Indian Journal of Science and Technology, Vol 10(10), DOI: 10.17485/ijst/2017/v10i10/109294, March 2017
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Anonymizing Sensitive Based Qualities Utilizing Hadoop Framework

Shivaprakash Ranga*, Balakrishnan and Nageswara Guptha

Department of ISE, Sri Venkateshwara College of Engineering, Vidyanagara Cross, Bangalore International Airport Road, Bettahalasur Post Yelahanaka, Bengaluru – 562157, India; shivaprakashranga@gmail.com, Bala.k.btech@gmail.com, mnguptha@yahoo.com

Abstract

Objectives: To propose an approach thereby only data owner can view the records by fetching from cloud using decryption and the anonymized records can be made available for the researchers to understand the patterns for a particular disease. **Methods**: Proposed work is implemented by using Hadoop Map-Reduce Framework. To offer personal privacy, it uses two attributes: Sensitive Disclosure Flag (SDF) and Sensitive Weight (SW) which is marked on the basis of personal privacy. **Findings**: Only data owner can view the records by fetching from cloud using decryption, the anonymized records are made available for the researchers to understand the patterns for a particular disease. Our proposed method is the enhancement of previous anonymity techniques, initially each patient IDs in the medical record table is changed using encryption method and then the records are generalized to make anonymity and at the end, original and anonymized data records are stored in cloud which eliminates the big data storage problem at local system.

Keywords: BigData, Map-Reduce, Personal Anonymization

1. Introduction

Big data[1] contrasts from conventional information warehousing endeavors. It performs information examination on any sort record including pictures, sounds, recordings, and information created from web-based social networking. Security safeguarding of information is critical point with regards to Enormous Information. Distributed computing has turned into a practical, standard answer for huge information handling. Processing a lot of informational collections back and forth from the cloud prompts to a security issues among associations. To enhance the distinguishing proof of the analysis persistent information are available in various healing facilities and it is used by researchers to understand the pattern for disease and to treat it in a better way to the society.

The medicinal information which is available in hospital comprises of name, DOB, zip code, address, symptoms and disease. To safeguard the character or protection of a patient, healing facility administrator erases the subtitle elements of a patient which prompt to the security break after conveying the patient information into the cloud. Consider the Understanding Information Distributed by the healing facility which does not contain insights with respect to name and address. Presently, assailant can utilize the openly accessible outer information and base Table which can execute join question that may uncover individual points of interest.

Table 1. Patient published data

PID	Age	Zipcode	Disease
1	26	556171	Brain tumor
2	28	556176	Stomach Cancer
3	32	556362	Flu
4	36	556175	Heart Disease
5	42	556178	Heart Disease
6	46	55S6177	Stomach Cancer

Joining Table 1 and Table 2 gives the value Rajesh from zip code 589171 and age 26 is having Brain tumor which results to a Record Level Disclosure. Attributes

*Author for correspondence

present in Patient Data which is published that can be linked to external publicly data bases like ZIP, DOB are called Quasi-Identifier (Q) attributes.

Table 2. External voter's data base.

Name	Age	Zipcode
Rajesh	26	556171
Mani	28	556176
Seema	32	556362
Indu	36	556175
Sneha	42	556178
Suraj	46	556177

Generalization[3,4] resultant tables has duplicate records there by restricting the disclosure of the patient. In[6] introduced k-anonymity which satisfies every set of quasi-identifier attributes if every record in the table is indistinguishable from at least k-1 other records.

Table 3. 2-anonymus table

Zipcode	Age	Disease
556***	[20-30]	Brain Tumor
556***	[20-30]	Stomach Cancer
556***	[30-40]	Flu
556***	[30-40]	Flu
559***	[40-50]	Flu
556***	[40-50]	Heart Disease

Table 3 shows a generalization of 2-anonymous Table 1. If attacker uses publicly available data base and finds Rajesh zip code is 556171 and his age is 26 and wants to know the disease of Rajesh, now the attacker observes the anonymized Table 3 but generalized to 556*** and age [20-30] which cannot be linked with voters database record and hence the disease cannot be revealed. But drawback results in Attribute Level Disclosure if all the diseases indicated in a group are related to the same disease.

The primary thought of this strategy is to add two attributes to the first table i.e., Sensitive Disclosure Flag (SDF) decides sensitive data of the patient is to be uncovered or not from the information proprietor. The second quality is Sensitive Weigh (SW) which indicates how much delicate information from other ailment.

The rest of the paper is organized as follows. In Section 2, the prior works on privacy preserving techniques are illustrated. The proposed scheme is defined in Section 3. Section 4 presents results and analysis. Section 5 concludes the paper.

k-anonymity technique[5,6] is used for anonymization but disadvantage of this technique results in record level disclosure. Many techniques evolved like l-diversity[7] says Privacy beyond k-anonymity but disadvantage leads to Skewness and Back ground Knowledge Attack. t-closeness[8] is used for a larger information loss compare to other approaches. In[8] the record owner uses the guarding node to indicate his/her privacy, but the drawback of this method leads to several iterations based on the guarding node.

The inconvenience of SW-SDF based anonymization proposed[2] is that, it doesn't give information protection or security on database in local system from outer assaults and furthermore nearby database doesn't bolster preparing an extensive arrangement of anonymized records. Additionally putting away the first records and also anonymized records in nearby database builds the local system stockpiling. The proposed work defeats these inconveniences, once the first information records are handled; local data base system contains just patient IDs and name of the patients which expels the weight of information stockpiling in nearby database. The rest of the fields of anonymized records, scrambled patient IDs are spared in cloud. These encoded understanding IDs secures the anonymized records from the outer connecting assaults. For preparing the extensive arrangement of anonymized records proposed work is executed on Hadoop Framework. At that point the both anonymized information and partial anonymized information are put away in cloud and made accessible to analysts.

2. Proposed Work

The propose approach is to make customized anonymization of the extensive scale well being records relying upon the way that, couple of patients are prepared to uncover their sensitive information and a few patients are not prepared to uncover. Handling huge size of information it depend on Hadoop MapReduce Framework which has two levels of parallelization i.e., job level and task level.

The proposed work consists of four modules as shown in Figure 1. They are Admin Module, Mapper and Anonymizer Module, MapBased Reducer Module and Cloud Database Module.

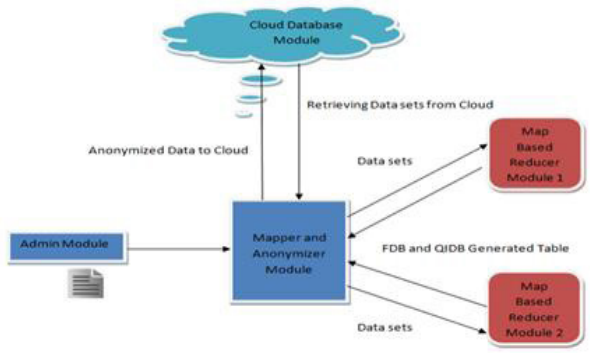

Figure 1. Architecture of the proposed work.

2.1 Admin Module

Doctor's facility manager i.e., Hospital administrator includes two attributes SDF and SW to the original dataset. In the event that SDF = 0 implies the patient is not prepared to uncover his/her sensitive disease taken from the patient while SDF = 1 wouldn't fret uncovering his/her affect ability of the sickness. In proposed plot every patient IDs scrambled as appeared in Table 4. The changed dataset is brought by Mapper and Anonymizer Module and these records are deployed in cloud without SW and SDF section. Administrator keeps in healing facility neighborhood database just a unique IDs and Name. To acquire unique records administrator utilizes an unscrambling strategy.

Table 4. Modified dataset table with SW and SDF values

PID	Age	Zipcode	SW	SDF
er34	24	584170	0	1
rt43	34	584171	0	1
tg48	39	560017	1	1
hy27	35	584173	1	0
bf24	36	584171	0	1
ik25	36	584175	1	0
ik46	25	560018	1	0

2.2 Mapper and Anonymizer Module

The Mapper module segments the dataset among the quantity of Reducers and sends to the Map-Based Reducer Module. The outcome from the diverse MapBased Reducer Module is FDB produced Table 5 and QIDB created Table 6. These two tables are consolidated and final anonymization is finished by running SW-SDF anonymization calculation, taken from[2]. SDF-SW gives personal privacy anonymization by isolating anonymized and partial anonymized information.

Table 5. FDB generated table

Disease	FDB Probability
Flu	0.4
Stomach Caner	0.4
Heart Disease	0.2

Table 6. QIDB generated table

Disease	QIDB Probability
Flu	0
Stomach Caner	0.5
Heart Disease	0.5
Disease	QIDB Probability
Flu	0
Stomach Caner	0.5
Heart Disease	0.5

In paper[2] unveiling the information implies gives no speculation consequently regardless of the possibility that the information is delicate there is no individual security. However, in our proposed conspire unveiled information is summed up with generalized with basic Anonymity and it is called partial anonymized information and the suppressed data is the anonymized data which is generalized with higher Anonymity, appeared in Table 11 and 12 individually. The anonymized and incomplete anonymized set of information with encrypted PIDs is given to Cloud Database module where they are stored in cloud server.

3.3 MapBase Reducer Module

This module is the slave of Hadoop framework. The partitioned datasets from Mapper and Anonymizer module is processed at each MapBased Reducer modules to create FDB and QIDB. FDB contains distribution of each disease as for unique private information. For each record with SW = 1 and SDF = 0 QIDB is made. This module first checks the aggregate number of columns in the datasets table and chooses the remarkable Ailment illustration like Gastric Ulcer, Stomach Growth, influenza, Heart Illnesses and so on. FDB probability is calculated the resultant

probability is included into the FDB table. In the wake of producing FDB table, QIDB table is created just for delicate ailments with SW = 1.

$$FDB\ probability = \frac{\textbf{Number of count on each disease}}{\textbf{Total number of count}}$$

The result from the different MapBased Reducer Module is FDB generated table QIDB generated table are shown in Table 5 and Table 6 which is given to the Mapper and Anonymizer Module for further processing.

$$\text{QIDB probability} = \frac{Number of Count Statisfying. SW = 1, SDF = 0}{Total Number Count of Sensitive Diseases} \quad (2)$$

3.4 Cloud Database Module

Paper[2] all storage was in local system which expands the odds of security misfortune if any assaults occur at nearby framework and further if crash local system may lead in entire information misfortune. In our proposed plot Cloud Database module is included which is implemented on Amazon[10] S3 public cloud. This module stores the original health record with encrypted IDs as shown in Table 7 which is just accessible to the Administrator and furthermore stores the anonymized and partial anonymized records as appeared in Table 8 and Table 9 individually which are accessible for the researchers.

Table 7. Partial anonymized data

Age	Zipcode	Disease	Records
24	554***	Flu	1
34	554***	Flu	1
36	554***	Flu	2
39	560***	Stomach Cancer	2
28	554***	HeartDisease	1

Table 8. Modified dataset table with permuted ID's

PID	Age	Zipcode	Disease
er34	24	584170	Flu
rt43	34	584171	Flu
tg48	39	560017	Stomach Cancer
hy27	35	584173	Heart Disease
bf24	36	584171	Flu
ik25	36	584175	Stomach Cancer
ik46	25	560018	Stomach Cancer

Table 9. Anonymized data

Age	Zipcode	Disease	Records
[30-40]	554***	Heart Disease	1
[20-30]	550***	Stomach Cancer	1
[30-40]	554***	Stomach Cancer	1
[30-40]	554***	Heart Disease	1

4. Result and Analysis

Our proposed approach accomplishes the entire protection for every individual patient by executing individual information anonymization. The information records is smothered and afterward summed up with more nameless i.e., two columns are generalized as appeared in Table 9. Data records which is to be disclosed are summed up with less namelessness i.e., one segment is summed up as appeared in Table 8 These two tables are made accessible for researchers. Just anonymized data sets are conveyed in cloud. Patient unique Ids (PID) and name are just present local database system. The encrypted IDs with every single other detail are put away in cloud as appeared in Table 7. The administrator can unscramble the IDs utilizing the first IDs to get the entire records at whatever point required. In this way, the local database is free from expansive stockpiling of information and inside assault.

5. Conclusions

In working with cloud and Bigdata, Personalized privacy is a critical research heading and SW-SDF is a superior outcome for customized Security. Despite the fact that when disease is normal the persistent necessities security, the records are generalized with basic anonymity. For the delicate diseases, the records are generalized with more anonymity which beats the linkage issue. As the private database contains just patient IDs and name it unravels the burden of storage. PIDs in cloud are put away in encrypted form which conquers record linkage and attribute linkage, thus giving complete individual protection. The anonymized records are made openly accessible for

the researchers to understand the patterns for a particular disease.

6. References

1. Patel AB, Birla M, Nair U. Addressing big data problem using Hadoop and Map Reduce. Nirma University International Conference on Digital Object Identifier; 2012. p. 1–5. Crossref.
2. Kiran P, Kavya NP. SW-SDF based personal privacy with QIDB-anonymization method. IJACSA. 2012; 3(8):60–6.
3. Lefevre K, Dewitt D. Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. Proceedings of ACM SIGMOD ACM; New York. 2005. p. 49–60.
4. Fung BC, Wang K, Yu PS. Anonymizing classification data for privacy preservation. IEEE Trans Knowledge. 2007; 711–25.
5. Samarati P. Protecting respondents privacy in microdata release. IEEE Trans on Knowledge and Data Engineering TKDE. 2011; 13(6):1010–27. Crossref.
6. Sweeney L. k-Anonymity a model for protecting privacy. International J Uncertain Fuzz. 2002; 10(5):557–70. Crossref.
7. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ICDE Conference; 2012. p. 1–12.
8. Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy beyond k-anonymity and l-diversity. Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE); 2012. p. 1–15.
9. Xiao X, Tao Y. Personalized privacy preservation. Proceedings of the ACM SIGMOD Conference; New York. 2006. doi: 10.1145/1142473.1142500.
10. 10. Amazon Elastic, MapReduce. Available from: http://aws:amazon:com/elasticmapreduce