ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 3, March 2016

"Design of Modern Secure Distributed Deduplication Systems with Improved Reliability"

Miss. Gore Swapnali , Miss. Gore Supriya ,Miss. Tengale Kanchan, Miss. Tengale Varsha,

Prof. S.B.Bandgar
Asst. Prof. in Department of Computer engineering.
S B Patil College Of Engineering, Indapur Dist - Pune.
Savitaribai Phule Pune University.

***Abstract*— For removing replication copies of data we use data deduplication process. As well as it is used in cloud storage to reduce memory space & upload bandwidth. Data deduplication techniques is the most important Data compression techniques & it is used for to eliminate duplicate copy of data. It is used in cloud storage to reduce the storage space on memory ,memory utilization is reduse with the help of this techniques.for the purpose of confidentiality we use this data deduplication techniques ,in which we use the encryption and decryption techniques with help of this our system get secure. only one copy for each file stored in cloud that can be used by number of users.Deduplication process helps to improve storage reliability. One more challenge of privacy for sensitive data also arises when they are outsourced by users to cloud. The most important aim of this paper is to the make in first attempt formalize the idea of distributed reliable deduplication system to world. In our proposed deduplication system we are going to develop new distributed deduplication systems which is highly reliable and secure. In deduplication process data chunks are distributed across multiple cloud servers those are the CS1,CS2,CS3 instead of using convergent encryption as in previous deduplication systems.we use deterministic secret sharing scheme in distributed storage systems for the purpose of eliminate duplicate copy in the system. So that we can achieve the required concepts for security and reliability that are data confidentiality and tag consistency. In the proposed security model, Security analysis demonstrates that our deduplication systems are secure.**

Index Terms*—Deduplication, reliability, distributed storage system, secret sharing, encryption, security, *Block level deduplication

I. INTRODUCTION

In our proposed system we are going to use data deduplication process. First we must know what is data deduplication, 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 methed system reliability is improved as well as it avoide wastage of memory space. Secure means in our system we use encryption & decryption techniques. Encryption means convert plane text into ciphertext this techniques called as a encryption. This ciphertext is transfer to server CS1,CS2,CS3.Again this ciphertext is converted into plane text called as decryption techniques. Distributed means we create server CS2,CS2,CS3 from this server user choose any one, from server selection data was distributed. Deduplication means delete duplicate copy in which use two techniques first one is file level deduplication and second one is block level deduplication. In file level duplication checking duplicate copy file name wise which discover redundancy between file & remove this redundancy. In block level duplication checking duplicate copy block wise which discover redundancy between different block. File is divided into smaller fix size or variable size block in this block contain 10 no. of line. Reliablity means maintain integrity. In our system we use 3 tier architecture which are web browser web server & data base. Our actual system work in between web server & data base. Today's commercial cloud storage service such as Google drive, Mosy have been apply deduplication to save network bandwidth. With the exquisite growth of digital data, deduplication techniques are widely used to backup data and minimize network and storage overhead by detecting and eliminating unnecessary 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 copy.

Deduplication system is mostly used in both industry & academic because it can save storage space on memory and more increces storage utilization on memory, specially for those applications which have high deduplication ratio such as accession storage systems. When use this techniques eliminate duplicate copy. For redusing storage space and uploading bandwidth in large amount it has used,in cloud storage. A different different type of deduplication systems has proposed those are based on number of strategies those strategies such as client-side or server-side deduplications, file-level or block-level deduplications. In the first attempt to describe the safe distributed deduplication system. The main Aim of our system is to describe the distributed reliable deduplication system with more security from this security is provideds.

We are proposed new distributed deduplication system, which has more and more reliability. In that chunks are distributed across multiple cloud servers. Deduplication technique can used for to save the memory space on the memory for the cloud storage service providers; this is reduces the reliability of the system.Security analysis 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 mostly improves storage utilization & it saves storage space .That's why the deduplication system 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.The Most commercial storage to the No of service providers are oppose to apply encryption over the data because it is impossible to make deduplication. The reason of that system is the traditional encryption mechanism.

II. Literature Survey

Data deduplication used for removing duplicate copies ofdata.Thesetechniquesareveryinteresting techniques. The Reliability mean Consistency and validity of test results. It produces consistent results. They only focused on files without encryption, without considering the reliable deduplication over ciphertext. Cipher text is also known as encrypted or encoded information. In 1997 M. Bellare explained the idea of security and scheme for symmetric encryption.

They give different idea of security and analyze the reduction among them. They provide method of encryption using a block cipher, cipher block chaining and counter mode. Its have two goals .First is to study the idea of security for symmetrical encryption and second is to provide concrete security analysis of fixed symmetric encryption device. Convergent encryption provides data security in deduplication. Bellare explains the message locked encryption system and give its application in secure outsourced storage [8]. Encryption is used to achieve the data privacy. Encrypted data is called cipher text. Li et al explained block level deduplication having some key management problems, through several servers [10].

Bellare et al. displayed how to protect private data through the conversion of predictable message into the unpredictable message [8]. In their system, another third party knew the key server. It was introduced to produce the file tag to check the replicate copies. Stanek et al. cultivated better efficiency and security of data storage [9]. They provide different security for all types of data.

III. Proposed System

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 also be calculated and send to storage server as the 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. Retrieve data copies owner must access a minimum number of storage server by a validation and obtain the secret shares to alter the data.In different way, the authorized uses will access the secret shares data copy. Another distinguishable feature of our proposal is that data completeness incloses tag 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 our knowledge 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.

A. 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 server for 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 their relation 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 occurs.

Web client having two services data storage service and security service. Data storage server contains all the uploaded files and Security service provide security to that files. DB profiler store all the metadata of the file

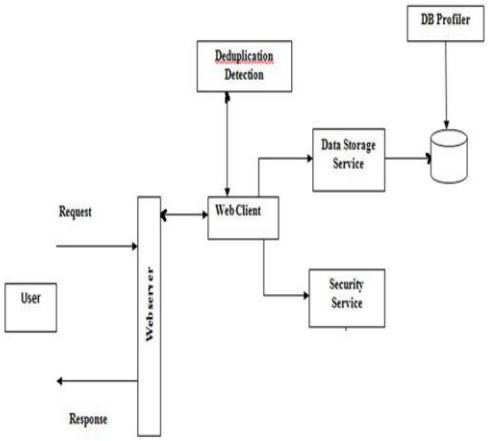

Fig: System Architecture

Workflow For File Upload /Download
Authorized user can access the file from cloud storage.

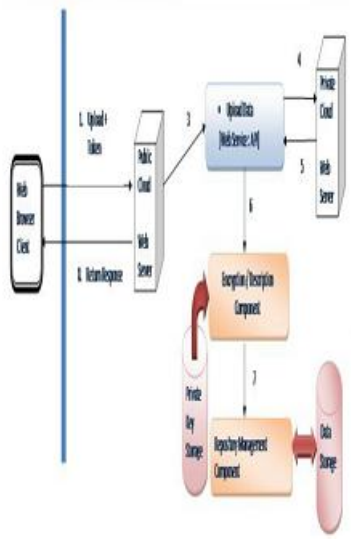

Fig: Workflow of upload/download

A. UML Diagram

a. Data flow Diagram

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.The data flow diagram (DFD) is one of the most important modeling tools is used model the system These

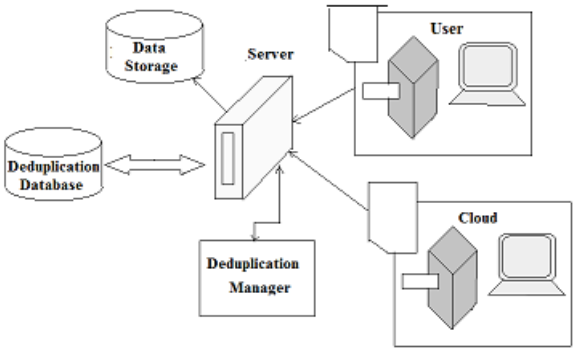

Fig: Data Flow Diagram

components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information and functional detail.

b. Use Case Diagram

A Use case diagram capture use-cases and actors interaction .It describes the fuctional requiredment of the system the manner that outside interact at the system boundary and the response of the system. Provide an overview of all or part of the usage requiredments for a system or organization in the form of an essential model or business model.Communicate the scope of a development project. Model the analysis of usage requiredment in the form of a system use case model. A use case model comparises one or more use case diagram of any supporting documentation such as use case specification and actor definition.

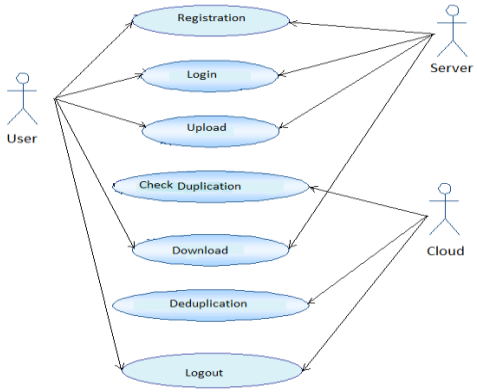

Fig: Use Case Diagram

c. Activity Diagram

An activity diagram is used for modeling dynamic feature of the system an activity diagram consist of flowchart, which shows the flow of control from of flow chart from one activity to another acitivity.In a computational process, sequencial or the concurrent steps are present,to model these sequential or the concurrent process we use activity diagram Also the flow an object from one state to another at different points in any transition can be modeled by using activity diagram. Acitivity diagram done can the entire task specified by UML i.e visualize, specify, construct and document the dynamic aspect of an object.

Notation

1) Action state
2) Activity state
3) Transition
4) Branching

Class diagram show a set of classes interfaces and collaboration and their relationship. Class diagram are important not only for visualizing, specifying and document structural model, but also for constructing executable system through forword and reverse engineering contents.

1) Class or object.
2) Interfaces.
3) Collaborations.
4) The relationship like dependency, generalization and association. The structure of a system by showing the systems classes, there attribute operations and the relationship among the classes.

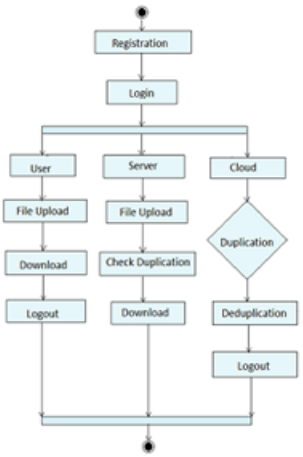

Fig: Activity Diagram

IV. Modules

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

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, which produces 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,…,idn respectively. 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.

V. Conclusion

We implement the secure distributed deduplication systems to improve the reliability of data while achieving the secret of the clients outsourced data. Four constructions were proposed to support file-level and fine-grained block-level data deduplication. The security of tag consistency and integrity were achieved. We implemented our deduplication systems using the Ramp secret sharing scheme and

demonstrated that it incurs small encoding/decoding overhead compared to the network transmission overhead in regular upload/download operations.

VI. ACKNOWLEDGMENT

We have great pleasure in delivering the design paper on the topic Modern Secure Distributed Deduplication Systems with Improved Reliability". We take this opportunity to thank all those who have contributed in successful completion of this paper. We would like to take this opportunity to thank our internal guide Prof. S. B. Bandgar for giving me all the help and guidance we needed. We are really grateful to them for their kind support. Their valuable suggestions were very helpful.

REFERENCES

[1] J. S. Plank and L. Xu, "Optimizing Cauchy Reed-solomon Codes for fault-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 at Untrusted stores," in *Proceedings of the 14th ACM conference on Computer and communications security*, ser. CCS '07 . New York, 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, "Secure data deduplication," in *Proc. of StorageSS*, 2008.

[5] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems." in ACM Conference on Computer 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: Serveraided encryption for deduplicated storage," in *USENIX Security Symposium*, 2013.

[9] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data 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.