# Development of a Data Warehouse for Lymphoma Cancer Diagnosis and Treatment Decision Support

TEH YING WAH, ONG SUAN SIM
Department of Information Science, Faculty of Computer Science and Information Technology
University Malaya
Lembah Pantai, 50603, Kuala Lumpur
MALAYSIA
Email: tehyw@um.edu.my, ongsuansim@yahoo.com

*Abstract*: - Data warehousing is becoming an indispensable component in data mining process and business intelligence. Data warehouses often act as a data collector, data integrator and data provider in the data mining process. This paper reviews the development and use of a clinical data warehouse specific to the Lymphoma or Lymph Node cancer, which could be used by doctors, physicians and other health professionals, in conjunction with a Clinical Decision Support System (DSS), to support the clinical process as well as to formulate the appropriate model to improve the quality of diagnosis and treatment recommendation decision making. This paper proposes a 5-stage sequential methodology for the clinical data warehouse development. Research on the evaluation of the developed data warehouse and how it would support the data mining process will be discussed in a separate paper.

*Keywords:* Clinical data warehouse, Clinical Decision Support System (DSS), Lymphoma or Lymph Node cancer, diagnosis and treatment recommendation decision making.

## 1 Introduction

Lymphoma is one of the diseases that uses Clinical DSS for diagnosis and treatment recommendation, and has achieved significant results. The Image Guided Decision Support System developed by Siemens Corporate Research is one of those Clinical DSS to assist pathologists in discriminating among malignant lymphomas and chronic lymphocytic leukemia directly from microscopic specimens [1].

There are 30 to 40 types of Lymphoma that have been discovered so far. The disease was first discovered in the 1960s, however, its clinical management and support has yet to be established completely due to the reasons below:

1. It is not easy to do early detection since the causes of the majority of Lymphoma are still unknown.
2. It is sometimes hard to diagnose Lymphoma because the disease has imprecise symptoms and similar syndrome as other medical problems (for example infections and "Cat Scratch Fever") [2].
3. It is difficult to determine an optimal treatment for the patient, because there are many uncertainties associated with the selection of a treatment for Lymphoma.

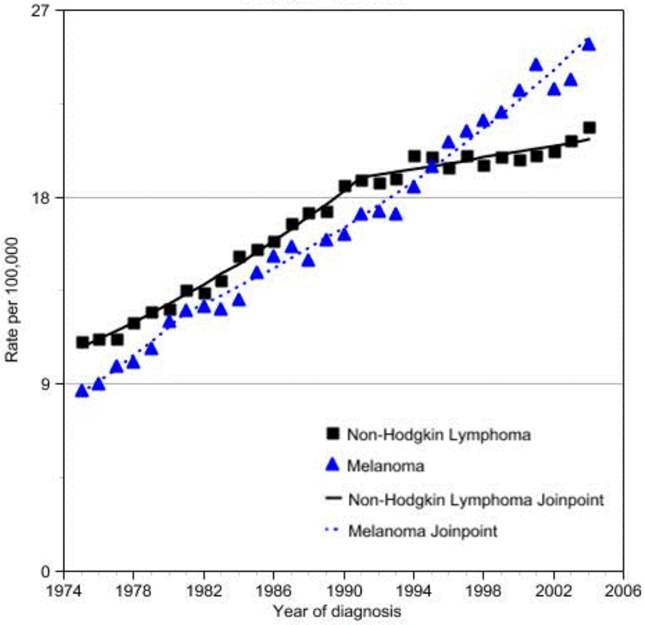


Figure 1: Cancers that are increasing: 1975 – 2006
Source: SEER program, National Cancer Institute

Since the early 1970s, incident rates for Lymphoma cancer, especially Non-Hodgkin's lymphoma, have nearly doubled. As shown in Figure 1, according to the “Cancer Trends Progress Report – 2007” [3] published by U.S. National Cancer Institute, Lymphoma is one of the cancers on the rise and require greater efforts at control. The report had also estimated 71,380 new cases of Lymphoma will occur in the year 2007, including 8,190 cases of Hodgkin's Lymphoma and 63,190 cases of Non-Hodgkin's Lymphoma.

Hence, the development of a clinical DSS and data warehouse to improve the diagnosis and treatment decision making process of Lymphoma, so as make it to be more precise and accurate, is very important and useful.

## 2 Literature Review

The concept of "data warehousing" arose in mid 1980s with the intention to support huge information analysis and management reporting. Data warehouse was defined as a "subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process" by W. H Inmon, father of the data warehouse, in year 1990 [4].

Today, data warehouses are not only deployed extensively in banking and finance, consumer goods and retail distribution and demand-based manufacturing, it has also became a hot topic in non-commercial sector, mainly in medical fields, government, military services, education and research community etc.

The growingly application of clinical information system and electronic medical records (EMR) in medical field in the past few years, has led to the evolution of clinical data warehouses.

There are quite a few clinical data warehouses currently exceed 150 terabytes in size. In a research done by Deloitte Healthcare College in year 2006, as shown in Figure 2, based on the rate of clinical data growth every year, It is estimated that the first petabyte database (i.e. 1000 terabytes) might appear by the year 2010 [5].

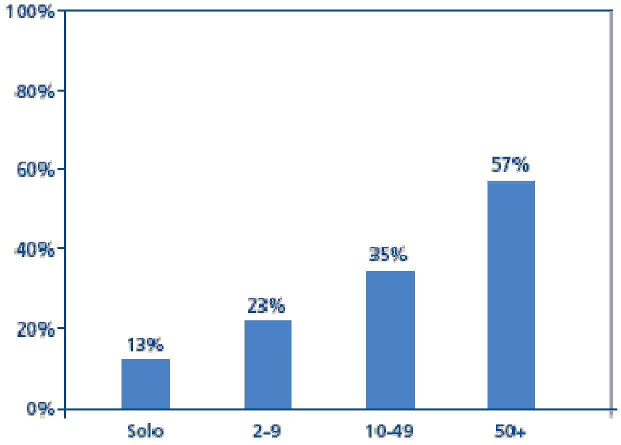


Figure 2: Rate of clinical data growth every year
Source: Deloitte Healthcare College 2006

Clinical data warehouse is normally built to validate assumptions and to discover trends on large amount of patient data [6]. It contains not only alphanumeric administrative data, but also images or signals such as X-ray pictures, echography, electrocardiogram, etc… [7].

Torben Bach Pedersen and Christian S. Jensen also identified in their "Research Issues in Clinical Data Warehousing" that clinical data warehouse needs to support for "complex-data modeling features, advanced temporal support, advanced classification structures, continuously valued data, dimensionally reduced data, and the integration of very complex data" [8]. Hence, clinical data warehouse requires advanced data modeling than conventional multidimensional data warehousing approaches.

From the above, we summarize that clinical data warehouse is different from a commercial data warehouse in the 4 aspects below (Table 1):

| Aspect | Commercial data warehouse | Clinical data warehouse |
|---|---|---|
| Usage of data stored in the data warehouse | To identity patterns from the enormous amount of data in the operational database for better management decisions. | To validate assumptions, find indicators, descriptors and risk factors in order to understand and characterize the complex medical data and trends. |

| Aspect | Commercial data warehouse | Clinical data warehouse |
|---|---|---|
| Data types support | Deals with simple data types:<br>▪ String, text<br>▪ Numeric, decimal<br>▪ Boolean | Supports both simple conventional data types a advanced data types that suit to medical data specificity:<br>▪ Advanced temporal support;<br>▪ Advanced classification structure;<br>▪ Continuously valued data;<br>▪ Image data, e.g. X-ray, electro-cardiogram. |
| Semantic of the data | The semantic of the data is clear and explicit. | The semantic of the data is implicit, and special tools or features are often required for data understanding and semantic derivation. |
| Data processing method | The Extract, Load and Transform (ETL) processes are simple and straight forward. | Complex algorithms based on signal processing, pattern recognition, statistical methodologies, are often required to extract and transform the raw data into relevant information as well as to validate them. |

Table 1: Differences between commercial data warehouses and clinical data warehouses

# 3 Data Warehouse Architecture

According to Laura Hadley in her “Developing a data warehouse architecture” article, the architecture of a data warehouse is “a description of the elements and services of the warehouse, with details showing how the components will fit together and how the system will grow over time…., it is a set of documents, plans, models, drawing and specifications, with separate sections for each key component area and enough detail to allow their implementation by skilled professionals….” [9].

An architecture is very critical in the development of data warehouse, it shows what, how and why a data warehouse is developed, and it should be driven by the business needs. This paper proposes a three-layer architecture for the clinical data warehouse development as illustrated in Figure 3.

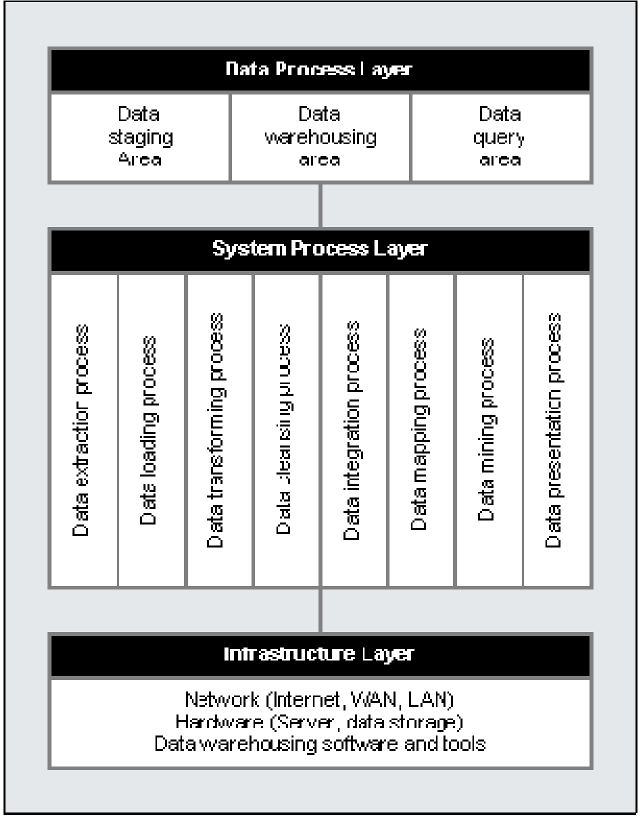


Figure 3: Clinical data warehouse architecture

## 3.1 Data process layer

This layer presents the structure and contents of the data model which represents the business requirements. It can be further divided into 3 areas:

- ***Data staging area*** – a temporary area to keep the pre-processing data before they are stored permanently in the data warehouse area.

- ***Data warehousing area*** – the back-end repository that stores the foundation data. The data structure is normally represented in a Star, Snowflake or Chen's Entity-Relationship schema.

- ***Data query area*** – the result area that carries the queried data for the data mining process. Data structure is presented in multi-dimensional model.

### 3.2 System process layer

This layer encompasses several processes to be performed in the data warehouse. They are briefly explained in the table 2 below:

| Process | Description |
|---|---|
| Extract, Transform and Load (ETL) | ▪ *Extract*: to dig out useful data from sources;<br>▪ *Transform*: to cleanse, integrate, de-normalize, convert, aggregate, summarize, and reformat data to an appropriate form for data mining;<br>▪ *Load*: to import and store data;<br>▪ *Security*: to handle data compression and encryption when required;<br>▪ *Control*: to schedule and monitor data batch processing and to log, notify and handle exceptions and errors. |
| Data mapping | ▪ To map and transform information from source to target;<br>▪ To enrich the semantics of metadata using techniques like neural network, machine learning etc. |
| Data mining | ▪ To analyze multi-dimensional data to identify unknown information, using data mining techniques, such as neural network, fizzy logic, machine learning, clustering, feature distribution etc. |
| Data Present-ation | ▪ To present flexible visualization of the mining results<br>▪ To reformat and present the interlinked, multi-dimensional data model visually, using scrolling, drill-down, roll up techniques. |

Table 2: System process layer of a data warehouse

### 3.3 Infrastructure layer

This layer specifies the data warehousing considerations from the IT perspectives. The data process and system process designs determine the architectural requirements of this layer. This layer hosts the application, software and hardware used in system and data architecture layers, such as servers, network devices, operating system, relational data management system (RDBMS), and all the front-end tools used in data warehouse processing.

This layer also specifies the data access methods (e.g. JDBC, ODBC, OLE, OLE DB, DCE, ORBs, FTP, SFTP, XML, MSMQ, web services etc) and network connectivity methods (e.g. LAN, WAN, DNS, LDAP etc). Together the data process, system process and infrastructure layers will provide an organizing framework for actual creation of the clinical data warehouse.

## 4 Development of Data Warehouse

Like all other Information system developments, the design and implementation of a data warehouse development use methodologies. There are numerous data warehouse development frameworks and best practice methodologies, the appropriate approach to a data warehouse development varies depending on the objectives it desires to achieve and organization it supports.

Some adopt MIDEA, a multidimensional data warehouse development methodology based on a multidimensional data model [10]; some follow conventional Software Development Life Cycle model; others apply tailor-made business process oriented development strategy to specific data warehouse projects [11]. This paper proposes a clear-cut sequential 5 stages approach for the development of the Lymphoma specific data warehouse, as illustrated in Figure 4.

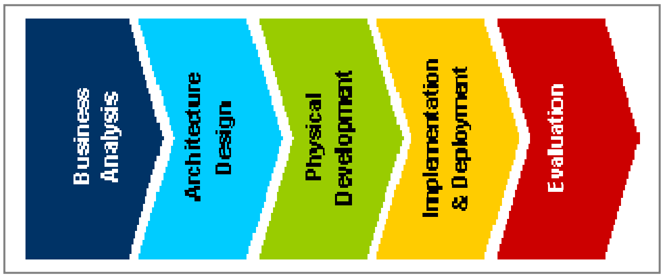


Figure 4: Data warehouse development methodology

## 4.1 Business Analysis

In the business analysis stage, the following two aspects are studied in detail to produce a high-level view of the desired data warehouse, together with its goals and acceptance criteria:

- ***Business process analysis*** –the existing process is studied and analyzed.

- ***Business requirement analysis*** – to collect and understand business requirements which state the business value of the data warehouse and drive the architecture of the data warehouse.

### 4.1.1 Business process analysis

The use case diagram as shown in Figure 5 depicts a high level overview of system functionality provided by a typical Clinical DSS that uses a clinical data warehouse for Lymphoma diagnosis [12].

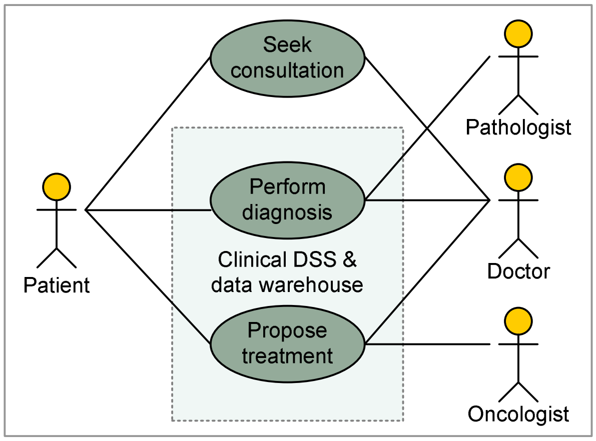


Figure 5: Clinical DSS use case diagram

There are 4 actors in the process; they are Patient, Doctor, Pathologist, and Oncologist. The interactions between the actors and activities are demonstrated in the following paragraphs.

Activity 1: Seek consultation

Patient seeks consultation from doctor when certain prolonged symptoms are noticed, e.g. lumps in neck, armpits, groin, weight loss, fever, loss of appetite, itchiness all over body, excessive sweating etc (Figure 6).

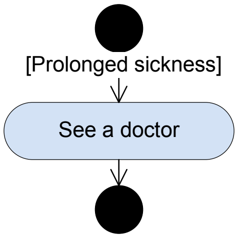


Figure 6: Activity 1 – Seek consultation

Activity 2: Perform diagnosis

Upon confirming the warning signs and symptoms of Lymphoma, the doctor together with pathologist will perform a series of tests on the patient to determine the type, stage and prognosis of the disease (Figure 7).

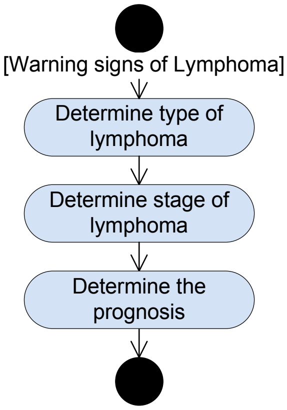

[Warning signs of Lymphoma]
Determine type of lymphoma
Determine stage of lymphoma
Determine the prognosis

Figure 7: Activity 2 – Perform diagnosis

Activity 2.1: Determine the type of Lymphoma

First, the doctor will perform an excisional biopsy, by taking tissue sample from the affected organs for further examination. If other organs e.g. skin, brain, stomach are affected, a biopsy from these organs is also required. Next, pathologist will inspect the physical appearance of the biopsy sample under a microscope, or identify the special molecules on the cancer cells using markets, that result to determining the type of Lymphoma (Figure 8).

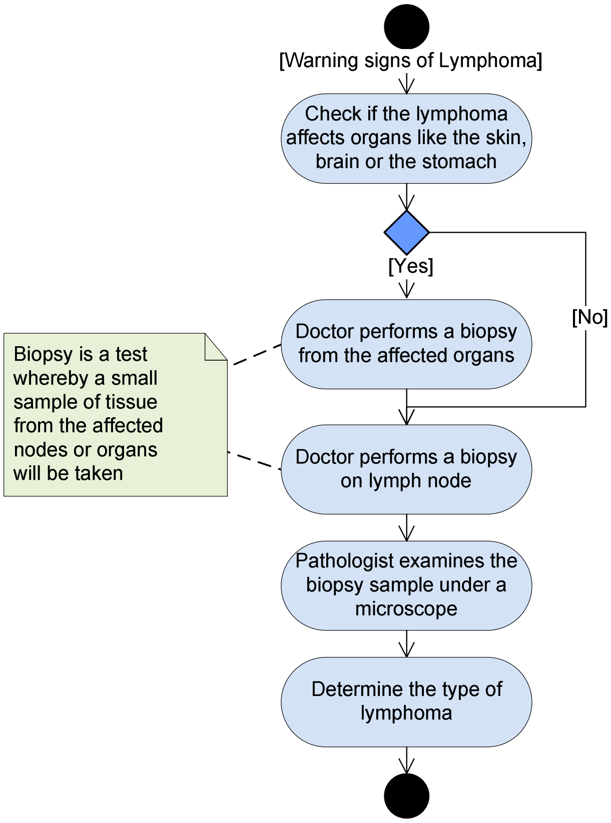

[Warning signs of Lymphoma]
Check if the lymphoma affects organs like the skin, brain or the stomach
[Yes]
[No]
Doctor performs a biopsy from the affected organs
Biopsy is a test whereby a small sample of tissue from the affected nodes or organs will be taken
Doctor performs a biopsy on lymph node
Pathologist examines the biopsy sample under a microscope
Determine the type of lymphoma

Figure 8: Activity 2.1: Determine type of Lymphoma

Activity 2.2: Determine the stage of Lymphoma
Once the exact type of lymphoma is determined, the doctor will proceed with a number of tests to see how advanced the cancer is and how far it has spread. These tests include blood tests, chest x-ray, bone marrow aspiration, PET scan, CT scan, MRI scan of the chest/abdomen/pelvis, lumbar puncture etc (Figure 8).

Activity 2.3: Determine the prognosis of Lymphoma
Next, the doctor will look into other factors to determine the prognosis of the disease.

Activity 3: Propose treatment
When the above investigations are completed, the doctor and oncologist will counsel the patient regarding the best treatment options available, based on the type and the stage of the disease and some prognostic factors (Figure 9).

There are four main types of treatment normally used to cure Lymphomas:

- Chemotherapy – Using drugs as infusions into the patient's veins.
- Radiotherapy – Using high energy rays over the affected areas.
- Biological therapy or antibiotic therapy – Using drugs like Rituximab to target special molecules on the cancer cells.
- Bone marrow or stem cell transplant – Using high doses of chemotherapy or radiation to kill all cancer cells while saving the bone marrow with transplantation of marrow or stem cells.

Besides proposing the treatment options to patient, the doctor also need to explain to the patient about the risks of taking the particular treatment and chance of recovery. Upon endorsement by the patient, the doctor will schedule the treatments for the patient.

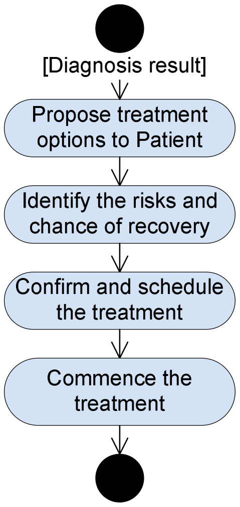

[Diagnosis result]
Propose treatment options to Patient
Identify the risks and chance of recovery
Confirm and schedule the treatment
Commence the treatment

Figure 9: Activity 3: Propose treatment of Lymphoma

**4.1.2 Requirement analysis**
Table 3 below lists some of the standard requirements for a clinical DSS to support Lymphoma diagnosis and treatment recommendation:

| # | Description |
|---|---|
| 1. | The medical diagnosing function requires understanding of patient medical history, symptoms, drug-drug interactions, knowledge of diseases in general as well as the general population. |
| 2. | The system should auto generate sequential record number for each clinical record. |
| 3. | Minimal level of demographic details about the patient need to be captured in the patient master record include: Full Name, Race, Gender, Date of Birth, Marital Status, Address, |

| # | Description |
|---|---|
| | Occupation etc. |
| 4. | All patient master records can be recognized by a unique patient identification number. The system must be able to detect duplication of patient master records with the same new or old IC Number. The system shall automatically merge the patient records, if duplication of master records for the same patient is found. |
| 5. | Treatments details need to be captured include: Procedures, Duration, and Outcomes. System should auto generate sequential record number for each treatment record. |
| 6. | The system should be able to utilize varying levels of data in order to diagnose a patient. |
| 7. | The system should provide reasoning for the medical diagnosis. Such feature would allow the medical professionals to understand the reasons for a specific decision that may have been made. |
| 8. | The system must able to display data at both summary and detail levels, and allow users to drill down to analyze a specific result. |
| 9. | The system must facilitate a multi-dimensional data model presentation. Patient and medical details can be queried by multiple factors, such as location, gender, age range, blood group, race etc. |
| 10. | The system should have to be updateable and adapted to constant changes that accompany the scientific development. |

Table 3: Requirements of a clinical data warehouse

## 4.2 Architecture Design

Figure 10 illustrates the overall architecture of the clinical data warehouse. Data is imported from several sources and transformed within a staging database before it is integrated and stored in the production data warehouse for further analysis.

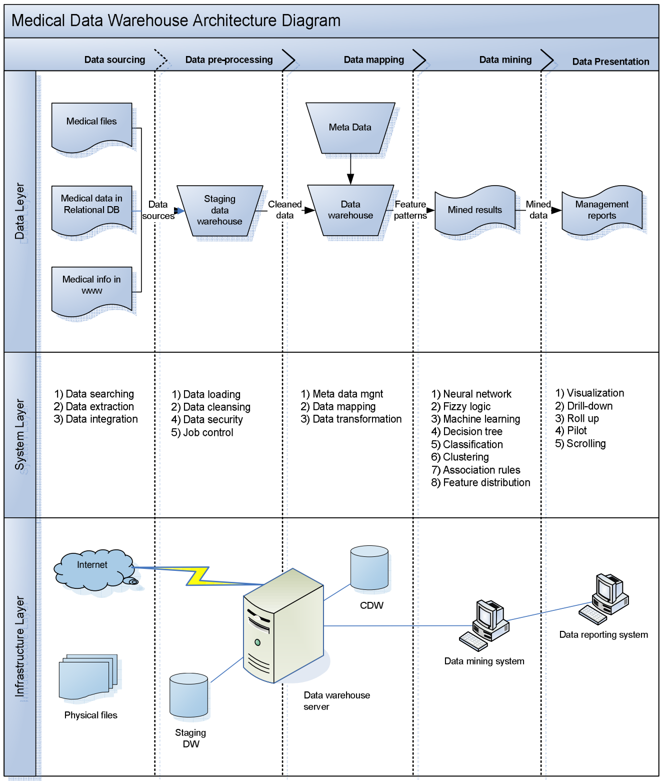


Figure 10: Clinical data warehouse architecture

### 4.2.1 Data architecture

The detailed star schema, as shown in Figure 11, demonstrates the data layer architecture of the clinical data warehouse. The designed clinical data warehouse uses a de-normalized schema, as shown in the star schema, the dimensional tables, such as diseaseType_Dimension and diseaseStage_ Dimension, contain de-normalized or redundant data. Such de-normalization may facilitate data mining and Business Intelligence techniques.

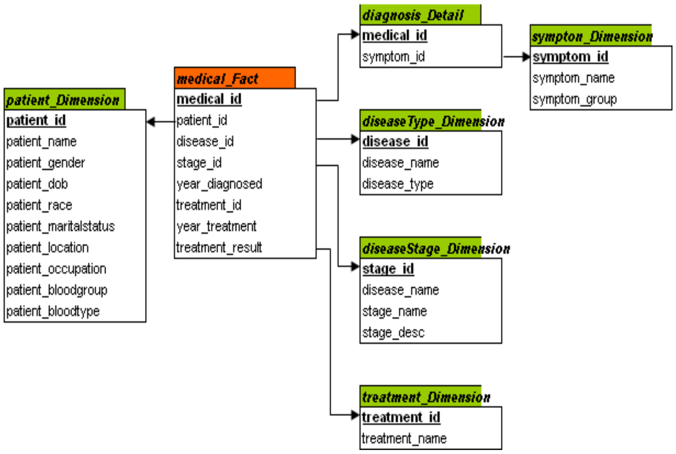


Figure 11: Clinical data warehouse star diagram

From the star diagram, a multidimensional view of the clinical data can be represented as illustrated in Figure 12. A multidimensional database can then be projected in a relational database or object-oriented database.

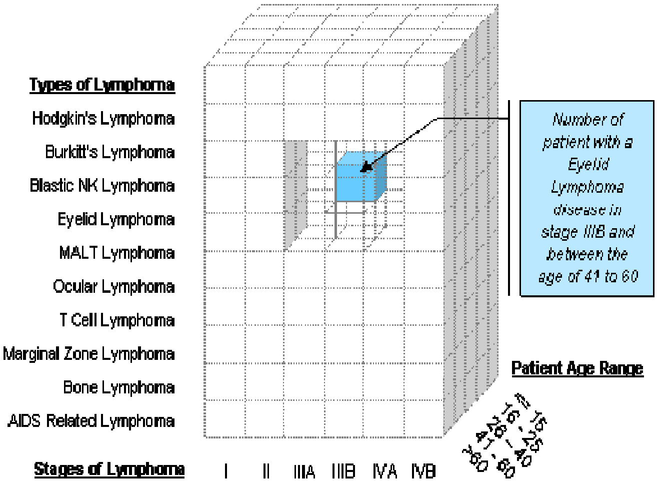


Figure 12: Multidimensional view of the clinical data

Below describes the fact and dimension tables of the clinical data warehouse in detail:

Fact table
The Fact table that describes the subject matter is named *medical_fact*. The table consists of the case ID, patient ID, disease ID, disease stage ID, medical check-up date, treatment date, treatment given, treatment result etc. The symptoms a patient possess and the diagnosed conditions are decomposed into another table named *diagnosis_ detail.*

Dimension tables
The table 4 below listed the 6 Dimension tables that detail each entity in the *medical_Fact* table.

| Table name | Table description |
|---|---|
| *patient_ dimension* | A table that stores patient information, such as patient name, gender, age, race, location, occupation etc. The data is used to show demographic data for the Lymphoma disease. |
| *diseasetype_ dimension* | A table that stores all the diseases and the types of the diseases. In this project only information pertaining to the subject matter, i.e. Lymphoma disease will be created. |
| *diseasestage _dimension* | A table that stores the staging systems which are specific for each type of disease. In this project only information pertaining to the subject matter, i.e. Lymphoma disease will be created. |
| *treatment_ dimension* | A table that stores all the treatment options. |
| *symptom_ dimension* | A table that stores all the symptoms, the normal condition value and abnormal condition value. |

Table 4: Dimension tables of the clinical data warehouse

### 4.2.2 System architecture

Looking at the scope of this project, the following data warehousing processes are required:

1. Data Extract, Transform and Load (ETL) process
   - To identify data pertaining to Lymphoma.
   - To extract the data from sources, such as web pages, electronic medical data, physical files.
   - To load the data onto staging database.

2. Data Cleansing process
   - To cleanse the data and remove noise data.
   - To convert the data to common format.

3. Data Migration process
   - To migrate cleansed and processed data from staging database to the data warehouse.

### 4.2.3 Infrastructure architecture

The designed clinical data warehouse is built with:

- Relational database management system: MySQL community server 5.0 which consists of MySQL Administrator, MySQL Query browser, MySQL Data migration toolkit.

- Operating system: Windows XP professional with Service pack 3 running on a Dell PC, Intel Pentium IV CPU 2.8GHz, 512MB RAM, 30GB hard disk space.

- ETL tool: self-developed PL/SQL scripts to extract, load, transform and cleanse the data onto the data warehouse, Microsoft Office Professional, Microsoft Visio Professional.

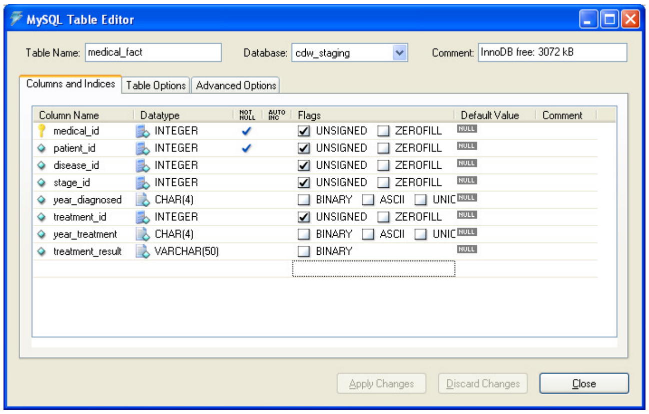
## 4.3 Physical Development

This stage involves:

- ***Data warehouse creation*** – create the data warehouse using the MySQL RDBMS based on the architecture designed.
- ***Data provisioning*** – extracting and loading data from source into data warehouse.
- ***Data cleansing*** – cleansing and transforming raw data to ensure data quality and integrity.

### 4.3.1 Data Warehouse Creation

Upon completion of the infrastructure setup, a new schema called "cdw_staging" is created using MySQL, based on the star schema diagram. Sample screenshot of the fact and dimension tables created are included in Figure 13 and Figure 14.

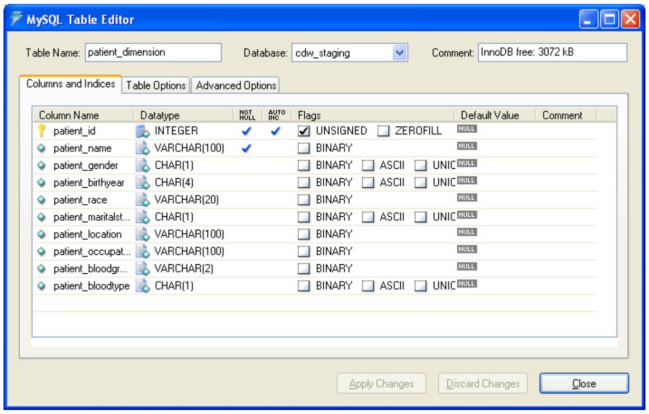


MySQL Table Editor

Table Name: medical_fact   Database: cdw_staging   Comment: InnoDB free: 3072 kB

| Column Name | Datatype | NOT NULL | AUTO INC | Flags | Default Value | Comment |
|---|---|---|---|---|---|---|
| medical_id | INTEGER | ✓ | | ☑ UNSIGNED ☐ ZEROFILL | NULL | |
| patient_id | INTEGER | ✓ | | ☑ UNSIGNED ☐ ZEROFILL | NULL | |
| disease_id | INTEGER | | | ☑ UNSIGNED ☐ ZEROFILL | NULL | |
| stage_id | INTEGER | | | ☑ UNSIGNED ☐ ZEROFILL | NULL | |
| year_diagnosed | CHAR(4) | | | ☐ BINARY ☐ ASCII ☐ UNIC | NULL | |
| treatment_id | INTEGER | | | ☑ UNSIGNED ☐ ZEROFILL | NULL | |
| year_treatment | CHAR(4) | | | ☐ BINARY ☐ ASCII ☐ UNIC | NULL | |
| treatment_result | VARCHAR(50) | | | ☐ BINARY | NULL | |

Figure 13: Medical_Fact table



MySQL Table Editor

Table Name: patient_dimension   Database: cdw_staging   Comment: InnoDB free: 3072 kB

| Column Name | Datatype | NOT NULL | AUTO INC | Flags | Default Value | Comment |
|---|---|---|---|---|---|---|
| patient_id | INTEGER | ✓ | ✓ | ☑ UNSIGNED ☐ ZEROFILL | NULL | |
| patient_name | VARCHAR(100) | ✓ | | ☐ BINARY | NULL | |
| patient_gender | CHAR(1) | | | ☐ BINARY ☐ ASCII ☐ UNIC | NULL | |
| patient_birthyear | CHAR(4) | | | ☐ BINARY ☐ ASCII ☐ UNIC | NULL | |
| patient_race | VARCHAR(20) | | | ☐ BINARY | NULL | |
| patient_maritalst... | CHAR(1) | | | ☐ BINARY ☐ ASCII ☐ UNIC | NULL | |
| patient_location | VARCHAR(100) | | | ☐ BINARY | NULL | |
| patient_occupat... | VARCHAR(100) | | | ☐ BINARY | NULL | |
| patient_bloodgr... | VARCHAR(2) | | | ☐ BINARY | NULL | |
| patient_bloodtype | CHAR(1) | | | ☐ BINARY ☐ ASCII ☐ UNIC | NULL | |

Figure 14: patient_Dimension table

### 4.3.2 Data Provisioning

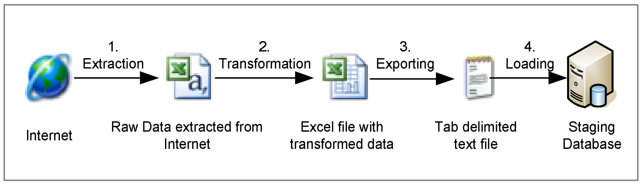
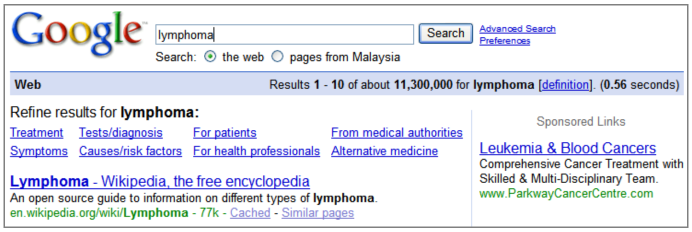


Figure 15: ETL process of the clinical data warehouse

1. Data Extraction

There are two main extractions will be involved: (i) general Lymphoma information extraction, such Lymphoma stage, Lymphoma types, symptoms etc, and (ii) patient medical records extraction.

Web search engines, such as Google, Yahoo etc., are used to dig relevant data from the Internet [13]. General Lymphoma information can be found and extracted easily from the web. Figure 16 shows millions of information resulted from the key word search of "Lymphoma" using Google search. Once identified, data are manually copied and pasted to MS Excel files.



Results **1** - **10** of about **11,300,000** for **lymphoma** [definition]. (**0.56** seconds)

Refine results for **lymphoma:**

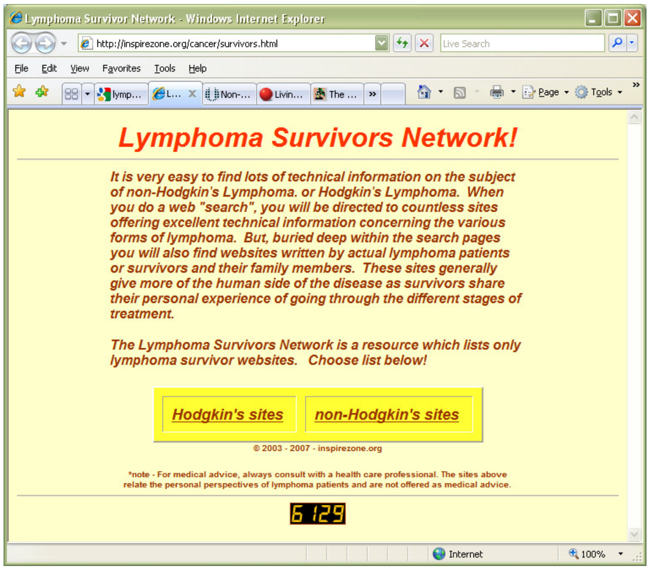
Treatment Tests/diagnosis For patients From medical authorities
Symptoms Causes/risk factors For health professionals Alternative medicine

**Lymphoma** - Wikipedia, the free encyclopedia
An open source guide to information on different types of **lymphoma**.
en.wikipedia.org/wiki/**Lymphoma** - 77k - Cached - Similar pages

Sponsored Links
Leukemia & Blood Cancers
Comprehensive Cancer Treatment with Skilled & Multi-Disciplinary Team.
www.ParkwayCancerCentre.com

Figure 16: Web search result of "Lymphoma"

Real world patient medical records are somewhat difficult to be seized due to patient data confidential and security policies,. This paper uses information from the web, such as Lymphoma survivors network, Lymphoma patient blogs etc, as shown in Figure 17 and Figure 18, as well as mock-up data, instead of from a hospital medical database.


Lymphoma Survivor Network - Windows Internet Explorer
http://inspirezone.org/cancer/survivors.html
Live Search
File Edit View Favorites Tools Help
lymp... L... Non-... Livin... The ...
Page Tools
Internet
100%

Lymphoma Survivors Network!

*It is very easy to find lots of technical information on the subject of non-Hodgkin's Lymphoma. or Hodgkin's Lymphoma. When you do a web "search", you will be directed to countless sites offering excellent technical information concerning the various forms of lymphoma. But, buried deep within the search pages you will also find websites written by actual lymphoma patients or survivors and their family members. These sites generally give more of the human side of the disease as survivors share their personal experience of going through the different stages of treatment.*

*The Lymphoma Survivors Network is a resource which lists only lymphoma survivor websites. Choose list below!*

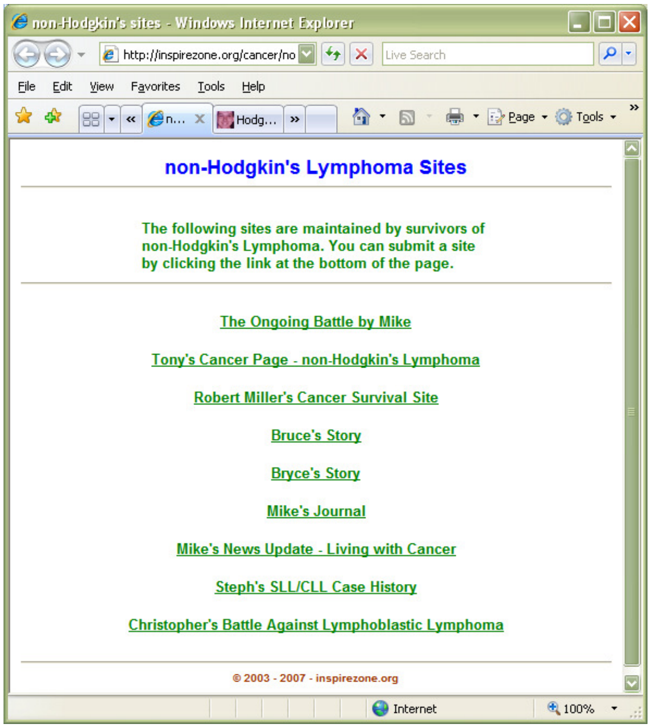
Hodgkin's sites | non-Hodgkin's sites

© 2003 - 2007 - inspirezone.org

*note - For medical advice, always consult with a health care professional. The sites above relate the personal perspectives of lymphoma patients and are not offered as medical advice.

6 129

Figure 17: Lymphoma Survivor Network Website


non-Hodgkin's sites - Windows Internet Explorer
http://inspirezone.org/cancer/no
Live Search
File Edit View Favorites Tools Help
n... Hodg...
Page Tools
Internet
100%

non-Hodgkin's Lymphoma Sites

The following sites are maintained by survivors of non-Hodgkin's Lymphoma. You can submit a site by clicking the link at the bottom of the page.

The Ongoing Battle by Mike

Tony's Cancer Page - non-Hodgkin's Lymphoma

Robert Miller's Cancer Survival Site

Bruce's Story

Bryce's Story

Mike's Journal

Mike's News Update - Living with Cancer

Steph's SLL/CLL Case History

Christopher's Battle Against Lymphoblastic Lymphoma

© 2003 - 2007 - inspirezone.org

Figure 18: Lymphoma Survivor List

## 2. Data Transformation

The staging database is created in the most de-normalized form. Apart from the primary key, all other fields allow NULL value. Primary keys are filled up with incremental numbers before they are loaded onto database.

## 3. Data Loading

Data that are ready to be loaded to the staging database are saved as a tab delimited text file (.txt) and imported to database:

```
LOAD DATA INFILE 'raw_symptom.txt'
IGNORE INTO TABLE symptom_dimension
FIELDS TERMINATED BY '\t'
OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 LINES;
```

Upon completion, a success message of the task will be displayed:

```
# rows affected by the last command, no
resultset returned.
```

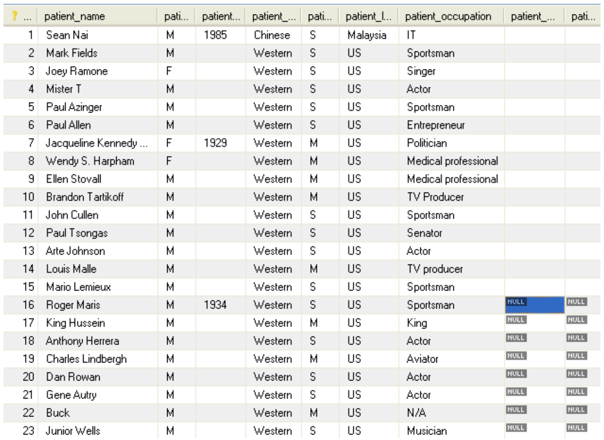
Below listed the problems encountered during data loading to staging database and the resolutions.

1. Problem description:
   Auto increment did not work for primary key column(s) during data import.

   Error message:
   Incorrect integer value: 'Sean' for column 'patient_id' at row 1.

   Action taken:
   Insert the primary key into the column in Excel file before exported to tab delimited text file.

2. Problem description:
   Import failed on a CHAR column that contains NULL data, because the LOAD DATA command does not read the CHAR size.



| ... | patient_name | pati... | patient... | patient_... | pati... | patient_l... | patient_occupation | patient_... | pati... |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Sean Nai | M | 1985 | Chinese | S | Malaysia | IT | | |
| 2 | Mark Fields | M | | Western | S | US | Sportsman | | |
| 3 | Joey Ramone | F | | Western | S | US | Singer | | |
| 4 | Mister T | M | | Western | S | US | Actor | | |
| 5 | Paul Azinger | M | | Western | S | US | Sportsman | | |
| 6 | Paul Allen | M | | Western | S | US | Entrepreneur | | |
| 7 | Jacqueline Kennedy ... | F | 1929 | Western | M | US | Politician | | |
| 8 | Wendy S. Harpham | F | | Western | M | US | Medical professional | | |
| 9 | Ellen Stovall | M | | Western | M | US | Medical professional | | |
| 10 | Brandon Tartikoff | M | | Western | M | US | TV Producer | | |
| 11 | John Cullen | M | | Western | S | US | Sportsman | | |
| 12 | Paul Tsongas | M | | Western | S | US | Senator | | |
| 13 | Arte Johnson | M | | Western | S | US | Actor | | |
| 14 | Louis Malle | M | | Western | M | US | TV producer | | |
| 15 | Mario Lemieux | M | | Western | S | US | Sportsman | | |
| 16 | Roger Maris | M | 1934 | Western | S | US | Sportsman | NULL | NULL |
| 17 | King Hussein | M | | Western | M | US | King | NULL | NULL |
| 18 | Anthony Herrera | M | | Western | S | US | Actor | NULL | NULL |
| 19 | Charles Lindbergh | M | | Western | M | US | Aviator | NULL | NULL |
| 20 | Dan Rowan | M | | Western | S | US | Actor | NULL | NULL |
| 21 | Gene Autry | M | | Western | S | US | Actor | NULL | NULL |
| 22 | Buck | M | | Western | M | US | N/A | NULL | NULL |
| 23 | Junior Wells | M | | Western | S | US | Musician | NULL | NULL |

Table 5: Data with NULL value

Error message:
Row 6 doesn't contain data for all columns.

Action taken:

a. Make sure all CHAR columns are filled with data or a " " (empty space)..

b. Modify the LOAD DATA command with a IGNORE statement, so that the CHAR column in data import is read as "NULL" when a field separator or a line separator is reached.

### 4.3.3 Data Cleansing

Data loaded to the staging database are normally contains certain types of logical or structural problems, such as data anomaly, repetition of data, full facts etc. Several data cleansing steps, as demonstrated in Figure 19, are performed to ensure data loaded are of certain qualities, i.e. Accuracy, Consistency, Completeness, Integrity, Validity, Timeliness, for further data processing.

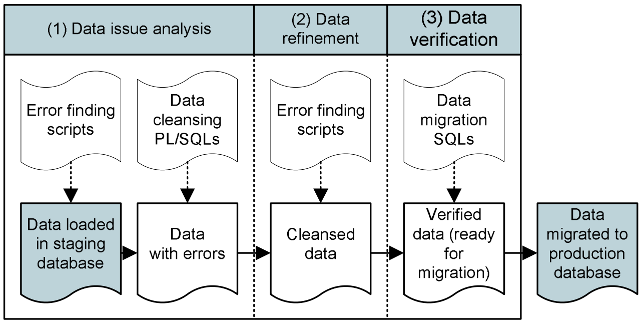


Figure 19: Data cleansing process

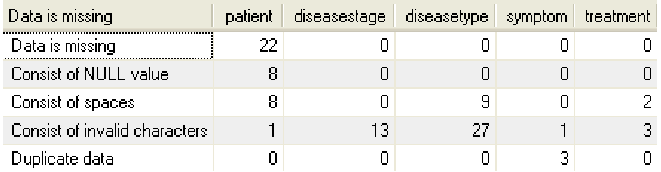
1. Data Issue Analysis

Data are being further studied, PL/SQL scripts are written to analyze the data issues in 3 areas:

i) Noisy data, inconsistent or irrelevant data, and duplicate data are reported with the SQL command:

```
...
WHERE symptom_name IS NULL
OR t.treatment_name = ""
OR t.treatment_name REGEXP
"[[...]|[.'.]|[.,.]|[.~.]|[.*.]|[.#.]|
[.-.]|[.+.]|[.#.]|[.%.]|[.@.]|[.(.]|
[.).]]")
```

The query result is as in table 6 below:



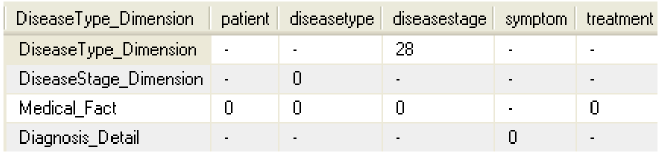
| Data is missing | patient | diseasestage | diseasetype | symptom | treatment |
|---|---|---|---|---|---|
| Data is missing | 22 | 0 | 0 | 0 | 0 |
| Consist of NULL value | 8 | 0 | 0 | 0 | 0 |
| Consist of spaces | 8 | 0 | 9 | 0 | 2 |
| Consist of invalid characters | 1 | 13 | 27 | 1 | 3 |
| Duplicate data | 0 | 0 | 0 | 3 | 0 |

Table 6: Noisy, inconsistent or irrelevant data, and duplicate data

ii) Missing value, false data or conflicts in data are identified using the SQL command:

```
...
SELECT COUNT(*) AS 'diseasestage'
FROM diseasestage_dimension ds
WHERE disease_name NOT IN
 (SELECT DISTINCT dt.disease_name
  FROM diseasetype_dimension dt
 )
...
```

The query result is shown as in table 7 below:



| DiseaseType_Dimension | patient | diseasetype | diseasestage | symptom | treatment |
|---|---|---|---|---|---|
| DiseaseType_Dimension | - | - | 28 | - | - |
| DiseaseStage_Dimension | - | 0 | - | - | - |
| Medical_Fact | 0 | 0 | 0 | - | 0 |
| Diagnosis_Detail | - | - | - | 0 | - |

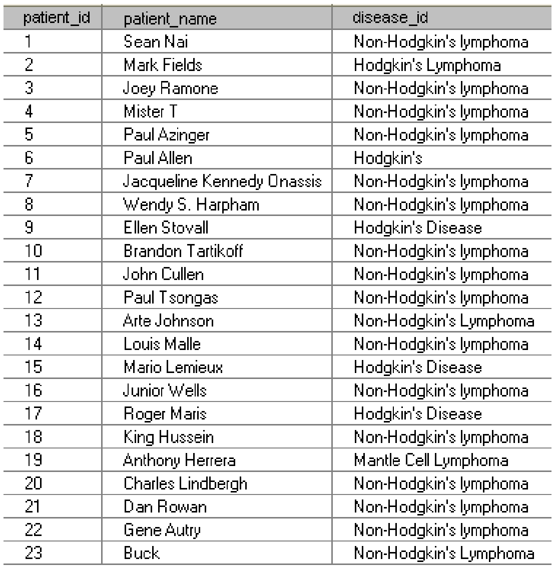
Table 7: Missing value, false data or conflicts in data

iii) Data is in inappropriate forms for mining, for instance, patient's DOB is in dd-mmm-yyyy format, and number of incident by age range is not possible.

2. Data Refinement

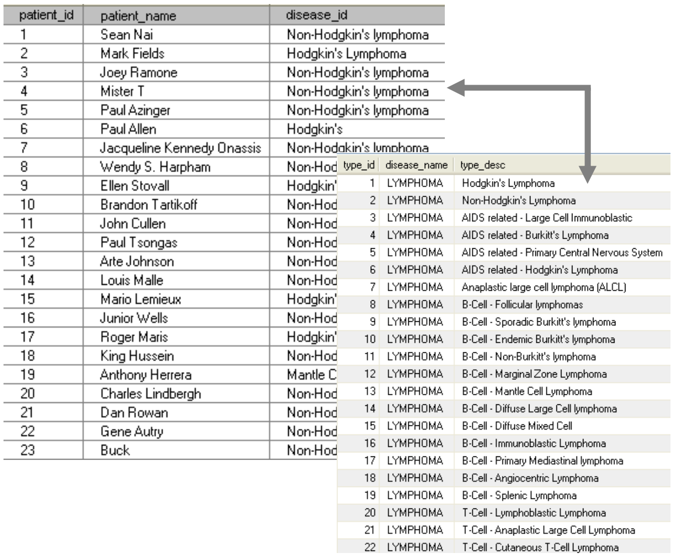
- *Fix noise, duplicate, inconsistent or irrelevant data* – a PL/SQL script is generated to remove the empty space at the end of each column, which has a VARCHAR column type, using `TRIM` command. Another PL/SQL is written to remove duplicate records from the dimension tables to ensure data uniqueness.

- *Rectify false data or conflicts in data* – by mapping certain duplicate data against a known list of entities to reduce data anomaly or integrity problem. For example, as shown in Figure 20 and Figure 21, cross reference the patients' diagnosed disease to the diseaseType_Dimension table. Multiple instances of the same information will be represented by a single instance in the file.



| patient_id | patient_name | disease_id |
|---|---|---|
| 1 | Sean Nai | Non-Hodgkin's lymphoma |
| 2 | Mark Fields | Hodgkin's Lymphoma |
| 3 | Joey Ramone | Non-Hodgkin's lymphoma |
| 4 | Mister T | Non-Hodgkin's lymphoma |
| 5 | Paul Azinger | Non-Hodgkin's lymphoma |
| 6 | Paul Allen | Hodgkin's |
| 7 | Jacqueline Kennedy Onassis | Non-Hodgkin's lymphoma |
| 8 | Wendy S. Harpham | Non-Hodgkin's lymphoma |
| 9 | Ellen Stovall | Hodgkin's Disease |
| 10 | Brandon Tartikoff | Non-Hodgkin's lymphoma |
| 11 | John Cullen | Non-Hodgkin's lymphoma |
| 12 | Paul Tsongas | Non-Hodgkin's lymphoma |
| 13 | Arte Johnson | Non-Hodgkin's Lymphoma |
| 14 | Louis Malle | Non-Hodgkin's lymphoma |
| 15 | Mario Lemieux | Hodgkin's Disease |
| 16 | Junior Wells | Non-Hodgkin's lymphoma |
| 17 | Roger Maris | Hodgkin's Disease |
| 18 | King Hussein | Non-Hodgkin's lymphoma |
| 19 | Anthony Herrera | Mantle Cell Lymphoma |
| 20 | Charles Lindbergh | Non-Hodgkin's lymphoma |
| 21 | Dan Rowan | Non-Hodgkin's lymphoma |
| 22 | Gene Autry | Non-Hodgkin's lymphoma |
| 23 | Buck | Non-Hodgkin's Lymphoma |

Figure 20: Pre-transformed data (data anomaly)



| patient_id | patient_name | disease_id |
|---|---|---|
| 1 | Sean Nai | Non-Hodgkin's lymphoma |
| 2 | Mark Fields | Hodgkin's Lymphoma |
| 3 | Joey Ramone | Non-Hodgkin's lymphoma |
| 4 | Mister T | Non-Hodgkin's lymphoma |
| 5 | Paul Azinger | Non-Hodgkin's lymphoma |
| 6 | Paul Allen | Hodgkin's |
| 7 | Jacqueline Kennedy Onassis | Non-Hodgkin's lymphoma |
| 8 | Wendy S. Harpham | Non-Hod |
| 9 | Ellen Stovall | Hodgkin' |
| 10 | Brandon Tartikoff | Non-Hod |
| 11 | John Cullen | Non-Hod |
| 12 | Paul Tsongas | Non-Hod |
| 13 | Arte Johnson | Non-Hod |
| 14 | Louis Malle | Non-Hod |
| 15 | Mario Lemieux | Hodgkin' |
| 16 | Junior Wells | Non-Hod |
| 17 | Roger Maris | Hodgkin' |
| 18 | King Hussein | Non-Hod |
| 19 | Anthony Herrera | Mantle C |
| 20 | Charles Lindbergh | Non-Hod |
| 21 | Dan Rowan | Non-Hod |
| 22 | Gene Autry | Non-Hod |
| 23 | Buck | Non-Hod |

| type_id | disease_name | type_desc |
|---|---|---|
| 1 | LYMPHOMA | Hodgkin's Lymphoma |
| 2 | LYMPHOMA | Non-Hodgkin's Lymphoma |
| 3 | LYMPHOMA | AIDS related - Large Cell Immunoblastic |
| 4 | LYMPHOMA | AIDS related - Burkitt's Lymphoma |
| 5 | LYMPHOMA | AIDS related - Primary Central Nervous System |
| 6 | LYMPHOMA | AIDS related - Hodgkin's Lymphoma |
| 7 | LYMPHOMA | Anaplastic large cell lymphoma (ALCL) |
| 8 | LYMPHOMA | B-Cell - Follicular lymphomas |
| 9 | LYMPHOMA | B-Cell - Sporadic Burkitt's lymphoma |
| 10 | LYMPHOMA | B-Cell - Endemic Burkitt's lymphoma |
| 11 | LYMPHOMA | B-Cell - Non-Burkitt's lymphoma |
| 12 | LYMPHOMA | B-Cell - Marginal Zone Lymphoma |
| 13 | LYMPHOMA | B-Cell - Mantle Cell Lymphoma |
| 14 | LYMPHOMA | B-Cell - Diffuse Large Cell lymphoma |
| 15 | LYMPHOMA | B-Cell - Diffuse Mixed Cell |
| 16 | LYMPHOMA | B-Cell - Immunoblastic Lymphoma |
| 17 | LYMPHOMA | B-Cell - Primary Mediastinal lymphoma |
| 18 | LYMPHOMA | B-Cell - Angiocentric Lymphoma |
| 19 | LYMPHOMA | B-Cell - Splenic Lymphoma |
| 20 | LYMPHOMA | T-Cell - Lymphoblastic Lymphoma |
| 21 | LYMPHOMA | T-Cell - Anaplastic Large Cell Lymphoma |
| 22 | LYMPHOMA | T-Cell - Cutaneous T-Cell Lymphoma |

Figure 21: Transformed data (data mapped to cross reference table)

- *Reduce dimensionality of the database* to enable better data mining process. For example, by repeating the disease_name value in both diseaseType_Dimension and diseaseStage_ Dimension tables.

- *Transform data into appropriate forms for data mining* – Data transformation operations such as grouping on sums, percentages of overall totals, date/month/year format conversion are performed, for instance, Patient's DOB is converted to year to allow age to be calculated easily.

3. Data Verification

The same PL/SQL scripts created for data cleansing are executed again to verify and ensure that the data issues have been resolved after the data cleansing process.

## 4.4 Implementation and Deployment

Next, the cleansed and transformed data are migrated onto an operational data warehouse.

### 4.4.1 Data Warehouse creation

The clinical data warehouse is created in 4 steps:

1. Migration of database schema

A new schema called "cdw_production" is created. The `CREATE` statements of all the finalized tables in the staging schema will be captured, modified and executed in the production schema.

2. Verification of database schema

Next, the production data warehouse is refreshed and all table structures are verified against the staging database to ensure that the correct schema has been migrated.

3. Migration of production data

All data in the staging database are exported to tab delimited text files (.txt) and subsequently loaded to production data warehouse using the `LOAD DATA INFILE` command.

4. Verification of production data

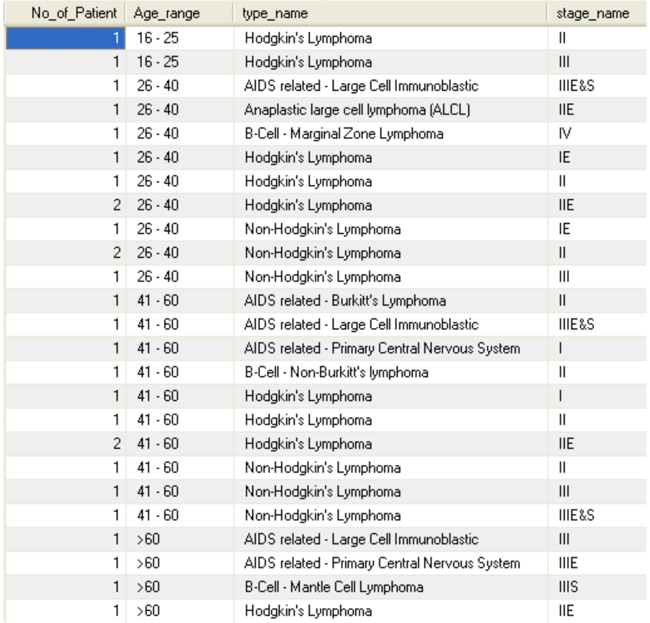
Finally, all migrated data are verified to ensure they are of certain "qualities" (i.e. Accuracy, Completeness, Consistency, Timeliness, and Validity) for data mining. Reports are generated to compare production data against the staging database.

### 4.4.2 Data Query

The ultimate objective of this stage is to put the data warehouse into operation, whereby data can be queried and feed into a data mining software for processing and reporting.

The data should be able to be queried and presented in a multidimensional view as illustrated in figure 12: Multidimensional view of the clinical data.

The query result is as in Figure 18 below:



| No_of_Patient | Age_range | type_name | stage_name |
|---|---|---|---|
| 1 | 16 - 25 | Hodgkin's Lymphoma | II |
| 1 | 16 - 25 | Hodgkin's Lymphoma | III |
| 1 | 26 - 40 | AIDS related - Large Cell Immunoblastic | IIIE&S |
| 1 | 26 - 40 | Anaplastic large cell lymphoma (ALCL) | IIE |
| 1 | 26 - 40 | B-Cell - Marginal Zone Lymphoma | IV |
| 1 | 26 - 40 | Hodgkin's Lymphoma | IE |
| 1 | 26 - 40 | Hodgkin's Lymphoma | II |
| 2 | 26 - 40 | Hodgkin's Lymphoma | IIE |
| 1 | 26 - 40 | Non-Hodgkin's Lymphoma | IE |
| 2 | 26 - 40 | Non-Hodgkin's Lymphoma | II |
| 1 | 26 - 40 | Non-Hodgkin's Lymphoma | III |
| 1 | 41 - 60 | AIDS related - Burkitt's Lymphoma | II |
| 1 | 41 - 60 | AIDS related - Large Cell Immunoblastic | IIIE&S |
| 1 | 41 - 60 | AIDS related - Primary Central Nervous System | I |
| 1 | 41 - 60 | B-Cell - Non-Burkitt's lymphoma | II |
| 1 | 41 - 60 | Hodgkin's Lymphoma | I |
| 1 | 41 - 60 | Hodgkin's Lymphoma | II |
| 2 | 41 - 60 | Hodgkin's Lymphoma | IIE |
| 1 | 41 - 60 | Non-Hodgkin's Lymphoma | II |
| 1 | 41 - 60 | Non-Hodgkin's Lymphoma | III |
| 1 | 41 - 60 | Non-Hodgkin's Lymphoma | IIIE&S |
| 1 | >60 | AIDS related - Large Cell Immunoblastic | III |
| 1 | >60 | AIDS related - Primary Central Nervous System | IIIE |
| 1 | >60 | B-Cell - Mantle Cell Lymphoma | IIIS |
| 1 | >60 | Hodgkin's Lymphoma | IIE |

Figure 22: Multidimensional query result

### 4.4 Evaluation

The finalized production data will be queried and feed into a data mining software for benchmarking, reporting and analysis. The data will be evaluated against some acceptance criteria, such as applicability, novelty, understandability, representative, provability, validity etc. Research on how the evaluation of the developed data warehouse and how it would support the data mining process will be further discussed in a separate paper.

## 5 Conclusions

Several challenges were encountered during the data warehouse development:

1. Data warehouse development requires specialized skills that are very different from a typical database development.

2. Data cleansing plays the most critical role in a data warehouse development. It is time-consuming yet must be handled thoroughly and with full effort, to avoid no garbage in garbage out.

3. In an ideal solution, data should be extracted from physical medical files, such as patient medical records, blood tests, urine test results, x-ray results, CT scan results etc, or retrieved directly from the operational medical system. However, due to data security issue, real world medical data cannot be seized easily. Hence, this project uses data from the Internet or mocked-up data. However, the proposed methodology and architecture framework can be applied to an operational Clinical data warehouse development. Additionally, the structure of the data warehouse is designed in such a flexible way that it can be extended to other diseases.

Having said that, this paper has reached it goals to propose and develop a methodology and architecture for a specific disease clinical data warehouse. The potential benefits of the developed data warehouse are:

1. The data warehouse can be used as a fundamental building block of a Clinical DSS to improve the quality of Lymphoma diagnosis and treatment decision support.

2. The data warehouse can also be used with data mining tools such as Integral Solution's Clementine, Thinking Machines' Darwin, Cognos' Scenario, IBM's Intelligent Miner, SAS' Enterprise Miner, Data Mind's DataCrunche, open source tools such as MIDAS, Tyson software etc.

3. The structure of the data warehouse is designed in such a flexible way that it can be extended to other diseases.

*References:*

[1] Dorin Comaniciu1, Peter Meer, David J. Foran, Image-guided decision support system for pathology. Machine Vision and Applications, Report No. 11, August 16, 1999.

[2] Harvey Simon and David Zieve, Hodgkin's Lymphoma, In-Depth From A.D.A.M. The New York Time. 25 June 2008.

[3] The "Cancer Trends Progress Report – 2007", U.S. National Cancer Institute. 2007.

[4] W. H Inmon, R. D.Hackethorn, Using the Data Warehouse. New York: John.

[5] Wiley & Sons(1994) Sami Benmechiche, Carol Chouinard, Ross Christen, Richard Kupcunas, Deepak Goyal, Ajit Kumar, “Clinical data is gold. Data warehouse are Ft Knox”, Using IT to turn your data into a strategic information asset, Deloitte, 2006.

[6] Anne Tchounikine, Maryvonne Miquel, André Flory, Information Warehouse for Medical Research, Data Warehousing and Knowledge Discovery, Volume 2114/2001, Springer Berlin / Heidelberg, January 01, 2001.

[7] Anne-Muriel Arigon1 , Maryvonne Miquel, Anne Tchounikine, Multimedia data warehouses: a multiversion model and a medical application, Multimedia Tools and Applications, Springer Netherlands, Volume 35, Number 1 / October, 2007.

[8] Torben Bach Pedersen, Christian S. Jense, Research Issues in Clinical Data Warehousing, Proceedings of SSDBM’98, July 1-3 1998.

[9] Laura Hadley, Developing a Data Warehouse Architecture, 2002.

[10] Jose Maria Cavero, Mario Piattini, Esperanza Marcos, Multidimensional modeling using MIDEA, 5th WSES CSCC, 3rd WSES MCP, 3rd WSES MCME International Conferences, 2001.

[11] Maris Klimavicus, Data warehouse development with EPC, Proceedings of the 5th WSEAS Int. Conf. on DATA NETWORKS, COMMUNICATIONS & COMPUTERS, Bucharest, Romania, October 16-17, 2006.

[12] Detailed Guide: Lymphoma, Non-Hodgkin Type, How is Non-Hodgkins Lymphoma staged, American Cancer Society Inc., August 29, 2007 Retrieved from http://www.cancer.org/.

[13] Teh Ying Wah, Ng Hooi Peng, Ching Sue Hok, Development of Specific Disease Data Warehouse for Developing Content from General Guide for Hypertension Screening, Referral and Follow Up, 7th WSEAS International Conference on APPLIED COMPUTER SCIENCE, Venice, Italy, November 21-23, 2007.