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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

Privacy Preservation of Medical Datasets using Hadoop - A Survey

Balaji Bodkhe*, Meghana Rao, Monish Moryani, Mervyn Chacko, Varada Dharmadhikari

Department of Computer Engineering, MES College of Engineering, Pune, India*

ABSTRACT: As human dependency on computing services has increased in the past decade, privacy of datasets is a major concern. Privacy Preservation techniques need to be implemented in order to protect the datasets from malicious users. Moreover, this needs to be done in fields such as medical sciences and data analytics. Techniques such as data anonymization which includes generalization, bucketization, tuple partitioning are few of the novel methods used for the same. This paper focuses on privacy preservation of medical datasets using Hadoop. This is a survey regarding anonymization techniques used in data processing phase. It also talks about k-anonymity and l-diversity and their respective attacks

KEYWORDS: Big data, Hadoop, Slicing, Medical datasets, Data Anonymization.

I. INTRODUCTION

In the past decade, there has been a colossal increase in medical facilities all over the world. The healthcare industry is a major contributor to Big Data. Patient health data produced by the use of sensors, analysis of insurance claims for detection of anomalies and frauds, human genome data, predictive analysis are a few such contributors. Due to this, Big Data analytics has to be used to cater to its needs. Privacy Preservation of this medical data is also required. Big Data analytics platforms such as Apache Hadoop, YARN and Hive are best suited for the same. Many algorithms have been implemented in various stages of Big Data life cycle. These stages are production of Big Data, storing the data and processing the data. It must also be taken into consideration that the utility of the data is not lost. Various algorithms such as Classification, Clustering, Regression and Artificial Intelligence are used for analysis of Big Data.

Every database consists of Quasi-Identifiers which when used together during analysis can reveal the identity of a person.

Sensitive attributes (SA) are the attributes that need to be preserved in the dataset, for example the disease of a person.

We use a Java based framework for processing the large datasets called Hadoop. Hadoop consists of two main components that are; HDFS and Map-Reduce. Large datasets on commodity hardware can run on HDFS.

HDFS uses a 64MB block. This can be increased to 128MB or 256MB.

HDFS consists of NameNode, DataNodes and Secondary NameNode.

NameNode stores filesystem metadata.

Secondary NameNode performs internal NameNode transactions log check pointing, it acts as a help to the NameNode but not as a replacement.

DataNode stores data in the filesystem in the form of blocks.

HadoopMapReduce is used to process large amounts of data on any commodity hardware in parallel.

Data is split into independent blocks which are processed by the “Map” in a parallel execution. The output from the “Map” jobs is given as the input to “Reduce” function.

Map-Reduce framework consists of a single master – JobTracker, one per cluster. This schedules the jobs and then assigns it to slaves. It also monitors jobs and also re-executes them in case of failure.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

II. RELATED WORK

In [1] the author talks about *Generalization* for Privacy preserving of data publishing. Along with this it also throws light upon various drawbacks of Generalization. It develops ANGEL that is a new anonymization technique which is as effective as Generalization but can retain significantly more useful correlations in the database called Microdata.

Generalization is regarded as a point to rectangle transformation in the space formed by QI attributes. Various Generalizations may provide drastically different Privacy protection and therefore Generalization needs to be guided by anonymization principles such as k-anonymity, l-diversity and l-closeness.

In [2] the author talks about Generalization and Bucketization along with its drawbacks. The paper presents a novel technique called *Slicing* which is used for Privacy preserving data publishing. It also talks about various advantages of Slicing when compared with Bucketization and Generalization. Data utility is better preserved in Slicing than in Generalization. It also preserves more attribute correlation than Bucketization. It handles high dimensional data well without clear separation of SA and QI.

Slicing divides the data both horizontally and vertically. Vertical partitioning is facilitated by grouping attributes into columns based on correlations among each attribute. Every column contains a subset of attributes which are highly correlated. Tuples are grouped into buckets in horizontal partitions. Within each bucket, the values of each column are randomly permutated to break the linking between different columns. In Slicing the association across different columns is broken but association within each column is preserved. Due to this, dimensionality of the data is considerably reduced and utility is preserved better than Generalization and Bucketization.

In this technique, the output table also satisfies l-diversity. The associations between attributes which are uncorrelated are broken.

In [3] the author talks about updates made to confidential and anonymous databases where a database owner A needs to determine whether a database when inserted with a tuple owned by a person B is still secure. It talks about two protocols which contain Generalization based and Suppression based k-anonymous confidential databases.

In [4] the author talks about focusing on Data Publication in a dynamic dataset. It talks about k-anonymity and its vulnerabilities such as homogeneity attacks and background knowledge attacks. They propose a stronger model which is l-diversity which needs every QI group to have at least one well represented SA. They are throwing light upon *m-invariance* that limits the risks of privacy disclosure in re-publication. It considers insertions, updates and deletions in the microdata.

In [5] the author talks about enhanced slicing models due to the drawback of Slicing i.e. when more number of similar attribute values are present along with their sensitive values in different tuples, the output maybe the original tuple despite performing random permutation. Also, the utility of the data set may be lost due to the generation of the fake tuples. Enhanced Slicing models such as suppression slicing which is done by suppressing any one of the attribute value in the tuple. Only a very few values are suppressed and random permutation is used to maintain privacy. It also throws light upon the *MondrainSlicing* in which random permutation is done not within a single bucket but within all buckets.

III. MOTIVATION

Privacy Preservation of datasets is useful in various feilds such as healthcare, customer transactions etc. and hence efficient methods are necessary to provide the same.

For privacy preservation of healthcare in big data [], there is need for real-time security and privacy analysis. The data sources must be secure and threats related to these sources must be predicted. Incomplete data, problematic handwriting and incorrect or missing data must be handled with appropriate identification algorithms. When the data is anonymized and generalized, the quality and the utility of the data must also be preserved. The data should be such that it is usable for analytics.

Emphasis is given to the fact that valuable information must not be lost while implementing various techniques for privacy preservation.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

IV. TECHNIQUES USED FOR PRIVACY PRESERVATION

Generalization:
Generalization methods prevents linkingattacks. It works by replacing QI-values in the microdatawith fuzzier forms. In this method, the first tuples are grouped into buckets and then foreach bucket, values of one attribute is replaced with ageneralized value, because same attribute value may begeneralized differently when they appear in different buckets.

Slicing:
Slicing, partitions the data both horizontally and vertically. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. It is a more secure method in compare togeneralization and bucketization, because all the following steps are involved in slicing process.

Slicing process is carried out using the following steps:
-Extract the data set from the database.
-Anonymity process divides the records into two.
-Interchange the sensitive values.
-Multiset values generated and displayed.
-Attributes are correlated and secure data is displayed.

Bucketization:
In Bucketization, separate buckets are created and it is used to separate the quasi identifier and sensitive attribute.
Although Bucketization has better data utility than generalization, it does not prevent membership disclosure. An adversary can find out whether an individual has a record in the published data or not, because bucketization publishes the QI values in their original forms.

V. TOOLS USED

Apache Hadoop:
Apache Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. One of the main reasons for using Apache Hadoop is due to its ability to store massive amounts of data (structured, semi-structured, unstructured). It also has a very high processing power. It is possible to build large and complex applications atop the Hadoop platform.
It contains two modules:
• MapReduce:
The algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The Reduce task takes the output from a map as an input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map job.
• Hadoop Distributed File System (HDFS). It breaks up the input data and stores them on compute nodes. Hence, the data is processed in parallel using all the machines in the cluster.
In the healthcare domain, Hadoop is used to store a large amount of patient records. Further, analytics are performed on such records using tools such as Hive, SerDe, Spark etc.

Apache Hive:
Apache Hive is a Big Data tool built on top of Apache Hadoop used to query, summarize and analyse the stored data (which is not appropriate for a relational database). Hive provides a Query Language called the Hive Query Language (HQL) which is similar to SQL. Using this query language, we can perform various functions on the data stored in the Hive tables.
In the healthcare domain, tables are built to store the patient records using the command-line interface. Querying is then performed on the records so as to obtain various results depending on the attributes specified.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

XMLSerDe:
SerDe stands for Serializer and Deserializer.
It is an interface which allows us to instruct Hive as to how a record should be processed.
The Deserializer interface takes a string or binary representation of a record, and translates it into a Java object that Hive can manipulate. The Serializer, takes a Java object that Hive has been working with, and turns it into something that Hive can write to HDFS. Deserializers are used at query time to execute 'Select' statements, and Serializers are used when writing data, such as through an 'Insert-Select' statement.
Node

VI. DEFINITIONS

K-anonymization:

A property of anonymized data, K-anonymity makes sure that the data of persons in a released dataset is such that the identity of the person/persons is not released, despite this, it makes sure that the data is useful for further analysis. K-anonymity can be defined as : "A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release".

K-anonymization can be done in the following ways:

- Suppression: In this method, one attribute (one column) is completely replaced by *. Values such as name, gender, zipcode can be suppressed.
- Generalization: Here, the border categories are replaced with different values. For example, Age can be written as 20-30 rather than giving absolute values.

L-diversity:

A group-based data anonymization technique, l-diversity, is used for the preservation of data privacy. In this technique, the granularity of data is reduced, thereby helping privacy preservation. Reducing the granularity of data acts as a trade off between privacy preservation and data integrity since there is a significant loss of data.It must be taken care that the SA must be diverse within each QI equivalence class. That means, each equivalence class has at least *l* well represented Sensitive values.

VII. FUTURE SCOPE AND CONCLUSION

The total amount of data generated in the last 7 years is equal to the total data ever generated and hence it is essential to preserve this data. This section consists on different methods to preserve privacy.

-The cryptography techniques like ABE and IRE are used in end to end communications. The problem arises since the entire data needs to be encrypted or decrypted before any operations are performed on the data. Hence not allowing data owners to share the data for analysis. The data needs to be shared to obtain accurate values but this also increases the possibility of data leakage since different companies haver different keys for encryption. Hence there is a need for a solution that avoids encryption my multiple parties.

-Techniques like data anonymization are used to preserve privacy, but the existing techniques are ineffective for large volumes of data. Most of the techniques are for static data whereas most of the data is dynamic. Thus, there is a need for new privacy metrics.

-Since all the data is stored in a single cloud there is a risk since a single point failure would indicate the loss of the centralised server. Personalised clouds can be used to avoid this issue. Similar works have been implemented on projects like OwnCloud and IndieWeb.

REFERENCES

[1] "Angel: Enhancing The Utility Of Generalization For Privacy Preserving Publication" Ieee Transactions On Knowledge And Data Engineering, Yufei Tao, Hekang Chen, Xiaokui Xiao, Shuigeng Zhou, Vol 21, July 2009.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

[2] "Slicing: A New Approach For Privacy Preserving Data Publishing" Ieee Transactions On Knowledge And Data Engineering, Tiancheng Li, Ninghui Li, Vol 24, No. 3, March 2012.
[3] "Privacy-Preserving Updates To Anonymous And Confidential Databases", Ieee Transactions On Dependable And Secure Computing, Alberto Trombetta, Wei Jiang, Vol. 8, No. 4, July/August 2011.
[4] "Privacy Preserving Research For Re-Publication Multiple Sensitive Attributes In Data", Xiaolin Zhang, Lifeng Zhang, Ieee 2011.
[5] "Enhanced Slicing Models For Preserving Privacy In Data Publication", International Conference On Current Trends In Engineering And Technology, Icctet'13, S. Kiruthika, Dr. M. Mohamed Raseen.
[6] "A Data Anonymous Method Based On Overlapping Slicing", Jing Yang, Ziyun Liu, Yangyue, Jianpei Zhang, Proceedings Of The 2014 Ieee 18th International Conference On Computer Supported Cooperative Work In Design.
[7] "Anonymization Techniques Through Record Elimination To Preserve Privacy Of Published Data", Proceedings Of The 2013 International Conference On Pattern Recognition, Informatics And Mobile Engineering, February 21-22, R. Mahesh, T Meyyappan.
[8] "Safe Realization Of The Generalization Privacy Mechanism", 2011 9th Annual International Conference On Privacy, Security And Trust, Tristan Allard, Benjamin Nguyen, Philippe Pucheral.
[9] "Privacy Preservating Classification In Two-Dimension Distributed Data", 2010 Second International Conference On Knowledge And Systems Engineering, Luong The Dung, Ho TuBao, Nguyen The Binh And Tuan-Hao Hoang.
[10] "Medical Application Of Privacy Preservation By Big Data Analysis Using Map Reduce Framework", Ashish P, Tejas S, Srinivasa J G, Sumeet G, Sunanda Dixit And Mahesh Belur, International Journal On Advanced Computer Theory And Engineering.
[11] "Protection Of Big Data Privacy", AbidMehmood, IynkarangNatgunanathan, Yong Xiang, GuangHua.