A PROPOSAL OF DATA QUALITY FOR DATA WAREHOUSES ENVIRONMENT

Leo Willyanto Santoso, Kartika Gunadi
Faculty of Industrial Technology, Informatics Department, Petra Christian University
e-mail : leow@petra.ac.id, kgunadi@petra.ac.id

ABSTRACT: The quality of the data provided is critical to the success of data warehousing initiatives. There is strong evidence that many organisations have significant data quality problems, and that these have substantial social and economic impacts. This paper describes a study which explores modeling of the dynamic parts of the data warehouse. This metamodel enables data warehouse management, design and evolution based on a high level conceptual perspective, which can be linked to the actual structural and physical aspects of the data warehouse architecture. Moreover, this metamodel is capable of modeling complex activities, their interrelationships, the relationship of activities with data sources and execution details.

Keywords: data quality, metamodel, data warehouse.

INTRODUCTION

Many organisations are currently developing data warehouses in order to reduce the costs associated with the provision of data, to support a focus on complete business processes, and to achieve high estimated returns on investment. A key factor in the success of data warehousing initiatives is the quality of the data provided [5, 11]. It is essential therefore that data quality is understood and that data quality assurance procedures are developed and implemented. Whilst many organisations are aware of the importance of data quality for their ability to compete successfully in the market place, research and industry surveys indicate that organizations are increasingly experiencing problems with data quality [14] and that these have substantial economic and social impacts [8]. There has been a lack of methods and frameworks for measuring, evaluating, and improving data quality [14], however, and little discussion of the management, economic or organisational aspects of data quality [12].

A number of researchers have identified various data quality dimensions. However, these dimensions are often overlapping, vaguely defined and not soundly based in theory [9]. Some frameworks have been developed which organise important concepts for defining and understanding data quality [14, 9], and support methodical approaches to improving data quality processes within organisations [11].

This paper describes a study which explores modeling of the dynamic parts of the data warehouse. This metamodel enables data warehouse management, design and evolution based on a high level conceptual perspective, which can be linked to the actual structural and physical aspects of the data warehouse architecture. Moreover, this metamodel is capable of modeling complex activities, their interrelationships, the relationship of activities with data sources and execution details. Organisations are often aware of data quality problems. However, their improvement efforts generally focus narrowly on only the *accuracy* of data, and ignore the many other data quality attributes and dimensions that are important [13].

The paper first defines data quality and reviews existing research in data quality. The next section describes the research approach adopted in this study. The general framework for the treatment of data warehouse metadata in a metadata repository is then described. The framework requires the classification of metadata in at least two instantiation layers and three perspectives. Implications of the case study findings for data quality practice are discussed, and the paper concludes with some suggestions for future research.

DATA WAREHOUSE

A Data Warehouse (DW) is a collection of technologies aimed at enabling the knowledge worker (executive, manager, analyst, etc) to make better and faster decisions. Many researchers and practitioners share the understanding that data warehouse architecture can be formally understood as layers of materialized views on top of each other. Data warehouse architecture exhibits various layers of data in which data from one layer are derived from data of the lower layer. Data sources, also called operational databases, form the lowest layer. They may consist of structured data stored in open database systems and legacy systems, or unstructured or semi-structured

data stored in files. The central layer of the architecture is the global (or primary) Data Warehouse. The global data warehouse keeps a historical record of data that result from the transformation, integration, and aggregation of detailed data found in the data sources. Usually, a data store of volatile, low granularity data is used for the integration of data from the various sources: it is called Operational Data Store (ODS). The Operational Data Store, serves also as a buffer for data transformation and cleaning so that the data warehouse is populated with clean and homogeneous data. The next layer of views is the local, or client warehouses, which contain highly aggregated data, directly derived from the global warehouse. There are various kinds of local warehouses, such as the data marts or the OLAP databases, which may use relational database systems or specific multidimensional data structures.

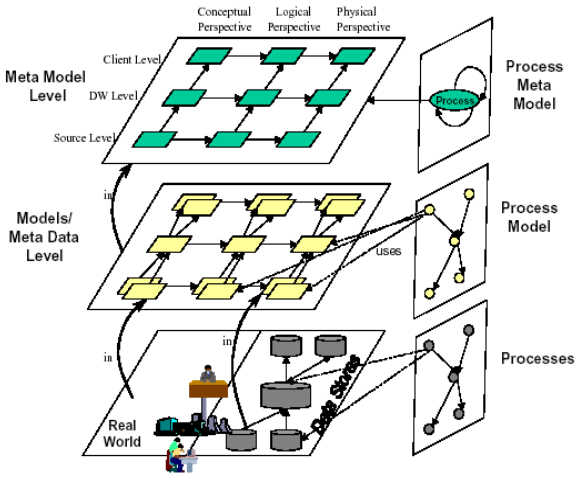

Figure 1. Data Warehouse Metadata Framework

All the data warehouse components, processes and data are -or at least should be- tracked and administered from a metadata repository. The metadata repository serves as an aid both to the administrator and the designer of a data warehouse. Indeed, the data warehouse is a very complex system, the volume of recorded data is vast and the processes employed for its extraction, transformation, cleansing, storage and aggregation are numerous, sensitive to changes and time varying. The metamodel is reproduced in Figure 1.

The metadata repository serves as a roadmap that provides a trace of all design choices and a history of changes performed on its architecture and components. For example, the new version of the Microsoft Repository [2] and the Metadata Interchange Specification (MDIS) [6] provide different models and application programming interfaces to control and manage metadata for OLAP databases. In Figure 2, a architecture for a data warehouse is depicted.

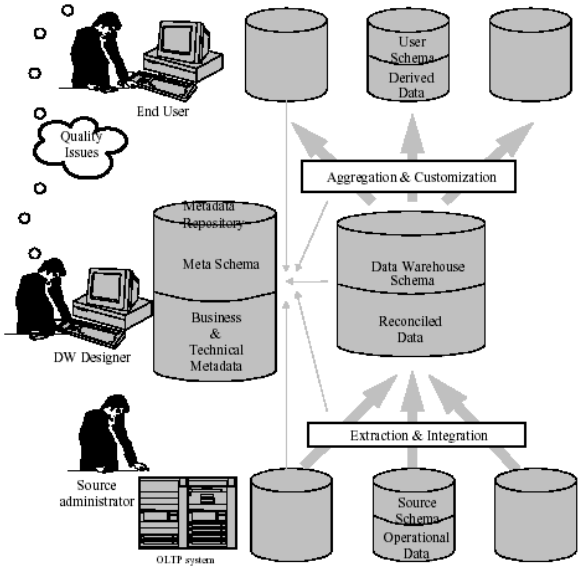

Figure 2. Architecture for a Data Warehouse

Data warehouses have proved their value to serve as repositories for integrated, homogenized and clean data. In other words, they do not serve only as information buffers for answering complex questions quickly but also as intermediate stages in the processing of information within the information system of an enterprise, where the information becomes more accurate and useful. Thus, at the end of the data processing chain, at the front end of an OLAP/ DW application, is ultimately the overall quality of the information which is provided to the end user.

QUALITY AND DATA WAREHOUSES

Data quality has been defined as the fraction of performance over expectancy, or as the loss imparted to society from the time a product is shipped [1]. We believe, though, that the best definition is the one found in [10, 7, 13]: data quality is defined as "fitness for use". The nature of this definition directly implies that the concept of data quality is relative. For example, data semantics is different for each distinct user.

As a decision support information system, a data warehouse must provide high level quality of data and quality of service. Coherency, freshness, accuracy, accessibility, availability and performance are among the quality features required by the end users of the data warehouse. Still, too many stakeholders are involved in the lifecycle of the data warehouse; all of them seem to have their quality requirements. As

already mentioned, the Decision Maker usually employs an OLAP query tool to get answers interesting to him. A decision-maker is usually concerned with the quality of the stored data, their timeliness and the ease of querying them through the OLAP tools. The Data Warehouse Administrator needs facilities such as error reporting, metadata accessibility and knowledge of the timeliness of the data, in order to detect changes and reasons for them, or problems in the stored information. The Data Warehouse Designer needs to measure the quality of the schemata of the data warehouse environment (both existing and newly produced) and the quality of the metadata as well. Furthermore, he needs software evaluation standards to test the software packages he considers purchasing. The Programmers of Data Warehouse Components can make good use of software implementation standards in order to accomplish and evaluate their work. Metadata reporting can also facilitate their job, since they can avoid mistakes related to schema information. Based on this analysis, we can safely argue that different roles imply a different collection of quality aspects, which should be ideally treated in a consistent and meaningful way.

From the previous it follows that, on one hand, the quality of data is of highly subjective nature and should ideally be treated differently for each user. At the same time, the quality goals of the involved stakeholders are highly diverse in nature. They can be neither assessed nor achieved directly but require complex measurement, prediction, and design techniques, often in the form of an interactive process. On the other hand, the reasons for data deficiencies, non-availability or reach ability problems are definitely objective, and depend mostly on the information system definition and implementation.

Furthermore, the prediction of data quality for each user must be based on objective quality factors that are computed and compared to users' expectations. As the number of users and the complexity of data warehouse systems do not permit to reach total quality for every user, another question is how to prioritize these requirements in order to satisfy them with respect to their importance. This problem is typically illustrated by the physical design of the data warehouse where the problem is to find a set of materialized views that optimize user requests response time and the global data warehouse maintenance cost at the same time. The interpretability of the data and the processes of the data warehouse are heavily dependent on the design process and the expressive power of the models and the languages which are used. Both the data and the systems architecture are part of the interpretability dimension. Furthermore, query optimization is related to the accessibility dimension, since the sooner the queries are answered, the higher the transaction availability is. The extraction of data from the sources is also influencing the availability of the data warehouse. Consequently, one of the primary goals of the update propagation policy should be to achieve high availability of the data warehouse.

ARCHITECTURE AND QUALITY META MODELS

The framework requires the classification of metadata in at least two instantiation layers and three perspectives. The metamodel layer constitutes the schema of the metadata repository and the metadata layer the actual meta-information for a particular data warehouse. We linked this framework to a well-defined approach for the architecture of the data warehouse [4]. Then, we presented our proposal for a quality metamodel, which builds on the widely accepted Goal-Question-Metric approach for the quality management of information systems. The exploitation of the quality model can be performed in versatile ways. It is important to note that as far as the lifecycle of the data warehouse is concerned, this usage can be done in a dual fashion. Notice, that not only do we provide an initial specialization of the quality metamodel for common data warehouse processes, but the data warehouse stakeholder can further detail this provision with his own templates for the quality management of his specific data ware-house, in a similar fashion. Secondly, the use of the Concept Base metadata repository can be exploited, due to its querying facilities. Following [3] we give a small example of a query upon the metadata repository.

Figure 3. Design and Administration Quality Factors

The following two queries detect objects with trace quality problems.

```
QualityQueryLeo Range isA Integer
with parameter
    q: EstimatedMeasurement
constraint
    c: $ exists n/Interval, u/Integer, l/Integer
        (n upper u) and (n lower l) and (q hasValue n) and
        (this le u) and (this ge l) $
end

QualityQueryLeo ObjectsWithQualityProblems isA
DW_Object
with constraint
    c: $ exists q1 ActualMeasurement ,
    q2/EstimatedMeasurement,
        m/Metric, t/Timestamp
        (m actual q1) and (m expescted q2) and
        (q1 timestamp t) and (q2 timestamp t) and
        not(q1 in range[q2]) and
        (q1 for this)$
end
```

Third, the quality metamodel is coherent with the generic metadata framework for data warehouses. Thus, every new data warehouse object can be linked to metrics and measurements for its quality, without any change to the schema of the repository.

DATA WAREHOUSE PROCESSES

Metamodel enables data warehouse management, design and evolution based on a high level conceptual perspective, which can be linked to the actual structural and physical aspects of the data warehouse architecture.

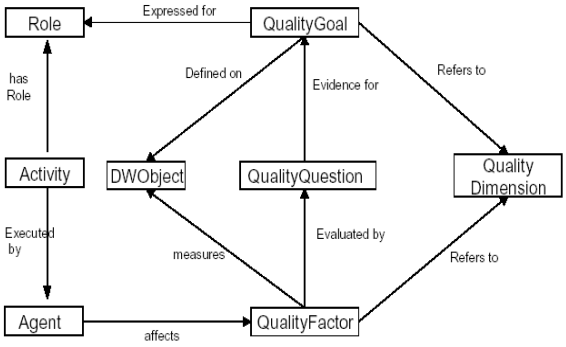

Figure 4. Relationships Between Processes and Quality

The proposed metamodel is capable of modeling complex activities, their interrelationships, the relationship of activities with data sources and execution details. This is shown in Figure 4.

Finally, the metamodel complements proposed architecture and quality models in a coherent fashion, resulting in a full framework for data warehouse metamodeling.

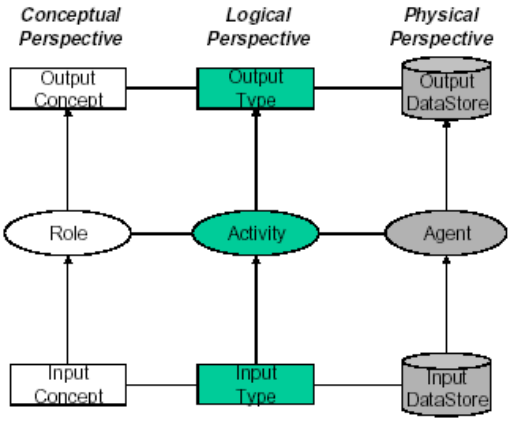

Figure 5. Three Perspectives of The Process Metamodel

Yet, there are also design processes in such an environment, which do not seem to fit this model so smoothly. We used the global-as-view approach for the data warehouse definition, i.e., we reduce the definition of the data warehouse materialized views to the data sources.

TESTING OF DATA WAREHOUSE REPOSITORY

Following the approach of previous work [3, 4], we store semantically rich meta-information of a data warehouse in a metadata repository concerning the conceptual, logical and physical perspective of the data warehouse. In addition, the information on the quality of the stored objects is recorded in this repository.

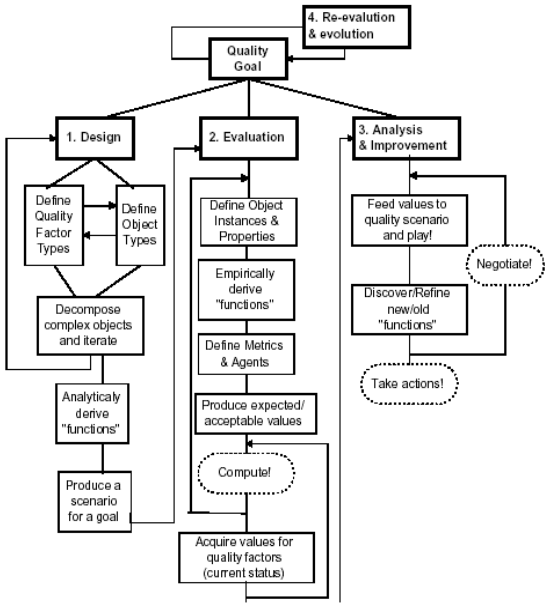

Figure 6. Methodology for Data Warehouse Quality Management

Our approach extends GQM, based on the idea that a goal is operationally defined over a set of questions. Thus, we provide specific "questions" for the full lifecycle of a goal: this way the data warehouse metadata repository is not simply defined statically, but it can be actually exploited in a systematic manner. These questions are expressed as a set of steps aiming, in one hand, to map a high-level subjective quality goal into the measurement of a set of interrelated quality factors, and, in the other hand, to propose improvement actions which may help in achieving the target quality goal.

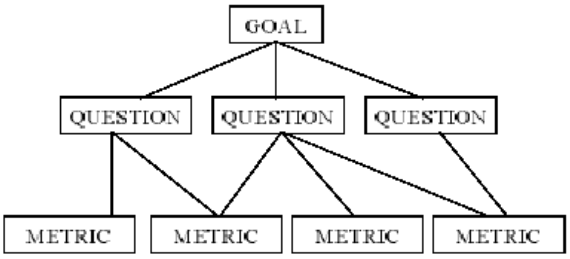

Figure 7. Goal Question and Metrics

The benefit from the use of the methodology is not only the obtained solution to a specific problem. Maybe of greater importance is the fact that the involved stakeholder gets a more clear view of the data warehouse interdependencies. This is achieved through the systematic application of the methodological steps, which convert a subjective problem, expressed in a high-level vocabulary, to specific measurable factors that affect the solution to the problem.

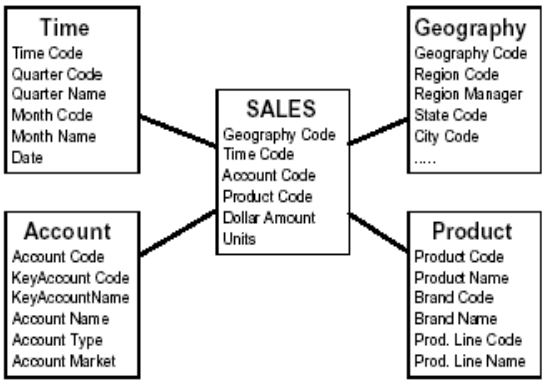

Figure 8. Star Schema of the Model

We have verified our methodology in a set of case studies. We believe that the full application of the methodology in a wider extent in the future will provide the academic community with the insight for further tuning.

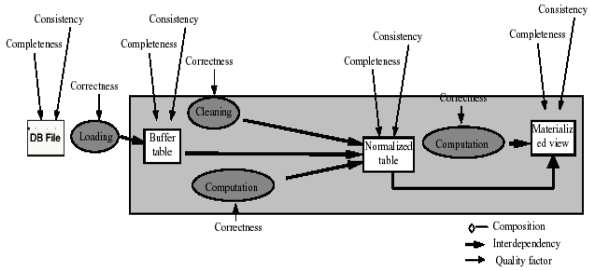

Figure 9. The Scenario of The Case Study

CONCLUSION AND FUTURE WORKS

This paper presented a set of results towards the effective modeling and management of data warehouse metadata with special treatment to data warehouse quality. The first major result that we presented was a general framework for the treatment of data warehouse metadata in a metadata repository. The framework requires the classification of metadata in at least two instantiation layers and three perspectives. The metamodel layer constitutes the schema of the metadata repository and the metadata layer the actual meta-information for a particular data warehouse.

Research can continue in several ways using our results. As far as the process metamodel is concerned, we have dealt only with the operational processes of a data warehouse environment. Yet, there are also design processes in such an environment, which do not seem to fit this model so smoothly. It is in our future plans to investigate the modeling of design processes and to capture the trace of evolution in a data warehouse.

REFERENCES

1. Besterfield, D. H., Besterfield-Michna, C., Besterfield, G. and Besterfield-Sacre, M.. *Total Quality Management*. Prentice Hall, 1995.

2. Bernstein, P.A., Bergstraesser, Th., Carlson, J., Pal, S., Sanders, P. and Shutt, D., *Microsoft Repository Version 2 and the Open Information Model. Information Systems*, vol. 24, no. 2, 1999.

3. Jeusfeld, M.A., Quix, C., Jarke. M., *Design and Analysis of Quality Information for Data Warehouses*. Proceedings of the 17th International Conference on Conceptual Modeling (ER'98), Singapore, 1998.

4. Jarke, M., Jeusfeld, M.A., Quix, C., Vassiliadis, P., Architecture and quality in data warehouses: An extended repository approach. *Information Systems*, 24(3), 1999. pp. 229-253,

5. Kimball, R., *The Data Warehouse Toolkit*. John Wiley and Sons, 1996.

6. Coalition, M., Metadata Interchange Specification (MDIS v.1.1). http://www.metadata.org/ standards/toc.html 1997.

7. Orr, K., Data quality and systems theory. *Communications of the ACM*, 41(2), 1998. pp. 66-71.

8. Strong, D.M., Lee, Y.W. and *Wang,* R.Y., *Beyond Accuracy: How Organizations are Redefining Data Quality" (No. TDQM-94-07*), Cambridge Mass. Total Data Quality Management Research Program, MIT Sloan School of Management, 1994.

9. Svanks, M.I., "Integrity Analysis: Methods for automating data quality assurance". *EDP Auditors Foundation*, vol. 30, no. 10, 1984.

10. Tayi, G.K., Ballou, D.P., Examining Data Quality. *In Communications of the ACM*, 41(2), 1998. pp. 54-57

11. Wang, R.Y., A product perspective on total data quality management. *Communications of the ACM (CACM),* vol. 41, no. 2, February 1998.

12. Wang, R.Y., Storey, V.C., Firth, C.P. *A Framework for Analysis of Data Quality Research. IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, No. 4, August 1995.

13. Wang, R.Y., Strong, D., Guarascio, L.M., *Beyond Accuracy: What Data Quality Means to Data Consumers. Technical Report TDQM-94-10*, Total Data Quality Management Research Program, MIT Sloan School of Management, Cambridge, Mass., 1994.

14. Wand, Y. and Wang, R.Y., Anchoring data quality dimensions ontological foundations. *Communications of the ACM (CACM),* vol. 39, no. 11, November 1996.