ISSN: 2321-7782 (Online)
Volume 3, Issue 8, August 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

An Improved and Efficient Data Privacy in Big Data with K-Anonymity and Alpha Deassoaciation

Salini. S[1]
Mtech in CSE
Marian Engineering College
Trivandrum- India

Sreetha. V. Kumar[2]
Asst.Proffesor in CSE
Marian Engineering College
Trivandrum- India

Neevan. R[3]
Asst,Professor in CSE
College of Engineering, Kottarakara
Kollam, India

Abstract: ***Big data concerns of large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, big data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Information sharing is an ultimate goal for all systems involving multiple parties. To protect privacy, two common approaches are used first is restrict access to the data, and second is anonymize data fields such that sensitive information cannot be pinpointed to an individual record. Here data privacy can protect using K- Anonymity technique and data security is implemented using authentication method. K-Anonymity is the method that anonymized data fields such that sensitive information cannot be pinpointed to an individual record. K-Anonymity privacy preserving model is good solutions to links attack, but without restriction on sensitive data, an attacker can use background knowledge about the k-anonymity data, So leakage of sensitive data are still there, so there must need a better privacy preserving technique. In this paper proposing an alternate method using Alpha, K- Anonymity in order to obtain a better privacy in big data environment, and adding a security engine using AES encryption.***

Keywords: K-Anonymity; Alpha deassociation; Data privacy; Big data; Background knowledge Attack

I. INTRODUCTION

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. Information sharing is an ultimate goal for all systems involving multiple parties, so data privacy is an important factor for data mining with big data. It cannot manage them with our current methodologies or data mining software tools. Big data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big data challenge is becoming one of the most exciting opportunities for the next years. K- Anonymity is the method that anonymized data fields such that sensitive information cannot be pinpointed to an individual record. One of the major benefits of the data anonymization based information sharing approaches is that, once anonymized, data can be freely shared across different parties without involving restrictive access controls. But K- Anonymity have some drawbacks such that, back ground knowledge attack is possible. So need a better privacy preserving technique, here proposing an another privacy preserving algorithm Alpha, K- Anonymity with classification.

II. Problem Statement

Big data characteristics are HACE theorem. This theorem states that Big Data starts with large-volume, Heterogeneous; Autonomous sources with distributed and decentralized control, and seeks to explore Complex and Evolving relationships among data. Then the challenges of big data are classified in to three tiers and they are big data mining platform, big data semantics and inside big data semantics there are two issues information sharing and data privacy and domain and application knowledge and the tier III is big data mining algorithms. In tier III there are three steps local learning and model fusion for multiple information sources, mining from sparse uncertain, and incomplete data, mining complex and dynamic data. The challenges at [1] Tier I focus on data accessing and arithmetic computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing. The challenges at Tier II centre on semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III) [1][7].

But implementing these all tiers are not feasible so concentrate in data privacy and also in security related to big data mining. Privacy preservation has become a major issue in many data mining applications. When a data set is released to other parties for data mining, some privacy-preserving technique is often required to reduce the possibility of identifying sensitive information about individuals. This is called the disclosure-control problem [1][7]. While the motivation for sharing is clear, a real world concern is that Big Data applications are related to sensitive information, such as banking transactions and medical records. Simple data exchanges or transmissions do not resolve privacy concerns. But public disclosure of an individual's information can have serious consequences for privacy. To protect privacy, two common approaches are to 1) Restrict access to the data, such as adding certification or access control to the data entries, so sensitive information is accessible by a limited group of users only, and 2) Anonymized data fields such that sensitive information cannot be pinpointed to an individual record.

For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be misconduct by unauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals. For example, the most common k-anonymity privacy measure is to ensure that each individual in the database must be indistinguishable from k-1 others. Common anonymization approaches are to use suppression, generalization, perturbation, and permutation to generate an altered version of the data, which is, in fact, some uncertain data. Privacy relates to what data can be safely disclosed without leaking sensitive information regarding the legitimate owner.

In the literature of K- Anonymization, there are two main models. One model is global recoding and the other is local recoding. K- Anonymity uses generalization and suppression techniques. Generalization replaces lower level domain values with higher level domain values. In global recoding, all values of an attribute come from the same domain level in the hierarchy. One advantage is that anonymous view has uniform domain but it may lose more information .Global recoding suffers from over generalization. With local recoding values may be generalized to different levels in the domain.

III. K- Anonymization

A large number of privacy models were developed most of which are based on the k-anonymity property. The K-anonymity model was proposed to deal with the possibility of indirect identification of records form public databases, k-anonymity means each released record has at least (k-1) other records in the release whose values are indistinct. For example, Hospital contains large database in such a way that identity of individual cannot be revealed. It helps to reveal public databases without compromising privacy. Thus, it prevents database linkages. In k-annonymity the granularity of data representation is reduced by using techniques such as generalization and suppression. The granularity is reduced to such a level that any given record maps

onto a least K other records in the dataset. A general method widely used for masking initial micro data to conform to the k-anonymity model is the generalization of the quasi identifier attributes.

One way to enable effective data mining while preserving privacy is to anonymize the data set that includes private information about subjects before being released for data mining. One way to anonymized data set is to manipulate its content so that the records adhere to k-anonymity. Two common manipulation techniques used to achieve k-anonymity of a data set are generalization and suppression. Generalization refers to replacing a value with a less specific but semantically consistent value, while suppression refers to not releasing a value at all. Generalization is more commonly applied in many domains since suppression may dramatically reduce the quality of the data mining results if not properly used.

One of the major benefits of the data anonymization based information sharing approaches is that, once anonymized, data can be freely shared across different parties without involving restrictive access controls. This naturally leads to another research area namely privacy preserving data mining where multiple parties, each holding some sensitive data, are trying to achieve a common data mining goal without sharing any sensitive information inside the data [5].Similar to all other fields of security, database security uses authentication, authorization, and accounting to ensure that only authenticated users perform authorized activities at authorized points in time. Database security includes many layers of security, which can be classified in topics such as physical security, network security, encryption, and authentication. K-anonymity is one of the most important concepts in data anonymity through re-identification. Although there are many data sets available for linking to external attackers, k- anonymity does not make any assumptions regarding them. The re-identification algorithm is an implementation of k-annonymity, and is triggered after every change of the data set (insertion, deletion or update). The k factor is dynamically decided based on the number of records existing in the data ware house.

K-Anonymity privacy preserving model is a good solutions to link attack, but without restriction on sensitive data , an attacker can use background knowledge attack. Attacker can also use sensitive information of a certain equivalence class is basically the same in K- Anonymity data, combined with the quasi identifier attribute to determine the equivalence class belongs to the individual privacy treat this is homogeneity attack. This is because K- Anonymity model doesn't protect both identifications and relationships to sensitive information in data. So in order to further improve the ability of K- anonymity privacy preserving model against background knowledge attacks and homogeneity attacks, a better privacy preserving model is needed. So proposing an another better privacy model Alpha ,K- Anonymity with classification

IV. ALPHA, K- ANONYMITY WITH CLASSIFICATION

The k-anonymity model requires that every value set for the quasi-identifier attribute set has a frequency of zero or at least k. Consider a large collection of patient records with different medical conditions. Some diseases are sensitive, such as HIV, but many diseases are common, such as cold and fever. Only associations with sensitive diseases need protection. Here the α-deassociation requirement for the protection. If a data set D is alpha deassociated, an attribute set Q and a sensitive value s in the domain of attribute $S \in Q$. Let (E, s) be the set of tuples in equivalence class E containing s for S and α be a user specified threshold, where $0 < \alpha < 1$. Data set D is α-deassociated with respect to attribute set Q and the sensitive value s if the relative frequency of s in every equivalence class is less than or equal to α. That is $|(E, s)| / |E| \leq \alpha$ for all equivalence classes E [3].

Local recoding typically distorts the values in the tuples in a data set. It can define a measurement for the amount of distortion generated by a recoding, which is call the recoding cost. If a suppression is used for recoding of a value which modifies the value to an unknown *, then the cost can be measured by the total number of suppressions, or the number of *'s in the resulting data set. The aim is to find local recoding with a minimum cost. The corresponding decision problem is defined as follows.

(α, k)-ANONYMIZATION: Given a data set D with a quasi- identifier Q and a sensitive value s, is there a local recoding for D by a function c such that, after recoding, (α, k)-anonymity is satisfied and the cost of the recoding is at most C.

There are four steps in Alpha, K –Anonymity they are Attribute Classification, Hidden and Generalization, K-Anonymity Privacy Preserving Model , (Alpha, K) – Anonymity, (Alpha, K) – Anonymity using Classification**.** In attribute classification, data anonymity will deal with the raw data divided in to four categories. Identifier attribute, which is used to uniquely identify an unique individual, such as ID number, Social Security Numbers and so on. Quasi Identifier attribute , which is a set of attributes , links with back ground knowledge can only identify the unique individual, such as age, address, nationality, postcode. Sensitive Attribute , which is the attribute that contains sensitive information, such as pay, illness and so on and Non Sensitive attribute, which is public attribute. Generalization and hidden are two common anonymity method. Hidden means sensitive attribute is hidden, that is the sensitive attribute not publication, and the attacker cannot see the hidden attributes. But in some cases there must be the need to disclose the sensitive attribute. For that cases generalization and suppression are applied to the quasi identifiers. Generalization is a summary of the raw data, for example, range instead of the specific of data values, although generalization to protect privacy, but there the problem of data missing. So here used suppression to the quasi identifiers and not all the data are suppressed, this will prevent distortion of the data.

Next is K-Anonymity privacy preserving model requires each quasi-identifier attribute equivalence group consists at least K- value records in the publication of data table. Using the quasi identifier set firstly generate the hierarchy. In the hierarchy K-Anonymity is checked. Although K- Anonymity model protects privacy leak to certain extent, but if the attacker knows the background detail then he/she will accurately pinpoint the individual. So here relationships to sensitive attribute must be also considered. If the frequency of sensitive attribute of the equivalence group less than alpha (alpha $\in (0,1)$) based on k-Anonymity. Firstly the alpha value is set as 0.5. when the alpha value decreases as much the privacy increases.

Definition : Alpha Constraint : The raw data set D can be divided into n equivalence class E_i (i $\in$[1,n]), the equivalence class E_i containing N_i records, N_i records containing n_i sensitive attributes, the frequency $(E_i , S) = n_i / N_i$. Alpha is a parameters that user defined data, alpha $\in$ (0,1), if the equivalence class E_i is alpha constraint, then the equivalence class E_i has frequency (E_i , S) < alpha. In (Alpha, K) – Anonymity using Classification, Alpha is the user specified threshold. It has the occurrence between 0 and 1. For example if alpha is set as 0.5 means that the frequency of sensitive attribute of the equivalence group is less than 0.5. That is if there are 10 rows in the equivalence class then the frequency of the sensitive attribute is less than 5. For different groups of people sensitive attributes are different, But after studying with different alpha values here give a classification to the sensitive attributes and each classification have different alpha values. For highly sensitive attributes alpha must be less than or equal to 0.5. But for medium sensitive attributes alpha is set to be greater than 0.5. Inside the alpha, K – Anonymity the local recoding algorithm is used. Local recoding algorithm is otherwise called as top down approach. The architecture is given in figure 1.1

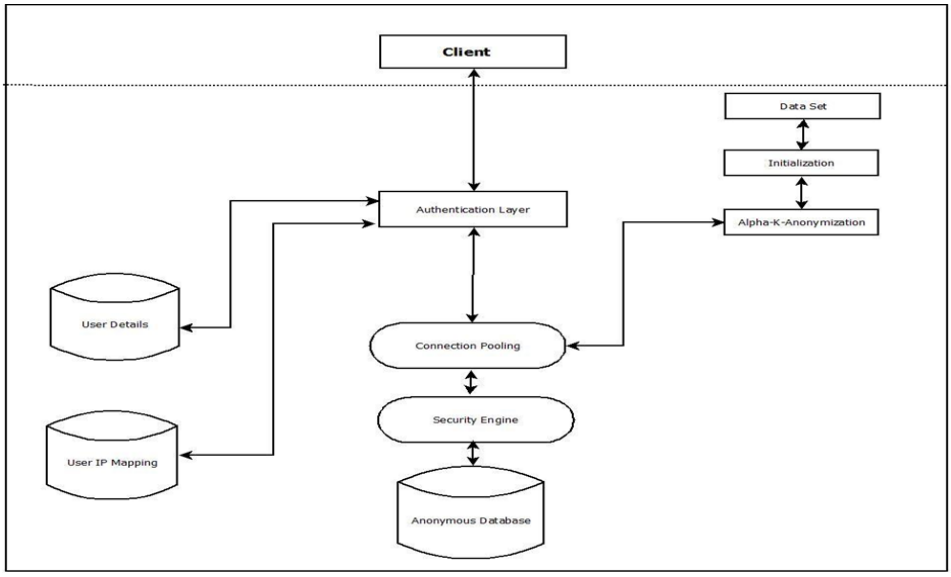

Figure 1.1 Architecture of Proposed system with Alpha, K – Anonymity

V. LOCAL RECODING ALGORITHM

The extended Incognito algorithm is an exhaustive global recoding algorithm which is not scalable and may generate excessive distortions to the data set. Here use a scalable local-recoding algorithm called top-down approach. A top-down approach to tackle the problem. First present the approach for a quasi-identifier of size 1. Then, the method is extended to handle quasi identifiers of size greater than 1. The idea of the algorithm is to first generalize all tuples completely so that, initially, all tuples are generalized into one equivalence class. Then, tuples are specialized in iterations [3]. During the specialization, we must maintain (α, k)-anonymity. The process continues until we cannot specialize the tuples anymore. Let us illustrate with an example in Table 1.1. Suppose the quasi- identifer contains Postcode only.

Gender	Birth	Postcode	Sensitive value
male	May 1990	4351	HIV
male	Jun 1988	4351	Cancer
male	Jul 1985	4351	HIV
male	Aug 1985	4352	Cancer

Table 1.1 Sample Data Set

Gender	Birth	Postcode	Sensitive value
male	May 1990	4351	HIV
male	Jun 1988	4351	Cancer
male	Jul 1985	435*	HIV
male	Aug 1985	435*	Cancer

Table 1.2 Generalized Data set

Assume that α = 0.5 and k = 2. Initially, generalize all four tuples completely to an equivalence class with Postcode = **** (Figure 1.2 (a)). Then, specialize each tuple one level down in the generalization hierarchy. We obtain the branch with Postcode = 4*** in Figure 1.2 (b). In the next iterations, obtain the branch with Postcode = 43** and the branch with Postcode = 435* in Figure 1.2 (c) and Figure 1.3 (d), respectively. As the Postcode of all four tuples starts with the prefix ”435”, there is only one branch for each specialization of the postcode with prefix ”435”. Next, it can further specialize the tuples into the two branches as shown Figure 1.3 (e). Hence the specialization processing can be seen as the growth of a tree. If each leaf node satisfies (α, k)-anonymity, then the specialization will be successful. However, it may encounter some problematic leaf nodes that do not satisfy (α, k)-anonymity. Then, all tuples in such leaf nodes will be pushed upwards in the generalization hierarchy. In other words, those tuples cannot be specialized in this process. They should be kept unspecialized in the parent node. For example, in Figure 1.3 (e), the leaf node with Postcode = 4352 contains only one tuple, which violates (α, k)-anonymity, where k = 2. Thus, we have to move this tuple back to the parent node with Postcode = 435*. See Figure 1.3 (f). After the previous step, move all tuples in problematic leaf nodes to the parent node. However, if the collected tuples in the parent node do not satisfy (α, k)-anonymity, we should further move some tuples from other leaf nodes L to the parent node so that the parent node can satisfy (α, k)-anonymity while L also maintain the (α, k)-anonymity. For instance, in Figure 1.3 (f), the parent node with Postcode = 435* violates (α, k)-anonymity, where k = 2. Thus, should move one tuples upwards in the node B with Postcode = 4351 (which satisfies (α, k)-anonymity). In this example, move tuple 3 upwards to the parent node so that both the parent node and the node B satisfy the (α, k)-anonymity. Finally, in Figure 1.4 (g), obtain a data set where the Postcode of tuples 3 and 4 are generalized to 435* and the Postcode of tuples 1 and 2 remains 4351. Then call the final allocation of tuples in Figure 1.4 (g) the final distribution of tuples after the specialization. The results can be found in Table 1.2. In this approach, sometimes it have to unspecialize some tuples which have already satisfied the (α, k)-anonymity.

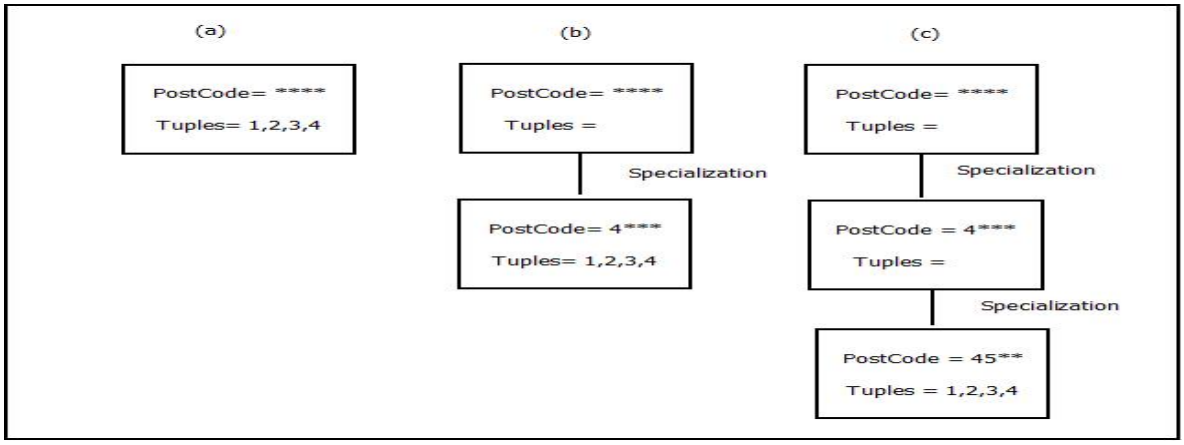

Figure 1.2: First Three Steps of Top Down Algorithm For Quasi Identifier Size1

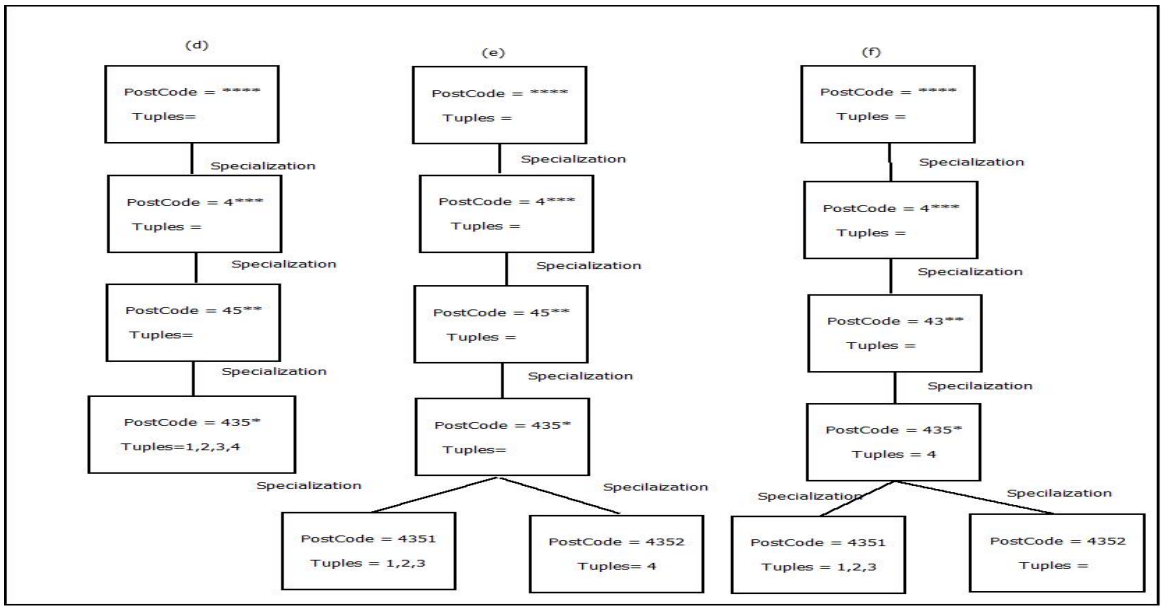

Figure 1.3: Next Three Steps of Top Down Algorithm For Quasi Identifier Size 1

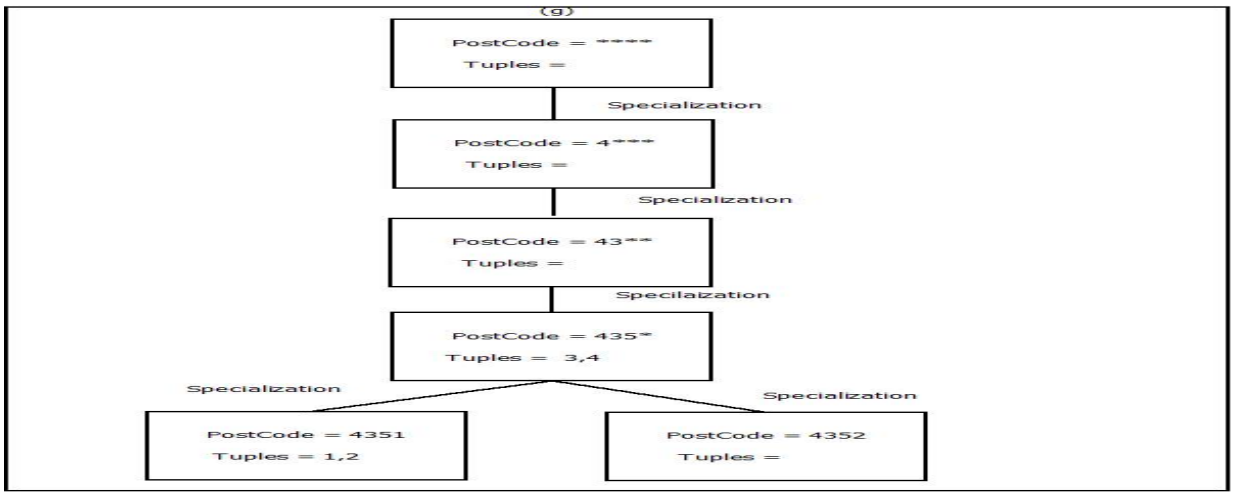

Fig 1.4: Final Step of Top Down Algorithm with Quasi Identifier size 1

For getting better security adding a security engine between the database and connection pooling. For security engine use AES encryption. AES stands for Advanced Encryption Standard Algorithm that is based on several substitution , permutations and linear transformations.

VI. APPLICATION – BANK DATA SET

While the motivation for sharing is clear, a real-world concern is that Big Data applications are related to sensitive information, such as banking transactions and medical records. Here bank data is taken as the synthetic data set. In the bank data set there are 45211 tuples. That is the details of 45211 customers are there. In the bank data set post code is set as Quasi Identifier that is quasi identifier of size 1. The sensitive attribute of the bank data set is the Cibil Value. The Credit Information Report (CIR) summarizes your payment history of loans and credit cards borrowed from all banks and financial institutions. Based on this credit history, a 'Credit Score' is generated. The CIBIL TransUnion Score is a 3-digit number ranging from 300-900. The closer your score is to 900, the stronger your credit profile. A Credit Score plays a critical role in the loan and credit card approval process. This is the first screening criterion applied by banks and financial institutions when reviewing your loan application.

When the cibil value is high that is greater than 800, that person has high transaction roles in bank and that person is up to date in the banking transactions. So the details regarding that person have much importance and privacy is an aspect for that person so while applying Alpha,K – Anonymization with classification on bank data set, here classifying the cibil values to a range such that values between 800 – 900 comes under Class A category and for that class the alpha value set is 0.4. The next classification is 700 – 799 comes under Class B and for that category the alpha value set is 0.5. The third classification is between 600 – 699 comes under Class C and for that class the alpha value set is 0.6. The fourth classification is between 500 – 599, which comes under Class D and for that class the alpha value set is 0.7.For the value of K, here there are one quasi identifier and the data set contains 45211 tuples. For big data the tuples are of huge amount. So the K value set for the data set is 250. So for class A the value for alpha and K are (0.4,250) , For class B the value are (0.5, 250), for class C the value are (0.6, 250) and for class D the values are (0.7,250).

Here the application is created as a web service, where the user has the provision to register. While registering, the ip address is taken automatically and a user has the provision to add 5 Ip addresses. Accessing the anonymized data base is limited to these ip addresses only. User has the provision to add and delete the Ip addresses. After registering the user has the provision to enter the system and view the anonymized data base. The server part is deployed in cloud. Client part is on client system. The data are stored in the cloud database as encrypted format for better security. Only authorized people have the provision to see the anonymized database. This application is applicable in different fields such as medical data set, survey data sets, bank data sets and so on.

The hardware requirement for the application is 1GB cloud storage in any cloud, minimum intel core i3 processor with 2 GHz and 2GB RAM. The data base used is MongoDB. MongoDB is an open source NoSQL scalable, distributed database used in the big data platforms. MongoDB focuses on storage and efficient retrieval of data. MongoDB has its own MapReduce frame work. The IDE used is eclipse and server used is Tomcat 7.0. Cloud computing provides flexible infrastructure and high storage capacity for BigData applications.

VII. EXPERIMENTAL ANALYSIS

Intel core i5- 480M 2.66GHz with 4GB RAM was used to conduct the experiment. The algorithm was implemented in Jave. For experiment, here adopted the publicly available bank data set. We eliminated the records with unknown values. The resulting data set contains 45,211 tuples. One of the attribute was chosen as the quasi-identifier (Postcode). On default, we set k = 250 and α = 0.5. The sensitive value is set as Cibil value. Here evaluated the proposed algorithm in terms of one measurement: execution time. Here conducted the experiments five times and took the average execution time.

In fig 1.5 the X axis represents different K values and the Y axis represents Execution time in seconds. Here the proposed algorithm takes less execution time than the existing K- Anonymity algorithm. Also proposed Alpha, K – Anonymity with

classification takes less execution time than original Alpha, K –Anonymity Algorithm. But when the k value increases the execution time also increases.

In fig 1.6 the X axis represents different alpha values and the Y axis represents Execution time in seconds. Here K is set to be 250 as default. Here the proposed algorithm takes less execution time than the existing K- Anonymity algorithm when alpha value increases. This is because, when α increases, the number of candidates, (representing the generalization domain) increases, and thus the execution time increases. Also proposed Alpha, K – Anonymity with classification takes less execution time than orginal Alpha, K –Anonymity Algorithm.

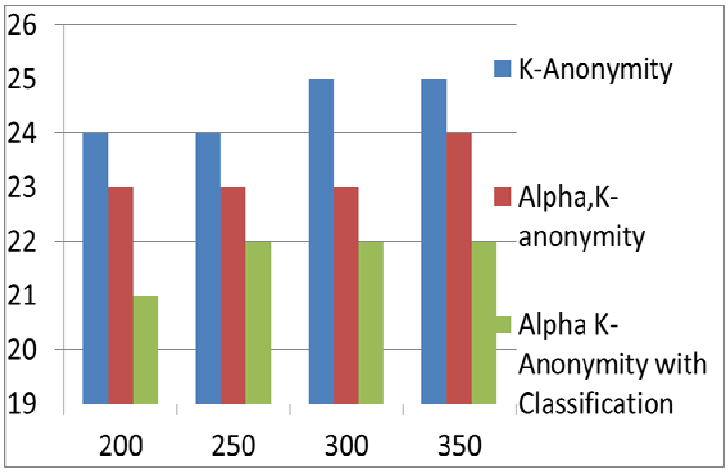

Fig 1.5 K values V/s Execution Time

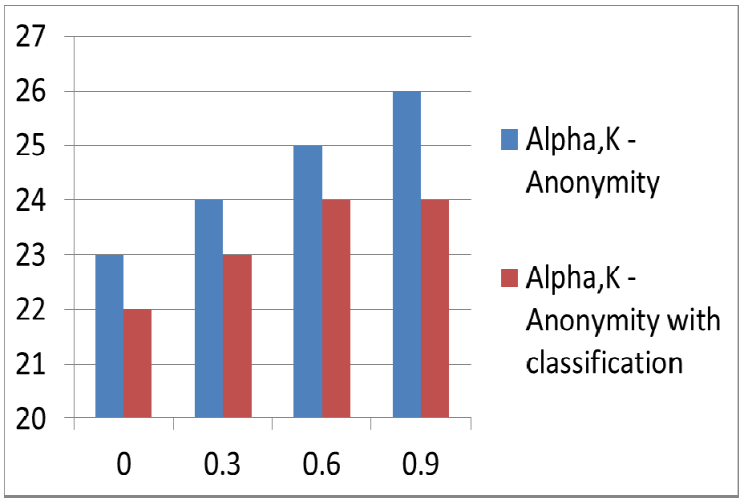

Fig 1.6 Alpha values V/s Execution Time

VIII. Conclusion And Future Enhancement

With the rapid development of information technology and the wide application of networks, large-scale of digital information is stored and published, and knowledge discovery and data mining applications in information retrieval have played an active role gradually, which greatly contributed to the various departments from the massive data mining of useful information needs. At the same time it also brings many problems regarding the privacy, the disclosure of sensitive information has become prominent nowadays, and privacy preservation has become a research hotspot in the field of data privacy and security of big data environment. Big Data mining is the ability of extracting constructive information from huge streams of data or datasets, that due to its variability, volume, and velocity. Data mining includes exploring and analyzing big quantity of data to locate different molds for big data. So while doing data mining in big data, data privacy is an important issue. Among all the algorithms of privacy preservation in data mining, K-anonymity is a kind of common and valid algorithm in privacy preservation, which can effectively prevent the loss of sensitive information under linking attacks, and it is widely used in

various fields recent years. But the recent studies prove that K-anonymity has not as much efficiency to prevent the sensitive information, it has the chances to lose of original data while publishing, so we need a better algorithm for the sensitive data publishing. So in this paper reach a conclusion that there is better algorithm which is more powerful than K-Anonymity, is (∞,K)- Anonymity with classification and while implementing that algorithm for proposed work it increase the efficiency as well as privacy than K- Anonymity. Also adding the security engine helps to improve the data security in database.Here K-Anonymity and Alpha, K –Anonymity algorithm assumes quasi identifier size as one. But also quasi identifier size can be taken as a set of attributes. So in future the quasi identifier can be set as 2 or 3.When the quasi identifier size increases the complexity and running time of the algorithm also increases. The complexity of the algorithm increases because for big data environment handles huge amount of data sets. If the size increases as much the hierarchy generation and correlation of algorithm will take much complex. Also classification can be given to K values as such as Alpha values. This will also increase the complexity.

References

1. "Data Mining With Big Data "Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, And Wei Ding, Senior Member, IEEE, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.
2. "Providing Data Anonymity for a Secure Database Infrastructure", Traian Popeea, Anca Constantinescu, Razvan Rughinis "Politehnica" University of Bucharest Bucharest, Romania, 2012.
3. "(α, k) Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing" Raymond Chi Wing Wong ,Department of Computer Science and Engineering , The Chinese University of Hong Kong , Jiuyong Li and Ada WaiChee Fu, Department of Mathematics and Computing , The University of Southern Queensland and Ke Wang Department of Computer Science, Simon Fraser University, Canada. 2006
4. "Generalization Based Approach to Confidential Database Updates" Neha Gosai, S.H.Patil, International Journal of Engineering Research and Applications, June 2012.
5. "Efficient Multidimensional Suppression for K-Anonymity" Slava Kisilevich, Lior Rokach, Yuval Elovici, Member, IEEE, and Bracha Shapira, IEEE Transaction march 2010.
6. "Privacy-Preserving Updates to Confidential and Anonymous Databases" Alberto Trombetta, Wei Jiang, Elisa Bertino and Lorenzo Bossi, Purdue University 2007.
7. "Supporting Pattern-Preserving Anonymization for Time-Series Data" Lidan Shou, Xuan Shang, Ke Chen, Gang Chen, and Chao Zhang, IEEE transaction April 2013.
8. "Anonymization by Local Recoding in Data with Attribute Hierarchical Taxonomies", Jiuyong Li, Member, IEEE, Raymond Chi-Wing Wong, Student Member, IEEE, Ada Wai-Chee Fu, Member, IEEE, and Jian Pei, Senior Member, IEEE, September 2008.
9. "K-Anonymity for Crowdsourcing Database", Sai Wu, Xiaoli Wang, Sheng Wang, Zhenjie Zhang and Anthony K.H. Tung, IEEE Transaction 2013.
10. "Anonymizing Classification Data for Privacy Preservation",Benjamin C.M. Fung, Ke Wang, and Philip S. Yu, Fellow, IEEE, may 2007.
11. "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using Map Reduce on Cloud", Xuyun Zhang, Laurence T. Yang, Senior Member, IEEE, Chang Liu, and Jinjun Chen, Member, IEEE , February 2014.
12. "Anonymous Publication of Sensitive Transactional Data" Gabriel Ghinita, Member, IEEE, Panos Kalnis, and Yufei Tao, IEEE transaction February 2011.
13. "Accuracy-Constrained Privacy-Preserving Access Control Mechanism for Relational Data", Zahid Pervaiz, Walid G. Aref, Senior Member, IEEE, Arif Ghafoor, Fellow, IEEE, and Nagabhushana Prabhu, IEEE Transaction April 2014.
14. " k-Anonymization with Minimal Loss of Information", Aristides Gionis and Tamir Tassa, IEEE Transactions on knowledge and data engineering, February 2009.
15. Review on "Data Mining with Big Data" Vitthal Yenkar, Prof.Mahip Bartere, IJCSMC, Vol. 3, Issue. 4, April 2014.
16. Enhancing Data Privacy in Data Extraction with Big Data", Melwin Devass, Gera.Praveen Kumar, ICETETS 2014.
17. "Scalable Privacy Preservation in Big Data A Survey", Vennila.S, Priyadarshini. J, Department of computing science and Engineering ,Vellore Institute of Technology, Chennai, India. Science direct, 2015.
18. "Survey on Data Privacy in Big Data with K- Anonymity", Salini. S, Sreetha .V.Kumar, Neevan.R, Department of computer science and engineering, Marian Engineering college, Trivandrum and College of engineering, Kottarakara, Kollam, IJIRCCE,volume-3Iissue 5 May 2015.

