International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Storage Management System using Block Level Deduplication Technique in Cloud Computing

Mr.Amol Bapurao Suroshe and Prof. Abhijit Patankar

Department of Computer Engineering,Alard College of Engineering and Management

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Cloud storage as one of the most wanted cloud computing services helps cloud users overcome the bottleneck of limited resources and expand storage without upgrading their devices. To ensure the security and privacy of cloud users, data is always outsourced in encrypted form. However, encrypted data could generate a lot of storage waste in the cloud and complicate the exchange of data between authorized users. we are still facing challenges in storing and managing encrypted data with deduplication. Traditional deduplication schemes always focus on specific application scenarios, where deduplication is completely controlled by data owners or servers in the cloud. They cannot flexibly satisfy the different requests of data owners based on the level of data sensitivity. In this paper, author propose a heterogeneous information stockpiling the executives plot that deftly offers both deduplication the board and access control simultaneously over various cloud specialist co-ops (CSPs). creator assess execution with security examination, correlation and usage. The results show its safety, effectiveness and efficiency towards a possible practical use.

Keywords: *Cloud Computing, Data Deduplication, Access Control, Storage Management.*

Introduction

Despite the fact that cloud storage framework has been for the most part received, it neglects to oblige some significant developing needs, for example, the capacity of evaluating honesty of cloud documents by cloud customers and distinguishing copied records by cloud servers. creator uncover the two issues underneath. Thes cloud server can lighten clients from the monstrous load of limit the officials and backing. The most complexity of dispersed stockpiling from ordinary in-house accumulating is that the data is traveled through Internet and set aside in a faulty territory, not leveled out of the clients in any way shape or form, which unavoidably raises clients staggering stresses on the trustworthiness of their information.These stresses start from how the disseminated stockpiling is affected to security risks from both outside and inside the cloud, and the uncontrolled cloud servers may inactively hide a couple of data setback events from the clients to keep up their reputation. What is continuously certified is that for putting aside money and space, the cloud servers may even adequately and purposefully discard hardly found a good pace having a spot with a standard client. Considering the huge size of the redistributed information records and the customers' obliged asset capacities, the principal issue is summed up as by what means can the customer proficiently perform normally trustworthiness confirmations even without the nearby duplicate of information document. Distributed computing will be figuring in which enormous gatherings of remote servers are arranged to permit incorporated information stockpiling and online access to PC administrations or assets.

In cloud computing, enormous pools of assets can be associated through private or open system. Openly cloud, administrations (for example applications and capacity) are accessible for general use over the web. A private cloud is a virtualized server farm that works inside a firewall. Distributed computing gives calculation and capacity assets on the Internet. Expanding measure of information is being put away in the cloud and it is shared by clients with determined benefits, which characterizes unique rights to get to put away information. Dealing with the exponential development of ever-expanding volume of information has become a basic test. As indicated by IDC cloud report 2014, organizations in India are making a steady move from on premise heritage to various types of cloud. While the procedure is steady, it has begun by relocating certain application outstanding tasks at hand to cloud. To make adaptable administration of put away information in distributed computing, de-duplication has been notable method which turns out to be progressively well known as of late. De-

duplication is a particular information pressure strategy, which diminish extra room and transfer data transfer capacity in distributed storage. In de-duplication, just a single exceptional occasion of the information is very the server and excess information is supplanted with a pointer to the one of a kind information duplicate. Deduplication can occur either at record level or square level. From the client viewpoint, security and protection concerns are emerge as information are helpless to both insider and outcast assault. We should appropriately authorize classification, trustworthiness checking, and get to control systems the two attacks.

De-duplication does not work with traditional encryption. User encrypts their files with their individual encryption key, different cipher text would emerge even for identical files. Thus, traditional encryption is incompatible with data de duplication. Convergent encryption is a widely used technique to combine the storage saving of de-duplication to enforce confidentiality. In convergent encryption, the data copy is encrypted under a key derived by hashing the data itself. This convergent key is used for encrypt and decrypt a data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since encryption is deterministic, indistinguishable information duplicates will produce the equivalent concurrent key and a similar figure content. This permits the cloud to perform de duplication on the figure writings. The figure writings must be unscrambled by the relating information proprietors with their joined keys. Differential approval copy check is an approved de-duplication strategy where every client is given a lot of benefits during framework instatement. This arrangement of benefits indicates what sort of clients is permitted to perform copy check and access the records.

Literature Survey

" D. Meister, J. Kaiser, and A. Brinkmann[1] created Characteristics of reinforcement remaining tasks at hand underway systemsThe creator shows a total portrayal of reinforcement outstanding burdens by investigating measurements and substance metadata gathered from a huge arrangement of EMC Data Domain reinforcement frameworks underway use. This examination is finished (it covers the measurements of more than 10,000 frameworks) and inside and out (it utilizes point by point hints of the metadata of various creation frameworks that store practically 700TB of reinforcement information). We contrasted these frameworks and a nitty gritty investigation of Microsoft's essential stockpiling frameworks and exhibited that back-up capacity varies fundamentally from the essential stockpiling remaining burden as far as information amounts and limit necessities, just as the measure of information stockpiling limit. excess inside the information. These properties offer one of a kind difficulties and openings when planning a plate based record framework for reinforcement remaining tasks at hand.

M. Lillibridge, K. Eshghi, and D. Bhagwat[2] developed Primary data deduplication-large scale study and system design The author introduces a huge scale investigation of essential information deduplication and utilizations the outcomes to manage the structure of another essential information deduplication framework executed in the Windows Server 2012 working framework. The file data were analyzed by 15 servers of globally distributed files that host data for over 2000 users in a large multinational company. The results are used to achieve a fragmentation and compression approach that maximizes deduplication savings by minimizing the metadata generated and producing a uniform distribution of the portion size. Deduplication processing resizing with data size is achieved by a frugal hash index of RAM and data partitioning, so that memory, CPU and disk search resources remain available to meet the main workload of the IO service.

"V. Tarasov, A. Mudrankit, W. Buik, P. Shilane, G. Kuenning, and E. Zadok[3]:Propose a new storage reduction scheme that reduces data size with comparable efficiency to the most expensive techniques, but at a cost comparable to the fastest but least effective. The plan, called REBL (Block Level Redundancy Elimination), misuses the benefits of pressure, cancellation of copy squares and delta encoding to dispose of a wide range of repetitive information in a versatile and effective manner. REBL by and large encodes more minimalistically than pressure (up to a factor of 14) and a blend of pressure and concealment of copies (up to a factor of 6.7). REBL is additionally coded also to a method dependent on delta encoding, which essentially diminishes the general space for a situation. Likewise, REBL utilizes super unique mark, a strategy that lessens the information expected to recognize comparable squares by definitely diminishing the computational prerequisites of the coordinating squares: it converts the comparisons of O (n2) into searches of hash tables. As a result, the use of super fingerprints to avoid enumerating the corresponding data objects decreases the calculation in the REBL resemblance phase of a couple of orders of magnitude.

D. T. Meyer and W. J. Bolosky[4]:The data and the private cloud where the token generation will be generated for each file. Before uploading the data or file to the public cloud, the client will send the file to the private cloud for token generation, which is unique to each file. Private clouds generate a hash and token and send the token to the client. The token and hashes are kept in the private cloud itself, so that whenever the next token generation file arrives, the private clone can refer to the same token. Once the client gets the token for a given file, the public cloud looks for the token similar if it exists or not. If the public cloud token exists, it will return a pointer to the existing file, otherwise it will send a message to load a file. A system

that achieves confidentiality and allows block-level deduplication at the same time. Before uploading the data or file to the public cloud, the client will send the file to the private cloud for token generation, which is unique to each file. The private cloud generates a hash and token and sends them to the client. The token and the hash are kept in the private cloud itself so that whenever the next token generation file arrives, the private clone can refer to the same token.

G. Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu[5]:They are getting data deduplication by providing data evidence from the data owner. This test is used when the file is uploaded. Each file uploaded to the cloud is also limited by a set of privileges to specify the type of users who can perform duplicate verification and access the files. New duplication constructs compatible with authorized duplicate verification in the cloud hybrid architecture where the private cloud server generates duplicate file verification keys. The proposed system includes a data owner test, so it will help implement better security issues in cloud computing.

El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta[6]: They proposed recovery due to the fragmentation of the parts is a serious problem faced by data deduplication systems in one piece: the recovery speeds for the most recent backup can eliminate orders of magnitude during the life cycle of a system. We have studied three techniques: increase the size of the cache, limit the containers and use a direct assembly area to solve this problem. Limiting the container is a time-consuming task and reduces fragmentation of fragments at the cost of losing part of the deduplication, while using a direct assembly area is a new technique of recovery and caching in the recovery process which exploits the perfect knowledge of the future access to the fragments available during the restoration of a backup to reduce the amount of RAM needed for a certain level of caching in the recovery phase.

P. Shilane, M. Huang, G. Wallace, and W. Hsu[7]: Propose another methodology, called Block Locality Cache (BLC), which catches the past reinforcement execution fundamentally superior to existing methodologies and consistently utilizes around date data about the area and is in this way less inclined to maturing. We assessed the methodology utilizing a recreation dependent on the location of different arrangements of genuine reinforcement information. The reenactment contrasts the Block Locality Cache and the methodology of Zhu et al. also, gives a definite examination of the conduct and the IO design. What's more, a model execution is utilized to approve the reproduction.

P. Kulkarni, F. Douglis, J. D. LaVoie, and J. M. Tracey [8]:They collect data from the file system content of 857 desktop computers in Microsoft for a period of 4 weeks. We analyze the data to determine the relative efficiency of data deduplication, especially considering the elimination of complete file redundancy against blocks. We have found that full file deduplication reaches about three quarters of the space savings of more aggressive block deduplication for live file system storage and 87 of backup image savings. We also investigated file fragmentation and found that it does not prevail, and we have updated previous studies on file system metadata, and we have found that file size distribution continues to affect very large unstructured files.

Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou[9]: They built up a conventional model of record framework changes dependent on properties estimated in terabytes of genuine and diverse stockpiling frameworks. Our model interfaces with a conventional structure to copy changes in the record framework. In light of perceptions from explicit conditions, the model can produce an underlying document framework followed by constant changes that copy the conveyance of copies and record sizes, sensible changes to existing documents and record framework development.

Shweta D. Pochhi, Prof. Pradnya V. Kasture [10]: They found the improved WAN replication of reinforcement informational indexes utilizing delta pressure revealed by the stream Off-site information replication is basic for catastrophe recuperation reasons, yet the present tape move approach is lumbering and blunder inclined. Replication in a wide territory arrange (WAN) is a promising other option, yet quick system associations are costly or unrealistic in numerous remote areas, so better pressure is expected to make WAN replication commonsense. We present another procedure for reproducing reinforcement informational collections through a WAN that expels copy document districts (deduplication) yet in addition packs comparative record locales with delta pressure, which is accessible as an element of EMC DataDomain frameworks."

Proposed Methodology

In this paper,propose a new approach in the challenge of data ownership and cryptography to manage the storage of encrypted data with deduplication. Goal is to solve the problem of deduplication in the situation where the data owner is not available or it is difficult to get involved. Meanwhile, the data size does not affect the performance of data deduplication in our schema. author is motivated to save space in the cloud and to preserve the privacy of data owners by proposing a scheme to manage the storage of encrypted data with deduplication and then test safety and evaluate the performance of the proposed scheme through analysis and simulation. The results show its efficiency, effectiveness and applicability.

A. Architecture

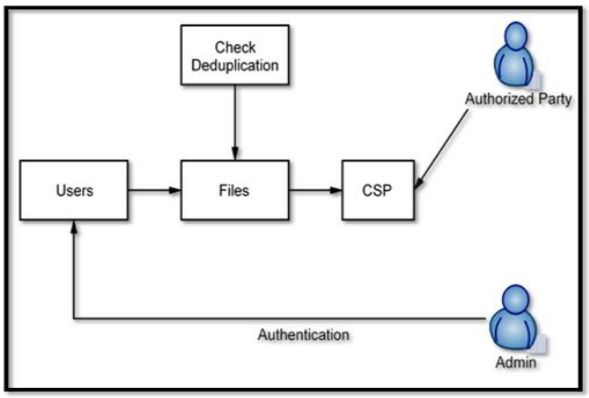

Figure 1. System Architecture

Cloud Service Provider: The CSP allows the owner of data for data storage services. You can not trust completely. This is why the content of the stored data is curious. It must be done honestly in the conservation of data for profit.

Data Holder: The data owner can upload and save his data and files in the CSP. In this system, it is possible that the number of data holders may store their files in raw cryptographic data in the CSP. The owner of the data that produces or creates the file considers the file as the owner of the data. The owner of the data is in normal form that the highest priority of the owner.

Authorized Party: An authorized party where data owners trust completely. Data holders to verify data ownership and manage data deduplication. It does not converge with the CSP. In this case, CSP does not need to know the user data in its memory.

B. Algorithms

1. AES Algorithm for Encryption and Decryption. AES (advanced encryption standard).It is symmetric algorithm. It used to convert plain text into cipher text .The need for coming with this algo is weakness in DES. The 56 bit key of des is no longer safe against attacks based on exhaustive key searches and 64-bit block also consider as weak..

Input:

128 bit /192 bit/256 bit input (0, 1) Secret key (128 bit) +plain text (128 bit).

Process:

10/12/14-rounds for-128 bit /192 bit/256 bit input

Xor state block (i/p)

Final round:10,12,14

Each round consists: sub byte, shift byte, mix columns, add round key.

Output:

cipher text(128 bit)

2. MD5 (Message-Digest Algorithm) The MD5 message digest calculation is a generally utilized cryptography hash work that creates a 128-piece (16-byte) hash esteem, commonly communicated as content in 32-digit hexadecimal numbers. MD5 has been utilized in a wide assortment of cryptographic applications and is likewise ordinarily used to check information honesty.

Steps:

1. A message digest calculation is a hash work thatacknowledges a succession of bits of any length and produces an arrangement of bits of a little fixed length.
2. The yield of a message digest is considered as anadvanced mark of the information data.
3. MD5 is a message digest calculation that produces 128 bits of data.
4. Utilize the constants got from the trigonometric sinefunction.
5. Experience the first message in squares of 512 bits

C. Mathematical Model

KeyGenCE (M): K is the key generation algorithm that maps a data copy M to a convergent key K;

EncryptCE(K, M): C is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs a ciphertext C;

DecryptCE(K,C): M is the decryption algorithm that takes both the ciphertext C and the convergent key K as inputs and then outputs the original data copy M

TagGenCE(M): T(M) is the tag generation algorithm that maps the original data copy M and outputs a tag T(M). We allow TagGenCE to generate a tag from the corresponding ciphertext, by using T(M)=TagGenCE(C), where C=EncryptCE(K,M).

Result and Discussion

Proposed System tested the time spent to encryption and decryption a file with different sizes by applying AES with 2 different key sizes, namely 128 bits and 256 bits and observe from graph that encrypting or decrypting a file of 10 to 20 megabytes (MB) with 128-bit AES takes about 100 seconds. It is a reasonable and practical choice to apply symmetric encryption for data protection.

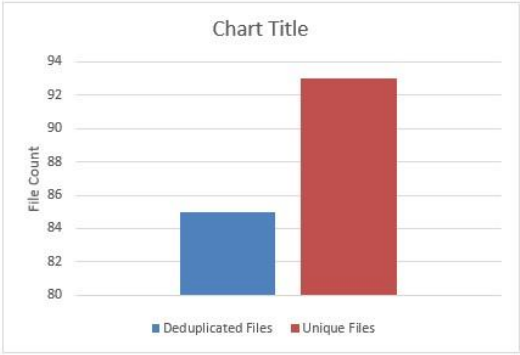

Figure 2.Graph

Table 1: Comparative Result

Parameter	Deduplicate Files	dUnique Files
Files	82	92

Conclusion

Data deduplication is important and significant in the practice of data storage in the cloud, in particular for

the management of big data. In this paper, proposed a heterogeneous data storage management scheme, which offers flexible data deduplication in the cloud and access control. This schema can be adapted to different scenarios and application requests and offers cost-effective management of big data storage across multiple CSPs. Data deduplication and access control can be achieved with different security requirements. Security analysis, comparison with existing work and implementation based performance evaluation have shown that our scheme is safe, advanced and efficient.

References

[1] Abhijit Janardan Patankar ,Dr. Kshama V. Kulhalli ,Dr. Kotrappa Sirbi,"Emotweet: Sentiment Analysis tool for twitter"2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT)

[2] Abhijit J. Patankar, Kotrappa Sirbi, Kshama V. Kulhalli"Preservation of Privacy using Multidimensional K-Anonymity Method for Non-Relational Data"IJRTE 2019.

[3] 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: PP Year 2014.

[4] D. Meister, J. Kaiser, and A. Brinkmann, "Block locality caching for data deduplication," in Proc. 6th Int. Syst. Storage Conf., 2013, pp. 1–12.

[5] M. Lillibridge, K. Eshghi, and D. Bhagwat, "Improving restore speed for backup systems that use inline chunk-based deduplication," in Proc. 11th USENIX Conf. File Storage Technol, Feb. 2013, pp. 183–197.

[6] V. Tarasov, A. Mudrankit, W. Buik, P. Shilane, G. Kuenning, and E. Zadok, "Generating realistic datasets for deduplication analysis," in Proc. USENIX Conf. Annu. Tech. Conf., Jun. 2012, pp. 261–272.

[7] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," ACM Trans. Storage, vol. 7, no. 4, p. 14, 2012.

[8] G. Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu, "Characteristics of backup workloads in production systems," in Proc. 10th USENIX Conf. File Storage Technol., Feb.2012,pp.33–48.

[9] 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.

[10] P. Shilane, M. Huang, G. Wallace, and W. Hsu, "WAN optimized replication of backup datasets using stream-informed delta compression," in Proc. 10th USENIX Conf. File Storage Technol.,Feb.2012,pp.49–64.

[11] P. Kulkarni, F. Douglis, J. D. LaVoie, and J. M. Tracey, "Redundancy elimination within large collections of files," in Proc.USENIXAnnu.Tech.Conf. Jun.2012, pp.59–72.

[12] 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