PRIVACY PRESERVATION FOR SCALABLE BIG DATA IN CLOUD

[1]Varsha Bais,[2] SayyadaFahmeeda Sultana

[1] PG student,[2] Assistant Professor
[1] Department Of Computer Science,
[1] PDA College of Engineering, Kalaburagi, India

Abstract : Big Data is term used to describe a collection of data that is huge in size and yet growing exponentially. Managing this big data with secure and privacy is major outline goal. Data mining is primarily done to extract and retrieve information. One of the major concern in big data mining approach is with security and privacy big data application such as online social media, mobile services , health care etc. Data sets in such application often contain privacy sensitive information which brings about privacy concerns potentially if the information is shared or released to third party .In this project anonymization technique is used via generalization to statisfy the given privacy model .the project proposed two phase clustering method. In first phase the dataset is collected and this dataset is partitioned into number of attributes this is done by t-ancestor clustering,and during this phase attributes are placed into three different identifiers those are General identifier, Sensitive identifier ,Quasi identifier .In second phase only sensitive identifiers are anonymized before sharing data. A variety of privacy models and data anonymization approaches has been proposed, however applying these traditional approach to big data anonymization poses scalability and efficiency challenges because of the “3Vs”, that is Volume, Velocity, Variety. Hence our approaches can preserve the proximity privacy, and can significantly improves the scalability and time-efficiency of local-recoding anonymization over existing approaches.

***IndexTerms* - Big Data, Attributes, Data Anonymization, Local Recoding, Proximity Privacy .**

I. INTRODUCTION

Big data is defined as data that is huge in size. Big Data is term used to describe a collection of data that is huge in size and yet growing exponentially with time . Managing such big data is done by outlining our goals such as securing the data, protecting data and also know the data thst you need to capture. Data mining is primarily done to extract and retrieve information from humongous quantity of data. One of the major concern in big data mining approach is with security and privacy big data application such as online social media, mobile services , health care etc. an enormous amount of data is generated based on various aspects of the individual without proper security and privacy protection in all aspects of computing environment. This can disclosed intentionally or unintentionally severe threats on the individual.

Agencies and other organization often need to publish big data such as medical data or census data, for research and other purpose. Such data is stored in a table and each record has a number of attributes ,which are classified into three identifiers (1)Attributes that clearly identify individual these are known as General identifiers, this includes social security numbers , address, name and so on (2) Attribute whose values are taken together can potentially identify the individual these are known as Quasi identifier, these includes Zip code , Date of Birth etc and so on (3)Attributes that are consider sensitive one are known as sensitive identifiers and includes salary, disease etc

When releasing microdata, it is necessary to prevent the sensitive information of the individual from being disclosed. Two types of information disclosure have been identified that is Identity disclosure and Attribute disclosure , Identity disclosure occurs when an individual is linked to a particular record in the release table. Attribute disclosure occurs when new information is added, when some individual revealed that release data makes it possible to infer the characteristics of an individual more accurately.

While the release table gives useful information to researches, it prevents disclosure risk of individual whose data are in table. Our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. Data anonymization refers to hiding identity and or sensitive data so that privacy of individual is effectively preserved while certain aggregated information can still be exposed to data users for diverse analysis and mining tasks.

To address the scalability problem, we propose a two phase clustering approach consisting of t-ancestor clustering and proximity-aware agglomerative clustering. The first phase splits an original data set into ‘t’ partitions. In second phase partition data are locally recoded by proximity-aware agglomerative clustering.

The major contribution of for project is put fourth, firstly an extended proximity privacy model is taken which allowing multiple sensitive values and categorical sensitive values (numerical values). Second, this project has model the problem of big data local recoding against proximity privacy breaches as a proximity-aware clustering problem. Third, a scalable and efficiency two-phase clustering approach is proposed which consist of t-ancestor clustering and proximity-aware agglomerative clustering approach this parallelize local recoding on multiple data partitions . fourthly, anonymization technique is used on the multiple sensitive values , the remaining attribute are kept for data mining tasks and research propose.

II. RELATED WORK

Balaji Palanisamy and Ling Liu in 2015[1]The authors presents an anonymization framework for publishing large association graph datasets with the goal of supporting multi-level access controlled query processing in shared storage systems.Author propose a suite of anonymization techniques and a utility-preserving grouping technique to support multi-level access controlled query processing on published datasets.Jian Xu, Wei Wang , Jian Pei in 2006[2]the author, study the problem of utility-based anonymization.First, he propose a simple framework to specify utility of attributes. The framework covers both numeric and categorical data. Second, he develop two simple yet efficient heuristic local recoding methods for utility-based anonymization.

X. Zhang, C. Liu, S.Nepal, S.Pandey in 2013 [3]Large volume of intermediate data set are generated to provide privacy of intermediate dataset becomes a challenging problem because adversaries may recover privacy-sensitive information by analyzing multiple intermediate data sets. They propose a novel upper bound privacy leakage constraint-based approach to identify which intermediate data sets need to be encrypted and which do not, so that privacy-preserving cost can be saved while the privacy requirements of holders can be satisfie . B. C. M. Fung, K. Wang, R. Chen in 2010 [4] Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy.Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy.

X. Zhang, L. T. Yang, C. Liu and J.Chen in 2014 [5]The author presents a practical and productive algorithm for determining a abstract version of data that masks sensitive information and remains useful for standardizing organization. G. Aggarwal, R. Panigrahy , T. Feder, D. Thomas in 2010[6]The author proposed a Dummy-Location Selection (DLS) algorithm to protect user's location privacy against adversaries with side information. Based on the obtained side information and the entropy metric, DLS carefully selects the dummy locations to achieve the optimal level of k-anonymity. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan in 2008 [7]the author provided algorithms for incorporating a class oftarget workloads, consisting of classification or regression models, as well as selection predicates, when generating an anonymous data recoding.

J. Xu, W. Wang, j.Pie,B. Shi in 2006[8]Anonymization of local recoding scheme from the utility perspective and put forth a bottom-up greedy approach and the top-down counter-part.the author used the agglomerative clustering technique while later employed the divisive hierarchical clustering technique both of which poses constraint on the size of the cluster.W. Byun, A. Kamra, E.Bertino and N.Li in 2007[9]author investigated the inconsistency issue of local-recoding anonymization in data with hierarchical attributes and proposed KACA(K-Anonymization by Clustering in Attributes hierarchies)algorithms.G. Aggarwal, R. Panigrahy, T.Feder,D.Thomas in 2010[10]author proposed a set of constant factors approximation algorithms for clustering based anonymization problem that is r-GATHER and r-CELLULAR CLUSTERING, where cluster center are published.

III. PROPOSED SYSTEM

In this project first select the big dataset. It is perprocessed the dataset. The process mode include the elimination of unwanted elements and unwanted symbols in the dataset. Then, split the attributes in the dataset. It has been classified into three attributes. They are General Identifier, Sensitive Attribute , Quasi Identifier. this above process is done by t ancestor clustering method. After the dataset has been partitioned anonymize the data using local recoding mechanism using generalization mechanism. the process is done by proximity-aware agglomerative clustering method. After that upload the anonymized data.

Data is essential for Data Science with tons of data being generated every second this most of data are unlabelled . To organise the data, Clustering comes into picture. This clustering is used for unlabelled data, the word cluster means grouping similar things together here clustering method used is T-Ancestor clustering which works like K-means here "T" means number of cluster. The T-Ancestor clustering work in following steps.

1. T-Ancestor clustering Technique

```
Input: unlabled database
Output: Partitioned database attributes
Step 1. Initialize cluster centroids
Step 2. Assign data point to clusters
Step 3. Update cluster centroids
Step 4.Repeat step 2 and 3 until all dataset are placed into respective clusters
Step 5. Name is given to the formed clusters assigned name of dataset present in cluster is considered as label
```

In step 1 as starting point, the model picks up K here K means number of datapoint, (let K=3) datapoint from dataset. These datapoint are called cluster centroids.

In second step the model calculate distance between datapoint and all centroids using Euclidean distance formula and assign cluster with the nearest centroid.

Dist=$(x2 - x1)^{2}$+$(y2 - y1)^{2}$+$(z2 - z1)^{2}$

In third step model updates the new cluster values. The new cluster value is updated cluster centroid which is taken as average or the mean value of all the datapoints within that cluster.

New centroid=Avg of data points

Stopping criterion is maintain to stop updating the cluster. The stopping condition are the distance of datapoints from their centroid is minimum threshold is maintain and centroid must remain same.

2. Proximity aware agglomerative clustering Technique

- Each data record is regarded as a cluster initially and then two clusters are picked to merge
- Two clusters with shortest distance are merged here we use leverage the complete-linkage distance
  D $(C_{x,\ Cy}) = \max_{rx\in Cx\ ,ry\in Cy} dist\ (r_{x,}\ r_{y)}$
- The distance between two clusters equals to the weighted distance between those two records that are farthest away from each other
- Merged clusters is used for anonymization
- Anonymization is done by k-anonymity this is shown in fig 2
- In k-anonymity each merged record is indistinguishable with at least k-1 other record with respective Sensitive Attributes.
- If the table satisfies k-anonymity for some value k, then anyone who knows only Quasi Attributes values of one individual cannot identify the record corresponding to that individual with confidence greater than 1/k

Dataflow Diagram:

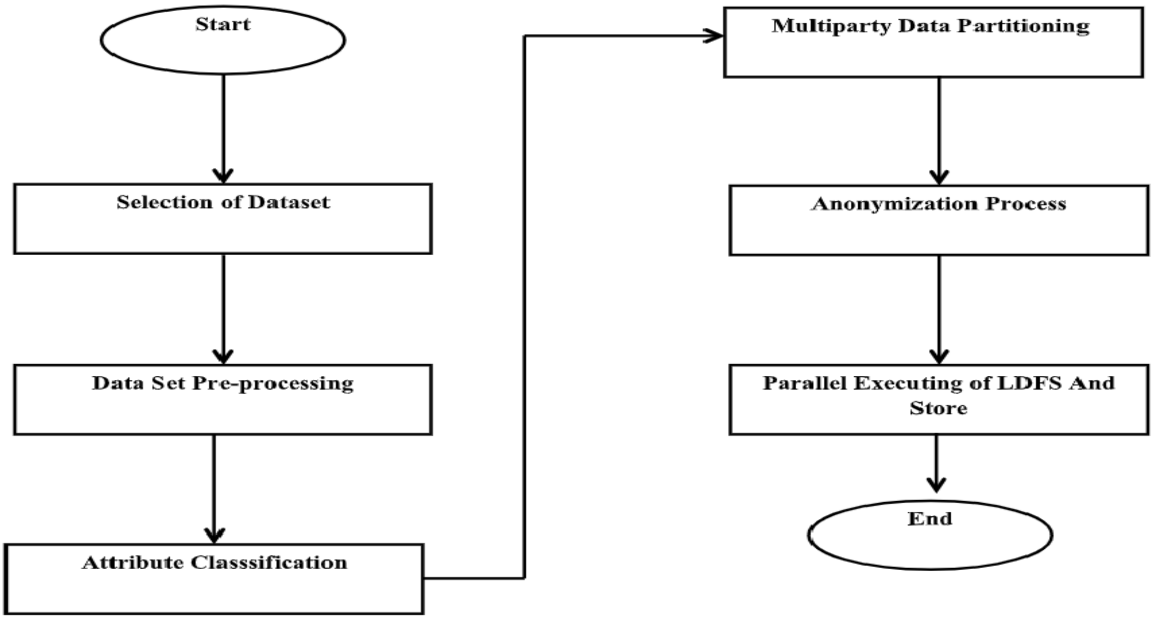
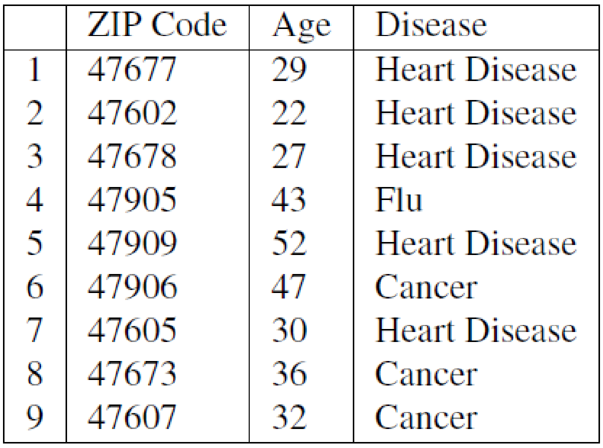

Start
Selection of Dataset
Data Set Pre-processing
Attribute Classsification
Multiparty Data Partitioning
Anonymization Process
Parallel Executing of LDFS And Store
End

Fig1: Flow chart of proposed system

IV. RESULTS AND DISCUSSION

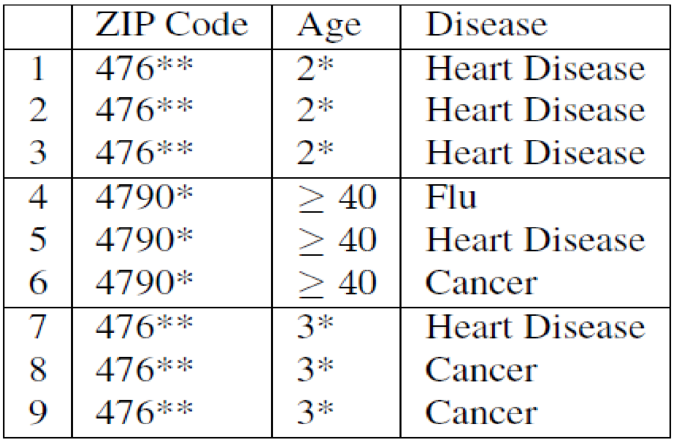
4.1Results of Anonymization Technique

Table 4.1: Original Dataset



	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table4.1.2:AK-AnonymousVersion



	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Table 4.1 is the original data table, and Table 4.1.2 isan anonymized version of it satisfying 3-anonymity. The Disease attribute is sensitive. Suppose Alice knows that Bob is a 27-year old man living in ZIP 47678 and Bob’s record is in the table. From Table

4.1.2, Alice can conclude that Bob corresponds to one of the first three records, and thus must have heart disease. This is the homogeneity attack. For an example of the background knowledge attack, suppose that, by knowing Carl's age and zip code, Alice can conclude that Carl corresponds to a record in the last equivalence class in Table 4.1.2.

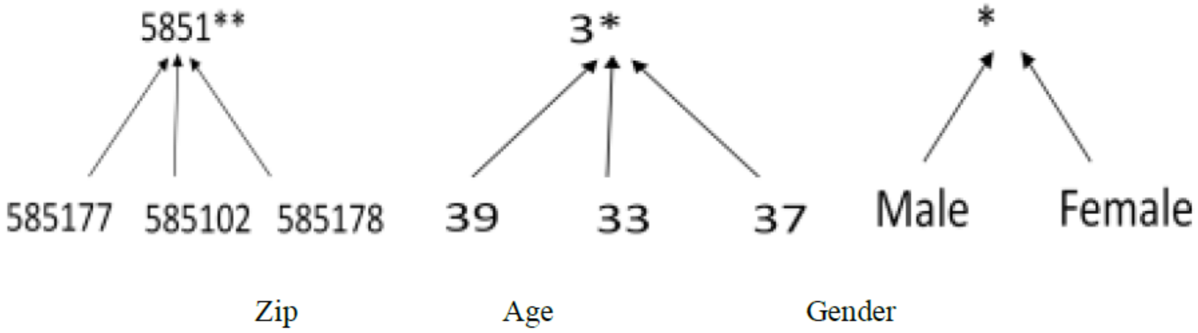


Figure2. Anonymization of attributes zip-code, age, gender

Generalization of attributes replaces with less specific but semantically consistent values Figure 2 shows the generalization of zip-code, age and gender.

V. CONCLUSION

local-recoding anonymization for big data in cloud has been investigated from the perspectives of capability of defending proximity privacy breaches, scalability and time-efficiency. This system will be proposing a proximity privacy model by allowing multiple sensitive attributes and semantic proximity of categorical sensitive values. And, also it proposed scalable two-phase clustering approach to address the exiting problem for time-efficiency. Extensive experiments on real-world data sets can be demonstrated that this above approach significantly improves the capability of defending proximity attacks, the scalability and the time-efficiency of local-recoding anonymization over existing approaches. In cloud environment, the privacy preservation for data analysis, share and mining is a challenging research issue due to increasingly larger volumes of datasets, thereby requiring intensive investigation.

References

[1] Balaji Palanisamy and Ling Liu," Privacy-preserving Data Publishing in the Cloud: Multilevel Utility Controlled Approach" in IEEE 8th International Conference on Cloud Computing 2015

[2] J Xu, Wei Wang, J. Pei, X. Wang, B. Shi, and A. W. C. Fu, "Utilitybased anonymization using local recoding," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data, 2006, pp. 785–790.

[3] X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, "A privacy leakage upper bound constraint-based approach for cost-effective privacy preserving of intermediate data sets in cloud," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1192–1202, Jun. 2013.

[4] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Survey, vol. 42, no. 4, pp. 1–53, 2010.

[5] G. Aggarwal, R. Panigrahy, T. Feder, D. Thomas, K. Kenthapadi, S. Khuller, and A. Zhu, "Achieving anonymity via clustering," ACM Trans. Algorithms, vol. 6, no. 3, 2010.

[6] G. Aggarwal, R. Panigrahy, T. Feder, D. Thomas, K. Kenthapadi, S. Khuller, and A. Zhu, "Achieving anonymity via clustering," ACM Trans. Algorithms, vol. 6, no. 3, 2010.

[7] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization techniques for large-scale datasets," ACM Trans. Database Syst., vol. 33, no. 3, pp. 1–47, 2008.

[8] J. Xu, W. Wang, j.Pie,B. Shi,and A. W. C. Fu,"utility-based anonymization using local recoding," in proc.12th ACM SIGKDD Int.Conf.Knowl.Discovery Data, 2006,pp.785-790

[9] W. Byun, A. Kamra, E.Bertino and N.Li,"Efficiency K-anonymization using clustering technique," in proc.12thInt.Conf.Knowl.Database Syst.Adv.Appl.2007,pp.188-200

[10] G. Aggarwal, R. Panigraphy, T. Feder D. Thomas, "Achieving anonymity via Clustering," ACM Trans. Algorithms, vol.6, no.3, 2010