International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 2 Issue 2 ǁ February. 2013 ǁ PP.49-55

Improving data integrity on cloud storage services

Miss. M.Sowparnika[1], Prof. R. Dheenadayalu[2]

[1](*Department of Information Technology, Saveetha Engineering College, India)*
[2](*Department of Information Technology, Saveetha Engineering College, India)*

ABSTRACT*: Cloud computing is an internet based computing which enables obtaining resources (hardware, software, platform and services) from the internet on a scalable basis. . Many users place their data in the cloud, so correctness of data integrity and security are prime concerns. This work studies the problem of ensuring the integrity and security of data storage in Cloud Computing. To ensure the correctness of data, we consider the task of allowing a Third Party Auditor (TPA), on behalf of the cloud client to verify the integrity of the data stored in the cloud. The technique of bilinear aggregate signature is used to achieve batch auditing. Batch auditing reduces the computation overhead. Unlike most prior works, the new scheme further supports secure and efficient dynamic operations on data blocks, including: data update, delete and append. For clustering high dimensional data, hyper graph model is used, where frequent item sets found from association rule algorithm are used as hyper edges. To achieve accuracy and promptness, user preference and server labeling is also included. To achieve accuracy and promptness, user preference and server labeling is also included.*

Keyword: *Batch auditing, Bilinear aggregate signature, Data integrity, Hyper graph, Third-Party Auditor (TPA)*

I. INTRODUCTION

Cloud Computing is the use of Internet for the tasks performed on the local machine, with the hardware and software demands maintained elsewhere. It represents a different way to architect and remotely manage computing resources. Cloud is widely used everywhere owing to its convenience, be it in simple data analytic program or composite web and mobile applications. Local computers no longer have to do all the heavy lifting when it comes to running applications. The network of computers that make up the cloud handles them instead. The only thing the user's computer needs to run is the cloud computing systems' interface software, which can be as simple as a Web browser and the cloud's network takes care of the rest. Cloud computing is being driven by many which includes Google, Amazon and Yahoo as well as traditional vendors including IBM, Intel and Microsoft [1]. The data should be available in the cloud for it to be accessed. There are four main types of cloud storage:

1.1 Mobile Cloud Storage

Mobile cloud storage stores the individual's data in the cloud and provides access to the data from anywhere.

1.2 Public Cloud Storage

There is no connection between the enterprise and storage service provider and the cloud resources are stored separately from the enterprise's data center. Management of resources is fully audited in the cloud storage provider's environment.

1.3 Private Cloud Storage

In private cloud storage, the storage provider has infrastructure in the enterprise's data center that is typically managed by the storage provider.

1.4 Hybrid Cloud Storage

Hybrid is a combination of public and private cloud storage where some critical data resides in the enterprise's private cloud while other data is stored and accessible from a public cloud storage provider.

Storing data in the cloud gives rise to the issue of data integrity verification at entrusted servers. If a client can log in from any location to access data and applications, it is possible that client's privacy could be compromised. Cloud computing companies will need to find ways to protect client privacy. To ensure client's

privacy, data file handling mechanism is audited by a secure third-party which was previously discussed as storage service provider.

The persona of Third Party Auditor (TPA) is listed as follows:

- Reduce data owner's burden in managing the data.
- Ensure the client that the data stored in the cloud is intact and data integrity is maintained.
- Aid in achieving high economies of scale through customer satisfaction.

There is a foreseeable increase in the tasks to be carried out by the TPA which is quite tedious. So to enable the TPA perform multiple tasks simultaneously, batch auditing is required. It reduces the communication and computation overhead without demanding the local copy of the data. Some of the prior works in the field of simultaneous auditing include HAIL, Store, Forget and Check Method, Algebraic signature method, Cooperative Internet Backup scheme. However, these schemes focus on static data rather than the live data. As a result, their capabilities of handling dynamic data remain unclear. To experience dynamic operations there should be an option to perform manipulations such as data update, delete and append. Since the data involved might be huge, there should also be a mechanism to cluster them meaningfully.

The following work discusses the storage of data in the cloud in a secure manner and the contributions are summarized as follows:

- Improve data integrity on cloud storage services in a dynamic environment through an external authorized auditor.
- Dynamic operations on the data: update, append and delete procedures.
- Simultaneous task performance through batch auditing.
- Clustering high dimensional data using frequent item set algorithm and hyper graph model.
- Increased accuracy through server labeling and user preference recording.

The rest of the paper is organized as follows: Section 2 briefs the design model of ensuring data integrity in the cloud. Section 3 explains the algorithm used for encryption and decryption. Section 4 explains the advantages in proposed work. Section 5explains the mechanism used for batch auditing in detail. Section 6 explains hyper clustering mechanism for managing high dimensional data. Section 7 briefs the dynamic operations on data [2] [3].

II. PROPOSED SYSTEM

2.1 System Entities

The different network entities involved in the entire data integrity system can be described as follows:

- User

  User is an entity whose responsibility is to store data in the cloud and rely on the cloud for data storage computation.

- Cloud Server

  Cloud Server is managed by the Cloud service provider (CSP) to provide data storage service including storage space and computation resources.

- Third Party Auditor (TPA)

  TPA has the authentication to assess and expose risk of cloud storage services on behalf of the users upon request.

In cloud data storage, user stores his data into the cloud server with the help of a Cloud Service Provider. [4] [5] Cloud servers are multiple in number and they run simultaneously in a cooperated and distributes manner.

2.2 Adversary Model

The third-party auditor also takes up the role of adversary in the proposed work. TPA has the following capabilities, which captures both the external and internal threats towards the cloud data integrity. It will check the compromised files and machine by doing the following [1]. It will corrupt the user's data files stored on individual servers thereby corrupting the server. Then the TPA corrupts the original data files by modifying or introducing its own fraudulent data to prevent original data from being retrieved by the user [4]. The TPA then becomes the valid auditor to make use of the hyper graph model to ensure that the compromised machines and fraudulent users can be eradicated completely [6].

2.3 System Model

The architecture for cloud data storage with TPA is illustrated [7] [3] in the **Fig.1**.

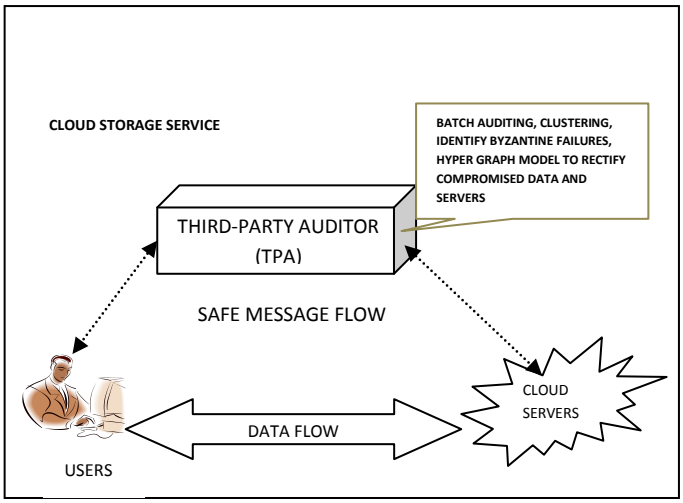

Fig.1 System design model

The cloud service storage comprises of users, set of cloud servers and optional Third-Party Auditor. Secure flow of message is ensured between them through the algorithms and procedures employed in the Auditing tasks. The TPA protects the entire system model from byzantine failures [8]. Byzantine failures are those that hide data errors from the client for its own benefit. TPA uses byzantine fault tolerance algorithm for overcoming them [7] [8]. To identify misbehaving servers, homomorphic token with erasure-coded data is used thereby achieving data integrity whenever data corruption is detected during storage [3]. Additional techniques employed by TPA can be summarized as:

Token pre-computation

- Before file distribution the user pre-computes a certain number of short verification tokens on individual vector.
- User wants to make sure storage correctness for the data in the cloud; he challenges the cloud servers with a set of randomly generated block indices.
- Each cloud server computes a short "signature" over the specified blocks and returns them to the user.

Error localization

- Integrates the correctness verification and error localization (misbehaving server identification) using challenge response protocol [4].
- The response values from servers for each challenge not only determines the correctness of the distributed storage but also contain information to locate potential data error(s)

Error Recovery

- The user can reconstruct the original file by downloading the data vectors from the first m servers, assuming that they return the correct response values
- Verification scheme is based on random spot-checking, so the storage correctness assurance is a probabilistic one [4] [5].
- The data corruption is detected. The comparison of pre-computed tokens and received response values can guarantee the identification of misbehaving server(s).

III. RSA ALGORITHM

The RSA algorithm is used for encryption and decryption of the data blocks. It is explained as follows: The RSA algorithm is a public key algorithm that can be used to send an encrypted message without a separate exchange of secret keys. It can also be used to sign a message. As the initial requirement the user and the TPA generates their own private key and public key with respect to the strong RSA algorithm. The public keys have been shared between them as the part of SLA or in some other ways. Then with respect to the protocol the message is encrypted as well as signed in a unique way. With Respect to the RSA Algorithm, the user selects two relative prime number p1 and q1 with these, the following values are computed.

n1 = p1 * q1
fn1 = (p1-1) * (q1-1)
Then, the public key α1 is selected. So, the Private Key of the TPA is:
β1 = (1/α1) % fn1

Similarly, the user selects his own relative prime numbers p2 and q2 with these the private key and the public key of the user is estimated as:

n2 = p2 * q2
fn2 = (p2-1) * (q2-1)
The public key of the user is declared as α2. So, the Private Key of the user is:
β2 = (1/α2) % fn2

Now, Key set of TPA is: {α1, n1}, {β1, n1}
Key set of User is: {α2, n2}, {β2, n2}

With the generated public key sets they get exchanged between the user and the TPA. At first the data is signed with the user's private key then the cipher is again encrypted with the TPA's public key. This package is now sent to the Cloud and also the TPA.

Data encrypted form is: {{{data} β2} α1}.

The TPA now decrypts the encrypted message with his private key and then de-signs the cipher with the user's public key to recognize the data. Then the same process of decryption is carried out in the cloud by the TPA to verify the correctness by comparing the data which he has with the stored one. Then as per the result the TPA indicates the user [2] [8].

IV. ADVANTAGES OF PROPOSED SYSTEM

By processing the integrity of data using data reading protocol and data management algorithm after and before entering data into the cloud, user can assure that all data in cloud are in protected condition for its trustworthiness. So the actual size of stored data in cloud is easily maintained even though the user himself has done any modification, deletion and update for his purpose by using proposed scheme. Here user takes full control and process on the data stored in cloud apart from TPA and he can give strong assurance and protection to the data stored in multiple cloud server environments. To avoid server failure and any unexpected error, one server restore point is put in cloud server database and efficient data back up or restore using multi server database. Server labeling is done to get quick access to the data .High dimensional data are clustered. This is the major advantage of our proposed work.

V. BATCH AUDITING AND TOKEN COMPUTATION

5.1 Batch Auditing

TPA handles multiple auditing tasks upon various users' requests. It is very time-consuming and tedious. In order to overcome this problem, TPA will delegate its auditing procedures to its users by encapsulating the underlying mechanism [3]. A delegation of audit is secure, if

- The data owner can verify whether TPA has indeed performed the audit task specified by the data owner.
- The data owner can verify whether TPA did perform the audit task at the right time specified by the data owner.
- The confidentiality of the data is protected against the TPA and/or the CSP.
- It uses a batch signature scheme called BLS signature algorithm to perform safe batching [3].

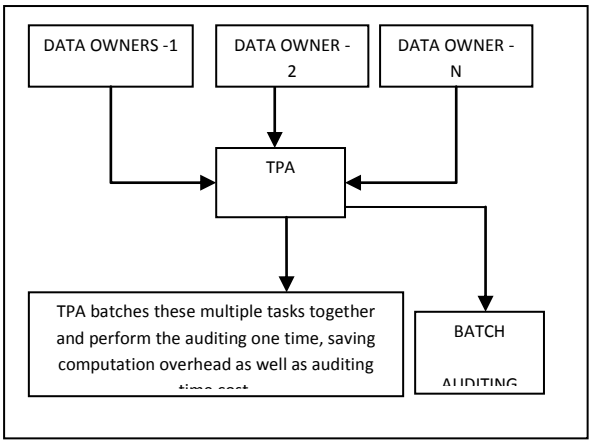

Fig. 2 Batch Auditing

Fig.2. illustrates the delegation process to TPA. The task of delegated users is to just report misbehaving servers to TPA. The BLS algorithm consists of three phases:

- Generation Phase
  In generation phase, sender chooses the private key x and public key y.
- Signing Phase
  Signing phase generates the hash function and generates a signature for message m.
- Verification Phase
  In verification phase, the receiver first computes hash function, compares it with the signature of m, if it matches, then the message is authentic.

5.2 Homomorphic Token

This technique is used to integrate the authenticator with a random masking technique [8] [3]. Four algorithms are used for generating this token:

- KeyGen to setup scheme
- SigGen to verify metadata
- GenProof to generate proof for storage correctness
- Verify Proof to audit the generated proof

The linear combination of sampled blocks in the server's response is masked with randomness generated by a Pseudo Random Function (PRF). With random masking, the TPA no longer has all the necessary information to build up a correct group of linear equations and therefore cannot derive the user's data content, no matter how many linear combinations of the same set of file blocks can be collected. Meanwhile, due to the algebraic property of the homomorphic authenticator, the correctness validation of the block-authenticator pairs will not be affected by the randomness generated [8].

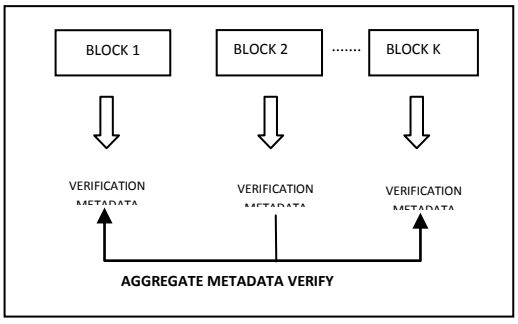

Fig.3 Homomorphic Authenticator

Fig.3 illustrates homomorphic authenticators which are unforgettable metadata generated from individual data blocks which can be securely aggregated in such a way to assure a verifier that a linear combination of data blocks is correctly computed by verifying only the aggregated authenticator. Using this technique requires additional information encoded along with the data before outsourcing. If enough linear

combinations of the same blocks are collected, the TPA can simply derive the sampled data content by solving a system of linear equations. These data blocks are sent to authenticator merely for verification [8].

VI. HYPER GRAPH MODEL

Hyper graph model is for clustering high dimensional data. In this frequent item sets found by the association rule algorithm is used as hyper edges. Also the user preference search is used to get more accurate results.

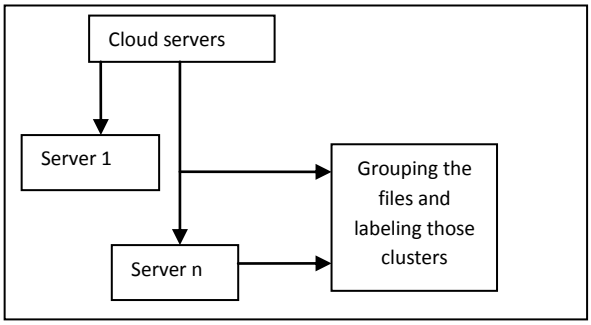

Fig.4 Labeling Servers

Here the labeling illustrated in **Fig.4** is done for each server which is situated in the cloud so that user can get access to the files quickly.

VII. DYNAMIC DATA OPERATION

The data stored in the cloud can be dynamic. User might have the requirement of performing various block-level operations such as update, delete and append to modify the existing data file. Storage correctness must also be assured at the same time. So for any dynamic data operation, the user must first generate the file blocks using the secret key. The cloud service provider checks if dynamic operations are performed correctly using verification tokens [7]. The process is summarized below as follows:

- Update Operation

  The user modifies some data blocks stores in the cloud from its current value to new value.
- Delete Operation

  The user replaces the data blocks with zero or some special reserved keywords or predetermined special blocks.
- Append Operation

  The user increases the size of data by adding blocks at the end of the data file

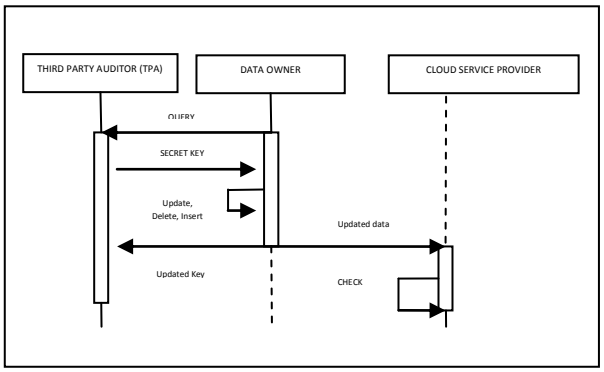

Fig.5 Dynamic Operations

To ensure the security, dynamic data operations are only available to data owners or authorized users who hold the secret key. [3] Here, all operations are based on data blocks. Moreover, to implement audit services, index-hash table needs to be updated. It is necessary for TPA and CSP to check the validity of updated data. In **Fig. 5**, we describe the process of dynamic data operations and audit. First, the data owner obtains the public verification information from TPA. Second, the data owner invokes the Update, Delete and Insert algorithms and then sends the new verification information to TPA and CSP respectively. Finally, the CSP makes use of an efficient algorithm check to verify the validity of updated data. Note that, the Check algorithm is important to ensure the effectiveness of the audit materials.

VIII. CONCLUSION

This paper focuses on auditing mechanisms to ensure data integrity where users can safely delegate the integrity checking tasks to Third Party Auditors and be worry-free to use the cloud storage services. Here both public and batch auditing is done. Batch auditing not only enables simultaneous verification from multiple clients but also reduces the computation cost on the TPA side. It effectively locates the malfunctioning server when data corruption has been detected. By relying on erasure correcting code in the file distribution preparation, data availability is guaranteed. Along with storage correctness verification, the misbehaving servers are also identified. Same level of storage correctness is maintained even if users modify, delete or append their data files in the cloud. In addition, hyper graph model is used for clustering high dimensional data. To achieve accuracy, server labeling and user preference is also included.

REFERENCES

[1]. C. Wang, Q. Wang, K. Ren, and W. Lou, "Ensuring Data Storage Security in Cloud Computing," Proc. 17th Int'l Workshop Quality of Service (IWQoS '09), pp. 1-9, July 2009.

[2]. K.D. Bowers, A. Juels, and A. Oprea, "HAIL: A High-Availability and Integrity Layer for Cloud Storage," Proc. ACM Conf. Computer and Comm. Security (CCS '09), pp. 187-198, 2009.

[3]. C. Wang, K. Ren, W. Lou, and J. Li, "Towards Publicly Auditable Secure Cloud Data Storage Services," IEEE Network Magazine, vol. 24, no. 4, pp. 19-24, July/Aug. 2010.

[4]. K. Ren, C. Wang, and Q. Wang, "Security Challenges for the Public Cloud," IEEE Internet Computing, vol. 16, no. 1, pp. 69-73, 2012.

[5]. G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z.Peterson, and D. Song, "Provable Data Possession at Untrusted Stores," Proc. 14th ACM Conf. Computer and Comm. Security (CCS '07), pp. 598-609, Oct. 2007.

[6]. G. Ateniese, R.D. Pietro, L.V. Mancini, and G. Tsudik, "Scalable and Efficient Provable Data Possession," Proc. Fourth Int'l Conf. Security and Privacy in Comm. Netowrks (SecureComm '08), pp. 1-10, 2008.

[7]. Q. Wang, C. Wang, J. Li, K. Ren, and W. Lou, "Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing," Proc. 14th European Conf. Research in Computer Security (ESORICS '09), pp. 355-370, 2009.

[8]. T. Schwarz and E.L. Miller, "Store, Forget, and Check: Using Algebraic Signatures to Check Remotely Administered Storage," Proc. IEEE Int'l Conf. Distributed Computing Systems (ICDCS '06), pp. 12 12, 2006.