ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

A Survey on Secure Distributed Deduplication Systems with Improved Reliability

Snehal Bhusal[1], Pratiksha Shende[2], Raju Pal[3], Gaikwad Pooja[4], Prof. Shaikh U. F.[5]

B.E Student, Department of Computer Engineering, PREC, Loni, Ahmednagar, Maharashtra, India[1,2,3,4]

Asst. Professor, Department of Computer Engineering, PREC, Loni, Ahmednagar, Maharashtra, India[5]

ABSTRACT: Deduplication is one of the latest technologies in the current market because it hasability to reduce costs. Data deduplication technique is one of the important data compression technique for eliminating redundant copies. Distributed data duplication system is used in cloud storage to reduce memory space & upload bandwidth only one copy for each file stored in cloud even if that file can be used by number of users. Main purpose of this paper is to makes the first attempt to formalize the idea of distributed reliable deduplication system. In this system data chunks are distributed across multiple servers. In distributed storage systems, instead of convergent encryption as used in previous deduplication systems security requirements of data tag consistency and confidentiality are achieved by using a deterministic secret sharing scheme the Security analysis demonstrate that our scheme is secure in terms of the definitions specified in the proposed security mode.The aim of this paper is to make the first attempt formalize the idea of distributed reliablededuplication system. In our proposed system we are going to develop new distributed deduplication systems whichis highly reliable. In deduplicationprocess data chunk are distributed across multiple cloud servers. instead of using convergent encryption as in previous deduplication systems we usedeterministic secret sharing scheme in distributed storage systems. So that we can achieve the required concepts for security that are dataconfidentiality and tag consistency.

KEYWORDS: Deduplication, reliability, distributed storage system, secretsharing, encryption

I. INTRODUCTION

In our proposed systemwearegoing to use data deduplication process. First of all we must know what data deduplication is, it reduces the amount of data that needs to be physically stored by eliminating extra information and replacing after repetition of it with a pointer to the original. In data deduplication we remove unwanted copy of data and save the memory space.With the help of data deduplication method system reliabilityis improved as well as it avoid wastage of memory space. With the exquisite growth of digital data, deduplication techniques are widely used to backup data and minimize network and storage overhead by detecting and removing unwanted among data. Instead of keeping multiple data copies with the same content, deduplication eliminates unwanted data by keeping only one physical copy and referring other unwanted data to that Deduplication has received much attention from both academic and industry because it can more improves storage utilization and save storage space, especially for the applications with high deduplication ratio such as accession storage systems. For eliminating duplicate copies of data we use data deduplication technique. To reduce storage space and for uploading bandwidth mostly it has been used, In cloud storage. A various deduplication systems has been proposed based on number of policies such as client-sideor server-side deduplications, file-level or block-level deduplications. The first attempt to describe the notion of distributed safe deduplication system. The main Aim of our proposed system is to describe the notion of distributed reliable deduplication system with more security. We implement new distributed deduplication system, which has more reliability. In that data chunks are distributed across multiple cloud servers. Deduplication technique can save the memory space for the cloud storage service providers; it reduces the reliability of the system. Securityanalysis indicate that our deduplication systems are secure in terms of the definitions specified in this security model. As a proof of concept, we implement the proposed systems that indicate the acquired aerial is very limited in actual environments. Deduplication process improves storage utilization & it saves storage space .That’s why it is useful in industry as well as in academic.It is useful in such application which has high deduplication ratio like as archival storage system. Furthermore, for the data privacy challenge is also arises more. The more sensitivedata are redistributed by the users tocloud.Encoding have been usually Utilized, for to provide protection confidentiality before the redistributed data into cloud. Most commercial storage No of service providers are oppose to apply encryption over the data because it is impossible to make deduplication. The reason of that is the traditional encryption mechanism. In which including the public key encryption and symmetric key encryption have requirenumber of users to encrypt their data with own key. For the result of similar data copy of the number of users will lead to the different Data has been encrypted.To solve the problems of confidentiality and deduplication, for solving the problem of deduplication we implement notation of the convergent encryption.

II. LITERATURE SURVEY AND RELATED WORK

In [1], author canLiterature survey is the process of presenting the summary of the journal articles, conferencepapers and study resources. So in this section I have studied the related topics summarized it below.

In 2002 John R Douceur[1] gives mechanism to address the problems of identifying and coalescing identical files in the Farasite.Farasite-gives advantages of high availability by distributing multiple encrypted replicas of each file among a multiple desktop computers. Due to replication consumes significant storage space it is necessary to reclaim used space as possible. This paper gives the solution for control replication. Author presents two mechanisms:

- ConvergentEncryption
- SALAD (Self Arranging Lossy AssociativeDatabase)

A) CONVERGENT ENCRYPTION:

It enables duplicate files to coalesce into space of single files, even if the files are encrypted with different user'skeys. It produces identical cipher text files from identical plaintext files irrespective of encryption keys. Convergent encryption enables identical encrypted filestoberecognized as identical but there remains the problem of performing this identification across large no of machines in decentralized manner. This problem solved by storing location of file & content information in distributed data structure it is nothing but SALAD.

B) SALAD-SELF ARRANGING LOSSY ASSOCIATIVE DATABASE:

It aggregate file content and location information in decentralized, scalable, fault tolerant manner.Collectively these components called as DFC(Duplicate File Coalescing)sub system of Farasite.

In 2008 Mark W. Storer[2] developed a solution that provides both data security and space efficiency in single-server storage and distributed storage systems to solve the problem such that deduplication exploits identical content, while encryption tries to make all content appear random ,the same content encrypted with two different keys results in very different cipher text. Deduplication and encryption are opposed to one another. Deduplication takes benefit of data similarity to achieve a reduction in storage space & the goal of cryptography is to make cipher text indistinguishable from theoretically random data. The goal of a secure deduplication system is to provide data security, against both inside and outside adversaries. Storer developed two models for secure deduplicated storage authenticated and anonymous in both of these authenticated and anonymous model, an inside adversary at the chunk store would not be able to modify data without being detected. Since the chunk'sname is based on the content, a user would not be able to request the modified chunk, or at the very least could tell that the chunk they have requested is different from the chunk that was returned to them.

In 2010 p.anderson [3] presents an algorithm which takes benefits of the data which is common between users to reduce the storage requirements, and increasethe speed of backups. This algorithm supports client-end per-user encryption which is important for confidential personal data, also supports a unique feature that allows immediate detection of common sub trees, avoiding the necessity to query the backup system for every file. This system has shown that a community of laptop users shares a considerable amount of data in between. This gives the potential to significantly decrease backup times and storage requirements. However, they have shown that manual selection of the relevant data backing up only home directories is a poor strategy; this become fails to take backup of important files, at the same time as unnecessarily duplicating other files. This exploits a novel algorithm to reduce the number of files which have to be scanned and therefore decreases backup times.

In 2013 M.Bellare [4] Cloud storage service providers like Drop box, Mozy, and others perform deduplication to save space by only storing one copy of each file uploaded. They propose an architecture that provides secure deduplicated storage to resist brute-force attacks, and realize it in a system called DupLESS. In DupLESS, clients encrypt under message-based keys which obtained from a key-server via an oblivious PRF protocol. It enables clients to store encrypted data with an existing available service, have the service perform deduplication on their behalf, and achieves strong confidentiality guarantees.

In 2014 Jin Li and Yan Kit Li makes [5] the first attempt to address the problem of authorized data deduplication. The system present new deduplication constructions to support authorized duplicate checking. This paper shows that authorized duplicate check method incurs minimal overhead as compared to conversion encryption.

III. PROPOSED SYSTEM

In propose structure to protect private data the secret sharing technique is used which is corresponding to distributed storage systems.In this paper the secret sharing technique is used for protection of private data.In detail a file is divide and encode into sections by using secret sharing technique. These sections will be distributed over many independant storage servers. A cryptanalysis hash value of the content will alsobe calculated and send to storage server asthe mark of the fragment stored at each server. only the data user who first upload the data is required to calculate and distribute such secret shares and following users own same data copy do not need to calculate and stores these shares. Retrievedata copies owner must access a minimum number of storage server by a validation and obtain the secret shares to alterthe data.In different way, theauthorized uses will access the secret shares data copy. Another distinguishable feature of our proposal isthat data completeness inclosestag consistency, can be derived. To explain further if the same value is stored in various cloud storage then deduplication check by methods.It cannot oppose the collision attack established by many servers. To ourknowledge no related work on secure deduplication can rightly address, the reliability and tag consistency problem. The file level and block level deduplication is used for higher reliability.The secret splitting technique is used for protect data. Our proposed structure support both traditional deduplication methods.Privacy,credibility and integrity can be achieved in our proposed system.In solution to kind of secret aggrement attacks are considered.These are the attack on the data and the attack against servers. The data is secure when the opponent control limited number of storage servers.

Block Diagram/Architecture of Proposed System:
When the user wants to upload and download the file from cloud storage at that time first user request to the web serverfor uploading file. It means only approved user can upload the file to web server for that purpose it use the proof of ownership algorithm . User to prove theirrelation of an owner to the thing possessed of data copies to the storage server. When file is uploaded it splits into blocks i.e by default size of block is 4KB.According to file size the block occurs. After that deduplication detection occur.

1.Secrete sharing scheme
2.The File-level Distributed Deduplication system
3.The Block-level Distributed Deduplication system

DOI: 10.15680/IJIRCCE.2016. 0409107

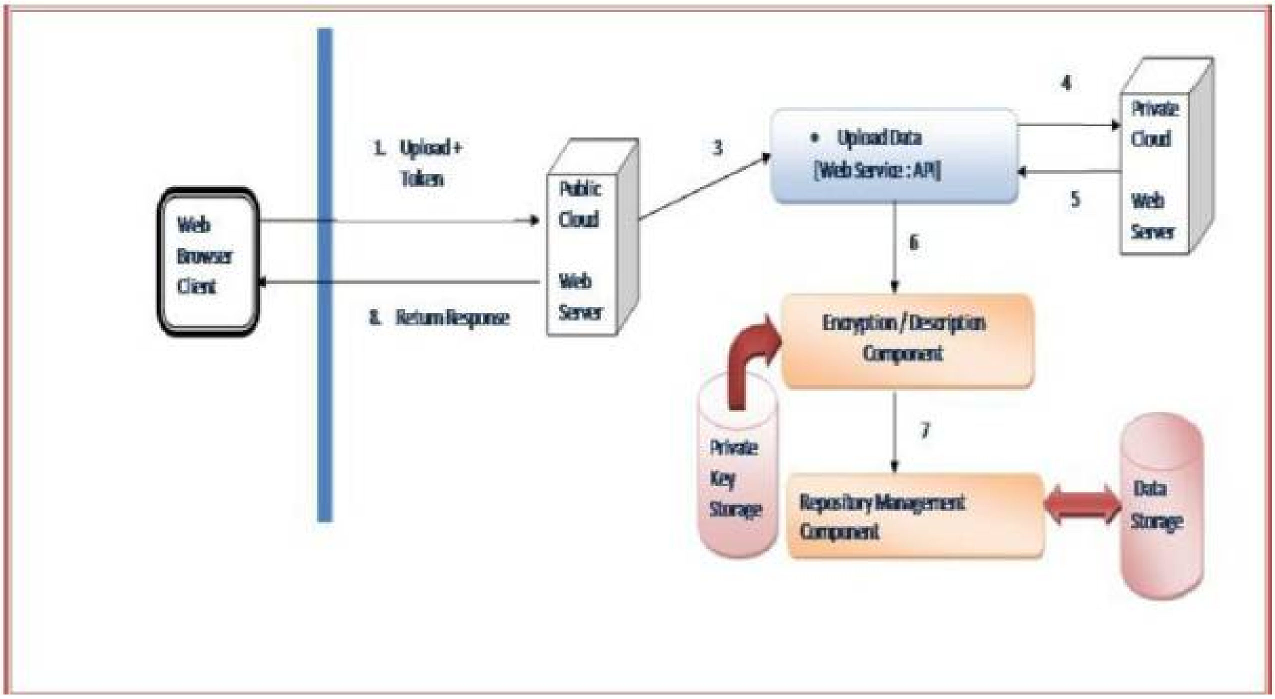

Fig: Workflow of upload/download

1) SECRETE SHARING SCHEME:-

In this module two algorithms are used which are Share and Recover. Share algorithm is used for partitioned and shared secret. With sufficient shares, Extracted and retrieved the secret with the help of Recover algorithm. Share divides secret S into (k-r) fragments of same size, whichproduces r for random fragments of the equal size, and translate into simple language the k fragments using a non-systematic k-of–n erasion code into n shares of the similar size. Out of n shares the Recover adopts k from n shares as inputs. After that outputs the original secret S.A message authentication code (MAC) is a small section of knowledge used to authenticate a message and to provide integrity and authenticity certainty on the message. In our structure, the MAC is applied to derive the bonafides of the external sourced stored files.

2) FILE-LEVEL DISTRIBUTED DEDUPLICATION SYSTEM:-

It support capable duplicate check, tags for each file will be calculated and send to storage cloud service provider. To prevent alignment invasion organized by the S-CSPs,tag collected at different storage servers. System Setup:In our structure, the storage cloud service provider is considered to be n with identities denoted by id1,id2,...,idnrespectively. To upload file F, the client communicate with S-CSPs to perform the elimination of duplicate data .For downloading file F, the client downloads the secret shares of the file from k out of storage servers.

3) BLOCK-LEVEL DEDUPLICATION SYSTEM:-

In this part,we appear how to derive the fine grained block level distributed deduplication. In this system, the client also demands to perform the file level deduplication before uploading file. The user partition this files into blocks, if no duplication is found and performs block-level deduplication system. The system set up is similar to file-level deduplication and also block size parameter will be defined.

IV. CONCLUSION AND FUTURE WORK

Using Secure distributed deduplication technique for the IT Industries provides lot of benefits with the use of both public and private clouds and also provide storage benefits at lower costs. The main idea is that we can limit the damage of stolen data if we success to decrease the value of that stolen information to the attacker. We can achieve this with the help of preventive disinformation attack. We posit that secure deduplication services can be implement given additional security features insider attacker and outsider attacker by using the detection of masquerade activity. Making confusion of the attacker and the additional costs incurred to distinguish real information from bogus information, which plays a significant role in preventing masquerade activity. We posit that the combination of these security features give unprecedented levels of security for the deduplication.

REFERENCES

[1]J. S. Plank and L. Xu, “Optimizing Cauchy Reed-solomon Codes forfault-tolerant network storage applications,” in NCA-06: 5th IEEE International Symposium on Network Computing Applications, Cambridge, MA, July 2006

[2]G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner,Z. Peterson, and D. Song, “Provable data possession atUntrustedstores,” inProceedings of the 14th ACM conferenceon Computer and communications security, ser.CCS’07. NewYork, NY, USA: ACM, 2007,

[3] A. Juels and B. S. Kaliski, Jr., “Pors: proofs of retrievability for large files,” in Proceedings of the 14th ACM conference on Computer and communications security, ser. CCS ’07. New York, NY, USA: ACM, 2007,

[4]M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, “Securedata deduplication,” in Proc. of StorageSS, 2008.Swapnaliet al., International Journal of Advanced Research in Computer Science and Software Engineering 5(10),October-2015, pp. 77-80© 2015, IJARCSSE All Rights Reserved

[5]S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, “Proofsof ownership in remote storage systems.” in ACM Conference onComputer and Communications Security, Y. Chen, G. Danezis, and V. Shmatikov, Eds. ACM, 2011, pp. 491–500.

[6]A. Rahumed, H. C. H. Chen, Y. Tang,P. P. C. Lee, and J. C. S. Lui, “A secure cloud backup system with assured deletion and version control,” in 3rd International Workshop on Security in Cloud Computing, 2011

[7]W. K. Ng, Y. Wen, and H. Zhu, “Private data deduplication protocols in cloud storage.” in Proceedings of the 27th Annual ACM Symposium on Applied Computing, S. Ossowski and P. Lecca, Eds. ACM, 2012, pp. 441–446.

[8]M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Serveraidedencryption for deduplicated storage,” in USENIX SecuritySymposium, 2013

[9]J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, “A securedata deduplication scheme for cloud storage,” in Technical Report,2013.

[10]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.