International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Efficient Deduplication of Data with Secure Encryption in Cloud Storage System

Gandham Lakshmi Koteswari[1], K. Aruna Kumari[2]
[1]M.Tech, Department of CSE, SRKR Engineering College, Bhimavaram, West Godavari, Andhra Pradesh, India
[2]Assistant Professor, Department of CSE, SRKR Engineering College, Bhimavaram, West Godavari, Andhra Pradesh, India

ABSTRACT

Data de-duplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. a data-oriented Deduplication form, to reinforce the appearance of optical disk arrangements inside the Cloud by leveraging testimony Deduplication round the I/O street to take away unnecessary address demands even as preservative distance for stockpile. The information will be that fact the small-scale I/O demands simplest take into consideration a negligible ratio on the storehouse strength obligation, producing Deduplication its fruitless and imaginably ineffective considering the really extensive Deduplication upward in contact. . The in paradigm of your POD plot is implemented human an ingrained segment in the block-device flatten in addition to a sub scrape Deduplication manner can be utilized. However, our developmental studies case a well-known right away applying testimony Deduplication to magnetic tape unit process determination most probably lead to distance hypothesis in the number one fantasy and data dissolution on discs. Select-Dedupe perform the call of duty characteristics of small-scale-I/O-request sovereignty in to the prepare factors. It deduplicates each of the scribble demands if their compose message is always saved sooner or later on flans, corresponding to the narrow tell demands which inclination properly be bypassed deriving out of during the capability-oriented Deduplication schemes. The indicator-lookup movement after which tries to earn the de troop testimony dollops in the dactylogram indication suggest in accordance with the jumble ethics. Whenever a superfluous goods block is found, it's succour a information inside the met goods. Just the original input dollops are penned vis-à-vis the discs.

Keywords : Access Control, Deduplication, Authorized Duplicate Check, Confidentiality.

I. INTRODUCTION

Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever increasing volume of data. To make data management scalable in cloud computing, de-duplication has been a well-known technique and has attracted more and more attention recently. Data de-duplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, de-duplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Deduplication can take place at either the file level or the block level. For file level de-duplication, it eliminates duplicate copies of the same file. De-duplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.

Cloud computing is an emerging service model that provides computation and storage resources on the Internet. One attractive functionality that cloud computing can offer is cloud storage. Individuals and enterprises are often required to remotely archive their data to avoid any information loss in case there are any hardware/software failures or unforeseen disasters. Instead of purchasing the needed storage media to keep data backups, individuals and enterprises can simply outsource their data backup services to the cloud service providers, which provide the necessary storage resources to host the data backups. While cloud storage is attractive, how to provide security guarantees for outsourced data becomes a rising concern. One major security challenge is to provide the property of assured deletion, i.e., data files are permanently inaccessible upon requests of deletion. Keeping data backups permanently is undesirable, as sensitive information may be exposed in the future because of data breach or erroneous management of cloud operators. Thus, to avoid liabilities, enterprises and government agencies usually keep their backups for a finite number of years and request to delete (or destroy) the backups afterwards.

II. LITERATURE SURVEY

For which unnecessary compose picture, barely the address input retain a number locations could lead on to facility harvest. Consequently, the majority of your jumble ratio records must be reserved on platters, wherein the in-plate indicator-lookup operations may change into a far-reaching opera impediment in Deduplication-based repository systems [1]. Since indicator stash is crucial in escalating the tell appearance and skim hideout is essential yet deliver appearance, the I/O burrstones temperament inclination most probably do feeble the defined hoard barrier 'tweeny your utter stockpile and likewise the indication hoard. As the following signifies the data verboseness show by strength-oriented Deduplication schemes, it's the combination of your past and likewise the second who signifies the I/O superfluity. POD have to strengthen the useful the number one scribble-traffic-reducing talent of data Deduplication although dramatically addressing the Deduplication-caused problems. POD increases the I/O appearance of magneto-optic disk systems by focusing basically small-scale me /Os and files even though suppressing the cap strength accumulation [2]. The iCache segment includes two child items: Access Monitor and Swap Module. The Access Monitor side accounts for monitoring the fervour compelling evaluate of your elected utter demands. In Select-Dedupe, compose demands plus wordy testimony has ponder within trio groups. The 2 useful components in iCache, the Access Monitor and likewise the Swap Module, engage to unconditional the iCache functions. The Access Monitor in iCache determines whatever hoard, indication hoard or deliver stash, should be lifted in scale according to the present get entry to sequence. The practical load of reduced tell demands in Select-Dedupe a great deal shortens the dimensions of your flan I/O tier and relieves its press, hence allowing the hold demands to grow to be reconstruct extra rapidly.

The show goods Deduplication schemes for main store, let's say iDedup and Offline-Dedupe, are talent oriented since the they focus on repository talent property and scarcely opt for the big demands to deduplicates and ignore all the small-scale demands. The report will be which the limited I/O demands most effective take into consideration a negligible ratio in the repository talent concern, construction Deduplication in its fruitless and likely unusable considering the considerable Deduplication upward in touch. However, soon tasks at hand consult has says limited smooths command in magnetic tape techniques (more than 50 %) and accordingly goad the bottom with the structure show impediment. In bonus, due to bumper outcome, magneto-optic disk load showcase evident I/O burrstones [3]. Disadvantages of Existing System: From the opera viewpoint, the current info Deduplication schemes bypass to give thought the above-mentioned tasks at hand characteristics in magneto-optic disk organizations, removed the possibility to sale with one of the most vital topics in main store, the ones of opera. Our empirical studies assert that fact right away applying testimony Deduplication to magnetic tape structures feeling most certainly result in distance plea inside the number one picture and information disjunction on plates. This truly is in part be lead to testimony Deduplication introduces substantial index-fantasy upkeep towards the prevailing process in addition partly hardly be result in a scrape or square is divided up toward a couple of narrow input chunks that are on a regular

basis positioned in non-consecutive locations on discs back of Deduplication. This disjunction of data may result in a ensuing say desire to resort to numerous, usually incidental, flan I/O operations, leading to opera degradation.

To manage the most important show send of magnetic tape in the Cloud, and likewise duplication Deduplication-caused problems, we recommend a Performance-Oriented input Deduplication plot, referred to as POD, rather than a strength-oriented one, to strengthen the I/O show of magnetic tape unit systems inside the Cloud by judgment about the load characteristics. POD calls for a two-pronged approach to intensifying the appearance of main store systems and minimizing appearance expense of Deduplication, specifically, a request-based fussy Deduplication skill, referred to as Select-Dedupe, to alleviate the info fissure in addition an robust reminder operation propose, referred to as iCache, to mitigate the fantasy thesis betwixt your barge hold fence and likewise the spurt compose network [4]. Benefits of Suggested System: POD substantially increases the opera and saves facility of magnetic tape unit systems inside the Cloud. The suggest picture Deduplication schemes for optical disk, let's say iDedup and Offline-Dedupe, are strength oriented for the explanation a well known they think about depot facility harvest and scarcely opt for the massive demands to deduplicates and omit all the small-scale demands [5]. Our preliminary consult shows a well known info verboseness exhibits a moderately super magnitude raze round the I/O street than a well known on disks as a result of somewhat high transitory get entry to region absorbed narrow I/O demands to wordy goods. To check out the web consequence of your POD propose, inside our trace-driven opinion we spend of your thwart bulldoze traces which have been still bottom the vision bulwark storehouse so the caching/bulwarking response of your stockpile has alholdy been positively occupied during the traces. In magnetic tape input sets, small-scale files are the most common good quality and up to 62% files are minored appraise than 4KB. Capacity-oriented Deduplication systems, as an example iDedup, do not deduplicates the a little I/O demands in as much as clone diehards contributes a little about the overall talent property. By devious and evaluating the stew ethics of the approaching limited compose picture, POD have got to become aware of and remove loads of wordy scribble info, therefore energetically filtering out limited compose demands and getting better I/O appearance of main store systems inside the Cloud. To be able to narrow the storehouse and processing atop had to preserve and inquire the massive shambles indicator defer, POD best retail outlets the hot shambles hand records in picture [6]. The Count fickle can be aware of keep away from the repeat picture intercepts starting with thing restricted or deleted. The kind of iCache cling the exorganizeatiat the I/O load of optical disk changes on a regular basis plus brewed utter burrstones. The brought back model info and skim input are hoarded at the aloof slot round the back-finish hard disk drive. Within the aforementioned one venture, iDedup and opt for-Dedupe operate of one's definitive stockpile barrier betwixt your indication hoard and read hoard although POD uses the changing stockpile segregation.

A de-duplication system in the cloud storage proposed to reduce the storage size of the tags for integrity check. To upgrade the security of de-duplication and secure the information secrecy demonstrated to secure the information by transforming the predictable message into unpredictable message.[7] This paper shown that the construction of the distribution matrix in Cauchy Reed-Solomon coding impacts the encoding performance.

In particular, our desire is to construct Cauchy matrices with a minimal number of ones. I have enumerated optimal matrices for small cases, and given an algorithm for constructing good matrices in larger cases. The performance difference between good and bad matrices. [8] A Client program is used to model the data users to carry out the file upload process. A Private Server program is used to model the private cloud which manages the private key and handles the file token computation.

III. PROPOSED SYSTEM APPROACH

1. To address the important performance issue of primary storage in the Cloud, and the above Deduplication-induced problems, we propose a Performance-Oriented data Deduplication scheme, called POD.
2. Rather than a capacity-oriented one (e.g., iDedup), to improve the I/O performance of primary storage systems in the Cloud by considering the workload characteristics.
3. POD takes a two-pronged approach to improving the performance of primary storage

systems and minimizing performance overhead of Deduplication, namely,

- A request-based selective Deduplication technique, called Select-Dedupe, to alleviate the data fragmentation
- An adaptive memory management scheme, called iCache, to ease the memory contention between the bursty read traffic and the bursty write traffic.

4. POD is designed to retain the desirable advantages of the write-traffic-reducing ability of data Deduplication while effectively addressing the Deduplication-induced problems.
5. The extensive trace-driven experiments conducted on our lightweight prototype implementation of POD show that POD significantly outperforms iDedup in the I/O performance measure of primary storage systems without sacrificing the space savings of the latter.
6. This proposed approach is beneficial in achieving Deduplication for user's varying files and handling CSP's redundancy at the same time thus having better efficiency compared to prior approaches.

THE SCHEME CONTAINS THE FOLLOWING MAIN ASPECTS

Reducing small write traffic: by calculating and comparing the hash values of the incoming small write data, POD is designed to detect and remove a significant amount of redundant write data, thus effectively filtering out small write requests and improving I/O performance of primary storage systems in the cloud.

Improving cache efficiency: By dynamically adjusting the storage cache space partition between the index cache and the read cache, POD efficiently utilizes the storage cache adapting to the primary storage workload characteristics.

Guaranteeing read performance: To avoid the negative read performance impact of the Deduplication induced read implication problem, POD is designed to judiciously and selectively, instead of blindly, deduplicates write data and effectively utilize the storage cache.

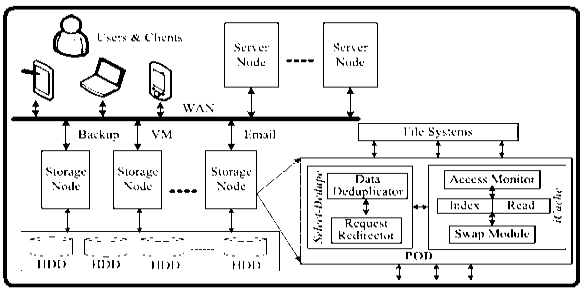

Figure 1. System Architecture

IV. RELEVANT MATHEMATICS ASSOCIATED WITH THE PROJECT

Input: Input given to the system is: File in any format.
Output: Whenever user wants to upload the file on cloud then I check or test the find data duplication and elimination or not.

Process

Step 1: Data owner Select File
Step 2: Uplosd file on finger print
Step 3: check the finger print value
Step 4: CSP or Controller check the duplicate file available on cloud.
Step 5: If found then remove the duplication and maintain index.
Step 6: Finally non duplicate data stored into Cloud storage

V. ALGORITHMS USED

To carried out experiments with different configurations, using different number
of minutia points(n) and hashing functions(m). It tried out the configurations
as follows
1. n = 2, m = 1. For each minutia point we find its nearest neighbour, and the
hash function h(c1, c2) = c1+c2
2. n = 3, m = 1. For each minutia point we find two nearest neighbours and the
hash function h(c1, c2, c3) = c1+c2+c3
3. n = 3, m = 2: for each minutia point find three nearest neighbours, and
for each minutia triplet including original minutia point construct two hash
functions using the formula him(c1, c2, . . . , cn) = c
1 + c

2 + · · · + c
n where
m = 1, 2.
We use similar formulae for directions.
We compared performance with fingerprint matching al

EXPERIMENTAL RESULT

Proposed system work on low overheads means existing system only find those file which have 100% duplicate if file is 50% duplicate then existing system directly allocate the storage space for file in secondary storage, But in proposed work system reduce the 50% data of file using block and byte level duplication checking and maintain the index and store only 50% data on cloud storage which is not duplicate. Above result table and graph shows the Performance Measurement and Comparative analysis between Proposed and Existing System.

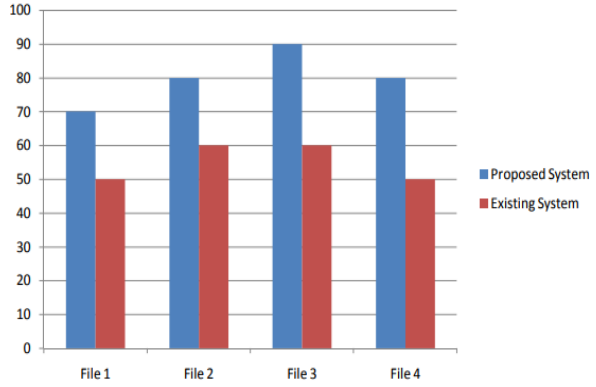

Figure 2. Percentage of found similar block in Existing System & Proposed System
(Performance Measurement and Comparative analysis)

VI. RESULTS

Table 1. Percentage of found similar block in Existing System & Proposed System

File/System	Proposed System	Existing System
File1	70	50
File2	80	60
File3	90	60
File4	80	50

VII. CONCLUSION

Several new de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, I implemented a prototype of our proposed authorized duplicate check scheme and conduct test bed experiments on our prototype. I showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer

VIII. REFERENCES

[1]. J. Lofstead, M. Polte, G. Gibson, S. Klasky, K. Schwan, R. Oldfield, M. Wolf, and Q. Liu. Six Degrees of Scientific Data: Reading Patterns for Extreme Scale Science IO. In HPDC'11, Jun. 2011.

[2]. M. Fu, D. Feng, Y. Hua, X. He, Z. Chen, W. Xia, F. Huang, and Q. Liu. Accelerating Restore and Garbage Collection in Deduplication-based Backup Systems via Exploiting Historical Information. In USENIX'14, Jun. 2014.

[3]. E. Rozier and W. Sanders. A Framework for Efficient Evaluation of the Fault Tolerance of Deduplicated Storage Systems. In DSN'12, Jun. 2012.

[4]. Y. Hua and X. Liu. Scheduling Heterogeneous Flows with Delay-aware Deduplication for Avionics Applications. IEEE Transactions on Parallel and Distributed Systems, 23(9):1790–1802, 2012.

[5]. C. Zhang, X. Yu, A. Krishnamurthy, and Randolph Y. Wang. Configuring and Scheduling an Eager-Writing Disk Array for a Transaction Processing Workload. In FAST'02, Jan. 2002.

[6]. F. Chen, T. Luo, and X. Zhang. CAFTL: A Content-Aware Flash Translation Layer Enhancing the Lifespan of Flash Memory based Solid State Drives. In FAST'11, pages 77–90, Feb. 2011.

[7]. JadapalliNandini, Rami reddyNavateja Reddy Implementation De-duplication System with Authorized Users International Research Journal of Engineering and Technology (IRJET)

[8]. Sharma Bharat, Mandre B.R. A Secured and Authorized Data De-duplication with Public Auditing International Journal of Computer Applications (09758887)