ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(40), DOI: 10.17485/ijst/2016/v9i40/98622, October 2016* ISSN (Online) : 0974-5645

A Comprehensive Study on Big Data Security and Integrity Over Cloud Storage

J. Raja[1] and M. Ramakrishnan[2]

[1]Department of Computer Science and Engineering, Sathyabama University, Rajiv Gandhi Salai, Jeppiaar Nagar, Chennai - 600119, Tamil Nadu, India; rajdharshun@gmail.com
[2]School of Information Technology, Madurai Kamaraj University, Palkalai Nagar, Madurai - 625021, Tamil Nadu, India; ramkrishod@gmail.com

Abstract

In this paper, a study on cloud storage security issues and challenges such as authentication, access control, policy integration, service management, trust management, data security, regulatory compliance, privileged user access, etc are presented. In the first half of the paper, a survey on generic security issues in cloud storage is presented which suggests that third party auditor based security is best for cloud environments. In the second half, data security issues specific to Big Data is presented, which can be achieved again by using third party authentication process.

Keywords: Big Data, Cloud Storage, IaaS, PaaS, Public Auditing, SaaS, Security and Integrity, Security Issues

1. Introduction

Cloud computing is dynamically scalable and resources are provided as services over the internet[1]. It is a service model based on internet which provides an infrastructure to share collection of resources such as applications, storage servers, data, software and hardware with low cost. Cloud environments provide blending of virtual techniques in an efficient way on the minute[1].

Cloud computing provides many services on demand and all these requests are broadly grouped into three important service models. They are

- Software as a Service (SaaS)
- Platform as a Service(PaaS)
- Infrastructure as a Service (IaaS)

SaaS is the top layer of cloud architecture designed to distribute software applications. By using web browsers, these software applications can be accessed without installing them in the host machines. With this setup, the user is free from installation and maintenance of softwares[2].

PaaS allows users to build their own software applications. PaaS provides program development tools, frameworks, templates, platforms and a container to run user's components. All the above services are offered to users with low cost over the internet. Example of PaaS is Google App Engine which allows users to develop software programs in Python, Java and Go. Another example is Apprenda which provides users with. NET programming environments. PaaS can be extended in order to provide customers ready features[2].

The bottom layer, IaaS, provides infrastructure (computing resource) services to the customers[3]. This includes virtual computers, firewalls, intrusion detection systems, hardwares, network devices, etc. Apart from these three dominant service models, we have other ones also such as Storage as a Service (STaaS), Data as a Service (DaaS), Security as a Service (SECaaS)[3].

Though the primary objective of cloud computing is to provide efficient service over internet, there are several security issues pertaining to storage of data in cloud. Data storage in cloud should be protected from unauthorized access, modification and deletion thereby ensuring data integrity[4]. This survey paper presents a review on various data storage security issues and available solutions. The paper has been organized as follows: Introduction to cloud computing and data storage in cloud is discussed in section 1 and section 2 respectively. Section 3 deals with generic data security techniques used by cloud storage providers. Section 4 deals with storing Big Data

* *Author for correspondence*

in cloud and section 5 provides a study on security issued of Big Data under cloud environment. The paper ends by narrating the findings in section 6 as conclusion.

2. Data Storage in Cloud

Security breaches in cloud computing seriously threaten the trust between customer and service provider. Cloud service provider has to ensure security issues such as authentication of users, correctness of data, availability of the data, non-leakage of data, maintenance of data, avoiding loss of data, etc[5].

User's data are stored in the Cloud Service Provider (CSP) set of servers that are running in a distributed and concurrent fashion. Ensuring data integrity and confidentiality are important[6]. Few general techniques adopted to ensure integrity and confidently of the data stored at the CSP that are:

- Ensure limited access to the users' data by the CSP employees.
- Strong authentication mechanisms to ensure that only legitimate employees gain access and control CSP servers.
- The CSP should use well defined Data backup and redundant data storage to make data recovery possible.

3.1 Third Party Auditor (TPA)

Secure Socket Layer, Point to Point Tunneling Protocol, Virtual Private Networks are the general concepts used by cloud service providers to ensure security[7]. But malicious users and attackers have gained unauthorized access over cloud stored data. Hence third party authentication is used. This mechanism is implemented in both the ends; user and cloud service provider[7].

TPA works by using Service Level Agreement, a legal understanding between client and cloud service provider. TPA monitors both ends of data transmission and it follows auditing norms and techniques designed for it[8]. TPA periodically audits the user data in cloud without affecting the integrity of the users.

Auditing is done in three phases viz. planning, execution and reporting. In the first phase, planning the audit is carried out in which details of audit such as auditing content, schedule of audit, duration of audit, area of audit, audit team size and members, etc are finalized. All these activities are carried out in execution phase.

It analyses the security threats to cloud storage analyses the security threats to cloud storage, examines previous threats and valuating the level of data integrity[9]. All the observations are reported to both users and cloud service providers in the final phase. Based on the report, cloud service provider can get details of user activities as well as their performance. Malicious users will be removed from their service. This report also helps improve cloud service provider to improve their performance.

3.2 Encryption Based Storing of Data

In this type of security provision, cryptographic techniques are used. Cloud storage servers encode and forward messages. Key servers individually perform partial encryption. Any encryption standards can be used like RSA, DSA, etc. To ensure security all the data owners are provided with access key and using the key, they can decrypt the data during data retrieval from cloud[10].

The new users have to register with cloud service provider. The data owner computes a message using D5 algorithm and send it to new user. The new user can access the data using this key.

3.3 Privacy Preserving Public Auditing

This method uses the concepts of TPA. It audits the data without getting a copy. Homomorphic authenticators are used which are securely aggregated to ensure an auditor that liner combination of data blocks are correctly computed. Along with homomorphic authenticators, random mask technique is used.

This method uses combination of four important algorithms and each one is having specific impact in maintaining data storage security. Keygen is an algorithm which generates key that is used by the user. Singen generates verification metadata that includes digital signature. Genproof algorithm generates proof that the data stored is intact and secured, and this proof is supplied to cloud storage provider. Verify proof is the final algorithm that verifies the proof generated by Genproof algorithm.

This method works in two phases viz. setup phase and audit phase. Using keygen algorithm, public and secret parameters are initialized, data files are preprocessed and digital signature which is mean for meta data are done in setup phase[11]. Setup phase also allows to delete the local copy of data and alternation of data files. In audit phase, reports pertaining to the audit conducted are generated

and supplied to cloud service provider. TPA verifies the response generated by Genproof algorithm.

3.4 Secure and Dependable Storage

Error localization is the real need for eliminating problems in data storage in cloud. This method integrates the correctness verification and error localization to ensure data security. This method uses homomorphic token with distributed verification of erasure-coded data. This method easily locates the error and identifies the misbehaving servers[12]. It also allows secure and efficient operations on data blocks, data update and append. The problem with this method is that it monitors only one server at a time and it does not provide data availability guarantee against server failures.

3.5 Non Linear Authentication

It is also a security technique for cloud storage which uses homomorphic non linear authenticator with random masking. In order to encrypt and decrypt the data, RSA algorithm is used which follows digital signatures for authentication. Slight modification of this method, called Extensible Authentication Protocol, uses three way handshaking scheme along with RSA[13]. This version of protocol is more light weight and efficient compared to other third party authentication protocols.

Client first requests for a service with cloud service provider and service provider authenticator sends handshaking challenge to the client as request. For this request, client sends response which is calculated hash function. Service provider authenticator verifies the response and calculates the value. If both the value matches, it sends authentication rights to the client. This setup provides security against IP spoofing attacks, data tampering and DOS attacks.

3.1 Comparison of Various Data Storage Techniques

Methods	Advantages	Disadvantages
Third Party Authentication	• Auditing job is performed simultaneously for different users. • Applied in both the ends; user and cloud service providers.	• Does not support both dynamic data correctness and public verification.
Encryption based techniques	• Data are coded that makes difficult for the intruders to understand the data • Any encryption standard can be used.	• Encoding and decoding of data adds additional overhead for end users. • Computationally this method is costly.
Privacy Preserving Public Auditing	• Homomorphic authenticators securely aggregates the data blocks. • Four sub algorithms viz. keygen, Singen, GenProof and VerifyProof are used which adds security to high level	• Dependent on the basic concepts of TPA
Secure and Dependable Storage	• This method easily locates the error and also identifies the misbehaving nodes in cloud. • Allows secure and efficient operations on data blocks.	• Adds gross overhead to the nodes. • It monitors only one server at a time and fails to provide data availability guarantee against server failures.
Non-Linear Authentication	• Using RSA algorithm, it encrypts and decrypts the data that adds more security. • Scheme is more light weight and efficient. • Secure against IP spoofing attacks, data tampering and DOS attacks	• Three way handshaking is used which delays the response time from both the parties.
File Assured Deletion	• Supports dynamic data operations. • Storage of meta data is less.	• Lot of memory is needed to implement this scheme.
Optimal Cloud Storage	• Used as a generic architecture for optimal storage controller. NubiSave is freely available	• It needs interaction with frontends for future research.

4. Cloud and Big Data

Big data architecture supports high velocity data capture, storage and analysis. Big data requires huge amount of storage. Data in Big data may be in unstructured format, without standard formatting, and data sources can be beyond the traditional corporate database.

Storing small and medium sized business organization's data in cloud as Big Data is a better option for data analysis work. An in-house storage model used to store Big Data in Network-Attached Storage (NAS). The architecture of NAS constitutes several computers attached to each other[14]. Clustered NAS storage is not feasible for small and medium size business.

The Big Data stored in cloud can be analyzed using a programming methodology called MapReduce in which query is passed and data are fetched. The extracted query results are then reduced to the data set relevant to query. This query processing is simultaneously done using NAS devices. Though MapReduce algorithm usage in Big Data is well appreciated by many researchers as it is schema free and index free, it requires parsing of each record at reading point[15]. This is the biggest disadvantage of MapReduce algorithm usage for query processing in cloud computing.

A feature integrated framework that combines the merits of MapReduce algorithm and DBMS was proposed called Hadoop, which substantially increases the task processing time in cloud environment[16]. Another technique which combines the good features of Local Sensitive Hashing (LSH) and MapReduce called Rank Reduce, which performs k-nearest neighbours search for high dimensional spaces.

4.1 Big Data Optimization

Since Big Data deals with huge amount of data, it is necessary to store the data in cloud with more efficient and cost effective way. It is necessary to optimize the data to be stored in cloud with lower price. In this section, we will discuss about the data optimization tools used in cloud storage.

Deduplication is the process which removes duplicate copies of redundant data. Also known as single-instance storage, deduplication tool analyzes the data to track and store unique chunks of data[17]. It repeatedly compare new chunk of data with stored one, and if a match is found, redundant data is discarded. It is observed by researchers that a same pattern of data may occur many times and the smaller the storage block size, greater the ability to deduplicate the data. Cloud storage environments are generally the targets of duplication and this method removes the redundant data thereby enhancing the cloud storage efficiency.

Data compression is a familiar method to achieve efficiency in storing data in any environment and for cloud also. It encodes the data into few bits than the original content. Categorized as lossy and lossless compression, cloud environment prefers lossless compression to achieve efficient storage.

A virtualization concept, called thin provisioning, can also be used for optimized storage in cloud[18]. It differs from data reduction technique and it is a type of resource provisioning. It creates virtual appearance of physical resources than available. Storage can be created and adjusted based on demand. It is applied to large scale disk-storage and storage virtualization systems[18].

Database optimization is another technique used in clouds for storage efficiency. This is done by splitting the database into logical pieces called partitioning. It can be vertical partitioning where data columns are splitted into separate pieces and then store it in other database tables. Horizontal partitioning splits different rows into different tables. It is generally meant as partition per database schema. If multiple rows are removed, it is termed as sharding which reduces index size to improve search performance[19]. The demerit of sharding is that it heavily relies on server interconnections and inability to provide consistency among data.

Another method to achieve optimized storage in clouds is to use specific software[20]. It is termed as software defined storage and it is a better method for cloud storage issues. It abstracts logical storage services from physical storage systems, with multiple implementations that are different from each other. It uses storage virtualization, virtual volumes, API for storage interaction and NFS[20].

4.2 Big Data Management in Cloud

The cost and scalability of the database server is high and hence it cannot be used for Big Data processing. One option is to use classic multi-tier database application architecture for processing Big Data[21]. In general, different business models are used for different applications of Big Data. Majority of the cloud service providers are following hybrid architecture for handling Big Data storage.

Distributed File System architecture that supports fault-tolerance by data partitioning and replications. Google's cloud computing platform and Hadoop are utilizing this distributed file system[22]. Web data sets are generally semi-structured and it cannot satisfy big service providers. Big table is used as distributed storage system that can scale to very large amount of data. Bit table provides clients a data model that supports dynamic control over data layout and format[22].

4.3 Challenges in Big Data Management

There are many challenges in managing Big Data in cloud. We need to provide mechanisms to handle every increasing volume of data, data that are unstructured. These varieties of data need to be extracted quickly along with the provisions of aggregating and correlating data if they are from multiple sources.

How to organize, store and extract unstructured data is a biggest challenge in cloud environments. Since we have large volumes of data, timely retrieval of data is expected. Protocols and interfaces are needed for integrating data of different nature viz. structured, semi-structured and unstructured from different sources.

To manage resources and efficient data processing, new programming methodologies and paradigms are needed with improved backend engines to manage optimized file system architecture[23].

5. Securing Big Data in Cloud

There are several methods that can be employed to secure big data in cloud environments. In this section, we will examine few methods.

Source Validation and Filtering - Data are coming from different sources, with different formats and vendors. The storage authority should verify and validate the source before storing the data in cloud storage. The data are filtered in the entry point itself so that security can be maintained.

Application Software Security - The primary concern of Big Data is to store huge volume of data and not about security. Hence it is advisable to use originally secure versions of software to access data. Though open source software and freewares might be cheap, it may result in security breaches.

Access Control and Authentication - The cloud storage provider must implement secure access control and authentication mechanisms. It has to provide several request's of the user's with their roles. That difficulty in imposing these mechanisms is that requests may be from different locations. Few secure cloud service providers give authentication and access control only on registered IP addresses thereby ensuring security vulnerabilities[24]. Securing privileged user access requires well-defined security controls and policies.

Secure Data Management - From the requirements phase, security issues should be addressed and based on this a business model should be selected. The applications should be built in-house completely. The infrastructure under which the Big Data is stored need to be secure and this is achieved by imposing strong network firewalls, security devices and traffic monitoring devices.

6. Conclusion

A study on data storage security issues is presented and discussions are made on general methods that are adopted to ensure security in clouds storage viz. third party auditor, encryption based security mechanisms, privacy preserving public auditing, non-linear authentication, etc. Third party authentication mechanisms simultaneously performs auditing job for different users but fails to adapt dynamic data correctness. Both encryption based techniques and secure dependable storage methods add computational overhead to the nodes. On the other hand, memory requirement is high in file assured deletion method but it supports dynamic data operations. Non linear authentication method offers various advantages and provides security against various attacks compared to other methods. But whole procedure is delayed because of three way hand shaking process. Overall, we endure that cloud storage security is still under development and more enhancements are needed in near future. Though cloud systems provide few secure storage mechanisms, it is yet to become the accepted model for implementation. The paper also presents security issues related to storing Big Data in clouds.

7. References

1. Durairaj M and Manimaran A. A study on security issues in cloud based e-learning. Indian Journal of Science and Technology. 2015; 8(8):757-65.

2. Verma A, Kaur I and Arora N. Comparative Analysis of Information Extraction Techniques for Data Mining. Indian Journal of Science and Technology. 2016; 9(11).
3. Paul Pocatilu, Boja Catalin and Ciurea Cristian. Syncing Mobile Applications with Cloud Storage Services. Informatica Economica. 2013; 17.2:96.
4. Dimitrios Zissis and Dimitrios Lekkas. Addressing cloud computing security issues. Future Generation computer systems 2012; 28.3:583-92.
5. Srinivasan S. Is security realistic in cloud computing? Journal of International Technology and Information Management. 2013; 22.4:3.
6. Irfan Gul and Hussain M. Distributed cloud intrusion detection model. International Journal of Advanced Science and Technology. 2011; 34:71-82.
7. Hassan Takabi, Joshi James BD and Ahn Gail-Joon. Security and privacy challenges in cloud computing environments. IEEE Security & Privacy. 2010; 6:24-31.
8. Youssef Ahmed E and Alageel Manal. A Framework for A Framework for Secure Cloud ure Cloud ure Cloud Computing Computing Computing. 2012.
9. Mather Tim, Kumaraswamy Subra and Latif Shahed. O'Reilly Media, Inc.: Cloud security and privacy: an enterprise perspective on risks and compliance. 2009.
10. Jambhekar ND, Misra S and Dhawale CA. Cloud Computing Security with Collaborating Encryption. Indian Journal of Science and Technology. 2016; 9(21).
11. Wang Cong, et al. Privacy-Preserving Public Auditing for Secure Cloud Storage. IEEE Transactions On Computers. 2013; 62.2.
12. Wang Cong, et al. Toward Secure and Dependable Storage Services in Cloud Computing. IEEE Transactions on Services Computing. 2012; 2.5:220-32.
13. Maheswari MI, Revathy S and Tamilarasi R. Secure Data Transmission For Multi sharing in Big Data Storage. Indian Journal of Science and Technology. 2016; 9(21).
14. Melekhova Anna and Vinnikov Vladimir. Cloud and Grid Part I: Difference and Convergence. Indian Journal of Science and Technology. 2015; 8.29.
15. Liu Chao, et al. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce. ACM, Proceedings of the 19th International conference on World Wide Web. 2010.
16. Doulkeridis Christos and Norvag Kjetil. A survey of large-scale analytical query processing in MapReduce. The VLDB Journal. 2014; 23.3:355-80.
17. Meyer Dutch T and Bolosky William J. A study of practical deduplication. ACM Transactions on Storage (TOS). 2012; 7.4:14.
18. Mishra Ratan and Jaiswal Anant. Ant colony optimization: A solution of load balancing in cloud. International Journal of Web & Semantic Technology. 2012; 3.2:33.
19. Wu Sai, et al. Efficient B-tree based indexing for cloud data processing. Proceedings of the VLDB Endowment. 2010; 3.1-2:1207-18.
20. Jensen Meiko, et al. On technical security issues in cloud computing. 2009 IEEE International Conference on Cloud Computing, CLOUD'09. IEEE, 2009.
21. Kossmann Donald, Kraska Tim and Loesing Simon. An evaluation of alternative architectures for transaction processing in the cloud. ACM, Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. 2010.
22. Zhang Qi, Cheng Lu and Boutaba Raouf. Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications. 2010; 1.1:7-18.
23. Thiriveni GV and Ramakrishnan M. Distributed Clustering based Energy Efficient Routing Algorithm for Heterogeneous Wireless Sensor Networks. Indian Journal of Science and Technology. 2016; 9(3).
24. Gollmann Dieter. Computer security. Wiley Interdisciplinary Reviews: Computational Statistics. 2010; 2.5:544-54.