A

Project-I Report

on

PERTURBATION METHOD FOR EFFECTIVE DATA PRIVACY IN DATA MINING

Submitted in Partial Fulfillment of
the Requirements for the Degree
of

Bachelor of Engineering

in

Computer Engineering

to

North Maharashtra University, Jalgaon

Submitted by

Bhavesh Patil
Anita Patil
Sujata Patil
Gayatri Pawar

Under the Guidance of

Prof. Dr. Girish K. Patnaik

DEPARTMENT OF COMPUTER ENGINEERING
SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY,
BAMBHORI, JALGAON - 425 001 (MS)

2016 - 2017

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY, BAMBHORI, JALGAON - 425 001 (MS)

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the Project-I entitled *Perturbation Method For Effective Data Privacy In Data Mining*, submitted by

Bhavesh Patil
Anita Patil
Sujata Patil
Gayatri Pawar

in partial fulfillment of the degree of *Bachelor of Engineering* in *Computer Engineering* has been satisfactorily carried out under my guidance as per the requirement of North Maharashtra University, Jalgaon.

Date: November 10, 2016
Place: Jalgaon

Prof. Dr. Girish K. Patnaik
Guide

Prof. Dr. Girish K. Patnaik
Head

Prof. Dr. K. S. Wani
Principal

Acknowledgements

First of all we would like to extend our deep gratitude to almighty God, who has enlightened us with power of knowledge. We wish to express our sincere gratitude to the principal Prof. Dr. Kishor S. Wani for giving us such a great opportunity to develop this project. A special gratitude to our Head of Department and project guide, Prof. Dr. Girish K. Patnaik, whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project especially in writing this report. He has invested his full effort in guiding the team in achieving the goal.

Furthermore we would also like to acknowledge with much appreciation the crucial role of the staff of our department, who gave the permission to use all required equipment and the necessary materials to complete this project report. We would like to thanks to our parents and all friends for their co-operation and supports in making this project report successful. We acknowledge our sincere gratitude to all who have directly or indirectly helped us in completing this report successfully.

Bhavesh Patil
Anita Patil
Sujata Patil
Gayatri Pawar

Contents

List of Figures

Abstract

Data mining is the process of finding correlations or patterns among the dozens of fields in large database. The main challenge in data mining is to maintain the privacy of confidential information. In order to share data while preserving privacy, data owner must achieve the dual goal of privacy preservation as well as accuracy of data mining. Data perturbation is an efficient and effective approach that has been used to protect privacy of sensitive information. The proposed approach uses tuple values of sensitive attribute to generate normalized value which produces perturbed data.

Chapter 1

Introduction

Data mining technology has been developed with the goal of providing tools for automatically and intelligently transforming large amount of data in knowledge relevant to users. The extracted Knowledge, often expressed in form of association rules, decision trees or clusters, allows one to find interesting patterns and regularities deeply buried in the data that are meant to facilitate decision making processes. Such a knowledge discovery process, however, can also return sensitive information about individuals, compromising the individuals right to privacy. Moreover, data mining techniques can reveal critical information about business transactions, compromising the free competition in a business setting. Thus, there is a strong need to prevent disclosure not only of confidential personal information, but also of knowledge which is considered sensitive in a given context. For this reason, recently much research effort has been devoted to addressing the problem of privacy preserving in data mining.

In Section 1.1 background is presented. Motivation is presented in Section 1.2. In Section 1.3 problem definition is presented. Scope is presented in Section 1.4. In Section 1.5 Objective is presented. Organization of the Report is presented in Section 1.6. Finally summary is presented in the last section.

1.1 Background

A number of recently proposed techniques address the issue of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to perform the mining. The typical additive reconstruction technique is column-based additive randomization. This type of techniques relies on the facts that Data owners may not want to equally protect all values in a record, thus a column-based value distortion can be applied to reconstruct some sensitive columns.

The condensation approach is a typical multidimensional reconstruction technique, which

aims at preserving the covariance matrix for multiple columns. Thus, some geometric properties such as the shape of decision boundary are well preserved. Different from the randomization approach, it regenerate multiple columns as a whole to generate the entire reconstructed data set. As the reconstructed data set preserves the covariance matrix, many existing data mining algorithms can be applied directly to the reconstructed data set without requiring any change or new development of algorithms.

In data swapping technique confidentiality protection can be achieved by selectively exchanging a subset of attributes values between selected record pairs. Data swapping preserves the privacy of original sensitive information available at record level. If the records are picked at random for each swap then it is called random swaps. It is difficult for an intruder to recognize particular person or entity in database, because all the records are altered to the maximum level. The enviable properties of swapping technique are that it is simple and can be used only on sensitive data without disturbing non sensitive data [2].

1.2 Motivation

Data hiding tries to remove confidential or private information from the data before its disclosure. In this case, many randomization methods have been addressed. The randomization method has been traditionally used in the context of distorting data by probability distribution. In this case, the data miner does not know the raw data and also can get the similar result. Which the key point for data miner is how to reconstruct the raw data distribution. Many privacy preserving data mining methods are inherently limited by the curse of dimensionality in the presence of public information. The technique in analyzes the kd-tree recursive partitioning technique in the presence of increasing dimensionality.

1.3 Problem Definition

Develop a system that generates perturbed data from a database for privacy preserving of sensitive data, Based on the query by an authenticated user, The query handler fetches the required data from database. Data perturbation is performed on the fetched data by value distortion. The value distortion approach perturbs the data by divide and conquer. The resultant dataset with perturbed sensitive attribute values is likely presrves statistical characteristics of original dataset.

1.4 Scope

The individuals need to protection and privacy of sensitive and private data. To do this work kd tree based algorithm using various different approaches, this approaches are concentrating the mining of data as sensitive with confidential and non-confidential data sets. To enhance privacy of the data without changing the original data set. To reduce the information loss.

1.5 Objective

Vary the confidential data from entire data is risk and the data of confidential rules changes on the data access vendor. It is very costly operation on mining the data from databases and handling the sensitive data.

1.6 Organization of the Report

Chapter 1 presents the basic introduction to the Proposed System, Problem Statement, Problem Definition, Objective and Future Scope.

The system analysis is presented in second chapter, which describes Literature Survey, Existing System, Proposed System, Feasibility Study, Risk Analysis.

Chapter 3 describes the system requirements and specifications which includes Software Requirement, Functional Requirement, Non Functional requirement.

The system designing concepts are described in Chapter 4 including Proposed System Flow, Data Flow Diagrams, E-R Diagram, UML Diagrams.

1.7 Summary

In this chapter, Introduction is presented. In the next chapter, System Analysis is presented.

Chapter 2

System Analysis

Analysis is a software engineering task that bridges the gap between system level requirements engineering and software design. Requirements engineering activities result in the specification of softwares operational characteristics (function, data, and behavior), indicate software's interface with other system elements, and establish constraints that software must meet. System analysis allows the software engineer (sometimes called analyst in this role) to refine the software allocation and build models of the data, functional, and behavioral domains that will be treated by software.

In Section 2.1 Literature Survey is presented. Proposed System is presented in Section 2.2. In Section 2.3 presents Feasibility Study. Risk Analysis is described in Section 2.4. Project Scheduling is presented in Section 2.5. Effort Allocation is described in Section 2.6. Finally summary is presented in the last section.

2.1 Literature Survey

A number of recently proposed techniques address the issue of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to perform the mining. The work presented in [1] addresses the problem of building a decision tree classifier in which values of individual records have been perturbed using randomization method. While it is not possible to accurately estimate original values in individual data records, the propose are construction procedure to accurately estimate the distribution of original data values.

There are many approaches which have been adopted for privacy preserving data mining. Classification is based on the following dimensions:

- Data distribution
- Data modification

- Data mining algorithm
- Data or rule hiding
- Privacy preservation

The first dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution.The second dimension refers to the data modification scheme.

In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization.

The third dimension refers to the data mining algorithm, for which the data modification is taking place. This is actually something that is not known beforehand, but it facilitates the analysis and design of the data hiding algorithm. various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.

The fourth dimension refers to whether raw data or aggregated data should be hidden. The complexity for hiding aggregated data in the form of rules is of course higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as rule confusion. The last dimension which is the most important refers to the privacy preservation technique used forth selective modification of the data.

2.2 Proposed System

The proposed approach, aim to achieved better result for privacy on database than existing system. The Proposed approach uses value distortion method for data perturbation. The proposed mechanism of reconstruction tree, the tree will handing the data partitioning the data sets and subsets. The each subset must satisfy the some minimum conditional values will store and from as leaf of the tree. This subset partitioning is combination of the confidential and non-confidential data. The proposed mechanism works and implements the approach of reconstruction tree, as one of the general method like divide and conquer method. This method will divide the data in the data sets and subsets, this data sets and subsets are

conquering in the tree set approaching. This tree leaf sets are linked in from of average squared distances. The mechanism result of reconstruction tree delivers the tree set having the leaf as a distance relation linkage. The distances of linkage will find based on the average square distances.

2.3 Feasibility Study

The feasibility study is carried out to test whether the proposed system is worth being implemented. Feasibility study is a test of system proposed regarding its work ability, its impact on the organization ability to meet user needs and effective use of resources. The key consideration involve in the feasibility study are:

- Economical Feasibility
- Operational Feasibility
- Technical Feasibility

2.3.1 Economical Feasibility

This is the main factor in the feasibility study. When product is economically affordable then it can be used. So project must be cost saving. Establishing the cost effectiveness of the proposed system i.e. if the benefits do not out weighs the costs then it is not worth going ahead. The project involves the utilization of open source tools and softwares which indirectly decreases the release cost of system. The system is economically feasible.

2.3.2 Operational Feasibility

Operational feasibility is the ability to utilize, support and perform the necessary tasks of a system or program. It includes everyone who creates, operates or uses the system. Provide summary statistical information without disclosing individuals confidential data. This makes the system operationally feasible.

2.3.3 Technical Feasibility

Technical feasibility centers on the existing computer system (hardware, software etc) and to what extent it can support the proposed system addition. The front end used in the system java. The platform used for developing the applications is Linux which is easily available. There is no more hardware required other than the personal system for its execution. Data mining technologies have enabled organizations to extract useful knowledge from data in

order to better understand and serve their customers and, thus, gain competitive advantages. While successful applications of data mining are encouraging, there are increasing concerns about invasions to the privacy of personal information.

2.4 Risk Analysis

Risk analysis and management are a series of steps that help a software team to understand and manage uncertainty. Many problems can plague a software project. A risk is a potential problem might happen, it might not. But, regardless of the outcome, it is really a good idea to identify it, assess its probability of occurrence, estimate its impact, and establish a contingency plan should the problem actually occur. Everyone involved in the software process, managers, software engineers, and customers participate in risk analysis and management.

2.4.1 Software Risk

The Record linkage will be finding based on the perturbed data and the original data. The distance between the perturbed data and original data is a disclosure risk of the data.

2.4.2 Project Risk

Project risks threaten the project plan. That is, if project risks become real, it is likely that project schedule will slip and that costs will increase. Project risks identify potential budgetary, schedule, personnel (staffing and organization), resource, customer, and requirements problems and their impact on a software project. Project risk can occur if any one of member allocated is unavailable according to project plan and estimation. If project is not completed within time in this situation project risk can occurs.

2.4.3 Technical Risk

Threaten the quality and timeliness of the software to be produced. If a technical risk becomes a reality, implementation may become difficult or impossible. Technical risks identify potential design, implementation, interface, verification, and maintenance problems. In addition, specification ambiguity, technical uncertainty, technical obsolescence are also risk factors. Technical risks occur because the problem is harder to solve than thought it would be. If any module does not work properly according to expectation then technical risk may occur. The records if not maintained properly may affect the quality and accuracy.

2.5 Project Scheduling

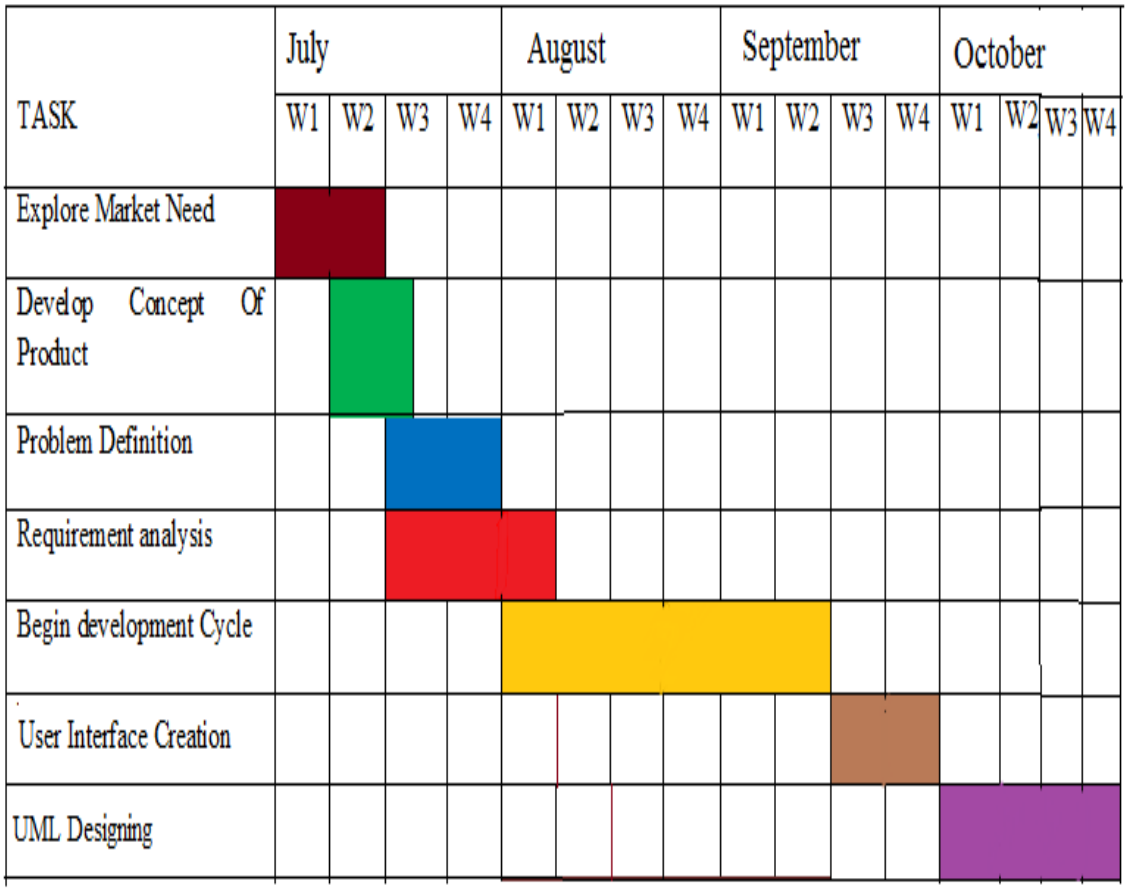

Figure 2.1: Gantt Chart

2.6 Effort Allocation

Following figure shows the efforts and involvement of every member in this project

Table 2.1: Effort Allocation

	Bhavesh Patil	Anita Patil	Sujata Patil	Gayatri Pawar
Project Planning	25%	25%	25%	25%
Requirement Gathering	30%	20%	20%	30%
Design	25%	20%	30%	25%

2.7 Summary

In this chapter, System Analysis is presented. In the next chapter, System Requirement Specification is presented.

Chapter 3

System Requirement Specification

It provides requirements, needs of project and those things which help to complete project. System requirement describe a system from a technical perspective, which describe the essential characteristics of the hardware and software that will meet those needs. It should specify the capabilities, capacities and characteristics of the system in both qualitative and quantitative terms.

In Section 3.1 Hardware Requirements are presented. Software Requirements is presented in Section 3.2. Finally summary is presented in the last section.

3.1 Hardware Requirements

The hardware requirement includes a system with following configurations:

- Processor : Pentium IV or higher.
- RAM : 1 GB.
- Hard Disk : 40 GB.
- Input device : Standard Keyboard and Mouse.
- Output device : VGA and High Resolution Monitor

3.2 Software Requirements

- Operating System : WINDOWS XP sp3 , Windows 7 , Windows 8 or above
- Front End : Java Jdk 1.6
- Web Server : Tomcat Web Server 6.0
- Back End : MySQL

3.3 Summary

In this chapter, System Requirement Specification are presented. In the next chapter, System design is presented.

Chapter 4

System Design

System design provides the understanding and procedural details necessary for implementing the system. Design is an activity concerned with making major decisions, often of a structural nature. Design builds coherent, well planned representations of programs that concentrate on the interrelationships of parts at the higher level and the logical operations involved at the lower levels. Software design is the first of the three technical activities designs, coding and test which are required to build and verify the software.

In Section 4.1 System Architecture are presented. E-R Diagram is presented in Section 4.2. In Section 4.3 Database Design is presented. Data Flow Diagram is presented in Section 4.4. Finally summary is presented in the last section.

4.1 System Architecture

The System Architecture provides the details of how the components or modules are integrated [8].

Fig 4.1 is indicating the system architecture of the tree based data perturbation process. This architecture will give complete description of input and outputs of each process. This process have several modules. They are

- Query Handler
- Privacy Preserving
- Data Perturbation
- Result Evaluation

4.1.1 Query Handler

The Query handler is accepting the query data from the client and process the query with the data base and fetching the data sets from the data base.

4.1.2 Privacy Preserving

The privacy preserving is a process of providing the security for sensitive data. The sensitive data like an employee salary, annual income of company, transferring the money from one account to another account etc. Providing the security to this data is very important. To do this work some existing system are there and one system is proposed here. Those approaches are implemented by following sub modules of this mechanism. The proposed approach is Perturbation Tree and the existing systems are tuple value based multiplicative data perturbation approach.

4.1.3 Tree Based Approach

Perturbation tree is proposed approach. This approach is using the divide and conquer technique. This technique will be using the following process, this approach accept the data sets as input. This data sets will divided in subsets by using above mention technique and storing in tree format up to in tree each child of leaf node having the attributes as the user mention equals or less values. After completion of the division process, each leaf node attributes sensitive data will replacing with the average value and sending to sharable person or other requested client.

4.1.4 Result Evaluation

The result evaluation ISA process to finding the error rate of different states in the data perturbation of original data and the perturbed data. In the result evaluation considering several processes those are time complexity, record linkage, regression and classification. The time complexity will evaluated based on the processing time delay of the input acceptance to producing the output to client. To do this work first find the start time and end time of process, subtracting the end time and start time we get the time of evaluation process in milliseconds. The Record linkage will be finding based on the perturbed data and the original data. The distance between the perturbed data and original data is a disclosure risk of the data.To calculate the bias in standard deviation proposed system is using of the original data and perturbed data. By finding this value,get the loss of information in perturbed data. The regression error rate will finding based on the mean average error rate. These values will giving the information of error rate of this approach on the data perturbation.

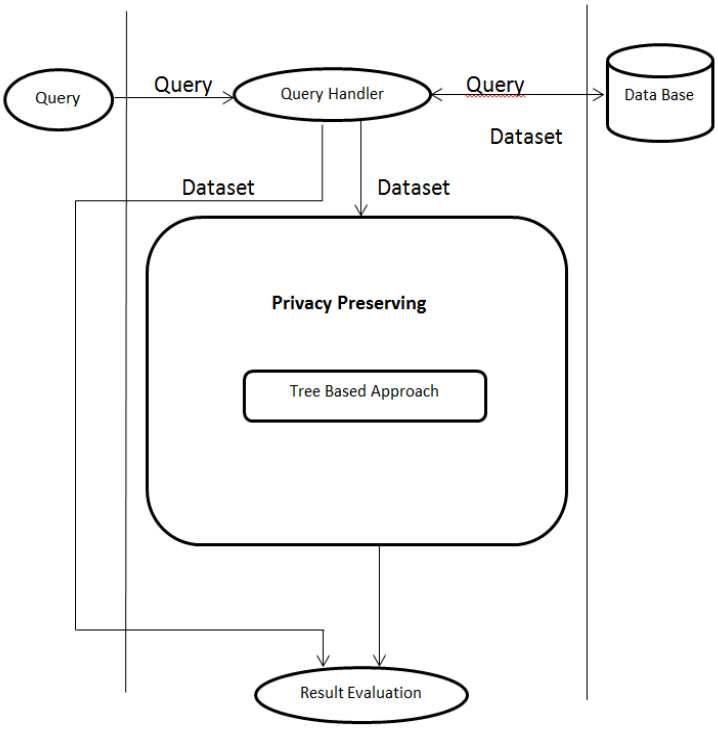

Figure 4.1: System Architecture

4.2 E-R Diagram

In software engineering, an entity relationship model (ER model) is a data model for describing the data or information aspects of a business domain or its process requirements, in an abstract way that lends itself to ultimately being implemented in a database such as a relational database. The main components of ER models are entities (things) and the relationships that can exist among them.

4.3 Data Flow Diagram

A data flow diagram (DFD) is a graphical representation of the 'flow' of data through an information system, modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design). A DFD shows what kind of information will be input to and output from the system, where the data will come from and go to, and where the data will be stored.

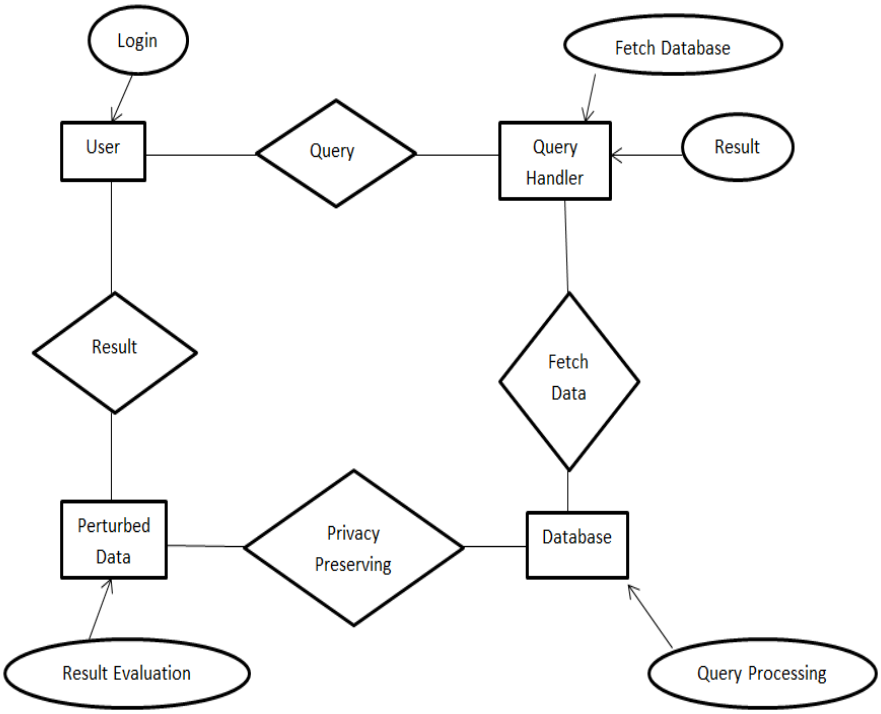

Figure 4.2: E-R Diagram

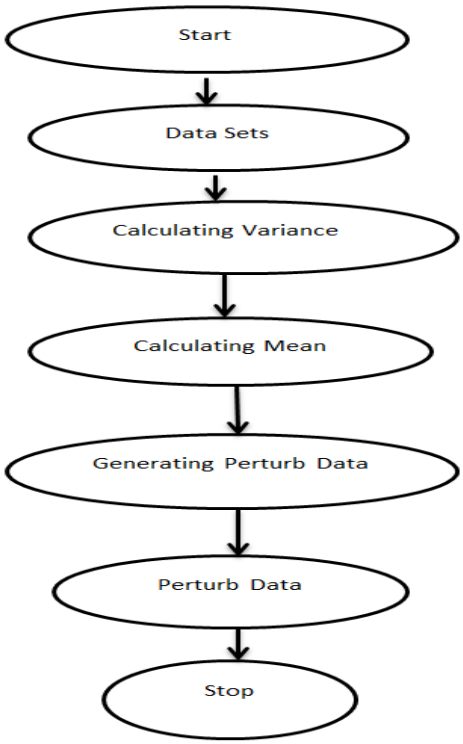

Figure 4.3: Data Flow Diagram for Tree Based Approach

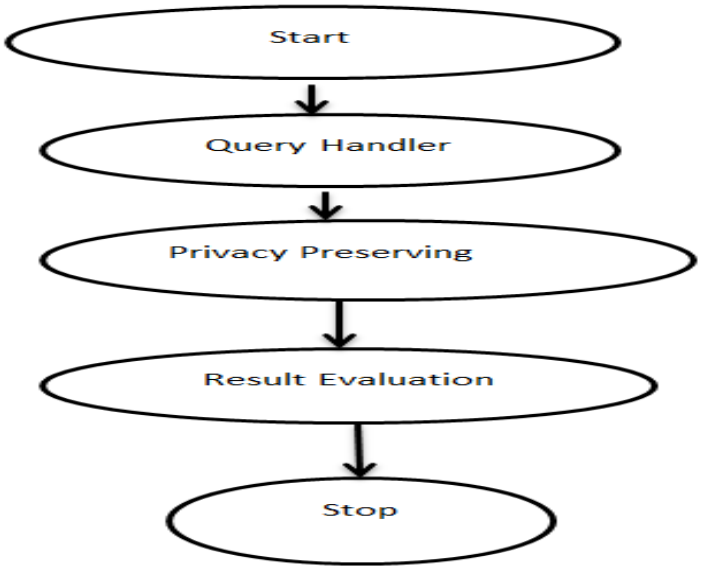

Figure 4.4: Data flow Diagram for Data Perturbation Approach

4.4 UML Diagrams

The UML is a language for: Visualizing, Specifying, Constructing and Documenting.

4.4.1 Use Case Diagram

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well.

4.4.2 Sequence Diagram

A Sequence diagram is a structured representation of behavior as a series of sequential steps over time. It is used to depict workflow, message passing and how element in general cooperate over time to achieve a result.

4.4.3 Activity Diagram

Activity diagram is another important diagram in UML to describe dynamic aspects of the system. Activity diagram is basically a flow chart to represent the flow form one activity to

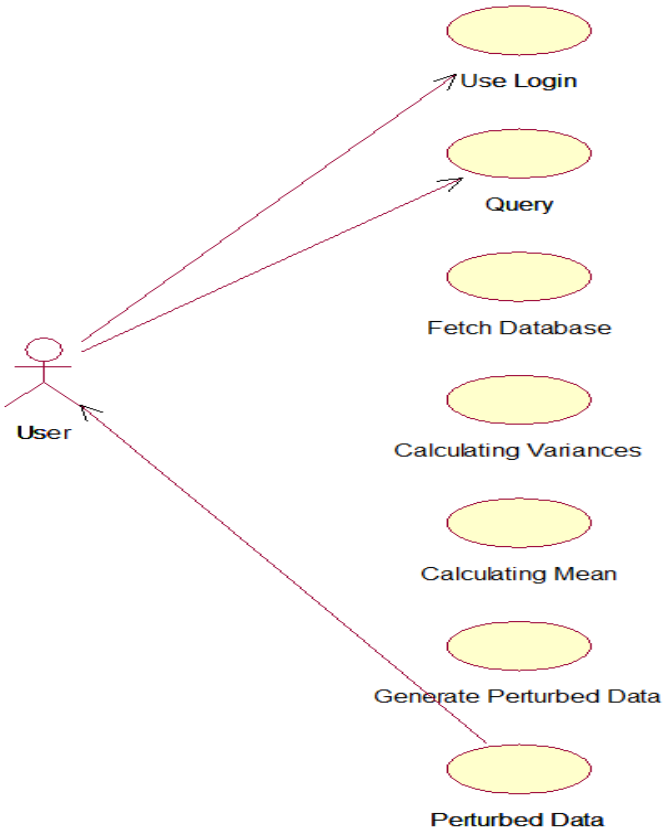

Figure 4.5: Use Case Diagram for Tree Based Approach

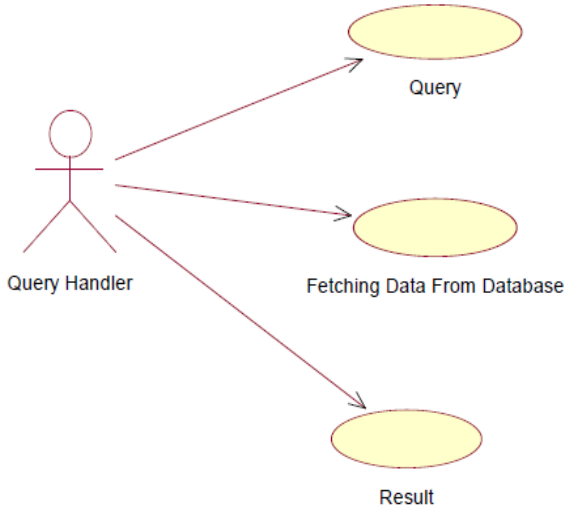

Figure 4.6: Use Case Diagram for Query Handler

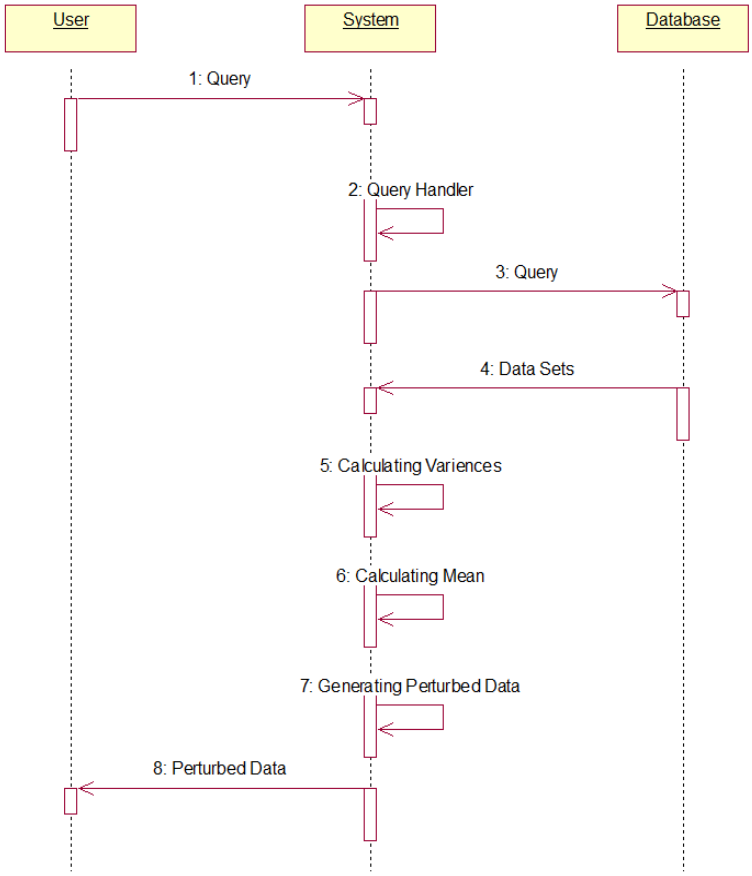

Figure 4.7: Sequences Diagram for Tree Based Approach

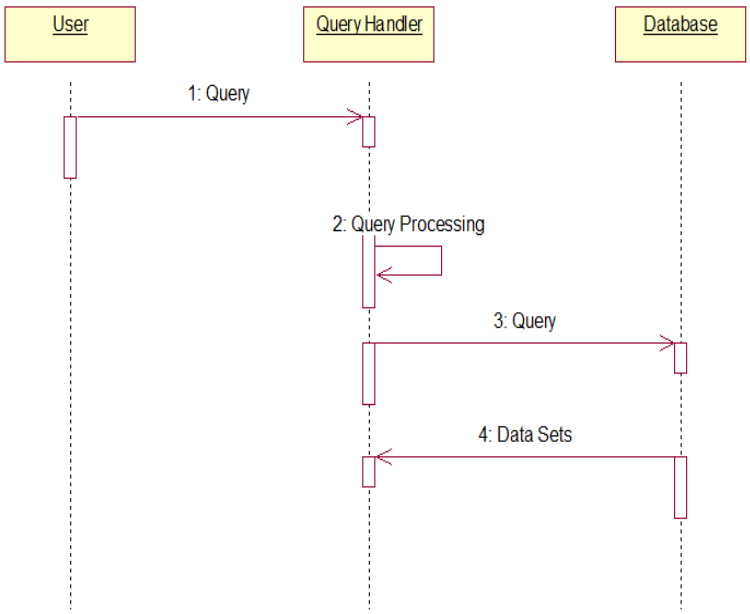

Figure 4.8: Sequences Diagram for Query Handler

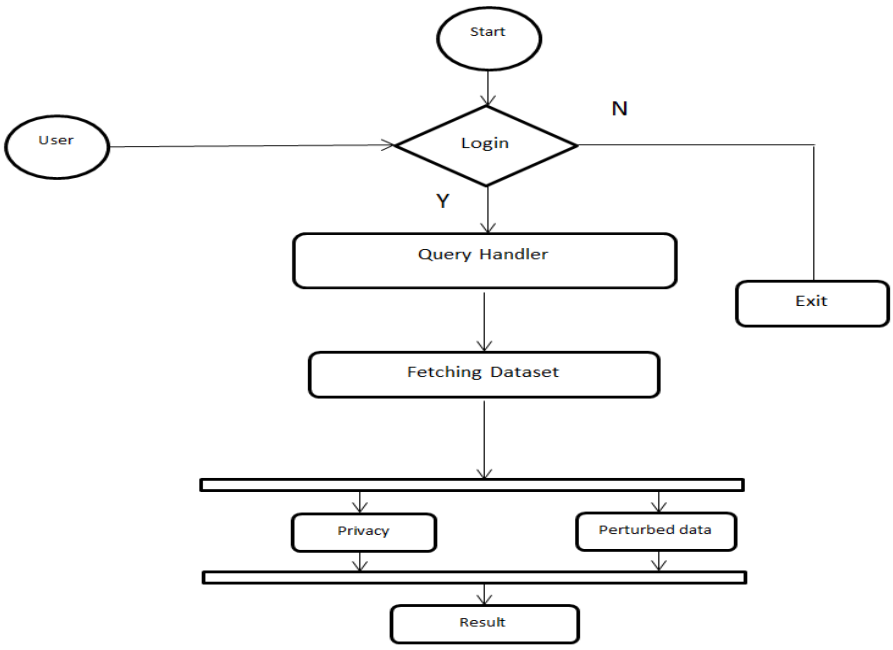

Figure 4.9: Activity Diagram

another activity. The activity can be described as an operation of the system. So the control flow is drawn from one operation to another.

4.4.4 State Diagram

A state diagram is a type of diagram used in computer science and related fields to describe the behavior of systems. State diagrams require that the system described is composed of a finite number of states; sometimes, this is indeed the case, while at other times this is a reasonable abstraction. Many forms of state diagrams exist, which differ slightly and have different semantics.

4.4.5 Component Diagram

In the Unified Modeling Language, a component diagram depicts how components are wired together to form larger components and or software systems. They are used to illustrate the structure of arbitrarily complex systems.

4.4.6 Deployment Diagram

Deployment diagram is a structure diagram which shows architecture of the system as deployment of software artifacts to deployment targets. Artifacts represent concrete elements

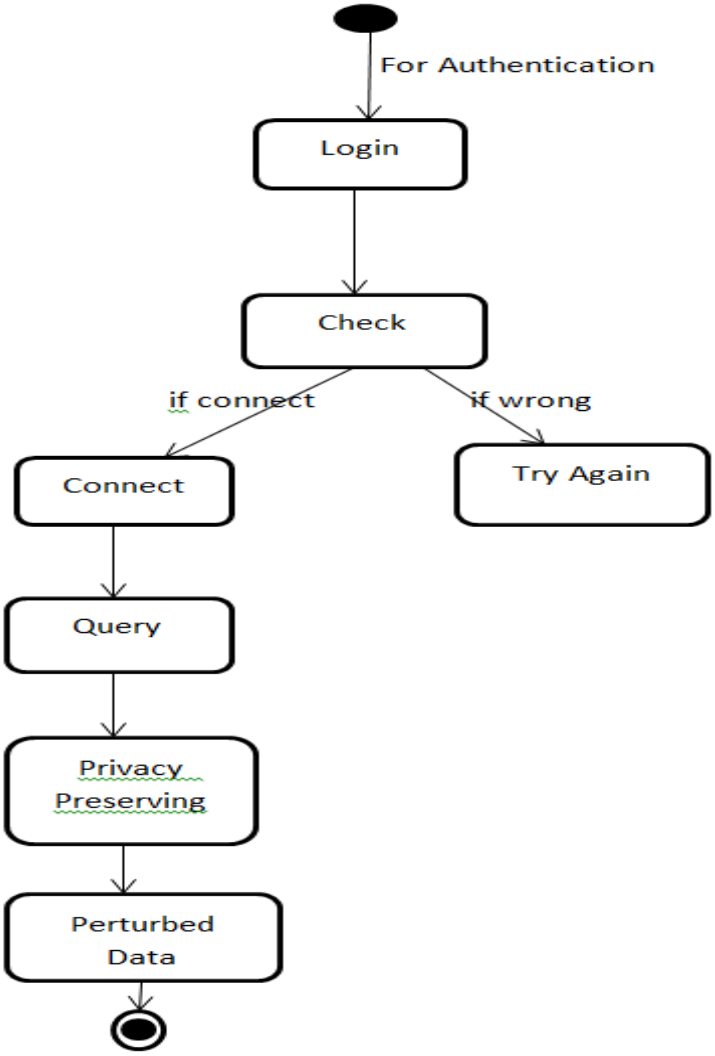

Figure 4.10: State Diagram

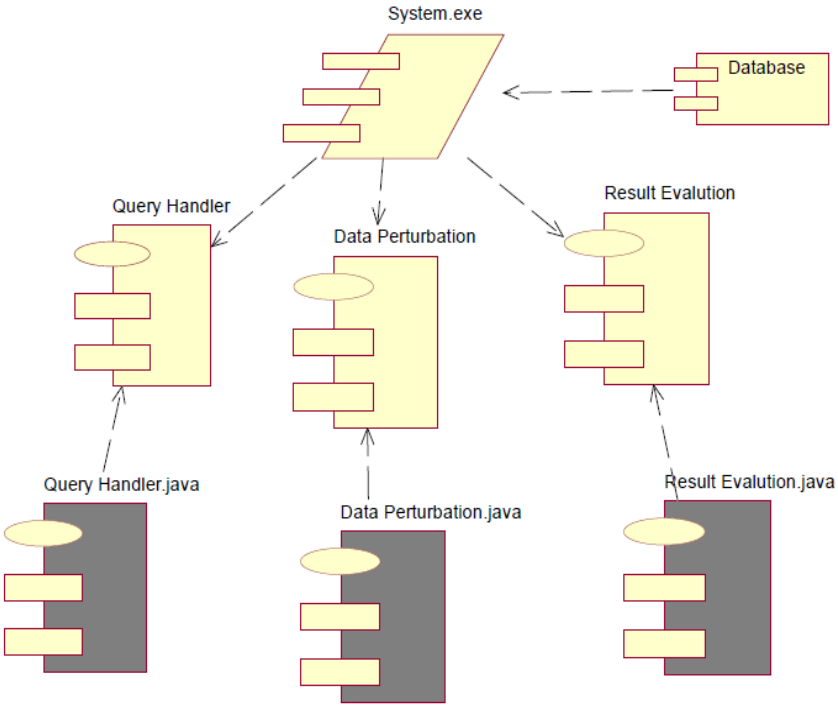

Figure 4.11: Component Diagram

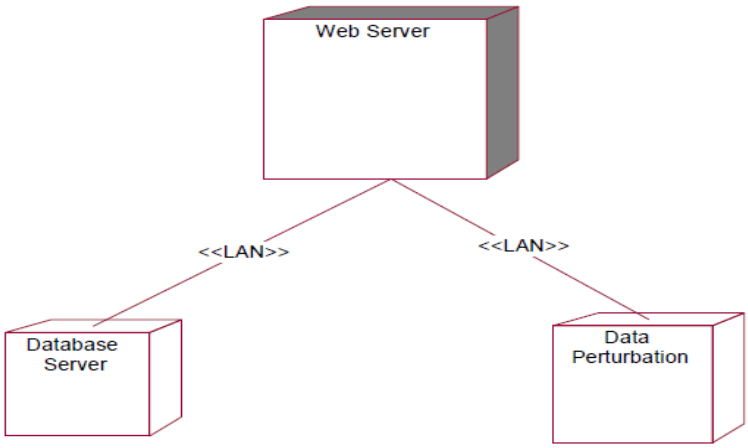

Figure 4.12: Deployment Diagram

in the physical world that are the result of a development process.

4.5 Summary

In this chapter, System Design is presented. In the next chapter, conclusion is presented

Chapter 5

Conclusion

Proposed approach focused on data perturbation by randomization noise addition to preserve privacy of sensitive attributes. The proposed mechanism will give very high performances and low error rate compared with existing methods. Typical challenge of mining the confidential data (sensitive data) from datasets will be solved by perturbation tree.

Bibliography

[1] Agrawal R., Srikant R, 'Privacy Preserving Data Mining' , ACM SIGMOD Conference, 2009.

[2] Lambodar Jena, Ramkrushna Swain, IEEE, 'Comparative study on Privacy Preserving Association Rule Mining Algo' , International Journal of Internet Computing, Vol.1, 2011.

[3] R. Agrawal and R. Srikant, 'Privacy Preserving Data Mining' , Proc. 2000 ACM SIGMOD Intl Conf. Management of Data, pp. 439- 450, 2000.

[4] N. R. Adam and J. C. Wortmann, 'Security Control Methods for Statistical Databases: A Comparative Study' , ACM Computing Surveys, vol. 21, no. 4, pp. 515-556, 1989.

[5] H. Kargupta, S. Datta, Q. Wang and K. Sivakumar, (2003) 'On the privacy preserving properties of random data perturbation techniques' , In Proceeding of the IEEE International Conference on Data Mining, Melbourne, FL. November, pp: 99-106.

[6] Polat H and Wenliang D, (2003) 'Privacy preserving collaborative filtering using randomized perturbation techniques' , In Proceedings of the 3rd IEEE International Conference on Data Mining, pp 625-628.

[7] Z. Huang, W. Du and B. Chen, (2005) 'Deriving private information from randomized data' , In Proceeding of SIGMOD, pp. 37-48, Baltimore, Maryland, USA.

[8] Xiao-Bai Li and Sumit Sarkar, 'A Tree Based Data Reconstruction Approach for Privacy Preserving Data Mining' , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 18, NO. 9.