International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 12 265 – 270

A Comprehensive Survey on Data Integrity Proving Schemes in Cloud Storage

Miss. Moakumla
M.Tech (CSE) Student
Assam Down Town University
Panikhaiti, Assam, India
aierakum@yahoo.co.in

Dr. Lakshmi Prasad Saikia
Professor & HOD, Computer Science &Engineering
Assam Down Town University
Panikhaiti, Assam, India
lp_saikia@yahoo.co.in

***Abstract*—**Cloud computing requires broad security solutions based upon many aspects of a large and lightly integrated system. The cloud data storage service releases the users from the burden of huge local data storage and their preservation by out- sourcing mass data to the cloud. However, the fact that users no longer have physical possession of the possibly large size of outsourced data makes the data integrity protection in Cloud Computing a very challenging and potentially formidable task, especially for users with constrained computing resources and capabilities. One of the significant concerns that need to be spoken is to assure the customer of the integrity i.e. rightness of his data in the cloud.

The data integrity verification is done by introducing third party auditor (TPA) who has privileges to check the integrity of dynamic data in cloud on behalf of cloud client. Cloud client can get notification from TPA when the data integrity is lost. These systems have sustenance data dynamics via the data operation such as data modification, insertion, deletion. Many work has been done but it lacks the support of either public auditability or active data processes

To securely introduce an effective third party auditor (TPA), the following two fundamental requirements have to be met: (i) TPA should be able to efficiently audit the cloud data storage without demanding the local copy of data, and introduce no additional on-line burden to the cloud user; (ii) The third party auditing process should bring in no new vulnerabilities towards user data privacy. Here, a proposed scheme is discussed in which gives a proof of data integrity in the cloud which the customer can employ to check the correctness of his data in the cloud. This proof can be agreed upon by both the cloud and the customer and can be incorporated in the Service level agreement (SLA). This scheme ensures that the storage at the client side is minimal which will be beneficial for the organization. In this paper, we define a survey on Cloud computing and provide the architecture for creating Clouds, characteristics, deployments, and integrity proofs etc.

Keywords- *Data Integrity, Cryptography, TPA, Cloud Computing, Cloud Storage, POR, PDP.*

*****

I. INTRODUCTION

A. Cloud Computing

In the simplest terms, cloud computing means storing and accessing data and programs over the Internet instead of your computer's hard drive.

In a cloud computing system, there is significant workload

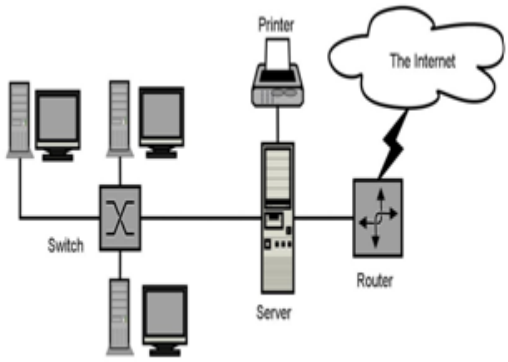

Figure 1. Cloud Computing Network

shift. 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. Hardware and software demands on the user's side decrease. The only thing the user's computer needs to be able 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.

Many companies are delivering services from the cloud. Some notable examples are as follows:

i) Google — has a private cloud that it uses for delivering many different services to its users, including email access, document applications, text translations, maps, web analytics, and much more.

ii) Microsoft — Has Microsoft SharePoint online service that allows for content and business intelligence tools to be moved into the cloud and Microsoft currently makes its office applications available in a cloud.

iii) Salesforce.com — runs its application set for it customers in a cloud, and it's Force.com and Vmforce.com products provide developers with platforms to build customized cloud services. The following sections note cloud and cloud computing characteristics, services models, deployment models, benefits, and challenges.

B. Characteristics

The characteristics of cloud computing include on-demand self service, broad network access, resource pooling, rapid elasticity and measured service. On-demand self service means that customers (usually organizations) can request and manage their own computing resources. Broad network access allows services to be offered over the Internet or private networks. Services can be scaled larger or smaller; and use of a service is measured and customers are billed accordingly.

Cloud computing has a variety of characteristics, with the main ones being:

i) Shared Infrastructure — Uses a virtualized software model, enabling the sharing of physical services, storage, and networking capabilities. The cloud infrastructure, regardless of deployment model, seeks to make the most of the available infrastructure across a number of users.

ii) Dynamic Provisioning — Allows for the provision of services based on current demand requirements. This is done automatically using software automation, enabling the expansion and contraction of service capability, as needed. This dynamic scaling needs to be done while maintaining high levels of reliability and security.

iii) Network Access — Needs to be accessed across the internet from a broad range of devices such as PCs, laptops, and mobile devices, using standards-based APIs (for example, ones based on HTTP). Deployments of services in the cloud include everything from using business applications to the latest application on the newest smartphones.

iv) Managed Metering — uses metering for managing and optimizing the service and to provide reporting and billing information. In this way, consumers are billed for services according to how much they have actually used during the billing period. In short, cloud computing allows for the sharing and scalable deployment of services, as needed, from almost any location, and for which the customer can be billed based on actual usage.

C. Service Models

Once a cloud is established, how its cloud computing services are deployed in terms of business models can differ depending on requirements. The primary service models being deployed are commonly known as:

I. Software as a Service (SaaS) — Consumers purchase the ability to access and use an application or service that is hosted in the cloud. A benchmark example of this isSalesforce.com, where necessary information for the interaction between the consumer and the service is hosted as part of the service in the cloud. Also, Microsoft is expanding its involvement in this area, and as part of the cloud computing option for Microsoft Office 2010, its Office Web Apps are available to Office volume licensing customers and Office Web App subscriptions through its cloud-based Online Services.

II. Platform as a Service (PaaS) — Consumers purchase access to the platforms, enabling them to deploy their own software and applications in the cloud. The operating systems and network access are not managed by the consumer, and there might be constraints as to which applications can be deployed.

III. Infrastructure as a Service (IaaS) — Consumers control and manage the systems in terms of the operating systems, applications, storage, and network connectivity, but do not themselves control the cloud infrastructure.

D. Deployment of cloud services

Deploying cloud computing can differ depending on requirements, and the following four deployment models have been identified, each with specific characteristics that support the needs of the services and users of the clouds in particular ways.

I. Private Cloud — the cloud infrastructure has been deployed, and is maintained and operated for a specific organization. The operation may be in-house or with athird party on the premises.

II. Community Cloud — the cloud infrastructure is shared among a number of organizations with similar interests and requirements. This may help limit the capital expenditure costs for its establishment as the costs are shared among theorganizations. The operation may be in-house or with a third party on the premises.

III. Public Cloud — the cloud infrastructure is available to the public on a commercial basis by a cloud service provider. This enables a consumer to develop and deploy a service in the cloud with very little financial outlay compared to the capital expenditure requirements normally associated with other deployment options.

IV. Hybrid Cloud — the cloud infrastructure consists of a number of clouds of any type, but the clouds have the ability through their interfaces to allow data and/orapplications to be moved from one cloud to another. This can be a combination of private and public clouds that support the requirement to retain some data in anorganization, and also the need to offer services in the cloud.

E. Benefits

The following are some of the possible benefits for those who offer cloud computing-based services and applications:

I. Cost Savings — Companies can reduce their capital expenditures and use operational expenditures for increasing their computing capabilities. This is a lower barrier to entry and also requires fewer in-house IT resources to provide system support.

II. Scalability/Flexibility — Companies can start with a small deployment and grow to a large deployment fairly rapidly, and then scale back if necessary. Also, the flexibility of cloud computing allows companies to use extra resources at peak times, enabling them to satisfy consumer demands.

III. Reliability — Services using multiple redundant sites can support business continuity and disaster recovery.

IV. Maintenance — Cloud service providers do the system maintenance, and access is through APIs that do not require application installations onto PCs, thus further reducing maintenance requirements.

V. Mobile Accessible — Mobile workers have increased productivity due to systems accessible in an infrastructure available from anywhere.

II. INTRODUCTION OF DATA STORAGE IN CLOUD

Cloud Storage is a crucial service of cloud computing, that permits information house owners (owners) to maneuver data from their native computing systems to the cloud. More and additional house owners begin to store the information within the cloud. However, this new paradigm of information hosting service additionally introduces new security challenges. Data owners would worry that the information can be lost within the cloud.

This is because information loss might happen in any infrastructure, no matter what high degree of reliable measures cloud service providers would take. Sometimes, cloud service suppliers could be dishonest. They may discard the data that haven't been accessed or seldom accessed to save the cupboard space and claim that the information area unit still correctly hold on within the cloud. Therefore, house owners have to be compelled to be convinced that the information area unit properly holds on within the cloud. Traditionally, house owners will check the information integrity based mostly on two-party storage auditing protocols. In cloud storage system, however, it is inappropriate to let either facet of cloud service providers or house owners conduct such auditing, as a result of none of them can be sure to give unbiased auditing result. During this scenario, third-party auditing could be a natural choice for the storage auditing in cloud computing. A third-party auditor (auditor) that has experience and capabilities can do a additional economical work and persuade each cloud service suppliers and house owners.

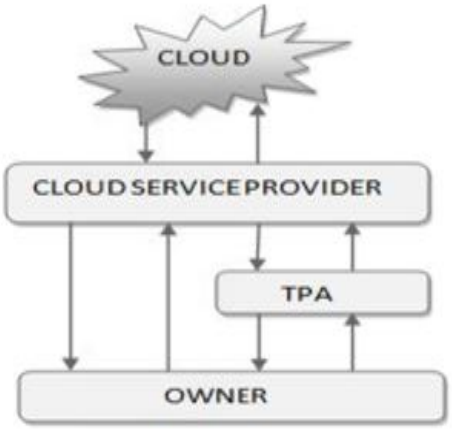

Figure 2. Architecture of Cloud Storage

For the third party auditing in cloud storage systems, there are units many necessary needs that are projected in some previous works. The auditing protocol ought to have the subsequent properties:

1. Confidentiality. The auditing protocol ought to keep owner's information confidential against the auditor.
2. Dynamic auditing. The auditing protocol ought to support the dynamic updates of the data within the cloud.
3. Batch auditing. The auditing protocol ought to even be able to support the batch auditing for multiple house owners and multiple clouds.

The cloud storage is better than all traditional storage methods. Because of the following reasons

i. Companies do not need to install physical storage devices in their own data centre nor offices.

ii. Storage maintenance tasks, such as backup, and purchasing additional storage devices are offloaded to the responsibility of a service provider, allowing organizations to focus on their core business.

iii. Companies need only pay for the storage they actually use.

Some of the characteristics of cloud storage are as follows:

I. Performance

II. Manageability

III. Availability

Some of the best examples for cloud storage are Amazon S3, Windows Azure Storage, EMC Atmos, FilesAnywhere, Google Cloud Storage, Google App Engine Blobstore, iCloud by Apple. Recently, many remote integrity checking protocols were projected to permit the auditor to envision the information integrity on the remote server. Table 1 provides the comparisons among some existing remote integrity checking schemes in terms of the performance, the

privacy protection, the support of dynamic operations and also thebatch auditing for multiple owners and multiple clouds. From Table 1, they will notice that many of them aren't privacy conserving or cannot support the information dynamic operations, in order that they cannot be applied to cloud storage systems. In [11] the authors projected a dynamic auditing protocol that may support the dynamic operations of the data on the cloud servers, however this technique might leak the data content to the auditor as a result of it needs the server to send the linear mixtures of information blocks to the auditor.

Table 1.Comparison of remote integrity checking schemes

Parameters	SPDP(interactive provable data possession)	DPDP
type of guarantee	probabilistic/Deterministic	Probabilistic
data dynamics	append only	Yes

In [12] the authors extended their dynamic auditing scheme to be privacy conserving and support the batch auditing for multiple house owners. However, as a result of the massive number of information tags, their auditing protocols might incur a heavy storage overhead on the server. In proposed work a cooperative demonstrable information possession theme that can support the batch auditing for multiple clouds and also extend it to support the dynamic auditing. However, their theme cannot support the batch auditing for multiple house owners. That's as a result of parameters for generating the information tags employed by every owner area unit completely different, and thus, they cannot mix the information tags from multiple owners to conduct the batch auditing. Another downside is that their theme needs a further sure organizer to send a commitment to the auditor throughout the multi cloud batch auditing, as a result of their theme applies the mask technique to confirm the information privacy. However, such additional organizer isn't sensible in cloud storage systems.

A. Overview of the Network and Cloud Security

Network Security is an organization's strategy and provisions for ensuring the security of its assets and of all network traffic. Network security is manifested in an implementation of security policy, hardware, and software. For the purposes of this discussion, the following approach is adopted in an effort to view network security in its entirety Policy, Enforcement, Auditing, Policy .The IT Security Policy is the principle document for network security. Its goal is to outline the rules for ensuring the security of organizational assets. Employees today utilize several tools and applications to conduct business productively. Policy that is driven from the organization's culture supports these routines and focuses on the safe enablement of these tools to its employees. The enforcement and auditing procedures for any regulatory compliance an organization is required to meet must be mapped out in the policy as well.

1) Enforcement

Most definitions of network security are narrowed to the enforcement mechanism. Enforcement concerns analyzing all network traffic flows and should aim to preserve the confidentiality, integrity, and availability of all systems and information on the network. These three principles compose the CIA triad:

i. Confidentiality - involves the protection of assets from unauthorized entities
ii. Integrity - ensuring the modification of assets is handled in a specified and authorized manner
iii. Availability - a state of the system in which authorized users have continuous access to said assets.

B. Overview of the System

Data outsourcing to cloud storage servers is raising trend among many firms and users owing to its economic advantages. This essentially means that the owner (client) of the data moves its data to a third party cloud storage server which is supposed to - presumably for a fee – faithfully store the data with it and provide it back to the owner whenever required.

As data generation is far outpacing data storage, it proves costly for small firms to frequently update their hardware whenever additional data is created. Also maintaining the storages can be a difficult task. Storage outsourcing of data to cloud storage helps such firms by reducing the costs of storage, maintenance and personnel. It can also assure a reliable storage of important data by keeping multiple copies of the data thereby reducing the chance of losing data by hardware failures.

Storing of user data in the cloud despite its advantages has many interesting security concerns which need to be extensively investigated for making it a reliable solution to the problem of avoiding local storage of data. The problem of implementing a protocol for obtaining a proof of data possession in the cloud sometimes referred to as Proof of retrievability (POR).This problem tries to obtain and verify a proof that the data that is stored by a user at a remote data storage in the cloud is not modified by the archive and thereby the integrity of the data is assured.

Such verification systems prevent the cloud storage archives from misrepresenting or modifying the data stored at it without the consent of the data owner by using frequent checks on the storage archives. Such checks must allow the

data owner to efficiently, frequently, quickly and securely verify that the cloud archive is not cheating the owner.

III. OVERVIEW OF DATA INTEGRITY

As the word suggests itself data integrity means completeness or wholeness and it is basic requirement of information technology. Data integrity refers to maintaining and assuring the accuracy and consistency of data over its entire life-cycle. Data corruption is a form of data loss and data integrity is opposite of data corruption. Data integrity ensures the data is the same as it was when it was originally recorded.

A. Physical Vs. Logical Integrity

Data integrity can be roughly divided into two overlapping categories Physical integrity and logical integrity Physical integrity deals with challenges related to storing and fetching of the data. Challenges for the physical integrity may include electromechanical faults, design flaws, material fatigue, corrosion, power outages, natural disasters, acts of war and terrorism. Physical integrity makes use of error detecting algorithms known as error correcting codes. Logical integrity is related with the correctness or rationality of a piece of data. Types of integrity constraints are as referential integrity, entity integrity and domain integrity.

Entity integrity is an integrity rule which states that every table must have a primary key and that the column or columns chosen to be the primary key should be unique and not null. The referential integrity rule states that any foreign-key value can only be in one of two states. Domain integrity specifies that all columns in relational database must be declared upon a defined domain. User-defined integrity refers to a set of rules specified by a user and which do not belong to the domain, entity and referential integrity categories. Data Integrity is necessary in databases and it is also necessary in data Stored in the cloud. Data integrity is a factor that affects on the performance of the cloud.

B. Data Integrity Proving Schemes

1) Provable Data Possession (PDP)

G. Ateniese, R. Curtmola, R. Burns, J. Herring, L. Kissner, Z. Peterson, and D. Song introduced a model for provable data possession (PDP) [7].This model allows to check integrity of the data without retrieving it which is stored by the client [7]. To reduce input and output cost this model generates the probabilistic proof of possession by sampling random sets of blocks [7].This PDP scheme does not include any error correcting code [7]. This scheme works in two phases' setup phase and challenge phase [7]. Ateniese et al. [7] developed first PDP scheme in which they considered public audiability in their model for ensuring possession of data on untrusted storage. In this scheme Homomorphic Variable tags are used to for auditing outsourced data. But this scheme is beneficial for only static data they do not consider dynamic data storage. Later Ateniese et al. [7] proposed dynamic version of PDP scheme but it does not support fully dynamic data operations. This scheme offers only limited functionality and very basic blocks of operations. Erway et al. [13] were who proposed a scheme for dynamic PDP. They extended the PDP scheme proposed by the Ateniese et al. [7].This scheme is later improved by Feifei Liu[14]. This newly proposed scheme reduces the computational and communication complexity.

2) Proof of Retrievability

In the POR scheme the scheme using keyed hash function is the simplest scheme than any other scheme for proof of retrivability of data files. In this scheme the data file is stored in the cloud storage but before storing it in the cloud storage that file is pre-processed and cryptographic hash is computed. After calculating hash value the file is stored in the cloud storage. The cryptographic key which is used to calculate hash value is then released to the cloud storage and verifier ask to calculate hash value again. Then values calculated by the verifier and values calculated by the cloud storage are compared with each other. From that comparison the final conclusion is considered. The main advantage of this scheme is simple to implement. Limitation of this scheme is, it is computational burdensome for the devices like mobile phones, PDAs etc.

Another scheme for proof retrivability is using sentinels .This scheme is proposed by Ari Juels and Burton S. KaliskiJr [6]. Sentinels are the special blocks which are used in this scheme to verify the integrity. Sentinels are embedded in the data blocks randomly during setup phase by the verifier in the setup phase . The integrity of the data file is calculated by challenge and response. The verifier throws the challenge to the cloud storage by specifying the position of the collection of the sentinels and the cloud storage has to return the associated sentinels values to the verifier. If the file stored by the client is modified then the associated sentinels' values also get changed and the cloud will return wrong values to the verifier. From this integrity of the file is checked. Limitation of this scheme is that this scheme involves encryption of file so this is computationally cumbersome for the small devices like mobile phones; PDA etc .Also storage overhead will be there due to newly inserted sentinels and error correcting code.

Sravan Kumar R and Ashutosh Saxena present a scheme [15] which involves selection of random bits per blocks of data due to this computational overhead of the client is reduced. File is processed by the verifier before storing it in the cloud storage. After that verifier attaches some Meta data to the file .This Meta data is used at the time of verification of the integrity of the file. The limitation of this scheme is this scheme applies only static data.

IV. CONCLUSION

The operation of encryption of data generally consumes a large computational power. In our scheme the encrypting process is very much limited to only a fraction of the whole data thereby saving on the computational time of the client. Many of the schemes proposed earlier require the archive to perform tasks that need a lot of computational power to generate the proof of data integrity. But in our scheme the archive just need to fetch and send few bits of data to the client. It reduces the access time of the cloud server and reduces the cost for retrieving the file and bandwidth consumption across the network.

From this survey we observed that data integrity is very important part of the cloud storage because data is not locally stored. In cloud computing data is remotely stored it is also called as archive or server. PDP scheme allows a client to check that the server possesses the original data without retrieving it that has stored by client. By sampling random sets of blocks from server this model generates probabilistic proofs of possession. This scheme also has limitations because it takes only static data. From this survey we can say that there is vast scope in the field of data integrity field for cloud storage.

REFERENCES

[1] Cong Wang, Student Member, IEEE, Qian Wang, Student Member, IEEE,KuiRen,Member,IEEE,Ning Cao, Student Member, IEEE, and Wenjing Lou, Senior Member, IEEE, " Towards Secure and Dependable Storage Services in Cloud Computing.", 2011.

[2] Addressing cloud computing security issues, Future generation computer systems(2011) www.elsevier.com/locate/fgcs.

[3] T. Wobber, T. L. Rodeheffer, and D. B. Terry, "Policybased access control for weakly consistent replication," in ACM EuroSys, 2010.

[4] E. Mykletun, M. Narasimha, and G. Tsudik, "Authentication and integrity in outsourced databases," Trans. Storage, vol. 2, no. 2, pp. 107–138, 2013.

[5] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in SP '00: Proceedings of the 2000 IEEE Symposium on Security and Privacy. Washington, DC, USA: IEEE Computer Society, 2012, p. 44.

[6] A. Juels and B. S. Kaliski, Jr., "Pors: proofs of retrievability for large files," in CCS '07: Proceedings of the 14th ACM conference on Computer and communications security. New York, NY, USA: ACM, 2007, pp.584–597.

[7] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson,and D. Song, "Provable data possession at untrusted stores," in CCS'07: Proceedings of the 14th ACM conference on Computer and communications.

[8] Beginning ASP.NET 3.5 in C# 2008: From Novice to Professional, Second Edition by Matthew MacDonald.

[9] Ateniese, Giuseppe, et al. "Improved proxy re-encryption schemes with applications to secure distributed storage." ACM Transactions on Information and System Security (TISSEC) 9.1 (2006): 1-30.

[10] Lu, Rongxing, et al. "Secure provenance: the essential of bread and butter of data forensics in cloud computing." Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security. ACM, 2010.

[11] Marshall D.Abrams, Harold J.podell on Cryptography.

[12] Anoop Ms, "public key cryptography Applications Algorithms and Mathematical Explanations".

[13] C. Erway, A. Kupcu, C. Papamanthou, and R. Tamassia,"Dynamic Provable Data Possession," Proc. 16th ACM Conf. Computer and Comm. Security (CCS '09), 2009.

[14] Feifei Liu, Dawu Gu, Haining Lu," An Improved Dynamic Provable Data Possession Model," 978-1-61284-204-2/11/$26.00 ©2011 IEEE.

[15] Sravan Kumar R, Ashutosh Saxena," Data Integrity Proofs in Cloud Storage,"ISBN: 978-1-4244-8953-4/11/$26.00 c 2011 IEEE.