An Efficient Recommendation Of policies to Achieve Data Privacy Using Map Reduce

Nasreen Taj M B[1], Sahana K S [2], Sneha G D[3], Syed Meenaz[4],Tejaswini Y C[5]

[12345]*G M Institute of Technology, Davangere, Karnataka, India.*

(E-mail:nasreent@gmit.ac.in)

(E-mail:snehasahana66@gmail.com)

(E-mail:snehagd55@gmail.com)

(E-mail: syedameenaz31@gmail.com)

(E-mail:yctejaswini609@gmail.com)

Abstract—The data owners needs the various platform to release their data in the real world applications, and it is also important to discover the valuable information which is hiding behind those data. However, existing re-identification attack on those data causes the tremendous threads to the privacy. Thus, it is important resolve all these kinds of attacks and risks by recommending the de-identification policies to preserve the privacy of data. This de-identification policy is recommended in such a way that it is used for protecting the privacy of data. The skyline computation can be used to select such policies, but its a challenging for efficient skyline processing over large number of policies. In addition, little is known about Crime-as-a-Service (CaaS), a criminal business model that carry the cybercrime underground.

Keywords—*Crimeware-as-a-Service, De-identification, Hadoop MPCM, PPDP (privacy-preserving data publication), QI(quasi-identifier), Representational state transfer (REST)*

INTRODUCTION

In the age of big data, it is important to exchange and share data among different parties. However, publishing those data containing sensitive information could violate individual's privacy. In order to get sufficient protection while maintain high data utility, privacy-preserving data publication (PPDP) it is becoming an important and interesting research topic[1]. Not all attributes are sensitive, only some of them which are sensitive need to be protected, However, the published records may still contain quasi-identifiers, Even though the quasi-identifier (QI) attributes do not directly reveal individual's identity, but they may appear together with identification attributes in another published datasets, which may lead to linkage attacks to re-identify private information. Thus, re-identification becomes one of the most important privacy threats for public data tables that contain individual's records.

Many privacy preservation algorithms that rely on QI attributes generalization are proposed to solve this challenge[3]. They usually adopt syntactic sanitization approaches to disturb the data, and the utility of sanitized data is also measured syntactically. To protect personal sensitive information, various laws require that personal data which can be used to link one record in one table to another table containing explicit identifiers based on QI attributes, should be de-identified[4].For instance, to achieve de-identification, the Privacy Rule of the Health Insurance Portability and Accountability Act defines two approaches: Safe Harbor and Expert Determination. In this paper, we focus on the de-identification policy which is one common privacy-preserving approach and an important application of generalization. By using de-identification policies, a continuous balance between privacy protection and data utility can be achieved by choosing the appropriate policies.

PROPOSED SYSTEM

Besides the important test for PCM to guarantee person's security and additionally to held the information which utilizable for investigates. The main point of this examination is mainly to build up a structure it fulfills disparity protection norms and to ensure guarantee most extreme information ease of use to manage the arrangement step for acquaintance diggers. The center advantage of the proposed work is mainly to guarantee the simplicity of accessibility of large notch information to elevate collective logical research to accomplish new discoveries.

Numerous endeavors performed to acquire a decent harmony between information protection and information utility. One agent strategy is mainly to worry at security process, and scientists proposed a progression of protection guidelines those are k-anonymity, l assorted variety, t-closeness. Despite the fact that these models can enhance the step of security insurance, and the authorization and identification risk has been set sufficiently little concerning distinctive necessities, yet it is relating arrangements can just complete a limited range of the utilized region and hazard are contrasted with the control and regulation based strategies, it count almost all ascribes to get more hazard and helpfulness qualities.

Block diagram

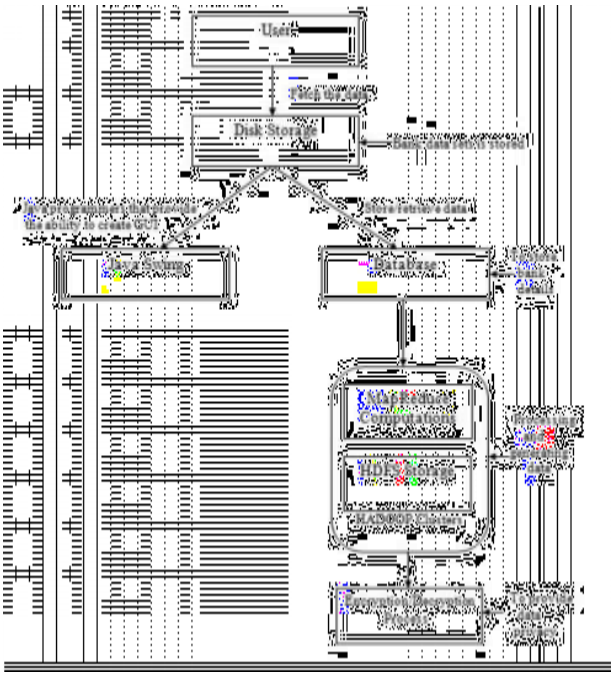


Fig 1: *Methodology diagram for the proposed system*

The fig 1 depicts the approach graph for the proposed framework. The accompanying advances depict it.
Stage 1: User gets the data.
Stage 2: Datasets is put away in the disk stockpiling.
Stage 3: The client's points of interest will be put away in database. It will be recovered as and when required.
Stage 4: The record is transferred to HDFS. At that point grouping is part data as some specific characteristics based or breaking down trait esteems through part or apportioning data separately, Map Reduce is a programming model and a related usage for preparing and creating big data sets with a parallel, conveyed calculation on a bunch. Here Hadoop stage with actualized the specific procedure as discovering general data with mapping and how much data is diminished.
Stage 5: with a specific end goal to give protection data ought to be encoded utilizing mystery key, at that point unscrambles scrambled data utilizing the key utilized for encryption.
Stage 6: Java program is a specialized knowledge makes graphical UI parts, for example, catches and scroll bars.

System Architecture

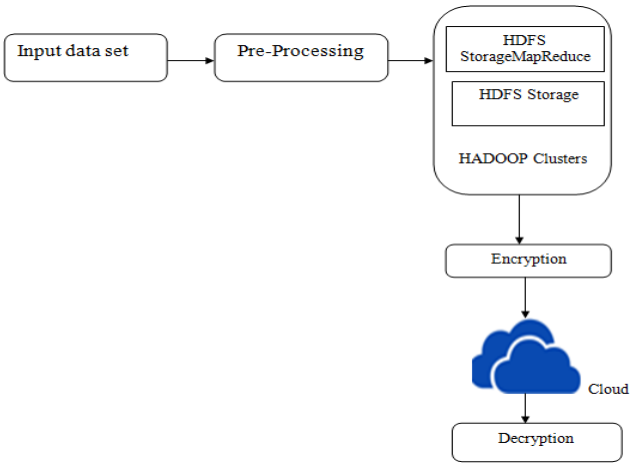


Fig 2: *Architecture Diagram*

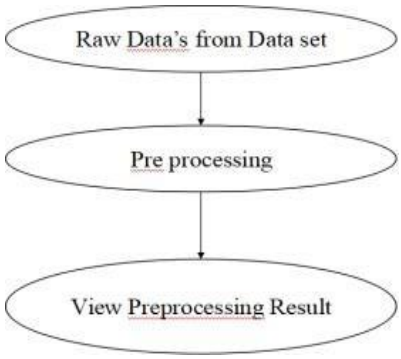
Fig 2 illustrates about the concept flow of the system. Also it explains about the well organized structured actions or behavior of the process. It provides the brief explanation about the steps which has been carried out during project execution. Overall it provides the flow of the information which has been forwarded from one step to other step. The following figure explains about the architectural design which has been utilized for the implementation of the selected process. The above figure depicts the architecture diagram. The following steps describe it.
Step 1: Here input is raw dataset. It means, dataset might consist of unwanted data, unnecessary data, and redundant data. This kind of raw dataset will be provided as an input. System uses the bank related dataset as the input dataset.
Step 2: Preprocess means the removal of unnecessary data from the raw dataset. Redundant data must be eradicated. And unwanted dataset must be eradicated. This step leads the dataset it consist of the required attributes and properties.
Step 3: Clustering process has been done after preprocess step. In this procedure, dataset will be splitting into few properties relied on few properties relative values. Clustering will follow the technique that is dividing and conquer policy. It means it will split based on some attributes and joins them based on some values.
Step 4: The modified dataset will be uploaded into the hadoop distributed file system. It will be used as a database for storing the dataset which has been resulted from the clustering process.
Step 5: Fifth step is using the mapreduce. It is a process of programming procedure and corresponding associative and relative implementation for further processing and producing the big dataset. Here hadoop platform and hadoop distributed file system components have been utilized.
Step 6: The modified dataset which has went through all the mentioned above steps will be either uploaded or save cloud. Here cloud can be any sort of freely available to the user cloud.

MODULES

- Preprocessing
- PCM Clustering
- Map Reduce
- Fetch Cloud

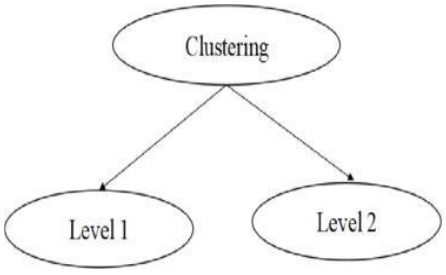
Preprocessing

Preprocess is one of main modules for data mining system. Here we are removing unwanted data or null values and unstructured data. So when we remove unstructured data's then only we get accurate results for given dataset.



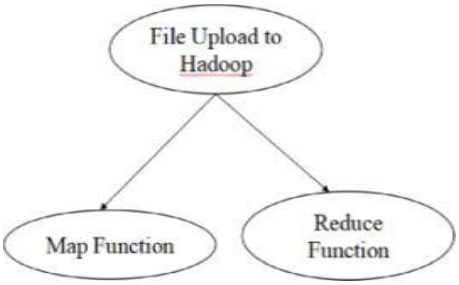
Clustering

Here clustering is splitting data as some particular attribute based or analyzing attribute values through we will splitting or partitioning data individually. First, the fuzzy and possibilistic c-means (FCM and PCM) clustering algorithms are analyzed and some drawbacks and limitations are pointed out. Second, based on the reformulation theorem, by means of modifying PCM model, an effective and efficient clustering algorithm is proposed here, which is referred to as a modified PCM clustering (MPCM).



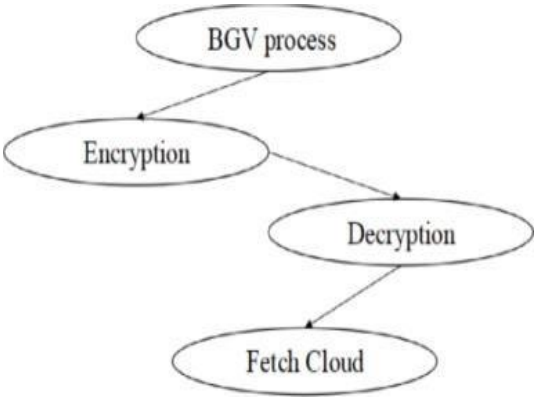
Map reduce

Map Reduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. The model is a specialization of the split-apply-combine strategy for data analysis.Here Hadoop platform with implemented the certain process as finding overall data with mapping and how much data will be reduced.



Fetch cloud

Fetch cloud is the extracting data from the cloud server through some security mechanism. Most cloud storage providers support Web architectures based on representational state transfer (REST) application programming interfaces (APIs).Some also support traditional block- and file-based data, and cloud storage gateway providers can help customers access data in major storage clouds. To aid the clustering process in this task, we performed pre-processing steps such as feature selection and Principal Component Analysis (PCA) and still, the choice of clustering method is not a trivial one. To find the best performing algorithm.



I. Flow diagram

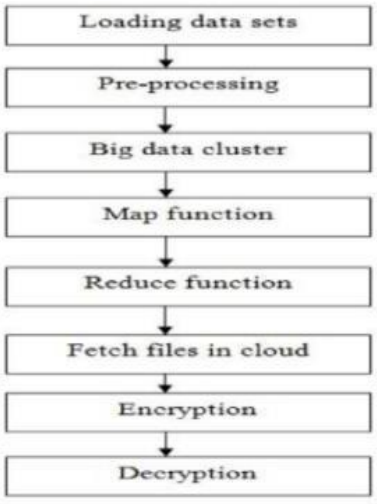


Fig 3:*Flow Diagram*

Fig 3 illustrates the Flow diagram. Here bank user's dataset will be loaded then Preprocessing is carried out. After preprocessing, data is divided into two clusters based on the selected attribute, and then the MapReduce function is applied in order to provide privacy. After that access files in cloud will happen and later encryption and decryption process will be carried out to provide the security for the uploaded file.

System Requirements Specification

A software requirement specification illustrates about the behavior of the system which has been implemented and the environment which has been used to develop the proposed system. In total view it completely says about the active functional requirements and the non functional requirements of the proposed system.

Purpose of SRS

The key need of software requirement illustrates that the way of implementing the application with the help of the functional environment requirements and non functional environment requirements.

Functional Requirements

It mainly depicts the activities which are performed by the system with the inputs provided and roles and actions performed by the inputs with different action sets. It also associates with the environment of the non functional need.

Customer should have the ability to get the data from database.

- To store all bank customer's information, a database is made using MySQL.
- User should have the ability to a viable treatment of systems.
- User should have the ability to exchange record to HDFS (Hadoop Distributed File System).
- The data should be taken care of by using MapReduce.
- Data should be mixed and unscrambled to give assurance.

Non Functional Requirements

The proposed system makes use of the following properties for the environment of non functional activities.

Accessibility Key reason illustrates about connectivity and availability of the main factor considered here are the internets. To get and fetch the developed framework anyplace and whenever it is conveyed on a worldwide server location.

Deployment The deployment indicates that by considering the environment of the system it will deploy the developed system.
Here system depends on the most of the requirements such as hardware as well as the software requirements.

Documentation Documenting is considered as the one of the important step to illustrate and explain what the developed system will do at different scenarios.

Efficiency Here the rate of efficiency can be measured by checking the size of the datasets. Here quantity of the dataset handles by this the key factor to measure the efficiency.

Scalability The system is adequately scalable to incorporate additional future request to plug-in in an easier way. An average channel, which can easily occupy the present system. System must contacted incorporate various more web based interconnections channels promotions bits of knowledge.

Conclusion

The projected system is build up with a reason on condition that suggestion of regulations mainly to accomplish information security by means of Map reduce programming model. Initially a successful path for the regulations and rules which has been applies for security reason. The novel projected system characterization, it will reduce the amount of instance of producing the regulations and the quantity of another regulations group artificially. Subsequently projected system illustrates about the usage of SKY FILTER MR. It is one of the form of using MapReduce programming model. It mainly used to provide solution for the regulations proficiently. It executes the most composite and tedious environment of dataset to provide the high security for the files which has been uploaded to the cloud. Whereas the proposed
system effectives uses the asymmetric encryption algorithm to achieve the high level security for the dataset. Results of the proposed system illustrates that caliber of the system is high.

Future Scope

Future scope of this paper will be:

- Right now the proposed system is using the single dataset, but it can enhance to utilize the more number of dataset.
- Projected system has been built basically for the bank database, but it can be expanded to perform calculations for any sort of dataset such as medical field dataset, finance dataset etc.,
- Clustering and classification can be improved by applying the enhanced algorithms.
- A policy which has been said for security of the dataset can be standardized by using an efficient encryption and decryption algorithm.

References

[1] [2016]Xiaofeng Ding, Li Wang, Zhiyuan Shao, and Hai Jin "A well organized advise of De recognition steps by using MapReduce"

[2] [2016]Weiyi Xia, Raymond Heatherly, Xiaofeng Ding, Jiuyong Li, "An Efficient reaseach of De-recognition steps for a hazard based outskirts"

[3] [2015]Jevin D West, Ian Wesley-Smith, Carl T. Bergstrom "An empirical study on security of data which is saved on cloud" [2015]Zhihua Xia, Xinhui Wang, Xingming Sun, Qian Wang "A protected and active different key based score look for Scheme about Encrypted information Data"

[4][2014]Jun Zhang, Graham Cormode,CeciliaM.Procopiuc, Divesh Srivastava Xiaokui "PrivBayes: Private Data Release via Bayesian Networks"

[5] [2013]Benjamin C. M. Fung, Ke Wang, Rui Chen, Philip S. Yu" Security-defend information exhibit: An empirical Survey of latest developments."