International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 1, January-February 2015 **ISSN 2278-6856**

Hybrid model for Authorized De-duplication in Cloud

Miss Prachi D. Thakar[1], Dr. D.G.Harkut[2]

[1]M.E. (C.E.) II Year
Prof.Ram Meghe College of Engineering and Management, Amravati, India

[2]Department of Computer Science and Engineering,
Prof.Ram Meghe College of Engineering and Management, Amravati, India

Abstract

Cloud computing provides a more cost effective environment to outsource storage computation. Many enterprises need to store and operate huge amount of data. One serious problem of today's cloud storage service is the management of ever-increasing volume of data. To make data management elastic in cloud computing de-duplication system is used. Data de-duplication is a compression technique that improves storage efficiency by eliminating redundant data. In this approach we have proposed a de-duplication technique which is different from traditional de-duplication system. In this system users with differential privileges are also consider in duplicate check which was not possible in previous de-duplication system. Again in this approach we also focus on the confidentiality of sensitive data in support of de-duplication. We are proposing a Hybrid cloud approach. In Hybrid cloud approach two cloud are maintained, public cloud and a private cloud. A private cloud is working as an interface between user and public cloud. Private cloud provides a set of private key to the user. We also presented several new de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture.

Index Terms- Hybrid Cloud, De-duplication

1. INTRODUCTION

Cloud storage is gaining popularity in recent years. Cloud computing is a model for delivering information technology services in which resources are retrieved from the internet through web-based tools and application, instead of direct connection to a server. Cloud provides scalable, location independent infrastructure for data management in storage. To make data management scalable in cloud computing, de-duplication has been a well-known technique and has attracted more and more attention recently. Data de-duplication is a technique that stores only a single copy of redundant data, and provides links to that copy instead of storing other actual copies of that data. By storing and transmitting only a single copy of duplicate data, de-duplication offers savings of both disk space and network bandwidth. De-duplication is either file level or block level. In file level de-duplication duplicate copy of same file is eliminated and in block level duplicate block of data is eliminated. The main issue with data de-duplication is security and privacy as users sensitive data are susceptible to both insider and outsider attacks. Tradition encryption, while provide data confidentiality, is incompatible with data de-duplication. In traditional encryption technique, each user encrypts their data with their own key. Thus, identical data copy of different user will lead to different cipher text which makes data de-duplication impossible. So in our proposed model we are using a technique which will take care of both security and scalability.

Again the traditional de-duplication system cannot support differential authorization duplication check [1] .In differential authorized duplicate check, users with different privileges on same file may also be consider in duplicate check ,i.e. if a file has two users and both user have different privilege still a single copy of file get stored.

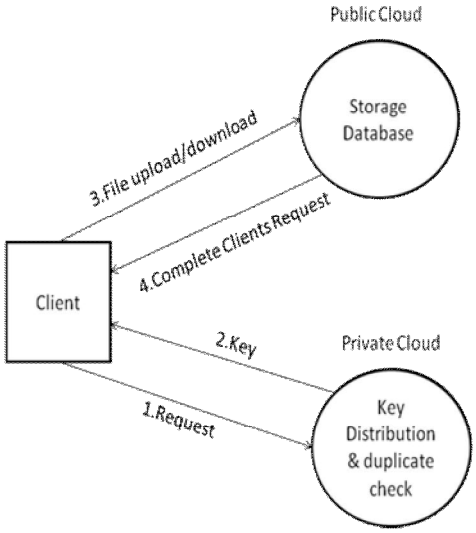

Figure 1 Autherized de-duplication model

2. MOTIVATION

Cloud storage services are becoming very popular now a day. Cloud provides a better way of storage with efficient cost. One major problem with cloud is to manage huge amount of data. In order to manage data de-duplication technique is used. Although, de-duplication has many advantages but it has some security issues. This motivates us to propose a model which manage the security issues of de-duplication and provide authorized de-duplication in cloud.

3. LITERATURE REVIEW

A Jan Stanek, Alessandro Sorniotti, Elli Androulaki, and Lukas Kencl [2] differentiate the data according to their popularity for secure data de-duplication. The unpopular data is consider as sensitive and given semantic security and the popular data that is used by many user is consider less sensitive so given weaker security and better storage. A multi layered cryptosystem has been proposed, convergent cryptosystem and threshold cryptosystem. The unpopular files are protected using two layers, whereas the popular files are protected using single layer. The main focus is on varying layer of security and applying de-duplication on data.

Shai Halevi, Danny Harnik, Benny Pinkas, and Alexandra Shulman-Peleg [3] mainly focuses on client-side de-duplication .In client-side de-duplication if a client want to upload any file on the server then it first send the hash value of the file to the server and the server check whether it is present or not, if present the it inform the client that there is no need to store the file, otherwise it store the file .But the client-side de-duplication introduces an problem. In both the cases the server marks the client as the owner of the file and at that point there is no difference between the owner and the client. Again if any person got the hash values that he/she can access that file. To overcome such problem proof-of-ownership is introduced. In this the owner has to prove to the server that he is the actual owner of that file without sending the actual file. A streaming protocol is introduced.

Mark W. Storer, Kevin , Darrell D. E. Long, Ethan L. Miller [4] mainly deal with the contradiction between the de-duplication and encryption .De-duplication save a single of same file while in case of encryption , same file or same content encrypted with two different key result in different cipher text . That means same file or content is getting store in the server. To solve this problem author propose an approach in which the encryption keys are generated in a consistent manner from the chunk of data. They present two approaches for secure de-duplication, authenticated and anonymous. Both the approaches can be applied to single server as well as distributed storage server. Both the model offer same feature with slight difference. In this model for security convergent encryption is used. Again in this model plan text is never sending to server all the encryption is done on client side.

Wee Keong Ng, Yonggang Wen, and Huafei Zhu [5] introduce de-duplication technique for private data storage. Here the client has to prove his identity by using proof-of-ownership protocol. A client who holds the private data proves the server that he/she is holding private data by just giving summary string of the file, and not reveling the entire file. This private data de-duplication protocol is secure in simulation based framework.

Sven Bugiel, Stefan N□urnberger, Ahmad-Reza Sadeghi, and Thomas Schneider [6] mainly dealing with the secure outsourcing of data. Cloud computing provide cost-effective and flexible data storage. The customer of the cloud has not only trust the security mechanism and configuration of the cloud provide, but also the cloud provider also. When data is outsource many security risks are there such a malicious code running on cloud, cloud provide can misuse their capability and leak data. The requirement of cloud client is confidentiality of their data. The author proposes two clouds (twin), a trusted loud and commodity cloud. The working of both clouds is different. The trusted cloud performs security critical operation such as encryption and the commodity cloud perform the time critical operation on the encrypted data. Client first interacts with the trusted cloud and then the commodity cloud.

John Douceur, Atul Adya, William J. Bolosky, Dan Simon and Marvin Theimer [7] mainly focus on Farsite [10] Distributed file system. To reclaim used space they have used convergent encryption, in which if a same file is encrypted with different technique still a single file get stored .They have also mentioned that this technique is very beneficial for many file system such as freenet [8] .

Ian Clarke, Oskar Sandberg, Brandon Wiley, and Theodore W. Hong [8] propose a peer-to-peer network application which provides an effective means of anonymous information storage and retrieval. Freenet system is designed to address the problem of data privacy and data availability.

Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou [9] propose an efficient and reliable way for secure de-duplication by using convergent key management .They proposed a new concept called as Dekey, in which there is no need for key management .There is secure distribution of convergent key shares across multiple servers. The dekey support client-side de-duplication with POW [3]. They implement dekey using Ramp secret sharing scheme.

Atul Adya, William J. Bolosky, Miguel Castro, Gerald Cermak, Ronnie Chaiken, John R. Douceur, Jon Howell, Jacob R. Lorch, Marvin Theimer, Roger P. Wattenhofer [10] describe farsite which is a serverless distributed file system that logically works as centralized file server but physically distributed among a set of untrusted computer. This file system provides the benefit of both, the centralized file server and the local desktop file system. Farsite is designed to support the file I/O workload of desktop computers.

Chun-Ho Ng and Patrick P. C. Lee [11] propose a model which is different from the other de-duplication model. They design a de-duplication system for VM disk image backup in virtualization model. De-duplication is useful for space storage but it also introduce fragmentation. To reduce this revDedup a deduplication system removes duplicate from old data and not from the new. If in de-duplication some repeated content are present then the old one is replaced by the new one.

Mihir Bellare , Sriram Keelveedhi and Thomas Ristenpart [12] came up with a new model Dupless , in which there is a group of affiliated clients encrypt there data with the help of key server , a server which is different from the storage server .The use of key server is to store the key

used for encryption . The client has to authenticate the key server without leaking any information of their data. Dupless is a system that combines a convergent encryption type based Message lock encryption scheme [13] with the ability to obtained message-derived keys with the help of key server. They also show that this system is easy to deploy: it is transparently work on top of any storage server such as Google drive. This system can also work in conjunction with technique such as POW [3] .

Mihir Bellare, Sriram Keelveedhi, Thomas Ristenpart [13] proposed a scheme in which the key is derived from message itself. This derived key is used for encryption and decryption purpose. This message lock encryption is symmetric encryption scheme, provides a way through which the privacy and integrity of the file can be preserve.

Danny Harnik, Benny Pinkas and Alexandra Shulman-Peleg [14] focuses on the cross-user de-duplication .They demonstrate that de-duplication can be a channel of privacy leakage in cross user de-duplication .It can be used as a side channel that reveal information of the file. They have three solutions on this issue. The first solution is stop encryption in cross user de-duplication , second is to perform de-duplication at server side and the third solution is to provide some random threshold for every file and perform de-duplication if and only if the number of copies of the file exceed the given threshold.

Jia Xu, Ee-Chien Chang, Jianying Zhou [15] came up with a approach for client-side de-duplication of encrypted file in cloud storage by taking a reference of POW [3].They enhanced and generalized the convergent encryption method. The proposed scheme protects data confidentiality against both the outside adversaries and honest-but-curious cloud server. This scheme allows a bounded amount one-time leakage of target file before it starts to execute as compared to POW.

Jiawei Yuan, Shucheng Yu [16] focuses on data integrity and storage efficiency for cloud storage. Proof of retrivability [17] and proof of data processing technique[18] assure data integrity and the POW[3] assure the storage efficiency .But the combination of these two technique increase the storage overhead which contradict the POW scheme[3]. To solve this problem a novel scheme is introduce which include polynomial based authentication tags and homomorphic linear authenticator .This scheme allow de-duplication of both the files and the corresponding authentication tag.

Giuseppe Ateniese, Randal Burns, Reza Curtmola, Joseph Herring, Lea Kissner, Zachary Peterson, Dawn Song [17] introduce a model that allow the cloud client that stores its data at an untrusted server to verify that the server possesses the original data without retrieving that data.

Ari Juels, Burton S. Kaliski Jr. [18] came up with a protocol that unable a user that is verifier to determine that a prover possesses a file and the verifier can retrieve that file in secure way.

Zhike Zhang, Preeti Gupta, Avani Wildani, Ignacio Corderi, Darrell D.E. Long [19] proposes a different way of de-duplication. In de-duplication, the number of shared data chunks increase due to this fragmentation of data occurs. This increase seek operation and degrades the system performance .To solve this they propose reverse de-duplication. In traditional de-duplication system as new data added the checking is done and if that same data exists in the storage then the newer data are not stored. In reverse de-duplication the newer data is written contiguously and old data segment that share chunks with the newer segment will reference those chunks.

Chun-Ho Ng, Mingcao Ma, Tsz-YeungWong, Patrick P. C. Lee, and John C. S. Lui [20] propose live DFS, live de-duplication file system that enable de-duplication storage of VM images in an open source cloud. This file system allows general I/O operation such as read, write, modify and delete, while enabling in-line de-duplication. It mainly target for VM image storage. They mainly focus on single storage partition.

Yu Hua, Xue Liu, Dan Feng [21]propose an in-network de-duplication for storage aware software defined network for the efficient handling of data and to reduce the overhead of network transmission . In source side de-duplication there is latency of communication of message from source to destination. In destination de-duplication there is heavy consumption of resources. To overcome this problem, in-network de-duplication scheme is introduce.

Deepa Karunakaran and Rangarajan Rangaswamy [22] have used the ABC (Artificial Bee Colony) algorithm for detection of the duplicate record in the data set. The ABC algorithm is used to generate the optimal similarity measure. By using this optimal similarity the de-duplication of remaining data set is done. They again evaluate the performance of ABC and genetic algorithm based technique.

Sean Quinlan and Sean Dorward [23] focus on the archival data. This approach adopt write-once policy which prevent the accidental destruction of data. This model also has the ability to coalesce duplicate copies of block.

Jingxin Feng, Jiri Schindler [24] focuses on content de-duplication in host side cached data. De-duplication in the host-side cache improves the cache hit rate. This cache de-duplication also shift load from networked storage system when most VM instances perform the same operation such as virus scan. This also improves cache effectiveness. They have also defined a term de-duplication metric which has three parameters, how to improve cache hit rate, how much space can be reclaimed by eliminating duplicate copy and how to structure cache meta-data.

Pasquale Puzio, Refik Molva, Melek □nen, Sergio Loureiro [25] designed a system which achieve confidentiality and block level de-duplication. There system is based on convergent encryption. Convergent encryption introduces a new overhead of key management. Author includes a new component for management of key for each block together with the actual de-duplication operation. This design mainly focuses on the two basic operations in cloud storage, storage and retrieval.

Nesrine Kaaniche, Maryline Laurent [26] proposes a scheme for securely storing and sharing outsourced data.

They propose a new cryptographic method. This scheme is based on convergent encryption and Merkle based tree.

Can Wang, Zhi-Guang Qin, Jing Peng Juan Wang [27] propose a novel approach in which basic encryption is transformed from the file to the chunk . Symmetric key is used for ciphering. This symmetric key is generated from the chunk content.

Yufeng Wang, Chiu C Tan, Ningfang Mi [28] came up with an algorithm known as elasticity aware de-duplication to improve de-duplication performance .It is an indexing algorithm that uses an ability to dynamically increase memory resource. This algorithm has the capability of dynamically adjust the computing resource.

K. Zhang, X. Zhou, Y. Chen, X. Wang, Y. Ruan [29] came up with the concept of hybrid cloud computing that support data intensive computing framework. They differentiate the cloud as private and public. The computation of private data is on private cloud and the computation on non-sensitive data is done on the public cloud. This technique uses the special feature of Mapreduce to automatically partition the computing job according to the security level of the data. To reduce the inter cloud communication this model automatically analyzed the reducer of a legacy mapreduce job to extract a combiner for aggregating the map outcome on the public cloud.

Qingji Zheng, Shouhuai Xu [30] develop a new scheme which is a combination of PDP [17],PORs [18] and POW [3] that provide both security and efficiency in cloud storage. Actually the PDP, POR are contradictory to POW, but the authors came up with the novel scheme of combining this technique and call it as POSD.

4. COMPARATIVE STUDY

Table 1: Comparisons of Existing Model

Sr. No.	Methodology	Performance Evaluation			
		Security	Storage Efficiency	Computation Overhead	Bandwidth Consumption
1	Multi-layered cryptosystem [2]	Moderate	Moderate	High	Low (for popular data)
2	Merkle-hash tree [3]	Low	Moderate	Low	Low
3	Convergent encryption & SALAD [7]	Moderate	High	High	-
4	Ramp secret sharing scheme [9]	Moderate	High	Low	Low
5	Authenticated and Anonymous model [4]	Moderate	High	High	-
6	2-party computational modal [5]	High	High	Moderate	-
7	Two cloud Architecture [6]	High	-	Low	Moderate
8	Client-Server Modal [11]	-	Moderate	High	High
9	CE based MLE scheme [12]	High	High	High	High
10	Message lock encryption [14]	High	High	High	-
11	Client-side de-duplication scheme [15]	Moderate	Moderate	-	-
12	PCAD scheme polynomial based authentication tags and homomorphic linear authenticator [16]	High	High	Moderate	Low
13	Reverse De-duplication [19]	Moderate	Moderate	High	-
14	LiveDFS [20]	Moderate	Moderate	Moderate	-
15	POSD [30]	Moderate	Moderate	High	High
16	Hybrid cloud [29]	High	-	Low	-

We have done the comparative study of different approaches regarding de-duplication in cloud. We have also identified their implementation techniques. We analyze the limitations of different techniques and we overcome some of the limitation to increase the performance and efficiency of system in our proposed model.

We proposed a model that can protect the data by including differential privileges of users in the duplicate check. The owner of the file decides what type of privileges given to other user who want to access that file. In our proposed model we can protect sensitive data by using the hybrid cloud concept. In hybrid cloud there are two clouds one private and one public. The client has to first interact with the private cloud. All the duplicate check and key distribution is done on private cloud only. In our proposed model system generated key is used, so the privacy is preserved. The public cloud is used for only storage purpose.

5. CONCLUSION

In this paper we have reviewed different existing models and techniques for authorized de-duplication in cloud storage. By analyzing the existing system we proposed a hybrid cloud approach for authorized de-duplication in which we differentiate two cloud as private and public and the de-duplication is done on private cloud .Public cloud is used only for storage purpose.

References

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, ”A Hybrid Cloud Approach for Secure Authorized De-duplication,”IEEE Transactions on Parallel and Distributed System, vol. PP, no.99, 2014.

[2] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl,” A secure data de-duplication scheme for cloud storage,” In Technical Report, 2013.

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

[4] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, “Secure data de-duplication,” In Proc. of StorageSS, 2008.

[5] W. K. Ng, Y. Wen and H. Zhu, “Private data de-duplication protocols in cloud storage, ”In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446, ACM, 2012.

[6] S. Bugiel, S. Nurnberger, A. Sadeghi and T. Schneider, “Twin clouds: An architecture for secure cloud computing, ” In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011

[7] 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 ICDCS, pages 617–624, , 2002.

[8] I. Clarke, O. Sandberg, B. Wiley, and T. Hong,” Freenet: A Distributed Anonymous Information

Storage and Retrieval System,"ICSI Workshop on Design Issues in Anonymity and Unobervability, Jul 2000

[9] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure de-duplication with efficient and reliable convergent key management, "In IEEE Transactions on Parallel and Distributed Systems, 2013

[10] Atul Adya, William J. Bolosky, Miguel Castro, Gerald Cermak, Ronnie Chaiken, John R. Douceur, Jon Howell, Jacob R. Lorch, Marvin Theimer, Roger P. Wattenhofer,"FARSITE: Federated, available, and reliable storage for an incompletely trusted environment," ACM SIGOPS Operating Systems Review, 2002.

[11] C. Ng and P. Lee,"Revdedup: A reverse de-duplication storage system optimized for reads to latest backups, "In Proc. of APSYS, Apr 2013.

[12] M. Bellare, S. Keelveedhi and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage, "In USENIX Security Symposium, 2013.

[13] Shulman-Peleg A. Harnik D. and Pinkas B, "Side Channels in Cloud Services: Deduplication in Cloud Storage, "In IEEE Security and Privacy Magazine, specialissue of Cloud Security, 2010.

[14] M. Bellare, S. Keelveedhi and T. Ristenpart,"Message-locked encryption and secure de-duplication. "In EUROCRYPT, pages 296– 312, 2013.

[15] J. Xu, E.-C. Chang, and J. Zhou,"Weak leakage-resilient client-side deduplication of encrypted data in cloud storage,"In Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, ASIA CCS '13, pages 195–206, New York, NY, USA, 2013.

[16] Jiawei Yuan and Shucheng Yu," Secure and Constant Cost Public Cloud Storage Auditing with Deduplication,"IEEE Conference, pages 145-153, 2013.

[17] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z, " Peterson and D. Song. Provable data possession at untrusted stores, "In ACM CCS '07, pages 598–609, 2007.

[18] Ari Juels and Burton S. Kaliski,"Pors: proofs of retrievability for large files, "In CCS '07: ACM conference on Computer and communications security, pages 584–597, 2007.

[19] Zhike Zhang, Preeti Gupta, Avani Wildani, Ignacio Corderi, Darrell D.E. Long,"Reverse Deduplication: Optimizing for Fast Restore, ".

[20] Chun-Ho Ng, Mingcao Ma, Tsz-YeungWong, Patrick P. C. Lee, and John C. S. Lu, " Live Deduplication Storage of Virtual Machine Images in an Open-Source Cloud, "Proceedings of ACM/IFIP/USENIX 12th International Middleware conference, 2011.

[21] Yu Hua, Xue Liu, Dan Feng,"Smart In-Network Deduplication for Storage-aware SDN, " Proceedings of ACM SIGCOMM, and ACM SIGCOMM Computer Communication Review, Volume 43, Issue 4,pages: 509-510, 2013.

[22] Deepa Karunakaran and Rangarajan Rangaswamy,"Optimization Techniques To Record Deduplication, "In Journal of Computer Science 8 (9): 1487-1495, 2012.

[23] S. Quinlan and S. Dorward,"Venti: a new approach to archival storage, "In Proc. USENIX FAST, Jan 2002.

[24] Jingxin Feng and Jiri Schindler, "A Deduplication Study for Host-side Caches in Virtualized Data Center Environments, "In IEEE computer, 2013.

[25] Pasquale Puzio, Refik Molva, Melek nen, Sergio Loureiro,"ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage, "In IEEE CloudCom, 2013.

[26] Nesrine Kaaniche and Maryline Laurent,"A Secure Client Side Deduplication Scheme in Cloud Storage Environments, "IFIP International Conference on New Technologies, Mobility and Security, 2014.

[27] C. Wang, Z. guang Qin, J. Peng, and J. Wang,"A novel encryption scheme for data deduplication system, "In Communications, Circuits and Systems (ICCCAS), 2010 International Conference on, pages 265–269, ,2010

[28] Yufeng Wang, Chiu C Tan, Ningfang Mi,"Using Elasticity to Improve Inline Data Deduplication Storage Systems, IEEE International Conference on Cloud Computing (CLOUD), AK, USA, 2014.

[29] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan, "Sedic: privacy aware data intensive computing on hybrid clouds, "In of the 18th ACM conference on Computer and communications security,CCS'11, pages 515–526, New York, NY, USA. ACM, 2011.

[30] Qingji Zheng, Shouhuai Xu,"Secure and Efficient Proof of Storage with Deduplication, "Proc. of ACM conference on Data and Application Security and Privacy, 2012.

AUTHOR

Prachi D Thakar received B.E (Computer Science & Engineering) from SGB Amravati University in 2010 and pursuing M.E.(Computer Engineering) from SGB University. Right now working as a lecturer in the Dept. of First Year Engineering at Prof. Ram Meghe College of Engineering & Management, Badnera –Amravati.

Dinesh G Harkut received B.E. (Computer Science & Engineering) & M.E. (Computer Science & Engineering) from SGB Amravati University in 1991 and 1998 respectively. He completed his masters in Business Management and obtained his Ph.D. from SGB Amravati University in Business Management in 2013 while serving as a full-time faculty in the Dept. of Computer Science & Engineering at Prof Ram Meghe College of Engineering & Management, Badnera – Amravati. His research interests are Embedded Systems and RTOS