A mechanism to achieve accuracy and privacy in Data Mining

By

Tariq Mehmood Khan Niazi

NUST201260760MSEECS60012F

Supervisor

Dr. Omar Arif

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology (MS-IT)

In

School of Electrical Engineering and Computer Science

National University of Sciences and Technology (NUST)

Islamabad, Pakistan.

(August 2016)

Approval

It is certified that the contents and form of the thesis entitled "**A mechanism to achieve accuracy and privacy in Data Mining**" submitted by **Tariq Mehmood Khan Niazi** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Omar Arif**

Signature: _____

Date: _____

Co-Supervisor: **Dr. Zeeshan Shafi**

Signature: _____

Date: _____

Committee Member 1: **Dr. Sharifullah Khan**

Signature: _____

Date: _____

Committee Member 2: **Dr. Shahzad Saleem**

Signature: _____

Date: _____

Dedication

I dedicate this thesis work to my family specially my wife who from the start encouraged, supported and indulged me. Also to my supervisor and co-supervisor who is an unending source of motivation and guidance.

It is also dedicated to my friends and to my teachers with whom I have an exceptional and admirable relationship.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Tariq Mehmood Khan Niazi**

Signature: _____

Acknowledgment

First and foremost, praises be to Allah, the Almighty, on whom ultimately we depend for direction and guidance.

I wish to express my deep sense of gratitude to my supervisor Dr. Omar Arif for giving me a chance to work with him, in a research environment and enhance my skills. Words are inadequate in offering my thanks to the members of advisory committee for their encouragement and valuable input whenever required.

Finally, yet importantly, I would like to thank my friend Kashif Kamran who helped me directly or indirectly whenever support required to accomplish thesis work.

Table of Contents

List of Figures

This page is intentionally left blank

Abstract

Today we are producing abundant data every day and need to extract useful information by applying data mining techniques. This is also known as Knowledge Discovery from Data (KDD). KDD cycle involves four key users including Data Provider, Data Collector, Data Miner and Decision maker. All these users have their own concerns, data privacy is the main concern of data provider, correct data collection is the concern of data collector, accurate results are the concern of data miner and decision maker.

Wide use of data mining techniques leads to data privacy issues and therefore need to address privacy preservation in KDD. Many techniques exist to cater this issue like Perturbation, Condensation, data hiding etc. Each tries to preserve privacy at maximum level but this leads to data loss resulting compromised data mining results. It has been observed that if privacy not compromised then accuracy of results is compromised hence data mining results get affected.

Most critical Quasi-identifier attributes need to identify in the data set as these can help attacker to identify individual and hence lead to privacy leak. Once identified such attributes, need to take appropriate measure and preserve privacy.

We are proposing a solution not only to preserve privacy with minimum accuracy loss which is concern of data provider but also taking care of concerns of other users as well by introducing trusted third party. Trusted third party will take care of the interests of users involved in KDD cycle from data provision to decision making.

Chapter 1: Introduction

This chapter apprises readers about the challenges faced, risks involved and consequences in sharing data with others. It also defines terminologies generally used in data publishing and data mining process.

An overall depiction of the proposed architecture for privacy protection of published data has been given in the later section of this chapter.

1.1 Background

Today we are producing abundant data every day and need to extract useful information by applying data mining techniques. This is also known as Knowledge Discovery from Data (KDD). KDD cycle involves four key users including Data Provider, Data Collector, Data Miner and Decision maker. All these users have their own concerns, data privacy is the main concern of data provider, correct data collection is the concern of data collector, accurate results are the concern of data miner and decision maker.

1.2 Privacy of Shared Data

It is of utmost importance to protect the data from its unintended use. Data collected by the companies usually relates to its customers and they are legally and morally bound to keep the privacy of their customers intact. If we consider the health sector, patients' data pertaining to their medical ailments demands secrecy in terms of their ailments.

One of the guaranteed solutions to safeguard the privacy of data is to restrict access to the datasets to the companies' / hospitals' staff. This approach definitely ensures data safety but at the cost of depriving scientists and researchers to conclude results which are further utilized for betterment of customers/patients ultimately.

1.3 Factors behind Data Sharing

There are multiple factors which compels organizations to share and publish their customers' data. Following factors can be categorized as main reasons behind data sharing:

- **Legal Bondage**
- **Social Surveys**

Legal Bondage

Many companies are bound legally to publish the data of their customers and patients which is utilized by the Regulatory authorities to monitor the QoS (Quality of Service) of companies being offered to their customers.

Social Surveys

Various government, Semi government and NGOs (Non-Government Organizations) conduct surveys to measure certain attributes. These surveys are analyzed by the data miners or data analysts to conclude meaningful information. The data collected in the surveys are usually published and easily accessible to public and decision makers to plan resources or conduct research.

It has been established from the above discussion that it is inevitable to keep the data unshared. Once we are convinced to share or publish data then we came across with the challenges of data privacy. One of the core challenges is to keep individual's identity hidden from others to whom the data is being shared with or published on the internet. Let us see how individual identity is compromised on published data.

1.4 Data Publishing Challenges

Once the data is published for general access, you lose control over the data. Now it is up to the data users or data miners that how they use the data. Genuine scientists and researchers seldom misuse the published data but other misusers try to extract the information for which the published data has not been provided.

Data Misusage

Data misusage can be defined as using published data for the purpose for which the publisher has not published the data.

The above definition can further be clarified by the following examples:

A health survey is being conducted in country and data is published for the donor organization to channelize their efforts for eradication of a specific disease in a community. Let us have a look on a snapshot of the published data in Table 1.1.

Age	Postal Code	Religion	Illness
35	45000	Islam	Diabetic
25	45000	Islam	HIV
20	45000	Islam	Diabetic
23	47000	Christian	Diabetic
67	47000	Hindu	Hypertensive
90	41000	Hindu	TB
80	47000	Christian	Diabetic

62	47000	Christian	Arthritis

Table 1.1

The surveyor purpose to publish the above data was to make the donors attentive towards the eradication of most frequent disease like Diabetes. It was the intended purpose of the data publisher and was well served if the data has been used for the same purpose.

Since data is publically available therefore a miscreant's elements may easily deduce that the area with postal code 47000 is populated with minorities considering the example of our dear homeland. Thus inferring information for which the data has not been published is called data misusage.

In the above example personal identity has not been revealed, now let us consider that ECP (Election Commission of Pakistan) published following data known as voter list:

CNIC	Name	Age	Postal Code
45602-9170630-7	Imran Hameed	35	45000
95608-1706398-8	Shaukat Mehmood	25	45000
54315-1472563-8	Altaf Khan	20	45000
68721-4698217-6	Iqbal Masih	23	47000
47815-1754256-2	Kirshan Laal	67	47000
95142-7951357-8	Rajesh Chaand	90	41000
76542-7514235-3	John Harry	80	47000
46549-7516751-5	Edward Masih	62	47000

Table 1.2

The publisher of Table1.1 is unaware about the publically available data as shown in Table 1.2. Now the data misuser can easily correlate records from Table 1.1 to Table 1.2 based upon age and postal code attributes and guesses the identity of individuals.

One can easily correlate record at row no. 3 in Table1 to row no. 3 in Table2 and can reveal the identity of the individual. Working on similar pattern will result in disclosure of identity of all individuals listed in Table 1.1 and Table 1.2.

The above examples have clarified that how easy it is for the miscreants to get the identity of individuals. Gaining personal identities is more dangerous as compared to other privacy pilferage.

In light of the above examples, we are convinced enough to focus our efforts for individuals' identity protection.

1.5 Definitions

In order to proceed further we have defined certain terms which will be frequently used in later discussion.

Data Provider

Data Provider is the person who is the actual owner of the data and mainly concerned about his privacy.

Data Collector

Data collector is directly linked with Data Provider who has collected the data either from their customers, students or through surveys.

Data User / Miner / Traverser

Data user, miner and traversers may be the same entities and basically meant to extract information from the published or provided datasets according to their own requirements.

Trusted Third Party

Trusted third party is in the pivotal role in entire data processing scenario. As its name suggests Trusted Third Party is enjoying the trust of both i.e. data owner and data user.

Trusted Third Party takes data from data collector, applies privacy protection algorithm and provide the protected data to the data users with following conditions:

I. It ensures the privacy preservation of data to data collector that data will be privacy protected prior to publishing.

II. It guarantees the data user that data provided does not deviate much from the actual data after applying privacy protection algorithms.

Trusted third party safeguard the interests of both parties i.e. data provider and data users.

There are some jargons frequently used in the datasets and needs to be elaborated as following:

Direct Identifiers

These are the attributes / columns in a table helping the data traverser to identify an individual in a dataset. Examples of direct identifiers are Name and CNIC. There may be other direct identifiers depending upon the datasets, therefore we cannot limit ourselves to Name and CNIC only. Other identifiers include Students' Roll Numbers, Patients IDs and Customers IDs etc.

Quasi Identifiers

Quasi Identifiers are those identifiers (Columns / Attributes) helping the data traverser to identify a person indirectly. These attributes usually relate to a person and are not unique enough to directly reveal the individual's identity.

Examples of Quasi Identifiers include age, sex, weight, region, religion, postal code etc. Quasi Identifier is an excellent tool for correlating multiple datasets and finally revealing the identity of an individual. We have already seen the example of Quasi Identifier correlation in Table 1.1 and Table 1.2. The data traverser easily correlated age and postal code from Table 1.1 to Table 1.2 and found individuals' identity.

Sensitive Identifiers

These are neither Direct nor Quasi identifiers. Sensitive identifier is purely relative to the dataset and owner of the dataset. It's up to the owner of the dataset defining the attributes as sensitive or otherwise. Let us see an example of patients' dataset, the disease of patients may be very sensitive to the owner of the hospitals and similarly CGPA of students may be termed as sensitive by Head of an educational institute.

Many data privacy techniques discourage to publish sensitive information but our proposed architecture opt otherwise after anonymizing the sensitive attributes.

1.6 Data Privacy and Identity Protection

Wide use of data mining techniques leads to data privacy issues and therefore need to address privacy preservation in KDD. Many techniques exist to cater this issue like Perturbation, Condensation, data hiding etc. Each tries to preserve privacy at maximum level but this leads to data loss resulting compromised data mining results.

Identity protection is the core objective of our proposed technique, since no other information can be more sensitive than the identity itself. Once the identity of a person is revealed, other information about the individuals can be correlated and accurately guessed.

We will see in the subsequent chapters that the proposed architecture for privacy protection will discourage the data miners to correlate different datasets for identity disclosure.

Chapter 2: Literature Review

This thesis not only introduces the concept of "Trusted 3rd party" but successfully proposes and implements "Data Anonymization Model" to be utilized by the "Trusted 3rd Party" for guaranteeing data privacy.

Data privacy is at stake when it is being shared by the data owner to the data user. Data privacy attacks like inference and reconstruction attacks jeopardize the data security leading to data misusage. "Data Misusage" can be defined as inference of information from the data set for which it has not been provided by the data owner.

There are multiple endeavors by the researcher to ensure data privacy while sharing to another party. This chapter encompasses the following models already proposed by the computer scientist in the domain of data privacy for data sharing with other parties or publishing the data publicly:

2.1. Privacy Preserving Data Mining Techniques

2.2. Utilizing Noise Addition for Data Privacy

2.3. Hybrid approach in Privacy Preserving

2.4. Task Independent Privacy Preservation

2.5. Chaotic Based Approach for Privacy Preservation

2.1. Privacy Preserving Data Mining Techniques [1]

The research study at subject highlights an important aspect of privacy compromise by linking quasi identifiers with other publically available data sets to recover personal identity.

K-Anonymity Model

K-anonymity model was proposed by Samarati and Sweeney to enforce data privacy for publically available data sets. A table is called K-anonymized if each record of a table is indistinguishable from K-1 other records considering the columns of Quasi Identifiers only. K-anonymity mode is viable in two scenarios

i) Knowledge of attributes available in the other / external data sources
ii) Type of data privacy attacks.

The Perturbation Approach

The data perturbation approach introduces two major techniques for data privacy. The first approach called the probability distribution approach interchanges the values within the data set. The second approach tends to add noise in data. The noise may be additive, multiplicative or combination of both.

Randomized Response Technique

RRT scrambles the data from individual users in such a way that no one can retrieve the exact information about a specific individual but integrity of the complete dataset remains intact. Thus information extracted from the entire dataset remains true while preserving the privacy of the individual records.

2.2. Noise Addition for Data Privacy [2]

This algorithm targets data privacy by noise addition to the actual data set. Adding noise results in changed information with least compromise on the desired results to be produced by the data set utilizer.

Proposed Algorithm

This technique has been decomposed in two sub-techniques to anonymize the actual data set for data privacy. These two sub-techniques are:

I. De-identification
II. Noise Addition

De-Identification

It is basically Personal Identity Information (PII) Identification and Removal. PII removal ensures removal of personal information from the data set. The personal information has also been termed as Personal / Direct Identifier in the relevant studies of data privacy. Personal Information includes

Name, SSN, and Phone Numbers etc. The proposed architecture believes in completing removing the personal identifier from the data set prior to noise addition.

Noise Addition

Noise addition is only possible in quantitative attributes like salary, CGPA etc. The research introduces multiples noises like additive noise, multiplicative noise and logarithmic multiplicative noise.

Noise addition always perturbed the actual data set. It is pertinent to keep the accuracy of perturbed data, close to the actual data set. The noise to be added is generated by randomization technique within the specified limits to keep the variance minimum between the anonymized data set and the actual data set. Final results comparison after application of the proposed algorithm shows correlation of 0.999 between actual data set and the anonymized data set. []

This approach works well only for data sets containing more quantitative attributes as compared to qualitative attributes. If a data set contains only qualitative attributes, then the proposed algorithm will not produce the desired results. []

2.3. Hybrid approach in Privacy Preserving [3]

The proposed approach applies two techniques to anonymize data for ensuring maximum data security. This hybrids approach first applies randomization technique and then applies modified K-anonymity model to protect the dataset from known attacks. []

Randomization Technique

The randomization technique proposed in this algorithm is quite different from the randomization technique already discussed in "Noise Addition Method". Randomization in noise addition method targets quantitative attributes of a data set while this technique randomizes the values irrespective of quantitative or qualitative aspect.

The proposed randomization techniques focus on relocating the values within a data set based on randomized Matrix (AxA) generated. Here "A" is the number of rows in a data set. The values of a tuples are interchanged to avoid linkage attack.

Modified K-anonymity Model

K-anonymity model advocates preserving data security by removing key attributes (direct identifier) and anonymizing the Quasi Identifiers. All the identifiers other than the direct identifiers are considered as sensitive attributes and need to be anonymized yielding in considerable information loss. The modified K-anonymity model categorizes the sensitive attributes in highly sensitive and low sensitive. The attributes categorized as highly sensitive are only anonymized to achieve minimal information loss as compared to information loss in K-anonymity model.

The proposed architecture successfully protects information if and only if the dataset is small rather the smallest. Since the technique generates a randomization matrix of Size AxA while A is the number of rows in a dataset, therefore this technique demands exorbitant computational powers and is not feasible with limited hardware resources.

2.4. Task Independent Privacy Preservation [4]

The aforesaid architecture targets preservation of medical data since authors have realized the sensitivity of medical records in context of privacy pilferage. The uniqueness of the proposed architecture of task independence considering data mining techniques. Since the database architecture containing all rows and columns remains intact by publishing all sensitive attributes during privacy preservation method therefore data miners can perform multiple tasks independently. The algorithm traverses the dataset in two cycles to preserve as following:

Numerical Data Transformation

Common privacy preservation techniques usually generalize the numerical data by classifying it in different domains. This technique however not only generalize the numerical data in different classes but assign numerical value to each class and then publishes the data. Publishing single value for one class eliminates the chances of inferring undesired information by the data miner.

Categorical Data Transformation

Most of the proposed techniques for ensuring privacy of individuals' medical records discourage to publish sensitive attributes. This technique rather encourages the publication of sensitive attributes but after mapping with dummy values. Any data miner can get the desired results with dummy values and he has to contact the data owner to know the original values against the mapped values. Thus the data owner can certainly share the original values after authenticating the genuine researchers and scientists.

This technique always keep the data owner as stake holder during the data mining process even he has published the dataset. Thus dependency on the data owner always remain inevitable. The commitment of the data owner for provision of original values against mapped values may increase manifolds if the published data is enormously utilized by the researchers.

2.5. Chaotic Based Approach for Privacy Preservation [5]

The algorithm in question addresses an interesting scenario of extracting the actual/original dataset from the perturbed data. The original dataset perturbed for the sake of publication uses different levels of Gaussian noise for different data users depending upon their trust levels. When multiple perturbed copies of the same data are present, it is easy to estimate / guess actual data.

Colpitts Oscillator as Chaotic System

Perturbation is usually achieved by adding noise generated by Gaussian PDF (Probability Distribution Function). This technique uses output of Gaussian PDF as input for the initial condition of the collpitts oscillator. The output of the collpitts oscillator is then used as noise to be added in the original dataset for generating multiple copies of the perturbed data from the original data set for various customers of different trust levels.

This technique works well with quantitative data but cannot be applied on qualitative intensive datasets. It is also computational hunger technique since it uses Gaussian PDF and collpits oscillator function to generate noise.

Chapter 3: Privacy Preservation Architecture

This chapter describes our proposed mechanism and its working with the help of simple example for better understanding of ultimate privacy preservation of an individual and accuracy of results as well. Data provider, data collector, data miner and decision makers are the core stake holders of the data mining process therefore we need to take care of their concerns, in Figure 5.2, core stake holders and their interaction with each other is illustrated.

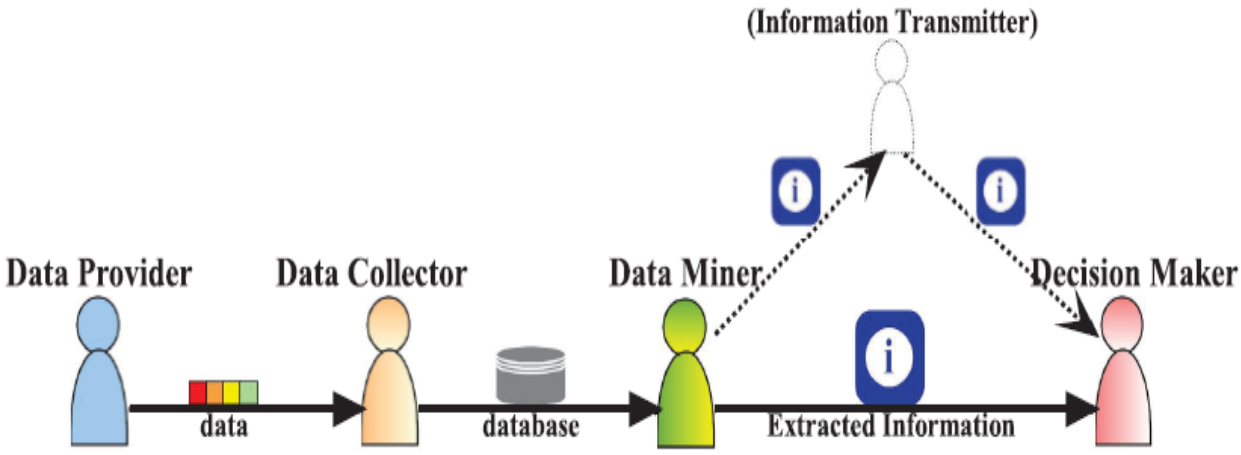

Figure 3.1 A simple illustration of the application scenario with data mining at the core

3.1. Trusted Third Party Mechanism

We are proposing a solution not only to preserve privacy with minimum accuracy loss which is concern of data provider but also taking care of concerns of other users as well by introducing trusted third party. Trusted third party will take care of the interests of users involved in KDD cycle from data provision to decision making. It has been observed that due to privacy concern data provider provides data false as incidents are reported where data of individuals have been misused, therefore they are reluctant to provide the correct data. Data Miner collects data from the data collector for analysis purpose but due to privacy concern they did not get all the required data which makes sometime incomplete analysis specially related to identity of individual. There are certain scenarios where data miner need to segregate number of individuals irrespective of their identity. For example, data miner wants to get number of customers who are affected by particular complaint from the complaints data where multiple complaints exist for specific customer. This is very simple example to illustrate that if identifier of individual is not specified then particular analysis may not be done on that data set. Based on the results of data miner, decision maker forms its strategy which impacts overall his business, continuing the mentioned example, without having identity he can't approach directly to his customer. Therefore, in Figure

3.2, Trusted Third Party has been introduced to take care of the concern of all stake holders, decision make can directly communicate to data collector who have detail information of individual or customer via Trusted Third Party and data collector provided complete information to trusted third party who make sure privacy of individual is preserved and results are not compromised. It's worth to note that there is a relationship between privacy and accuracy, as we increase the privacy, accuracy is compromised. In the next chapter we will present the graphical presentations to see the impact on accuracy as we increase privacy of the data.

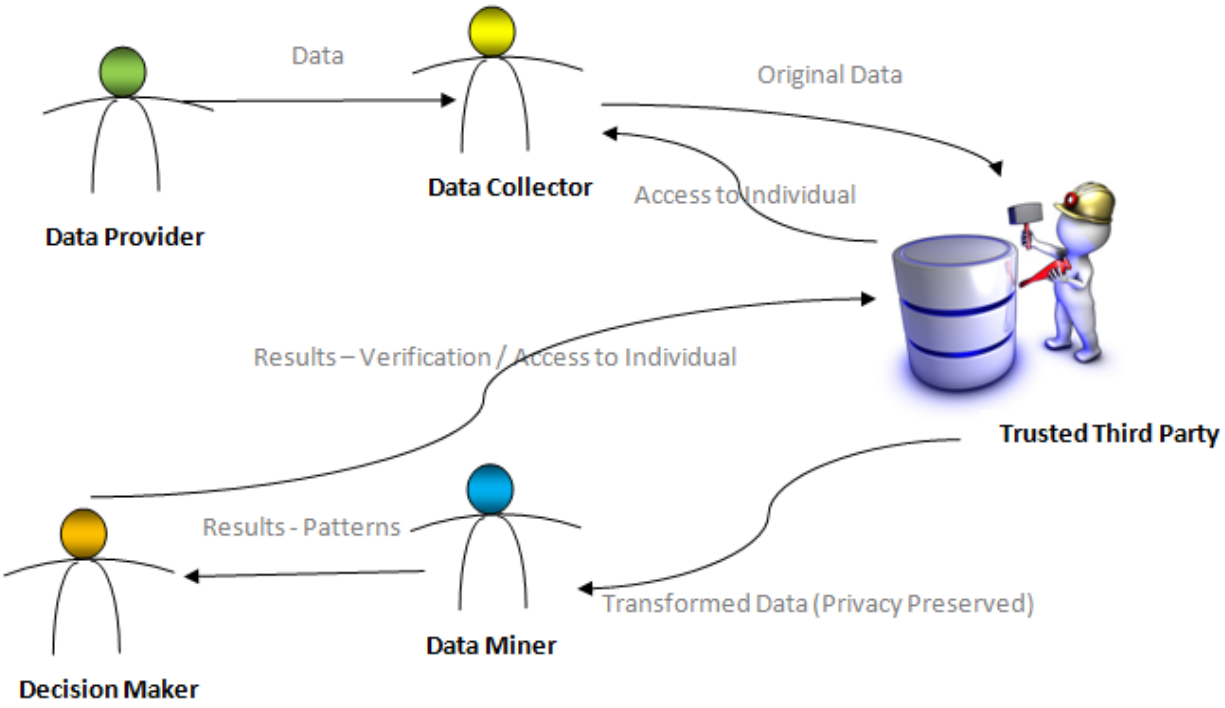

Figure 3.2 Trusted Third Party Concept

It has been observed that if privacy not compromised then accuracy of results is compromised hence data mining results get affected.

Most critical Quasi-identifier attributes need to identify in the data set as these can help attacker to identify individual and hence lead to privacy leak. Once identified such attributes, need to take appropriate measure and preserve privacy.

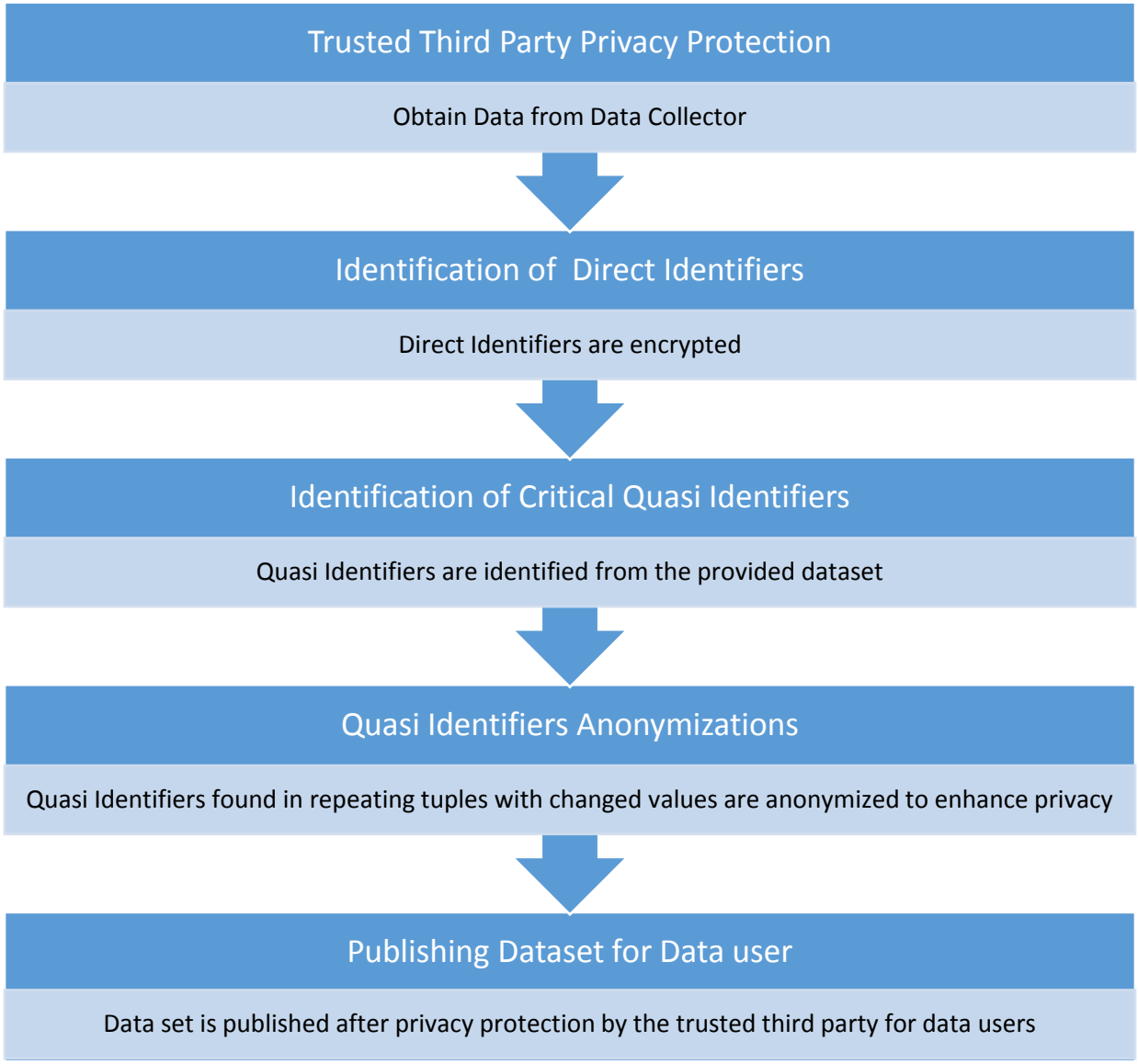

Figure 3.3 Trusted Third Party Mechanism

The above schematic diagram given in Figure 3.3 has been elaborated in the subsequent section of this chapter. The proposed architecture has been converted into working model of trusted third party using different implementation and analytical tools. The results in the Analysis chapter proves the success the of proposed architecture since variance in the actual and protected data after anonymization is negligible.

3.2. Proposed Privacy Preservation Architecture

In Figure 3.4, high level steps of proposed privacy preservation architecture have been given, starting from the dataset received from data collector, applying privacy preservation techniques and finally publishing dataset to data user for analysis purpose.

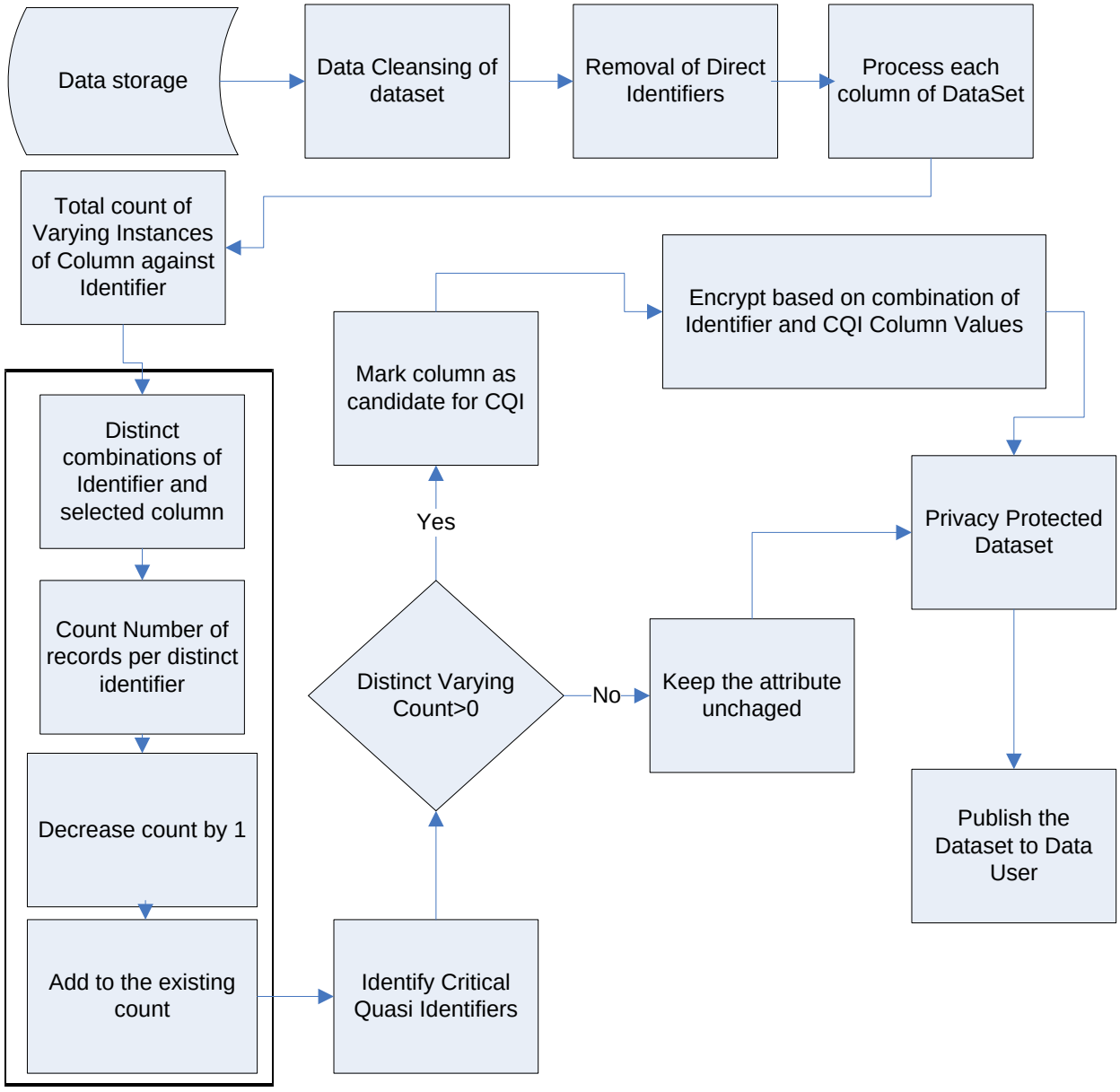

Figure 3.4 Privacy Preservation Architecture

3.3. Steps to achieve Privacy Preservation

i. Data Storage:

Data received from the data collector needs to be stored in the database for further processing. It will be fetched from the database by using sql scripts.

ii. Data cleansing of dataset:

Data may need further cleansing like removal of out liers, characters where need to be numbers, missing values etc. It depends on the quality of data received from data collector that how much cleansing required.

iii. Removal of direct Identifiers:

Before publishing dataset to data user, need to remove direct identifiers which are redundant and not utilized in analysis purpose. For example, in a dataset, we have two columns which can be used to identify individual, one Owner_Accnt_ID and CNIC_Number. Both columns can identify any individual therefore we need to remove CNIC_Number which is more sensitive than Owner_Accnt_ID. Our purpose can be fulfilled by using only Owner_Accnt_ID where data user need to segregate individuals and can do further analysis based on individuals.

iv. Process each column of dataset:

Each column from the dataset need to be processed one by one to check that how much it can be critical to expose privacy.

v. Total count of Varying Instances of Column against Identifier:

For each column, varying instances against identifier is calculated. This step is further elaborated in the figure 3.4, it can be understood by looking at the below example given in Table 3.1. We have sample dataset as given in Table 3.1, we have to calculate the instances where owner_accnt_id is same but occupation is varying. For example, we have owner_accnt_id 1-3JXL-161 and 1-YHW3-526 against which two different occupation exists. We are not concerned with the owner_accnt_id like 1-ZESF4-3221 where occupation is same.

Owner_Accnt_ID	**Occupation**	**Name**
1-3JXL-161	Labourer	Atif
1-3JXL-161	Technician	Atif
1-AV7-1443	Labourer	Jahangir
1-YHW3-526	Labourer	Khalid
1-YHW3-526	Technician	Khalid
1-ZESF4-3221	Programmer	Jameel
1-ZESF4-3221	Programmer	Jameel

Table. 3.1

Based on criteria mentioned against sample data set we got count of 2 against column ‘occupation’ as shown in Table 3.2. Its mean that there two instances against particular ‘owner_accnt_id’ where occupation is changed.

Column	Count
OWNER_ACCNT_ID	0
OCCUPATION	2
NAME	0

Table. 3.2

vi. Identify Critical Quasi Identifiers:

As sample dataset given in Table 3.1, if we encrypt ‘owner_accnt_id’, someone can short list the individuals who have gone through occupation change which is misuse of dataset and compromise on privacy. The purpose is to hide this fact as it is not the intended purpose of dataset, lower the count so less short listed individuals will be returned therefore that column will be more critical quasi identifier.

vii. Distinct Varying Count > 0: No: Keep the attribute unchanged

If distinct varying count is 0, its mean that there is no instance exist where identifier (owner_accnt_id) is same but column value (Name) is varying.

viii. Distinct Varying Count > 0: Yes: Mark column for Critical Quasi Identifier (CQI):

If distinct varying count is not 0 but higher, its mean that there is are instances exist where identifier (owner_accnt_id) is same but column value (Occupation) is varying.

ix. Encrypt based on combination of Identifier and CQI Column Values:

As Table 3.3 given, encryption has been applied on combination of Identifier (Owner_Accnt_ID) and occupation to make owner_accnt_id anonymized.

Owner_Accnt_ID	Occupation
DCB9B1E6A6	Labourer
0DB0D0983F	Technician
82423CFE0B	Labourer
19FC06D5F1	Technician
9F2C3BECC3	Labourer
62185E5136	Technician
3E62D5CD32	Labourer
BF4ACEDAE1	Technician

Table. 3.3

x. Privacy Protected dataset:

As Table 3.4 given based on Table 3.3, varying instance count became 0, its mean no one can extract information about the individuals who have change of occupations. In this way, privacy has been preserved by little manipulation but accuracy may be compromised a little. In Chapter 4, we will go into detail to compare that how much privacy has been preserved and accuracy compromised.

Column	Count
OWNER_ACCNT_ID	0
OCCUPATION	0

Table. 3.4

xi. Publish the Dataset to Data User:

This is the last step where privacy protected dataset can be shared with data user i.e. data miner.

Chapter 4: Evaluation and Testing

In this chapter details of dataset and results are discussed to understand and evaluate the technique we have proposed to increase privacy but their impact on the results. Briefly discussed about the compiled scripts to evaluate and test scenarios which fit into recommended scenario. Based on results generated by test scripts, graphical presentation in graphs aid to compare the results produced. Two datasets have been taken to double check our proposed technique and better compare the results.

4.1. Dataset-1

Sample dataset which is taken from one of the telecom company consists of around 50,000 rows. This dataset contains information related to the customers who are using Broadband as a service and periodically registered complaints to the company for the issues they have faced. It consists of the registered complaints where one or more rows relate to one customer i.e. customer may have registered more then one complaint. Oracle 11g is used create scripts and check different scenarios. This dataset can be utilized by the data miner to find hidden patterns which may help company to identify the major reason, their causes, resolution and ultimately enhancing customer satisfaction. Initially dataset has anomalies which were removed by writing and then applying transformation scripts.

4.2. Scripts Applied on Dataset-1

Various scripts containing Sql and Pl/Sql code have been written to fetch related result set. These scripts have been written considering performance as major concern. Performance of scripts found up to the mark to carry out required results. Below is the list of columns and their brief description used in dataset:

Column	Description
OWNER_ACCNT_ID	Customer Identifier
OCCUPATION	Occupation
RGN	Region
DILIVERY_METHOD	Billing delivery Method
DATA_RATE	Data Rate
EXCHANGE	Exchange
CUSTOMER_RATING	Customer Rating i.e. Silver, Platinum
TELEPHONE_NUM	Telephone Number
SERV_ACCT_ID	Service Account ID
BILL_ACCNT_ID	Billing Account ID
SR_NUM	Complaint Number

RESOLUTION_DAYS	Number of days taken to resolve the issue
NO_OF_INSTAL_DAYS	Number of installed days till complaint registered
NATURE_OF_FAULT	Type of Fault
STATUS_CD	Customer's Service Status
PRIMARY_CRED_VAL	Customer's Primary credential value
MOBILE_PHONE	Customer's mobile number
HOUSE_FLAT	Customer's Address, House / Flat number
SECTOR_AREA_HOUSING	Customer's Address, Sector / Area
LAST_NAME	Customer's Last Name
FIRST_NAME	Customer's First Name
EMAIL	Customer's Email address
ABM_NAME	Name of Person who is responsible for the service installed at customer's premises
SR_SEV_CD	Severity of Complaint
SR_PRIO_CD	Priority of the Complaint

Below is the result set generated by script [Annexure-A 2], it describes count for each column of the sample dataset which is the number of instances repeated against identifier column (Owner_Accnt_ID). For example, it will return 1 if two rows with different occupation against one owner_accnt_id returned, so overall it calculates and return final count for column 'occupation', in this way count calculated for each column as result presented Table 4.1:

Column	**Count**
C_OWNER_ACCNT_ID	0
C_OCCUPATION	4
C_RGN	11
C_DILIVERY_METHOD	15
C_DATA_RATE	31
C_EXCHANGE	64
C_CUSTOMER_RATING	157
C_TELEPHONE_NUM	68
C_SERV_ACCT_ID	68
C_BILL_ACCNT_ID	68
C_SR_NUM	26381
C_RESOLUTION_DAYS	16672
C_NO_OF_INSTAL_DAYS	26157
C_NATURE_OF_FAULT	14209
C_STATUS_CD	0
C_PRIMARY_CRED_VAL	0
C_MOBILE_PHONE	56
C_HOUSE_FLAT	34

C_SECTOR_AREA_HOUSING	21
C_LAST_NAME	0
C_FIRST_NAME	0
C_EMAIL	386
C_ABM_NAME	5974
C_SR_SEV_CD	8644
C_SR_PRIO_CD	8038

Table 4.1

Below is the result set generated by script [Annexure-A 3], it describes count for each column of the sample dataset which is the number of instances repeated against concatenated hash value generated with combination of identifier column (Owner_Accnt_ID) and most critical quasi identifier columns (New_Occupation, Rgn, Old_Dilivery_Method, Data_Rate, Exchange, Customer_Rating) as identified in result set presented in Table 4.1. Lower the count but not zero so the column is most critical quasi identifier, for further clarification refer to result presented in Table 4.1. As result presented in Table 4.2, generated hash value further anonymized the key generated with combination of C_OCCUPATION, C_RGN, C_DILIVERY_METHOD, C_DATA_RATE, C_EXCHANGE and C_CUSTOMER_RATING, therefore, result showing 0 against all critical quasi identifier columns.

Column	**Count**
C_OWNER_ACCNT_ID	0
C_OCCUPATION	0
C_RGN	0
C_DILIVERY_METHOD	0
C_DATA_RATE	0
C_EXCHANGE	0
C_CUSTOMER_RATING	0
C_TELEPHONE_NUM	27
C_SERV_ACCT_ID	27
C_BILL_ACCNT_ID	27
C_SR_NUM	23775
C_RESOLUTION_DAYS	15374
C_NO_OF_INSTAL_DAYS	23586
C_NATURE_OF_FAULT	12195
C_STATUS_CD	0
C_PRIMARY_CRED_VAL	0
C_MOBILE_PHONE	55
C_HOUSE_FLAT	32
C_SECTOR_AREA_HOUSING	19
C_LAST_NAME	0
C_FIRST_NAME	0
C_EMAIL	371

C_ABM_NAME	5030
C_SR_SEV_CD	8357
C_SR_PRIO_CD	7775

Table 4.2

Please not that result presented in Table 4.2,

Below is the result set generated by script [Annexure-A 4], it describes count for each column of the sample dataset which is the number of instances repeated against concatenated hash value generated with combination of identifier column (Owner_Accnt_ID), most critical quasi identifier columns (New_Occupation, Rgn, Old_Dilivery_Method, Data_Rate, Exchange, Customer_Rating) and SR_Num. By adding only SR_Num with the combination, result showing 0 against all columns, its mean repetition has not been found against the new combination.

Column	Count
C_OWNER_ACCNT_ID	0
C_OCCUPATION	0
C_RGN	0
C_DILIVERY_METHOD	0
C_DATA_RATE	0
C_EXCHANGE	0
C_CUSTOMER_RATING	0
C_TELEPHONE_NUM	0
C_SERV_ACCT_ID	0
C_BILL_ACCNT_ID	0
C_SR_NUM	0
C_RESOLUTION_DAYS	0
C_NO_OF_INSTAL_DAYS	0
C_NATURE_OF_FAULT	0
C_STATUS_CD	0
C_PRIMARY_CRED_VAL	0
C_MOBILE_PHONE	0
C_HOUSE_FLAT	0
C_SECTOR_AREA_HOUSING	0
C_LAST_NAME	0
C_FIRST_NAME	0
C_EMAIL	0
C_ABM_NAME	0
C_SR_SEV_CD	0
C_SR_PRIO_CD	0

Table 4.3

To understand better, data of customers who have more than one distinct regions as presented in Table 4.4, refer to script [Annexure-A 5]

Owner_Accnt_ID	Rgn
1-1317Q-323	KTR - 1
1-1317Q-323	KTR - 2

1-14IK9-12	KTR - 2
1-14IK9-12	KTR - 3
1-1F5KW-335	HTR
1-1F5KW-335	ITR
1-32C8-350	ITR
1-32C8-350	RTR
1-3AQNA06N	HTR
1-3AQNA06N	ITR
1-3AZWM44O	KTR - 2
1-3AZWM44O	KTR - 3
1-3E1-12	KTR - 3
1-3E1-12	STR
1-3W3L-361	GTR
1-3W3L-361	RTR
1-4GMB-327	KTR - 1
1-4GMB-327	KTR - 2
1-90B-719	KTR - 2
1-90B-719	KTR - 3
1-VAQZ-190	KTR - 1
1-VAQZ-190	KTR - 2

Table 4.4

To understand better, data of customers who have more than one distinct occupation as presented in Table 4.5, refer to script [Annexure-A 6]

Owner_Accnt_ID	**Occupation**
1-3JXL-161	Labourer
1-3JXL-161	Technician
1-4NV-4062	Labourer
1-4NV-4062	Technician
1-AV7-1443	Labourer
1-AV7-1443	Technician
1-YHW3-526	Labourer
1-YHW3-526	Technician

Table 4.5

To understand better, data of customers who have more than one distinct delivery method of billing as presented in Table 4.6, refer to script [Annexure-A 7]

Owner_Accnt_ID	**Delivery_Method**
1-116PO-199	General POST

1-116PO-199	TCS
1-17RNO-73	General POST
1-17RNO-73	TCS
1-1F5KW-335	General POST
1-1F5KW-335	TCS
1-1GHGN-304	General POST
1-1GHGN-304	TCS
1-1IMBW-455	General POST
1-1IMBW-455	TCS
1-1JKU5-444	General POST
1-1JKU5-444	TCS
1-1NF0C-1326	General POST
1-1NF0C-1326	TCS
1-1SEZ3-638	General POST
1-1SEZ3-638	TCS
1-4EG-358	General POST
1-4EG-358	TCS
1-4GMB-327	General POST
1-4GMB-327	TCS
1-8Q5-3018	CSR/Exchange Staff
1-8Q5-3018	General POST
1-9Q2-845	General POST
1-9Q2-845	TCS
1-VAQZ-190	General POST
1-VAQZ-190	TCS
1-XFKP-243	General POST
1-XFKP-243	TCS
1-Y2LC-325	General POST
1-Y2LC-325	TCS

Table 4.6

4.3. Results Discussion - Dataset-1

Summary of the data has been prepared based on the result set returned by scripts [Annexure-A 7, A-8 and A-9] as shown in Table 4.7, column A representing number and %age of customers affected based on per customer per row, column B representing number and %age of customers based on identifier (owner_accnt_id) and column C representing number and %age of customers affected based on hash value generated by combination of critical quasi identifiers.

Tabular Comparison	A	B	A-B	C	C-B
TOTAL_CUSTOMERS_AFFECTEE	**49591**	**37496**	12095	**38087**	591
Sync Loss	14939	10690	4249	10776	86
%Sync Loss	30.12	28.51	1.61	28.29	-0.22

Frequent disconnections	8690	5820	2870	6130	310
%Frequent disconnections	17.52	15.52	2.00	16.09	0.57
Customer Can't Configure CPE	8513	6996	1517	7027	31
%Customer Can't Config CPE	17.17	18.66	-1.49	18.45	-0.21
No Browsing	4931	3219	1712	3281	62
%No Browsing	9.94	8.58	1.36	8.61	0.03
Authentication Problem	3166	2748	418	2766	18
%Authentication Problem	6.38	7.33	-0.94	7.26	-0.07
Port Up/Down Gradation	2395	2078	317	2082	4
%Port Up/Down Gradation	4.83	5.54	-0.71	5.47	-0.08
Slow Browsing-Low Line Params	1917	1499	418	1550	51
%Slow Browsng-Low Line Params	3.87	4.00	-0.13	4.07	0.07
Slow Browsing - Network Issue	1658	1371	287	1386	15
%Slow Browsing - Network Issue	3.34	3.66	-0.31	3.64	-0.02
Modem and Splitter Problem	1078	955	123	965	10
%Modem and Splitter Problem	2.17	2.55	-0.37	2.53	-0.01
Freqt Discon but line Param ok	1023	898	125	899	1
%Freq Discon but line Param ok	2.06	2.39	-0.33	2.36	-0.03
Service Restoration	650	626	24	629	3
%Service Restoration	1.31	1.67	-0.36	1.65	-0.02
Data Rate Issue	631	596	35	596	0
%Data Rate Issue	1.27	1.59	-0.32	1.56	-0.02

Table 4.7

Tabular Comparison	A	%age (A)	B	%age (B)	(A)-(B)	C	%age (C)	(C)-(B)
Nature of Fault	**49591**		**37496**		**12095**	**38087**		591
Sync Loss	14939	30.12	10690	28.51	1.61	10776	28.29	-0.22
Frequent disconnections	8690	17.52	5820	15.52	2.00	6130	16.09	0.57
Customer Can't Configure CPE	8513	17.17	6996	18.66	-1.49	7027	18.45	-0.21
No Browsing	4931	9.94	3219	8.58	1.36	3281	8.61	0.03
Authentication Problem	3166	6.38	2748	7.33	-0.94	2766	7.26	-0.07
Port Up/Down Gradation	2395	4.83	2078	5.54	-0.71	2082	5.47	-0.08
Slow Browsing-Low Line Params	1917	3.87	1499	4.00	-0.13	1550	4.07	0.07
Slow Browsing - Network Issue	1658	3.34	1371	3.66	-0.31	1386	3.64	-0.02
Modem and Splitter Problem	1078	2.17	955	2.55	-0.37	965	2.53	-0.01
Freqt Discon but line Param ok	1023	2.06	898	2.39	-0.33	899	2.36	-0.03
Service Restoration	650	1.31	626	1.67	-0.36	629	1.65	-0.02

Data Rate Issue	631	1.27	596	1.59	-0.32	596	1.56	-0.02

Table 4.8

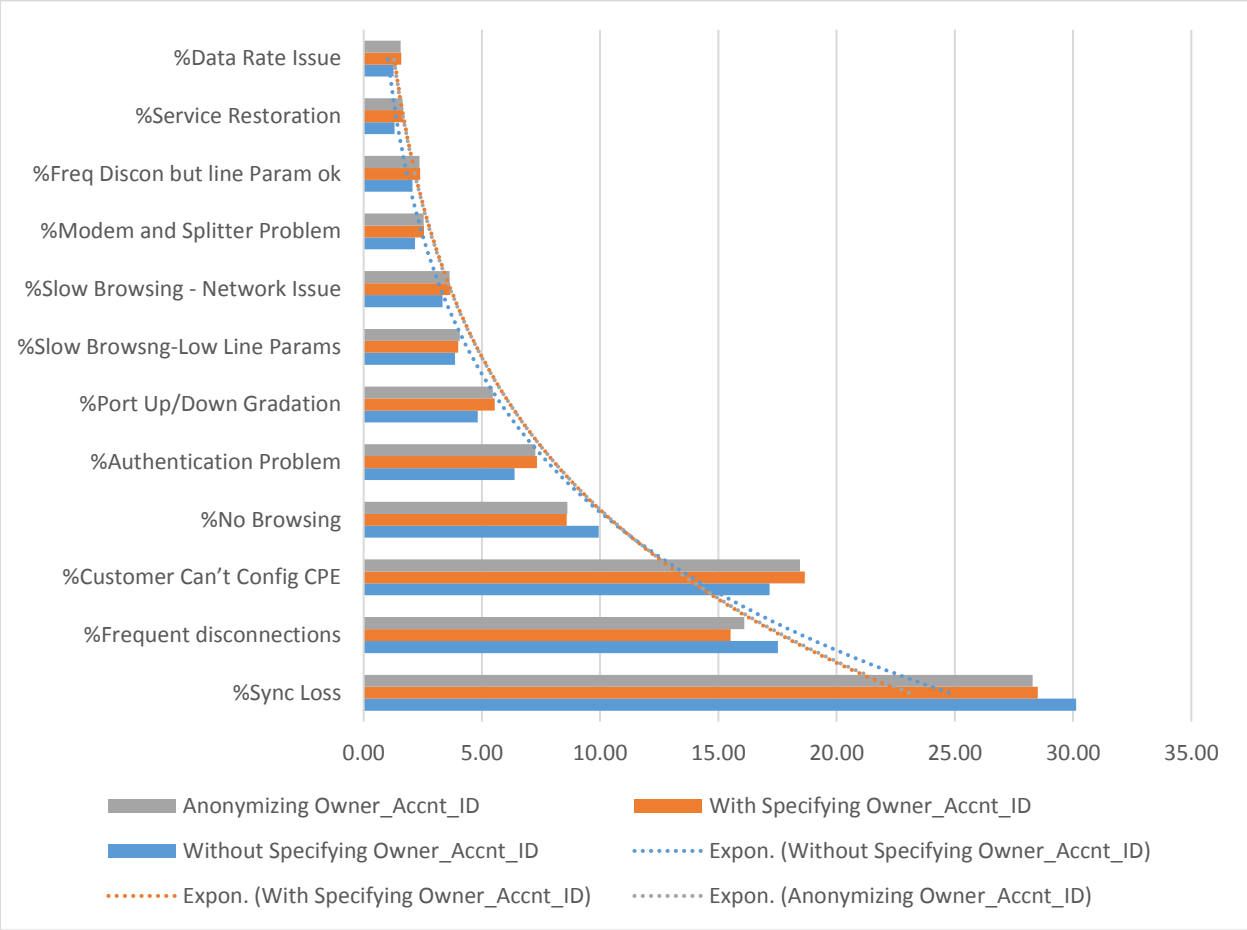

Figure 4.1 Result Accuracy Comparison based on Dataset-1

Diabetic Data set

4.4. Dataset-2

The second dataset has been used to cross validate the results produced by using dataset-1. This dataset has been taken online from *https://archive.ics.uci.edu/ml/datasets/Diabetes* dataset and it consists of around .1 million rows. This dataset contains information related to the admission of a patient in the hospital, tests taken and their result, and medication advised.

Below are the five conditions to restrict the data for analysis purpose:

1) Inpatient encounter
2) Diabetic was the initial diagnosis
3) Minimum time period of stay in the Hospital was 24 Hours but less than 14 Days
4) Laboratory Tests were taken by the Hospital during that period
5) Medications were taken care by the Hospital

Below are the four main categories used for further analysis of readmission probability:

1. HbA1c - No test was performed
2. HbA1c - Result was high and the diabetic medication was changed
3. HbA1c - Result was high but the diabetic medication was not changed
4. HbA1c - Normal result of the test

Oracle 11g is used create scripts and check different scenarios. This dataset can be utilized by the data miner to find hidden patterns which may help organizations related to health toward identifying the major reason which causing readmission rate higher and further to cater that reason to lower readmission rates [6]. For this purpose, following four categories have been used to analyze the results to identify the cause of readmission:

4.5. Scripts Applied on Dataset-2

Various scripts containing Sql and Pl/Sql code have been written to fetch related result set as detail given. These scripts have been written considering performance as major concern. Performance of scripts found up to the mark to carry out required results.

Result set given in Table 4.9 is the result of script given in [Annexure-A 11], it is supposed that if Patient Number is not given in the dataset, then each row will be considered as the unique instance of patient. In this script we have grouped our calculations based on HbA1c result, diabetic medication changes i.e. Y/N and readmitted status. %age of the population and Readmitted %in the group are calculated percentage fields based on no. of encounters and readmitted (no. of encounters).

	No. of encounters	**% of the population**	**Readmitted – No. of encounters**	**Readmitted - % in group**
	99353			
HbA1c - No test was performed	82518	83.1	9641	**11.7**
HbA1c - Result was high and the diabetic medication was changed	5306	5.3	546	**10.3**
HbA1c - Result was high but the diabetic medication was not changed	2831	2.8	263	**9.3**
HbA1c - Normal result of the test	8698	8.8	864	**9.9**

Table 4.9

Result set given in Table 4.10 is the result of script given in [Annexure-A 12], hash value of Patient Number is created, then distinct patient has been fetched based on distinct value of hash value created i.e. one instance against one hash value of patient number, the result criteria remains same used to produce results given in Table 4.9.

	No. of encounters	% of the population	Readmitted – No. of encounters	Readmitted - % in group
	69997			
HbA1c - No test was performed	57073	81.5	5343	**9.4**
HbA1c - Result was high and the diabetic medication was changed	4077	5.8	362	**8.9**
HbA1c - Result was high but the diabetic medication was not changed	2189	3.1	166	**7.6**
HbA1c - Normal result of the test	6658	9.5	589	**8.8**

Table 4.10

Result set given in Table 4.11 is the result of script given in [Annexure-A 13], hash value of Patient Number is created on basis of combination of Patient Number and weight, then distinct patient has been fetched based on distinct value of hash value created i.e. one instance against one hash value of patient number and weight, the result criteria remains same used to produce results given in Table 4.9. Creating hash value by combining patient number and weight, we are anonymizing Patients to further short list patients who has different weights in the dataset. It can be noticed that number of rows given in Table 4.11 are increased a little based on the combination with weight as compared to table 4.10.

	with weight			
	No. of encounters	**% of the population**	**Readmitted – No. of encounters**	**Readmitted - % in group**
	70126			
HbA1c - No test was performed	57189	81.6	5362	**9.4**
HbA1c - Result was high and the diabetic medication was changed	4079	5.8	362	**8.9**
HbA1c - Result was high but the diabetic medication was not changed	2190	3.1	166	**7.6**
HbA1c - Normal result of the test	6668	9.5	590	**8.8**

Table 4.11

Result set given in Table 4.12 is the result of script given in [Annexure-A 13], hash value of Patient Number is created on basis of combination of Patient Number and Age, then distinct patient has been fetched based on distinct value of hash value created i.e. one instance against one hash value of patient number and age, the result criteria remains same used to produce results given in Table 4.9. Creating hash value by combining patient number and age, we are anonymizing Patients to further short list patients who has different ages in the dataset. It can be

noticed that number of rows given in Table 4.12 are increased a little based on the combination with age as compared to table 4.11, it means that there are more patients who have varying age then weight.

	with age			
	No. of encounters	**% of the population**	**Readmitted – No. of encounters**	**Readmitted - % in group**
	71532			
HbA1c - No test was performed	58367	81.6	5514	**9.4**
HbA1c - Result was high and the diabetic medication was changed	4155	5.8	375	**9.0**
HbA1c - Result was high but the diabetic medication was not changed	2232	3.1	169	**7.6**
HbA1c - Normal result of the test	6778	9.5	604	**8.9**

Table 4.12

Result set given in Table 4.13 is the result of script given in [Annexure-A 14], hash value of Patient Number is created on basis of combination of Patient Number, Weight and Age, then distinct patient has been fetched based on distinct value of hash value created i.e. one instance against one hash value of patient number, weight and age, the result criteria remains same used to produce results given in Table 4.9. Creating hash value by combining patient number, weight and age, we are anonymizing Patients to further short list patients who has different age and weight in the dataset. It can be noticed that number of rows given in Table 4.13 are more then all related tables given i.e. Table 4.10, 4.11 and 4.12, it means that there are more patients who have varying either age or weight.

	with Weight and Age			
	No. of encounters	**% of the population**	**Readmitted – No. of encounters**	**Readmitted - % in group**
	71632			
HbA1c - No test was performed	58457	81.6	5529	**9.5**
HbA1c - Result was high and the diabetic medication was changed	4156	5.8	375	**9.0**
HbA1c - Result was high but the diabetic medication was not changed	2233	3.1	169	**7.6**
HbA1c - Normal result of the test	6786	9.5	605	**8.9**

Table 4.13

4.6. Results Discussion - Dataset-2

Table 4.14 is the summarized table to compare the results presented in table 4.9, 4.10, 4.11, 4.12 and 4.13 and column ‘Readmitted %age in group’. It can be observed the difference between

actual results given in column B (extracted from Table 4.10) and other columns A (extracted from Table 4.9), C (extracted from Table 4.11), D (extracted from Table 4.12) and E (extracted from Table 4.13). Moreover, variation of result is much higher as given in column A with respected to actual results B then the other columns C, D and E.

	A	B	C	D	E
HbA1c - No test was performed	11.68	9.36	9.38	9.45	9.46
HbA1c - Result was high and the diabetic medication was changed	10.29	8.88	8.87	9.03	9.02
HbA1c - Result was high but the diabetic medication was not changed	9.29	7.58	7.58	7.57	7.57
HbA1c - Normal result of the test	9.93	8.85	8.85	8.91	8.92

Table 4.14

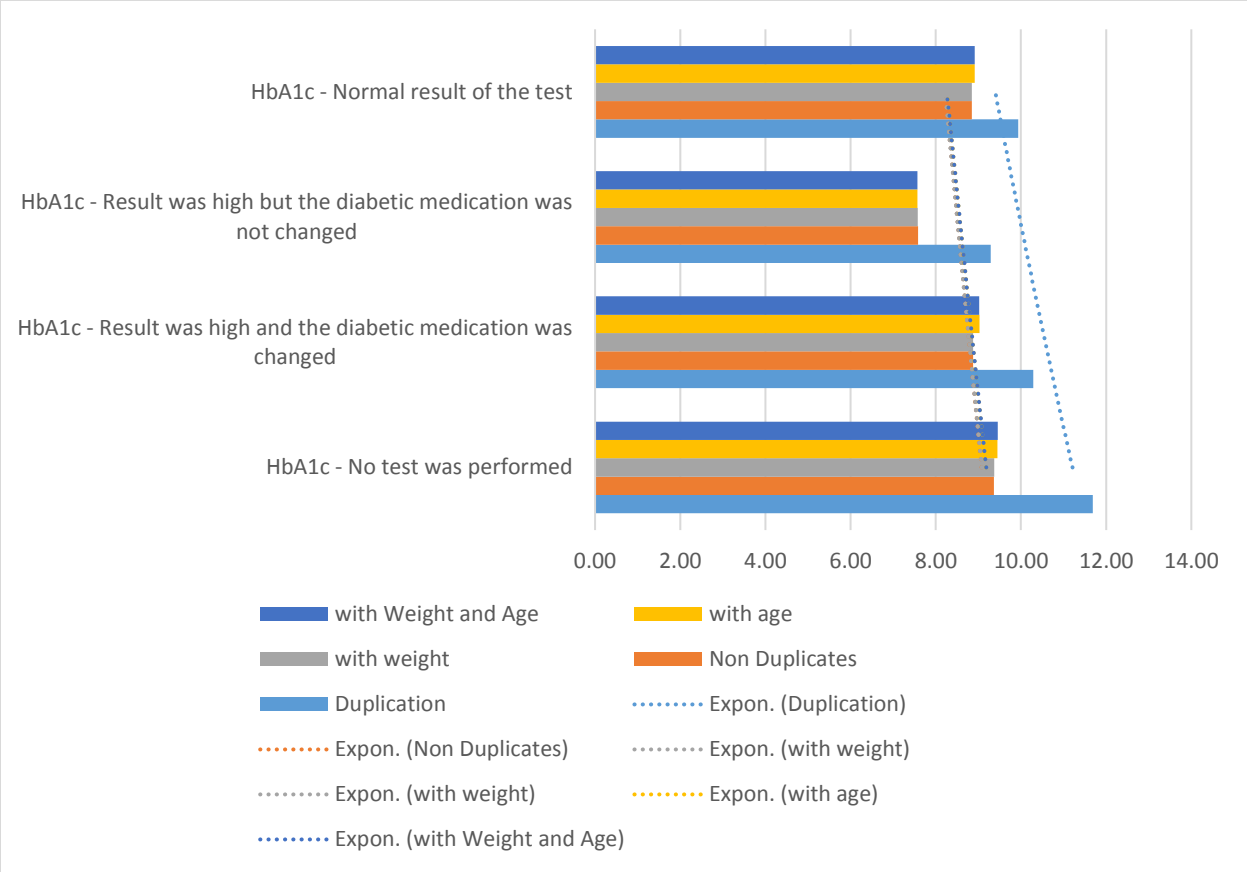

Figure 4.2 Result Accuracy Comparison based on Dataset-2

Figure 4.2 is the graphical representation of our results given Table 4.14. Variation of result is well depicted in the chart i.e. difference much higher as given in column A with respect to actual results B then the other columns C, D and E. In nutshell we have seen that anonymizing identifier key based on critical quasi identifiers, we have increased privacy on the cost of bit variation of results. The trend line shown for all i.e. A, B, C, D and E but we can see only two

trend lines, the reason is that trend line for B, C, D and E are so close and overlapped whereas trend line printed for column A is easily distinguishable and bit away from the overlapped trend line against B, C, D and E.

Conclusion

This chapter concludes our endeavor about the proposed privacy protection architecture. It has been established that privacy protection is inevitable prior to publishing data for data miners.

We have studied multiple architectures to address the privacy protection challenge during literature review phase. The techniques summarized in chapter 2 compelled us to propose architecture applicable in heterogeneous type of datasets.

Our proposed architecture has shown a new facet in the dimension of data protection by introducing the concept of trusted third party. It has won the confidence of both the parties i.e. data provider and data user.

The proposed architecture has not been limited to a theoretical model, rather it has been implemented to prove its efficacy on complex datasets. The empirical results on different datasets tabulated in result analysis section and graphs depicting the comparisons between actual dataset and anonymized data set are self-evident of its success. Moreover, the proposed architecture safeguards the data against correlation attacks of published dataset against external datasets.

The scripts written to convert the proposed architecture into working model are simple enough to be comprehendible by person having basic knowledge of SQL and PL/SQL. The implementation has been kept simple to accommodate future enhancements smoothly.

The proposed architecture in this thesis may be further enhanced by incorporating following additions:

a. Generalization may be added in parallel to suppression for anonymization.
b. The scripts may be optimized for performance tuning to handle large scale datasets consisting millions of records.
c. Trusted third party may provide the services of data mining to the data users to extract the required information from the datasets and publish the anonymized dataset along with results.

Privacy preservation is vast subject and still require further working. In future we can further extend our work to compare our results in terms of cost as security comes with cost. As we are applying algorithms to enhance security therefore further study needed to measure impact of cost. Moreover, we have found critical quasi identifiers based on single attribute, further study needed to experiment with combination of attributes rather only one attribute.

Annexure A: Scripts

1 Script

```
CREATE OR REPLACE FUNCTION Count_Col_d(Col_Name_A IN VARCHAR2, Col_ID_A
IN VARCHAR2) RETURN NUMBER IS
/*********************************************
Total Change Count per Col_ID_A
*********************************************/
  sql_stmt    VARCHAR2(2000);
  count_rec       NUMBER;

BEGIN
  sql_stmt := ' Select Sum(Count_CUST_RANK) Count_Cust_Rank
 From
  ( Select
    '||Col_Id_A||', Count(distinct UPPER('||Col_Name_A||'))-1 Count_CUST_RANK
    From BB_COMPLAINT_DATA_TBL cd
    Where Is_Deleted = "N" and '||Col_Name_A||' is not null--not null
    Group by '||Col_Id_A||
  ')
';

   EXECUTE IMMEDIATE sql_stmt INTO count_rec;
  DBMS_OUTPUT.PUT_LINE(count_rec);
    RETURN count_rec;
END Count_Col_d;
```

2 Script

```
/*****************************************************************
Refer to section 2. Script for definition of function utilized in section 1. Script
*****************************************************************/
SELECT
Count_Col_d('OWNER_ACCNT_ID','OWNER_ACCNT_ID') C_OWNER_ACCNT_ID,
Count_Col_d('NEW_OCCUPATION','OWNER_ACCNT_ID') C_OCCUPATION,
Count_Col_d('RGN','OWNER_ACCNT_ID') C_RGN,
Count_Col_d('OLD_DILIVERY_METHOD','OWNER_ACCNT_ID')
C_DILIVERY_METHOD,
Count_Col_d('DATA_RATE','OWNER_ACCNT_ID') C_DATA_RATE,
Count_Col_d('EXCHANGE','OWNER_ACCNT_ID') C_EXCHANGE,
Count_Col_d('CUSTOMER_RATING','OWNER_ACCNT_ID') C_CUSTOMER_RATING,
Count_Col_d('TELEPHONE_NUM','OWNER_ACCNT_ID') C_TELEPHONE_NUM,
```

```
Count_Col_d('SERV_ACCT_ID','OWNER_ACCNT_ID') C_SERV_ACCT_ID,
Count_Col_d('BILL_ACCNT_ID','OWNER_ACCNT_ID') C_BILL_ACCNT_ID,
Count_Col_d('SR_NUM','OWNER_ACCNT_ID') C_SR_NUM,
Count_Col_d('RESOLUTION_DAYS','OWNER_ACCNT_ID') C_RESOLUTION_DAYS,
Count_Col_d('NO_OF_INSTAL_DAYS','OWNER_ACCNT_ID') C_NO_OF_INSTAL_DAYS,
Count_Col_d('NATURE_OF_FAULT','OWNER_ACCNT_ID') C_NATURE_OF_FAULT,
Count_Col_d('STATUS_CD','OWNER_ACCNT_ID') C_STATUS_CD,
Count_Col_d('NEW_PRIMARY_CREDENTIAL_VALUE','OWNER_ACCNT_ID')
C_PRIMARY_CRED_VAL,
Count_Col_d('NEW_MOBILE_PHONE','OWNER_ACCNT_ID') C_MOBILE_PHONE,
Count_Col_d('NEW_HOUSE_FLAT','OWNER_ACCNT_ID') C_HOUSE_FLAT,
Count_Col_d('NEW_SECTOR_AREA_HOUSING','OWNER_ACCNT_ID')
C_SECTOR_AREA_HOUSING,
Count_Col_d('NEW_LAST_NAME','OWNER_ACCNT_ID') C_LAST_NAME,
Count_Col_d('NEW_FIRST_NAME','OWNER_ACCNT_ID') C_FIRST_NAME,
Count_Col_d('NEW_EMAIL','OWNER_ACCNT_ID') C_EMAIL,
Count_Col_d('X_ABM_NAME','OWNER_ACCNT_ID') C_ABM_NAME,
Count_Col_d('SR_SEV_CD','OWNER_ACCNT_ID') C_SR_SEV_CD,
Count_Col_d('SR_PRIO_CD','OWNER_ACCNT_ID') C_SR_PRIO_CD
FROM DUAL
```

3 Script

```
/*****************************************************************
Refer to section 2. Script for definition of function utilized in section 1. Script
Sys.CONV_HASH built-in function has been utilized to generate hash value with the
combination of critical quasi identifiers additional to Owner_Accnt_ID i.e. New_Occupation,
Rgn, Old_Dilivery_Method, Data_Rate, Exchange and Customer_Rating
*****************************************************************/

SELECT
Count_Col_d('OWNER_ACCNT_ID','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCU
PATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_OWNER_ACCNT_ID,
Count_Col_d('NEW_OCCUPATION','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCU
PATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_OCCUPATION,
Count_Col_d('RGN','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||O
LD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)') C_RGN,
Count_Col_d('OLD_DILIVERY_METHOD','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_
OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_DILIVERY_METHOD,
Count_Col_d('DATA_RATE','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATIO
N||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_DATA_RATE,
```

```
Count_Col_d('EXCHANGE','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION
||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)') C_EXCHANGE,
Count_Col_d('CUSTOMER_RATING','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCC
UPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_CUSTOMER_RATING,
Count_Col_d('TELEPHONE_NUM','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUP
ATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_TELEPHONE_NUM,
Count_Col_d('SERV_ACCT_ID','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPAT
ION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_SERV_ACCT_ID,
Count_Col_d('BILL_ACCNT_ID','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPA
TION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_BILL_ACCNT_ID,
Count_Col_d('SR_NUM','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rg
n||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)') C_SR_NUM,
Count_Col_d('RESOLUTION_DAYS','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCC
UPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_RESOLUTION_DAYS,
Count_Col_d('NO_OF_INSTAL_DAYS','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OC
CUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_NO_OF_INSTAL_DAYS,
Count_Col_d('NATURE_OF_FAULT','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCC
UPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_NATURE_OF_FAULT,
Count_Col_d('STATUS_CD','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION
||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_STATUS_CD,
Count_Col_d('NEW_PRIMARY_CREDENTIAL_VALUE','Sys.CONV_HASH(OWNER_ACC
NT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Custo
mer_Rating)') C_PRIMARY_CRED_VAL,
Count_Col_d('NEW_MOBILE_PHONE','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OC
CUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_MOBILE_PHONE,
Count_Col_d('NEW_HOUSE_FLAT','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCU
PATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_HOUSE_FLAT,
Count_Col_d('NEW_SECTOR_AREA_HOUSING','Sys.CONV_HASH(OWNER_ACCNT_ID||
NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rat
ing)') C_SECTOR_AREA_HOUSING,
Count_Col_d('NEW_LAST_NAME','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUP
ATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_LAST_NAME,
Count_Col_d('NEW_FIRST_NAME','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCU
PATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_FIRST_NAME,
```

```
Count_Col_d('NEW_EMAIL','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATIO
N||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)') C_EMAIL,
Count_Col_d('X_ABM_NAME','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATI
ON||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_ABM_NAME,
Count_Col_d('SR_SEV_CD','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION
||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)') C_SR_SEV_CD,
Count_Col_d('SR_PRIO_CD','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATIO
N||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating)')
C_SR_PRIO_CD
FROM DUAL
```

4 Script

```
/*****************************************************************
Refer to section 2. Script for definition of function utilized in section 1. Script
Sys.CONV_HASH built-in function has been utilized to generate hash value with the
combination of critical quasi identifiers additional to Owner_Accnt_ID i.e. New_Occupation,
Rgn, Old_Dilivery_Method, Data_Rate, Exchange, Customer_Rating and SR_Num
*****************************************************************/
SELECT
Count_Col_d('OWNER_ACCNT_ID','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCU
PATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NU
M)') C_OWNER_ACCNT_ID,
Count_Col_d('NEW_OCCUPATION','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCU
PATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NU
M)') C_OCCUPATION,
Count_Col_d('RGN','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||O
LD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_RGN,
Count_Col_d('OLD_DILIVERY_METHOD','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_
OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||S
R_NUM)') C_DILIVERY_METHOD,
Count_Col_d('DATA_RATE','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATIO
N||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_DATA_RATE,
Count_Col_d('EXCHANGE','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION
||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_EXCHANGE,
Count_Col_d('CUSTOMER_RATING','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCC
UPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NU
M)') C_CUSTOMER_RATING,
Count_Col_d('TELEPHONE_NUM','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUP
ATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)
') C_TELEPHONE_NUM,
```

```
Count_Col_d('SERV_ACCT_ID','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_SERV_ACCT_ID,
Count_Col_d('BILL_ACCNT_ID','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_BILL_ACCNT_ID,
Count_Col_d('SR_NUM','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_SR_NUM,
Count_Col_d('RESOLUTION_DAYS','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_RESOLUTION_DAYS,
Count_Col_d('NO_OF_INSTAL_DAYS','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_NO_OF_INSTAL_DAYS,
Count_Col_d('NATURE_OF_FAULT','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_NATURE_OF_FAULT,
Count_Col_d('STATUS_CD','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_STATUS_CD,
Count_Col_d('NEW_PRIMARY_CREDENTIAL_VALUE','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_PRIMARY_CRED_VAL,
Count_Col_d('NEW_MOBILE_PHONE','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_MOBILE_PHONE,
Count_Col_d('NEW_HOUSE_FLAT','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_HOUSE_FLAT,
Count_Col_d('NEW_SECTOR_AREA_HOUSING','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_SECTOR_AREA_HOUSING,
Count_Col_d('NEW_LAST_NAME','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_LAST_NAME,
Count_Col_d('NEW_FIRST_NAME','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)') C_FIRST_NAME,
Count_Col_d('NEW_EMAIL','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_EMAIL,
Count_Col_d('X_ABM_NAME','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_ABM_NAME,
```

```
Count_Col_d('SR_SEV_CD','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION
||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_SR_SEV_CD,
Count_Col_d('SR_PRIO_CD','Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATIO
N||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating||SR_NUM)')
C_SR_PRIO_CD
FROM DUAL
```

5 Script

```
/*********************************************************
Fetch data of customers who have more than one region in dataset
*********************************************************/
Select
 OWNER_ACCNT_ID, Count(distinct RGN)-1
From BB_COMPLAINT_DATA_TBL cd
Where Is_Deleted = 'N' and RGN is not null
Group by OWNER_ACCNT_ID
Having Count(distinct RGN)-1 <> 0;

Select Distinct owner_accnt_id, Rgn
From BB_COMPLAINT_DATA_TBL
Where Is_Deleted = 'N'
and owner_accnt_id in
(  Select
   OWNER_ACCNT_ID
   From BB_COMPLAINT_DATA_TBL cd
   Where Is_Deleted = 'N' and RGN is not null
   Group by OWNER_ACCNT_ID
   Having Count(distinct RGN)-1 <> 0
)
Order By owner_accnt_id;
```

6 Script

```
/*********************************************************
Fetch data of customers who have more than one occupation in dataset
*********************************************************/
  Select
   OWNER_ACCNT_ID, Count(distinct New_Occupation)-1
   From BB_COMPLAINT_DATA_TBL cd
   Where Is_Deleted = 'N' and New_Occupation is not null
   Group by OWNER_ACCNT_ID
   Having Count(distinct New_Occupation)-1 <> 0;
```

```
Select Distinct
owner_accnt_id, New_Occupation
From BB_COMPLAINT_DATA_TBL
Where Is_Deleted = 'N'
and owner_accnt_id in
(  Select
    OWNER_ACCNT_ID
    From BB_COMPLAINT_DATA_TBL cd
    Where Is_Deleted = 'N' and New_Occupation is not null
    Group by OWNER_ACCNT_ID
    Having Count(distinct New_Occupation)-1 <> 0
)
Order By owner_accnt_id;
```

7 Script

```
/*************************************************************************
Fetch data of customers who have more than one delivery method of billing in dataset
*************************************************************************/
  Select
    OWNER_ACCNT_ID, Count(distinct OLD_DILIVERY_METHOD)-1
    From BB_COMPLAINT_DATA_TBL cd
    Where Is_Deleted = 'N' and OLD_DILIVERY_METHOD is not null--not null
    Group by OWNER_ACCNT_ID
    Having Count(distinct OLD_DILIVERY_METHOD)-1 <> 0 ;

Select distinct owner_accnt_id, OLD_DILIVERY_METHOD
From BB_COMPLAINT_DATA_TBL
Where Is_Deleted = 'N'
and OLD_DILIVERY_METHOD is not null
and owner_accnt_id in
(Select
    OWNER_ACCNT_ID
    From BB_COMPLAINT_DATA_TBL cd
    Where Is_Deleted = 'N' and OLD_DILIVERY_METHOD is not null
    Group by OWNER_ACCNT_ID
    Having Count(distinct OLD_DILIVERY_METHOD)-1 <> 0 )
Order By owner_accnt_id;
```

8 Script

```
/*************************************************************************
Calculate the number and %age of Customers who are faced particular nature of Fault, each row
has been treated as one customer
*************************************************************************/
```

```
SELECT Count(1) Total_Customers_Affectee,
 SUM(CASE WHEN Nature_Of_Fault = 'Sync Loss' THEN 1 ELSE 0 END)  "Sync Loss",
 SUM(CASE WHEN Nature_Of_Fault = 'Sync Loss' THEN 1 ELSE 0 END)/Count(1)*100
"%Sync Loss",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent disconnections' THEN 1 ELSE 0 END)
"Frequent disconnections",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent disconnections' THEN 1 ELSE 0
END)/Count(1)*100  "%Frequent disconnections",
 SUM(CASE WHEN Nature_Of_Fault = 'Customer Can't Configure CPE' THEN 1 ELSE 0
END)  "Customer Can't Configure CPE",
 SUM(CASE WHEN Nature_Of_Fault = 'Customer Can't Configure CPE' THEN 1 ELSE 0
END)/Count(1)*100  "%Customer Can't Config CPE",
 SUM(CASE WHEN Nature_Of_Fault = 'No Browsing' THEN 1 ELSE 0 END)  "No
Browsing",
 SUM(CASE WHEN Nature_Of_Fault = 'No Browsing' THEN 1 ELSE 0 END)/Count(1)*100
"%No Browsing",
 SUM(CASE WHEN Nature_Of_Fault = 'Authentication Problem' THEN 1 ELSE 0 END)
"Authentication Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Authentication Problem' THEN 1 ELSE 0
END)/Count(1)*100  "%Authentication Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Port Upgradation/Down Gradation' THEN 1 ELSE 0
END)  "Port Up/Down Gradation",
 SUM(CASE WHEN Nature_Of_Fault = 'Port Upgradation/Down Gradation' THEN 1 ELSE 0
END)/Count(1)*100  "%Port Up/Down Gradation",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Low Line Parameters' THEN 1
ELSE 0 END)  "Slow Browsing-Low Line Params",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Low Line Parameters' THEN 1
ELSE 0 END)/Count(1)*100  "%Slow Browsng-Low Line Params",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Network Issue' THEN 1 ELSE 0
END)  "Slow Browsing - Network Issue",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Network Issue' THEN 1 ELSE 0
END)/Count(1)*100  "%Slow Browsing - Network Issue",
 SUM(CASE WHEN Nature_Of_Fault = 'Modem and Splitter Problem' THEN 1 ELSE 0 END)
"Modem and Splitter Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Modem and Splitter Problem' THEN 1 ELSE 0
END)/Count(1)*100  "%Modem and Splitter Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent Disconnections but line Parameters ok'
THEN 1 ELSE 0 END)  "Freqt Discon but line Param ok",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent Disconnections but line Parameters ok'
THEN 1 ELSE 0 END)/Count(1)*100  "%Freq Discon but line Param ok",
 SUM(CASE WHEN Nature_Of_Fault = 'Service Restoration' THEN 1 ELSE 0 END)  "Service
Restoration",
 SUM(CASE WHEN Nature_Of_Fault = 'Service Restoration' THEN 1 ELSE 0
END)/Count(1)*100  "%Service Restoration",
 SUM(CASE WHEN Nature_Of_Fault = 'Data Rate Issue' THEN 1 ELSE 0 END)  "Data Rate
Issue",
```

```
 SUM(CASE WHEN Nature_Of_Fault = 'Data Rate Issue' THEN 1 ELSE 0 END)/Count(1)*100
"%Data Rate Issue"
FROM ( Select Nature_Of_Fault
      From BB_COMPLAINT_DATA_TBL cd
      WHERE Is_Deleted = 'N' AND Nature_Of_Fault is not null
      AND Nature_Of_Fault IN ('Sync Loss', 'Frequent disconnections', 'Customer Can't
Configure CPE', 'No Browsing',
                        'Authentication Problem', 'Port Upgradation/Down Gradation', 'Slow
Browsing - Low Line Parameters',
                        'Slow Browsing - Network Issue', 'Modem and Splitter Problem', 'Frequent
Disconnections but line Parameters ok',
                        'Service Restoration', 'Data Rate Issue')
    );
```

9 Script

```
/*************************************************************************
Calculate the number and %age of Customers who have faced particular nature of Fault,
customer identified by identifier 'Owner_Accnt_ID'
*************************************************************************/
SELECT Count(Owner_Accnt_ID) Total_Customers_Affectee,
 SUM(CASE WHEN Nature_Of_Fault = 'Sync Loss' THEN 1 ELSE 0 END)  "Sync Loss",
 SUM(CASE WHEN Nature_Of_Fault = 'Sync Loss' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Sync Loss",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent disconnections' THEN 1 ELSE 0 END)
"Frequent disconnections",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent disconnections' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Frequent disconnections",
 SUM(CASE WHEN Nature_Of_Fault = 'Customer Can't Configure CPE' THEN 1 ELSE 0
END)  "Customer Can't Configure CPE",
 SUM(CASE WHEN Nature_Of_Fault = 'Customer Can't Configure CPE' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Customer Can't Config CPE",
 SUM(CASE WHEN Nature_Of_Fault = 'No Browsing' THEN 1 ELSE 0 END)  "No
Browsing",
 SUM(CASE WHEN Nature_Of_Fault = 'No Browsing' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%No Browsing",
 SUM(CASE WHEN Nature_Of_Fault = 'Authentication Problem' THEN 1 ELSE 0 END)
"Authentication Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Authentication Problem' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Authentication Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Port Upgradation/Down Gradation' THEN 1 ELSE 0
END)  "Port Up/Down Gradation",
 SUM(CASE WHEN Nature_Of_Fault = 'Port Upgradation/Down Gradation' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Port Up/Down Gradation",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Low Line Parameters' THEN 1
ELSE 0 END)  "Slow Browsing-Low Line Params",
```

```
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Low Line Parameters' THEN 1
ELSE 0 END)/Count(Owner_Accnt_ID)*100  "%Slow Browsng-Low Line Params",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Network Issue' THEN 1 ELSE 0
END)  "Slow Browsing - Network Issue",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Network Issue' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Slow Browsing - Network Issue",
 SUM(CASE WHEN Nature_Of_Fault = 'Modem and Splitter Problem' THEN 1 ELSE 0 END)
"Modem and Splitter Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Modem and Splitter Problem' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Modem and Splitter Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent Disconnections but line Parameters ok'
THEN 1 ELSE 0 END)  "Freqt Discon but line Param ok",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent Disconnections but line Parameters ok'
THEN 1 ELSE 0 END)/Count(Owner_Accnt_ID)*100  "%Freq Discon but line Param ok",
 SUM(CASE WHEN Nature_Of_Fault = 'Service Restoration' THEN 1 ELSE 0 END)  "Service
Restoration",
 SUM(CASE WHEN Nature_Of_Fault = 'Service Restoration' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Service Restoration",
 SUM(CASE WHEN Nature_Of_Fault = 'Data Rate Issue' THEN 1 ELSE 0 END)  "Data Rate
Issue",
 SUM(CASE WHEN Nature_Of_Fault = 'Data Rate Issue' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Data Rate Issue"
FROM ( Select OWNER_ACCNT_ID, Nature_Of_Fault
      From BB_COMPLAINT_DATA_TBL cd
      WHERE Is_Deleted = 'N' AND Nature_Of_Fault is not null
      AND Nature_Of_Fault IN ('Sync Loss', 'Frequent disconnections', 'Customer Can't
Configure CPE', 'No Browsing',
                        'Authentication Problem', 'Port Upgradation/Down Gradation', 'Slow
Browsing - Low Line Parameters',
                        'Slow Browsing - Network Issue', 'Modem and Splitter Problem', 'Frequent
Disconnections but line Parameters ok',
                        'Service Restoration', 'Data Rate Issue')
      Group By OWNER_ACCNT_ID, Nature_Of_Fault
    );
```

10 Script

```
/*****************************************************************************
Calculate the number and %age of Customers who have faced particular nature of Fault,
customer identified by the key generated by combination of OWNER_ACCNT_ID,
NEW_OCCUPATION, Rgn, OLD_DILIVERY_METHOD, Data_Rate, Exchange and
Customer_Rating
*****************************************************************************/
SELECT Count(Owner_Accnt_ID) Total_Customers_Affectee,
 SUM(CASE WHEN Nature_Of_Fault = 'Sync Loss' THEN 1 ELSE 0 END)  "Sync Loss",
```

```
 SUM(CASE WHEN Nature_Of_Fault = 'Sync Loss' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Sync Loss",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent disconnections' THEN 1 ELSE 0 END)
"Frequent disconnections",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent disconnections' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Frequent disconnections",
 SUM(CASE WHEN Nature_Of_Fault = 'Customer Can't Configure CPE' THEN 1 ELSE 0
END)  "Customer Can't Configure CPE",
 SUM(CASE WHEN Nature_Of_Fault = 'Customer Can't Configure CPE' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Customer Can't Config CPE",
 SUM(CASE WHEN Nature_Of_Fault = 'No Browsing' THEN 1 ELSE 0 END)  "No
Browsing",
 SUM(CASE WHEN Nature_Of_Fault = 'No Browsing' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%No Browsing",
 SUM(CASE WHEN Nature_Of_Fault = 'Authentication Problem' THEN 1 ELSE 0 END)
"Authentication Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Authentication Problem' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Authentication Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Port Upgradation/Down Gradation' THEN 1 ELSE 0
END)  "Port Up/Down Gradation",
 SUM(CASE WHEN Nature_Of_Fault = 'Port Upgradation/Down Gradation' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Port Up/Down Gradation",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Low Line Parameters' THEN 1
ELSE 0 END)  "Slow Browsing-Low Line Params",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Low Line Parameters' THEN 1
ELSE 0 END)/Count(Owner_Accnt_ID)*100  "%Slow Browsng-Low Line Params",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Network Issue' THEN 1 ELSE 0
END)  "Slow Browsing - Network Issue",
 SUM(CASE WHEN Nature_Of_Fault = 'Slow Browsing - Network Issue' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Slow Browsing - Network Issue",
 SUM(CASE WHEN Nature_Of_Fault = 'Modem and Splitter Problem' THEN 1 ELSE 0 END)
"Modem and Splitter Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Modem and Splitter Problem' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Modem and Splitter Problem",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent Disconnections but line Parameters ok'
THEN 1 ELSE 0 END)  "Freqt Discon but line Param ok",
 SUM(CASE WHEN Nature_Of_Fault = 'Frequent Disconnections but line Parameters ok'
THEN 1 ELSE 0 END)/Count(Owner_Accnt_ID)*100  "%Freq Discon but line Param ok",
 SUM(CASE WHEN Nature_Of_Fault = 'Service Restoration' THEN 1 ELSE 0 END)  "Service
Restoration",
 SUM(CASE WHEN Nature_Of_Fault = 'Service Restoration' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Service Restoration",
 SUM(CASE WHEN Nature_Of_Fault = 'Data Rate Issue' THEN 1 ELSE 0 END)  "Data Rate
Issue",
 SUM(CASE WHEN Nature_Of_Fault = 'Data Rate Issue' THEN 1 ELSE 0
END)/Count(Owner_Accnt_ID)*100  "%Data Rate Issue"
```

```
FROM ( Select
Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating) OWNER_ACCNT_ID,  Nature_Of_Fault
      From BB_COMPLAINT_DATA_TBL cd
      WHERE Is_Deleted = 'N' AND Nature_Of_Fault is not null
      AND Nature_Of_Fault in ('Sync Loss', 'Frequent disconnections', 'Customer Can't Configure CPE', 'No Browsing',
                  'Authentication Problem', 'Port Upgradation/Down Gradation', 'Slow Browsing - Low Line Parameters',
                  'Slow Browsing - Network Issue', 'Modem and Splitter Problem', 'Frequent Disconnections but line Parameters ok',   'Service Restoration', 'Data Rate Issue')
      Group By
Sys.CONV_HASH(OWNER_ACCNT_ID||NEW_OCCUPATION||Rgn||OLD_DILIVERY_METHOD||Data_Rate||Exchange||Customer_Rating), Nature_Of_Fault
);
```

11 Script

```
/*****************************************************************************
Calculate the number of patients and respective readmission count against below mentioned
four Categories for further analysis of readmission probability of patients, each row is considered one encounter of patient in the Hospital:
HbA1c - No test was performed
HbA1c - Result was high and the diabetic medication was changed
HbA1c - Result was high but the diabetic medication was not changed
HbA1c - Normal result of the test
*****************************************************************************/
Select
 SUM(CASE WHEN A1Cresult = 'None' THEN 1 ELSE 0 END) "No test was performed",
 SUM(CASE WHEN A1Cresult = 'None' AND readmitted Like '<30' THEN 1 ELSE 0 END) "NoTestWasPerformedAndReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' THEN 1 ELSE 0 END)  "ResultWasHiDiabMedChanged",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' AND readmitted Like '<30' THEN 1 ELSE 0 END)  "ResultWasHiDiabMedChangedReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' THEN 1 ELSE 0 END) "ResHiDiabMedWasNotChned",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' AND readmitted Like '<30' THEN 1 ELSE 0 END) "ResHiDiabMedWasNotChngdReadm",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') THEN 1 ELSE 0 END) AS "Normal result of test",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') AND readmitted Like '<30' THEN 1 ELSE 0 END) AS "NorResultoftestAndReadm"
From
( Select row_number() over (partition by Patient_Nbr order by number_inpatient, encounter_id asc) rownumber,
  DD.*
```

```
  From Sys.Diabetic_Data_Tbl dd
  Where discharge_disposition_id not in (11,13,14)
);
```

12 Script

```
/*************************************************************************
Calculate the number of patients and respective readmission count against four Categories (refer
11 Script) for further analysis of readmission probability of patients, patient is identified by
Patient No to take only one instance per patient.
*hash key generated by using built-in function
*************************************************************************/
Select
 SUM(CASE WHEN A1Cresult = 'None' THEN 1 ELSE 0 END) "No test was performed",
 SUM(CASE WHEN A1Cresult = 'None' AND readmitted Like '<30' THEN 1 ELSE 0 END)
"NoTestWasPerformedAndReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' THEN 1
ELSE 0 END)  "ResultWasHiDiabMedChanged",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' AND
readmitted Like '<30' THEN 1 ELSE 0 END)  "ResultWasHiDiabMedChangedReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' THEN 1 ELSE 0 END)
"ResHiDiabMedWasNotChned",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' AND readmitted Like '<30' THEN 1
ELSE 0 END) "ResHiDiabMedWasNotChngdReadm",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') THEN 1 ELSE 0 END) AS
"Normal result of test",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') AND readmitted Like '<30'
THEN 1 ELSE 0 END) AS "NorResultoftestAndReadm"
From
( Select row_number() over (partition by Sys.CONV_HASH(Patient_Nbr) order by
number_inpatient, encounter_id asc) rownumber,
  readmitted, A1Cresult, change, diabetesMed
  From Sys.Diabetic_Data_Tbl dd
  Where discharge_disposition_id not in (11,13,14)
) Where rownumber = 1;
```

13 Script

```
/*************************************************************************
Calculate the number of patients and respective readmission count against four Categories (refer
11 Script) for further analysis of readmission probability of patients, patient is identified by
distinct combination of Patient No and Weight to take only distinct instance of patient and
weight
*hash key generated by using built-in function with the combination of Patient No and Weight
*************************************************************************/
Select
 SUM(CASE WHEN A1Cresult = 'None' THEN 1 ELSE 0 END) "No test was performed",
```

```
 SUM(CASE WHEN A1Cresult = 'None' AND readmitted Like '<30' THEN 1 ELSE 0 END)
"NoTestWasPerformedAndReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' THEN 1
ELSE 0 END)  "ResultWasHiDiabMedChanged",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' AND
readmitted Like '<30' THEN 1 ELSE 0 END)  "ResultWasHiDiabMedChangedReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' THEN 1 ELSE 0 END)
"ResHiDiabMedWasNotChned",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' AND readmitted Like '<30' THEN 1
ELSE 0 END) "ResHiDiabMedWasNotChngdReadm",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') THEN 1 ELSE 0 END) AS
"Normal result of test",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') AND readmitted Like '<30'
THEN 1 ELSE 0 END) AS "NorResultoftestAndReadm"
From
( Select row_number() over (partition by Sys.CONV_HASH(Patient_Nbr||weight) order by
number_inpatient, encounter_id asc) rownumber,
  DD.*
  From Sys.Diabetic_Data_Tbl dd
  Where discharge_disposition_id not in (11,13,14)
) Where rownumber = 1;
```

14 Script

```
/*****************************************************************************
Calculate the number of patients and respective readmission count against four Categories (refer
11 Script) for further analysis of readmission probability of patients, patient is identified by
distinct combination of Patient No and Age to take only distinct instance per patient and Age
*hash key generated by using built-in function with the combination of Patient No and Age
*****************************************************************************/
Select
 SUM(CASE WHEN A1Cresult = 'None' THEN 1 ELSE 0 END) "No test was performed",
 SUM(CASE WHEN A1Cresult = 'None' AND readmitted Like '<30' THEN 1 ELSE 0 END)
"NoTestWasPerformedAndReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' THEN 1
ELSE 0 END)  "ResultWasHiDiabMedChanged",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' AND
readmitted Like '<30' THEN 1 ELSE 0 END)  "ResultWasHiDiabMedChangedReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' THEN 1 ELSE 0 END)
"ResHiDiabMedWasNotChned",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' AND readmitted Like '<30' THEN 1
ELSE 0 END) "ResHiDiabMedWasNotChngdReadm",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') THEN 1 ELSE 0 END) AS
"Normal result of test",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') AND readmitted Like '<30'
THEN 1 ELSE 0 END) AS "NorResultoftestAndReadm"
From
```

```
( Select row_number() over (partition by Sys.CONV_HASH(Patient_Nbr||Age) order by
number_inpatient, encounter_id asc) rownumber,
  DD.*
  From Sys.Diabetic_Data_Tbl dd
  Where discharge_disposition_id not in (11,13,14)
) Where rownumber = 1;
```

15 Script

```
/*****************************************************************************
Calculate the number of patients and respective readmission count against four Categories (refer
11 Script) for further analysis of readmission probability of patients, patient is identified by
distinct combination of Patient No, Weight and Age to take only distinct instance per patient,
weight and Age
*hash key generated by using built-in function with the combination of Patient No, Weight and
Age
*****************************************************************************/
Select
 SUM(CASE WHEN A1Cresult = 'None' THEN 1 ELSE 0 END) "No test was performed",
 SUM(CASE WHEN A1Cresult = 'None' AND readmitted Like '<30' THEN 1 ELSE 0 END)
"NoTestWasPerformedAndReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' THEN 1
ELSE 0 END)  "ResultWasHiDiabMedChanged",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'Ch' and   diabetesMed = 'Yes' AND
readmitted Like '<30' THEN 1 ELSE 0 END)  "ResultWasHiDiabMedChangedReadm",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' THEN 1 ELSE 0 END)
"ResHiDiabMedWasNotChned",
 SUM(CASE WHEN A1Cresult = '>8' AND change = 'No' AND readmitted Like '<30' THEN 1
ELSE 0 END) "ResHiDiabMedWasNotChngdReadm",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') THEN 1 ELSE 0 END) AS
"Normal result of test",
 SUM(CASE WHEN (A1Cresult = 'Norm' OR A1Cresult = '>7') AND readmitted Like '<30'
THEN 1 ELSE 0 END) AS "NorResultoftestAndReadm"
From
( Select row_number() over (partition by Sys.CONV_HASH(Patient_Nbr||Weight||Age) order by
number_inpatient, encounter_id asc) rownumber,
  DD.*   From Sys.Diabetic_Data_Tbl dd
  Where discharge_disposition_id not in (11,13,14)
) Where rownumber = 1;
```

References

[1] JIAN WANG, YONGCHENG LUO, YAN ZHAO AND JIAJIN LE, "A Survey on Privacy Preserving Data Mining," In Proc. IEEE. First International Workshop on Database Technology and Applications. Conf., 2009, pp. 111_114.

[2] KATO MIVULE, "Utilizing Noise Addition for Data Privacy, An overview"

[3] MANISH SHANNAL, ATUL CHAUDHAR, MANISH MATHURIA, SHALINI CHAUDHAR, SANTOSH KUMAR, In Proc. 2014 International Conference on Signal Propagation and Computer Technology (ICSPCT), pp. 244-249.

[4] E. POOVAMMAL, M. PONNAVAIKKO , "Task Independent Privacy Preserving Data Mining on Medical Dataset" In Proc. 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, pp. 814-818.

[5] ANIL PRATAP SINGH, ABHISHEK MATHUR, "A Chaotic Based approach for Privacy Preserving Data Mining Applications with Multilevel Trust " In Proc. 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE) , pp. 792-797.

[6] BEATA STRACK, JONATHAN P. DESHAZO, CHRIS GENNINGS, JUAN L. OLMO, SEBASTIAN VENTURA, KRZYSZTOF J. CIOS, AND JOHN N. CLORE, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records" In Proc. BioMed Research International, Volume 2014, Article ID 781670, 11 pages, Hindawi Publishing Corporation, http://dx.doi.org/10.1155/2014/781670

[7] R. AGRAWAL, R. SRIKANT, "Privacy-Preserving Data Mining", In Proc. SIGMOD, pp. 439–450, 2000.

[8] V. S. VERYKIOS, E. BERTINO, "et. al. State-of-the-art in Privacy Preserving Data Mining", In SIGMOD Record, 33(1), pp. 50–57, 2004.

[9] L. BRANKOVIC AND V. ESTIVILL-CASTRO, "Privacy issues in knowledge discovery and data mining," In Proc. Austral. Inst. Comput. Ethics Conf., 1999, pp. 89_99.

[11] LEI XU, CHUNXIAO JIANG, JIAN WANG, JIAN YUAN, AND YONG REN, "Information Security in big Data: Privacy and Data Mining," In Proc. IEEE Access. The journal for rapid open access publishing Volume 2., 2014, pp. 1149_1176.