IJCTA, 9(27), 2016, pp. 431-436 ISSN: 0974-5572
© International Science Press

An Efficient Protocol to Prevent Deducpilcation of Data in Cloud Environments in a Safe and Secure Manner

P. Thangaraju* and Logeshwari K.**

ABSTRACT

Due to the increasing usage of cloud humongous amount of data is getting stored in the cloud 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

1. 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 ofthe size: 1File-level deduplication, which finds the redundancies between different files and removes these redundancies 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 fixed size blocks by simplifying the computations of block boundaries meanwhilethe variable-size blocksprovides better deduplicationefficiency than the fixed size.

2. EXISTING SYSTEM

In existing data deduplication systems in cloud servers, the private cloud is involved as a proxy to allow data owner users to perform duplicate checks in 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 cipher texts thus rendering deduplication redundant while the security is also compromised.

* Associate Professor, Department of Computer Applications, Bishop Heber College, Tiruchirappalli, India, *Email: thangarajubhc@yahoo.co.in*

** Research Scholar, Department of Computer Science, Bishop Heber College, Tiruchirappalli, India, *Email: logasuki13@g.mail.com*

3. RELATED WORK

C. Liu, Y. Gu, et al [1] proposed the model R-Admad and addressed the reliability in deduplication hasHowever, theyfocused only on the traditional files without encryption and did not consider deduplication over cipher text 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-locked encryption 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.

Thomas Ristenpart et al [7] proposed a new cryptographic model called Message-Locked Encryption (MLE). Bugiel et al. [8] provided anovel architecture where twinclouds use secure data outsourcing and arbitrarycomputations to an untrusted data storage server. Zhang et al.[9] also presented the hybrid cloud techniques tosupport privacy-aware data-intensive computing. 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. Li et 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 aunpredictable data J.Stanek [13] it presented a novel encryption scheme that provided differential security for popular and unpopular data.For popular data that are not particularly sensitive, the conventional encryption is performed.M.Bellare[14] It predicable transform data is use predicatable message into a unpredicatable messages.with third party called key server generate the file tag for the duplicate check.

S. Halevi [15] proposed the notation of "proofs of owner ship"(POW) for deduplication system,so that a client can efficiently prove to the cloud storage server that he/she owns a file without uploading the file itself. R.D.Pietro and A.Sorniotti[16] proposed another efficient POW scheme by choosing the projection of a file onto some randomly selected bit-positions as the file proof. D.Harnik, B.Pinkas et al[17] presented a number of attacks that can lead to data leakage in a cloud storage system supporting client-side deduplication

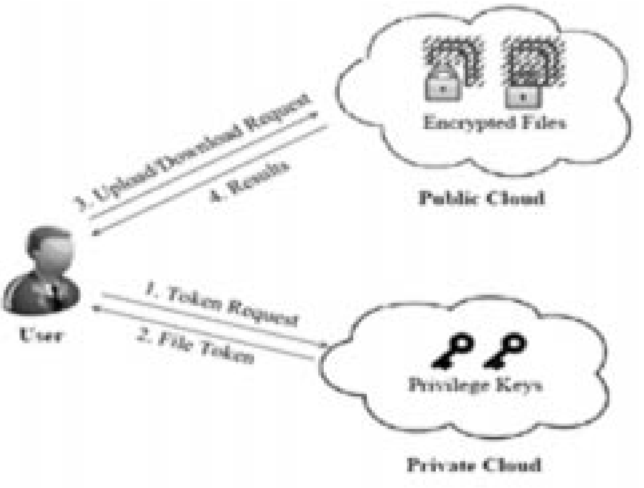
G. Ateniese,R.Burns [18] Introuced the concept of data possession (PDP).this notion was introduced to allow a cloud client to verify the integrity of its data outsourced to the cloud in a very efficient way.[19]presented a POW scheme that randomly choosen selected bit position in file proof. S.Keelveedhi [20] Formalized this primitive as message locked encryption and explored in use serveral implentions of convergent

4. PROPOSED MODEL

The proposed system is secure, takes less space and is efficient and inherently the model supports better security with differential privilege keys. In this way, the users without corresponding privileges cannot perform the duplicate check. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP. Security analysis demonstrates that our system is secure in terms of the definitions specified in the proposed security model.

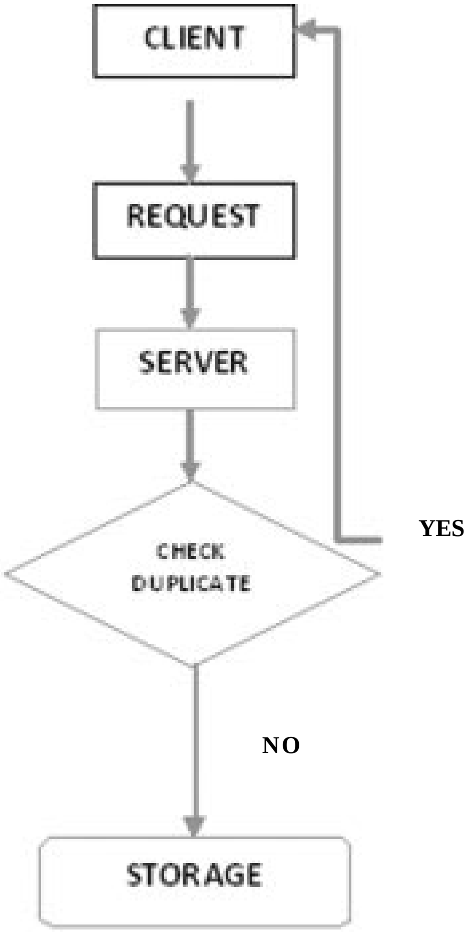
5. ADVANTAGES OF PROPOSED SYSTEM:

The user is only allowed to perform the duplicate check for files marked with the corresponding privileges. We present an advanced scheme to support stronger security by encrypting the file with differential privilege keys. Reduce the storage size of the tags for integrity check. To enhance the security of deduplication and protect the data confidentiality,



With the Distributed Benefit-based caching in cloud computing presented in Distributed cooperative caching DACHE section, when there is no enough space in the cache for accommodating a new object, the existing object with the minimum benefit is identified and replaced with the new one. It will be done if the new object shows more total benefit. The newly requested object is then checked in the downloaded cached and in case the content object is requested from another node present in the deduplication partition which means that one extra copy already exists in the cache then such a n object is labeled as duplicate. The new object is placed in the cache thereby serving all the requests placed by nodes avoiding duplicate request.

6. DATAFLOW



7. ALGORITHM STEPS (DEDUPLICATION)

```
Step 1:   Client Requests Cloud
Step 2:   Data is Indexed
Step 3:   Checks in the Index
Step 4:   Whether already available in
Step 5:   Requested index cache
Step 6:   If available
Step 7:   Take from cache
Step 8:   Send data to client via
Step 9:   Main server
Else
Step 10:  Request from storage Node
Step 11:  Get data
Step 12:  Store in cache
Step 13:  Update Index Node
Step 14:  Send to Client Via Server
```

8. COMPUTAION

$$Cd = Ri \ldots..\{d/Nn\}_{Dup}^{t}$$

For each of the Cache Data (Cd) requested the number of requests R for each data d by number of Nodes (Nn) identify the data content (d) with respect to the iterations find the duplicate data (Dup).

9. EVALUATION AND DISCUSSION

The model of benefit based object caching in cloud computing reduces duplicates. This resolves the space problems in the cache foraccommodating extra objects. The existing model places objects which are identified and replaced with the object only in case the new object has changes. The downloaded cache size is lesser than the original when the duplicate is avoided. When the object is downloaded from another node in the deduplication cache within the size allotted partition where a primary object j in the partition must be stored.

In certain rare situations the object status modification process fails to satisfy the above constraint. In certain models consider a situation in which only one node in the network generates requests and other nodes make no requests. In Deduplication at cloud levels, because of space problems only one node can only store objects and this object status modification process does not help that particular node to get rid of objects to the extra nodes in the client side. Thus the model of offloading extra objects to other nodes needs extra protocol syntax and requires additional power from the cloud and overhead in the algorithm. Thus two client nodes may consider an object as

Table 1
Represent of Parameter and Accuracy

Methods	*Accuracy*	*Security*
Existing	75	67
Proposed	95	98

primary copy while they are in the same deduplication partition. The caching may result in storing extra copies of some objects. Due to these inconsistencies Deduplication heuristics does not guarantee a cost-optimal object placement.

The availability of correlations among base stations with varying traffic volumes leads us to the fact that nodes requesting identical data can be identified on cache content and deduplication costs can be avoided.

The focus here in case of cellular optimization is to balance the loads of the nearby or proxmity base stations. If not balanced in a network one base station will have abnormally high load while its neighbor nodes have only small traffic. This directly leads to computational overheads, route inefficiency and bandwidth resource wasting.

The proposed protocol model reduces overheads, increaes accuracy and efficency by spatial positioning and the results and findings are discussed below.

First by verifying the long-range time-correlation of the call arrivals at the base station, it is found that call arrivals in a minute are uncorrelated to the number of call arrivals in another minute in short-term.

Second, it is found that correlation of time to call arrivals is governed by the arrival time and the present location of base stations.

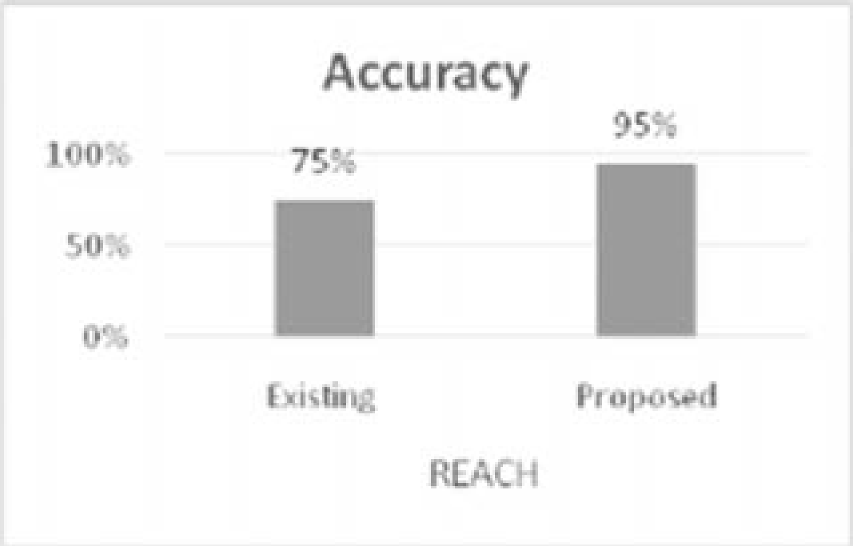


Figure 2: Representation of parameters in existing and proposed methodologies (Chart Showing Accuracy)

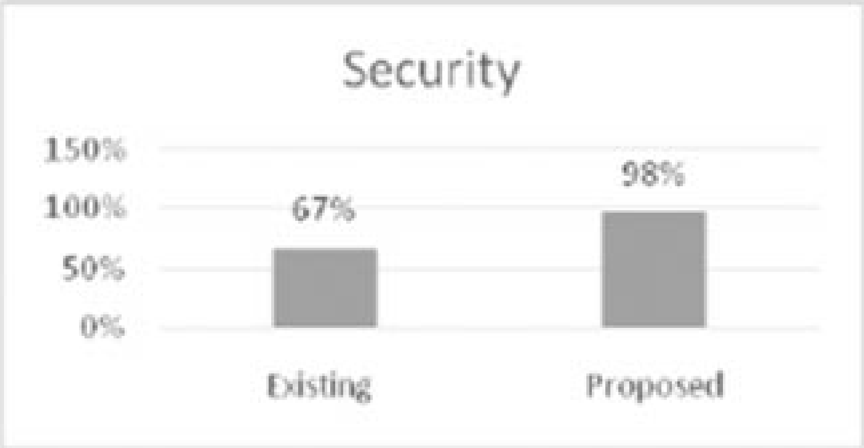


Figure 3: Representation of parameters in existing and proposed methodologies (Chart Showing Security)

10. 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-level data deduplication. The tag consistency and integrity 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 on Supercomputing,* 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,* vol.25(6), pp. 1615–1625,2014.

[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 server less distributed file system." in ICDCS, pp. 617–624,2002

[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, vol. 196, pp. 242,1985.

[6] S. Quinlan and S. Dorward. Venti: a new approach to Archival storage. In Proc. USENIX AST, Jan,2002.

[7]T. Ristenpart.Message-locked encryption and secure deduplication. In EUROCRYPT, pp. 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*,pp. vol.25(6), pp. 1615–1625,2014.

[12] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Server aided encryption for deduplicated storage," in *USENIX SecuritySymposium,* 2013

[13] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data deduplication scheme for cloud storage," in *Technical Report*, [14] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Serveraided encryption or deduplicated storage," in *USENIX Security Symposium*, 2013.

[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, ACMC conference on Computer and Communications Security, pp.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, pp. 81–82,2012.

[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, pp. 598–609, 2007. Available:http://doi.acm.org/10.1145/1315245.1315318

[19] J. Xu, E.-C. Chang, and J. Zhou, "Weak leakage-resilient client-side deduplication of encrypted data in cloud storage," in *ASIACCS,* pp. 195-206, 2013

[20] S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication", in Proc. IACR Cryptology Print Archive, 2012.