
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 5, pp. 3691–3700, DOI 10.12694/scpe.v25i5.3099

MULTI OBJECTIVE DATA TRANSFORMATION IN HYBRID CLOUDS NETWORKS FOR OFFLOADING DATA

V SRIDHAR REDDY*, N. JAYANTHI†, SHARON ROSE VICTOR JUVVANAPUDI‡, SRINIVAS BACHU§, AND MADIPALLI SUMALATHA¶

Abstract. Recently hybrid cloud solutions integrating public and private cloud is proposed to address the privacy and security concerns faced by Enterprises in their data offloading decisions. In these solutions, the transformed data is kept in public cloud while transformation keys are kept in private cloud. The existing works for data transformation used in hybrid clouds does not address multiple objectives of privacy, security, fine grained access control, utility preservation for mining and data retrieval efficiency. This work proposes a multi objective data transformation technique for hybrid cloud to address all these objectives. The proposed solution is built on attribute based hierarchical data access control with hierarchy selection based on joint consideration of security, utility preservation and retrieval efficiency. The proposed solution is able to provide 5% higher security strength, 1.34% higher clustering accuracy over perturbed data and 29.95 % higher data retrieval efficiency over perturbed data compared to existing works.

Key words: Multi objective data transformation, Hybrid cloud, Hierarchical data access control, Generalization control .

1. Introduction. Enterprises are adopting cloud for offloading both storage and computations. The adoption is triggered due to various benefits like CAPEX and OPEX reduction, high availability and mobility etc. With increasing cloud adoption rate, there is also increasing cloud security breaches. The recent security survey by IDC and Ermetic [1] reports that almost 98% of enterprises suffer atleast one security breach. The average total cost of data breach globally is estimated about 4.24 million USD. Data breaches and leakage can create huge financial loss for the Enterprise, lose to competitors and sometimes wipe out from market. Thus ensuring security and privacy of data has become a important requirement for enterprises in their cloud offloading and vendor selection decisions. Though there are various data protection mechanisms, enterprises are adopting multi cloud and hybrid clouds to reduce the risk. Flexera's 2021 State of the Cloud Report [2] found that almost nine out of ten enterprises are adopted multi cloud approach and eight in that nine enterprises are adopting hybrid cloud to reduce security risks. Hybrid cloud solutions are also not full proof. Though they have reduced risks and breach cost compared to public and private cloud, they could not completely eliminate the data breach cost as evident from the IBM and the Ponemon Institute's 2021 Cost of a Data Breach Report Figure 1.1

This data breach cost could be avoided in hybrid cloud with more effective security and privacy enforcement. The data must be prevented from compromise either directly or through inference. Many works have been proposed in category of anonymization, randomization, cryptography, diversification and aggregation to address the security and privacy concerts. In addition to security and privacy, the data transformation techniques must also address other requirements like differential access control, utility preservation and retrieval efficiency. Most solutions as discussed in Section II do not address all these requirements. This work proposes a multi objective data transformation technique which jointly addresses all the five requirements of privacy, security, differential

*Department of Information Technology, Vignana Bharathi Institute of Technology, Hyderabad, India (vsridharreddy19@gmail.com)
†Department of Computer Science Engineering, CMR Institute of Technology, Bengaluru, India (jayanthi.n@cmrit.ac.in)
‡Department of Electronics and Communication Engineering, Pragati Engineering College,Surampalem, AP, India (jsr.victor@gmail.com)
§Department of Electronics and Communication Engineering, Siddhartha Institute of Technology & Science, Hyderabad,Telangana, India (bachusrinivas@gmail.com)
¶Department of Electronics and Communication Engineering, Siddhartha Institute of Technology & Science, Hyderabad,Telangana, India (sumasriece@gmail.com)

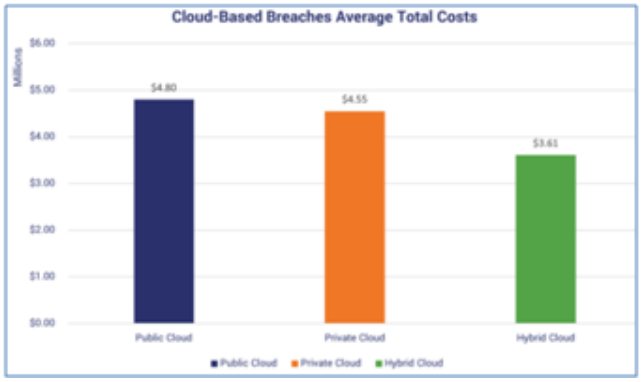

Fig. 1.1: Data breach report (Courtesy: IBM and Ponemon Report 2021)

access control, utility preservation and retrieval efficiency. Attribute based hierarchical data transformation with mix of anonymization, aggregation and diversification is adopted in this work with hierarchy selection fine tuned to address the five requirements.

The rest of paper is organized as follows. Section II provides the survey of data transformation techniques used in cloud and their shortcomings. Section III details the proposed multi objective data transformation technique. Section IV provides the results of the solution and its comparison to existing works. Section V provides the concluding remarks and the scope for future work.

2. Related Work. A survey of data transformation techniques for cloud is presented in this section. Yang et al [3] proposed a data transformation technique addressing security and privacy for data in cloud. The data is partitioned vertically and transformed using cryptographic primitives. The keys for transformation are kept at the private cloud and transformed data at public cloud. The transformation is not distance preserving and the method is not able to provide differential privacy to users. Kao et al [4] proposed a reversible privacy contrast mapping (RPCM) algorithm for data transformation.

Data is transformed by replacing two adjacent values by a new value. The mapping between adjacent values to new value is kept separately in private cloud. By grouping the values, anonymity is created among the records. But without consideration for distance preservation in transformation, the utility of data for data mining becomes infeasible. Yun et al [5] proposed a faster data perturbation algorithm using tree travel strategy.

A multi tier tree structure is built, which is able to transform a numeric attribute to another attribute. Though the retrieval efficiency is ensured, the differential privacy is not considered in this work. Zhang et [6] proposed a data transformation technique called as Cocktail.

This technique applied quasi identified partitioning with differential privacy strategy. The data transformation is lossless. But the retrieval efficiency is low in this approach. Zhou et al [7] proposed a data partition strategy. The strategy is application independent. The sensitive attributes are detected based on entropy with the identifier. The sensitive columns are kept in private cloud. The insensitive columns are moved to public cloud. Though retrieval efficiency is high in this approach, it distorts the data mining utilization. Lyu et al [8] transformed the data using repeated Gompertiz (RP) followed by random projection (RP). The data is transformed to less dimension space with distance preserving property so that utility is not affected. But the retrieval efficiency is poor and it is not possible to execute any query operations on transformed data. Security is strong this approach. The transformed data is secure against estimation and component analysis attacks. Chen et al [9] proposed geometric data perturbation scheme. The geometric perturbation has three steps of rotation, translation and noise addition.

The mechanism has tighter security and privacy, bit retrieval efficiency is low for higher dimensional datasets. The same author in [10] proposed a random projection perturbation extension to geometric perturbation. The method is able to achieve faster geometric perturbation for multi dimensional datasets. Yuan et al [11] used

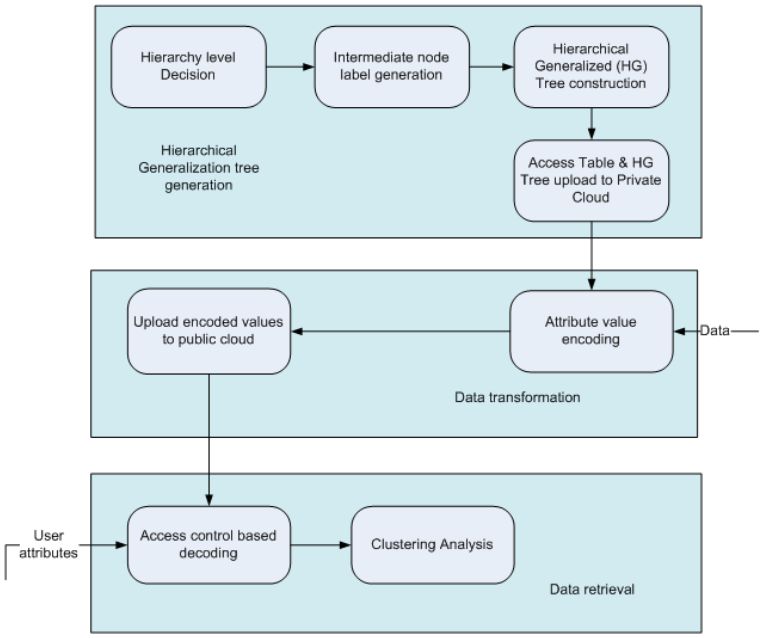

Fig. 2.1: Multi objective data transformation

compressive sensing based data transformation and fast indexing to improve retrieval efficiency. Due to violation of distance preservation, the transformed data becomes unsuitable for data mining operations like clustering and classification. Majeed et al [12] proposed a data transformation technique based on data anonymization. The attribute values in fixed interval are replaced with their averages. This method affects the original data and suitable only for certain data publishing requirements. Li et al [13] proposed two K-anonymity algorithm for data transformation. The data is transformed in way not to affect the classification capability. It is done by checking the entropy after transformation and choosing the level of transformation based on entropy. Begum et al [14] proposed data transformation scheme by removing sensitive items based on support and confidence values. The minimum number of items is removed in such a way to remove sensitivity.

Though the method is able to reduce the security leakages, it reduces the utility of data for data mining operations. Sridhar et al [15] clustered the data and passed the clustered data to geometric data perturbation. The solution considered security, privacy and utility preservation but did not consider retrieval efficiency. The solution did not consider differential privacy and query matching. Kodhai et al [19] proposed a secure fuzzy keyword search technique for data stored in cloud. But the technique cannot used for the case of differential privacy and fine grained access controlled search problem considered in this paper work. Gheisari et al [20] used four different techniques of data micro aggregation, sampling, swapping and random noise to perturb data. But these schemes do not consider fine grained access control and differential privacy. Jafar et al [21] used public key cryptosystem for providing security to data but the scheme does not support differential privacy, fine grained access control and search over perturbed data.From the survey, it can be seen that most of the data transformation technique focused on privacy and security, but they have not considered multi objectives of providing differential privacy, fine grained access control over data, retrieval over perturbed data and preservation of utility mining etc. In addition, the existing works have not considered efficiency in data transformation on hybrid cloud platform. This work considers this problem of multi objective data transformation and efficiency in data transformation for hybrid cloud environment.

3. Multi objective data transformation. The architecture of the proposed multi objective data transformation is given in Figure 2.1. As seen from figure, the proposed solution has three important functionalities: generation of hierarchical generalization tree, data transformation and retrieval. Each of the functionalities is

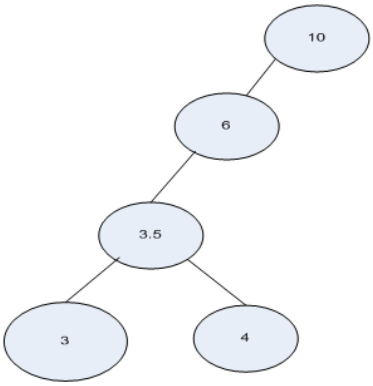

Fig. 2.2: Generalization tree for categorical variable

detailed below.

3.1. Generation of hierarchical generation tree. The data uploaded by data owner is in table format in which each column is an attribute. The attributes are marked as sensitive or in-sensitive by the data owner. For each of the sensitive attributes, a hierarchical generalization tree is constructed. This generalization tree is constructed to transform the data values of the corresponding attribute. The transformation using the generalization tree provides a differential view on the data after retrieval as desired by owner. Generalization is the key to the differential view. A sample generalization tree is shown in Figure 2.2. The attribute values are at the leaf nodes. The intermediate nodes in the tree are the generalization labels. It is not possible to construct a semantically correct generalization label for categorical variables as it requires domain knowledge. Thus data owner must provide the translation for each of the categorical variable in terms of numbers. The number must be provided in such way that if two categorical variables a and b are semantically close by degree d_1, then the distance between their corresponding numerical variables must be in proportion to d_1.

$$|N(a) - N(b)| \propto d_1 \tag{3.1}$$

Also if there is order in categorical variable, then if

$$a < b \text{ then } N(a) < N(b) \tag{3.2}$$

The generalization tree for attributes is constructed automatically by normalized the attribute values and repeated binary split till the maximum level allowed by data owner is achieved. The generalization tree is constructed in way to maximize the data mining utilization. The attribute value is fist normalized in range of 0 to 1 from their actual value to decide the hierarchy. The normalization is done as

$$NV = \frac{AV}{(MaxV - Minv + 1)} \tag{3.3}$$

where NV is the normalized value, AV is the actual value, MaxV is the maximum value of the attribute and Minv is the minimum value of the attribute.

A Gaussian kernel density function is plotted with the normalized values. The minima of kernel density estimation for normalized values are taken the partitions as shown in Figure 3.2. In the Figure 3.2, minima of kernel is at $< 0, 0.4, 0.5, 0.9, 1 >$. Thus 4 clusters need to created with values from (0 to 0.4), (0.4 to 0.5), (0.5 to 0.9), (0.9 to 1).

Once the normalized ranges are identified, they can again brought back to actual value.

$$AV = NV \times (MaxV - Minv + 1) \tag{3.4}$$

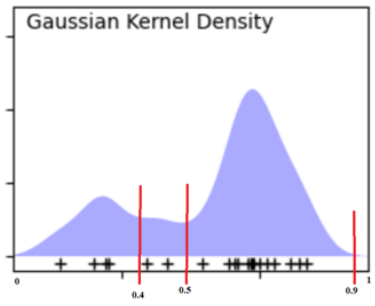

Fig. 3.1: Gaussian Kernal Density

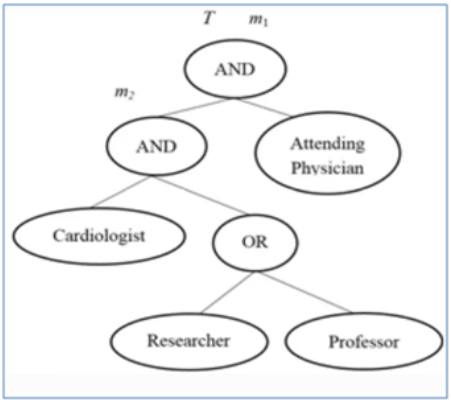

Fig. 3.2: Access tree

The cluster ranges are now in actual values. The mean of the values in those cluster is taken as next level generalization label. The generalization label is calculated for next level in same way till the maximum level specified by the owner is achieved. Once the generalization trees for the attributes are constructed, access control tree is provided by data owner for each level. A sample access tree for each generalization label is given in Figure 4.1.

The access tree is generated for each level based on the user attributes by data owner and access tree is uploaded to private cloud.

3.2. Data transformation. Each attribute value (V) is encoded into all its level generalization as En=V,L_1 (V),L_2 (V),...L_n (V) where L_x (V) is the level x generalization of V and n is the number of levels. A homomorphic encryption (HE) key (k) is generated for each data owner. The encoding En is then homomorphically encrypted using the k, the encrypted En is given as

$$E(En) = HE(V,k),(HE(L_1)(V),k),.(HE(L_n)(V),k) \tag{3.5}$$

The control to the particular level in the E(En) is enforced using AHAC (Attribute based Hierarchical Access Control) CP-ABE (Cipher policy Attribute Based Encryption) [16]. The access tree for each level and E(En) is passed as input to the AHAC CP-ABE encryption algorithm. The encryption algorithm provides a transformed E(En) as output. The algorithm for transformation of attribute value is given in Algorithm 1.

Algorithm 1: Encoding

Input: attribute value (V) , HG tree of Attribute, HE key k, access control tree(T)

1. En=V

Table 4.1: Clustering accuracy

K	RG+RP [8]	GP [15]	SFAC-SHC [17]	Proposed
2	66.23	65.89	68.52	70.12
3	66.56	71.25	75.62	77.14
4	73.33	74.59	78.85	79.89
5	67.22	79.58	82.34	83.56

2. for x=1: num of level(HG)

$$En = En \cup HE(L_x(V), k)$$

3. EV AH ACCP-ABE.*Encrypt*(En,T) [16]
4. upload EV to public cloud
5. upload HG, k,T to private cloud

Each of the attribute value in the data is transformed using Algorithm 1.The transformed data is then uploaded to private cloud.

3.3. Data retrieval. When user requests the data for data mining utilities like clustering, the transformed must be decoded at first stage. The user attributes and transformed attribute value as input and provides the value at index of E(En) corresponding the access provided for the user. Since the value is homomorphically encrypted , distance preservation is maintained and the data is suitable for data mining operations like clustering without any need for decryption. Since the $HE(L_x(V), k)$ is provided only for the level x matching the user access control credentials, it is difficult for user to learn any data characteristics beyond his access rights. The decoding algorithm for de-transformation of each attribute value in transformed data is given as Algorithm 2.

Algorithm 2: Decoding
Input: Query User (U).
Ouput: Decoded Value (DV).

1. $EV \leftarrow downloadfromprivatecloud$
2. QV ← Download user attributes from public cloud(U)
3. $DV \leftarrow AHACCP - ABE.Decrypt(EV, QV)$ [16]
4. return DV

Each of the attribute value in transformed data is decoded using Algorithm 2. This de-transformed data is then used for data mining utilities like clustering analysis.

4. Result. The performance of the proposed solution is tested against Arrhythmia dataset [18] in UCI machine learning repository. The performance is measured in terms of: (i) clustering accuracy (ii) data storage overhead (iii) retrieval efficiency and (iv) security against attacks. The performance of proposed solution is compared against geometric data perturbation (GP) [15], RG+RP [8] and searchable fine access control on secure hybrid clouds (SFAC-SHC) [17].

The clustering accuracy is calculated by measuring the differences between clusters of original and perturbed data. K-means algorithm is used for clustering the original and perturbed dataset.
The clustering accuracy is calculated as

$$ACC = \frac{1}{N}\sum_{i=1}^{N} | - Cluster_i(P)| - |Cluster_i(P')| \tag{4.1}$$

where P is the original data, is the transformed data, k is the number of clusters and N is the number of items in the dataset. The value of ACC is measured for different values of K and result is given in Table 4.1.

The value of ACC in proposed multi objective transformation is 9.3% higher compared to RG+RP , 4.87% higher compared to GP and 1.34% higher compared to SFAC-SHC . With the increase in K value, the ACC

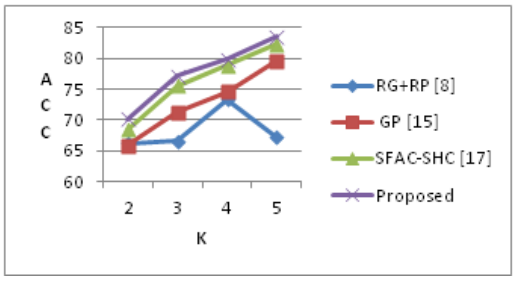

Fig. 4.1: Comparison of clustering accuracy

Table 4.2: Data upload time

Size (MB)	**Upload time (sec)**			
	Proposed	**SFAC-SHC[17]**	**GP [20]**	**RG+RP [8]**
20	13	14	19	22
40	18	24	35	38
60	24	43	61	64
80	30	78	112	122
Average	21.25	39.75	56.75	61.5

increases in the proposed solution. ACC value increases by 13.82 % for K value increase from 2 to 5. Distance preservation is ensured using Homomorphic encryption in proposed solution. This has increased the accuracy in the proposed solution.

The data storage time is measured as the time taken for transformation of data and uploading of transformed data to cloud. The data storage time is measured for various volumes of data and result is given in Table 4.2.

The storage time in proposed multi objective transformation is atleast 20.36% lower compared tp GP and 36.93% lower compared to RG+RP. While storage time increases exponentially with increase in data volume in existing works, it is linear in proposed solution. This is because generalization tree construction and data transformation are linear operation in proposed solution.

The data retrieval efficiency is measured by varying the volume of data and the result is given in Table 4.3. The data retrieval efficiency is on average 29.95% lower compared to GP and 35.36% lower compared to RG+RP. Like storage time, retrieval time also increases exponentially with increase in data volume in existing works, but it is linear in proposed solution. The data de transformation using AHAC CP-ABE is linear in proposed solution and due to this retrieval time increases linearly with increase in data volume.

The security strength is measured in terms of difficulty in predicting the original data from perturbed data, provided the attacker has access to the perturbed data. The difficulty level is estimated in terms of measure called Variance of difference (VoD).

Let X_i be a random variable representing the column i, X_i' be the estimated result of X_i and difference $D_i = X' - X$. Let mean of D be $E(D_i)$ and variance be $Var(D_i)$. VOD for column i is $Var(D_i)$. VOD is measured for each column and average VOD is given as privacy measure(pm)

$$pm = \sum_{i=1}^{N} \frac{VOD_i}{N} \tag{4.2}$$

A guess is launched for 5 hours on the perturbed data and the privacy measure (pm) is measured for every 1-hour interval and plotted in Figure 4.4.

Higher the value of VoD, the effort to predict the original data is difficult. VoD in proposed solution is very high in proposed solution. It is almost twice compared to GP and RG+RP and it is on average 5% higher

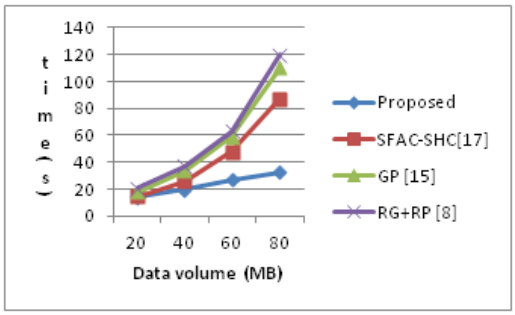

Fig. 4.2: Storage time

Table 4.3: Data retrieval time

Size (MB)	Upload time (sec)			
	Proposed	**SFAC-SHC[17]**	**GP [20]**	**RG+RP [19]**
20	14	15	18	21
40	20	26	34	37
60	27	48	59	63
80	33	87	110	120
Average	23.5	44	55.25	60.25

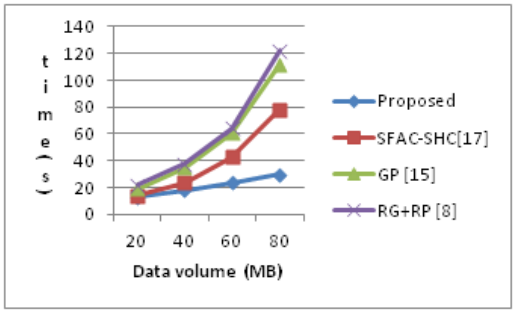

Fig. 4.3: Comparison of retrieval time

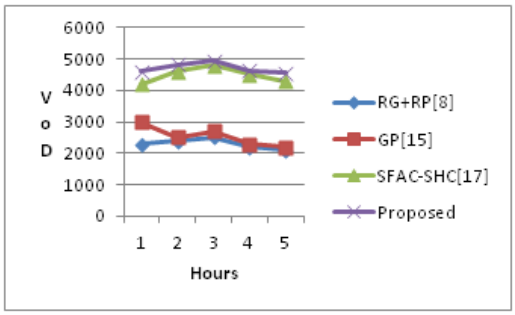

Fig. 4.4: VOD over time

compared to SFAC-SHC. Two level of encryption – first with Homomorphic encryption, followed by AHAC CP-ABE has made it difficult to predict the original data in the proposed solution.

5. Discussion. The data transformation technique proposed in this work addressed multiple objectives of privacy, security, differential access control, utility preservation and retrieval efficiency. The private data were perturbed using homomorphic encryption (HE) with access control over the keys for HE. The privacy and security was measured in terms of VOD and it is found that it is harder to infer any private information by launching brute force attacks. The security is higher by 5% compared to existing works. The higher security is due to differential privacy provided over data. Existing works used same keys without considering differential privacy. The proposed solution preserved distance based statistics in data even after perturbation. This has increased the accuracy for clustering operations on perturbed data by atleast 1.34% compared to existing works. Distance preservation in proposed solution is due to HE based transformation. The data retrieval efficiency is also higher in proposed solution as the transformation is light weight and easy to reverse the data. But existing works based geometric transformation had higher overhead during the stage of data retransformation. The proposed solution had fine grained access control over the data which was not considered in earlier works. The access control was using hierarchical access tree and the tree itself was secured by keeping it in private cloud. Thus the proposed solution performed better compared to existing works in terms of privacy, security, fine grained access control and retrieval efficiency.

6. Conclusion. A multi objective data transformation function for hybrid cloud is proposed in this work. The solution addressed multi objectives of privacy, security, differential access control, utility preservation and retrieval efficiency in data transformation specific to hybrid cloud environment. A generalized hierarchical tree is constructed from the data and data transformation is done based on generalization labels and access control rights for the users in the proposed solution. The proposed solution also provides differential privacy and access control to different users without affecting the utilization of data for data mining operations. The proposed solution provides atleast 5% higher data security, 29.95% higher data retrieval efficiency and 1.34% higher clustering accuracy over perturbed data compared to existing works. Adaption of solution for streaming data is in scope of future work.

REFERENCES

[1] Dan Yachin *"l.ermetic.com/wp-idc-survey-results"*, 2021.

[2] *"https://info.flexera.com/CM-REPORT-State-of-the-Cloud.'.*

[3] Yang, Ji-Jiang, Jian-Qiang Li, and Yu Niu., *"A hybrid solution for privacy preserving medical data sharing in the cloud environment."*, Future Generation computer systems 43 (2015): 74-86.

[4] Kao, Yuan-Hung, Wei-Bin Lee, Tien-Yu Hsu, Chen-Yi Lin, Hui-Fang Tsai, and Tung-Shou Chen, *"Data perturbation method based on contrast mapping for reversible privacy-preserving data mining."*, Journal of Medical and Biological Engineering 35 (2015): 789-794.

[5] Yun, Unil, and Jiwon Kim., *"A fast perturbation algorithm using tree structure for privacy preserving utility mining."*, Expert Systems with Applications 42, no. 3 (2015): 1149-1165.

[6] Zhang, Hongli, Zhigang Zhou, Lin Ye, and Xiaojiang Du. , *"Towards privacy preserving publishing of set-valued data on hybrid cloud."*, IEEE Transactions on cloud computing 6, no. 2 (2015): 316-329.

[7] Zhou, Zhigang, Hongli Zhang, Xiaojiang Du, Panpan Li, and Xiangzhan Yu. , *"Prometheus: Privacy-aware data retrieval on hybrid cloud"* , In 2013 Proceedings IEEE INFOCOM, pp. 2643-2651. IEEE, 2013.

[8] Lyu, Lingjuan, James C. Bezdek, Yee Wei Law, Xuanli He, and Marimuthu Palaniswami., *"Privacy-preserving collaborative fuzzy clustering."*, Data & Knowledge Engineering 116 (2018): 21-41.

[9] Chen, Keke, Gordon Sun, and Ling Liu. , *"Towards attack-resilient geometric data perturbation."*,In proceedings of the 2007 SIAM international conference on Data mining, pp. 78-89. Society for Industrial and Applied Mathematics, 2007.

[10] Chen, Keke, and Ling Liu., *"Geometric data perturbation for privacy preserving outsourced data mining."*,Knowledge and information systems 29 (2011): 657-695.

[11] Yuan, Xingliang, Xinyu Wang, Cong Wang, Jian Weng, and Kui Ren, *"Enabling secure and fast indexing for privacy-assured healthcare monitoring via compressive sensing."*,IEEE Transactions on Multimedia 18, no. 10 (2016): 2002-2014.

[12] Yuan, Xingliang, Xinyu Wang, Cong Wang, Jian Weng, and Kui Ren. , *"Enabling secure and fast indexing for privacy-assured healthcare monitoring via compressive sensing."*,IEEE Transactions on Multimedia 18, no. 10 (2016): 2002-2014.

[13] Li, Jiuyong, Jixue Liu, Muzammil Baig, and Raymond Chi-Wing Wong., *"Information based data anonymization for classification utility."*,Data & Knowledge Engineering 70, no. 12 (2011): 1030-1045.

[14] Sabin Begum, R., and R. Sugumar., *"Novel entropy-based approach for cost-effective privacy preservation of intermediate datasets in cloud."* ,Cluster Computing 22, no. Suppl 4 (2019): 9581-9588.

[15] Reddy, Vulapula Sridhar, and Barige Thirumala Rao., *"A Combined Clustering and Geometric Data Perturbation Approach for Enriching Privacy Preservation of Healthcare Data in Hybrid Clouds."* ,International Journal of Intelligent Engineering & Systems 11, no. 1 (2018).

[16] He, Heng, Liang-han Zheng, Peng Li, Li Deng, Li Huang, and Xiang Chen., *"An efficient attribute-based hierarchical data access control scheme in cloud computing."* ,Human-centric Computing and Information Sciences 10 (2020): 1-19.

[17] Vulapula, Sridhar Reddy, and Srinivas Malladi., *"Attribute-Based Encryption for Fine-Grained Access Control on Secure Hybrid Clouds."* ,International Journal of Advanced Computer Science and Applications 11, no. 10 (2020).

[18] , *"https://archive.ics.uci.edu/ml/datasets/Arrhythmia"* .

[19] Pachipala, Yellamma, and Jafar Alzubi., *"Managing the cloud storage using deduplication and secured fuzzy keyword search for multiple."* ,International Journal of Pure and Applied Mathematics 118, no. 14 (2018): 563-565.

[20] Gheisari, Mehdi, Hamid Esmaeili Najafabadi, Jafar A. Alzubi, Jiechao Gao, Guojun Wang, Aaqif Afzaal Abbasi, and Aniello Castiglione. , *"OBPP: An ontology-based framework for privacy-preserving in IoT-based smart city."* ,Future Generation Computer Systems 123 (2021): 1-13.

[21] Alzubi, Jafar A., Ramachandran Manikandan, Omar A. Alzubi, Issa Qiqieh, Robbi Rahim, Deepak Gupta, and Ashish Khanna., *"Hashed Needham Schroeder industrial IoT based cost optimized deep secured data transmission in cloud."* ,Measurement 150 (2020): 107077.

Edited by: Anil Kumar Budati
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Dec 31, 2023
Accepted: Apr 8, 2024