Modeling Data Warehouse using Quality Metrics: The Need of Software Process

Naveen Dahiya
Assistant Prof.,CSE Dept.
MSIT, C-4, Janakpuri
New Delhi

Vishal Bhatnagar, Ph.D
Associate Prof.,CSE Dept.
AIACT & R, Geeta Colony
Delhi

Manjit Singh, Ph.D
AssociateProf.,CSE Dept.
YMCAUST, Sector-6
Faridabad, Haryana.

ABSTRACT
Data warehouses play a powerful role in decision making in the organizations. Data warehouse provides most accurate and relevant information to improve strategic decisions making process. There exist several approaches for data warehouse design and their quality assurance to help designers choose among alternative schemas that are semantically equivalent. This paper focuses on the quality of the conceptual models of the data warehouses. The process of metrics creation is explained followed by validation of metrics along with a discussion on previously proposed metrics for data warehouse conceptual models.

Keywords*:* Conceptual Models, Metrics, Validation

1. INTRODUCTION
Data warehouse is an information delivery system that enables knowledge executives, managers and analysts to make better and fast decisions. Data warehouses provides managers with the most accurate and relevant information to improve strategic decisions. According to ISO 9126 (ISO 2001), quality can be defined as the extent to which a product satisfies stated and implied needs when used under specified conditions. Information quality of the data warehouse comprises of the data warehouse system quality and data presentation quality as shown in fig.1. Data warehouse quality can be influenced by data base management system quality, data quality and data model quality. Therefore, data warehouse model quality depends on the type of data model be it conceptual, logical or physical model.

The paper focuses on the quality of the conceptual models. The quality of data warehouse is measured in terms of quality metrics. Aim of these metrics is to help designers choose the best among alternative schemas that are semantically equivalent. Creation of the metrics is not only obtaining a valid set of metrics but also includes validation as well.

Section1 gives a brief introduction of data warehouse model quality. Section 2 and 3 discusses the importance of quality metrics and metrics proposed by various researchers. Section 4 aims at defining the complete process of metrics creation followed by validation of metrics. Section 5 provides future implication of the research work followed by section 6 that concludes the paper.

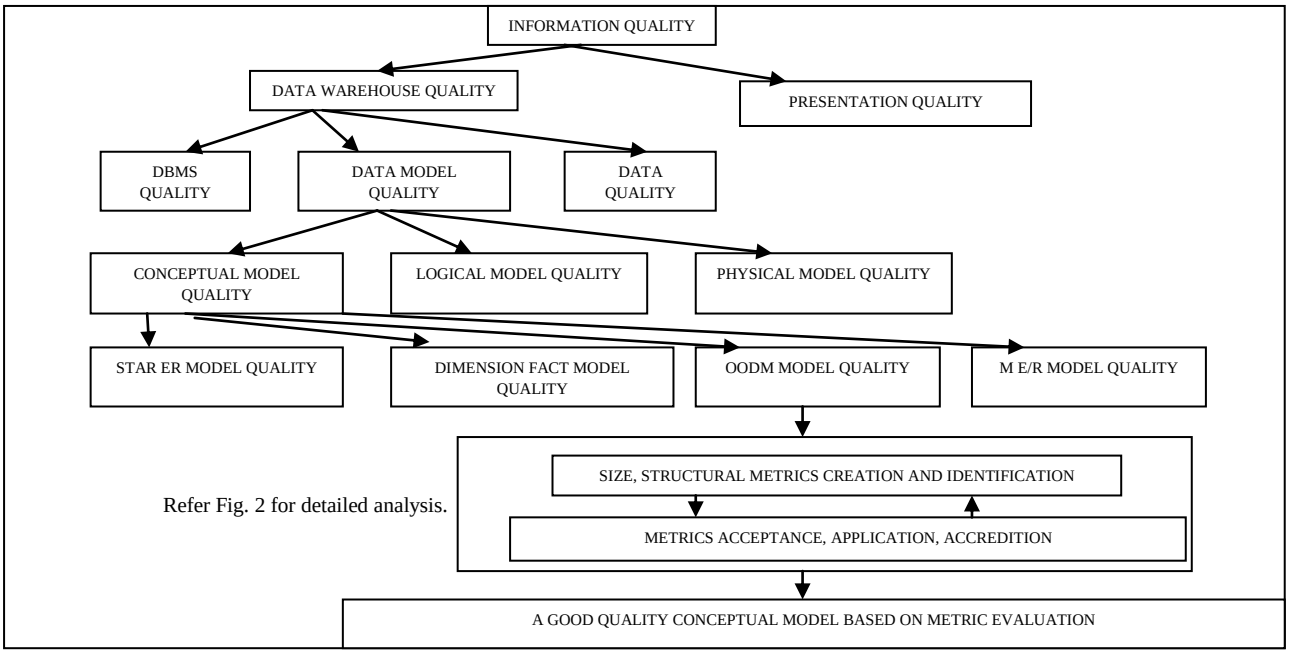

Fig. 1 Information and Data Warehouse Quality ref [1]

2. RELATED RESEARCH AND MOTIVATION

Several researchers have proposed valid set of metrics to measure the quality of data warehouse models. Since the conceptual phase is the first step in design of data warehouse model and it affects the quality of successive logical or physical phases, therefore our focus in the paper is on valid set of metrics proposed for conceptual models of data warehouses.

Gray, et al. [17] proposed a common standard for data model quality. The author proposes two metrics namely Entity Relationship (ER) metric. The proposed metrics are evaluated using entities, their attributes and the relationships existing between respective entities.

Similarly Kesh, et al. [18], described the methodology for aggregation of scores on various metrics to calculate an overall quality score for an E-R (entity relationship) model. The author proposes a method based on first calculating the scores of the individual ontological components, both structure components and content components. The behavioral scores are calculated based on scores of ontological components. Finally behavioral scores are added to calculate score of model quality.

Moody, et al. [19] proposed a comprehensive set of metrics for evaluating the quality of an E-R diagram. A total of twenty nine candidate metrics are proposed each of which measures one of the quality factors namely correctness,

Completeness, integrity, flexibility, understandability, simplicity, integration, integrity and implement ability.

Genero, et al. [20] proposed that the complexity of an E-R diagram could be highly influenced by the different elements such as entities, attributes, relationships, generalizations [37]. Several metrics for measurement of structural complexity of E-R diagrams are proposed. Broadly the metrics are classified into entity metrics, attribute metrics and relationship metrics which are then further sub classified accordingly.

Calero, et al. [3], focused on dimensional data model quality. Various metrics that can be applied at table, star and schema level are discussed. The author proposes 2 table metrics, 8 star metrics and 6 schema metrics.

Serrano et al. proposed [1], 8 quality metrics to assure the quality of the conceptual schema used in the early stages of a DW design. The author signifies that the proposed metrics helps in measurement of understandability and the efficiency of conceptual schemas [22].

Sahraoui et al. [10] defined a set of 4 structural metrics for assuring data warehouse quality. The aim of these metrics is to help designers choose among alternative schemas that are semantically equivalent.

The metric validation has been carried out in order to prove the practical utility of proposed metrics.

3. NEED OF QUALITY METRICS

Data warehouses are increasingly being used in the organizations to gain competitive advantage in the current market scenario as they enable strategic decision making along with analysis of past trends and prediction of future trends. Therefore, it becomes important for the organization to incorporate quality information for efficient data warehouse design from the early conceptual phases.

The information quality is influenced by various factors such as presentation quality and data warehouse quality as shown in fig.1. The information quality of data warehouse conceptual model is measured using metrics. The researchers have proposed quality metrics for multidimensional conceptual models. These metrics needs to be theoretically as well as empirically validated in order to prove their practical utility. Validation of metrics proves that the proposed metrics play an important role in measurement of quality of conceptual model of data warehouse. Thus, the metrics are significant in defining data warehouse quality during early phases of data warehouse development.

4. MODELING DATA WAREHOUSE

Conceptual phase is the initial phase in design of data warehouse. If a good quality conceptual model is designed in the initial phase then it can be assumed that quality of final product will also be good. Quality metrics are the indicators to judge the quality of the resulting product. There is a particular process to obtain valid and useful metrics which is described in following subsection:

4.1 Metrics Creation Process

In metrics creation process [1], there are five main phases as shown in fig. 2 starting from identification of goals to creation, application, acceptance and accreditation.

1) Identification: In this phase goals of the metrics are defined and hypotheses are formed. All the following phases will be based upon these goals and hypotheses.
2) Creation: In this phase metrics are defined and validated. This phase is divided into three sub phases:

a) Metrics definition: Metric definition is made taking into account the specific characteristics [2] of the system we wish to measure, the experience of the designers of these systems.

b) Theoretical validation: After defining the metrics, theoretical validation is applied to prove that the metric defined are correct. The formal (or theoretical) validation helps us to know when and how to apply metrics. There are two main tendencies in metrics theoretical validation [3]. Frameworks based on axiomatic approaches and framework based on measurement theory.

b.1) Framework Based on Axiomatic Approaches [4,5]

Briand's formal framework provides the framework based on axiomatic approaches. This framework provides the set of mathematical properties that characterize several important measurement concepts: size, length, complexity, cohesion and coupling [6].

b.2) Framework Based on Measurement Theory [7, 8]

In measurement theory, empirical conditions are formulated from which hypothesis of reality can be derived. There are two frameworks included in it.

b.2.i) Zuse's framework [3]:

Zuse apply a measurement-theoretic approach to complexity measures. The focus is on the conditions that should be satisfied by relational systems in order to provide them with additive ratio scale measures.

b.2.ii) DISTANCE framework [1]:

The framework is called DISTANCE as it concerns the concepts [9] of distance and dissimilarity. It consists of five steps:

- Step 1. Find a measurement abstraction
- Step 2. Model distances between measurement abstractions
- Step 3. Quantify distances between measurement abstractions
- Step 4. Find a reference abstraction
- Step 5. Define the software measure

c) Empirical validation: Empirical validation [10] is crucial for the success of any software measurement process. Basically, empirical validation can be divided into: experiments, case studies and surveys. The process is evolutionary and iterative and as a result of the feedback, the metric could be redefined or discarded depending on formal or empirical validation. A first empirical validation using basic statistical techniques was conducted and discussed in [11] and [12]. In empirical validation, the following analysis techniques may be applied:

c.1) Correlation: is used to check whether there exists a relationship between independent and dependent variables.

c.2) Regression: After correlation, the following step is to determine whether the relationship is linear or not using univariate or multivariate regression techniques.

c.3) Case Base Reasoning [13,14]: is used to find the most similar cases in order to group similar metrics and removal of redundant metrics.

c.4) Formal Concept Analysis [15]: This technique forms the subsets of the structural metrics that could be the best indicators of the understandability of the schemas.

c.5) Bayesian Classifier [16]: aims to determine the degree of participation of the metrics in the decision about understandability classification.

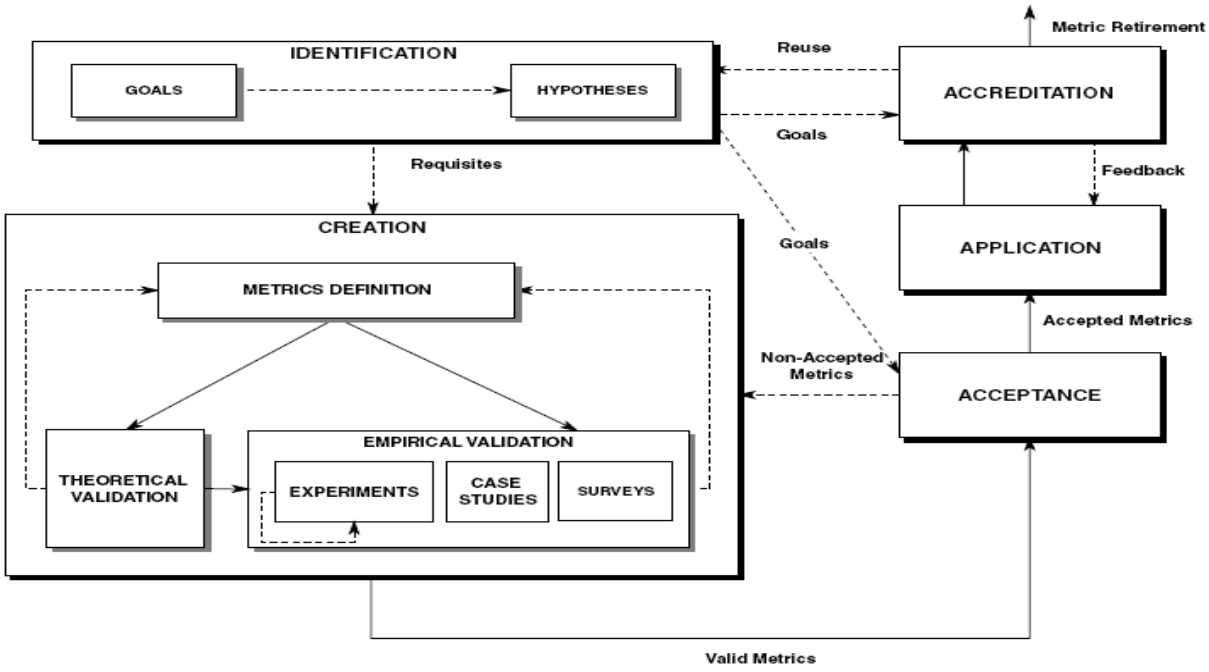

Fig. 2 Metrics Creation Process ref [1]

3) Acceptance: In this phase the metric is subjected to experimentation. The result of experimentation is used to verify system performance and defined characteristics according to stated goals/hypothesis.
4) Application: The accepted metric is applied to real users and business cases.
5) Accreditation: This is the final phase of the process. The phase proceeds together with the application phase. As a result of this phase the metric can be retired or reused for a new metric definition process.

5. RESEARCH IMPLICATION AND DISCUSSION

The Study of the paper will help the researchers to identify prospective direction of future research. It will open up the areas where efficient use of data warehousing can be made and how can a best data warehouse for a domain be developed considering the various metrics which will be created and further used for quality output.. The researchers who are already working in data warehouse domain will come to know about the latest trends and directions of future research. The in-depth study of such literature also helps to gain better understanding as what new metrics can be explored and how can it affect the future software output techniques for optimizing data warehouse development.

6. CONCLUSION

The work on identifying new metrics and how to create them in more efficient way is subject of interest. The importance of data models in software development and their influence in the final Information System quality and cost, emphasize the importance of data model quality. The paper discusses whole process of metrics creation and some recent proposals have been summarized. The authors have specified the metrics for OO data models using elements such as aggregation or other kind of relationships. Therefore the future work for conceptual data models may include defined metrics not only for measuring static diagrams like class diagrams but also metrics for dynamic diagrams, such as state diagrams and activity diagram.

7. REFERENCES

[1] Serrano, M.,Trujillo, J.,Calero, C., & Piattini, M. Metrics For Data warehouse Conceptual Models Understandability, Information and software technology, science direct,49,2007.

[2] Basili, V., Weiss, D., A Methodology For Collecting Valid Software Engineering Data, IEEE Transaction on Software Engineering, 1984.

[3] Calero C., Piattini, M., Pascual, C., & Serrano, M. Towards Data warehouse Quality Metrics, International Workshop on Design and Management of Data Warehouses (DMDW'01), 2001.

[4] Weyuker, E., Evaluating Software Complexity Measures, IEEE Transactions on Software Engineering, 1988.

[5] Briand, L., Morasca, S. & Basili, V., Property-based Software Engineering Measurement, IEEE Transcations on Software Engineering, 1996.

[6] Piattini, M., Genero, M., & Jimenez, L., Metrics Based Approach for Predicting Conceptual Data Model Maintainability, International Journal of Software Engineering and Knowledge Engineering vol:11,703-729, 2001.

[7] Whitmire, S., Object Oriented Design Measurement, Ed. Wiley, 1997.

[8] Zuse, H., A Framework of Software Measurement, Walter de Gruyter, 1998.

[9] Poels, G., Dedene, G., DISTANCE: A Framework Software Measure Construction, Research ReportDTEW9937 Dept. Applied Economics Katholieke University Leuven, Belgium, 1999.

[10] Sahraoui, H., Serrano, M., Calero, C. & Piattini, M., Empirical Studies to Assess the Understandability of Data warehouse Schemas Using Structural Metrics, Software Qual J, 2008.

[11] Serrano, M., Calero, C., & Piattini, M., Validating metrics for data warehouses. IEEE Proceedings SOFTWARE, 149(5), 161–166, 2002.

[12] Serrano, M., Calero, C., & Piattini, M., An experimental replication with warehouse metrics. International Journal of Data Warehousing & Mining, 1(4), 1–21, 2005.

[13] Grosser, D., Sahraoui, H. A., & Valtchev, P., An analogy-based approach for predicting design stability of Java classes. In International Symposium on Software Metrics,252–262, 2003.

[14] ilson, D., & Martinez, T., Improved heterogeneous distance functions, Journal of Artificial Intelligence Research, 6, 1–34,1997

[15] Godin, R., Mineau, G., Missaoui, R., St-Germain, M., & Faraj, N., Applying concept formation methods to software reuse. International Journal of Knowledge Engineering and Software Engineering, 5(1), 119–142,1995.

[16] Ramoni, M., & Sebastiani, P., Bayesian methods for intelligent data analysis. In: M. Berthold & D. J.,Hand (Eds.), An introduction to intelligent data analysis, 1999.

[17] R. Gray, B. Carey, N. McGlynn and A. Pengelly, Design metrics for database systems, BT Technology J., 9(4), 69-79, 1991.

[18] S. Kesh, Evaluating the Quality of Entity Relationship Models, Information and Software Technology, 37(12), 681-689, 1995.

[19] L. Moody, Metrics For Evaluating the Quality of Entity Relationship Models, Proceedings of the Seventeenth International Conference on Conceptual Modeling (ER ´98), 213-225, 1998.

[20] M. Genero, M. Piattini, C. Calero and M. Serrano, Assurance of Conceptual Data Model Quality Based on Early Measures, IEEE Transactions on Software Engineering, 2001.

[21] Boman, M., Bubenko, J., Johannesson, P. & Wangler, B., Conceptual Modeling, Prentice Hall, 1997.

[22] Serrano, M., Definition of a Set of Metrics for Assuring Data warehouse Quality, University of Castilla, 2004.