ISSN No. 0976-5697

Volume 4, No. 9, July-August 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Assessment of Data Warehouse Model Quality

Hunny Gaur
Student (M.Tech – Information Security)
Ambedkar Institute of Advanced Communication Technologies and Research
Geeta Colony, New Delhi, India
gaurhunny20@gmail.com

Rakhee
Student (M.Tech – Information Security)
Ambedkar Institute of Advanced Communication Technologies and Research
Geeta Colony, New Delhi, India
rakhee.sharma60@yahoo.com

Abstract: Data Warehouse (DW)is always used in making strategic decisions thus ensuring its quality is very much crucial for any organization. Data Warehouse consists of huge and unorganized data. The information in Data Warehouse is needed to make important decisions; its quality is thus a matter of concern. The Data Warehouse quality can be improved in many aspects for example it can be improved by improving the quality of information it holds, can be further improved by improving the data Warehouse design, here in this paper we have surveyed about the impact of conceptual model metrics on Data Warehouse quality. There have been so many approaches to design the Data Warehouse from the conceptual, logical, and physical perspectives. In our point of view there is lack of objective indicators to guide the designers in obtaining an outstanding model that allow us to guarantee the quality of the DW. However only M. Serrano and M. Piattini had provided a set of empirically validated metrics to help the designers. The paper summarizes the set of metrics defined for DW conceptual models and their formal and empirical validation to assure their correctness.

Keywords: Data Warehouse, Multidimensional Modeling, Conceptual Model, Metrics, UML.

I. INTRODUCTION

Data Warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making processes [1].It has become a strong aspect of any organization dealing with information technology. Many strategic decisions are made on the basis of the information hold in DW, thus its quality must be assured; lack of quality can result in disaster from both a technical and organizational point of view. Thus organization must guarantee the quality of information contained in DW. The quality of information in DW is determined by the quality of the system itself and the quality of the data presentation. Figure 1 represents the hierarchy of information quality of data warehouse. In fact, it is important not only that the data of the Data Warehouse correctly reflects the real world, but also that the data are correctly interpreted. Regarding data warehouse quality, three aspects must be considered: the quality of the DBMS (Database Management System) that supports it, the quality of the data models used in their design (conceptual, logical and physical) and the quality of the data themselves contained in the data warehouse.

In this paper, we will focus on the quality of the data models, and more concretely, on the quality of conceptual models. It is well known that empirical validation is crucial for the success of any software measurement project as it helps us to confirm and understand the implications of the measurement of the products. Many methods have been proposed by researchers for DW design, but they are not enough to assure the quality of DW. We cannot consider only subjective criteria, as different people can have different interpretation of the same concept. Thus some measurable criteria are needed to avoid arguments of style. Thus to start with, first of all metric must be defined and it is done in a methodological ways, some necessary steps must also be followed. First metrics must be defined. The definition must be made taking into account the specific characteristics of the multidimensional data model. Then it is formally or theoretically validated. Formal validation helps us to know when and how to apply metrics and at the end it is empirically validated. Here, the objective is to prove the practical utility of the proposed metrics. Empirical validation can be done in three ways, through surveys, experiments and case studies.

The remain of the paper is structured as follows: section 2 summarizes the multidimensional model for Data warehouses with their various quality indicator metrics proposed as well as their formal and empirical validation study. Sections 3 summarizes the conceptual model quality for data warehouse, based on the UML and describe the empirical validation performed with the proposed metrics. Section 4 defines the comparative study of the quality metrics proposed so far. Finally section 5 defines conclusion and proposes future work arising from this study.

II. MULTIDIMENSIONAL MODEL QUALITY

Dimensional data model is usually designed using the star schema modeling facility; which allows good response times and an easy understanding of data and metadata for both users and developers [2]. A multidimensional data model is a direct reflection of the manner in which a business process is viewed. Metrics can be used to understand and maintain the quality of the data warehouses. A first proposal of metrics for data warehouse was given in [3]. Figure 2 presents the method followed for the metrics proposal [4].

The work [3] considered only two steps related with definition and formal validation. Figure 3 shows a multidimensional data model design. The paper [3] proposed 2 metrics at table level (NA and NFK), 8 at star level (NDT, NT, NADT, NAFT, NA, NFK, RSA and RFK) and 14 at schema level (NFT, NDT, NSDT, NT, NAFT, NADT, NASDT, NA, NFK, RSDT, RT, RScA, RFK and

RSDTA). However, at schema level most of the metrics are derived metrics. Metrics formal validation was made using the framework proposed by [5]. This framework is a measurement theory based framework whose goal is to determine the scale to which a metric pertains. The framework works with three main mathematical structures namely, the extensive structure, the independence conditions and the modified relation of belief, defined in [3].

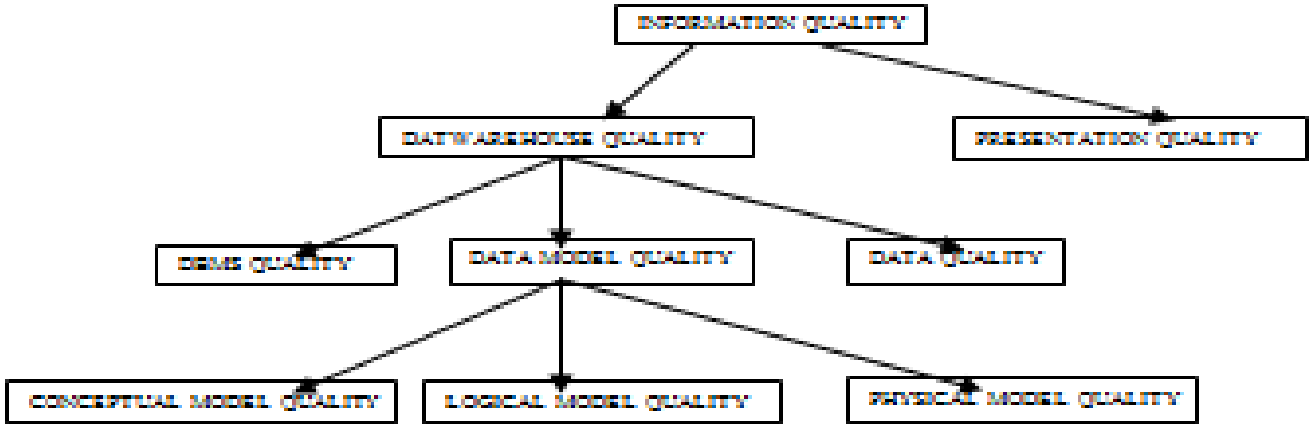

Figure 1.Data Warehouse Quality

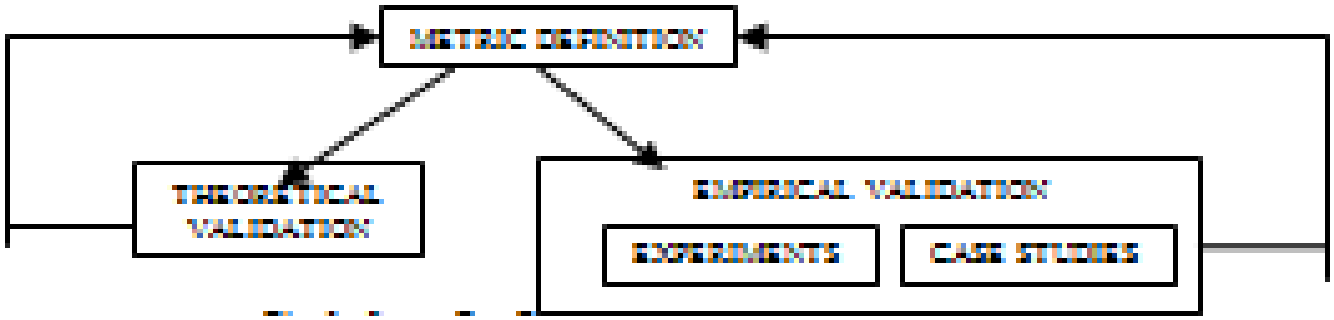

Figure 2. Steps For Definition And Validation Of Metrics

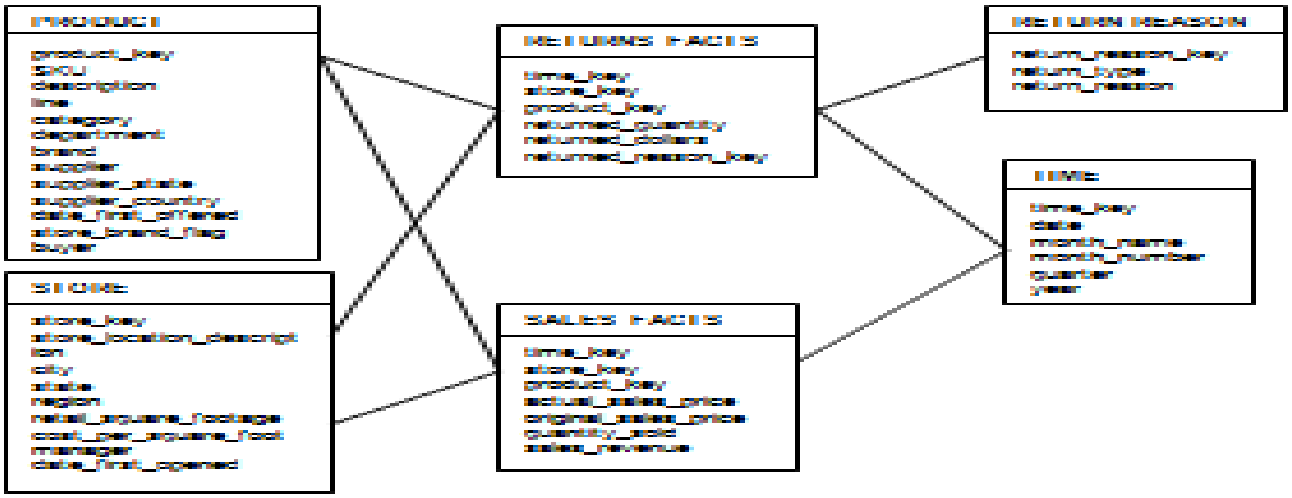

Figure 3. Multidimensional Data Model Design

The paper [3] proposed a large set of metrics for formal and empirical validation. This metrics were refined in [6] by considering only schema metrics that could be useful in order to measure the quality of a data warehouse. An empirical validation was done with the presented metrics, in order to know if they are useful as complexity mechanisms from a practical point of view. The paper also described a controlled experiment (a five step process) carried out for empirically validating the proposed metrics. Table 1 summarizes the steps involved in controlled experiment defined in [6].

Table I: Controlled Experiment Process Steps

Sr. No.	*Steps*	*Sub Steps*
1.	Definition	Experimental goal is defined
2.	Planning	Context selection, Selection of subjects, Variable selection, Instrumentation, Hypotheses formulation and Experiment design
3.	Operation	Preparation, Execution and Data validation
4.	Analysis and Interpretation	Use collected data to test the hypotheses formulated during planning
5.	Validity Evaluation	Threats to conclusion, construct, internal and external validity

Further, experimental validation of metrics for multidimensional model was conducted by [7]. Two metrics (No. of fact tables, NFT and No. of dimension tables, NDT) were presented and an experiment was developed in order to validate them as quality indicators. The process of experiment was kept same as [6] and one Null and 3 Alternative hypotheses [7] were tested. As a conclusion of the experiment, the number of fact tables seems to be a solid indicator of the complexity of multidimensional data models but the number of the dimensional tables is neither an indicator of this complexity nor can it modulate the complexity.

The metrics proposed for multidimensional data models act as objective indicators of the quality. However, they have not considered the structural complexity due to relationships among various elements present in the models. The paper [8] proposes a complexity metric which considers structural complexity present in multidimensional models. The advantage of the metric is that it is available during early phase of software development life cycle. The metric is proposed on the basis of GQM (Goal Question Metric) approach. This approach's first step is to define measurements goals according to the specific needs of an organization. Goals are refined in an operational, traceable way, into a set of quantifiable questions. Questions in turn imply as set of metrics and data for collection [8]. The complexity metric is defined considering complexity due to attributes or data variables presented in elements and complexity due to relationships among these elements. The metric is validated using a framework that gives the preliminary idea that the metric is actually measuring what is supposed to measure.

III. CONCEPTUAL MODEL QUALITY FOR DATA WAREHOUSE, BASED ON THE UML

One of the main issues that influence data warehouse quality lies on the models (conceptual, logical and physical) we use to design them. The paper [9] presents a set of metrics to measure the quality of conceptual models for DW's and validates them through an empirical experiment performed by expert designers in DW's. Figure 4 shows the object oriented data warehouse conceptual model using UML.

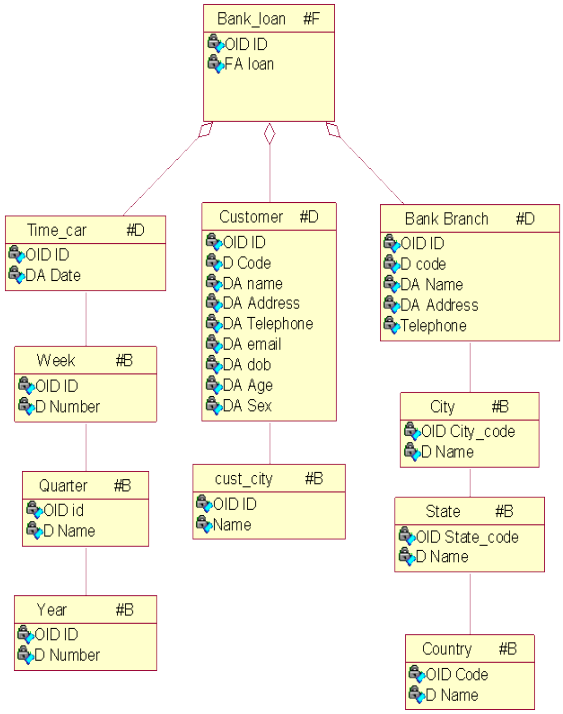

Figure 4: Object Oriented Data Warehouse Conceptual Model Using UML

Taking into account the metrics defines for data warehouse at a logical level [6] and metrics defined for UML class diagrams [10], authors proposed an initial set of metrics for data warehouse models at 3 different levels: class, star and diagram. For empirical validation a within-subject design experiment was selected. The documentation, for each design, included a data warehouse schema (as shown in fig 4) and a question/answer form that included the task to be performed. For each design, the subjects had to analyze the schema, answer some questions about the design and performed some modification on it. As a result of this experiment it seems that there exist correlation between several of the metrics and the understandability of the conceptual data warehouse models.

This empirical validation was extended in [11] by increasing number of subjects from 17 to 25. However, only star scope metrics were used for deriving the correlation between metrics and two new terms namely, effectiveness and efficiency was calculated for the experiments done by subjects. After these experiments [11] concluded that several metrics are correlated with the understandability of the models (mainly those measuring the number of elements

in the conceptual schema such as the number of classes, associations, attributes and so on) and with the efficiency of the subjects when dealing with those models (those measuring the number of classes, dimensions and the number of hierarchy levels defined in dimensions).

IV. COMPARATIVE STUDY

Various researchers have proposed a number of metrics for data warehouse star schema. Some focused only on formal and empirical validation of metrics and some focused on complexity metric. Table II summarizes the comparative study of various researches regarding the metrics and their affects.

Table 2: Comparative Study

Year	*Paper*	*No. of metrics used*	*Formal validation*	*Empirical validation*	*Results*
2001	Towards data warehouse quality metrics[3]	2(table metrics) 8(star metrics) 14(schema metrics)	Yes	No	Proposed set of metrics for data warehouse star design and provide formal validation.
2002	Validating metrics for data warehouse[6]	14 schema metrics	Yes	Yes	Showed correlation between metrics and schema complexity
2003	Experimental validation of multidimensional data models metrics[7]	2 metrics	Yes	Yes	Two metrics have been presented (no. of fact tables and no. of dimension tables) and between them no. of fact tables was considered as solid quality indicator.
2004	Empirical validation of metrics for conceptual models of data warehouse[9]	2(class metrics) 11(star metrics) 14(diagram metrics)	Yes	Yes	Showed high correlation between the metrics and schema understanding time.
2007	Metrics for data warehouse conceptual models understandability[11]	11(star metrics)	Yes	Yes	Showed high inverse relationship between metrics and efficiency. Also showed that understanding time and efficiency are related with no. of classes and hierarchy in these classes of the schema.
2011	Quality metrics for conceptual models for data warehouse focusing on dimensional hierarchies.[12]	5 metrics	No	No	Hierarchy of dimensional table affects structural complexity of model which in turn affects understandability and modifiability of the model.
2012	Complexity metric for multidimensional models for data warehouse.[8]	1 complexity metric	Yes	No	Metrics have been validated using a practical framework [13]. And identified complexity metrics on the basis of GQM paradigm.
2013	Empirical validation of structural metrics for predicting understandability of conceptual schemas for Data Warehouse[14]	11 metrics	Yes	Yes	Metrics have been validated, applied statistical and machine learning methods on the collected data to show the impact of schema metrics on its understandability.
2013	Empirical validation of metrics for object oriented multidimensional model for Data Warehouse[15]	12 metrics	No	Yes	Showed correlation between metrics and understandability, Metrics and efficiency.

V. CONCLUSIONS AND FUTURE RESEARCH

Many strategic decisions in an organization are taken on the basis of data stored in Data Warehouse. Thus assuring its quality is important matter of concern for any organization. To assure its quality one needs to guarantee the DBMS quality, data model quality and data quality itself. This survey paper deeply studies data model quality in order to improve the quality of data warehouse. Various metrics have been proposed and their formal as well as empirical validations are done to prove their practical utility, some of the proposed metrics are found to be real quality indicators. This guide the designers in obtaining an outstanding model that allow us to guarantee the quality of the DW. Our future work is to summarize the set of metrics defined for data warehouse conceptual models and will provide their formal validation to assure their correctness and then to prove their practical utility, empirical validation will also been done through conducting family of experiments on students. Another further work we will deal with is by applying data mining techniques; these techniques are used to validate the results which we will obtain after conducting series of experiments on students.

VI. REFRENCES

[1]. W.H Inmon.”Building Data Warehouse”, John Wiley & Sons.

[2]. R. Kimball, L. Reeves, M. Ross and W. Thornthwaite. “The Data Warehouse Lifecycle Toolkit”, John Wiley & Sons, 1998.

[3]. C.Calero, M. Piattini, C. Pascual and M.A Serrano.”Towards Data Warehouse Quality Metrics”, Proceedings of the international workshop on design and management of data warehouses, 2001.

[4]. C. Calero, M. Piattini and M. Genero.”Metrics For Controlling Database Complexity”, Chapter III in developing quality complex database systems: practices,

techniques and technologies, Becker(ed), Idea Group Publishing, 2001.

[5]. H. Zuse. "A Framework Of Software Measurement" Walter de gruyter, 1998.

[6]. M. Serrano, C. Calero and M. Piattini."Validating Metrics For Data Warehouses", IEE proc.-softw., vol.149, No.5, October 2002.

[7]. M.Serrano, C. Calero and M. Piattini. "Experimental Validation Of Multidimensional Data Model Metrics".proceedings of the 36th Hawaii international conference on system science, 2003.

[8]. S. Nagpal, A. Gosain and S. Sabharwal."Complexity Metric for Multidimensional Models For Data Warehouse", CUBE, 2012.

[9]. M. Serrano, C. Calero, J. Trujillo, S. L. Mora and M. Piattini."Empirical Validation Of Metrics For Conceptual Models Of Data Warehouses",CAiSE, LNCS, 2004.

[10]. M. Genero, J. Olivas, M. Piattini and F. Romero."Using Metrics to Predict OO Information Systems Maintainability", Proc.Of 13th international conference on advanced information systems engineering (CAiSE'01). Lecture Notes in Computer Science 2068, 2001.

[11]. M. Serrano, J. Trujillo, C. Calero and M. Piattini."Metrics for Data Warehouse Conceptual Models Understandability", Information and software technology 49, 2007.

[12]. A.Gosian, S.Nagpal, S.Sabharwal (2011) "Quality metrics for conceptual models for Datawarehouse focusing on dimension hierarchies" ACM SIGSOFT Software Engineering Notes, Volume 36 Number 4 DOI: 10.1145/, 2011.

[13]. C.Kaner, Software Engineering matrics: What do they measure and how? 2004. In Proceedings of the 10thInternational Software Metrics Symposium (Chicago, IL, USA, September 14-16, 2004). Metrics' 04, 1-10.

[14]. M. Kumar, A.Gosain, Y.Singh(2013) "Empirical validation of structural metrics for predicting understandability of conceptual schemas for Data Warehouse ",The society for reliability engineering, quality and operation management(SREQOM), India and The division of operation and maintenance, Lulea university of technology, Sweden.

[15]. A. Gosian, S. Mann(2013)"Empirical validation of metrics for object oriented multidimensional model for Data Warehouse"The society for reliability engineering, quality and operation management(SREQOM), India and The division of operation and maintenance, Lulea university of technology, Sweden.

