ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Privacy Preservation on Big Data using PK-Anonymization

Nikkath Bushra.S, Dr.A.Chandrasekar

Research Scholar, Dept. of CSE, Bharath University, Chennai & Associate professor, St.Joseph's College of Engineering Chennai, Tamil Nadu, India

Professor, Dept. of CSE, St.Joseph's College of Engineering,Chennai, Tamil Nadu, India

ABSTRACT: Anonymization technology is essential for achieving protection on privacy when using personal data. In the age of bigdata a great deal of information has been accumulated in the world. There are issues where in individual are Identified by matching with other data. Anonymization in bigdata is a challenge to convert personal data into non personal data. With the help of the map reducing framework the large number of companies and organizations to process huge-volume data sets. Privacy preservation and high utility of Data is possible due to the map reducing framework and the PK-Anonymization Technology.

K-Anonymity is a privacy property used to limit the risk of re-identification in a micro data set. A data set satisfying K-anonymity consists of groups of k records which are indistinguishable as far as their quasi-identifier attributes are concerned. Hence, the probability of re-identifying a record within a group is 1/k. Probabilistic k-anonymity property, which relaxes the in distinguishability requirement of k-anonymity and only requires that the probability of re- identification be the same as in k-anonymity. K-anonymity is achieved (mathematically guaranteed) only with stochastic displacement of data. Pk-anonymization is the world's first randomization method with safety equivalent to K- anonymization. Machine learning provides correct data by estimating data suitable for analysis, using the parameter of randomization.

KEYWORDS: K-Anonymization, PK-anonymization, Map Reducing, Bigdata, Quasi Identifiers.

I. Introduction

Cloud computing and Big Data [6] [7] are the two disruptive trends at present which poses significant influence on current IT industry and research communities. Cloud computing provides massive computation power and storage capacity which enable users to deploy applications. Big data is a broad term used for large and complex dataset that traditional data processing applications are inadequate to perform analysis, capture, duration, search, sharing, storage, transfer, visualization and information privacy .Data anonymization is a type of information sanitization whose intent is to protect data. Data anonymization enables to transfer information across a boundary, such as between two departments within an agency or between two agencies, while reducing the risk of unintended disclosure in certain environments in a manner that enables evaluation and analytics post anonymization. The map reducing framework has been widely adopted by a large number of companies and the organizations to process a huge-volume of datasets.

The most common technique being used to anonymize a given dataset is generalization [3] [4] [5] and suppression. In multidimensional space, the counter part of these operations is replacing a set of points with the minimum bounding box that covers the points. PK-anonymity is the world's first randomization method with safety equivalent to k- anonymization. Machine learning provides correct data by estimating the data suitable for analysis using the parameters of randomization. Big Data cannot be fit in the memory in one normal cloud computation node and usually stored across a number of nodes. Anonymity and utility of personal data is balanced in PK-anonymization.

II. Related work

In the paper Privacy Preservation over Big Data in Cloud Systems referred by Xuyun Zhang, Chang Liu, Surya Nepal, Chi Yang and Jinjun Chen they perform k - anonymation technique for privacy [2]. In the paper Balancing

DOI: 10.15680/IJIRCCE.2015. 0311241

Utility with Anonymity of Data Satoh and Takahashi suggests PK-anonymization for high data security. It provides high security rather than K anonymization .

The Several categories of anonymization techniques have been proposed including generalization and suppression. Generalization method is used for anonymization is widely investigated and adopted in existing algorithms. To encrypt and process on large data sets efficiently will be quite a challenging task. Scalability is necessary for current privacy- preserving approaches, because the scale of data sets is too large to be processed by existing centralized algorithms. This reduction is a trade off that results in some loss of effectiveness of data management or mining algorithms in order to gain some privacy.

III. PROPOSED ALGORITHM

The datasets are stored confidentially by data owners and can never be accessed by the data users. The data holders specify privacy requirements and submit them to the privacy preservation framework. Fig 1 specifies the overall performance that are done in the PK-Anonymization.

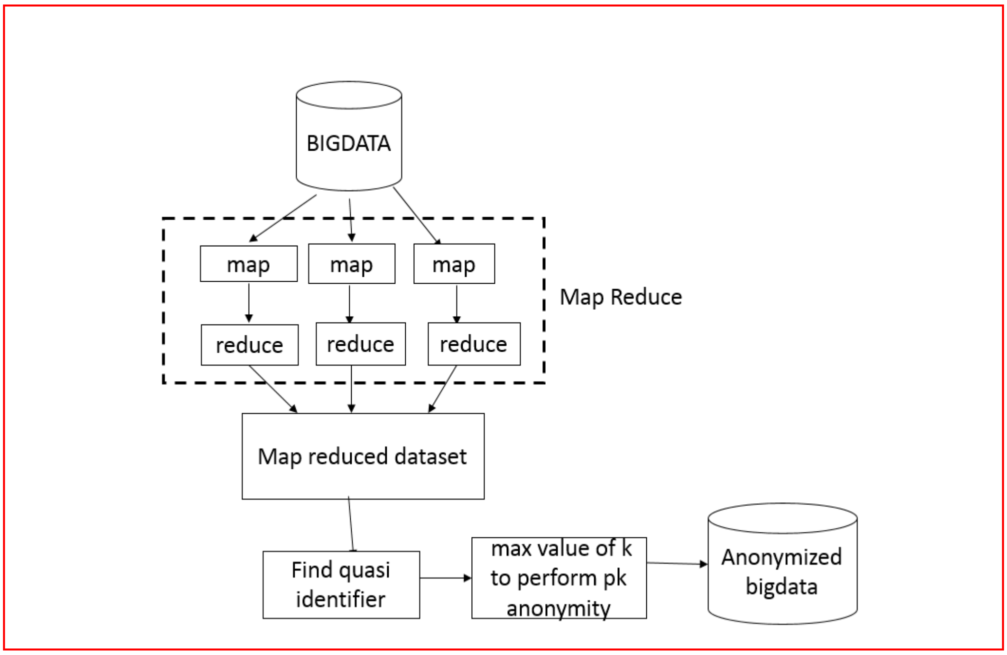


Fig 1. Architecture diagram

First the Bigdata undergo a process called map reducing, it provides the data will not allow the duplicate data it contains only the consistent data without duplication. And only the consistent data will go for the process of PK-anonymization. Map reducing [8] consists of three process to perform the consistent data. It includes mapping, shuffling and reducing the data. The resultant set of the map reducing data can undergo the process of the PK -anonymization. First to identify the quasi identifier. Quasi-identifiers [12] are pieces of information that are not of themselves unique identifiers, but sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier.

Quasi-identifiers can thus, when combined, become personally identifying information. This process is called re-identification. Motwani and Ying warn about potential privacy breaches being enabled by publication of large volumes of government and business data containing quasi-identifiers. The Quasi-identifier is used for the anonymization of the bigdata. Here anonymizing the data with the help of the PK- anonymization. PK-anonymization is the world’s first randomization method with safety equivalent to k- anonymization[1]. More than one quasi identifier can be used in the process of anonymization. But the anonymization level must be good for the better anonymization. The maximum value of k must be provided specified the value must be small to avoid the complexity of anonymization. If it becomes higher than it is difficult to anonymize the dataset.

Algorithm Pk –Anonymization

Randomization: k-anonymity is achieved (mathematically guaranteed) only with stochastic displacement of data. PK-anonymization is the world's first randomization method with safety equivalent to k-anonymization.[1] Machine learning provides correct data by estimating data suitable for analysis using the parameters of randomization. Anonymization in big data is requested under the current interpretation of the legal system, behind which are a variety of needs for free use of personal data by making them anonymous. It is possible to process personal data into a state that has almost entirely eliminated individuality included in data.

k-Anonymity guarantees that, for any combination of values of quasi-identifier attributes in the published micro dataset T0(A1,...,An), there are at least k records sharing that combination of values. Therefore, given an individual in an external non-anonymous data set, the probability of performing the right linkage back to the corresponding record in the published micro data set, and thus the probability of learning its confidential attributes, is at most 1/k. It is in this sense that probabilistic k-anonymity is defined.

B- Input Bigdata set q- quasi identifier Pi- probability value of i Pj- probability value of j
Pk- original probability value to perform anonymization
B*-anonymized Bigdataset

During anonymization there will be a loss of information. When performing anonymization the original value in the record will be changed to provide security. It must be minimized to provide good anonymization. More the loss of information means the anonymization level will be damaged. It also depends on the quasi identifier the possible value in the quasi identifier is more there will be more loss of information. Loss of information provides that how much data lost during the anonymization will be calculated by

Information loss: (IL)=a(d)/t(d) * 100

a(d) represents the anonymised data set

t(d) represents the total data set

If the loss of information value is below 50% then the anonymization is good. Balancing both information loss and the level of the anonymization is the key. Sometimes the loss of information is minimum but the level of anonymization is not sufficient. So that the security of the bigdata will be lost. To perform good anonymization maintain the level of anonymization and the hierarchy will be minimum to provide good encryption on the bigdata. Before the calculation on the loss of the information the elapsed time will be calculated to perform the operation.

IV. PSEUDO CODE

```
 Begin
1. Input the big data as B and the level to generalize
2. Specify the quasi identifier q(i)
3. If B(i) have duplicate value
       3.1 Call mapreduce()
4. Else
       4.1 Then sort the dataset
       4.2 Pi=quasi value of i/n
       4.3 pj=quasi value of j/n Pk=max(pi,pj)
 5. End if
 6. Change the pj value to pk value
```

Big data set having anonymised data set of B* value

End

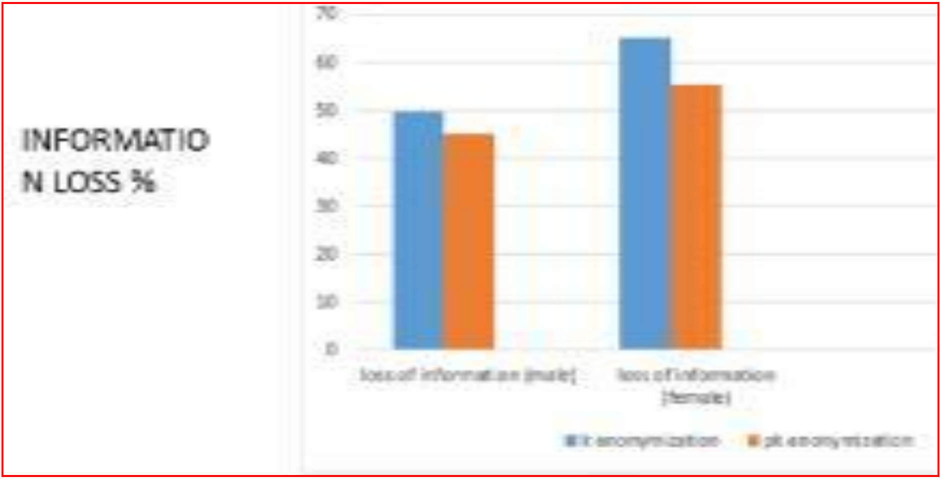
Here, first the data will be in the form of ascending order depends on the quasi identifier to provide meaningful data. After that the map reducing will be passed to avoid the duplicate data

V. Simulation Results

1. Experimental results on Data Anonymization:

By comparing both the approaches. PK-anonymization provides less information loss than the k anonymization. And run both the approaches from 25KB to 100KB. Hence, the privacy-preserving framework can significantly improve the capability and efficiency compared with existing state-of-the-art anonymization approaches. (k anonymization method). In this chart clearly demonstrates the comparison between the k and the PK-anonymization. And finally the loss of information must be below 50% to provide better anonymization. Here the y- axis will be the loss of information in %, x- axis will be the quasi identifier value. To evaluate the main components of the privacy preservation framework via conducting the experiments on real world data sets. Using the public hospital dataset for the process of the anonymization. Quasi identifier will be identified correctly for the better anonymization. Comparing with k anonymization, the PK-anonymiation having better utility of the data with the low loss of information. It will be shown on the below Fig 2. The centralized approach suffers from the memory insufficiency when the data size is more than 500MB.

Fig 2 Experimental results on Data Anonymization



2. Experimental results on Execution Time:

Execution time is the time during which a program is running (executing), in contrast to other phases of a program's lifecycle such as compile time, link time and load time. If the record in the bigdata is small then the run time of K-anonymization is better than the PK- anonymization but if the record becomes more and more the time will be greater than the PK- anonymization. By running both the approaches from 200KB to 600KB the PK- anonymization execution time is less when there is more number of datasets as shown in the Fig 3. The privacy preservation monetary cost is measured to evaluate the cost effectiveness. It is calculated by

ElapsedTime = system.nanoTime() – startTime()

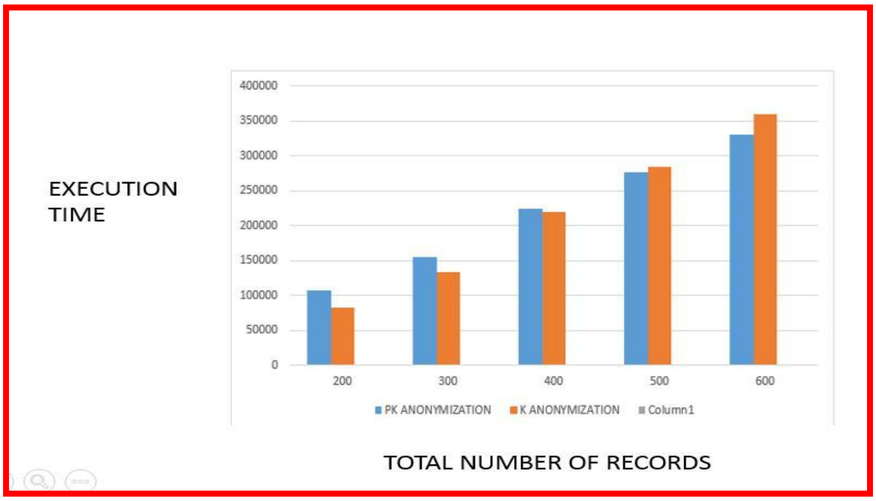


Fig 3. Experimental results on Execution Time

VI.CONCLUSION AND FUTURE WORK

Anonymization techniques result in distortions the data. Excessive anonymization may reduce the quality of the data making it unsuitable for some analysis, and possibly result in incorrect or biased results. Therefore, it is important to balance the amount of anonymization being performed against the amount of information loss. It is therefore important to understand precisely the types of re-identification attacks that can be launched on a data set and the different ways to properly anonymize the data before it is disclosed

Here we Propose a flexible, scalable, dynamical and cost-effective privacy-preserving framework based on Map Reducing on cloud called PK-Anonymity. The privacy-preserving framework can anonymize large-scale data sets and manage the anonymous data sets in a highly flexible, scalable, efficient and cost-effective fashion. This project provides flexible privacy framework on traditional Bigdata and not for streaming of data. With the help of the **storm** the streaming of data can be updated very effectively. And also the unique id will be anonymised with the help of the storm tool. Several data processing framework will be integrated to perform the anonymization more effectively. PK-anonymization can be used to anonymize different bigdata set in an effective manner.

REFERENCES

[1] http://www.nii.ac.jp/userdata/results/pr_data/NII_Today/6 4_en/p10-11.pdf

[2]. http://link.springer.com/chapter/10.1007/978-3-642-3858 6-5_8#page-1 Privacy Preservation over Big Data in Cloud Systems by Xuyun Zhang, Chang Liu, Surya Nepal, Chi Yang and Jinjun Chen

[3]. Fung, B.C.M., Wang, K., Chen, R.,Yu, P.S.:Anonymizing classification data for privacy preservation, IEEE Trans. Knowl. Data Eng. 19(5) (2007) 711–725.

[4]. P.Samarati, Protecting Respondents' Identities in Microdata Release, IEEE Transactions on Knowledge and Data Engineering, v.13 n.6, p.1010-1027, November 2001.

[5]. Xu, J., Wang, W.,Pei, J., Wang, X., Shi, B., Fu, and A.W.C.: Utility-based anonymization for privacy preservation with less information loss. ACMSIGKDD Explor. Newsl.8 (2) (2006)21–30.

[6]. Azza Abouzeid , Kamil Bajda-Pawlikowski , Daniel Abadi , Avi Silberschatz , Alexander Rasin, HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads, Proceedings of the VLDB Endowment, v.2 n.1, August 2009.

[7]. Wang, L., Zhan, J., Shi. W., Liang.Y: In cloud, can scientific communities benefit from the economies of scale?. IEEE Trans. Parallel Distrib.Syst.23(2) (2012) 296–303.

[8]. Dean, J., Ghemawat, S.:MapReduce: A flexible data processing tool.Commun. ACM 53(1) (2010) 72–77.

[9]. Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D,Zhu A: Approximation algorithms for k-anonymity. Journal of Privacy Technology (2005), paper number 20051120001.

[10]. Yufei Tao, Hekang Chen, Xiaokui Xiao, Shuigeng Zhou, Member, IEEE Computer Society, and Donghui Zhang, ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication. IEEE Transaction on Knowledge and Data Engineering. Vol 21.No.7.pp.1073-1087 (2009).

[11]. T. M. Truta, A. Campan and P. Meyer. “Generating Micro data with p-sensitive k-anonymity Property”,SDM

2007:124-141.

[12]. Zhang X, Liu C, Nepal S, Chen J (2013) An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud. J Comput Syst Sci 79(5):542–555

[13]. LeFevre, K.., DeWitt, D.J., Ramakrishnan, R.: Workload-aware anonymization techniques for large-scale datasets. ACM Trans. Database Syst.33(3) (2008)1–47.

BIOGRAPHY

S. Nikkath Bushra received her Bachelor's degree from the University of Madras, Chennai, Tamil Nadu ,India . M.C.A from the University of Madras, Chennai, Tamil Nadu, India, M.E degree (Computer Science and Engineering) from the Sathyabama University ,Chennai, Tamilnadu India and M.Phil from Mother Teresa University, Kodaikanal, Tamil Nadu, India from 2008 to 2009, she worked as a Database Administrator in IBM. She is currently an Associate Professor in the Department of Computer Science and application at St. Joseph's College Of Engineering, Chennai and She is a research scholar of Bharat University, Chennai, Tamil Nadu, India. Her research interests are in cloud computing, privacy preservation in cloud computing and Data mining.

Chandra Sekar A, received his B.E degree from Angala Amman College of Engineering and Technology affiliated to Bharathidasan University, M.E degree from A.K. College of Engineering affiliated to Madurai Kamaraj University, and Ph.D in Information and Communication Faculty (Computer science & Engineering) from Anna University, Chennai, India. He is currently working as a Professor in the Department of Computer Science & Engineering in St. Joseph’s College of Engineering, Chennai. His area of interest includes Network Security and Analysis of Algorithms. Life Member in ISTE and fellow membership from International Science congress Association.

DOI: 10.15680/IJIRCCE.2015. 0311241