Privacy Preservation in Big Data using K-Anonymity Algorithm with Privacy Key

Amit Kumar Gupta
M. Tech Scholar
Gyan Ganga College of Technology
Jabalpur MP (India)

Neeraj Shukla
HOD CSE Department
Gyan Ganga College of Technology
Jabalpur MP (India)

ABSTRACT
Big Data is a complex collection of huge amount of data records, everything around us is a source of big data that is too broad and too complex. Big data is generated by traditional data processing applications are inadequate. Many challenges are there in Big Data, they includes analysis of data in big data, capturing of data, data curtain, searching, sharing of data record, storage of data, data transfer, visualization, querying with database, updation of database and information privacy. Privacy preservation can be managed properly in case of limited amount of data, but in case of huge amount of data that is "Big Data" privacy preservation is a very big issue. Many algorithms are discussed here to solve the privacy preservation issue, generally Anonymization, Notice and Consent, Differential Privacy methods are used to resolve the issue of privacy preservation, each method have its merits and demerits, any method is not fully comfortable to provide required level of privacy having less amount of burden. In this paper there is discussed an advanced k-anonymity algorithm with privacy key, unique key is generated automatically by privacy key generation mechanism at the time of data creation. Then the information is stored in the database with that key, thus, user data is much secure than existing k- anonymity algorithm because of two level of security. The first level is by K-anonymity and the second is by privacy key.

General Terms
Privacy Preservation, Big Data Privacy and Security, Various privacy preserving algorithms and there comparison etc

Keywords
Big Data; Anonymization; Notice and Consent; Differential Privacy; K-anonymity; K-anonymity with privacy key, Suppression, Generalization etc.

1. INTRODUCTION
Today Society is experiencing exponential growth in the number and variety of data collections containing person-specific information, as computer technology, disk storage space and network connectivity become increasingly affordable. Big data is among emerging technologies that are beginning revolution in the world of data storage, and data analytics. Big Data has power to provide insights to unseen aspects of data analysis. Generally Big Data is defined with 3 V's [1]. The Volume, the Variety, and the Velocity. These 3 points explains that the Big Data has a very large volume, it has uncountable varieties, and its generation speed is very high. Thus it can be said that the Big Data is a very huge term which is growing second basis [5].

We create 2.5 quintillion bytes of data — so much that 90%of the data in the world today has been created in the last two years alone [7]. There are many sources of big data, like emails, phones, videos, search queries, mobile phones, and a lot of other communication systems. This shows that sources of big data are scattered all around internet. That's it can be said that big data is mostly depends on web and networks. Millions of data generated today and managed by big data software's, the big data software's provide several big data operations, like data creation, deletion, data searching, data retrieval etc. Many security issues included in these operations, big data security is a very big issue, not considering security issue may cause big loses also, and thus there must keep proper security features to big data applications to ensure secure operations in big data and secure big data analytics.
Big data may yields good analytics results but at same time can pose security threats also, hidden values in big data can be very useful values to hackers, that's why there is a tradeoff between big data availability and big data security. There is a need to maintain balance between big data availability and big data security, thus to protect the hidden information during the whole analytics process is the main issue.

Traditional algorithms are not too efficient to ensure privacy and security in big data, due to its large volume, variety, velocity and its complexity. Privacy preservation is very big issue because, big data generally contains person specific information, social profile details having sensitive information.

Traditional methods like cryptography (encryption/decryption) can be used for privacy and security, but these methods are not too efficient to handle the complex nature of data.

Anonymization is also useful to hide personal information in big data. Anonymization is the process of changing the private information to such a way that hides the key identification of data record. In Anonymization the original data changes by applying some method to convert it into an un-recognizable data record. Many variations of Anonymization is used namely K-anonymity, L-diversity, T-closeness etc.

Two other methods used to ensure privacy in big data, first is Notice and Consent, and the second is Differential Privacy.

In Notice and Consent method consumer information is shared only after obtaining consent from the user when using a new app or a new web service [2].

Differential Privacy is being widely used. It is a method enabling analysts to extract useful information from databases containing personal information while preserving strong individual privacy protections [2].

2. REVIEW OF EXISTING ALGORITHMS FOR PRIVACY PRESERVATION IN BIG DATA

Cryptography is one of the most common used algorithm for data privacy preservation. It refers to an encryption/ decryption algorithm. In this algorithm plaintext is converted to cipher text at sender end, again the cipher text is converted to plaintext at receiver end, the whole process called the cryptography, there are many methods used to implement cryptography like digital signatures, public key cryptography, RSA algorithm etc.

Cryptography alone is not sufficient to solve the problem of data privacy preservation, it can't enforce the required level of privacy preservation for common cloud computing and big data services, because of the large and complex nature of big data. As we discussed 3 V's regarding the big data that are Volume, Variety and Velocity which create a great level of complexity to scale the cryptography algorithms for data privacy preservation.

The challenge with encryption/ decryption algorithm all or nothing retrieval problem. During Data analytics process the less sensitive data can also be encrypted which is useful and may be not accessible to users. Privacy can also be destroyed if the attackers attacks before encryption [2].

Thus it can be stated that Cryptography cannot ensure the privacy preservation alone, cryptography methods can be used with other methods to ensure a great level of privacy preservation [2].

3. PRIVACY PRESERVATION FOR BIG DATA

Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data [9]. Today, Privacy preservation in big data is such a big issue that is not easy to resolve due to nature and varieties of data in big data which is very large and huge and having a great level of complexity. There is discussion on some methods used for privacy preservation in big data that are Data Anonymization, Notice and Consent and Differential Privacy preservation. Data Anonymization is a most common used method in various ways. Many variations are there in Anonymization method which can be applied easily and data privacy problem can be resolved to a level, many variations in Anonymization method is based on some improvements with each variation. Considering a sample data table to implement all the methods to see their effects on original data, merits, demerits and their level of data privacy preservation.

Table 1: Sample Data Set

Name	Age	Gender	Salary/Mon (Rs.)
Amit	23	Male	50000
Ram	31	Male	50000
Ajay	45	Male	60000
Abhilash	25	Male	40000
Sunita	26	Female	40000
Swara	32	Female	20000
Aksahra	35	Female	25000
Abhishek	41	Male	35000
Smaran	15	Male	25000
Hirendra	27	Male	28000
Sriram	31	Male	30000
Komal	32	Female	16000

3.1 Data Anonymization

Data Anonymization is a process to change the original data into some other form to store and publish in database that is un–identifiable to attackers [1]. It is also known as data de-identification. In Data Anonymization some parts of information are converted into another form or replaced by some symbols to hide the original information, this process is applied to big data to hide the key information of data to ensure the data privacy and to prevent data from attacks [8]. This is most commonly used privacy preservation method in big data, many variations of this method are also available. Each have some merits and demerits, these methods are described below.

3.2 K-anonymity

K-anonymity is the simplest form of Anonymization method, it is the simplest implementation of Anonymization method. A data set is called k-anonymized for any row with given attributes (fields) if there are at least k-1 other data records that match the attributes. Two techniques are used to implement k-anonymity. First is the Supersession and the second is the Generalization.

In Supersession some key attributes is replaced by some symbols like * or some constant values like 0 [4]. For example in above given Table 1 "Name" is Key attribute and Age is a quasi-identifier attribute, if Supersession is applied to "Name" and replace it by symbol '*' as Table 2, then key information can be hidden from database.

Table 2: Anonymized data after Suppression

Name	Age	Gender	Salary/Mon
*	23	Male	50000
*	31	Male	50000
*	45	Male	60000
*	25	Male	40000
*	26	Female	40000
*	32	Female	20000
*	35	Female	25000
*	41	Male	35000
*	15	Male	25000
*	27	Male	28000
*	31	Male	30000
*	32	Female	16000

Generalization is another technique to hide the original information, in this technique the quasi identifiers are modified to more general values to hide the key information from database [2]. For example in above Table 1 "Age" is a quasi-identifier field, that can be modified to most general values, like below table.

Table 3: Anonymized Data after Generalization

Name	Age	Gender	Salary/Mon (Rs.)
Amit	>20<30	Male	50000
Ram	>30<40	Male	50000
Ajay	>40<50	Male	60000
Abhilash	>20<30	Male	40000
Sunita	>20<30	Female	40000
Swara	>30<40	Female	20000
Aksahra	>30<40	Female	25000
Abhishek	>40<50	Male	35000
Smaran	>10<20	Male	25000
Hirendra	>20<30	Male	28000
Sriram	>30<40	Male	30000
Komal	>30<40	Female	16000

This can be understand by below table.

Table 4: Anonymized Dataset after both Suppression & Generalization

Name	Age	Gender	Salary/Mon (Rs.)
*	>20<30	Male	50000
*	>30<40	Male	50000
*	>40<50	Male	60000
*	>20<30	Male	40000
*	>20<30	Female	40000
*	>30<40	Female	20000
*	>30<40	Female	25000
*	>40<50	Male	35000
*	>10<20	Male	25000
*	>20<30	Male	28000
*	>30<40	Male	30000
*	>30<40	Female	16000

There are two more variations in Anonymization techniques to ensure a higher level of privacy, first is the L- Diversity, L-diversity technique of Anonymization of data tries to bring diversity in the sensitive attributes of data. It enables that each equivalence class of quasi identifiers has at least L different values of sensitive attribute. And the second is T- Closeness, An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the attribute distribution in the whole table is no more than a threshold value t [2].

Data Anonymization can be applied to big data to resolve the problem of privacy preservation but the problem lies in the reality that as size and variety of data increases, the chances of re-identification also increase. Thus, Anonymization has a limited potential to resolve the privacy problem in the field of big data.

3.3 Notice and Consent

The most common method for privacy preservation for web services is notice and consent, every time an individual accesses the new application a notice stating privacy concerns generated, the user needs to consent the notice before using the service. This method enhance an individual rights to secure its data, means it puts the burden of privacy preservation on individual.

Notice and Consent method when used for big data, many challenges generates, in most cases uses of big data are unexpected and unknown at the time when notice and consent is given. This requires notice to change every time big data is used for deferent purpose. Big data is processed and collected so rapidly that is notice is also a burden to individuals to consent each time. An optional notice and consent system can be used for big data to resolve the so rapid notice and consent [2].

3.4 Differential Privacy

Differential Privacy is another method for big data privacy preservation. Differential privacy is a new method to implement the privacy in big data. Differential privacy method enable the analysts to extract the useful information from databases containing private information, offering strong privacy preservation [2].

In this method users don't have the direct access of database. In this method there is an intermediate interface between users and database, that is users directly not access the database, that interface provide the needed information to users from database. The interface calculates the results and adds desired inaccuracies during providing the answer of any question. This interface used as a firewall. The added inaccuracies during answers are good enough to protect the data privacy. This method of privacy protection can be understand by following diagram:

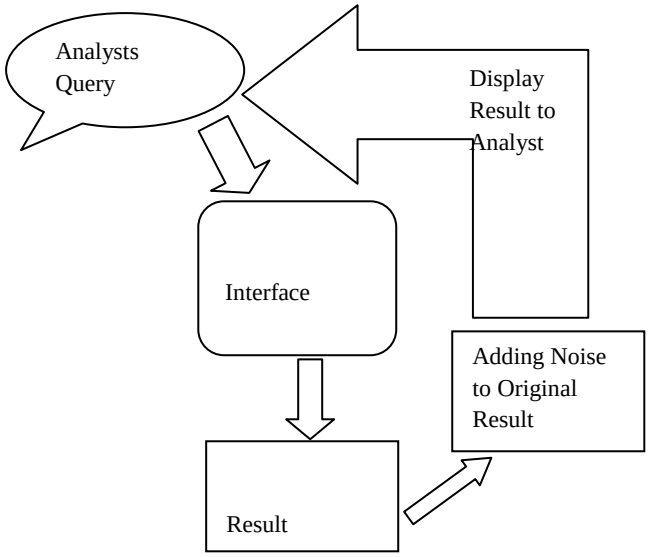

Figure 1: Differential Privacy Process

Benefits of Differential Privacy:

- Original Data set is not modified.
- No need for suppression and generalization techniques.
- No need to notice users.
- Inaccuracies added to data due to some mathematical calculations according to type of data.

3.5 (Alpha, k)-anonymity Algorithm (An Enhanced K-anonymity model)

(Alpha, k)-anonymity algorithm is extended form of k-anonymity algorithm. This method extends the simple k-anonymity model to multiple sensitive values. When there are two or more sensitive values and they are rare cases in database, they can be combined into one combined sensitive class and the simple k-anonymity algorithm is applicable. The inference confidence to each individual sensitive value is smaller than or equal to the confidence to the combined value which is controlled by α. anonymity algorithm. The (α, K) - anonymity model is an algorithm that mainly for the situation when there are few sensitive values in a data table [6],

That is, only consider the property value with high degree of sensitive in the sensitive attribute. It allows a confidence level from k-anonymous group to the reasoning of certain sensitive property value, less than a given value α *[6]*.

4. PROPOSED WORK

From the literature review it is clear that many algorithms are there to ensure privacy preservation in Big Data, each method have its merits and demerits, thus it can be said till now no any method is there that can fully ensure privacy preservation considering all aspects. Thus there is a need of an algorithm which can ensure high level of privacy, considering such complex and large volume of data i.e. Big Data. This work proposes an advanced method for privacy preservation in big data using anonymity technique with privacy key

In proposed algorithm it is based on k-anonymity algorithm, and a little bit on differential privacy method. In proposed algorithm first step is to apply the k-anonymity on the dataset after that publish this data record with a unique privacy key generated by privacy key generation mechanism, at the time of data creation, the whole data record stored with a privacy key, that is only known by the original user, thus it can be said that the data is secured with the provided key, and attackers do not know the privacy key i.e. they will be unable to make any attack on data without privacy key. During data extraction, the original user must have that key to extract the data from database. Without privacy key original user also cannot extract the data record. Only Anonymized data can be extracted by user without privacy key.

Thus security of data is increased in this method due to privacy key concept, data is two level secured, first is anonymity and the second is by privacy key. Thus data privacy and security is enhanced.

The proposed algorithm can be understand by following graphical explanation:

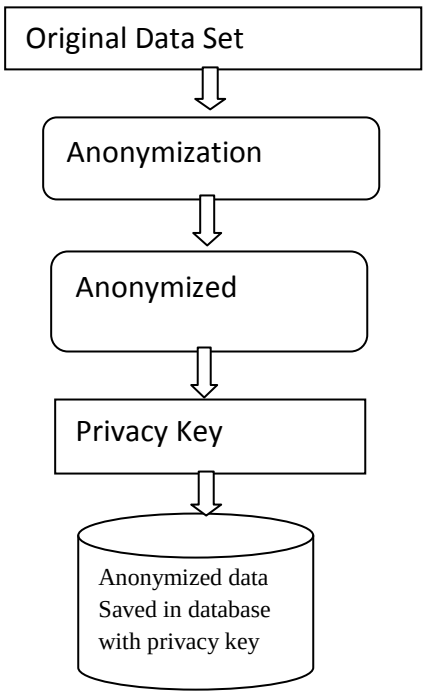

Figure: 2 Proposed Privacy Preservation Method

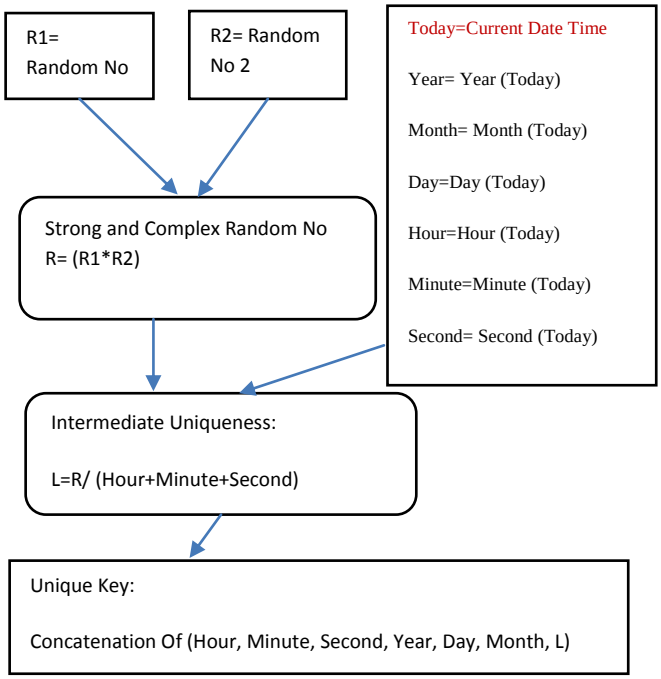

Figure: 3 Privacy key Generation Mechanism:

The whole algorithm can be well understand by an example.

Considering same dataset as Table 1, According to k-anonymity algorithm we can apply suppression and generalization on original dataset and can get anonymized dataset as following:

Table 5: Anonymized Dataset after both Suppression & Generalization

Name	Age	Gender	Salary/Mon (Rs.)
*	>20<30	Male	50000
*	>30<40	Male	50000
*	>40<50	Male	60000
*	>20<30	Male	40000
*	>20<30	Female	40000
*	>30<40	Female	20000
*	>30<40	Female	25000
*	>40<50	Male	35000
*	>10<20	Male	25000
*	>20<30	Male	28000
*	>30<40	Male	30000
*	>30<40	Female	16000

Now, there is anonymized data (Table 5) having no any identifiable attribute, now a privacy key is generated by privacy key generation mechanism to more secure user's data, the dataset will be stored and published in database with privacy key attribute. During data extraction if the user provides the privacy key then original data can be found and can be accessed only, otherwise only anonymized dataset can be extracted, thus data can be kept in more secure and private.

5. IMPLEMENTAION AND RESULTS

This section of paper implements the proposed "Privacy Preservation using K-anonymity with privacy key" method using Dot Net (C#) technology, this implementation is done on a small scale environment. This implementation is done on the basis of following minimum software and hardware resources:

Windows Operating System, 1GB RAM, 1GB Hard Drive Space, Dual Core Processor, Visual Studio, MySql etc., I have done the implementation work on Windows 10 Operating System, Having Visual Studio 13.0, MySQL 5.0 with Xampp server, and on the configuration of 4 GB RAM, 500GB Hard Drive, Intel i3 processor etc.

After Successful implementation of the proposed algorithm analysis of generated result is taken. It can be staed that using this algorithm data can be secured in two ways, first is by the anonymization and by private privacy key also. This two level of security is more reliable and more helpful to privacy preservation in Big Data.

Comparing to previous methods, this method is more fruitful in the sense of securing the private data, due to its nature of two level of security, it includes private key base protection with the power of anonymization methodology.

After reviewing existing methods their results and level of privacy preservation, each method have some merits and demerits, and a privacy preserving technique. A comparison chart is created for existing mehods. The comparison chart well describes each method with its result its level of security and its limitation, the comparison is as following table:

Table 6: Comparison between various privacy preservation algorithms:

Technique	Result	Limitation
K-Anonymity	It is good method based on static data set.[4]	It is based on static data set, there are some problems to be discussed.[4]
k-anonymity and (alpha, k) anonymity	(α, k)-anonymity model protect both identifications and relationships in data[6]	The k-anonymity model protects identification information, but Does not protect sensitive relationships in a data set.[6]
Notice and Consent	More secure than Anonymity[2]	Burden on user due to much more interactions[2]
Differential Privacy	Efficient and secure[2]	Not yet implemented[2]

Proposed "Privacy Preservation using k-anonymity with privacy key"	More secure than k-anonymity, due to two level of security	Database Space requirement may increase.

6. CONCLUSION

Big data privacy is a major issue because it is directly related to customers, hospital patients, students, employees and peoples from different sector organizations. It is very important for an organization to ensure privacy of any person during big data analytics. To provide proper privacy preservation to big data, Anonymization techniques have limited exposures. Notice and consent method is also not so powerful and user friendly. Proposed method is more powerful than existing k-anonymity method this method reduces chances of attacks. Differential Privacy may also can provide better solution for big data privacy. Differential privacy can be applied to big data without modification to original data. A combined procedure of Differential privacy and enhancements of proposed method can be a better future replacement.

7. REFERENCES

[1] Latanya Sweeney "k-Anonymity: A model for protectig data privacy". School of Computer Science, Carnegie Mellon University, Pittsburgh, USA 2002.

[2] Anjana Gosain, Nikita Chugh "Privacy Preservation in Big Data"USICT, Guru Govind Singh Indraprastha University, Delhi India 2014.

[3] "Security Issues associated with Big Data in Cloud Computing" Venkata Narsimha Inukollu[1], Sailja Arsi[1], Srinivas Rao Ravuri[3], [1]Department of Computer Science Engineering, Texas Tech University, USA, [3]Department of Banking and financial services, Cognigent Technology Solutions, India.

[4] "Research on privacy preserving on k-anonymity" Yun Pan, Xiao-ling Zhu, Ting-gui Chen, School of Computer Science & Information Engineering, Zhejiang Gongshang University, Hangzhou, China 2012.

[5] Brijesh B. Mehta, Uday Pratap Rao "Big Data Privacy: Issues and Challenges" Computer Science Engineering Department, Sardar Vallabh Bhai National Institute of Technology, Surat.

[6] "An Improved and Efficient Data Privacy in Big data with K-anonymity and alpha Dissociation" Salini S, MTech in CSE Marian Engineering College , Triendram India,Sreetha V Kumar Assistant Professor in CSE,, Marian Engineering College, Trivendram , IndiaNeevan R, Assistant Professor in CSE, College of Engineering Kottarkara Kollam, India.

[7] "Survey Paper on Big Data" Ms. Vibhavari Chavan , Prof Rajesh. N. Phursule.

[8] "Quasi and Sensitive Attribute based Perturbation Technique for Privacy Preservation" , Neha Patel, Prof. Srikant Lade, Prof. Ravindra Kumar Gupta 2015, Department of Computer Science & Engineering RKDF, IST, RGPV, University, Bhopal, India.

[9] "Challenges and Opportunities with Big Data" A Community white paper developed by leading researchers in across the united states.

[10] "Addressing Cloud Computing security issues" Dimitrios Zissis, Dimitrios Lekkas, Department of Product and System design engineering , University of the Aegon, Syros84100, Greece.