Journal of Theoretical and Applied Information Technology
20th April 2016. Vol.86. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REDUCED FILE HASH FOOT PRINTS FOR OPTIMIZED DEDUPLICATION IN CLOUD PLATFORMS

[1]M.JYOTHIRMAI, [2]Dr.K.THIRUPATHI RAO

[1]Student, Department of Computer Science & Engineering, K L University, Vaddeswaram

[2]Professor, Department of Computer Science & Engineering, K L University, Vaddeswaram

E-mail: [1]jyothimandelli@gmail.com, [2]kthirupathirao@kluniversity.in

ABSTRACT

Cloud Data Storages decreases colossal burden on clients as for their neighborhood stockpiles yet acquaints new issues with deference with information copies in the cloud. Albeit some prior methodologies managed the issue of actualizing a way to deal with handles cloud security and execution as for de-duplication by appropriately characterizing the concerned gatherings in the cloud and summoning document signature distinguishing proof procedure utilizing customary hash message validation code (HMAC). Because of these hash code calculations like SHA-1 and MD5 the document trustworthiness qualities are colossal prompting idleness variable at the de-duplication estimation. Because of this above issue the capacity exhibit obliges earlier trustworthiness hash codes prompting execution issues. In this paper, we propose a Genetic Programming way to deal with record deduplication that joins a couple of unmistakable bits of affirmation removed from the information substance to discover a deduplication limit that has the limit perceive whether two sections in a store are duplicates or not. As appeared by our trials, our methodology beats a current cutting edge technique found in the writing. Additionally, the proposed capacities are computationally less requesting since they utilize less confirmation. Furthermore, our hereditary programming methodology is prepared to do consequently adjusting these capacities to a given settled copy ID limit, liberating the client from the weight of choosing and tune this parameter.

Keywords: *Hybrid cloud computing, Cloud security, SHA, MD5, Message Authentication Codes, Genetic programming, Cross-over Mutation, Similarity Function, and Checksum.*

1. INTRODUCTION

Circulated registering gives obviously unlimited "virtualized" advantages for customers as organizations over the whole Internet, while hiding stage and use purposes of hobby. Today's cloud organization suppliers offer both exceedingly available limit and incredibly parallel figuring resources at modestly low costs. As appropriated processing gets the opportunity to be pervasive, a growing measure of data is being secured in the cloud and bestowed by customers to decided advantages, which describe the passageway benefits of the set away data. One fundamental test of conveyed stockpiling organizations is the organization of the continually growing volume of data.

To make data organization adaptable in conveyed processing, deduplication has been a comprehended strategy and has pulled in more thought starting late. Data deduplication is a specific data weight procedure for discarding duplicate copies of repeating data away. The framework is used to upgrade stockpiling utilize and can in like manner be associated with system data trades to reduce the amount of bytes that needs to be sent. In lieu of placing different copies of data with the similar substance, deduplication slaughters tedious data by keeping emerge physical copy and different overabundance data to that copy.

In spite of the truth that data deduplication brings a huge amount of favorable circumstances, security and insurance concerns rise as customers' fragile data are defenseless to both inside and untouchable attacks. Customary encryption, while giving data security, is opposite with data deduplication. Specifically, standard encryption requires assorted customers to encode their data with their own specific keys. Thusly, undefined data copies of assorted customers will provoke unmistakable figure writings, making deduplication unimaginable.

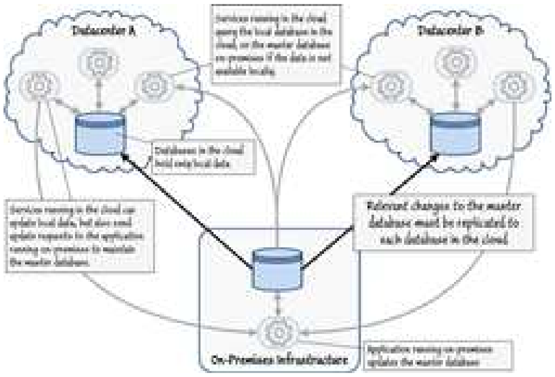
Datacenter A
Datacenter B
Services running in the cloud query the local database in the cloud, or the master database on-premises if the data is not available locally
Database in the cloud hold only local data
Services running in the cloud can update local data, but also send update requests to the application running on-premises to maintain the master database
Relevant changes to the master database must be replicated to each database in the cloud
Application running on-premises updates the master database
On-Premises Infrastructure

Figure 1: Secure Duplication System For Data Storage In Cloud.

To avert unapproved access, a protected check of ownership tradition is excessively expected, making it impossible to give the proof that the customer without a doubt has the same record when a duplicate is found. After the confirmation, coming about customers the similar archive will be given a pointer from the server without hoping to exchange the same record. A mixed record can be downloaded by the customer with the pointer from the server, which must be unscrambled by the contrasting data proprietors and their simultaneous keys. Along these lines, simultaneous encryption allows the cloud to perform deduplication on the figure writings and the check of proprietorship keeps the unapproved customer to get to the archive.

In this paper, we show a Genetic Programming (GP) approach to manage document deduplication. Our strategy solidifies a couple of exceptional bits of affirmation isolated from the data substance to convey a deduplication capacity that has the capacity distinguish whether two or more passages in an archive are imitations or not. Since record deduplication is a period expending errand notwithstanding for little archives, our point is to encourage a strategy that finds a legitimate blend of the best bits of proof, in this way yielding a deduplication capacity that amplifies execution utilizing somewhat illustrative section of the relating data for get ready purposes. At that point, this limit can be utilized on the remaining information or even connected to different archives with comparative qualities. Besides, new extra information can be dealt with likewise by the recommended capacity, the length of there are no sudden changes in the information examples, something that is extremely unrealistic in huge information vaults. It merits seeing that this (number juggling) capacity, which can be thought as a blend of an few successful deduplication standards, simple and quick to process, permitting an productive application to the deduplication of substantial vaults. A capacity utilized for record deduplication must achieve particular however clashing targets: it ought to proficiently augment the distinguishing proof of record copies while abstaining from committing errors amid the procedure (e.g., to perceive two records as propagations when they are definitely not). The reason we have picked GP method is its known ability to find exact reactions for a given issue, without looking the complete chase space down game plans, that is frequently broad, and when there is one or more objective as master. Actually, we and different specialists have effectively connected GP to a few data administration related issues, for example, positioning capacity revelation record order substance based picture recovery, and substance target publicizing, to refer to a couple, outflanking as a rule other best in class machine learning systems.

The fundamental commitment of this paper is a GP-based technique to manage data deduplication that:

. Beats a current best in class machine learning based system found in the written work;

. Gives arrangements less computationally escalated, since it recommends deduplication works that utilization the accessible confirmation all the more effectively:

. Liberates the client from the weight of picking how to join likeness capacities and archive traits. This recognizes our methodology from every current framework, since they require customer gave settings:

Liberates the client from the weight of picking the imitation distinguishing proof limit esteem, since it has the capacity consequently select the deduplication limits that better fit this deduplication parameter

The rest of this paper sorts out as takes after: area II clarifies related work of copies recognition in distributed storage server. Segment III introduces background approach for accessing duplicate files in hybrid cloud processing in real time cloud applications. Section IV introduces Genetic Programming approach for duplicate detection in cloud data storage. Section V introduces efficient experimental evaluation in finger print generation of detection of duplicates in cloud data storage. Section VI introduces overall conclusion of our proposed approach with duplicate detection in cloud.

2. RELATED WORK

Advancements in cloud computing lead the way to affordable resources and thus gradually leading to degradation of those resources especially storage space due to the presence of duplicate copies of the same file. Since this has become a potential problem the research community tried to address it. Yuan [2] being the first to address it and inspired from system level deduplication implementations suggested the idea of a deduplication system in cloud framework. But his implementation is limited to plain file formats but not ciphered ones'. Bellare [3] demonstrated the idea of data obfuscation which involves using an outside domain key server for file tag generation to protect its confidentiality which is not feasible for our implementation.

Stanek [4] suggested ciphering schemes that differentiates between important cipherable content and unimportant ignorable content. A two layered ciphering scheme involving a deduplication was implemented on the important content which we were inspired to adapt to our current approach. Although the distinguishable factors between important and unimportant are not applicable to our current context. We tend to apply it to all users files.

Li [5] suggested Key sustenance topic at file block level stages by transferring the keys between multiple key servers after securing file contents. Cohesive Encryption suggested in [6] implements data isolation during deduplication process. Bellare [3] used this concept in a secured encryption scheme and applied its usage on cloud stored data thus being the first to actually implement this scheme in clouds. Bitcasa, a real time Commercial CSH uses this cohesive encryption scheme. Halevi [7] suggested ownership proofs that acts as key identity for implementing deduplication which is a two stage communication process involving communication cost. It requires the client to prove to the CSH that they own the file and then really requiring to upload a file. Merkle Hash solutions suggested by Halevi attain deduplication by handling parameter leakages. Many other solutions do exist but they all end up ignoring data privacy (that means implemented on plain files). Some other cloud pairing solutions were also proposed by Bugiel [8] and Zhang [9] which can be implemented to hybrid cloud platforms and whose work comes close to that of ours.

To settle these irregularities it is essential to layout a deduplication limit that joins the information open in the data chronicles with a particular deciding objective to perceive whether two or three record entries insinuates the same bona fide component. In the area of bibliographic references, for case, this issue was comprehensively analyzed by Lawrence et al.

As more procedures for separating dissimilar bits of proof get to be accessible, numerous works have proposed new particular ways to deal with consolidate and utilize them. The thought of consolidating proof to recognize imitations has pushed the information administration research group to search for techniques that could profit by space particular data found in the real information and in addition for strategies taking into account general similitude measurements that could be adjusted to particular spaces. As a case of a system that adventures general similitude capacities adjusted to a particular area, we can specify. There are some inventors to propose a planning estimation that, given a record in a report (or vault), searches for other record in a reference record that helps in matching the previous record as showed by a resemblance limit. The organized reference records are chosen taking into account a client characterized least closeness edge. In this way, more than one applicant record may be returned.

The suggestions mostly identified with the work are those that helps in applying machine learning approaches for choosing similitude works in record level that join field-level closeness points of confinement, including the best errand of weights. These suggestions utilize a little information for get prepared. This readiness set of data is relied upon to have similar credits to those of the test data set, which makes conceivable to the machine learning frameworks to entirety up their answers for concealed information. The great results generally got with these strategies have exactly exhibited that those suppositions are legitimate. We propose a GP-based approach to manage push ahead results made by the Fellegi and Sunter's strategy. Particularly, we use GP to change the weight vectors conveyed by that genuine technique, to deliver a prevalent confirmation mix than the essential summation used by it. Our trial results with certified data sets show changes of 7 percent in precision concerning the standard Fellegi and Sunter's method. In connection with our past results, this paper shows a wider and upgraded GP-based philosophy for deduplication, which has the capacity consequently produce powerful deduplication capacities notwithstanding when a suitable likeness capacity for every record trait is not gave ahead of time. What's more, it likewise adjusts the recommended capacities to changes on the reproduction recognizable proof limit qualities used to order a couple of records as reproductions or not.

3. HYBRID ARCHITECTURE FOR SECURE DEDUPLICATION

At an abnormal state, our setting of hobby is a venture system, comprising of a gathering of associated customers (for instance, representatives of an organization) who will use the S-CSP and store data with deduplication technique. In this setting, deduplication can be as often as possible utilized as a part of these settings for information reinforcement and catastrophe recuperation applications while incredibly diminishing storage room. Such frameworks are across the board what's more, are frequently more suitable to client document reinforcement and synchronization applications than wealthier stockpiling reflections. There are three substances portrayed in our structure, that is, customers, private cloud and S-CSP out in the open cloud as showed up in Fig. 2. The S-CSP performs deduplication by checking if the substance of two records are the same and stores one and just of them. We will simply consider the record level deduplication for ease. In another word, we elude an information duplicate to be an entire record and document level deduplication which takes out the stockpiling of any excess records. Really, square level deduplication can be effortlessly derived from record level deduplication, which is comparative to as of now displayed record away framework. In particular, to transfer a document, a client first performs the record level copy check. On the off chance that the record is a copy, then every one of its squares must be copies also; something else, the client further performs the piece level copy check and recognizes the one of a kind squares to be transferred. Every information duplicate (i.e., a record or a square) is associated with a token for the duplicate check.

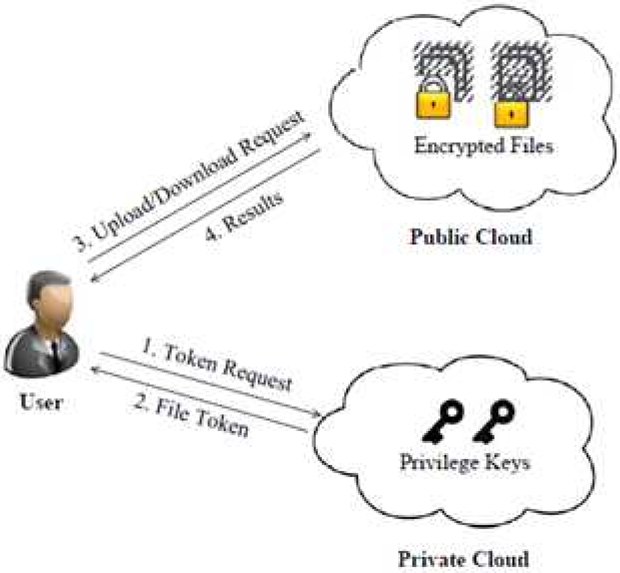

Figure 2: Hybrid Cloud Approach For Detecting Secure Data Duplication In cloud.

• S-CSP. This is an element that gives an information stockpiling administration out in the cloud. The S-CSP gives the data outsourcing organization and stores data for sake of the clients. To decrease the capacity cost, the S-CSP gets rid of the limit of dreary data through deduplication and keeps just one of kind information. In this paper, we expect that S-CSP is always online and has extensive limit farthest point and computation power.

• Data Users. A customer is a component that needs to outsource data stockpiling to the S-CSP and access the data later. In a limit system supporting deduplication, the client just transfers one of a kind information however does not transfer any copy information to spare the transfer data transfer capacity, which may be claimed by the same client or diverse clients. In the endorsed deduplication system, each customer is issued a course of action of advantages in the setup of the structure. Each record is guaranteed with the joined encryption key and advantage keys to make sense of it the endorsed deduplication with differential advantages.

Private Cloud. Separated and the standard deduplication assistant masterminding in appropriated figuring, this is another substance showed for engaging client's guaranteed use of cloud association. Specifically, since the selecting assets at information client/proprietor side are confined and people when in doubt cloud is not completely trusted basically, private cloud has the point of confinement give information client/proprietor with an execution area and establishment filling in as an interface amidst customer and individuals as a rule cloud. The private keys for the advantages are directed by the private cloud, who answers the document token requesting from the customers. The interface offered by the private cloud grants customer to submit records and inquiries to be securely secured and enrolled independently. Notice this is a novel auxiliary designing for data deduplication in dispersed registering, which includes a twin fogs (i.e., the overall public cloud and the private cloud).

We address the issue of security protecting deduplication in dispersed processing and propose another deduplication structure supporting for

- Differential Authorization. Each approved client has the capacity get his/her individual token of his document to perform copy check in view of his benefits. Under this suspicion, any client can't produce a token for copy look at of his advantages or without the aide from the private cloud server.

- Authorized Duplicate Check. Affirmed customer has the limit use his/her individual private keys to create question for certain record and the advantages he/she had with the help of private cloud, while the open cloud performs duplicate check clearly and tells the customer if there is any cloud

4. DEDUPLICATION USING GP

At the point when utilizing GP (or even some other transformative strategy) to take care of an issue, there are some essential prerequisites that must be completed, and rely upon the data structure used to identify with the game plan. For our circumstance, we have picked a tree-based GP representation for the deduplication limit, since it is a trademark representation for this sort of capacity.

In our methodology, every bit of proof (or essentially "proof") E is a couple <attribute; equivalence function> that addresses the use of a specific resemblance limit over the estimations of a specific quality found in the data being penniless down. For example, if we have to deduplicate a database table with four characteristics (e.g., forename, surname, area, and postal code) using a specific similarity limit (e.g., the Jaro limit [10]), we would have the going with summary of evidence: E1<name; Jaro>, E2<surname; Jaro>, E3<address; Jaro>, and E4<postal code; Jaro>. For this case, a to a great degree clear limit would be an immediate blend, for instance, FsðE1; E2; E3; E4þ = E1 þ E2 þ E3 þ E4 and a more eccentric one would be Fc=E1; E2; E3; E4þ =E1 (E3E2=E4)

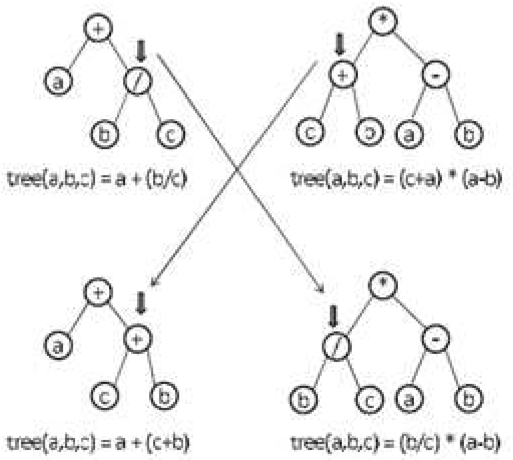

Figure 3: Replica Detection Using Tree Based Cross Over In Gp.

To model such limits as a GP tree, each verification is identified with by a leaf in the tree. Each leaf (the closeness between two properties) creates an institutionalized authentic number quality (some place around 0.0 and 1.0). A leaf can in like manner be a sporadic number some place around 1.0 and 9.0, which is picked at the moment that each tree is created. Such leaves (unpredictable numbers) are used to allow the transformative strategy to find the most attractive weights for each verification, when critical. The inside center points address operations that are joined with the gets out. In our model, they are clear experimental limits (e.g.;E1;E2 ; =; exp) that control the leaf values.

The tree info is an arrangement of proof examples, separated from the information being taken care of, and its yield is a genuine number worth. This quality is looked at against a copy distinguishing proof limit esteem as takes after: on the off chance that it is over the limit, the records are considered reproductions, otherwise, the records are viewed as unmistakable passages. It is essential to notice that this order empowers further examination, particularly with respect to the transitive properties of the copies.

$$P = \frac{NumberofCorrentlyIdentifiedDuplicatedPairs}{NumberOfIdentifiedDuplicatedPairs}$$

$$R = \frac{NumberofCorrentlyIdentifiedDuplicatedPairs}{NumberOfTrueDuplicatedPairs}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

This can improve the viability of collection counts, since it gives not only an estimation of the equivalence between the records being taken care of, furthermore a judgment of whether they are duplicates or not. we have used the F1 metric as our wellbeing limit. The F1 metric pleasingly unites the conventional accuracy (P) what's more, review (R) measurements normally utilized for assessing data recovery frameworks, as characterized beneath:

Here, this metric is used to express, as a lone quality, how well a specific individual performs in the endeavor of perceiving proliferations. In diagram, our GP-based approach tries to support these wellbeing qualities via looking for individuals that can settle on all the more right decisions with less blunders.

The time unpredictability of the preparation stage, in view of our displaying, is O(Ng Ni) Te, where Ng is the quantity of development eras, Ni is the quantity of people in the populace pool, and Te is the wellness assessment many-sided queue.

5. SYSTEM IMPLEMENTATION

We a model of the proposed endorsed deduplication system, in which we three components as segregated is used data customers to do the record exchange process. A Private Server venture is used to show the private cloud which manages the private keys and handles the record token estimation. A Storage Server venture is used to the S-CSP which stores and deduplicates archives. We execute cryptographic operations of hashing also, encryption with the OpenSSL library .We in like manner execute the correspondence between the substances in perspective of HTTP, using GNU Libmicrohttpd and libcurl. Along these lines, customers can issue HTTP Post requesting to the servers. Our utilization of the Client gives the going with limit calls to reinforce token period and deduplication along the record exchange process.

- **FileUploadReq(FileID, File, Token)**
  This module involves generating an upload request between user and CSH.
- **FileEncrypt(File)**
  This module involves mapping plain file contents for deduplication and encrypting the file itself later.
- **File_Unique_Identifier(File)**
  This module involves generating file identifier based on their cloud type and uploaded file id.
- **Token_Request(Fid, UserID)**
  This module involves wrapping up the above identifier with user id.
- **Duplicate_Check_Request(Token)**
  This module involves initiating the following genetic algorithm driven validation process.
- **Share_Token_Request(Fid, {Priv.})**
  This module involves sharing the genetic algorithm validations between different clouds.

The following parameters are usually defined in any genetic algorithm (GA) process and when adapting these to our hybrid cloud deduplication, their definitions are as follows:

Table 1: Parameters defined in Genetic Algorithm

Representation	Collection of a specific user (U) cloud files (F)
Recombination	Collection of all user files(C) along with the (U) files in hybrid cloud
Mutation	Tracking similarities between C and F and separating unique contents into a pool
Parent selection	Collection of all the original files sorted by their uploader and upload date , the root being the parent
Crossover	After the mutation aspect, the not so unique ones are pooled here.

Survivor selection	Identifying the candidate files for deduplication process with an exact chunk similarity match of 100% between C and F
Fitness	C+F=P. The fitness parameter implements a file chunk hash generation and comparisons to determine survivor aspect of genetic algorithm

These phases all converge to be a part of a deduplication process implemented in Share Token Request.

6. EXPERIMENTAL EVALUATION

We lead tested assessment on our model. Our assessment concentrates on looking at the overhead impelled by approval steps, including document token era and offer token era, against the united encryption also, document transfer steps. We assess the overhead by changing distinctive components, counting 1) Size of the File 2) Total number of files that are stored 3) Ratio of Deduplication 4) Size of the Privilege Set. We survey the model with a certifiable workload considering VM pictures. The design can be implemented on Standard machine with the following configurations Intel(R) D CPU 2.8 GHz, 2GB RAM. 32-bit Windows 7 Operating System

Apart from successfully identifying duplicate files, the performance difference is measured in share token generation aspect of prior approaches and our genetic algorithm based approaches. The following table illustrates the duration aspects.

Table 2: Table Illustrating Duration Aspects

FileSize(KB)	Plain Share Token Duration(sec)	GA-Based Share Token Duration(sec)
93	1.03	0.69
184	1.23	0.95
289	1.39	1.14
395	2.01	1.45
490	2.45	1.88

The performance difference illustrated in the following highlights the efficiency of our genetic algorithm based solution compared to prior approaches.

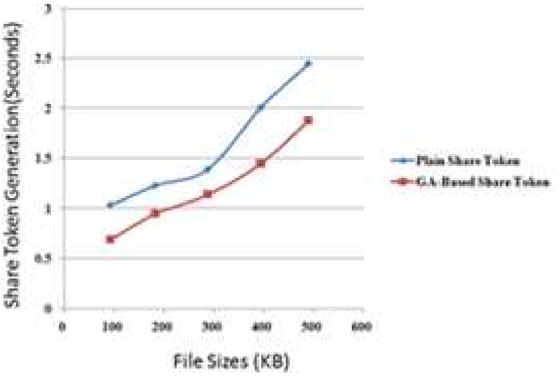

Figure 4: Comparison Specification Regarding Files Storage With Duplication In Cloud.

A basic perspective in regards to the viability of a few deduplication methodologies is to set the values that are limited and arrange a couple of records as imitations or not with admiration to the consequences of the deduplication capacity. In this last arrangement of tests, our goal was to ponder the capacity of our GP-based way to deal with the deduplication capacities for changes in the copy distinguishing proof limit, going for finding whether it is conceivable to utilize a already altered (or recommended) esteem for this parameter.

A conceivable clarification for this conduct can be drawn by the accompanying truths:

7. The copy distinguishing proof limit is dependably complete worth.

2. The estimations of the proof occasions (the effect that happened after applying a capacity of string to a pair of characteristic) fluctuate from 0 to 1.

3. On account of flawless match for each and every quality, the overall proof example qualities will be equivalent to the quantity of characteristics utilized as confirmation furthermore; its aggregate augmentation would be equivalent to 1.

4. There is no need that all quality sets must achieve a flawless similarity in request to be viewed as an imitation. Our GP-based methodology helps to join particular proof to expand the wellness capacity results, and one main consideration that

may affect the outcomes is the reproduction distinguishing proof limit esteem. Accordingly, if the picked limit worth is out of the scope of a conceivable compelling proof mix, this hopeful arrangement (deduplication capacity) will come up short in the assignment of recognizing imitations.

8. CONCLUSION AND FUTURE WORK

In this paper we explained the architecture of a hybrid cloud platform and the usual troubles they face with respect to storage space and the solutions to reclaim that space. The primary reasons for the loss is explained as several users uploading the same content knowingly or unknowingly in their accounts. Methods to reclaim this space by eliminating the storage redundancy are required and are technically termed as deduplication. Hence we propose a genetic algorithm based share token module that is implemented in a file chunk validation procedure deployed at CSH in the hybrid clouds. We demonstrated the results on a couple of public and private cloud partitions hosted within the hybrid cloud. A real time analysis performed by uploading files of varied sizes(depicted in the above table) and estimated round trip time consists of upload duration, hash generation, share token request generation.Our optimization procedure enhances this share token request using a genetic algorithm based approach and above results highlights the efficiency of our proposed approach. We furthermore showed the results of examinations on the duplicate recognizing verification utmost, using bonafide and made sets of data. This work shows that this GP-based procedure is prepared for changing the proposed deduplication abilities as far as possible qualities used to portray several records as impersonation or not. Also, the results recommend that the usage of a settled breaking point regard, as almost 1 as could be normal in light of the current situation, encourages the transformative effort moreover prompts better courses of action. As future work, we hope to lead additional exploration remembering the deciding objective to increase the extent of usage of our GPbased implementation for managing the deduplication. Many newly evolving digital signature schemes can replace our genetic algorithm based solutions but they are left for future work.

REFRENCES:

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEM VOL:PP NO:99 YEAR 2014.

[2] J. Yuan and S. Yu. “Secure and constant cost public cloud storage auditing with deduplication”,*IACR Cryptography* ePrint Archive,2013:149, 2013.

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

[4] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. “A secure data deduplication scheme for cloud storage”, *In Technical Report*, 2013.

[5] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. “Secure deduplication with efficient *and reliable convergent key management”,In IEEE Transactions on Parallel and Distributed Systems*, 2013.

[6] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. “Reclaiming space from duplicate files in a serverless distributed file system”, *In ICDCS*, pages 617–624, 2002.

[7] 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.

[8] 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.

[9] 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.

[10] Moise's G. de Carvalho, Alberto H.F. Laender, Marcos Andre' Gonc¸alves, and Altigran S. da Silva, "A Genetic Programming Approach to Record Deduplication", *IEEE Transactions on knowledge and data engineering, VOL. 24, NO. 3, March 2012.*