CANCER CACHE –
A CANCER DATA WAREHOUSE AND PATTERN IDENTIFICATION TOOL

BY:

PRIYANKA ARORA 071519 (SUCK MY DICK)

SHRADHA PANT 071522

Under the supervision of:

DR. PRADEEP KUMAR NAIK

MR.DIPANKAR SENGUPTA

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
juit
विद्या तत्व ज्योतिसम

MAY-2011

Submitted in partial fulfillment of the Degree of

Bachelor of Technology

In

BIOINFORMATICS

DEPARTMENT OF

BIOTECHNOLOGY & BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT

TABLE OF CONTENTS

TOPIC	PAGE.NO
CERTIFICATE	v
ACKNOWLEDGEMENT	vi
SUMMARY	vii-viii
LIST OF FIGURES	. ix
LIST OF TABLES	xi
CHAPTER 1: INTRODUCTION	1-17
1.1 Cancer	5-12
1.1.1 What causes cancer?	6
1.1.2 Treatment	7
1.1.3 Early detection	8
1.1.4 Types of Cancer	9
1.2 Brief Introduction of the Project	13-17
1.2.1 Objective	16
1.2.2 Project Plan	17
CHAPTER 2: TOOLS AND TECHNIQUES	18-26
2.1 Pentaho	18-22
2.1.1 Kettle	18
2.1.2 Kettle Architecture	19
2.2 MySQL (RDBMS Package)	22
2.3 MATLAB	23
2.4 HTML	24
CHAPTER 3 : DEVELOPMENT OF DATA WAREHOUSE: CANCER CACHE	27-60
3.1 Introduction to Data Warehouse	27
3.2 How Does a Data Warehouse Differ From a Database?	30

3.3 Data Marts 33

3.4 Creating a Data warehouse 33-39

3.4.1 Prerequisites - the data model 33-35

3.4.1.1 Why is Data Modeling Important? 34

3.4.1.2 Dimensional Modeling 35

3.4.2 Image Processing 36-38

3.4.2.1 Introduction 36

3.4.2.2 Role of Image processing in medical field 36

3.4.2.3 Advantages of Digital Processing for Medical Applications 38

3.4.3 Identification of Data Source 38

3.5 Methodology Used for Data Warehouse Construction 39-60

3.5.1 Step 1#: Data Collection 40

3.5.2 Step 2#: Data Modeling 43

3.5.3 Step 3#: Data Preprocessing 46

3.5.4 Step 4#: Processing of Images 55

3.5.5 Step 5#: Data processing to Functional Schema 57

CHAPTER 4 : APPLICATION OF DATA MINING TECHNIQUES 61-71

4.1 Introduction to Data Mining 61

4.2 Limitations of Data Mining 62

4.3 Uses of Data Mining 63

4.4 Data Mining Issues 63

4.5 Use of Data Mining in Cancer Cache 64

CONCLUSION & FUTURE SCOPE 72-73

APPENDIX 74-85

A - MATLAB PROGRAM-tumor.m 75

B - SQL QUERIES 81

REFERENCES 86-88

LIST OF PUBLICATIONS (COMMUNICATED/ACCEPTED) 89

BRIEF BIODATA OF THE STUDENTS 90

CERTIFICATE

This is to certify that the thesis entitled "*CANCER CACHE* – A CANCER DATA WAREHOUSE AND PATTERN IDENTIFICATION TOOL" submitted by **PRIYANKA ARORA (071519)** and **SHRADHA PANT (071522)** to the Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat 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.

Dr. Pradeep Kumar Naik	**Mr. Dipankar Sengupta**
Assistant Professor	**Associate Lecturer**
Dept.of	Dept.of
Biotechnology & Bioinformatics	Biotechnology & Bioinformatics
Jaypee University of	Jaypee University of
Information Technology	Information Technology
Waknaghat-173234	Waknaghat-173234
Solan, Himachal Pradesh	Solan, Himachal Pradesh

Date: 2011

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 **Dr Pradeep K. Naik** and **Dipankar Sengupta**, our Project Guides, for their guidance, encouragement and support. They have always encouraged us to put in our best efforts and deliver a quality and professional output. Their 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. I really thank them for their time & efforts.

Secondly, 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 also thank **Mrs. Somlata Sharma** (Bioinformatics Laboratory In-charge) and **Mr. Sree Krishna** for providing their full cooperation with keen interest.

We would also like to thank doctors from **Gandhi Memorial & Associated Hospitals (KGMC), Lucknow Cancer Institute and Raj Diagnostics, Lucknow** 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 this project 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.

All copyrights and trademarks that are cited in this document remain the property of their respective owners.

Priyanka Arora (071519)

Date:

Shradha Pant (071522)

Date:

SUMMARY

The current status of research being conducted in the field of Life Sciences is leading us to a phase, where huge chunk of data is generated on daily basis. It has thus become a necessity for efficient storage of the data that is being generated, further trying to correlate them and extracting out knowledge from them.

One of the forms of data is the one that is being generated in Hospitals, Clinics & Diagnostics centers with respect to health aspect of an individual. Patient's records in various hospitals are increasing at an exponential rate, thus adding to the problem of data management and storage. This has led to the evolvement of **Clinical Informatics**. In order to deal with rapidly growing data the hospitals have made use of computers to make data easily accessible, yet its use in clinical informatics is still not promoted. Another major problem being faced is the, varied dimensionality of the data, ranging from images to numerical data. Therefore there is a need for development of efficient solutions which can handle the multi-dimensionality of the data and can store all the historical aspect with respect to an individual.

In this project we have selected the stated problem lying in façade of clinical informatics and tried to develop a **Clinical Data Warehouse** called ***"Cancer Cache"*** which could store historical data of all the cancer patients concerned with a particular hospital. Also our aim is the data being stored could be correlated and a predictive tool can be designed for predicting the occurrence of a cancer when a user submits results of common biochemical or pathological tests.

Concepts of classical data warehouse have been implemented in this project involving use of ETL technology for extracting, cleaning & storing data in the warehouse. Application of data mining techniques has been planned to be implemented for designing the prediction tool. Also for easy user handling, a web interface has been made so that any individual can easily deposit their records,

and can use the prediction tool for finding the probability of occurrence of a particular disease currently which is restricted to occurrences of cancer.

Cancer Cache has been hosted by use of a graphical interface on JUIT web server and can be accessed at: **www.juit.ac.in/attachments/CancerCache/home.html**.

Signature of Student

Name Priyanka Arora (071519)

Date

Signature of Student

Name Shradha Pant (071522)

Date

Naik

Signature of Supervisor

Name Dr. P. K. Naik

Date

Signature of Supervisor

Name Mr. Dipankar Sengupta

Date

LIST OF FIGURES

Fig1: Cancer
Fig2: Diagnostics
Fig3: Colon Cancer
Fig4: Lung Cancer
Fig5: Project Plan
Fig6: Loading an Input file dialog box.
Fig7: Previewing the list of attributes.
Fig8: Preview of data will look like in a table.
Fig9: List of errors along with the description.
Fig10: MATLAB Interface
Fig11: Dreamweaver Interface
Fig12: A Schematic Representation of Data warehouse model
Fig13: Representation for flow of data.
Fig14: Image Processing
Fig15: Medical Imaging
Fig16: Different types of data
Fig 17: Performa
Fig 18: Schema
Fig19: Snapshot of a Transformation
Fig20: Snapshot of an Input File
Fig21: No .of samples to be analyzed to get fields
Fig22: Fields after analyzing the sample
Fig23: Data type of various fields
Fig24: The Output file
Fig25: Altering of the table in MySQL from Spoon
Fig26: MRI of Brain(Axial view)
Fig27: Example of a Function
Fig28: Select/Rename function

Fig29: Cancer Cache:Website Snapshot

Fig30: Cancer cache -Upload Data Browser

Fig31: Decision Tree example

LIST OF TABLES

Table 1: Tables in the staging schema "cancer"

Table 2: Tables in the working schema "working"

Table 3: Transformations for the Working Schema

CHAPTER 1

INTRODUCTION

Every day in a different country in a different state in a different city in a different hospital lands a new patient, a new case, a new heap of data but an old problem of data storage of nearly every hospital/research institute. Its time for a change, an advancement. The extraordinary explosion of medical knowledge, technologies, and ground-breaking drugs may vastly improve healthcare delivery for the welfare of its consumers, but the key is to implement these technologies, to extract as much as we can. **Healthcare informatics** utilizes computer hardware, specialized software, and communication devices to form complex computer networks to collect, analyze, and transmit medical processes. [1]

We are dealing more precisely with data storage & data correlation aspects of clinical informatics. It is a method of organizing information in the health care industry. It blends information technology, computer science and biomedical informatics. Clinical informatics is a field that is constantly striving to make information more accessible in the simplest way. It involves storing, managing and accessing important health records. Clinical informatics uses technology and computers to store data at an institution such as a hospital, doctor's office or other health care facility. 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. Examples of information stored in health informatics include disease research, patient backgrounds, statistics and treatment plans.

Clinical computing is typically the easiest way to store the required information. This use of technology allows not just for the entry of facts and figures, but for the automatic recording of a patient's vital health statistics, such as temperature or blood pressure, into his or her electronic medical records. Clinical informatics can also be used to communicate between doctors at different hospitals or clinics and even at different locations. Through a process known as telemedicine, doctors can exchange pictures of medical conditions across the globe.

Since clinical informatics is a multidisciplinary field, it combines data representation, cognitive science, policies, telemedicine and data discovery. The ability to quickly and efficiently retrieve information makes the creation of an organized database indispensable. Clinical informatics provides for this and makes the representation and interpretation of complex medical terms quite simple. Cognitive science comes into play to help those in the medical community, understand process and perceive artificial intelligence and computing. While telemedicine refers to the way patient data is transferred using information technology, policies evaluate this technology on the larger health care system. [2]

A stone for somebody might be gold for us, in other words what hospitals consider to be junk data might be as important and meaningful as a medicine/drug given to a patient. It's all about fetching 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 early detection of certain life threatening diseases.

In clinical informatics, we deal majorly with the clinical data concerned with a patient or a group of patients, this may include a patient's health records, and history with the disease, and treatment description etc. the kind/type of data varies from a needle to a sophisticated machine.

It's not only about discovering a drug to cure an epidemic. Technology allows clinical research and patient care to become more integrated and interactive. In so-called translational science, its a need that basic science and clinical researchers work together on interpretation and application of research data in clinical settings. Data sharing is necessary to improve the quality of healthcare and accelerate progress in biomedical sciences from bench to bedside to community. To go from clinical research to community practice, integrated data systems (IDSs) must be created to allow community researchers to easily access secure and confidential research data. These data can then be used to answer questions relevant to specific communities and can be extrapolated to a national level [Slim Prim Biomedical Database – University of Tennessee]. Furthermore, information can be assimilated for community education to help improve healthcare.

To address these clinical data integration issues, we at **Jaypee University of Information Technology** undertook the challenge of developing a solution called ***"Cancer Cache*: A Data Warehouse"**, to provide a solution of addressing the dimensionality issue concerned with the data and long term storage of patient's clinical data. Currently we are only trying to address the problem with respect to disease-cancer, but in nearby future we would like to expand the horizon and include all the other data for patients with various other diseases.

Cancer affects everyone – the young and old, the rich and poor, men, women and children – and represents a tremendous burden on patients, families and societies. Cancer is one of the leading causes of death in the world, particularly in developing countries. Yet, many of these deaths can be avoided. Over 30% of all cancers can be prevented. Others can be detected early, treated and cured. Even with late stage cancer, the suffering of patients can be relieved with good palliative care. [3]

The global burden of cancer continues to increase largely because of aging and growth of the world population alongside an increasing adoption of cancer-causing behaviors, particularly smoking, in economically developing countries. Based on the GLOBOCAN 2008 estimates, about 12.7 million cancer cases and 7.6 million cancer deaths are estimated to have occurred in 2008; of these, 56% of the cases and 64% of the deaths occurred in the economically developing world. Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer death among females, accounting for 23% of the total cancer cases and 14% of the cancer deaths.[4]

Lung cancer is the leading cancer site in males, comprising 17% of the total new cancer cases and 23% of the total cancer deaths. Breast cancer is now also the leading cause of cancer death among females in economically developing countries, a shift from the previous decade during which the most common cause of cancer death was cervical cancer. Further, the mortality burden for lung cancer among female in developing countries is as high as the burden for cervical cancer, with each accounting for 11% of the total female cancer deaths [4]. Although overall cancer incidence rates in the developing world are half those seen in the developed world in both sexes, the overall cancer mortality rates are generally similar. Cancer survival tends to be poorer in developing countries, most likely because of a combination of a late stage at diagnosis and limited access to timely and standard treatment. A substantial proportion of the worldwide burden of cancer could be prevented through the application of existing cancer control knowledge and by implementing programs for tobacco control, and early detection and treatment, as well as public health campaigns promoting physical activity and a healthier dietary intake. Clinicians, public health professionals, and policy makers can play an active role in accelerating the application of such interventions globally.

"Cancer scenario in India is not very comfortable and every year there is an increment of 10,000 new cancer patients and the number of total victims stands at about 25 lakh all over," Indian Council of Medical Research (ICMR) Director General Viswamohan Katoch. In last decade number of cancer cases have increased significantly and in some sub forms of cancer India actually top the chart. Two out of five Cancer cases in India are due to Tobacco and Breast Cancer cases have risen by 100% in Urban India. About 3 people die every 2 minutes because of cancer in India [5], there are about 25 lakhs(2.5 million roughly one every 400 persons) or one in 80 families have cancer) cancer patients in India - Indian Council of Medical Research (ICMR) [6]. Over 400,000 Indians die every year from cancer and the disease is growing by 11 percent annually. Every year about 8,50,000 new cancer cases are diagnosed in India resulting in about 5,80,000 cancer related death every year. India has the highest number of the oral and throat cancer cases in the world. Every third oral cancer patient in the world is from India.

Many researchers believe that change in life style of Indians is contributing mainly in the increase of this problem. Present life style of people is full of junk food, lack of exercise and stress. All these factors lead to the increase in free radicals in body which ultimately increase the chances of body acquiring cancer. Though, there are number of other reasons also which do contribute in cancer, like exposure to some radiation, drug etc; however, present life lifestyle is the main culprit as per most of scientists. It is possible for people to reduce chances of cancer from 70% to 30 % by opting healthy life style. Healthy Life style include eating healthy food full of all nutritional elements like vegetable, cereals, fruits, wheat etc and doing regular exercise and meditation. In various independent studies, it is proved scientifically that chance of cancer decreases significantly with healthy life style. However, in the present fast life style very less number of people get chance for living a healthy life. It is very important for all of us to understand that cancer is one of deadly diseases present in our world and it mostly gives no initial warning to people. Present trends of cancer in India only suggest that we all Indians are required to opt for a cautious approach on this problem and start living a healthy life style.

1.1 Cancer

Cancer (medical term: malignant neoplasm) is a class of diseases in which a group of cells display uncontrolled growth, Cancer starts in the cells that are the units of our physical structure. Generally, our physical structure builds fresh cells as we require them, substituting old cells that break down. Occasionally that procedure miscarries. Fresh cells develop even while we do not require them, and older cells do not die once they should. The following additional cells could build a mass addressed as a tumor. Tumor could be benign or malevolent/malignant. Invasion that intrudes upon and destroys adjacent tissues, and sometimes metastasis, or spreading to other locations in the body via lymph or blood are termed malignant properties. These three malignant properties of cancers differentiate them from benign tumors, which do not invade or metastasize. Researchers divide the causes of cancer into two groups: those with an environmental cause and those with a hereditary genetic cause. Cancer is primarily an environmental disease, though genetics influence the risk of some cancers. Common environmental factors leading to cancer include: tobacco, diet and obesity, infections, radiation, lack of physical activity, and environmental pollutants. These environmental factors cause or enhance abnormalities in the genetic material of cells. Cell reproduction is an extremely complex process that is normally tightly regulated by several classes of genes, including oncogenes and tumor suppressor genes. Hereditary or acquired abnormalities in these regulatory genes can lead to the development of cancer. [7]

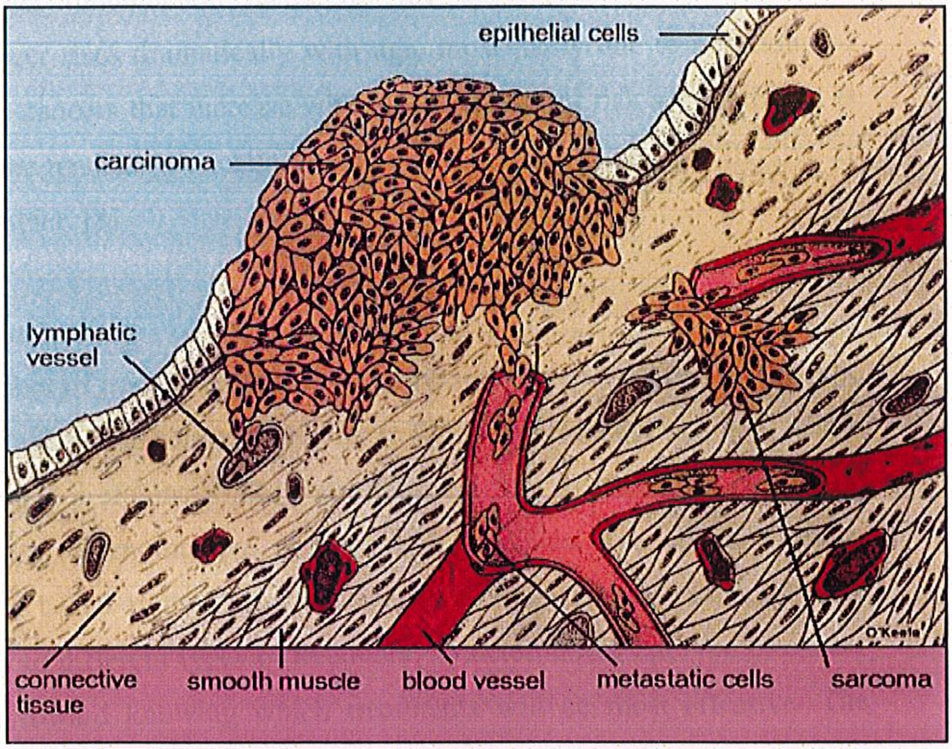


Fig1: Cancer

1.1.1 What causes cancer?

Cancer arises from one single cell. The transformation from a normal cell into a tumor cell is a multistage process, typically a progression from a pre-cancerous lesion to malignant tumors. These changes are the result of the interaction between a person's genetic factors and three categories of external agents, including:

- Physical carcinogens, such as ultraviolet and ionizing radiation;
- Chemical carcinogens, such as asbestos, components of tobacco smoke, aflatoxin (a food contaminant) and arsenic (a drinking water contaminant); and
- Biological carcinogens, such as infections from certain viruses, bacteria or parasites.

Ageing is another fundamental factor for the development of cancer. The incidence of cancer rises dramatically with age, most likely due to a buildup of risks for specific cancers that increase with age. The overall risk accumulation is combined with the tendency for cellular repair mechanisms to be less effective as a person grows older. [8]

1.1.2 Treatment

The number of treatment choices an individual has will depend on the type of cancer, the stage of the cancer and other individual factors such as age, health status, and personal preferences. The four major types of treatment for cancer are surgery, radiation, chemotherapy, and biologic therapies. The specific cancer treatment will be based on the individual's needs. Certain types of cancer respond very differently to different types of treatment, so determining the type of cancer is a vital step toward knowing which treatments will be most effective. The cancer's stage will also determine the best course of treatment, since early-stage cancers respond to different therapies than later-stage ones. The individual's overall health, lifestyle, and personal preferences will also play a part in deciding which treatment options will be best. About all cancers are diagnosed for where they begin. For instance, lung carcinoma begins in the lung, and breast carcinoma commences in the breast.

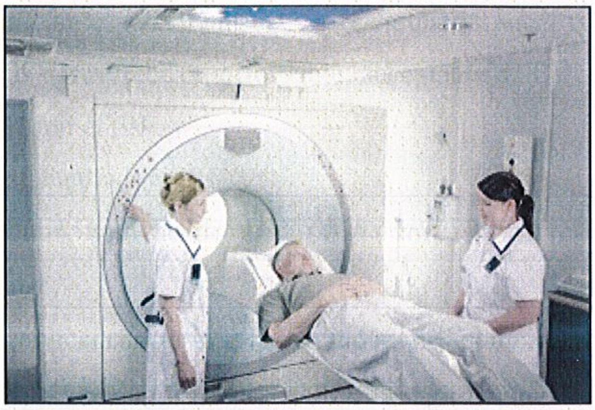


Fig2: Diagnostics

A cancer diagnosis is nearly always made by an expert looking at cell or tissue samples under a microscope. In some cases, lab tests of the cells' proteins, DNA, and RNA can help tell doctors if cancer is present. The procedure that takes a sample for this testing is called a biopsy, and the tissue sample is called the biopsy specimen. Lumps that might be malignant (cancer) may be found by imaging (radiology) studies or felt as masses (lumps) during a physical exam, but they still must be sampled and looked at under a microscope. More than 30% of the cancer deaths can be prevented. [9]

1.1.3 Early detection

Cancer mortality can be reduced if cases are detected and treated early. There are two components of early detection efforts:

o **Early diagnosis**

It is the awareness of early signs and symptoms (such as cervical, breast and oral cancers) in order to facilitate diagnosis and treatment before the disease becomes advanced. Early diagnosis programmes are particularly relevant in low-resource settings where the majority of patients are diagnosed in very late stages.

o **Screening**

It is the systematic application of a screening test in an asymptomatic population. It aims to identify individuals with abnormalities suggestive of a specific cancer or pre-cancer and refer them promptly for diagnosis and treatment.

Screening programmes are especially effective for frequent cancer types that have a screening test that is cost-effective, affordable, acceptable and accessible to the majority of the population at risk.

Examples of screening methods are:

- Visual inspection with acetic acid (VIA) for cervical cancer in low-resource settings;
- PAP test for cervical cancer in middle- and high-income settings;
- Mammography screening for breast cancer in high-income settings.[9]

Worldwide, the 5 most common types of cancer that kill men are (in order of frequency): lung, stomach, liver, colorectal and oesophagus and the 5 most common types of cancer that kill women are (in the order of frequency): breast, lung, stomach, colorectal and cervical.[9]

1.1.4 Types of Cancer

The four most common cancers are:

- **Breast Cancer**

Breast cancer is the most common cancer in women, being responsible for almost 20% of all cancer deaths in women. It ranks second after lung cancer. With increased awareness and increased use of routine mammograms, more women are diagnosed in the earlier stages of this disease, at which time a cure may be possible. For every 100 women, one man is diagnosed with this disease. The disease is more common in women after age 40. It is also more frequent in women of a higher social-economic class. [10]

- **Colon Cancer**

Colorectal cancer is the third most common cancer in men and women. Certain genetic factors play a role in the development of this cancer. The specific cause of colorectal cancer is unknown, however, environmental, genetic and familial factors have been linked to the development of this cancer. It is more common among African-Americans.

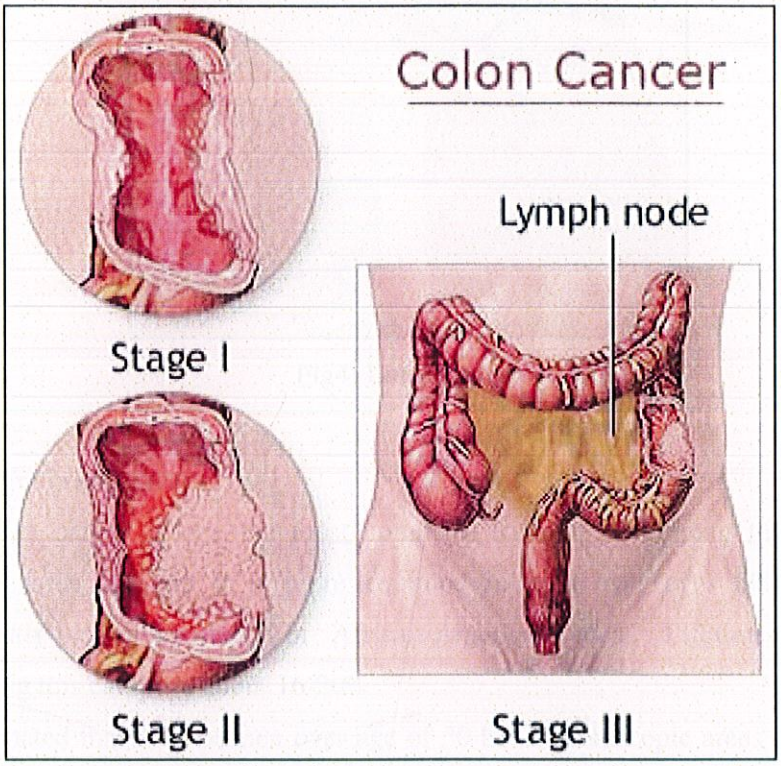


Fig3: Colon Cancer

- **Lung Cancer**

Lung cancer is the second most common malignancy affecting both sexes. It is considered the most rapidly increasing cause of death from cancer. Since 1987, lung cancer has been the leading cause of cancer death in women, surpassing breast cancer. And while lung cancer incidence has leveled off among men, it continues to rise steadily among women. The average age of patients with lung cancer is 60 years. It is more common in African-Americans and Hawaiians. [10]

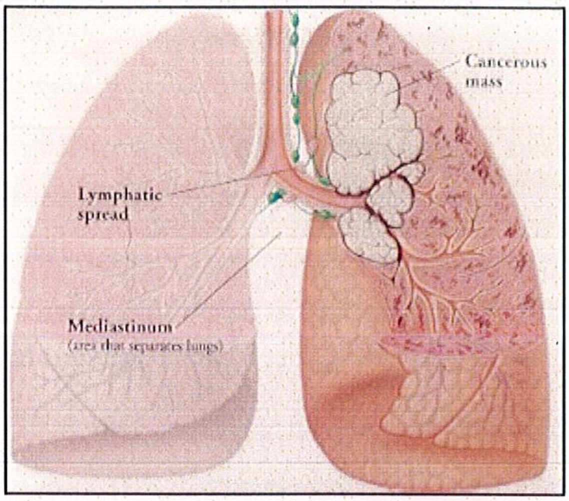


Fig4: Lung Cancer

- **Prostate Cancer**

Prostate cancer is one the most common. Risk of developing this cancer increases with age and it is more common in men over ages 60-65. It is significantly more common in African-American men. Lifetime risk of developing this cancer is about 16-20% .

It is estimated that 40% of men over age of 50 have microscopic areas of cancer in their prostate gland. However, only 8% of men will develop clinically significant disease and only 3% will die of this disease. Prostate cancer grows very slowly in older men and does not contribute to the cause of death in majority of cases. [10]

Other types include:

❖ **Cancers of Blood and Lymphatic Systems:**

Leukemia

Lymphomas

Multiple Myeloma

❖ **Skin cancers:**

Malignant Melanoma

Skin Cancer

- **Cancers of Digestive Systems:**

  Head and Neck Cancers

  Esophageal Cancer

  Stomach Cancer

  Cancer of Pancreas

  Liver Cancer

  Colon and Rectal Cancer

  Anal cancer

- **Cancers of Urinary system:**

  Kidney Cancer

  Bladder Cancer

  Testis Cancer

  Prostate Cancer

- **Cancers in women:**

  Breast Cancer

  Ovarian Cancer

- **Miscellaneous cancers:**

  Brain Tumors

  Bone Tumors

  Soft Tissue Tumors

  Thyroid Cancer

  Cancers of Unknown Primary Site

[10]

1.2 Brief Introduction of the Project

In all, about 40 different factors have been found to be contributory causes of the more common forms of cancer in man. Some of these, such as pollution or industrial chemicals and their effluents, are man-made and their toxic effects can be controlled by strict regulations. Others such as diet, alcohol consumption, smoking or sexual habits, are more personal and cannot be controlled in this way.

What affects one body tissue may not affect another. For example, tobacco smoke that we breathe in may help to cause lung cancer. Overexposing our skin to the sun could cause a melanoma on our leg. But the sun won't give us lung cancer and smoking won't give us melanoma.

On an average, cancers are diagnosed at a much later stages in India .A majority of deaths due to cancer in India can be prevented if the disease is diagnosed early. Around 99 per cent of cancer is curable in the initial stages but patients in India come to know about the disease very late, when it is not curable anymore.

We need to optimize patient care by presenting the right information at the right time to the right people.

Interpreting data across multiple systems is challenging, and various integration techniques, with varying levels of complexity, have been proposed to solve the problem of data integration and storage. [11–14] Nagarajan *et al.* introduced data-warehousing-based solutions utilizing relational database management systems (RDBMSs) for assembling and integrating data. A relational database model is composed of classes of data, with each class characterized by a set of attributes. This conventional design is ideal for data sets composed of classes with a limited and fixed number of attributes. When each instance has values for all attributes (or columns) within a class (or table), then the database is not filled with numerous null entries and memory is used efficiently. However, research reveals that this design is not efficient for data sets with large numbers of attributes that vary over time [15]. Because most database engines limit the number of columns per table, they cannot accommodate massive numbers of class attributes. Also, continuously changing the number and type of attributes necessitates frequent modification of the database structure. Inefficient use of memory because of the large number of null entries is also a legitimate concern. Recent research has proposed a knowledge-based terminology for identifying data dimensions in clinical informatics. [16] Other research has focused on the conceptual development of IDs using ontology-based systems for the design and integration of clinical trial data. [17] The inherent variation between databases due to the different demands on each system means that there is no consensus on ontology and metadata descriptions. It might therefore be necessary to define a new ontology for each database. Although this approach gives the database designer freedom at the outset, inexperienced designers can spend excess time researching previous knowledge, seeking an optimum design. Where possible, designers should use preexisting ontologies. These can be modified as necessary to improve accessibility. Wang et al. developed the BioMediator system to provide a theoretical and practical foundation for data integration across diverse biomedical domains via a "knowledge-base-driven centralized federated database" model. [18] However, the efficiency of query processing time and the need to filter out unnecessary query results still are concerns. The data architecture required for clinical data warehousing has been researched in applications such as clinical study data management systems (CDMSs) and clinical patient record systems (CPRSs). They both use an entity-attribute-value (EAV) system (i.e., row modeling) as opposed to conventional database design [19]. The EAV system has the advantage of remaining stable as the number of parameters increases when knowledge expands, a common situation in the basic sciences and in clinical trials. [20] The characteristics of clinical data as it originates during the process of clinical documentation, including issues of data availability and complex representation models, can make data mining applications challenging. Data preprocessing and transformation are required before one can apply data mining to clinical data. The application of classical data warehousing process should be thus able to answer the queries being raised and also be able to mitigate issues like appropriate storage structure of clinical data, able to handle varied sources of data, reduce the dimensionality constraint, and handling of multiple data variables. This it would make it easier for researchers and data analysts to acquire the data and information they need. The data warehouse for clinical data being developed should be able to render the data in appropriate structures, provide metadata that adequately records syntax/semantics of data and reference pertinent medical knowledge.

In this project we have built a Clinical Data warehouse for Cancer, which acts as a data collector, data integrator and data provider in the data mining process which could be used by doctors, physicians and other health professionals, in conjunction with a Prediction Tool, to support the clinical process as well as to formulate the appropriate model to improve the quality of diagnosis and treatment recommendation decision making.

As Clinical informatics is the study of information systems (computers and programs) used in the clinical practice of medicine, so our honest attempt might contribute in the following aspects:

- ✓ **Data Entry** - The hospitals can keep a complete record of their patients in the form of electronic health record (EHR).
- ✓ **Telemedicine** - The pathologist in X location can review the unusual tumor seen in a young male in rural Y location without the physical slides—Easy accessibility
- ✓ **Data Display** - Vital signs can be highlighted when abnormal mean or median values can be graphed with the raw numbers over time to simplify clinician review-This is done using the data mining techniques.
- ✓ **Decision Support** -Immediate feedback at the time of order entry about any correlation among various parameters of that patient can be shown to reduce both patient morbidity and healthcare costs.

Hospitals generate large amounts of data yet there is no knowledge on how to mine them to find any meaningful information or pattern for further research. Clinical data warehouses are complex and time consuming to review a series of patient records however they are one of the efficient data repositories existing to deliver quality patient care. Data integration tasks of medical data store are challenging scenarios when designing clinical data warehouse architecture.

A few decades ago, physicians knew pretty much everything that is to be known about medicines; most doctors could recollect the names of their patients. However, today, no doctor can keep up with the explosion of medical and health information. While health care organizations have recognized the use of computers in other industries, its application in healthcare have not been encouraging. This is because, among other factors, it takes too long to get information in many cases; there is no easy accessibility to data, and no uniform standard among various vendors. But once the data warehouse is ready, it's worth spending the time and money in it.

1.2.1 Objective

- Develop ***Cancer Cache***-A Clinical Data warehouse for efficiently storing the clinical health records of the cancer patients.
- Processing of images generated by clinical analysis and bringing the data at a common granular level.
- Further to analyze the "Histopathological & other parameters" so as to find correlation among them, in order to give a probabilistic prediction about the occurrence of cancer.

1.2.2 Project Plan

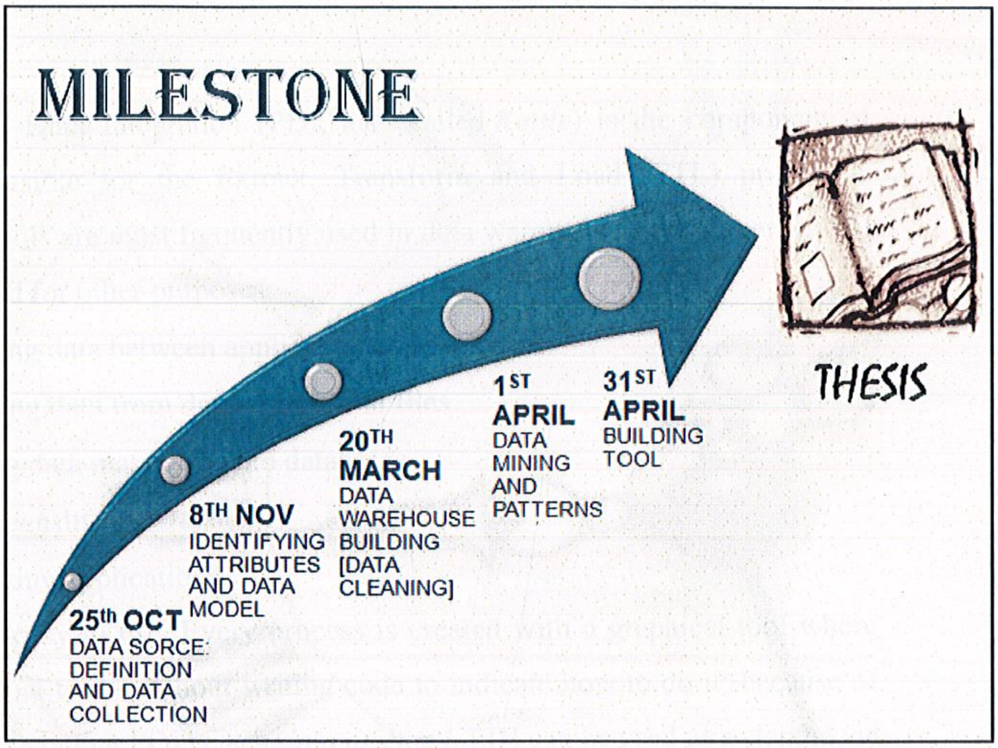


Fig5: Project Plan

CHAPTER 2

TOOLS AND TECHNIQUES

2.1 Pentaho

Pentaho Data Integration (PDI, also called *Kettle*) is the component of Pentaho responsible for the Extract, Transform and Load (ETL) processes. Though ETL tools are most frequently used in data warehouse environments, 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

PDI is easy to use. Every process is created with a graphical tool where you specify what to do without writing code to indicate how to do it; because of this, you could say that PDI is *metadata oriented.* PDI can be used as a standalone application, or it can be used as part of the larger Pentaho Suite. As an ETL tool, it is the most popular open source tool available. PDI supports a vast array of input and output formats, including text files, data sheets, and commercial and free database engines. Moreover, the transformation capabilities of PDI allow you to manipulate data with very few limitations.

2.1.1 Kettle

Kettle is a free, open source ETL (Extraction, Transformation and Loading) tool. The product name should actually be spelled as K.E.T.T.L.E, which is a recursive acronym for "Kettle Extraction, Transport, Transformation and Loading Environment".

Kettle was first conceived by Matt Casters, who a platform-independent ETL needed tool for his work as a BI Consultant. Matt's now working for Pentaho as Chief of Data Integration. [21]

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 modeled database called a data warehouse

Although most of these concepts are equally applicable to almost any data importing or exporting processes, ETL is most frequently encountered in data warehousing environments.

2.1.2 Kettle Architecture

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

Spoon

It 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

It is a command line tool that executes transformations modeled with Spoon.

Chef

It 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.

Model-driven

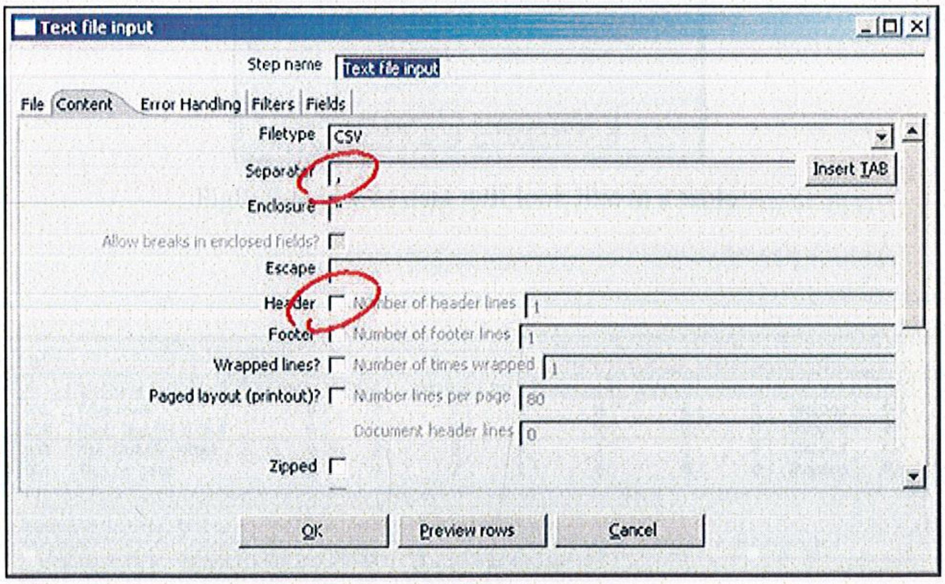
An interesting feature of Kettle is that it is model-driven. Both Spoon and Chef offer a graphical user interface to define the ETL processes on a high level. Typically, this involves no actual programming at all - rather, it's a purely declarative task which results in a model.

The command line tools Pan and Kitchen (or rather the underlying API) know how to read and interpret the models created by Spoon and Chef respectively. These tools actually execute the implied ETL processes. This is done all in one go: there is no intermediate code generation or compilation involved. [21]

Repository-Based

Models can be saved to file in a particular XML format, or they can be stored into a relational database: a repository. Using a repository can be a major advantage, especially when handling many models. Because the models are stored in a structured manner, arbitrary queries can be written against the repository. The repository may also be used to store the logs that are generated when executing transformations and jobs. Certain environments, such as banks, require that every manipulation that is performed with financial data be stored for longer periods of time for auditing purposes. The repository sure seems to be the place to do that, at least, as far as the ETL process is concerned.[21]

Pentaho Usage:



Step name: Text file input

Filetype: CSV

Separator: ,

Enclosure: "

Allow breaks in enclosed fields? □

Escape:

Header □ Number of header lines: 1

Footer □ Number of footer lines: 1

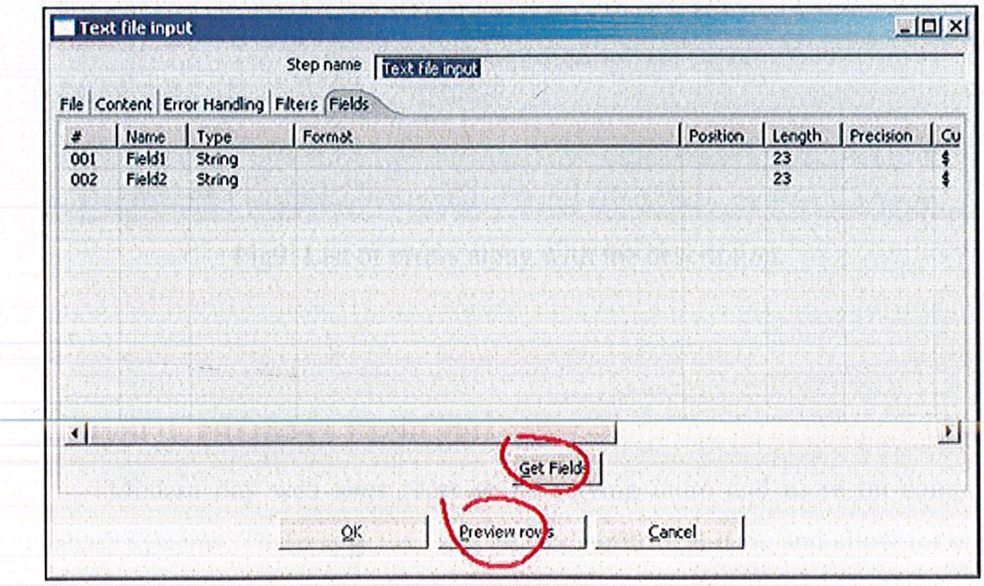
Wrapped lines? □ Number of times wrapped: 1

Paged layout (printout)? □ Number lines per page: 80

Document header lines: 0

Zipped □

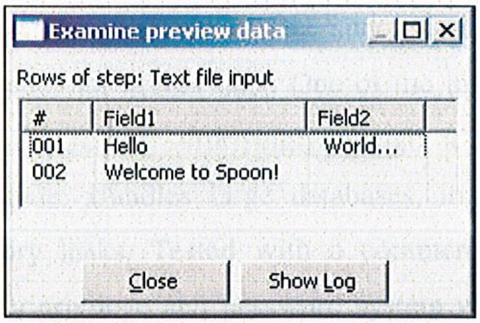
Fig6: Loading an Input file dialog box.



Step name: Text file input

#	Name	Type	Format	Position	Length	Precision	Cu
001	Field1	String			23		$
002	Field2	String			23		$

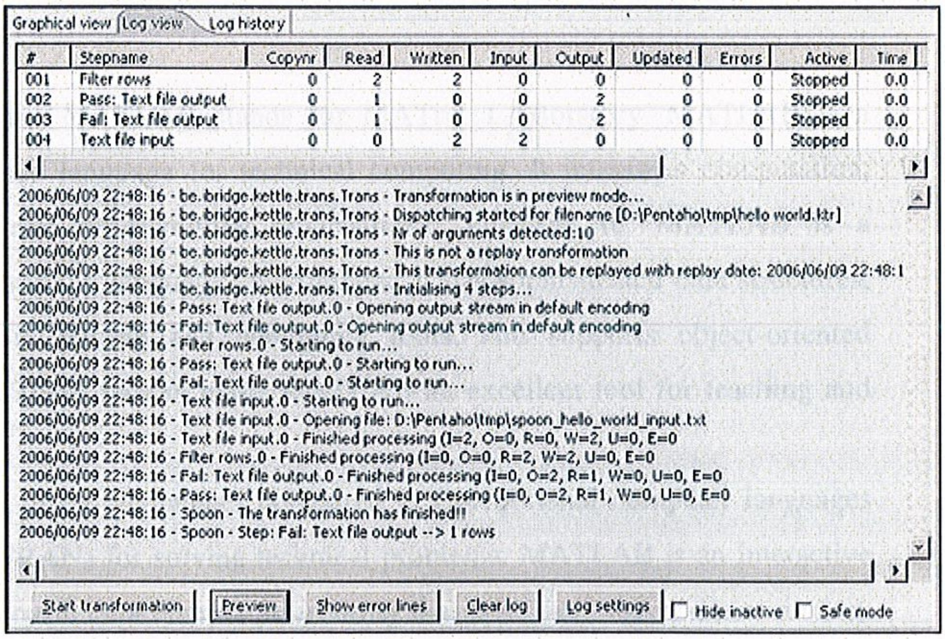
Fig7: Previewing the list of attributes.



Rows of step: Text file input

#	Field1	Field2
001	Hello	World...
002	Welcome to Spoon!	

Fig8: Preview of data will look like in a table.



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Errors	Active	Time
001	Filter rows	0	2	2	0	0	0	0	Stopped	0.0
002	Pass: Text file output	0	1	0	0	2	0	0	Stopped	0.0
003	Fail: Text file output	0	1	0	0	2	0	0	Stopped	0.0
004	Text file input	0	0	2	2	0	0	0	Stopped	0.0

```
2006/06/09 22:48:16 - be.ibridge.kettle.trans.Trans - Transformation is in preview mode...
2006/06/09 22:48:16 - be.ibridge.kettle.trans.Trans - Dispatching started for filename [D:\Pentaho\tmp\hello world.ktr]
2006/06/09 22:48:16 - be.ibridge.kettle.trans.Trans - Nr of arguments detected:10
2006/06/09 22:48:16 - be.ibridge.kettle.trans.Trans - This is not a replay transformation
2006/06/09 22:48:16 - be.ibridge.kettle.trans.Trans - This transformation can be replayed with replay date: 2006/06/09 22:48:1
2006/06/09 22:48:16 - be.ibridge.kettle.trans.Trans - Initialising 4 steps...
2006/06/09 22:48:16 - Pass: Text file output.0 - Opening output stream in default encoding
2006/06/09 22:48:16 - Fail: Text file output.0 - Opening output stream in default encoding
2006/06/09 22:48:16 - Filter rows.0 - Starting to run...
2006/06/09 22:48:16 - Pass: Text file output.0 - Starting to run...
2006/06/09 22:48:16 - Fail: Text file output.0 - Starting to run...
2006/06/09 22:48:16 - Text file input.0 - Starting to run...
2006/06/09 22:48:16 - Text file input.0 - Opening file: D:\Pentaho\tmp\spoon_hello_world_input.txt
2006/06/09 22:48:16 - Text file input.0 - Finished processing (I=2, O=0, R=0, W=2, U=0, E=0
2006/06/09 22:48:16 - Filter rows.0 - Finished processing (I=0, O=0, R=2, W=2, U=0, E=0
2006/06/09 22:48:16 - Fail: Text file output.0 - Finished processing (I=0, O=2, R=1, W=0, U=0, E=0
2006/06/09 22:48:16 - Pass: Text file output.0 - Finished processing (I=0, O=2, R=1, W=0, U=0, E=0
2006/06/09 22:48:16 - Spoon - The transformation has finished!!
2006/06/09 22:48:16 - Spoon - Step: Fail: Text file output --> 1 rows
```

Fig9: List of errors along with the description.

2.2 MySQL (RDBMS Package)

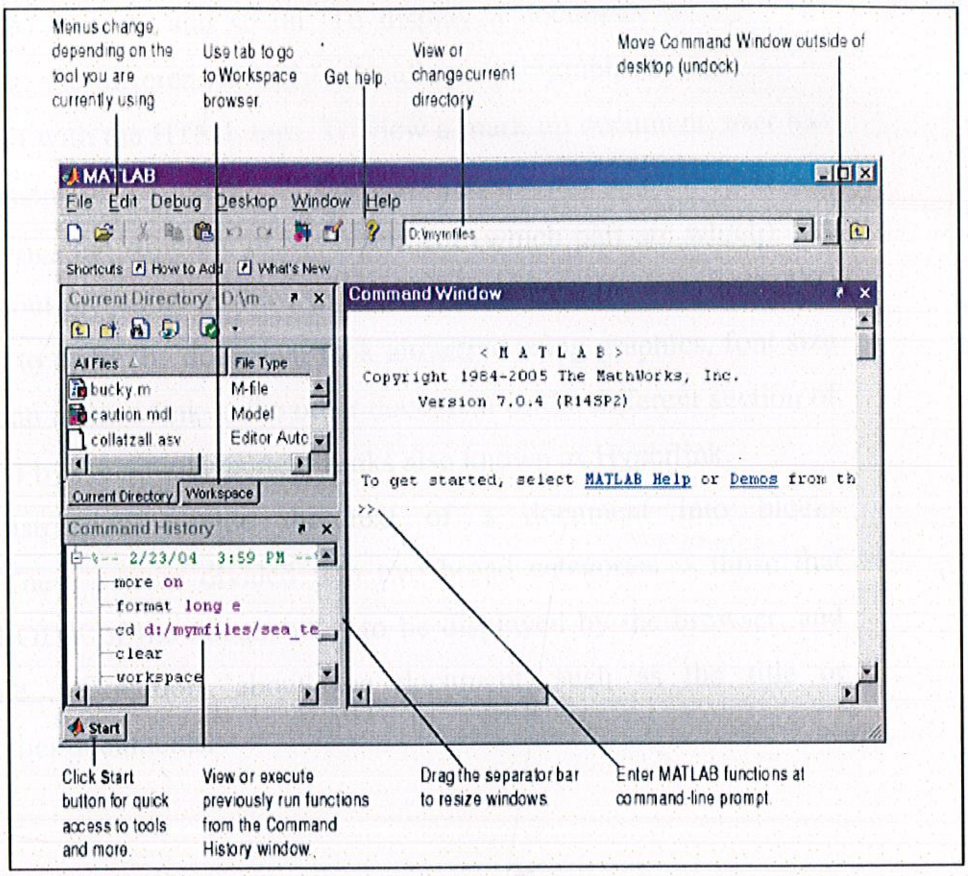
Modern day web sites seem to be relying more and more on complex database systems. These systems store all of their critical data, and allow for easy maintenance.

MySQL is a powerful Relational Database Management System (RDBMS), which uses the principles of database and data manipulation using Structured Query Language (SQL) statements. SQL is a database language that is used to retrieve, insert, delete and update stored data and its standardization makes it quite easy to store, update and access data. One of the most powerful SQL servers out there is called MySQL and surprisingly enough, it's free. Some of the features of MySQL Include: Handles large databases, in the area of 50,000,000+ records. No memory leaks. Tested with a commercial memory leakage detector (purify). It has a privilege and password system which is very flexible and secure, and which allows host-based verification. Passwords are secure since all password traffic when connecting to a server is encrypted.

2.3 MATLAB

The name MATLAB stands for MATrix LABoratory. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming environment. Furthermore, MATLAB is a modern programming language environment: it has sophisticated data structures, contains built-in editing and debugging tools, and supports object-oriented programming. These factors make MATLAB an excellent tool for teaching and research.

MATLAB has many advantages compared to conventional computer languages (e.g., C, FORTRAN) for solving technical problems. MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. [22]



Menus change depending on the tool you are currently using
Use tab to go to Workspace browser
Get help
View or change current directory
Move Command Window outside of desktop (undock)
MATLAB
File Edit Debug Desktop Window Help
D:\mymfiles
Shortcuts How to Add What's New
Current Directory - D:\m...
All Files
File Type
Current Directory
Workspace
Command History
Command Window
Start
Click Start button for quick access to tools and more.
View or execute previously run functions from the Command History window.
Drag the separator bar to resize windows
Enter MATLAB functions at command-line prompt

All Files	File Type
bucky.m	M-file
caution.mdl	Model
collatzall.asv	Editor Auto

```
%-- 2/23/04  3:59 PM --%
  more on
  format long e
  cd d:/mymfiles/sea_te
  clear
  workspace
```

```
                    < M A T L A B >
      Copyright 1984-2005 The MathWorks, Inc.
               Version 7.0.4 (R14SP2)


To get started, select MATLAB Help or Demos from th
>>
```

Fig10: MATLAB Interface

M-files

Matlab can also be used as a programming language. To program in Matlab you simply create a text file containing Matlab commands exactly as you would type them interactively in the Matlab window. The file may have any legal UNIX name, and should end with a .m extension. These files may be placed in your root directory, or a directory named /matlab. (Any other directories would have to be explicitly added to your Matlabpath.

2.4 HTML

HTML (Hypertext Markup Language) is used to create document on the World Wide Web. It is simply a collection of certain key words called 'Tags' that are helpful in writing the document to be displayed using a browser on Internet. It is a platform independent language that can be used on any platform such as Windows, Linux, Macintosh, and so on. To display a document in web it is essential to mark-up the different elements (headings, paragraphs, tables and so on) of the document with the HTML tags. To view a mark-up document, user has to open the document in a browser. A browser understands and interpret the HTML tags, identifies the structure of the document (which part are which) and makes decision about presentation (how the parts look) of the document. HTML also provides tags to make the document look attractive using graphics, font size and colors. User can make a link to the other document or the different section of the same document by creating Hypertext Links also known as Hyperlink.

HTML instructions divide the text of a document into blocks called *elements*. These can be divided into two broad categories -- those that define how the BODY of the document is to be displayed by the browser, and those that define information `about' the document, such as the title or relationships to other documents.

Html-CSS

CSS- CSS is the abbreviation for Cascading Style Sheet. A style sheet simply holds a collection of rules that we define to enable us to manipulate our web pages. CSS is a style language that defines layout of HTML documents. For example, CSS covers fonts, colors, margins, lines, height, width, background images, advanced positions and many other things. HTML can be (mis-)used to add layout to websites. But CSS offers more options and is more accurate and sophisticated. CSS is supported by all browsers today. [23]

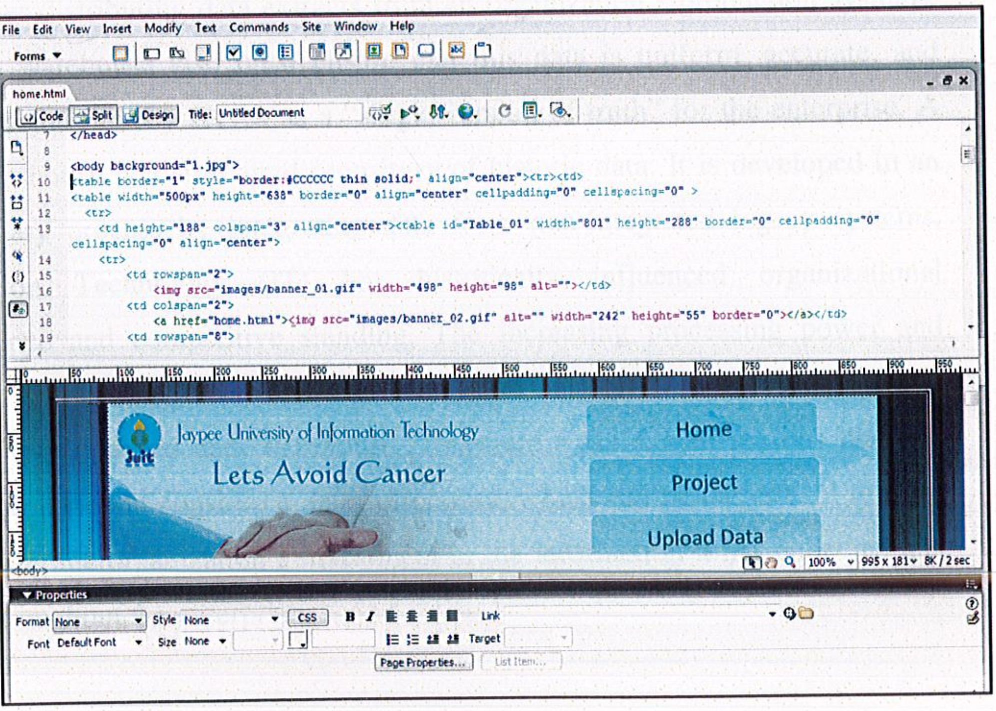
What is the difference between CSS and HTML?

HTML is used to structure content. CSS is used for formatting structured content. The language HTML was only used to add structure to text. CSS are a way to control the look and feel of your HTML documents in an organized and efficient manner. CSS was a revolution in the world of web design. The concrete benefits of CSS include:

- Control layout of many documents from one single style sheet;
- More precise control of layout;
- Apply different layout to different media-types (screen, print, etc.);
- Numerous advanced and sophisticated techniques.

Dreamweaver

Macromedia Dreamweaver is one of the top web design programs on the market, used by professionals and beginners alike. Dreamweaver is called a WYSIWYG (What You See Is What You Get) design environment, which means that the program will take care of converting your ideas into HTM, leaving you with time to do more important things. Dreamweaver is a powerful web page creation and web site management tool. It offers numerous, sophisticated functions that can be used to create professional quality web sites. Because of this, it's one of the most popular web authoring tools among web designers. You can use Dreamweaver to create a brand new web site, or to update a current site, even if it wasn't created in Dreamweaver.



```
</head>

<body background="1.jpg">
<table border="1" style="border:#CCCCCC thin solid;" align="center"><tr><td>
<table width="500px" height="638" border="0" align="center" cellpadding="0" cellspacing="0" >
  <tr>
    <td height="188" colspan="3" align="center"><table id="Table_01" width="801" height="288" border="0" cellpadding="0"
cellspacing="0" align="center">
    <tr>
        <td rowspan="2">
            <img src="images/banner_01.gif" width="498" height="98" alt=""></td>
        <td colspan="2">
            <a href="home.html"><img src="images/banner_02.gif" alt="" width="242" height="55" border="0"></a></td>
        <td rowspan="8">
```

Fig11: Dreamweaver Interface (Design of Cancer Cache Web Page)

CHAPTER 3

DEVELOPMENT OF DATA WAREHOUSE: *CANCER CACHE*

3.1 Introduction to Data Warehouse

To understand the data warehouse, it is important for us to realize that it is not a single object. It is more of a strategy or a process, an integration of various support systems and programs that are knowledge based. The goal of using a data warehouse is to allow businesses and organizations to make strategic decisions. The data warehouse concept originated in an effort to solve data synchronization problems and resolve data inconsistencies that resulted when analysts acquired data from multiple operational or production systems. One of the most important functions of a data warehouse is to serve as a collection point for consolidating and further distributing data extracts from an organization's production systems. The data warehouse also must ensure that this data is uniform, accurate, and consistent and thereby serves as a "single version of truth" for the enterprise. A Data Warehouse is a structured repository of historic data. It is developed in an evolutionary process by integrating data from non-integrated legacy systems. Information Technology (IT) has historically influenced organizational performance and competitive standing. The increasing processing power and sophistication of analytical tools and techniques have put the strong foundation for the product called data warehouse. There are a number of reasons that any organization should consider a data warehouse, which can be the critical tool for maximizing the organization's investment in the information it has collected and stored throughout the enterprise.

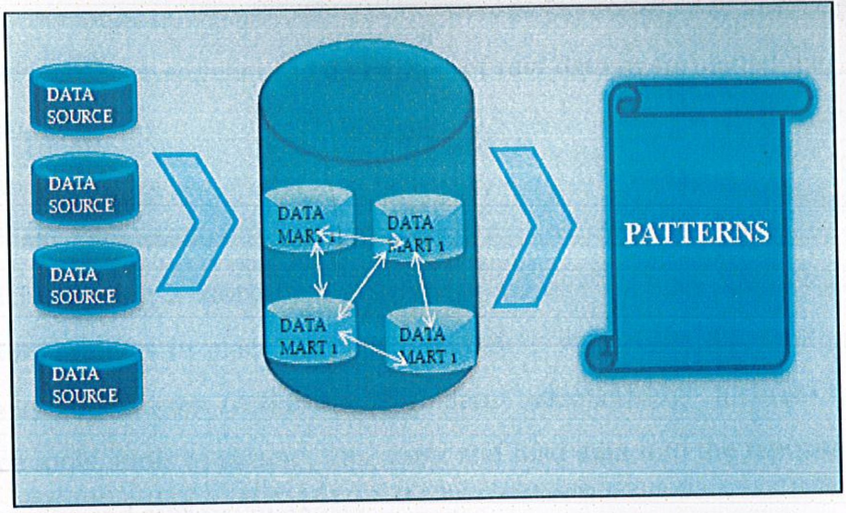


Fig12: A Schematic Representation of Data warehouse model

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

The meanings of the key terms are as defined below:

1. **Subject-Oriented**: Organization of data in a warehouse is around the key subjects (or high-level entities) of the enterprise. For instance: patients, students, and products.

2. **Integrated**: The data is assumed to be using consistent naming conventions, formats, encoding structures, and related characteristics for sharing and usability.

3. **Time-variant:** Data contain a time dimension so that they can be used for historical purposes.

4. **Nonvolatile**: Data are refreshed from operational data, and cannot be updated by users. Considering the above key terms data warehousing could be defined as the process by which an organization extract meaningful information from historical data.

Data warehouses provide access to data for complex analysis, knowledge discovery, and decision-making to build a Decision Support System (DSS). DSS is a technique used by organizations to come up with facts, trends or relationships that can help them make effective decisions or create effective strategies to accomplish their organizational goals. Usually a data warehouse is either a single computer or many computers (servers) tied together to create one giant computer system.

Data in a Data warehouse can consist of raw data or formatted data. It can be on various types of topics including organization's sales, salaries, operational data, summaries of data including reports, copies of data, human resource data, inventory data, external data to provide simulations and analysis, etc.

Besides being a store house for large amount of data, Data warehouse must possess systems in place that make it easy to access the data and use it in day to day operations. Definition for data warehousing includes business intelligence tools, tools to extract, transform and load data into the repository, and tools to manage and retrieve metadata.

While operational databases played an important role in the past, they are not used directly for information processing within modern data warehouses. In most cases, they will merely act as a repository for data, and they will also play a fundamental role in information processing. There are a number of reasons why a company should want to separate operation data bases from those that are information based. One of the most important reasons for this is because the users of both forms of data are different. While analysts will often be responsible for working with informational data, administrative employees will spend more time working with operational data.

The data must be cleaned and transformed in a way that allows it to remain accurate. The consistency of data within the data warehouse is extremely important. Summarization also plays an important role in the function of the data warehouse, and it is important for the user to make sure the correct level is created, thus allowing the organization to make important business decisions. Another issue with many data warehouses is user friendliness. Most data warehouses are subject oriented. This means that the information that is in the data warehouse is stored in a way that allows it to be connected to objects or events which occur in reality. Another characteristic that is frequently seen in data warehouses is called a time variant. A time variant will allow changes in the information to be monitored and recorded over time. The information that exists in data warehouses is non-volatile. This means that it cannot be deleted, and must be held to be analyzed in the future. All of the programs that are used by a particular institution will be stored in the data warehouse, and it will be integrated together. The first data warehouses were developed in the 1980s. As societies entered the information age, there was a large demand for efficient methods of storing information. [24]

The quality of decisions that are facilitated by a Data warehouse is only as good as the quality of the Data.

3.2 How Does a Data Warehouse Differ From a Database?

There are a number of fundamental differences which separate a data warehouse from a database. The biggest difference between the two is that most databases place an emphasis on a single application, and this application will generally be one that is based on transactions. If the data is analyzed, it will be done within a single domain, but multiple domains are not uncommon.

Some of the separate units that may be comprised within a database include payroll or inventory. Each system will place an emphasis on one subject, and it will not deal with other areas. In contrast, data warehouses deal with multiple domains simultaneously. Because it deals with multiple subject areas, the data warehouse finds connections between them. This allows the data warehouse to show how the company is performing as a whole, rather than in individual areas. Another powerful aspect of data warehouses is their ability to support the analysis of trends. They are not volatile, and the information stored in them doesn't change as much as it would in a common database. The two types of data that we will want to become familiar with is operational data and decision making data. The purpose, format, and structure of these two data types are quite different. In most cases, the operational data will be placed in a relational database. While operational data will deal mostly with transactions that are made daily, decision support data will give meaning to the data that is operational. The differences between decision support data and operational data can be split into three categories, and these are dimensionality, time span, and granularity.

Dimensionality is a concept which shows that the data is connected in various ways. The data that is stored in a data warehouse will often be multidimensional, and it is much different than the simple view that is often seen with operational data. Many data analysts are concerned with the many dimensional aspects of data. The time span deals with transactions that are atomic, or current. These transactions will deal with things such as the inventory movement, or the purchase of an order. Generally, operational data will deal with a short time frame. However, decision support data tends to deal with long time frames. Granularity is the third concept that separates operational data from decision support data. Operational data will deal with transactions that have occurred within a certain period of time. However, the decision support data must be broken down into different parts of aggregation. While it may be summarized, it may also be more current. Data warehouses have become more important in the Information Age, and they are a necessity for many large corporations, as well as some medium sized businesses. They are much more elaborate than a mere database, and they can find connections in data that cannot be readily found within most databases.

Data warehouses are useful because they collect data and remodel it. The information is placed in a single unit, and we can get a clear picture of how their company is performing. Most importantly, they will be able to make decisions with a great deal of confidence. Data will be stored in the warehouse from multiple sources. Once the data is stored, it must be cleaned and transformed.

The process of cleaning and transforming data is known as ETL, or Extraction, Transformation, and Loading. Properly caring for the data is an important part of maintaining a successful data warehouse. Most companies store data for the long term, and they follow set rules and procedures. The data warehouse is specifically designed to give information as a single entity. Data will be placed in the warehouse periodically, and it will be done in batches. In most cases, the data will be stored at times when the company isn't extremely busy. The data is considered to be non-volatile. One of the most powerful benefits of a data warehouse is the fact that operational forms of data can be optimized for a certain level of efficiency. [24]

Another important concept that we should be familiar with is metadata. Metadata can be defined as the information on the data that is stored in the warehouse. In other words, it's the data about data. Metadata can be broken down into three categories, and this is operational, business, and administrative. The administrative is related to the columns and tables of the warehouse, and it also deals with the rules by which the data is maintained. As the second name implies, business metadata deals with various business terms. This data is especially important to those who will be making the key decisions. The operational metadata deals with the errors, history, and usage. As the name suggests, it deals with the operational issues surrounding the data. There may be different needs for the data; therefore we may construct smaller data warehouses that are tailored towards certain subjects. These small data warehouses are referred to as being **data marts**. The data mart will get its information from the central data warehouse that is being used. The last part of the data warehouse is the decision support program. These programs will get their information from the data warehouse, as well as the data marts. They will take the information they are given, and they will use it for querying purposes.

The decision support programs will fall under one of three categories, and these are data mining, SQL, or OLAP. These applications are designed in a way that will allow the type of user getting important answers to their questions. These answers can assist them in the decision making process. The data can be presented in such a way, that it allows the decision makers to look at summarized data before looking for information that is much more specific. Data mining is quite powerful because it allows an AI or neural network to sift through the data looking for important trends or relationships, connections that are impossible for humans to find within a short time period. Data mining will typically use logistic regression or specific algorithms.

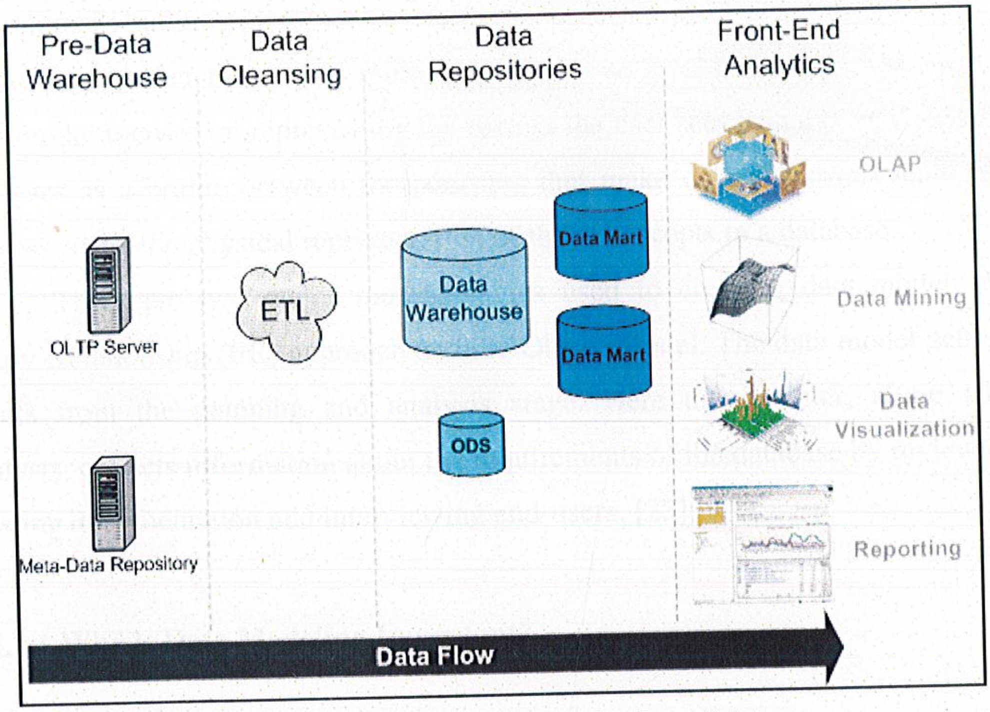


Fig13: Representation for flow of data.

3.3 Data Marts

A **data mart** (DM) is the access layer of the data warehouse (DW) environment that is used to get data out to the users. A data mart is a focused subset of a data warehouse that deals with a single area (like different department) of data and is organized for quick analysis

3.4 Creating a Data warehouse

3.4.1 Prerequisites - the data model

A data model is a conceptual representation of the data structures that are required by a warehouse. The data structures include the data objects, the associations between data objects, and the rules which govern operations on the objects. As the name implies, the data model focuses on what kind of data is required and how it should be organized rather than what operations will be performed on the data. To use a common analogy, the data model is equivalent to an architect's building plans. A data model is independent of hardware or software constraints. Rather than try to represent the data as a database would see it, the data model focuses on representing the data as the user sees it in the "real world". It serves as a bridge between the concepts that make up real-world events and processes and the physical representation of those concepts in a database.

There are two major methodologies used to create a data model: the Entity-Relationship (ER) approach and the Object Model. The data model gets its from the planning and analysis stage. Here the modeler, along with inputs from the planning and analysis stage. Here the modeler, along with analysts, collects information about the requirements of the database by reviewing existing documentation and interviewing end-users. [25]

3.4.1.1 Why is Data Modeling Important?

Data modeling is probably the most labor intensive and time consuming part of the development process. Why bother especially if you are pressed for time? A common response by practitioners who write on the subject is that you should no more build a database without a model than you should build a house without blueprints. The goal of the data model is to make sure that the all data objects required by the database are completely and accurately represented. Because the data model uses easily understood notations and natural language, it can be reviewed and verified as correct by the end-users. The data model is also detailed enough to be used by the database developers to use as a "blueprint" for building the physical database. The information contained in the data model will be used to define the relational tables, primary and foreign keys, stored procedures, and triggers. A poorly designed database will require more time in the long term. Without careful planning you may create a database that omits data required to create critical reports, produces results that are incorrect or inconsistent, and is unable to accommodate changes in the user's requirements.

There are different data modeling concepts like ER Modeling (Entity Relationship modeling) DM (Dimensional modeling) Hierarchal Modeling Network modeling. But popular are ER and DM only.

3.4.1.2 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.

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.

3.4.2 Image Processing

3.4.2.1 Introduction

Image processing is any form of signal processing for which the input is an image, such as a photograph or video frame; the output of image processing may be either an image or, a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it. [26]

Fig14: Image Processing

Image processing typically attempts to accomplish one of three things:

- Restoring images: Restoration takes a corrupted image and attempts to recreate a clean original
- Enhancing images : Enhancement alters an image to makes its meaning clearer to human observers
- Understanding images: Understanding usually attempts to mimic the human visual system in extracting meaning from an image.

3.4.2.2 Role of Image processing in medical field

Medical Imaging

Imaging technology in Medicine made the doctors to see the interior portions of the body for easy diagnosis. It also helped doctors to make keyhole surgeries for reaching the interior parts without really opening too much of the body. CT scanner, Ultrasound and Magnetic Resonance Imaging took over x-ray imaging by making the doctors to look at the body's elusive third dimension. With the CT scanner, body's interior can be bared with ease and the diseased areas can be identified without causing either discomfort or pain to the patient. MRI picks up signals from the body's magnetic particles spinning to its magnetic tune and with the help of its powerful computer, converts scanner data into revealing pictures of internal organs. Image processing techniques developed for analyzing remote sensing data may be modified to analyze the outputs of medical imaging systems to get best advantage to analyze symptoms of the patients with ease.

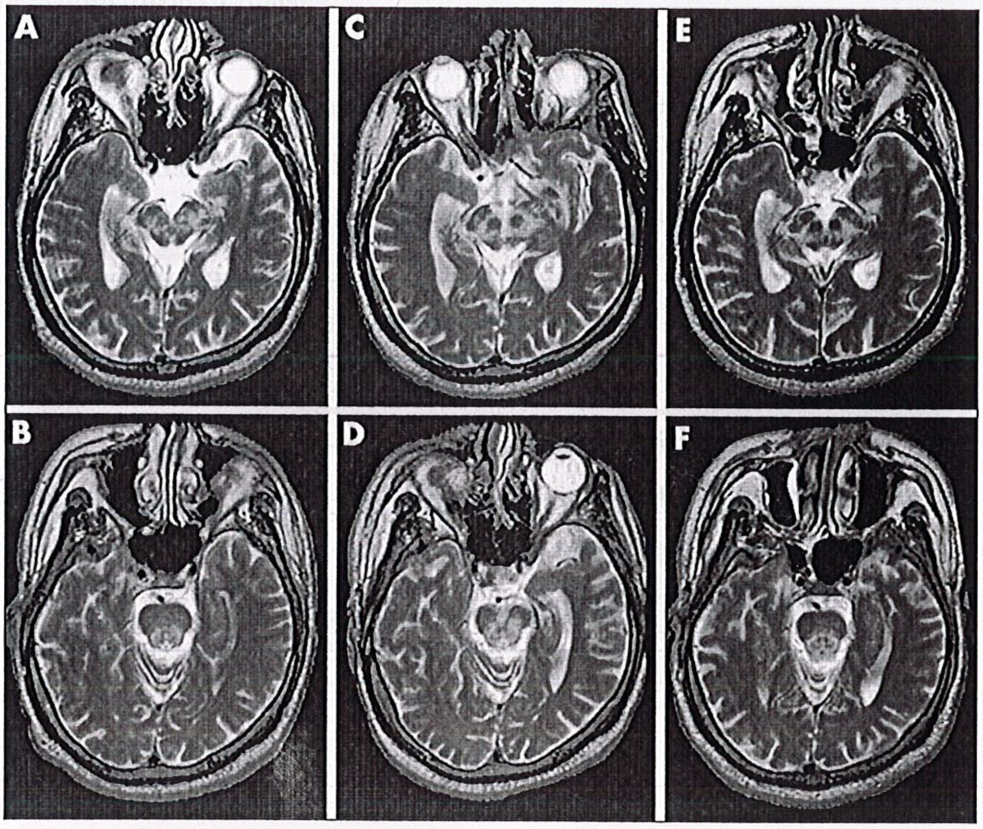


Fig15: Medical Imaging

3.4.2.3 Advantages of Digital Processing for Medical Applications

- Digital data will not change when it is reproduced any number of times and retains the originality of the data.
- Offers a powerful tool to physicians by easing the search for representative images
- Displaying images immediately after acquiring;
- Enhancement of images to make them easier for the Physician to interpret;
- Quantifying changes over time
- Providing a set of images for teaching to demonstrate examples of diseases or features in any image;
- Quick comparison of images

The Digital Imaging and Communications in Medicine (DICOM) standard was created by the National Electrical Manufacturers Association (NEMA). Its aim is to support the distribution and viewing of medical images from CT, MRI and other medical modalities [27].

3.4.3 Identification of Data Source

The data that we managed to gather varied a lot. Some of the data was in the electronic form, some in hardcopy written by the doctors, some were printed reports, and even the images were in different formats. Eg. Jpeg, dicom etc. The following figure gives a glimpse of the kind of data collected:

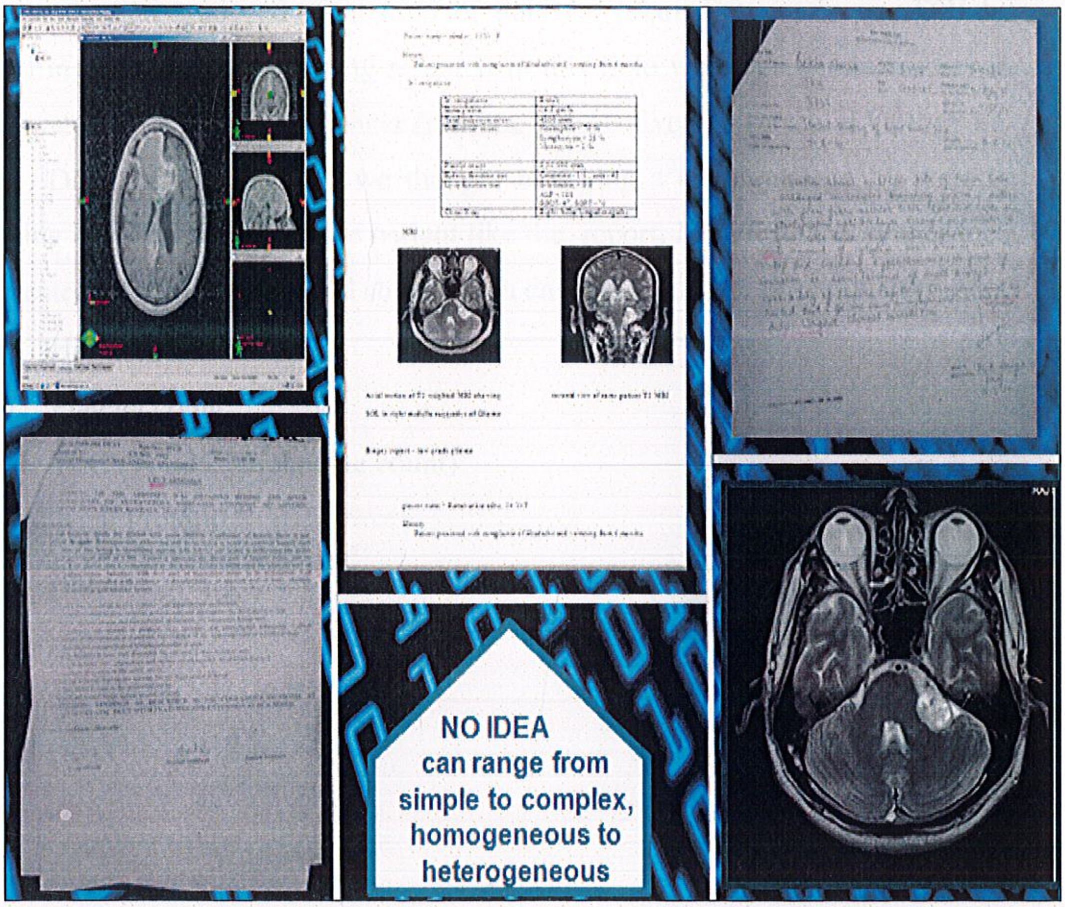


Fig16: Different types of data

3.5 Methodology Used for Data Warehouse Construction

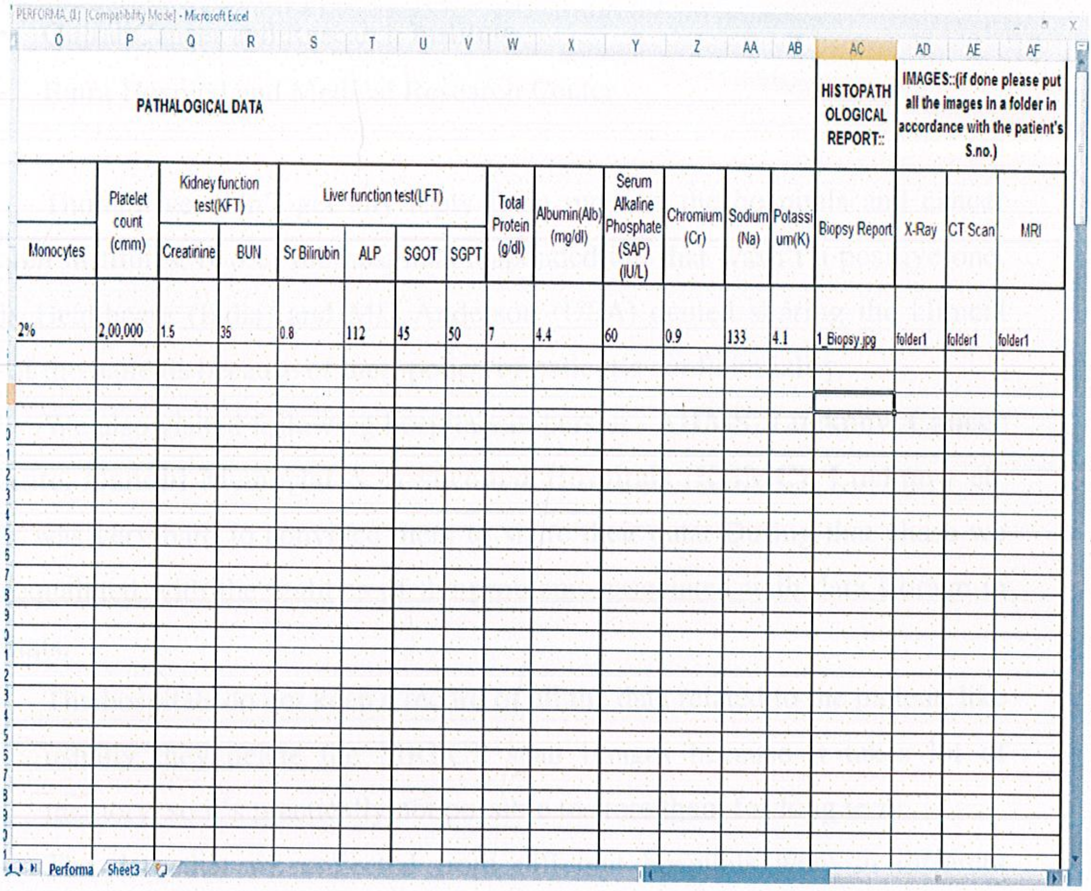
Cancer may not be the reason for the highest number of deaths, but almost everybody surrounding us knows someone or the other suffering from cancer. Not all cancer cases ends up in death, some can be prevented if they are detected at an early stage i.e. early detection can help us to bring down the cancer death statistics. This inspired us to work on cancer. We made a Cancer Data warehouse, which has a collection of the cancer patient's clinical health records, from which we mined useful information so as to draw something beneficial from the huge amounts of data. The methodology is explained below:

3.5.1 Step 1#: Data Collection

We first analyzed what kind of data we require to put in our DW by consulting a doctor. After doing research in this field we came across studies in which research was done in cancer from the perspective of Molecular biology i.e. genes, DNA markers etc. So we thought of taking a different lane by majorly focusing on Clinical data of the patient like the report, MRI/CT scan images etc., we created a Performa in Excel sheet which enlisted a number of parameters, like:

- Hemoglobin
- Platelet count
- TLC (Total Lymphocyte count)
- Total Protein
- Etc.

Performa used for collection of data:



PERFORMA_(1) [Compatibility Mode] - Microsoft Excel

O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
PATHALOGICAL DATA														HISTOPATHOLOGICAL REPORT::	IMAGES::(if done please put all the images in a folder in accordance with the patient's S.no.)		
	Platelet count (cmm)	Kidney function test(KFT)		Liver function test(LFT)				Total Protein (g/dl)	Albumin(Alb) (mg/dl)	Serum Alkaline Phosphate (SAP) (IU/L)	Chromium (Cr)	Sodium (Na)	Potassium(K)	Biopsy Report	X-Ray	CT Scan	MRI
Monocytes		Creatinine	BUN	Sr Bilirubin	ALP	SGOT	SGPT										
2%	2,00,000	1.5	35	0.8	112	45	50	7	4.4	60	0.9	133	4.1	1_Biopsy.jpg	folder1	folder1	folder1

Fig 17: Performa

The Performa was uploaded on Google Docs. Afterwards we tried to contact a number of cancer hospitals, Research Institutes and Path Labs by sending our Performa through mail requesting for the data. Some of the Centers where we contacted were:

- Dharamshila Cancer Foundation and research center
- Fortis Hospital
- Apollo Hospitals
- Amala Cancer Institute
- King George Medical College(KGMC)
- Jawaharlal Nehru Cancer Hospital and research Center

- Lucknow Cancer Institute
- MD Anderson Cancer Center
- Mohan Dai Oswal Cancer Treatment and Research Foundation
- Gujrat Cancer and Research Institute
- Batra Hospital and Medical Research Center

Though we didn't get any reply from most of the hospitals and cancer research institutes, two of the hospitals responded but that wasn't a positive one. Fortis Healthcare (India) and MD Anderson (USA) denied sharing the clinical data of the patients because of their policy of patient's confidentiality.

We also visited following hospitals in person - **AIIMS, Lucknow Cancer Institute, Gandhi Memorial & Associated Hospitals (KGMC), Lucknow** etc and it was very hard to convince them to share their data. During that phase we got acquainted with the realities of the problems associated with data storage in hospitals.

- The hospitals do not keep a record of all the data related to the patient, like usually they delete the MRI/CT scan images because it takes lot of memory, so it's practically not possible to store them for long term.
- The data that we collected from different hospitals were in different formats. Like Electronic records, Images, Hardcopies (Printed) and handwritten, adding to the complexity even the formats of the images varied (for eg. DICOM, jpeg, png etc.). Even in some cases we had to load the data from the reports into an excel sheet manually.
- It was very difficult to get integrated data i.e. to collect same patient's blood report, MRI/CT scan etc. In one of the hospitals the ids were given in particular to the department for eg., If a patient is prescribed for biopsy then an id is assigned corresponding to this dept., now if the same patient was prescribed for MRI then they assigned a different id respective to MRI-this created a lot of chaos in keeping the record of the same patient as the same patient gets multiple ids. We have also tried to address this problem.

- Out of all, the major problem still remains the confidential policy of the health care institutes in the process of data collection.

Stating the above problem it has been learning experience knowing about the real-time issues associated which if appropriately dealt can lead us to development of effective systems & solutions for clinical informatics.

We were successfully able to obtain information with respect to around **100** patients, among whom **42** were suffering from cancer while remaining are healthy individuals with respect to the disease.

3.5.2 Step 2#: Data Modeling

After getting the data from multiple sources we then designed a Data model for our data warehouse. This data model is basically a blueprint of how our data warehouse is structured. We used Dimensional modeling to create our data model i.e. Schema. A schema is a diagrammatic representation of the tables (Fact tables and Dimension tables) and the relationships between them. The following diagram displays the schema of our data warehouse:

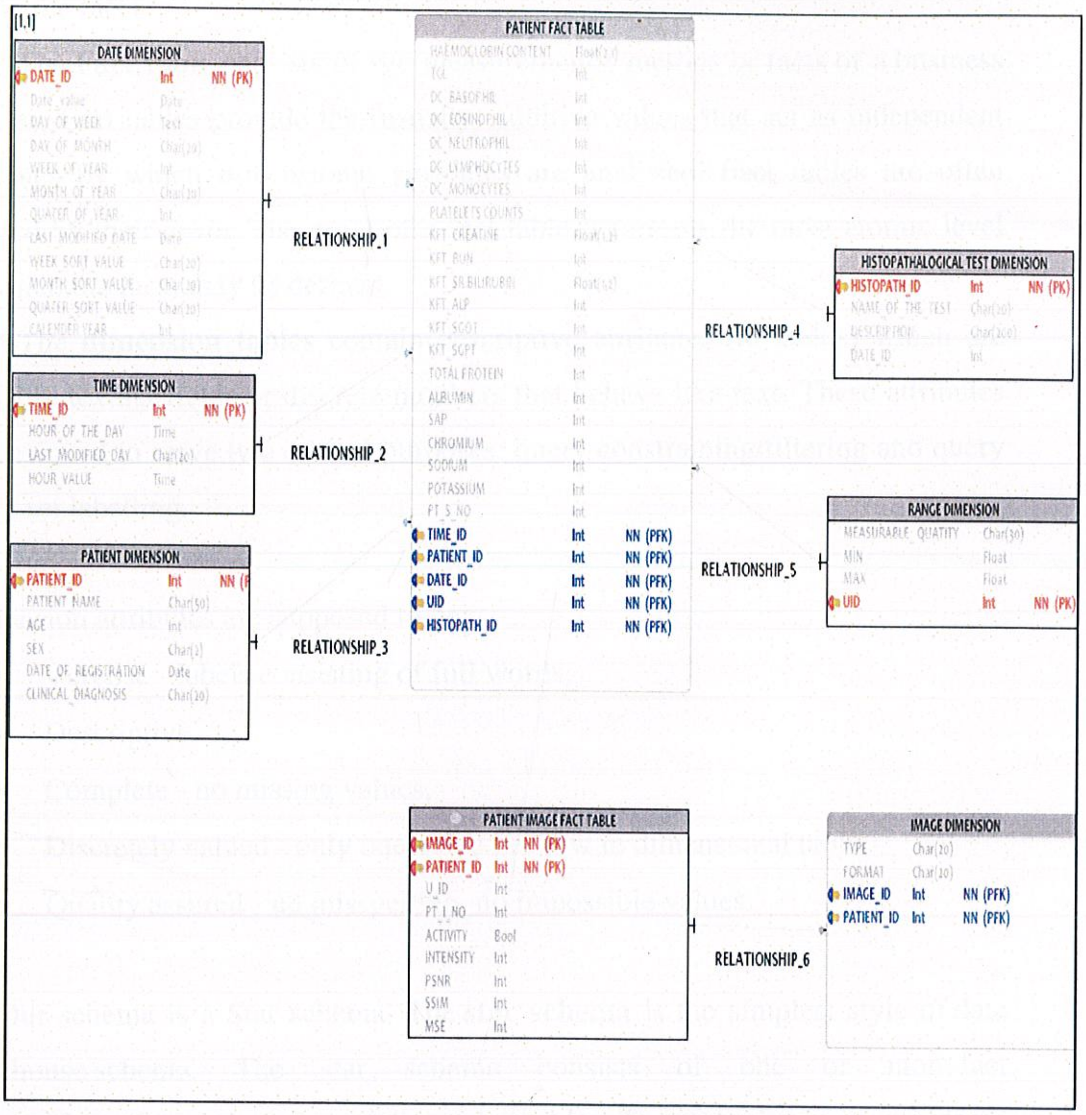


Fig 18: Dimensional Model (Star Schema Design)

A **schema** of a database system is its structure described in a formal language supported by the database management system (DBMS) and refers to the organization of data to create a blueprint of how a database will be constructed (divided into database tables). The formal definition of database schema is a set of formulas (sentences) that specify integrity constraints imposed on the database.

Dimensional modeling always uses the concepts of facts (measures), and dimensions (context). Facts are typically (but not always) numeric values that can be aggregated, and dimensions are groups of hierarchies and descriptors that define the facts.

The **Fact table** consists of the measurements, metrics or facts of a business process. Fact tables provide the (usually) additive values that act as independent variables by which dimensional attributes are analyzed. Fact tables are often defined by their *grain*. The grain of a fact table represents the most atomic level by which the facts may be defined.

The **dimension tables** contain descriptive attributes (or fields) which are typically textual fields or discrete numbers that behave like text. These attributes are designed to serve two critical purposes: query constraining/filtering and query result set labelling.

Dimension attributes are supposed to be:

- Verbose - labels consisting of full words,
- Descriptive,
- Complete - no missing values,
- Discretely valued - only one value per row in dimensional table,
- Quality assured - no misspelling, no impossible values.

Our schema is a Star schema. The **star schema** is the simplest style of data warehouse schema. The star schema consists of one or more fact tables referencing any number of dimension tables. The star schema is considered an important special case of the snowflake schema, and is more effective for handling simpler queries.

The star schema is a way to implement multi-dimensional database (MDDB) functionality using a mainstream relational database: given most organizations' commitment to relational databases, a specialized multi-dimensional DBMS is likely to be both expensive and inconvenient. Another reason for using a star schema is its simplicity for users: queries are never complex because the only joins and conditions involve a fact table and a single level of dimension tables, without the indirect dependencies to other tables that are possible in a better normalized snowflake schema.

In our schema, the Patient fact table and the Patient Image fact table, which are the fact tables, are referencing four and one dimension table respectively. Patient_Id serve as the primary key of the Patient dimension table, which is the unique id that is provided to each patient and this is the id which is majorly linking all other information related to that patient. Patient fact table is connected to Date dimension, Patient Dimension, Test Dimension and Range dimension. While creating a data warehouse Date dimension tables are created by default. Date_Id, Patient_ID, Histopath_ID and uid are the primary keys of Date Dimension, Patient Dimension, Dimension and Range Dimension respectively and they are the foreign keys for the Patient Fact Table. The Image Dimension is linked to Patient Image Fact Table; here Patient_ID and Image_ID (in combination) act as the composite key.

3.5.3 Step 3#: Data Preprocessing

The data was in form of MRI/CT scan images and histopathological parameters, which were filled in the excel sheet Performa. We first converted the data in the excel sheet into comma separated file format i.e. in the CSV format. After which it was processed into the 'staging database'.

The **staging database** is a separate data cache (storage area) that helps users in continuous access to application data. Their access continues even when data is being imported from the various external sources and prepared for loading. This minimizes the downtime that users experience during data loading or data refreshing. Here the data is dumped as it is, without any changes being made to it i.e. the data here is in its original form.

We then extracted the data from the staging schema and then processed it, cleaned it so that it could further be used for analysis.

The process of cleaning and transforming data is known as ETL, or Extraction, Transformation, and Loading. Properly caring for the data is an important part of maintaining a successful data warehouse.

The data that we collected from different hospitals could not be used for analysis because "data in the real -world is dirty" i.e. have errors, unusual values, and inconsistencies. Data quality can be assessed in terms of accuracy, completeness and consistency. The data that we have was:

- **Incomplete**: Some of the records were lacking some attribute values. E.g. few of the patient's records did not have their 'BASOPHIL' count while few had 'Sr Bilirubin'. This leads to incompleteness in the data
- **Noisy**: Means that the data contains errors or outliers E.g. One record had value 10 in platelet count which was an error
- **Inconsistent**: Containing discrepancies in codes or names or format E.g.. The date in few records was in mm/dd/yy format and in few dd/mm/yyyy format

There are many reasons which might lead to such kind of data. Incomplete, noisy, and inconsistent data are commonplace properties of large, real-world databases and data warehouses. Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, such as Kidney function test data may not be done by. Other data may not be included simply because it was not considered important at the time of entry. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the history or modifications to the data may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.

Data can be noisy, having incorrect attribute values, owing to the following: The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Errors in data transmission can also occur. There may be technology limitations, such as limited buffer size for coordinating synchronized data transfer and consumption. Incorrect data may also result from inconsistencies in naming conventions or data codes used. Level of redundancy is another important factor in data processing .It is useful to know how much of the data is repeated from the various sources .Redundant data can slow down or confuse the knowledge discovery process. Data reduction and cleaning methods, carefully employed, can aid in removing duplicated data prior to its usage.

Major Tasks in Data Preprocessing

- Data cleaning: Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data integration: Integration of multiple databases, data cubes, or files.
- Data transformation: Normalization and aggregation.
- Data reduction: Obtains reduced representation in volume but produces the same or similar analytical results.
- Data discretization: Part of data reduction but with particular importance, especially for numerical data.

We have used **Pentaho 3.0 (Kettle)** and **MySQL 5.019** to preprocess and store the data respectively. A staging database in MySQL by the name **"Cancer"** was created which holds several tables each corresponding to the source from where we got the data; for e.g. kgmc_data1 contains the data collected from King George Medical College (KGMC).

Information about various tables in staging schema along with type of information being stored in them is as follow:

Table 1: Tables in the staging schema "cancer"

TABLE NAME	FUNCTION OF THE TABLE
Date	This table stores the date records. Here each date has been given a unique id and other details like week of the year, in which quarter of the year is the dates falling are given.
kgmc_data1	It stores the data collected from KGMC with details of patients records
kgmc_new	Most recent data from kgmc has been dumped in this table
lady	It stores the data collected from another hospital.
lady1	Most recent data from the same hospital
lci_control	Stores the records of healthy patient which do not have cancer , and act as a control
lucknow _cancer _institute	Stores cancerous patients health records

The next step was to designs mappings (transformation) in Pentaho. Transformation applies a series of rules or functions to the extracted data from the source to derive the data for loading into the end target. Some data sources required very little or even no manipulation of data. In other cases, one or more of the following transformation types was designed to meet the needs of the target database. [28]

We made a transformation by the name "kgmc data" in which we processed data which was in the .csv format was loaded into the kgmc_data1 table in cancer database.

The transformation looked like:

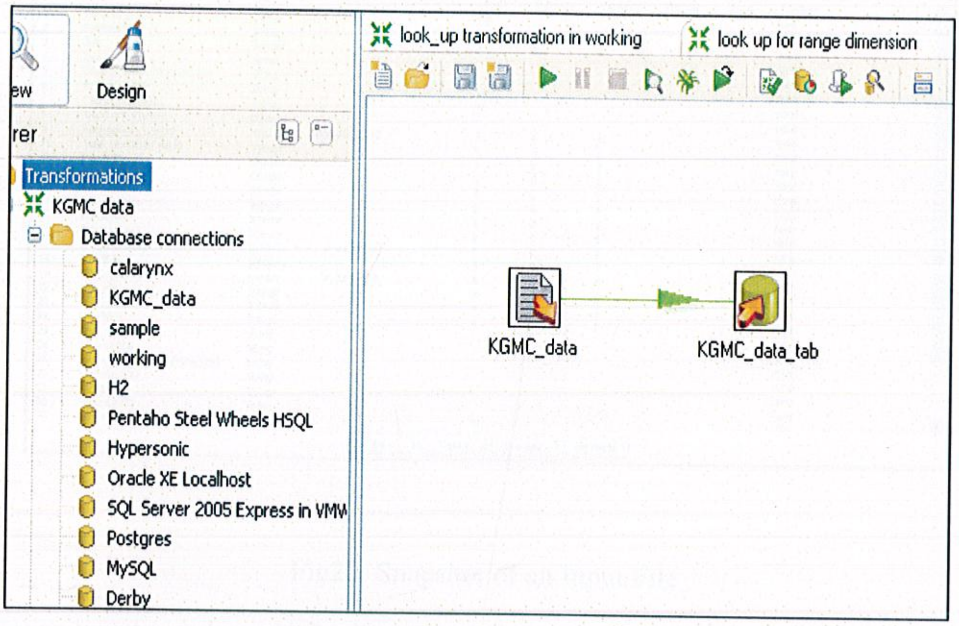
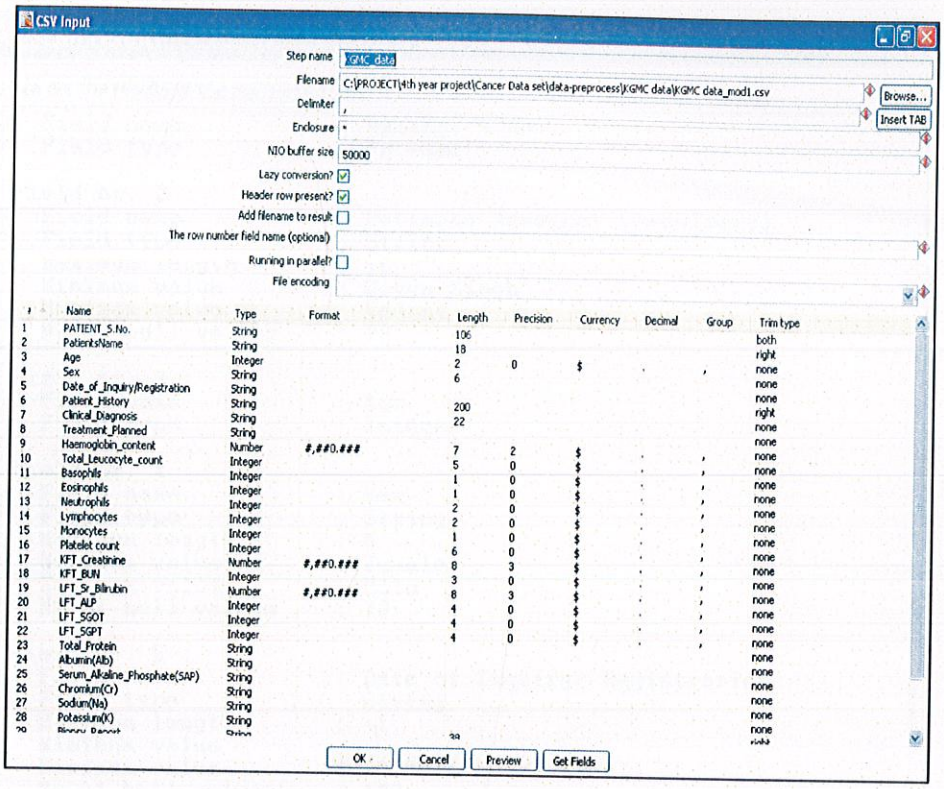


Fig19: Snapshot of a Transformation

The “kgmc_data” contains the csv format file where we mentioned various details such as the csv file name, name of this input etc.

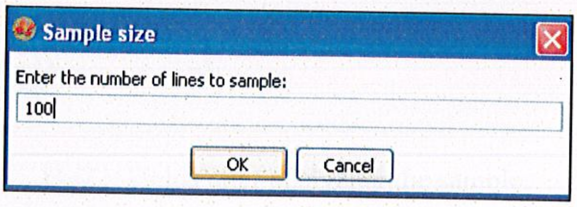


Step name: KGMC data
Filename: C:\PROJECT\4th year project\Cancer Data set\data-preprocess\KGMC data\KGMC data_mod1.csv
Delimiter: ,
Enclosure: "
NIO buffer size: 50000
Lazy conversion? ☑
Header row present? ☑
Add filename to result □
The row number field name (optional)
Running in parallel? □
File encoding

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	PATIENT_S.No.	String		106					both
2	PatientsName	String		18					right
3	Age	Integer		2	0	$	.	,	none
4	Sex	String		6					none
5	Date_of_Inquiry/Registration	String							none
6	Patient_History	String		200					right
7	Clinical_Diagnosis	String		22					none
8	Treatment_Planned	String							none
9	Haemoglobin_content	Number	#,##0.###	7	2	$	.	,	none
10	Total_Leucocyte_count	Integer		5	0	$	.	,	none
11	Basophils	Integer		1	0	$	.	,	none
12	Eosinophils	Integer		1	0	$	.	,	none
13	Neutrophils	Integer		2	0	$	.	,	none
14	Lymphocytes	Integer		2	0	$	.	,	none
15	Monocytes	Integer		1	0	$	.	,	none
16	Platelet count	Integer		6	0	$	.	,	none
17	KFT_Creatinine	Number	#,##0.###	8	3	$	.	,	none
18	KFT_BUN	Integer		3	0	$	.	,	none
19	LFT_Sr_Bilirubin	Number	#,##0.###	8	3	$	.	,	none
20	LFT_ALP	Integer		4	0	$	.	,	none
21	LFT_SGOT	Integer		4	0	$	.	,	none
22	LFT_SGPT	Integer		4	0	$	.	,	none
23	Total_Protein	String							none
24	Albumin(Alb)	String							none
25	Serum_Alkaline_Phosphate(SAP)	String							none
26	Chromium(Cr)	String							none
27	Sodium(Na)	String							none
28	Potassium(K)	String							none
29	Biopsy Report	String		39					right

Fig20: Snapshot of an Input File

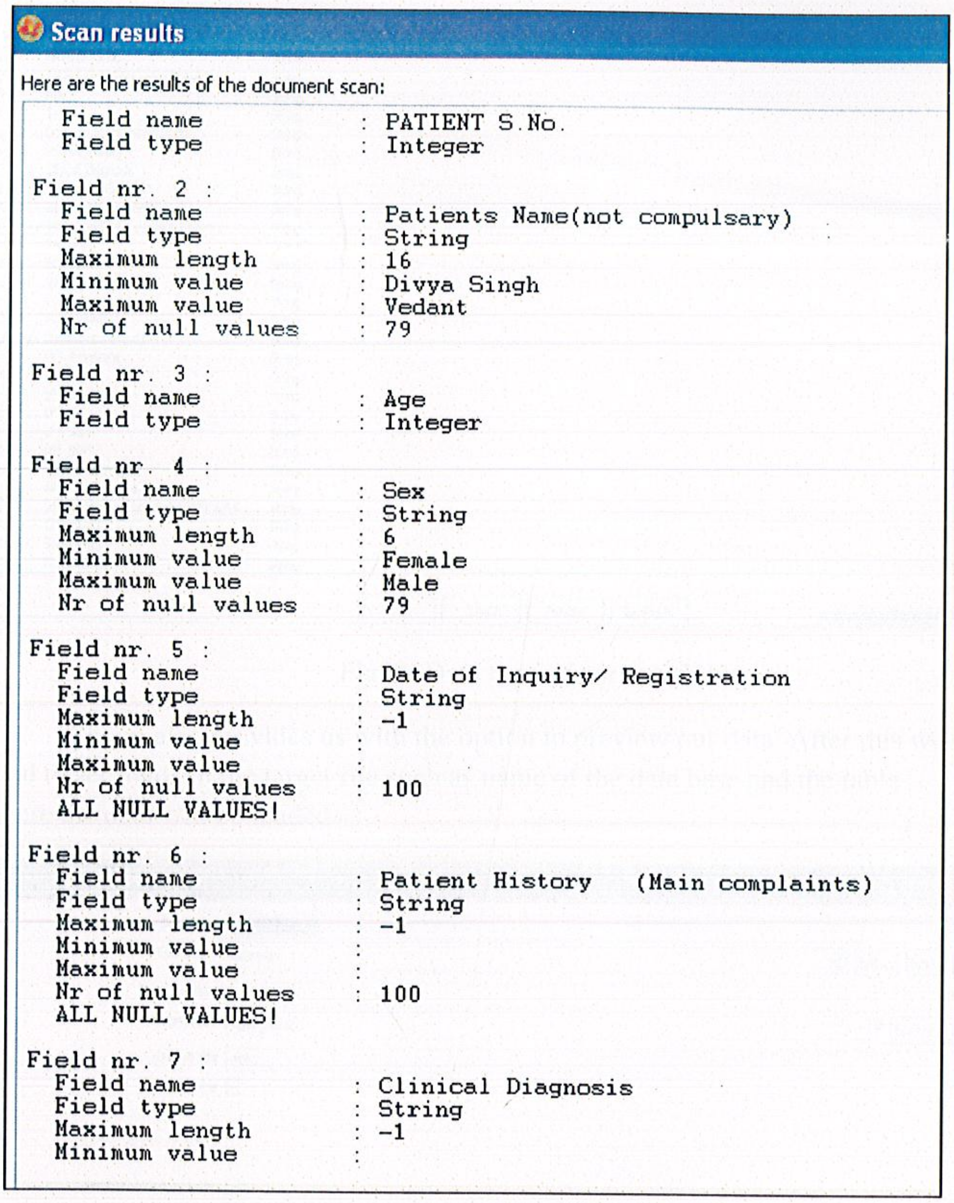
Spoon has the option of "Get Fields" where we can get all the attributes from the csv file by analyzing a few number of records from our data .It asks the user the number of sample it wants to be analyzed in order to get the attributes.



Enter the number of lines to sample:
100

Fig21: No .of samples to be analyzed to get fields

It then shows you a window where all the possible attributes which the spoon has found are listed with details.

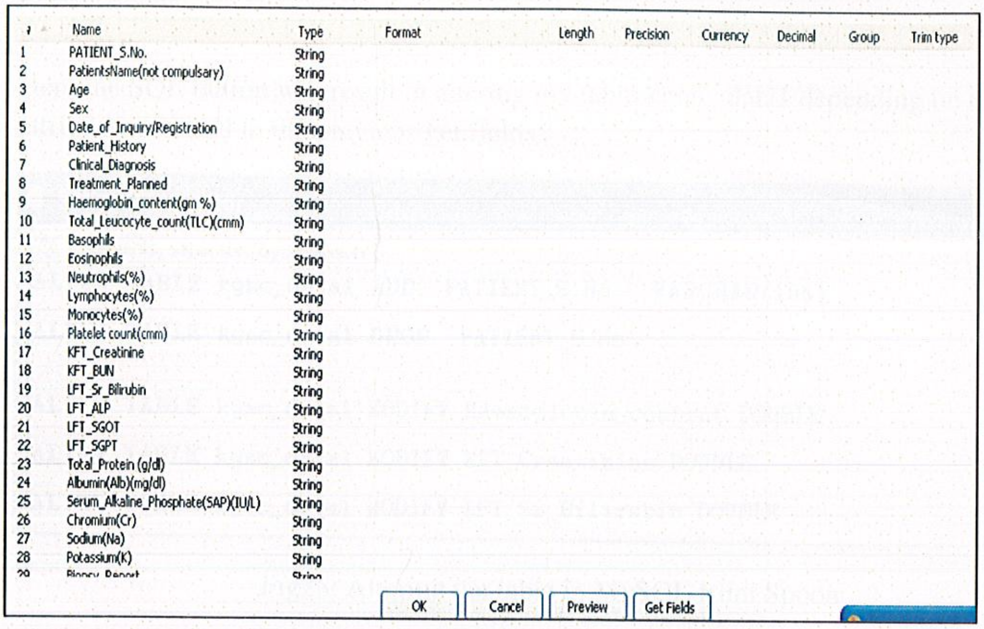


Here are the results of the document scan:

```
   Field name                 : PATIENT S.No.
   Field type                 : Integer

Field nr. 2 :
   Field name                 : Patients Name(not compulsary)
   Field type                 : String
   Maximum length             : 16
   Minimum value              : Divya Singh
   Maximum value              : Vedant
   Nr of null values          : 79

Field nr. 3 :
   Field name                 : Age
   Field type                 : Integer

Field nr. 4 :
   Field name                 : Sex
   Field type                 : String
   Maximum length             : 6
   Minimum value              : Female
   Maximum value              : Male
   Nr of null values          : 79

Field nr. 5 :
   Field name                 : Date of Inquiry/ Registration
   Field type                 : String
   Maximum length             : -1
   Minimum value              :
   Maximum value              :
   Nr of null values          : 100
   ALL NULL VALUES!

Field nr. 6 :
   Field name                 : Patient History   (Main complaints)
   Field type                 : String
   Maximum length             : -1
   Minimum value              :
   Maximum value              :
   Nr of null values          : 100
   ALL NULL VALUES!

Field nr. 7 :
   Field name                 : Clinical Diagnosis
   Field type                 : String
   Maximum length             : -1
   Minimum value              :
```

Fig22: Fields after analyzing the sample

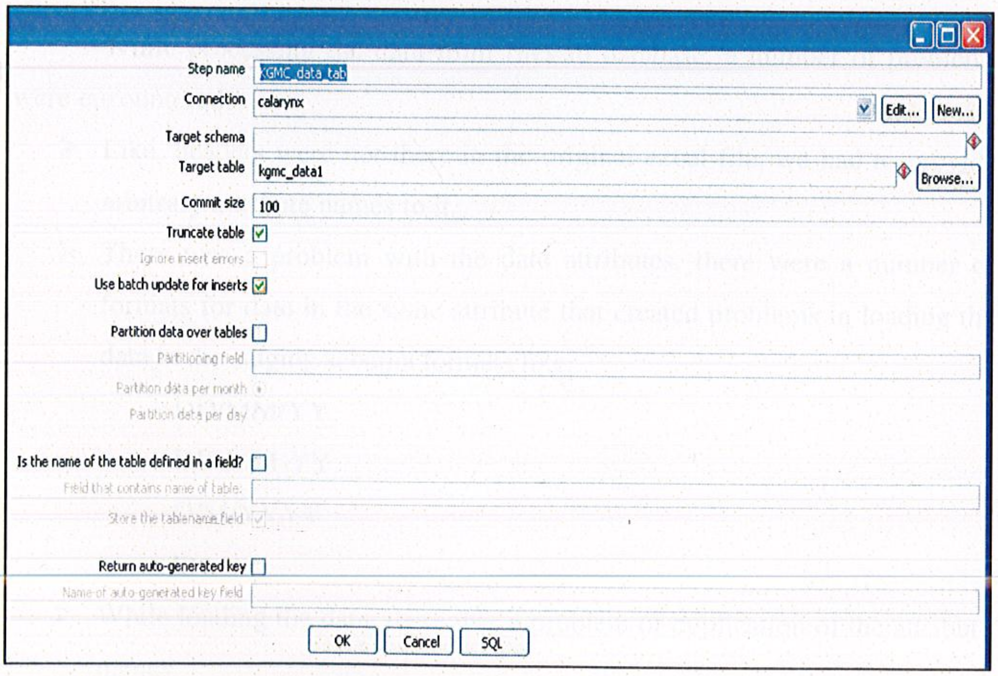
After the fields were selected we assigned them a specific data type depending on the kind of data contained in that attribute.



#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	PATIENT_S.No.	String							
2	PatientsName(not compulsary)	String							
3	Age	String							
4	Sex	String							
5	Date_of_Inquiry/Registration	String							
6	Patient_History	String							
7	Clinical_Diagnosis	String							
8	Treatment_Planned	String							
9	Haemoglobin_content(gm %)	String							
10	Total_Leucocyte_count(TLC)(cmm)	String							
11	Basophils	String							
12	Eosinophils	String							
13	Neutrophils(%)	String							
14	Lymphocytes(%)	String							
15	Monocytes(%)	String							
16	Platelet count(cmm)	String							
17	KFT_Creatinine	String							
18	KFT_BUN	String							
19	LFT_Sr_Bilirubin	String							
20	LFT_ALP	String							
21	LFT_SGOT	String							
22	LFT_SGPT	String							
23	Total_Protein (g/dl)	String							
24	Albumin(Alb)(mg/dl)	String							
25	Serum_Alkaline_Phosphate(SAP)(IU/L)	String							
26	Chromium(Cr)	String							
27	Sodium(Na)	String							
28	Potassium(K)	String							
29	Biopsy Report	String							

Fig23: Data type of various fields

Spoon also provides us with the option to preview our data. After this we need to set fields in the target file such as name of the data base and the table where the data will be loaded



Step name: KGMC_data_tab

Connection: calarynx

Target schema:

Target table: kgmc_data1

Commit size: 100

Truncate table ☑

Ignore insert errors □

Use batch update for inserts ☑

Partition data over tables □

Partitioning field:

Partition data per month ○

Partition data per day ○

Is the name of the table defined in a field? □

Field that contains name of table:

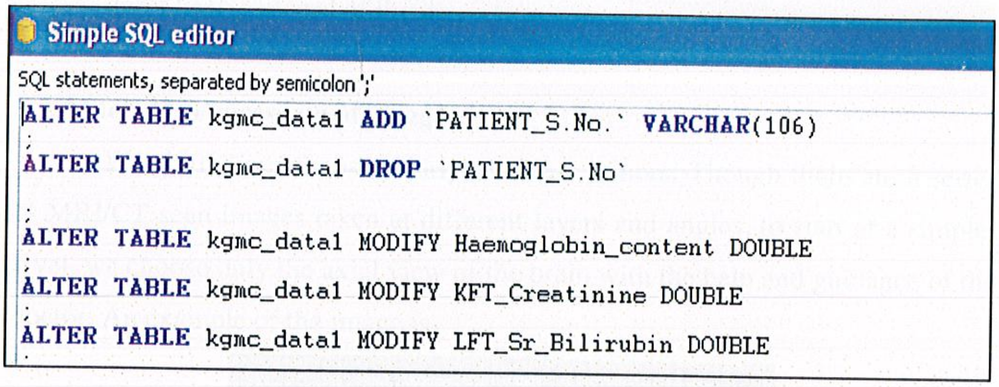
Store the tablename field ☑

Return auto-generated key □

Name of auto-generated key field:

Fig24: The target description

Here the SQL button will result in altering my table kgmc_data1 depending on the attributes selected in the previous get fields



```
ALTER TABLE kgmc_data1 ADD `PATIENT_S.No.` VARCHAR(106)
;
ALTER TABLE kgmc_data1 DROP `PATIENT_S.No`

;
ALTER TABLE kgmc_data1 MODIFY Haemoglobin_content DOUBLE
;
ALTER TABLE kgmc_data1 MODIFY KFT_Creatinine DOUBLE
;
ALTER TABLE kgmc_data1 MODIFY LFT_Sr_Bilirubin DOUBLE
;
```

Fig25: Altering the table in MySQL from Spoon

After the transformation has been completed we need to save it and then execute it. In case if there is any error it will give a message, otherwise the data will be processed and records will be stored in the appropriate table in the database based on rules defined.

While processing the data from files to database, a number of problems were encountered:

- ➢ Like, headers were not there in the original excel file, we had to provide arbitrary attribute names to it.
- ➢ There was a problem with the date attributes, there were a number of formats for date in the same attribute that created problems in loading the data in the staging schema formats like:

    DD/MM/YY

    DD-MM-YY

    D/M/YYYY

    Etc.

- ➢ While loading the data, there was a problem of duplication of the attribute names.

- Another problem was how to store the images-in which formats (*.jpg ,dicom etc.)? and how?(directly or in some other way-because the MRI/CT scan images take a lot of memory that is the reason why hospitals don't keep a record of them).

3.5.4 Step 4#: Processing of Images

We shifted our focus majorly on Brain tumors. Though there are a series of MRI/CT scan images taken at different layers and angles, to start at a simpler level, we choose only the axial view of the brain with the help and guidance of the doctor. An example of the image is:

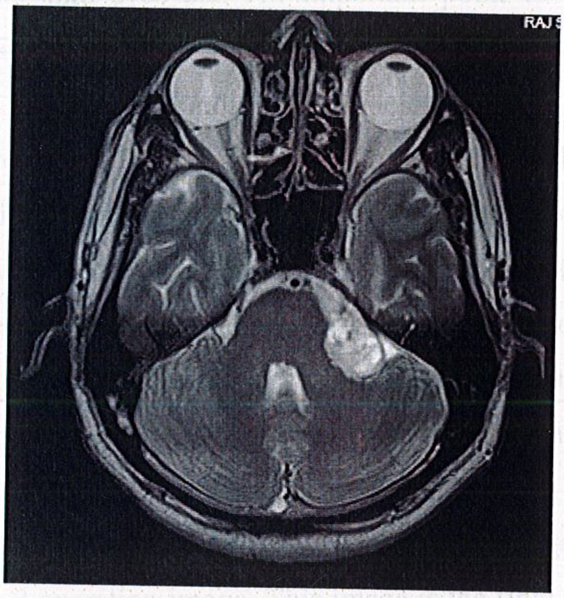


Fig26:MRI of Brain(Axial view)

Medical images are usually in the DICOM format. DICOM stands for The Digital Imaging and Communications in Medicine. The sources shared the data precisely the images in the jpg format. Our objective was to detect the tumor in the MRI images and to calculate certain parameters related to the tumor, like: Average Intensity of the tumor, standard Deviation, number of tumors, Region of Interest (ROI), Maximum Intensity and Minimum Intensity.

We used MATLAB for image processing in order to detect the tumor in this view. This is a secondary task of our project. The logic of our program lies in detecting the tumor based on its intensity, as the intensity of the tumor is fairly high compared to the intensity of the rest of the regions of brain. The programs in matlab are written in M files with *.m extension. For detail of program refer to APPENDIX A.

The program tumor.m reads the MRI/CT scan images using the function 'imread' (if we use direct Dicom images then 'dicomread' function should be used instead). In the next step we converted our *.jpg image into grayscale image, and then we cropped our image by analyzing the portions where there is no chance of brain tumor like eyes etc. An important point to note is that this program is specific for the axial view images of brain. Next a track of different intensity values and their corresponding frequencies is kept, using which we calculated the total frequency. This is then used to set a threshold intensity value which will help identify the tumor in the image. We took those intensity values as the tumor intensity values that are above 95% of the total intensity. Then we converted our grayscale images into binary image and found the total number of objects (connected components). After this we applied the Dilation and Erosion functions by first using the structure element as disk with 5 as the radius and then a disk with 3 as the radius. Then we found the list of the objects that are common in disk 5 operation and disk 3 operations and displayed the image with along with all the objects of disk 3 operation image. This image is the final image displaying the tumor. After this we calculated 6 parameters namely:

- The average intensity of the tumor
- The standard deviation of the tumor
- The region of Interest(ROI)
- Number of Tumors
- Maximum Intensity
- Minimum Intensity

The above parameters are then displayed.

3.5.5 Step 5#: Data processing to Functional Schema

After the data was dumped in the staging database, we had to clean process and store it in functional schema which can be used further used for analysis and to find correlation. Appropriate mappings designs were developed to process and store the data based on the dimensional model designed (**refer dimensional model in figure 18**). For example values of a particular test coming from different sources should be in same format. The tables like patient dimension have only selected attributes so we accordingly use appropriate function to load the required attributes in the table.

Details of various tables in the functional schema along with type of information stored in them are as follow:

TABLE 2: Tables in the working schema "working"

TABLE NAME	FUNCTION OF THE TABLE
date_dimension	This tables stores the date information which have been processed from the 'Date' staging table
test dimension	This table contains name of all the common tests that a patient gets done e.g. platelet count, TCL, KFT, LFT etc. and their description. Along with this all the tests have been given a unique id.
patient_dimension	Patient's personal details such as name, patient_id , age, sex etc.. are stored. For each unique patient an unique id is being autogenerated (given as patient id) whenever information is entered for the first time.
patient_fact_table	This table contains the patient id and their corresponding pathological data only. Here we also have the date_id and

	uid which act as foreign key here. The patient_id, date_id,uid act as primary key here for this table. It helps to store all the historical information corresponding to any number of tests being conducted.
range dimension	This table stores minimum and maximum range of all the tests
image dimension	All the parameters calculated by processing an MRI image are stored in this table

Based on the requirement various modifications were made that we require to make our data ready for analysis or to load in a table with special requirements e.g.

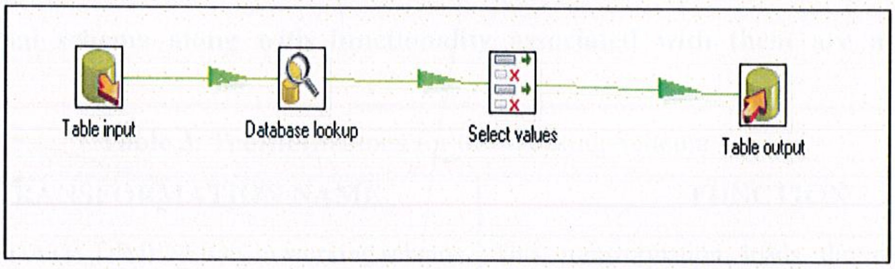
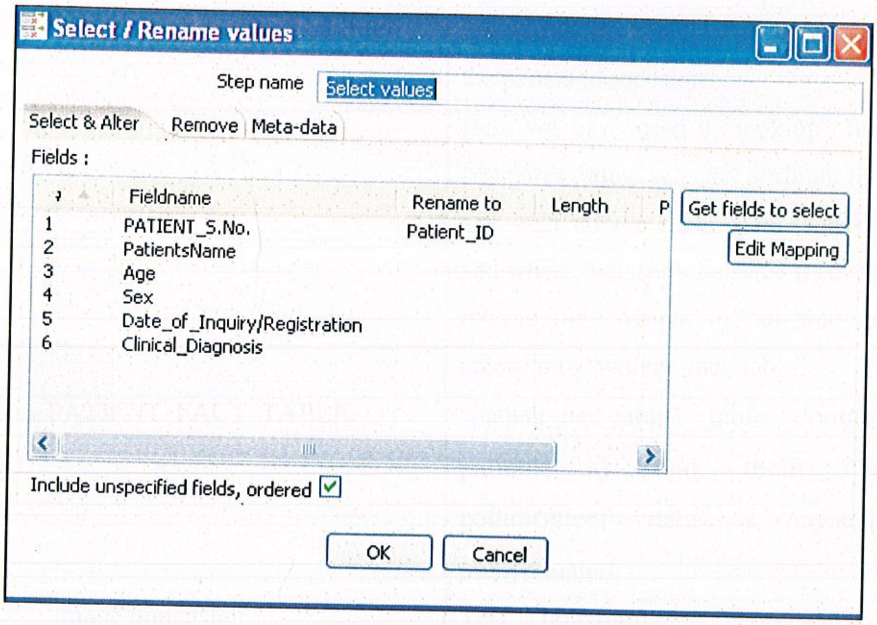


Fig27: Example of a Function

In this image the data is being loaded finally into the working database. The target table in "Table Output" contains only lesser attributes from the source table so we use the "select value" function which gives us the option to select the attribute that we want to be loaded in the target table, and even to rename those attributes.



Select / Rename values

Step name: Select values

Select & Alter | Remove | Meta-data

Fields :

#	Fieldname	Rename to	Length	P
1	PATIENT_S.No.	Patient_ID		
2	PatientsName			
3	Age			
4	Sex			
5	Date_of_Inquiry/Registration			
6	Clinical_Diagnosis			

Include unspecified fields, ordered ☑

Fig28: Select/Rename function

List of various transformations designed for processing data into the functional schema along with functionality associated with them are as follow:

Table 3: Transformations for the Working Schema

TRANSFORMATION NAME	FUNCTION
creating DATE_DIMENSION in working schema	This transformation loads the data from the DATA_DIMENSION in "cancer" database into working database
creating patient dimension with selected attribute	A table is created in this transformation by the name "patient_dimension" with few selected attributes like name, sex, age etc.
_TEST_DIMENSION	In this " test dimension" table is created which contains name of all the common tests that a patient gets done
loading kgmc_new into patient dimension	Data from the "kgmc_new" is loaded into the patient_dimension
loading lady into working	Data from the "lady" is loaded into the patient_dimension

Loading lci_control into working	Data from the “LCI_CONTROL” is loaded into the patient_dimension
lookup transformation patient id	Here we have used a “look-up” function which compares some selected attribute like name, sex, age with all the records in “patient_dimension” and where they find the same records the function returns the “patient id” of that corresponding record into “patient_fact_table”
making PATIENT FACT TABLE	“patient_fact_table” table containing the patient id and their corresponding pathological data is created in this transformation
range dimension	This transformation creates a “ Range dimension” table in working

CHAPTER 4

APPLICATION OF DATA MINING TECHNIQUES

4.1 Introduction to Data Mining

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line.

It is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers), classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes). They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Due to its complex capabilities, two precursors are important for a successful data mining exercise; a clear formulation of the problem to be solved, and access to the relevant data.

Reflecting this conceptualization of data mining, some observers consider data mining to be just one step in a larger process known as knowledge discovery in databases (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, (data mining), pattern evaluation, and knowledge presentation. Data mining has become increasingly common in both the public and private sectors. Organizations use data mining as a tool to survey customer information, reduce fraud and waste, and assist in medical research. [29]

4.2 Limitations of Data Mining

While data mining products can be very powerful tools, they are not self-sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primary data or personnel related, rather than technology-related.

Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. Similarly, the validity of the patterns discovered is dependent on how they compare to "real world" circumstances. [30]

4.3 Uses of Data Mining

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. [30]

4.4 Data Mining Issues

As data mining initiatives continue to evolve, there are several issues related to implementation and oversight. [30]

Data Quality

Data quality is a multifaceted issue that represents one of the biggest challenges for data mining. Data quality refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presence of duplicate records, the lack of data standards, the timeliness of updates, and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to subtle differences that may exist in the data. To improve data quality, it is sometimes necessary to "clean" the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database accounting for missing data points, removing unneeded data fields, identifying anomalous data points and standardizing data formats (e.g., changing dates so they all include MM/DD/YYYY).

Interoperability

Related to data quality, is the issue of interoperability of different databases and data mining software. Interoperability refers to the ability of a computer system and/or data to work with other systems or data using common standards or processes. Interoperability is a critical part of the larger efforts to improve interagency collaboration and information sharing.

Mission Creep

Mission creep is one of the leading risks of data mining cited by civil libertarians, and represents how control over one's information can be a tenuous proposition. Mission creep refers to the use of data for purposes other than that for which the data was originally collected. This can occur regardless of whether the data was provided voluntarily by the individual or was collected through other means.

Privacy

Concerns about privacy focus both on actual projects proposed, as well as concerns about the potential for data mining applications to be expanded beyond their original purposes (mission creep).[30]

4.5 Use of Data Mining in Cancer Cache

As the data warehouse was developed, so data will be uploaded periodically and thus can be managed and retrieved efficiently. Further adding value to it we would be using Data Mining techniques, through which we would be making correlation analysis among the various histopathological parameters. This will help us in giving a probabilistic prediction about the occurrence of cancer once the user has entered his histopathological details. These results can be further used by physicians or other research associates to improve the quality of the treatment both with respect to time and medication.

We are using **Statistica 9** software for finding the correlation among the parameters. 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. [31]

The whole dataset will be imported in Statistica 9 software for implementing the data mining techniques. In the data mining techniques the dataset will be first partitioned into two sets: **Training** and **Testing**. Most of the data will be used for training, and a smaller portion of the data for testing analysis. There are techniques that randomly sample the data, to help ensure that the testing and training partitions are similar. By using similar data for training and testing, we can minimize the effects of data discrepancies and better understand the characteristics of the model.

The **training dataset** is used to train or build a model. For example, in a linear regression, the training dataset is used to **fit** the linear regression model, i.e. to compute the regression coefficients.

Once a model is built on training data, we need to find out the accuracy of the model on unseen data. For this, the model should be used on a dataset that was not used in the training process - a dataset where you know the actual value of the target variable. The discrepancy between the actual value and the predicted value of the target variable is the error in prediction. Some form of average error (MSE of average % error) measures the overall accuracy of the model. If we were to use the training data set to compute the accuracy of the model fit, we would get an overly optimistic estimate of the accuracy of the model. This is because the training or model fitting process ensures that the accuracy of the model for the training data is as high as possible - the model is specifically suited to the training data. To get a more realistic estimate of how the model would perform with unseen data, we need to set aside a part of the original data and not use it in the training process. This dataset is known as the **Test dataset.** After fitting the model on the training dataset, we should test its performance on the Test dataset. The test dataset is often used to fine-tune models. [32]

We have used Random Partitioning approach for dividing our data into Training and Test dataset. We have around 100 records for cancer and healthy patients, which we randomly partitioned into a ratio of **70:30** for testing and testing dataset respectively.

There are a number of data mining techniques available, but it depends on the kind and size of the dataset as well as the purpose of our study. As the size of our dataset is quite small so it's difficult to choose an appropriate data mining technique, as the performance may not be so good. Few of the other data mining techniques include: Artificial Neural Network (ANN), Decision Tree, Genetic Algorithm etc.

Classification is another form of data analysis which can be used to extract models describing important data classes. Classification predicts categorical labels (or discrete values). It is a two step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attributes. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. Since the class label of each training sample is provided, this step is also known as supervised learning (i.e., the learning of the model is 'supervised' in that it is told to which class each training sample belongs). It contrasts with unsupervised learning (or clustering), in which the class labels of the training samples are not known. Learned model is represented in the form of classification rules, decision trees, or mathematical formulae.

One of the basic techniques for data classification is **decision tree induction**. A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The topmost node in a tree is the root node. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node which holds the class prediction for that sample. Decision trees can easily be converted to classification rules. The basic algorithm for decision tree induction is a greedy algorithm which constructs decision trees in a top-down recursive divide-and-conquer manner.

The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules. One rule is created for each path from the root to a leaf node.

An example of classification of data (available in *Cancer Cache)* based on Platelet count, Hemoglobin & Protein Content.

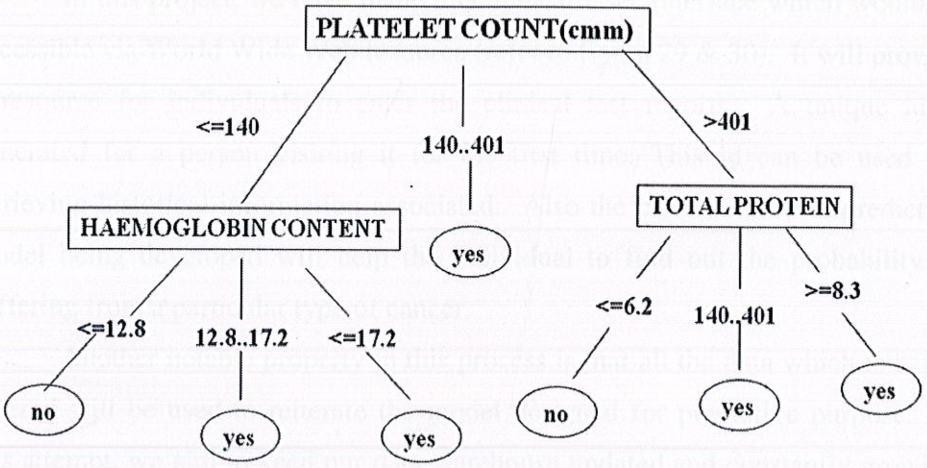


Fig31: Decision Tree example

Further its necessary to validate the model or the classification technique finalized for usage. In this process we would check for sensitivity and specificity of the predictions of the model developed. [33]

Sensitivity

Sensitivity is the proportion of true positives that are correctly identified by the test.

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

Specificity

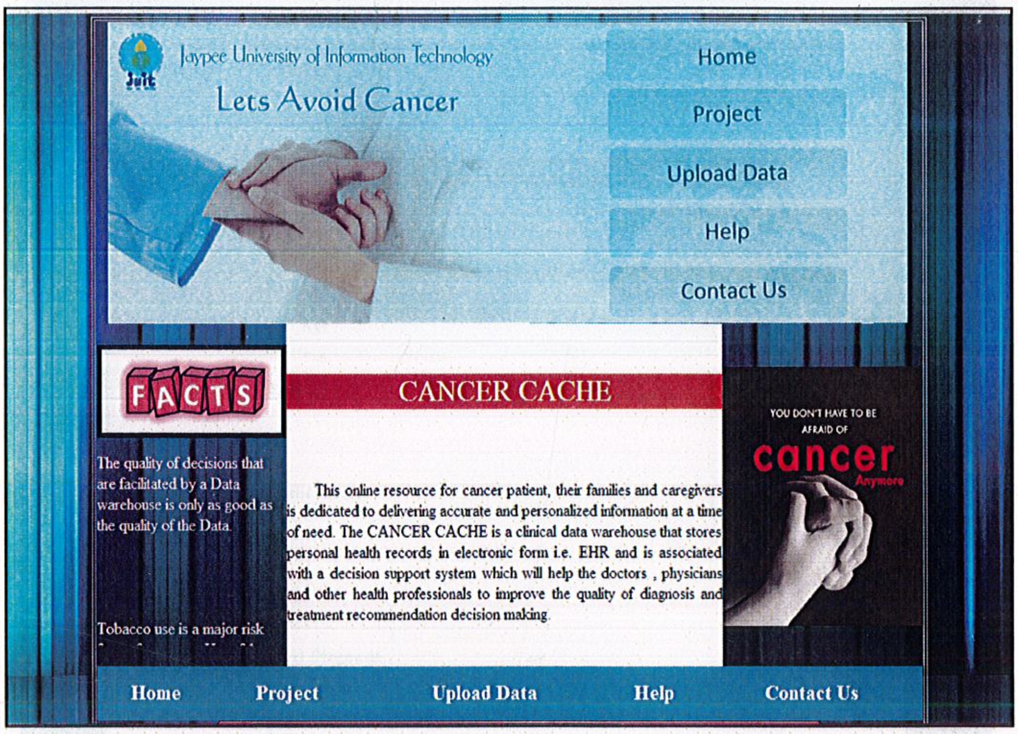
Specificity is the proportion of true negatives that are correctly identified by the test.

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

After the validation step, the model would be ready to use.

In this project, we have made a Graphical User Interface which would be accessible via World Wide Web resource (refer to figure 29 & 30). It will provide a resource for individuals to enter the clinical test reports. A unique id is generated for a person visiting it for the first time. This id can be used for retrieving historical information associated. Also the classification & predictive model being developed will help the individual to find out the probability of suffering from a particular type of cancer.

Another notable property in this process is that all the data which is being entered will be used to reiterate the model designed for predictive purpose. By this attempt, we aim to keep our data warehouse updated and constantly growing with newly added patient data.



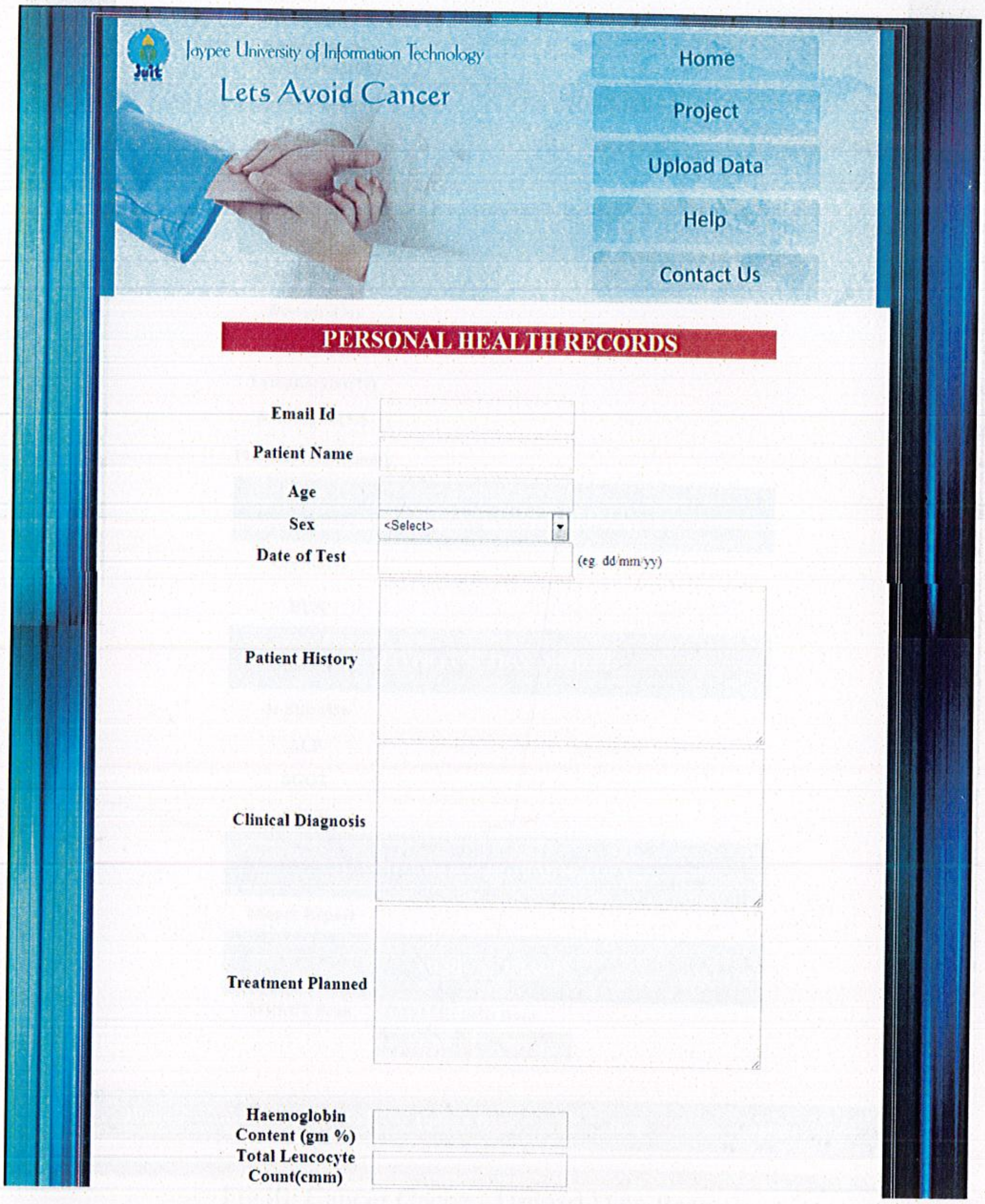
CANCER CACHE

The quality of decisions that are facilitated by a Data warehouse is only as good as the quality of the Data.

Tobacco use is a major risk

This online resource for cancer patient, their families and caregivers is dedicated to delivering accurate and personalized information at a time of need. The CANCER CACHE is a clinical data warehouse that stores personal health records in electronic form i.e. EHR and is associated with a decision support system which will help the doctors , physicians and other health professionals to improve the quality of diagnosis and treatment recommendation decision making.

Fig29: Cancer Cache-Website Snapshot

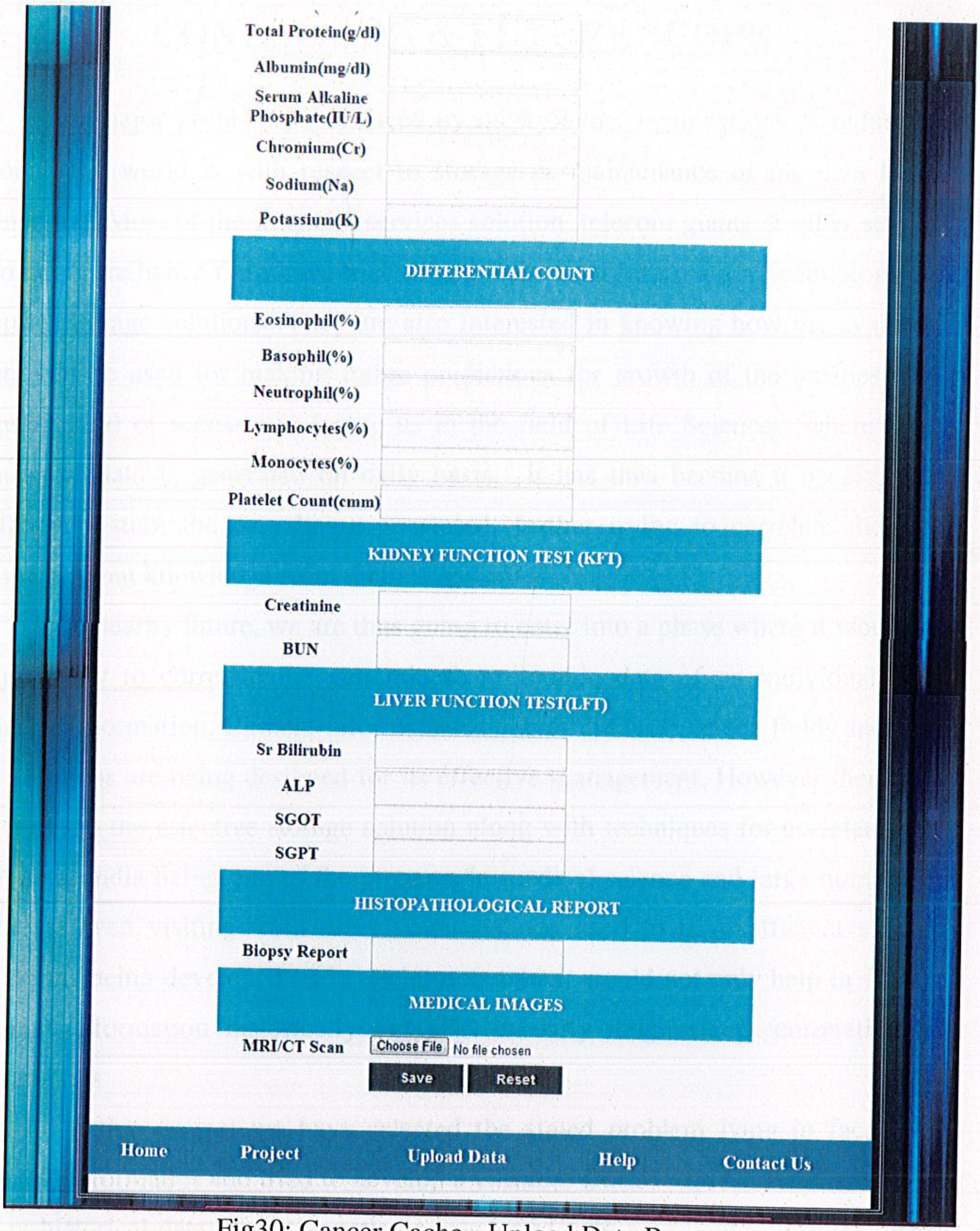


Jaypee University of Information Technology

Lets Avoid Cancer

PERSONAL HEALTH RECORDS

Email Id	
Patient Name	
Age	
Sex	<Select>
Date of Test	(eg. dd/mm/yy)
Patient History	
Clinical Diagnosis	
Treatment Planned	
Haemoglobin Content (gm %)	
Total Leucocyte Count(cmm)	



Total Protein(g/dl)

Albumin(mg/dl)

Serum Alkaline Phosphate(IU/L)

Chromium(Cr)

Sodium(Na)

Potassium(K)

DIFFERENTIAL COUNT

Eosinophil(%)

Basophil(%)

Neutrophil(%)

Lymphocytes(%)

Monocytes(%)

Platelet Count(cmm)

KIDNEY FUNCTION TEST (KFT)

Creatinine

BUN

LIVER FUNCTION TEST(LFT)

Sr Bilirubin

ALP

SGOT

SGPT

HISTOPATHOLOGICAL REPORT

Biopsy Report

MEDICAL IMAGES

MRI/CT Scan

Fig30: Cancer Cache - Upload Data Browser

CONCLUSION & FUTURE SCOPE

A major problem being faced by most of the organizations & industries around the world is with respect to storage & maintenance of the data being generated. Most of the financial services solution, telecom giants & other service providers are hence forth have tried to take help from Information Technology for getting storage solutions. They are also interested in knowing how the available data can be used for making future predictions for growth of the business. A similar kind of scenario is facing us in the field of Life Sciences, where huge chunk of data is generated on daily basis. It has thus become a necessity to efficiently store the data being generated, further trying to correlate them & extracting out knowledge from them.

In nearby future, we are thus going to enter into a phase where it would be a necessity to correlate the genomic & proteomic data of an individual with clinical information. Clinical Informatics is one of the most versed fields and new IT solutions are being designed for its effective management. However there still is a gap in the effective storage solution along with techniques for correlation of the data. India being one of the pioneers in medical science and large number of patients even visiting from other countries, we need to have efficient storage solutions being developed and certain tools which would not only help in storing patient information historically but also in early diagnosis & correlation of parameters.

In this project we have selected the stated problem lying in façade of clinical informatics and tried to develop a **Clinical Data Warehouse** which could store historical data of all the patients concerned with a particular hospital. Also we have addressed the major problem being faced of varied dimensionality of the data, ranging from images to numerical data. We have designed an appropriate dimensional model for the structure of the warehouse to have the data at a single granular level. Currently the warehouse designed is **Cancer Data Mart** having clinical data of cancer patients. We tried contacting several hospitals for obtaining data but due to certain technical constraints and policies most of them refused to share the data. Currently it includes the processed & cleaned-up data obtained from **Gandhi Memorial & Associated Hospitals (KGMC), Lucknow Cancer Institute and Raj Diagnostics, Lucknow.** The data extraction, cleaning & processing process has been carried out using ETL technology by use of an open source tool called **SPOON** (Pentaho product). **MySQL** RDBMS package was used to design the data warehouse. Currently it is storing information of 100 individuals.

As planned currently we are applying data mining technique to find some correlation among various attributes which would help us in the development of a prediction tool which could be used by doctors, physicians, other health professionals and even by a common man who has got knowledge about how to use computer & internet. It would be basically a prediction tool which could assist anyone using it, in predicting the occurrence of cancer based on the values obtained by routine biochemical test & tests. We are in phase of validating the model designed based on information of cancer suffering & non-cancer suffering people which has been designed by having the data in 70:30 ratio for the training & testing dataset.

For easy accessibility and usage a web interface has been created so that an individual can easily deposit their records, which would be stored in the warehouse as well as they can use the prediction tool for finding the probability of occurrence of a particular disease currently which is restricted to occurrences of cancer. ***Cancer Cache*** been hosted by use of a graphical interface on JUIT web server and can be accessed at: **www.juit.ac.in/attachments/CancerCache/home.html**.

In future we plan to expand the warehouse to include information about other diseases also and come up with a complete solution which can be used by hospitals for storage purpose. We dream of an era in which all the genetic information of an individual will be correlated with clinical information aspect.

APPENDIX

A - MATLAB PROGRAM-tumor.m

```
i=imread('image.jpg');
iz=imread('image.jpg');
[xo,yo]=size(iz);

%disp(xo);
%disp(yo);
ix=imread('image.jpg');
j=rgb2gray(i);
j1=imcrop(j,[50 70 220 220]);
figure,imshow(j1);
[x,y]=size(j1);
%disp(x);
%disp(y);

for i=1:255
   freq(i)=0;
end

for i=1:x
   for j=1:y
      if (j1(i,j)~=0)
      freq(j1(i,j))=freq(j1(i,j))+1;
      end
   end
end
total=0;
for i=1:255
   total=total+freq(i);
   % disp(freq(i));
end
thresh=0;
for i=255:-1:1
   sum=0;
   %disp(i);
   for k=1:i
      sum=sum+freq(k);
   end
   %disp(sum);

   if (sum < (.95*total))
      thresh=i+1;
      break;
```

```
    end
 end
 %disp(thresh);

 for i=1:x
    for j=1:y
        if j1(i,j)<=thresh
            j1(i,j)=1;
        end
    end
end

%figure,imshow(j1);
se=strel('disk',7);
j2=imerode(j1,se);
%figure,imshow(j2);
j3=imdilate(j2,se);
%figure,imshow(j3);

%converting into binary
th=graythresh(j3);
p=im2bw(j3,th);
p=bwareaopen(p,50);
%figure,imshow(p);

%finding the objects
ob=bwconncomp(p,8);

%disp(ob.NumObjects);
%disp('####');
for i=1:ob.NumObjects
object=false(size(p));
object(ob.PixelIdxList{i})=true;
%figure,imshow(object);
%finding the cordinates

for j=1:x
    for k=1:y
         f(i,j,k)=0;
        if (object(j,k)==1)
           f(i,j,k)=1;
         end
    end
end
end
```

```
se1=strel('disk',3);
j21=imerode(j1,se1);
%figure,imshow(j21);
j31=imdilate(j21,se1);
%figure,imshow(j31);

%converting into binary
th1=graythresh(j31);
p1=im2bw(j31,th1);
p1=bwareaopen(p1,50);
%figure,imshow(p1);


%finding the objects
ob1=bwconncomp(p1,8);

for i=1:ob1.NumObjects
object1=false(size(p1));
object1(ob1.PixelIdxList{i})=true;
%figure,imshow(object1);
%finding the cordinates

for j=1:x
  for k=1:y
    if (object1(j,k)==1)
      f1(i,j,k)=1;
    end
  end
end
end


%finding the list of objects where the disk5 thing matches with disk 3
%thing n enlist those objects of disk 3 in list array

ii=1;
list(1)=0;
ii2=1;
for i=1:ob.NumObjects
  for j=1:x
    for k=1:y
      if (f(i,j,k)==1)
        for m=1:ob1.NumObjects
```

```
               if (f1(m,j,k)==1)
                  ref=0;

                  for ii1=1:ii
                     if(list(ii1)==m)
                        ref=1;
                     end
                  end
                  if(ref==0)

                  list(ii2)=m;
                  ii2=ii2+1;
                  ii=ii2-1;

                  end

               end
            end
         end
      end
   end
end

%display all the objects/tumor

for i=1:ii
  % disp(list(i));
end

object2=false(size(p1));
for i=1:ii

object2(ob1.PixelIdxList{list(i)})=true;

end
%below is the final image displaying the tumor
figure,imshow(object2);


%displaying the perimeter of the tumor in the original image

tumor=bwperim(object2);
%c=uisetcolor([1 0 0]);
for i=1:x
   for j=1:y
      if(tumor(i,j)==1)
```

```
%         ix(i+50,j+70)='green';
        end
    end
end
%figure,imshow(ix);

%calculating the parameters
area=0;
c=0;

for i=1:xo
   for j=1:yo
      if(i>50 && j>70 && i<270 && j<290)
       if(object2(i-50,j-70)==1)

          intensity(i,j)=iz(i,j);
          area=area+1;
          c=c+1;
       end
      else
          intensity(i,j)=0;
      end

    end
end
%disp('$$$$$$$');
%disp(c);

%calculating the average intensity
avg=0;
s=0;
for i=1:xo
   for j=1:yo
        s=s+intensity(i,j);
    end
end
%disp(s);
%disp('gfccsagg');
avg=s/c;

disp('the average is:: ')
disp(avg);
disp('the region of interest is:: ');
disp(area);
a=0;
for i=1:xo
```

```
    for j=1:yo
        if(i>49&& j>69 && i<270 && j<290)
        if (object2(i-49,j-69)==1)
            a=a+((avg-intensity(i,j))^2);
        end
        end
    end
end
var=a/c;
sd=sqrt(var);

disp('the standard deviation is ::  ');
disp(sd);

disp('total number of tumors:: ');
disp(ii);

%maximum intensity
mx=0;
mx=intensity(1,1);
for i=1:xo
    for j=2:yo
        if(mx<intensity(i,j))
            mx=intensity(i,j);
        end
    end
end


disp('maximum intensity:: ');
disp(mx);

%minimum intensity
mn=0;
mn=intensity(1,1);
for i=1:xo
    for j=2:yo
        if(i>50 && j>70 && i<270 && j<290)
        if(object2(i-50,j-70)==1)
        if(mn>intensity(i,j))
            mn=intensity(i,j);
        end
        end
        end

    end
```

```
end

disp('minimum intensity:: ');
disp(mn);
%u=input('prompt');
```

B- SQL QUERIES

CREATION OF STAGING SCHEMA (Cancer)

```
Create Schema Cancer;
```

CREATION OF TABLES IN CANCER SCHEMA

1) kgmc_data1

```
CREATE TABLE kgmc_data1
(
 `PATIENT_S.No.` VARCHAR(106)
, PatientsName VARCHAR(18)
, Age INT

, Sex VARCHAR(6)
, `Date_of_Inquiry/Registration` TINYTEXT
, Patient_History VARCHAR(200)
, Clinical_Diagnosis VARCHAR(22)
, Treatment_Planned TINYTEXT
, Haemoglobin_content DOUBLE
, Total_Leucocyte_count INT
, Basophils INT
, Eosinophils INT
, Neutrophils INT
, Lymphocytes INT
, Monocytes INT
, `Platelet count` INT
, KFT_Creatinine DOUBLE
, KFT_BUN INT
, LFT_Sr_Bilirubin DOUBLE
, LFT_ALP INT
, LFT_SGOT INT
, LFT_SGPT INT
, Total_Protein TINYTEXT
, `Albumin(Alb)` TINYTEXT
, `Serum_Alkaline_Phosphate(SAP)` TINYTEXT
, `Chromium(Cr)` TINYTEXT
, `Sodium(Na)` TINYTEXT
, `Potassium(K)` TINYTEXT
, Biopsy_Report VARCHAR(38)
, `X-Ray` VARCHAR(64)
, CT_Scan TINYTEXT
, MRI TEXT
);
```

2)kgmc_new

```
CREATE TABLE kgmc_new
(
 `PATIENT S.No.` VARCHAR(46)
, PatientsName VARCHAR(11)
, Age VARCHAR(2)
, Sex VARCHAR(4)
, `Date of Inquiry/ Registration` TINYTEXT
, Patient_History VARCHAR(155)
, Clinical_Diagnosis TINYTEXT
, Treatment_Planned TINYTEXT
, Haemoglobin_content VARCHAR(2)
, Total_Leucocyte_count VARCHAR(4)
, Basophils TINYTEXT
, Eosinophils TINYTEXT
, Neutrophils TINYTEXT
, Lymphocytes TINYTEXT
, Monocytes TINYTEXT
, Platelet_count VARCHAR(39)
, KFT_Creatinine VARCHAR(6)
, KFT_BUN TINYTEXT
, LFT_Sr_Bilirubin VARCHAR(1)
, LFT_ALP TINYTEXT
, LFT_SGOT TINYTEXT
, LFT_SGPT TINYTEXT
, Total_Protein TINYTEXT
, `Albumin(Alb)` TINYTEXT
, `Serum_Alkaline_Phosphate(SAP)` TINYTEXT
, `Chromium(Cr)` TINYTEXT
, `Sodium(Na)` TINYTEXT
, `Potassium(K)` TINYTEXT
, Biopsy_Report TINYTEXT
, `X-Ray` TINYTEXT
, CT_Scan TINYTEXT
, MRI VARCHAR(1)
);
```

3)lady and lady1

```
CREATE TABLE  lady
(
 `PATIENT S.No.` VARCHAR(46)
, PatientsName VARCHAR(11)
, Age VARCHAR(2)
, Sex VARCHAR(4)
, `Date of Inquiry/ Registration` TINYTEXT
, Patient_History VARCHAR(155)
, Clinical_Diagnosis TINYTEXT
, Treatment_Planned TINYTEXT
, Haemoglobin_content VARCHAR(2)
, Total_Leucocyte_count VARCHAR(4)
, Basophils TINYTEXT
, Eosinophils TINYTEXT
, Neutrophils TINYTEXT
, Lymphocytes TINYTEXT
, Monocytes TINYTEXT
, Platelet_count VARCHAR(39)
, KFT_Creatinine VARCHAR(6)
, KFT_BUN TINYTEXT
, LFT_Sr_Bilirubin VARCHAR(1)
, LFT_ALP TINYTEXT
, LFT_SGOT TINYTEXT
, LFT_SGPT TINYTEXT
, Total_Protein TINYTEXT
, `Albumin(Alb)` TINYTEXT
, `Serum_Alkaline_Phosphate(SAP)` TINYTEXT
, `Chromium(Cr)` TINYTEXT
, `Sodium(Na)` TINYTEXT
, `Potassium(K)` TINYTEXT
, Biopsy_Report TINYTEXT
, `X-Ray` TINYTEXT
, CT_Scan TINYTEXT
, MRI VARCHAR(1)
);
```

CREATION OF FUNCTIONAL SCHEMA (WORKING)

```
Create Schema Working;
```

CREATION OF TABLES IN WORKING SCHMEA

1) Date Dimension

```
CREATE TABLE `date dimension`
(
 DATE_ID INT
, DATE_VALUE INT
, DAY_OF_WEEK VARCHAR(3)
, DAY_OF_MONTH VARCHAR(8)
, WEEK_OF_YEAR VARCHAR(14)
, WEEK_SORT_VALUE INT
, MONTH_OF_YEAR VARCHAR(10)
, MONTH_SORT_VALUE INT
, QTR_OF_YEAR VARCHAR(9)
, QTR_SORT_VALUE INT
, CALENDER_YEAR INT
);
```

2)Patient Dimension

```
CREATE TABLE `patient dimension`
(
 Patient_ID INT
, PatientsName VARCHAR(18)
, Age INT
, Sex VARCHAR(6)
, `Date_of_Inquiry/Registration` VARCHAR(255)
, Clinical_Diagnosis VARCHAR(22)
, Patient_History VARCHAR(200)
);
```

3)Range Dimension

```
CREATE TABLE `range in sample`
(
 UID VARCHAR(2)
, `MEASURABLE QUANTITY` VARCHAR(37)
, MINIMUM DOUBLE
, MAXIMUM DOUBLE
);
```

4) Test Dimension

```
CREATE TABLE ` test dimension`
(
 HISTOPATH_ID VARCHAR(3)
, DATE_ID INT
, NAME_OF_THE_TEST VARCHAR(29)
, DESCRIPTION TEXT
);
```

5)Patient Fact Table

```
CREATE TABLE `patient fact table`
(
 PATIENT_ID TINYTEXT
, DATE_ID TINYTEXT
, UID TINYTEXT
, HISTOPATH_ID TINYTEXT
, `Haemoglobin content` DOUBLE
, `Total Leucocyte count(TLC)` DOUBLE
, DC_Basophils DOUBLE
, DC_Eosinophils DOUBLE
, DC_Neutrophils DOUBLE
, DC_Lymphocytes DOUBLE
, DC_Monocytes DOUBLE
, `Platelet count` DOUBLE
, KFT_Creatinine DOUBLE
, BUN DOUBLE
, `Sr Bilirubin` DOUBLE
, LFT_ALP DOUBLE
, LFT_SGOT DOUBLE
, LFT_SGPT DOUBLE
, `Total Protein` DOUBLE
, `Albumin(Alb)` DOUBLE
, `Serum Alkaline Phosphate(SAP)` DOUBLE
, `Chromium(Cr)` DOUBLE
, `Sodium(Na)` DOUBLE
, `Potassium(K)` DOUBLE
);
```

REFERENCES

1. Wikipedia. “Healthcare Informatics.” Internet: http://en.wikipedia.org/wiki/Health_informatics[Aug 15, 2010].
2. Rachel Burkot, Bronwyn Harris. “What-is-clinical-informatics.” Internet: www.wisegeek.com/what-is-clinical-informatics.htm, [Sept. 12, 2010].
3. “10 facts about cancer” Internet: www.who.int/features/factfiles/cancer/en/index.html, [Aug. 21, 2010].
4. Publications of the WHO/ICO Information Centre on HPV and Cervical Cancer. “Human Papillomavirus and Related Cancers” Internet: apps.who.int/hpvcentre/statistics/dynamic/ico/country_pdf/IND.pdf, Sept. 15, 2011[Nov. 29, 2010].
5. “Cancer Research in ICMR Achievements in Nineties” Internet: www.icmr.nic.in/cancer.pdf, [Nov. 12, 2010].
6. “Cancer statistics” Internet: http://westcancertrust.org/.../cancer-status-india-1-in-80-families-affected-by-cancer [Sept. 20, 2010].
7. Kosmix™ Corporation. “Statistics of Cancer” Internet: www.wisegeek.com/what-is-clinical-informatics.htm, [Dec.3, 2010].
8. Globocan 2008, IARC, 2010 “Cancer” Internet: www.who.int/mediacentre/factsheets/fs297/en/index.html, [Sept. 7, 2010].
9. “Cancer Control” Internet:http://whqlibdoc.who.int/publications/2007/9241547111_eng.pdf, [Nov 20, 2010].
10. “What are the different kinds of Cancer?” Internet:www.tirgan.com/catyps.htm, [Sept. 12, 2010].
11. Brazhnik, O., and J. Jones. “Anatomy of Data Integration.” *Journal of Biomedical Informatics* 40, no. 3 (2007): 252–269.
12. Geisler, S., A. Brauers, C. Quix, and A. Schneink. “Ontology-based System for Clinical Trial Data Management.” *Proceedings: Annual Symposium of the IEEE/EMBS Benelux Chapter*. Heeze, the Netherlands: IEEE 2007. pp. 53-55. 2007.

13. Wang, K., P. Tarczy-Hornoch, R. Shaker, P. Mork, and J. F. Brinkley. "BioMediator Data Integration: Beyond Genomics to Neuroscience Data." *AMIA Annual Symposium Proceedings* (2005): 779–783.

14. Nagarajan, R., M. Ahmed, and A. Phatak. "Database Challenges in the Integration of Biomedical Data Sets." *Proceedings of the Thirtieth International Conference on Very Large Data Bases.* Toronto: VLDB Endowment, 2004.

15. Dinu, V., and P. Nadkarni. "Guidelines for the Effective Use of Entity–Attribute–Value Modeling for Biomedical Databases." *International Journal of Medical Informatics* 76, no. 11 (2007): 769–779.

16. Brazhnik, O., and J. Jones. "Anatomy of Data Integration."

17. Geisler, S., A. Brauers, C. Quix, and A. Schneink. "Ontology-based System for Clinical Trial Data Management."

18. Wang, K., P. Tarczy-Hornoch, R. Shaker, P. Mork, and J. F. Brinkley. "BioMediator Data Integration: Beyond Genomics to Neuroscience Data."

19. Deshpande, A. M., C. Brandt, and P. M. Nadkarni. "Metadata-driven Ad Hoc Query of Patient Data." *Journal of the American Medical Informatics Association* 9 (2002): 369–382.

20. Anhøj, J. "Generic Design of Web-Based Clinical Databases." *Journal of Medical Internet Research* 5, no. 4 (2003): e27. 12. Hughes, R. "Optimal Data Architecture for Clinical Data Warehouses." *Information Management,*

21. Pentaho Corporation. "Pentaho Data Integration (Kettle)" Internet: www.kettle.pentaho.com/,[Jan 15,2011].

22. scribd Inc. "INTRODUCTION TO MATLAB FORENGINEERING STUDENTS" Internet: www.scribd.com/doc/55075554/1/Introduction,[Jan 26,2011].

23. UniSIM. "HTML" Internet: sst.unisim.edu.sg:8080/dspace/bitstream/123456789/171/1/09_Fan%20Ying%20Xin.pdf,[Jan 15,2011].

24. Scott Arnett,"Data Warehousing." Internet: http://www.pbinsight.com/files/resource-library/resource-files/pbbi-data-warehousing-keys-to-success-wp-usa.pdf[Feb 15, 2011].

25. "Data Modeling" Internet:http://www.liberty.edu/media/1414/%5B6330%5DERDDataModeling.pdf, [Aug 24,2010].

26. Wikipedia. "Image Processing" Internet: en.wikipedia.org/wiki/Image_processing,[Jan 25, 2011].

27. K.M.M. Rao, V.D.P. Rao "Medical Image Processing" Internet: www.drkmm.com/resources/MEDICAL_IMAGE_PROCESSING_25sep06.pdf,[Mar 20,2011].

28. Wikipedia. "Extract, transform, load" Internet:http://en.wikipedia.org/wiki/Extract,_transform,_load,[feb 10, 2011].

29. "An Introduction to Data Mining" Internet:www.thearling.com/text/dmwhite/dmwhite.htm,[Mar 30, 2011].

30. Jeffrey W. Seifert Analyst in Information Science and Technology Policy Resources, Science, and Industry Division "Data Mining: An Overview" Internet: www.fas.org/irp/crs/RL31798.pdf,[Mar 30, 2011].

31. Wikipedia "Statistica" Internet: http://en.wikipedia.org/wiki/STATISTICA

32. Google "Standard Data Partitioning" Internet: http://www.resample.com/xlminer/help/Partition/Partition.htm

32. Jiawei Han and Micheline Kamber. "Classification and Prediction" in "Data Mining: Concepts and Techniques", 2000 (c) Morgan Kaufmann Publishers

LIST OF PUBLICATIONS (COMMUNICATED/ACCEPTED)

BRIEF BIODATA OF THE STUDENTS

PRIYANKA ARORA

She is pursuing her B.Tech in Bioinformatics from Jaypee University of Information Technology and will be completing her degree in June 2011. Her Technical and Research interests include: Data warehousing and data mining techniques, C, C++, HTML, and MATLAB etc. She is placed in Accenture as an Associate Software engineer. She is interested in pursuing higher studies and planning for M.Tech studies in data mining.

SHRADHA PANT

She is pursuing her B.Tech in Bioinformatics from Jaypee University of Information Technology and will be completing her degree in June 2011. Her Technical and Research interests include Data mining, C, C++, and HTML etc. She is placed in Accenture as an Associate Software engineer and will be joining them in June 2011.