ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

A Comprehensive Survey on Data Integrity Proving Schemes in Cloud Storage

Patil Nikhil N[1], Mapari Rahul B[2]

M.Tech Scholar, Department of Computer Science & Engineering, Maharashtra Institute of Technology (MIT), Aurangabad, Maharashtra, India [1]

Assistant Professor, Department of Computer Science & Engineering, Maharashtra Institute of Technology (MIT), Aurangabad, Maharashtra, India [2]

Abstract: Cloud Computing is the technology which is gaining wide acceptance day by day .Cloud computing is a technology which provides different services like software as a service (Saas), Platform as a service (Paas) , Infrastructure as a service (Iaas).These services are provided through the internet. It is a challenge for computer user to store all the data that they have acquired because storage space is limited. So people prefer to buy hard disk and invest their money in it. Others use external devices like USB drives .But some prefer to store their data on cloud storage. There are many advantages of cloud storage such as easy scale up, easy scale down and many more. Along with advantages, it also brings some security related challenges. Clients move their data to remotely located server called cloud storage. But in cloud computing there is no provision to check the integrity of the data which is stored by the client. So there is need to overcome this challenge for efficient use of this technology. In this paper we are focusing on the basics of cloud storage, working of cloud storage system, various concepts related to data integrity like overview of data integrity, physical vs. logical integrity, Types of integrity constraints and survey on basic data integrity proving schemes. The basic data integrity proving schemes are Provable Data Possession (PDP) and Proof of Retrievability (PoR). On the basis of these schemes various techniques are proposed by different researchers. We also observe these techniques in this paper.

Keywords: Cloud Computing, Cloud Storage, Challenges, data Integrity, PDP, PoR

I. INTRODUCTION

Cloud computing is a technology where different services are hosted by third party also called as cloud service provider and cloud storage is a visualized pool. In cloud storage user's application and data are stored [1].The users or clients of the cloud service pay for the service that the clients get from the cloud service provider. The costing is on pay per use basis means based on the clients' usage. All the cloud services are provided through the internet. Due to rapid growth of the internet and advanced networking technology it has leads to a trend of data outsourcing and needs of information technology to external Service providers.

By outsourcing data organizations can concentrate on their core task rather than wasting their time on data management, hardware and software maintenance [2]. Due to improving network bandwidth and reliability it reduces the user's dependence on the local resources [3]. It also reduces the energy and labor costs and also complexity of the computing system is moved towards the centralized administration of the hardware [3]. Users who store their data and applications on cloud storage reside thousands of miles on machines. Users are totally unknown of the location of data stored [3]. For example, every internet user very well known the storage service of Google is Google Drive. It provides free storage service up to 1GB and for more storage space users have to subscribe the service for storage and have to pay for that. The leading cloud service providers are Amazon, Google, Microsoft etc .Amazon simple storage service is a cloud storage service provided by Amazon also called as Amazon S3 [4].

Storing of client's data to the cloud storage has many advantages but along with advantages it is facing many security concerns like data authentication and integrity [5]. Authentication and security ensure that cloud storage server returns correct and complete results in response of queries of client. It is needs to be investigated to make it reliable solution [5]. To solve the problem of data integrity checking various schemes are proposed on the basis of different systems and models.

In this paper we surveyed many integrity proving schemes along with different methods used in it. The rest of the paper is organized as follows: section 2: basics of cloud storage section 3: overview of data integrity, physical vs. logical integrity, various types of integrity constraints and section 4: review of data integrity proving schemes in which PDP scheme and POR scheme. We conclude our work in section 5.

II. BASICS OF CLOUD STORAGE

It is a challenge for computer user to store all the data that they have acquired because storage space is limited. So people prefer to buy hard disk and invest their money in it.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

Others use external devices like USB drives .But some prefer to store their data on cloud storage. Cloud storage refers [6] to storing data to remote location and it is maintained by third party. In this you store the data in remote location instead of storing it to your local hard disk. The data or information which is stored at remote location is the accessible through the internet [6].

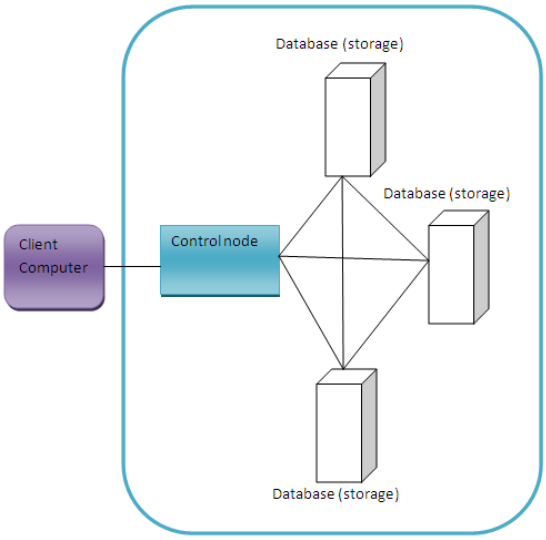

Fig 1: How cloud storage works [6]

Cloud storage has many advantages over traditional storage system. If you store data on the cloud storage you are able to extract it from anywhere you just require a internet connection. There is no need to carry your physical storage device and use of the same computer to save or retrieve information. It is convenient and offers more flexibility. Now the question arises how does it work?

A. Working Of Cloud Storage

There are number of various cloud storage systems. Some systems are designed for a specific reason such as storing information like multimedia, email messages and all forms of digital data [6]. At its basic level a cloud storage can consist of only one data server connected to the internet. A client which is a subscriber of cloud storage service sends data to the data centre over the internet [6]. When the client wishes to retrieve the information, he accesses the data server through the web interface [6]. Generally Cloud storage system relies on number of data servers. Data needs to be stored on multiple locations because computer requires maintenance occasionally [6]. This multiplication of information ensures that clients can access data at any time [6].

There are number of cloud storage service providers on the Web and the number .is increasing every day. There is lot of competition between various cloud service providers [6]. You are very well familiar with the several cloud service providers such as Google Docs, email service providers such as Gmail, Hotmail, and yahoo mail. Sites like Flickr and Picasa for hosting photographs [6]. You Tube for uploading videos. Social networking site such as Facebook, Twitter allow post, share photos and thoughts on it [6]. Services like Google Drive offers storage space for any kind of data. Some of these services are free and some are paid services [6]. The various concerns about cloud storage are reliability and security. To secure data various techniques are used like Encryption, Authorization and Authentication [6]. Encryption means encoding information by using complex algorithm [6]. By using encryption key the information is then decoded to original form. Authentication process requires creating username and password [6]. Only authorized clients can access the information stored on the cloud storage. The main concern of the cloud system is data integrity, which we will discuss next section.

III. OVERVIEW OF DATA INTEGRITY

As the word suggests itself data integrity means completeness or wholeness and it is basic requirement of information technology [7]. Data integrity refers to maintaining and assuring the accuracy and consistency of data over its entire life-cycle [8]. Data corruption is a form of data loss and data integrity is opposite of data corruption [8]. 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 [8]. Challenges for the physical integrity may include electromechanical faults, design flaws, material fatigue, corrosion, power outages, natural disasters, acts of war and terrorism [8]. Physical integrity makes use of error detecting algorithms known as error-correcting codes [8]. Logical integrity is related with the correctness or rationality of a piece of data [8]. Types of integrity constraints are as referential integrity, entity integrity and domain integrity [8].

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 [8]. The referential integrity rule states that any foreign-key value can only be in one of two states [8]. Domain integrity specifies that all columns in relational database must be declared upon a defined domain [8]. 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 [8]. Data Integrity is necessary in databases and it is also necessary in data Stored in the cloud [7]. Data integrity is a factor that affects on the performance of the cloud [7].

IV. DATA INTEGRITY PROVING SCHEMES

A. 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) [2].This model allows to check integrity of the data without retrieving it which is

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

stored by the client [2]. To reduce input and output cost this model generates the probabilistic proof of possession by sampling random sets of blocks [2].This PDP scheme does not include any error correcting code [2]. This scheme works in two phases' setup phase and challenge phase [2]. Ateniese et al. [2] developed first PDP scheme in which they considered public audiability in their model for ensuring possession of data on untrusted storage [11]. 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 [11]. Later Ateniese et al. [2] proposed dynamic version of PDP scheme but it does not support fully dynamic data operations [11]. This scheme offers only limited functionality and very basic blocks of operations [11]. Erway et al. [12] were who proposed a scheme for dynamic PDP. They extended the PDP scheme proposed by the Ateniese et al. [2].In this scheme rank based rank based authenticated skip lists are used to support dynamic data operations [11][12].This scheme is later improved by Feifei Liu[13]. This newly proposed scheme reduces the computational and communication complexity[11][13].

B. 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 [5]. 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 [5]. After calculating hash value the file is stored in the cloud storage [5]. 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 [5]. Then values calculated by the verifier and values calculated by the cloud storage are compared with each other [5]. From that comparison the final conclusion is considered [5]. 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 [5].

Another scheme for proof retrivability is using sentinels [3].This scheme is proposed by Ari Juels and Burton S. Kaliski Jr [3]. 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 [3]. 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[3][5]. 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 [3][5]. 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 [5].Also storage overhead will be there due to newly inserted sentinels and error correcting code [5].

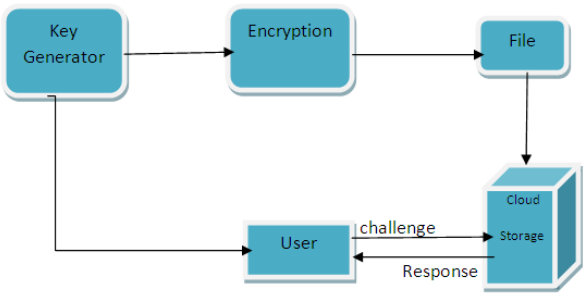

Fig 2: A Diagrammatic view of a proof of retrievability based on inserting random sentinels in the data files F[5]

Sravan Kumar R and Ashutosh Saxena present a scheme [5] 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 [5]. After that verifier attach some meta data to the file [5].This meta data is used at the time of verification of the integrity of the file [5]. The limitation of this scheme is this scheme applies only static data [5].

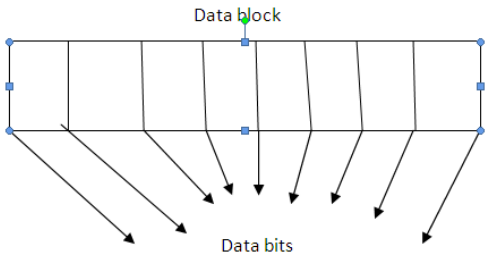

Fig 3: A block of data of file F with random bits selected in it [5]

Fig 4: The encrypted file eF which will be stored in the cloud. [5]

V. CONCLUSION

In this paper we confronted, overview of data integrity in which we studied physical vs. logical integrity, types of integrity constraints, objectives of data integrity and data integrity proving schemes that are PDP and PoR. 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

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

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. Later we surveyed another scheme that is Proof of retrievability PoR in this scheme we surveyed hash based technique , sentinel based technique and then technique base on selecting random blocks. The hash based technique and sentinel base technique has limitation of computational and storage overhead. And after that technique based on selecting random blocks. 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] Daliya Attas , Omar Batrafi "Efficient integrity checking technique for securing client data in cloud computing", IJECS –IJENS Vol:11 No:05,2011
[2] 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 security. New York, NY, USA: ACM, 2007, pp. 598–609.
[3] A. Juels and B.S. Kaliski Jr., "Pors: Proofs of Retrievability for Large Files," Proc. 14th ACM Conf. Computer and Comm. Security (CCS '07), pp. 584-597, 2007.
[4] Amazon.com. Amazon simple storage service (Amazon S3), 2007.
[5] Sravan Kumar R, Ashutosh Saxena," Data Integrity Proofs in Cloud Storage,"ISBN: 978-1-4244-8953-4/11/$26.00 c 2011 IEEE
[6] http://computer.howstuffworks.com/cloud-computing/cloud-storage.htm
[7] Satyakshma Rawat, Richa Chowdhary, Dr. Abhay Bansal," Data integrity of cloud data storage (CDSs) in cloud" ijarcsse Vol. 3, Issue 3, March 2013.
[8] Wikipedia site:http://en.wikipedia.org/wiki/Data_integrity.
[9] Saranya Eswaran, Dr. Sunitha Abburu "Identifying Data integrity in cloud storage", IJCSI Vol. 9, Issue 2, No 1, March 2012.
[10] E. Mykletun, M. Narasimha, and G. Tsudik, "Authentication and integrity in outsourced databases," Trans. Storage, vol. 2, no. 2, pp. 107–138, 2006.
[11] T S Khatri, Prof G B Jethava," Survey on Integrity Approaches used in the Cloud Computing", IJERT Vol. 1 Issue 9, November-2012.
[12] C. Erway, A. Kupcu, C. Papamanthou, and R. Tamassia,"Dynamic Provable Data Possession," Proc. 16th ACM Conf. Computer and Comm. Security (CCS '09), 2009.
[13] Feifei Liu, Dawu Gu, Haining Lu," An Improved Dynamic Provable Data Possession Model," 978-1-61284-204-2/11/$26.00 ©2011 IEEE.

BIOGRAPHY

Patil Nikhil N. Pursuing M. Tech From MIT, Aurangabad. He is Redhat certified expertise in cloud storage and openstack administration. He has accomplished B.E, (Computer Engineering) from SGDCOE, Jalgaon.