ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

A Review on Remote Data Auditing in Cloud Computing

[1] Arjun U, [2] Vinay S
[1] Asst. Professor, [2] Professor
[1] Dept. of ISE, PESITM Shivamogga, [2] Dept. of CSE, PESCE, Mandya

Abstract: Cloud computing has emerged as a computational paradigm and also as an alternative to the conventional computing. Cloud computing aims at providing reliable, resilient infrastructure with the high quality of services for cloud users in both academic and business environments. The outsourced data in the cloud and the computation results are not always trustworthy because of the insufficiency in physical possession and control over the data for data owners. Ever since security protection threats to outsourced data have become an exigent task in cloud computing. Many researchers have focused on refining this problem and enabling public auditability for cloud data storage security using remote data auditing techniques. This paper presents a survey on the remote data storage auditing and presents remote data auditing approaches. The intent of this paper is to highlight issues and challenges of RDA protocols in the cloud and the mobile cloud computing. The state-of-the-art RDA approaches are also analysed and classified into two groups of provable data possession, proof of retrievability.

Index Terms —Cloud computing, Auditing, third party audit

INTRODUCTION

According to the NIST cloud computing is defined as "A model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (network, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort"[1].

The main idea of cloud computing is to outsource the management and delivery of software and hardware resources to third-party companies (cloud providers), which specialize in that particular service and can provide much better quality of service at lower costs in a convenient fashion. For example, now an enterprise can purchase the access of hardware resources according to its actual demands and without upfront costs. If the demand decreases, the enterprise can decrease the amount of remote hardware resources for which it is paying. If demand increases, the enterprise can easily adjust the resources to the demand.

Even though cloud computing offers several advantages for users, there are some security concerns that prevent a full adoption of the new technology. Hence, an autonomous reviewing and auditing facility is necessary to guarantee that the information is effectively accommodated and used in the cloud.

II. BACKGROUND

In this section the concepts of cloud distributed storage systems and the mechanism of the RDA are described.

A. Distributed Storage Systems

Distributed storage system is "storing data on multitude of standard servers, which behave as one storage system although data is distributed between these servers". They are created to allow users to remotely store data and provide services, such as archiving, publishing, federation, and anonymity. The main reason behind using distributed storage system is that the current approach to storage does not work anymore as it is not flexible enough, fast enough or due to its high cost [4].

B. Remote Data Auditing

Remote data auditing is an important &useful technique for auditing the responsibility & integrity of external sources of data to a single server or distributed servers. This kind of assurance is essential to ensure long-term reliability of data outsourced at data centers or at cloud storage providers (CSPs). The remote data auditing service contains a set of protocols to prove the intactness of the remote data that resides in cloud storage more reliably and efficiently, devoid of downloading the entire data. Furthermore, the outsourced data is also subject to administration by unreliable third-party cloud providers. The RDA frameworks use a spot-checking technique to validate the outsourced data, in which only a small

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

fragment of the whole data is required to be accessed by the auditor.

The RDA methods are applicable for both single and distributed cloud servers. In the single cloud server, algorithms are responsible only for preventing unauthorized parties from altering the outsourced data but when data corruption is detected, a majority of the single-server RDA techniques do not have the required capabilities to recover data. Therefore, an RDA technique is accompanied with data storage redundancy because the data owner is able to restore the corrupted data by using the remaining servers. In the design and implementation of the RDA technique some of the significant properties such as efficiency, public verifiability, frequency, detection probability, recovery, dynamic update should be considered.

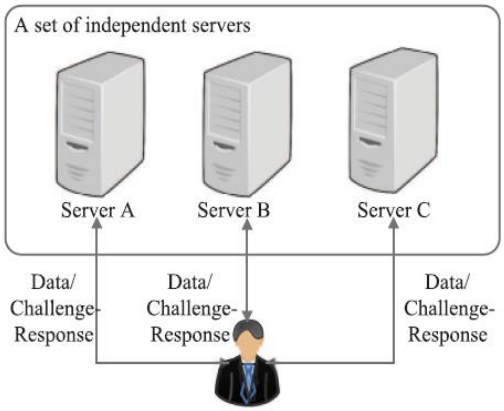
III. ARCHITECTURES OF REMOTE DATA AUDITING

Most individuals and organizations are motivated to reduce the cost and time involved in procurement and maintenance of local data storage infrastructure by outsourcing the data to the cloud. In cloud computing, the Cloud Service Provider (CSP) is in charge of managing the cloud storage services. As a result, the DOs are unable to maintain their possession and direct control over the uploaded data and instead the data is exclusively managed by an untrustworthy third party. On the other hand, the CSP or any insider adversary is able to maliciously manipulate data content without user knowledge[3].The remote data auditing technique samples data on the cloud and analyses these sampled data on the basis of integrity, correctness, and validity as benchmarks to ensure the reliability and trustworthiness of cloud service providers.

A. Architecture of RDA for Distributed Servers

The RDA schemes for distributed cloud servers consist of four main entities:

(1) Data Owner: the person who uploads his or her data to the cloud space and later might perform delete, insert, and append operations on the outsourced data.
(2) Cloud Service Provider: Has a tremendous amount of computing resources and stores and manages the DO's data. The CSP is also responsible for managing cloud servers.
(3) Third Party Auditor: In order to reduce the computation burden on the DO's side, the auditing process is often assigned to a TPA with adequate skills and capabilities to accomplish the auditing task on behalf of the DO. The TPA's role is particularly important when DOs possess relatively poor computing devices in terms of processing power, storage space, and bandwidth. During data auditing the TPA must be prevented from obtaining knowledge of the DO's data content and to protect privacy of data.
(4) User (individual or enterprise): Is enrolled and authenticated by the DO and permitted to have a predetermined type of access to the outsourced data. The RDA architecture for distributed storage systems is classified into three categories such as:[7]



(1) Multiserver model: In multiserver model, the data owner distributes multiple copies of the data among several servers and separately checks each of them. Fig.3.1 shows the architecture of the multiserver data auditing model.
(2) Single cloud and multiserver: In single cloud and multiserver model, all of the servers are distributed within a single cloud where the CSP is in charge of managing the servers. As is shown in Fig3.2 the data owner and the TPA are directly connected to CSP rather than all of the servers.

Fig 3.2. Single cloud and multiserver audit architecture.
(3) Multicloud and multiserver: The data owner outsources the data among multiple clouds instead of a

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

single cloud. Similar to the single cloud and multiserver model, one of the CSPs, namely, the organizer, is responsible for managing all of the servers and the other CSPs. the organizer that is directly connected to the owner receives data and a challenge from the data owner to distribute among the clouds and the servers. Moreover, the organizer aggregates the received proofs from the servers and sends them to the DO.

A typical RDA service works according to the following essential response-challenge procedure: First, the data owner performs a preliminary process on his or her file to generate some metadata to be passed to the TPA. Hereafter, the data owner does not need to be engaged in the rest of the auditing process. To verify the integrity and correctness of the remote data residing on the cloud, the TPA selects a random index of the outsourced data as a challenge message and directs that message to either the organizer or the CSP. When the organizer or the CSP receives the challenge, it is distributed among the servers, and then the organizer computes the corresponding response by aggregating the received messages from the servers. After receiving a response from the organizer or the CSP, the verification is carried out by the auditor to ensure the reliable placement of the file in the cloud storage.

B. Single server versus Multiserver

In a single server, the RDA techniques are classified into three groups:

The first category of RDA methods in the single servers is called the integrity-based schemes, in which the auditor is only permitted to validate the correctness of the outsourced data directlyor by using a third party.The second category of RDA schemes is the recovery-based models that are capable of verifying the data integrity and recovering the corrupted data when an error is detected.Deduplication-based facilitate the integrity and efficiency of data in a single server by removing data redundancy and increasing data storage optimization. Currently, individuals and organizations prefer to store data on distributed servers, because the single-server setting has no capability to recovery the data properly when data corruption is detected.

C. Categories of RDA

The two categories in remote data auditing are proof of retrievability (POR) and Provable data possession (PDP).

The Proof of retrievability tries to obtain and verify a proof that the data which is stored by a user at a remote data storage in the cloud (called cloud storage archives or simply archives) is not modified by the integrity of data is assured. This type of verification systems prevents 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 storage archives and these 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[6]. Provable Data Possession (PDP) allows data owner to periodically and remotely audit integrity of the data stored in the cloud storage, without repossessing the file and keeping a local copy. The system model of PDP is illustrated in Figure 3.3[5]. The users or data owners store their data in the remote cloud servers (CSP) and delegate them the maintenance of their data. The auditing process can be done by a third party (TPA) on behalf of the user upon request. PDP technique involves two phases which are setup and audit. The setup phase includes a key generation in which users can negotiate the keys with CSP and TPA and an authentication generation where users can compute authenticators as data tags of their data. However, the audit phase is usually done via a challenge-response procedure which follows challenge, response and verify steps. In the challenge step, a challenge message which includes indexes of randomly selected data blocks will be sent by the TPA to a CSP or multiple CSPs. Then, in response step, CSP receives the message and accordingly sends a response message includes both a data proof and the authenticator proof to the TPA. At last, the TPA, in turn, verifies the correctness of the proof to complete the verify step.

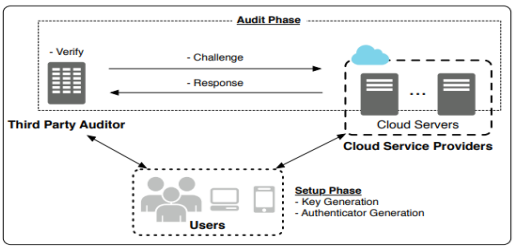


Fig 3.3. PDP-based remote data auditing system model

IV. STATE-OF-THE-ART RDA SCHEMES FOR DISTRIBUTED CLOUD SERVERS

The RDA schemes make use of various techniques to protect the integrity of the outsourced data for distributed storage systems. This section surveys the state-of-the-art RDA methods for distributed storage systems and classify the survey of RDA algorithms based on the data redundancy feature. The simplest and most common way

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

to achieve the reliability against data failures is to use a replication technique in which multiple copies of data are outsourced within the distributed storage systems. Whenever a data corruption is detected, the client can use an intact copy of the file. The main disadvantage of the replication method is that the storage cost. This is because during the repair phase, the client must retrieve a replica of size, and the communication overhead of replication in the recovery mode is equal to one[2]. Some of the replication-based RDA methods are:

• Multiple Replica Provable Data Possession (MR-PDP)

This method extends the PDP scheme to generate multiple replicas for distributed servers without encoding each replica separately. This was a provably secure scheme to store a large number of copies of files

In this scheme, the client generatesunique replicas by encrypting the original file and masking the blocks of the encrypted file by using a random value for each of the replicas. Thereafter, the client uses a decrypted file to create a tag for each block. The client then outsources a generated replica and a set of tags on each server.

Although the MR-PDP method is suitable for checking the integrity and availability of distributed servers, the data owner is unable to entrust the auditing to the TPA because the MR-PDP only supports private verification. Moreover, to update a block of the file, the DOs must retrieve the entire data, which imposes a huge computation and communication overhead on the client and server.

• Efficient Multi-Copy Provable Data Possession (EMC-PDP)

This method supports dynamic auditing, resilient against colluding servers attack and less storage overhead than MR-

PDP scheme. The EMC-PDP is introduced in two different versions: deterministic (DEMC-PDP) and probabilistic (PEMC-PDP). In the deterministic version, all of the file blocks are verified. The probabilistic scheme relies on the spot-checking approach in which only a random fraction of the file is checked. Even though the DEMC-PDP provides a stronger security guarantee, it is achieved at the expense of a higher storage overhead on the client and the server.

The main idea behind the EMC-PDP method is to generate a unique replication of file by attaching a replica number to the original file. Therefore, the generated replica is encrypted with a strong diffusion property of an encryption scheme. The DO also generates a distinctive tag for each block of replicas and distributes them along with the replicas among the servers. Finally, the authorized users are able to validate the data possession of all of the replicas or a random subset by using a challenge-response protocol.

The EMC-PDP is more efficient than the MR-PDP scheme in the following ways:
(1) It supports authorized users.
(2) The storage cost for the EMC-PDP is six times less than that of the MR-PDP.
(3) The required bandwidth is much less than the MR-PDP due to the application of the aggregation strategy.
(4) The PEMC-PDP is the most efficient protocol in terms of computational cost.

• Cooperative Provable Data Possession (CPDP)

Cooperative Provable Data Possession is a replication-based remote data auditing framework for distributed systems. This method uses the homomorphic verification response (HVR) and hash index hierarchy (HIH). The hash index hierarchy is a hierarchical structure which presents the relationships among the data blocks of various storage service providers and includes three layers such as the Express, Service, and Storage Layers which supports the batch auditing for auditing multiple clouds and the dynamic data auditing. The HVR is another fundamental technique in the C-PDP scheme which takes care of combining the generated responses from numerous cloud providers into one response based on the sum of the challenges. As a result, the communication overhead is reduced and the privacy of data is preserved by hiding the outsourced data location in the distributed storage system.

In the architecture of the C-PDP scheme, an independent server or one of the existing CSPs is assumed as an organizer, who has responsibility for managing all of the CSPs, initiating and organizing the verification process, and communicating directly with the client. Moreover, after a challenge is issued by the client, the organizer aggregates all of the responses received from the CPSs into one response by using the HVR technique, to be sent to the client. Even though the C-PDP scheme has several advantages, it must be assumed that the organizer is a trusted entity. Moreover, a heterogeneous structure of the proposed scheme leads to a high communication load due to the intercommunication between various cloud servers.

• Tree-Based Dynamic Multi-Copy Provable Data Possession (TB-DMCPDP)

In this method, the original form of the MHT is used for each of the replicas, and then the root of each of the trees

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

is placed as a leaf to construct a unique tree, namely, the directory of the Merkle Hash Tree(MHT). The main concept behind such an approach is to verify the integrity of all of the replicas in a hierarchical manner using a directory MHT in which the leaf nodes of the tree are the root node of each file copy's MHT. The drawbacks of this method are more storage and communication overhead than MB-DMCPDP and Cloud needs to store MHT for each file, which affects system performance.
Some of the modify, insert, and delete operations impose sufficiently great overhead on the server side in the server side in the TB-DMCPDP method. The CSP must rebalance all of the MHT structures to perform modification, insertion and deletion. On the contrary, storing several MHTs on the servers suffer from an enormous storage cost when the size of files is dramatically increasing.

• ***Map-Based Dynamic Multi-Copy Provable Data Possession (MB-DMCPDP)***
To address the storage and computation overhead, a novel data structure known as the map-version table was implemented. The table that is used to check the outsourced data integrity contains three columns: Serial Number (SN), Block Number (BN), and Version Number (VN). The SN basically represents the actual (or physical) position of the block in the file, while the BN shows the logical location of the block in the file. The VN for a block indicates the number of dynamic operations applied to that block so far.
The map-version table needs to be stored in the local storage of the DO, who is responsible for updating the table during the modify, insert, and delete operations. For example, when the DO decides to insert a data block after position i, a new row must be appended to the table (after the last entity of the table as an actual position) with these characteristics (SN,BN,VN) = (i+1, Max (BN) + 1, 1). Meanwhile, to delete a data block from the outsourced data, the DO is only required to delete the requested block from the map-version table. In the map-version-based approach, the update operation is performed in an efficient way that leads to fewer computational and communication costs. This method is also efficient when there are many verifiers connected to the CSP, as the challenge-response phase requires a lower computational time. The main disadvantage of the map-version table is that the required storage to keep the table is more than the MHT directory.

• ***Transparent, Distributed, and Replicated Dynamic Provable Data Possession (DRDP)***
In the architecture of the DR-DPDP, one of the servers is considered as a logical entity, called the organizer, who is in charge of connecting servers to the clients. The servers are only able to communicate with the organizer, and there is no internal communication among the servers. The central idea behind such architecture is to break a large authenticated skip list into several smaller sublists. The top sublist is provided to the organizer, and the other low-level parts are stored by the other servers, which causes improvement in the scalability. Each sublist may also be copied to more than one server to enhance availability and reliability. In the course of the upload phase, the client must divide the input file into n blocks and generate a unique tag for each of the blocks before transmitting the data to the organizer. When the file is received, the organizer splits the file in some partitions and sends them to an agreed-upon number of servers. Each server then constructs the corresponding part of the rank-based authenticated skip list and returns the root value as a response to the organizer. In the uploading phase, the organizer reconstructs a rank-based authenticated skip list and returns the root value to the auditor

V. CHALLENGING ISSUES FOR DISTRIBUTED BASED RDA TECHNIQUES.

A. Dynamic Data Update
As online word processing intrinsically deal with a dynamic form of data or are involved with dynamic log files, dynamic data update is an important characteristics of RDA[7]. During the update operations, such as modify, delete, insert, and opened in static mode, the clients must completely download the outsourced data from the cloud and upload it after performing the corresponding operations. If the auditing method supports the dynamic data update, then the client needs to download the number of blocks that are to be updated which in turn reduces the computation and communication overhead of updating data on the client and servers.

B. Batch Auditing
This feature enables TPA to process multiple auditing tasks received from different users at the same time rather than performing each of the tasks separately. Because of the redundancy characteristic of the RDA algorithms, addressing batch auditing in the distributed storage systems is more challenging. Only a few existing RDA methods focus on the batch auditing issue in the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

distributed storage systems. A simple way to achieve such a goal is to use a bilinear aggregate signature to combine the proof messages into a single and unique signature that can be verified by the auditor.[8]

C. Data Deduplication

Data deduplication basically removes duplicate data copies in order to facilitate a cost-effective storage. It is a kind of data compression technique (as a single-instance data storage) that is employed to avoid data redundancy.

D. Lightweight Data Auditing Approach

Developing lightweight remote data auditing approaches to improve the security of mobile users without any further limitation and requirement is a significant challenge in mobile cloud computing environment. Dividing the huge files into some blocks, generating the specific tag for each block, computing a challenge, and verifying the proof message are particular tasks in data auditing mechanisms that noticeably increase overall execution time and decrease the lifetime of resource constrained devices such as smart phones and tablets. A feasible approach to decrease the side effect of remote data auditing approach on mobile devices is to utilize the efficient public verification approach.

VI. CONCLUSION

In this paper the concept of cloud computing and distributed storage systems and the RDA technique to protect the outsourced data in cloud servers is explained and it focuses mainly on the architecture of the distributed-based remote data auditing techniques. The fundamental differences between distributed and single auditing approaches is mentioned. The state-of-the-art RDA techniques were compared for multiple cloud servers. The issues and the challenges concerning the security requirements to offer an efficient and lightweight security mechanism is mentioned. Furthermore, numerous open challenges, were introduced as prominent upcoming challenges for further investigation.

VII. REFERENCES

][1]. Peter Mell, Timothy Grance: The NIST Definition of Cloud Computing (2011).

[2]. Mehdi Sookhak, HamidTalebian, EjazAhmeda, AbdullahGani, Muhammad KhurramKhan on Remote Data Auditing in Cloud Computing Environments: A Survey, Taxonomy, and Open Issues (2015).

[3] Geeta C M, Raghavendra S, RajkumarBuyya, Venugopal K R, S SIyengar, L M Patnaik on Data Auditing and Security in Cloud Computing: Issues, Challenges and Future Directions (2018).

[4]Jia Xu, Ee-Chien Chang, and Jianying Zhou: Towards E_cient Provable Data Possession in Cloud Storage (2007)

[5]. Haritha Nuthi, HemalathaGoli, Ramakrishna Mathe: Data Integrity Proof for Cloud Storage (2014)

[6]. Ms. Sneha Shete, Mrs. NilimaDongre: Analysis & Auditing of Network Traffic in Cloud Environment (2017)

[7] A. F. Barsoum, M. A. Hasan, Provable multicopy dynamic data possession in cloud computing systems, IEEE Trans. on Information Forensics and Security, 10(3): 485–497, 2015.

[8] J. Yu, K. Ren, C. Wang, V. Varadharajan, Enabling cloud storage auditing with key-exposure resistance, IEEE Trans. on Information Forensics and Security, 10(6): 1167–1179, 2015.