ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

A Survey on Authorized Deduplication Technique for Encrypted Data with DARE scheme in a Twin Cloud Environment.

Rasika V. Gode, Prof. Rupali Dalvi

Student, Dept. of Computer Engg, M.M.C.O.E., Pune, Savitribai Phule Pune University Pune, India

Dept. of Computer Engg, M.M.C.O.E., Pune, Savitribai Phule Pune University Pune, India

ABSTRACT: Data reduction has become increasingly important in storage systems due to the explosive growth of digital data in the world that has ushered in the big data era. One of the main challenges facing large-scale data reduction is how to maximally detect and eliminate redundancy at very low overheads. In this paper, we present DARE, a low-overhead deduplication-aware resemblance detection and elimination scheme in a Twin Cloud environment that effectively exploits existing duplicate-adjacency information for highly efficient resemblance detection in data deduplication based backup/archiving storage systems as well as supports authorized deduplication technique for encrypted data. The main idea behind DARE is to employ a scheme, call Duplicate-Adjacency based Resemblance Detection (DupAdj), by considering any two data chunks to be similar (i.e., candidates fordelta compression) if their respective adjacent data chunks are duplicate in a deduplication system, and then further enhance theresemblance detection efficiency by an improved super-feature approach. In existing system DARE Deduplication technique is used only in-house computer, in our proposed system you can use DARE Deduplication technique in cloud storage also and you can perform DARE Deduplication technique on encrypted data. Our experimental results based on real-world and synthetic backup datasets show that DARE only consumes about 1/4 and 1/2 respectively of the computation and indexing overheads required by the traditional super-feature approaches while detecting 2-10 percent more redundancy and achieving a higher throughput, by exploiting existing duplicate-adjacency information for resemblance detection and finding the “sweet spot” for the super-feature approach.

KEYWORDS: Data deduplication, delta compression, storage system, index structure, performance evaluation

I. INTRODUCTION

DATAdeduplication is the technique which compresses the data by removing the duplicate copies of identical data and it is extensively used in cloud storage to save bandwidth and minimize the storage space. In cloud computing, users outsource their data to external cloud servers and that data may contain sensitive privacy information, such as personal photos, emails, etc. If there is no any efficient protection, then it leads to severe confidentiality and privacy violations. It is therefore necessary to encrypt the sensitive data before outsourcing them to the cloud. This issue in mobile cloud computing motivates to protect the confidentiality of sensitive data while supporting de-duplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, our technique makes the first attempt to formally address the problem of authorized data de-duplication andto maximally detect and eliminate redundancy at very low overheads. We presents a low-overhead Deduplication-Aware Resemblance detection and Elimination (DARE) scheme for data reduction with low overhead.

The amount of digital data is growing explosively, as evidencedin part by an estimated amount of about 1.2 zettabytesand 1.8 zettabytes respectively of data produced in2010 and 2011. As a result of this “data deluge”, managingstorage and reducing its costs have become one of themost challenging and important tasks in mass storage systems.According to a recent IDC study, almost 80 percentof corporations surveyed indicated that they were exploringdata deduplication technologies in their storage systems to increase storage efficiency.

Data deduplication is an efficient data reductionapproach that not only reduces storage space by eliminating duplicate data but also minimizesthe transmission of redundant data in low-bandwidthnetwork environments. In

DOI: 10.15680/IJIRCCE.2016. 0411171

general, achunk-level data deduplication scheme splits data blocks ofa data stream (e.g., backup files, databases, and virtualmachine images) into multiple data chunks that are eachuniquely identified and duplicate-detected by a secureSHA-1 or MD5 hash signature (also called a fingerprint). Storage systems then remove duplicates of data chunksand store only one copy of them to achieve the goal of spacesavings. While data deduplication has been widely deployed instorage systemsfor space savings, the fingerprint-based deduplicationapproaches have an inherent drawback: they oftenfail to detect the similar chunks that are largely identicalexcept for a few modified bytes, because their secure hashdigest will be totally different even only one byte of a datachunk was changed. It becomes a bigchallenge when applying data deduplication to storage datasetsandworkloads that have frequently modified data,whichdemands an effective and efficient way to eliminate redundancyamong frequently modified and thus similar data.

Delta compression, an efficient approach to removingredundancy among similar data chunks has gained increasingattention in storage systems. Forexample, if chunk A2 is similar to chunk A1 (the basechunk),the delta compression approach calculates and thenonly stores the differences (delta) and mapping relationbetween A2 and A1. Thus, it is considered a promising techniquethat effectively complements the fingerprint-baseddeduplication approaches by detecting similar data missedby the latter.One of the main challenges facing the application of deltacompression in deduplication systems is how to accurately detect the most similar candidates for delta compression with lowoverheads. The state-of-the-art solutionsdetect similarity for delta compression by computing severalRabin fingerprints as features and grouping them intosuper-fingerprints, also referred to as super-features (SF). Nevertheless, to index a dataset of80 TB and assuming an average chunk size of 8 KB and 16bytes per index entry, for example, about 200 GB worth ofsuper-feature index entries must be generated, which willstill be too large to fit in memory. Since the randomaccesses to on-disk index are much slower than that toRAM, the frequent accesses to on-disk super-features willcause the system throughput to become unacceptably lowfor the users.

II. BACKGROUND

1. Convergent Encryption Module:

Traditional encryption, while providing data confidentiality, is incompatible with data de-duplication. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identical data copies of different users will lead to different cipher texts, making de-duplication impossible. Convergent encryption has been proposed to enforce data confidentiality while making de-duplication feasible. It encrypts/ decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since the encryption operation is deterministic and is derived from the data content, identical data copies will generate the same convergent key and hence the same cipher text.

2. Proof Of Ownership:

The notion of proof of ownership (POW) enables users to prove their ownership of data copies to the storage server.Specifically, POW is implemented as an interactive algorithm (denoted by POW). The verifier derives a short value Φ(M) from a data copy M. To prove the ownership of the data copy M, the prover needs to send Φ’to the verifier such that Φ’= Φ(M).

3. Resemblance Detection Based Data Reduction:

Data deduplication is becoming increasingly popular indata-intensive storage systems as one of the most efficientdata reduction approaches in recent years. Fingerprintbaseddeduplication techniques eliminate duplicatechunks by checking their secure-fingerprints (i.e., SHA-1/SHA-256 signatures), which has been widely used in commercialbackup and archiving storage systems.Another challenge for data deduplication is how to maximallydetect and eliminate data redundancy in storage systemsby determining appropriate data chunking schemes.In order to find more redundant data, the content-definedchunking (CDC) approach was proposed in LBFS to findthe proper cut-point of each chunk in the files and addressthe boundary-shift problem. Re-chunking approaches were also proposed to divide those non-duplicatechunks into smaller ones to expose and detect moreredundancy.Resemblance detection with delta compression as another approach to data reduction in storage systems, was proposed more than 10 years ago but was laterovershadowed by fingerprint-based deduplication due to the former’s scalability issue. Table 1 comparesthese two data reduction approaches. Resemblance detectiondetects redundancy among similar data at the byte levelwhile duplicate detection finds totally identical data at thechunk level, which makes the latter much more scalablethan the former in mass storage systems.

4. **Fact of Duplicate Adjacency:**

The modified chunks may bevery similar to their previous versions in a backup system while unmodified chunks will remain duplicate and areeasily identified by the deduplication process. For those non-duplicate chunks that are location-adjacent to known duplicatedata chunks in a deduplication system, it is intuitive and quite possible that only a few bytes of them are modified from thelast backup, making them potentially excellent delta compression candidates.

5. **Rethinking of the Super-Feature Approaches:**

Similar data, like duplicate data, are in wide existence inbackup systems. Meister and Brinkmann find that small semantic changes on documents may result in big modifications in the binary representation of files, and delta compression is more effective in eliminating redundancyin such cases. To support delta compression, resemblance detection will be required for selecting suitable similar candidates.

III. LITERATURE SURVEY

Sr. No.	Author & Title	Proposed Scheme	We have referred
1	B. Zhu, K. Li, and R. H. Patterson, “Avoiding the disk bottleneck in the data domain deduplication file system,” in Proc. 6th USENIX Conf. File Storage Technol., Feb. 2008, vol. 8, pp. 1–14.	Describes three techniques employed in the production Data Domain deduplication file system to relieve the disk bottleneck. These techniques include: (1) the Summary Vector, a compact in-memory data structure for identifying new segments; (2) Stream-Informed Segment Layout, a data layout method to improve on-disk locality for sequentially accessed segments; and (3) Locality Preserved Caching, which maintains the locality of the fingerprints of duplicate segments to achieve high cache hit ratios. Together, they can remove 99% of the disk accesses for deduplication of real world workloads. These techniques enable a modern two-socket dual-core system to run at 90% CPU utilization with only one shelf of 15 disks and achieve 100 MB/sec for single-stream throughputand 210 MB/sec for multi-stream throughput.	(1) A compact in-memory data structure for identifying new segments; (2) Stream-Informed Segment Layout, a data layout method to improve on-disk locality for sequentially accessed segments; and (3) Locality Preserved Caching, which maintains the locality of the fingerprints of duplicate segments to achieve high cache hit ratios.
2	D. Meister, J. Kaiser, and A. Brinkmann, “Block locality caching for data deduplication,” in Proc. 6th Int. Syst. Storage Conf., 2013, pp. 1–12.	Propose a novel approach, called Block Locality Cache (BLC), that captures the previous backup run significantly better than existing approaches and also always uses up-to date locality information and which is, therefore, less prone to	Simplified example of chunk index entries and block recipes and Simplified example of the cache data structuresof the BLC approach.

DOI: 10.15680/IJIRCCE.2016. 0411171

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

		aging. We evaluate the approach using a trace-based simulation ofmultiple real-world backup datasets. The simulation compares the Block Locality Cache with the approach of Zhu et al. [37] and provides a detailed analysis of the behavior and IO pattern. Furthermore, a prototype implementation is used to validate the simulation.	
3	A. El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, “Primary data deduplication-large scale study and system design,” in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2012, pp. 285–296.	We present a large scale study of primary data deduplicationand use the findings to drive the design of a new primary data deduplication system implemented in the Windows Server 2012 operating system. File data was analyzed from 15 globally distributed file servers hosting data for over 2000 users in a large multinational corporation. The findings are used to arrive at a chunking and compression approach which maximizes deduplication savings while minimizing the generated metadata and producing a uniform chunk size distribution. Scaling of deduplication processing with data size is achieved using a RAM frugal chunk hash index and data partitioning – so that memory, CPU, and disk seek resources remain available to fulfill the primary workload of serving IO.	Architecture of a new primary data deduplication system and evaluate the deduplication performance and chunking aspects of the system.
4	S. Quinlan and S. Dorward, “Venti: A new approach to archival storage,” in Proc. USENIX Conf. File Storage Technol., Jan. 2002, pp. 89–101.	describes a network storage system, called Venti, intended for archival data. In this system, a unique hash of a block’s contents acts as the block identifier for read and write operations. This approachenforces a write-once policy, preventing accidental or malicious destruction of data. In addition, duplicate copies of a block can be coalesced, reducing theconsumption of storage and simplifying the implementation of clients. Venti is a building block for constructing a variety of storage applications such as logical backup,	Choice of Storage Technology , A tree structure for storing a linear sequence of blocks and Build a new version of the tree.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

		physical backup, and snapshot file systems. Built a prototype of the system and present some preliminary performance results. The system uses magnetic disks as the storage technology, resulting in an access time for archival data that is comparable to non-archival data. The feasibility of the write-once model for storage is demonstrated using data from over a decade's use of two Plan 9 file systems.	
5	A. Venish and K. Siva Sankar, "Study of Chunking Algorithm in Data Deduplication " Proceeding of the International Conference on Soft Computing System, ICSCS , Volume 2.	In the deduplication technology, data are broken down into multiple pieces called "chunks" and every chunk is identified with a unique hash identifier. These identifiers are used to compare the chunks with previously stored chunks and verified for duplication. Since the chunking algorithm is the first step involved in getting efficient data deduplication ratio and throughput, it is very important in the deduplication scenario. In this paper, we discuss different chunking models and algorithms with a comparison of their performances.	Chunking algorithm Data deduplication Fixed and variable chunking method
6	Bo Mao, Member, IEEE, Hong Jiang, Fellow, IEEE, Suzhen Wu, Member, IEEE, and Lei Tian, Senior Member, IEEE "Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud" IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016	propose a performance-oriented I/O deduplication, called POD, rather than a capacity-oriented I/O deduplication, exemplified by iDedup, to improve the I/O performance of primary storage systems in the Cloud without sacrificing capacity savings of the latter. POD takes a two-pronged approach to improving the performance of primary storage systems and minimizing performance overhead of deduplication, namely, a request-based selective deduplication technique, called Select-Dedupe,to alleviate the data fragmentation and an adaptive memory management scheme, called iCache, to ease the memory contentionbetween the bursty read traffic and the bursty write traffic. We have implemented a prototype	Select the every requests to deduplicate or do not bypass small requests (e.g., 4 KB, 8 KB or less).

DOI: 10.15680/IJIRCCE.2016. 0411171

		of POD as a module in the Linuxoperating system.	
7	Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou "A Hybrid Cloud Approach for Secure Authorized De-duplication" IEEE Transactions on Parallel and Distributed Systems: VOL. 26 NO.5 MAY2015.	In the proposed system we are achieving the data de-duplication by providing the proof of data by the data owner. This proof is used at the time of uploading of the file. Each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files.	New de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Proposed system includes proof of data owner so it will help to implement better security issues in cloud computing.
8	Shweta D. Pochhi, Prof.Pradnya V. Kasture "Encrypted Data Storage with De-duplication Approach on Twin Cloud " International Journal of Innovative Research in Computer and Communication Engineering	The data and the Private cloud where the token generation will be performed for each file. Before uploading the data or file to public cloud, the client will send the file to private cloud for token generation which is unique for each file. Private clouds then generate a hash and a token and send the token to client. Token and hash keep in the private cloud itself so that whenever next file comes for token generation, the private clod can refer the same token. Once Client gets token for a particular file, public cloud search for the similar token if it exists or not. If the token exist public cloud will return a pointer of the already existing file otherwise it will send a message to upload a file.	A system which achieves confidentiality and enables block-level de-duplication at the same time.Before uploading the data or file to public cloud, the client will send the file to private cloud for token generation which is unique for each file. Private cloud then generate a hash and a token and send the token to client. Token and hash keep in the private cloud itself so that whenever next file comes for token generation, the private clod can refer the same token.
9	BhushanChoudhary, AmitDravid "A Study On Secure Deduplication Techniques In Cloud Computing" International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) Volume 3, Issue 12, April 2014	A de-duplication system in the cloud storage proposed to reduce the storage size of the tags for integrity check. To upgrade the security of de-duplication and secure the information secrecy demonstrated to secure the information by transforming the predictable message into unpredictable message.	A de-duplication system in the cloud storage proposed to reduce the storage size of the tags for integrity check. To upgrade the security of de-duplication and secure the information secrecy demonstrated to secure the information by transforming the predictable message into unpredictable message.

10	Wee Keong Ng SCE, NTU Yonggang Wen SCE, NTU Huafei Zhu "Private Data De-duplication Protocols in Cloud Storage" SAC12 March 2529, 2012, Riva del Garda, Italy. Copyright 2011 ACM 9781450308571/12/03	This paper studies private data de-duplication technique for cloud storages. Intuitively, a private data de-duplication protocol allows a client who holds a private data proves to a server who holds a summary string of the data that he/she is the owner of that data without revealing further information to the server.	The proposed private data de-duplication protocol is provably secure in the simulation based framework assuming that the underlying hash function is collision resilient, the discrete logarithm is hard and the erasure coding algorithm E can erasure up to fraction of the bits in the presence of malicious adversaries.

IV. EXISTING SYSTEM APPROACH

The existing solutions to the indexing issue of delta compression either record the resemblance information for files, instead of data chunks, so that similarity index entries can fit in the memory or exploit the locality of backup data streams in deduplication-based backup/archiving systems, which avoids the global indexing on the disk. The first approach faces an implementation difficulty in large-scale data deduplication systems since it is hard to record all the resemblance or version information of files in such systems. The second approach often fails to detect a significant amount of redundant data when the workloads lack locality. Another challenge facing the super-feature method is the high overhead in computing the super-features. According to a recent study of delta compression and our experimental observation, the throughput of computing uper-features is about 30 MB/s, which may become a potential bottleneck for deduplication-based storage systems, particularly if most indexentries are fit in memory or partially on SSD-based storage for which the throughput can be hundreds of MB per second or higher.

In existing system DARE Deduplication technique is used only in-house computer, in our proposed system you can use DARE Deduplication technique in cloud storage also and you can perform DARE Deduplication technique on encrypted data. In the existing data de-duplication systems on a cloud, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges. Such architecture is practical and has attracted much attention from researchers. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud.

DISADVANTAGES OF EXISTING SYSTEM

1. Indexing issue of delta compression either record the resemblance information for files, instead of data chunks.
2. In a hybrid cloud approach we can achieved authorized deduplication on data but we cannot eliminate redundancy at very low overheads.

V. PROPOSED SYSTEM

The system will provide an authorized deduplication on encrypted data which can be in the form of text file.The system effectively achieves the storage space management in a secure and authorized manner as well as the system enables to maximally detect and eliminate redundancy at very low overheads by using DARE scheme.
Propose system is divided into three parts :

1. Data user authentication scheme is used in a private cloud server to prevent system from attackers.
2. Convergent encryption is used to encrypt the data and perform duplicate check.
3. DARE maximally detect and eliminate redundancy at very low overheads.

DARE, a low-overhead Deduplication-Aware Resemblance detection and Elimination scheme for deduplication based backup and archiving storagesystem. The main idea of DARE is to effectively exploit existing duplicate-adjacency information to detect similardata chunks (DupAdj), refine and supplement the detectionby using an improved super-feature approach (Low-OverheadSuper-Feature) when the existing duplicate-adjacencyinformation is lacking or limited. In addition, we present ananalytical study of the existing super-feature approach witha mathematic model and conduct an empirical evaluation of this approach with several real-world workloads in datadeduplication systems. In existing system DARE Deduplication technique is used only in-house computer, in our proposed system you can use DARE Deduplication technique in cloud storage also and you can perform DARE Deduplication technique on encrypted data. Our experimental evaluation results, based on real-worldand synthetic backup datasets, show that DARE significantly outperforms the traditional Super-Feature approach.More specifically, the DupAdj approach achieves a similardata reduction efficiency to the pure super-feature approachand DARE detects 2-10 percent more redundant data while achieving a higher throughput of data reduction than thepure super-feature approach. Meanwhile, DARE only consumes about 1/4 and 1/2 respectively of the computationand indexing overheads required by the traditional superfeatureapproach for resemblance detection. It is importantto note that our evaluation also demonstrates the superiordata-restore performance of the DARE-enhanced deduplicationsystem over the deduplication-only systems via deltacompression, where the former outperforms the latter by a factor of 2 (2X).

ADVANTAGES OF PROPOSED SYSTEM

1. The proposed system can flexibly support access control on encrypted data with deduplication.
2. Low Cost of Storage.
3. Large-scale data reduction.
4. Removing redundancy among similar data chunks has gained increasing attention in storage systems.
5. Accurately detect the most similar candidates for delta compression with low overheads

VI. SYSTEM ARCHITECTURE

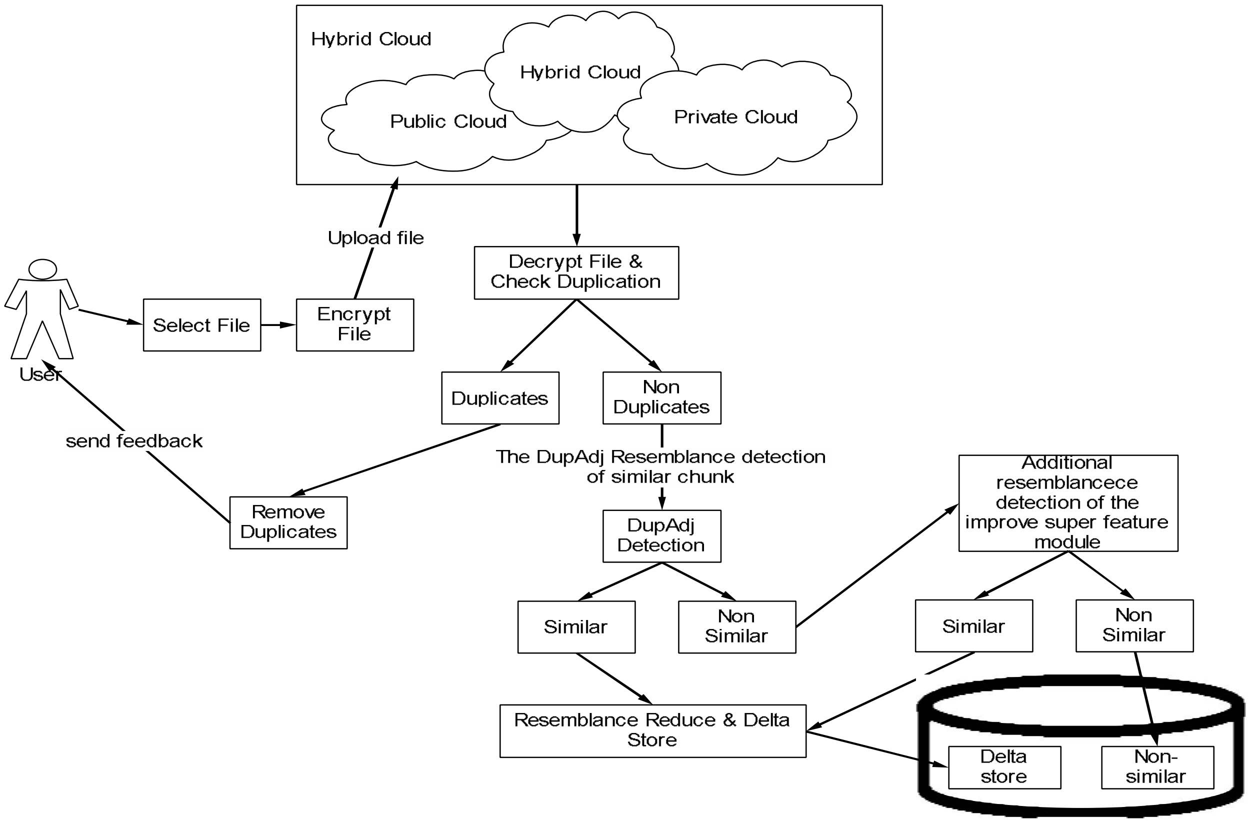

Fig.1. Proposed System Architecture

The data reduction workflow of DARE, showing an example of resemblance detection for delta compression first by the DupAdj approach and then by the super-feature approach.

VI. CONCLUSION

The system will provide an authorized deduplication on encrypted data which can be in the form of text file.The system effectively achieves the storage space management in a secure and authorized manner. And the system enables to maximally detect and eliminate redundancy at very low overheads by using DARE scheme.DARE uses a novel approach, DupAdj, whichexploits the duplicate-adjacency information for efficient resemblance detection in existing deduplication systems, and employs an improved super-feature approach to further detecting resemblance when the duplicate-adjacency information is lacking or limited. Results from experiments driven by real-world and synthetic backup datasets suggest that DARE can be a powerful and efficient tool for maximizing data reduction by further detecting resembling data with low overheads. Specifically, DARE only consumes about 1/4 and 1/2 respectively of the computation and indexing overheads required by the traditional super-feature approaches while detecting 2-10 percent more redundancy and achieving a higher throughput.

REFERENCES

1. B. Zhu, K. Li, and R. H. Patterson, “Avoiding the disk bottleneck in the data domain deduplication file system,” in Proc. 6th USENIX Conf. File Storage Technol., Feb. 2008, vol. 8, pp. 1–14.
2. D. Meister, J. Kaiser, and A. Brinkmann, “Block locality caching for data deduplication,” in Proc. 6th Int. Syst. Storage Conf., 2013, pp. 1–12.
3. A. El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, “Primary data deduplication-large scale study and system design,” in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2012, pp. 285–296.
4. S. Quinlan and S. Dorward, “Venti: A new approach to archival storage,” in Proc. USENIX Conf. File Storage Technol., Jan. 2002, pp. 89–101.
5. A. Venish and K. Siva Sankar, “Study of Chunking Algorithm in DataDeduplication ” Proceeding of the International Conference on Soft Computing System, ICSCS , Volume 2.
6. Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou “A Hybrid Cloud Approach for Secure Authorized De-duplication” IEEE Transactions on Parallel and Distributed Systems: VOL.26 N0.5 MAY 2015.
7. Bo Mao, Member, IEEE, Hong Jiang, Fellow, IEEE, Suzhen Wu, Member, IEEE, andLeiTian, Senior Member, IEEE “Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud ” IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016
8. Mr Vinod B Jadhav Prof Vinod S Wadne Secured Authorized De-duplication Based Hybrid Cloud Approach International Journal of Advanced Research in Computer Science and Software Engineering
9. JadapalliNandini, Rami reddyNavateja Reddy Implementation De-duplication System with Authorized Users International Research Journal of Engineering and Technology (IRJET)
10. Backialakshmi. N Manikandan. M SECURED AUTHORIZED DE-DUPLICATION IN DISTRIBUTED SYSTEM IJIRST International Journal for Innovative Research in Science and Technology— Volume 1 — Issue 9 — February 2015

DOI: 10.15680/IJIRCCE.2016. 0411171