IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 12 (December. 2018), ||V (II) || PP 17-25

A Novel Improvised Anonymization with Map Reduce Framework for Big Data Privacy Preservation in Cloud

Priyadarshini. D[1], Jamsheena Nellisseri[2]
Assistant Professor, Department Of Computer Science, Sree Narayana Guru College, K.G.Chavady, Coimbatore, India[1]
M.Phil Scholar, Department Of Computer Science, Sree Narayana Guru College, K.G.Chavady, Coimbatore, India[2.]
Corresponding Author: Priyadarshini. D

Abstract: Cloud computing provides promising scalable dynamic infrastructure to maintain various processing of a variety of big data real time applications in sectors such as healthcare and business and other area. Data sets like electronic health records in such applications often contain privacy-sensitive information, which brings about privacy concerns potentially shared to third-parties in cloud. Privacy-preserving data mining (PPDM) refers seeks to preserve sensitive information from unsolicited or unsanctioned disclosure. Protecting the individual privacy within the sensitive data is becoming a prominent research focus data in big data environment. Local-Recoding Anonymization plays major role in privacy preservation. Data anonymization refers to hiding identity of sensitive data so that the privacy of an individual is effectively preserved. Most of the existing privacy model fails to maintain effective privacy Preservation due to their insufficiency or poor scalability. The system proposes an innovative framework to effectively preserved privacy of an individual and to keep published data secret. The proposed system aims at achieve highly scalable privacy. The proposed system aims at achieve highly scalable privacy through the combination of Improvised k-anonymity along with novel SBDT model. It preserves better data utility than generalization. This aims provide better anonymize along with Minimizing the information loss.

Keywords: Big data analytics, Machine Learning, Healthcare, Disease Detection, Medical Data Analytics.

Date of Submission: 29-11-2018 Date of acceptance: 13-12-2018

I. INTRODUCTION

The Cloud computing and the big data is most preference stage on current trend. Most common and widely used this exclusive concepts is move on current information technology industry and problem solving or researching communities are like that. However, nowadays a huge number of big data based applications and based services are transfer or moved into the cloud for data mining, that is big data focus the cloud for improve the data mining area oriented processing and sharing. The best appearance of cloud computing is, (i) high scalability, (ii) easy to accessing capability, it is based on payment, if make the design of fashion then easily access with cheap rate. Various numbers of organizations is grouped and it access through public cloud infrastructure. The cloud oriented infrastructure and its methodologies below description is clearly expressed. The big data used data set is which contains its applications oriented personal privacy-sensitive data such as, (a) electronic health records, (b) financial transaction records. The electronic data is not same as normal data, because the normal is only viewed they cannot modified that is, cannot pass the query, but the electronic data is easily move on the different types of data types such as, (i) easily apply the query and its constraints, (ii) easily search that information is very quickly, (iii) easy to express the clear data ways like, (a) Bar Chart, (b) Histogram, (c) Line Chart, (d) pie Chart, (e) Area Chart, etc., (iv) easy to transfer because use various encoding techniques are easily apply particularly, privacy protection that is, authentication related purposes; To analyze these type of data sets is provides the decision making knowledge that is similar KDD, and into a number of small areas of the society oriented applications, are (i) healthcare, (ii) medical, (iii) government services, (iv) e-research. This information is how to get, which means it first with coupled the big data and with public cloud environments disabled, so some basic traditional privacy protection measures in the cloud area, because every privacy protection is based on traditional privacy protection methods, some more variations occur in the developed techniques but the trackers know the traditional privacy protection ways so some logics are apply to get the actual content easily and recoding easily. So this is considers the problems are, (i) economic data loss, (ii) Difficult social data losses to data owners. These data loss is come from shared and released data access by the third-parties, so need the robust privacy preservation. The data anonymization is a technology and it is

widely accepts the potential data privacy protection. The privacy preservation is applied into the non-interactive data sharing and releasing data, so the data anonymization is to take the correct action and it based on the privacy oriented tasks. This term is first hiding the identity or the sensitive data, so the individual information is effectively protected, then it will moved into aggregate functions or aggregate data, it exposed into the data users. This aggregated information is used into different analysis and many mining tasks, also.

II. PROBLEM DEFINITION

In this paper [1] proposed l-diversity: Privacy Beyond k-anonymity. In this system mainly focus two problems, (i) the attacker attacks the sensitive values so little diversity is occurred. This technique is first show that the attacker and discover the little diverse sensitive attributes. (ii) Then, the second one is show that the attacker using background knowledge. The l-diversity is simply describes in this system is, first finds and show the attacked sensitive data and then secondly finds the attacker knowledge that is the sensitive data is attacked by using which knowledge use the attacker this means. These two problems are solved by using novel and powerful privacy preserved concept with definition included term called, "l-diversity".

In this paper has [3] proposed the concept is, "Anonymization by Local Recoding in Data with Attribute Hierarchical Taxonomies". In this system concept is mainly focused on the individual privacy, which means the published data set is not proper de-identification. De-identification means prevent the data set identification from related with information. The k-anonymity is a technique to de-identify the give data set. This concept is only focus on two major issues in local-recoding with k-anonymization in attribute hierarchical taxonomies. The issues are: (i) proper distance metric and achieve local-recoding generalization with little distortion, (ii) to control the inconsistency generalized attribute domain by local-recoding. This system gives the higher quality of k-anonymity tables. This table quality is measured by three terms are: (a) global-recoding anonymization method, (b) incognito, (c) multidimensional-recoding anonymization method. This system considers the drawback is large volume of data is not match this concept.

In the paper has [4] proposed the useful concept is "utility-based anonymization using local recoding". In this concept use the local-recoding concept it simply and easily describes the global recoding difficulties. In the paper describes the global-recoding is maps the domains, this domains includes the quasi-identifier attributes. The global-recoding is changed the values or generalized the data by using the identifier attribute namely, 'quasi'. The globally recoded is successfully run and performed easily but may not achieve the effective and efficient anonymization. So this problem is overcome the local-recoding technique. The global-recoding use the anonymized data but this technique is use the categorized data. The local-recoding is simple framework and it specified the utility of data, and it covers the numerical and categorical data. Then, additionally covers the method namely, 'heuristic local-recoding'. This method is used for utility based anonymization. This system takes the theme like, global-recoding or constrained local recoding technique, but overcome this technique based difficulties. This system uses mostly take the k-anonymity privacy preservation model by using quasi-identifier. It follows the bottom-up and top-down approaches from greedy methods. Form the cluster (tuples) and apply these two greedy methods.

In the paper has [5] proposed the privacy preservation concept is, "Efficient k-Anonymization Using Clustering Techniques". In this system is mainly focus on the k-anonymization method. This method requires the anonymized data at same time it minimized the data loss, this type of result are get from the data modification time. In this system is develops the k-anonymization method by using the clustering technique. The major merit of this system is good data quality. This system is gives the best solution for k-member clustering problem. This problem is comes from NP-Hard presented greedy heuristic method, it gives the difficulty is considered that is in $O(n^2)$. This system is estimates the information loss by using new introducing concept generalization. This concept is applied both types of data, that is numerical and categorical data. The k-anonymization is called, k-anonymity. This concept uses the attributes, so the set of attributes is called quasi-identifier.

In this paper has [6] proposed the concept, "Mondrian Multidimensional K-anonymity". In the system is give the privacy preservation into microdata publishing and using the recoding technique and it achieves the k-anonymity. The proposed system is provides the flexibility and only use multidimensional model. The multidimensional model is includes both global and local recoding schemes. This concept is introduced by the simple greedy approximation algorithm. The microdata is published various number of organizations, so this system gives the privacy preservation by removing the identifiers. The k-anonymity is reduces the risk of attack, because the main aim of k-anonymization is firstly, give the individual privacy.

In the Paper has [8] proposed the "(α, k)-Anonymity: an enhanced k-anonymity model for privacy preserving data publishing". This system is give the protection to both types of sensitive data, the protection is related to that data identification and relationships. This concept is overcomes the NP-hard problem. In this system use the recoding schemes like, global and local; first apply the global-recoding scheme then, secondly

apply the local-recoding scheme to gets the scalability and less distortion. The global-recoding scheme is depends on the monotonicity property, this property is holds the generalized (α, k)-Anonymity.

III. PROPOSED SYSTEM

The chapter discusses about the proposed methodology and the steps involved in this proposed system. The system proposes a new effective privacy Preservation approach, which concentrates on the effective Improvised k-anonymity along with novel SBDT model. This chapter discuss about the algorithms and methodologies.

- The system implements a new Improvised k-anonymity algorithm for effectively complete anonymization. The system introduces a new novel SBDT model to perform the sequential clustering for improving the accuracy in anonymization. .

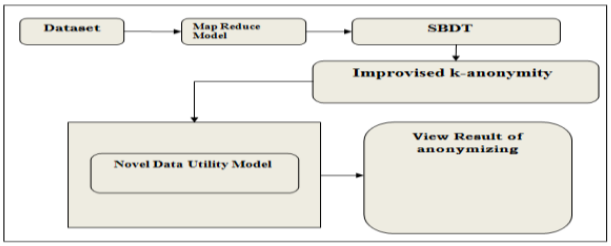


Fig 3.1 Proposed system architecture

Contribution of The Proposed Work

The followings are the contributions of the proposed system.

- The existing Data anonymization is difficult in handling of large datasets in cloud applications .It is very challenged to achieve privacy preservation techniques and insufficiency of scalability to address this problem, the system introduces an Improvised k-anonymity with SBDT Model.
- A Hadoop Map Reduce framework proposes to quantify the information Loss during the anonymization process.
- An Improvised k-anonymity process has been applied for the effective protection techniques to preserve the data integrity effectively.
- SBDT (Sequential Background Data Transition) unit has been applied this is a new probabilistic defense technique based on random probability divergence.
- Novel Data Utility Model been used for Measure of information loss or loss in the functionality of data in providing the results.

IV. METHODOLOGIES

Through an experimental evaluation on a large data set, that shows the effectiveness of the defense under different methods used to extract background knowledge; the results also show that the proposed algorithm provides a very good tradeoff between privacy and data utility. The list of following methodology explains in this chapter.

1. Hadoop Map Reduce framework
2. SBDT Model
3. Improvised k-anonymity
4. Novel Data Utility Model

Hadoop Map Reduce framework

Map Reduce is one of the programming models and software framework. Apache Hadoop is an open-source framework. It allows to store and process big data in a distributed environment across clusters of computers using simple programming models. Map Reduce is the core component for data processing in Hadoop framework. In layman's term Map reduce helps to split the input data set into a number of parts and run a program on all data parts parallel at once.

Map Reduce program works in two phases.

1. Map phase – This converts the incoming data into key and value.
2. Reduce phase - Key / Value pairs provided to reduce are sorted by key

In between Map and Reduce, there is small phase called Shuffle and Sort in Map Reduce.

Steps for Map Reduce
Step 1: Input Splits:
Input to a Map Reduce job is divided into fixed-size pieces called input splits Input split is a chunk of the input that is consumed by a single map.

Step 2: Mapping
The mapping is the first phase. It executes the execution of map-reduce program. The first phase data in each split is passed to a mapping function to produce output values. The following example shows the first phase process, job of mapping phase is to count number of occurrences of each word from input splits (more details about input-split is given below) and then prepares a list in the form of <*word, frequency*>.

Step 3: Shuffling
This phase consumes output of mapping phase. Its process is to consolidate the relevant records from Mapping phase output. The following example is describes the second phase process level, same words are clubed together along with their respective frequency.

Step 4: Reducing
The phase reducing is takes the output values from the "Shuffling phase" are aggregated. The thrid phase combines values from Shuffling phase and returns a single output value. Finally short out, this phase summarizes the complete dataset then it calculates total occurrences of each word.

ALGORITHM 1: Map Reduce

Map Step: First give the input dataset and is stored on HDFS in the format of < key; value > pairs, each of which represents a record in the data set. The key is combination of the quasi-identifiers and the value is content of the tuple. The data set is split and broadcasted to all mappers.

Input: (key1: quasi-identifiers; value1: text of a record)
Output: key2: a string representing a cell, value2: the value in current dimension
Parse the string value;
Set string out keys and out Value as null;
Key set of quasi-identifiers;
Value value in current dimension;
OutKey sorted key based on quasi-identifiers;
Out Value data [current Dimension];
Output (outKey, out Value);

Reduce Step: The input of the reduce function is the result obtained from the map function as output. Since the objective of this Map Reduce programme is to sort the input data, the Reduce function does only operate as a combiner.

Input: (key2: a string representing a cell, value2: the value in current dimension)
Output: key3: text, value3: the value in current dimension
Out Key sorted key based on quasi-identifiers;
Out Value data [current Dimension];
Output (out Key, out Value);

SBDT

The proposed rule based SBDT modal contains three portions which enhanced the existing anonymization modal with background, sequential knowledge gathering and rule forming methods. Although this approach is simple, it requires the computation of all the generalized tables. Most of the computation is avoids, the concept of distance vector between tuples is introduced and exploited. Let T be a table and x; y 2 T be two tuples such that x = hv to the domain at which they generalize to the same value vi). Initially the SBDT performs the sequential clustering for improving the accuracy in anonymization.

Algorithm for SBDT:
Input: History of original views Hr=(V1,….Vn) a sequence of sensitive values seq, and a sensitive value s.
Output: the conditional probability p(s|seq), ehich corresponds to the frequency of sequence in Hr.
Step 1: begin with the original history and seq,s

Step 2: for h=1 to r do
Step 3: for all respondent u of a tuple in Vh do
Step 4: for j=h to 1 do
Step 5: seq_j=seq.of past j sensitive values of u in Hh.
Step 6: seq_j.numocc= seq_j.numocc + 1
Step 7: End
Step 8: if(seq.numocc==0) then return()
Step 9: else
Step 10: sequence=(seq,s)
Step 11: return sequence.numocc/seq.numocc
Step 12: end

Improvised k-anonymity
Initially unique class IDs are assigned in a specific fashion. Then the vertical partitioning of the table is performed to split the table into two, of which one will contain sensitive data and its class ID and the other containing non-sensitive data and its class ID. Then finally cross join is performed on the vertically partitioned tables to obtain the privacy preserved anonymous table.

Improvised k-anonymity Steps

Input: Raw cleaned dataset.
Output: Anonymized dataset.
Algorithm
Step 1: Input dataset with n no. of fields
Step 2: Separate sensitive and non sensitive fields //vertical partitioning Store in two different tables t1 and t2;
Step 3: Anonymize the sensitive fields to (k-1) records each;
Step 4: Update table1;
Step 5: Assign class_id to each record in t1 and t2;
Step 6: sort records with class id as the quasi-identifiers;
Step 7: Join both table's t1 and t2 by join function;
Step 8: Store in a new table t3;
Step 9: Update table t3 as output table.

Novel Data Utility Model

Novel Data Utility Model been used for Measure of information loss or loss in the functionality of data in providing the results. This implementation steps are list out below.

Algorithm: Data Utility Measurement

Input: ds0: Dataset Initial
pAlg: Perturbation Algorithm
NumIter: Number of iterations
Output: MCRi: Metric Change Rate for data set
CMCR: Cumulative Metric Change Rate
Steps:
Step 1: CMCR: = 0; //assign the value
Step 2: select Metrics:= {m1, m2, m3, …};
Step 3: for mi: = 1 to | selMetricss | do metricValue0:= evalueteMetric (mi, ds0);
Step 4: dsnew: = ApplyPerturbarion (ds0, pAlg); //assign the count values
Step 5: metric Value [selGraphMetrics [mi]] := evalueteMetric(mi,dsnew);
Step 6: computeMetricChangeRate (metricValue0, metricValuenew);
cmtr := cmtr + MCRi;
/* Cumulative Metric Change Rate Average */
Step 7: CMCR: = CMCR / Output: CMCR;
End;

V. RESULT AND ANALYSIS

Data Sets

In the experiments, the system use patient Health Records data set from UCI repository. Dynamic datasets have been used to evaluate the proposal. The proposed experiment shows the difference between the existing systems such as two-phase clustering approach. Anonymized result has been created with the new set of data which is ready to publish the data. The patient Health Records has been created and that were used in the experiments. The first real data set used in the experiments is the PHR (patient Health Record and Patient Personal Record) set from the data repository, which has in some ways become the data of measuring the algorithms. This data set has 8 attributes (Age, name, Gender, address, postal code, disease, phon and reports the actual census data. The data set can contain any number of tuples. the SQL Database server has been used for the data storage.

Dataset Description:

Dataset Name: Patient Health Record
URL: http:\\www.UCIrepostory.org\HealthRecord

Name	Age	Gender	Address	Phone	Pincode	Disease
anu	45	female	coimbatore	7778889994	641012	Cancer
karthick	34	male	coimbatore	8889997744	641020	Flu
priya	34	male	coimbatorenorth	9998887776	641012	cancer
Tamil	34	male	coimbatore	8889997774	641010	diabeties
john	45	female	coimbatore	7778889994	601012	Cancer
hema	34	female	coimbatore	7676999774	641020	Flu
lilli	48	female	coimbatorenorth	8888887776	631012	cancer
jasmine	52	female	coimbatore	9995557774	621010	diabeties

Experiment Process

The data set has 7 attributes considered are: Name, Age, Gender, Address, Phone and Pin code along with Disease. There are a total of 150 patient records in the database.

Experimental Results

This section describes the implementation process. Implementation is the realization of an application, or execution of plan, idea, model, design of a research. This section explains the software, datasets and modules which are used to develop the research. The algorithms are implemented in Java and are run under Windows operating platform.

The implementation considered *Age*, gender, pin code, and disease as the set of key attributes. We applied generalization for the key attributes using the generalization domains as described in the above table.

We use the following notations:

- *Ai* (*i* = 0, 1, 2, 3) for the domain generalization of *Age*;
- G*j* (*j* =0) for gender;
- P*k* (*k* = 0, 1, 2, 3) for *pin code*;
- *Dp* (*p* = 0, 1, 2, 3) for disease.

Generalization:

The technique generalization is to generalizing attribute values into "*less-specific but semantically consistent values*," generalization offers some protection against membership disclosure. It was shown in that generalization alone (e.g., used with k-anonymity) may leak membership information if the target individual is the only possible match for a generalized record. The perception is similar to our rationale of fake tuple. Although, if suppose a generalized tuple is does not introduce fake tuples (i.e., none of the other combinations of values are reasonable). So, there will be only one the original tuple that matches with the generalized tuple and the information of membership is can be still be inferred. Nergiz et al. has presents the concept and defines. A large background table is a set of all "possible" tuples. In that tuples in order to estimates the probability whether a tuple is in the data or not presence). The major problem with is that it can be difficult to define the background table and in some cases the data publisher may not have such a background table. The protection against membership disclosure depends on the option of the background table with the careful anonymization, generalization can offers some level of membership disclosure protection.

Improvised k-anonymity

Just removes the name and generalize the zip code and date of birth we have a less anonymized set. Consider k=2 for this set.

SBDT MODAL
Generalization method:

Age	Gender	Pin code	Disease
34	*	641020	Flu
34	*	641012	cancer
34	*	641010	diabetics
34	*	641020	Flu
45	*	641012	Cancer
45	*	601012	Cancer
48	*	631012	cancer
52	*	621010	diabetics

In this experiment, it evaluates the effect of the Extended K-set Anonymization based on clustering accuracy. Column age considered as, gender as b, and disease as c. Table 1.2 shows the results on learning the sensitive attribute by using generalization Improvised k-anonymity shows better accuracy than generalization.

The experiments are basically designed so that the different parts of the work could be evaluated easily and effectively. The performance of this proposed work Scheme was compared with the existing algorithms based on the following parameters.

- **Information Loss**- measures the proportion of amount of information loss during generalization
- **Execution Time** – Determines the Time interval
- **Clustering accuracy**–accuracy measurement.
- **Time taken for Clustering –** Determines the processing time involved for clustering.

Performance Evaluation:

This experiment has been done through the personal health care Dataset. The dataset is preprocessed by Improvised k-anonymity with SBDT Model.

Measurement of Privacy Preservation Rate

In Improvised k-anonymity with SBDT model, the privacy preservation rate is defined the amount of data that are must be kept secret from people who have no direct access to the original to the total number of available data. The privacy preservation rate is measured in terms of percentage (%) and formulated as,

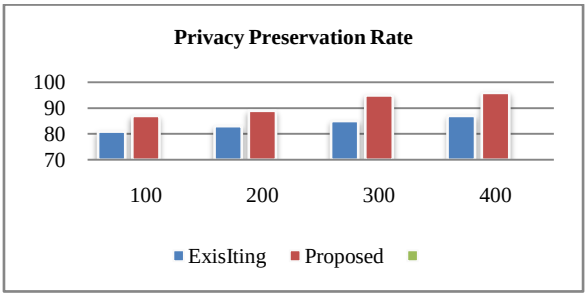
Privacy preservation rate=amount of data kept secret /total number of available.

When the privacy preservation rate is higher, the method is said to be more efficient.

Record size (MB)	**Privacy Preservation Rate (%)**	
	Existing two-phase clustering approach	**Proposed Improvised k-anonymity with SBDT**
100	81	87
200	83	89
300	85	95
450	87	96
500	89	98

Table 5.1: Tabulation for Privacy Preservation Rate

Privacy Preservation Rate of proposed Improvised k-anonymity with SBDT existing approaches based on dataset Preservation Rate.



From the results shown in the graphs, it can be observed that the proposed approaches provide better Privacy Preservation Rate. When, it is analyzed with different number of datasets.

Measurement of Execution Time to Preserve Privacy

The execution time to preserve privacy is mathematically formulated as given below,

Execution Time =Record size ∗ Time (εj)

The execution time to preserve privacy is measured in terms of milliseconds (ms). When the execution time to preserve privacy is lower, the method is said to be more efficient.

Table 5.2 Tabulation for Execution Time to Preserve Privacy

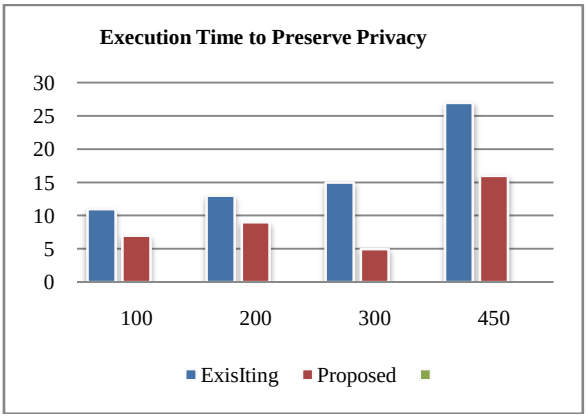


Figure 5.1 Measurement of Execution Time to Preserve Privacy

We consider the framework with different number of record size in the range of 100 to 150 is taken for experimental purpose using Java language. From the table value, it is illustrative that the execution time to preserve privacy using model Improvised k-anonymity with SBDT is reduced when compared to the other existing methods.

Measurement of Information Loss

In proposed model, information loss measures the information loss of an anonymous data table. The concept of information loss or data distortion often is used to reflect the data quality in privacy-preserving publishing. Information loss usually decreases the quality of the data and affects data utility. When information loss is low, the method is said to be more efficient.

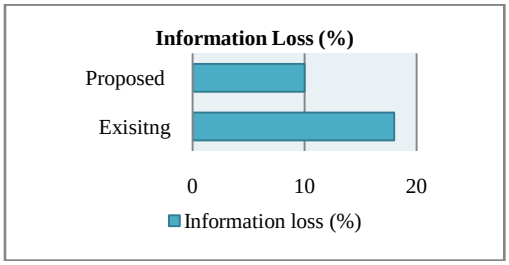


Figure 5.2 Measurement of Information Loss

Clustering Accuracy

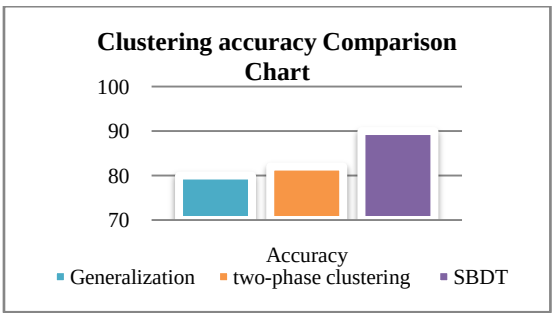


Figure 5.3 Comparison between SBDT and existing techniques

VI. CONCLUSION

This proposed Improvised k-anonymity with SBDT with hybrid techniques presents a novel approach that modifies the database to hide sensitive details with limited side effects. This SBDT and other techniques propose or mainly focused to hide all the valid sensitive attributes even in multi user access. Proposed in this thesis, the transactions can be randomized in an order so that both the numbers of original sensitive attributes and modified entries are considered. The experimental results show that the existing approaches results and the proposed approach. Here the proposed system is scalable in terms of database size and effective in terms of optimization. Moreover, this approach and the efforts taken to the avoidance of undesired side effects in sensitive attribute preserving are effective in two well-designed experiments. In most cases, all the sensitive attributes and **its** privacy are hidden without false complete anonymity. The experimental results are evaluated using the Java. The experimental result shows that integrated extended proposed algorithm shows better privacy preserving compared to traditional clustering techniques. This privacy preserving technique has been implemented in medical dataset however this consideration and enhancements that can be adapts to any domain. This is the major advantages of the proposed system. But the future work may also extend the discovery of the full set of rules for cost effective privacy preserving. This research study emphasis efficient mechanisms are required to speed up the privacy preserving process with dynamic datasets. Another issue is the fast recognition of privacy and sensitive rules that cannot be hidden according to the user-specified constraint. An ideal solution or goal is to build a system that can aid the database administrator to find the sensitive rules for hiding.

REFERENCES

[1]. Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan, Venkitasubramaniam. "l-diversity: Privacy Beyond k-anonymity". Proceedings of the 22nd International Conference on Data Engineering (ICDE'06). (2006).

[2]. Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Samir Khuller, Rina Panigrahy, Dilys Thomas, An Zhu. "Achieving Anonymity via Clustering". Encyclopedia of Database Systems. Springer, Boston, MA. 2136-2137. (2009).

[3]. Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, Ada Wai-Chee Fu. " utility-based anonymization using local recoding". Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.

[4]. Ji-Won Byun, Ashish Kamra, Elisa Bertino, and Ninghui Li. "Efficient k-Anonymization Using Clustering Techniques". International Conference on Database Systems for Advanced Applications. Springer, Berlin, Heidelberg, 2007.

[5]. Kristen LeFevre, DeWitt D J, Raghu Ramakrishnan. "Mondrian Multidimensional K-anonymity". Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. IEEE, 2006.

[6]. Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, Ke Wang. "(α, k)-Anonymity: an enhanced k-anonymity model for privacy preserving data publishing". Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.

Priyadarshini. D. " A Novel Improvised Anonymization with Map Reduce Framework for Big Data Privacy Preservation in Cloud."IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 12, 2018, pp. 17-25.