

Jaypee University of Information Technology
Waknaghat, Distt. Solan (H.P.)

Learning Resource Center

CLASS NUM:

BOOK NUM.:

ACCESSION NO.: SP09055/SP0913056

This book was issued is overdue due on the date stamped below. If the book is kept over due, a fine will be charged as per the library rules.

Due Date	Due Date	Due Date

Learning Resource Center

ASSOCIATION MINING BASED STUDY FOR IDENTIFICATION OF CLINICAL PARAMETERS AKIN TO DIABETES

BY:

**SHRUTI KAPIL (091507)
SHAGUN THAKUR (091520)**

Under the supervision of:

MR.DIPANKAR SENGUPTA



MAY-2013

Submitted in partial fulfillment of the Degree of Bachelor of Technology

**DEPARTMENT OF
BIOTECHNOLOGY & BIOINFORMATICS
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
WAKNAGHAT**



**ASSOCIATION MINING BASED STUDY FOR
IDENTIFICATION OF CLINICAL PARAMETERS AKIN
TO DIABETES**

BY:

**SHRUTI KAPIL (091507)
SHAGUN THAKUR (091520)**

Under the supervision of:

MR.DIPANKAR SENGUPTA



MAY-2013

Submitted in partial fulfillment of the Degree of Bachelor of Technology

**DEPARTMENT OF
BIOTECHNOLOGY & BIOINFORMATICS
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
WAKNAGHAT**

TABLE OF CONTENTS

<u>TOPIC</u>	<u>PAGE.NO</u>
CERTIFICATE	IV
ACKNOWLEDGEMENT	V
ABSTRACT	VI
LIST OF FIGURES AND TABLES	VII
CHAPTER 1: INTRODUCTION	1-8
1.1 Diabetes mellitus	3
1.1.1 Types of Diabetes mellitus	3-4
1.1.2 Symptoms of Diabetes mellitus	4
1.1.3 Global scenario	5
1.1.4 Scenario in India	5
1.1.5 Diagnosis of Diabetes mellitus	5
1.1.6 Treatment of Diabetes mellitus	7
1.4 Brief Introduction of the Project	7
1.4.1 Objective	8
1.4.2 Project Plan	8
CHAPTER 2: DATAWAREHOUSE DEVELOPMENT	9-22
2.1 Data Warehousing	9
2.2 Data Marts	10
2.3 Tools and Techniques	11
2.4 Design Methodology	14
2.5 Methodology used	16

CHAPTER 3: DATA MINING – A Knowledge discovery process	23-28
3.1 Introduction	23
3.2 Data Mining Process	23
3.3 Types of Data Mining	24
3.4 Association Rule Mining	24
3.4.1 Process	26
3.4.2 Result	27
FUTURE WORK	29
APPENDIX	30-31
SQL QUERIES	30
REFERENCES	32-34
BRIEF PROFILE OF STUDENTS	35

CERTIFICATE

This is to certify that the thesis entitled “ASSOCIATION MINING BASED STUDY FOR IDENTIFICATION OF CLINICAL PARAMETERS AKIN TO DIABETES” submitted by SHRUTI KAPIL (091507) and SHAGUN THAKUR (091520) to the Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Wagnaghat in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Bioinformatics is a record of bona fide research work carried out by them under our supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Date: ...27/05/2013



Dipankar Sengupta

Associate Lecturer

Dept. of Biotechnology and Bioinformatics

Jaypee University of Information Technology

Wagnaghat, Solan (H.P.)

ACKNOWLEDGEMENT

As we conclude our project, we have many people to thank for all the help, guidance and support they lent us, throughout the course of our endeavor.

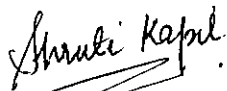
First and foremost, we would like to thank **Mr. Dipankar Sengupta**, our Project Guide, for his guidance, encouragement and support. He has always encouraged us to put in our best efforts and deliver a quality and professional output. His methodologies of making the system strong from inside has taught us that output is not the END of project; it's the learning that is important. We really thank him for his time & efforts.

We would like to express gratefulness to Ms. Manya and PHD scholar Ms. Charu Suri for their invaluable guidance and great support because this project was not possible without them.

We would like to pay our most sincere thanks to **Prof. R.S.Chauhan**, Head of Department, Department of Biotechnology and Bioinformatics, for providing us with opportunities and facilities to carry out the project.

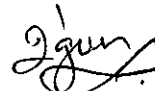
We would also like to thank doctors from **Indra Gandhi Medical College (IGMC), Shimla and Solan Hospital** who have helped us by not only providing the data but also spending enough time to answer out our queries and playing a major role in successful completion of our project.

Apart from these, countless events, countless people and several incidents have made a contribution to these projects which are indescribable. We again express our gratitude to them. We are indebted to all those who provided their reviews and suggestions for improvising our project and extend our apologies to any one whom we have failed to recognize for their contribution.



Shruti Kapil (091507)

Date: 27/05/2013



Shagun Thakur (091520)

Date: 27/05/2013

ABSTRACT

About a decade back Ralph Kimball identified domains which can be associated with the Business Intelligence model for a hospital care system and can be associated with a patient's life cycle corresponding to billing aspect. However much thought has not been given towards clinical data being generated during this life cycle which can be used for correlation and predictive based researches.

Vast quantities of data are generated through health care processes. While technological advancements in the form of computer-based patient record software and personal computer hardware are making the collection of and access to health care data more manageable, few tools exist to evaluate and analyze this clinical data after it has been captured and stored.

Evaluation of stored clinical data may lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management. Techniques are needed to search large quantities of clinical data for these patterns and relationships.

In our project we tried to achieve a similar task at Jaypee University of Information Technology which addressed the above mentioned challenges. We developed a Data Warehouse for efficient integration, access, storage and retrieval of vast data, that includes diagnostic test results of patient suffering from Diabetes. Then we further performed Data Mining via Association Mining technique for identification of relevant clinical parameters related to diabetes which will be used to carry out predictive studies, which would help in creation of a predictive model and a tool for same.

LIST OF FIGURES AND TABLES

Figure 1: Project Planning

Figure 2: Datawarehouse Architecture

Figure 3: Representation for flow of data in Kimball model

Figure 4: Representation for flow of data in Inmon model

Figure 5: Dimensional Model (Star Schema Design).

Figure 6: Snapshot of a Transformation of Disease excel file into table

Figure 7: Snapshot of a Transformation of Diagnosis excel file into table

Figure 8: Snapshot of a Transformation of Patient excel file into table.

Figure 9: Snapshot of a Transformation of Date excel file into table.

Figure 10: Snapshot of a Transformation of Diagnosis csv file into table.

Figure 11: Snapshot of a Transformation staging_disease' to table 'DIM_DISEASE'.

Figure 12: Snapshot of a Transformation Processing of data from table 'staging_diagnosis' to table 'DIM_DIAGNOSIS'.

Figure 13: Snapshot of a Transformation Processing of data from table 'staging_patient' to table 'DIM_PATIENT'.

Figure 14: Snapshot of a Transformation Processing of data from table 'staging_date' to table 'DIM_DATE'.

Figure 15: Snapshot of a Transformation Processing of data table 'staging_diagnosis_level' to table 'FACT_DIAGNOSIS_LEVEL'.

Figure 16: Data Mining Process

Figure: 17 Association mining rules

Table: 1 Example of association rules.

Table: 2 Final result set of 12 association rules.

CHAPTER 1

INTRODUCTION

As the population is increasing, there is a rising tide of data in various sectors of Clinical Sciences like healthcare, clinical trials, genomics etc. The huge amount of data generated is not being properly stored and managed by the hospitals /research institute. Medical practitioners usually make diagnostic decisions and treatment recommendations based on history, medical imaging, lab results and other text or multimedia records of patients. This data is used only for diagnosis purpose and afterward considered as junk. There is a wealth of knowledge to be gained from this data. Healthcare is a major focus for big data companies and data scientists because there's large amount of data involved and the problems associated with management & analysis. The right analytic tools could end up saving lives or saving billions of dollars in an industry where just about everyone agrees that costs are out of control. [1, 2]

Along the development of information technology since 1990s, healthcare providers realised that the information could transfer significant benefits to improve their services by computerised cases and data, for instance of gaining the information for directing patient care and assessing the best patient care for specific clinical conditions. Information technology (IT) has the potential to improve the quality, safety, and efficiency of health care. So, we are trying to meld machine learning, IT and health care for their more efficient working in the field of healthcare.[3]The major challenge is storing the data, better management and analysis of this data. To overcome this challenge a new and emerging field of bioinformatics comes into play "Healthcare Informatics" that will be utilized to analyze the healthcare data.

Healthcare informatics combines the fields of information technology and health to develop the systems required to administer the expansion of information and improve medical processes. We are dealing more precisely with data storage and correlation aspects of clinical informatics. Health Informatics is a rapidly growing field. With the aging population on the rise in developed countries and the increasing cost of healthcare, governments and large health organizations are becoming very interested in the potential of Health Informatics to save time, money, and human lives. Health informatics allows doctors to have faster access to more relevant information, and thus make more optimal decisions.[4,5]

Healthcare informatics is a field that is constantly striving to make information more accessible in the simplest way. Since there are so many papers and files to process at any medical setting, an efficient system for keeping track of it all is required. Medical informatics becomes a way to

organize and process the information. Examples of information stored in health informatics include disease research, patient backgrounds, statistics and treatment plans.

The information which hospitals consider to be junk data might be as important and meaningful as drug/medicine. It's all about fetching useful information from this raw data. This information may help a patient in his own case as well as when studied on a larger scale, this information can help in prevention proactive treatments and an early detection of certain life threatening diseases at population level.[3,5]

In this project, we are addressing how clinical data warehousing in combination with data mining can help clinical, research and educational aspects of Diabetes. Applying data mining techniques on the centralized database will give doctors analytical and predictive tools that go beyond what is apparent from the surface of the data. With the aim of providing substantial solution to the problems concerning clinical data management and analysis, we sought to create a clinical data warehouse that would store the data and would also make data handling much easier and data querying much efficient. Our focus here is diabetes, as its one of the most prevalent disease among the population in India and across the globe. [8]

In this study we are performing association based study to deduce rules which can be used to akin occurrence of diabetes with other clinical parameters & are related to healthy functioning of organs like kidney, liver, etc. The analysis of integrated data from the warehouse could greatly help to extract useful information for the assessment of health care delivery process. The repository stores healthcare data, mainly concerning the data that is related to blood profiling reports of the patients. analysis of such healthcare datawarehouse, which integrate clinical data, could greatly help to gain a deeper insight into the health condition of the population and to extract useful information for the assessment of health care delivery process. The analysis helped us to uncover some new meaningful information from this seemingly unrelated clinical parameters in the form of association rule. These association rules will be used to build a predictive model that will predict the state of given clinical parameter and it could help the medical practitioners in their efficient diagnosis.

1.1 DIABETES MELLITUS

Diabetes mellitus is a group of metabolic diseases in which a person has high blood sugar, either because the pancreas does not produce enough insulin, or because cells do not respond to the insulin that is produced. People with diabetes mellitus have high levels of glucose in their blood because of a lack of insulin or resistance to the effects of insulin.

1.1.1 Types of Diabetes mellitus:-

There are three main types of diabetes mellitus (DM).

Type 1 diabetes

In type 1 diabetes, beta cells in the pancreas that make insulin are destroyed by the immune system, causing severe, usually complete deficiency of insulin. This form of diabetes is more common in childhood and young adulthood but can occur at any age. People with type 1 diabetes must receive daily insulin injections to sustain life, and must do regular finger prick blood glucose tests to monitor their diabetes. Progressive improvements in management and introduction of new technology have resulted in reduced rates of complications and greatly improved life expectancy. This was also referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes".

Type 2 diabetes

People with type 2 diabetes are resistant to the action of insulin and also have relative insulin deficiency because of progressive failure of the pancreatic beta cells.

An inherited susceptibility to Type 2 diabetes is aggravated by abdominal obesity. Type 2 diabetes usually occurs in middle aged and elderly people, but is becoming more common in younger adults, adolescents and children, particularly in Aboriginal populations.

Type 2 diabetes and diabetic complications are often asymptomatic, making early diagnosis difficult. In Aboriginal and other high risk, remote and under-resourced groups, diabetes is often undiagnosed until advanced complications have developed. This was also referred to as non insulin-dependent diabetes mellitus (NIDDM) or "adult-onset diabetes".

Gestational diabetes

In women predisposed by genetic factors and abdominal obesity, hormonal changes during pregnancy increase blood glucose levels, resulting in increased rates of congenital abnormalities, foetal weight gain and perinatal complications. Gestational diabetes is usually asymptomatic, and is detected by screening tests. Treatment consists of dietary modification and, in some cases, insulin injections. Up to 50% of women who have had gestational diabetes subsequently develop type 2 diabetes.

Other forms of diabetes mellitus include congenital diabetes, which is due to genetic defects of insulin secretion, cystic fibrosis-related diabetes, steroid diabetes induced by high doses of glucocorticoids, and several forms of monogenic diabetes. [9]

1.1.2 Symptoms of Diabetes Mellitus

Symptoms of Type 1 Diabetes Mellitus:

- Above average thirst.
- Feeling tired.
- Frequent urination.
- Losing weight.
- Skin infections.
- Genital itchiness.

Symptoms of Type 2 Diabetes Mellitus:

- Feeling tired during the day, particularly after meals.
- Often feeling hungry, particularly if you feel hungry shortly after eating.
- Urinating more often than normal, particular needing to do so during the night.
- Feeling abnormally thirsty.
- Blurring of vision.
- Itching of the skin, particularly itchiness around the genitals.
- Slow healing of cuts or wounds.
- Having regular yeast infections (thrush).
- Having a skin disorder such as psoriasis or acanthosis nigricans.
- Sudden weight loss or loss of muscle mass.[10]

1.1.3 Global Scenario of Diabetes Mellitus

Globally, as of 2010, an estimated 285 million people had diabetes, with type 2 making up about 90% of the cases. Its incidence is increasing rapidly, and by 2030, this number is estimated to almost double. Diabetes mellitus occurs throughout the world, but is more common (especially type 2) in the more developed countries. The greatest increase in prevalence is, however, expected to occur in Asia and Africa, where most patients will probably be found by 2030. The increase in incidence in developing countries follows the trend of urbanization and lifestyle changes, perhaps most importantly a "Western-style" diet. This has suggested an environmental (i.e., dietary) effect, but there is little understanding of the mechanism(s) at present, though there is much speculation, some of it most compellingly presented.

Diabetes is the sixth leading cause of death in Australia (NHPAC 2005) and is the world's fastest growing disease (Diabetes Australia 2007). It is estimated that the number of people with diabetes in Australia will double by 2010, making the prevention of diabetes a national priority (International Diabetes Federation and International Association for the Study of Obesity 2004). Diabetes (diabetes mellitus) is one of Western Australia's most significant health issues. Type 2 diabetes is the most common form, comprising 85 to 90% of those with diabetes. [13]

1.1.4 Scenario in India

India has more diabetics than any other country in the world, according to the International Diabetes Foundation, although more recent data suggest that China has even more. The disease affects more than 50 million Indians - 7.1% of the nation's adults - and kills about 1 million Indians a year. The average age on onset is 42.5 years. The high incidence is attributed to a combination of genetic susceptibility plus adoption of a high-calorie, low-activity lifestyle by India's growing middle class. [8,13]

1.1.5 Diagnosis of Diabetes Mellitus

Diabetes mellitus is characterized by recurrent or persistent hyperglycemia, and is diagnosed by demonstrating any one of the following:

- Fasting plasma glucose level ≥ 7.0 mmol/l (126 mg/dl).
- Plasma glucose ≥ 11.1 mmol/l (200 mg/dL) two hours after a 75 g oral glucose load as in a glucose tolerance test.

- Symptoms of hyperglycemia and casual plasma glucose ≥ 11.1 mmol/l (200 mg/dl).
- Glycated hemoglobin (Hb A1C) $\geq 6.5\%$.

A positive result, in the absence of unequivocal hyperglycemia, should be confirmed by a repeat of any of the above methods on a different day. It is preferable to measure a fasting glucose level because of the ease of measurement and the considerable time commitment of formal glucose tolerance testing, which takes two hours to complete and offers no prognostic advantage over the fasting test. The two fasting glucose measurements above 126 mg/dl (7.0 mmol/l) is considered diagnostic for diabetes mellitus.

People with fasting glucose levels from 110 to 125 mg/dl (6.1 to 6.9 mmol/l) are considered to have impaired fasting glucose. Patients with plasma glucose at or above 140 mg/dL (7.8 mmol/L), but not over 200 mg/dL (11.1 mmol/L), two hours

There are several different types of blood glucose tests.

- **Fasting blood sugar (FBS)** measures blood glucose after you have not eaten for at least 8 hours. It is often the first test done to check for prediabetes and diabetes.
- **2-hour postprandial blood sugar** measures blood glucose exactly 2 hours after you start eating a meal. This is not a test used to diagnose diabetes.
- **Random blood sugar (RBS)** measures blood glucose regardless of when you last ate. Several random measurements may be taken throughout the day. Random testing is useful because glucose levels in healthy people do not vary widely throughout the day. Blood glucose levels that vary widely may mean a problem. This test is also called a casual blood glucose test. Random testing is not used to diagnose diabetes.
- **Oral glucose tolerance test** is used to diagnose prediabetes and diabetes. An oral glucose tolerance test is a series of blood glucose measurements taken after you drink a sweet liquid that contains glucose. This test is commonly used to diagnose diabetes that occurs during pregnancy (gestational diabetes). This test is not commonly used to diagnose diabetes in a person who is not pregnant.
- **Glycohemoglobin A1c** measures how much sugar (glucose) is stuck to red blood cells. This test can be used to diagnose diabetes. It also shows how well your diabetes has been controlled in the last 2 to 3 months and whether your diabetes medicine needs to be changed. The result of your A1c test can be used to estimate your average blood sugar level. This is called your estimated average glucose, or eAG.[11]

1.1.6 Treatment of Diabetes Mellitus

Treatment of type 1 DM

People with type 1 diabetes must take insulin, which is the hormone they lack. Most people take several insulin injections every day or use an insulin pump—a device worn outside the body that pumps insulin through a flexible tube to a small needle inserted under the skin. The pump can be set to give small amounts of short-acting insulin continuously through the day and additional doses before meals.

Treatment of Type 2 Diabetes:

Most people with type 2 diabetes can be treated with diet and exercise and oral antidiabetic agents. Some people may need insulin injections one or more times each day to control their diabetes. Different types of oral antidiabetic agents work in different ways. They can be used alone or in combination with other agents or insulin. The most common types of oral antidiabetic drugs are: Biguanides (metformin), Sulfonylureas (glipizide, glyburide, glimepiride), Thiazolidinediones (pioglitazone, rosiglitazone). DPP-4 inhibitors (sitagliptin, saxagliptin). In addition to oral medications, two injectable antidiabetic agents (exenatide and pramlintide acetate) help control blood sugar levels. These medications help the pancreas produce insulin more efficiently. They may also lead to a decrease in appetite and weight loss. [14]

1.4 Brief Introduction of the Project

With the aim of providing a substantial solution to the problems concerning clinical data management and analysis, we sought to create a clinical datawarehouse that would not only simplify data storage problems but would also make data handling much easier and data querying much efficient. The disease in focus for our study is diabetes mellitus. We are focussing on storing all kind of parameters w.r.t it in a self designed & developed datawarehouse based on the tests conducted on an individual for blood profiling, Kidney based disorders, liver based disorders. We would also try to find out if any association exists between various clinical parameters that are not related with each other based on which we would design a model which would help in early prediction of the disorders.

1.4.1 Objective

- Design and development of a Data Warehouse for Diabetes.
- Association mining among various parameters.
- Model development – for predictive purposes.

1.4.2 Project Plan

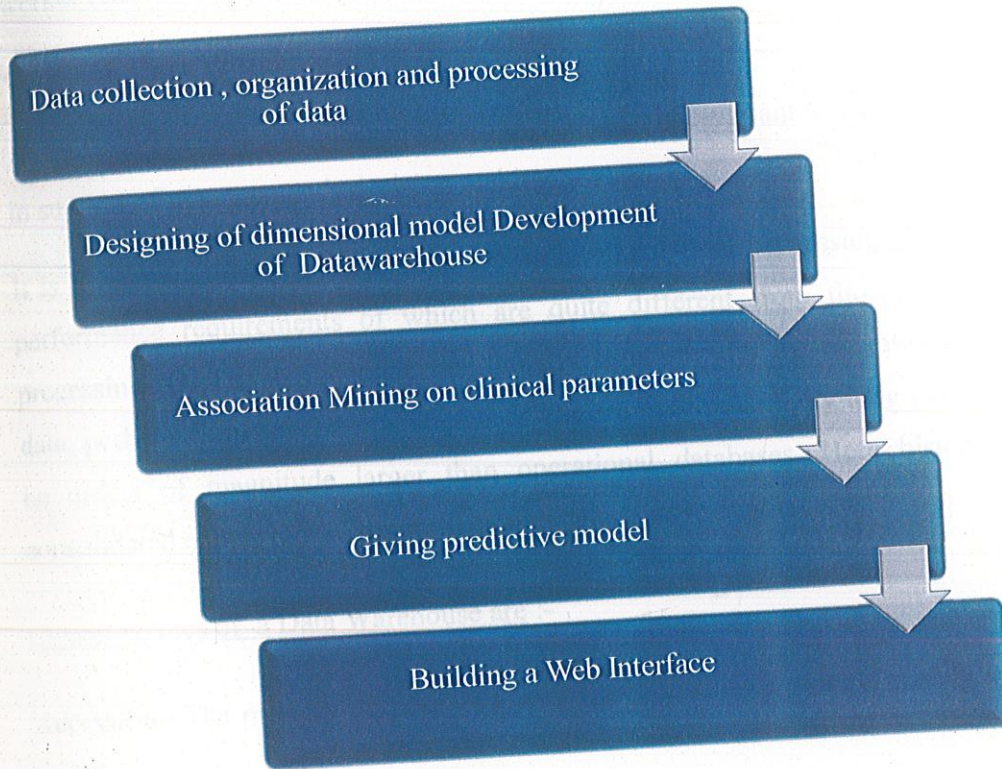


Figure1: Project Planning

CHAPTER 2

DATAWAREHOUSE DEVELOPMENT

2.1 Data Warehousing

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions.

What is a Data Warehouse??

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

It is a relational database that supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases. They store data, perhaps from several operational databases, over potentially long periods of time and tend to be orders of magnitude larger than operational databases. Here historical, summarized and consolidated data is more important than detailed, individual records.

Characteristics of a Data Warehouse are :-

Accessible:-The primary purpose of a data warehouse is to provide readily accessible information to end-users.

Process-Oriented:-It is important to view data warehousing as a process for delivery of information. The maintenance of a data warehouse is ongoing and iterative in nature.

Subject Oriented:-Data is transformed from the application orientation of the operational systems to a normalized, subject oriented structure.

Integrated:- The Data warehouse gathers and "normalizes" data from applications and data stores across the enterprise.

Time Variant:-Data in the warehouse is only meaningful when associated with a time. A date / time indicator often participates in a record's primary key.

Non-volatile:- Data is added to the warehouse, but is rarely updated or deleted. Implies that a date-bounded query will always return the same result.

A data warehouse is populated through a series of following steps :-

- 1) Remove data from the source environment (extract).
- 2) Change the data to have desired warehouse characteristics like subject-orientation and time-variance (transform).
- 3) Place the data into a target environment (load). [15,16]

2.2 Data Mart

A data mart is a subset of the data in the data warehouse, focused on a particular type of data or portion of the information in a warehouse. Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data.

Why a Data Mart ?

- This format is developed to meet the needs of a particular type of analysis and to improve query performance.
- Frequently used data-aggregations or calculations can be preformed when the data mart is loaded.
- Lower cost than implementing a full data warehouse.
- Easy access to frequently needed data.[15]

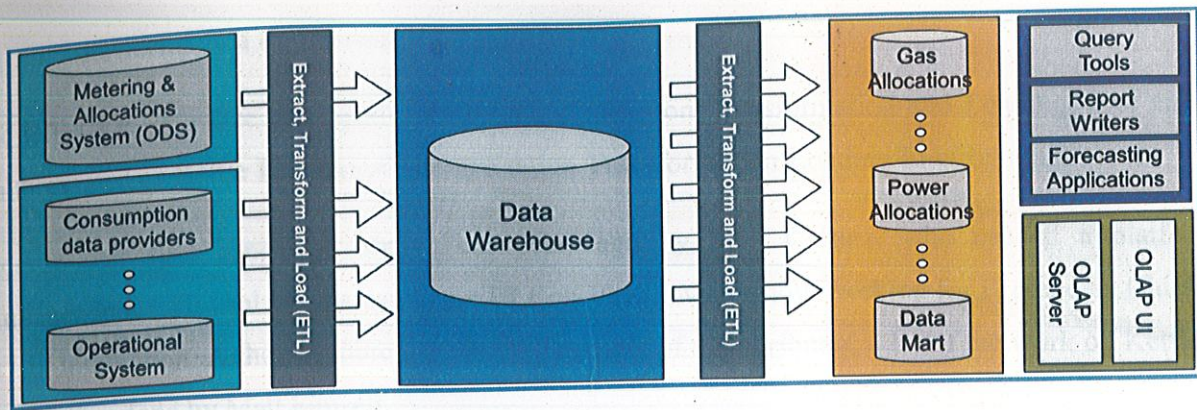


Figure 2: Datawarehouse Architecture

2.3 Tools and Techniques

1. Pentaho Data Integration (Kettle)

Pentaho Data Integration (PDI) delivers powerful Extraction, Transformation and Loading (ETL) capabilities using an innovative, metadata-driven approach. With an intuitive, graphical, drag and drop design environment, and a proven, scalable, standards-based architecture, Pentaho Data Integration is increasingly the choice for organizations over traditional, proprietary ETL or data integration tools.

PDI can also be used for other purposes :-

- Migrating data between applications or databases
- Exporting data from databases to flat files
- Loading data massively into databases
- Data cleansing
- Integrating applications

Pentaho Data Integration's intuitive and rich graphical designer allows you to do exactly what the most skilled code developers can accomplish, in a fraction of the time, and without requiring you to manually code.

Pentaho Data Integration's graphical designer includes :-

- Intuitive, drag and drop designer
- Rich library of pre-built components

- Powerful data transformation mappings

K.E.T.T.LE. is a free and open source ETL (Extraction, transformation and Loading) tool and is abbreviated as **Kettle Extraction Transformation Loading Environment**.

Kettle was first conceived about four years ago by Matt Casters, who needed a platform-independent ETL tool for his work as a BI Consultant. Matt's now working for Pentaho as Chief of Data Integration. Although there's a growing number of a contributor, a lot of the work on Kettle is still being done by Matt himself.

Being an ETL tool, Kettle is an environment that's designed to:

- Collect data from a variety of sources (extraction).
- Move and modify data (transport and transform) while cleansing, denormalizing, aggregating and enriching it in the process.
- Frequently (typically on a daily basis) store data (loading) in the final target destination, which is usually a large, dimensionally modelled database called a data warehouse

Kettle Architecture

Kettle is built with the java programming language. It consists of four distinct applications.

Spoon-is a graphically oriented end-user tool to model the flow of data from input through transformation to output. One such model is also called a *transformation*.

Pan-is a command line tool that executes transformations modelled with Spoon.

Chef-is a graphically oriented end-user tool used to model *jobs*. Jobs consist of job entries such as transformations, FTP downloads etc. that are placed in a flow of control.

Kitchen-is a command line tool used to execute jobs created with Chef. [17]

2. My SQL5.0.19 (RDBMS package)

MySQL is the world's most popular open source relational database management system (RDBMS) as of 2008 that run as a server providing multi-user access to a number of databases. With its superior speed, reliability, and ease of use, MySQL has become the preferred choice for Web.It

eliminates the major problems associated with downtime, maintenance and administration for modern, online applications.

It is a key part of LAMP (Linux, Apache, MySQL, PHP / Perl / Python), the fast-growing open source enterprise software stack. More and more companies are using LAMP as an alternative to expensive proprietary software stacks because of its lower cost and freedom from platform lock-in.

Founded and developed in Sweden by two Swedes and a Finn: David Axmark, Allan Larsson and Michael "Monty" Widenius, who had worked together since the 1980's. It is named after co-founder Michael Widenius' daughter, My.[18]

The SQL phrase stands for Structured Query Language MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation.

Some important features :-

- Cross-platform support.
- Stored procedures.
- Information schema.
- Independent storage engines.
- Query caching.
- Embedded database library.
- Partitioned tables with pruning of partitions in optimizer.
- Hot backup under certain conditions.[18]

3. CA ERwin Data Modeller 8.1.00.2808

CA ERwin Data Modeler (CA ERwin DM) is a database design tool that raises the level of data quality in transactional and data warehouse systems. It provides the tools to design and implement databases for transactional business, E-commerce, and data warehousing applications.

You can create and maintain graphical models that represent databases, data warehouses, and enterprise data models. CA ERwin DM provides a modeling platform where corporate data requirements and related database designs can be defined, managed, and implemented across a wide variety of database platforms.

Features:-

- Logical Data Modeling

- Physical Data Modeling
- Logical-to-Physical Transformation
- Forward engineering
- Reverse engineering
- Model-to-model comparison
- An "Undo" feature. [19]

2.4 Design Methodology

Data Warehouse Design

Schemas

- Staging Schema: - For dumping the data obtained from various data sources.
- Functional or Working Schema:-Where the processed data is stored and being accessed for any business query. Data from the staging schema is being processed based on the given business definition and then stored in this schema.

Methodologies

- Kimball Model (bottom-up approach):-Data marts are created first and implemented such that they make up the data warehouse when their information is joined.

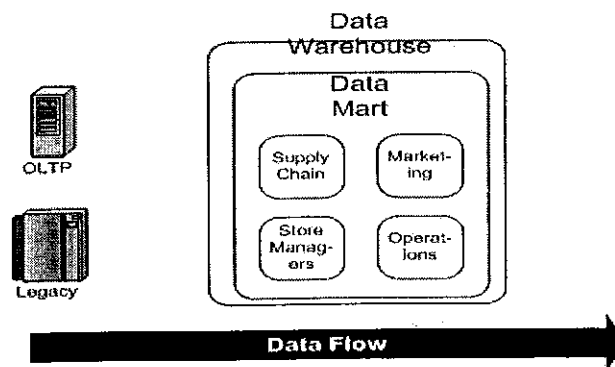


Figure 3: Representation for flow of data in Kimball model

- Inmon Model (top-down approach):-A central data warehouse can be developed and implemented first with data marts created later.

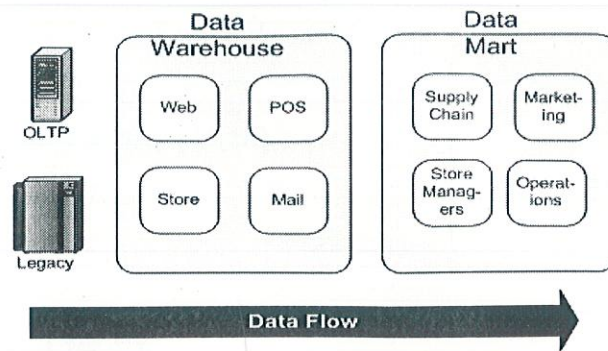


Figure 4: Representation for flow of data in Inmon model

In either approach, design must be centralized so that all of the organization's data warehouse information is consistent and usable. Data marts that adhere to central design specifications produce reports that are consistent even though the data resides in different places. For example, a sales data mart must use the same product table arranged in the same way as the inventory data mart or summary information will be inconsistent between the two.

Dimensional Modeling

Dimensional modeling is a technique for conceptualizing and visualizing data models as a set of measures that are described by common aspects of the business.

Dimensional modeling has two basic concepts :-

Facts:

- A fact is a collection of related data items, consisting of measures.
- A fact is a focus of interest for the decision making process.
- Measures are continuously valued attributes that describe facts.
- A fact is a business measure.

Facts are stored in a **FACT_TABLE**.

Dimension:

- The parameter over which we want to perform analysis of facts.
- The parameter that gives meaning to a measure number of customers is a fact, perform analysis over time. Since a dimensional model is visually represented as a fact table surrounded by dimension tables, it is frequently called Star schema. Another kind of schema include Snowflake schema. A Snowflake schema is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake in shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions. A Data

warehouse needs to integrate the data of multiple operational systems and disparate sources and establish a common format.

Dimensions are stored in a DIM_TABLE. [15,16]

2.5 Methodology Used

1) Development of a Dimensional Model:

We used the tool CA ERwin Data Modeler 8.1.00.2808 for the development of the dimensional model. It provides a simple, visual interface to manage your complex data environment.

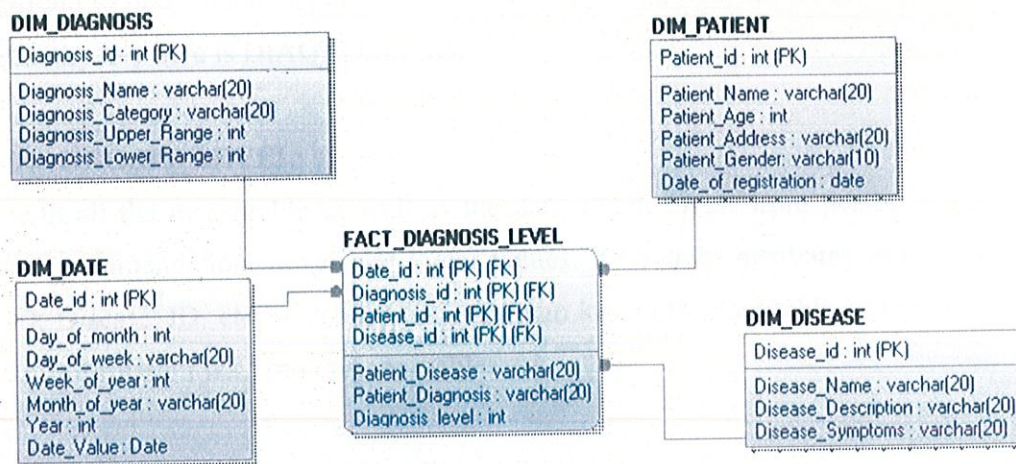


Figure 5: Dimensional Model (Star Schema Design).

The dimensional tables created in the dimensional model were DIM_DATE, DIM_PATIENT, DIM_DIAGNOSIS, DIM_DISEASE. There is one fact table created named FACT_DIAGNOSIS_LEVEL.

DIM_DATE:

Basically, it helps the user to access the patient's information date wise. It helps query retrieval easier. It consists of the following attributes:

DATE_ID: It is the primary key of this dimensional table. It provides a unique value to each record.

DIM_PATIENT:

It stores all the information concerned to a patient like patient name, patient address, his gender, his age, date of registration. Details concerning patients are necessary in the Warehouse and therefore it is stored in this dimension table.

DIM_DISEASE:

Everything related to a disease like disease name, symptoms, treatment, etc is stored in this dimension table. This information is required for the proper diagnosis of the disease. Also adding this information in a separate dimensional table solves the problem of data redundancy.

DIM_DIAGNOSIS:

It consists of the lower and upper standard ranges of the various test conducted. This information is very important to find whether a patient is normal or suffering from disease.

- Primary key here is DIAGNOSIS_ID.

FACT_DIAGNOSIS_LEVEL:

It deals with all the measurable as well as the data which varies from person to person like the diagnosis, recommendation, upper and lower values. It contains attributes from other tables like Patient_ID, Disease_ID, etc. which acts as a foreign key in the fact table and in turn helps in the retrieval of information w.r.t dimensional tables.

2) Datawarehouse Development:

Firstly, the data related to diseases and diagnostic tests was stored in the excel sheets. After that the data was processed into the MySQL databases with the help of ETL tool 'Kettle'. The transformation process required various steps and vital debugging steps were carried out from the log file.

TRANSFORMATIONS

Transformations were created in 'Kettle' and CSV / EXCEL files were processed to the staging schema.

The transformations are as following :-

- a) Creation of **STAGING SCHEMAS:**
 - (i) Disease excel file to the staging schema:

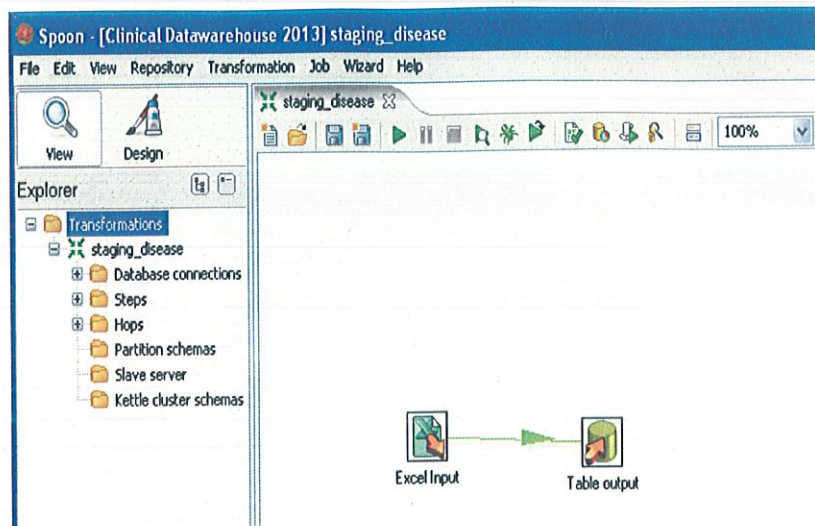


Figure 6: Snapshot of a Transformation of Disease excel file into table.

Table output of disease warehouse: A table output in MySQL was generated corresponding to it in which all the data was processed and stored.

(ii) Diagnosis excel file to the staging schema:

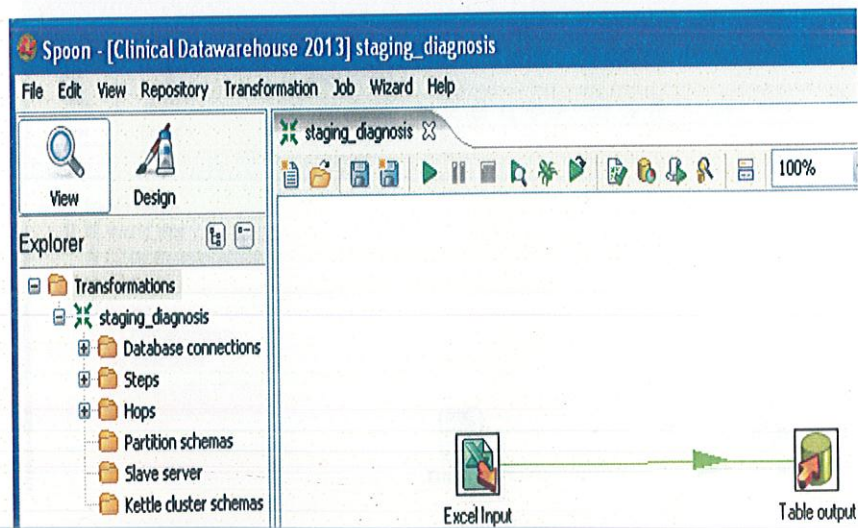


Figure 7: Snapshot of a Transformation of Diagnosis excel file into table.

Table output of diagnosis warehouse: A table output in MySQL was generated corresponding to it in which all the data was processed and stored.

Likewise, other transformations were carried and tables generated corresponding to them.

(iii) Patient excel file to the staging schema:

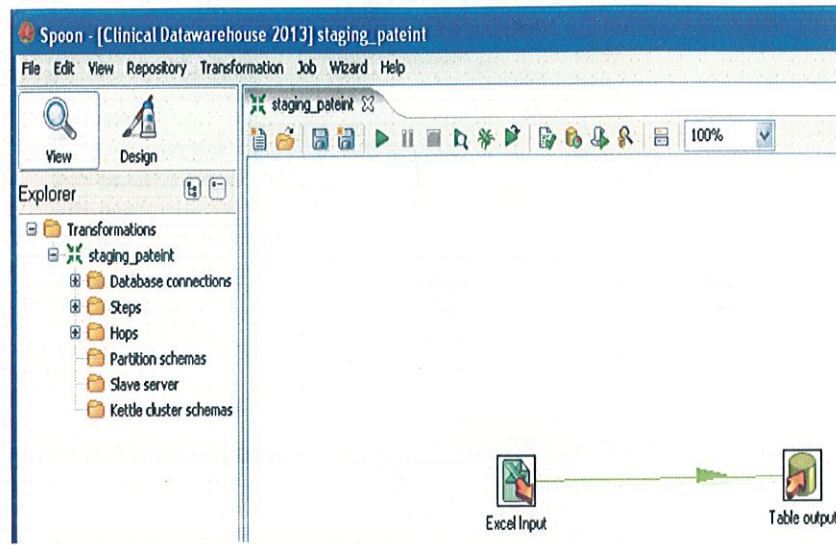


Figure 8: Snapshot of a Transformation of Patient excel file into table.

(iv) Date excel file to the staging schema:

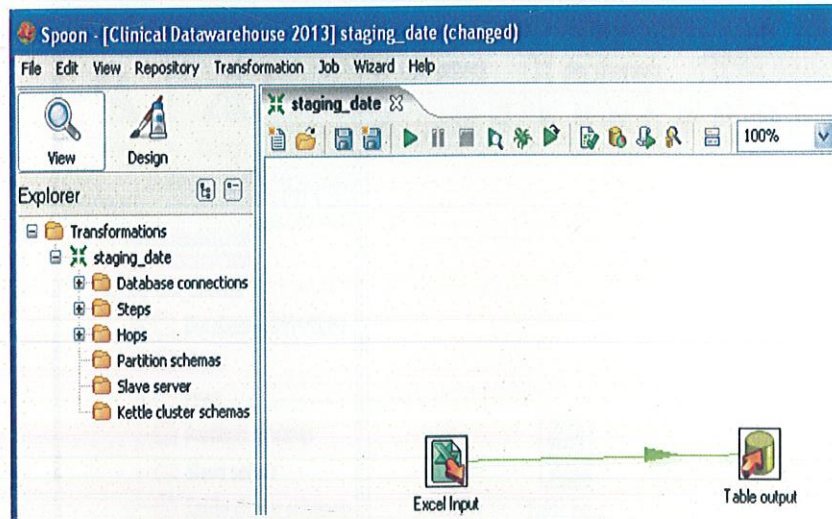


Figure 9: Snapshot of a Transformation of Date excel file into table.

(v) Diagnosis_level CSV file to the staging schema:

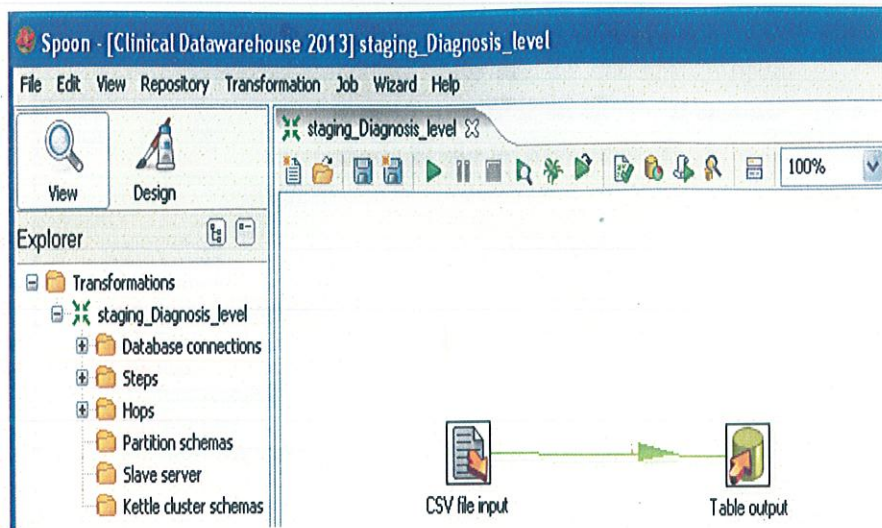


Figure 10: Snapshot of a Transformation of Diagnosis csv file into table.

b) Creation of FUNCTIONAL_SCHEMAS:

(i) Processing of data from table 'staging_disease' to table 'DIM_DISEASE':

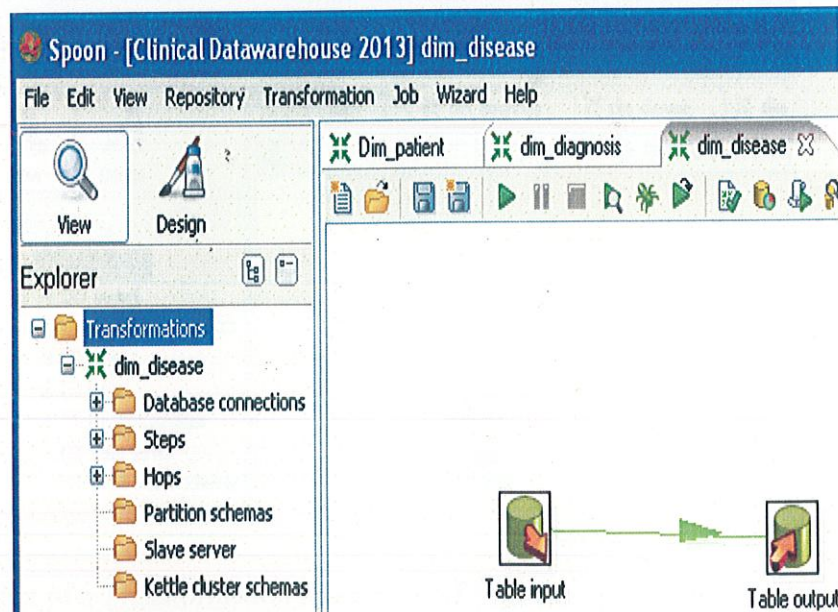


Figure 11: Snapshot of a Transformation 'staging_disease' to table 'DIM_DISEASE'.

(ii) Processing of data from table 'staging_diagnosis' to table 'DIM_DIAGNOSIS':

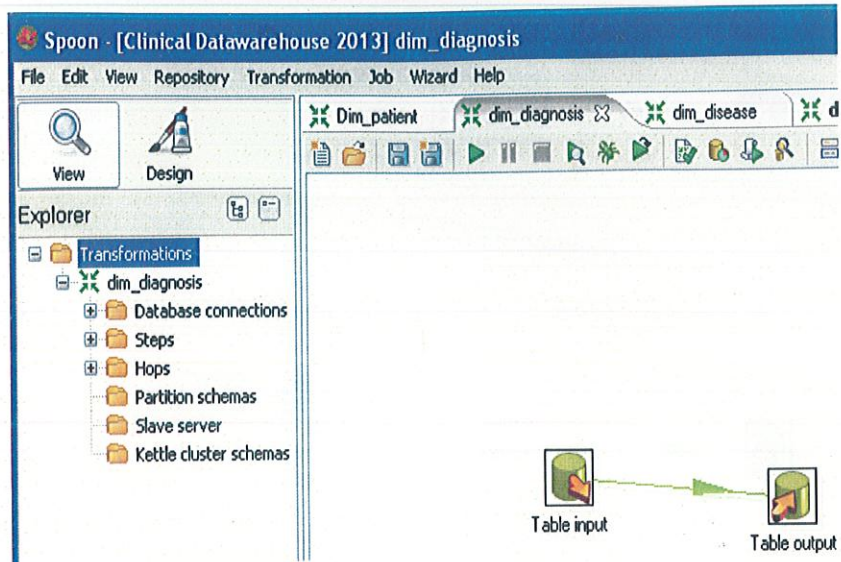


Figure 12: Snapshot of a Transformation Processing of data from table 'staging_diagnosis' to table 'DIM_DIAGNOSIS'

(iii) Processing of data from table 'staging_patient' to table 'DIM_PATIENT':

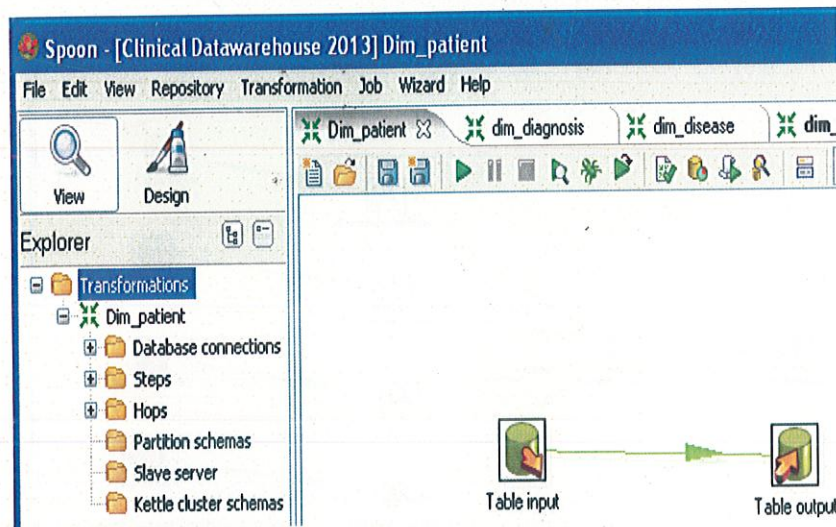


Figure 13: Snapshot of a Transformation Processing of data from table 'staging_patient' to table 'DIM_PATIENT'

(iv) Processing of data from table 'staging_date' to table 'DIM_DATE':

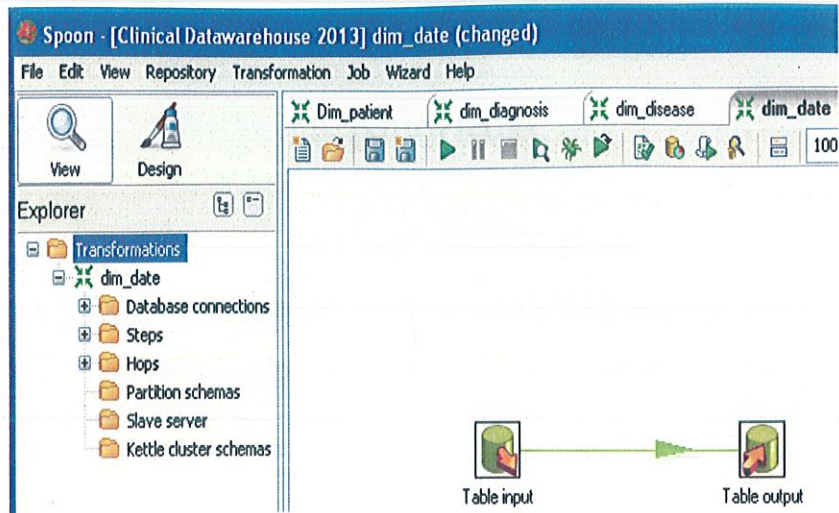


Figure 14: Snapshot of a Transformation Processing of data from table 'staging_date' to table 'DIM_DATE'

(v) Processing of data from table 'staging_diagnosis_level' to table 'FACT_DIAGNOSIS_LEVEL'

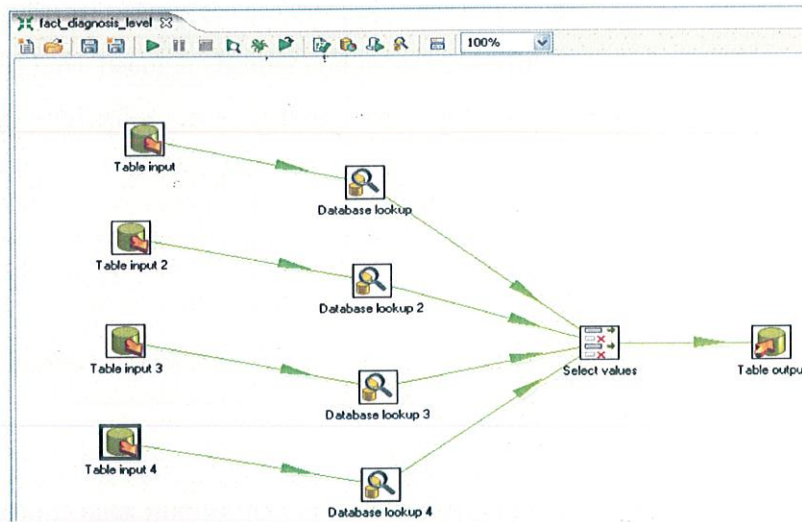


Figure 15: Snapshot of a Transformation Processing of data table 'staging_diagnosis_level' to table 'FACT_DIAGNOSIS_LEVEL'

CHAPTER 3

DATA MINING- A Knowledge Discovery Process

3.1 Introduction

Data mining is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc.

Why Mine Data?

The following are the reasons to understand the need of data mining :-

- Loads of data is collected, stored and warehoused at enormous speeds.
- Help in classifying and segmenting data
- Help Scientists in hypothesis formation
- When traditional techniques are infeasible for raw data
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong like providing better, customized services etc.

3.2 Data Mining Process

There are namely three processes :-

Pre-processing

It is executed before data mining techniques are applied to the right data. The pre- processing includes data cleaning (removing noises or make some compromises), integration and selection of data.

Data Mining

In this process different algorithms are applied to produce hidden knowledge.

Post-processing

Evaluates the mining result according to users requirements and domain knowledge.

This knowledge can be presented if the result is satisfactory, otherwise we have to run some or all of those processes again until we get the satisfactory result.[20, 28]

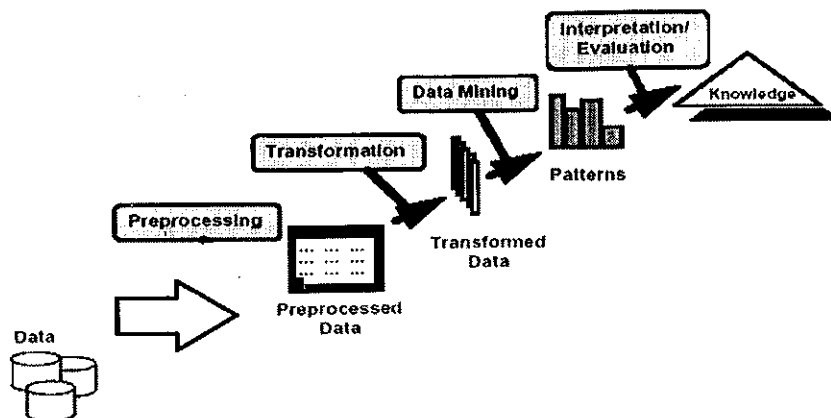


Figure 16: Data Mining Process

3.3 Types of Data Mining

- Predictive Methods :-

A predictive model makes a prediction about values of data using known results from different data set. Use some variables to predict unknown or future values of other variables.

- Descriptive Methods :-

A descriptive model identifies pattern or relationship in data. It serves as a way to explore the properties of data examined and not to predict new properties. And find human-interpretable patterns that describe the data.

In general, there are various types of mining techniques such as association rules, classifications and clustering. In this project we have used Association Mining Techniques.

3.4 Association Rule Mining

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.

It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories.

Association rules are widely used in various areas such as telecommunication networks, market and risk management, Web usage mining, intrusion detection and bioinformatics.

Mathematical definition :-

Let $I = \{I_1, I_2, \dots, I_m\}$ be a set of m distinct attributes, T be transaction that contains a set of items such that T is a subset of I , D be a database with different transaction records T_s

An association rule is an implication in the form of $X \Rightarrow Y$, where X, Y subsets of I are sets of items called itemsets, and $X \cap Y = \text{empty set}$. X is called antecedent while Y is called consequent, the rule means X implies Y . [21]

The two basic parameters of Association Rule Mining (ARM) are:-

- Support- The support (x) of an itemset x is defined as the proportion of transactions in the data set which contain the itemset.
- Confidence- The confidence of a rule is defined $\text{conf}(x \Rightarrow y) = \frac{\text{supp}(x \cup y)}{\text{supp}(x)}$. [13,23]

The association mining can be better understood with the help of an example:

Transaction ID	Items
1	Beef, Chicken, Milk
2	Beef, Cheese
3	Cheese, Boots
4	Beef, Chicken, Cheese
5	Beef, Chicken, Clothes, Cheese, Milk
6	Chicken, Clothes, Milk
7	Chicken, Milk, Clothes

Table: 1 Example of association rules.

{Chicken, Clothes, Milk} [sup = 3/7]

Association rules from the item set:

Clothes → Milk, Chicken [sup = 3/7, conf = 3/3] ...

Clothes, Chicken → Milk, [sup = 3/7, conf = 3/3]

3.4.1 Process

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two subproblems:-

- 1) One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets.
- 2) The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence. [28]

METHODOLOGY:-

For carrying out association mining, we used a tool named Statistica 8. It is a statistics and analytics software package developed by StatSoft. The software includes an array of data analysis, data management, data visualization, and data mining procedures; as well as a variety of predictive modeling, clustering, classification, and exploratory techniques. [26]

The quantitative file data is first converted into qualitative data then the file containing qualitative data of the patients was used as an input file. In data mining module of the statistica we opted for association rules option. A number of clinical variables were selected as state variables and number of clinical parameters as categorical variables. Association mining was done in many different combinations of all the clinical parameters. Then the process of association mining was performed using each and every clinical parameter.

1. The frequency and support corresponding to each item set was calculated by the software.
2. Then 40062 association rules are formed. The support, confidence and the correlation is also being calculated corresponding to each and every association rules that being formed.

S.No	Association rule	Support (%)	Confidence (%)	Correlation (%)
1	High PP → High BUN	59.0909	100	87.4475
2	High BUN, High Cholesterol → High PP	59.0909	100	76.8706
3	High LDL → High PP, High BUN, High Cholesterol	54.5455	100	96.0769
4	High PP, High LDL → High BUN, High Cholesterol	54.5455	100	96.0769
5	High BUN, High LDL → High PP, High Cholesterol	54.5455	100	96.0769
6	High Cholesterol, High LDL → High PP, High BUN.	54.5455	100	84.0168
7	High VLDL → High PP, High BUN, High Cholesterol	50	100	91.9866
8	High FBS, High Cholesterol → High PP, High BUN	50	100	80.44
9	High PP, High VLDL → High BUN, High Cholesterol	50	100	91.9866
10	High BUN, High Cholesterol → High FBS, High PP	50	84.6154	68.2191
11	High BUN, High VLDL → High PP, High Cholesterol	50	100	91.9866
12	High VLDL, High Cholesterol → High PP, High BUN	50	100	80.44

Table 2: Final result set of 12 association rules.

FUTURE WORK

- **Predictive model :**

The correlation studies are performed on the data by using tool called STATISTICA and association rules are generated, on the basis of the rules, a predictive model can be developed using normalised regression approach. For normalised regression, we can develop a program in perl, C, C++ or in any other programming language.

- **Development of the tool:**

Finally, a tool will be developed which will be integrated with this predictive model, which will predict the state of these clinical parameters when the user will input the results of its blood profiling report. The user can be a medical practitioner, patient, clinical research associate etc.

Probable Outcomes

- Development of a predictive tool that can be used to predict the occurrence of diabetes mellitus when a user submits results of common blood profiling diagnostic tests.

APPENDIX

SQL QUERIES

ALTER QUERY

```
alter table dim_diagnosis_level_n3  
add Disease varchar(30);
```

SELECT QUERY

```
SELECT * FROM dim_diagnosis_level d;  
select p_id, name, pp, fbs, rbs,t3, t4, tsh, disease from functional.dim_diagnosis_level;
```

QUERY FOR DIABETES

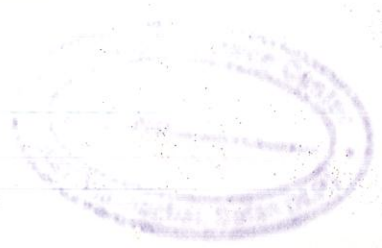
```
update functional.dim_diagnosis_level  
set Disease="Diabetes"  
where FBS>110 OR RBS>150 OR PP>150;
```

QUERY FOR HYPOTHYROID

```
update functional.dim_diagnosis_level  
set Disease="Hypothyroid"  
where T3<0.6 AND T4<5.2 AND TSH<0.35;
```

QUERY FOR HYPERTHYROID

```
update functional.dim_diagnosis_level  
set Disease="Hyperthyroid"  
where T3>1.81 AND T4>12.7 AND TSH>5.5;
```



COUNT QUERY

DIABETES

```
select count(p_id) from dim_diagnosis_level
```

```
where Disease="Diabetes";
```

OUTPUT

Number of Diabetic patients = 172

HYPOTHYROID

```
select count(p_id) from dim_diagnosis_level
```

```
where Disease="Hypothyroid";
```

OUTPUT

Number of Hypothyroid patients = 9

HYPERTHYROID

```
select count(p_id) from dim_diagnosis_level
```

```
where Disease="Hyperthyroid";
```

OUTPUT

Number of Hyperthyroid patients = 1



REFERENCES

1. Wikipedia. "Healthcare Informatics." Internet: http://en.wikipedia.org/wiki/Health_informatics [Aug 15, 2012].
2. By Ruben D. Canlas Jr. "Data mining in healthcare: Current application and issues MSIT MBA", Internet: biworld.co.kr/files/mc/board/0/289/Data_Mining_Health.pdf [July 5, 2012].
3. Corporate Information Factory. "Data Warehousing In The Healthcare Environment." Internet: <http://inmoncif.com> [September 10 ,2012]
4. Logicalis. "Data storage in healthcare." Internet: us.logicalis.com/.../Storage%20in%20Healthcare%20Feature%20Story [November 3,2012].
5. Medpac. "Information technology in health care." Internet: <http://www.medpac.gov> [September 17 , 2012].
6. Yi-An Chen, Lokesh P. Tripathi, Kenji Mizuguchi, "TargetMine an Integrated Data Warehouse for Candidate Gene Prioritisation and Target Discovery", *PlosONE*, Volume 6 , Issue 3 , 8March 2011.
7. Anthony C. Smith, James A. Blackshaw and Alan J. Robinson, "MitoMiner: a data warehouse for mitochondrial proteomics data", *Nucleic Acids Research*, Vol. 40, Database issue, November 2011.
8. Wikipedia. "Diabetes Mellitus." Internet: http://en.wikipedia.org/wiki/Diabetes_mellitus [Aug 28, 2012].
9. Diabetes Model of Care, Endocrine Health Network Working Party, Department of Health , January 2008.
10. Diabetes association UK, "Type1 diabetes." Internet: <http://www.diabetes.co.uk/type1-diabetes.html>. [August 18,2012].
11. Diabetes association UK, "Blood glucose", Internet: <http://diabetes.webmd.com/blood-glucose>. [October 8 ,2012].
12. Diabetes association UK, "Diabetes." Internet: <http://www.medicalnewstoday.com/info/diabetes>. [October 8,2012].
13. Sarah Wild, Gojka Roglic, Anders Green, Richard Sicree, Hilary King , "Global Prevalence of Diabetes Estimates for the year 2000 and projections for 2030", *Diabetes Care* ,Volume 27, Issue 1047, December 2004.
14. Hormone Health Network, "Treatment of diabetes." Internet: <http://www.hormone.org/Diabetes/treatment.cfml>. [November 8, 2012].

15. Alejandro Gutiérrez, Adriana Marotta, Instituto de Computación, Facultad de Ingeniería, "An Overview of Data Warehouse Design Approaches and Techniques", Volume 30 , Issue 1134, October 2000.
16. Surajit Chaudhuri Umeshwar Dayal Microsoft Research, "An Overview of Data Warehousing and OLAP Technology", *ACM Sigmod Record*, March 1997.
17. Pentaho Corporation. "Pentaho Data Integration (Kettle)" Internet: www.kettle.pentaho.com/ [Jan 15,2012].
18. Wikipedia "My SQL." Internet :<https://en.wikipedia.org/wiki/MySQL>. [September 17,2012]
19. CA ERwin Modelling. Internet: <http://erwin.com/products/>. [November ,2012]
20. "An Introduction to Data Mining." Internet: www.hearling.com/text/dmwhite/dmwhite.htm. [Mar 30, 2011].
21. Qiankun Zhao Nanyang Technological University, Singapore and Sourav S. Bhowmick Nanyang Technological University, Singapore, "Association Rule Mining: A Survey. Technical Report", Volume 2003116 , 2003.
22. S. A. P. Chubb, W. A. Davis, and T. M. E. Davis, "Interactions among Thyroid Function, Insulin Sensitivity, and Serum Lipid Concentrations: The Fremantle Diabetes Study", *The Journal of Clinical Endocrinology & Metabolism*, Volume 1185,doi: 10.1210/jc, 2005.
23. Wynne Hsu Mong Li Lee Bing Liu Tok Wang Ling, "Exploration Mining in Diabetic Patients Databases: Findings and Conclusions", 2003.
24. Liangjiang Wang and Aidong Zhang, "BioStar models of clinical and genomic data for biomedical data warehouse design" , *International Journal of Bioinformatics Research and Applications* , Volume 1, Issue 1, April 2005,pp 63-80 .
25. Sohrab P Shah, Yong Huang, Tao Xu, Macaire MS Yuen, John Ling and BF.Francis Ouellette, "Atlas – a data warehouse for integrative bioinformatics , *BMC Bioinformatics* , Volume 6, doi:10.1186 ,21 February 2005.
26. Wikipedia "Statistica." Internet: <http://en.wikipedia.org/wiki/STATISTICA> [February, 2013].
27. Stephen T. C. Wong, PHD, Kent Soo Hoo, Jr, Robert C. Knowlton, MD, Kenneth D. Laxer, MD, Xinhau Cao, PHD, Randall A. Hawkins, MD, PHD, William P. Dillon, MD, Ronald L. Arenson, MD, "Design and Applications of a Multimodality Image Data Warehouse Framework", *Journal of the American Medical Informatics Association* , Volume 9 ,Number 3, May / Jun 2002.

28. CONCARO a,b,1, Lucia SACCHI a, Carlo CERRA b, Riccardo BELLAZZI a Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy. Mining Administrative and Clinical Diabetes Data with Temporal Association Rules. *European Federation for Medical Informatics*. Volume1854, doi:10.3233/978-1-60750-044-5-574, 2009.

BRIEF PROFILE OF THE STUDENTS

SHRUTI KAPIL

She is pursuing her B.Tech in Bioinformatics from Jaypee University of Information Technology and will be completing her degree in June 2013. Her Technical and Research interests include: Data warehousing and data mining techniques, Perl, C++, HTML etc. Her biotechnological interests include: Cancer biology and Genetic Engineering. She is interested in pursuing higher studies and planning for post graduate studies in Clinical informatics or Cancer Biology field.

SHAGUN THAKUR

She is pursuing her B.Tech in Bioinformatics from Jaypee University of Information Technology and will be completing her degree in June 2013. Her Technical and Research interests include Data mining, C, C++, and HTML etc. She is interested in pursuing higher studies in Data mining and its applications.