International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 2, 2018 ISSN: 2348-6600
Reference ID: IJCS-332 PAGE NO: 2203-2208

# An Efficient Way to Avoid Redundancy Record in Big Data

V.Geetha[#1], Dr.P.Mohammed Shareef[*2], R.Sathiyavani[*3]

[#]*Head of the Department of Computer Science, Sengamala Thayaar Educational Trust Women's College, Mannargudi, Thiruvarur, Tamil Nadu*

[*]*Department of Computer Science, Sengamala Thayaar Educational Trust Women's College, Mannargudi, Thiruvarur, Tamil Nadu*

[3]pvaruna1995@gmail.com

*Abstract*— **Distributed computing offers another method for benefit arrangement by re-organizing different assets over the Internet. The most essential and prevalent cloud benefit is information stockpiling. Keeping in mind the end goal to save the security of information holders, information are frequently put away in cloud in an encoded frame. Nonetheless, encoded information present new difficulties for cloud information deduplication, which winds up essential for huge information stockpiling and preparing in cloud. In this thesis, our propose build a hybrid deduplication mechanism named Stream Locality Aware Deduplication to avoid de-duplicate encrypted data stored in cloud based on ownership challenge and proxy re-encryption. This is, unsurprising, due to the digitization of our society. In this research, achieved high security, low error rate and low computational complexity of the scheme for potential practical deployment.**

***Index Terms*— Deduplication, hybrid mechanism, Stream Locality Aware Deduplication**

## I. INTRODUCTION

A distributed computing offers another method for Information Technology benefits by revising different assets (e.g., capacity, figuring) and giving them to clients in light of their requests. Distributed computing gives a major asset pool by connecting system assets together. It has attractive properties, for example, adaptability, versatility, adaptation to internal failure, and pay-per-utilize. Consequently, it has turned into a promising administration stage.

Deduplication has demonstrated to accomplish high cost reserve funds, e.g., decreasing up to 80-85 percent stockpiling requirements for reinforcement applications and up to 68 percent in standard record frameworks. Clearly, the funds, which can be passed back specifically or by implication to cloud clients, are critical to the financial matters of cloud business. The most effective method to oversee encoded information stockpiling with deduplication in a proficient way is a commonsense issue. Nonetheless, current mechanical deduplication arrangements can't deal with encoded information. Existing answers for deduplication experience the ill effects of animal power assaults. They can't adaptably bolster information get to control and disavowal in the meantime. Most existing arrangements can't guarantee unwavering quality, security and protection with sound execution. Practically speaking, it is difficult to enable information holders to oversee deduplication because of various reasons. To start with, information holders may not be constantly on the web or accessible for such an administration, which could cause stockpiling delay. Second, deduplication could turn out to be excessively convoluted as far as correspondences and calculations to include information holders into deduplication process. Third, it might barge in the protection of information holders during the time spent finding copied information. Forward, an information holder may have no clue how to issue information get to rights or deduplication keys to a client in a few circumstances when it doesn't know other information holders because of information super-circulation. Hence, CSP can't coordinate with information holders on information stockpiling deduplication as a rule.

There are two different ways to distinguish duplications in a distributed storage framework. One is contrasting squares or documents bit with bit, and the other is looking at obstructs by hash esteems. The benefit of contrasting squares or documents bit with bit is that it is precise, however it is additionally tedious. The benefit of looking at squares or records by hash esteem is that it is quick, yet there is a possibility of unintentional crash. The possibility of unintentional crash relies upon the hash calculation. Be that as it may, the shot is extremely little. In this manner, the

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 2, 2018 ISSN: 2348-6600
Reference ID: IJCS-332 PAGE NO: 2203-2208

blend of MDS and SHA1 will incredibly decrease the likelihood. The current methodologies for recognizing duplications dependably take a shot at two distinct levels. One is the document level; the other is the piece level. On the lump level, information streams are separated into pieces, each piece will be hashed, and all these hash esteems will be kept in the list. The benefit of this approach is that it is advantageous for an appropriated record framework to store lumps, yet the downside is the expanding amount of hash esteems. It implies that hash esteems will possess more RAM utilization and increment the query time. On the document level, the hash capacity will be executed for each record, and all hash esteems will be kept in the file. The benefit of this approach is that it diminishes the amount of hash esteems fundamentally. The downside is that, when the hash work needs to manage an expansive record, it will end up being somewhat moderate.

In general, this research makes the accompanying commitments: 1) construct a crossover deduplication component named SLADE (Stream Locality Aware Deduplication) that breakers a proficient inline deduplication process and a post-handling deduplication process together for essential stockpiling frameworks shared by numerous applications. 2) The execution of "apparition reserve" construct calculations with respect to the square layer deduplication frameworks for the essential stockpiling and uncover their non-insignificant memory overhead in unique mark store administration. 3) An area estimation based reserve substitution calculation which essentially enhances the unique mark reserving effectiveness in essential stockpiling deduplication frameworks. The estimation technique can abuse stream-based territory for better store space allotment.

## II. Related Work

*A.Blind and Anonymous Identity-Based Encryption and Authorized Private Searches on Public Key Encrypted Data*

The research can be viewed as an augmentation of their hilter kilter conspire with the likelihood to absently look through the scrambled database. In the symmetric case, in which the review log server knows all the data required for decoding the database, the issue of performing unmindful inquiries is secured by. The issue of unmindful seeking on open key encoded information is more troublesome. Blueprint of our answer. In the unbalanced accessible encryption plot depends on character based encryption (IBE). The catchphrases themselves are utilized to scramble the database, i.e., they are the character strings of the IBE conspire. The namelessness property of Boneh-Franklin IBE plot guarantees that a figure content does not release the character string used to produce the encryption. The security server holds the ace mystery key that is utilized to infer the mystery keys relating to the watchwords that are required for seeking. A comparable method for accessible encryption was formalized as open key encryption with catchphrase look (PEKS). In PEKS, the determined keys are alluded to as pursuit trapdoors, which can be given to outsiders to concede them look rights.

*B.Searchable Encryption Revisited: Deduplication, Consistency Properties, Relation to Anonymous IBE, and Extensions*

In this research, they additionally investigate one of the variations of this objective, specifically open key encryption with watchword seek (PEKS) as presented. They start by talking about consistency-related issues and results, at that point consider the association with unknown character based encryption (IBE) every year examine a few augmentations. Any cryptographic crude must meet two conditions. One is obviously a security condition. The other, which they will here call a consistency condition, guarantees that the crude completely its capacity. For instance, for open key encryption, the security condition is protection. (This could be formalized from various perspectives, eg. IND-CPA or IND-CCA.) The consistency condition is that decoding inverts encryption, implying that if M is encoded under open key pk to bring about figure content C, at that point unscrambling C under the mystery key comparing to pk brings about M being returned. It is normal to inquire as to whether BDOP-PEKS meets some consistency condition that is weaker than theirs yet at the same time sufficient practically speaking. To answer this, they give some new definitions.

*C.DFA-Based Functional Encryption: Adaptive Deduplication of Dual System Encryption*

In the double framework technique, semi-practical parts for ciphertexts and keys generally imitate the structure of the typical ciphertexts and keys separately. In any case, these are created utilizing some mystery components with the goal that their dispersion is factually escaped the foe. Since there is a solitary gathering component (hash $H\sigma$) for every image $\sigma$, there will be a relating scalar in the semi-practical bit for every image amid reenactment. On the off chance that images are rehashed, at that point so are these scalars. Be that as it may, giving out an excessive number of duplicates of these qualities will uncover them data hypothetically to the assailant which crushes the double framework evidence. This

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 2, 2018 ISSN: 2348-6600
Reference ID: IJCS-332 PAGE NO: 2203-2208

holds for the two strings and automata. Utilizing the double framework procedure, they have gotten a DFA-based useful encryption conspire that has versatile security under static presumptions in composite request pairings. The cost of accomplishing this is an expansion in the sizes of the ciphertext and keys alongside limited usefulness. It is fascinating to get versatile security without confining the quantity of events of images in either the strings or changes of automata. Another characteristic inquiry is whether specific security can be accomplished utilizing static suppositions yet without forcing the confinements on the framework.

*D.Deterministic and Efficiently Searchable Encryption and avoid deduplication data*

In the symmetric setting, deterministic encryption is caught by figures including square figures. They present the thought of proficiently accessible encryption plans. These are plans allowing quick (i.e. logarithmic time) look. Encryption might be randomized, yet there is a deterministic capacity of the plaintext that can likewise be figured from the figure content and fills in as a "tag," allowing the standard thing (quick) examination based pursuit. Deterministic encryption plans are an uncommon case and the idea of security continues as before. (Our PRIV definition does not really expect encryption to be deterministic.) The advantage of the speculation is to allow plans with more adaptable protection to look time exchange offs. In particular, they break down a plan from the database writing that they call "Hash-and-Encrypt." It encodes the plaintext with a randomized plan yet additionally incorporates into the figure message a deterministic, impact safe hash of the plaintext.

*E.Conjunctive, Subset, and Range Queries on Encrypted Data*

They build open key frameworks that help examination inquiries on scrambled information and in addition more broad questions, for example, subset inquiries. These frameworks bolster self-assertive conjunctive inquiries without spilling data on individual conjuncts. What's more, they exhibit a general structure for developing and breaking down open key frameworks supporting inquiries on scrambled information. Questions on scrambled information are most effortless to clarify with a case. Consider a charge card installment entryway that watches a surge of scrambled exchanges, say encoded under Visa's open key. The portal needs to hail all exchanges fulfilling a specific predicate P. Say, all exchanges whose esteem is finished. Putting away Visa's mystery key on the door is an awful thought for both security and protection concerns. Rather, Visa wishes to give the entryway a token TKP that empowers the portal to distinguish exchanges fulfilling P without getting the hang of whatever else about these exchanges. Obviously, producing the token TKP will require Visa's mystery key.

## III. METHODOLOGY

*A.Secure Data Deduplication*

Movement past a specific breaking point mines the execution and the nature of administration saw by the clients. Amid top circumstances in swarmed metropolitan situations, clients as of now encounter long latencies, low throughput, and system blackouts because of the clog and over-burden at level. Shockingly, this pattern can just worsen in future due to the anticipated versatile information blast. The issue concerns essentially organize administrators since they need to exchange off consumer loyalty with business productivity, given the pattern toward about level rate plans of action. As it were, the exponential increment in rush hour gridlock streaming in their does not create enough extra incomes to be designated into additionally updates. The previously mentioned conditions cultivated the enthusiasm for elective strategies to moderate the weight on the cell arrange. Hunt over encoded information is a basically critical empowering procedure in Cloud processing, where encryption before outsourcing is a crucial answer for securing client information protection in the untrusted cloud server condition.

Numerous protected inquiry plans have been concentrating on the single-supporter situation, where the outsourced dataset or the safe, accessible record of the dataset are scrambled and overseen by a solitary proprietor, normally in light of symmetric cryptography. In this theory, they center around an alternate yet all the more difficult situation where the outsourced dataset can be contributed from different proprietors and are accessible by various clients, i.e. Multi-client multi-donor case. Motivated by certain quality based encryption, they display the principal evident characteristic based watchword look conspire with productive client denial that empowers versatile fine-grained (i.e. Record level) look approval. Our plan enables numerous proprietors to scramble and outsource their information to the cloud server freely. Clients can produce their own particular pursuit capacities without depending on a constantly online confided in the specialist. Fine-grained seek approval is additionally executed by the proprietor authorized access arrangement on the record of each document. Further, by fusing intermediary re-encryption and apathetic re-encryption procedures, they can designate overwhelming framework refresh workload

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 2, 2018 ISSN: 2348-6600
Reference ID: IJCS-332 PAGE NO: 2203-2208

amid client repudiation to the ingenious semi-confided in cloud server. They formalize the security definition and demonstrate the proposed technique conspire specifically secure against picked catchphrase assault.

In propose a crossover deduplication component that wires an effective inline deduplication process and a post-preparing deduplication process together for essential stockpiling frameworks shared by various applications. Our assess the execution of "phantom reserve" construct calculations in light of the square layer deduplication frameworks for essential stockpiling and uncover their non-insignificant memory overhead in unique mark store administration. They propose an area estimation based reserve substitution calculation which altogether enhances the unique mark reserving effectiveness in the essential stockpiling deduplication frameworks. The estimation strategy can abuse stream-based area for better reserve space portion. Our assess SLADE utilizing follows produced from genuine applications. The outcome demonstrates that SLADE beats the cutting edge inline and post-handling deduplication systems.

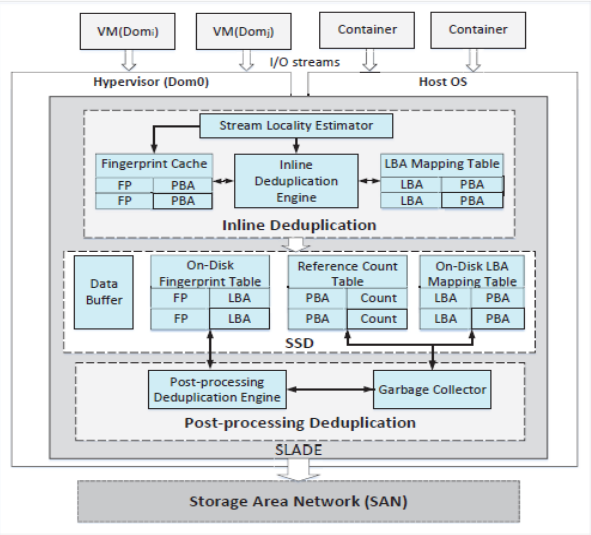


Fig.1 Stream storage architecture

$f_{ij}$ is the $j^{th}$ from stream i.
$S_i$ is the segment representing stream i in the Priority Segment Tree.
$C_i$ is the cache space for stream i.
Procedure Threshold Process()

$T_u$ = Threshold for updating Priority Segment Tree;
Upon Event Interval End && ($S_i$) > $T_u$
Update Priority Segment Tree();
M = Total Cache Size;
Threshold Count ($f_{ij}$);
If len ($S_i$) > Cache Threshold then
If $C_{ij}$ < M then
Insert $f_{ij}$ into $C_i$;
Else
$f_{pq}$;$C_p$ = Threshold To Evict();
Evict $f_{pq}$ from $C_p$;
Insert $f_{ij}$ into $C_i$;
End
End

To decrease workloads because of copy records, proposed Index Name Servers to oversee not just the document stockpiling, information deduplication, upgraded hub choice, and server stack adjusting, yet in addition document pressure, lump coordinating, ongoing criticism control, IP data, and occupied level file checking. To oversee and advance stockpiling hubs in light of a customer side transmission status by the proposed INS, all hubs must evoke ideal execution and offer reasonable assets to customers. Along these lines, not exclusively can the execution of a capacity framework be enhanced, however the documents can likewise be sensibly dispersed, diminishing the workload of the capacity hubs. Notwithstanding, this work can't deduplicate scrambled information. A cross-breed information deduplication system that gives a functional arrangement fractional semantic security. This arrangement underpins deduplication on plaintext and ciphertext. Be that as it may, this component can't bolster scrambled information deduplication exceptionally well. It works in light of the supposition that CSP knows the encryption key of information.

## IV. Result and Discussion

In proposed SLADE gives a safe way to deal with secure and deduplicate the information put away in cloud by covering plaintext from both cloud specialist co-op and approved gathering. The security of the proposed Scheme is guaranteed by pre hypothesis, symmetric key encryption and elliptic curve cryptography hypothesis. The clients don't have to process information encryption if another client has transferred the information as of now. In this way, our plan can spare a great deal of calculation load and correspondence cost for cloud clients. Moreover, the information possession challenge in the proposed plot is exceptionally lightweight,

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 2, 2018 ISSN: 2348-6600
Reference ID: IJCS-332 PAGE NO: 2203-2208

which does not include much weight to cloud clients. This infers our plan is exceptionally effective and extremely reasonable for deduplication keep an eye on huge information. Extra presumptions include: information holders sincerely give the encoded hash codes of information for possession check. The information proprietor has the most noteworthy need. An information holder ought to give a substantial declaration keeping in mind the end goal to ask for an exceptional treatment. Clients, cloud service provider and authorized party convey through a protected channel with each other. CSP can verify its clients during the time spent cloud information stockpiling. It also expect that the customer approach for data storing, sharing and deduplication is given to CSP in the midst of customer enlistment. In proposed method provide high security, low error rate and low computational complexity compared with existing technique.

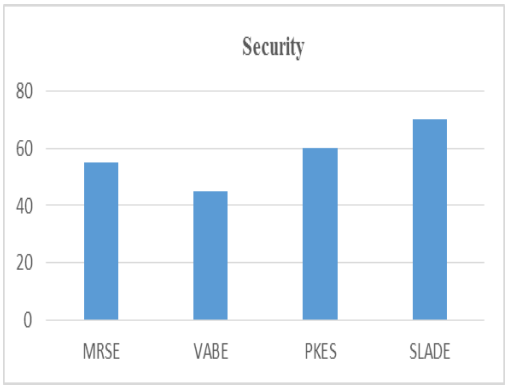

Fig. 2 Analysis of security

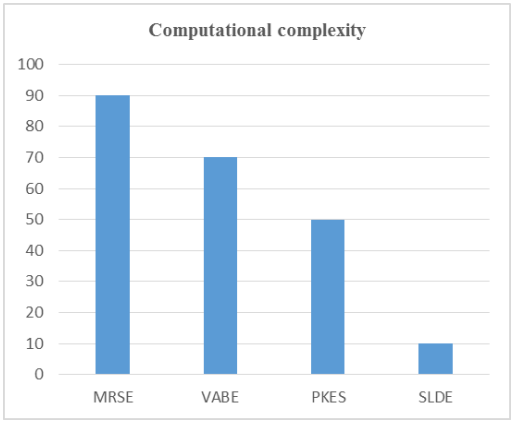

Fig. 3 Computational complexity

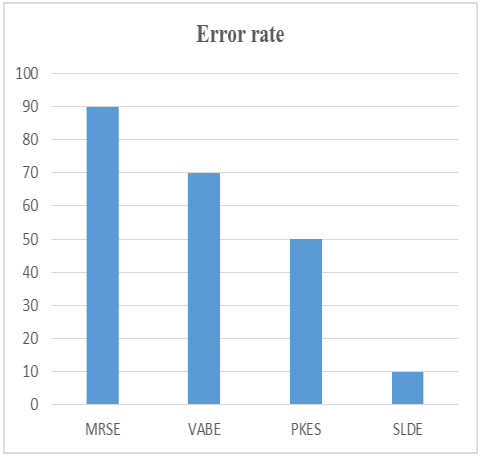

Fig. 4 Error rate analysis

## V. Conclusions

Overseeing scrambled information with deduplication is critical and huge by and by for accomplishing a fruitful distributed storage benefit, especially for huge information stockpiling. In this research, proposed SLADE a handy plan to deal with the encoded huge information in cloud with deduplication in view of proprietorship test and PRE. It built a hybrid deduplication system named SLADE. SLADE achieved high inline cache efficiency and reduced the deduplication workload in the post-processing deduplication phase. SLADE improved the inline deduplication ratio by up to 39.70% compared with iDedup in our experiments. Meanwhile, SLADE reduced up to 45.08% disk capacity requirement compared with the post-processing deduplication mechanism in our evaluation. Broad execution examination and test demonstrated that our plan is secure and proficient under the depicted security display and exceptionally reasonable for huge information deduplication. The after effects of our PC reproductions additionally demonstrated the practicability of our plan.

## References

[1] 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 Proc. IEEE Int. Conf. Distrib. Comput. Syst., 2002, pp. 617–624.

[2] G. Wallace, et al., “Characteristics of backup workloads in production systems,” in Proc. USENIX Conf. File Storage Technol., 2012, pp. 1–16.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 2, 2018 ISSN: 2348-6600
Reference ID: IJCS-332 PAGE NO: 2203-2208

[3] G. Ateniese, K. Fu, M. Green, and S. Hohenberger, "Improved proxy re-encryption schemes with applications to secure distributed storage," ACM Trans. Inform. Syst. Secur., vol. 9, no. 1, pp. 1–30.
[4] D. T. Meyer and W. J Bolosky, "A study of practical deduplication," ACM Trans. Storage, vol. 7, no. 4, pp. 1–20.
[5] J. Pettitt, "Hash of plaintext as key?" (2016). [Online]. Available: http://cypherpunks.venona.com/date/1996/02/msg02013.html.
[6] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. Cryptology—EUROCRYPT, 2013, pp. 296–312.
[7] D. Perttula, B. Warner, and Z. Wilcox-O'Hearn, "Attacks on convergent encryption." (2016).
[8] C. Y. Liu, X. J. Liu, and L. Wan, "Policy-based deduplication in secure cloud storage," in Proc. Trustworthy Comput. Serv., 2013, pp. 250–262.
[9] P. Puzio, R. Molva, M. Onen, and S. Loureiro, "ClouDedup: Secure deduplication with encrypted data for cloud storage," in Proc. IEEE Int. Cof. Cloud Comput. Technol. Sci., 2013, pp. 363–370.
[10] Z. Sun, J. Shen, and J. M. Yong, "DeDu: Building a deduplication storage system over cloud computing," in Proc. IEEE Int. Conf. Comput. Supported Cooperative Work Des., 2011, pp. 348–355.
[11] Z. C. Wen, J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li, "A verifiable data deduplication scheme in cloud computing," in Proc. Int. Conf. Intell. Netw. Collaborative Syst., 2014, pp. 85–90.
[12] J. Li, Y. K. Li, X. F. Chen, P. P. C. Lee, and W. J. Lou, "A hybrid cloud approach for secure authorized deduplication," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 5, pp. 1206–1216.
[13] P. Meye, P. Raipin, F. Tronel, and E. Anceaume, "A secure two phase data deduplication scheme," in Proc. HPCC/CSS/ICESS, 2014, pp. 802–809.
[14] J. Paulo and J. Pereira, "A survey and classification of storage deduplication systems," ACM Comput. Surveys, vol. 47, no. 1, pp. 1–30.
[15] Y.-K. Li, M. Xu, C.-H. Ng, and P. P. C. Lee, "Efficient hybrid inline and out-of-line deduplication for backup storage," ACM Trans. Storage, vol. 11, no. 1, pp. 2:1-2:21.
[16] M. Fu, et al., "Accelerating restore and garbage collection in deduplication-based backup systems via exploiting historical information," in Proc. USENIX Annu. Tech. Conf., 2014, pp. 181–192.
[17] M. Kaczmarczyk, M. Barczynski, W. Kilian, and C. Dubnicki, "Reducing impact of data fragmentation caused by in-line deduplication," in Proc. 5th Annu. Int. Syst. Storage Conf., 2012, pp. 15:1–15:12.
[18] M. Lillibridge, K. Eshghi, and D. Bhagwat, "Improving restore speed for backup systems that use inline chunk-based deduplication," in Proc.USENIX Conf. File Storage Technol., 2013, pp. 183–198.
[19] L. J. Gao, "Game theoretic analysis on acceptance of a cloud data access control scheme based on reputation," M.S. thesis, Xidian University, State Key Lab of ISN, School of Telecommunications Engineering, Xi'an, China, 2015.