International Journal of Engineering & Technology, 7 (3.6) (2018) 55-58

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Survey on Data Integrity, Recovery, and Proof of Retrievability Techniques in Cloud Storage

Neha Narayan Kulkarni[1*], Shital Kumar A. Jain[2]

[1]MIT Academy Of Engineering, Alandi, Pune.
[2]MIT Academy Of Engineering, Alandi, Pune. E-Mail:Sajain@Comp.Maepune.Ac.In
**Corresponding Author E-Mail:Nehakulkarni281@Gmail.Com*

Abstract

Recently the technologies are growing fast, so they have become the point of source and also the sink for data. Data is generated in large volume introducing the concept of structured and unstructured data evolving "Big Data" which needs large memory for storage. There are two possible solutions either increase the local storage or use the Cloud Storage. Cloud makes data available to the user anytime, anywhere, anything. Cloud allows the user to store their data virtually without investing much. However, this data is on cloud raising a concern of data security and recovery. This attack is made by the untrusted or unauthorized user remotely. The attacker may modify, delete or replace the data. Therefore, different models are proposed for a data integrity check and proof of retrievability. This paper includes the literature review related to various techniques for data integrity, data recovery and proof of retrievability.

Index Terms: *Data integrity, data recovery, proof of retrievability, third party auditor.*

1. Introduction

Cloud computing

The subsequent step in the Internet's progression is of the cloud which grant the computing capability to computing base, custom methods to peculiar collaboration produced assistance anywhere and whenever needed. The Cloud can be explained as the collection of device, interfaces, services, and storage that link to provide the perspective of various services. The data is managed properly using the cloud storage [1]. It allows managing the information virtually. Users do not own to acquire the expensive hardware and create strategies to enhance the data.

Cloud computing is a cloud-based internet computing which provides internet access on demand. The resources get hold up by the users. Cloud computing is a facility rendered via provider which allows the user to save files on the server to decrease the investment cost and its support cost [2]. As the data is increasing rapidly, it is required to increase the storage size. The cloud storage is more favored by the user because of rising storage issues. Various universal and vast services are provided by the cloud to the user. Cloud storage gives accessible services for the information stored online. This data creates the virtualized groups that are presented by the TPA. As the data is stored virtually the security risk and data access risk increases, which needs to be adequately controlled to reduce the data loss. Cloud computing service works with following service paradigms 1) SaaS (Software as a Service)- Cloud model that provides access to the software through the web whereas the user does not has control over the infrastructure. Example-Google Docs, Microsoft Office, DropBox. 2) PaaS (Platform as a Service)-The Cloud Service Provider (CSP) provides the necessary hardware and platform for the developers. Example-Flexiscale, Gizmox. 3) IaaS (Infrastructure as a Service) –This model allow the user to access virtual and physical environment. The software, application, operating system is manageable to the user. Example-Amazon Web Service, Cisco cloud verse. The deployment models for cloud computing are 1) Public cloud-Open access to network publicly. 2) Private cloud-Open to the authorized end users. 3) Community cloud-This cloud is accessible to a category of the users who satisfies the policy for authorization. 4) Hybrid cloud-It integrates the features of public, private, community cloud. Cloud computing provides the pay-as-per-use service. The aspect of the cloud is shared so, the user loses its control on the data when it gets stored virtually. The user is not permissible to the substantial access [3]. Also, the users are unaware to manage the unknown attacks. The significant concern related to the data on the cloud are security and privacy as the data is out of control and managed remotely at the same instance. The integrity of remotely stored data needs verification. Here, various techniques for security concerns are proposed.

Pros and cons of cloud computing

The data is being generated at the large amount which has subsequently increased the platform for managing data. Various organizations are trying for a quick, frequent, reliable and systematic provision for the user data. Maintaining company's data at their local server increase the infrastructure cost, employment cost, and maintenance cost. Now cloud has offered the minimum-cost solution for the IT progressing organizations [4]. Cloud made access to the data simple, accessible, adaptable, inexpensive and flexible.

1. Cost-Cloud storage helps to reduce maintenance charge, security cost, operational price, and licensing cost.
2. Time-Cloud makes it possible to access data independently.
3. Compatibility- Cloud services are compatible with different operating systems.
4. Accessibility-Data is available to all users, and it provides flexibility to work simultaneously with same data.
5. Storage-Cloud offers the unlimited storage and networking.

Drawbacks of the cloud computing are as follow

1. Unauthorized access- As the data is available to a user at any place the unauthorized user may also access data to corrupt data available.
2. Downtime- If the downtime for cloud service increase it incurred the more cost for the users.
3. Privacy-Though the service provider provides the security layer it is important to verify the security of data.
4. Data Location-Physical location is unknown which reduces the transparency in data access.
5. Integrity-Integrity is the primary concern to be taken care of to verify the outsourced file.

Data integrity

The essential parameters considered for checking integrity of the data are validity, consistency and accuracy. The integrity of the stored data can get affected because of malicious attacks or unconsciousness errors of the individual. Going down to the information, utilizing security systems, applying mistake discovery, and amendment programming insufficient for information honesty. Now, accuracy and accessibility of the data turn into a noteworthy inquiry for clients. With a specific end goal to evaluate the test of information trustworthiness monitoring and inspecting, many models and frameworks are launched. CSP must convince the customers about their information which is preserved uninfluenced and remained careful of debasement, alteration or unapproved exposure using some of the existing strategies.

Data recovery

In prior days electronic information is expanded in enormous sum. This data requires extensive space for the accommodation means to save information. Consequently, the span of Hard Disk broadened to Terabyte[6]. The clients want the cloud to keep the substantial measure of intelligence because of capacity measure issue. The problem is emerged because of information security if there can be an occurrence of cloud harmed or on the other hand that can be degenerate. In this circumstance valuable information may be lost, to stay away from this circumstance there need few systems to provide reinforcement of cached data and recover such information if the condition occurs. There are diverse methods which are known as the data recovery methods. These methods are having numerous dependability and security issues. Moreover, also they are not advantageous and stable. To defeat this downside from the information reinforcement and restoration concerns, it needs a compelling and robust framework.

Proof of retrievability

Proof of Retrievability is a test reaction verification given to a customer by the prover that the put away record in the cloud is safe and necessary that the client can ultimately recover it. The favorable primary position of Proof of Retrievability over different methods is productivity. The reply can be smaller and utilize little part of file F; the verifier can get the accuracy of the document F. If the file is unrecoverableor not detected it is not useful. Subsequently, Proof of Retrievability is deployed, for the file designated over a few servers [5]. The record F is put away in various servers in a repetitive frame. There exists multiple evidence of retrievability (POR) assuring storage accuracy with cryptographic means, which come inconsistently with the deduplication innovation. Among competitive cloud suppliers, Amazon S3 and Google drive these days offer unlimited storage nearly to the user. As the number of clients and the volume of generated information keeps on expanding, the clients outsource their information. The clients loan the full control of their data to cloud suppliers and have no way to confirm the integrity of their data; then again, cloud specialist co-ops are offering an exponential storage utilization which turns out hard to control.

2. A Historical Perspective for Cloud Computing

Around fifties, the processing unit was shared by the multiple user which is considered as a leading thought for cloud computing. J.C.R. Licklider imagined a thought that every client will manage the data independently of the place. The book "The Challenge of the Computer Utility" is the first book which has stated the properties of the cloud. The virtual machines were used by the clients in 1970. The concept of the cloud emerged with the further progress [4]. In 2002, Amazon Web Services were in use. In 2009, Google and Web2.0 existed as an important component of this progress. Presently, Google, Microsoft, Amazon are leading providers of cloud service to the user.

3. Background Concept

a. Data owner: The master of the data is the owner who can be the client or the organization. Data maintenance is the primary service on which the owner depends for support.
b. Cloud Server: The owner receives the storage service from the server. The cloud server is capable of managing data.
c. Third party Auditor (TPA): The skillful object possessing means and power to manage the service from the server and request from the client is termed as TPA. The client request TPA to verify the security.
d. Bilinear Pairing: Let A1, A2 denote cyclic groups of order 'm,' generated by S, where A1 is an additive A2 is a multiplicative group A pairing defined as a map e: A1 × A1 → A2 satisfying the properties below [5]:
    1. Bilinearity: for all a, b belongs Z, for all S belongs A1, for all T belongs A1: e (Sa, Tb) = e (S, T)ab
    2. Non-degeneracy: e (S, T) not equal one where P, Q belongs A1
e. Homomorphism: Let A1, A2 denote groups working with (.) and (*) operators respectively [5]. If f (a.b)→f(a)*f(b) then it is called homomorphism.

4. Methods for Checking Data Integrity in the Cloud

Rivest-shamir-adleman (RSA)

The author of the paper [6] has discussed the RSA based hash and RSA based tag. The prime numbers are used for operating on the data. It uses KeyGen algorithm, SigGen algorithm, GenProof algorithm, Verify Proof algorithm. RSA algorithm works by partitioning the data into the blocks. These blocks are used as a signed metadata. The proof is generated by the server and evaluated by the TPA.

Message digest 5(MD5)

The author has also discussed the MD5 algorithm. The file is compacted. The MD5 function produces the message digest when the compacted file is provided as input to the function [6]. The produced output is coded and added to the original file. This appended file is used by MD5 to produce a hash of the file. The comparison of the available record and the generated hash is undertaken. When the hash value match, then the file is correct otherwise file tampers.

Message authentication code(MAC)

The encryption algorithm MAC gives message authentication to check integrity. MAC accepts various size blocks and a cryptographic key which helps in generating the authenticated code [6]. The owner possesses the secret key and message which is used to verify the integrity.

Encryption algorithm

An encryption algorithm is designed with the components such as a trusted party, the client and the service provider. The authorized client outsources the data to the cloud. The third party divides the file into the partitions, accepts the request and verifies the integrity [6]. Tags get generated for each segment and compute the hash. When the hash value matches then the file is verified else tampered.

Identity-based remote data integrity checking (ID-RDIC)

This scheme approaches existing complex key management issue which is dependent on public key infrastructure for checking data integrity. The author has proposed the new system using the key homomorphic cryptographic method. The author has contributed to formalize zero-knowledge privacy against TPA. Many constraints and algorithm are used to build the system such as Setup, Extract, TagGen, Challenge, ProofGen, ProofCheck. Param is the standard parameter setting of "pbc" library. Setup and Extract algorithm works very fast, and TagGen is an expensive algorithm. [7] Provides detail study of the technique.

Identity-based cloud data integrity checking(ID-CDIC)

ID-CDIC addresses the key management issue and complex certificate management. The author proposed the protocol based on RSA signature and supported variable size file block and public auditing. This protocol consists of six algorithms such as private key generator generate RSA module using Setup. Extract compiles secret key using user identity and the master key. TagGen stores file to cloud with various parameters signature, signing key, public key, blocks, and tags. Challenge sends a challenge to cloud from TPA, then ProofGen generates proof and sends it to the TPA. ProofCheck executes and checks whether the proof matches or not. This system is verified using variable block size. [8] Provides descriptive for ID-CDIC.

Identity-based distributed provable data possession (ID-DPDP)

ID-DPDP verifies the data integrity externally for the whole data. This protocol eliminated the complex certificate management and supported multi-cloud scenario. The implementation results showed that it improves the soundness of PDP as it adds security to the PDP. It uses BLS signature scheme for security. It is efficient. It involves Setup, Queries (Extract and Store), Prove and Output. Setup and Extract computational cost is negligible. [9] Provides detail study regarding ID-DPDP protocol.

Identity-based proxy oriented data uploading and remote data integrity checking in public cloud (ID-PUIC)

ID-PUIC verifies whether the data is uploaded and intact correctly. It works with the hard code of Diffie Hellman and bilinear pairing. This protocol addresses the private virtual integrity checking, delegate checking, and public integrity verification. It uses public key cryptology. It comprises Setup, Extract, Proxy KeyGen, TagGen, Proof.[10] provides detail study of the ID-PUIC scheme.

Proof of storage (POS)

POS is the important cryptographic method for outsourced data. Existing POS work with single user environment. The author has introduced new features deduplication for the cross-user using Homomorphic Authentication Tree it solves private key generation problem. It supports unlimited verification and updates operation. It includes building block phase, preprocess phase, upload, deduplication, update, POS. [11] describes complete procedure for DPOS.

Remote data possession

This method verifies the privacy of the authenticator for the cloud. It supports block-less verifications. Used algorithms are KeyGen, AuthGen, AuthVer, Challenge, ProofGen, ProofVerf, ExtAuth. [12] Provide detail study of the technique.

5. Methods for Data Recovery in the Cloud

The paper [13], describes different algorithms developed for recovery of the data such as Cold back-up, Hot back-up, PCS, HSDRT, Linux Box, ERGOT. Comparing PCS with the other methods we can observe that PCS is manageable, facile and reliable for recovery. PCS is executed depending on congruence service. PCS provide high likelihood of recovering stored data. The system creates an implicit disk and generate equivalence group which stores the parity data. This method is efficient, but it is hard to control the complexity during its implementation. HSDRT is useful for mobile client. This method needs excessive execution cost. This scheme is not resistant to duplication of the data. HSDRT is unsuccessful because of its high cost and duplication issue. This method works with the distributed data which uses encryption mechanism, providing a high speed and efficient file recovery process.

Another method is the Efficient Routing Grounded on Taxonomy (ERGOT). The execution of ERGOT is based on grammatical evaluation. This method is deprecated because of its required time and difficulty in its execution. ERGOT helps in locating service because of its semantic characteristic. ERGOT execute on the basis of service illustration and invocation for data retrieval. Linux Box model provides the simple backup method and low-cost implementation. This model lacks the security provided for the data recovery. It provides the efficiency for migration of the client from one cloud to other. When the storage request to sync and backup it works for full storage rather than the single file this is the drawback of this method. Because of this nature the frequency of the server is used extravagantly. As the recovery time increases with the increase in the size of data so Linux Box need more time to backup as this algorithm works for whole data.

6. Methods for Proof of Retrievability

Erasure coded authenticated logs (ECAL)

This method is a dynamic POR which stores encoded log, and these logs are garbled, to secure them from the server. Generic

ECAL uses buffer consisting of Initialization, Setup, and Challenge. The proposed ECAL uses equi-buffer to store log and LInit, LAppend, LAudit where retrievability involves PInit, PUpdate, PRead, PAudit. [14] Provides detail study of ECAL.

Message locked proof of retrievability

It addresses the problem of data retrievability and the cross-user deduplication. The author has proposed some modification in the existing POR which lack the deduplication. The author uses the message lock key technique to provide security. [15] Provides detail concept overview of message lock POR.

Static

Basic-This scheme encrypts only small bits of data which reduces overhead on the client side. Storage overhead on the client is also minimized [16]. This scheme reduces the size of proof and network bandwidth utilization. It limits the capacity and computational cost on both server and customer side. This limit is quite large and might be sufficient if the amount of information storage is small. It is a challenge to broaden the number of queries with this scheme.
POR on large file-In this scheme the prover stores a single cryptographic key. The verifier stores key irrespective of the dimension and attributes of the files. This method needs that the verifier accesses a small part of a file F within the limit of a POR. The part of file modified by the verifier is independent of the length of the file and includes fewer numbers of blocks [16]. This retrievability scheme encrypts the file and embeds a group of random value block.
HAIL-High Availability and Integrity Layer (HAIL) checks correctness and provide high availability of storage. The storage blocks are tested and reallocated if the system is stopped here the algorithm uses POR as a building block. It avoids the server redundancy. The verification of the file is dependent on the verifier who interacts with the server.

Dynamic

Data Correctness-The Data Correctness scheme involves encryption of small amount of data instead of working on the whole file. Data correctness reduces the execution load on the client side as well as the bandwidth use [16]. Hence this type is more proper for limited resource means. The third-party auditor needs to keep one key and two functions generating random sequences, no concern about dimensions of the file. The third-party auditor adds metadata to the file and uploads it to the cloud. Hence during verification TPA uses this metadata to verify the correctness of the stored data.
Public Auditability-There is two types of auditing methods, private and public. Although private does high potential auditing, it is overhead on the client side. In public verification, the auditor has a key and efficiently audits the outsourced data to test the integrity of cloud storage data outwardly and not using personal information related to the clients [16]. This method helps for stateless verification considering no status information. Thus public audit permits the clients to handover the verification to third-party auditors reducing the burden.
DPOR-DPOR has three stages as follow: Pre-process: The client preprocesses the data and generates metadata of the data, before uploading the file to the server [16]. Then the client will upload the file to the cloud and keeps the metadata. Verification stage: The client periodically checks the data for correctness. The client requests the server for the proof. It verifies with the metadata to check security and integrity. Update stage: The server updates the file according to the request made by the user. After every update, the server will send proof to the client to verify that the file is uploaded correctly.

7. Conclusion

In this survey paper, we discussed various methods for checking integrity of the data, proof of retrievability, and data recovery for a cloud environment. Many research papers are discussed related to the data integrity, retrievability, and data recovery. From the above survey, we can summarize that cloud services are required to be dynamic, adjusted, efficient, relevant and retrievable. Cloud computing services are widely used, so especially data security is an open issue for the researchers.

References

[1] Razaque A & Rizvi SS, "Privacy-preserving model: a new scheme for auditing cloud stakeholders", *Journal of Cloud Computing: Advances, Systems and Applications*, (2017).
[2] Kiraz MS, "A comprehensive meta-analysis of cryptographic security mechanisms for cloud computing", *Journal of Ambient Intelligence and Humanized Computing*, Vol.7, No.5,(2016), pp.731-760.
[3] Garg N & Bawa S, "Comparative analysis of cloud data integrity auditing protocols", *Journal of Network and Computer Applications, ELSEVIER*, Vol.66, (2016), pp.17-32.
[4] Seviş KN & Şeker E, "Survey on Data Integrity in Cloud", *IEEE 3rd International Conference on Cyber Security and Cloud Computing*, (2016).
[5] Worku SG, Ting Z & Zhi-Guang Q, "Survey on Cloud Data Integrity Proof Techniques", *Seventh Asia Joint Conference on Information Security Information Security (Asia JCIS)*, (2012).
[6] Desai CV & Jethava GB, "Survey on Data Integrity Checking Techniques in Cloud Data Storage", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol.4, No.12, (2014).
[7] Yu Y, Au MH, Ateniese G, Huang X, Susilo W, Dai Y & Min G, "Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage", *IEEE Transactions on Information Forensics and Security*, Vol.12, No.4, (2017), pp.767-778.
[8] Yu Y, Xue L, Au MH, Susilo W, Ni J, Zhang Y, Vasilakos AV & Shen J, "Cloud data integrity checking with an identity-based auditing mechanism from RSA", *Future Generation Computer Systems*, Vol.62, (2016), pp.85-91.
[9] Liu H, Mu Y, Zhao J, Xu C, Wang H, Chen L & Yu Y, "Identity-based provable data possession revisited: Security analysis and generic construction", *Computer Standards & Interfaces*, Vol.54, (2017), pp.10-19.
[10] Wang H, He D & Tang S, "Identity-Based Proxy-Oriented Data Uploading and Remote Data Integrity Checking in Public Cloud" *IEEE Transactions on Information Forensics and Security*, Vol.11, (2016), pp.1165–1176.
[11] He K, Chen J, Du R, Wu Q, Xue G & Zhang X, "DeyPoS: Deduplicatable dynamic proof of storage for multi-user environments", *IEEE Transactions on Computers*, Vol.65, No.12, (2016), pp.3631-3645.
[12] Shen W, Yang G, Yu J, Zhang H, Kong F & Hao R, "Remote data possession checking with privacy-preserving authenticators for cloud storage", *Future Generation Computer Systems, ELSEVIER*, Vol. 76, (2017), pp.136-145.
[13] Pophale K, Patil P, Shelake R & Sapkal S, "Seed Block Algorithm: Remote Smart Data-Backup Technique for Cloud Computing", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol.4, No.3, (2015).
[14] Etemad M & Küpçü A, "Generic Efficient Dynamic Proofs of Retrievability", *Communication of ACM*, (2016), pp.85-96.
[15] Vasilopoulos D, Önen M, Elkhiyaoui K & Molva R, "Message-Locked Proofs of Retrievability with Secure Deduplication", *Communication of ACM*, (2016), pp.73-83.
[16] Hegde RA & Prakash M, "A Survey on Proof of Retrievability and its Techniques", *International Journal of Engineering and Techniques*, Vol.2, No.2, (2016).