A FRAMEWORK FOR DATA QUALITY ANALYSIS IN DATA WAREHOUSE SYSTEM ARCHITECTURE

AZWA BIN ABDUL AZIZ

MASTER OF SCIENCE

FACULTY OF SCIENCE AND TECHNOLOGY
UNIVERSITI MALAYSIA TERENGGANU

2010

A FRAMEWORK FOR DATA QUALITY ANALYSIS IN DATA WAREHOUSE SYSTEM ARCHITECTURE

AZWA BIN ABDUL AZIZ

MASTER OF SCIENCE
FACULTY OF SCIENCE AND TECHNOLOGY
UNIVERSITI MALAYSIA TERENGGANU
2010

DEDICATIONS

" To my love ones...
My Lovely Wife, Nurul &
sweetest Daughter, echa ...

To the important people in my life...
Abah, Mak dan Keluarga...
I appreciate all the love,
understanding and support from
each one of you...

In Memory of
Mohd Amali Darwisy... "

Abstract of thesis presented to the Senate of University Malaysia Terengganu in fulfillment of the requirement for the degree of Master of Science

A FRAMEWORK FOR DATA QUALITY ANALYSIS IN DATA WAREHOUSE SYSTEM ARCHITECTURE

AZWA BIN ABDUL AZIZ

October 2010

Chairperson	:	**Professor Md. Yazid Mohd Saman, Ph.D.**
Member	:	**Associate Professor Mohd. Pouzi Hamzah, Ph.D**
Faculty	:	**Science and Technology**

Information provided by any applications systems in organization is vital in order to obtain a decision. Due to this factor, the quality of data provided by Data Warehouse (DW) is really important for organization to produce the best solution for their company to move forwards. Many people see DW as a solution to collect information from various sources. It performs forecasting and data analysis based on information available in DW. They believe DW has every important data that have been extracted from source systems. However, the quality of data plays major roles to determine whether their decisions are good or vice versa.

Most of the DW developers are more concentrated on designing DW and how to integrate and extract from source systems to DW. Designing data mart which includes facts tables and dimensions are crucial in a DW practice. Extraction, Transform, Load (ETL) jobs are the major processes in DW developments. It easily consumes 70% of the resources needed

for implementation and maintenance of typical DW. Therefore, developers tend to focus on these two activities rather Data Quality (DQ) process. Those processes are vital in DW development. Thus, leaving DQ activities may lead to failure in DW projects and to avoid "Garbage In Garbage Out "(GIGO) phenomenon.

Therefore this study emphasizes how importance DQ process in DW projects. A framework for DQ process is developed as a guidelines for DW applications developments. The framework include a set of phases to perform systematic DQ activities for improve the quality of data in DW. The framework introduces three type of analysis: Metadata Analysis, Base Analysis and Gap Analysis. Metadata Analysis is the process of trying to understand what data should be (target values) by analyzing both business and technical metadata store in metadata repository. In contrast, Base Analysis is a process to gather what data look likes (actual values). Finally, Gap Analysis is a set of strategy to reduce a gap between both values. A prototype known as Data Quality Analysis System (DQAS) is developed using PHP to support Base Analysis. To prove the framework, the implementation of the DW end to end processes is shown. Open Source System (OSS) technology that integrated with DQAS is made. Talend Open Studio (TOS) is use as ETL tools and Business Intelligence Reporting Tools (BIRT) as BI tools in this research.

The proposed framework and the DQAS application has been demonstrated to selected organizations (Universiti Malaysia Terengganu & Universiti Sultan Zainal Abidin). It is to test the prototype usability and to gain feedback from IT personnel. The evaluation is conducted by providing questionnaire to respondents. The research questions have been

divided into seven factors. They are "Understandable of Framework", "System Functionality", "DQAS Interface", "Commercial Values", "User Interest", "Technology Used in DQAS" and "OSS DW Interest". The highest score in the *Positive* statement for the factor "OSS DW Interest" is 4.00 which mean the respondents are interested in applying DW applications using open source technology. The lowest score is for "Technology Used" factor. The mean score 3.0 shows respondents not very familiar with client-server architecture and PHP programming language used to develop DQAS. As conclusion, a result has shows by applying DQ framework in an organizations, it will improve the quality of data. It avoiding data anomalies such as inconsistent data contain in the system applications.

Abstrak tesis yang dikemukakan kepada Senat Universiti Malaysia Terengganu sebagai memenuhi keperluan untuk Ijazah Master Sains

RANGKA KERJA UNTUK ANALISIS DATA KUALITI DI DALAM PEMBANGUNAN APLIKASI GUDANG DATA

AZWA BIN ABDUL AZIZ

Oktober 2010

Pengerusi : **Prof. Md. Yazid Mohd Saman, Ph.D.**

Ahli : **Associate Professor Dr. Mohd. Pouzi Hamzah, Ph.D**

Fakulti : **Sains dan Teknologi**

Maklumat yang diperoleh daripada setiap sistem di sesebuah organisasi memainkan peranan penting di dalam proses membuat keputusan. Oleh kerana faktor tersebut, kualiti yang terdapat di dalam sesebuah Gudang Data (GD) amat penting untuk setiap organisasi membuat keputusan dalam meningkatkan pencapaian organisasi. Ramai pengguna melihat GD sebagai satu penyelesaian di dalam proses mengambil maklumat daripada pelbagai sumber sistem dan digunakan untuk menganalisa data. Mereka percaya bahawa gudang gata mempunyai segala maklumat penting yang tersimpan daripada setiap sistem aplikasi. Walau bagaimanapun, Data Kualiti (DK) memainkan peranan yang utama di dalam keputusan yang diambil, samada baik atau sebaliknya. DK meningkatkan prospek kejayaan dalam sesuatu keputusan yang diambil berdasarkan data yang ada di dalam GD.

Kebanyakan pembangun GD lebih menumpukan langkah bagaimana untuk merekabentuk GD dan bagaimana untuk mengekstrak data daripada sistem sumber. Merekabentuk *Data*

Mart yang mengandungi jadual fakta dan jadual dimensi amat penting di dalam projek GD. Seterusnya, membina proses kerja "Extract, Transform, Load (ETL)" memainkan peranan utama di dalam pembangunan GD. ETL merangkumi 70 peratus proses kerja perlaksanaan GD. Disebabkan kepentingan kedua-dua proses tersebut, pembangun GD lebih menumpukannya dari membincangkan perlaksanaan kerja DK. Tidak dinafikan, proses ETL dan rekabentuk *Data Mart* amat penting di dalam pembangunan GD. Meskipun begitu, tanpa proses kerja DK yang berkesan, kemungkinan besar ianya akan mengakibatkan projek GD mengalami kegagalan. DK amat penting dilaksanakan untuk mengelakkan fenomena "Sampah Masuk, Sampah Keluar".

Oleh kerana itu, kajian ini menekankan kepentingan proses DK dalam projek GD. Satu kerangka kerja yang sistematik untuk proses DK telah dicadangkan. Kerangka kerja tersebut akan menjadi panduan untuk setiap projek GD. Ianya mengandungi fasa penuh aktiviti DK. Antara aktiviti penting termasuklah memperkenalkan tiga jenis analisis utama iaitu Analisis Metadata, Analisis Asas dan Analisis Jurang. Satu prototaip dibangunkan untuk melaksanakan Analisis Asas yang dikenali sebagai "Data Quality Analysis System (DQAS)". Selain itu, beberapa proses dari awal dan akhir pembangunan GD ditunjukkan menggunakan teknologi sumber terbuka untuk membuktikan keberkesanan cadangan kerangka kerja. Ia termasuklah perlaksanaan ETL proses menggunakan aplikasi "Talend Open Studio (TOS)" dan aplikasi "Business Intelligence Reporting Tools (BIRT)" untuk memaparkan laporan.

Kerangka kerja yang dicadangkan dan DQAS telah dibentangkan kepada organisasi terpilih (Universiti Malaysia Terengganu & Universiti Sultan Zainal Abidin) untuk mendapatkan maklum balas mengenai kesesuaian DQAS bagi dilaksanakan di organisasi mereka. Maklum balas ini dilakukan dengan penerangan dan demonstrasi mengenai DQAS. Kemudian, satu set soalan berkenaan kebolehgunaan sistem diedarkan kepada responden untuk dijawab. Soalan kajian dibahagikan kepada beberapa faktor iaitu "kefahaman tentang kerangka kerja", "kesesuaian fungsi DQAS", "antaramuka DQAS", "nilai komersil", "minat pengguna","teknologi yang digunakan DQAS" dan "minat pengguna pada sistem sumber terbuka". Hasil kajian, faktor "pandangan pengguna pada sistem sumber terbuka" mendapat markah tertinggi. Ini membuktikan penguna berminat mengaplikasi teknologi sumber terbuka di dalam organisasi mereka. Faktor teknologi yang digunakan DQAS" mendapat markah paling rendah yang membuktikan pengguna mempunyai kurang pengetahuan dalam teknologi yang dicadangkan. Kesimpulannya, perlaksanaan kerangka kerja DK dapat meningkatkan kualiti data dalam sesebuah organisasi dan mengelakkan masalah data berlaku.