International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Efficient Methods for Data Integrity Protection and Fault Tolerant Storage System in Cloud

Dr. M. Ramanan[1], Dr. G. Ganesh Kumar[2]

[1]Department of Physical Sciences and IT, Tamilnadu Agricultural University, Coimbatore, India

[2]Department of Computer Science and Engineering, PARK College of Engineering and Technology, Coimbatore, India

Abstract: ***Cloud computing is a paradigm that provides computing, communication and storage resources as a service over internet. Cloud Storage constitutes data that is mostly heterogeneous and dynamic in nature. In Cloud storage systems the data is to be protected from corruptions, redundant data to tolerate failures of storage and lost data should be repaired when storage fails. By distributing data across several servers, traditional code regeneration techniques for failure recovery are effective only in the case of using less network traffic. Also remotely checking the integrity of the data using code regenerated corruption methods is a complicated process in real time cloud storage. To overcome the problem of reducing data corruption and data retrieval, a novel method is proposed which implements a Integrity Protection Scheme (IPS). Unlike other approaches, we have considered both proof of integrity of data as well as proof of data possession. This makes only a fraction of data needed to analyzed for any corruption and not the entire file has to be reviewed. Therefore this IPS effectively protects the properties of the data stored in cloud and regardless of bandwidth of the cloud storage, the availability of data is improved. This proposed method when compared with other methods shows 17.27% average recovery time for 250 nodes, 12.12% average recovery time for 500 nodes and 8.57% average recovery time for 500 nodes.***

Keywords: Data Integrity, Fault Tolerance, Cloud Integrity Management, Replica Creation, Data Corruption

1. Introduction

Cloud computing is one of the types of a large scale distributed computing paradigms ; In the past several years, owing to its innovating and promising potential, it has become a driving force for Information and Communications Technology (ICT) . The way in which hardware and software are designed, purchased and used and improvement in Information Technology (IT) systems management has been facilitated by cloud computing (Jouini& Rabai2016). From content management to specialist applications for various activities, there are a range of applications that can be delivered to the users using the cloud computing models. Cloud users can access almost “infinite” and “ubiquitous” information services using the cloud computing. Cloud storage is one of the most important services offered by cloud computing wherein, users can move their data from the local servers to the cloud servers (Yu et al., 2016).

Even though a more secure and reliable environment is promised to the users of cloud by the cloud service providers, the data on the cloud is likely to be compromised as these systems are prone to human errors as well as hardware/software failures. Several schemes are suggested for protecting the integrity of data on the trusted cloud. In these mechanisms, a signature is associated to every data block and the accuracy of these signatures determines the data integrity. Public auditing is one important feature of these mechanisms- they allow the data owners as well as a public verifier like a Third Party Auditor to check for the data integrity on the cloud without the need to completely download the data (Wang et al., 2015).

A technique that is used often in the cloud is replication like, Google File System (GFS), Hadoop Distributed File System (HDFS). Nonetheless, due to the rapid growth in the size as well as the number of the cloud data centres, dynamically scalable as well as completely virtualized resources have been provided as services over the internet. Data replication, in most real clouds is attained using a data resource pool; here on the basis of history, the number of data replicas is statistically set and is normally lesser than 3. While at most times this strategy works well, it may fail at critical junctures. Also, it is normally not required to replicate all data files, more so, the non-popular ones. It is required to dynamically adjust the popular data files, the number of data copies and the sites for placing the new replicas as per the current cloud environments, so that the requirements of high availability, high fault tolerance and high efficiency are met.

Nodes in cloud computing system are heterogeneous because of which the data of a high- Quality of Service (QoS) application may be replicated in a low performance node (the node with slow communication and disk access latencies). Later, in case if data in the node running the high-QoS application is corrupted, the application data is retrieved from the low performance node. The QoS requirement for the high QoS application may be violated as the low performance node has slower communication as well as disk access latencies. It is to be noted that the QoS requirement of a high QoS application is defined from the perspective of the requested data. For instance, the response time of a data object access is defined as the QoS requirement of an application in the content distribution system (Lin et al., 2013).

2. Related Works

Cloud Storage has many benefits and great advantages. Some of the advantages are like provide better accessibility, one can easily access their data from anywhere by using Internet. Another benefit is we should not take the hardware storage with us for the data it also enhances the team work because one can easily be share their work and a group can collaborate with each other easily and many more

advantages. Beside all the benefits and advantages of the Cloud there are some of the limitations of the Cloud. The limitation discuss in terms of public cloud. In the public Cloud the data stored in the cloud is visible to all or accessible to all and one can easily get the data because in public cloud data is open to the public. Cloud Integrity in [1] proposed a benchmark for transmit of the data. Here protection of the data during migration through benchmark is discussed for the Encryption overhead and security. For more security, more powerful encryption is required.

Data Integrity in [2] discuss among threats involved in insecure APIs are anonymous access and/or reusable tokens or passwords, clear-text authentication, improper authorization and API dependencies. In the paper [3] the author proposed the new version of AES- 512 bit encryption algorithm. The author presents the architecture for AES-512 and efficient hardware that requires to implementation of the Algorithm is also discussed. In this paper the author uses the 512 bit key size and same bit block size also uses which makes the algorithm more resistant towards the attacks.

According to the user this algorithm provides the more security to the data with more throughputs. The author in [4] studied the security as a part of survey. According to the survey of the author various other security issues should also be considered beside the main issues and their solution also. Here different data security concerns are analysed and solved by classification of the data. Different security and protection is provided according to the degree of values of the data.

The problems of confidentiality and privacy of the data in cloud is addressed by using a proper data integrity scheme. To overcome these two problems

They design the framework which solves the problem of unauthorized access. Different mechanisms are used for different task like Key Management mechanism, Data Encryption mechanism, Multi-way Tree index mechanism etc. These mechanisms are different in client and the server side. An effective security is achieved when the users is aware of the state of the data [6]. Data exists in one of the three states: at rest, at process and in transit. In the paper author convey that in all the three states data require different security. All the three states require different and unique security protection. For example if the data is considered as sensitive then it remains sensitive in all the states i.e. at rest, at process and in transit.

As the use of the cloud increases different algorithm are proposed for the protection of the data. In [7] author proposed an optimized technique for the security of the data by using encryption process. Here symmetric block cipher algorithm (CHis-256) to protect the data in the efficient manner. In the paper [8] the author shows us the efficient manner of the implementation of the AES-512 algorithm with proper and efficient utilization of the resources used in it. Here the comparison between the AES-128/256 and AES-

3. Our Contribution

The manner in which the QoS requirements can be effectively taken into account is considered by The QoS Aware Data Replication (QADR) problem. Reducing the cost of data replication is the objective of the QADR. This reduction will lead to the decrease in the rate of corrupting the data. The data replicas in some applications may be stored in nodes that have lower performance because of the constrained space of replication in a storage node. This may result in some of the replicas not meeting 512 is done and shows us the reasons to use AES-512 for more security with better throughput.

There are several types of issues [11] that cloud users face and consumer may face during the use of the services of the cloud. Most of the issues are with security to the data. The data is confidential and available when it is business. In paper [12] data classification technique is used for providing security to the data. Here the classification is done on the basis of the confidentiality of the data and according to the respective domain of classification security provided. Classification levels are Basic level, confidential level and High Confidential level. Here the best security technique which is used for the security is AES-256 with SHA.

Before applying security to the data it is important to understand and identify the various security challenges which are going to be faced. In [11] the author shows us the different security challenges other than the basic security issues. Here not only the author displays the security challenges but also focus on the percentage of importance of that challenges in the cloud computing. The requirements of the QoS as well as their applications. These are the copies that violate the QoS and the amount of such of the replicas is assumed to be a small number. This work proposes an SDS algorithm for resolving the issue of QADR.

4. Methodology

Many schemes are being proposed for checking the data integrity as well as the owner's public key prior to the usage of data as the cloud environment is prone to security threats. Better availability, scalability and durability is assured by the replication of data on cloud servers using multiple data centres. The security of Public Key Infrastructure (PKI) determines the accuracy of selecting the right type of public key in the prior schemes. Despite the prevalent usage of the conventional PKI in building the public key cryptography, it lacks many aspects of handling security risks, especially management of certificates.

4.1 Model

As shown in Fig 1, there are four different entities in this system model" the cloud, the data owner, data users and finally Key Generation Centre (KGC). Data services are provided by the cloud to both the owner and the user. The data that is outsourced by the owner to the cloud will be saved on local devices. This data is effectively partitioned into data blocks for enabling effective modification. This data is used by the user. Search operation may also be

performed by the user on the cloud data, for some reasons. A trusted source in this prototype is the KGC that is required within the framework of schemes that lack certificates and can generate partial and private key for the entity/ data owner on the basis of identification such as the name or the email address. The remainder of the private key is generated by this very entity itself (Wang et al., 2013).

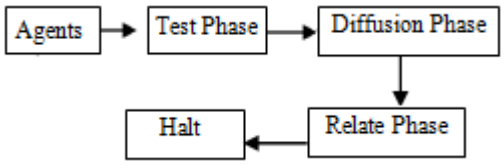

Figure 1: System Model for Data Integrity

Cloud data can be corrupted due to two reasons: Firstly, data corruption by an external malicious interference that can pre-empt the owner as well as the user from making use of the data accurately. Secondly, the data may be corrupted due to the human errors caused by the cloud service provides because of which the owners and the data users are inhibited from completely trusting the cloud.

A signature is attached to the owner's private key, and this aids in protecting the data integrity. The integrity of the cloud data comprising the search, computation and data mining has to be examined by the data user; this will be followed by the cloud generating a proof of possession. Finally, based on the auditing outcome, the integrity of the data will be verified by the data user. It has to be taken into account that in order to check for the data as well as its integrity, the owner needs to play the role of verifier as well by means of following the protocol.

There are three goals of formulating this public auditing mechanism: 1)Correctness: A public verifier (i.e. data user) can ascertain the data integrity on the cloud precisely. 2) Public Auditing: The accuracy of the data is properly audited by a public verified without retrieving the entire data from the cloud and 3) Certificate less: This means that the public auditing has been accurate and there is no requirement for a public verifier to manage certificates.

All of the agents that are set to be inactive, during initialization stage arbitrarily choose a hypothesis from a search space. The initialization is influenced by : 1) That at least one single agent is duly initialized with one of the best hypothesis is assured by the actual ratio of the size of this model to that of the search space and is greater than 1.2) The knowledge of the prior location of this model that will aid the performing of successive searches or such of the similar search spaces, like, the successive video frames.

These agents, during the test phase would determine if they have to adjust themselves to be more active or inactive. Applying a single test function to its current hypothesis will help achieve this. The position of hypothesis is partially evaluated by this type of a test function. Depending on the domain of the application as well, the test function varies. If this in-part analysis of the hypothesis returns success, the agents will be set to active, else they will remain inactive (al-Rifaie et al., 2011).

The knowledge of the hypothesis is exchanged by the agents during the diffusion phase. This aids in the active agents distributing their hypothesis to those that are not active. Three recruitment strategies are followed in distribution of hypothesis: the active recruitment, the passive recruitment, combination of active and passive recruitment. An accurate hypotheses in recruiting the inactive agents is resulted from this type of an information exchange which will lead to several agents aggregating in the vicinity of the available hypothesis. A standard SDS strategy of recruitment will be deployed and passive.

An optional stage that is incorporated if there are several extant models within the search space is a relate phase which can aid the dynamic re- assignment of the agents. This also enables the re- alignment of the dynamic search space by making use of the appropriate hypotheses. The context free mode and the context sensitive mode are the two modes that are comprised in a relate phase.

The criteria to halt when certain percentage of agents, irrespective of the hypothesis are active, is stated by the Weak halting criteria. A stabilization exists that can be viewed as an active agent set which is stable with certain tolerance margin. One this is met, the search process halts. The halt state instead of being defined as being associated with the number of active agents in a large population using the tolerance rule or the threshold as the weak halting state is seen as the percentage of the agents that are active within this large population.

The goal in equation (1) is minimizing the first minimum term that refers to the total replication expense of all the data replicas, and the second minimum terms would be the actual amount of such replicas that violated QoS. The subsequent minimum terms will be before that of the first minimum term. For assuring that the actual number of the QoS violated data replicas are one of the first to be minimized, a coefficient k is used (Lin et al., 2013). The main reason will be explained subsequently. In the equation (1), in case the requested node ri will put one data block replica within the node qj, this particular event will be recorded by means of setting 1 in (□□, □□). However, in case this replica is one of a QoS-violated data block type of a replica, (□□, □□) it will also be set as 1. By means of adding all these values of y, the number of QoS violated data replicas will be obtained. This number will be expected to be small by means of associating with one constant coefficient. With the actual setting of k, each (□□, □□) has a larger coefficient than (□□, □□). This is called the values of the (□□, □□) and (□□, □□) which are either 0 or 1.

With the objective of finding a state having a fitness that is as high or as low as possible, the algorithm that is used in this work, searches the discrete space S. For some current state, the algorithm makes successive improvements and achieves it. The manner in which the formulator of the SH algorithm selects the encoding of the solutions to the problems to be solved determines the form of states in S , just like in GAs. The solution encoding can be done as bit strings, permutations or some other form. The neighbourhood structure and the fitness function f that is imposed on S in formulating the algorithm will determine

the local improvement impacted by the SH algorithm (Juels& Wattenberg 1996).

The neighbourhood structure can be considered as a undirected graph G on a vertex set S. The algorithm attempts to improve its current state by making a transition to one of the neighbours of in G. Specifically, according to some suitable probability distribution on the neighbours of, the algorithm chooses a state. Becomes the new current state of its fitness is at least as much as, else, the latter is retained. The procedure is repeated. Random walk is performed by the algorithm wherein any mode that results in decreased fitness of the current state is rejected. More formally, in the form of a piece of pseudo code performing algorithm for M iterations:

5. Result

The figure 2 shows the average recovery time for Integrity Management algorithm.

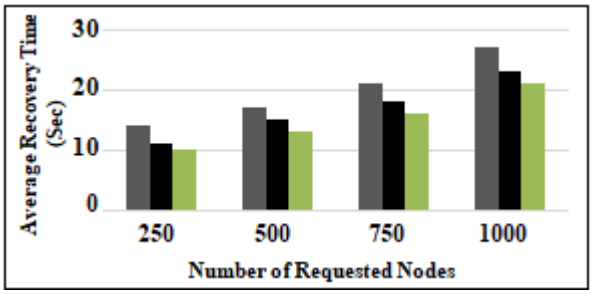

Figure 2: Average Recovery Time

From the figure 4, it can be observed that the Data integrity Management algorithm has lower average recovery time by 33.33% & 9.52% for 250 number of requested nodes, by 26.66% & 14.28% for 500 number of requested nodes, by 27.02% & 11.76% for 750 number of requested nodes and by 25% & 9.09% for 1000 number of requested nodes when compared with random and data integrity management algorithm.

6. Conclusion

Verification of the data integrity is a critical issue in cloud storage service. As cloud data auditing enables the cloud users to ascertain the integrity of their outsourced data effectively, cloud data auditing is paramount in obtaining cloud storage. The first certificate-less public auditing scheme for verifying the data integrity on cloud has been proposed in this work. This enables the public verifier to verify the data integrity on the cloud and also to preempt any security threats due to the PKI in the previous solutions. A solution that is seen as promising is data replication that brings the data (i.e. databases) closer to the data consumers (i.e. cloud applications).

References

[1] Wang, B., Li, B., Li, H., & Li, F. (2013, October). Certificateless public auditing for data integrity in the cloud. In Communications and Network Security (CNS), 2013 IEEE Conference on (pp. 136- 144). IEEE.
[2] Williams, H., & Bishop, M. (2014). Stochastic diffusion search: a comparison of swarm intelligence parameter estimation algorithms with ransac. Algorithms, 7(2), 206-228.
[3] El-henawy, I. M., & Ismail, M. M. (2014). A Hybrid Swarm Intelligence Technique for Solving Integer Multi-objective Problems. International Journal of Computer Applications, 87(3).
[4] al-Rifaie, M. M., Bishop, M. J., & Blackwell, T. (2011, July). An investigation into the merger of stochastic diffusion search and particle swarm optimisation. In Proceedings of the 13th annual conference on Genetic and evolutionary computation (pp. 37-44). ACM.
[5] Lin, J. W., Chen, C. H., & Chang, J. M. (2013). QoS-aware data replication for data- intensive applications in cloud computing systems. IEEE Transactions on Cloud Computing, 1(1), 101-115
[6] Juels, A., & Wattenberg, M. (1996). Stochastic hillclimbing as a baseline method for evaluating genetic algorithms. In Advances in Neural Information Processing Systems (pp. 430-436).
[7] De, M. K., Slawomir, N. J., & Mark, B. (2006). Stochastic diffusion search: Partial function evaluation in swarm intelligence dynamic optimisation. In Stigmergic optimization (pp. 185-207). Springer, Berlin, Heidelberg.
[8] Yu, Y., Xue, L., Au, M. H., Susilo, W., Ni, J., Zhang, Y., ... & Shen, J. (2016). Cloud data integrity checking with an identity- based auditing mechanism from RSA. Future Generation Computer Systems, 62, 85-91.
[9] Wang, B., Li, B., & Li, H. (2015). Panda: Public auditing for shared data with efficient user revocation in the cloud. IEEE Transactions on services computing, 8(1), 92-106.
[10] Wang, B., Li, B., Li, H., & Li, F. (2013, October). Certificateless public auditing for data integrity in the cloud. In Communications and Network Security (CNS), 2013 IEEE Conference on (pp. 136- 144). IEEE.
[11] Lin, J. W., Chen, C. H., & Chang, J. M. (2013). QoS-aware data replication for data- intensive applications in cloud computing systems. IEEE Transactions on Cloud Computing, 1(1), 101-115.
[12] Jouini, M., &Rabai, L. B. A. (2016). A Security Framework for Secure Cloud Computing Environments. International Journal of Cloud Applications and Computing (IJCAC), 6(3), 32-44.