UDD based Procedure for Record Deduplication over Digital Storage Systems

Shaik Anjumun Jabeen, Y Prasanth, Gudapati Syam Prasad

Abstract: ***Digital libraries, E-commerce brokers and similar vast information-oriented systems rely on consistent data to offer high-quality services. But presence of duplicates, quasi replicas, or near-duplicate entries (Dirty Data) in their repositories asperses their storage resources directly and delivery issues indirectly. Significant investments in this field from interested parties prompted the need for best methods for removing replicas from data repositories. Prior approaches involved using SVM classifiers, approaches to handle these dirty data. New distributed deduplication systems with higher reliability in which the data chunks are distributed across multiple cloud servers. The security requirements of data confidentiality and tag consistency are also achieved by introducing a deterministic secret sharing scheme in distributed storage systems, instead of using convergent encryption as in previous deduplication systems. So propose to use Unsupervised Duplicate Detection (UDD) Mechanism a query-dependent record matching method that requires no pre trained data set. UDD uses two cooperating classifiers that is, a weighted component similarity summing (WCSS) classifier and an SVM classifier that iteratively identifies duplicates in the query results from data sources. Achieves the same efficiency in terms of Deduplication results but significantly at a better performance rate (time) compared to GP systems. A practical implementation of the proposed approach validates the claim***

Index Terms: Cloud Computing Environment, Genetic Programming, Active Learning, Deduplication and Cloud Security.

I. INTRODUCTION

The measure of advanced information on the planet is developing violently, as prove to a limited extent by the huge increment in the evaluated measure of information produced in 2010 and 2011 from 1.2 zettabytes to 1.8 zettabytes, individually [1, 2], and the anticipated measure of information to be created in 2020 is 44 zettabytes [3, 4]. As a aftereffect of this "information storm," how to oversee capacity cost-effectively has turned out to be a standout amongst the most testing and vital undertakings in mass stockpiling frameworks in the enormous information time. The outstanding burden contemplates led by Microsoft [5, 6] furthermore, EMC [7, 8] propose that about half and 85% of the information in their creation essential and auxiliary capacity frameworks, separately, are repetitive. Concurring to an ongoing IDC examine [9], practically 80% of partnerships reviewed showed that they were investigating information deduplication advances in their capacity frameworks to lessen repetitive information and accordingly increment stockpiling productivity and lessen capacity costs.

Information deduplication is an effective information decrease approach that not just lessens extra room [5– 7, 10–13] by taking out copy information yet in addition limits the transmission of excess information in low-transfer speed organize conditions [8, 14, 15]. When all is said in done, an average lump level information deduplication framework parts the information stream (e.g., reinforcement records, database previews, virtual machine pictures, and so forth.) into numerous information "lumps" that are each remarkably recognized and copy identified by a cryptographically secure hash signature (e.g., SHA- 1), additionally called a unique finger impression [11, 14]. These lumps can be fixed in size [11], similar to record squares, or variable-sized units dictated by the substance itself [14]. Deduplication frameworks at that point expel copy information lumps and store or exchange just a single duplicate of them to accomplish the objective of sparing extra room or system transfer speed. To increase the identification of deduplication in secure cloud for different resources, traditionally more number of approaches was used to support deduplication in outsourced data in cloud. For deduplication, use some machine learning like genetic programming approach and others effectively. Interestingly with the techniques already said, the strategy proposed in this paper takes after a semi-regulated approach in view of dynamic figuring out how to help diminishing the cost of naming cases while developing grouping models with hereditary programming. The propose Adaptive and Secure Active Learning & Extensive GP (ASAL & EGP), starting now and into the foreseeable future alluded as Adaptive And Secure Active Learning & Extensive GP (ASAL & EGP), works with an advisory group of people that chooses which cases ought to be sent to the client to mark. It additionally executes a fortification learning approach, which helps assessing the certainty of board of trustee's individuals in their groupings. Adaptive And Secure Active Learning & Extensive GP (ASAL & EGP) was custom fitted to tackle a testing database issue: information deduplication. The principle objective of information deduplication is to recognize distinctive records in a database alluding to a similar genuine element.

Revised Manuscript Received on April 25, 2019.

Shaik Anjumun Jabeen, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P, India.

Dr. Y Prasanth, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P, India.

Dr. Gudapati Syam Prasad, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P, India.

This issue was picked on the grounds that, given the measure of the vaults required (in the request of a great many records), the way toward marking information can be to a great degree costly or even unconventional. Moreover, sometimes it is hard notwithstanding for people to choose if two records are copies or not without enough data. The proposed ASAL & EGP plot is portrayed by following properties:

1) ASAL & EGP can safely "deduplicate" the verification labels by totaling the labels of a similar document from various proprietors, and consequently maintain capacity low representations free to quantity of proprietors of the record;

2) Correspondence data representation consists our ASAL & EGP plot is made steady because of our novel outline of multi-nominal confirmation labels and secure information conglomeration;

3) Computational procedure on cloud clients is likewise steady in light of the fact that most computational errands can be safely offloaded to the cloud server;

4) ASAL & EGP bolsters open inspecting, i.e., the information uprightness evaluating file attributes rather than recommended files;

5) ASAL & EGP permits bunch evaluating, i.e., numerous reviewing solicitations can be safely totaled, which significantly diminishes the inspecting cost for concurrent solicitations;

6) ASAL & EGP information respectability reviewing and privacy with respect to replication attributes with existing approaches.

II. LITERATURE REVIEW

Considering just reliability analyzing for data outsourced to cloud servers, different POR designs [5], [6], [7], [8] and PDP designs [9], [10], [11], [12] have been proposed. Among those ref.[5] has the best execution which achieves open assessing at a steady correspondence cost. Like other POR or PDP designs, customers in ref.[5] still need to perform O(k) increment and development tasks over the concealed field, where k is the amount of checking data squares. Cluster assessing for different requesting circumstances isn't maintained in ref.[5]. For secure limit deduplication, Halevi et al. [13] exhibited the foremost POW plot in light of the Merkle hash tree. Pietro et al. [14] enhanced ref.[13] and proposed a protected POW plot which lessens the computational cost to an unfaltering number of pseudo-random work activities. Regardless, these POW designs don't consider data dependability assessing. To achieve the two data respectability investigating and limit deduplication, one insignificant course of action is to explicitly join a current POR/PDP plan with a POW plot. This inconsequential course of action, in any case, will achieve an O(W) storing overhead for each record, where W is the amount of proprietors of this report. This is in light of the fact that the data proprietors, lacking shared place stock in, require to autonomously store their own particular affirmation marks in cloud for report reliability looking at. Since these names are made for assessing a comparative archive, securing O(W) such copies addresses a kind of duplication which disavows the objective of POW for saving storing cost. For viable affirmation of limit with deduplication (POSD), Zheng et al. [15] proposed an arrangement going for giving both open data reliability checking on and secure limit deduplication. In ref.[15] the correspondence cost and computational cost on the customer side are straight to the amount of parts in each datum block and furthermore the amount of checking frustrates in the midst of the trustworthiness evaluating process. With an extending masses of convenient customers, who get the chance to cloud through flexible applications (e.g., iAWS, iCloud, et cetera.) and have obliged computational resources and exchange speed (e.g., mobile phones with compelled data plan), such a correspondence and computational capriciousness could address an impediment to getting to the circulated stockpiling advantage. In a perfect world, computational cost and correspondence cost on the customer side may be reliable. Moreover, ref.[15] has been exhibited not secure [16]. Specifically, by setting the parts in riddle keys to some unprecedented regards, a data proprietor who outsources data to the cloud server can use the server as a malware appointment arrange. Thusly, in any case it requires another response for help beneficial and secure data genuineness examining with limit deduplication for disseminated capacity.

In [12], de Carvalho et al. given the primary procedure in light of GP for the data deduplication process. In their work, GP is utilized to discover record-level resemblance includes that union single-characteristic similarity highlights, attempting to build the acknowledgment of duplicate data and, in the meantime, avoiding botches. This strategy was last broad in [13]. While in [12] single-characteristic similarity highlights are picked from the earlier by the purchaser, in [12] the single-quality highlights are likewise picked by the GP. The procedure proposed here symbolizes its kin similarly as in [12] and [13], however while these methods adhere to an observed methodology, assuming that all brands of wellbeing and wellness circumstances are known previously, the strategy gave here knows just the brands of some in the preparation set.

III. BACKGROUND APPROACH

Discussed in above section, there are some machine learning techniques, approaches. At the point if you use GP (or even some other major strategy) to take care of a problem, there are some essential requirements to be implemented, and rely on require details file parameters to identify redundant process. For our situation, we have selected a tree-based GP reflection for the redundancy restrict, since it is a signature reflection for this kind of capacity.

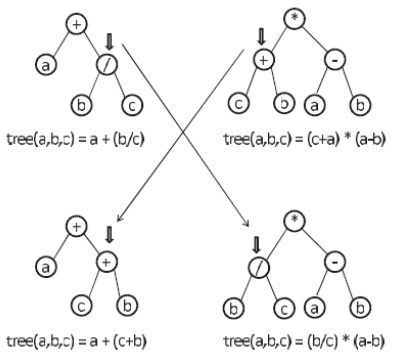

Fig .3. GP approach procedure to handle deduplication.

In our technique, every bit of proof (or basically "proof") E is several <attribute; equivalence attributes with functions> that details at particular similar restricts to estimates of a particular top quality seen in the information being broke down. If we have to perform reduplicated information source desk with four features (e.g., first name, last name, place, and mailing code) using a particular likeness restrict (e.g., the Jaro [2]), we present review of evidence: E1<name; First name >, E2<last name; Jaro>, E3<communication address; Jaro>, and E4<postal code; Jaro>. For this situation, a to an excellent level obvious restrict would be an immediate combination, for example, File with similarity = E1; E2; E3; E4þ = E1 þ E2 þ E3 þ E4 and a more unusual one would be redundant files =E1; E2; E3; E4þ =E1 (E3E2=E4).

This implemented model consists limits as a GP tree structure, each attribute 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 1-9; these users are used to continue transformative strategy to identify most attractive weights for each verification, when critical. The inside center points address operations that are joined with the gets out. In our genetic approach , 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.

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

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

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

We can improve the ability of iteration counts, it gives estimation equivalent between the original and duplicate records being taken care of, furthermore a judgment of whether they are duplicates or not. We used F1 and F0 metric representations wellbeing limit. F1 metric formations pleasingly unites the conventional accuracy (P) what's more, review (R) measurements normally utilized for assessing data recovery frameworks, as characterized beneath: the metrics used to exposed based on low && relative data, how well a specific individual performs in the endeavor of perceiving proliferations. In above diagram GP machine learning techniques tries to support these wellbeing qualities with respect to independent attributes on all the more right decisions with fewer blunders.

IV. PROPOSED SYSTEM IMPLEMENTATION & DESIGN PROCEDURE

A. *Construction of UDD & WCSS*

Let G and G1 be the multiplicative groups of same with different order specification q, the bilinear map is a connection for all $g, h \in G$ & $x, y \xleftarrow{R} Z_q^*, e(g^x, h^y) = e(g,h)^{xy}$. For linear operations based on computational with compute efficiency e of $e(g,g) \neq 1$.

Let H(.) be the one-way hash operate, G be a multiplicative cyclic number of primary purchase q, g be the generator of G and $u \xleftarrow{R} G.$. F0 is the erasure written computer file to be contracted and is divided into n prevents, each of which has s elements $\{n_{ij}\}, 1 \le k \le m, 0 \le l \le f - 1.$ $s_{x(a)}$ is considered to be a sequential polynomial presentations with different parameters $x = (x_0, x_1, ..., x_{f-1})$.

B. *Basic preliminaries of ASAL & EGP*

keyGen: TA is dynamic and general numbers $x \xleftarrow{R} Z_g^*$ & defines key representations for all data variations $\{g^{a^j}\}_{j=0}^{s+1}$, as is the master key specification on TA. The TA will define after posting its community important factors. Given a protection attribute λ , the information proprietor defines a deciding upon key-pair $((server, client) \xleftarrow{R} Verf())$, then public, master and secret keys are:

$$PK = \{q, k, v, spk, u, \{g^{a^j}\}_{j=0}^{s+1}\}$$

Setup-Environment: To delegate a information which consists file F, the information proprietor retrieves F0 by implementing erasure rule, where F0 includes n information detects at each and every basic elements:

$$\{n_{ij}\}, 1 \le k \le m, 0 \le l \le m-1$$

Then random parameter sequences for owner authentication tags are shown in below.

$$\sigma_i = (v^{H(pronoun||i)} . \prod_{j=0}^{s-1} p^{m_{ij} x^{j+2}})^{\in} = (v^{H(pronoun||i)} . p^{f_{\beta_i}^{tw}(x)}) \in$$

Basic challenge: Ensure the reliability of F0, a user get file from cloud server & confirms the verification key from server representations. If the trademark is not legitimate, the user refined & halts; otherwise, the user gets back computer file number n. Then the consumer arbitrarily selects a k-elements subset K of [1; n] and two unique numbers.

Prove: Based on challenging variables CM={K,r}, then cloud server generates $\{\{p_i = r^i \bmod q\}, i \in K\}$ and polynomial data representation $\overset{u}{A} = \{0, 0, \sum_{i \in K} p_i m_i 0, ..., \sum_{i \in K} p_i m_{i,s-1}\}$ & then finally generates security information $\sigma = \prod_{i \in K} \sigma_i^{p_i}$ and information $\Pr f = \{\sigma, \psi, y\}$ to each user relations.

Verify: After receiving $\Pr f$, user computes $v = \sum_{i \in K} p_i H(name \| i)$ then security integration for different data formations as follows:

$$e(\eta, j).e(\psi, u, j^{-r}) = e(\sigma, p).e(j^{-y}, p)$$

It defines security aspect for each user data with different parameters for file auditing.

Deduplication: In data replication, client send upload his data with different features F0 store in cloud environment, it consists same features F1 related to cloud server, then check cloud server identify whole F0 and F1 with mediate document elements D relates [1,n] for different clients. Based on above considerations formation of correctness of our proposed construction as follows:

$$e(\eta, j).e(\psi, u.j^{-r})$$
$$= e(v.p)^{\in(\sum_{i \in J} p_i H(pronoun \| i)).p^{\in f(x)}, p).e(j^{-y}, p)}$$
$$= e(n', j).\psi'$$

C. *Deduplication Calculation*

Compared to existing data redundancy approaches that can be accessed to determining replications, WCSS consists 3 basic phases: (1) produces all the related images for evaluation of different files, extensively or through preventing methods. (2) Determines a likeness measurement for two chromosomes based on relative attributes. In this stage, each feature is communicated with defined range measurement relates to measurement of different parameters (i.e., mathematical, string & attributes). This area concentrates on stage 3, semi-trained approach with inherited development & effective and re-arrangement studying discovers a panel (set) of multi functional parameters that categorizes a set is redundant or not. Observe that, although we concentrate on the information redundancy with respect to files, the process suggested procedure to any other program sector where marking illustrations are important and do the procedure. Procedure of the active and adaptive GP shown in algorithm 1.

Input: Data records.
Output: Similarity Measure.
Step 1. Let us consider D be the records of deduplicate.
Step 2. Generate set of different variables P=p1, p2, ..., pn from source data base.
Step 3. Compute sim for each variable p€P.
Step 4. R ← position P according to sim(p) principles ;
Step 5. Let T be the different k-pairs in R, evaluate gen_0 based on weight w_0
Step 6. for i = 0 to a predetermined variety of years do.
Step 7. For each pair p€P, compute each label info for L_b € {Tr, Fa, D}
Step 8. Update each variable based on sim for each label.
Step 9. Compute similarity index based on similarity parameters with pairs.
Step 10. Similarity Index.

Algorithm 1. AGP procedure to evaluate data deduplication.

In data deduplication, independent symbolizes a likeness functional parameters with respect to data accessibility. The plants that signify the likeness features are produced basic and statistical providers. The devices are unique integers from 1 to 9, and the of likeness between each feature present in the data source. Alg. 1 explains the procedure suggested to learn multi-attribute features that categorize sets of records as copies or not

D. *Implementation Procedure*

We numerically examine our UDD & EGP conspire what's more, contrast it and. For effortlessness, remaining of this document, we utilize EXP1 and MUL to signify the multifaceted nature of one increase operations with exponentials group operations independently. Blending is a bilinear blending operation.

1) Computational: In ASAL & EGP plot, the correspondence implemented procedure is represented by the testing message CM = fK; rg and the confirmation data Prf = f ; ; yg. The CM comprises of a key set K with respect mean square k-ids and an regularization static number. As we talked about in Sections, the client can haphazardly challenge k = 600 information pieces to guarantee no less than 99:999% blunder recognition likelihood. In the event that an mistake discovery likelihood a represented parameter, the span of K can be considered as steady and the many-sided quality of output message for CM with respect to time complexity is O(1). The evidence data is formed a polynomial y and two gathering components also, . In this manner, the aggregate correspondence unpredictability of evaluating process in our ASAL & EGP plot is additionally O (1). In deduplication User send encoded information and then service provider process that information and then access with different presentation. Discussed in traditional sections, the cloud server just needs difficult 300 squares or on the other hand 460 squares to accomplish 95% or 99% certainty whether the client really possesses the entire information record.

Along these lines, the extent of D can be limited and the aggregate correspondence multifaceted nature of the Deduplication procedure in our plan is O(1).

2) Public Auditing with Security: we talk about communication and implementation cost with respect to different representations spared by our clump evaluating outline for various solicitations situations. Assume a TPA is enlisted by T information proprietors to enable them to review the trustworthiness of service provider intermittently. In the event that the third party auditor forms these L evaluating demands step by step, it needs different operations for verification to different relational attributes for implementation, MUL, EXP and other computational components, L arbitrary numbers and L polynomials for correspondence. With our bunch inspecting outline, the cloud server can total L i into one gathering component and utilize one arbitrary operations with less than selected file. Along these lines, contrasted and handling demands successively, our cluster inspecting plan via third party activity from server to spare around half correspondence cost. Extensive view of implementation cost, our group reviewing configuration empowers the third party auditor to decrease pair operations from source, which is substantially more costly contrasted with different multiplicative operations. In this way, around 25% computational assignments are put something aside for the TPA with our clump inspecting plan. Accept c% records are from same information properties; our cluster reviewing configuration can spare extra with respect to communication cost.

V. Experimental Evaluation

To implement our approach, enhanced active learning and secure approach define effective and adaptability like Amazon EC2cloud computing environment. To develop this application, use Java 1.8 and Net beans 8.0 with advanced versions. Using these software's, we develop interface for communication to cloud and access different services to share features and other parameters to cloud. To verify our proposed approach with existing approach GP in terms of communication cost with respect to encryption and decryption and other parameters present in outsourced data. Then communication cost at user side based on number of blocks stored data in cloud server. Table 1 shows communication cost for different approaches.

Data Blocks	UDD	GP
100	601	805
200	624	814
300	632	821
400	637	832
500	639	836
600	641	845
700	643	874
800	647	889
900	649	898
1000	652	912

Table 1. Communication values for different data block specifications.

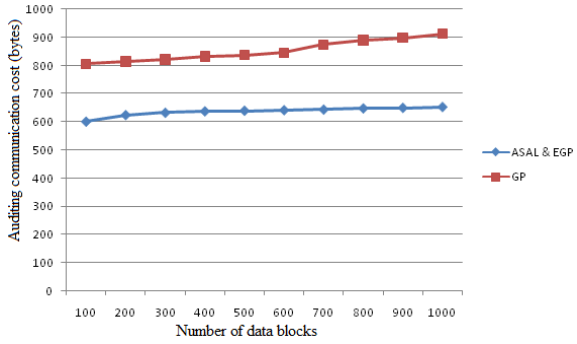

Fig 4. Auditing communication cost with respect to data blocks on user sides.

Figure 4 shows the computational price of different users to perform stability review process different steps with variety of data stops in the audit details information enhances. With respect to the interaction cost, it also continues to be continuous with different details information increases.

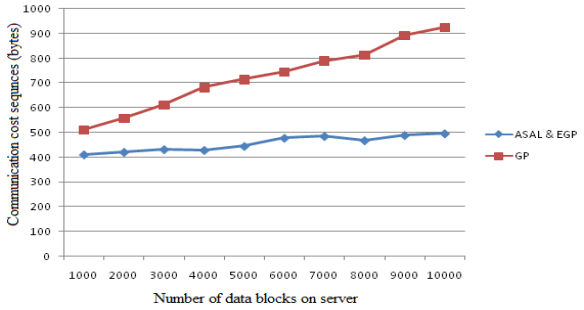

Fig 5. Auditing cost sequences based on file data blocks on server.

As we showed over, the quantity of information hinders in each document does not impact the execution of our plan, we set the quantity of information pieces to 5000 in each inspecting errand. Contrasted and playing out these examining assignments, Fig.5 demonstrates third party auditor around 58% reviewing time with cluster evaluating. From the point of view of correspondence cost, Fig.5 demonstrates our group evaluating spares around 60% transfer speed for the TPA.

Data Blocks	ASAL & EGP	GP
1000	411	511
2000	421	558
3000	432	612
4000	429	682
5000	445	714
6000	478	745
7000	485	789
8000	468	814
9000	489	892
10000	496	924

Table 2. Average data values with respect to communication cost in server side.

Considering the normal cost per undertaking, which is processed by separating all out reviewing time and aggregate evaluating transmission capacity with respect to quantity of assignments separately, Cons cutely, above table shows communication cost representations with respect to time efficiency.

Originally, a pre-processing produces Set of similarity pairs like P present in data source, being it is replicated with respect to different information whether it is relevant or not for all other representations shown in relevant scenarios with trimming similarity sets.

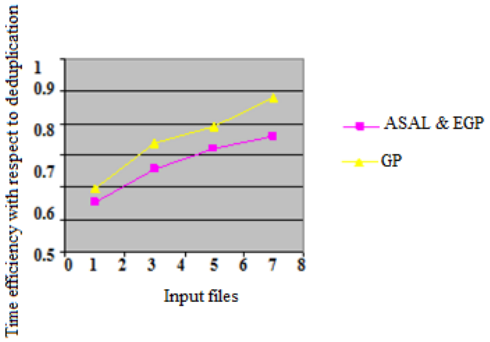

Fig 6. Time efficiency with respect to proposed and existing approaches.

The outcome results come for implemented approach i.e. machine learning scenarios information redundancy shown in above figure 6. Based on these futures, finally proposed approach gives better and efficient deduplication with respect to time and other perspective parameters in cloud computing environment. When compare to traditional machine learning like genetic algorithm will give better communication results in distributed environment.

VI. SUMMERY

Data deduplication is a scalable and efficient redundant data reduction technique for large-scale storage systems, which addresses the challenges imposed by the explosive growth in demand for data storage capacity. We propose and develop Adaptive & secure machine learning approach to record deduplication on various types of applications. Mainly in digital libraries, they have different types of dirty (replicate) data to maintain multiple copies for single file with different scenarios. Our approach follows efficient secure based identification of deduplication with respect to different attributes. This approach consists enhanced machine learning feature to extract and compare homogonous attributes from original files present in cloud computing environment, It follows multi-predictive representation to maintain low resource utilization to process files to different users in distributed manner. Our experimental results show efficient secure data deduplication to maintain different attribute scenario. Further improvement of our proposed approach is to support integrating fingerprint security for deduplication in real time cloud oriented application in outsourced data.

REFERENCES

1. Jiawei Yuan, Shucheng Yu, "Secure and Constant Cost Public Cloud Storage Auditing with Deduplication", Trans. Inf. Theory., vol. 22, no. 6, pp. 644–654, Sep. 2015.
2. Junio de Freitas, Gisele L. Pappa, Altigran S. da Silva, "Active Learning Genetic Programming for Record Deduplication", in *ICML '09: Proc. of the 26th Annual Int. Conf. on Machine Learning*. New York, NY, USA: ACM, 2009, pp. 49–56.
3. M. G. Carvalho, A. H. F. Laender, M. A. Gonc¸alves, and A. S. da Silva, "Replica identification using genetic programming," in *Proc. of the ACM Symposium on Applied computing*, 2008, pp. 1801–1806.
4. J. Gantz and D. Reinsel, "The digital universe decade – are you ready?" http://www.emc.com/collateral/analyst-reports/idc-digitaluniverse-are-you-ready.pdf, May 2010.
5. J. Yuan and S. Yu, "Proofs of retrievability with public verifiability and constant communication cost in cloud," Proceedings of the ACM ASIACCS-SCC'13, 2013.
6. H. Shacham and B. Waters, "Compact proofs of retrievability," in Proceedings of the 14th International Conference on the Theory and Application of Cryptology and Information Security, ser. ASIACRYPT '08, Berlin, Heidelberg, May 2008, pp. 90–107.
7. A. Juels and B. S. Kaliski, Jr., "Pors: proofs of retrievability for large files," in Proceedings of the 14th ACM conference on Computer and communications security, ser. CCS '07. New York, NY, USA: ACM, 2007, pp. 584–597.
8. Y. Dodis, S. Vadhan, and D. Wichs, "Proofs of retrievability via hardness amplification," in Proceedings of the 6th Theory of Cryptography Conference on Theory of Cryptography, ser. TCC '09, Berlin, Heidelberg, 2009, pp. 109–127.
9. G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable data possession at untrusted stores," in Proceedings of the 14th ACM conference on Computer and communications security, ser. CCS '07. New York, NY, USA: ACM, 2007, pp. 598–609.
10. G. Ateniese, R. Di Pietro, L. V. Mancini, and G. Tsudik, "Scalable and efficient provable data possession," in Proceedings of the 4th international conference on Security and privacy in communication netowrks, ser. SecureComm '08. New York, NY, USA: ACM, 2008.
11. C. Erway, A. Küpçü, C. Papamanthou, and R. Tamassia, "Dynamic provable data possession," in Proceedings of the 16th ACM conference on Computer and communications security, ser. CCS '09. New York, NY, USA: ACM, 2009, pp. 213–222.
12. Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li, "Enabling public auditability and data dynamics for storage security in cloud computing," IEEE Trans. Parallel Distrib. Syst., vol. 22, no. 5, pp. 847–859, 2011.
13. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in Proceedings of the 18th ACM conference on Computer and communications security, ser. CCS '11. New York, NY, USA: ACM, 2011, pp. 491–500.
14. R. Di Pietro and A. Sorniotti, "Boosting efficiency and security in proof of ownership for deduplication," in Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, ser. ASIACCS '12. New York, NY, USA: ACM, 2012, pp. 81–82.
15. Q. Zheng and S. Xu, "Secure and efficient proof of storage with deduplication," in Proceedings of the second ACM conference on Data and Application Security and Privacy, ser. CODASPY '12. New York, NY, USA: ACM, 2012, pp. 1–12.
16. K. K. Youngjoo Shin, Junbeom Hur, "Security weakness in the proof of storage with deduplication," Cryptology ePrint Archive, Report 2012/554, 2012, http://eprint.iacr.org/.
17. D. Boneh and X. Boyen, "Short signatures without random oracles," in EUROCRYPT, 2004, pp. 56–73.
18. I. S. Reed and G. Solomon, "Polynomial Codes Over Certain Finite Fields," Journal of the Society for Industrial and Applied Mathematics, vol. 8, no. 2, pp. 300–304, 1960.
19. X. Jia and C. Ee-Chien, "Towards efficient provable data possession," in Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, ser. ASIACCS '12, Seoul, Korea, 2012.
20. "A Hybrid Cloud Approach for Secure Authorized Deduplication" by Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, procedures in IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEM VOL:PP NO:99 YEAR 2014.
21. Gonc¸alves, and Altigran S. da Silva "A Genetic Programming Approach to Record Deduplication", by Moise's G. de Carvalho, Alberto H.F. Laender, Marcos Andre', procedures in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012.
22. M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-bolted encryption and secure deduplication. In EUROCRYPT, pages 296– 312, 2013.
23. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Confirmations of proprietorship in remote stockpiling frameworks. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.
24. J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with proficient and solid merged key administration. In IEEE Transactions on Parallel and Distributed Systems, 2013.
25. C. Ng and P. Lee. Revdedup: A converse deduplication stockpiling framework advanced for peruses to most recent reinforcements. In Proc. of APSYS, Apr 2013.

AUTHORS PROFILE

Prasanth Yalla received his B.Tech Degree from Acharya Nagarjuna University, Guntur (Dist), India in 2001, M.Tech degree in Computer Science and Engineering from Acharya Nagarjuna University in 2004, and received his Ph.D. degree in CSE titled "A Generic Framework to identify and execute functional test cases for services based on Web Service Description Language" from Acharya Nagarjuna University, Guntur (Dist), India in April 2013.

He was an associate professor, with Department of Information Science and Technology in KL University, from 2004 to 2010. Later he worked as Associate professor, with the department of Freshman Engineering from 2011 in KL University. Presently he is working as Professor in the department of Computer Science & Engineering in KL University. Till now he has published 9 papers in various international journals and 4 papers in conferences. His research interests include Software Engineering, Web services and SOA. He taught several subjects like Multimedia technologies, Distributed Systems, Advanced Software Engineering, Object Oriented Analysis and design, C programming, Object-Oriented programming with C++, Operating Systems , Database management systems, UML etc. He is the Life member of CSI and received "Active Participation- Young Member" Award on 13-12-13 from CSI.

Dr. G. Syam Prasad is Currently working as Professor in CSE at KLEF(Deemed to be University), Vaddeshwaram, Guntur . Andhra Pradesh, India.He received the B.Tech and M.Tech degrees from the Department of computer Science and Engineering , Acharya Nagarjuna University, Guntur, India in 1999 and 2004 respectively, and Ph.D. degree from the Department of Computer Science and Systems Engineering , Andhra University at Visakhapatnam, India , in 2015. His research interests include network Security, cryptography, security and privacy, image processing, Data Mining, compilers and algorithms.