Data Integrity Issues in Cloud Servers

Arsalan Iqbal[1], Hina Saham[2]

[1]Computer Networks Program
Ryerson University, ON, Canada

[2]Department of Computer Science
Quaid-e-Azam University, Islamabad, Pak

Abstract

In the past few years cloud has become the buzzword in computing. But this wide acceptance and ease of use exposed the new IT based technology into a number of data integrity (correctness of data) and security issues. Integrity of user data in the cloud servers is one of the most important concerns of users nowadays. In this paper we will analyze different methodologies and protocols, which the customer/users can use to check the correctness of their data with the simplest possible way and less overhead at the customer side and to overcome the challenges faced by cloud servers for the security and integrity of users data.

Keywords: *Cloud Computing, Data Integrity, Data Availability, Security, Cloud Servers, Users, Encryption.*

1. Introduction

Data outsourcing [1] 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. Many problems like data authentication and integrity (i.e., how to efficiently and securely ensure that the cloud storage server returns correct and complete results in response to its clients' queries [2]), outsourcing encrypted data and associated difficult problems dealing with querying over encrypted domain [3] were discussed in research literature.

2. Cloud Computing Services

Cloud computing is an internet based computing. It dynamically delivers everything as a service over the internet based on user demand, such as network, operating system, storage, hardware, software, and resources. These services are classified into three types: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). Cloud computing is deployed as three models such as Public, Private, and Hybrid clouds [4].

Cloud data storage (Storage as a Service) is an important service of cloud computing referred as Infrastructure as a Service (IaaS). Data storage in cloud offers so many benefits to users:

1) It provides unlimited data storage space for storing user's data.

2) Users can access the data from the cloud provider via internet anywhere in the world not on a single machine.

3) We do not buy any storage device for storing our data and have no responsibility for local machines to maintain data.

Amazon's Elastic Compute Cloud (EC^2) and Amazon Simple Storage Service (S3) ([5], [6]) are well known examples of cloud data storage. On the other side along with these benefits' cloud computing faces big challenge i.e. data storage security problem, which is an important aspect of Quality of Service (QoS). Once user puts data on the cloud rather than locally, he has no control over it i.e. unauthorized users could modify user's data or destroy it and even cloud server collude attacks. Cloud users are mostly worried about the security and reliability of their data in the cloud. Amazon's S3 is such a good example [5].

3. System Model

The network representative architecture[7] for cloud data storage, which contains three parts as shown in Figure 1, viz Users, Cloud Service Provider (CSP) and Third Party Auditor (TPA).

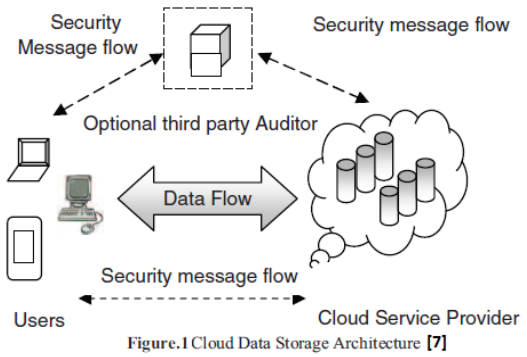

Figure.1 Cloud Data Storage Architecture **[7]**

As shown in Figure 1, the brief descriptions of these parts as follows:

Users: - Users who have data to be stored and interact with the cloud service provider (CSP) to manage their data on the cloud. They are typically, desktop computers, laptops, tablet computers, mobile phones, etc.

Cloud Service Provider (CSP):- Cloud service provider (CSP) has major resources and expertise in building and managing distributed cloud storage servers. A CSP offers storage or software services to user's available via the Internet.

Third Parity Auditor (TPA):- An optional TPA, who has expertise and capabilities that users may not have, is monitoring the risk of cloud data storage services on behalf of users.

4. Data Integrity and its POR

4.1 Data Integrity

In terms of a database data integrity [8] refers to the process of ensuring that a database remains an accurate reflection of the universe of discourse it is modeling or representing. In other words there is a close correspondence between the facts stored in the database and the real world it models.

4.2 Proof of Retrievability (POR)

Proof of Retrievability (POR) mechanism [1] tries to obtain and verify a proof that the data that is stored by a user at remote data storage in the cloud (called cloud storage archives or simply archives) is not modified by the archive and thereby the integrity of the data is assured.

The simplest Proof of retrievability (POR) scheme can be made using a keyed hash function $h_k(F)$. In this scheme the verifier, before archiving the data file F in the cloud storage, pre-computes the cryptographic hash of F using $h_k(F)$ and stores this hash as well as the secret key K.To check if the integrity of the file F is lost the verifier releases the secret key K to the cloud archive and asks it to compute and return the value of $h_k(F)$[1].

Such kinds of proofs are very much helpful in peer-to-peer storage systems, network file systems, long term archives, web-service object stores, and database systems. 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. Cheating, in this context, means that the storage archive might delete some of the data or may modify some of the data. It must be noted that the storage server might not be malicious; instead, it might be simply unreliable and lose or inadvertently corrupt the hosted data. But the data integrity schemes that are to be developed need to be equally applicable for malicious as well as unreliable cloud storage servers. Any such proofs of data possession schemes do not, by itself, protect the data from corruption by the archive. It just allows detection of tampering or deletion of a remotely located file at an unreliable cloud storage server. To ensure file robustness other kind of techniques like data redundancy across multiple systems can be maintained.

5. Work done to obtain POR

In our survey we will divide the work done by different researchers in two categories: data placed on a single server and data stored on multiple servers.

5.1 Data placed on single server at cloud

Ari Juels and Burton S. Kaliski Jr [9] proposed a scheme called Proof of retrievability for large files using 'sentinels'. A single key can be used irrespective of the size of the file. The archive needs to access only a small portion of the file F. Special blocks (called sentinels) are hidden among other blocks in the data file F. In the setup phase, the verifier randomly embeds these sentinels among the data blocks. During the verification phase, to check the integrity of the data file F, the verifier challenges the prover (cloud archive) by specifying the positions of a collection of sentinels and asking the prover to return the associated sentinel values as shown in fig 2.

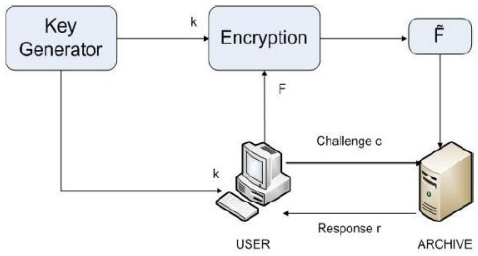

Fig. 2. Schematic view of a proof of retrievability based on inserting random sentinels in the data file F [9]

Ateniese et al. described a Provable Data Possession (PDP) to verify the integrity of outsourced data; it detects the large fraction of file corruption, but no guaranty of file retrievability [7].

In their subsequent work R.D.Pietro et al. proposed a Scalable Data Possession (SDP), this scheme overcomes all problems in the PDP scheme.

Shacham et al. introduced a new model of POR, which enables unlimited no of queries for public verifiability with less overhead.

Kennadi D et al. proposed a theoretical framework for the design of POR.

In the scheme proposed by S Kumar and A Saxena [1] the data is not encrypted as a whole. Instead, they encrypt only few bits of data per data block thus reducing the computational overhead on the clients. Client storage overhead is also minimized as it does not store any data with itself. In this data integrity protocol the verifier needs to store only a single cryptographic key - irrespective of the size of the data file F. The verifier does not store any data with it, as shown in fig 3 and fig 4.

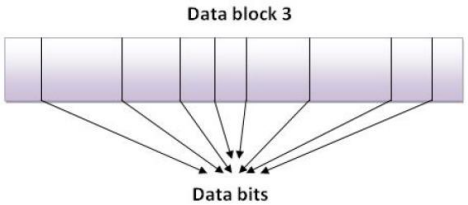

Fig. 3. A data block of the file F with random bits selected in it [1]

Fig. 4. The encrypted file $\widetilde{F}$ which will be stored in the cloud. [1]

5.2 Data placed on multiple servers

All the schemes discussed above produce weak security, because they work only for single server.

Later, in their subsequent work Kennadi Brow et al. Introduced a HAIL protocol, which extended the POR schemes on multiple servers. HAIL achieves the integrity and availability of data in cloud [7].

Curtomola et al. [10] described a Multiple Replica-Provable Data Possession (MR-PDP), which is an extension of PDP to ensure data availability and reliability of outsourced data on multiple servers.

Shah et al. proposed [11] a new scheme, which allows Third Party Auditor (TPA) to keep on-line storage honesty with hash values computed by user on encrypted file. However, this scheme works only for encrypted files.

The recent work by Wang et al. described [12] a homomorpic distributed verification scheme using Pseudorandom Data to verify the storage correctness of user data in cloud. This scheme achieves the guaranty of data availability, reliability, and integrity. However, this scheme was also not providing full protection to user data in cloud computing, since pseudorandom data would not cover the entire data.

Based on the work of Wang [7] proposed a distributed verification protocol to guaranty the data storage security in cloud computing. This scheme uses Reed-Solomon erasure code for the availability and reliability of data and utilizes the token pre-computation using Sobol Sequence rather than pseudorandom data to check the integrity of erasure coded data in cloud data storage. A comparison of Pseudorandom data and sobol sequence is shown in figure 5.

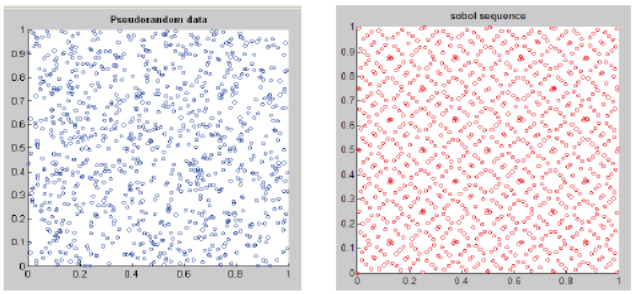

Fig 5 Comparison of Pseudorandom Data and Sobol Sequence.

6. Comparison

In this survey we provide a comparison of several POR protocols. Different factors are considered like single servers, multi-servers, entire data protection, TPA requirement, thin users, encryption of data, etc. as shown in table 1.

Table 1: Comparison Analysis of Different POR schemes

Methodology	Single Server	Multi Server	Require a TPA (Third Party Auditor)	POR (Proof of retrievability)	Encrypted	Thin Users	Entire data
Simplest POR	Yes	No	No	Yes	Yes	No	Yes
POR using sentinels	Yes	No	No	Yes	Yes	No	Yes
PDP	Yes	No	No	No	NS	No	NS
SDP	Yes	No	No	Yes	NS	No	NS
Kumar & Saxina Proposed model	Yes	No	No	Yes	Yes(Partial)	Yes	Yes
Shacham	Yes	No	No	Yes	NS	Yes	NS
Kennadi's HAIL protocol	Yes	Yes	No(Optional)	Yes	Yes	No	Yes
MR-PDP	Yes	Yes	No(Optional)	Yes	NS	No	NS
Shah	Yes	Yes	Yes	Yes	Yes	Maybe	Yes
Wang	Yes	Yes	No(Optional)	Yes	No	No	No
Sobol Sequence	Yes	Yes	No(Optional)	Yes	No	No	Yes

7. Conclusion

Issues related to data integrity on clouds servers need to be addressed properly. In this paper we tried to cover some of the important aspects that need to be considered while storing users data on cloud servers. If the servers are secure and reliable enough to store the data properly, then the simple scheme proposed by Kumar and Saxina, is a good choice. Byzantine failure is one of the main reasons of corrupting users data. Due to this failure the servers begin to behave improperly. To solve the issue, nowadays, data is distributed on multiple servers in the cloud for availability and reliability. So that if one server fails to respond then data is available on other servers to respond to users queries/requests. In this case, if the availability, reliability, security and integrity are all the factors to be considered then the scheme proposed by [7] using Sobol sequence is the best choice for customers.

Acknowledgments

I would like to thank my Professor, Dr. Ngok-Wah, for the patient guidance, encouragement and advice he has provided throughout my course as his student. I have been extremely lucky to have a professor who cared so much about my work, and who responded to my questions and queries so promptly.

References

[1] Sravan Kumar, R and Saxena, A. "Data Integrity Proofs in Cloud Storage", Third International conference on Communication Systems and Networks (COMSNETS), pp 1-4, IEEE-2011

[2] E. Mykletun, M. Narasimha, and G. Tsudik, "Authentication and integrity in outsourced databases," Trans. Storage, vol. 2, no. 2, pp. 107–138, 2006.

[3] 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, 2000, p. 44.

[4] Cloud Computing FOR DUMMIES by Judith Hurwitz, Robin Bloor, Marcia Kaufman, and Fern Halper. WILEY INDIA EDITION.

[5] Amazon.com, "Amazon Web Services (AWS)," Online at http://aws.amazon.com, 2008.

Cong Wang,Qian wang and Kui Ren and Wenjing Lou, "Ensuring Data Storage Security in Cloud Computing , Quality of Service, 2009, IWQoS IEEE 17th International workshop ,pp 1-9,2009.

[6] P. Syam Kumar, R. Subramanian and D. Thamizh Selvam," Ensuring Data Storage Security in Cloud Computing using Sobol Sequence" 1st International Conference on Parallel, Distributed and Grid Computing, pp 217-222, IEEE-2011.

[7] Ran Canetti, Oded Goldreich and Shai Halevi, The Random Oracle Methodology Revisited, STOC 1998, pp. 209-218pp. 529-551, April 1955

[8] 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.

[9] R. Curtmola, O. Khan, R. Burns, and G. Ateniese, "MR-PDP: Multiple-Replica Provable Data Possession," Proc. of ICDCS '08, pp. 411–420,2008

[10] M. A. Shah, M. Baker, J. C. Mogul, and R. Swaminathan, "Auditing to Keep Online Storage Services Honest," Proc. 11th USENIX Workshop on Hot Topicsin Operating Systems (HOTOS'07),pp.1–6, 2007

[11] Cong Wang,Qian wang and Kui Ren and Wenjing Lou, "Ensuring Data Storage Security in Cloud Computing ,Quality of Service, 2009, IWQoS IEEE 17th International workshop ,pp 1-9

[12] M. A. Shah, M. Baker, J. C. Mogul, and R. Swaminathan, "Auditing to Keep Online Storage Services Honest," Proc. 11th USENIX Workshop on Hot Topicsin Operating Systems (HOTOS'07),pp.1–6, 2007

Arsalan Iqbal received his B.Sc degree in Computer Science from University of Peshawar, Pakistan in 2003. He received his Masters degree in Computer Science with distinction from University of Peshawar, Pakistan in 2006. He worked as a lecturer in Department of Computer science, King Khalid University, Kingdom of Saudi Arabia from 2007 till 2012. He is now a graduate student in Computer Networks Program, Ryerson University, ON, Canada. His current research interests include Cloud Computing, Virtualization and data integrity issues on Cloud servers.

Hina Saham received her B.Sc degree in Computer Science from University of Peshawar, Pakistan in 2004. She received her Masters degree in Computer Science from Quaid-e-Azam University, Pakistan in 2007. She served as a lecturer and assistant department coordinator in Department of Information Systems, Community College for girls, Abha, King Khalid University, Kingdom of Saudi Arabia. Her research area spans around operating systems, information security, routing protocols and MPLS.