Cloud based Secure Tag-flexible De-duplication supported Integrity Check Protocol (TDICP) Protocol to Eliminate Duplicate Files

***Miss.V.Nirmala[1], Mr.D.Sudhakar[2]**

[1]Assistant Professor, Dept .of MCA, Annamacharya PG College of Computer Studies, Rajampet ,India,

[2]Assistant Professor, Dept.of MCA, Annamacharya PG College of Computer Studies,Rajampet,India,

ABSTRACT: Since cloud computing is a metaphor for the Internet, using cloud-based services for large-scale information delivery, processing, and storage has become popular in recent years. To reduce volume in the storage area network, the data de duplication technique finds recurring data forms and reduces them to a single instance. This study paper discusses secure de duplication for text in order to better provide secure de duplication in the cloud. In particular, we introduce a novel verification tag called note set, which permits multiple users holding the same file to generate their unique verification tags and still supports tag de duplication at the CSP. This allows us to propose a novel Tag-flexible De duplication-supported Integrity Check Protocol (TDICP) based on Private Information Retrieval (PIR). Additionally, we introduce a novel User Determined Duplication Check Protocol (UDDCP) based on Private Set Intersection (PSI) that can prevent a CSP from giving users a false duplication check result, marking the first attempt to ensure the accuracy of data duplication check. Security study verifies the accuracy and stability of our plan. The effectiveness and efficiency of our suggested method, as well as its notable advantages over previous arts, are demonstrated by simulation experiments based on actual data.

KEYWORDS: Integrity Check, Duplication Check, Private Information Retrieval, Data De duplication, Cloud Computing, Verifiable Computation.

I.INTRODUCTION

Cloud computing has become a popular information technology service by providing a huge amount of resources (e.g., storage and computing) to end users based on their demands. Among all cloud computing services, cloud storage is the most popular. Since the volume of data in the world is increasing rapidly, saving cloud storage becomes essential. One of the key reasons that causes storage waste is duplicate data storage. Multiple users may save the same files or different files containing the same pieces of data blocks at the cloud. Obviously, duplicate data storage in the cloud introduces a big waste of storage resources. Data de duplication provides a promising solution to this issue. In a de duplication scheme, the CSP can cooperate with the cloud user to first check whether a pending uploaded file has been saved already or not, and then provide the user whose pieces of file data are checked duplicate a way to access the file without storing another copy at the cloud. However, since the CSP cannot be fully trusted, the cloud users may suffer from some security and privacy issues. Notably, a semi- trusted CSP may modify, tamper or delete the uploaded data driven by some profits. The damage of de duplicated data could cause huge loss to all related users

(e.g., data owners and holders). Thus, the integrity of the data stored at the cloud should be verified, especially for duplicate data storage with de duplication. Several Proof of Re trievability (PoR) schemes have been proposed to address the issue of integrity check on cloud data storage in recent decades. In such schemes, a user adds verification tags along with a file. During the verification, the user creates a random challenge and sends it to the CSP, the CSP has to use all the data in the user's corresponding files it stored as inputs to compute a response back to the user. The user then checks the integrity of the stored file by verifying the response. However, existing PoR solutions mainly aim to improve the performance at the user side and assume that the CSP has infinite computation and storage resources. While, in practice, the CSP performs data de duplication in order to achieve the most economic usage of its storage. Unfortunately, existing solutions mentioned above are incompatible with de duplication. This is because the verification tags of these schemes are created with user individual private keys unknown to each other, thus different verification tags are generated, given the same file held by different users. But these verification tags cannot be de duplicated at the CSP.

II. LITERATURE SURVEY

[1] Z. Yan, L. F. Zhang, W. X. Ding, and Q. H. Zheng, “Heterogeneous data storage management with de duplication in cloud computing,” IEEE Transactions on Big Data, pp. 1–1, 2017. Previous work cannot check integrity of de duplicated encrypted data at the cloud and ensure the correctness of duplication check during data upload. In this paper, we propose a verifiable de duplication scheme called VeriDedup to address the above problems. It can guarantee the correctness of duplication check and support flexible tag generation for integrity check over encrypted data de duplication in an integrative way. Concretely, we propose a novel Tag-flexible De duplication-supported Integrity Check Protocol (TDICP) based on Private Information Retrieval (PIR) by introducing a novel verification tag called note set, which allows multiple users holding the same file to generate their individual verification tags and still supports tag de duplication at a Cloud Storage Provider (CSP). Furthermore, we make the first attempt to guarantee the correctness of data duplication check by introducing a novel User Determined Duplication Check Protocol (UDDCP) based on Private Set Intersection (PSI), which can resist the CSP from providing a fake duplication check result to users. Security analysis shows the correctness and soundness of our scheme. Simulation studies based on real data show the efficacy and efficiency of our proposed scheme and its significant advantages over prior arts.

[2] Z. Yan, W. X. Ding, and H. Q. Zhu, “A scheme to manage encrypted data storage with de duplication in cloud,” in International Conference on Algorithms and Architectures for Parallel Processing, 2015. Secured de-duplication, with the approach of distributed computing, has pulled in much consideration as of late from research group. The most imperative and well known cloud administration is information stockpiling. With a specific end goal to save the security of the information holder, the information is normally put away in the encoded frame. Conventional de duplication doesn't take a shot at scrambled information. This paper speaks to private information de duplication, one of the effective pressure strategies, in half and half cloud. This strategy is generally utilized as a part of cloud

to limit the information storage room and transmission capacity with a specific end goal to ad lib the effectiveness. The secrecy is the greatest test in the de duplication system. The delicate information must be shielded from outer assaults, hence these information needs to experience encryption before outsourcing. To defeat the security shortcoming, this paper gives the model to secured de duplication in a group cloud.

[3] Z. Yan, M. J. Wang, Y. X. Li, and A. V. Vasilakos, "Encrypted data management with deduplication in cloud computing," IEEE Cloud Computing, vol. 3, no. 2, pp. 28–35, 2016. In Cloud computing the most significant services are cloud big data storage and cloud computing Facilitates cloud users to expand the data storage without upgrading their devices. In this paper, we propose a flexible heterogeneous big data storage management scheme to offer both de duplication management and access control simultaneously across multiple Cloud Service Providers (CSPs). We evaluate the performance of the proposed scheme with security analysis, comparison and implementation. The implementation and analysis results show its effectiveness, security and efficiency towards potential practical usage. The cloud computing based public auditing data preserving refers to the process of conducting audits on data stored in the cloud to ensure its integrity, authenticity, and confidentiality are maintained. It involves allowing external entities, such as auditors or regulators, to verify the correctness of data without compromising its security. In this paper, we proposed a scheme for shared data that supports privacy, identity traceability and group dynamics. Cloud storage auditing is an extremely important technique for resolving the problem of ensuring the integrity of stored data in cloud storage. This scheme is secure against collusion attacks between CSPs and revoked users. This concept is for public auditing with secure group management. This scheme is useful for auditing purposes. In this scheme the auditing details the auditor will upload the files to the cloud and the one group management will maintain a detail in the secure management. This group contains a Third party auditor one who uploads the auditing details, a Group manager one who manages the group, Group members who work in that group and Cloud Service Provider that is providing the cloud service to the group management and will maintain the data. In this, the security will be provided to the group manager and group members so that the information will be secure before as by OTP process and after the data will be secure through attribute-based encryption of the process. As one of the first large enterprises to move entirely to the public cloud, we've invested in our in-house cloud software tech for almost a decade and continue to build at the cutting edge of cloud technology. Our use of the most advanced cloud services enables our developers to focus less on managing infrastructure and more on building great applications, data products, and machine learning solutions for customers and our business.

Existing Algorithm

The cloud computing based public auditing data preserving refers to the process of conducting audits on data stored in the cloud to ensure its integrity, authenticity, and confidentiality are maintained. It involves allowing external entities, such as auditors or regulators, to verify the correctness of data without compromising its security. In this paper, we proposed a scheme for shared data that supports privacy, identity traceability and group dynamics. Cloud storage auditing is an extremely important technique for resolving the

problem of ensuring the integrity of stored data in cloud storage. This scheme is secure against collusion attacks between CSPs and revoked users. This concept is for public auditing with secure group management. This scheme is useful for auditing purposes. In this scheme the auditing details the auditor will upload the files to the cloud and the one group management will maintain a detail in the secure management. This group contains a Third party auditor one who uploads the auditing details, a Group manager one who manages the group, Group members who work in that group and Cloud Service Provider that is providing the cloud service to the group management and will maintain the data. In this, the security will be provided to the group manager and group members so that the information will be secure before as by OTP process and after the data will be secure through attribute-based encryption of the process. As one of the first large enterprises to move entirely to the public cloud, we've invested in our in-house cloud software tech for almost a decade and continue to build at the cutting edge of cloud technology. Our use of the most advanced cloud services enables our developers to focus less on managing infrastructure and more on building great applications, data products, and machine learning solutions for customers and our business.

III.METHODOLOGY AND DISCUSSION

VeriDedup offers guarantee on the correctness of duplication check and supports the integrity check of de duplicated encrypted data in cloud storage. Our target system contains three types of entities: a Data holder who owns data and saves its data that consists of multiple blocks at CSP. It is possible that a number of eligible data holders share the same encrypted data blocks in the CSP. In particular, the data holder that first uploads the data blocks to the CSP is denoted as a data owner with regard to the same blocks. CSP who provides a data storage service with de duplication to data holders. Only one data copy is stored at the CSP, which can be accessed by all data holders with authority. Authenticated auditor (AA) who serves as a third party to check data ownership, authorize data access and cooperate with other two types of entities aiming to audit the whole procedure of data duplication check. The system model of VeriDedup. We perform our research based on the following assumptions. We assume that the data holder is honest. We assume the CSP is semi-trusted. It may raise the following three security threats: Snooping the private data of the data holders cheating the data holders by providing a wrong duplication check result in order to ask for a higher storage fee Causing data loss due to carelessness of data maintenance. In VeriDedup, we focus on the last two issues since many existing solutions of the first issue can be found in the literature. Thus, we assume that the first issue has been solved, e.g., through data encryption. In addition, we assume AA and CSP do not collide. However, AA is semi-trusted, which is curious about the data stored at the cloud, thus private data should be kept away from AA. We assume data holders, CSP, and AA communicate with each other through secure channels by applying some security protocol (e.g., Open-Secure Sockets Layer (SSL)). And all system parameters are shared with all related parties during system setup or initialization phase in a secure way. VeriDedup follows the construction of our previous de duplication scheme and improves it by using PSI and PIR to ensure both data integrity and duplication check correctness over encrypted data de duplication. Specifically, compared with previous work, we introduce a PSI based challenge and response mechanism to the duplication check

procedure in order to let the data holder first tell whether the uploaded blocks are duplicate or not instead of the CSP. In addition, we employ AA to verify the computations of the CSP during the duplication check, so that the CSP cannot cheat the users to upload data blocks that have been stored already. Furthermore, we propose a note insertion mechanism based on PIR to let the data holder insert a specific set called note set that contains several randomized bit sequences, which conform to a hidden function, as verification tags into the encrypted blocks of an uploaded file. The data owners/holders who are proved to have the ownership of the corresponding blocks can verify the integrity of the uploaded blocks through a challenge on whether the notes conform to the hidden function. Attention need be given that the verification tags generated by multiple data holders with various notes can also be de duplicated in VeriDedup, so that the CSP will no longer be required to maintain multiple pieces of verification tags from the same block of different data holders for integrity check, which reduces storage consumption of performing de duplication. In what follows, we first introduce the two proposed novel protocols (i.e., TDICP and UDDCP) and then detail the whole construction of VeriDedup.

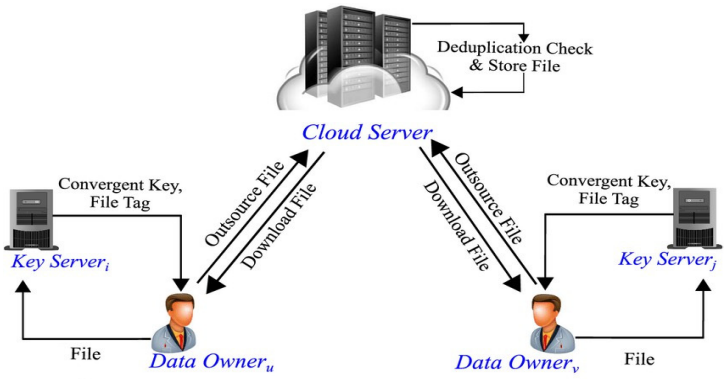

*Fig 1:**SystemArchitecture***

VeriDedup offers guarantee on the correctness of duplication check and supports the integrity check of de duplicated encrypted data in cloud storage. The Proposed Tag-flexible De duplication-supported Integrity Check Protocol (TDICP) is aimed at enhancing the security of De duplication processes in cloud environments, specifically for text data. In response to the increasing reliance on cloud-based services for content storage, processing, and distribution, the protocol addresses the need for efficient and secure data de duplication strategies. At the TDICP lies the concept of Private Information Retrieval (PIR), which allows users to retrieve information from a database without revealing which specific items they are accessing. TDICP introduces a novel verification tag termed as the “note set”, which enables multiple users holding the same file to generate individual verification tags. This innovation facilitates tag de duplication at the Cloud Service Provider (CSP), thereby optimizing storage efficiency while preserving data integrity. Moreover, to ensure the accuracy of data duplication checks, TDICP incorporates the User Determined Duplication Check Protocol (UDDCP) based on Private Set Intersection (PSI). This protocol prevents the CSP from furnishing falsified duplication check results to users, thus bolstering the overall

security of the de duplication process. Moreover, to ensure the accuracy of data duplication checks, TDICP incorporates the User Determined Duplication Check Protocol (UDDCP) based on Private Set Intersection (PSI). This protocol prevents the CSP from furnishing falsified duplication check results to users, thus bolstering the overall security of the de duplication process.

Proposed Algorithm

Step 1: Initialize Variables and Parameters

Step 2: Receive Data for De duplication

Step 3: Determine Duplication Check Method

Step 4: Duplicate Check If hash-based duplication check if content-based duplication check

Step 5: Store or Discard Data

Step 6: Update Metadata

Step 7: Monitor and Maintain System

IV. RESULTS

Impact of note ratio: Fig 2 to Fig 4 shows note insertion cost, integrity check cost, and note removing cost of our scheme with the note ratio varying from 0.02 to 0.10 and notes size of 32 KB, 64 KB, and 128 KB, respectively. As we can see, the larger the note size is, the higher the note insertion cost, integrity check cost, and note removing cost, which is the same as our expectation. When the note ratio increases, all these costs increase linearly since our meta verification block is a note set that contains 4 notes that conform to the hidden function. The increase of note ratio causes the increase of operation time regarding inserting, verifying, and removing those similar verification blocks.

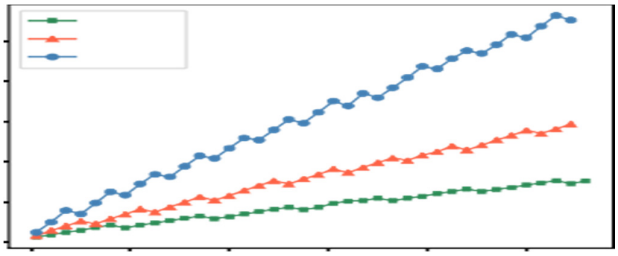

Fig.2: Inserting Note Cost

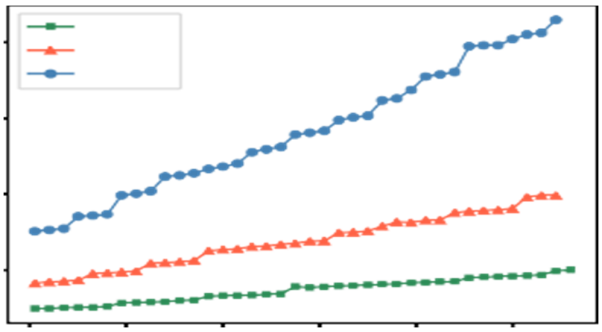

Fig.3:IntegrityCheck Cost

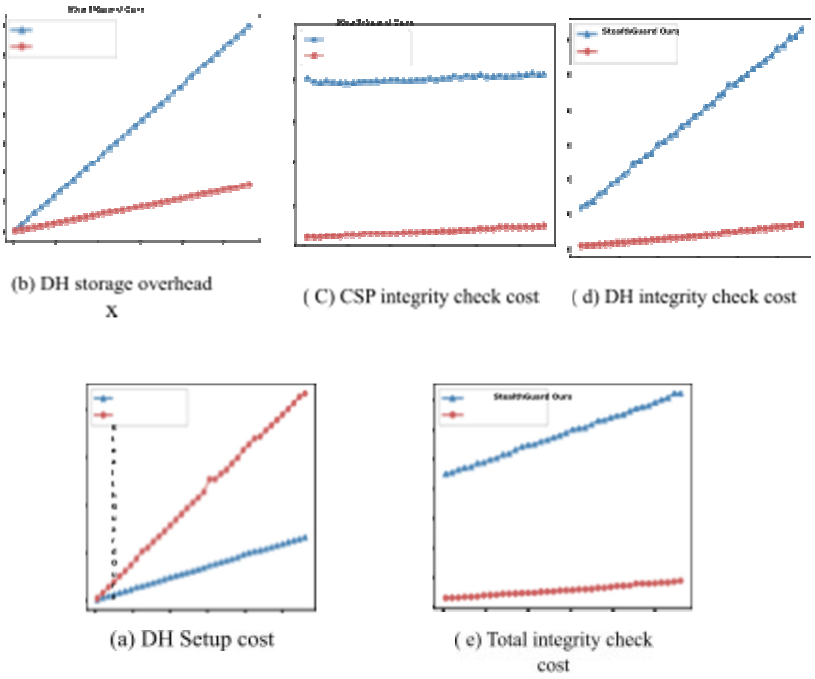

Fig 4 (a) to Fig 4 (e): Costs and storage over head

Fig 4 (a) to Fig 4 (e) shows the setup cost, the data holder storage overhead, the CSP integrity check cost, the data holder integrity check cost , and the total integrity check cost of TDICP with regard to the size of notes (watchdogs) varying from 2 KB to 14 KB compared with StealthGuard.

Fig 4 (a) compares the setup cost of our scheme with StealthGuard. The setup cost increases as the size of tag increases in both schemes as expected. As we can see, TDICP incurs a higher computation cost than the StealthGuard at the setup phase. The reason is that TDICP needs to additionally perform multiple HASH operations and permutations than the StealthGuard.

Fig 4 (b) compares the storage overhead of TDICP with Stealth Fraud at the data holder. StealthGuard incurs higher storage overhead since it requires the data holder to record all the

watchdogs and TDICP requires the data holder to store the position index P of the notes whose size is smaller than that of the watchdogs.

Fig 4 (c) compares the CSP cost of TDICP with StealthGuard. StealthGuard incurs higher computation cost since it requires the CSP to transfer the data into an 80 bits matrix, which increases the times of multiplication executed at the CSP.

Fig 4 (d) compares the data holder cost of TDICP with StealthGuard. We can see that StealthGuard incurs higher computation cost since Stealth Gerard requires the data holder to perform more computations on extracting the verification tags from the response.

As a total, Fig 4 (e) concludes and compares the total cost of checking the integrity of a 128KB file with StealthGuard. We can see that TDICP outperforms StealthGuard with respect to the computation cost in both the CSP and the data holder side, and all of those costs increase as the size of notes (watchdogs) increases.

V.CONCLUSION

In this paper, we introduced VeriDedup to check the integrity of an outsourced encrypted file and guarantee the correctness of duplication check in an integrated way. The integrity check proto- col TDICP of VeriDedup allows multiple data holders to verify the integrity of their outsourced file with their own individual verification tags without interacting with the data owner. On the other hand, we employed a novel challenge and response mechanism in the duplication check protocol UDDCP of VeriDedup to let the data holder instead of the CSP first tell whether a file is duplicate in order to guarantee the correctness of the duplication check. Security and performance analysis show that VeriDedup is secure and efficient under the described security model. The result of our computer simulation further shows its efficiency compared with highly related prior arts.

REFERENCES

[1] Z. Yan, L. F. Zhang, W. X. Ding, and Q. H. Zheng, “Heterogeneous data storage management with deduplication in cloud computing,” IEEE Transactions on Big Data, pp. 1–1, 2017.

[2] Z. Yan, W. X. Ding, and H. Q. Zhu, “A scheme to manage encrypted data storage with deduplication in cloud,” in International Conference on Algorithms and Architectures for Parallel Processing, 2015.

[3] Z. Yan, M. J. Wang, Y. X. Li, and A. V. Vasilakos, “Encrypted data management with deduplication in cloud computing,” IEEE Cloud Com- puting, vol. 3, no. 2, pp. 28–35, 2016.

[4] W. Shen, Y. Su, and R. Hao, “Lightweight cloud storage auditing with de duplication supporting strong privacy protection,” IEEE Access, vol. 8, pp. 44 359–44 372, 2020.

[5] Q. Zheng and S. Xu, “Secure and efficient proof of storage with de duplication,” in CODASPY ’12, New York, NY, USA, 2012, p. 1–12.

[6] A. Giuseppe, R. Burns, and C. Reza, “Provable data possession at un- trusted stores,” in Proceedings of the 14th ACM Conference on Computer and Communications Security, 2007, pp. 598–609.

[7] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, Z. Peterson, and D. Song, “Remote data checking using provable data possession,” ACM Transactions on Information and System Security, vol. 14, pp. 1– 34, 2011.

[8] Z. Wen, J. Luo, H. Chen, J. Meng, X. Li, and J. Li, “A verifiable data deduplication scheme in cloud computing,” in INCOS ’14, USA, 2014, p. 85–90.

[9] P. Meye, P. Raïpin, F. Tronel, and E. Anceaume, “A secure two-phase data deduplication scheme,” in HPCC ’14, CSS ’14, ICESS ’14, 2014, pp. 802–809.

[10]D. Vasilopoulos, M. Önen, K. Elkhiyaoui, and R. Molva, “Message- locked proofs of retrievability with secure deduplication,” in Proceedings of the 2016 ACM on Cloud Computing Security Workshop, 2016, pp. 73– 83.

[11] M. Bellare, R. Canetti, and H. Krawczyk, “Keying hash functions for message authentication,” in CRYPTO ’96, Berlin, Heidelberg, 1996, pp. 1–15.

[12]X. Q. Liang, Z. Yan, X. F. Chen, L. T. Yang, W. J. Lou, and Y. T. Hou, “Game theoretical analysis on encrypted cloud data deduplication,” IEEE Transactions on Industrial Informatics, vol. 15, no. 10, pp. 5778–5789, 2019.

[13]X. Q. Liang, Z. Yan, R. H. Deng, and Q. H. Zheng, “Investigating the adoption of hybrid encrypted cloud data deduplication with game theory,” IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 3, pp. 587–600.