International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 2, April 2020, pp. 2011~2022
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i2.pp2011-2022

An extensive research survey on data integrity and deduplication towards privacy in cloud storage

Anil Kumar G., Shantala C. P.
Department of Computer Science and Engineering, Channabasaveshwara Institute of Technology,
Visvesvaraya Technological University, India

Article Info

Article history:

Received Jun 1, 2019
Revised Oct 23, 2019
Accepted Nov 4, 2019

Keywords:

Cloud computing
Data deduplication
Data integrity
Data privacy
Data security

ABSTRACT

Owing to the highly distributed nature of the cloud storage system, it is one of the challenging tasks to incorporate a higher degree of security towards the vulnerable data. Apart from various security concerns, data privacy is still one of the unsolved problems in this regards. The prime reason is that existing approaches of data privacy doesn't offer data integrity and secure data deduplication process at the same time, which is highly essential to ensure a higher degree of resistance against all form of dynamic threats over cloud and internet systems. Therefore, data integrity, as well as data deduplication is such associated phenomena which influence data privacy. Therefore, this manuscript discusses the explicit research contribution toward data integrity, data privacy, and data deduplication. The manuscript also contributes towards highlighting the potential open research issues followed by a discussion of the possible future direction of work towards addressing the existing problems.

Corresponding Author:

Anil Kumar G.,
Department of Computer Science and Engineering,
Channabasaveshwara Institute of Technology,
Visvesvaraya Technological University,
NH 206, Gubbi, Tumkur, Karnataka 572216, India.
Email: aaniltumkur@gmail.com

1. INTRODUCTION

Cloud computing has offered various forms of application in the form of services, while a distributed storage system in the cloud is one of the most significant contributions [1]. With a large scale of network of cloud clusters over its global datacenters, now it is feasible to access and store the data and make it available anytime [2]. Different forms of challenges associated with cloud computing are data availability, data access, data integrity, data location, and network load [3]. Apart from these, there are various other forms of security threats that act as a challenge to the cloud storage system. The primary challenge is to ensure data integrity, which is about offering access rights by legitimate members only. Therefore, a data integrity mechanism is usually implemented in the form of validating the data. There could be various reasons that lead to a vulnerable situation towards data integrity over the cloud storage system. At present, there are various schemes of data integrity [4-6], but they suffer from i) lack of dynamic support, absolutely no preservation of privacy, zero codes of error correction for solving data corruption issue. Majority of the existing schemes of data integrity suffers from more number of problems, e.g., limited numbers of secret keys are utilized in the verification process, and so they are not applicable for large scale data. Another related problem of the existing system is that a user will need to have physical access to complete data in order to generate a new security token and hence, they are not applicable for the larger file system. Apart from this, the quantity of updating operation, which is very important in security, is highly limited for the clients. Moreover, a scheme like scalable provable data possession doesn't offer insertion of the block.

Apart from this, such a mechanism of data outsourcing increases privacy issues [7]. In the direction of the security of data, data deduplication is also frequently used for distributed data storage in the cloud. The prime task of the deduplication operation is to retain the highest possible security information and retain optimality of storage space [8]. Once the data is encrypted, it is subjected to a deduplication process which maintains more security and redundant data management. In order to offer a secured deduplication process, it is necessary to offer encryption process [9, 10]. The process performs permutation of the data that is replicated with a specific secret key where the elements of the replicated data are obtained by applying conventional hash function. The client obtains the secret keys after the encryption process, and the encrypted data is forwarded to the client after that. According to the conventional theory, it is stated that applying secure deduplication will result in optimization of channel capacity, more data reliability, up-scaling performance, etc. However, it is very difficult to ascertain this fact in the practical situation as there are various forms of threats widely available over cloud ecosystem which is more potential and their attack behavior has never been studied in the past. At present, there is no such evidence of a standard model which claims that data cannot be accessed by the illegitimate member and thereby causing a breach to a distributed data storage system.

Therefore, the present manuscript offers a discussion of some recent trends of research contribution towards data security over cloud storage system in order to visualize the existing scenario. The core goal of this paper is to offer current state of condition of the existing solution towards security problems in the cloud storage system. Section-2 discusses data integrity problem while discussion of data privacy problem is carried out by Section-3. The research work towards data deduplication is carried out by Section-4, followed by highlights of open research issues in Section 5. Section 6 briefs of possible future work direction towards addressing the existing security problem in cloud storage. Finally, Section 7 discusses the contribution of the existing paper.

2. STUDY TOWARDS DATA INTEGRITY PROBLEM

Data integrity is one of the primary security problems over the distributed storage system in the cloud ecosystem. The concept of data integrity lets the original user access and offer complete control of managing their intellectual property and bar other illegitimate users. However, there is less evidence about it. By making the replicates of the data over distributed cloud servers, the service providers have the nearest access to such data. Therefore, there is always uncertainty about the ownership of the data from the security aspect, which directs a question mark over data integrity over the cloud storage system. Various conventional mechanisms Figure 1 has been evolved out in order to address the problem of data integrity over cloud storage system viz. i) provable data possession [11], ii) Message Authentication Codes integrated with provable data possession scheme [11], iii) usage of symmetric encryption in provable data possession scheme, and iv) Proof of Retrievability [11], etc.

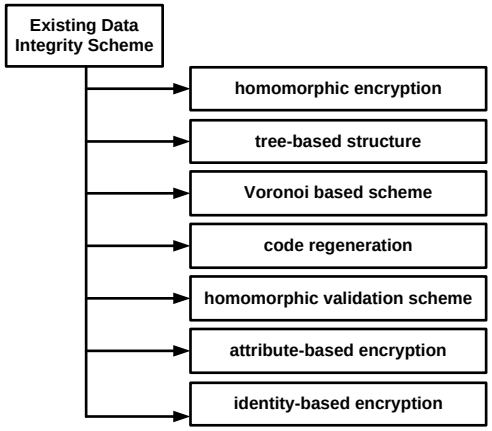

Figure 1. Existing approaches towards data integrity

In the existing system, the data integrity problem is investigated by remotely accessing cloud storage units. However, such a mechanism of assessing data integrity is also challenging owing to the distributed nature of the cloud storage units. * The problem of data integrity is more complex in the area

of Internet-of-Things (IoT) as massive generation of data. The existing mechanism is not functional over assuring IoT data integrity as their applicability is restricted over a single data block. This problem is sorted out by a tree-based data structure design for supporting the parallel update of multiple data blocks, as seen in the work of He et al. [12]. The authors have used a *homomorphic encryption* mechanism for seamless data transmission and for supporting enhanced updating process. However, such schemes are quite ineffective against sensitive file whose integrity cannot be ascertained. It is also essential that there should be run-time check towards such forms of the file system. Study towards such direction is carried out by Shi et al. [13] where the integrity of such dynamic data is made possible to be verified. An effective resistance towards illegal access of files is constructed by tracking operations associated with cache and input-output.

Blockchain is another mechanism to offer data integrity, considering the data via a virtual machine. Zhao et al. [14] have constructed a network on the basis of blockchain, followed by developing a partially constructed block that is distributed to other nodes for ensuring data integrity. The technique also uses *attribute-based encryption* for further securing the network of the data block. Apart from this, an *identity-based encryption* mechanism is also reported to offer remote checking of data integrity as seen in the work of Wang et al. [15]. Adoption of data auditing mechanism is also another mechanism assisting in the identification of the state of data integrity. However, they too suffer from key management problems that render the possibility of intrusion in storage units. The work of Li et al. [16] has constructed an auditing model where fuzzy logic has been used along with the secret sharing process for ascertaining robust data integrity with fault tolerance. Auditing method to offer data integrity has been presented by Shao et al. [17] where the vehicular network has been considered as a case study. The technique uses *the tree-based structure* with multiple branches for facilitating authentication as well as the technique also jointly uses a digital signature as well as bilinear pairing scheme. It is because *the bilinear scheme* has been found to reduce the overhead of the meta-data generation, as claimed by Shuang et al. [18]. Apart from this, the usage of enhanced signatures scheme is also proven helpful for offering data integrity of multiple clients with the same data. Such work was carried out by Wang et al. [19] where a public verification process has been presented with a data block being signed by multiple owners.

Essential information could also be in the form of a query system which is currently found to be vulnerable in terms of authentication of over outsourced cloud data. This problem has been addressed by Hu et al. [20], where a *Voronoi* based scheme has been introduced to understand the relationship between the spatial data and the query system. Apart from data, service integrity is another problem over cloud ecosystem when associated with the distributed architecture of the cloud. This problem has been solved by Du et al. [21] where the graph-based approach has been adopted for offering identification of malicious user followed by offering quarantined operation. The graph-based approach towards distributed cloud storage has also been presented by Lu and Hu [22] where the authentication is supported publically by Voronoi diagram over graph along with the enhanced hash tree. The author has also used *homomorphic validation* scheme to ensure data integrity.

According to Chen and Lee [23], *code regeneration* is one effective mechanism to ensure fault tolerance over a distributed storage unit. A model has been developed, which considers the mobility aspect of the Byzantium adversary and offers an enhanced capability to the client to perform a remote check of data integrity using a mathematical model. Study towards remotely checking of integrity has been carried out by Fan et al. [24] in order to protect the integrity proof using a non-conventional *cryptographic* means of handshaking mechanism. Adoption of *erasure-code*d while constructing a cloud storage system is also considered to protect data integrity. Integrity checking scheme presented by Shen et al. [25] using *homomorphic validation scheme*. Adoption of a *trust factor* over the operational platform is another mechanism to address this problem. The approach of Du et al. [26] has used a virtualized platform where trust computation is carried out towards access attempts over the cloud storage units. Apart from this, other popular existing schemes include joint usage of identity and homomorphic encryption (Yu et al. [27]) and obfuscation-based approach (Zhang et al. [28] and Zhu et al. [29]). These schemes address the data integrity problems with its specific cryptographic approach - the next section briefs of schemes to protect privacy factor.

3. STUDY TOWARDS DATA PRIVACY PROBLEM

Irrespective of the potential privilege of cloud storage in distributed order, there is always a risk of privacy factor associated with the data. The primary reason for this is the higher degree of dependency of the third party vendors to offer data security, which may not be appropriate to the exact business demands leading to loopholes in privacy. The root causes of privacy problems in the cloud are because of following–ineffective control over the data (especially while performing file sharing by the third party), illegitimate leakage of data (even by the service provider as well as by malicious hackers), accessibility of

the data/service by diversified devices (or service provider), higher risk of data interception over internet, poor key management, storage of user credential over cloud that can be fairly compromised. Therefore, there are various pitfalls of the existing system that is not so robust in protecting privacy factor of the data stored over cloud storage unit. In order to address the privacy problem, there has been an evolution of various research-based schemes and techniques. Out of various schemes Figure 2, the *encryption-based scheme* is one potential scheme to resist adversary to leak data privacy factor. The work carried out by Alabdulatif et al. [30] has used *homomorphic encryption* for retaining the privacy factor for sensor data reposited over the cloud. According to this scheme, the transmitted data over the cloud is encrypted while forwarding to the cloud servers. Apart from encryption, recent approaches have also witnessed the usage of *watermarking* approaches towards strengthening data privacy. The work of Tang et al. [31] has utilized adaptive watermarking scheme that is capable of encapsulating the data securely. The technique also uses *Diffie-Hellman* as a standard *key-exchange* mechanism for resisting replay attack. The mechanism of data embedding is fixed while applying the adaptive watermarking operation. The technique uses a *consensus mechanism* with simplified challenge and response based intrusion resistance technique for preventing replay attack.

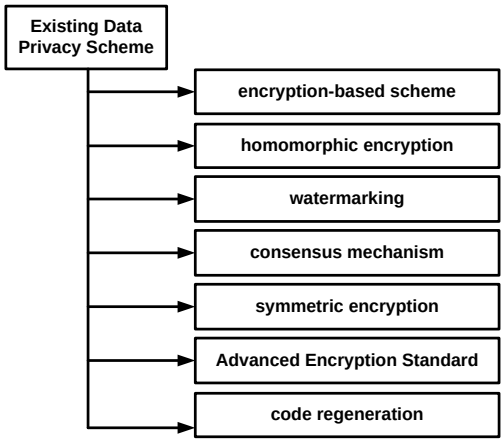

Figure 2. Existing approaches towards data privacy

Another recent work carried out by Du et al. [32] has used *a symmetric encryption* mechanism for resisting data leakage issues. The authors have presented an indexing mechanism for the privacy factors as well as protect multiple query processes which are claimed to be resistive against keyword-based intrusion. However, the approach could only offer forward privacy and not backward privacy factor, which is also essential. All these studies have been carried out with respect to hypothetical data and cannot be claimed to be secured if the data type is changed. It is because various biometric-based applications are running over the cloud system whose morphological information is protected in distributed storage units. The work carried out by Hu et al. [33] has used key agreement over the specific session as well as an encryption scheme for facilitating data privacy. The implementation of the study has been carried out, considering the fog computing environment where *SHA-1* and *Advanced Encryption Standard (AES)* has been mainly used. Work in the equivalent direction towards adopting fog computing was also seen in the case study considered by Wang et al. [34]. According to the author, the existing encryption techniques that are frequently adopted in offering data privacy are incapable of resisting threats within the cloud storage units. Therefore, a multi-layered based cloud storage system is formulated on fog computing. The technique has also used Hash-Solomon code for splitting the data as well as for assisting in decoding operation. Just like the capability to deal with the problem of data integrity, the *code regeneration* technique is found to resist data privacy problem too. By integrating auditing scheme with code regeneration approach, Liu et al. [35] have developed a system to ascertain robust data privacy. Auditing scheme has also been found to offer a solution towards privacy protection. Unfortunately, existing privacy protection scheme cannot be helpful much over the distributed nodes in the cloud. This problem has been discussed by Wang et al. [36] where the *ring signature* has been utilized for constructing metadata associated with verification demanded to assess the appropriateness of distributed shared data. According to this scheme, the information connected with the user identity is kept private from other users without any dependency over complete data.

User information in terms of identity is highly variable term and can be used for protecting data integrity. Therefore, usage of user identity information integrated with lightweight encryption scheme can be considered as a good option for protecting data privacy. Study in this concern has been carried out by Yu et al. [37] where the authors have used user identity information integrated with the joint usage of *the key-based* and *homomorphic-based encryption mechanism*. The authors claim of good control of computational complexity as well as reduced cost of operation using this cryptographic approach. According to the study, the authors highlighted that frequently used public key infrastructure is not a good option as it suffers from computational complexity. The technique also claims that data privacy is ensured without leading any private information associated with the stored data over the cloud. Work of Li et al. [38] has developed an auditing scheme considering the concern of low-end computational devices. The technique uses a *digital signature* as well as the mechanism offers better data dynamic with a wide range of supportability towards batch auditing. Study towards facilitating the public assessment of the data privacy is also carried out by Wang et al. [39]. According to the scheme, the verification towards the data privacy can be carried out without any dependency to access the original data content. Such claims are also offered in the work of Hao et al. [40]. **Research Gap:** The approaches towards ensuring data privacy have been discussed by various researchers where the majority of the approaches are found to have a common claim, i.e., ensuring data privacy without any dependency of accessing the original data from the verifier viewpoint - the next section briefs of data deduplication approach.

4. STUDY TOWARDS DATA DEDUPLICATION PROBLEM

Owing to the distributed nature of the cloud storage units and the presence of the virtualized environment, duplicated, and redundant data always exists in multiple sources. Such presence of duplicated data results in error-prone query processing as well as could also result in a security breach over the cloud storage system. One of the recent techniques to mitigate this problem of redundant data is called data deduplication resulting in minimization of storage overhead as well as optimized better data integrity. According to the standard process, the input file is subjected to hashing for extracting hash value followed by comparing the obtained hash value with that maintained over the index table of hash. Upon finding a positive match, the pointer is set to the existing location of data or else it reposits the novel data on its memory system and allocates a new hash on it. Irrespective of various methods Figure 3, the process of data deduplication can take place in both target and source. *Source deduplication* results in zero hardware dependency along with minimization usage of storage and network resources. However, *target-based deduplication* in expensive even if it ensures performance benefits over large data scale [41]. In present times, the deduplication process can be carried out by inline deduplication, post-processing deduplication, block-or-file level deduplication [41]. However, this standard technique of deduplication suffers from various loopholes too viz. i) large expense with data center management, ii) inadequate performance for catering up operating system and backup demands, iii) non-practical capacity planning for deduplication process, iv) not so efficient usage of hashing over large scale environment for waste resource processing, and v) poorly planned life-cycle control process [41].

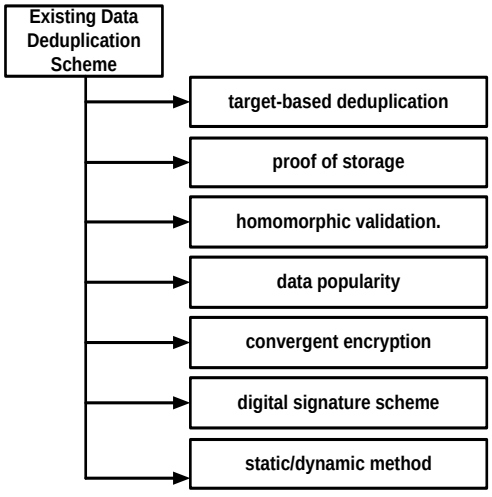

Figure 3. Existing approaches towards data deduplication

Apart from the issues mentioned above, there is a strong connection between the deduplication processes with the security factor in cloud storage units. When the cryptographic algorithms (which are majorly used in cloud storage security) are deployed than the original data is transformed to the encrypted data which is a very different format quite difficult even to identify its original form. This process is very different from data deduplication, and hence, there is a potential conflict between existing data encryption and data deduplication process flow. Therefore, when deduplication is applied over encrypted data, it will be extremely challenging even to identify the target data. It is because there can be the generation of multiple encrypted data forms of the same data when applied with different levels of the secret key. This causes failure in the deduplication process eventually.

In this regards, two standard techniques have been evolved in existing system viz. convergent encryption and proof of ownership [42]. The existing system offers feasibility for the direct client to monitor the deduplication process of their data, which also facilitates them to check the data integrity. Therefore, the existing system has jointly investigated data auditing process with deduplication. Study of Youn et al. [43] has applied a *digital signature scheme* as well as *homomorphic validation* approach. This operation is outsourced to a third party system in order to perform unbiased validation of data integrity of deduplicated data. Therefore, such a mechanism performs deduplication of data prior to the outsourcing process to the cloud storage system in order to retain better privacy factor. However, such schemes use equivalent encryption key for the same content, making it vulnerable for man in the middle attack. This problem is addressed by Hur et al. [44], where deduplication takes place with the server for managing the access rights for the dynamic data being uploaded by the users.

The presented scheme offers minimal data leakage and maximal data integrity. The proof-based approach also uses encryption while performing deduplication; however, their applicability is limited to a single user. Study towards the similar proof concept of multi-users has been presented by He at al. [45] where the authors have used *proof of storage* in its dynamic form for assisting the deduplication process for cross-users. The technique also constructs a *tree* using *homomorphic validation*. Irrespective of better execution formulation, the work suffers from computational complexity problem as there is an additional need of identifying all the duplicated encrypted files. This problem has been discussed in the work of Jiang et al. [46] using both *static/dynamic method* for complexity reduction.

A unique data deduplication scheme has been presented by Stanek et al. [47], which is based on a *data popularity* score of the data. According to this technique, the deduplication process is applied only when the data becomes popular. Study towards secure data deduplication over the multimedia file is discussed in the work of Zheng et al. [48], which encrypts the deduplicated file and uploads it on the specific media center. However, the strategy to offer defense against attacks is put forward by the third party server. Study towards deduplication concerning about reliability factor is carried out by Li et al. [49] over multiple servers of the cloud. The technique also implements secret sharing over a distributed storage system. The combined study of data integrity and deduplication process is presented by Li et al. [50] where a secured cloud system has been introduced. The mechanism calls for performing auditing operation over the conventional distributed software framework. This process generates an index of specific data prior to uploading process to ensure better data integrity. The approach presented by Yan et al. [51] has used re-encryption over proxy sources as well as challenges of ownership in order to perform deduplication of ciphered data over cloud storage system. The technique also establishes associated between access control systems with data deduplication. The similar direction of the work has also been carried out by Fan et al. [52] where *convergent encryption* process is mainly applied along with hashing/public encryption usage. **Research Gap:** Irrespective of the approaches mentioned above, the studies towards data deduplication are quite less in contrast to other associated problems with data security of the cloud. The next section outlines other auditing processes.

5. EXISTING SECURITY AUDITING SCHEMES

Auditing is a procedure to investigate the performance effectiveness of the services hosted over cloud environment. Generally, auditing is carried out by third parties in order to extract data associated with various operational performances of cloud-based application/services. The prime objectives of performing auditing are viz. i) formulating the data architecture, ii) controlling IT risk, iii) strategically constructing an IT plan, iv) communication management, and v) security controls. Therefore, auditing scheme relates to the operational assessment of cloud where security is just one factor to be assessed along with many other functional factors [53-56]. Hence, most recently Figure 4, various researchers have investigated the connection of security factor with auditing schemes.

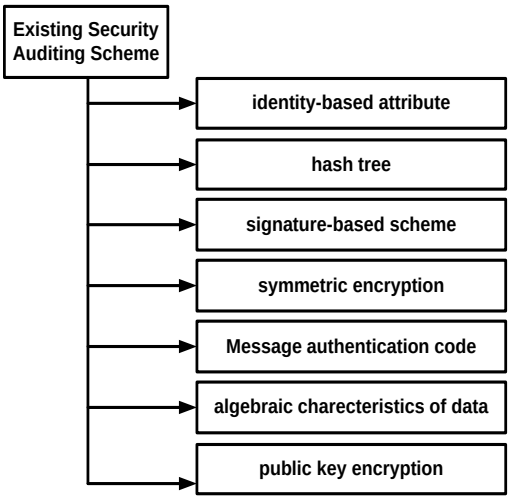

Figure 4. Existing approaches towards security auditing scheme

Usage of the private key for validating the user by the third party is one of the common techniques of security auditing system (Zhang et al. [57]). Such scheme offer benefits on processing time, which is required to assess the scalability of the auditing scheme. As seen from the previous section, deduplication is also witnessed to be frequently used as a standard auditing mechanism over cloud storage (Aujla et al. [58]). Auditing scheme can also be enhanced by using messaging factor as well blockchain. It was found most recently that blockchain offers more privacy and better form of data integrity in the existing auditing scheme (Esposito et al. [59]). Apart from the messaging system, usage of identity factor for auditing offers more data hiding capabilities without affecting data availability. Study towards *identity-based attribute* for cloud auditing has also been carried out by Wang et al. [60] where the technique has been used for outsourcing data. Such scheme facilitates the user to select a secured proxy in order to outsource the data over the server. Identity of such proxy nodes is used for verification, which discards the utilization of certificate over the server (He et al. [61]). A scheme discussed by Shen et al. [62] has used a *signature-based scheme* for validating data integrity while performing remote auditing. Such schemes can be more inclined towards a single attribute of security while multiple attributes of security consideration are highly mandatory to offer data security over distributing storage (Yang et al. [63]). Existing approaches towards auditing scheme as also been focused on using *symmetric encryption* with the capability to verify the outcomes. Such techniques, when integrated with *the hash tree*, offers robust building evidence. It was also noticed that the existing auditing scheme claims of supportability for public users where *public key encryption* plays a dominant role, and thereby publically auditing tool has evolved (Yu et al.[64], Jiang et al.[65]). Usage of the hash tree was also found useful in auditing distributed software framework, e.g., MapReduce (Wang et al. [66]). Constructing a hash table dynamically also facilitates public cloud auditing, but they still suffer from dependencies from third parties (Tian et al. [67]). A recent study carried out by Wang et al. [68] has used public key encryption without any certificate while the scheme is claimed to offer provable possession of data. A similar form of adoption of provable possession of data was also seen in the work of Wu et al. [69]. However, such schemes also suffer from the disclosure of the public key. Yu and Wang [70] present a study addressing this problem. Apart from this, such schemes only support static attribute while the dynamic attribute is highly demanded (Ni et al. [71]). Incorporating flexibility to such an auditing scheme offers more capability to extend its verification process over multiple nodes, too (Jian et al. [72], Ren et al. [73], Zhu et al. [74]). Existing studies have also been carried out considering the mobile users where auditing is facilitated *without any dependency of a third party* (Zhang et al. [75]). It was noted that the usage of proxy re-encryption is quite good enough for resisting threats if they are well defined. Literature has also witnessed a unique approach where the *algebraic charecteristics of data* is computed for carrying out remote auditing of data over cloud storage (Sookhak et al. [76]). Enhancement to the existing data structure in this regard also assists in dynamic auditing data. A similar line of methodology was also carried out by Yuchuan et al. [77] where the algebraic properties of data are dynamically computed for facilitating remote auditing process. The study considers formulating the model using proxy node, cloud, and user where signatures are used in proxy nodes, and data is maintained in cloud storage. Adoption of the trust factor is another evolved scheme facilitating secured auditing procedure over cloud storage. The works of Gonzales et al. [78] have developed a reference model using multi-tenancy. An effective auditing scheme is also presented by controlling

the degree of exposure to the secret key (Yu et al. [54]). Such a scheme offers enhanced forward security and better security assessment model. Adoption of encryption based approach for performing remote auditing of data is more prevalent in the existing literature. *The message authentication code* is reportedly used alongside with homomorphic validation method for data auditing. Utilization of proof of retrievability is another data auditing scheme in the existing system [79]. Existing literature has also explored that if the updates among the storage units over cloud could be securely updated than it could offer better-secured reposition of distributed data over the cloud. This fact was proven by Liu et al. [80] where a *signature*, as well as *the hash tree*, has been used. The work of Yang et al. [81] has taken the shape of a protocol emphasizing over privacy actor while performing auditing while the work of Wang et al. [82] discusses data dynamicity associated with auditing. **Research Gap:** It can be noticed that there has been extensive research contribution focusing on public auditing mechanism, which is mainly carried out remotely. Majority of the schemes offers such verification privilege to users where different encryption and signature schemes are used to secure the auditing operation over cloud storage.

6. OPEN RESEARCH ISSUES

The open research issues are as follows:

- *Unrealistic Assumptions*: Almost majority of the solution towards data integrity problem is carried out by public verification by the user only and not by the service provider. This assumption bounds the user to involve in the verification process with higher communication overhead consistently. Moreover, user cannot be assumed to always possess high configuration computational device and good network resource availability. Another unrealistic assumption of all the approaches of public auditing scheme is that the users (or auditors) are a non-malicious node. It is not always possible to confirm this as normally the users will be have more exposure to the threats compared to service providers and hence if the auditors are from user side than there is no guarantee of its legitimacy.
- *Non-Applicability towards External Intruder*: A closer look into all the existing approaches towards data *integrity*, data privacy, and data deduplication method for secured cloud storage will show that they have been experimented with respect to specific forms of threats. Such threats are mainly internal, and hence, privacy cannot be protected for such data. All these forms of threats are highly capable of bypassing the existing auditing mechanism as it is not cost effective feasibility to construct a secure communication channel during the ongoing auditing process.
- *Computational Cost not Considered*: Practically speaking, all the outsourced data cannot be considered to be safe, which is not discussed in the existing system due to the lack of sufficient physical control over the *outsourced* data. Researchers have also claimed that remote auditing schemes can solve it, but they are not much applicable to the massive scale of data owing to the involvement of large cost. Some of the presented technique claims of supporting updating operation over dynamic data, but such operations are carried out at the cost of the extensive computational burden.
- *Deduplication not focused on Data Integrity*: The existing approaches towards data deduplication have used file level as maximum approaches. All these approaches are found to use convergent encryption as a standard. By doing so, data integrity cannot be ascertained as performing deduplication over the encrypted *file* will require some dependency on the metadata information which was never considered by any researchers. It will mean that the deduplication process in the existing system will only retain privacy to some level at the high computational cost but not the data integrity.

In order to offer better data security over cloud storage, it is necessary to incorporate data integrity, data privacy, and secure data deduplication combined. None of the existing research work is found to offer benchmarked outcome of secured distributed cloud storage till date.

7. FUTURE LINE OF RESEARCH

From the prior section, it was seen that it is quite a challenging process to jointly achieve data integrity and data deduplication in order to incorporate better data privacy over cloud storage. Therefore, better feasibility of implementation of the secure cloud storage system can be carried out using divide and conquer rule. Figure 5 highlights the future line of research to secure distributed cloud storage system.

Following are the brief information of implementation:

- *The strategy of Implementation*: The core strategy of implementation will be to develop two different sub-framework viz. i) framework for offering robust data integrity and ii) framework for secure data deduplication. Both the framework will have a common goal of data privacy incorporated within it. Apart from this, the proposed system also targets to resists the majority of lethal threats over the cloud storage server.

- *The flow of Execution*: The primary step will be to develop the first sub-framework, where users will be offered authority to cross-check the integrity of the data stored in the distributed cloud. A simplified encryption scheme can be developed to store the indexed data, followed by a unique preventive measure. A challenge-based message could also be used for preventing any form of access by the intruders, thereby protecting data integrity and privacy. The secondary step will be to enhance the standard approach of proof of ownership. A novel indexing mechanism can be formulated that maintains consistency over the secure data deduplication process. The existing tree structure can also be modified for facilitation better encryption process over the key. This will assist in the generation of the secret key to be used for data uploading over storage servers resulting in better privacy control. The indexing mechanism can be carried out over block levels, which offers unique deduplication process along with privacy preservation.
- *Anticipated Outcomes*: The anticipated outcome of the proposed study will be to retain a good balance between dynamic intruder resistivity capability and optimal service delivery. The model is also expected to offer both forward and backward secrecy with less computation overhead, unlike any existing system.

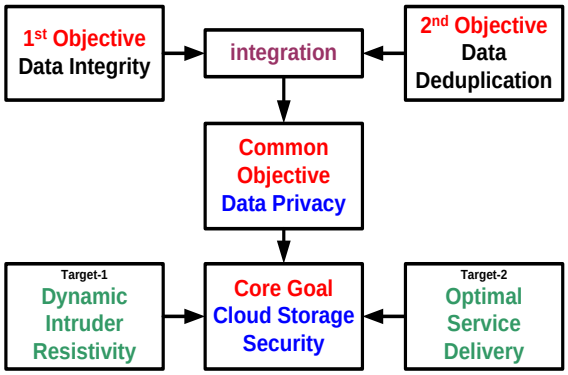

Figure 5. Scheme of feasible implementation

8. CONCLUSION

Offering a higher degree of protection over split data in the storage servers of the cloud system is yet to be achieved. At present, there is various works being carried out towards ensuring data security, but the approaches towards securing a data storage system are quite scattered. It is because efficient and robust cloud data storage will mandatory required to ensure optimal data integrity, data privacy, and data deduplication, which are some elementary operation carried out. Unfortunately, the existing research work is not found to incorporate all the above three points towards evolving up for better storage solution. Therefore, the existing solution always lacks one out of these three points towards a secure data storage system. This manuscript discusses the contribution of recent work being carried out in this direction and briefs of all the open end problems followed by a discussion of a possible way to carry out further research work.

REFERENCES

[1] Naresh Kumar Sehgal, Pramod Chandra P. Bhatt, "Cloud computing: Concepts and practices," *Springer*, 2018.
[2] Dac-Nhuong Le, Raghvendra Kumar, Gia Nhu Nguyen, Jyotir Moy Chatterjee, "Cloud computing and virtualization," *John Wiley & Sons*, 2018.
[3] Zaigham Mahmood, "Cloud computing: Challenges, limitations and R&D solutions," *Springer*, 2014.
[4] H. Zhao and X. Zheng, "A survey on the integrity checking of outsourced data in cloud computing," *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, pp. 1650-1656, 2015.
[5] C. B. O. M. E. Moctar and K. Konaté, "A survey of security challenges in cloud computing," *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 843-849, 2017.
[6] K. N. Sevis and E. Seker, "Survey on data integrity in cloud," *2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud)*, pp. 167-171, 2016.
[7] M. El-Zoghby and M. A. Azer, "Cloud computing privacy issues, challenges and solutions," *2017 12th International Conference on Computer Engineering and Systems (ICCES)*, pp. 154-160, 2017.
[8] Meikang Qiu, Keke Gai, "Mobile cloud computing: Models, implementation, and security," *CRC Press*, 2017.
[9] A. N. Jaber and Mohamad Fadli Bin Zolkipli, "Use of cryptography in cloud computing," *2013 IEEE International Conference on Control System, Computing and Engineering*, pp. 179-184, 2013.

[10] C. Moore, M. O'Neill, E. O'Sullivan, Y. Doröz and B. Sunar, "Practical homomorphic encryption: A survey," *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2792-2795, 2014.
[11] Sooyeon Shin and Taekyoung Kwon, "A survey of public provable data possession schemes with batch verification in cloud storage," *Journal of Internet Services and Information Security (JISIS)*, vol. 5(3), pp. 37-47, 2015.
[12] He Jialing, Zijian Zhang, Meng Li, Liehuang Zhu, and Jingjing Hu. "Provable data integrity of cloud storage service with enhanced security in the internet of things." *IEEE Access*, vol. 7, pp. 6226-6239, 2019.
[13] Sh, Bin, Bo Li, Lei Cui, and Liu Ouyang. "Vanguard: A cache-level sensitive file integrity monitoring system in virtual machine environment," I*EEE Access*, vol. 6, pp. 38567-38577, 2018.
[14] Zhao Bo, Peiru Fan, and Mingtao Ni. "Mchain: A blockchain-based VM measurements secure storage approach in iaas cloud with enhanced integrity and controllability," I*EEE Access*, vol. 6, pp. 43758-43769, 2018.
[15] Wang Huaqun, Debiao He, and Shaohua Tang. "Identity-based proxy-oriented data uploading and remote data integrity checking in public cloud," *IEEE Transactions on Information Forensics and Security*, vol. 11(6), pp. 1165-1176, 2016.
[16] Li Yannan, Yong Yu, Geyong Min, Willy Susilo, Jianbing Ni, and Kim-Kwang Raymond Choo, "Fuzzy identity-based data integrity auditing for reliable cloud storage systems," *IEEE Transactions on Dependable and Secure Computing*, 2017.
[17] Shao Bilin, Genqing Bian, Yue Wang, Shenghao Su, and Cheng Guo, "Dynamic data integrity auditing method supporting privacy protection in vehicular cloud environment," *IEEE Access*, vol. 6, pp. 43785-43797, 2018.
[18] Tan Shuang, Lin Tan, Xiaoling Li, and Yan Jia, "An efficient method for checking the integrity of data in the cloud," *China Communications*, vol. 11(9), pp. 68-81, 2014.
[19] Wang Boyang, Hui Li, Xuefeng Liu, Fenghua Li, and Xiaoqing Li, "Efficient public verification on the integrity of multi-owner data in the cloud," *Journal of Communications and Networks*, vol. 16(6), pp. 592-599, 2014.
[20] Hu Ling, Wei-Shinn Ku, Spiridon Bakiras, and Cyrus Shahabi, "Spatial query integrity with voronoi neighbors," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25(4), pp. 863-876, 2013.
[21] Du Juan, Daniel J. Dean, Yongmin Tan, Xiaohui Gu, and Ting Yu, "Scalable distributed service integrity attestation for software-as-a-service clouds," *IEEE Transactions on parallel and distributed systems*, vol. 25(3), pp. 730-739, 2014.
[22] Lu Yu, and Fei Hu, "Secure dynamic big graph data: Scalable, low-cost remote data integrity checking," *IEEE*, vol. 7, pp. 12888-12900, 2019.
[23] Chen Henry CH, and Patrick PC Lee, "Enabling data integrity protection in regenerating-coding-based cloud storage: Theory and implementation," *IEEE transactions on parallel and distributed systems*, vol. 25(2), pp. 407-416, 2014.
[24] Fan Xinyu, Guomin Yang, Yi Mu, and Yong Yu, "On indistinguishability in remote data integrity checking," *The Computer Journal*, vol. 58(4), pp. 823-830, 2013.
[25] Shen Shiuan-Tzuo, Hsiao-Ying Lin, and Wen-Guey Tzeng, "An effective integrity check scheme for secure erasure code-based storage systems," *IEEE Transactions on reliability*, vol. 64(3), pp. 840-851, 2015.
[26] Du, Ruizhong, Wangyang Pan, and Junfeng Tian, "Dynamic integrity measurement model based on vTPM," *China Communications*, vol. 15(2), pp. 88-99, 2018.
[27] Yu, Yong, Man Ho Au, Giuseppe Ateniese, Xinyi Huang, Willy Susilo, Yuanshun Dai, and Geyong Min, "Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage," *IEEE Transactions on Information Forensics and Security*, vol. 12(4), pp. 767-778. 2017.
[28] Zhang,Yuan, Chunxiang Xu, Xiaohui Liang, Hongwei Li, Yi Mu, and Xiaojun Zhang, "Efficient public verification of data integrity for cloud storage systems from indistinguishability obfuscation," *IEEE Transactions on Information Forensics and Security*, vol. 12(3), pp. 676-688, 2017.
[29] Zhu Yan, Hongxin Hu, Gail-Joon Ahn, and Mengyang Yu, "Cooperative provable data possession for integrity verification in multicloud storage," *IEEE transactions on parallel and distributed systems*, vol. 23(12), pp. 2231-2244, 2012.
[30] Alabdulatif Abdulatif, Heshan Kumarage, Ibrahim Khalil, Mohammed Atiquzzaman, and Xun Yi, "Privacy-preserving cloud-based billing with lightweight homomorphic encryption for sensor-enabled smart grid infrastructure," *IET Wireless Sensor Systems*, vol. 7(6), pp. 182-190, 2017.
[31] Tang Xin, Yongfeng Huang, Chin-Chen Chang, and Linna Zhou, "Efficient real-time integrity auditing with privacy-preserving arbitration for images in cloud storage system," *IEEE Access*, 2019.
[32] Du Minxin, Qian Wang, Meiqi He and Jian Weng, "Privacy-preserving indexing and query processing for secure dynamic cloud storage," *IEEE Transactions on Information Forensics and Security*, vol. 13(9), pp. 2320-2332, 2018.
[33] Hu Pengfei, Huansheng Ning, Tie Qiu, Houbing Song, Yanna Wang, and Xuanxia Yao, "Security and privacy preservation scheme of face identification and resolution framework using fog computing in internet of things," *IEEE Internet of Things Journal*, vol. 4(5), pp. 1143-1155, 2017.
[34] Wang Tian, Jiyuan Zhou, Xinlei Chen, Guojun Wang, Anfeng Liu, and Yang Liu, "A three-layer privacy preserving cloud storage scheme based on computational intelligence in fog computing," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2(1), pp. 3-12, 2018.
[35] Liu Jian, Kun Huang, Hong Rong, Huimei Wang, and Ming Xian, "Privacy-preserving public auditing for regenerating-code-based cloud storage," *IEEE transactions on information forensics and security*, vol. 10(7), pp. 1513-1528, 2015.
[36] Wang Boyang, Baochun Li, and Hui Li, "Oruta: Privacy-preserving public auditing for shared data in the cloud," *IEEE transactions on cloud computing*, vol. 2(1), pp. 43-56, 2014.

[37] Yu Yong, Man Ho Au, Giuseppe Ateniese, Xinyi Huang, Willy Susilo, Yuanshun Dai, and Geyong Min, "Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage," *IEEE Transactions on Information Forensics and Security*, vol. 12(4), pp. 767-778, 2017
[38] Li Jiangtao, Lei Zhang, Joseph K. Liu, Haifeng Qian, and Zheming Dong, "Privacy-preserving public auditing protocol for low-performance end devices in cloud," *IEEE Transactions on Information Forensics and Security*, vol. 11(11), pp. 2572-2583, 2016.
[39] Wang Cong, Sherman SM Chow, Qian Wang, Kui Ren, and Wenjing Lou, "Privacy-preserving public auditing for secure cloud storage," *IEEE transactions on computers* vol. 62(2), pp. 362-375, 2013.
[40] Hao Zhuo, Sheng Zhong, and Nenghai Yu, "A privacy-preserving remote data integrity checking protocol with data dynamics and public verifiability," *IEEE transactions on Knowledge and Data Engineering*, vol. 23(9), pp. 1432-1437, 2011.
[41] Daehee Kim, Sejun Song, Baek-Young Choi, "Data deduplication for data optimization for storage and network systems," *Sprnger*, 2016
[42] Robert Deng, Jian Weng, Kui Ren, Vinod Yegneswaran, "Security and privacy in communication networks," *Springer*, 2017
[43] Youn Taek-Young, Ku-Young Chang, Kyung-Hyune Rhee, and Sang Uk Shin, "Efficient client-side deduplication of encrypted data with public auditing in cloud storage," *IEEE Access*, vol. 6, pp. 26578-26587, 2018.
[44] Hur Junbeom, Dongyoung Koo, Youngjoo Shin, and Kyungtae Kang, "Secure data deduplication with dynamic ownership management in cloud storage," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28(11), pp. 3113-3125, 2016.
[45] He Kun, Jing Chen, Ruiying Du, Qianhong Wu, Guoliang Xue, and Xiang Zhang, "Deypos: Deduplicatable dynamic proof of storage for multi-user environments," *IEEE Transactions on Computers*, vol. 65(12), pp. 3631-3645, 2016.
[46] Jiang Tao, Xiaofeng Chen, Qianhong Wu, Jianfeng Ma, Willy Susilo, and Wenjing Lou, "Secure and efficient cloud data deduplication with randomized tag," *IEEE Transactions on Information Forensics and Security*, vol. 12(3), pp. 532-543, 2017.
[47] Stanek Jan, and Lukas Kencl, "Enhanced secure thresholded data deduplication scheme for cloud storage," *IEEE Transactions on Dependable and Secure Computing*, vol. 15(4), pp. 694-707, 2018.
[48] Zheng Yifeng, Xingliang Yuan, Xinyu Wang, Jinghua Jiang, Cong Wang, and Xiaolin Gui, "Toward encrypted cloud media center with secure deduplication," *IEEE Transactions on Multimedia*, vol. 19(2), pp. 251-265, 2017.
[49] Li Jin, Xiaofeng Chen, Xinyi Huang, Shaohua Tang, Yang Xiang, Mohammad Mehedi Hassan, and Abdulhameed Alelaiwi, "Secure distributed deduplication systems with improved reliability," *IEEE Transactions on Computers*, vol. 64(12), pp. 3569-3579, 2015.
[50] Li Jingwei, Jin Li, Dongqing Xie, and Zhang Cai, "Secure auditing and deduplicating data in cloud," *IEEE Transactions on Computers*, vol. 65(8), pp. 2386-2396, 2016.
[51] Yan Zheng, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, "Deduplication on encrypted big data in cloud," *IEEE transactions on big data*, vol. 2(2), pp. 138-150, 2016.
[52] Fan Chun-I., Shi-Yuan Huang, and Wen-Che Hsu, "Encrypted data deduplication in cloud storage," *2015 10th Asia Joint Conference on Information Security*, pp. 18-25, 2015.
[53] Madria Sanjay K., "Security and risk assessment in the cloud," *Computer*, vol. 49(9), pp. 110-113, 2016.
[54] Jian X. U., L. I. Mingjie, L. I. Fuxiang, Y. A. N. G. Qingsong, and Z. H. O. U. Fucai, "Optimized algorithms for flexible length-based authenticated skip list," *China Communications*, vol. 13(1) pp. 124-138, 2016.
[55] Yu Jia, Kui Ren, Cong Wang, and Vijay Varadharajan, "Enabling cloud storage auditing with key-exposure resistance," *IEEE Transactions on Information forensics and security*, vol. 10(6), pp. 1167-1179, 2015.
[56] Nepal Surya, Rajiv Ranjan, and Kim-Kwang Raymond Choo, "Trustworthy processing of healthcare big data in hybrid clouds," *IEEE Cloud Computing*, vol. 2(2) pp. 78-84, 2015.
[57] Zhang Yue, Hanlin Zhang, Rong Hao, and Jia Yu, "Authorized identity-based public cloud storage auditing scheme with hierarchical structure for large-scale user groups," *China Communications*, vol. 15(11), pp. 111-121, 2018.
[58] Aujla Gagangeet Singh, Rajat Chaudhary, Neeraj Kumar, Ashok Kumar Das, and Joel JPC Rodrigues, "SecSVA: Secure storage, verification, and auditing of big data in the cloud environment," *IEEE Communications Magazine*, vol. 56(1), pp. 78-85, 2018.
[59] Esposito Christian, Francesco Palmieri, and Kim-Kwang Raymond Choo, "Cloud message queueing and notification: Challenges and opportunities," *IEEE Cloud Computing*, vol. 5(2), pp. 11-16, 2018.
[60] Wang Yujue, Qianhong Wu, Bo Qin, Wenchang Shi, Robert H. Deng, and Jiankun Hu, "Identity-based data outsourcing with comprehensive auditing in clouds," *IEEE transactions on information forensics and security*, vol. 12(4), pp. 940-952, 2017.
[61] He Debiao, Sherali Zeadally, and Libing Wu, "Certificateless public auditing scheme for cloud-assisted wireless body area networks," *IEEE Systems Journal*, vol. 12(1), pp. 64-73, 2018.
[62] Shen Wenting, Jing Qin, Jia Yu, Rong Hao, and Jiankun Hu, "Enabling identity-based integrity auditing and data sharing with sensitive information hiding for secure cloud storage," *IEEE Transactions on Information Forensics and Security*, vol. 14(2), pp. 331-346, 2019.
[63] Yang Yuli, Rui Liu, Yongle Chen, Tong Li, and Yi Tang, "Normal cloud model-based algorithm for multi-attribute trusted cloud service selection," *IEEE Access*, vol. 6, pp. 37644-37652, 2018.
[64] Yu Yong, Jianbing Ni, Man Ho Au, Yi Mu, Boyang Wang, and Hui Li, "On the security of a public auditing mechanism for shared cloud data service," *IEEE Transactions on Services Computing*, vol. 8(6), pp. 998-999, 2014.
[65] Jiang Tao, Xiaofeng Chen, and Jianfeng Ma, "Public integrity auditing for shared dynamic cloud data with group user revocation," *IEEE Transactions on Computers*, vol. 65(8), pp. 2363-2373, 2016.

[66] Wang Yongzhi, Yulong Shen, Hua Wang, Jinli Cao, and Xiaohong Jiang, "MtMR: Ensuring mapreduce computation integrity with merkle tree-based verifications," *IEEE Transactions on Big Data*, vol. 4(3), pp. 418-431, 2018.
[67] Tian Hui, Yuxiang Chen, Chin-Chen Chang, Hong Jiang, Yongfeng Huang, Yonghong Chen, and Jin Liu, "Dynamic-hash-table based public auditing for secure cloud storage," *IEEE Transactions on Services Computing*, vol. 10(5), pp. 701-714, 2017.
[68] Wang Feng, Li Xu, and Wei Gao, "Comments on "SCLPV: Secure certificateless public verification for cloud-based cyber-physical-social systems against malicious auditors'," *IEEE Transactions on Computational Social Systems*, vol. 99, pp. 1-4, 2018.
[69] Wu Tsu-Yang, Yuh-Min Tseng, Sen-Shan Huang, and Yi-Chen Lai, "Non-repudiable provable data possession scheme with designated verifier in cloud storage systems," *IEEE Access*, vol. 5, pp. 19333-19341, 2017.
[70] Yu Jia, and Huaqun Wang, "Strong key-exposure resilient auditing for secure cloud storage," *IEEE Transactions on Information Forensics and Security*, vol. 12(8), pp. 1931-1940, 2017.
[71] Ni Jianbing, Yong Yu, Yi Mu, and Qi Xia, "On the security of an efficient dynamic auditing protocol in cloud storage," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25(10), pp. 2760-2761, 2014.
[72] Jian X. U., L. I. Mingjie, L. I. Fuxiang, Y. A. N. G. Qingsong, and Z. H. O. U. Fucai, "Optimized algorithms for flexible length-based authenticated skip list," *China Communications*, vol. 13(1), pp. 124-138, 2016.
[73] Ren Zhengwei, Lina Wang, Qian Wang, and Mingdi Xu, "Dynamic proofs of retrievability for coded cloud storage systems," *IEEE Transactions on Services Computing*, vol. 11(4), pp. 685-698, 2018.
[74] Zhu Yan, Gail-Joon Ahn, Hongxin Hu, Stephen S. Yau, Ho G. An, and Chang-Jun Hu, "Dynamic audit services for outsourced storages in clouds," *IEEE Transactions on Services Computing*, vol. 6(2), pp. 227-238, 2013.
[75] Zhang Jiang, Zhenfeng Zhang, and Hui Guo, "Towards secure data distribution systems in mobile cloud computing," *IEEE Transactions on Mobile Computing*, vol. 16(11), pp. 3222-3235, 2017.
[76] Sookhak Mehdi, F. Richard Yu, and Albert Y. Zomaya, "Auditing big data storage in cloud computing using divide and conquer tables," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29(5), pp. 999-1012, 2018.
[77] Yuchuan Luo, Fu Shaojing, Xu Ming, and Wang Dongsheng, "Enable data dynamics for algebraic signatures based remote data possession checking in the cloud storage," *China Communications*, vol. 11(11), pp. 114-124, 2014.
[78] Gonzales Dan, Jeremy M. Kaplan, Evan Saltzman, Zev Winkelman, and Dulani Woods, "Cloud-trust-A security assessment model for infrastructure as a service (IaaS) clouds," *IEEE Transactions on Cloud Computing*, vol. 5(3), pp. 523-536, 2017.
[79] Li Jin, Xiao Tan, Xiaofeng Chen, Duncan S. Wong, and Fatos Xhafa, "OPoR: Enabling proof of retrievability in cloud computing with resource-constrained devices," *IEEE Transactions on cloud computing*, vol. 3(2), pp. 195-205, 2015.
[80] Liu Chang, Jinjun Chen, Laurence T. Yang, Xuyun Zhang, Chi Yang, Rajiv Ranjan, and Ramamohanarao Kotagiri, "Authorized public auditing of dynamic big data storage on cloud with efficient verifiable fine-grained updates," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25(9), pp. 2234-2244, 2014.
[81] Yang Kan, and Xiaohua Jia, "An efficient and secure dynamic auditing protocol for data storage in cloud computing," *IEEE transactions on parallel and distributed systems*, vol. 24(9), pp. 1717-1726, 2013.
[82] Wang Qian, Cong Wang, Kui Ren, Wenjing Lou, and Jin Li, "Enabling public auditability and data dynamics for storage security in cloud computing," *IEEE transactions on parallel and distributed systems*, vol. 22(5), pp. 847-859, 2011.

BIOGRAPHIES OF AUTHORS

Mr. Anil Kumar G is Research Scholar in Computer Science and Engineering department of Channabasaveshwara Institute of Technology at Visvesvarahya Technological University. He perused his bachelor degree in Computer Science & Engineering from Gulburga University, Karnataka, India and masters in Computer Science & Engineering from Dr. MGR Educational Research Institute, Chennai, India. Mr. Anil Kumar is having good academic and research experience in the area of Computer Networks, Unix Systems Programming, Cloud Computing with good number of publications.

Dr. Shantala C P is Professor & HOD in Computer Science and Engineering department of Channabasaveshwara Institute of Technology at Visvesvaraya Technological University. She is vice principal of Channabasaveshwara Institute of Technology. She has completed her PhD in the area of Data Security and Masters in Computer Science & Engineering. Her research interests lie in the areas of Network & Data Security, Cloud Storage, Data Mining & Brain Computer Interface. Her research works brought her various awards like Seed Money for Young Scientist from VGST & Women Achiever Award from IEI.