cloud. After receiving hash code it checks whether hash code is present. If that file is stored then user will discard the file and its file will be connected by pointer. Otherwise file will be uploaded. [11]

2. Motivation

Data stored over cloud and flows through network in the plain text format is a serious security threat. So client encrypts the data and sends it to cloud storage provider. It is found that deduplication technique can save up to 90% storage, dependent on applications. [17] Convergent encryption has security flaws with regard to ownership revocation and tag consistency. [3] Big data workloads have redundancy. On average, 44% of the active data set in our big data workloads is redundant. Two Schemes of deduplication are: Inline Scheme and Offline Scheme. In inline scheme deduplication gets performed before the data is saved so that both disk traffic and space are saved. In offline scheme deduplication gets performed after the data is saved on hard drives. It is used to save disk space.

3. Background

3.1 Architecture of deduplication

Source based approach of deduplication: Here, before uploading any data to cloud the user firstly sends an identifier or the tag of his/her data (such as the hash value) to the storage server before uploading it to the cloud. The server performs the data redundancy check operation. If the data has not been previously stored to the cloud, then user uploads his new data and server stores the complete data. But, if the same data is already stored then the user needs to upload its metadata only, and then server makes a pointer pointing to the already stored data. So, this approach improves both bandwidth and the storage space.

Target based approach of deduplication: Unlike the previous approach, here user is not aware of any de-duplication that may occur to his uploaded data. The user just uploads the data to the cloud without any redundancy check and then the server performs the de-duplication process after receiving the data. So, this approach only improves the storage space but is unable to handle transmission volume. [15][16]

3.2 System Description

Data deduplication system model contains three kinds of objects:-

Owner of data and Data Holders: The owner is one who owns data and upload that to the cloud service provider. Whenever the owner uploads any data to the CSP, their passwords and identifications should be verified. Then, they encrypt data and upload it to the CSP alongwith its information of the corresponding index, known as the tag and this tag generation is done by hashing. The person who uploads the data for the first time is known as data owner and the other users are known as data holders. Data owner have high priority than data holders.

CSP: The CSP or the cloud service provider is that unit whose function is to provide the storage services in cloud. It comprises of further two units which are cloud storage and the cloud server. It basically acts as a database. The cloud server maintains data structure to achieve access control to identify the owner.

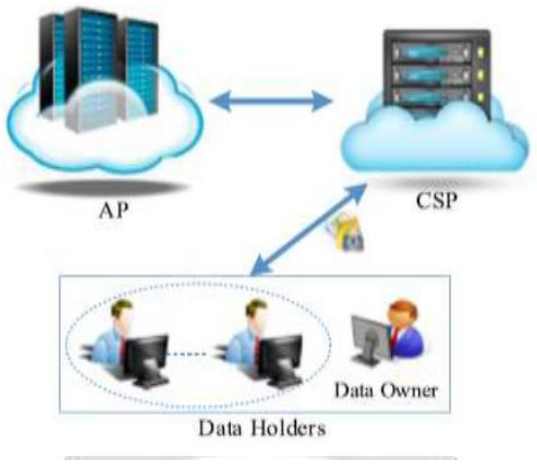

Fig.1 Deduplication system model [1]

AP: The main function of AP or the authority provider is to manage the ownership lists comprising of the tag and identities of all the owners for the stored data. It also controls the access to data and manages dynamic ownership. It helps in verification as well. [2]

3.3 Security Requirements

The requirements needed for the security of data deduplication in the system can be itemized as follows:-

Data privacy: Unauthorized data owners or the data holders who are unable to prove their realness must not be allowed to access data residing in cloud.

Tag consistency: Consistency of the tag must be checked against the attacks by the algorithm like the deduplication algorithm should provide security against identical attack in which a valid file is changed by a fake file with the same tag, and an attacker who only has the tag of the data gets the corresponding file. This leads to unlawful access of data which create trouble.

Forward and backward secrecy: In file-level deduplication, forward secrecy describes that after any request for deletion or updation of the data, the users must not be allowed to access the data stored in cloud whereas backward secrecy describes that whenever the data owner uploads data already existing in the storage, the user must not be allowed to access data before the ownership check.

Away from the security requirements, efficiency is most important for data deduplication. As the volume of stored data is humungous, parameters like storage effectiveness, transmission effectiveness and computation effectiveness must be achieved. [2]

4. Literature Survey

Various researchers have developed several schemes which work on deduplication mechanism. Some of the schemes involve deduplication on encrypted data. The traditional schemes had some security issues. The research work of different researchers has been summarized below.

In paper [1] a scheme is proposed to deduplicate encrypted format data stored in cloud and it deals with the ownership challenge and proxy re-encryption. When the data holders are offline it also supports deduplication.

A secure data deduplication scheme with efficient Proof-of-Ownership process for dynamic ownership management has been proposed in [2]. Cross-user file-level and Inside-user block-level data deduplication are also gets supported over this scheme.

In [4] they have given a DeDup app which is easy to use application interface based on the challenges of ownership and re-encryption. The efficiency of scheme is a result of the computer simulation and the analysis performed. DeDup app which is a client-side application for deduplication supports different tasks like registration uploading, downloading and deletion of file in a graphical user interface.

In [6] a system is proposed in which data gets uploaded onto cloud in the encrypted format. At the same time, duplication is checked before storing data onto cloud so that the space on cloud should not get wasted and rent that the user has to pay to CSPs would also get reduced.

5. Key Operations and Procedures

System comprises the following key operations:-

Upload: Data is uploaded at server on encrypted format. If there is no duplicacy found, the data user encrypts the data to guarantee the confidentiality and safety, and upload the data at CSP with the hash code.

Data Deduplication: It takes place when data holder wants to upload the same data that was already stored in the CSP. For identifying the duplicated data, CSP will do the hash comparison using hashing algorithm. If check is positive, CSP contacts AP for deduplication and AP challenges data ownership and data holder eligibility also checked by AP, and then it issues a re-encryption key which convert the encrypted data to a form that will be decrypted by the authorized data holder.

Data Deletion: If data holder wants to delete data from server, Firstly Cloud Service Provider manages the records of duplicated data holders and then it remove the duplication record of the requested user. If the records are left, then in this case CSP will not remove the encrypted data stored, it only deletes the block of requested user. If there is nothing, the encrypted data should be removed at CSP.

Updation of Encrypted Data: If a data owner updates a key *DEK* with *DEK'* and then provides this new encrypted data to the CSP for replacing the previous data stored in order to accomplish better security, so the CSP provides the latest re-encrypted *DEK'* to all the data holders.

Ownership Management: Whenever any data owner uploads his data after the data holder, re-encryption takes place at cloud service provider for appropriate holders.

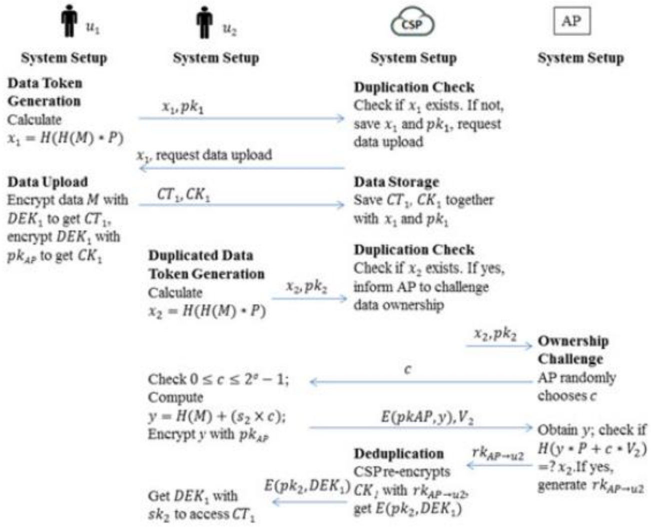
u_1
u_2
CSP
AP
System Setup
System Setup
System Setup
System Setup
Data Token Generation
Calculate
$x_1 = H(H(M) \cdot P)$
x_1, pk_1
Duplication Check
Check if x_1 exists. If not, save x_1 and pk_1, request data upload
x_1, request data upload
Data Upload
Encrypt data M with DEK_1 to get CT_1, encrypt DEK_1 with pk_{AP} to get CK_1
CT_1, CK_1
Data Storage
Save CT_1, CK_1 together with x_1 and pk_1
Duplicated Data Token Generation
Calculate
$x_2 = H(H(M) \cdot P)$
x_2, pk_2
Duplication Check
Check if x_2 exists. If yes, inform AP to challenge data ownership
x_2, pk_2
Ownership Challenge
AP randomly chooses c
c
Check $0 \le c \le 2^{\sigma} - 1$;
Compute
$y = H(M) + (s_2 \times c)$;
Encrypt y with pk_{AP}
$E(pkAP, y), V_2$
Obtain y; check if $H(y \cdot P + c \cdot V_2)$ =? x_2. If yes, generate $rk_{AP \to u2}$
$rk_{AP \to u2}$
Deduplication
CSP re-encrypts CK_1 with $rk_{AP \to u2}$, get $E(pk_2, DEK_1)$
$E(pk_2, DEK_1)$
Get DEK_1 with sk_2 to access CT_1

Fig.2 Deduplication process [1]

Fig. 2 defines the process of data deduplication at Cloud Service Provider and also at Authority Provider. It is supposed that user upload its data at Cloud Service Provider with protection using hash key, whereas another user is a data holder who wants to upload the duplicate data at Cloud Service Provider.

The complete process of data deduplication is defined as:
Step 1 – Setting up the System
Step 2 – Data token generation
Step 3 – Duplication check
Step 4 – Checking and uploading of duplicated data
Step 5 – Ownership challenge
Step 6 – Deduplication

Conclusion

The prevailing solutions available for deduplication cannot openly provide revocation as well as access control at one time. Most of those deduplication solutions cannot ensure security, reliability and data privacy. Further, it is difficult for any data holder to undertake deduplication due to several reasons in real time. At first, the data holders may not always be available to manage the data. Also, deduplication becomes a little complex in terms of the various calculations involved in the process. Next, it can also disrupt the data privacy in the process of the analysis of the De-duplicated data. Lastly, a shared key is being provided to all the users making them capable for accessing the same data. Whenever a user is removing data from his/her account, the keys are being reassigned to all the remaining users who still have access to the data. This reassignment which is a complicated task, if not carried out properly, the removed user may have illegal access to that data. [7] Finally the solution is data deduplication technique, which is a perfect hit. Duplicated data occupy gigantic storage space and handling power of system. When performing data deduplication over data increase the performance of system in relations of deduplication ratio, speed of read/write operations will be improved due to a reduced amount of overhead, storage space utilization will be advanced, data transmission speed will be improved and power consumption will be truncated and it also provide efficient storage cost of data.

References

[1] Yan, Z., Ding, W., Yu, X., Zhu, H., & Deng, R. H., “Deduplication on encrypted big data in cloud”. *IEEE Transactions on Big Data*, 2016,*2*(2), 138-150.
[2] S. Jiang, T. Jiang and L. Wang, "Secure and Efficient Cloud Data Deduplication with Ownership Management," in IEEE Transactions on Services Computing, 2017, vol. PP, no. 99, pp. 1-1.
[3] J. Hur, D. Koo, Y. Shin and K. Kang, "Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage (Extended Abstract)," 2017 IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, 2017, pp. 69-70.
[4] H. Kamboj and B. Sinha, "DEDUP: Deduplication system for encrypted data in cloud," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 795-800.
[5] H. Kamboj and B. Sinha, "Secure data deduplication mechanism based on Rabin CDC and MD5 in cloud computing environment," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2017, pp. 400-404.
[6] A. R. Deshmukh, R. V. Mante and P. N. Chatur, "Cloud Based Deduplication and Self Data Destruction," 2017 International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT), Warangal, 2017, pp. 155-158.
[7] R. Vidhya, P. G. Rajan and T. A. Lawrance, "Elimination of Redundant Data in Cloud with Secured Access Control," 2017 International Conference on Technical Advancements in Computers and Communications (ICTACC), Melmaurvathur, 2017, pp. 141-143.
[8] Y. Shin, D. Koo, J. Yun and J. Hur, "Decentralized Server-aided Encryption for Secure Deduplication in Cloud Storage," in IEEE Transactions on Services Computing, vol. PP, no. 99, pp. 1-1, 2017
[9] X. Yang, R. Lu, K. K. R. Choo, F. Yin and X. Tang, "Achieving Efficient and Privacy-Preserving Cross-Domain Big Data Deduplication in Cloud," in IEEE Transactions on Big Data, vol. PP, no. 99, pp. 1-1, 2017
[10] J. Hur, D. Koo, Y. Shin and K. Kang, "Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 11, pp. 3113-3125, Nov. 1 2016.
[11] R. B. Sirsat and N. R. Talhar, "Deduplication in cloud storage on the basis of proof of ownership," 2016 International

Conference on Computing Communication Control and automation (ICCUBEA), Pune, 2016, pp. 1-5.

[12] H. Cui, R. H. Deng, Y. Li and G. Wu, "Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud," in IEEE Transactions on Big Data, vol. PP, no. 99, pp. 1-1.

[13] Z. Yan, M. Wang, Y. Li and A. V. Vasilakos, "Encrypted Data Management with Deduplication in Cloud Computing," in IEEE Cloud Computing, vol. 3, no. 2, pp. 28-35, Mar.-Apr. 2016.

[14] C. M. Yu, C. y. Chen and H. c. Chao, "Proof of ownership in deduplicated cloud storage with mobile device efficiency," in IEEE Network, vol. 29, no. 2, pp. 51-55, March-April 2015.

[15] C. I. Fan, S. Y. Huang and W. C. Hsu, "Encrypted Data Deduplication in Cloud Storage," 2015 10th Asia Joint Conference on Information Security, Kaohsiung, 2015, pp. 18-25.

[16] R. Chen, Y. Mu, G. Yang and F. Guo, "BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication," in IEEE Transactions on Information Forensics and Security, vol. 10, no. 12, pp. 2643-2652, Dec. 2015.

[17] P. Prajapati and P. Shah, "Efficient cross user data deduplication in remote data storage," International Conference for Convergence for Technology-2014, Pune, 2014, pp. 1-5.

[18] Z. Wen, J. Luo, H. Chen, J. Meng, X. Li and J. Li, "A Verifiable Data Deduplication Scheme in Cloud Computing," 2014 International Conference on Intelligent Networking and Collaborative Systems, Salerno, 2014, pp. 85-90.

[19] J. Li, X. Chen, M. Li, J. Li, P. P. C. Lee and W. Lou, "Secure Deduplication with Efficient and Reliable Convergent Key Management," in IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 6, pp. 1615-1625, June 2014.

[20] N. Kaaniche and M. Laurent, "A Secure Client Side Deduplication Scheme in Cloud Storage Environments," 2014 6th International Conference on New Technologies, Mobility and Security (NTMS), Dubai, 2014, pp. 1-7.

[21] P. Puzio, R. Molva, M. Önen and S. Loureiro, "ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage," 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, Bristol, 2013, pp. 363-370.

[22] F. Rashid, A. Miri and I. Woungang, "Secure Enterprise Data Deduplication in the Cloud," 2013 IEEE Sixth International Conference on Cloud Computing, Santa Clara, CA, 2013, pp. 367-374.

[23] C. I. Ku, G. H. Luo, C. P. Chang and S. M. Yuan, "File Deduplication with Cloud Storage File System," 2013 IEEE 16th International Conference on Computational Science and Engineering, Sydney, NSW, 2013, pp. 280-287.

[24] X. Jin, L. Wei, M. Yu, N. Yu and J. Sun, "Anonymous deduplication of encrypted data with proof of ownership in cloud storage," 2013 IEEE/CIC International Conference on Communications in China (ICCC), Xi'an, 2013, pp. 224-229.

[25] F. Rashid, A. Miri and I. Woungang, "A secure data deduplication framework for cloud environments," 2012 Tenth Annual International Conference on Privacy, Security and Trust, Paris, 2012, pp. 81-87.

[26] S. Maddodi, G. V. Attigeri and A. K. Karunakar, "Data Deduplication Techniques and Analysis," 2010 3rd International Conference on Emerging Trends in Engineering and Technology, Goa, 2010, pp. 664-668

[27] Qinlu He, Zhanhuai Li and Xiao Zhang, "Data deduplication techniques," 2010 International Conference on Future Information Technology and Management Engineering, Changzhou, 2010, pp. 430-433.

[28] D. Harnik, B. Pinkas and A. Shulman-Peleg, "Side Channels in Cloud Services: Deduplication in Cloud Storage," in IEEE Security & Privacy, vol. 8, no. 6, pp. 40-47, Nov.-Dec. 2010.

[29] C. Wang, Z. G. Qin, J. Peng and J. Wang, "A novel encryption scheme for data deduplication system," 2010 International Conference on Communications, Circuits and Systems (ICCCAS), Chengdu, 2010, pp. 265-269.