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 4, April 2016

A Survey on Secure Auditing and Deduplicating Data in Cloud

Tejaswini Jaybhaye[1]; D. H. Kulkarni[2]

PG Student, Dept. of Computer Engineering, SKNCOE, Pune, India[1]

Assistant Professor, Dept. of Computer Engineering, SKNCOE, Pune, India[2]

ABSTRACT: Since last decade, cloud computing is one of the biggest innovative technologies; it provides the facility of heavy data maintenance and management by improving data sharing and data storing capabilities. The main threat for this cloud data storage is data security in terms of maintains data integrity and data deduplication on cloud. Handling both issue sane time is the difficult task. SecCloud and SecCloud+ are two new cloud auditing systems which help in maintaining cloud data integrity with efficient data deduplication,

In SecCloud system, user can able to generate data tags before storing data on cloud which helps during performing audit to check integrity of data, other side SecCloud+ system provide encryption of data before uploading it, which enables integrity check and secure deduplication of encrypted data

KEYWORDS: Cloud computing, Data integrity, Auditing, Data deduplication.

I. INTRODUCTION

[A] Cloud Computing

Cloud computing is internet based technology which has advanced computational power and which provided data sharing and data storing facility. Cloud computing is a shared pool of configurable computing resources, on-demand network access and provisioned by the service provider [1].It is cost saving but other hand it has major concern of security. Cloud storage is one of the attractive trend which provide benefits to the customer like cost saving, mobility and scalable service.

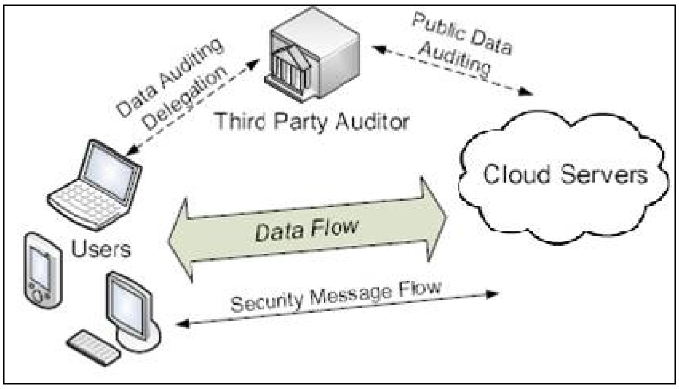

Fig.1 Architecture of Cloud Data Storage Service

A cloud data storage service involves 3 main entities.

i. Administrator Controls – Control over file insertion, file access, file deletion and at the time of user presents in the network and trying to access the cloud data's.
ii. Third Party Auditor (TPA) checks – TPA check on the correctness of cloud data.
iii. Users Access – Availability of the cloud data as per demand services.

Data security is a major concern related to cloud computing. Most of the cloud service provider provides some facility for security and privacy which mainly includes 4 types of data items –

DOI: 10.15680/IJIRCCE.2016. 0404117

i) usage data
ii) sensitive data
iii) Personally identifiable information
iv) Uniques device identities

[B] Data integrity

Integrity is another name consistency. It is a major factor which affects the cloud performance. Data integrity has a protocol for writing of the data in a reliable manner to the persistent data storages which when retrieved is in the same format without any changes. Maintaining integrity of shared data is difficult task. There are number of mechanisms have been proposed [2], [3], [4], [5], [6], & [7] to maintain integrity of data. Integrity is most important of all the security issues in cloud data storages as it ensure completeness of data as well as that the available data is correct, easily accessible to authorized user, consistent and of high quality.

There are three types of integrity constraints:

- Domain integrity
- Referential integrity
- Entity integrity

The correctness of data storage and computation compromised on the cloud due to the less of the control of data owners on data. Secure cloud computing always focuses on the cloud data storage security and cloud computing security. Safety of the data stored on the cloud has been compromised in many cases for monetary profit. To avoid this it is essential to maintain security and privacy of data and cloud computing by using different techniques and mechanisms.

[C] Deduplication

Data deduplication is one of the special techniques which used as a data compression technique, which helps in eliminating duplicate copies of repetitive data, here cloud server stores single copy of data file. This technique is provide benefits like saving network bandwidth but on other hand in hybrid cloud which is combination of public and private cloud deduplication may lead to loss of sensitive information.

Deduplication technique has two categories, which are based data units –

1. File Level Deduplication -

Here file is considered as a one data unit, hash value of file is used as its identifier. During deduplication check if two or more files have similar has value, then they consider that files with same contents and only one copy will be stored.

2. Block Level Deduplication -

Here file is divided into small data blocks, these data blocks are fixed-size or variable size, to check deduplication hash value is computed on each data block.

One more categorization criteria is the location if data are deduplicated at the client location, then duplication done is known as source-based deduplication else target-based. In source-based deduplication, the client first hashes each data segment and then uploads and sends these results to the storage provider to check whether such data are already stored.

II. Literature Review

[A] Mechanisms used for Cloud data integrity check

There are some different techniques used in different auditing mechanisms which provide integrity check on cloud.

1. HLA Based Solution

Chandinee Saraswathy *et al* explained this technique in [2] which uses linear combination for authentication which helps in performing auditing without retrieving data block and to check integrity of the data. HLA is nothing but verification meta data that authenticate. It checks integrity of data block by authenticating data using linear combination of the individual data blocks. This mechanism allows efficient data auditing and consuming only constant bandwidth, but it has disadvantage as it is time consuming because of use of linear combination for authentication.

2. Compact Proofs of Retrievability

Hovav Shacham and Brent Watersy [3] introduced proof-of-retrievability system. In this integrity check system, data storage centre provide proof to a verier that it is actually storing all of a client's data. Here they have explained two homomorphic authenticators the first authenticator is based on PRFs, gives a proof-of-retrievability scheme secure in the standard model. The second, based on BLS signatures [4], which give a proof-of-retrievability scheme with public variability secure in the random oracle model. Frameworks explained can allow arguing about the systems unforgeability, extractability, and retrievability with these parts based on cryptographic, combinatorial, and coding-theoretical techniques respectively.

3 Provable Data Possession: PDP at Untrusted Stores

Giuseppe Ateniese et all proposed a model which based on provable data possession (PDP) [5]. This PDP is used for verifying that server is processing the original data available on cloud without retrieving it, without reading its content. This model generates probabilistic proof of possession by sampling random sets of data blocks from the server. This is cost saving technique it helps to reduces I/O cost.

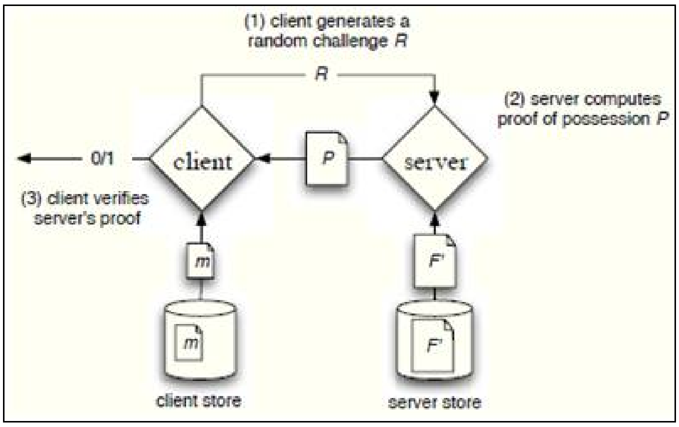
(1) client generates a random challenge R
R
(2) server computes proof of possession P
0/1
client
P
server
(3) client verifies server's proof
m
F'
m
F'
client store
server store

Fig.2 Provable Data Possession at Untrusted Stores

As demonstrated in Fig.2 client maintains a constant amount of metadata which helps in verification of the proof. The challenge/response protocol transmits a small, constant amount of data, which help to minimize network communication. PDP model for remote data checking mainly supports large data sets in widely-distributed storage systems. Key component of this technique is the homomorphic verifiable tags.

4. Robust Data Possession

This mechanism [6] integrates Forward Error Correcting codes (FEC) into Provable Data Possession (PDP).First file is encoded using FEC code and then PDD is applied on encoded file instead of original file. Here it has two benefits as it prevents corruption of a whole file when there is a change in the blocks and another it prevents corruption of small portion of file as changes within file are also detected due to FEC code is used.

5. Privacy Preserving Public Auditing

Cong Wang introduced Privacy Preserving Public Auditing technique [7].Here public auditing allows TPA and user to check the integrity of the outsourced data stored on a cloud & Privacy Preserving allows TPA to perform auditing without requesting data. TPA then able to audit the data by maintaining cloud data privacy. They have used 2 techniques first is the homomorphic linear authenticator and second is random masking to guarantee that the TPA would not learn any knowledge about the data content present on the cloud server during the efficient auditing process, which eliminates the burden of cloud user from the tedious and possibly expensive auditing task and also prevent the users from fear of the outsourced data leakage.

It is based on 4 algorithms:

- Keygen: Key generation algorithm for setup the scheme.
- Singen: Used by the user to generate verification metadata which may consist of digital signature
- GenProof: Used by CS to generate a proof of data storage correctness.
- Verifyproof: Used by TPA to audit the proofs

6. SecCloud: Protocol for security and privacy of data storage and cloud computing

SecCloud is the best protocol proposed by Jingwei Li et all in [8] which ensures security and privacy of data stored on cloud and it's computing. Here it uses cryptography, it first encrypt data and then send it on cloud. This data is then decrypted and stored on a cloud.

SecCloud introduces an auditing facility with maintenance of a MapReduce cloud; it helps clients for generating data tags before uploading and auditing the integrity of data stored on cloud. This technique helps to overcome previous work issues like computational load at user or at auditor which is huge for tag generation. For completeness of fine-grained, the functionality of auditing used in SecCoud is supported on both block level and sector level. In addition, SecCoud also enables secure deduplication.

[B] Techniques used for Data-deduplication

There are some different techniques used in different auditing mechanisms which provide integrity check on cloud

1. Dupless: Serveraided encryption for deduplicated storage.

M. Bellare *et al* [9] designed a system, DupLESS which combines a CE-type scheme which has ability to obtain message-derived keys with help of a key server (KS) which shared with group of clients. Then clients interact with the KS by a protocol for oblivious PRFs, ensured that the KS can use cryptography and mix in secret material to the per-message keys while learning nothing about files stored by clients. This system provide facility of data deduplication and same time it provides strong security against external attacks.

2. Secure Dedup: Secure deduplication with efficient and reliable convergent key management.

J. li *et al* [10] introduced "Dekey", an efficient and reliable convergent key management mechanism for secure de-duplication. Dekey apply de-duplication on convergent keys and distributes convergent key shares across multiple key servers, while preserved semantic security of convergent keys and also maintain confidentiality of outsourced data. They implement Dekey using the Ramp secret sharing mechanism and demonstrate that it incurs small encoding/decoding overhead compared to the network transmission overhead in the regular data uploading and downloading operations

3. Revdedup: A reverse deduplication storage system optimized for reads to latest backups.

C. Ng *et al* [11] presented RevDedup, a de-duplication system specially designed for VM disk image backup which is present in virtualization environments. RevDedup system has several design goals like low memory usage, high storage efficiency, and high backup and restores performance for latest backups. Reverse de-duplication is the core design component of RevDedup, which removes duplicates of old backups and mitigates fragmentation of latest backups. They extensively evaluate RevDedup prototype using different workloads and validate our design goals.

4. Secure data deduplication.

M. W. Storer [12] designed two models for secure de-duplicated storage: authenticated and anonymous. These two designs model demonstrate that cloud security can be combined with data de-duplication such a way that it help to provide a diverse range of security characteristics. In their model they have used convergent encryption for providing security. This technique, first used in the context of the Farsite system, which provides a deterministic way of generating an encryption key, by that two different users can encrypt data to the same cipher text. In authenticated and anonymous models, for each file a map is created that describes how to reconstruct a file from chunks. This file is encrypted using a unique key. In the authenticated model, sharing of this key is managed by using asymmetric key pairs. Other hand in the anonymous model, storage is immutable, and file sharing is conducted by sharing the map key offline and creating a map reference for each authorized user.

5. DWFD: Enhanced Dynamic whole file De-duplication(DWFD) for space optimization in private cloud storage backup

In paper [13] authors have presented how to optimize the private cloud storage backup which provides high throughput to the users of the organization by increasing the de-duplication efficiency? Disadvantages of this technique - It is not sufficient used to development of chunk level deduplication and block level reduplication. So it is highly effective for improving the private cloud backup storage efficiency by reducing the de-duplication time

6. Se-duplication: A secure data deduplication scheme for cloud storage.

J. Stanek *et al* [14] had proposed a new encryption scheme that guarantees semantic security for unpopular data which provides weaker security but better storage and also provides bandwidth benefits for popular data. Due to this encryption scheme, data se-duplication can be effective for popular data, while semantically secure encryption protects unpopular content, preventing its de-duplication. Transitions from mode to other mode take place seamlessly at the storage server side if a file becomes popular.

7. SecCloud+ :

SecCloud+ protocol proposed by Jingwei Li *et al* in [7] supports integrity auditing and secure deduplication both which also provides the guarantee of file confidentiality. It involves an additional trusted entity, namely key server, which assigns client the secret key (according to the file content) for encrypting files. It has 3 steps similar to SecCloud the file uploading, the integrity auditing and the proof of ownership protocol. The only difference SecCloud+ is the file uploading protocol. SecCloud+ involves an additional phase for communication between cloud client and key server which the client needs to communicate with the key server to get the convergent key for encrypting the uploading file before the phase 2 in SecCloud.

III. CONCLUSIONS

Cloud computing is world's biggest innovation which has advanced computational power and improved data sharing and data storing capabilities. It increases the ease of usage by giving access through any kind of internet connection. As every coin has two sides it also has some drawbacks. Data privacy and data security are the main issues for cloud storage. To ensure that the risks of privacy have been mitigated a variety of techniques that may be used in order to achieve privacy. This paper showcase some privacy techniques which introduced to maintain integrity of data and different methods for overcoming the issues data deduplication on untrusted data stores in cloud computing.There are still some approaches which are not covered in this paper. This paper categories the different methodologies in the literature as encryption based methods, access control based techniques, query integrity, keyword search schemes, and auditability schemes. Even though there are many techniques in the literature for considering the concerns in data integrity and data deduplication, no approach is highly developed to overcome both issue at a time. Thus to handle all these privacy concerns, we need to develop privacy–preserving framework which handle all the worries related to cloud data storage and strengthen cloud storage services.

REFERENCES

[1] P. Mell and T. Grance, "Draft NIST working definition of cloud computing".

[2] Chandinee Saraswathy K. , Keerthi D. , Padma G. "HLA Based Third Party Auditing For Secure Cloud Storage" International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1526-1532

[3] H. Shacham and B. Waters, "Compact Proofs of Retrievability,"in the Proceedings of ASIACRYPT 2008. Springer-Verlag, 2008, pp.90–107.

[4] K. Kiran Kumar, K. Padmaja, P. Radha Krishna, "Automatic Protocol Blocker for Privacy-Preserving Public Auditing in Cloud Computing", International Journal of Computer science and Technology, vol. 3 pp, ISSN. 0976-8491(Online), pp. 936-940, ISSN: 2229-4333 (Print), March 2012

[5] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson,and D. Song, “Provable Data Possession at Untrusted Stores,”in the Proceedings of ACM CCS 2007, pp. 598–610.

[6] Reza Curtmol, Osama Khan, Randal Burns " Robust Remote data Checking"

[7] C. Wang, Q. Wang, K. Ren, and W. Lou, “Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing,”in the Proceedings of IEEE INFOCOM 2010, 2010, pp. 525–533.

[8] Jingwei Li, Jin Li, Dongqing Xie and Zhang Cai "Secure Auditing and Deduplicating Data in Cloud" IEEE TRANSACTIONS ON COMPUTERS VOL: PP NO: 99 YEAR 2015

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

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

[12] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. In Proc. of StorageSS, 2008

[13] M. Shyamala Devi, V.Vimal Khanna, Naveen Balaji”Enhanced Dynamic Whole File De-Duplication (DWFD) for Space Optimization in Private Cloud Storage Backup”, IACSIT, August, 2014.

[14] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. In Technical Report, 2013.