A

Project-I Report

on

A NOVEL METHOD FOR LINKING EXISTING HEALTH-RELATED DATA AND MAINTAINING PARTICIPANT CONFIDENTIALITY

Submitted in Partial Fulfillment of
the Requirements for the Degree
of

Bachelor of Engineering

in

Computer Engineering

to

North Maharashtra University, Jalgaon

Submitted by

Mayuri D. Patil
Vrushali S. Patil
Swati S. Patil
Rupali S. Baviskar

Under the Guidance of

Mr. Manoj E. 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 *A Novel method for linking existing health-related data and maintaining participant confidentiality*, submitted by

Mayuri D. Patil
Vrushali S. Patil
Swati S. Patil
Rupali S. Baviskar

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 8, 2016
Place: Jalgaon

Mr. Manoj E. Patil
Guide

Prof. Dr. Girish K. Patnaik
Head

Prof. Dr. K. S. Wani
Principal

Acknowledgements

Apart from the my efforts the success of any work depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this special study work. I would like to express my heartfelt gratitude towards our principal Prof. Dr. K. S. Wani, our Guide Mr. Manoj E. Patil and Head of Computer Department Prof. Dr. Girish K. Patnaik for his support and valuable guidance which resulted in the successful completion of this report. I would also like to express my sincere gratitude for his valuable guidance and encouragement during the work. I would like to take opportunity to sincerely thanks to all the concern individuals, family members, friends, who made my case study success.

Mayuri D. Patil
Vrushali S. Patil
Swati S. Patil
Rupali S. Baviskar

Contents

List of Figures

Abstract

Develop a Secure System for a record linking of existing individual health care data and maintaining participant confidentiality. Reuse of individual health-related data faces several problems: Either a unique personal identifier, like social security number, is not available or non-unique person identifiable information, like names, are privacy protected and cannot be accessed. A solution to protect privacy in probabilistic record linkages is to encrypt these sensitive information. To overcome these challenges,develope the Privacy Preserving Probabilistic Record Linkage (P3RL) method. The Privacy Preserving Probabilistic Record Linkage method apply a three-party protocol, with two sites collecting individual data and an independent trusted linkage center as the third partner. The proposed method consists of three main steps: pre-processing, encryption and probabilistic record linkage. Privacy Preserving Record Linkage expands record linkage facilities in setting where a unique identifier is unavailable and/or regulations restrict access to the non-unique person identifiable information needed to link existing health-related data sets. Automated pre-processing and encryption fully protect sensitive information ensuring participant confidentiality. This method increase the security and confidentiality of data.

Chapter 1

Introduction

Medical databases of people usually contain identifiers like surnames, given names, date of birth, and address information. The problem of finding records that represent the same individual in separate databases without revealing the identity of the individuals is called "privacy-preserving record linkage". Initially, the obvious solution for privacy-preserving record linkage seems to be the encryption of the identifiers with a standard cryptographic procedure. The aim is to describe a new method for the calculation of the similarity between two encrypted strings for use in probabilistic record linkage procedures [3].

Section 1.1 describes the background of the privacy preserving probablistic record linkage.Motivation is discussed in section 1.2.Section 1.3 describes the problem definition.The scope of privacy preserving probablistic record linkage is discussed in Section 1.4.Section 1.5 describes the objectives of the system. The overall organization of the system is described in Section 1.6.

1.1 Background

Record linkage of existing individual health care data is an efficient way to answer important epidemiological research questions. Reuse of individual health-related data faces several problems: Either a unique personal identifier, like social security number, is not available or non-unique person identifiable information, like names, are privacy protected and cannot be accessed. A solution to protect privacy in probabilistic record linkages is to encrypt these sensitive information. Unfortunately, encrypted hash codes of two names differ completely if the plain names differ only by a single character. Therefore, standard encryption methods cannot be applied. To overcome these challenges, the Privacy Preserving Probabilistic Record Linkage (P3RL) method is develop[4].

1.2 Motivation

The existing problem is important to solve in order to get stronger privacy protection and security for data sending. The existing system involves the incomplete privacy protection. This may result in loss of data due to incomplete privacy protection. Hence Strong security is provided through standard Encryption Algorithm (i.e. AES) and by maintaning a Lock Table. The proposed system will help to provide the complete security.

1.3 Problem Definition

The Privacy Preserving Probabilistic Record Linkage method apply a three-party protocol, with two sites collecting individual data and an independent trusted linkage center as the third partner. The proposed method consists of three main steps: pre-processing, encryption and probabilistic record linkage. Privacy Preserving Record Linkage expands record linkage facilities in setting where a unique identifier is unavailable and/or regulations restrict access to the non-unique person identifiable information needed to link existing health-related data sets. To guarantee similar quality and format of variables and identical encryption procedure at each site, the linkage center generates semi-automated pre-processing and encryption templates. To retrieve information for the creation of templates without ever accessing plain person identifiable information, introduce a novel method of data masking. Sensitive string variables are encrypted using Advanced Encryption Algorithm(AES), which enables calculation of similarity coefficients. For date variables, develope special encryption procedures to handle the most common date errors. The linkage center performs probabilistic record linkage with encrypted person identifiable information and plain non-sensitive variables. This method increade the security and confidentiality of the data.

1.4 Scope

Privacy preserving probablistic record linkage provide security related to health-related data and maintaining participant confidentiality. There are many related problems such as insecurity, loss of data. Exact sense of security of the proposed system is up to some limit which are disambiguated in the project. Furthermore, improvements can be implemented in order to increase the accuracy of the Privacy protection[3]. Furthermore, improvements can be implemented in order to enhance the performance of system and secure the data effectively.

1.5 Objectives

The main objective of the proposed system involves stronger privacy protection. P3RL can well protect user privacy against both inside and outside attackers. It Provide an understanding of record linkage applications, challenges, and techniques. Understand the record linkage process, and key techniques employed in each step of this process. Appreciate the privacy and confidentiality challenges that record linkage poses[5].

1.6 Organization of project report

Chapter 1, titled Introduction,presents introduction about a efficient method for linking existing health-related data and maintaining participant confidentiality. The simulation results show stronger privacy protection than existing schemes.

Chapter 2, titled System Analysis, presents the development of proposed system, Feasibility Study, Risk Management, Effort Allocation, project scheduling.

Chapter 3, titled System Requirements, presents the software requirements, functional requirements and non-functional requirements.

Chapter 4, titled System Design, presents the E-R diagram, Data flow diagrams, Interface design and UML Diagrams.

Chapter 5, titled Conclusion, presents Concludes this dissertation and provides direction for further work in this area.

1.7 Summary

The chapter focuses on the basic introduction of the existing and proposed system. The next chapter will focus on the system analysis of the proposed system.

Chapter 2

System Analysis

The main purpose behind the development of proposed system is to overcome the drawbacks of existing system. Hence, easy and safe access is provided.Strong security is provided through standard encryption algorithm (i.e. AES) and by maintaning a lock table.

Section 2.1 describes the Literature Survey. Proposed system is describe in section 2.2. Section 2.3 describes feasibility study of the system. The risk analysis of the project includes in section 2.4. Section 2.5 describes project scheduling. The effort allocation study is include in the section 2.6.

2.1 Literature Survey

Record linkage of existing individual health care data is an efficient way to answer important epidemiological research questions. Reuse of individual health-related data faces several problems: Either a unique personal identifier, like social security number, is not available or non-unique person identifiable information, like names, are privacy protected and cannot be accessed. A solution to protect privacy in probabilistic record linkages is to encrypt these sensitive information by using the advanced encryption algorithm. Unfortunately, encrypted hash codes of two names differ completely if the plain names differ only by a single character. Therefore, standard encryption methods cannot be applied. Hence to overcome these challenges, develop the Privacy Preserving Probabilistic Record Linkage (P3RL) method[4].

The Privacy Preserving Probabilistic Record Linkage method use a three-party protocol, with two sites collecting individual data and an independent trusted linkage center as the third partner. Our method consists of three main steps: pre-processing, encryption and probabilistic record linkage. Data pre-processing and encryption are done at the sites by local personnel. To guarantee similar quality and format of variables and identical encryption procedure at each site, the linkage center generates semi-automated pre-processing and encryption templates[5].

2.2 Proposed System

In Privacy Preserving Probabilistic Record Linkage method apply a three-party protocol, with two sites collecting individual data and an independent trusted linkage center as the third partner. The method consists of three main steps: pre-processing, encryption and probabilistic record linkage. Data pre-processing and encryption are done at the sites by local personnel. To guarantee similar quality and format of variables and identical encryption procedure at each site, the linkage center generates semi-automated pre-processing and encryption templates. To retrieve information (i.e. data structure) for the creation of templates without ever accessing plain person identifiable information, we introduced a novel method of data masking. Sensitive string variables are encrypted using Advanced Encryption Algorithm, which enables calculation of similarity coefficients. For date variables, develope special encryption procedures to handle the most common date errors. The linkage center performs probabilistic record linkage with encrypted person identifiable information and plain non-sensitive variables[10].

2.3 Feasibility Study

The proposed system is built in order to more privacy protection. Three key considerations are involved in the feasibility study Economic, Operational, and Technical.

2.3.1 Economic Feasibility

The project involves the utilization of softwares like netbeans. The proposed system is economically feasible which will help improve the privacy of health-related data and maintaining participant confidentiality.

2.3.2 Operational Feasibility

Operational feasibility determines the proposed system will be beneficial to provide the security for health-related data. The system is user friendly which involves easy steps to providing the privacy[6]. The Privacy of health related data with AES Encryption techniques can be justified if the proposed system satisfies the user objectives and can be fitted in to current system operation. This system can be justified as operationally feasible based on the following :

- The methods of processing and presentation are completely acceptable by the users because they meet all their requirements.

- The methods of processing and presentation are completely acceptable by the users because they meet all their requirements.
- The system will certainly satisfy the user objectives and it will also enhance their capability.
- The system will certainly satisfy the user objectives and it will also enhance their efectively.

The proposed system will be beneficial to provide the security for health-related data. The system is user friendly which involves easy steps to providing the privacy[6].

2.3.3 Technical Feasibility

The project involves Advanced Encryption Algorithm to provide encryption and decryption. Hence further utilization of the code can be done to enhance the performance of the system. This test includes a study of function, performance and constraints that may affect the ability to achieve an acceptable system. This test begins with an assessment of the technical viability of the proposed system.

2.4 Risk Analysis

Risk Analysis and management are a series of steps that help a software team to understand and manage uncertainty. As developing Unit converter, if input is given which is out of range, the result may be wrong. Exponential conversion may also sometimes go wrong.

▪ *Introduction of Risk Analysis*

The goal of risk assessment is to prioritize the risks so that attention and resources can be focused on the more risky items. Risk identification is the last step in risk assessment, which identifers all the different risks for a particular project[6]. The problems or risks that commonly faced are listed below:

A Estimation and Scheduling The unique nature of individual software projects creates problems for developers in estimating and scheduling development time. By referring existing project experience to overcome this problem.

Sudden growth in requirements There can be a sudden growth in resources that not thought earlier while project planning. This sudden growth can also lead in being late for project completion.

Breakdown of specification At the initial stage of integration or coding, requirements and specifications are incomplete or insufficient. As coding got progressed, requirement of specification was fulfilled. These risks are project-dependent and identifying them is an exercise in envisioning what can go wrong. Methods that can aid risk identification include checklists of possible risks, surveys, meetings and brainstorming, and reviews of plans, processes, and work products.

Components of Risk Analysis

Everyone involved in the software process managers, software engineers, and customers participate in risk analysis and management.

Needs 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. It’s for a reason which being prepared understanding the risks and taking proactive measures to avoid or manage them-is a key element of good software project management[7].

Software Risk

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

- Uncertainty: The risk may or may not happen; which is, there are no 100 percent probable risks.
- Loss: If the risk becomes a reality, unwanted consequences or loss will occur. When risks are analyzed, it is important to quantify the level of uncertainty and the degree of loss associated with each risk[8]. To accomplish this, different categories of risks are considered.

Project Risks

Threaten the project plan. Which is, if project risks become real, it is likely that project schedule will slip and the 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. In the project, project risk occurs if requirement of technical member means technical team is unavailable according to the project plan and estimation and if the project is not completed within time then situation project risk can occurs[8].

- *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 thought it would be. In the project if any module of resume builder is not worked properly according to developer expectation then technical risk may occur.

2.5 Project Scheduling

In project management, a schedule is a listing of a project's milestones, activities, and deliverables, usually with intended start and finish dates. Those items are often estimated in terms of resource allocation, budget and duration, linked by dependencies and scheduled events. A schedule is commonly used in project planning and project portfolio management parts of project management. Elements on a schedule may be closely related to the work breakdown structure (WBS) terminal elements, the Statement of work, or a Contract Data Requirements List[9].

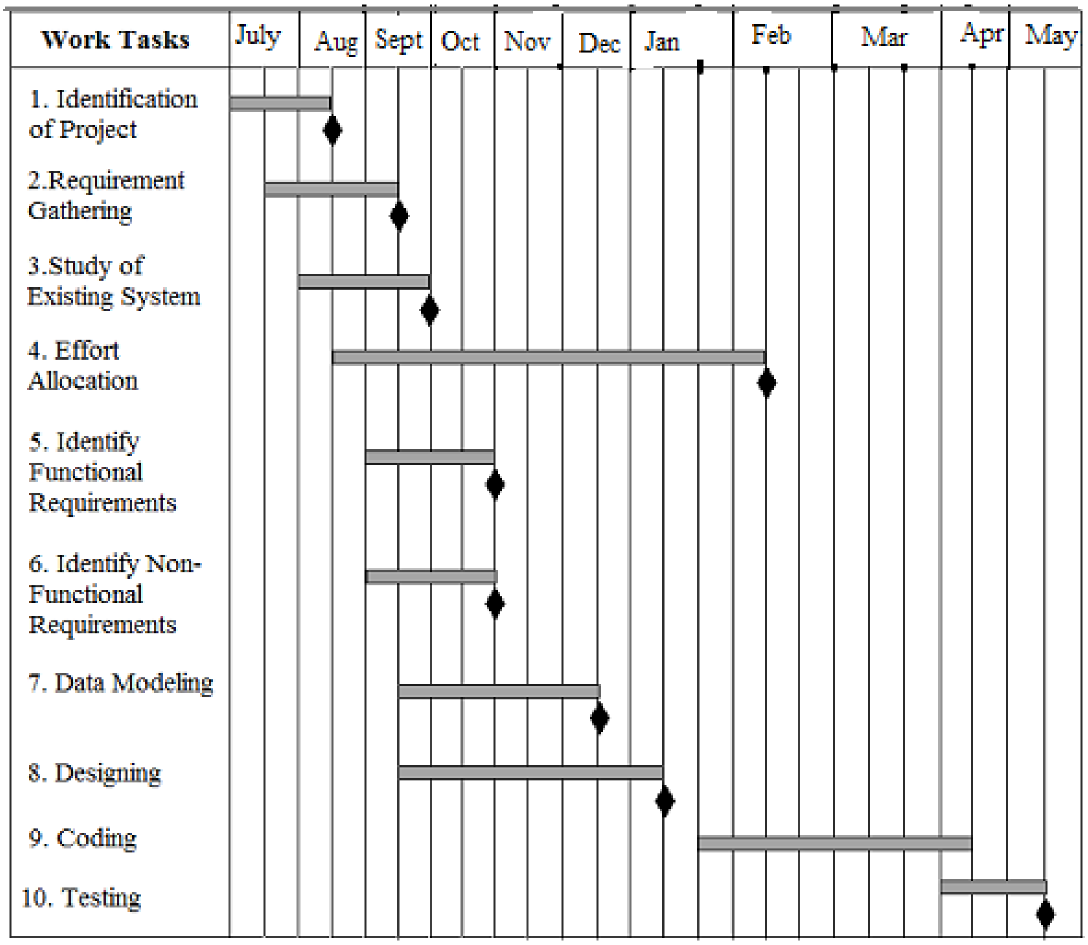

Figure 2.1: Project Scheduling

2.6 Effort Allocation

A criterion was presented to define the most eficient strategy for the exploration and maintenance of plant genetic resources. All of the three factors composing the efficiency. i.e., multiplicity of target populations. The amount of expenses, and goodness for individual populations of the conservation manipulation adopted, were incorporated in the present criterion. Sample size per target population for field collection was investigated on the basis of this criterion, leading to the conclusion that the number of visited populations rather than sample size per population determines the overall efficiency of a collection project as a whole. Without any particular reason, intensive sampling for a limited number of populations is not logical. A sample size as small as ten plants per site or population was estimated reasonable to cover a large target area[10].

Table 2.1: Effort Allocation

Work Tasks	Mayuri	Rupali	Swati	Vrushali	Work in %
Identification of Project	Yes	Yes	Yes	Yes	10
Requirement Gathering	Yes			Yes	10
Study of Existing System		Yes		Yes	10
Identify Requirements		Yes	Yes	Yes	5
Data Modeling					
Designing					
Coding					
Testing					

2.7 Summary

In this chapter the system analysis discuss in detail. In next chapter discuss about the system requirements specification.

Chapter 3

System Requirements Specification

The chapter focuses on the various requirements of the system includes its software requirements, functional requirements and non-functional requirements.

The Hardware Requirements are include in section 3.1.Section 3.2 describes the software requirements of the system. The functional requirements of the system are discussed in Section 3.3.Section 3.4 describes the non-functional requirements of the system.

3.1 Hardware Requirements

The various hardware requirements of the system can be summarized here:

- Windows-compatible with LAN Network
- Processor : Intel (R) Core(TM) i3 CPU
- Installed memory (RAM): 2GB
- System Type: 64-bit/32-bit operating system.
- Front end : Java
- Back end : Database

3.2 Software Requirements

The various software requirements of the system can be summarized here [**?**]:

- Operating system: Windows 7/8
- Jdk1.6
- MySQL

- JCreator
- Apache Tomcat Server
- Adobe Dreamweaver

3.3 Functional Requirements

Requirement Analysis is dependent on three aspects (Data, Function and Be- haviour. Requirement Analysis of data is a process of inspecting, cleaning, transforming, and modelling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encom- passing diverse techniques under a variety of names, in different business, science, and social science domains. Requirement Analysis of function is providing services to user as they expect in the sense of Java Integrated Environment. Function Analysis is one of important aspect of any project to determine project efficiency, integrity, user friendly etc[1].

3.4 Non-Functional Requirements

In Non-functional Requirements of this project implements those functions which does not effect on function and behaviour of project for desired goal and objective of project. Non-functional Requirement just provides user friendliness and notifications that are not most necessary for this project.

3.5 Summary

The chapter includes the various requirements of the system. The next chapter discusses about the design of the proposed system.

Chapter 4

System Design

The chapter focuses on the design of the system. Flowchart, how the data flows from the system architecture, and all the UML diagrams including class diagram, use case diagram, sequence diagram, activity data diagram, component diagram, deployment diagram of the project[1][2].

Section 4.1 describes System Architecture of P3RL.The E-R digram describe in Section 4.2.Section 4.3 describes data flow diagrams.The UML diagrams are discussed in Section 4.4.

4.1 System Architecture

The flow of data between sites and the sites responsible for the individual steps included in our P3RL method. P3RL consists of three main steps: preprocessing,encryption and probabilistic record linkage. Data pre-processing and encryption are done at the data custodian sites (site A and B) by authorized local personnel. Creating pre-processing and encryption templates,encryption validation and probabilistic record linkage are done at the linkage site (site C). The system architectureis shown in Fig.4.1.

- Masking
  Masking is used to disclose the individual site data structures to site C without revealing PII. The masked data are used to create the site-specific pre-processing templates. Data for building pre-processing templates are exported to site C as masked alone or masked and additionally shuffled depending on site restrictions.

- Pre-processing
  Pre-processing is a crucial step in record linkage. The aim of pre-processing is to harmonize the linkage variables at each site to make them directly comparable and thus easier to link. Pre-processing includes three steps: masking (site A and B), creating pre-processing templates (site C) and data cleaning (site A and B).Using masking, the

linkage center creates custom pre-processing templatesbased on the data structure at each site. The templates are supplied to the individual sites allowing them to perform standardized data cleaning procedures that result in linkage variables with similar data quality and harmonized formats.The aim of masking is to alter the plain text variables so they are no longer readable. Masking replaces numeric characters between 1 and 9 with 9, lower case alpha characters a to z with z and upper case A to Z with Z. Some characters are left untouched. For example, first characters of fields, numeric character zero, special or language specific characters (e.g., -()) and spaces are unchanged. Masking of linkage variables is performed at the individual sites based on a pre-determined sample number of records or the entire population (depending on project-specific restrictions). Masking informs the data cleaning procedures by hinting at data errors, like numbers in name fields, characters in a numeric field or special codes for missing data (e.g., 9, 99, , .) and reveals language of text, number of names (surnames, first names) in a single variable, separators, special characters (e.g. language specific) and date formats.

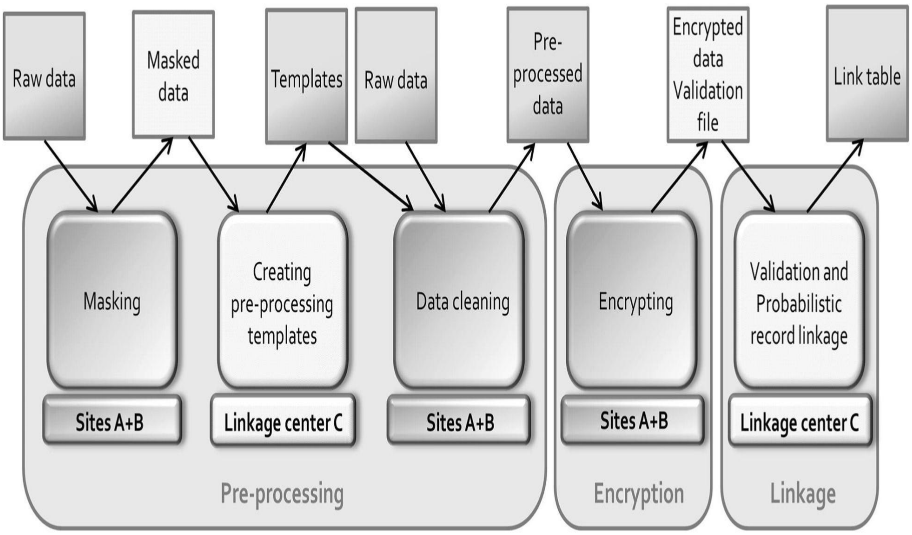

Figure 4.1: System Architecture

- Pre-processing Templates

  The aim of using templates for pre-processing is to guarantee similar quality and format of linkage variables after numbers instead of names for month, types of delimiters, and leading zero for numeric month and day. Although not particular to P3RL, checking the expected order of numeric day and month is not possible if both are less than 13.

- Data Cleaning

  Data cleaning is required because data from independent sources may differ in many aspects. For example, the format of variables may differ or string variables such as names can be inconsistent due to typographical errors, use of nicknames or abbreviations, changes due to marriage or pre- and postfixes. Therefore, the application of consistent data cleaning rules is crucial for any data warehouse generally and for record linkage particularly. In our P3RL workflow data cleaning is based on pre-processing templates and takes place at sites A and B, before encryption. This step is critical as non-pre-processed linkage variables result in a decreased linkage proportion because true matches are more frequently missed [3].

- Encryption

  The aim of encryption is to protect participant privacy and data confidentiality. Encryption is done at the individual sites using an automated encryption tool developed in-house specifically for our P3RL projects. All linkage variables deemed to be confidential (e.g. names, DOB) are encrypted while all other non-sensitive PII (e.g. marital status) are not. In this paper we focus solely on encryption of name and date variables. Nevertheless, the basic method of P3RL is applicable to other variable types. Levels of security and variables to be encrypted will differ from project to project.

- Probabilistic record linkage

  The last step in the P3RL method is probabilistic record linkage. However, before linkage begins site C must verify that the encryption was performed uniformly at site A and B. If the validation files from site A and B do not match the encryption must be redone before linkage can begin.

4.2 E-R Diagrams

Fig 4.2 describes the flow of proposed system in which the sites send the data. After sending data the server receives the data and preprocessing is perform on this data,and then encryption is provide for data and linkage record is take place.

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. Figure shows the E-R diagram for the disambiguation System.

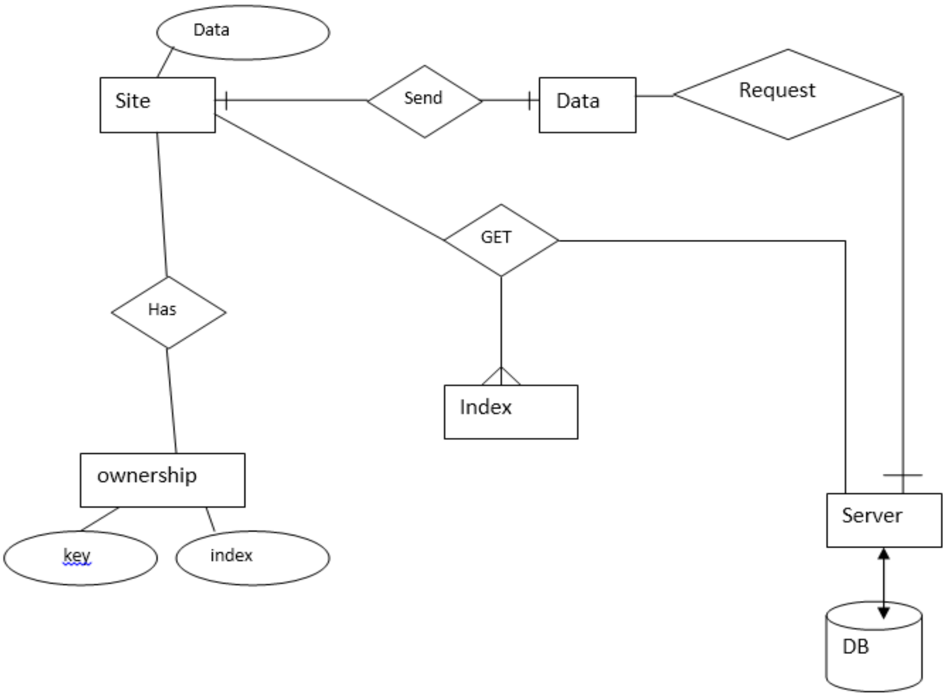

Figure 4.2: Entity Relationship Diagram

4.3 Data Flow Diagrams

A DFD is a graphical technique that depicts the information flow and the transformation that 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 describes 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. It makes whole procedure like a good document and makes simpler and easy to understand for both programmers and non-programmers by dividing into the sub process. The data flow diagram serves two purposes:

- To provide an indication of how data are transform as the moves through the system.

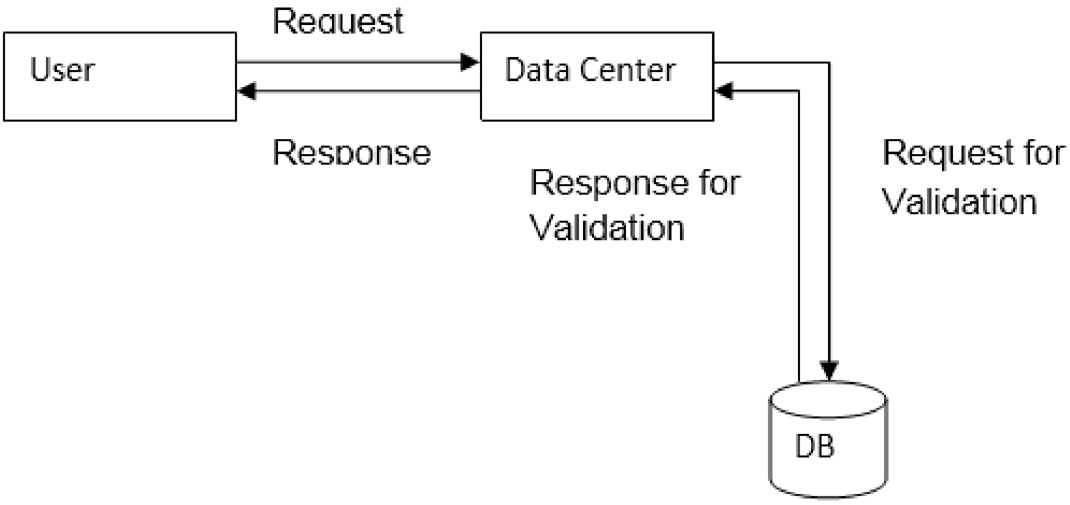

Figure 4.3: DFD Level 0 for Disambiguation System

- To depict the function that transforms the data flow.

The Figure 4.3 and 4.4 shows the level 0 DFD for disambiguation system in which input is given and after processing output is received.

4.4 UML Diagrams

The UML is a language for[2]:

- Visualizing- The structures which are transient can be represented using the UML.
- Specifying- The UML addresses the specification of all the important analysis, design and implementation decisions that must be made in developing and deploying a software- intensive system.
- Constructing- The UML is not a visual programming language, but its models can be directly connected to a variety of programming languages.
- Documenting- The UML addresses the documentation of a system's architecture and all of its details.

The various stuctural and behaivioural diagrams are discussed in the chapter.

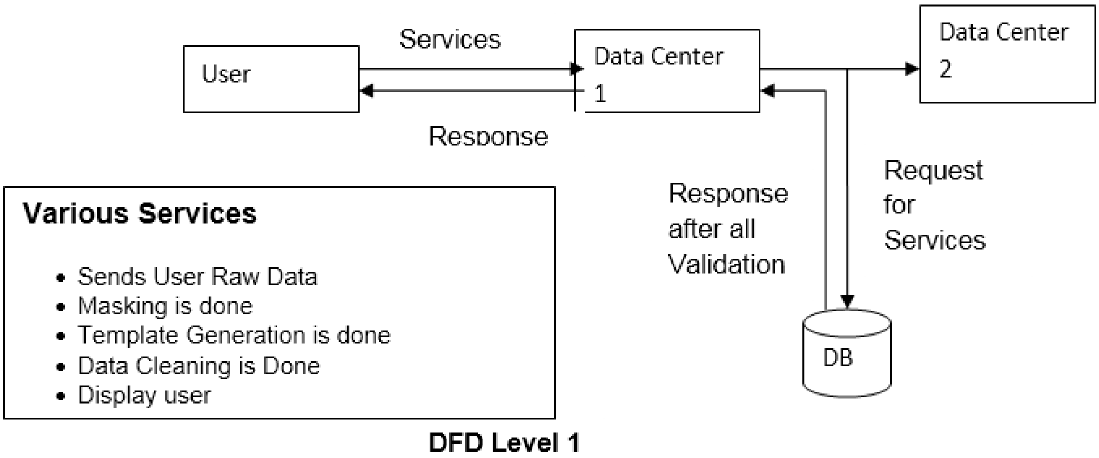

Figure 4.4: DFD Level 1 for Disambiguation System

4.4.1 Use Case Diagrams

Use case diagram shows a set of use cases and actors and their relationships. Use case diagrams address the static use case view of a system. These diagrams are especially important in organizing and modeling the behaviors of a system. The Use Case diagram of the proposed system is shown in Figure 4.5.

4.4.2 Class Diagram

A Class diagram shows a set of classes, interfaces and collaborations and their relation- ships. These diagrams are the most common diagram found in modeling object-oriented sys- tems. Class diagram address the static design view of a system. Figure 4.6 shows the class diagram for the proposed system.

4.4.3 Interaction Diagrams

Both sequence and collaboration diagrams are kinds of interaction diagrams. An interaction diagram shows an interaction, consisting of a set of objects and their relationships. They address the dynamic view of a system. Figures 4.7 and 4.8 are the sequence and collaboration diagrams of the system respectively.

- A sequence diagram is an interaction diagram that emphasizes the time-ordering of messages.
- A collaboration diagram is an interaction diagram that emphasizes the structural organization of the objects that send and receive massages.

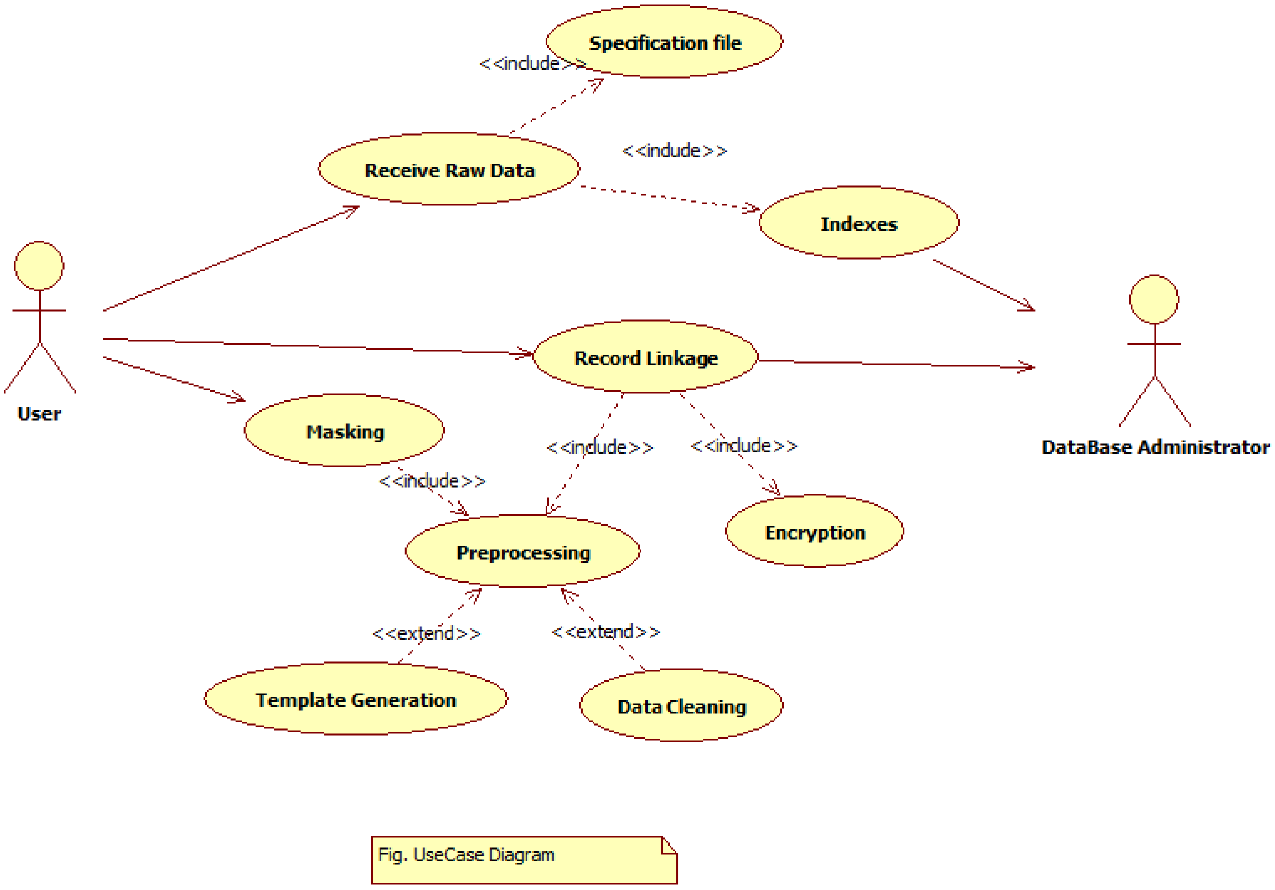

Figure 4.5: Usecase Diagram

Sequence diagram and collaboration diagrams are isomorphic i.e one can be transformed into other. A sequence diagram for disambiguation which shows the sequential process flow. It contain system and user actor. Collaboration diagram for disambiguation system are shown in Fig. 4.7 and 4.8.

4.4.4 Activity Diagram

Activity diagram for disambiguation which provide a way to model various states and one more addition is decision making in states (initial and final states) in which the object is exists. The diagram is used to express dynamic behaviour of system. The diagram shows the behaviour of an object. A condition enclosed in square box is called as "Guard" condition. A dark horizontal bar denotes the possibility is called fork. Black dot indicates start point and black dot with circle indicates stop i.e. terminate state. In a activity diagram contains Enter text, Check for ambiguous words, Display Message, Search the Meaning, Generate output in which diagram dynamic flow is mention. It is shown as in Fig. 4.9.

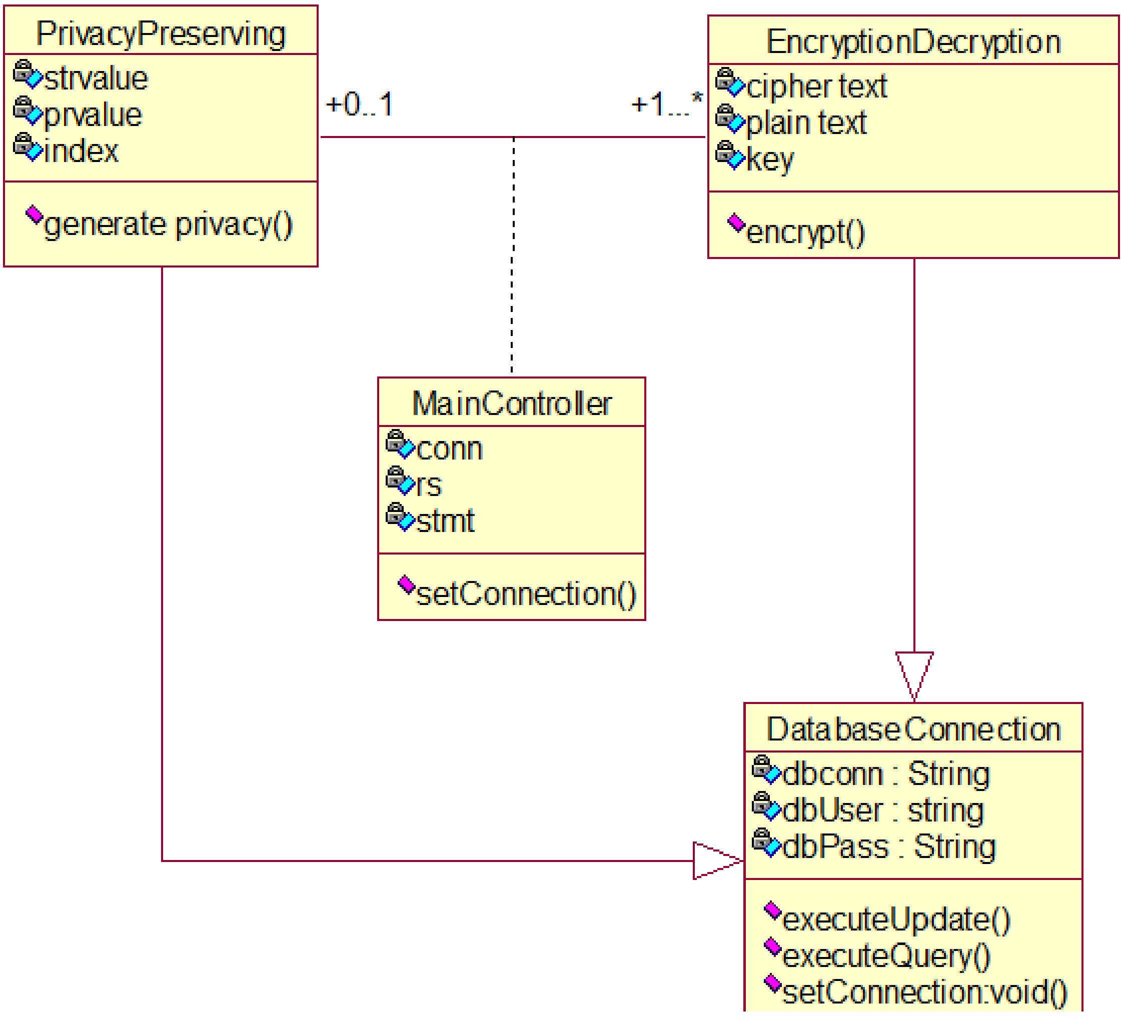

Figure 4.6: Class Diagram

4.4.5 Component Diagram

A component diagram shows the organization and dependencies among a set of components.Component diagrams address the static implementation view of a system. The component diagram for the proposed system is shown is Figure 4.10.

4.4.6 Deployment Diagram

A deployment diagram shows the configuration of run-time processing nodes and the components that live on them. Deployment diagram address the static deployment view of an architecture. Deployment diagram for disambiguation which shows the physical structure of system. By using diagram user can easily understand the physical layout of system. Deployment diagram contain nodes which are database, system, user in which interact with each other at the level of hardware part. Deployment diagram for the proposed system is shown

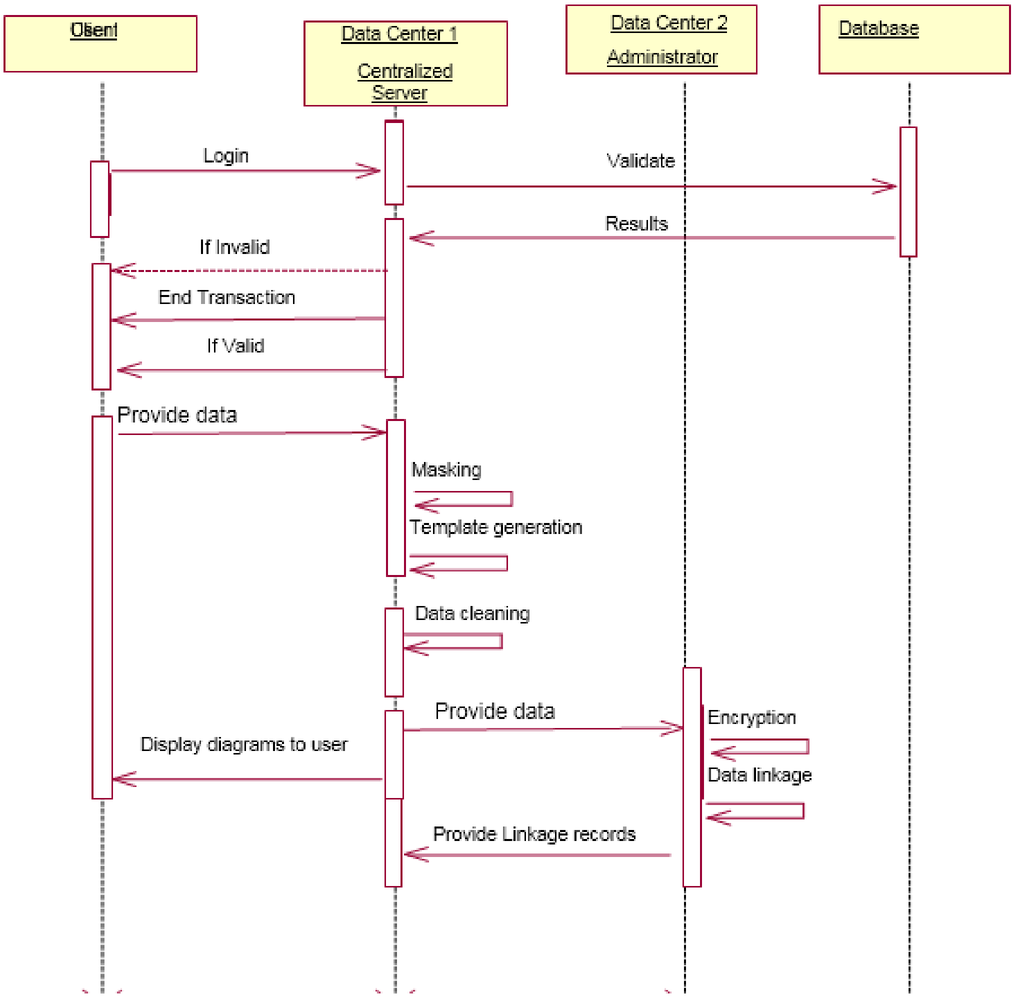

Figure 4.7: Sequence Diagram

in Figure 4.11.

4.4.7 State Transition Diagram

State transition diagram for Disambiguation which provides a way to model various states in which the object is exists. They are used to model more dynamic behaviour of system. The diagram shows the behaviour of an object. A condition enclosed in square box is called as 'Guard' condition. Black dot indicates start point and black dot with circle indicates stop i.e. terminate state. In the state transition diagram Input Text, Check for Ambiguous Words, Search Meaning of Word, Generate Output, takes in a square box because its shows transition from one state to other state in a system as shown in Fig. 4.12.

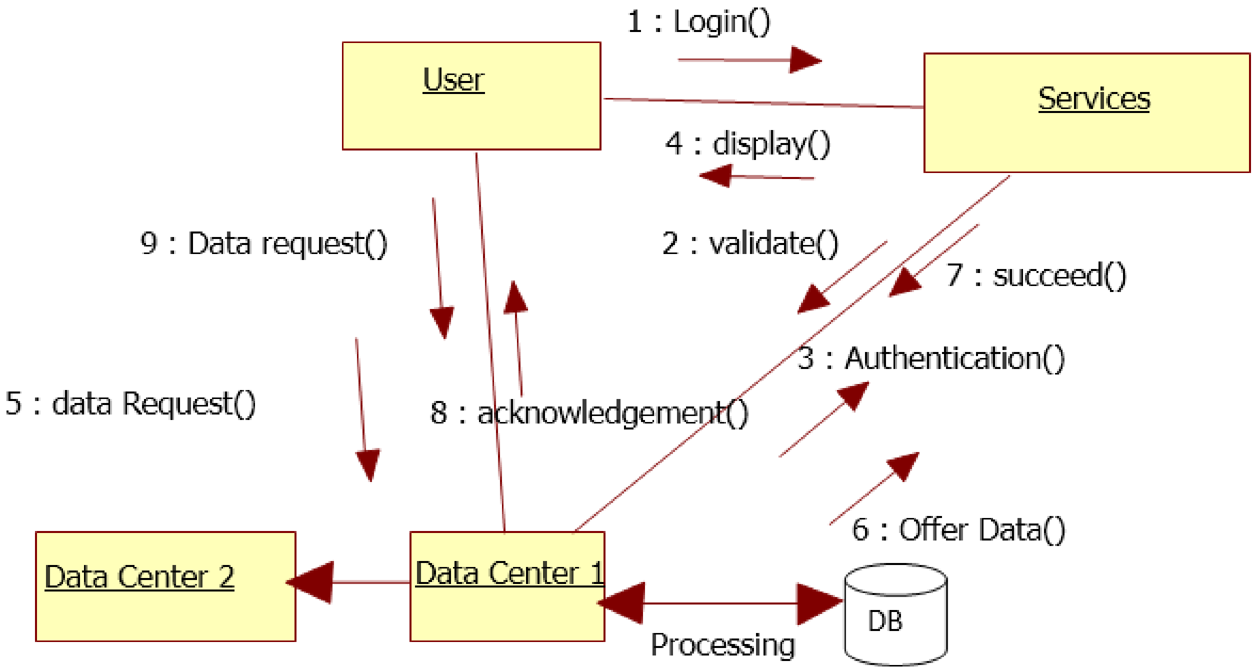

Figure 4.8: Collaboration Diagram

4.5 Summary

The chapter includes the designing of the proposed system including E-R diagrams, database schemas, UML diagrams, data flow diagrams, etc.

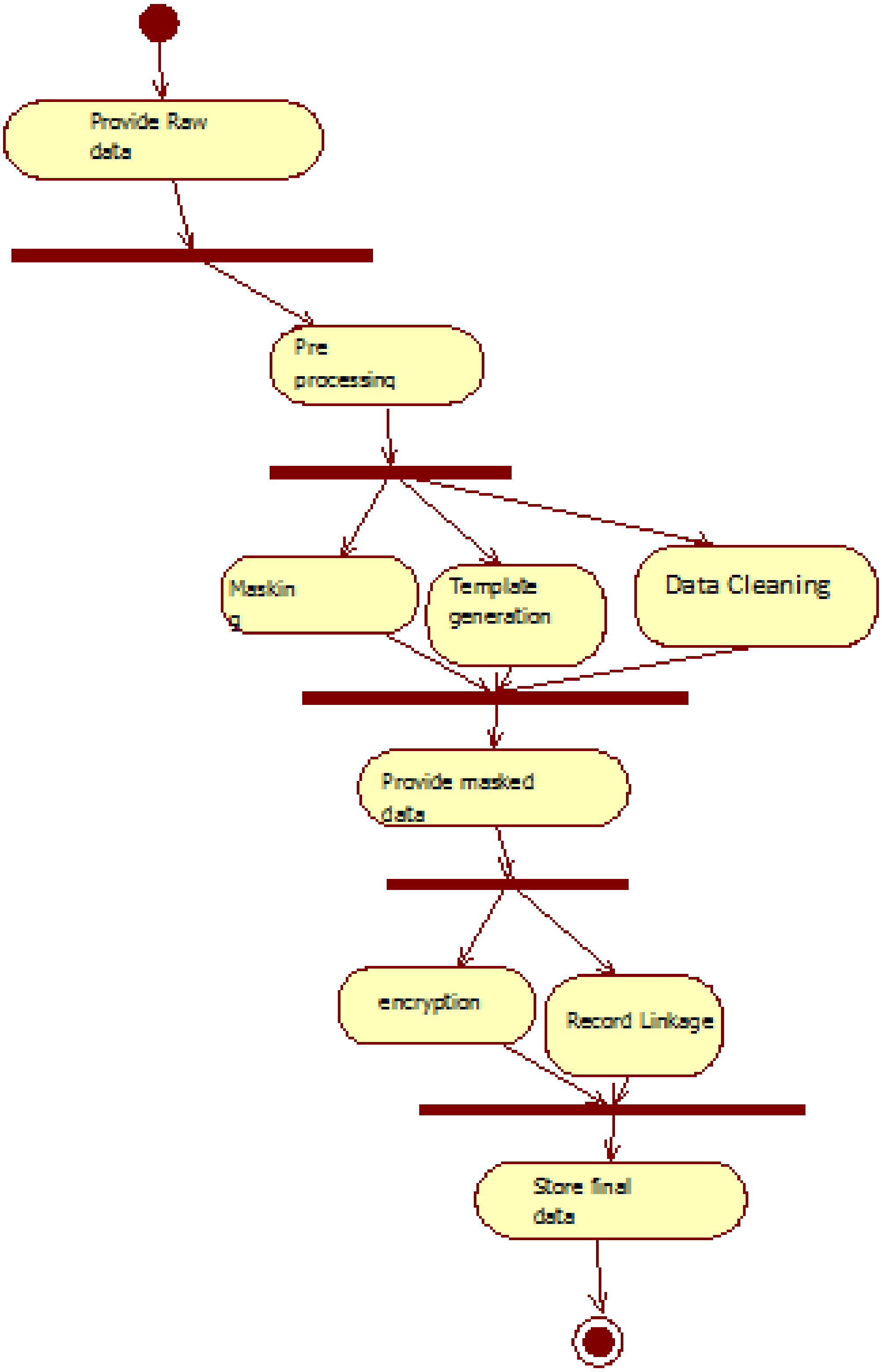

Figure 4.9: Activity Diagram

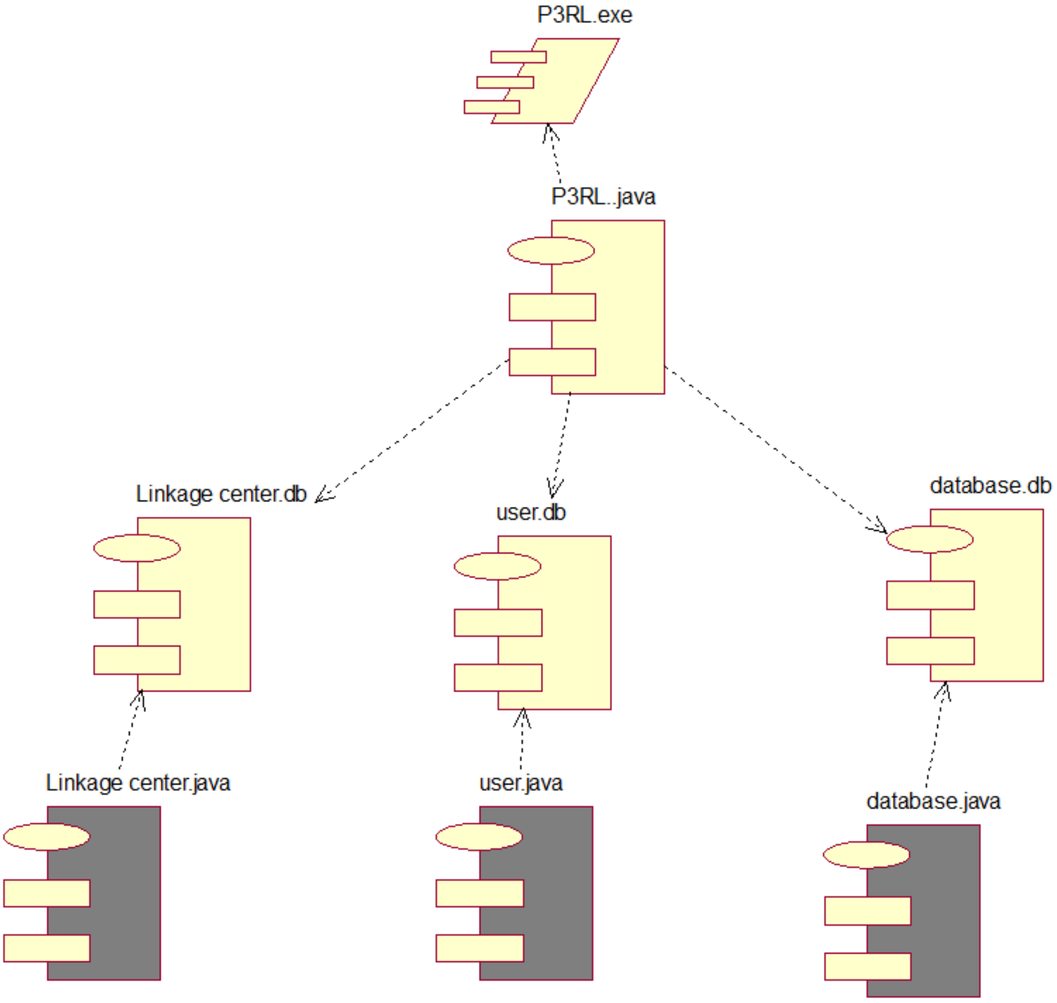

Figure 4.10: Component Diagram

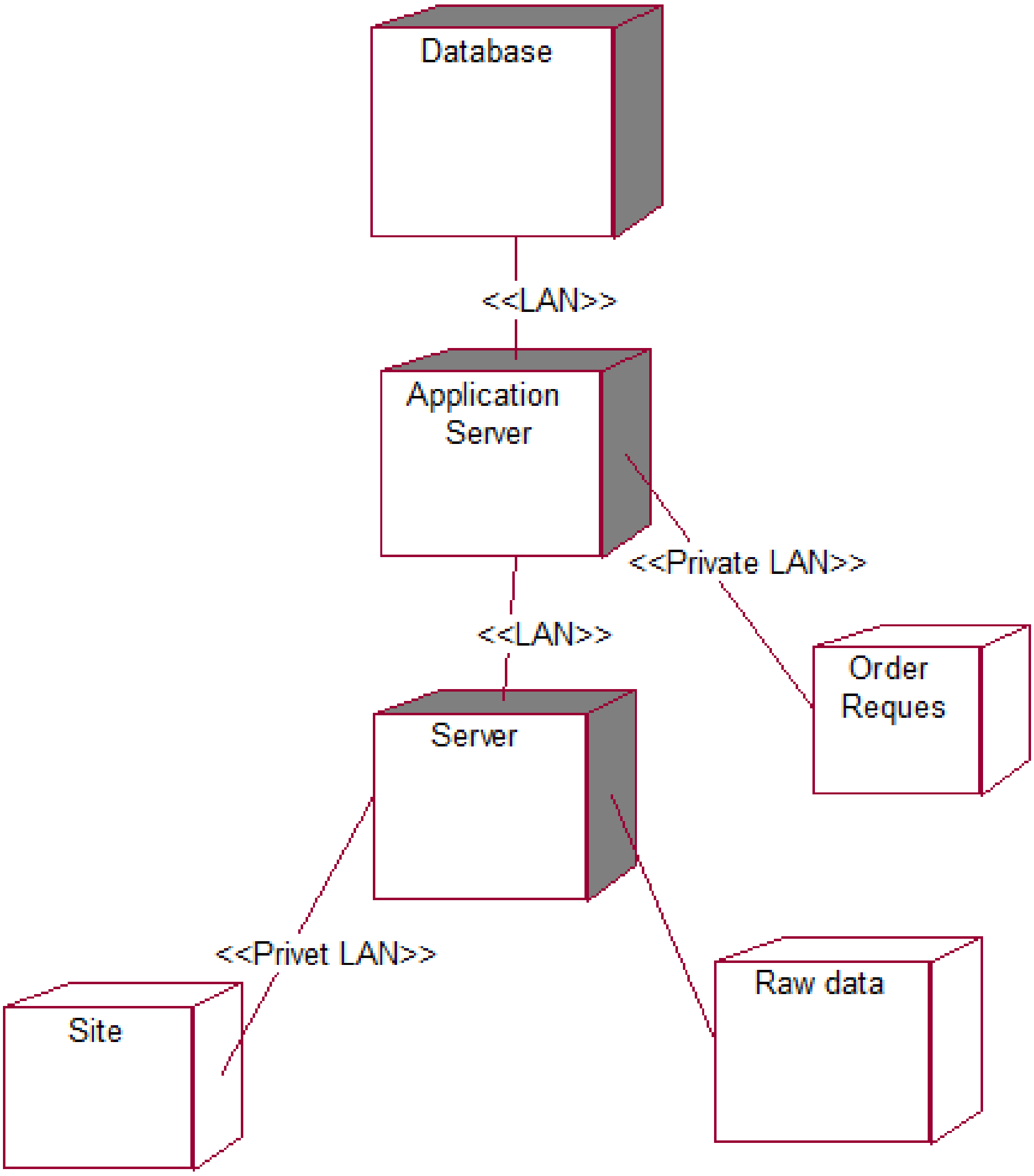

Figure 4.11: Deployment Diagram

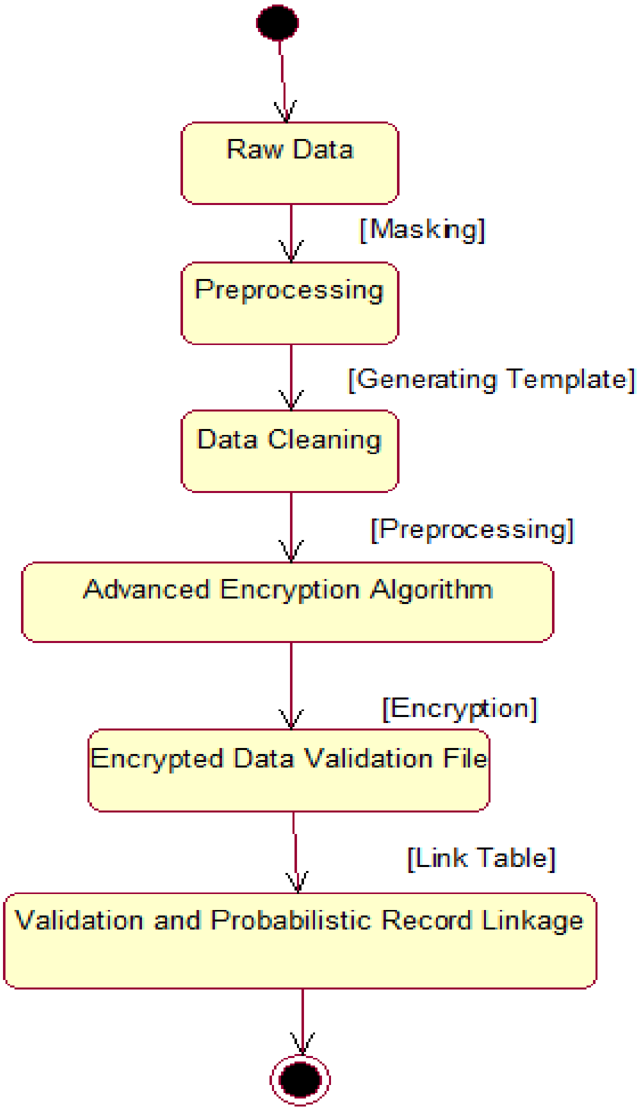

Figure 4.12: State Diagram

Chapter 5

Conclusion

Privacy Preserving Probabilistic Record Linkage facilitates the linkage of existing datasets in healthrelated research settings using automated pre-processing and encrypting to fully protect personal identifying information. Finally the analysis, requirements specifcation and designing part is conclued.

Bibliography

[1] Roger Pressman, Software Engineering: A Practioner's Approach, McGraw-Hill, Seventh Edition, pp.449-490. ISBN-10: 0073375977 ISBN-13:978-007337597, retrieved 5 Aug, 2016.

[2] Grady Booch, James Rumbaugh, Ivar Jacobson, The Unified Modeling Language User Guide, Pearson, Second Edition, 2005, pp. 225-284. ISBN: 978-03-2126-79-79, retrieved 10 Aug, 2016.

[3] Christen P," Privacy-Preserving Data Linkage of Health Related Information: Current Approaches and Research Directions", In: Data Mining Workshops, ICDM Workshops 2013 Sixth IEEE International Conference on: Dec. 2013, p. 497501.

[4] Trepetin S, "Privacy-Preserving String Comparisons in Record Linkage Systems: A Review", Inform Sec J, 2015;17(5/6):253266.

[5] Vatsalan D, "Christen P. An iterative two-party protocol for scalable privacy-preserving record linkage", In: Tenth Australasian Data Mining Conference: 2012; Sydney, 2012, p. 12738.

[6] Randall SM, Ferrante AM, "Privacy-preserving record linkage on large real world datasets", J Biomed Inform. 2014;50:20512.

[7] Vatsalan D, Christen P and Verykios VS, "A taxonomy of privacy-preserving record linkage techniques", Inform Syst, 2013;38(6):94669.

[8] A. Z. Broder, "On the resemblance and containment of documents", Compression and Complexity of Sequences: Proceedings IEEE, (2013), 21-29.

[9] D. Smith and N. Shlomo, "Privacy Preserving Record Linkage, Data Without Boundaries Deliverable D11", Report 2014-01, CMIST Working Paper (2014).

[10] Pang C, Hansen D, "Improved record linkage for encrypted identifiying data", In: Annual Health Informatics Conference: 2012; Sydney, 2012. p. 1648.