International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

A Review of Literature on Security and Data Deduplication In Cloud Computing

J. Selvam[1], Dr.P.Srivaramangai[2]

[1]Ph. D.Research scholar, Department of Computer Science, Maruthu pandiyar College of Arts & Science, Thanjavur, Tamilnadu, India

[2]Associate Professor, Department of Computer Science, Maruthu pandiyar College of Arts & Science, Thanjavur. Tamilnadu, India

ABSTRACT

Cloud computing is an internet technology that operates both internet and central remote servers to control the applications and data. Cloud computing refers to the delivery of storage capacity and computing as a service to a heterogeneous community of end-receivers. Nowadays, Cloud storage systems are popularly improving a technology basic and that provides highly available storage. Deduplication is a technique of removing or abolishing duplicate data or files from the database or storage. It has utilized within the cloud computing environment to reduce storage space and also minimize network bandwidth. Cloud storage service providers like Amazon, Dropbox and others perform de-duplication process to reduce space allocation by storing only individual copy of each file uploaded. A Detailed review of Deduplication techniques is clearly presented in this paper.

Keywords *: Cloud Computing, Data duplication, MD5, SHA-1,Bigdata*

I. INTRODUCTION

Data deduplication is essentially means that reducing storage space and it works by removing redundant data and ensuring that only one unique example of the data is truly preserved on storage media, like tape or disk. Redundant data is swapped to the unique data copy by using a pointer support. Data deduplication, is often utilized in conjunction with other forms of data reduction, every so often called single-instance storage or intelligent compression. In traditional compression, we discusses about three decades, applying mathematical algorithms to data in order to reduce repetitious parts for creating a file smaller in an effective manner. Likewise, delta differencing minimizes the total volume of stored data with balancing the old and new iteration of a file so as to saving only the data that had modified. Altogether, the usage of storage space is effectively optimized with these techniques. Data deduplication is slow-down the amount of storage space needed so as to low cost on disk expenditures. For longer disk retention periods that allows to an effective usage of disk space, which provides enhanced recovery time objective (RTO) for a longer time and reduces the necessitate for tape backups. Data deduplication also minimizes the data and that should be transmitted across a WAN for disaster recovery, replication data and remote backups.

Data deduplication platforms ought to challenge with the problem of "hash collisions." Each large piece of data is processed by using a hash algorithm, like SHA-1 or MD5, after that generating a unique number for each piece. An index of the existing hash numbers are compared by the result of the hash

number, if it is already in the index then the piece of data is a duplicate and does not require to be stored over again. The new hash number included to the index after that the new data is stored. In exceptional cases, the hash algorithm may possibly generate the same hash number for two different chunks of data. When such a hash collision" occurs, the system fails to store the new data because it sees that hash number already. This is called a false positive and can result in data loss. Some vendors combine hash algorithms to reduce the possibility of a hash collision. Several vendors are also examining metadata to identify data and prevent collisions. Data deduplication primarily works at the file, block and also bit level. In this deduplication process is relatively trouble-free to understand, other than two files are exactly located at the same place, one copy of the file is stored and following iterations receive pointers to the saved file, although, the file deduplication is not efficient for modifying even a single bit results were completely different copy of the entire file being stored. As a result of comparison, bit and block level deduplication appears within a file and saves unique iterations of each block after that a file is updated and saved only the modified data. This performances generates the block and bit deduplication far more efficient.

II. LITERATURE REVIEW

Deepavali Bhagwat, *et al.* [1] introduced a new technique; Extreme Binning is mainly suited for workloads consisting of each file with low locality and also utilized for parallel and scalable deduplication process. It makes only one disk access for chunk lookup per file instead of per chunk, so as to reducing the disk bottleneck problem. Extreme Binning splits the chunk index into two tiers resulting in a low RAM footprint after that allocates the system to maintain throughput for a larger data set than a flat index scheme. In Partition phase, the data chunks and the two tier chunk index is easy to use and clean. Files are sequentially allocated into a single node for deduplication and storage by using a stateless routing algorithm. Maximum parallelization can be accomplished because of the one file-one backup node distribution. Backup nodes can be added to improve the redistribution of indices and throughput level and chunks is a clean operation because there are no dependencies between chunks or between the bins attached to different bins. In autonomy of backup nodes, the data management tasks like integrity checks, garbage collection and data restore requests efficient. De-duplication loss is developed as easy to use and also small process compensated by the gains in RAM usage and scalability.

G.Prashanthi, *et al.* [2] proposed new de-duplication process, the duplicate-check tokens of files are produced by the private cloud server within the private keys and also productions supporting authorized duplicate check in hybrid cloud architecture. This concept is proved and implemented a prototype of authorized conduct tested experiments and duplicate check scheme on our prototype. Authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer. Security analysis demonstrates of this scheme is securely protected in terms of outsider and insider attacks identified in the proposed security model.

Bhushan Choudhary, *et al.* [3] researched from this paper; the thought of authorized information deduplication was suggested to assure the information security by counting differential advantages of clients within the duplicate copy check. Security check demonstrates that the techniques are securely protected against insider and outsider assaults detailed in the proposed security model. That model illustrated the authorized duplicate copy check method experience minimum overhead comparing data transfer and convergent encryption process.

M. Bellare, *et al.* [4] introduced, Confidentiality can be protected by converting predictable message into unpredictable form. The Server aided encryption for de-duplicated storage recommends different security mechanisms. One new concept established to generate the file tag for duplicate check and also for Key server (Third party auditor).

Mark Lillibridge, *et al.* [5] referred from this paper; deduplication is an important aspect of D2D backup that is increasingly becoming the backup solution of choice. This approach utilizes a lot of locality within backup data at the small number of megabytes scale to solve the chunk-lookup disk bottleneck issue. Through content-based segmentation, sparse indexing, and sampling, the author split incoming streams into segments, identify similar existing segments, and de-duplicate against them, yielding excellent deduplication and throughput while requiring little RAM. Compared with the BFPFI, the author exploits less than half of the RAM for an equivalent high level of deduplication.

C.Ng, *et al.* [6] proposed deduplication approach improves to latest VM image snarls consuming an inkling baptized reverse deduplication. The deduplication focuses for enhancing fingerprint indexing to triumph high snarl concert by utilizing backup storage process. It eradicates duplicates from hoary data, so as to fluctuating breakup to old data even though observance the draught of innovative data as chronological as imaginable. RevDedup smears coarse-grained comprehensive deduplication to remunerate disk pursues over large size data units and it preserves high deduplication efficiency. While eradicate process handles any redundant data of the old forms and stating it to the duplicate data of the latest form. It triumphs high deduplication efficiency, and in the meantime reduces crumbling and triumphs high read enactment. The RevDedup eradicates the data from the old to new data. It does not afford that much security.

Pasqualo Puzio, *et al* [7] proposed, the cost-wise deduplication with latest privacy challenges and security update in deduplication performance is quite effective. ClouDedup is mainly related to a secure protection and efficient storage service. This storage service ensures data confidentiality and block-level deduplication performs at the same time and it is depending on convergent encryption. An access control mechanism and an additional encryption operation both are implemented by ClouDedup that remains secure because of the definition of a component.

Wen Xia, *et al.* [8] presented DARE, low-overhead resemblance detection, a deduplication aware and elimination method for improving data reduction in backup/archiving storage systems. DARE uses a fresh approach, DupAdj, which employs an improved super-feature approach to additional detecting resemblance as soon as the duplicate adjacency information is limited or lacking that utilizes the duplicate-adjacency information for efficient resemblance detection in existing deduplication systems. DARE may be an efficient and powerful tool for maximizing data reduction by additionally detecting resembling data with low overheads, which results from experiments driven by synthetic and real-world backup datasets propose. Particularly, DARE detects 2-10% more redundancy and achieving a higher throughput that consumes only about 1/4 and 1/2 respectively of the computation and indexing overheads for the traditional super-feature approaches. Additionally, the DARE improved data reduction approach is exposed to be capable of enhancing the data-restore performance, rapid up the deduplication-only approach by an aspect of 2(2X) by utilizing delta compression to additional process for eliminating redundancy and effectively enlarge the logical space of the restoration cache.

Mihir Bellare, *et al* [9] proposed Cloud storage service offers like Mozy, Dropbox and others perform

deduplication to reduce space via storing single copy of each file uploaded. Message-locked encryption (remarkable manifestation is convergent encryption) determines this tension and also it has inherently vulnerable to brute-force attacks that may recover files falling into a known set. An architecture proposed that requires secure de-duplicated storage opposing brute-force attacks, and perceive in a system is known as DupLESS that clients encrypts the under message-based keys attained from a key-server by using an oblivious PRF protocol. Demonstrated that encryption for de-duplicated storage can achieve performance and space reducing for utilizing the storage service with plaintext data. PRF protocol enables clients to store encrypted data with a current service and yet achieves strong confidentiality guarantees.

W. K. Ng, *et al* [10] reviewed that deduplication technique is utilized for performing private cloud and it allocates the user who holds the private data. The security of private cloud is founded on the simulation-based framework. The private cloud is secure for underlying hash function that is discrete logarithm and collision resilient. The discrete logarithm is rigid and the expurgation coding algorithms E may expurgation up to α-fraction of the jiffs in the manifestation of malevolent antagonists. The collision-resilient muddle function is a polynomial time reckonable function H representing dualistic cords of whimsical length into judiciously diminutive ones so as to computationally infeasible to determine any impact, that is any two different twines x and y for which H(x)=H(y). These processes are derived from the private cloud deduplication and it does not afford high security.

Iuon-Chang , *et al* [11] proposed as enhances the rapid performance of data deduplication. Up-loaded file is confirmed or verified for integrity by specially computed signature. To implement this Zhang Fault Tolerant digital Signature Scheme as proposed by Zhang can correct and detect errors efficiently in digital signature and it has based on top of Zhangs scheme. A novel data deduplication technique is to improve not only the utilization of cloud storage capacity but also enhances the rapid performance of deduplication in cloud storage.

Wee Keong Ng, *et al.* [12] introduced a new notion that the author calls private data deduplication protocols are formalized in the context of two-party computations. The private data deduplication protocols has been analyzed and proposed as a feasible result. The simulation based framework is provably secure protected by the proposed private data deduplication protocol. The hash function is collision-resilient, and the discrete logarithm is hard and the erasure coding algorithms E can erasure up to α-fraction of the bits in the presence of malicious adversaries.

S.Sadeghi, *et al.* [13] suggested a novel encryption scheme that affords the vital security for both unpopular data and popular data. For unpopular data suggested another two-layered encryption scheme with stronger security while supporting deduplication process. The traditional conventional encryption is mainly performed for popular data that are not specifically sensitive. Thus, they achieved enhanced trade between the security and efficiency of the out-sourced data. In block-level deduplication, Li et al. addressed the key management risk resolved by distributing these keys across multiple servers after that encrypting the files.

S. Halevi, *et al* [14] proposed the deduplication systems as the notion of "proofs of ownership" (PoW) in which a client can prove to a server depending on Merkle trees and the error-control coding that it indeed has uploading without a copy of a file but their scheme may not assurance the freshness of the proof in every challenge. Additionally, this scheme has to make Merkle Tree on the encoded data and it has inherently inefficient and not consider about data privacy.

S. Bugiel, *et al* [15] reviewed from this research, the confident outsourcing of data and haphazard reckonings which is determined to an untrusted commodity cloud. The combination of the trusted cloud and Commodity cloud is considered as a form of twin cloud. In trusted Cloud, the security precarious maneuvers are accomplished, however the Commodity cloud the enactment precarious maneuvers are accomplished on scrambled data. A trustworthy cloud that verifies and encrypts the data stowed in the untrusted commodity cloud (either a private cloud or numerous secure hardware sections). The outsourced data is based on integrity and confidentiality protected. The client verified the exactness of the outsourced reckonings and it transfers the data depending upon the query.

M. Armbrust, *et al* [16] analyzed to point out a few security problems with convergent encryption, whereas, recommending two protocols and a security model for protecting data deduplication. Whereas the two models are like same and slightly vary in security properties. These two approaches protect via authenticated and anonymous to secure deduplication data and that can be applied to single server storage and distributed storage. In the earlier, single server storage, clients interact with a single file server that stores both metadata and data. Afterward, metadata is stored on an independent metadata server, and data is stored on a series of object-based storage devices (OSDs).

Shai Halevi, *et al* [17] suggested the cloud storage systems are becoming more popular concept. In Client-side deduplication, that attempts to recognize with deduplication opportunities previously at the client side and keep the bandwidth for uploading copies of existing files to the server. More particularly client-side deduplication process, an attacker to gain access to arbitrary-size files who identifies the hash signature of a file can ensure the storage service that it owns that file; therefore the server lets the attacker download the entire file. Deduplication stores only a single copy of duplicating data that remains cost effective.

Q. He, *et al* [18] presented that the deduplication preserves only one copy of the duplicate data afforded with a pointer to point to the duplicate blocks and that can be complete at file level, block level or byte level. The new data are evaluated with old data at byte level, if it is match then marked as duplicate. The redundant copy is deleted and data pointers are updated.

Yuan, *et al.* [19] proposed a deduplication system is to save the storage size of the tags for integrity check in the cloud storage. Another third party called key server is set up to generate the file tag for performing duplicate check process. To improve the security process of deduplication and protect the data confidentiality, Bellare et al. explained how to secure and protect and the data confidentiality by transforming the predictable message into unpredictable message in an effective manner.

Z. Li, *et al.* [20] discussed various cloud storage systems with respect to data deduplication. It has recommended only the unique instance of the data, therefore, saving data storage volumes. The data deduplication engine has created an index of the digital signature for the data segment along with the signature of a given repository to identify data blocks, if it is already present after that a pointer is provided by the index otherwise not able to provide. Moving data from one storage system to another which are at different geographical locations is referred to as data migration and it aims at keeping and cooperating load balance in cloud storage system. Data migration into other cloud storage units have to occur and the pointers to be kept in the old stored positions intact, or modify and update the index as changes occur. However, this may bring overhead to access bottleneck and network bandwidth to concurrent access of clients.

D. Harnik, *et al.* [21] presented a cloud storage services generally utilize deduplication, which eradicates redundant data by storing only a single copy of each block or file. Deduplication saves the bandwidth and space requirements of data storage services and it have most effective when applied across multiple users. Deduplication can be utilized as a covert channel by malicious software communicates with its control center and firewall settings at the attacked machine. Cloud storage providers are suspect to stop using this technology because of the high savings offered by cross-user deduplication so that they suggest easy mechanisms that enable cross-user deduplication while greatly saving the risk of data leakage.

Paul Anderson, *et al.* [22] suggested a unique feature allows immediate detection of common subtrees. It avoids the need to query the backup system for every file. Encrypted deduplication, the security of data and users to enhance the speed of backup and to support client-end confidentiality and reduce the storage requirement.

Camble, *et al.* [23] proposed the "Sparse Indexing" deduplication system which uses a different approach to avoid the chunk lookup disk bottleneck. Sparse Indexing permits to save a chunk multiple times if the similarity based system is not able to detect the segments, which already have stored the chunk. At this point, the chunks are consecutively grouped into segments and that segments are utilized to search similar existing segments by using a RAM based index, whereas stores only a small fraction of the already stored chunks. Hence, Sparse Indexing is considered a member of the class of approximate data deduplication systems.

G. Neven, *et al.* [24] analyzed security attacks or proofs for a large number of signature schemes and identity-based identification described either implicitly or explicitly in existing literature. A framework handles the one hand supports and how to explain these schemes are derived and on other hand allows modular security analyzes, thereby easy to understand, and unify previous work. It also may analyze a generic folklore construction in particular yields signature schemes and identity-based identification without random oracles.

Mohamed Adel Serhani, *et al.* [25] proposed a Quality of Big Data evaluation scheme to generate a set of actions to enhance and improve the data quality of Big data set. A Big Data quality evaluation algorithm developed based on a bootstrap sampling and BLB. By saving computation time and resources, the BLB sampling supported for achieving an efficient DQ evaluation. Hence, the results are analyzed a set of based generated proposals and the data quality scores. These proposed actions are improved on the source data set to enforce and increase its quality.

S. Quinlan, *et al.* [26] briefly described the archival data storage, which imposes a write-once policy, malevolent obliteration or thwarting fortuitous of data. It is an edifice wedge for fabricating an assortment of stowing solicitations like Polaroid dossier, corporeal backup, and coherent backup. Venti ascertains data lumps through a hotchpotch of their innards. Adequately enormous output has determined the hash of a data block as irreplaceable for utilizing the collision-resistant muddle so that have been resolved. The whorl of a block is baptized by the inimitable hash after that the address utilizes a read and write operations. The functionality utilities tar and zip, Vac of archives and almanacs as a solitary object. As tree of lumps on a Venti server are stockpiled the innards of nominated archived. Therefore that process affords security protection and archive of data storage.

J.R. Douceur, *et al.* [27] proposed the Farsite distributed file system that affords availability for replicating each file onto multiple desktop computers. This replication consumes considerable storage space,

it is necessary to reclaim used space where possible way. Measurement of over 500 desktop file systems illustrates that duplicate files is engaged almost half of all consumed storage space. Our mechanism comprises convergent encryption that allows duplicate files to combine into a single file, even though the files are encrypted with different users.

III. CONCLUSION

De-duplication is very important in live virtual machine migration to transfer of VMs with the available bandwidth and less migration time. Data De-duplication eliminates the redundant data by performing only the single copies of data in storage space and it is also essential for utilizing cellular networks, wired communication, backup services process, wireless communication, etc., to save the amount of data in storage and to rapid up the backup process. This duplicate data elimination has great benefits for cloud storage space by using deduplication. We have reviewed different existing techniques are suggested by researchers for deduplication process.

IV. REFERENCES

[1]. Deepavali Bhagwat, Kave Eshghi, Darrell D. E. Long,Mark Lillibridge,"Extreme Binning: Scalable, Parallel Deduplication for Chunk-based File Backup, 17th IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'2009), London, UK, September 2009

[2]. G.Prashanthi, Z.Shobarani (2015), A Hybrid Cloud Approach for Secure Authorized Deduplication, International Journal of Innovative Research in Computer and Communication Engineering Vol.3, Special Issue 4.

[3]. BhushanChoudhary, Amit Dravid (2014), A Study on Authorized Deduplication Techniques in Cloud Computing, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 12.

[4]. M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless:Server aided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

[5]. Mark Lillibridge, Kave Eshghi, Deepavali Bhagwat, Vinay Deolalikar, Greg Trezise, and Peter Camble,"Sparse Indexing: Large Scale, Inline Deduplication Using Sampling and Locality", 7th USENIX Conference on File and Storage Technologies

[6]. C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.

[7]. Wen Xia, Hong Jiang, Dan Feng,Lei Tian, "DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads", information: DOI 10.1109/TC.2015.2456015, IEEE Transactions on Computers.

[8]. Pasqualo Puzio, Refik Molva ,MelekOnen ,"CloudDedup: Secure Deduplication with Encrypted Data for Cloud Storage", SecludIT and EURECOM,France.

[9]. S. Keelveedhi, M. Bellare, and T. Ristenpart, —Dupless: Serveraided encryption for deduplicated storage,‖ in *Proceedings of the 22Nd USENIX Conference on Security*, ser. SEC'13. Washington, D.C.: USENIX Association, 2013, pp. 179–194. [Online]. Available:https://www.usenix.org/conference/usenixsecurity13/technicalsessions/presentation/bellare

[10]. Wee Keong Ng, Yonggang Wen, Huafei Zhu,"Private Data deduplication Protocols in Cloud Storage", *SAC'12* March 2529, 2012, Riva del Garda, Italy.

[11]. W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012.

[12]. Iuon –Chang Lin, Po-ching Chien ,"Data Deduplication Scheme for Cloud Storage" International Journal of Computer and Control(IJ3C),Vol1,No.2(2012)

[13]. Bugiel, S., N¨urnberger, S., Sadeghi, A.-R., Schneider,T.: Twin Clouds: An architecture for secure cloud computing (Extended Abstract). In: Workshop on Cryptography and Security in Clouds (WCSC 2011), March 15-16 (2011)

[14]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[15]. S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.

[16]. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski,G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M.Zaharia, "A view of cloud computing," Communication of the ACM, vol. 53, no. 4, pp.50–58, 2011.

[17]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, —Proofs of ownership in remote storage systems,‖ in *Proceedings of the 18th ACM Conference on Computer and Communications Security* . ACM, 2011, pp. 491–500.

[18]. Q. He, Z. Li, and X. Zhang, Data deduplication techniques, in International Conference on Future Information Technology and Management Engineering,pp.431-432,2010.

[19]. P. Yuan and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.

[20]. Z. Li, X. Zhang, and Q. He, Analysis of the key technology on cloud storage, in International Conference on Future Information Technology and Management Engineering, 2010, pp. 427428.

[21]. D. Harnik, B. Pinkas, and A. Shulman-Peleg. Side channels in cloud services: Deduplication in cloud storage. IEEE Security & Privacy, 8(6), 2010.

[22]. Mohamed Adel Serhani, Chafik Bouhaddioui, Rachida Dssouli," Big Data Quality: A Quality Dimensions Evaluation",ResearchGate, DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122

[23]. P. Anderson and L. Zhang. "Fast and secure laptop backups with encrypted de-duplication". In *Proc. of USENIX LISA*, 2010.

[24]. K. Camble and E. Miller, "The effectiveness of deduplication on virtual machine disk images" In Proc. SYSTOR 2009: The Israeli Experimental Systems Conference..

[25]. M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009.

[26]. S. Quinlan and S. Dorward. Venti: a new approach to archival storage. In Proc. USENIX FAST, Jan 2002.

[27]. Douceur JR, Adya A, Bolosky WJ, Simon D, Theimer M. "Reclaiming Space fromDuplicate Files in a Serverless Distributed File System," in Proc. ICDCS, 2002, 617-624.