DATA INTEGRITY TECHNIQUES IN CLOUD: AN ANALYSIS

[1] MS. R. K. PANDYA, [2] PROF. K. K. SUTARIA

[1]M.E.[Cloud Computing] Student, Computer Engineering Department, V. V. P. Engineering College, Rajkot, Gujarat
[2]Asst. Professor, Computer Engineering Department, V. V. P. Engineering College, Rajkot, Gujarat

riddspandya90@gmail.com, kamal.sutaria@gmail.com

***ABSTRACT*: *C*loud computing is the new technology in the field of information and technology. It provides so many things in terms of "As-A-Service" basis. For cloud storage privacy and security are the burning issues. When users store their data on the cloud then there may be a risk of losing the data, or sometimes data may be modified or updated. It may not be fully trustworthy because client doesn't have copy of all stored data. Cloud storage moves the user's data to large data centers, which are remotely located, on which user does not have any control. In this paper we will discuss the and privacy concerns of cloud environment. This paper mainly focuses on the survey of the privacy techniques that has been proposed so far for the data integrity like POR (Proof Of Retrivability), PDP (Provable Data Possession), HAIL (A High-Availability and Integrity Layer For Cloud Storage), Static PDP, Dynamic PDP, etc in the cloud environment for the data integrity.**

Keywords— Cloud computing, privacy, PDP(Provable Data Possession),POR(Proof of Retrivability), HAIL (A High-Availability and Integrity for Cloud Storage), Static PDP, Dynamic PDP, Data Integrity, Mobile Computing.

I: INTRODUCTION

Cloud computing is a new paradigm and dimension of the information and technology (IT), which aims to provide reliable, customized and guaranteed computing dynamic environment on "Pay-per-use" or "Pay-as-you-go" basis for the end users. "Clouds are a large pool of easily available, usable and accessible virtualized resources (such as hardware, development platform and/or services). These resources can dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization."[1] Cloud computing is a technology which provide you a service through which you can use all the computer hardware and software sitting on your desktop, or somewhere inside your company's network but they are not actually installed on your computer, it is provided for you as a service by another company and accessed over the internet. End users can access these services available in the internet without knowing how these resources are managed and where. This is transparent to end users. [2]

It is always available. It is highly mobile and available across platforms. Cloud reduced the cost of deployment. It allows unlimited storage. Cloud increased computing power with rapid scalability as and when needed. It allows easier workgroup collaboration in real time. It reduced risk of data loss. Clouds require fewer maintenance issues as there is no need to install or upgrade software and hardware. Cloud improves compatibility between operating system. [3][14]

It requires always on and high-speed internet connectivity. This technology is still at a nascent stage. Still, in cloud computing, there is unresolved, security and privacy issues. There is a lack of industry standards and inter-operability among applications. It is having limited features. In clouds users are subject to many terms and conditions. It is not environmentally sustainable. Sometimes your data will be lost from the server. [3][14]

II. SERVICE MODELS OF CLOUD COMPUTING

Depending on the services provided by the cloud, it is divided into main three categories:

A. INFRASTRUCTURE AS A SERVICE (IAAS)

IAAS is a service which provides an access to the hardware resources such as storage or computing hardware as pay per use basis. Example of this type of service is suppose you pay monthly or yearly subscription to hosting company which in turn stores your files on their server.[1][2][14]

B. SOFTWARE AS A SERVICE (SAAS)

SAAS provides a software services to the end user. Web-based email and Google documents are best example of this service. End user gets access to this software service but he/she cannot modify this software utility. Software is configured on cloud utility not installed on end user computer. [1][2][14]

C. PLATFORM AS A SERVICE (PAAS)

This service provides a platform or an environment on which end user can develop his own application. User is transparent about the location of the platform whether it is hosted on cloud or not. Google App Engine is an example of PAAS. [2][14]

Some of the best examples for cloud storage are Amazon S3, Windows Azure Storage, EMC Atmos, Files Anywhere, Google Cloud Storage, Google App Engine Blobstore, iCloud by Apple.

III. DEPLOYMENT MODELS OF CLOUD COMPUTING

Fig. 1 shows the types of cloud deployment models. These are describing in the following section.

A. PUBLIC CLOUD

In a public cloud the computing infrastructure is used by the organization or end user through cloud service providers or vendors. Public clouds are typically offered through virtualization and distributed among various physical machines. [1][2] [14]

B. PRIVATE CLOUD

In a private cloud the computing infrastructure is dedicated to the particular organizations and not shared with other organization. Private clouds are more secure than public clouds. [2][14]

C. HYBRID CLOUD

This is a combination of the other two types of cloud. In hybrid cloud organizations may host critical application on private clouds and applications which are having less security concerns hosted on public clouds. It is also known as cloud bursting. [2][14]

D. COMMUNITY CLOUD

It involves sharing of computing infrastructure in between organizations of the same community. For example all Government organizations within the state of Gujarat may share computing infrastructure on the cloud to manage data related to the citizens residing in Gujarat. [2][14]

IV. PRIVACY

Privacy and security are the burning issues of any technology, and cloud environment is also not an exception. As cloud is still at its nascent stage, privacy and security get more concern. Privacy is the fundamental right of the human being. Privacy of the user data and personal information can be provided by the cryptographic function and technology. As cloud computing is the virtual environment in which the autonomous systems are connected in a network and this will create the cloud. Now this cloud will serve “as a service” basis, so user have to registered himself to the cloud server or to the third party which provide the cloud service. So the privacy of the data and security need to be considered.

Literature represents the jeopardize of cloud security and privacy that is a client side security, security concerns from cloud service providers, data ownership and data location, lack of control over the data, network security, data recovery on cloud environment, securing data in cloud environment, installation and maintenance of firewalls, data encryption, data sanitization, certification and auditing, backup and recovery, and identity and access management. [4][14]

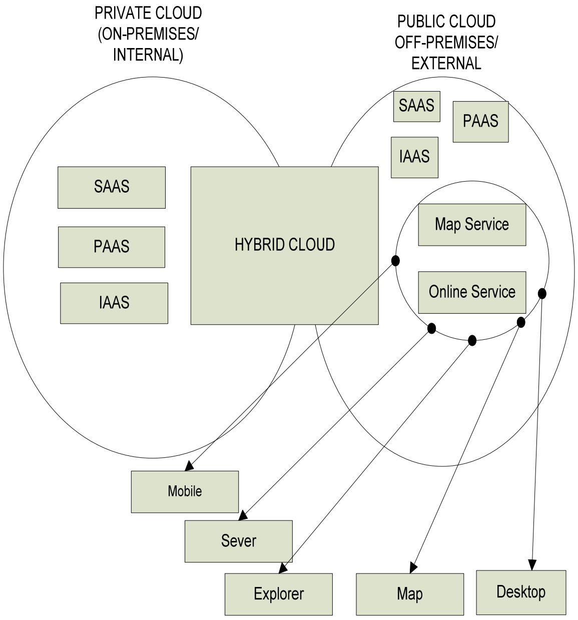

Fig. 1 Cloud Types

V. PRIVACY TECHNIQUES

Author of paper [5] suggest some defense strategies for the data integrity. Integrity checking on is a wide topic for research.

Traditional methods cannot be directly applied for the integrity checking, because the main issue for integrity checking of data is that tremendous amount of data are remotely stored on cloud server which are untrustworthy. Sometimes it is not feasible to download entire file and perform the integrity check due to the fact that it is computationally expensive and bandwidth consuming.

The following section describes the privacy techniques for data integrity.

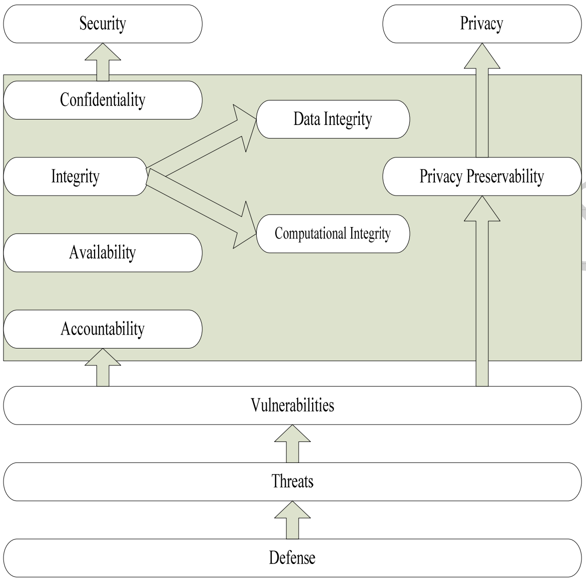

Fig. 2 Cloud security and privacy

VI.PROVABLE DATA POSSESSION

This technique is used to check the integrity of the data that is stored on the cloud server. These all techniques are used for the client to periodically check their data that is stored on a server. So this technique is for the customer to ensure that their data is secure on the server or not. This PDP includes the following number of techniques to perform integrity check on the data.

A. A NAIVE METHOD

The main idea behind this algorithm is to compare the data. In this method client will compute the hash value for th file F and having key *K* (i.e. *h*(*K*,F)) and subsequently it will send the file F to the server. Clients are having different collection of keys and hash values so it can check multiple check on the file F. Whenever client wants to check the file it release key *K* and sends it to the server, which is then asked to recomputed the hash value, based on F and *K.* Now server will reply back to the client with hash value for comparison.

This method gives the strong proof that server is having the original file F. But this method has high overhead as every time hashing process is run over the entire file. It is having very high computation cost.

B. ORIGINAL PROVABLE DATA POSSESSION

In this technique, the data is pre-processed before sending it to the cloud server. Here the data is filled with some tag value or say meta-data for the verification at the client side. Now entire data is sent over to the server and at the client side meta-data is stored. This meta-data is used for the verification when user need for it. When user wants to check for integrity it will sends the challenge to the server at that time server will respond with the data. Now the client will compare the reply data with the local meta-data. In this way client will say that the data is modified or not. Original PDP has low computation and storage overhead.

It supports both encrypted data and plain data. It offers public verifiability. It is efficient because small portion of the file needs to be accessed to generate proof on the server. This technique is only applicable to the static files (i.e. append-files only). Probabilistic guarantees may result in false positive.

Homomorphic hashing technique is used to compose multiple block inputs into a single value to reduce the size of proof. [5][7][14]

C. PROOF OF RETRIVABILITY (POR)

Proof of retrievability means Verify the data stored by user at remote storage in the cloud is not modified by the cloud. POR for huge size of files named as sentinels. The main role of sentinels is cloud needs to access only a small portion of the file (F) instead of accessing entire file.

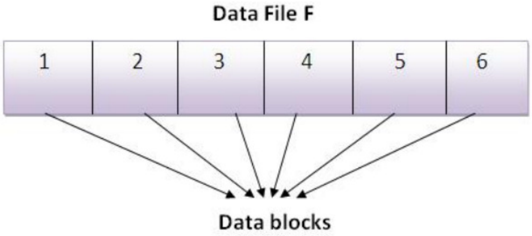

Fig. 3 A data file with 6 blocks

The above fig. 3 shows the data file which is having 6 data blocks and these entire 6 data block contains individual sentinels.

This technique uses the auditing protocol when solving the problem of integrity. Here any client who wants to check the integrity of the outsourced data then there is no need to retrieve full content.

Here user stores only a key, which is used to encode a file F which gives the encrypted file F'. This procedure leaves the set of sentinel values at the end of the file F'. Server only stores F'. Server doesn't know that where the sentinel value are stored because they indistinguishable from regular and it is randomly stored in the file F'.

Author of paper [13] proposed a Schematic view of a proof of retrievability based on inserting random sentinels in the data file. Semantic view of POR is shown in Fig. 4.

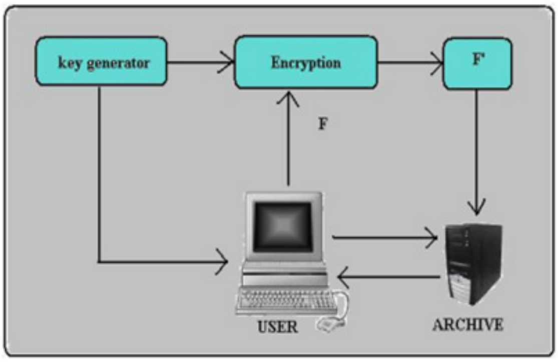

Fig.4 Schematic view of a POR

The above architecture describes that, user (cloud client) likes to store a file (F) in the cloud server (archive). Before storing the file to the cloud, owner needs to encrypt the file in order to prevent from the unauthorized access. Juels and Kaliski [13] proposed a scheme called Proof of Retrievability (POR). Proof of retrievability means Verify the data stored by user at remote storage in the cloud is not modified by the cloud. POR for huge size of files named as sentinels. The main role of sentinels is cloud needs to access only a small portion of the file (F) instead of accessing entire file.

When client send the challenge to the server to check for integrity, at that time in challenge response protocol server will ask to return a subset of sentinels in F'. If the data is tampered or deleted the sentinels may get corrupted or lost and so the server is unable to generate the proof of the original file. In this way client can prove that server has modified or corrupted the file.

POR is designed to be lightweight; it attempts to provide minimum storage in client and server side, number of data blocks accessed, the communication complexity of an audit. POR also provides the error-correcting codes to recover file having small fraction being corrupted. POR can be applied to the static files only. File needs to be encrypted before uploading to the server. It requires additional space to hide sentinels. [5][6][8][12][14]

D. SCALABLE PDP

Scalable PDP is an improved version of the original PDP. The difference is that Scalable PDP uses the symmetric encryption while original PDP uses public key to reduce computation overhead. Scalable PDP can have dynamic operation on remote data. Scalable PDP has all the challenges and answers are pre-computed and limited number of updates.

Scalable PDP does not require bulk encryption. It relies on the symmetric-Key which is more efficient than public-Key encryption. So it does not offer public verifiability. [5][9][14]

E. DYNAMIC PDP

As the name suggest this is the dynamic PDP so it supports full dynamic operations like insert, update, modify, delete etc. Here in this technique the dynamic operation enables the authenticated insert and delete functions with rank-based authenticated directories and with a skip list. Although DPDP has some computational complexity, it is still efficient. For example, to generate the proof for 1GB file, DPDP only produces 415KB proof data and 30ms computational overhead.

This technique offers fully dynamic operation like modification, deletion, insertion etc. as it supports fully dynamic operation there is relatively higher computational, communication, and storage overhead. All the challenges and answers are dynamically generated. [5][10][14]

F. A HIGH AVAILABILITY AND INTEGRITY LAYER FOR CLOUD STORAGE (HAIL)

HAIL is different from the other techniques those have been discussed so far. HAIL allows the client to store their data on multiple servers so there is a redundancy of the data. And at the client side only small amount of data is stored in local machine. The threats that can be attacked on HAIL is mobile adversaries, which may corrupt the file F.

This technique is only applicable for the static data only (i.e. append data). It is possible to check data integrity in the distributed storage via data redundancy. Here proof is generated that is independent of the data size and it is compact in size. HAIL uses the pseudorandom function, message authentication codes (MACs), and universal has function for the integrity process. [5][11][14]

VII.CONCLUSION AND FUTURE WORK

Privacy and security are the burning issues of any technology. Literature says that security and privacy are not the two different things but privacy gets concern within the security. As cloud is mainly used for the storage of the data, data integrity is the main issue of the client. After uploading data to the server, client will lost the control of the data. So at that time data can be modified or updated or sometimes deleted by the unauthorized access or by server. There are so many techniques available in the literature out of which we have analyzed some of the techniques and compare them. According that PDP techniques are very useful for the integrity checking. PDP supports fully dynamic operations so it is possible to verify data in case of modification or deletion. This techniques can be manipulated to reduce the computational and storage overhead of the client as well as to minimize the computational overhead of the remote storage server. New techniques can be invented to minimize the size of the proof of data integrity so as to reduce the network bandwidth consumption.

REFERENCES

[1] J. Ruiter and M. Warnier, Privacy regulation for cloud computing, compliance and implementation in theory and practice, article.

[2] P. Metri and G. Sarote, Privacy Issues and Challenges in Cloud Computing, International journal of Advanced Engineering Sciences and Technologies, Vol. No. 5, Issue No. 1,001-006.

[3] A white paper on , Cloud Computing : What How and Why.

[4] A. Methew, Security and Privacy Issues of Cloud computing: solution and secure framework, International Journal of Multidisciplinary Research Vol.2 Issue 4, (2012), ISSN 2231 5780

[5] Z. Xiao and Y. Xiao, Senior Member, IEEE, Security and Privacy in Cloud Computing, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, ACCEPTED FOR PUBLICATION.

[6] H.shancham and B.Waters, Compact proofs of Retrivability, Advances in Cryptology- ASIACRYPT (2008) 90-107.

[7] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, Provable data possession at untrusted stores, ACM CCS, (2007) 598-609.

[8] A. Juels and B. S. Kaliski, PORs: Proofs of retrievability for large
files, ACM CCS, (2007) 584-597.

[9] G. Ateniese, R. D. Pietro, L. V. Mancini, and G. Tsudik, Scalable and efficient provable data possession, SecureComm, 2008.

[10] C. Erway, A. K ¨ upc¸¨u, C. Papamanthou, and R. Tamassia, Dynamic provable data possession, Proc. 16th ACM conference on Computer and communications security, (2009) 213-222.

[11] K.D. Bowers, A. Juels, and A. Oprea, HAIL: A high-availability and integrity layer for cloud storage, Proc. 16th ACM conference on Computer and communications security, (2009) 187-198.

[12] S. Eswaran, Dr. S. Abburu, Identifying Data Integirty in the Cloud Storage, International Journal Of Computer Science Issues, Vol.9 Issue 2, (2012), ISSN 1694-0814

[13] R. Sravan and Saxena ,"Data integrity proofs in cloud storage" in IEEE 2011.

[14] R. Pandya, K. Sutaria, "An analysis of privacy techniques for data integrity in the cloud environment", International Journal of Computer and Electronics Engineering,(Dec 2012) ISSN: 0975-4202