Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 7, July 2016, pg.558 – 564

A SURVEY ON EFFICIENT PROTOCOL TO PREVENT DEDUCPILCATION OF DATA IN CLOUD ENVIRONMENTS IN A SAFE AND SECURE MANNER

[1] P. Thangaraju, [2] K. Logeshwari

[1] Associate Professor, [2] Research Scholar

PG Department of Computer Applications

Bishop Heber College, Tiruchirappalli, India

ABSTRACT: Due to the increasing usage of cloud humongous amount of data is getting stored in the cloud data servers. But this also leads to lots of duplicate data over a period of time. While existing systems have focused on the duplication removal also known as deduplication, they have done so only in one server and also not considered safety and security. This means that residual data left in data servers will lead to privacy breaches and leakage of sensitive data. The objective here is to develop a secure deduplication model for data removal from distributed servers in a safe and secure manner.

Keywords: Cloud, Deduplication, Security

INTRODUCTION

Data deduplication in cloud data storage servers is one of important problems in today's era. There is a need for eliminating duplicate and repeating data, which will reduce cloud storage space and more importantly bandwidth is saved. Another problem is that if this process is not safely done residual data will lead to sensitive data being leaked and privacy breached. So in order to protect the confidentiality of the sensitive data in the server during deduplication, the data is encrypted before outsourcing. Also the model proposed only authorized persons to perform deduplication who can be accounted for later. The model is a hybrid cloud architecture which is quite varying from traditional deduplication systems. There are different privileges of users involved in duplicate check. Traditionally two types of deduplication exist when considered in terms of the size: (1) File-level deduplication, which finds the redundancies between different files and removes theseredundancies to reduce capacity demands, and (2) blocklevel deduplication, which finds and removes redundancies between data blocks. The entire file is divided intosmaller fixed-size or variable-size blocks and use

the fixedsize blocks by simplifying the computations of block boundaries meanwhilethe variable-size blocksprovides better deduplicationefficiency than the fixed size.

EXISTING SYSTEM

In existing data deduplication systems in cloud servers, the private cloud is involved as a proxy to allow data ownerusers to perform duplicate checksin a safe and secure way i.e. using differential keys. The data owners only outsource their data storage by utilizing public cloud or website but their data which is managed by the private cloud is outsources to third parties where safety is an issue. But the traditional encryption systems while providing data confidentiality cannot be used with data deduplication because the identical data copies of different users normally leads to different ciphertexts thus rendering deduplication redundant while the security is also compromised.

RELATED WORK

C. Liu, Y. Gu, et al [1] proposed the model R-Admad and addressed the reliability in deduplication has However, they focused only on the traditional files without encryption and did not consider deduplication over ciphertext as is normally implemented. Li et al.[2] showed how to achieve reliable keymanagement in deduplication, but did not mention about the encryption reliability.

Later M. Li, C. Q [3]proposed "Convergent dispersal" model where they showed how toextend the method in for the construction of reliable deduplication for cloud user files stored in the data servers. But all the above mentioned works have not considered tag consistencyand integrity in the construction.

J. R. Douceur [4] proposed the Convergent encryption model which ensures data privacy indeduplication. Bellare etal.proposed formalized a message-lockedencryption scheme and explored its application in spaceefficient secure third party data servers. G. R. Blakley and C. Meadows [5]proposed Bitcasaand deployed convergent encryption, which is used in commercial cloud storage providers.

Quinlan [6] states that Data deduplication is a specialized technique for eliminating repeating or duplicate data found physically in data storage servers where both private and public cloudswhere some critical data is stored while other data is stored in andaccessible from all public cloud servers.

Thomas Ristenpart et al [7]proposed a new cryptographic model called Message-Locked Encryption (MLE), where both encryption and decryption are performed and provides a secure novel way to achieve deduplication which is both space-efficient and occupies less overheadsBut however considering the practically schemes that includes deployed schemes. On the theoretical side the challenge is hash functions secure on correlated inputs and the sample-then-extract model.

Bugiel et al. [8] provided a novel architecture where twin clouds use secure data outsourcing and arbitrarycomputations to an untrusted data storage server. Zhang etal. [9] Also presented the hybrid cloud techniques tosupport privacy-aware data-intensive computing which

considers the authorized deduplicationproblem in public cloud which assumes to be honest approach model.

Santis [10] et al proposed ramp scheme which is nothing but a protocol to distribute a secret s among n users in such a way that where the sets of participants of cardinality greater than orequal to k can reconstruct the secret s; and the sets of participants of cardinality less than orequal to t have no information on s, whereas the sets of participants of cardinality greater than t and have "some" information on s. This model proved a tight lower bound on the size of the shares held by each and every user.

Liet al.[11] in "Secure deduplication with efficient and reliable convergent key management," addressed the key-management issue inblock-level deduplication by spreading these keysacross multiple cloud servers after file encryption. Bellareet al.[12]proposed "Dupless" which showed how to protect data by transforming the predictable data into aunpredicatable dataand using a separate thirdpartydeploying key and also generatesthe tag for that particular stored file in the server. Stanek et al. [13] presented a novel encryption scheme that provideddifferential security for popular and unpopular data stored in the cloud servers.

M. Bellare, C. Namprempr[14] al proposed Convergent encryption, provides data confidentiality indeduplication by providing a convergentkey from the data copy and encrypts the data using this, further user also receives atag for the this tag will be used to detectduplicates and holds, i.e., if two data copies are the same, then their tagsare the same for duplicate detection the user first sends the tagto the server side to check if the identical copy has beenalready stored.

Halevi et al [15]proposed POW - "proofs of ownership" for cloud deduplication systems, so that any client can prove to the cloud storage server that owner owns a data file without actually uploading the physical file to the server. Several POW constructions based on the Merkle Hash Tree enable client-side deduplication. Pietro and Sorniotti et al [16]proposed efficient PoW scheme by choosing the projection of afile on randomly selected bitpositions which served as thefile proof POW, but did not foresee or consider the users privacy at all.

Harnik et al [17] proposed counter attack models so that data leakage in cloud storage servers implement client-side deduplication in the server model itself.Ateniese et al.[18] in "Provable data possession atuntrusted stores"proposed the conceptof proof of data possession (PDP),which allowed the client attached to the cloud data server to verify the data integrity outsourced to the third party server cloud in an effective way.

Xu et al. [19] proposed POWmodel which allows the clientside deduplication detection ina compounded setting. Ng et al. extended the PoWmodel for encryptedfiles stored in the data server but had very high system and communication overheads.

Gaurav Kakariya et al Deduplication [20] modeled Microsoft's Single Instance Server (SIS) and EMC's Centera which used a file level deduplication to eliminate duplicate blocks of data that occur in non-identical files. Dropbox a popular cloud storage uses a fixed-size very large block-level about 4 MB deduplication model. Daehee Kim et al. [21] Deduplication can

occurs at Inline, Post-process, Client-side and Target-based Pooja S Dodamani [22] et al. In Inline deduplication, it occurs before data stored on cloud i.e. performed at the data storage time in the server. It reduces the disk spaceneeded in the system and post the deduplication filters all irrelevant and redundant data from the data set after transferred to a secure storage server.

Whereas the Client-side deduplication model where deduplication occurs at the user side duplicate datais first identified before being sent to the cloud or others requesting but creates heavy overheads. This will definitely create burden on the CPU but atthe same time reduces the load on the network by reducing minimize bandwidth and space needed to uploadand store duplicated data.

NagaMalleswariet al. [23] Almost all Cloud based storage services like Dropbox, Memopal, , wuala, JustCloud, Mozy, Megaupload etc use data deduplication at the source side itself. This is probably done to savenetwork bandwidth and improve the transmission speed apart from conserving space. In Target-based de-duplication model the service will remove all deduplication redundancies from a backup taken already the cloud source and the target client server. However unlike cloud source deduplication, the Target deduplication does not minimize the total amount of data that is transferred instead reduces the storage space based on availability at the storage server Boga

Venkatesh et al. [24] says that data deduplication brings a lot of benefits apart from offering security where sensitive data aresusceptible to attacks from many other users inside and outside the network. Standout traditionally used encryption are not enough because it is not compatible with data duplication finding. This is mainly because users may use different techniques with different keys so making it impossible to detect duplication.

S. Keelveedhi et al. [25] the solution for balancing confidentiality and efficiency in deduplication was described by M.Bellare et al called convergent encryption. It has been proposed to enforce data confidentiality while makingdeduplication. It encrypts/decrypts a data copy with a convergent key, which is derived by computing thecryptographic hash value of the content of the data copy itself Xiaofeng Chen [26]. To prevent unauthorized access, a secure proof ofownership protocol Jin. Li, Yan Kit Li [27]proposed a model where users provide proof, and if the proof is verified, other users with the same file will be provided a pointer from the cloud server and the file will not be uploaded again.Existing deduplication systems does not find differential authorization duplicate checks easily.In anauthorized deduplication system, each user is issued a set of privileges during system initialization. Each fileuploaded to the cloud is also bounded by a set of stating which users are allowed to access and perform duplicate checks before uploading. The privileges of each user are stored separately.

Juels et al. presented the POR concept - proof ofretrievability which allowedcloud users to get the outsourced data stored across many data servers throughdynamic interactions. Shacham and Waters [29] extends POR with “Compact Proofs of Retrievability”which utilized thedifference between the two notions by proposing PoR that usesError CorrectionCodes to be tolerant towards the outsourced data which are damaged.

CONCLUSION

Thus the proposed the distributed deduplication model improves the reliability of data without security breach and also ensures the confidentiality of the users' outsourced data. The mode supports both file-level and fine-grained block-leveldata deduplication. The tag consistency andintegrity checks are achieved in the proposed deduplication model using the Ramp secret sharing scheme and also the overheads are small compared to all the existing models.

References

[1] C. Liu, Y. Gu, L. Sun, B. Yan, and D. Wang, "R-admad: High reliability provision for large-scale de-duplication archival storagesystems," in Proceedings of the 23rd international conference onSupercomputing, pp. 370–379.

[2] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with Efficient and reliable convergent key management," inIEEE Transactions on Parallel and Distributed Systems, 2014, pp.vol.25(6), pp. 1615–1625.

[3] M. Li, C. Qin, P. P. C. Lee, and J. Li, "Convergent dispersal:Toward storage-efficient security in a cloud-of-clouds," in The 6thUSENIX Workshop on Hot Topics in Storage and File Systems, 2014

[4]. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer,"Reclaiming space from duplicate files in a serverless distributedfile system." in ICDCS, 2002, pp. 617–624.

[5] G. R. Blakley and C. Meadows, "Security of ramp schemes," inAdvances in Cryptology: Proceedings of CRYPTO '84, ser. LectureNotes in Computer Science, G. R. Blakley and D. Chaum, Eds.Springer-VerlagBerlin/Heidelberg, 1985, vol. 196, pp. 242–250

[6] S. Quinlan and S. Dorward. Venti: a newapproachtoArchival storage. In Proc. USENIX AST, Jan,2002.

[7]T. Ristenpart.Message-locked encryption and securededuplication. In EUROCRYPT, pages 296– 312,2013

[8] S. Bugiel, S. Nurnberger, A. Sadeghi, and T.Schneider. Twin clouds: An architecture for securecloudcomputing. In Workshop on Cryptographyand Security in Clouds (WCSC 2011), 2011.

[9] P. Anderson and L. Zhang. Fast and secure laptopBackups with encrypted de-duplication. In Proc. OfUSENIX LISA, 2010

[10]A. D. Santis and B. Masucci, "Multiple ramp schemes," *IEEETransactions on Information Theory*, vol. 45, no. 5, pp. 1720–1728,Jul. 1999.

[11]J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management," in*IEEE Transactions on Parallel and Distributed Systems*, 2014, pp. vol.25(6), pp. 1615–1625

[12] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Serveraided encryption fordeduplicated storage," in *USENIX SecuritySymposium*, 2013

[13]J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A securedata deduplication scheme for cloud storage," in *Technical Report*,2013

[14] M. Bellare, C. Namprempre, and G. Neven, Security proofs for identity-based identification and signature schemes.J Cryptology, 22(1):1–61, 2009

[15]S. Halevi, D. Harnik, B. Pinkas, and A. ShulmanPeleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM,2011.

[16]R. D. Pietro and A. Sorniotti, "Boosting efficiency and securityin proof of ownership for deduplication." in *ACM Symposiumon Information, Computer and Communications Security*, H. Y. Youmand Y. Won, Eds. ACM, 2012, pp. 81–82.

[17]D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels incloud services: Deduplication in cloud storage." *IEEE Security &Privacy*, vol. 8, no. 6, pp. 40–47, 2010.

[18] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner,Z. Peterson, and D. Song, "Provable data possession atuntrusted stores," in Proceedings of the 14th ACM conferenceon Computer and communications security, ser. CCS '07. NewYork, NY, USA: ACM, 2007, pp. 598–609. Available:http://doi.acm.org/10.1145/1315245.1315318

[19] J. S. Plank and L. Xu, "Optimizing Cauchy Reed-solomon Codesfor fault-tolerant network storage applications," in NCA-06: 5thIEEE International Symposium on Network Computing Applications,Cambridge, MA, July 2006

[20] Gaurav Kakariya, Prof. Sonali Rangdale, "A Hybrid Cloud Approach For Secure AuthorizedDeduplication", International Journal ofComputer Engineering and Applications, Volume VIII, Issue I, October 14.

[21]Daehee Kim, Sejun Song , Baek-Young Choi, "SAFE: Structure-Aware File and Email Deduplication for Cloud-based Storage Systems".

[22] Pooja S Dodamani, Pradeep Nazareth, "A Survey on Hybrid Cloud with De-Duplication", International Journal of Innovative Research inComputer and Communication Engineering, December 2014

[23] T.Y.J. NagaMalleswari, D.Malathi, "Deduplication Techniques: A Technical Survey", International Journal for Innovative Research in Science & Technology, December 2014.

[24] Boga Venkatesh, Anamika Sharma, Gaurav Desai, Dadaram Jadhav, "Secure Authorized Deduplication by Using Hybrid Cloud Approach",November 2014.

[25] S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication", inProc. IACR Cryptology ePrintArchive, 2012

[26] Xiaofeng Chen, M. Li, J. Li, P. Lee, and W. Lou., "Secure Deduplication with Efficient and Reliable Convergent KeyManagement", In IEEE Transactions on Parallel and Distributed Systems,June- 2014

[27] Jin. Li, Yan Kit Li, Xiaofeng, P. Lee, and W. Lou., "A Hybrid Cloud Approach for Secure Deduplication" ,In IEEE Transactions onParallel and Distributed Systems, 2014.

[28] A. Juels and B. S. Kaliski, Jr., "Pors: proofs of retrievabilityfor large files," in Proceedings of the 14th ACM conferenceon Computer and communications security, ser. CCS '07. NewYork, NY, USA: ACM, 2007, pp. 584–597.

[29]H. Shacham and B. Waters, "Compact proofs of retrievability," inASIACRYPT, 2008, pp. 90–107.

