An enhanced privacy preserving approach with enforcing policies for processing big data in spark framework

S. Revathy 1*, Dr. Arunkumar Thangavel 1

1 School of Computer Science and Engineering, VIT, Vellore
*Corresponding author E-mail: revathy.swaminathan@vit.ac.in arunkumar.thangavelu@gmail.com

Abstract

Ensuring the privacy for the big data stored in a cloud system is one of the demanding and critical process in recent days. Generally, the big data contains a huge amount of data, which requires some security measures and rules for assuring the confidentiality. For this reason, different techniques have been developed in the traditional works, which intends to guarantee the privacy of the big data by implementing key generation, encryption, and anonymization mechanisms. But it limits the issues of increased time consumption, computational complexity, and error rate. Thus, the proposed work aims to design an enhanced mechanism for a secure big data storage. Here, the user’s bank dataset is considered as the input, which is protected from the unauthorized users by guaranteeing both the privacy and secrecy of the data. Here, the raw dataset is preprocessed to increase the data quality and correctness. Then, the security policies (i.e., rules) are generated for allowing the restricted access on the data by using an Improved FP-Growth (IFP-G) algorithm. Consequently, the sensitive and non-sensitive data attributes are classified based on the extracted features by using an Enhanced Random Forest (ERF) classification technique. At last, the privacy of user’s personal information and other details are protected with the use of a Modified Incognito Anonymization based Privacy Preservation (MIA-PP) algorithm. These enhanced mechanisms guarantee the security and confidentiality of the big data with reduced time consumption and increased accuracy. During experimental evaluation, the results of the proposed privacy mechanism is analyzed and compared by using different measures. Also, some of the existing anonymization and classification techniques have been considered to prove the betterment of the proposed technique.

Keywords: Big Data privacy; Enhanced Random Forest (ERF) Classification; Modified Incognito Anonymization based Privacy Preservation (MIA-PP); Improved FP-Growth (IFP-G) and Confidentiality.

1. Introduction

Cloud is the most prominent technology in recent days, due to its wide range of features and enormous services [1]. Also, it offers a shared pool of resources for handling different types of applications. It has the capability to execute millions of instructions per second, so it has been extensively used in many application areas [2]. In this environment, the cloud providers could follow many privacy and security policies for ensuring the data confidentiality [3]. For this reason, some security controls have been developed that analyzes the threats and provides the security to the data [4]. In which, the privacy of the data should be ensured for avoiding the unauthorized access on the data [5]. The following states are considered during privacy preservation that includes: personal sensitive data transmission to the cloud server, data transmission from the server to clients, and personal data storage on the cloud servers [6]. In cloud, preserving the information of the user, identity, and data are essential while performing public auditing. The issues arise during the attainment of privacy preservation are listed in the following Table 1.

<table>
<thead>
<tr>
<th>Issue</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insufficient user control</td>
<td>The data owner has lacks in their control over the data, specifically during the data accessing and processing.</td>
</tr>
<tr>
<td>Information disclosure</td>
<td>The disclosure of sensitive data, which may be user’s identity and data usage.</td>
</tr>
<tr>
<td>Unauthorized secondary storage</td>
<td>Accessing the retrieving the sensitive information from the cloud storage.</td>
</tr>
<tr>
<td>Uncontrolled secondary storage</td>
<td>The data flow is uncontrollable and unpredictable.</td>
</tr>
<tr>
<td>Dynamic provision</td>
<td>Due to the dynamic nature of cloud, it has the responsibility to assure the privacy.</td>
</tr>
</tbody>
</table>

In the traditional works, different privacy preservation techniques are used for big data storage, which includes public auditing for data storage, and public auditing for cloud storage. In which, the third party auditing, and data correctness are performed to assure the privacy preservation. These techniques have the major benefits of data correctness, secured batch auditing, and identity preservation. Moreover, the policy based privacy preservation techniques are also implemented for preserving the user’s data with controlled access. The techniques such as Policy Decision Points (PDP) and Policy Enforcement Points (PEP) are considered the most suitable measures for making the authorization decisions. Then, it preserves the data from the internal and external attacks by using the components of interface engine, rule engine and cloud database.
Also, big data [7] has been attracted by many professionals from different industries, because it contains a vast amount of data collected from different sources [8], which is categorized based on the characteristics of volume, velocity, variety, veracity, and value. So, providing security to big data is one of the demanding and crucial task in current days [9]. The architecture of big data privacy preservation is shown in Fig 1.

![Privacy Preservation](image)

1.1. Problem identification

In the traditional works, different techniques such as encryption, key generation, anonymization, rule generation and privacy preservation are developed for ensuring the security and confidentiality of the data stored in a cloud system. In which, protecting the personal identity information is highly difficult due to the distributed data storage [10]. To avoid this privacy concern, the anonymization mechanism is developed for hiding the personal data. Then, the key generation and distribution plays an essential role and it’s also a big security issue [11]. So, a dynamic authentication protocols are required to ensure the secure key generation and sharing. In a cloud system, a secure audit must be performed to extract the privacy policies while data access. In order to process the big data in a secured way, different security mechanisms generate the rules for defining the policies to restrict the data access. Typically, the big data requires additional requirements during data collection, analysis, data transferring, and storing. Thus, ensuring the privacy, security and confidentiality of the big data is more essential in many application areas.

1.2. Objectives

This paper focuses on the following objectives:

- To estimate the support and confidence values for generating the rules, an Improved FP-Growth (IFP-G) based rule generation technique is developed.
- To classify the sensitive and non-sensitive attributes based on the extracted features, an Enhanced Random Forest (ERF) classification technique is deployed.
- To anonymize the sensitive data for ensuring the privacy, a Modified Incognito Anonymization based Privacy Preservation (MIA-PP) algorithm is introduced.

1.3. Organization

The remaining sections present in the paper are formatted as follows: the standard algorithms and methodologies related to cloud data security and privacy preservation are surveyed in Section II. The working procedure of the proposed methodology is presented with its flow illustration in Section III. The performance results of the security mechanisms are evaluated and compared by using various performance measures in Section IV. Finally, the paper is summarized with the outcomes, and the future enhancement are stated in Section V.

2. Literature review

Ferrag, et al [12] surveyed various privacy preservation mechanisms for improving the security of smart grid communications. The techniques have been surveyed under the following categories: advanced metering infrastructures, data aggregation, marketing architecture, and smart community. Moreover, different types of attacks were classified by using various game theoretic approaches and privacy countermeasures. The attacks discussed in this study were key based, data based, impersonation based, and physical based attacks. Liu, et al [13] suggested some cryptographic solutions for outsourcing the big data in a secured manner. The measures such as privacy and security were considered as two key concerns in big data processing and analysis. The cryptographic mechanisms investigated in this paper were Functional Encryption (FE), Searchable Encryption (SE), and Order Preserving Encryption (OPE). Here, it was stated that the Fully Homomorphic Encryption (FHE) provided an efficient results with reduced complex mathematical structure. Anjum, et al [14] developed a trust based hybrid data privacy preservation approach, namely, MIDR angelization for ensuring the privacy of the patient’s health information. Here, the sensitive attributes were specified for reducing the maximum instance disclosure risk and average instance disclosure risk factors. Then, the random sampling was applied to reduce the instance disclosure risk of the micro-data. Moreover, the privacy risk assessment test was conducted to analyze the performance of the cloud system.

Anisetar, et al [15] balanced the quality of life and privacy for developing public health policies in smart cities. The contributions have been mainly focused on this work were as follows: the traditional policy making process was improved, and the privacy aware big data campaign was defined. The health policy making process includes the steps of situation analysis, action plan, implementation, evaluation, and monitoring. During data partitioning, the partitioning of retrospective data was defined by using the big data platform. Then, a set of relevant features were selected with respect to the policy goal. Jiang, et al [16] introduced a high performance and Privacy Preserving Query (P2Q) for ensuring the privacy of the multidimensional Big metering data. Moreover, an enhanced Ciphertext Policy Attribute based Encryption (CP-ABE) mechanism for providing control access on the search results. Then, the data confidentiality was attained and data owner’s privacy was preserved by using the P2Q scheme in a semi trusted cloud. However, it required to fine tune the system by implementing an advanced scheme. Quazziani and Bhakkali [17] utilized a k-anonymity mechanism for increasing the privacy of the Bid data. It incorporates the functionalities of both randomization and generalization approaches. In this mechanism, a set of quantifiers have been utilized for anonymization. Also, this technique intended to maintain both the data utility and privacy with reduced computational complexity. But, it failed to process the sensitive attributes for the anonymization of a set of data composed by the use of quasi identifier. Guan, et al [18] employed a privacy preserving data aggregation mechanism for increasing the privacy of data without compromising the privacy. Also, the substitution strategy was used to perform the fault
tolerance by relating the user’s data in other groups. This system includes the following entities: Control Center (CC), Key Initialization Center (KIC), Data Aggregation (DA), and residential users. Then, the Paillier encryption algorithm was used to encrypt the data based on the homomorphism properties. However, it does not use the real time data for evaluating the performance of the scheme.

Li, et al [19] suggested a distributed authentication and authorization mechanism for reducing the traffic overhead in a distributed shared data. In this paper, an Identity Based Cryptography (IBC) mechanism was used to verify the identities of the publishers and users. Then, the secure registration and efficient authentication mechanisms were utilized to develop a trustworthy registration framework. Moreover, this mechanism provided the following properties: big data integrity, trust, identity verification, symmetric key establishment, automatic attribute updation and flexible authorization. The performance results confirmed that the suggested mechanism efficiently reduced the bandwidth cost for both authentication and attribute retrieval processes. Li, et al [20] introduced a new algorithm, namely, Composite modular Exponentiation (CExp) for ensuring the secure outsourcing of the data. Here, the exponentiation operation was performed only a single server, which eliminates the potential of a collusion attack. Then, the mathematical division operation was performed for hiding the base and exponent of the outsourced data. This mechanism has the ability to detect the misbehavior of the server in a secured and efficient way.

Wu, et al [21] used a Scalable Privacy Preserving Big Data Aggregation (Sca-PBDA) mechanism for increasing the security of big data. In this scheme, both the inter cluster and intra cluster aggregation mechanisms have been utilized to reduce the energy consumption. Here, a network topology was determined by using an Agradient based equal network clustering algorithm, which estimated the node energy consumption. Then, the data aggregation algorithm was employed to ensure the privacy preserving data configuration and aggregation. This system has an increased computational complexity, which must be reduced by implementing an efficient technique. Terzi, et al [22] surveyed various privacy and security issues in big data processing. The mechanisms have been investigated in this work were, Hadoop security, monitoring, auditing, anonymization, cloud security and key management. Liang, et al [23] introduced a cipher text multi-sharing protocol for guaranteeing the confidentiality of the Big data. This mechanism integrated the benefits of proxy re-encryption and anonymous techniques for secure big data storage service. This cipher text mechanism intended to solve the following properties: anonymity, multiple receiver update and conditional sharing. Moreover, a multi-hop identity based conditional proxy re-encryption mechanism was utilized to preserve the anonymity of the sender and receiver.

Xu, et al [24] analyzed the privacy preserving machine learning algorithms for efficiently discovering the valuable knowledge and hidden information of the big data systems. Here, a MapReduce framework was utilized to conceal the training data. Then, the linear Support Vector Machine (SVM) technique was used to separate the hyper-plane between two classes of data. Moreover, a private preserving machine learning technique to disclose the sensitive information. Yu, et al [25] reviewed the data clustering and privacy frameworks for investigating the challenges and opportunities of the Big data storage system. The entities involved in this system were data generator, data curator, data user, and data attacker. Also, it performs the major operations such as collection, anonymization, and communication. The authors stated that privacy preservation was one of the important considerations in big data processing. Samuel, et al [26] developed a hybrid approach for analyzing the privacy requirements of the multimedia big data. In this framework, only the user had the access to compose the conflict free policies on the online multimedia data. Then, a new methodology was presented to assure both the logical consistency and correctness of the data. Furthermore, a new prototype, namely, Intelligent Privacy Manager (IPM) was implemented to share the multimedia big data in a secured way.

Based on the above investigations, it is analyzed that the existing security and privacy preservation frameworks contain both the advantages and disadvantages. Still, it lacks with the following limitations:

- Unavailability of the analytical models
- Increased computational complexity
- Limits with the data freshness
- Requires more storage and increased searching complexity

To solve these problems, this paper aims to develop a new privacy preservation mechanism by implementing an enhanced methodologies.

3. Proposed methodology

In this division, the working methodology of the proposed privacy preservation framework for big data processing is explained, and its flow illustration is depicted in Fig 1. The theme of this paper is to develop an enhanced privacy preservation framework by deploying some advanced techniques. This system incorporates the following stages:

- Data preprocessing
- Rule based feature extraction
- Sensitive and non-sensitive data classification
- Privacy preservation

3.1. Data preprocessing

At first, the input dataset is preprocessed for eliminating the irrelevant attributes and normalizing the data. Due to the variety of data sources, the collected dataset may contain the noise, redundancy and consistency, which must be eliminated before processing the data. Here, the data cleaning is also performed for improving the quality of data, which includes the following processes:

- The error types are determined and defined
- Error searching and identification
- Error corrections
- Modification of data entry procedures
3.2. Rule based feature extraction

After preprocessing the data, an Improved FP Growth (IPF-G) technique is implemented to extract the features. Based on these, the rule is generated that restricts the access on the data stored in cloud. Typically, IPF-G is a scalable and efficient algorithm that mines the frequent patterns based on the prefix tree structure. Then, it stores the compressed and crucial information for improving the accuracy analysis. Moreover, it has the ability to identify the frequent item-sets without candidate generation. In this algorithm, the support and confidence values are extracted, in which the minimum support value is used to construct the tree in an iterative manner. The major benefit of using this technique is to reduce the searching cost by identifying the shorter patterns, which are further concatenated for providing good selectivity. Due to the repetitive scanning of database and storage space, it is very difficult to process the Big data with the use of Apriori algorithm. Thus, the FP growth technique has been utilized in this work, and also it consumes less time compared than the Apriori technique.

In this stage, the preprocessed bank dataset is taken as the input, where the data array matrix is constructed for identifying the high sensitive attributes. Then, the redundancy can be avoided by performing the hash mapping. Then for each frequent itemset, the item consequence of the rule is generated. After that, the length of the bank dataset is estimated, if it is not equal to null, the minimum support value is computed and compared with the Apriori rule. If it is greater than less than the Apriori rule, the highly sensitive score is estimated; otherwise, low sensitive score is estimated and assigned as the overall sensitive score. Based on this score value, the sensitive and non-sensitive attributes are classified.

Algorithm I – Improved FP Growth based Rule Generation
Input: BkData ← Bank Dataset, MaxRand[ ] = {rand (max value )}
MinRand[ ] = {min value }, Msup =0.2;
Output: Fcont[ ] ← Variable (Integer) Conversion;
Where, Fcont[ ] ← FP Variable
ARule[ ] ← Apriori Rule
Hsen[ ] ← High sensitivity score;
Sence[ ] ← Overall sensitivity score
Lsen[ ] ← Low sensitivity score
CFvalue[ ] ← Confident value
Svalue[ ] ←Confident value
Step 1: ini i = 0;
Step 2: D Rand[i] = BkData; // D Rand[ ] ← Data Reading Array
Step 3: if (D Rand[i] equals (Hsup));
//Where, Hsup[ ] ← High sensitivity attribute
Step 4: Replace D Rand[i] =MinRand[ ]; i++;
Step 5: else
Step 6: D Rand[i] = MaxRand[ ]; i++;
Step 7: Fcont[i] = (D Rand[i] //
Step 8: break;
Step 9: for k=0 to Fcont[length ];
Step 10: Rd[k] = Fcont[k]; k++;
Step 11: Hashmap[ ]= 0(k, k= Rd); //Hash mapping to avoid redundancy content
Step 12: For each frequent k-itemset f/k, k≥ 2 do;
Step 13: H = { i | i ∈ f/k } {1-item consequence of the Rule}//m++;
Step 14: ARule[m] = Call Apriori-genrule (f/k, H )
Step 16: end for
Step 17: while (BkData. length! =null) i=0;
Step 18: if (BkData contain ARule[i])
Step 19: i++;
Step 20: if (ARule[i] ≥ Msup)PRINT;
Step 21: H = Sense[ ];
Step 21: else
Step 21: Lsen[ ] = ARule[i];
Step 22: Sense[ ] = Lsen[ ];
Step 23: end if;
Step 24: end if;

3.3. Sensitive and non-sensitive attributes classification

After extracting the features, an Enhanced Random Forest (ERF) classification technique is implemented to classify the sensitive and non-sensitive attributes based on the extracted features. This classifier is constructed by using various decision trees, which provides more accurate classification results. It uses a random method for building the forest that is composed by using a set of decision trees. The major reason of using this technique is, it has ability to process the large scale datasets and the high dimensional data. Furthermore, it solves the de- anonymization problem by providing the robust solution to the decision making. In this stage, both the sensitive and non-sensitive data are taken as the input for classification, where the patent node and child node are initialized at first. If the data is null, it can be assigned to the child node; otherwise, it can be assigned to the patent node. Based on this, the decision tree is constructed, then k number of features are considered and picked at randomly. After that, the out of bag error can be reduced by pruning the tree. These processes are performed until classifying all the sensitive and non-sensitive data attributes.

Algorithm II – Random Forest Classification
Input: Take Sense and Non-sense Data Sense[ ];
Initialize Fnode[m] ← Patent Node; Cnode[m] ← Child Node,
Tnode[m] ← Tree Node
Output: Rf[ ] // Random Classification Records
Step 1: ini i=Sense[ ]; length; l=0; ini m=0;
Step 2: while (l< i)
Step 3: if (Sense[ ] == 0)
Step 4: Cnode[m] = Sense[ ];
Step 5: else
Step 6: Pnode[m] = Sense[ ];
Step 7: l++;
Step 8: Break;
Step 9: for i to T do
Step 10: Draw n point = Sense[ ]; with replacement from Pnode[m] center node
Build full decision/regression tree on Dl;
Split consider k-features, picked uniformly at random;
Replacement from Cnode[m] center node to Sense[ ];
Step 11: else: Prune tree to minimize out-of-bag error
Step 12: End for;
Step 13: Average all T trees
Step 14: End procedure;

3.4. Modified incognito anonymization-based privacy preservation

After classification, a Modified Incognito Anonymization Based Privacy Preservation (MIA-PP) technique is developed to generate the anonymized data. Generally, the data anonymization is considered as the widely accepted method for privacy preservation. It is mainly performed to prevent the sensitive data for mitigating the unidentified risk. Moreover, the privacy preservation is provided at the cloud user side, where the user’s personal information and other information details are protected against the unauthorized access. In the proposed system, the user’s banking information is protected by applying anonymization. For this reason, the MIA-PP
mechanism is implemented, which guarantees the privacy and confidentiality for the user’s data. In this stage, the node data, generalization rule, privacy parameter and utility parameter are taken as the input. At first, the hierarchical lattice is created for all the possible generalization cases, where the degree of generalization is higher than the child node. For each node in the tree, the generalization process is performed between the parent node and child node. Then, the list of classes in the generalized tree are estimated, based on this the grouping can be performed for generating the anonymized data and catalog for the counter field record.

Algorithm III – Privacy Preservation using Modified Incognito Anonymization

Input: Node Data \( N_{data}\), Generalization Rule G, Privacy Parameter \( P \), Utility parameter \( h \)

Output: Anonymized Data \( A{T} \), Catalog for counter field record \( C \)

Step 1: Create hierarchical lattice \( P_{node} \) for all possible generalization cases, except for the cases where the degree of generalization is more than \( C_{node} \)

Step 2: \( m = \text{Maximum value of } B_k \)

Step 3: for each node \( n_i \in h \) do

Step 4: \( T = \text{generalization ( } P_{node}, C_{node} \) )

Step 5: \( E_{m1} = \text{list of equivalent class in } T \)

Step 6: if \( |E_m| < k \) then

Step 7: Add Counterfield\( (E_j, \text{Temp c}) \)

Step 10: end;

Step 11: end;

Step 12: end;

Step 13: end;

Step 14: \( C_{eq} = \text{grouping ( } T, \text{Temp c} \) )

Step 15: Result=Calculate \( (T^*, \text{Temp C}) \)

Step 16: if \( \text{min } > \text{result } \& \& \text{result } \neq \text{null} \) then

Step 17: \( AT^* = T^* \)

Step 18: \( \text{min } = \text{result} \)

Step 19: end;

Step 20: end;

Step 21: return \( AT \) and \( C \);

4. Performance analysis

In this sub-section, the performance results of the existing and proposed techniques are analyzed and compared by using various measures. It includes ranking precision, accuracy, execution, anonymization time, data analysis and error rate. Also, different existing techniques have been considered to demonstrate the effectiveness of the proposed system.

4.1. Ranking precision

The ranking precision is the mostly used measure for evaluating the performance of the cloud security mechanisms. It is defined as the positive predictive value that increases the efficiency of classification for ensuring the security. In Figure 2, the precision of the proposed security mechanism is evaluated with respect to varying dataset size (KB) and recall values (i.e. 20% and 40%). This analysis stated that the precision of the proposed technique with 20% recall is greater than the 40% recall value. Moreover, the performance of the big data storage system is highly depending on the measure of ranking precision.

4.2. Accuracy

Accuracy is defined as the overall correctness of the model that is calculated based on the sum of the correct classification divided by the total number of classifications.

\[
\text{Accuracy} = \frac{(TN+TP)}{(TN+TP+FN+FP)}
\]  

Where, TP – True Positive, TN – True Negative, FP – False Positive, FN – False Negative. In Figure 3, the accuracy of the existing and proposed classification techniques are validated with respect to varying dataset size (KB). The classification techniques considered in this evaluation are Bayes Net, AIRS, Support Vector Machine, C4.5, and CBA. This results stated that the proposed ERF provides the increased accuracy compared than the other techniques. Because, the ERF classifies the sensitive and non-sensitive attributes based on the set of extracted rules, which leads to increased classification efficiency.
4.3. Execution time

Execution time is defined as the amount of time required to execute the overall detection and prediction process, which is measured in terms of seconds. Figure 4 evaluates the execution time of the existing and proposed classification techniques with respect to varying file size (KB). The time complexity of the security and data storage system is determined based on the performance of rule generation and classification techniques. The classification techniques considered in this evaluation are Bayes Net, AIRS, Support Vector Machine, C4.5, and CBA. The evaluation results stated that the execution time of the proposed ERF technique consumes less time, when compared to the other classification techniques. Because, it takes the decision on classification by forming the tree with the set of decision making nodes. In which, each node has the ability validate the correctness of the data.

\[
F1 - score = \frac{(1+b^2) \times \text{recall} \times \text{precision}}{b^2 \times \text{recall} + \text{precision}}
\]

\[
G - \text{mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TP}{TN+FP}}
\]

4.4. Data analysis time comparison

Table 2 shows the data analysis of existing and proposed techniques, which includes the measures of time, F1 score, and G-mean value. In which, the F1 score is mainly used to analyze the efficiency of the classification model. Then, the measure of G-mean is estimated with respect to the geometric mean of precision and recall. These two measures are considered as the evaluation predictors for the ensemble of classification, and it is calculated as follows:

The techniques have been considered in this evaluation are Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), LR-SMOTE, RF-SMOTE and SVM-SMOTE. Due to the imbalance distribution of the user features, the existing classification techniques are failed to identify the decision boundary. So, the proposed ERF provides an efficient results compared than the other classification algorithms.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time(s)</th>
<th>F1</th>
<th>G-mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>58.7</td>
<td>0.22</td>
<td>0.24</td>
</tr>
<tr>
<td>RF</td>
<td>41.6</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>SVM</td>
<td>66.23</td>
<td>0.18</td>
<td>0.21</td>
</tr>
<tr>
<td>LR-SMOTE</td>
<td>102.1</td>
<td>0.25</td>
<td>0.28</td>
</tr>
<tr>
<td>RF-SMOTE</td>
<td>67.9</td>
<td>0.28</td>
<td>0.26</td>
</tr>
<tr>
<td>SVM-SMOTE</td>
<td>140.5</td>
<td>0.22</td>
<td>0.25</td>
</tr>
<tr>
<td>Proposed ERF</td>
<td>38.56</td>
<td>0.16</td>
<td>0.24</td>
</tr>
</tbody>
</table>
4.5. Anonymization time analysis

The time cost analysis of the anonymization process is estimated for the existing and proposed privacy preservation techniques with respect to different item sets as shown in Table 3. The techniques considered in this evaluation are traditional anonymization, ARH and proposed MIA-PP. Due to the increase in the minimum support value, the proposed MIA-PP provides the reduced time consumption compared than the other techniques.

![Table 3: Anonymization Time Analysis](image)

4.6. Error rate

The measure of error rate is used to evaluate the efficiency of the big data storage system. The intention of the big data storage system is to reduce the error rate by providing the more relevant results with respect to the user query. In Figure 5, the error rate is estimated for the existing and proposed privacy preservation mechanisms with respect to various item sets. The techniques have been considered in this analysis are k-anonymity and Association Rule Hiding (ARH). This evaluation results stated that the proposed MIA-PP provides reduced error rate compared than the other mechanisms.

![Fig. 5: Error Rate.](image)

4.7. Sensitive score level

Table 4 shows the privacy score generation level of the proposed privacy preservation mechanism with respect to various client IDs. Here, the sensitive level of 1 indicates the high score, and 0 indicates the low score. Also, the privacy score rate is increased up to 0.98 for the cloud users. It states that the proposed preservation mechanism provides an increased privacy for the user’s data. Then, its graphical illustration is provided in Figure 6, where the privacy score is generated with respect to different client IDs.

![Table 4: Sensitivity Score Level](image)

![Fig. 6: Privacy Score.](image)
5. Conclusion and future work

This research paper develops a new privacy mechanism by deploying various security measures. This system comprises the working stages of data preprocessing, rule generation, classification, and privacy preservation. At the initial stage, the given bank dataset is preprocessed by eliminating the unwanted and irrelevant attributes, which increases both the data correctness and quality. After that, the association rules generated for enabling a restricted access on the data, which is done by using an IFP-G algorithm. Based on the extracted rules, the classification technique, named as, ERF classifies the sensitive and non-sensitive data attributes from the dataset. Because, the sensitive data must be protected from the unauthorized access. Finally, the anonymization is provided for both the user’s personal information and other banking details, which is performed by the use of MIA-PP. These enhanced mechanisms provides the security to big data by assuring both the privacy and confidentiality. The performance of these techniques are validated with the use of several measures such as accuracy, ranking precision, execution time, anonymization time, and error rate. The evaluation stated that the proposed techniques outperforms the other techniques by efficiently generating the rules for restricting the access.

References