Evaluating a Data Warehouse for Lymphoma Diagnosis and Treatment Decision Support System

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

Abstract: - The Data Warehouse for Lymphoma Decision Support System is a data warehouse containing relevant Lymphoma medical information and patient data. The data warehouse provides the source for all current and historical clinical data to help doctors and other health professionals to improve the quality of diagnosis, clinical care and treatment recommendation decision making for Lymphoma patients.

This paper focuses on the evaluation of the architecture, data inputs as well as data warehousing processes and techniques of the Lymphoma data warehouse. An evaluation model based on Bill Inmon’s definition of data warehouse is proposed to evaluate the Lymphoma data warehouse.

Keywords: Data warehouse evaluation, Clinical data warehouse, Clinical Decision Support System, Lymphoma or Lymph Node cancer.

1 Introduction
Evaluation is the final stage in the development of data warehouse where the cleansed and finalized data are evaluated against some acceptance criteria, such as uniqueness, applicability, representative, provability, validity, understand-ability etc.

Several approaches have been adopted to evaluate a data warehouse. Some of them look at a broader view where the entire data warehousing project is assessed; others focus in a specific view where only the data warehouse is examined.

Approaches that evaluate the entire data warehousing project, such as the one proposed by David Heise [1], consider the following elements as listed in Table 1 below:

<table>
<thead>
<tr>
<th>#</th>
<th>Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Methodology</td>
<td>Does the approach and framework used to develop and implement the data warehouse project follow any industry best practices and well documented?</td>
</tr>
<tr>
<td>2</td>
<td>Databases</td>
<td>Does the database created supports huge data indexing and querying?</td>
</tr>
<tr>
<td>3</td>
<td>Metadata Repositories</td>
<td>Does the metadata repository facilitates integration and accommodate change management?</td>
</tr>
<tr>
<td>4</td>
<td>Design Tools</td>
<td>Does the tools and techniques employed to design and develop the data warehouse user friendly and ease to use?</td>
</tr>
<tr>
<td>5</td>
<td>Extract, Transfer and Loading</td>
<td>How effective and efficient the ETL tool and process are?</td>
</tr>
<tr>
<td>6</td>
<td>Reporting</td>
<td>Does the solution features any ad hoc queries or reporting tool?</td>
</tr>
<tr>
<td>7</td>
<td>OLAP</td>
<td>Does the solution provide any inline or sequential analytics that can be leveraged as part of a business process?</td>
</tr>
<tr>
<td>8</td>
<td>Data mining</td>
<td>Does the solution offers key data mining functions such as data modelling, multidimensional analysis, scoring and visualisation?</td>
</tr>
<tr>
<td>9</td>
<td>Data Mart Suites</td>
<td>Does the solution integrated with other components of a data mart?</td>
</tr>
</tbody>
</table>

Table 1: Elements to evaluate a data warehouse
On the other hand, Marc Demarest suggested that the evaluation criteria for data warehousing shall begin with a look back to where the data warehouse market began, i.e. to *Bill Inmon's Building The Data Warehouse*. Demarest proposed all the operational evaluation criteria apply to an online transaction processing (OLTP) system may also apply equally to a data warehouse, including the following [2]:

i. Boundary and capability of the data warehouse technology's physical storage capability.

ii. Loading and indexing performance of the data warehouse system.

iii. Operational integrity, reliability and manageability of the data warehouse system.

iv. Data connectivity support of the data warehouse system.

v. Query processing performance of the data warehouse system.

Milicevic and Batos [3] also suggested that the basic parameters for evaluation of a data warehouse are Extract, Transform and Loading (ETL) process speed, disk space consumption, query performances and user friendliness.

In this research work, an evaluation model based on Bill Inmon’s definition of data warehouse is proposed to evaluate the Lymphoma data warehouse.

2 The Evaluation Model

Data warehouse is defined as a “subject-oriented, integrated, time-variant and non-volatile collection of data in support of management’s decision making process” by the father of the data warehouse, Bill Inmon [4], in which:

i. **Subject-oriented**: All relevant data concerning a particular subject are gathered and stored in a single database.

ii. **Integrated**: Data that is gathered from a variety of data sources and merged into the data warehouse must be consistent in format, naming and other aspects.

iii. **Time-variant**: Data in a data warehouse are tagged with time to support both current and historical perspective measurements. Data are stored in the data warehouse for a long period of time, to facilitate the analysis of trends and relationship between data, in an attempt to improve decision making process.

iv. **Non-volatile**: Data in a data warehouse always stay static or stable to enable a highly consistent dimensional view of data. There is no modification or deletion performed against the data after it has been loaded into the database. Due to non-volatility, data is maintained in a consistent fashion, and a data warehouse can be heavily optimised for query processing.

Holding Inmon’s definition closely, the developed Lymphoma data warehouse is examined in 2 aspects, i.e. (a) data characteristic and (b) operational perspectives. The evaluation criteria are explained in Table 2 below.

<table>
<thead>
<tr>
<th>Perspectives</th>
<th>Subject-oriented</th>
<th>Integrated</th>
<th>Time-variant</th>
<th>Non-volatile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
<td>Reliability</td>
<td>Integrity</td>
<td>Capability</td>
<td>Manageability</td>
</tr>
<tr>
<td>How does the data evolve around a particular subject?</td>
<td>Does the data consistent in format?</td>
<td>Does the data tagged with a time?</td>
<td>How is the data maintained in the data warehouse?</td>
<td></td>
</tr>
<tr>
<td>How reliable are the data stored in the data warehouse?</td>
<td>How effective does the developed data cleansing technique cleanses the raw data and converts them into the proper format for data mining process?</td>
<td>Can the data warehouse host both historical and current data?</td>
<td>Is there any tool to manage and fine tune the data warehouse?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the maximum physical storage of the data warehouse?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Operational</th>
<th>Data extraction</th>
<th>Data cleansing</th>
<th>Data loading</th>
<th>Data querying</th>
</tr>
</thead>
<tbody>
<tr>
<td>How useful does the data extraction technique searches the relevant data?</td>
<td>How effective does the developed data cleansing technique cleanses the raw data and converts them into the proper format for data mining process?</td>
<td>How helpful does the developed data loading technique captures raw data from various data sources?</td>
<td>How well does the developed data querying technique performs on table selection, complex SQL queries etc.?</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Evaluation criteria of the clinical data warehouse
3 The Lymphoma Data Warehouse

3.1 The Lymphoma data warehouse
This paper reviews the evaluation of a clinical data warehouse for the Lymphoma disease. The Lymphoma or Lymph node cancer is selected as the subject matter of this clinical data warehouse research work, as it is one of the cancers on the rise in the past 10 years and requires greater efforts at control [5].

The development of the clinical data warehouse focuses in 3 areas:

i. Data stored in the clinical data warehouse
This research work aims to develop an architecture and methodology to build a clinical data warehouse for the Lymphoma cancer data. Data is derived from several sources and transformed into a de-normalized dimensional model.

Due to patient data confidential and privacy policies, all data are sourced from the Internet, instead of from a hospital medical database. Patient information from Lymphoma survivor’s network, Lymphoma patient blogs etc, as well as mock-up data, are used in this research work, to simulate the behaviour of the data warehouse framework.

The Lymphoma data warehouse has the following characteristics that allow it to be able to use mocked-up data without biasing on the performance of the proposed data warehouse model:

- The data warehouse does not exist yet and may change behaviour during actual implementation;
- The data warehouse includes information and methods exclusively for testing purposes;
- The mocked-up data supplies non-deterministic results, which mean the data structure allows return of different results depending on different retrieval parameters [6].

ii. Data warehousing processes and techniques
This research work only involves the related data warehousing processes, including Data Extraction, Transformation, Loading, Cleansing, Integration and querying processes. Data mining and Business Intelligent (BI) are not provided in this research work.

iii. Outcome
The built data warehouse can be used as a backend building block for a Clinical DSS, and result from the data querying process would help in knowledge presentation. However, using the data to achieve the data mining objective, for example, for the verification and validation of previously discovered knowledge or discovery of new knowledge are not the scope of this research work.

3.2 The Lymphoma data warehouse architecture
The clinical data warehouse development adopted a three-layer “Data – System – Infrastructure” architecture [7].

3.2.1 The data layer architecture
The detailed star schema in Figure 1 below illustrates the data layer architecture of the clinical data warehouse.

Figure 1: Lymphoma clinical data warehouse schema

3.2.2 The system layer architecture
The development of the Lymphoma clinical data warehouse involves the following data warehousing processes:

i. Data Extract, Transform and Load (ETL) process
Web search engines, such as Google, Yahoo etc. and blogs, such as Lymphoma survivor’s network, Lymphoma patient blogs etc, are used to source and extract both fact and dimension data.

PL/SQL procedural language provided by the MySQL database system is used to transform and load the data to a staging database for data cleansing process.

ii. Data Cleansing process
PL/SQL procedural language provided by the MySQL database system is used to query data with issue and to cleanse the data, as well as to convert the data to formats that required for data reporting and analysis.
iii. Data Migration process
PL/SQL procedural language provided by the MySQL database system is used to migrate the cleansed and pre-processing data from the staging database to the actual data warehouse.

3.2.3 The infrastructure layer architecture
The clinical data warehouse is built using Open source relational database management system MySQL community server 5.0. The main platform for this clinical data warehouse project is a Dell XPS machine with the following specifications:
- Intel duo core 2.10GHz processor
- 3 Gigabytes of memory
- 130 Gigabytes of hard disk space
- Windows Vista operating system with Service pack 2

4 Evaluation Result

4.1 Evaluation of data characteristics
The developed Lymphoma data warehouse is evaluated against the 4 data characteristics below:

i. Reliability (Subject-oriented)
In this clinical data warehouse research work, the Lymphoma or Lymph node cancer is selected as the subject matter. All data that are related to the diagnosis and treatment recommendation decision making process for Lymphoma disease, such as the Lymphoma types, phases, stages, treatments as well as patients’ details, are stored in the data warehouse. This subject orientation promotes an easy-to-understand data presentation format for doctors and other health professionals.

ii. Integrity (Integrated)
To maintain a minimum level of integrity in the Lymphoma data warehouse, the following techniques are applied:
- Use of cross reference tables to represent multiple instances of the same information with a single instance in the database. For instance, instead of storing symptoms in a text format, a Symptom Dimension table is maintained as the reference table. Other cross reference tables are Disease Stage, Disease Type, and Treatment etc.
- Auto-generate and store primary keys in Integer, such as patient_id, disease_id, stage_id, to avoid human error.

iii. Capability (Time-variant)
To facilitate both current and historical perspective measurements, a “record-datetime” field is added to tag the patient’s medical data (i.e. the fact data table). Furthermore, from the technical perspective, the volume capacity of MySQL as a Clinical data warehouse is huge. According to MySQL official website [8], theoretically, MySQL database can support up to 8 million terabytes (TB) data. In practical, the database performance is limited by the underlying operating system, in normal circumstances, database with a size <=100 GBs may function smoothly, but exceeding that amount could lead to tradeoffs in performance.

iv. Manageability (Non-volatile)
Data in the Lymphoma data warehouse are stable, they are appended to the clinical data warehouse without deletion. There is no restriction to the insertion of data, and all primary keys are automatically created using the AUTO_INCREMENT value provided by the MySQL database.

From the technical point of view, MySQL provides several user-friendly tools, such as MySQL query browser, MySQL administrator, to enable data warehouse administrator to manage the data warehouse easily.

4.2 Evaluation of data operations
The developed data warehousing techniques are assessed from the 4 data warehouse operational perspectives below:

i. Data extraction performance (Subject-oriented)
Data are manually searched and extracted from the Internet using web search engines, such as Google, Yahoo etc. This manual data extraction approach is time consuming. A lot effort is given to identify relevant, vital and valuable information as well as to filter out incomplete or erroneous information within the mass of unstructured data from Internet.

To enhance the data extraction operation, an advanced data searching tool should be employed to automatically explore and locate relevant information from various sources. Some sophisticated data searching tools provide data transformation and loading functions.
ii. Data cleansing performance (Integrated)
MySQL query language is used as a tool to cleanse and transform the extracted raw data to ensure data quality and integrity, before they are loaded onto the data warehouse.

Extensive SQL commands and PL/SQL scripts are used throughout the data cleansing process, i.e. Data issue analysis and identification, data refinement, and data verification, and signification results are achieved. The rectified or refined data are verified to be complete, reliable, relevant, valid, noise-free, consistent and most importantly they are in an appropriate form for data mining.

PL/SQL stored procedures are developed to cleanse the medical data and transform them to the appropriate format for data mining, however, the stored procedures need to be manually triggered in the MySQL query browser, and this is prone to human errors. Scheduled and automated cleansing procedures are required to support an actual data warehouse operation.

iii. Data loading performance (Time-variant)
There are 2 data loading processes in this data warehouse project:
   a. Loading data from its source, i.e. Internet, to staging database.
   b. Loading data from staging database to production data warehouse after the data cleansing process.

   In both processes, data are manually converted to text files, verified and then loaded to the destination. The process is executed without any automated data loading technique for the following reasons:
   a. Data are kept in one source, i.e. Excel spreadsheet, rather from various sources. This has made the data loading process pretty straightforward.
   b. MySQL’s LOAD DATA INFILE command requires data to be uploaded in tab or space delimited text file format, which is just a simple one-click “save as” function in Ms Excel.
   c. MySQL’s LOAD DATA INFILE command requires the source file location to be identified in the SQL statement. An automated process is not viable here.

iv. Data querying performance (Non-volatile)
In order to discover new knowledge from data, data mining tools or algorithms generate substantial queries to the data warehouse to produce vast data for data analysis and mining. Inefficient processing of these queries will inevitably have significant impact on the performance of data mining. The data querying performance is evaluated from the aspects below:

a. Data response time
Data querying response time often depends on the (i) size of data, (ii) complexity of query and (iii) specification of hardware. The larger the data processed or the more complex the queries, the slower the response time.

In this clinical data warehouse project, a satisfactory response time (<0.1 sec) is obtained for a conventional query to retrieve data from a 500MB data warehouse resided on a machine with single processor and 512MB memory.

Materialised projection views, indexing and MPSV are some of the capabilities to improve query processing performance and data reliability. In this clinical data warehouse project, as show in Figure 2 B-tree indexes are used for (a) all primary keys and (b) columns frequently queried in a WHERE clause to enable MySQL query engine to quickly determine the position of data and seek the rows efficiently.

Figure 2: Indexing used in the clinical data warehouse


August 2010

b. **Data visualisation**

By using pure MySQL query language as the data query and presentation tool, the visual interpretation of complex relationships of the multidimensional data can only be presented in a tabular format.

However, the data query techniques and drill down capability only facilitate advanced users who have basic SQL knowledge.

To extend the data query facilities to other users and to enhance data visualisation advanced data presentment and graphics tools should be considered to illustrate data relationships and provide a graphical presentation of information in histograms, pie charts and bar graphs, maps etc.

The evaluation results are summarized in table 3 below.

<table>
<thead>
<tr>
<th>Perspectives</th>
<th>Subject-oriented</th>
<th>Integrated</th>
<th>Time-variant</th>
<th>Non-volatile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
<td>Reliability</td>
<td>Integrity</td>
<td>Capability</td>
<td>Manageability</td>
</tr>
<tr>
<td>Lymphoma disease</td>
<td>Only data pertaining to Lymphoma disease are stored in the data warehouse.</td>
<td>Auto-generated primary keys and cross reference tables such as Symptoms, Disease Stage, Disease Type, and Treatment etc. are used to ensure data integrity.</td>
<td>A “record-datetime” field is used to tag the patient’s medical data to support both current and historical perspective measurement.</td>
<td>No restriction to data insertion. Primary keys are automatically created with AUTO_INCREMENT value. MySQL provides query browser and administratortool to allow users to manage and fine tune the data warehouse.</td>
</tr>
</tbody>
</table>

Table 3: Evaluation results of the clinical data warehouse

5 Conclusion and Future Works

5.1 Potential benefits of clinical data warehouse

The potential benefits of the developed data warehouse can be claimed as below.

i. **As a building block of a Clinical DSS**

The developed clinical data warehouse stores Lymphoma patients’ records and provides to help doctors and other health professionals to improve the quality of disease diagnosis and treatment decision support for Lymphoma cancer.
The clinical data warehouse is built with open source technologies (MySQL and stored procedures). It supports any OS and hardware platform and can be used as a fundamental building block of a Clinical DSS built in any proprietary or open source software.

ii. As data input to data mining tool
The developed data warehouse can be used in conjunction with data mining tools such as Integral Solution’s Clementine, Thinking Machines’ Darwin, Cognos’ Scenario, IBM’s Intelligent Miner, SAS’ Enterprise Miner, open source tools such as MIDAS, Tyson software etc.

5.2 Future Works
The research work may be continued in several areas. Below presents some of the possible areas.

i. Materializing the Lymphoma data warehouse
The developed medical data warehouse is the foundation of clinical Decision Support System. It should be extended with actual patient and medical records to support Lymphoma data mining and decision making processes. To materialize this, we can collaborate with the Clinical Research Centre (CRC) under the Ministry of Health (MOH), Malaysia to obtain the information, or to establish patient registries in this area if the data is not available.

ii. Evolving the clinical data warehouse architecture
The Clinical data warehouse schema presented in figure 1 is a framework for understanding the Lymphoma data warehousing and how the components of the Lymphoma data warehouse fit together. It provides a roadmap that can be used to design towards. The structure of the data warehouse is designed in a flexible way that it can be extended to support other diseases with changes in the Disease dimension table.

iii. Automating the data warehousing processes
The developed clinical data warehouse does not really consider production operations for extracting, transforming and loading data into the data warehouse. However, these processes are essential in order to integrate the data warehouse into any decision support system. The ETL process should be automated to accelerate the data warehousing as well as data mining processes.

In addition, to facilitate a full-fledged data warehouse, advanced data presentation tool, such as Business Intelligent reporting software and technologies should be explored and incorporated into the data warehouse to enable multidimensional data presentation and querying.

4. Refining the data warehouse infrastructure
The developed clinical data warehouse was installed and running with default settings in MySQL. To run the data warehouse in full operation mode as well as to gain incremental performance improvements, some of the global parameters in the database, for instance, key_buffer_size, table_cache, query_cache_size innodb_log_file_size, innodb_buffer_pool_size etc, need to be reconfigured when the amount of indexes, data size and workload are getting larger.

References:
[7] Development of a Data Warehouse for Lymphoma Cancer Diagnosis and Treatment Decision Support - 10th WSEAS Int. Conf. on MATHEMATICS AND COMPUTERS IN BIOLOGY AND CHEMISTRY (MCBC'09) – Teh Ying Wah, Ong Suan Sim – March 2009.