International Journal of Research in Signal Processing, Computing & Communication System Design
4 (2), 2018, 36-41
http://www.publishingindia.com/ijrsp

Prevention of Big Data by Secure Deduplication in Cloud Storage

Rashmi Gopalrao Mate[1] and Mohd. Saif Wajid[2*]
[1]M.Tech Student, CSE Department, Babu Banarasi Das University, Lucknow, Uttar Pradesh. India.
Email: rashmimate89@gmail.com
[2]Assistant Professor, CSE Department, Babu Banarasi Das University, Lucknow, Uttar Pradesh. India.
Email: mohdsaif06@gmail.com
*Corresponding Author

Abstract: As we seen Big data include a huge amount of data or massive data. Nowadays generation big data increases everywhere like microblogs and cloud storage also. Big data have three basic characteristics, velocity, volume, and variety. It is important to think about the storage capacity of cloud, speed and time while uploading and downloading the data in the cloud. Deduplication scheme utilizes to reduce the area and bandwidth requirements of services by removing redundant data and storing a single instance of them in the cloud storage. The powerfulness of deduplication method has occurred when multiple utiliser outsource the same data to turn out to cloud storage server. But somewhere we can get a problem of security and ownership. Recently several deduplication techniques have been seen to solve this matter of security, ownership, and operation on big data can generate in the cloud. Consideration to this we devise a server-side deduplication method to prevent big data in cloud storage for encrypted data. It permits utiliser to outsource the data even when ownership changes dynamically by exploiting randomized convergent encryption and secure ownership group key distribution. By using deduplication we try to reduce space and save the time while uploading and downloading the data in cloud storage. The proposed scheme is almost effective, preventing big data and saving time. In this paper we use this method in cloud storage onwards we try to use anywhere if there is big data available.

Keywords: Big data, Cloud storage, Deduplication, Dynamic ownership, Encryption.

I. Introduction

We hear the term big data nowadays everywhere like news article to professional magazines, from tweets to cloud storage data. The word originated by Roger Magouras from O'Reilly media in 2005 [1], shows a vast range of huge data sets almost impossible to organize and process [2]. The real issue is not that you can acquire huge amounts of data. It's what you do with data that counts, how can manage a data properly [3]. As we seen a generation of data increases rapidly in a cloud. So need to think about this big data storage in a cloud.

Cloud computing provides a low cost, scalable and location independent online services ranging from simple backup services to cloud storage infrastructure. The fast growth of data in the scene of big data in large volume, high velocity stored in cloud storage has lead to an increased demand for techniques for saving disk space and network bandwidth, to reduce resource utilization, large number of cloud storage system such as Google drive [7], Dropbox [4], Waula [5], Mozy [6] employ the deduplication scheme, where the cloud server save only a solitary instance of deduplicate data and give linkage to the copy instead of storing other true copies of that data, the saving are significant [8], and report to business approaches can gain disk and bandwidth saving of more than 90%. However, from security percept the shared usage data increases new challenges. As users are worried about their confidential data, they may encrypt their data before outsourcing in order to secure data privacy from uncertified outmost adversaries, as well as from the cloud service provider [10], [11], and [12].

Be that as it may, due to conventional encryption deduplication ended up inconceivable for taking after reason. Deduplication procedures take advantages of data proportionality to recognize the similar data and diminish the capacity region. In contradiction, encryption algorithms randomized the encrypted files in order to create ciphertext not distinguishable from conceptually random data. Encryption of similar data by unlike user with unlike encryption keys results in different ciphertext, which creates it hard for the cloud server to check whether the plain data are the similar and deduplicate them. Say Rashmi encrypts a file M under the secret key SkA and stores it corresponding ciphertext CA. Boby would save CB, which is encryption of M under his secret key SkB. Then two issues arise. First is how can cloud server detect that the underlining

file M is same and second is even if it can detect this, how can it allow both parties to recover the store data, based on their separate secret keys?

Straight forward client side encryption that is secure against a chosen plain text attached with randomly chosen encryption keys prevents deduplication [13], [14].

One credulous arrangement is to permit each client to scramble the information with an open key of the cloud capacity server. At that point, the server is able to deduplicate the distinguished by unscrambling it with its private key match. In any case, this arrangement permits the cloud capacity server to get the outsourced plain information, which may abuse the protection of the information in case the cloud server cannot be completely trusted [15], [16].

Convergent encryption [17] rotates this issue effectively. A convergent encryption algorithm scrambles input records with the with the hash esteem of the input record as an encryption key. The ciphertext is given to server and the client holds the encryption key. Since concurrent encryption is deterministic, indistinguishable records continuously encrypted into indistinguishable ciphertext, notwithstanding who scramble them. In this way, the cloud capacity server can execute deduplication, over the ciphertext, and each owner of the record can download the ciphered content (after the confirmation of possession handle alternatively) and unscramble it, afterward it since they have the same encryption key for the record. Long time since merged encryption has been profoundly considered in the commercial framework and has diverse encryption factors for secure deduplication [11], [18], [19], [20] which was formalized as message bolted encryption [21]. Indeed so, concurrent encryption endures from security blemishes with respect to tag consistency and ownership.

As an example of tag consistency attack issue, suppose Rashmi and Boby have the same data M, and Rashmi generates ciphertext CA from M and then maliciously generate another ciphertext C'A from M'(=! M). Next, she uploads C'A with an honesty generate a tag T (CA) = H (M) for cryptographic hash function H, which plays a role of data index.

When Boby produces cipher content CB from M and tries to transfer CB, the cloud server checks T (CA) =T (CB) at that point, it erases CB and keeps as it were C'A. A short time later, when Boby downloads and decrypt it, the information would be M', not M, which implies the integrity of his information, has been compromised.

Recently, message-locked encryption (MLE) [21] and leakage-reselient deduplication [22] schemes have been proposed to solve this problem by introducing addition integrity check phase for decrypted data.

As of late, message-locked encryption (MLE) [21] and leakage-reselient deduplication [22] plans have been proposed to illuminate this issue by presenting expansion integrity check stage for decoded data. In case of ownership revocation, assume different clients have ownership of ciphertext outsourced in cloud capacity, as time passes, a few of these clients may ask the cloud server to erase or adjust their information, and at that point, the server erases the ownership data of the clients from the ownership list for the comparing information. At that point, the repudiated client ought to be anticipated from getting to information put away in the cloud capacity after the cancellation or alteration asks (forward secrecy). On another hand, when the client transfers a information that as of now exist in cloud capacity, the client ought to be deuterated from getting to a information that was put away sometime recently the gotten the proprietorship by uploading it (in reverse secrecy).

These dynamic ownership changes may happen exceptionally habitually in a down to earth cloud framework, and in this way, it ought to be appropriately overseen in arrange to avoid the degradation of cloud benefit. Be that as it may, the past deduplication plot could not accomplish secure get to control beneath an energetic ownership changing the environment, in show disdain toward of its significance to secure deduplication, since the encryption key is determined deterministically and once in a while upgraded after the beginning key derivation. Therefore, for long as denied clients keep the encryption key, they can get to the comparing information in cloud capacity at any time, in any case of the validity of their ownership. This is the issue we endeavor to solve in this study

II. Literature Survey

On the basis of extensive literature survey related to the data deduplication with dynamic ownership management in cloud storage has been taken into consideration in this chapter.

A. Study about the Practical Deduplication

D. T. Meyer, and W. J. Bolosky [23] has proposed that File frameworks regularly contain repetitive duplicates of data: indistinguishable documents or sub-record locales, perhaps put away on a solitary host, on a mutual stockpiling group, or moved down to optional capacity. Deduplicating stock piling frameworks exploit this excess to decrease the fundamental space expected to contain the record frameworks (or reinforcement pictures thereof).

Deduplication can work at either the sub-document or entire record level. All the more fine-grained deduplication makes more open doors for space reserve funds, yet fundamentally lessens the successive format of a few records, which may have critical execution impacts when hard plates are utilized for

capacity (and at times requires confused strategies to enhance execution.

B. Learning Data Deduplication Ratios

M. Dutch [9] has stated that with respect to the comprehension of information deduplication proportions that information deduplication brings down business dangers, builds income openings, and diminishes stockpiling level expenses, bringing about an ideal tempest for organizations sending a versatile stockpiling foundation. Capacity flexibility advances, for example, RAID or RAIN, shield the reduplicated information to guarantee high accessibility of utilizations getting to the information.

The financial aspects of information deduplication make it more than convincing; it is required for any business looking to expand their client benefit levels. Information deduplication proportions are anything but difficult to over-break down and credit advantages to, that could conceivably exist.

C. Cloud Storage Having Private Data Deduplication Protocol

W. K. Ng, *et al.* [10] has proposed approximately another thought which we call private data deduplication tradition, a deduplication strategy for private data stockpiling is displayed and formalized. Instinctively, a private information deduplication convention permits a customer who holds private information demonstrates to a server who holds an outline string of the information that he/she is the proprietor of that information without uncovering additional data to the server.

D. Safe Data Deduplication

M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller [11] has proposed about the Businesses and buyers are ending up noticeably progressively aware of the estimation of secure, chronicled information stockpiling. In the business field, information safeguarding is regularly commanded by law, and information mining has ended up being an aid in forming business system.

For people, recorded capacity is being called upon to save nostalgic and authentic ancient rarities, for example, photographs, films and individual archives. Further, while few would contend that business information calls for security, protection is similarly imperative for people; information, for example, medicinal records and authoritative reports must be kept for drawn out stretches of time however should not be openly available. Incomprehensibly, the expanding estimation of authentic information is driving the requirement for cost-productive capacity; reasonable capacity permits the conservation of all information that may in the long run demonstrate helpful.

E. Merging End to End Privacy

N. Baracaldo, E. Androulaki, J. Glider, A. Sorniotti [12] represented that Cloud reckoning has developed as exceptionally useful for organizations that hope to decrease their expenses, send new applications quickly or that do not have any need to stay up their own process framework. In any case, late info ruptures in clear distributed storage suppliers have created customers be increasingly troubled concerning the classification of their (outsourced) info.

There are things wherever client info was conferred to and spilled by cloud provider representatives that had physical access to the capability medium, and moreover, wherever cloud clients accessed alternative customer's info within the wake of getting been distributed physical warehousing quality already alloted to a different customer e.g., then alternative client had worn out its distributed storage membership (in this paper we tend to suggest to that as indweller off boarding).

F. A Original Scheme of Encryption for Data Deduplication System

C. Wang, Z. Qin, J. Peng, and J. Wang [14] has stated another information pressure innovation, which is known as information deduplication, come in locate in pace with the gigantic increment of electronic information. Information deduplication partitions information into settled or variable size pieces and the cryptographic hash estimation of each lump is utilized as the piece's worldwide one of a kind 10 with the end goal that excess information might be distinguished.

These redundancies are utilized to either decrease stockpiling limit needs or to lessen organize activity. Not at all like conventional pressure technique, deduplication recognizes normal groupings of bytes both inside and amongst records, and just stores a solitary occasion of each piece paying little mind to the quantity of times it happens.

G. Reclaiming Space from Duplicate Files

J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer [17] in this it's address regarding the problems of recognizing and mixture indistinguishable documents within the Farsite circulated record framework, with the top goal of sick stowage eaten up by unexpectedly repetitive substance. Farsite may be a protected, versatile, server less document framework that coherently capacities as a focused record server but that's physically disseminate among associate organized accumulation of desktop workstations. Since desktop machines don't seem to be typically on, not halfway oversaw, and not physically secured, the area recovery method should endure a high rate of framework disappointment, work while not focal coordination, and capability couple with science security.

H. Safe and Fast Backup of Laptop with Encrypted

P. Anderson, L. Zhang [18] had stated that a run of the mill cluster of moveable digital computer shoppers share plenty of knowledge in like manner. This provides the chance to essentially diminish reinforcement times, and capability conditions. In any case, we've got incontestable that manual determination of the many info- as an example, moving down simply home registries - may be a poor methodology; this neglects to reinforcement some important records, within the meanwhile as superfluously repeating totally different documents. We've got exhibited a model reinforcement program that accomplishes a perfect level of sharing within the mean while as taking care of secrecy.

This adventures a unique calculation to diminish the number of documents that ought to be checked and later diminishes reinforcement times. We've got incontestable that run of the mill cloud interfaces, as an example, Amazon S3 don't seem to be acceptable to the present reasonably use, due to the time and price of run of the mill exchanges, and therefore the absence of multi-client verification to shared info.

We have delineated a usage utilizing a neighborhood server which might be from these problems by storing and prepreparing info before sending to the cloud. This can be looked as if it would accomplish noteworthy price funds 2.9: Safe backup of cloud system with guaranteed deletion.

I. A Secure Cloud Backup System with Assured Deletion

A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, J. C. S. Lui [20] has pictured that Cloud registering could be a rising administration demonstrate that offers calculation and capability assets on the net. One appealing utility that distributed computing can give is distributed storage. Folks and ventures area unit often needed to remotely document their info to dodge any knowledge misfortune within the event that there are a unit any equipment / programming disappointments or unforeseen calamities. Rather than getting the desired warehousing media to stay info reinforcements, folks and undertakings will simply source their info reinforcement administrations to the cloud specialist co-ops, which provide the important ware housing assets to possess the knowledge reinforcements. Whereas distributed storage is appealing, a way to provide security assurances to outsourced info turns into a rising concern.

J. Leakage of Resilient Client Side Deduplication

J. Xu, E. Chang, and J. Zhou [24] has expressed that Cloud reposition administration is learning prominence as recently. To diminish quality utilization in each system transmission capability and capability, several distributed storage administrations as well as Dropbox one and Wuala a pair of utilizes client aspect deduplication. That is, the purpose at that a consumer tries to transfer a document to the server, the server checks whether or not this specific record is as of currently within the cloud (transferred by some consumer beforehand), and spares the transferring procedure within the event that it's as of currently within the distributed storage.

Along these lines, every and each record can have simply one duplicate within the cloud (i.e. Single Instance Storage). SNIA study declared that the deduplication procedure will put aside to ninetieth capability, dependent on applications. As per Halevi *et al.* what is a lot of, Drop ship, a current usage of client aspect deduplication is as beneath: Cloud consumer Alice tries to transfer a record F to the distributed storage.

III. Contribution

We introduce a secure deduplication over encrypted data to prevent big data in cloud storage. The propose scheme ensure that prevention of big data which can generates in cloud storage by using deduplication technique, while doing this we will try to reduce the time, when uploading and downloading from the cloud storage.

This scheme promises that only authorized access to the shared data is possible which is considering being the most important challenge [21] in the environment where ownership changes dynamically.

IV. Proposed System

A few deduplications have been proposed to handle this issue by empowering each owner to have similar encryption key for comparative data. In any case, the larger part of the arrange involvement the sick impact of the security imperfections, since they don't consider the dynamic changes in the duty for the data that happen as regularly as conceivable in a down to business disseminated capacity advantage. In this proposal, we introduce a novel server-side deduplication method for encrypted information; in this research, we attempt to anticipate the Huge Information by deduplication strategy which is in cloud storage. It empowers the cloud server to control get to outsourced data the proposed conspire to ensure that security and verification of proprietorship. In proposed conspire we break the content into little blocks based on given block measure and the comparison done to each other.

Advantages of proposed system:

- Produce data labels sometime recently uploading as well as review the integrity of information has been put away in a cloud.
- Enabling the secure deduplication through presenting a POW protocol and anticipating the leakage of side channel data in information duplication.

- Integrity reviewing and secure deduplication straightforwardly on scrambled information.

V. Implementation

We try to run this pseudo code.

Begin.

Get Block Signature (BS).

Converts block signature into three parts-

First Bit (FB), Second Bit (SB), Third Bit (TB) to Last Bit (LB).

Get the Index of Code (IC) FB from zero Level BS.

Identifying the corresponding next level BS index using IC.

If (presence of SB in second level BS Index).

If (get (TB to LB) and check the presence of byte in third level).

Block Exist.

Else.

Else Block not exist.

End.

VI. Result

The accompanying depictions layout the outcomes or yields that we are going to get once regulated execution of the considerable number of modules of the framework.

Upload Process

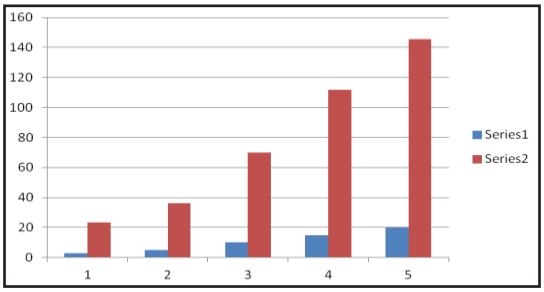

Fig. 1: Upload Process Result

While uploading the file, first step is break the file in small blocks based on given block size after that hash code get generated for all blocks, while generating hash code it will check whether it is new block of data or duplicate block of data based on hash code if hash code matched with existing hash code means it is duplicate block of data and if it is not matching means it is new data, all new block of data we will encrypt using AES encryption then we will upload to the cloud drive. As graph showing the result if file size is less it will take less time to upload and if file size is big it will take more time to execute.

Download Process

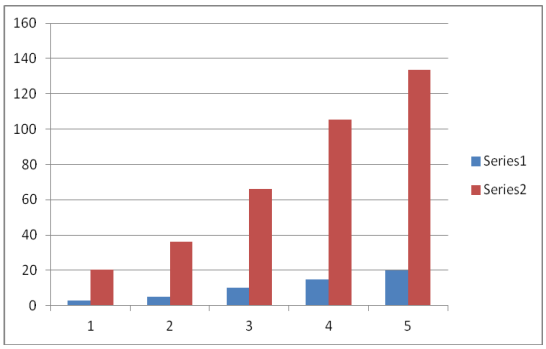

Fig. 2: Download Process Result

While downloading the file, first it will check how many blocks is there, after that it will start downloading that that block from cloud drive. While downloading blocks from cloud drive it will decrypt block content and after downloading the all blocks it will merge all block, to make a single file. So if file size is less it will take less time to download and file size is big it will take more time to download.

VII. Conclusion

This part for the most part talks about the papers that are eluded while making this thesis report. Every one of these papers give data identified with learning of aggregate conduct, their current arrangements, and strategies utilized furthermore their focal points and constraints.

VIII. Future Enhancement

In future enhancement we can add to upload many more file format text, images (png, jpg, gif, etc.), video files and if file size is very big (big data) that also we can use in this application, we can improve performance by reducing the time while uploading the file.

References

[1] http://strata.oreilly.com/2010/01/roger-magoulas-on-big-data.html

[2] G. Halevi, and H. F. Moed, "Overview of literature: The evaluation of big data as research and scientific topic," *Bibliometrics*, no. 30, September 2012.

[3] K. Shirudkar, and D. Motwani, "Big-data security," *International Journal of Advance Researcher in Computer Science and Software Engineering*, vol. 5, no. 3, pp. 1100-1109, March 2015.

[4] Dropbox, http://www.dropbox.com/

[5] Waula, http://www.waula.com/

[6] Mozy, http://www.mozy.com/

[7] Google drive, http://drive.google.com

[8] D. T. Meyer, and W. J. Bolosky, "A study of practical deduplication," *Proc. USENIX Conference on File and Storage Technologies*, 2011.

[9] M. Dutch, "Understanding data deduplication ratio," *SNIA, Data Management Forum*, 2008.

[10] W. K. Ng, W. Wen, and H. Zhu, "Private data deduplication protocols in cloud storage," *Proc. ACM SAC*, 2012.

[11] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure data deduplication," *Proc. StorageSS'08*, 2008.

[12] N. Baracaldo, E. Androulaki, J. Glider, and A. Saniotti, "Reconcilling end-to-end confidentiality and the data reduction in cloud storage," *Proc. ACM Workshop on Cloud Computing Security*, pp. 21-32, 2014.

[13] D. Harnik, B. Pinkaj, and A. Shulman-Peleg, "Side channels in cloud services, the case of deduplication in cloud storage," *IEEE Security and Privacy*, vol. 8, no. 6, pp. 40-47, 2010.

[14] C. Wang, Z.-G. Qin, J. Peng, and J. Wang, "A novel encryption scheme for data deduplication system," *Proc. International Conference on Communications, Circuits and Systems (ICCCAS),* pp. 265-269 2010.

[15] "Malicious insider attacks to rise." Available: http://news.bbc.co.uk/z/hi/7875904 stm

[16] "Data theft linked to ex-employees." Available: http://www.theaustation.com.au/austnation-it/data-theptlinked-toex-employees/story-e6frgakx-1226572351953

[17] J. R. Dauceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files a serverless distributed file systems," *International Conference on Distributed Computing Systems (ICDCS),* pp. 617-624, 2002.

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

[19] Z. Wilcox-O'Hearn, and B. Warner, "Tahoe: The least authority file systems," *Proc. ACM StorageSS'08*, 2008.

[20] A. Rahumed, H. C. H. Chen, Y. Yang, P. P. C. Lee, and J. C. S. Lui, "A secure cloud backup system with assured deletion and version control," *Proc. International Workshop on Security in Could Computing*, 2011.

[21] M. Mulazzani, S. Schritwieser, M. Leithner, and M. Huber, "Dark clouds on the horizon: Using cloud storage as attack vector and online slack space," *Proc. USENIX Conference on Security*, 2011.

[22] R. Mate, and Md. S. Wajid, "Big data analysis: Data management in microblogs," *ICICCI2017-IEEE*, 2017.

[23] D. T. Meyer, and W. J. Bolosky, "A study of practical deduplication," *Proc. USENIX Conference on File and Storage Technologies*, 2011.

[24] J. Xu, E. Chang, and J. Zhou, "Leakage-resilient client-side deduplication of encrypted data in cloud storage," ePrint, IACR. Available: http://eprint.iacr.org/2011/538