Volume: 9 Issue: IV Month of publication: April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33533

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Review Paper on Distributed De-Duplication System using File and Block Level

Miss. Sneha Lande[1], Dr. M. M. Bartere[2]
[1, 2]G.H. Raisoni University, Amravati, India

Abstract: Data consolidation is a challenging issue in data integration. The usefulness of data increases when it is linked and fused with other data from numerous (Web) sources. The process of eliminating the repeated or duplicates copies of data is called as Data deduplication. Modern backup storage systems adopt deduplication to save space by eliminating data duplicates whereas impairing the storage reliability. As the world moves to digital storage for archival purposes, there is an increasing demand for systems that can provide secure data storage in a cost-effective manner. This data deduplication process is widely used in in cloud storage to decrease storage space and upload bandwidth. By using, deduplication system progress of storage utilization and reliability is increases. In addition, the dare of privacy for sensitive data also take place when they are outsourced by users to cloud. Planning to address the above security test, this paper constructs the first effort to celebrate the idea of scattered reliable deduplication system. This paper recommends a new distributed deduplication systems with upper dependability in which the data chunks are distributed from corner to cornering multiple cloud servers. The safety needs of data privacy and tag stability are also accomplish by introducing a deterministic secret sharing scheme in distributed storage systems, instead of using convergent encryption as in previous deduplication systems. A deduplication technique, on the other hand, can reduce the storage cost at the server side and save the upload bandwidth at the user side.

I. INTRODUCTION

With the growing amount of data worldwide in recent years, data deduplication is widely deployed and becomes increasingly important in backup storage systems. Data deduplication reduces redundant data by dividing files into multiple chunks which are then uniquely identified by fingerprint with a secure hash signature, so files can be broken up into chunks that can be shared. Chunkbased deduplication is an effective method for saving storage space, but it fundamentally affects storage system reliability compared to without deduplication, as deduplication makes one-chunk loss corrupt multiple files that share it. Thus it is a critical for deduplicated storage systems to ensure high data reliability [1].

The complicating factors resulting in the mounting digital forensic backlog include: (i) the increasing number of cases involving digital investigation; (ii) the number of digital devices requiring analysis per case; (iii) the increasing storage volume of each device; (iv) the diversity of digital devices, storage formats, file systems, and physical data locations, e.g., Internet-of-Things devices, wearables, cloud storage, remote storage, peer-to-peer file synchronization services, etc. Data reduction techniques can aid in decreasing the volume of data to be analyzed. Data deduplication is a data reduction technique used to optimize data storage and is particularly efficient when common data is encountered [2].

A number of deduplication systems have been projected based on various deduplication scheme such as client-side or server-side deduplication, file-level or block-level deduplications. Specially, with the advent of cloud storage, data deduplication procedure grow to be more gorgeous and essential for the management of ever-increasing quantity of data in cloud storage services which inspires Endeavour and club to outsource data storage to third-party cloud providers. Today's commercial cloud storage services, such as Dropbox, Google Drive and Mozy, have been applying deduplication to save the network bandwidth and the storage cost with client-side deduplication.

Data security is another area of increasing importance in modern storage systems and, unfortunately, deduplication and encryption are, to a great extent, diametrically opposed to one another. Deduplication takes advantage of data similarity in order to achieve a reduction in storage space. In contrast, the goal of cryptography is to make ciphertext indistinguishable from theoretically random data. Thus, the goal of a secure deduplication system is to provide data security, against both inside and outside adversaries, without compromising the space efficiency achievable through single-instance storage techniques [3].

Data security is another area of increasing importance in modern storage systems and, unfortunately, deduplication and encryption are, to a great extent, diametrically opposed to one another. Deduplication takes advantage of data similarity in order to achieve a reduction in storage space.

In contrast, the goal of cryptography is to make ciphertext indistinguishable from theoretically random data. Thus, the goal of a secure deduplication system is to provide data security, against both inside and outside adversaries, without compromising the space efficiency achievable through single-instance storage techniques [4].

Data backup has been an important issue ever since computers have been used to store valuable information. There has been a considerable amount of research on this topic, and a plethora of solutions are available which largely satisfy traditional requirements. However, new modes of working, such as the extensive [5].

Deduplication efficiency is an important metric for deduplication backup systems. Though significantly reducing the storage space of backup systems, chunk-level deduplication fails to remove the redundancy between non-duplicate but very similar chunks [6].

Deduplication has received much attention from both academic and industry because it can more improves storage utilization and save storage space, especially for the applications with high deduplication ratio such as accession storage systems. For eliminating duplicate copies of data we use data deduplication technique. To reduce storage space and for uploading bandwidth mostly it has been used.

The aim of this paper is to make the first attempt formalize the idea of distributed reliable deduplication system. In our proposed system we are going to develop new distributed deduplication systems which is highly reliable. In deduplication process data chunks are distributed across multiple cloud servers. instead of using convergent encryption as in previous deduplication systems we use deterministic secret sharing scheme in distributed storage systems. So that we can achieve the required concepts for security that are data confidentiality and tag consistency. In the proposed security model, Security analysis demonstrates that our deduplication systems are secure.

The rest of the paper is organized as follows: Section I Introduction. Section II discusses Background and Related Work. Section III discusses existing methodologies. Section IV discusses proposed method. Finally section V Conclude this review paper.

II. BACKGROUND AND RELATED WORK

Reliable De-duplication systems Data de-duplication techniques are very interesting techniques that are widely employed for data backup in enterprise environments to minimize network and storage overhead by detecting and eliminating redundancy among data blocks. There are many de-duplication schemes proposed by the research community.

In paper[1] propose a coding scheme that operates on each object which is formed by packed deduplicated chunks (inner-object coding) rather than on multiple objects (inter-object coding), such that the storage overhead can be saved and the degraded read performance can be improved. We also leverage the rewriting algorithm to accelerate the recent backup read throughput. We build an erasure-coded deduplicated backup storage system prototype EEC-Dedup, which realizes inner-object coding scheme and the rewriting algorithm.

The research [2], presented in this paper is to revolutionize the current digital forensic process through the leveraging of centralized deduplicated acquisition and processing approach. Focusing on this first step in digital evidence processing, acquisition, a system is presented enabling deduplicated evidence acquisition with the capability of automated, forensically-sound complete disk image reconstruction. As the number of cases acquired by the proposed system increases, the more duplicate artifacts will be encountered, and the more efficient the processing of each new case will become. This results in a time saving for digital investigators, and provides a platform to enable non-expert evidence processing, alongside the benefits of reduced storage and bandwidth requirements.

In research [3], propose an architecture that provides secure deduplicated storage resisting brute-force attacks, and realize it in a system called DupLESS. In DupLESS, clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol. It enables clients to store encrypted data with an existing service, have the service perform deduplication on their behalf, and yet achieves strong confidentiality guarantees.

Paper [6], proposed SDC, a scheme selectively performing delta compression after chunk-level deduplication. On one hand, SDC simulates a restore cache during backup to identify nonbase- fragmented chunks and only performs delta compression for these non-base-fragmented chunks, thus avoiding the base chunk fragmentation. On the other hand, SDC adopts HAR rewriting algorithm to reduce the normal chunk fragmentation. The experimental results based on real-world datasets demonstrate the robustness of simulated restore cache and the better performance of SDC in terms of improving restore performance.

In research [7], propose Fo-DSC, a fogbased deduplicated spatial crowdsourcing framework to achieve precise task allocation and secure data deduplication. Specifically, by integrating fog computing, we design a two-step task allocation mechanism to improve the accuracy of tasks allocation in spatial crowdsourcing. The fog nodes can detect and erase the repeated data in crowdsensing reports without learning any information about the reports.

III. EXISTING METHODOLOGIES

A. EEC-Dedup: Efficient Erasure-Coded Deduplicated Backup Storage Systems

EEC-Dedup, an Efficient Erasure Code for Deduplicated backup storage systems that have three design goals that correspond to the above three challenges : (Goal 1) varied-size chunking; (Goal 2) inner-object coding; (Goal 3) optimized degraded read performance. For Goal 1, EEC-Dedup is designed based on AE chunk algorithm, which is a state-of-the-art varied-size chunking scheme. For Goal 2, EEC-Dedup operates erasure coding on each single object (we called inner-object coding) instead of inter-object coding so as to reduce the number of zero-byte paddings. For Goal 3, EEC-Dedup exploits the inner-object coding scheme and packs chunks into objects based on the state-of-theart rewriting algorithm such that backup degraded read performance can be improved significantly [1].

B. Deduplicated Evidence Acquisition

In the proposed system, prior to acquisition, files and data are hashed and compared with a centralized, known-file database to eliminate common files. The data acquired from the evidence devices include the files, slack space (at the disk level and block level) and unallocated space. The data collection process only collects the unique artifacts encountered in the evidence, saving bandwidth and storage space. This acquisition process acquires more than just a copy of evidence artifacts, it has also completed part of the analysis, i.e., the calculation of the artifact hashes and the indexing of associated metadata for future examination [2].

C. DupLESS: Server-Aided Encryption for Deduplicated Storage

DupLESS starts with the observation that brute-force ciphertext recovery in a CE-type scheme can be dealt with by using a key server (KS) to derive keys, instead of setting keys to be hashes of messages. Access to the KS is preceded by authentication, which stops external attackers. The increased cost slows down brute-force attacks from compromised clients, and now the KS can function as a (logically) single point of control for implementing rate-limiting measures [3].

D. Selective delta compression (SDC)

This module is designed to perform the delta encoding and ensure that no extra read operation will be required by base chunks during restore. Two data structures are designed for this module. First, a sketch index that records the sketches of the previously stored chunks is needed for resemblance detection. Sketches are compact representation of the corresponding chunks. Similar chunks have identical sketches. Note that SDC only allows 1-level delta compression. To ensure deltas will not serve as base chunks, SDC does not compute sketches for deltas. Second, a simulated restore cache is employed to identify the non-base-fragmented chunks. Belady’s optimal replacement scheme can be applied to restore cache for higher cache hit ratio [6].

E. Secure and Deduplicated Spatial Crowdsourcing: A Fog-Based Approach (FO-DSC)

Fo-DSC, a fogbased deduplicated spatial crowdsourcing framework to achieve precise task allocation and secure data deduplication. Specifically, by integrating fog computing, we design a two-step task allocation mechanism to improve the accuracy of tasks allocation in spatial crowdsourcing. The fog nodes can detect and erase the repeated data in crowdsensing reports without learning any information about the reports. Furthermore, Fo-DSC efficiently records the contributions of mobile users whose data are reduplicated and deleted. Fo-DSC satisfies the properties of fog-based task allocation and secure data deduplication with low computational and communication overheads [7].

IV. PROPOSED METHODOLOGY

To protect private data the secret sharing technique is used which is corresponding to distributed storage systems. In this paper the secret sharing technique is used for protection of private data. In detail a file is divides and encode into sections by using secret sharing technique. These sections will be distributed over many independent storage servers. A cryptanalysis hash value of the content will also be calculated and send to storage server as the mark of the fragment stored at each server. only the data user who first upload the data is required to calculate and distribute such secret shares and following users own same data copy do not need to calculate and stores these shares. Retrieve data copies owner must access a minimum number of storage server by a validation and obtain the secret shares to alter the data. In different way, the authorized uses will access the secret shares data copy. Another distinguishable feature of our proposal is that data completeness encloses tag consistency, can be derived. To explain further if the same value is stored in various cloud storage then deduplication check by methods. It cannot oppose the collision attack established by many servers.

To our knowledge no related work on secure deduplication can rightly address, the reliability and tag consistency problem. The file level and block level deduplication is used for higher reliability. The secret splitting technique is used for protect data. Our proposed structure supports both traditional deduplication methods. Privacy, credibility and integrity can be achieved in our proposed system. In solution to kind of secret agreement attacks are considered. These are the attack on the data and the attack against servers. The data is secure when the opponent control limited number of storage servers. Following figure Fig 1.shows the proposed architecture of system.

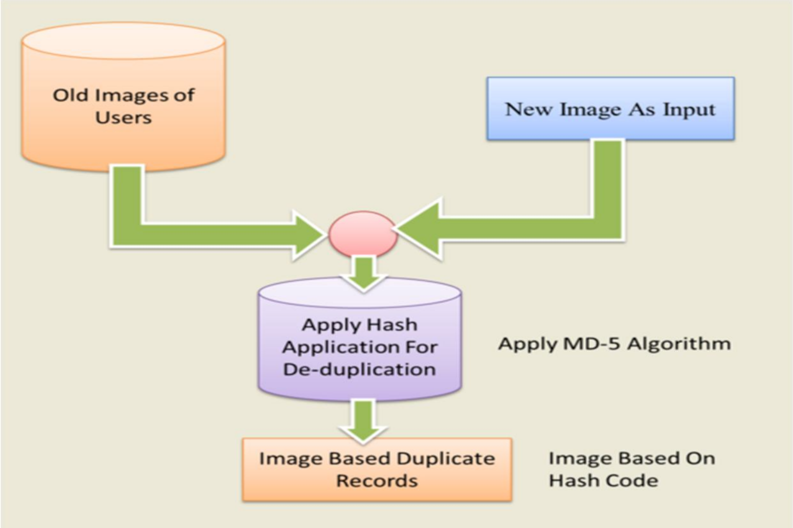

Fig 1: Proposed Architecture of System

V. CONCLUSION

Deduplication has received much attention from both academic and industry because it can more improves storage utilization and save storage space, especially for the applications with high deduplication ratio such as accession storage systems. For eliminating duplicate copies of data we use data deduplication technique. To reduce storage space and for uploading bandwidth mostly it has been used. In this paper proposed system we are going to develop new distributed deduplication systems which is highly reliable. In Proposed paper we describe a system which applying hash application for deduplication records and encryption using the MD 5 algorithm. In this paper we discuss different technologies ,methods for deduplication and secure data. In deduplication process data chunks are distributed across multiple cloud servers. instead of using convergent encryption as in previous deduplication systems we use deterministic secret sharing scheme in distributed storage systems.

REFERENCE

[1] Wenxiang Chen, Yuchong Hu, Siyang Yin, and Wen Xia,” EEC-Dedup: Efficient Erasure-Coded Deduplicated Backup Storage Systems”, IEEE International Symposium on Parallel and Distributed Processing with Applications and IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), 2017.

[2] Xiaoyu Du, Paul Ledwith and Mark Scanlon,” Deduplicated Disk Image Evidence Acquisition and Forensically-Sound Reconstruction”, 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering, 2018.

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Serveraided encryption for deduplicated storage,” in USENIX Security Symposium, 2013.

[4] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, “Secure data deduplication,” in Proc. Of StorageSS, 2008.

[5] P. Anderson and L. Zhang, “Fast and Secure Laptop Backups with Encrypted De-duplication,” in Proc. of USENIX LISA, 2010.

[6] Yucheng Zhang ; Dan Feng ; Yu Hua ; Yuchong Hu ; Wen Xia ; Min Fu ; Xiaolan Tang ; Zhikun Wang,” Reducing Chunk Fragmentation for In-Line Delta Compressed and Deduplicated Backup Systems”, International Conference on Networking, Architecture, and Storage (NAS), Aug 2017.

[7] Jianbing Ni ; Xiaodong Lin ; Kuan Zhang ; Yong Yu,” Secure and Deduplicated Spatial Crowdsourcing: A Fog-Based Approach”, IEEE Global Communications Conference (GLOBECOM), Dec 2016.

[8] Pengfei Zuo, Yu Hua, Ming Zhao†, Wen Zhou, Yuncheng Guo,” Improving the Performance and Endurance of Encrypted Non-volatile Main Memory through Deduplicating Writes”, 51st Annual IEEE/ACM International Symposium on Microarchitecture, 2018.

[9] Chan-I Ku, Guo-Heng Luo, Che-Pin Chang & Shyan-Ming Yuan,” File Deduplication with Cloud Storage File System”, IEEE 16th International Conference on Computational Science and Engineering, 2013.

[10] Yongquan Dong, Eduard C. Dragut, and Weiyi Meng”, Normalization of Duplicate Records from Multiple Sources”, IEEE Transactions On Knowledge And Data Engineering, Oct 2017.