International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 03, March 2017
www.ijates.com
ijates
ISSN 2348 - 7550

ELIMINATING DUPLICATE COPIES OF DATA USING DATA DEDUPLICATION TECHNIQUE TO IMPROVE BANDWIDTH EFFICIENCY

Dr.N.Venkatesan[1], C.Krubakaran.[2], M.Rathan Kumar[3]

[1]Associate Professor, [2]Assistant Professor, Dept. of IT,
Bharathiyar College of Engineering and Technology, Karaikal, Puducherry State
[]Research Scholar, PRIST University, Thanjavur, TN

ABSTRACT

Data deduplication is a technique for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space and upload bandwidth. However, there is only one copy for each file stored in cloud even if such a file is owned by a huge number of users. As a result, deduplication system improves storage utilization while reducing reliability. Furthermore, the challenge of privacy for sensitive data also arises when they are outsourced by users to cloud. Aiming to address the above security challenges, this paper makes the first attempt to formalize the notion of distributed reliable deduplication system. We propose new distributed deduplication systems with higher reliability in which the data chunks are distributed across multiple cloud servers. The security requirements of data confidentiality and tag consistency are also achieved by introducing a deterministic secret sharing scheme in distributed storage systems, instead of using convergent encryption as in previous deduplication systems. Security analysis demonstrates that our deduplication systems are secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement the proposed systems and demonstrate that the incurred overhead is very limited in realistic environments.

Keywords: Cloud storage, Distributed deduplication, Cloud Security, Data reliability

I. INTRODUCTION

Data deduplication is a technique for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space and upload bandwidth. Promising as it is, an arising challenge is to perform secure deduplication in cloud storage. Although convergent encryption has been extensively adopted for secure deduplication, a critical issue of making convergent encryption practical is to efficiently and reliably manage a huge number of convergent keys. One critical challenge of today's cloud storage services is the management of the ever increasing volume of data. To make data management scalable deduplication we are use convergent Encryption for secure deduplication services.

Businesses, especially startups, small and medium businesses (SMBs), are increasingly opting for outsourcing data and Computation to the Cloud. Today's commercial cloud storage services, such as Dropbox, Mozy, and Memopal, have been applying deduplication to user data to save maintenance cost. From a user's point of view,

data outsourcing raises security and privacy concerns. We must trust third party cloud providers to properly enforce confidentiality, integrity checking, and access control mechanisms against any insider and outsider attacks. However, deduplication, while improving storage and bandwidth efficiency, is compatible with Convergent key management. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Many proposals have been made to secure remote data in the Cloud using encryption and standard access controls. It is fair to say all of the standard approaches have been demonstrated to fail from time to time for a variety of reasons, including insider Attacks, mal configured services, faulty implementations, buggy code, and the creative construction of effective and sophisticated attacks not envisioned by the implementers of security procedures.

Building a trustworthy cloud computing environment is not enough, because accidents continue to happen, and when they do, and information gets lost, there is no way to get it back. One needs to prepare for such accidents. The basic idea is that we can limit the damage of stolen data if we decrease the value of that stolen information to the attacker. We can achieve this through a „preventive" disinformation attack. We posit that secure deduplication services can be implemented given two additional security features:

1.1 USER BEHAVIOUR PROFILING

It is expect that access to a user's information in the Cloud will exhibit a normal means of access. User profiling is a well-known technique that can be apply here to model how, when, and how much a user accesses their information in the Cloud. Such normal user's behaviour can be continuously checked to determine whether abnormal access to a user's information is occurring. This method of behaviour base security is commonly used in fraud detection applications. Such pro files would naturally include volumetric information, how many documents are typically read and how often. These simple user specific features can serve to detect abnormal Cloud access based partially upon the scale and scope of data transfer.

1.2 DECOYS

Decoy information, such as decoy documents, honey files, honey pots, and various other bogus information can be generated on demand and serve as a means of detecting unauthorized access to information and to poison the thief's infiltrated information. Serving decoys will confound and confuse an attacker into believing they have bogus useful information, when they have not. Whenever abnormal access to a cloud service is notice, decoy information may be return by the Cloud and deliver in such a way as to appear completely legitimate and normal. The true user, who is the owner of the information, would readily identify when decoy information is being return by the Cloud, and hence could alter the Cloud's responses through a variety of means, such as challenge questions, to inform the Cloud security system that it has inaccurately detect an abnormal access. In the case where the access is correctly identified as an abnormal access, the Cloud security system would deliver unbounded amounts of bogus information to the adversary, thus securing the user's true data from unauthorized disclosure.

The structure of this paper is organised as follows. In section 2 discuss the related works and the various problems in existing system. In section 3 describes the proposed solution with architecture. In section 4 performance analysis is given. In section 5 concluded the paper.

II. RELATED WORKS

With the explosive growth of digital data, deduplication techniques are widely employed to backup data and minimize network and storage overhead by detecting and eliminating redundancy among data. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Though deduplication technique can save the storage space for the cloud storage service providers, it reduces the reliability of the system. Data reliability is actually a very critical issue in a deduplication storage system because there is only one copy for each file stored in the server shared by all the owners. If such a shared file/chunk was lost, a disproportionately large amount of data becomes inaccessible because of the unavailability of all the files that share this file/chunk.

M. O. Rabin [3] Randomly chosen irreducible polynomials p(t) є are used to "fingerprint" bit-string. This method is applied to create a very simple real time string matching algorithm and technique for securing files against unauthorized variations. The method is proved as highly reliable and efficient for each input.

A. Adya , D. Simon ,J. R. Douceur , W. J. Bolosky , and M. Theimer [4], In this paper, authors present a mechanism to reclaim space from incidental duplication for controlled file replication to make it available. Authors mechanism includes 1) Convergent encryption, which encrypt the file by using hash function then hash value is encrypted using the public key of user, 2)Self Arranging, Lossy, Associative Database (SALAD) it is used for aggregating file content and information location in a decentralized, scalable, fault-tolerant. Huge scale reenactment examinations demonstrate that the duplicate-file merging system is scalable, fault-tolerant , and very effective.

M. Bellare, S. Keelveedhi, and T. Ristenpart [5], In this paper an architecture is proposed by authors which gives secure deduplicated storage struggling brute-force spasms, and identify it in a system called "DupLESS". In DupLESS, clients encrypt message-based keys took from a key-server by an unaware PRF protocol. It allows clients to use an available service to store encrypted data and have the service accomplish deduplication on their behalf, and still provides strong confidentiality guarantees. Using the storage service with plaintext data they show that encryption for deduplicated storage can reach performance and space savings close to these techniques.

A. D. Santis and B. Masucci [8]: Here (t; k; n; S) ramp structure is a protocol to distribute a secret 's' chosen in S amongst a set P of 'n' contributors in a particular way such as: 1) sets of contributors of cardinality are equal to or greater than k can restructure the secret 's'; 2) sets of contributors of cardinality are equal to or less than 't' have no information on s, while 3) sets of contributors of cardinality are less than k and greater than t so they might have some information of 's'. In this correspondence author examine numerous ramp schemes, which are protocols to share lots of secrets amongst a set P of contributors, using diverse ramp schemes. Specifically they verify a tight lower bound on the size of the shares held by every participant and on the dealer's randomness in numerous ramp schemes.

Jin li,wenjing lou [11], In this paper author propose by using Dekey, secure deduplication with an efficient and reliable convergent key management scheme. In this paper author introduce disadvantages of a baseline approach. To store convergent keys, Dekey uses Ramp secret sharing scheme.

2.1. PROBLEMS OF EXISTING SYSTEM

Issues in the existing techniques

- Data reliability is low.
- Shared file/chunk was lost, a large amount of data becomes inaccessible
- Previous deduplication systems only considered in a single-server setting
- Encryption over the data makes deduplication impossible.

III. APPLICATION AWARE LOCAL GLOBAL SOURCE DEDUPLICATION

3.1 Proposed System

Application aware Local Global source deduplication scheme that not only exploits application awareness, but also combines local and global duplication detection, to achieve high deduplication efficiency by reducing the deduplication latency to as low as the application-aware local deduplication while saving as much cloud storage cost as the application-aware global deduplication. Our application-aware deduplication design is motivated by the systematic deduplication analysis on personal storage.

3.2 Architecture Diagram

System architecture is the conceptual model that defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviours of the system.

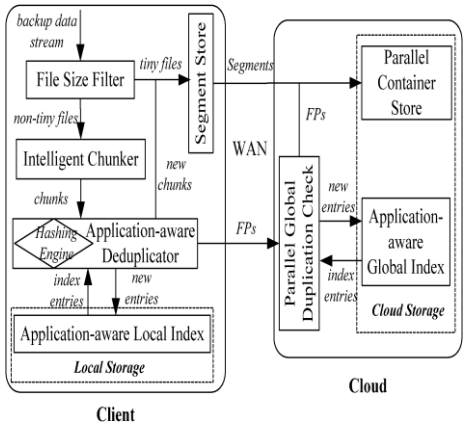

Fig 3.1: System Architecture

IV. PERFOMANCE ANALYSIS

The system is successfully analysed and performance metrics are evaluated. The following are the observation made by us

- Data reliability is achieved.
- Eliminating duplicate copies of data.
- Achieved better fault tolerance.
- Protect data confidentiality
- Storage space and upload bandwidth reduced.
- Reducing the deduplication latency

V. CONCLUSION

We proposed the ALG-Dedupe systems to improve the reliability of data while achieving the confidentiality of the users' outsourced data without an encryption mechanism. ALG-Dedupe are an application aware local-

global source-deduplication scheme for cloud backup in the personal computing environment to improve deduplication efficiency. An intelligent deduplication strategy in ALG-Dedupe is designed to exploit file semantics to minimize computational overhead and maximize deduplication effectiveness using application awareness. In our prototype evaluation, ALG-Dedupe is shown to improve the deduplication efficiency of the state-of-the-art application-oblivious source deduplication approaches by a factor of 1.6X,2.3X with very low system overhead.

REFERENCES

[1] Amazon, "Case Studies," https://aws.amazon.com/solutions/casestudies/# backup.

[2] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digi tal shadows, and biggest growth in the far east," http://www.emc.com/collateral/analyst-reports/idcthe-digital-universe-in-2020.pdf, Dec 2012.

[3] M. O. Rabin, "Fingerprinting by random polynomials," Center for Research in Computing Technology, Harvard University, Tech. Rep. Tech. Report TR-CSE-03-01, 1981.

[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 ICDCS, 2002, pp. 617–624.

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

[6] "Message-locked encryption and secure deduplication," in EUROCRYPT, 2013, pp. 296–312.

[7] G. R. Blakley and C. Meadows, "Security of ramp schemes," in Advances in Cryptology: Proceedings of CRYPTO '84, ser. Lecture Notes in Computer Science, G. R. Blakley and D. Chaum, Eds. Springer-Verlag Berlin/Heidelberg, 1985, vol. 196, pp. 242–268.

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

[9] M. O. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance," Journal of the ACM, vol. 36, no. 2, pp. 335–348, Apr. 1989.

[10] A. Shamir, "How to share a secret," Commun. ACM, vol. 22, no. 11, pp. 612–613, 1979.

[11] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, ''A View of Cloud Computing,'' Commun. ACM, vol. 53, no. 4, pp. 49-58, Apr. 2010.

[12] H. Biggar, ''Experiencing Data De-Duplication: Improving Efficiency and Reducing Capacity Requirements,'' Enterprise Strategy Grp., Milford, MA, USA, White Paper, Feb. 2007.

[13] C. Liu, Y. Lu, C. Shi, G. Lu, D. Du, and D.-S. Wang, ''ADMAD: Application-DrivenMetadata Aware De-Deduplication Archival Storage Systems,'' in Proc. 5th IEEE Int'l Workshop SNAPI I/Os,

[14] A. Katiyar and J. Weissman, ''ViDeDup: An Application-Aware Framework for Video De-Duplication,'' in Proc. 3rd USENIX Workshop Hot-Storage File Syst., 2011, pp. 31-35.

[15] Y. Tan,H. Jiang, D. Feng, L. Tian, Z. Yan, and G. Zhou, ''SAM: A Semantic-Aware Multi-Tiered Source De-Duplication Frame-Work for Cloud Backup,'' in Proc. 39th ICPP, 2010, pp. 614-623.