International Journal of Engineering & Technology, 7 (3.24) (2018) 422-426

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Integrated Privacy Preserving Data Deduplication Method using Third Party Auditor over Cloud Storage

Dr. Antony Xavier Bronson.F[1*], Dr. Rajagopalan.S[2], Dr. Ramamoorthy.S[3]

[1]Research scholar, Department of computer Science and Engineering,
Dr. M.G.R Educational and Research Institute, Chennai, Tamil Nadu, India.
[2]Professor Emeritus, Department of Computer Applications,
Dr. M.G.R Educational and Research Institute, Chennai, Tamil Nadu, India.
[3]Professor and Dean, Department of Computer Applications,
Dr. M.G.R Educational and Research Institute, Chennai, Tamil Nadu, India.
**Corresponding author E-Mail: antonyxavierbronson@gmail.com*

Abstract

Nowadays the deduplication process plays major role in the field of information technology by handling bigdata and cloud data at the cloud storage. This process occupies less storage because the redundant data are removed and only one copy of the data is kept with high level of trust with security. Traditional method of the handling cloud data is in only one provider, so the same data is available in the provider which causes excessive storage. Existing deduplication algorithms performs the elimination process but it suffers in the performance due to the factors like compression ratio, delay in the process. This problem is overcome by using integrated way of keeping the data with high availability with multi provider interactions through Third Party Auditor (TPA). These interactions are happens through agreement. The privacy level of the data is maintained properly with multi layered structure in order to prevent unauthorized disclosure. The proposed algorithm is compared with various algorithms by considering performance parameters. Analyzes of the deduplication algorithms are carried in efficient manner. The main objective of the proposed method is to integrate more than one cloud provider and maintains only one copy of the data at the cloud storage in highly reliable manner.

Keywords: *Cloud Computing, Storage, Privacy, DeDuplication, TPA*

1. Introduction

Data deduplication is the process of eliminating the redundant copies of the data at the cloud in order to reduce the overhead in storage and bandwidth. There are various algorithms exists to eliminate the data duplication with encryption mechanism. The Cloud Service Provider (CSP) handles the customer data in most efficient way without any problem. The data are outsources to any CSP for storage with encryption technique. The existing algorithms suffer the problem in handling of same plaintext with different cipher text. The data privacy is a challenging task while migrate the data among the CSP. The existing system not provides the protection of duplicate data disclosure. There are two types of dedupilcation process namely pre-process and post-process. In pre deduplication process removes the duplication information i.e. blocks prior to the cloud storage which achieves maximum performance because the processed data alone stored in to the storage. In post deduplication process perform the corresponding operation after storing the data into data storage. The huge amount of data is stored in the cloud which needs the high storage and processing capabilities with various requirements. The cloud needs cost effective way of handling the data at the storage during runtime, so global way of optimization technique will be used for better usage [1]. The file system based data handling requires the namespace method with efficient data access capability. This problem is overcome by using tree structure management but it lacks in dynamic data searching operation.

The above problems are addressed by implementing SANE system with novel storage methods [2]. Packet level deduplication method uses bandwidth saving, link access and 60%, 50% respectively. This technique restricted to single protocol, so independent protocol management is needed. This type of management technique can be extended to all type of protocol with recent technology [3]. Normally the storage devices are outsourced or rented to the third party in order to handle the file content. There are plenty of techniques are used to integrated with various parameters through virtualization. Index Name Servers performs all the basic deduplication operation with advanced features like compression, real time control, index monitoring with workload sharing features over the data [4]. The duplicated data are handled with various documents over the distributed computing needs more efficient access with same category of the server. Similarity-Aware Partitioning algorithm achieves more optimized data management in real time applications from various resources [5]. The main challenges in the real time data has more redundancy with high overhead which is eliminated by using DARE method. It has the features like awareness in location, backup of data, storage archival with duplication detection property. This method suffers data in different location which leads the performance problem [6]. The cloud stores vast amount information because of bigdata technology with variety of information needs reduced IO interaction problem. There are various types of interaction which classifies small IO operation to larger IO operation with sufficient level of redundancy.

Dedup is a deduplication technique which achieves high performance with effective traffic rate but it suffers variable bandwidth and IO operation [7]. There is a gap in performance during the cloud storage operation because of solid state devices which placed over the memory. Hypervisor creates huge number of virtual machines causes the data replication with various Virtual Machines (VMs). The caching mechanism is introduced in order to utilize the resource more effective way. There are various replacement techniques available over the VM for improving cache hit ratio without any redundancy [8]. In deduplication process requires more security because the data threats occur in various levels. This problem is overcome by using privacy measurement with preservation technique. The dynamic updating of the data handling in storage suffers due to redundant data in same location [9]. VM is secured by using crypto keys which is used to prevent the attacks in cache over the cloud computing. There are various channels are identified and implemented for better security on data. Two types of implementations are analyzed namely hardware and software based [10]. The above problem is overcame and addressed by using centralized data deduplication method with enough level of redundancy.. Multiple providers can interact each other by establishing agreement and keeps only one copy of the data with maximum reliability and efficiency. The rest of the paper organized as follows. Section 2 represented as problem formulation .Section 3 describes that the related work of the proposed method. Section 4 depicts that the proposed methodology with conceptual diagram. Section 5 represented as data privacy. Section 6 shows that an algorithm. Section 7 proposes the comparison of various algorithms. Finally section 8 provides the conclusion and future work.

2. Problem Formulation

Normally the data are maintained in the cloud storage with different variety with attributes. The data maintenance places the vital role in the cloud. Traditionally the same data are places in the cloud which leads the performance problem because of the excess storage space and inconsistency during data handling process. Traditional way of handling the data at the cloud storage is restricted to the single cloud provider with their own protocol. The deduplication process done at the single location i.e single cloud vendor. This type of deduplication process lacks in the performance problem due to the same data available in different cloud location. This problem is overcome by using more than one cloud service vendor interacts with other vendor by keeping the one copy of the data. The interaction is done by using the agreement which established between various vendors in order to achieve efficient contribution towards the data maintenance at the cloud. The third party auditor monitored and manages all type of interactions with suitable attributes like indexing, managing access rights, retrieval of data and location identification and so on.

3. Proposed Work

The privacy over the data is maintained properly in order to secure the inappropriate disclosure. The cloud vendors are categorized in to non-malicious and malicious vendor. Non-malicious vendor can access all types of services which belong to the cloud whereas the malicious vendor cannot access the services. This problem is addressed by using the privacy preserved deduplication technique with high level of security. The data are maintained in the cloud storage with deduplication technique is restricted to the single provider is shown in figure 1. The problem is that the same data can be available in other provider which leads the inconsistency problem. This problem is overcome by using more than one provider performs deduplication of data at one location is depicts in figure 2.

The interaction among provider establishes the agreement for sharing the deduplicated properties in order to maintain the single copies of the data.

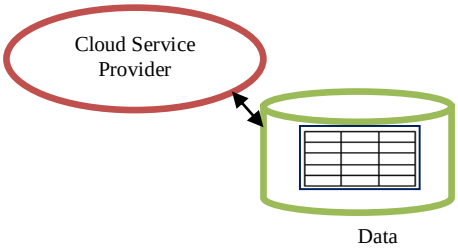

Figure 1: Single provider with Deduplication Process

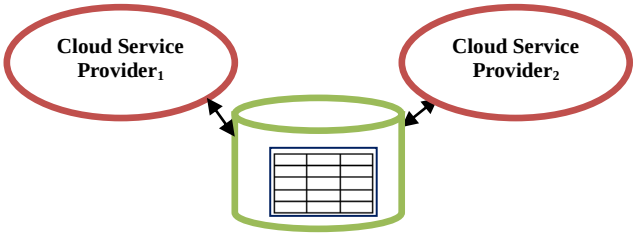

Figure 2: Multiple Providers with Deduplication Process

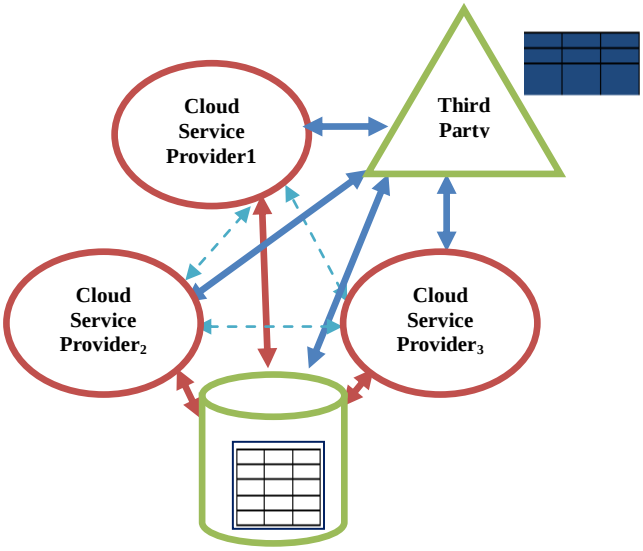

Figure 3: conceptual diagram of the proposed method

4. Proposed Integrated Privacy Preserving Data Deduplication (Ippdd)

The proposed technique uses the Third Party Auditor (TPA) with agreement for identifying the malicious provider for interaction. The provider interact each other for service sharing. The data are kept confidential at one location. The non malicious provider has the access rights to access in the data. The third party is a trusted auditor which manages entire provider and also have a full control over the data. The main objective of the proposed technique is that only one copy of the data are maintained by different provider in order to reduce the storage constraints and data inconsistency. There are various certificates are maintained during the interaction

namely provider certificate, agreement, TPA certificate and also the indexing mechanism for identifying the data location.

There are different phases of the operation is carried out for achieving the efficiency. The first phases is Service Level Agreement (SLA) establishment phases in which the providers are sharing their services, attributes and properties for better handling of deduplication process. All provider has to registered its own identity to the TPA for getting the access rights in order to handled the deduplicated data in more secure manner. TPA receives the provider request and verifies their identity then it will give the approval. This approval plays the vital role for keeping the provider as a non-malicious otherwise malicious. The idea behind this approval is that only the approved provider only accesses the data in centralized location. After approval process the TPA maintains the Index table for provider identification and also allocates the access rights to the provider in case of inserting new data, removing already existing data and updating the data. The main objective of the proposed techniques is to maintain the data in common location without any redundancy in the data with more secure manner with proper preservation of the privacy in all level. The overall architecture of the proposed method is shown in figure 3. Single provider handles the data in cloud storage with their own deduplication standards and protocol is represented in figure 4. The same data is available in the different provider which can handle properly and effectively. This method problem is addressed by using the collaboration of provider with common data maintenance in an efficient manner which is represented in figure 5.

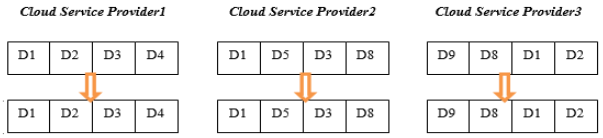

Figure 4: Single Provider with Data Deduplication

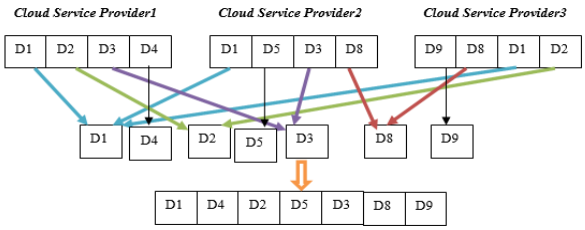

Figure 5: Multi Provider with Data Deduplication

5. Data Privacy Regions

The privacy maintenance is a process of preventing the data from unauthorized disclosure of the malicious user. The privacy boundaries are divided into four levels they are customer level, TPA level, Provider level and storage level. The data are kept confidential because it will travel in various boundaries so preserving the privacy is very important. There are different level of data leakage occur from customer level to data storage level. The privacy boundaries are represented in figure 6.

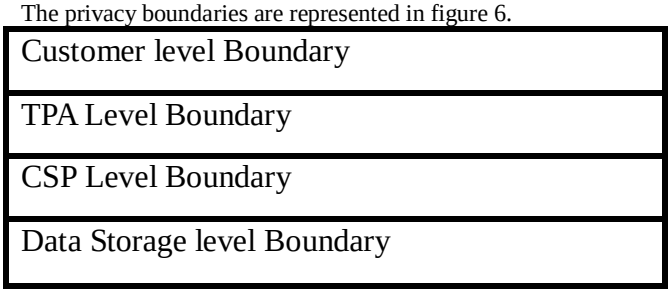

Figure 6: Regions of Privacy

Lemma 1:

All providers store the data in to common cloud storage which leads the better deduplication process.

$$CSP\_DATA_1 \cup CSP\_DATA_2 \cup \ldots CSP\_DATA_N \rightarrow Data\_Cloud\_Storage \quad \textbf{(1)}$$

Lemma 2:

If there is any insertion of data in the cloud storage reflects the indexing table at the TPA.

$$New\_Data \rightarrow TPA\_Index\_Table \cup Cloud\_Storage \quad \textbf{(2)}$$

Lemma 3:

If there is any updation in the existing data reflects the cloud storage and indexing table with all providers.

$$New\_Data \rightarrow (Cloud\_Storage) - (Old\_Data) \quad \textbf{(3)}$$

Lemma 4:

If there is any deletion in the existing data at the storage which reflects the data in the TPA indexing table as well as cloud storage.

$$Delete\_Data \rightarrow (Cloud\_Data) - (Old\_Data) \quad \textbf{(4)}$$

6. Algorithm

6.1 Algorithm for Integrated_deduplication ()

Begin

Let Provider as CSP1, CSP2…CSPn;
Let TPA is a third party auditor;
L1: for each provider ϵ CSP do
Prepare the credentials as USERID;
Identify the agreement attributes ATT;
Establish the SLA;
If (SLA U USERID U ATT) then
Generate the request for TPA registration;
TPA verifies the identity of the provider;
If (verification==Success) then
Provider is non-malicious;
TPA approves the data access for deduplication;
Make an entry to the indexing table;
Allocate the access rights to the providers based on their request;
Else
Provider is malicious;
Deny the access rights to the data;
Exit
Else
Goto L1;

End:

6.2 Algorithm for Privacy_Region_Perservation ()

Begin

For each data ϵ DATA do
Identify the privacy regions as PERregion;
If (PERregion = = User_Level) then
Add the privacy standard at the client region;
Else if (PERregion = = TPA_Level) then
Add the privacy standard at the TPA region;
Else if (PERregion = = Provider_Level) then
Add the privacy standard at the provider region;
Else
Keep the privacy in the deduplication at storage level;

End;

7. Comparisons of Various Algorithms

The deduplication process has various properties which are supported by the cloud storage in order to satisfy the customer with fast response time without any delay. There are different latencies are considered from the client end to storage end namely read latency, write latency and communication latency. Whenever the latency goes minimum level over the data or files at the cloud storage achieves better throughput. The compression ratio is another property which provides the efficient support during the data transformation. There are two ways to analyze the compression ratio in customer and in storage without any redundant data. Compression ratio of various algorithms such as DARE, iDeDup, RCE is compared with proposed method (IPPDD). The deduplication rate of two levels is analyzed over cloud storage. DARE is a deduplication technique with data similarity detection with minimum overhead. The similarity identification ratio is 0.890%. iDeDup is method which is used to optimize the deduplication process with less IO interaction and access time with the average percentage of latency is 13%. RCE is a Randomized Convergent Encryption method for performing deduplication process with more secure manner with the average latency of 16%. The latency comparison related various operations are shown in figure 7, figure 8 and figure 9.

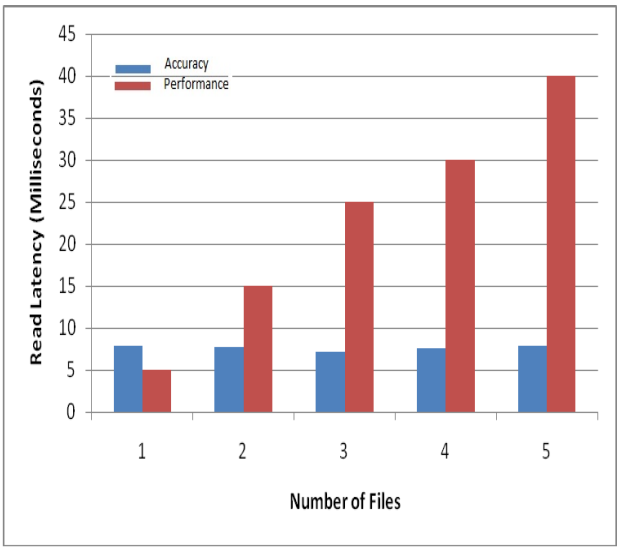

Figure 7: Comparison of Read Latency

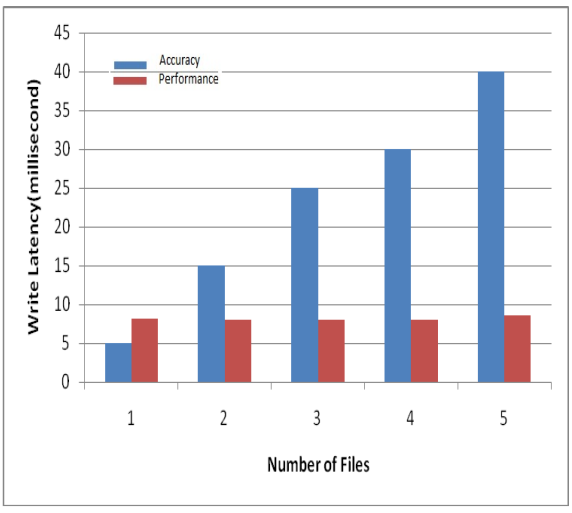

Figure 8: Comparison of Write Latency

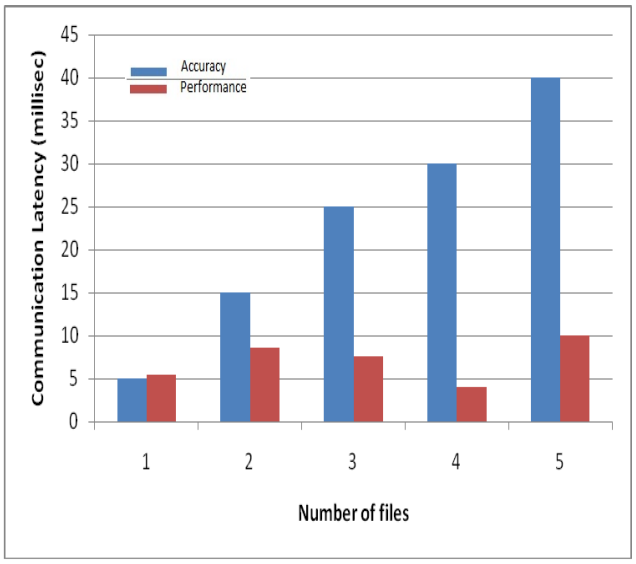

Figure 9: Comparison of Comunication Latency

The compression ration of the deduplication process with different files in clustomer end is shown in figure 10. The compression ration of cloud storage is shown in figure 11. Various algorithms which are related to the deduplication process with compression ratio is depicted in figure12.

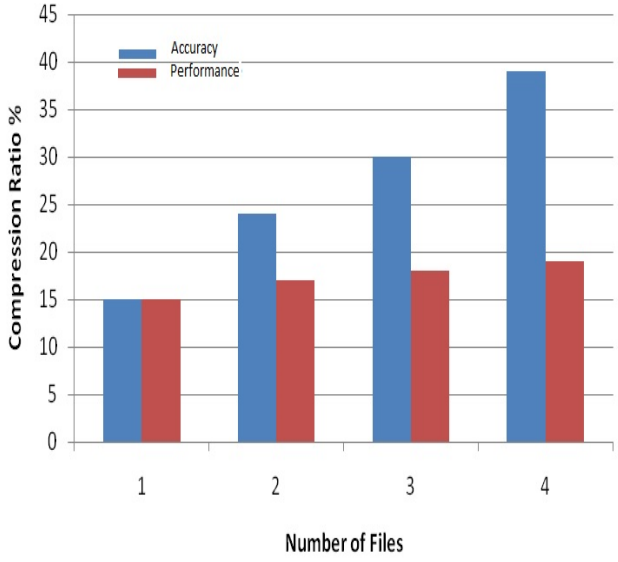

Figure 10:.Client Side Compression Radio

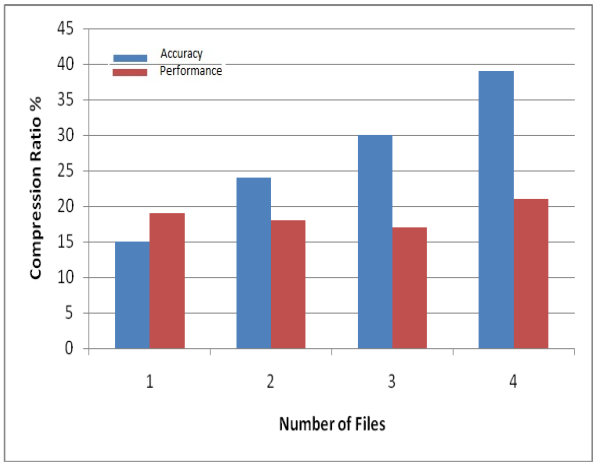

Figure 11: Compression Radio of Cloud Storage

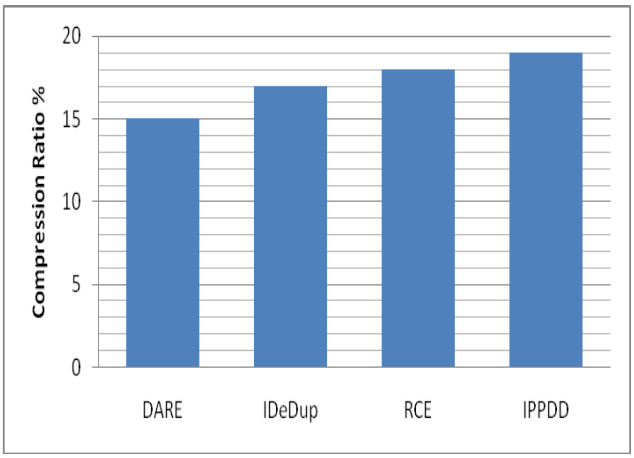

Figure 12: Integrated Compression Ratio

The data deduplication rate of both single provider and integrated provider with different parameters are represented in figure 13 and figure 14 respectively. From the graph analysis the proposed work provides efficient deduplication rate when compared to existing algorithms.

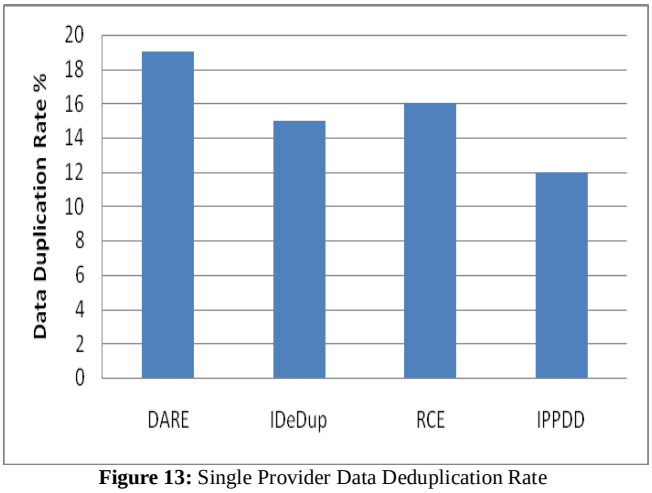

Figure 13: Single Provider Data Deduplication Rate

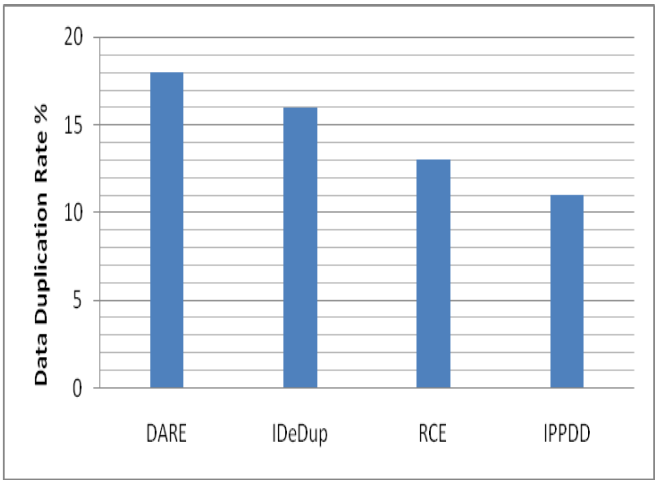

Figure 14: Integrated Data Deduplication Rate

7. Conclusion and Future Work

Deduplication process has been carried out by eliminating redundant data which are collected from different sources to the cloud storage. There may be same set of data is generated by real world object is handled carefully without any crash in the data. There are various algorithms are available but that are restricted to the one cloud storage. Performance attributes has been analyzed and also compared with the reliability over the data. The privacy boundaries are identified as a region for preventing the data from third party disclosure. This boundary is organized from the customer side to storage side for achieving more security. Integration of various providers with their protocol are analyzed with the establishment of agreement. The algorithms are generated based on the integration and privacy over the deduplicated data. The main objective of the proposed method to keep only one copy of the data which common to all provider with high performance and reliability due to less cloud storage. In future the proposed work can be extended to scientific application and education domain.

Reference

[1] Dong Yuan, Yun Yang et al, "A Highly Practical Approach toward Achieving Minimum Data Sets Storage Cost in the Cloud", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO. 6, JUNE 2013, Pp. 1234-1244.

[2] Yu Hua,Hong Jiang, Yifeng Zhu, Dan Feng,"SANE: Semantic-Aware Namespace in Ultra-Large-Scale File Systems",IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 5, MAY 2014,Pp. 1328-1338.

[3] Yan Zhang and Nirwan Ansari,"On Protocol-Independent Data Redundancy Elimination",IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 16, NO. 1, FIRST QUARTER 2014,Pp.455-472.

[4] Tin-Yu Wu,Jeng-Shyang Pan and Chia-Fan Lin, "Improving Accessing Efficiency of Cloud Storage Using De-Duplication and Feedback Schemes",IEEE SYSTEMS JOURNAL, VOL. 8, NO. 1, MARCH 2014, Pp. 208-218.

[5] Jianjiang Li, Jie Wu, and Zhanning Ma,"Frequency and Similarity-Aware Partitioning for Cloud Storage Based on Space-Time Utility Maximization Model",TSINGHUA SCIENCE AND TECHNOLOGY,ISSNll1007 0214ll02/10ll,Volume 20, Number 3, June 2015,pp233-245.

[6] Wen Xia,Hong Jiang,Dan Feng and Lei Tian,"DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for DataReduction with Low Overheads",IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016,Pp.1692-1705.

[7] Bo Mao,Hong Jiang,Suzhen Wu and Lei Tian,"Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud",IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016, Pp.1775-1788.

[8] Xian Chen,Wenzhi Chen,Zhongyong Lu, Peng Long, Shuiqiao Yang, and ZonghuiWang, "A Duplication-Aware SSD-Based Cache Architecture for Primary Storage in Virtualization Environment", IEEE SYSTEMS JOURNAL, VOL. 11, NO. 4, DECEMBER 2017,Pp.2578-2589.

[9] Mi Wen, Kaoru Ota, He Li,Jingsheng Lei, Chunhua Gu, and Zhou Su,"Secure Data Deduplication With Reliable Key Management for Dynamic Updates in CPSS",IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, VOL. 2, NO. 4, DECEMBER 2015,Pp.137-147.

[10] Berk Gulmezoglu, Mehmet Sinan _Inci, Gorka Irazoqui, Thomas Eisenbarth, and Berk Sunar,"Cross-VM Cache Attacks on AES", IEEE TRANSACTIONS ON MULTI-SCALE COMPUTING SYSTEMS, VOL. 2, NO. 3, JULY-SEPTEMBER 2016, Pp.211-222.

[11] K. Vijayakumar C. Arun, Continuous security assessment of cloud based applications using distributed hashing algorithm in SDLC, Cluster Computing DOI 10.1007/s10586-017-1176-x,Sept 2017

[12] R.Joseph Manoj, M.D.Anto Praveena, K.Vijayakumar, "An ACO–ANN based feature selection algorithm for big data", Cluster Computing The Journal of Networks, Software Tools and Applications, ISSN: 1386-7857 (Print), 1573-7543 (Online) DOI: 10.1007/s10586-018-2550-z, 2018.

[13] K. Vijayakumar, C.Arun, Automated risk identification using NLP in cloud based development environments Ambient Intell Human Computing, DOI 10.1007/s12652-017-0503-7, Springer May 2017.