Survey on Deduplication in Cloud Environment

Dr. P. Kumar,
Professor,
Department of computer science and Engineering
Rajalakshmi Engineering college
Thandalam, Chennai, India

E. V. Pavithra,
PG Student,
Department of Computer Science and Engineering
Rajalakshmi Engineering College
Thandalam, Chennai, India

***Abstract - A* desktop can store data and run applications. Everything can also host on the Cloud. Cloud is a best platform to store different types of data. Cloud Computing act has data centre and offers pay as per use of service. User had outsourced their data file after encryption. There are various encryption and decryption algorithms available for user's privacy. While the user outsourcing a data file into Cloud, which increases the volume of information in Cloud storage that causes biggest challenge. It also raised the issue of data replication. The duplicate copies occupy more space. Several ideas are found and first it is done with file level. Based on the survey of previous papers the deduplication schemes need to focus on block level deduplication to save space and improve security.**

Key words - Cloud Computing, encryption, decryption and block level deduplication.

I INRODUCTION

Cloud Computing services has huge amount of computational resources on demand by using pay-per-use. It provides computational resources with the help virtualization technology. It has the capability to store data and run applications. It enables us to access all the documents and run applications from anywhere in the world via the Internet. Cloud Computing enables network access to a shared pool of configurable Computing resources. Under Cloud Computing, multiple users has right to use on its own server to retrieve and update their data.

Cloud Computing categorized into three types that are the Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS). In which user has to select right type of service to avoid heavy lifting problem. The SaaS performs licensure of application to customer. The rights are supplied on demand basis. IaaS services involve delivering everything from operating system to servers and storage through IP based connectivity. PaaS has three layers of Cloud Computing. It is similar to SaaS, only the primary difference is being that instead of delivery software, it provides platform for creating software. There are four types of Cloud. The first type is private Cloud, the Cloud infrastructure works only for single organization re-estimates which was managed by third party. It requires organization existing resource decision. The public Cloud is the one in the services are provided to public over internet. One of the best examples of public Cloud is Amazon elastic compute Cloud (EC2). Both the Public and private Clouds are called as hybrid Cloud. Last one is community Cloud in which the data centre is owned by third party.

Cloud services platform facilitates fast and flexible access to data. It does not need any upfront payment in hardware and more time on it. Hence server, storage, database and broad set of applications through internet are easily access by Cloud Computing. One of Cloud services platform is Amazon web Services. It manages network connected hardware needed for application services. It supplies what the user need via web applications.

II RELATED WORK

Wen Xia and Min Fu [1] explained that cross- user redundant data are arising from duplicate files. They encrypted using method of convergent encryption. Its main aim is to backup the Cloud storage, performs deduplication to save space and network bandwidth. The solution is to achieve minimum storage space compared with existing. So user aware of convergent key encryption and multilevel key management has been done under this technique. The experimental result is to provide better performance.

Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee and Wenjing Lou [2] introduced the base line approach where user just keeps the master keys. The proposed scheme is used for efficient and convergent key management. They use different constraints to achieve its target using proposed Dekey. Due to Dekey user need not manage the key by own. The overall result of this experiment is the convergent keys are distributed across multiple servers. Therefore it has been partially succeeded in key management.

MihirBellare, sriram Keelveedhi and Thomas Ristenpart [3] stated about message locked encryption for to resolve the duplication of files. This encryption is used to give increase the strong confidentiality of outsourced file and guarantee. They handled it with storage plain text by knowing its structure and size. In Cloud it is used to give an optimal solution for the proposed work.

Pasquale Puzio, RefikMolva and MelekOnen [4] publish additional encryption operation and access control mechanism. Their goal is to get security and privacy challenges. They propose Cloudedup to handle different constraints. They used to reduce the storage space and save the storage space. The result of this experiment had being partially succeeded.

Junbeom Hur, Dongyoung Koo, Youngjoo Shin, and Kyungtae Kang [5] presented reduction of replicas with different quality constraints by using deduplication scheme to get the good performance and efficiency. The performance has been done with dynamic ownership management. They used to minimize the cost and bandwidth. The experimental results are based on the ownership management.

MihirBellare, Sriramkeelveedhi and Thomas Ristenpart [6] introduced encryption and decryption performed from message. They presented it for the purpose of achieve secure deduplication. They extract paradigm to deliver schemes under different assumptions and for various classes of message sources. But the result does not provide deduplication to expected level. It suffered in brute force attack.

Dimitrios vasilopoulos and Melek Onen [7] they presented proof of retrievability with MLE. Hence the data used is identical. It is performed on the setup phase with uploaded material. It introduces new encoding algorithm ML encode. But it fails because of current POR.

Dipti Bansode and Amar Buchade [8] the study of deduplication technique describes how to secure the data on Cloud. This system has two components front end and back end. It proposed uses application aware index structure. The result of these experiments achieves reliability in deduplication. In future need to focus on data acess and deletion.

Xinyi Huang, Shaohua Tang and Yang Xiang [9] their first attempt to formalize the notion of distributed reliable deduplication system. They proposed new distributed deduplication in which data distributed across multiple Cloud servers. It shows that the incurred overhead is very limited in realistic environments.

Pyla. Naresh, K. Ravindra, Dr. A. Chandra Sekhar [10] they deal with the danger of data stockpiling the data security as well as data integrity and data deduplication on Cloud. They proposed framework of D-Cloud .It create hash estimate before transferring, auditing, integrity of data put into Cloud.

Arthur Rahumed, Henry C.H chen, Yang Tang, Patrick P.C.Lee and John C.Lui [11] Their goal is to take backup for outsourced data with low cost. They used the fade version which eliminates redundancy among the data. Fade version had minimal performance overhead than other traditional Backup Service.

TABLE I. RESEARCH ISSUES

Algorithm	Constraints				
	Storage space	Key usage	Time overhead	Cost	Security
AES	High	High	High	Medium	Low
SHA-1	Medium	High	Low	High	Low
MD5	Medium	Low	Medium	Low	Medium

Vishalakshi N S and S.Sridevi [12] they used convergent key encryption to encrypt data before outsourcing. In which they address the problem of authorized data deduplication and follows method different from other traditional deduplication system. They implements prototype of authorized duplicate check scheme.

Vishalakshi N S and S.Sridevi [13] they proposed Cloudedup its target is to provide secure, efficient storage service and data confidentiality. It introduced additional encryption operation with convergent keys and access control mechanism.

Shweta D. Pochhi, Prof. Pradnya and V. Kasture [14] in this they proposed the data compression technique. To protect the outsourced data it encrypt before data put into Cloud and support authorized duplicate checking. They used LFSR (linear feedback shift register) for to reduce convergent key encryption weakness.

K.Kanimozhi and N.Revathi [15] here they implements secure proof of ownership. In which the keys are derived from content of data itself for convergent key encryption. And it uses hash functioning so the file where it is located is unknown to others.

III RESEARCH ISSUES

According to existing paper, the analysis of various research issues is described in the TABLE I. It is classified as three types.

High indicates the work has been completed in that area. There is an algorithm solving these types of problem.

Medium- it shows which achieved half the successes in that constraints.

Low- it depicts that there is need to explore optimized algorithm for the particular domain focus on different aspects such as storage, key usage and mainly on security.

An advanced encryption algorithm is faster than DES. It is a popular symmetric encryption algorithm. While using AES in deduplication it covers more storage space, key usage and time overhead with low security. SHA is secure hash algorithm. It is considered has stronger encryption and most preferred algorithm used by government. But usage of this algorithm causes high cost in deduplication. MD5 is secured hashing algorithm. The message authentication protocol verifies content of the message.

TABLLE II. COMPARISON OF EXISTING WORK

Authors and year	Method used	Parameter considered	Description	Environment	Tools
Wen xia and Lin Fu in 2015	Convergent Encryption	Security, reliablity	It hashes the data as a key and reliablity in key usage.	Cloud environment	Java
Patrick P.C.Lee and Wenjing Lou in 2014	Dekey	Realistic environment, Key usage	Dekey using ramp secret sharing scheme used to handle maximum limited keys.	Cloud environment	Java
Mihir Belare and Sriram keelveedhi in 2013.	Message locked encryption	Security and storage space	Dupless uses message based keys from key servers through PRF protocol for encryption	Cloud environment	Java
Pasquale Puzio and Retikmolva in 2016	Convergent key	Efficiency	Used to check if a given plaintext has already stored.	Cloud environment	Java
Jun beonHur and Denfoung in 2016	Data reencryption	Data Privacy, confidentiality	Deduplication is effective when user outsource their data in Cloud storage even the owner getting changed.	Cloud environment	Java
MihirBellare, Sriramkeelveedhi and Thomas Ristenpart in the year 2013.	MLE	Privacy	Both encryption and decryption performed itself from it message.	Cloud environment	Java
Dimitrios Vasilopoulous in the year 2016	POR	Guarantee for storage correctness	To reconcile proof of retrievability with file based cross user deduplication.	Cloud environment	Java
Dipit Bansode,Amar Buchade in year 2015	Application aware index structure		Identify deduplication using this structure	Cloud environment	Java
Jin Li and Xiaofeng Chen in the year 2015	Secret sharing scheme	Bandwidth and reliablity	Data Distributed across multiple Cloud server	Cloud environment	Java
Pyla. Naresh, K. Ravindra and Dr. A. Chandra Sekhar in 2016	D-Cloud	Reliablity and security	It encrypt the data before transferring to Cloud	Cloud environment	Java
Arthur Rahumed, Henry C.H chen, Yang Tang, Patrick P.C.Lee and John C.Lui in 2011	Fade version	Cost and security	Layered encryption approach	Cloud environment	Java
Vishalakshi N S and S.Sridevi (2016)	Convergent key encryption	Bandwidth and storage space	Implements authorized duplicate check scheme to identify redundancy.	Cloud environment	Java
Vishalakshi N S and S.Sridevi in the year 2017	Cloudedup		The proposed Cloudedup target is to provide secure and efficient storage services.	Cloud environment	Java
Shweta D. Pochhi, Prof. Pradnya and V. Kasture 2015	Data compression technique and LFSR	Security	It also support authorized duplicate checking.	Cloud environment	Java
K.Kanimozhi and N.Revathi 2016	Secure Proof of ownership and hash function	Confidentiality of data	Encryption is done based on content of the data.	Cloud environment	Java

According to the TABLE II comparison of existing work, it is clear that deduplication have been done with various algorithms in the same Cloud environment. They had tried to achieve success in the following parameter such as space storage, security, reliability, efficiency. But it had struggle in authority of correctness and still had the problem in storage space wastage and security. Therefore, in this proposed work need to focus on above mentioned factors.

IV CONCLUSION

In this paper, the survey on deduplication work with various algorithms tabulated them on the basis of algorithm, objective criteria, environment to which the works being performed. From the literature survey it is clear that, lot of work had been done already in deduplication but still it needs further development. (i.e)Deduplication need to establish with high level security and minimum space wastage.

REFERENCES

[1] Wen Xia, Min Fu, Fungting Huang and Chunguang Li, "A User-Aware Efficient Fine-Grained secure de-duplication scheme with Multi-Level key management," Huazhong University of science and technology in the year 2015.

[2] Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patric P.C Lee and Wenging Lou, "Secure De-duplication with Effiecient and Reliable Convergent Key Management," IEEE Transactions on parallel and distributed systems, vol .25, No.6, on June 2014.

[3] Mihir Belare, sriram Keelveedhi and Thomas Ristenpart, "Dupless: server-Aided Encryption for de-duplicated Storage," appeared at 2013 USENIX Security Symposium.

[4] Pasquale Puzio, Refik Molva, Melek Onen, "Secure De-duplication with Encrypted Data for Cloud Storage," secured project supported by French Government in the year 2013.

[5] Junbeom Hur, Dongyoung koo, Youngjoo shin and Kyungtae Kang, "Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage," IEEE Transaction on knowledge and data engineering/ TKDE-2016.

[6] Mihir Bellare, Sriram keelveedhi and Thomas Ristenpart "Message-Locked Encryption and Secure De-duplication," In EUROCRYPT , LNCS 7881, pp. 296–312, 2013.

[7] Dimitrios vasilopoulos , Melek Onen , kaoutar Elkhiyaoui and Refik molva , " Message- locked proofs of retrievability with secure deduplication," on October 28,2016.

[8] DiptiBansode and Amar Buchade "study on secure data deduplication system with application awareness over Cloud storage system," international journal of advanced computer engineering and networking volume-3, Issue-1 on Jan 2015.

[9] Xinyi Huang, Shaohua Tang and Yang Xiang, "secure distributed deduplication with enhanced reliability in Cloud storage system," IEEE Transaction on computer volume, Augest 2015.

[10] Pyla. Naresh, K. Ravindra and Dr. A. Chandra Sekhar, "The Secure Integrity Verification in Cloud Storage Auditing with Deduplication," on IJCST vol.7, Issue 4, 2016.

[11] Arthur Rahumed, Henry C.H chen, Yang Tang, Patrick P.C.Lee and John C.Lui, " A Secure Cloud Backup System with Assured Deletion and Version Control," At Chinese university of Hong Kong Symposium, 2014.

[12] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee and Wenjing Lou , "A Hybrid Cloud Approach for Secure Authorized Deduplication," IEEE transactions on Parallel and Distributed System, vol 26, No.5,May 2016.

[13] Vishalakshi N S and S.Sridevi, "Survey on Secure De-duplication with Encrypted Data for Cloud Storage," international journal of advanced science and research, Vol. 4, Issue 1, January 2017.

[14] Shweta D. Pochhi, Prof. Pradnya and V. Kasture, "Encrypted Data Storage with Deduplication Approach on Twin Cloud," vol.3 Issue-6 published at pune university in the year of June 2015.

[15] K.Kanimozhi and N. Revathi (2016) "Secure Deduplication on Hybrid Cloud Storage with Key Management," IRJET Volume: 03 Issue: 06/June 2016.