Using Metadata Analysis and Base Analysis Techniques in DQ Framework for DW

[1]Azwa Abdul Aziz, Md Yazid Mohd Saman, Mohd Pouzi Hamzah

[1]Faculty of Science and Technology,
Universiti Malaysia Terengganu,
21030, Kuala Terengganu, Terengganu MALAYSIA

ABSTRACT

Data Qualities (DQ) issues become a major concern over decade. Even with an enhancements and introduction of new technologies, it is stills not good if systems are lack of qualities of data. Data warehouses (DW) are complex systems that have to deliver highly-aggregated, high quality data from heterogeneous sources to decision makers. It involves a lot of integration of sources system to support business operations. This paper propose a framework for implementing DQ in DW systems architecture using Metadata Analysis Technique and Base Analysis Techniques to perform comparison between target value and current value gain from the systems. A prototype using PHP is develops to support Base Analysis Techniques. This paper also emphasize on dimension need to be consider to perform DQ processes.

Keywords
Data Warehouse (DW), Data Quality (DQ), Metadata, DQ Dimension

1.0 INTRODUCTION

The qualities of data contain in the Enterprise Information Systems have a significant impact and crucial to the decision maker. Many researchers have proved a lack of DQ may results failures in respective subject areas. As example, researchers have conducted the DQ impact in healthcare area, information system, IT management and many more. (Sadiq, Zhou & Maria, 2007) have stated that the hurricane protection in New Orleans failed because inadequate and incomplete using outdated elevation data. It has proved that the DQ problems may not just lead a lost of billion dollars in business transactions but can has a loss of hundred human life due to the decision made using a poor data.

As the scale in diversity of data grows in information system, the complexities of data grow multifold with it. The last several years has introduce many of new technologies and tool to support the business process – grid systems, ETL applications, semantic web namely as a few. However, this technology can utilize and successes if a data resides in those of systems are qualities one. It is synonym with a sport car using a water to move its. The well known principles of "Garbage In Garbage Out (GIGO)" indicates that regardless how much intelligence and science is dedicated in new software solutions, data quality is still a major factor in the successful operation of IT systems.

Information systems within and between organizations are often highly distributed and heterogeneous. DW is the process of taking data from legacy and transaction database system and transforming it into organized information in a user friendly format to encourage data analysis and support fact-based business decision making (Kimball & Casserta, 2004). DW involves a lot of integration process of many databases into one large database. The DQ problems often been ignored in the process of data warehouse construction and utilization (Chen & Weng, 2009). Many DW projects are discontinued due to the insufficient DQ (English, 1999). It is not doubt that the successful of DW project depend to the qualities of data provided. This paper proposes a DQ framework in DW architecture.

2.0 DEFINITION OF DATA QUALITY AND ITS MEASUREMENT

Some researcher refer DQ as Information Quality (IQ) (Madnick, Lee & Zhu, 2009)(Diane et al., 1997)(Wang, 1998)(Lee et. al, 2002). There are tendency to use DQ to technical issues of DQ while IQ related to nontechnical issues. However in this paper the term DQ refer to both technical and nontechnical issues of DQ.

The definition of DQ has subjective definitions across various fields. The simple definition of DQ is a data are "fit to use" (Wang, 1998). Some defines DQ as an accuracy of data or the freshness of data. Most of the DQ researcher and practitioner agree that DQ consist of several dimension to measure its. In early 1990's MIT propose Total Data Quality Management (TQDM) framework for measuring DQ. This becomes a pioneer work in DQ research which later most of the researcher using this framework to solve DQ problems (Madnick et al.). TQDM has classified the DQ dimension into 4 main categories which is Intrinsic, Accessibility, Contextual and Representational (Wang, 1998). Each category

contains a several DQ dimension. As example accuracy and believability dimension is under intrinsic categories. However the basic sets of DQ dimension are accuracy, completeness, consistency and timeliness, which constitute the focus of the majority author (Batini, Cappielo, Francalanci & Maurino, 2009).

The dimensions propose in the framework of this paper is emphasize on correctness (syntactic and semantic accuracy), consistency, completeness and timeliness (currency and volatility) of data. Those dimensions are important criteria to ensure the qualities of data in DW.

3.0 A DQ FRAMEWORK FOR DW

Data Warehouse (DW) application systems are normally used to help top management of an organization to make a better decision for the company so as to survive in the competitive market. Business Intelligence applications have become a favourite tool that helps decision makers. The DQ problem often been ignored in the process of DW construction and utilization (Chen & Weng, 2009). Assessing ad improving DQ is a still complex task to do, especially in modern organizations where data are ubiquitous and diverse (Batini et. al., 2009). This framework is develops to improve the DQ in DW.

Figure 1.0 shows the propose framework. Basically the framework is following steps taken in TDQM method which are Define, Measure, Analyze and Improve. However the framework will use Metadata Analysis to gain the target qualities value and Base Analysis Techniques to view actual values in data sources. A gap analysis technique will provide the strategies to reduce the gap between the target and actual values. This paper also proposes a DQ matrix strategy in DW design.

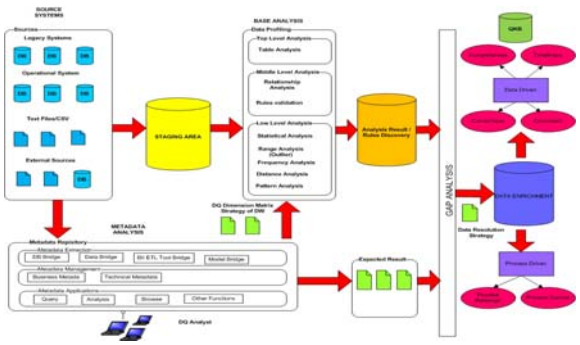

Figure 1: DQ Framework

3.1 Defines Source Systems

Data sources refer to data that exist in legacies systems, OLTP (Online Transaction Processing) systems or an unstructured and semi structured data stores in files or external sources. OLTP record the daily transactions into operational databases. Data from sources will be extracted to the staging area. Staging area is a temporary area to perform cleansing, matching and transformation process. This is a standard procedure to be performs in any DW projects.

3.2 Analysis Phase

Analysis Phase is the critical process to ensure only qualities data available in DW. DQ researchers have proposed different idea how to perform data analysis. ETL tools available in market also provide several features to perform DQ assessment. (Chen & Weng, 2009) have proposed the analysis will be performs after the data are available in DW. A set of algorithm is created to verify integrity assessment and accuracy assessment. (Yu et. al., 2009) are suggested to clean the unwanted data base on user model framework.

In this paper the analysis phase is applies before a data are extract to DW. The analysis phase is divides into three type of analysis which are Metadata Analysis, Base Analysis and Gap Analysis. Metadata Analysis is the process of trying to understand what the data should be (target values) by analyzing both business metadata and technical metadata stores in metadata repositories. In contrast, the Base Analysis process is to gather what data look likes (actual values). Then a set of strategies to reduce the gap between these two values is gain from Gap Analysis.

3.2.1 Analysis Phase: Metadata Analysis

The most simplistic definition of metadata is “data about data”. Metadata is all physical data and knowledge containing information about the business and technical processes, and data used by corporation (Marco, 2004). Metadata Management Environment (MME) needs to be created in order to manage metadata efficiently. MME represents the architectural components, people and processes that are required to properly and systematically gather retain and disseminate metadata throughout the enterprise (Santoso & Kartika, 2006).

Metadata repository propose in this paper is summarize of what be done by (Chen & Weng, 2009). It contains three layers which is metadata extraction, metadata management and metadata applications. Metadata extraction acts as bridge to provide a driver connected to any Database Management System (DBMS) platform, text files, ETL tools or data modeler tool. Metadata management stores the technical (user report, technical structure data mapping and transformation, etc) and business metadata (business definition, subject area, etc). Metadata application provides interface and functions

for DQ Analyst to access metadata repository and perform metadata analysis.

DQ analyst tries to understand the flow and process of data both from technical and business perspective without looking on how the actual data been stored. The target values will be derived from Metadata Analysis. There are only four critical DQ dimensions propose in this framework. Accuracy dimension refer to the data values stored in the database correspond to real world values (Ballaou & Pazer, 1985). Syntactic accuracy defines the closeness of value *v* to the elements of corresponding definition domain, D (Batini et. al., 2009). Semantic accuracy is focus on same value which has different meaning. Completeness often related to missing values and null values. Consistency refers to the violation of semantic rules defined over a set of data items (Batini et. al., 2009). Timeliness dimension divide into two categories which is currency and volatility. Currency factor captures the gap between the extractions of data from the sources and it delivery to users (Mokrane, 2004). Volatility is how frequent data been updated in sources system. DQ framework proposes a DQ dimension matrix strategies for DW design. As example, a figure 2 shows a star schema design for product information.

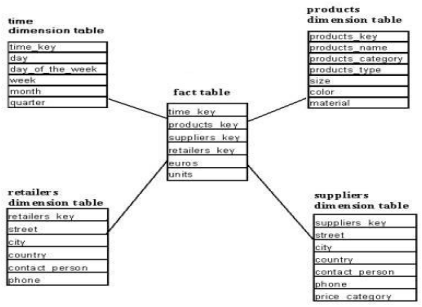

Figure 2: DW Star Schema Design

Metadata Analysis provides a DQ strategy on design above as shown in Table 1.

Table 1: DQ Dimension Matrix Strategy

Table Name: Fact Tables

Attributes	Syntactic Acc	Semantic Acc	Complete	Consistent	Currency	Volatility
Time Key						
Product Key						
Supplier Key						
Retailer Key						
Euros	/				/	/
Unit	/				/	/

Table Name: Supplier Dimension Tables

Attributes	Syntactic Acc	Semantic Acc	Complete	Consistent	Currency	Volatility
Supplier Key	/		/			
Street	/	/		/		
City		/		/		
Country	/			/		
Contact Per.						
Phone	/			/		/

This will become a guideline during Base Analysis to verify the DQ dimension for selected attributes. Fact table contains Euros and Unit as fact unit. The most important information needed to verifies the quality of fact unit is whether the values are in the ranges (outlier analysis) or how frequency the values are changes (volatility). Attributes in dimension tables such street and contact person need to be a single meaning (semantic) and consistent. (Loshin, 2009) has proposed the prioritization matrix to provide clarity for deciding relative importance, getting agreement on priorities matrix and determining the best measure to handle the problems.

3.2.2 Analysis Phase: Base Analysis

Base Analysis also can be referring as data analysis. This is the process to determine the actual values store in the sources system using data profiling techniques. It will use a data store in staging area to perform data profiling. It will take a lot of time and cost if the analysis is done by scanning every single item in the tables. Therefore, a DQ Dimension Matrix Strategies is important document to ensure analysis is carrying efficiently. The researcher has develops a prototype known as Data Quality Analysis System (DQAS) to support Base Analysis using PHP programming and Oracle databases (Azwa & Yazid, 2009).

Generally, Base Analysis consists of three main analyses. It is a Top-Level Analysis (TLA), Middle-Level Analysis (MLA) and Low-Level Analysis (LLA). TLA investigating and verifying the upper view of data model such as table definition. MLA is checking the relationship between tables and the validation rules provide by sources system. LLA relates to the values of attributes profiling.

a) Top-Level Analysis (TLA)

TLA will verify every tables or files which has been extract into staging area. This to ensure the tables contain in sources system are consistent in metadata technical definitions.

b) Middle-Level Analysis (MLA)

Two major sub analyses will be carry out in MLA are relationship analysis and rules validation analysis. Relationship analysis objective is to ensure the integrity rules of tables are not been violated. As example a foreign key exist in child table is exist in parent table. In contrast, rules validation ensures the logic of relationship between tables. (Mokrane, 2004) has given an example which stated the number of rows of a table depends on the number of rows of another table e.g., the number of accounts is equal or bigger than the number of customers.

c) Low-Level Analysis (LLA)

LLA related to the value analysis of an attributes. Some major process to be performs in LLA is statistical analysis, range analysis (outlier), frequency analysis, distances analysis and pattern analysis. Statistical Analysis provides mathematical information of data such as minimum/maximum values, mean, median, mode and standard deviation. This is to ensure the validity of the data and normally use against the metric in the tables. Table 2 shows an example of statistical analysis result using metric known as price.

Table 2: Statistical Analysis

Table Name: Account

Attributes Names	Prices
Data Type	Double
Min Value	-32990
Max Value	898989
Mean	114234.1973
Median	4848499.5

Range Analysis determines whether the values contain in an attributes is within the range. Some refer as outlier analysis. As example a value of price items need to have maximum and minimum number. Age values are not appropriate if having a value which is more than 130 years. Frequency Analysis is the process to calculate percentage of value been enter into the system. This will helps to solve the semantic problems. However, it still needs human intervention to determine which values are correct as been proposed by. Instead, the correct value also can be store in Quality Knowledge Base (QKB) which acts as a synonym dictionary. The same concept of QKB has been mention by (Feng et. al., 2008) using Unified Traditional Chinese Medical Language System (UTCMLS) to tackle data inconsistency. Table 3 shows example of the result from frequency analysis using attributes city.

Table 3: Frequency Analysis

Attributes Names	City	Per. Freq
	Kuala Terengganu	70%
	K.T	10%
	Kuala Trg	9%
	K. Trg	8%
	Kla Trenganu	3%

Distance Analysis is has been discussed by (Sadiq et. al., 2007) to overcome the problems of consistency and completeness. This framework proposes the correct values are store in QKB. Then a comparison will be applied to the value in store in sources system. The difference between this two value (e.g ahmad , ahmd has distance values 1) will be verify. A set of algorithm is created to state that the number of distance value can be accepts as a same data. Distance Analysis is quick similar with a Hamming Code techniques in data communication to detect data error during transmission. The distance function can be successfully resolve the typo mismatch but there is and evident risk of matching two different term which are similar in spelling, e.g,"Irin" and "Iris" should not be match as they may be intentionally). Here, the framework proposes a subject expert intervention to solve the problems arises.

Pattern Analysis is used to determine if the data values in a field are in the expected format. It will helps in developing the set of business rules for standardization. As an example, consider a pattern report for telephone numbers. There are many formats has been enter by users, but the valid formats should consist of three sets of numbers (three numbers for area code, three numbers for exchange, and four numbers for station). The examples of formats enter by user are:

- 9999999999 - (999) 999-9999 - 999-999-9999
- 999-999-AAAA - 999-999-Aaaa - 99AA99999

(9- represents any digit, A- represents any upper case alpha (letter) character, a - represents any lower case alpha character.)

3.2.3 Analysis Phase: Gap Analysis

Gap Analysis is verifying the comparison between target value gather from metadata analysis against actual value store in the sources system which is derived from base analysis. A set of improvement strategies is develop to closer the gap or ensure actual value same as target value. It can divide into data gap and rules gap analysis. Example of result from data gap analysis and rules gap analysis is shows in table 4 and 5 respectively.

Table 4: Data Gap Analysis

Table Name: Student Table

Attributes	Target Value	Actual Value	Difference	Improve Strategy
State	Kuala Lumpur	K.L, K. Lumpur, Kuala Lmpr	3 difference value gather from frequency report or distance report	Change two incorrect value to correct value

Table 3: Rules Gap Analysis

Table Name: Customer Table and Account Table

Target Rules	Actual Value	Difference	Improve Strategy
Number of Account Table is should be less or equal than number of Customer Table	Number of Account is greater than number of Customer	Account has content 100 record more compare to the Customer gather from MLA	Find the PK exist in Account tables not have FK in Customer Table. Apply integrity constraint in sources system

3.3 Improvement Phase: Data Enrichment

Data enrichment is the process to improve data and eliminate the fuzzy data exists in sources system. The improvements of the quality in the DW help to guarantee the reliability and help the enterprise to summarize and use the data for long period of time (Zhang, Jie, Zhang & Hua, 2009). (Donald & Tayi, 2009) have explores the factors that should be considered, such as current level of DQ, the level of DQ, the level of quality needed by the relevant decision process, and the potential benefit of project designed to enhance DQ. The framework proposed classified the data enrichment process into two main categories which is data-driven and process-driven.

3.3.1 Improvement Phase: Data-Driven

Data-driven strategies improve a quality of data by directly modifying the value of data while process-driven strategies improve quality by redesigning the process that create or modify data. Data driven accentuate on how to improve four main dimensions that have been analyzed.

The first dimension is the accuracy with contains syntactic accuracy and semantic accuracy. Syntactic accuracy refer to the correctness of data such as redundant primary key refer to same objects, inconsistency attributes for similar records, relationship problem and human typo errors at sources system. Standardization techniques and transformation data to correct value is two steps can be taken to handles this problems. In additions, redesign the entity relationship diagram (ERD) at sources system another measure can be taken to improve the quality of data. Of course the costs taken, the impacts of error and the priority of the improvement process needs to be study first.

A semantic problem refers to the same value which has different meaning. For handle this type of anomalies, subject expert need to add a new value that can be differentiate between those data. As example a two placed known as a “Taman Sentosa “exists in Kuala Terengganu area. This placed can be identifies if “Taman Sentosa “combines with postcode that make the value unique. Subject expert also can provide taxonomies as guidelines to recognize semantic values.

Inconsistency problems can be handle develops a complete dictionary that store the value expected such as Quality Knowledge Base (QKB). An algorithm can be develops to detect inconsistent value as mention before. Those values can transform into correct value by using Extract, Transform, and Load (ETL) value available in market. A null value, missing values and uncompleted records can be solved once again by interferences from subject experts. They need to provide correct values or default value to transform the data.

Improving ETL jobs and schedule will helps to handle timeliness dimension problems. ETL developers can design a job the captures the information needed for currency dimension such as a time taken to extract data from sources systems to DW. Volatility problem which deal with data that rapidly changes can be improved by increase the time to trigger ETL jobs. ETL tools provides by vendors has a capabilities to handle complex jobs with helps from powerful infra architecture and technologies. Now there are demand to provide report from DW which are not monthly basis or weekly basis, but a daily reports or even hour reports.

3.3.2 Improvement Phase: Process-Driven

Process-Driven proposes the improvement by changing the process to creating or modifying data at the sources system. It also can be refers as Business Process Reengineering (BPR). BPR is best method to handle DQ problem, however it will take a lot of cost and time to perform the activities. The worse case, redesign the sources system may lead to the failure of existing systems. Process- Driven techniques should be avoided in improving DQ in DW systems.

4.0 CONCLUSION

This paper has discussed a framework for DQ measurement in DW architecture. To prove a concept in the framework, a prototype is develops using PHP known as DQAS. DQAS can use to perform Base Analysis techniques to any Oracle databases. Then, DQAS is integrates with Talend Open Studio (TOS) ETL and Eclipse Business Intelligence Reporting Tools (BIRT) to implement the DQ framework. Both TOS and BIRT are open sources tools. A schema provided by Oracle known as HR schema is selected as data sources. Then a multidimensional model is created in DW schema base on HR schema. A mapping document between sources to target table is prepared. Finally the differences applying the DQ framework and without using it is shown.

ACKNOWLEDGEMENT

A fellowship of University Darul Iman Malaysia (UDM) to carry out this research is fully acknowledged.

REFERENCES

Sadiq, S., Zhou, X., & Maria, O. (2007). *Data Quality – The Key Success For Data Driven Engineering,* In Proceeding of the International Conference on Network and Parallel Computing – Workshops.

Kimball, R.,Caserta, J. (2004). *The Data Warehouse ETL Toolkit*, Wiley and Sons.

Chen, B., Wang, B. (2009). Analysis and Solution of Data Quality in Data Warehouse of Chinese Materia Medica, In Proceeding of the *International Conference on Computer Science & Education.*

English, L. (1999). *Improving Data Warehouse and Business Information Quality. Wiley*, New York et al.

Madnick, S.,E., Lee, Y., W., & Zhu, H. (2009). Overview and Framework for Data and Information Quality Research, *ACM Journal of Data and Information Quality*, *1*(1), Article 2.

Diane, M., S., Lee, Y., W., & Wang, R., W. (1997). *10 Potholes in the Road to Information Quality*,Cybersquare.

Wang, R., Y., (1998). *A product perspective on total data quality management. Comm. ACM 41, 2.*

Lee Y.W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). *AIMQ: A methodology for information quality assessment*. Inform. Manage. 40, 2, 133–460.

Batini, C., Cappielo, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assesment and Improvement, *ACM Computing Surveys*, 41(3), Article 16.

Cartaci, T., & Scannapieco, M. (2002). *Data Quality under Computer Science Perspective*. Archive Computer 2.

Batini, C., Cabitza, F., Cappiello, C., & Francalanci, C. (2006). *A Comprehensive Data Quality for Web and Strurured Data*, University Degli Studi di Milano Bococca.

Yu, H., Xiao-yi, Zhang. Zhen, Y., & Guao-quan, J. (2009). A Universal Data Cleaning Framework Based on User Model, In Proceeding of the *International Colloquium on Computing, Communication, Control and Management* (ISECS).

Marco, D. (2004). *Metadata & Knowledge Management: Managed Metadata Environment*, DM Review, http://www.dmreview.com.

Leo Willyanto Santoso, & Kartika Gunadi. (2006). *A proposal of DQ for data warehouses environment,* Petra Christian University.

Ballaou D., & Pazer, H. (1985). *Modeling data and process quality in multi-input, multi-output information system*, Management. Sci. 31, 2.

Mokrane, B. (2004). *A Framework for Analysis Data Freshness.* In Proceedings of the 2004 international Workshop on information Quality in information Systems. IQIS '04 University de Versailles: ACM.

Feng, Y., Wu, Z., Chen, H., Yu, T., Mao, Y., & Jiang, X. (2008). *Data Quality in Traditional Chinese Medicine*. Paper presented at the Proceedings of the 2008 International Conference on BioMedical Engineering and Informatics - Volume 01.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Manage. Sci., 35*(8), 982-1003. doi: http://dx.doi.org/10.1287/mnsc.35.8.982.

Azwa., A., & Md Yazid, S. (2009). *Data Warehouse System Overview and Measuring Data Quality*, In Proceeding of the International Conferences Software Engineering and Computer System (ICSECS 09).

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. (2004). The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. *Inf. Syst., 29*(7), 551-582. doi: http://dx.doi.org/10.1016/j.is.2003.12.004.

Zhang, J., Wen, Q., & Zhang, H. (2009). *The research in improving the quality of DW data: the job-scheduling and checking based program in upgrading DW performance*. Paper presented at the Proceedings of the 5th International Conference on Wireless communications, networking and mobile computing, Beijing, China.

Ballou, D. P., & Tayi, G. K. (1999). Enhancing data quality in data warehouse environments. *Commun. ACM, 42*(1), 73-78. doi: http://doi.acm.org/10.1145/291469.291471.

Loshin, D. (2009). *Data Quality Remediation*, Information System Management.