A

Project I Report

on

PRESERVING THE PRIVACY OF USER BY USING ANONYMIZATION TECHNIQUE

Submitted in Partial Fulfillment of
the Requirements for the Degree
of

Bachelor of Engineering

in

Computer Engineering

to

North Maharashtra University, Jalgaon

Submitted by

Tanaya Gajanan Marathe
Madhura Nitin Raverkar
Sneha Gopalkrishna Suryawanshi
Minakshi Fakira More

Under the Guidance of

Mr. Sandip S. Patil

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 *Preserving The Privacy Of User By Using Anonymization Technique*, submitted by

Tanaya Gajanan Marathe
Madhura Nitin Raverkar
Sneha Gopalkrishna Suryawanshi
Minakshi Fakira More

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: October 3, 2016
Place: Jalgaon

Mr. Sandip S. Patil
Guide

Prof. Dr. Girish K. Patnaik
Head

Prof. Dr. K. S. Wani
Principal

Acknowledgements

No work can be accomplished unless it has evolved as a result of co-operating, assistance and understanding of some knowledgeable group of people. We take the opportunity to thank our Principal **Prof.Dr.K.S.Wani** and Head of Department **Prof. Dr.Girish K. Patnaik** for providing all the necessary facilities, which were indispensable in the completion of project. We would like to thank my guide **Mr.Sandip S. Patil** for providing to be a great help by giving us guidance through their vast experience and intellectual skills. We are also thankful to all the staff members of the Computer Engineering Department. We would also like to thank the college for providing the required magazines, books and access to the internet for collecting information related to the project. Finally, we would like to thank our parents.

Tanaya Gajanan Marathe
Madhura Nitin Raverkar
Sneha Gopalkrishna Suryawanshi
Minakshi Fakira More

Contents

List of Tables

List of Figures

Abstract

Today we are living in the complex world. In this, sensitive information privacy is the main issue. Many algorithms are used to protect sensitive information in mined data which is not efficient because resulted output can be easily linked with public data so it reveals user identity. There are various techniques to protect privacy in data mining. Data mining approaches aim to avoid direct use of sensitive data. Proposed system uses anonymization techniques which help to eliminate privacy risk in data preparation. Existing anonymization methods are only apt for single sensitive and low dimensional data to keep up with privacy specifically like generalization and bucketization. In the proposed work, an anonymization technique is given that is a combination of the benefits of anatomization, and enhanced slicing approach adhering to the principle of k-anonymity and l-diversity for the purpose of dealing with high dimensional data along with multiple sensitive data. The anatomization approach dissociates the correlation observed between the quasi identifier attributes and sensitive attributes and yields two separate tables with non-overlapping attributes. In the enhanced slicing algorithm, vertical partitioning does the grouping of the correlated sensitive attributes in sensitive table together and thereby minimizes the dimensionality by employing the advanced clustering algorithm. The experimental outcomes indicate that the proposed method can preserve privacy of data with numerous sensitive attribute. The anatomization approach minimizes the loss of information and slicing algorithm helps in the preservation of correlation and utility which in turn results in reducing the data dimensionality and information loss.

KEYWORDS: Privacy, anonymization approaches, mined data .

Chapter 1

Introduction

Todays health care providers store stores a huge amount of sensitive data as a content of their business. The sensitive information can be personally recognizable information from the clients. So in this any kind of misuse of this information creates a critical risk in their business. When making the sensitive information available to the public, it is necessary for them to protect it from any abuse and risk. Data anonymization is the one and only popularly used approach. It modifies and changes information, keeping in mind to make it difficult to link individuals with their data. This methodology tries to ensure the identity along with the sensitive information of the data subjects when data is shared for diverse purposes. There are multiple anonymization methods that prevail for retaining privacy. They are namely generalization, suppression, anatomization, bucketization, permutation, and perturbation. Generalization and suppression concentrate on quasi-identifier attributes, whereas bucketization is focused on splitting sensitive attributes from quasi-identifier attributes with a description that is less specific. Anatomization and permutation dissociate the correlation between quasi-identifier attributes and sensitive attributes the by collection and rearrangement of sensitive values in a qid group. Perturbation tampers the data by the addition of noise, aggregation of values, swapping of values, or generation of artificial data or by the encryption of the data, in the light of few measurable characteristics of the first information. Slicing is a technique that can tackle with high dimensional data and hence preserve privacy and improve utility. The majority of the strategies above focus on anonymizing the micro data with only single sensitive attributes. As they are not suitable for functional usage, the current challenge is to preserve the multiple sensitive attributes efficiently in the high dimensional data.

This chapter is of 7 sections. First section 1.1 describes background of the project, motivations behind the project explains in section 1.2, section 1.3 will give problem definition of our project, scope of the project is defined in section 1.4, section 1.5 gives objectives of

the project, organization of the whole project is given in section 1.6 & section 4.5 gives summary.

1.1 Background

Privacy preserving data mining is a rapidly growing research area aiming at eliminating privacy breaches which may happen during the mining of data (Verykios et al. 2004; Kantarcioglu et al. 2004; Clifton 2009). The goal of privacy preserving data mining algorithm is to alter the original data for the purpose of maintaining privacy, leading to a low degree of data leakage. This will give way for obtaining good mining results. The work introduced in (Verykios et al. 2004) observes the privacy preserving data mining approach in the light of five different dimensions. They are data distribution, data modification, data mining algorithm, data or rule hiding and privacy preservation. Data distribution represents the organization of data that can be centralized or in a distributed fashion. The second step data modification refers to modifying the data. The work introduced in (Friedman et al. 2008) yields the possibilities for the construction of k-anonymous data models with k-anonymous data sets. More commonly, the k-anonymity concept is utilized by the PPDM algorithms in order to guarantee privacy (El Emam and Dankar 2008). It is a problem to be able to find optimal k-anonymous datasets through generalization and is rated as NP-Hard (Gedik and Liu 2008; Meyerson and Williams 2004). The work provided in Fung et al. (2007) introduces a generalization method for classification through the application of k-anonymity and it is a top down specialization algorithm. Anatomization (Xiao and Tao 2006) in contrast to generalization and suppression does not make modifications to the quasi-identifier or the sensitive attributes, but rather dissociates the relationship between the two. The greatest benefit of anatomy is that there is no modification of data in both quasi-identifier and sensitive attributes. Xiao and Tao proved that the anatomized tables can answer aggregate queries dealing with domain values of the quasi-identifier and sensitive attributes more accurately compared to the generalization approach. Tao et al. (2009) proposed an approach referred to as a permutation, sharing the same kind of spirit of anatomization. But all of the above methods suit single sensitive data only. In order to deal with multiple sensitive attribute a multiple sensitive bucketization (MSB) (Yanget al. 2008) is suggested. But it is appropriate for attributes less than three only. In our project, an anonymization technique is proposed that is a combination of the benefits of anatomization, and enhanced slicing approach adhering to the principle of k-anonymity and l-diversity for the purpose of dealing with high dimensional data along with multiple sensitive data.

1.2 Motivation

The generalization for k-anonymity and bucketization for l-diversity are popularly understanding privacy preservation strategies. Generalization for k-anonymity ignores a huge measure of data in case they are high dimensional data. In order to get over deformities in generalization, an inventive anatomization methodology is brought into use It reduces the data loss, although it is capable of preserving privacy for single sensitive data only. The anatomization preserves privacy as it is not representative of the sensitive value corresponding to any tuple, which might be assumed randomly from sensitive table. A larger l indicates more privacy. So taking the significance of both individuals privacy and utility into consideration, an algorithm, called as kl-redInfo, is proposed which enhances the anatomy algorithm. This is performed by presenting new approaches with the systematic integration of the remaining records, cell-based generalization in place of separation of the table into two parts, and sorting the records as per their quasi-identifier attributes for the purpose of reducing the total amount of information loss. Motivated by means of these works, we focuses on the preservation of the privacy of data with numerous sensitive attributes with lesser information loss and better data utility. In this work an anatomization approach is employed to minimize the information loss by releasing the quasi-identifier attributes directly. On the contrary, slicing maintains the correlation in the column and then carries out the break of correlation across the columns by means of vertical and horizontal partitioning. Slicing does the permutation of the sensitive attributes within each bucket in order to carry on the correlation break across the columns, and assures privacy. Every attribute in a column can be considered in the form of a sub table. This removes the dimensionality with respect to the data. Additionally, the research work functions in accordance with the principle of k-anonymity and l-diversity that does not impact the quasi-identifier values which are directly released by means of anatomization. This provides the way for preventing membership, identity and attributes disclosures.

1.3 Problem Definition

Occasionally, the data collector and data miner are different entities, with the data collector processing data from its original owners and then making it available to the data miner. Processing must be done in a way that makes it impossible for the data miner to identify the data owners identities and other sensitive information, yet still produce data that the data miner finds useful.This is the main need of the project. For that purpose we used privacy preserving techniques such as anonymization technique.Thus we mainly focus on anatomization with slicing anonymization technique.

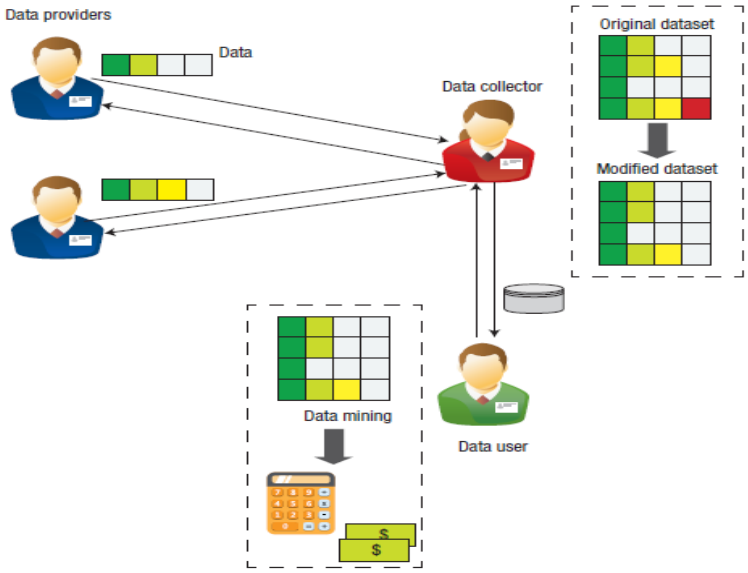

Figure 1.1: Scenario of Data Collection, Data Publishing, Data Mining

1.4 Scope

1.4.1 Need of Privacy Preserving

Protection of oneself from being disclosed to unauthorized people, is the main aim of privacy preservation. Privacy preservation is considered as a vital factor for adequate usage of the large volume of data. It preserves privacy during data collection from various sources, and when retrieving knowledge in mining operations. The figure 1.2 is shown below.

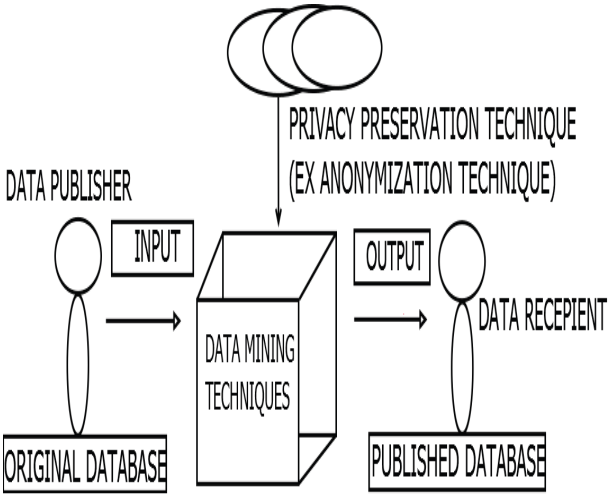

Figure 1.2: Flow of Data Collection and Publishing Data

1.5 Objective

The main aspect of this project is given below:

- To introduce a novel data anonymization technique called anatomization with slicing to improve the current state of the art.
- The anatomization approach dissociates the correlation observed between the quasi identifier attributes and sensitive attributes and yields two separate tables with non-overlapping attributes.
- To develop an efficient algorithm for computing the sliced table that satisfies l-diversity. Our algorithm partitions attributes into columns, applies column generalization, and partitions tuples into buckets. Attributes that are highly-correlated are in the same column.
- The anatomization approach minimizes the loss of information and slicing algorithm helps in the preservation of correlation and utility which in turn results in reducing the data dimensionality and information loss.

1.6 Organization of Report

- Chapter 2 describes the overall system analysis. It includes literature survey, proposed system, feasibility study, risk analysis, project scheduling and effort allocation.
- Chapter 3 gives system requirement specification such as hardware requirements, software requirements, functional requirements, non-functional requirements and other requirements and constraints.
- Chapter 4 shows system design by drawing system architecture, E-R diagram, Data flow diagram, user interface design and all important UML diagrams.

1.7 Summary

In this chapter, an overview of the problem statement along with its solution for the work contained in this dissertation is provided. In the next chapter, related work in the area of system analysis is presented.

Chapter 2

System Analysis

System analysis is the states interacting entities, including computer system analysis . This field is closely related to requirement analysis or operation research. Here we discussed system analysis briefly.

In this chapter, section 2.1 will discuss literature survey, proposed System will discuss in section 2.2, section 2.3 will discuss feasibility study, economical feasibility, operational feasibility, and technical feasibility, risk analysis will discuss in section 2.4, section 2.5 will discuss project scheduling, effort allocation will discuss in section 2.6, section 4.5 gives summary.

2.1 Literature Survey

2.1.1 Slicing: A New Approach for Privacy Preserving Data Publishing

Several anonymization techniques, such as generalization and bucketization, have been designed for privacy preserving microdata publishing. Recent work has shown that generalization loses considerable amount of information, especially for high-dimensional data. Bucketization, on the other hand, does not prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. In this paper, we present a novel technique called slicing, which partitions the data both horizontally and vertically. We show that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. They show how slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the l-diversity requirement.

2.1.2 Privacy Preserving Approaches for Multiple Sensitive Attributes in Data Publishing

Current privacy preserving data publishing techniques concentrate on tables with only one sensitive attribute. However, most of the real-world applications contain multiple sensitive attributes. Directly applying the existing single-sensitive-attribute privacy preserving techniques often causes unexpected private information disclosure. This paper firstly discusses the problem of secure publishing data when sensitive data contains multi attributes, and then propose a multi-dimensional bucket grouping approach on the idea of lossy join, called Multi-Sensitive Bucketization (MSB). In order to avoid exhausting search, three specific line-time greedy based MSB algorithms are proposed, which are maximal-bucket first algorithm (MBF), maximal single-dimension-capacity first algorithm (MSDCF), and maximal multi-dimension-capacity first algorithm (MMDCF). In addition, according to the differences among published data, a weighted MSB approach is further proposed. Experimental results on the real-world datasets show that the addition information loss of the proposed MSB methods were not more than 0.04 and the suppression ratios were less than 0.06.

2.1.3 A General Survey of Privacy-Preserving Data Mining Models and Algorithms

In recent years, privacy-preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. A number of algorithmic techniques have been designed for privacy-preserving data mining. In this paper, a review of the state-of-the-art methods for privacy is given . This paper also gives methods for randomization, k-anonymization, and distributed privacy-preserving data mining. The cases in which the output of data mining applications needs to be sanitized for privacy-preservation purposes are discussed in paper.The information about the computational and theoretical limits associated with privacy-preservation over high dimensional data sets is also given in this paper.

2.1.4 Existing System

First, many existing clustering algorithms (e.g., k- means) requires the calculation of the centroids. But there is no notion ofcentroidsin our setting where each attribute forms a data point in the clustering space. Second, k-medoid method is very robust to the existence of outliers (i.e., data points that are very far away from the rest of data points). Third, the order in which the data points are examined does not affect the clusters computed from the k-medoid method. bf Disadvantages of existing system

1. Existing anonymization algorithms can be used for column generalization, e.g.,Mondrian . The algorithms can be applied on the subtable containing only attributes in one column to ensure the anonymity requirement.

2. Existing data analysis (e.g., query answering) methods can be easily used on the sliced data.

3. Existing privacy measures for membership disclosure protection include differential privacy and presence.

2.2 Proposed System

We present a novel technique called anatomization with slicing, which partitions the data both horizontally and vertically. We show that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of this, is that it can handle high-dimensional data. We show how slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the l-diversity requirement. Our workload experiments confirm that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. Also the anatomization approach minimizes the loss of information.

2.2.1 Module Description

- Original Data: We conduct extensive workload experiments. Our results confirm that slicing preserves much better data utility than generalization. In workloads involving the sensitive attribute, slicing is also more effective than bucketization. In some classification experiments, slicing shows better performance than using the original data.

- Generalized Data: Generalized Data, in order to perform data analysis or data mining tasks on the generalized table, the data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible, as no other distribution assumption can be justified. This significantly reduces the data utility of the generalized data.

- Bucketized Data: We show the effectiveness of slicing in membership disclosure protection. For this purpose, we count the number of fake tuples in the sliced data. We also compare the number of matching buckets for original tuples and that for fake

tuples. Our experiment results show that bucketization does not prevent membership disclosure as almost every tuple is uniquely identifiable in the bucketized data.

- Multiset-based Generalization Data: We observe that this multiset-based generalization is equivalent to a trivial slicing scheme where each column contains exactly one attribute, because both approaches preserve the exact values in each attribute but break the association between them within one bucket.

- K-anonymity: In K-anonymity, within an anonymized table each testimony must be tantamount with at least k-1 other testimony within the data file, with respect to a set of quasi-identifier attributes or if one testimony in the table has a few quasi-identifier, at least k-1 other testimony also have the value quasi-identifier. quasi-identifier should have at least k minimal group size value. Originality disclosure is protected by K-anonymity.

- One-attribute-per-Column Slicing Data: We observe that while one-attribute-per-column slicing preserves attribute distributional information, it does not preserve attribute correlation, because each attribute is in its own column. In slicing, one groups correlated attributes together in one column and preserves their correlation.

- Sliced Data: Another important advantage of slicing is its ability to handle high-dimensional data. By partitioning attributes into columns, slicing reduces the dimensionality of the data. Each column of the table can be viewed as a sub-table with a lower dimensionality. Slicing is also different from the approach of publishing multiple independent sub-tables in that these sub-tables are linked by the buckets in slicing.

2.3 Feasibility Study

The feasibility analysis shows the developers all the aspects of the project and they can know that whether the project is practically possible to develop worth limited resources and time. There are few types of feasibility are exist so developer should take care of these feasibility or developer must aware about these feasibility.

- Economical Feasibility

- Operational Feasibility

- Technical Feasibility

2.3.1 Technical Feasibility

At first it is necessary to check that system (proposed system) is technically feasible or not. Also to determine the technology and skill it is necessary to carry out the project.If they are not available then find out solutions for them.In our project few types of tools used such as Tomcat, MySQL 5.5, Smart Draw tool, JDK kit to develop the system. Hence the project is technically feasible.

2.3.2 Operational Feasibility

Operational feasibility is beneficial if they can be turned into information system will meet the organization operating requirement.The system is user friendly. This feasibility in which only one machine is require. This feasibility is operational feasibility because no cost is required.Propose system has high operational feasibility.It provides security and privacy by using anonymization techniques.

2.3.3 Economical Feasibility

Economic feasibility is a cost benefit. In our project, we require Tomcat software and also we require java development tool kit this is also freeware software. Also we require MySQL database for storage which freeware software, so our project is feasible for developer and client. In our project java and MySQL are connected to connector.jar file. Hence the project is cost efficient.

2.4 Risk Analysis

2.4.1 Risk

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 problemit might happen, it might not. But, regardless of the outcome, its a really 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.

2.4.2 Need of Risk Analysis

Think about the Boy Scout motto: Be prepared. Software is a difficult undertaking. Lots of things can go wrong, and frankly, many often do. Its for this reason that being prepared

understanding the risks and taking proactive measures to avoid or manage themis a key element of good software project management.

2.4.3 Software Risk

Although there has been considerable debate about the proper definition for software risk, there is general agreement that risk always involves two characteristics

Uncertainty The risk may or may not happen; that is, there are no 100 percent probable risks.

Loss If the risk becomes a reality, unwanted consequences or losses will occur.

2.4.4 Project Risk

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.

2.4.5 Technical Risks

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, and "leading-edge" technology are also risk factors. Technical risks occur because the problem is harder to solve than we thought it would be.

2.4.6 Business Risk

Threaten the viability of the software to be built. Business risks often jeopardize the project or the product. Candidates for the top five business risks are:

1. Building a excellent product or system that no one really wants.
2. Building a product that no longer fits into the overall business strategy for the company.
3. Building a product that the sales force doesn't understand how to sell.
4. Losing the support of senior management due to a change in focus or a change in people.

5. Losing budgetary or personnel commitment. It is extremely important to note that simple categorization won't always work. Some risks are simply unpredictable in advance.

2.5 Project Scheduling

Software project scheduling is an activity that distributes estimated effort across the planned project duration by allocating the effort to specific software engineering tasks.It is important to note, however, that the schedule evolves over time. During early stages of project planning, a macroscopic schedule is developed. This type of schedule identifies all major software engineering activities and the product functions to which they are applied. As the project gets under way, each entry on the macroscopic schedule is refined into a detailed schedule.Project scheduling can be done by Gantt chart shown in figure 2.1

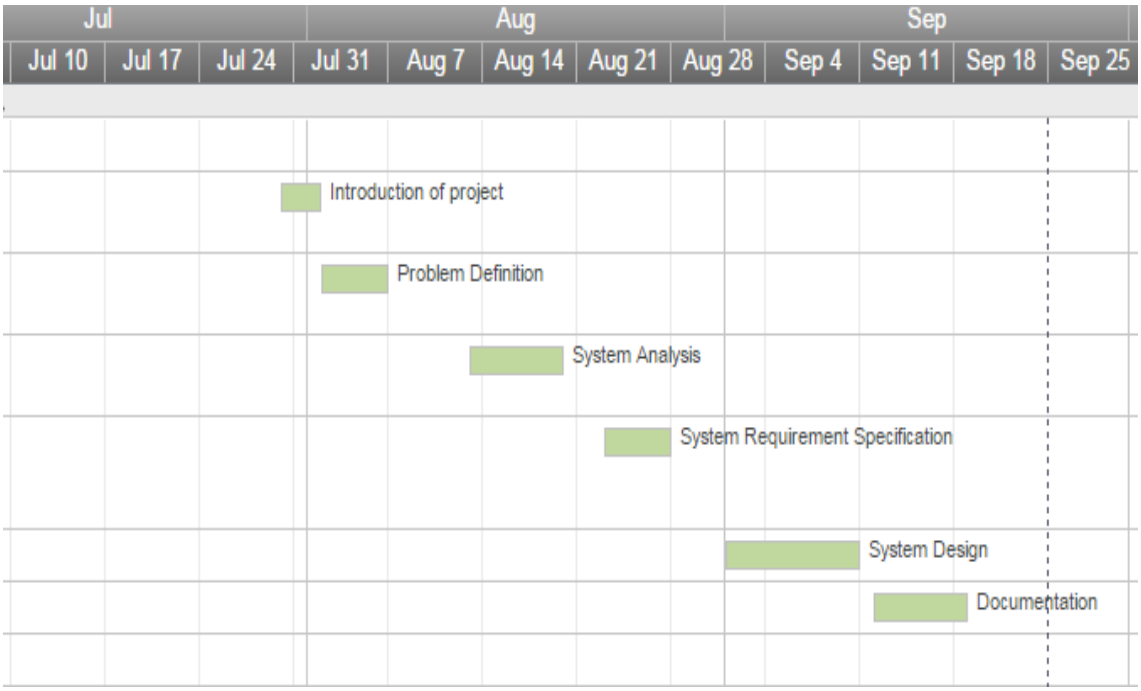

Figure 2.1: Gantt Chart

2.6 Effort Allocation

Software engineer or a team of engineers must incorporate a development strategy that encompasses the process, methods, and tools layers described. This strategy is often referred to as a process model or a software engineering paradigm. A process model for software engineering is chosen based on the nature of the project and application, the methods and

tools to be used, and the controls and deliverables that are required. There are many process models in the software engineering, but we have chosen Water Fall Model because the project is totally dependent on previous modules. Another Reason for choosing this model is to provide better user satisfaction.

Table 2.1: Effort Allocation Chart

	Tanaya	Madhura	Sneha	Minakshi
Introduction	✓			✓
System Analysis	✓	✓	✓	
System Requirement Specification	✓			✓
System Design		✓	✓	

Allocation of efforts to all project partners- Project means team work; Project is developed by combination of effort of team. So whole project is divided into modules and number of modules is allotted to team members. After completion of each module, it will be link from one module to another module to form a complete project.This effort allocation should be used as a guideline only. The characteristics of each project must dictate the distribution of effort. Work expended on project planning rarely accounts for more than 2-3 percent of effort, unless the plan commits an organization to large expenditures with high risk. Requirements analysis may comprise 1025 percent of project effort. Effort expended on analysis or prototyping should increase in direct proportion with project size and complexity. A range of 20 to 25 percent of effort is normally applied to software design.

2.7 Summary

In this chapter, section 2.1 discuss literature survey, proposed system explained in section 2.2, section 2.3 discuss feasibility study, economical feasibility operational feasibility, and technical feasibility discuss, risk analysis discuss in section 2.4, section 2.5 discuss project scheduling, Effort Allocation discuss in section 2.6.

Chapter 3

System Requirement Specification

In this chapter, section 3.1 will explain Hardware requirements, Software requirements will explain in section 3.2, section 3.3 will explain Functional requirements, Non-Functional requirements will explain in section 3.4 and section 4.5 gives summary.

3.1 Hardware Requirements

Computer system is require for project, other than that no need of any extra hardware requirement in our project.

- Processor - Pentium III
- Speed - 1.1 Ghz
- RAM - 256 MB(min)
- Hard Disk - 20 GB
- Floppy Drive - 1.44 MB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

3.2 Software Requirements

For these project, few softwares are required to be installed on computer system. As we are used Java language for coding, we need java runtime environment(jre) for execution. Also for storing data we require database server,for our project we are using Mysql. Software that are required for our project are as follows:

- Operating System :Windows 7 or minimum configuration os.
- Application Server : Tomcat5.0/6.X
- Front End : HTML, Java, Jsp
- Scripts : Java Script.
- Server side Script : Java Server Pages.
- Database Connectivity : Mysql.

3.3 Functional Requirements

In general, functional requirement defines what a system is supposed to do. List of functional requirements:-

- Original Data
- Generalized Data
- Bucketized Data
- Multiset-based Generalization Data
- One-attribute-per-Column Slicing Data
- Sliced Data

3.4 Non-Functional Requirements

Non-functional requirements that specify criteria that can be used to judge the operation of a system, rather than specific behaviours. This should be contrasted with functional requirements that specify specific behaviour or functions. In general non-functional requirement define how a system is supposed to be. List of Non-functional requirements :- Accessibility, Data Backups and Maintainability

3.5 Summary

In this chapter, section 3.1 explained Hardware requirements, Software requirements explained in section 3.2, section 3.3 explained Functional requirements, Non-Functional requirements explained in section 3.4. In the next chapter we will discuss about overall system design of our project.

Chapter 4

System Design

System design is a process of designing the architecture, components, modules, interfaces and data for a system to satisfy specified requirement. In this chapter, section 4.1 discussed system architecture. Section 4.2discussed E-R diagram. Data flow diagram discussed in section 4.3. All UML diagrams discussed in section 4.4.And last section 4.5 shows summary.

4.1 System Architecture

Software architecture is the development work product that gives the highest return on investment with respect to quality, schedule and cost. Software architecture alludes to the overall structure of the software and the ways in which that structure provides conceptual integrity for a system. The architectural design description should address how the design architecture achieves requirements for performance, capacity, reliability, security, adaptability, and other characteristics. The system architectural is shown in given figure 4.1

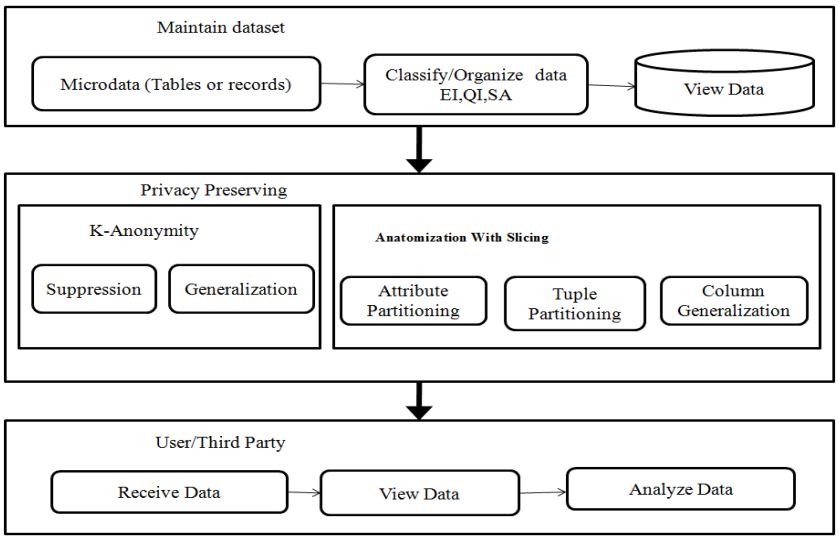

Figure 4.1: System Architecture

4.2 E-R Diagram

The entity relationship data model is based on a perception of a real world that consist of a collection of basic objects called entities, and relation among these objects.E-R diagram is shown in following figure 4.2

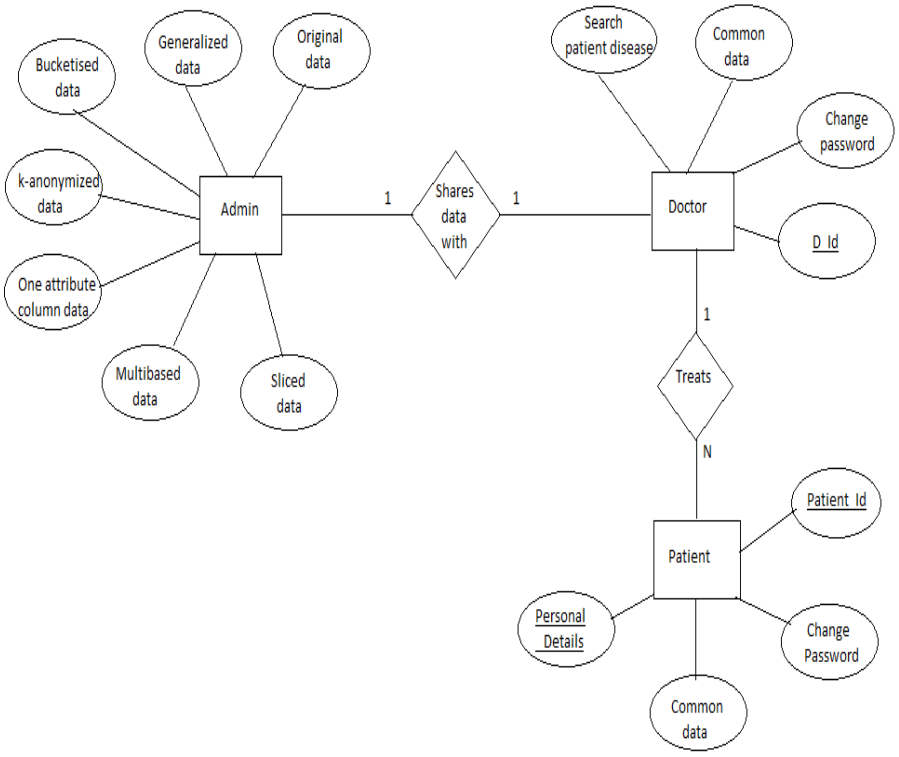

Figure 4.2: E-R Diagram

4.3 Data Flow Diagram

A DFD is a graphical technique that depicts the information flow and the transformation that we have applied as the data moves from input to output. The data flow diagram also known as data flow graph or bubble chart. A data flow diagram may be used to represent a system or software at any level of abstraction. The data flow diagram can be completed using only four simple notations i.e. special symbols or icons and the annotation that with a specific system.A data flow diagram (DFD) is a graphical technique that despites information about flow and that are applied as data moves from input to output. The DFD is also called as data flow graph or bubble chart. Named circles show the processes in DFD or named arrows entering or leaving the bubbles represent bubbles and data flow. A rectangle represents a source or sink and is not originate or consumer of data. Data flow diagrams are the basic building blocks that define the flow of data in a system to the particular destination and

difference in the flow when any transformation happens. The data flow diagram serves two purposes: (1) To provide an indication of how data are transform as the moves through the system. (2) To depict the function that transforms the data flow.

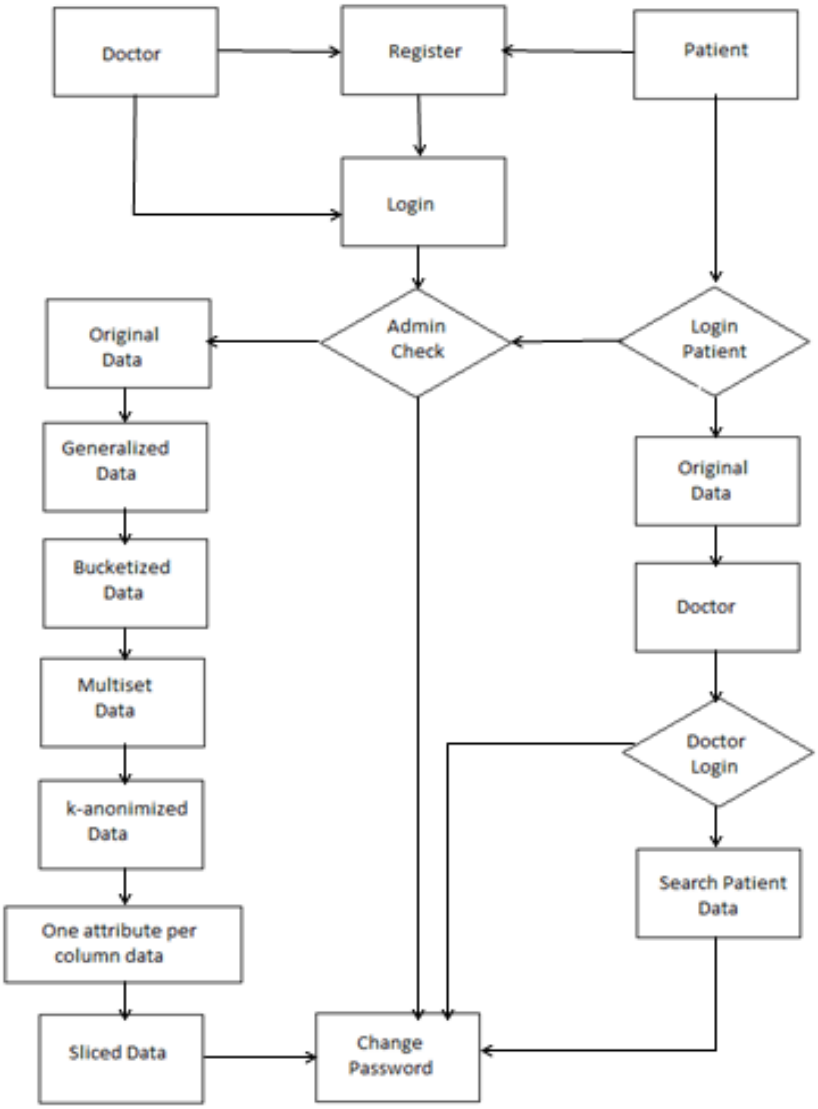

Figure 4.3: Data Flow Diagram-Level 2

4.4 UML Diagrams

The Unified Modeling Language (UML) is a standard language for writing software blueprints. UML may be used to visualize, specify, construct, and document the artifacts of a software-intensive system.Grady Booch, Jim Rumbaugh, and Ivar Jacobson developed UML in the mid 1990s with much feedback from the software development community. UML provides different diagrams for use in software modeling.

4.4.1 Use Case Diagram

UML use-case diagram 4.4 help you determine the functionality and features of the software from the users perspective. A use case describes how a user interacts with the system by defining the steps required to accomplish a specific goal.

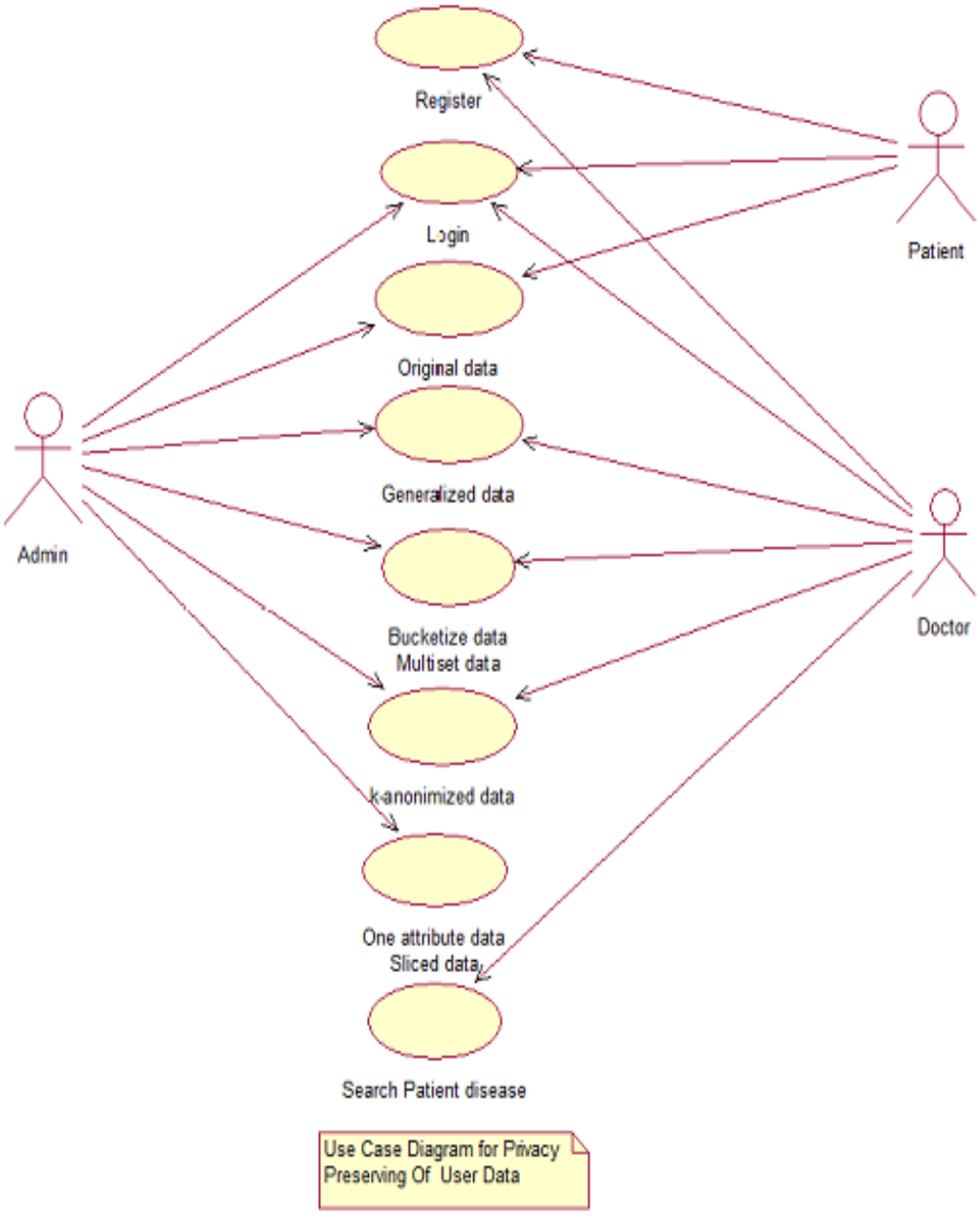

Figure 4.4: Use Case Diagram

4.4.2 Class Diagram

The main elements of a class diagram 4.5 are boxes, which are the icons used to represent classes and interfaces. Each box is divided into horizontal parts. The top part contains the name of the class. The middle section lists the attributes of the class. An attribute refers to something that an object of that class knows or can provide all the time.The third section of the class diagram contains the operations or behaviors of the class.

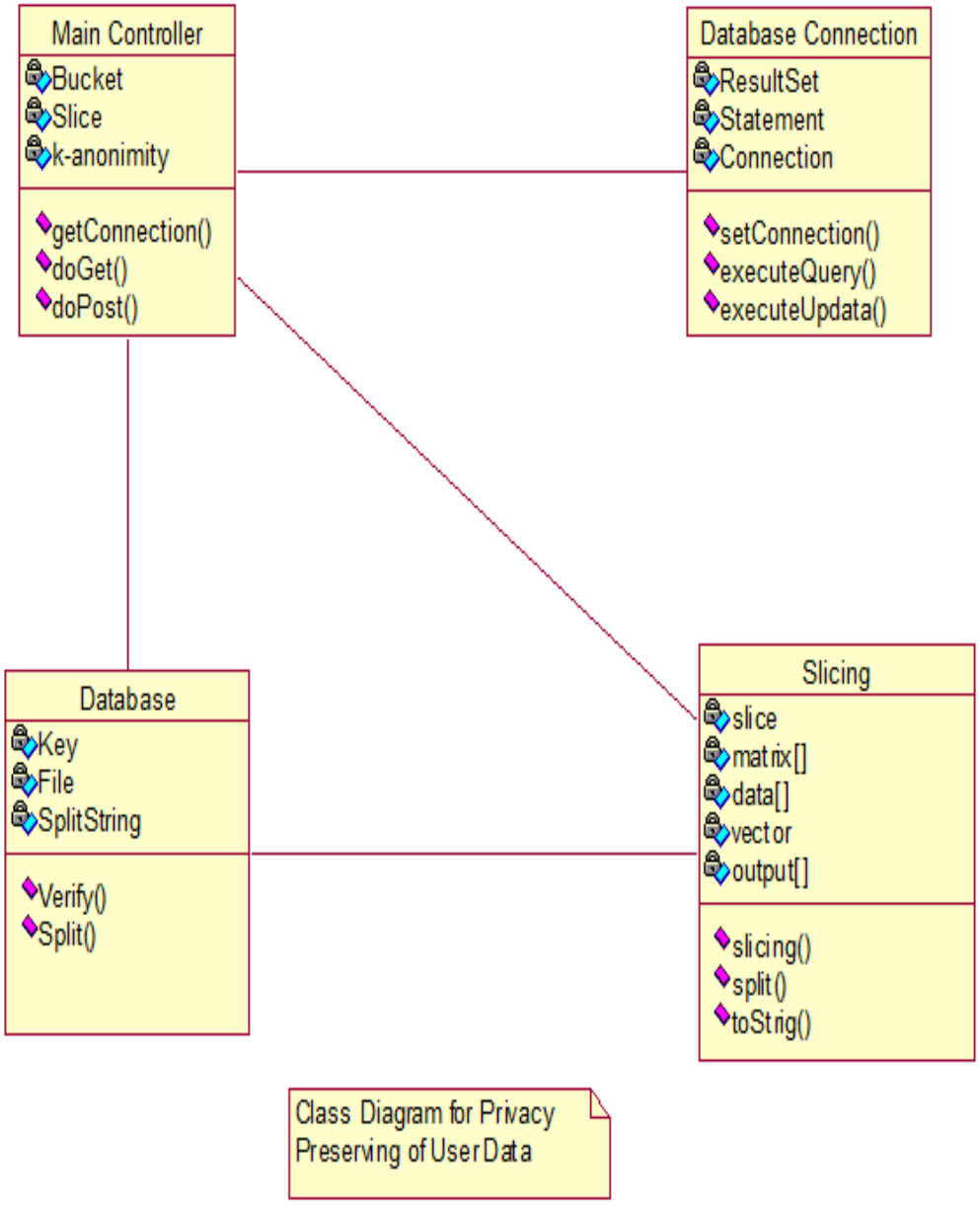

Figure 4.5: Class Diagram

4.4.3 Sequence Diagram

A sequence diagram 4.6 is used to show the dynamic communications between objects during execution of a task. It shows the temporal order in which messages are sent between the objects to accomplish that task. One might use a sequence diagram to show the interactions in one use case or in one scenario of a software system.

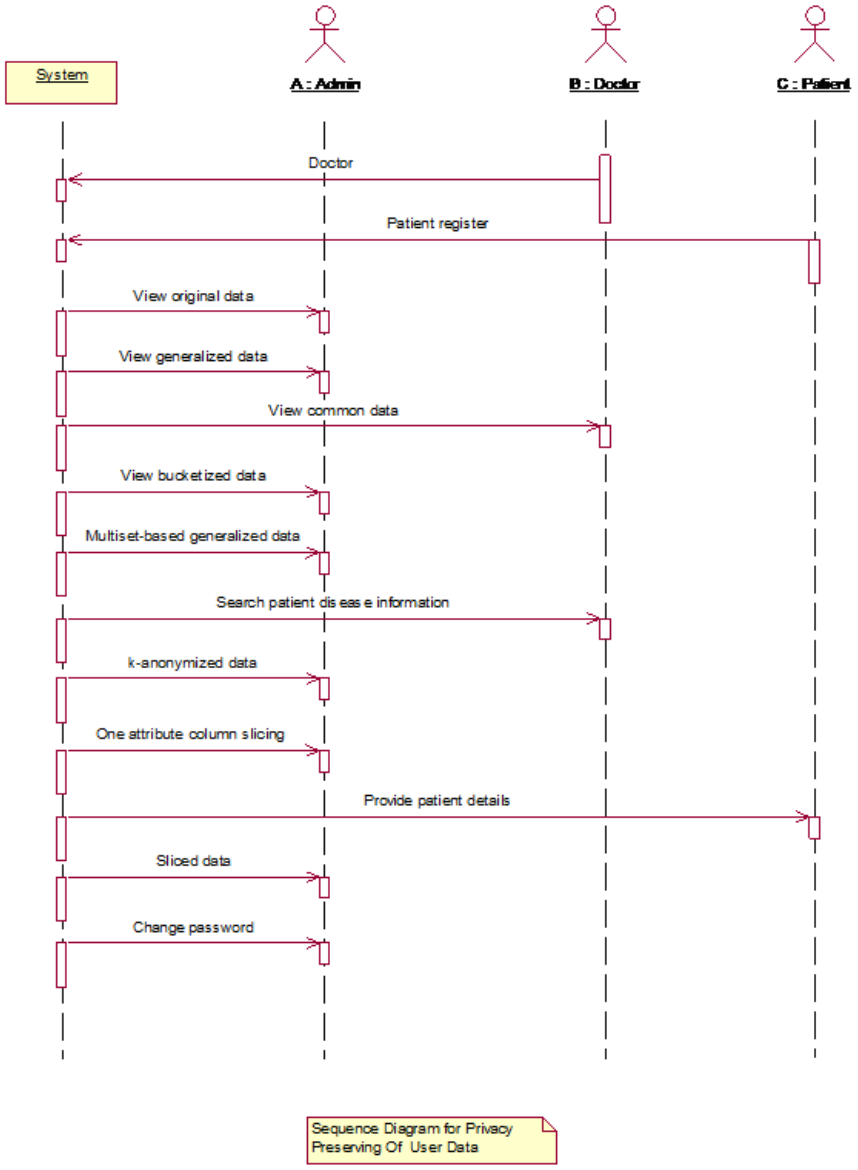

Figure 4.6: Sequence Diagram

4.4.4 Activity Diagram

A UML activity diagram 4.7 depicts the dynamic behavior of a system or part of a system through the flow of control between actions that the system performs. It is similar to a flowchart except that an activity diagram can show concurrent flows.

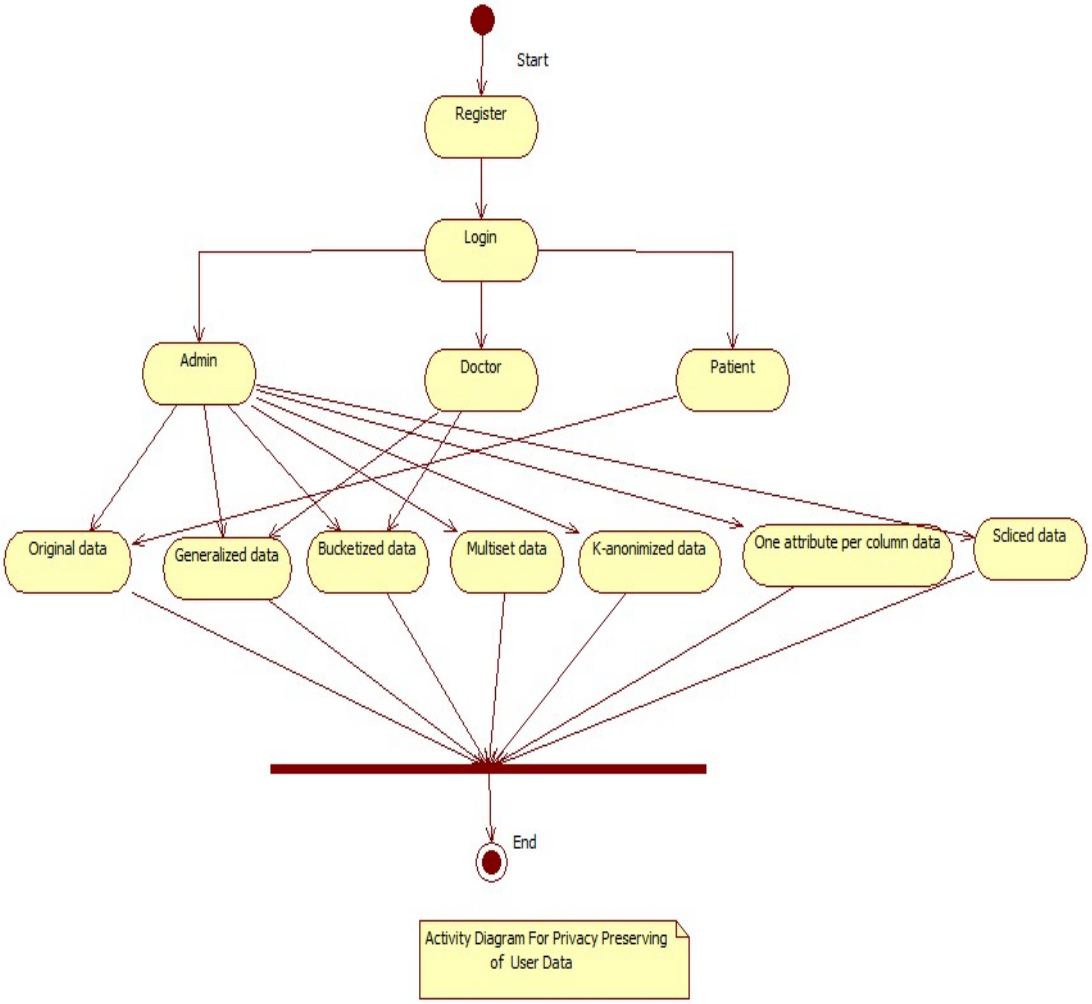

Figure 4.7: Activity Diagram

4.4.5 Component Diagram

A UML component diagram 4.8 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.

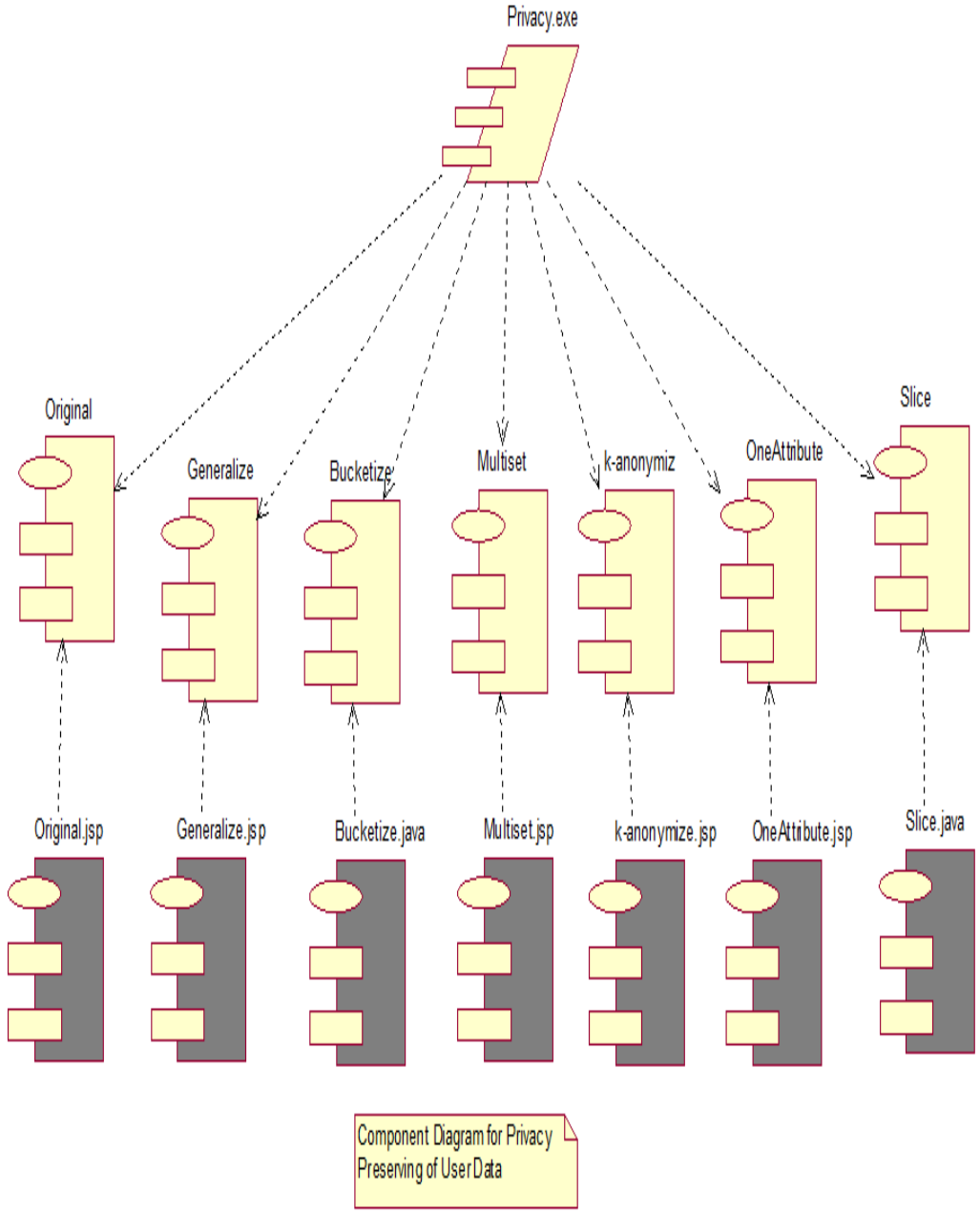

Figure 4.8: Component Diagram

4.4.6 Deployment Diagram

A UML deployment diagram 4.9 focuses on the structure of a software system and is useful for showing the physical distribution of a software system among hardware platforms and execution environments.

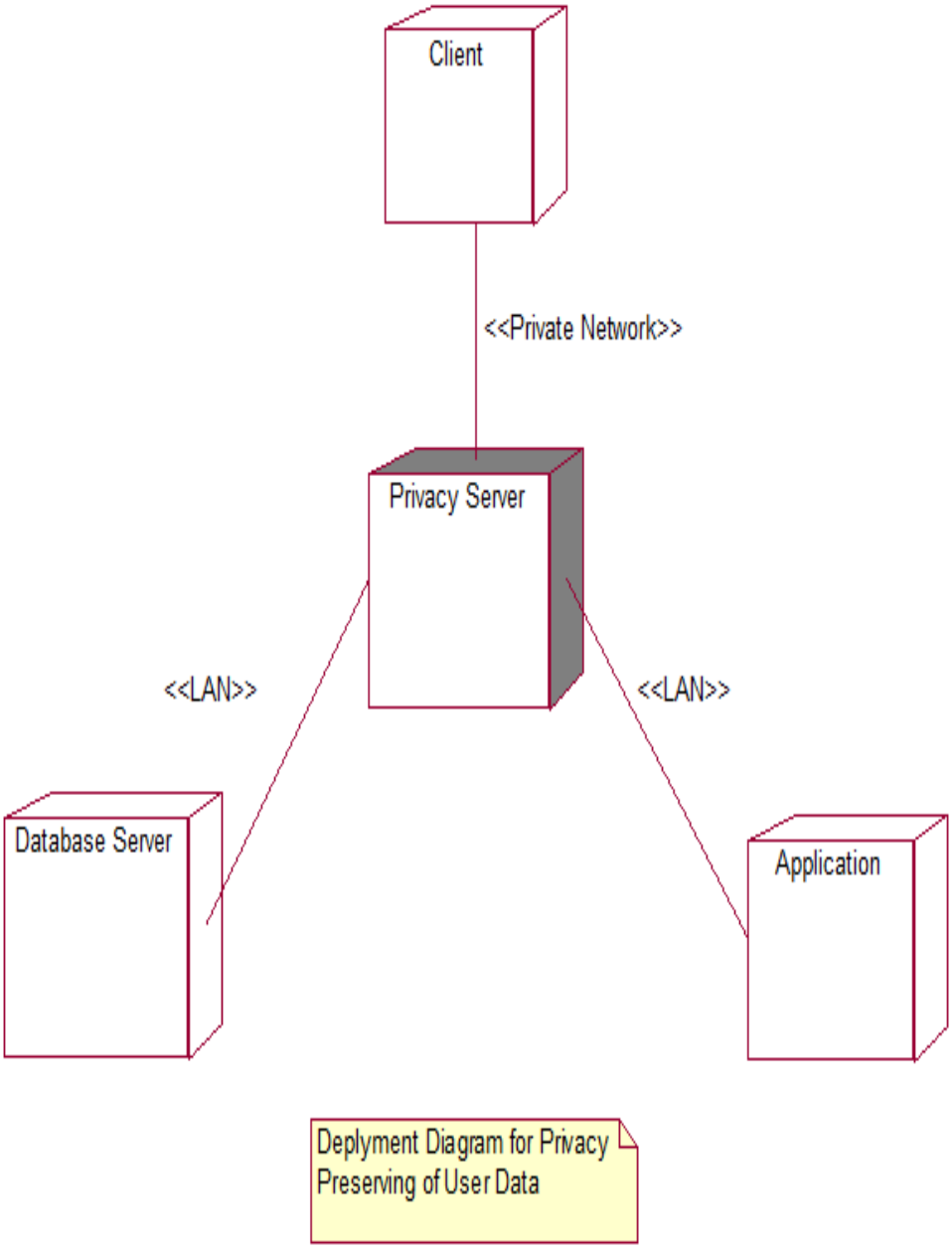

Figure 4.9: Deployment Diagram

4.4.7 Collaboration Diagram

The UML collaboration diagram (called a communication diagram) 4.10 provides another indication of the temporal order of the communications but emphasizes the relationships among the objects and classes instead of the temporal order.

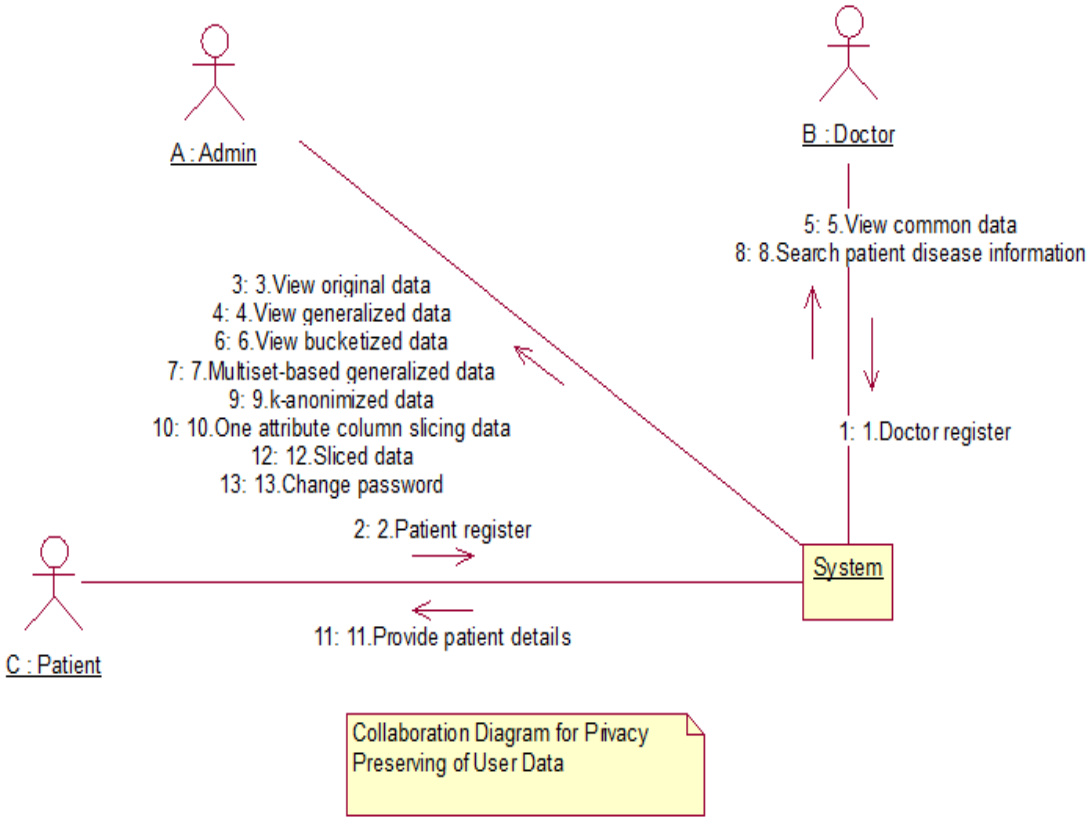

Figure 4.10: Collaboration Diagram

4.5 Summary

In this chapter, section 4.1 discuss system architecture, E-R diagram explained in section 4.2, section 4.3 discuss data flow diagram, UMLs discuss in section 4.4. In the next chapter we will discuss implementation of our project.

Bibliography

[1] Lei Xu and Chunxiao Jiang, Yan Chen, Jian Wang and Yong Ren," A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining". Feb.2016

[2] V. Shyamala Susan and T. Christopher, " Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes". 2016.

[3] Abid Mehmood, Iynkaran Natgunanathan, Yong Xiang, Guang Hua, and Song Guo, "Protection of Big Data Privacy", IEEE Access 2016.

[4] Preet Chandan Kaur, Tushar Ghorpade, Vanita Mane, "Analysis of Data Security by using Anonymization Techniques". IEEE, 2016.