A Secure Protocol for High Dimensional Bigdata Providing Data Privacy

Prasad S P[1], Dr. Anitha J[2]
[1]Asst. Professor, Dept. of ISE, DSATM, Bangalore, Karnataka, India
[2]Professor, Dept. of CSE, DSATM, Karnataka, Bangalore, India

***Abstract:*-Due to recent technological development, a huge amount of data generated by social networking, sensor networks, Internet, healthcare applications and many other companies, which could be structured, semi-structured or unstructured, adds more challenges when performing data storage and processing tasks. During Privacy Preserving Data Processing (PPDP), the collected data may contain sensitive information about the data owner. Directly releasing this information for further processing may violate the privacy of the data owner, hence data modification is needed in such a way that it does not disclose any personal information about the owner. On the other hand, the modified data should still be useful, not to violate the original purpose of data publishing. The privacy and utility of data are inversely related to each other. Existing privacy preserving techniques like k-anonymity, t-closeness are focusing on anonymization of data which have a fixed scheme with a small number of dimensions. There are various types of attacks on the privacy of data like linkage attack, homogeneity attack and background knowledge attack. To provide an effective technique in big data to maintain data privacy and preventing linkage attacks, this paper proposes a privacy preserving protocol - UNION, for multi-party data provider with KCL anonymization. Experiments show that this technique provides a better data utility to handle high dimensional data, and scalability with respect to the data size compared with existing anonymization techniques.**

***Keywords:* Big Data, Anonymization, k-Anonymity, t-Closeness, Privacy Preserving Protocol.**

I. INTRODUCTION

The term big data is defined as ‘‘A new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis’’[1]. Based on this definition, the properties of big data are reflected as volume, velocity, variety, veracity and value. Volume refers to the amount of data generated. With the emergence of social networking, there is dramatic increase in the size of the data. The rate at which new data are generated is often characterized as velocity. A big data may contain text, audio, image, or video etc. This diversity of data is denoted by variety. Veracity refers to the data that are generated uncertain in nature. It is hard to know which information is accurate and which is out of date. Finally the Value of data is valuable for society or not.

The life cycle of the big data has various phases like data generation, data storage and data processing. In data generation phase, large, diverse and complex data are generated by human and machine. Data storage phase refers to storing and managing large data sets. In data processing phase, various computations and transformations takes place on data set.

Data processing phase is the process of data collection, data transmission, pre-processing and data extraction. Data collection is needed because data may be coming from different diverse sources i.e., sites that contains text, images and videos. In data transmission phase, after collecting raw data from a specific data production environment, a high speed transmission mechanism to transmit data into a proper storage for various types of analytic applications. The pre-processing phase aims at removing meaningless and redundant parts of the data so that more storage space could be saved. Finally in data extraction phase only useful information are retrieved from data sets.

The data processing phase includes Privacy Preserving Data Publishing (PPDP). During PPDP, the collected data may contain sensitive in formation about the data owner. Directly releasing the information for further processing may violate the privacy of the data owner, hence data modification is needed in such a way that it does not disclose any personal information about the owner [2]. On the other hand, the modified data should still be useful, not to violate the original purpose of data publishing. The privacy and utility of data are inversely related to each other. Intrusion Detection Scheme (IDS) schemes have been implemented in wired and semi-wired networks. These systems look for certain misbehavior patterns in the network which would give a whiff of a malicious act and thereby trigger attack mitigating mechanism [3]. Many studies have been conducted to modify the data before publishing or storing them for further processing.

The most basic form of PPDP the data holder which has a table of the form D(Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes), where Explicit Identifier is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners; Quasi Identifier is a set of attributes that could potentially identify record owners; Sensitive Attributes consist of sensitive person-specific information such as disease, salary, and disability status; and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories[2].

Anonymization refers to the Privacy Preserving Data Publishing approach that seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis. The data are anonymized by removing the identifiers and modifying the quasi-identifiers before publishing or storing for further processing. As a result of anonymization, identity of the data owner and sensitive values are hidden from the adversaries. How much data should be anonymized mainly depends on how much privacy need to preserve in that data. Before publishing, the original table is modified according to the specified privacy requirements. To preserve the privacy, one of the following anonymization operations is applied to the data [3].

- **Generalization**: Generalization works by replacing the value of specific QID attributes with less specific description. In this operation some values are replaced by a parent value in the taxonomy of an attribute. The types of generalization techniques include full domain generalization, sub tree generalization, multidimensional generalization, sibling generalization, and cell generalization.
- **Suppression**: In suppression, some values are replaced with a special character or symbols (e.g., ``*"), which indicates that a replaced value is not disclosed.
- **Anatomization**: Instead of modifying the quasi-identifier or sensitive attributes, anatomization works by de-associating the relationship between QID and SA. The data on QID and SA are released in two separate tables, one table contains quasi-identifier and the other table contains sensitive attributes. Both tables contain one common attribute which is often called GroupID.
- **Permutation**: In permutation, the relationship between quasi-identifier and numerically sensitive attribute is de-associated by partitioning a set of records into groups and shuffling their sensitive values within each group.
- **Perturbation**: In perturbation, the original data values are replaced by some synthetic data values, so that the statistical information computed from modified data does not differ significantly from the statistical information computed from the original data. Some examples include adding noise, swapping data, and generating synthetic data.

The privacy models are basically classified into two categories based on the ability of an attacker to identify an individual. The first category is based on the assumption that the attacker is able to identify the records of a specific user by linking the records with external data sources. The second category is based on the assumption that the attacker has enough background knowledge to conduct probabilistic attacks.

II. DATA ANONYMIZATION

Data anonymization is an information sanitization whose intent is to provide privacy protection. It is the process of either encrypting or removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous. The data privacy technique, seeks to protect private or sensitive data by deleting or encrypting personally identifiable information from a database. Data anonymization is done for the purpose of protecting an individual's or companies private activities while maintaining the integrity of the data gathered and shared.

2.1 K-Anonymity

This model was developed because of the possibility of indirect identification of records from public database. This is because combinations of record attributes can be used to exactly identify individual records. A data set complies with K anonymity protection if each individual's record stored in the released table cannot be distinguished from at least K-1 individual records. Let Release Table RT have attributes of A1, A2…A be a table and QI, be the quasi identifier associated with it. The RT is said to be K-anonymity if and only if each sequence of values in the RT appears with at least K occurrences in RT [QI]. This method protects the data against identity disclosure.

Drawbacks of K-Anonymity:

- Unsorted matching attack against K-anonymity i.e., tuple position within the table reveals the sensitive attribute
- Complementary release attack against K-anonymity
- Temporal attack against K-anonymity
- Homogeneity attack

- Background Knowledge attack

2.2 L-Diversity

Information about an individual could not be published without revealing the sensitive attribute of the table. In case of K-anonymity, the data was not protected because of the homogeneity and background knowledge. L-diversity techniques describes that there are at least L- well represented values for the sensitive attribute would have the same frequency. The adversary needs (L-1) damaging pieces of background knowledge to eliminate (L-1) possible sensitive values and infer a positive disclosure. The advantages of this method are it no longer requires knowledge of full distribution of SA and NSA, it does not require the publisher to have as much information as the adversary has and removes the drawbacks of K anonymity.

Drawbacks of L-diversity:

- L-diversity may be difficulty and unnecessary to achieve it.
- L-diversity is insufficient to prevent attribute disclosure, skweness attack and similarity attack.

2.3 T-Closeness

It has been proposed to describe the distribution of sensitive attribute with equivalence class. An equivalence class is said to have t-closeness if the distance between the distribution of the sensitive attribute in the class and the distribution of the attributes in the whole table is no more than threshold t. This technique is useful when it is important to keep the data as close as possible to the original one to that end, a further constraint is placed on the equivalence class, namely that not only at least l different values should exist within each equivalence class, but also that each value is represented as many times as necessary to mirror the initial distribution of each attribute.

2.4 Slicing

The generalization technique loses significant amount of information particularly for high dimensional data. The bucketization technique does not prevent membership. Slicing was a popular data anonymization technique which could be formalized by comparing with generalization and bucketization. In Slicing the data set is partition into both horizontally and vertically. The vertical partition is done by grouping attribute into columns contains a subset of attributes that are highly correlated. The horizontal partition is done by grouping tuples into buckets, within each bucket; the values in each column are randomly sorted to break the linkage between different columns. This technique provides the protection against membership disclosure attack.

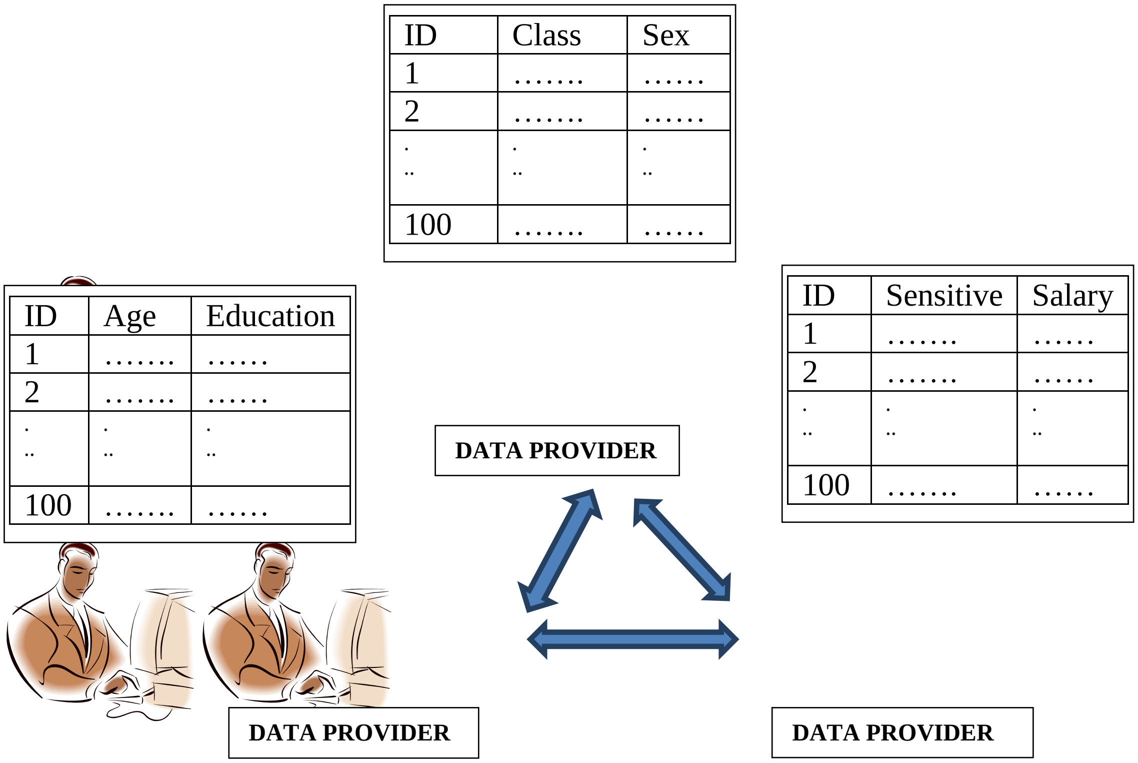
III. PROPOSED SYSTEM

The data privacy in the era of big data is mainly reflected in digging data under the premise of not exposing sensitive information of the user. Existing privacy preserving techniques are focusing on anonymization of data which have a fixed scheme with a small number of dimensions. So there is a need of new anonymization technique for dealing with high dimensional data. There are various types of attacks on the privacy of data like linkage attack, homogeneity attack and background knowledge attack. The anonymization techniques like generalization has following issues: There is huge amount of information loss particularly for high dimensional data and there is a significant decrease in the data utility of the generalized data. The existing anonymization techniques are insufficient to prevent the attribute disclosure. So, there is a need of a privacy model to overcome the above mentions issues.

This paper presents a novel technique called UNION, for integrating the distributed person specific data while preserving both privacy and information utility. The main idea is to slice the given dataset vertically into set of columns and each column is distributed into multiple parties. Because of slicing the data set into vertical columns, it makes very easier to handle the high dimensional data. Slicing preservers better data utility than the generalization.

Consider the below Table 1 which has the raw data. This raw data is vertically sliced and distributed between three data providers P1, P2 and P3 which provide the data during integrating the data. These data providers own the different set of attributes about the same individuals and P2 owns the Class Attribute. Figure 1 shows the distribution data into various data providers.

The data being integrated is in the form of a relational table that is vertically partitioned into sub tables and each of which is owned by one separate data provider. Each party Pi owns a table which has UID, explicit identifier (EID), quasi identifier (QID), sensitive attribute (SA) and class which is a categorical target class attribute for classification analysis.



ID	Class	Sex
1	…….	……
2	…….	……
. ..	. ..	. ..
100	…….	……

ID	Age	Education
1	…….	……
2	…….	……
. ..	. ..	. ..
100	…….	……

ID	Sensitive	Salary
1	…….	……
2	…….	……
. ..	. ..	. ..
100	…….	……

Figure 1: Privacy Preserving Data Distribution among Data Providers

TABLE 1: Raw data owned by different data providers

UID	Data Provider P1		Data Provider P2		Data Provider P3	
	Age	Education	Class	Sex	Sen	Salary
1	54	11th	Yes	Male	S2	15K
2	27	Master	Yes	Female	S1	35K
3	39	7th	No	Male	S2	5K
4	67	Doctorate	No	Female	S1	95K
5	29	Bachelor	Yes	Male	S2	29K

TABLE 2: Generalized data set values

UID	Data Provider P1		Data Provider P2		Data Provider P3	
	Age	Education	Class	Sex	Sen	Salary
1	1-99	Secondary	Yes	Any_Sex	S2	(10-70)K
2	1-99	University	Yes	Any_Sex	S1	(10-70)K
3	1-99	Secondary	No	Any_Sex	S2	(10-70)K
4	1-99	University	No	Any_Sex	S1	(70-100)K
5	1-99	University	Yes	Any_Sex	S2	(10-70)K

Now implement the anonymization technique of generalization to make the uniform distribution supposition that each value in a generalized interval is equally possible, as no additional distribution assumptions can be made. It contains generalizing the range which will substitute attribute value with more semantically consistent but minimum precise value. When using generalization, the original values are replaced by the more general ones of it. The below Table 2 show the generalized data set values.

IV. UNION: A SECURE PROTOCOL FOR HIGH DIMENSIONAL

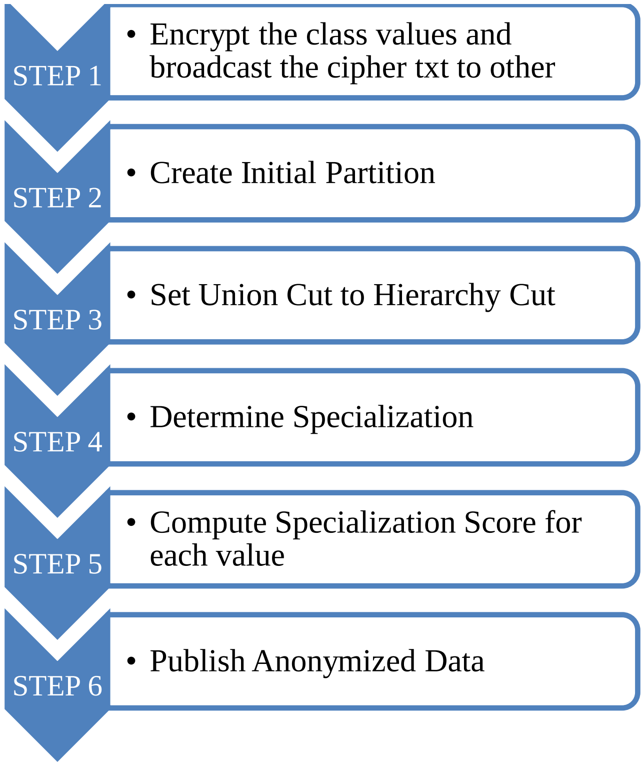
4.1 Solution Overview

This paper proposed a multi-party protocol named as Union, for integrating the person specific data distributed among various data providers. The main idea is to anonymize the raw data by generalizing all the raw data records to a general state and then performs a sequence of specializations such that in each specialization step we choose the specialization with highest score to maintain the highest possible information usage. Then we will use a distributed hierarchical approach for integrating the high dimensional data from multiple data providers, which preserves the data quality.

4.2 Multi Party Protocol for Data Integrity

The general idea is to initially generalize and assign all the records to a partition and then apply specialization process to specialize the records and assign then to disjoint child partitions. The partition is a data structure that consists of hierarchy cut and records. A record R can be assigned to a partition Part if for each attribute of R can be generalized to Part.Hcut. A where R.A is the value in R and Part.Hcut.A is hierarchy cut of the attribute A.

A specialization is valid if after the child partition is created, the leaf partitions as a whole in the partitioning. The specialization process terminates when there is no more valid specialization is available. The mash up data for final released are constructed from the hierarchy cut of the leaf partitions where each hierarchy cut is duplicated |Rec| times. The algorithm for the data integration is as shown below:



STEP 1
- Encrypt the class values and broadcast the cipher txt to other

STEP 2
- Create Initial Partition

STEP 3
- Set Union Cut to Hierarchy Cut

STEP 4
- Determine Specialization

STEP 5
- Compute Specialization Score for each value

STEP 6
- Publish Anonymized Data

V. EXPERIMENTAL RESULTS

Experiments were carried out on adult data set taken from UC Irvine Machine Learning Repository - UCI Machine Learning. (https://archive.ics.uci.edu/ml/datasets.html). The data set contains 48842 instances with 14 attributes both categorical and integer. The data contains sensitive and non-sensitive (quasi identifier) attributes. The data was cleansed and formatted and made into sets of 40000, 80000, 160000, 320000 and 640000 with random replication. The Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data 251 experiments are conducted on a machine with Intel ® Core TM i5-2120 CPU @ 3.30 GHZ, 4 GB RAM, Window 7, JAVA –JDK 8.0. The objective of the experiment is to find out performance metrics such as execution time, data utility and privacy of the various privacy preservation models applied to big data.

5.1. Execution Time

The following Table 3 and Figure 2show the execution time – the time taken by the algorithm to perform the task by various models with different data size.

Table 3: Execution Time

Sl N o.	Model/Data size	40000	80000	160000	320000	640000
1	K-anonymity	105	200	410	800	1450
2	L-diversity	130	240	430	860	1600
3	t-closeness	170	300	600	1300	2700
4	Differential privacy	160	280	590	1250	2500
5	Slicing	150	275	560	1200	2400

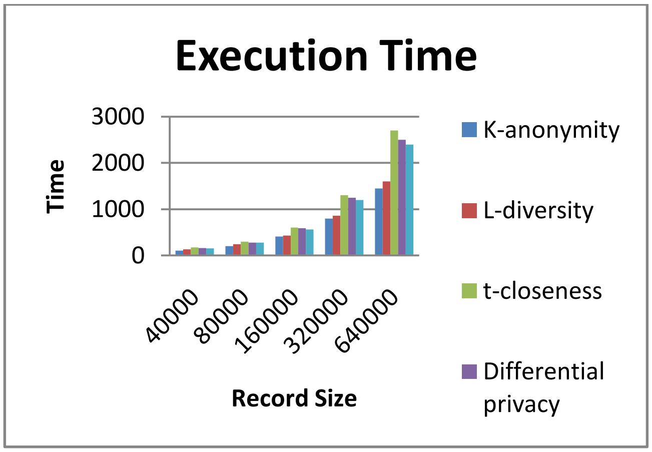


Figure 2: Execution Time

5.2 Data Utility & Complexity

Data utility is measured by the accuracy of the queries MIN, MAX, COUNT on the original data and the transformed data after applying the privacy preserving techniques as shown in the table 4.

Table 4: Data Utility and Complexity of Data Models

Sl. No	Models	Data Utility	Complexity
1	k-anonymity	Low	Very Low
2	l-diversity	High	Low
3	t-closeness	High	Very high

4	Differential privacy	Medium	high
5	Slicing	Medium	high

VI. CONCLUSIONS AND FURTHER WORK

In this paper, we present a secure protocol for data integration in a distributed setting. The protocol is privacy preserving, while the output is a mash up data for data mining. We empirically show that the mash up data contains higher information utility, and that the protocol is scalable with respect to the number of records as well as the number of attributes in the mash up data. For future work, we plan to address the privacy-preserving data mash up problem in a malicious adversarial model with public verifiability.

REFERENCES

[1]. J. Gantz and D. Reinsel, “Extracting Value from Chaos," in *Proc. IDC IView*, 2011.

[2]. B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, ``Privacy-preservingdata publishing: A survey of recent developments," *ACM Comput. Surv.*,vol. 42, no. 4, Jun. 2010, Art. no. 14.sing

[3]. Anitha J and Anil Kumar, ``Detection of Intrusion through Big data Analytics for Wireless Sensor Networks," Int. Journal of Science, Engineering and Technology,2017

[4]. Abid Mehmood, Iynkaran Natgunanathan, Yong Xiang, Guang Hua, “Protection of Big Data Privacy”,Theoretical Foundations For Big Data ApplicationsOpportunities, 2016.

[5]. L. Sweeney, “k-anonymity: A Model for Protecting Privacy," Int. J. Uncertainty, Fuzziness Knowl. Based Syst., vol. 10, no. 5, pp. 557570, 2002.

[6]. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam,"L-diversity: Privacy beyond k-anonymity,"*TKDD*, vol. 1, no. 1, 2007.

[7]. Yong Xu, Tinghuai Ma, Meili Tang,Wei Tian, “A survey of Privacy Preservation Data Publishing using Generalization and Suppression”, Int. Journal Applied Mathematics & Information Sciences, 1103-1116,2014.

[8]. Girish Agarwal, Pragati Patil, “Privacy Preserving for High Dimensional Data using Anonymization Techniques”, Int. Journal of Advanced Research in CS, June 2013.

[9]. Snehal M Nargundi, Rashmi Phalnikar, “k-anonymization using Multidimensional Suppression for Data De-identification”, Int. Journal of computer application, 2012.

[10]. Freny Presswala, Amit Thakkar, Nirav Bhatt, “Survey on Anonymization in Privacy Preserving Data Mining”, IJIERE 2015.