Simplified Approach for Data Warehouse Quality Measurement

Md. Ilyas Khan
Technocrats Inst. Of Technology
Anand Nagar, Piplani
Bhopal (MP)-India

Dr. R. K. Singh
ITM University
Gurgaon
Haryana-India

P.K. Dey
Jodhpur National University
Jhanwar Road, Narnadi
Jodhpur-India

ABSTRACT
Data Mining and warehousing is very challenging and interesting field in Computer Science and Information Technology. This field is still in its infant stage and thus a limited documentation on Data Warehousing processes available. As we know that now a days a number of organizations are switching from DBMS to Warehouse to manage their large volume of data .The only thing which can be deal with large, Huge Academic Data is Data Warehouse. In general Quality can be defined as Measure of excellence or state of being free from defects, deficiencies, and significant variations. ISO 8402-1986 standard defines quality as "the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs." A data warehouse is also a software product and designed to meet the requirements of the user. The user only appreciate Quality Products .So it is very much needed to device such methods which can be used to measure the Data Warehouse . For Measurement of quality of a Data Warehouse is very much needed and we must need an understanding about this Technology. We have presented a simplified approach in this paper for measurement of Warehouse..

General Terms
Date warehouse Quality parameters Measurement.

Keywords
Quality ,Data warehouse, Data Mining, Evaluation Parameters, and Measurement

1. INTRODUCTION
The concept of data warehousing dates back to the late 1980s. Data warehouse is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis. Data mining and warehousing is a field having vast potential for growth. It is advantageous to apply data mining and warehousing in educational institutions. Such institutions are feeling the heat of competition due to liberalization and globalization. To remain competitive and grow educational institutions need to exploit large amount of data lying in their repositories. As this field is experiencing changes in the environment forced by globalization, a need is felt to design a data warehouse especially suitable for educational institutions.

1.1 Data Warehouse define
Bill Inmon is universally recognized as the "father of the data warehouse.". The term Data Warehouse was coined by Bill Inmon in 1990, which he defined in the following way: "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". He defined the terms in the sentence as follows:

1.1.1 Subject-oriented
Data that gives information about a particular subject instead of about a company's ongoing operations.

1.1.2 Integrated
Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.

1.1.3 Time-variant
All data in the data warehouse is identified with a particular time period.

1.1.4 Non-volatile
Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.

Ralph Kimball provided a much simpler definition of a data warehouse. "It is a copy of transaction data specifically structured for query and analysis". This definition provides less insight and depth than Mr. Inmon's, but is no less accurate.

A database designed to support decision making in an organization. Data from the production databases are copied to the data warehouse so that queries can be performed without disturbing the performance or the stability of the production systems.

1.2 Data warehouses Structure
Data warehouses consist of several components: data sources, data warehouse, and end-user applications . Where the data warehouse component includes staging area, detail data, summarized data, data marts, and meta data. Figure shows the basic elements of a data warehouse. Brief descriptions for each data warehouse component are given below.

1.2.1 Data Source
Data source is the origin of the data in the data warehouse. One feature of data warehouses is integrating data from multiple autonomous and heterogeneous data sources. Data sources of a data warehouse could be structured (e.g., relational DB), semi-structured (e.g., XML, RDF files), or flat files . Furthermore,

warehouse data could come from either remote or local data sources. Such arbitrary make challenges to data warehouse builders for creating an uniform repository to store these multi-structure data, and designing an easy-understanding modeling language to express the schemas of data sources and data warehouse repositories, and the transformations between these schemas.

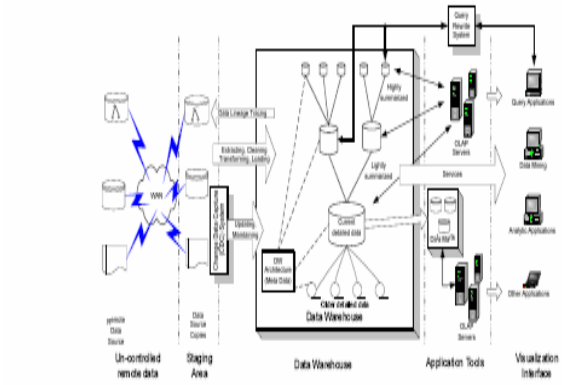
1.2.2 Data Warehouse

The actual data warehouse is the database which contains the integrated collection of data used to support strategic decision-making processes. It contains several components [5].

1.2.3 Staging Area

Staging area keep whole copies of the source data and brings them under the control of the data warehouse administrator. Naturally, staging area stores heterogeneous data and may contain "dirty" (i.e., duplicate, inconsistency) data. Data in staging area is the direct source of data in the data warehouse (the detailed data or summarized data which will be discussed below).

However, the warehouse data has uniformed structure and been cleaned (we can see below), thus data cleansing, data structure transforming, and change data capturing (CDC) processes



normally happen in this stage. In addition, no user querying service is in this area that is ender-user of the data warehouse can not access data in staging area

1.2.4 Detailed Data

Detailed data includes current detailed data and older detailed data. Far and away the current detailed data is the major concern of data warehousing. It is the exact lowest level source of the information supporting DSS processing. Normally, data here is stored in a singular, globally acceptable fashion (Such as RDB, or OODB) [5]. From the staging area to the detail data repository, data need to be extracted, cleaned, transformed, loaded and integrated. Such activities are the main processes of data warehousing. So far, current detailed data is almost always stored on disk storage, which is fast to access, but expensive and complex to manage.

1.2.5 Summarized Data

Summarized data is the data that is distilled from the low level detail data. It is divided into two levels, lightly and highly summarized data. Both of these could be treated as virtual or materialized views over the detailed data or other views. Mostly, DSS processing is based on these views. Maintaining these views, especially materialized views, is a main topic of data warehousing issues.

1.2.6 Data Marts

The issue of data marts is an important part of data warehousing research. Data warehouse users have different information requirements from different departments. Data flows from the data warehouse to various departments for their customized DSS usage. These departmental DSS databases are called data marts. A data mart is actually a body of DSS data for a department that has an architectural foundation of a data warehouse. Different data marts contain different combinations and selections of the same detailed data found at the data warehouse. Although data marts are very important parts in many data warehouses, they are not in the scope of this report.

1.2.7 Meta Data

Meta data plays a special and very important role in the data warehouse and is used as a directory of the structure of the contents of the data warehouse, and a guide to the mapping of the data which is transformed either from data source to the data warehouse, or from detailed data to summarized data in the data warehouse [5]. However, in many ways meta data sits in a different dimension than other data warehouse data, because meta data contains no data directly taken from the operational environment. In this report, my focus is that how the Auto Med approach can be used to create such meta data, and how these meta data can be used for data warehousing processes, especially data warehouse maintenance and date lineage tracing.

1.2.8 End-User Applications

End-user application is the interface used by data warehouse user to access warehouse data. It contains a series of tools, such as OLAP servers, query applications, analytic applications, and data mining tools, and so on [1]. Recently, the problem of query rewriting (a.k.a. answering queries using views) has received significant attention because of its relevance to a wide variety of data management problems: query optimization, maintenance of physical data independence, data integration and data warehouse design also extends the AutoMed approach in this area.

1.3 Operational Data vs. Informational Data

To compare data warehouse and operational databases [11], we also need to distinguish two kinds of data, operational data and informational data. Operational data is the data you use to run our business, and is typically stored in relational databases, but may be stored in legacy hierarchical or flat file formats as well.

Informational data is the data stored in data warehouse, and it's typically in a format that makes analysis much easier. Analysis can be in the form of decision support, report generation, executive information systems, and more in-depth statistical analysis. Informational data comes from operational data, after some preprocessing, like data cleaning, integrating.

1.4 Updated at the End of a Period

Data warehouses are generally batch updated at the end of the day, week or some period. Its contents are typically historical and static and may also contain numerous summaries.

1.5 Operational Data Stores

The data warehouse is structured to support a variety of analyses, including elaborate queries on large amounts of data that can require extensive searching. When databases are set up for queries on daily transactions, they are often called "operational data stores" rather than data warehouses

2. OBJECTIVE AND SCOPE

The basic objective of this work is to measure the quality of data warehousing. Scope of the study is limited to the perspective of end user from educational institution point of view.

3. DESCRIPTION OF WORK

It has been observed that there are various methods available to measure the quality of data warehouse on technical grounds. In almost all of these methods of evaluation participation of end-user is minimal. But, we feel is that, only the end user can evaluate any product most effectively and his / her opinion about the product definitely needs to be taken into consideration. Keeping this in mind it was decided to formulate the research problem in such a way so as to develop a new approach for measuring the quality of data Warehouse. To measure the quality of a data warehouse, the Data Warehouse Quality (DWQ) project, used the Goal-Question-Metric (GQM) software engineering methodology. Quality questionnaire is then formulated to relate the goals to the metrics. Metrics are defined to measure some property of each object in terms of quality. The solution methodology adopted is represented in the following algorithm.

4. PROCESS FOR EVALUATING QUALITY OF DATA WAREHOUSE

1. Identify required data warehouse software
2. Determine what is already available
3. Develop a shortlist
4. Create your evaluation criteria
5. Perform the evaluation

4.1.1 Establish what data warehouse software you need.

The core set of tools: database; extract, transform and load (ETL); and business intelligence (BI). I also strongly suggest a data modeling tool. Some companies may also need to examine data cleansing software -- but note that most of data quality is performed in the ETL code that you write.

4.1.2 Determine what you already have, or what you have standardized on database software.

Also, are you using other tools, such as BI tools, with other applications? A word of caution, though: Make sure you have the expertise to use the software you already have in-house AND that people are happy with that software.

4.1.3 Develop a shortlist

There are a few ways to get this list: look at industry analyst research reports, i.e. Forrester or Gartner; see what is bundled with software you already have such as ETL capabilities bundled with your database or BI tool; ask peers; and read articles by "independent" columnists.

4.1.4 Create your evaluation criteria

Find checklists that you could use as a start for your criteria. Try to avoid lengthy feature packed checklists, though, as you can easily lose sight of the forest for the trees. The criteria should be Does the tool work for what you plan to do? Much of what these tools offer is mature technology that would satisfy most people's evaluation criteria, so you should have more than one choice in each software category.

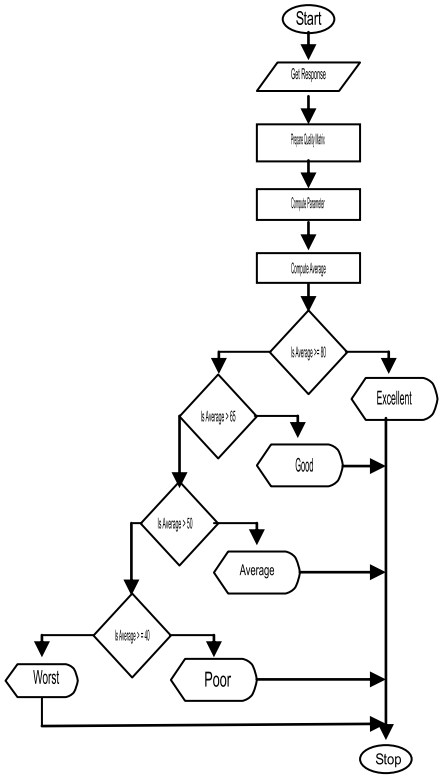
4.1.5 Perform the data warehouse software evaluation.

Two words of advice: keep it short and if you can structure it as a Proof-of-Concept (POC) that can be used in your project, all the better.

5. PROPOSED ALGORITHM FOR MEASUREMENT OF QUALITY

Step 1. Develop parameters.
Step 2. Prepare questionnaire.
Step 3. Distribute questionnaire to End Users.
Step 4. Collect data.
Step 5. Prepare data for analysis.
Step 6. Use defined parameters.
Step 7. Develop metric for Parameters.
Step 8. Evaluate parameters.
Step 9. Grade the warehouse.

6. PROPOSED FLOW CHART FOR MEASUREMENT OF QUALITY



7. PARAMETERS FOR MEASUREMENT

Data Warehouse can be evaluated on the basis of following attributes

- Accessibility

  Data Sources

  DW Design

  DW Processes

-Interpretability

  DW Design

  2. Models & Languages
  Query Processing

  DW Data & Processes

-Usefulness

  Update Policy

  DW Evolution

  DW Processes

-Believability

  Data Sources

  DW Design

  DW Processes

-Validation

  DW Processes

8. EVALUATING THE DATA WAREHOUSE

The best person who will decide that the data warehouse design is good or not is 'End User ' . The best way to evaluate the system quality is to ask the end user.

The Quality features are as follows (as per jarke and Vassiliou 1997)

- Accessibility
- Usefulness
- Interpretability
- Validation
- Believability

8.1 Interpretability

Suitability for interpretation with respect to answering adequately; requirements on a given type of target in terms of quality and scale. It can be measured in terms of drill-down capability.

8.2 Accessibility

Accessibility can be measured in terms of the ability to obtain at least as much information from the data warehouse as from current system.

8.3 Believability

Believability can be measured in terms of drill-down capability.

8.4 Validation

Validation can be measured by analyzing the users who were surveyed indicated that documentation of data sources was an important feature for validating data warehouse reports.

8.5 Usefulness

The metric chosen to measure the usefulness factor of the data warehouse in terms of the ease of use of the end user tools. It is the total of the number of steps required to produce a report.

9. CONCLUSION

A methodology to measure for quality of data warehouse has been suggested and evolved. The technique provides the framework and modeling formalisms for quality measurement tool that can be tuned to particular application domains and various levels of quality.

10. REFERENCES

[1] S. Adelman, Measuring the Effectiveness of Your Data Warehouse May, 2002. http://www.ambeo.com/acrobat/Sid%20Measuring%20the%20Effectiveness.pdf.

[2] Corpo Ulisse Di, How to prepare a questionnaire or a form. Syntropia 2005, 2, 64-68. http://www.sintropia.it/english/2005-eng-2-2.pdf

[3] L. Greenfield, The case against data warehousing.2001 http://dwhandbi.com/articles/20080623-The-Case-Against Data-Warehousing.php

[4] L. Hadley, Data Warehouse Quality Management http://www.users.qwest.net/~lauramh/resume/dwqual.htm2

[5] L. Hadley, Developing a Data Warehouse Architecture. http://www.users.qwest.net/~lauramh/resume/thorn.htm

[6] D. Henderson, Performance Measurement: The Data Warehouse Supports Best Practices. 1999

[7] Li Xiao and Dasgupta Subhasish, measurement of user satisfaction with web-based information systems: an empirical study, 2002

[8] M. Jarke and Y. Vassiliou, Data Warehouse Quality Design: A Review of the DWQ Project. 2nd Conference on Information Quality, Massachusetts Institute of Technology, Cambridge, 1997

[9] M. Schneider, Well-formed data warehouse structures, DMDW 2003 Germany. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-77/02_Schneider.pdf.

[10] T. Tobey, Core principles of data warehouse design, 27 Feb 2006. http://searchoracle.techtarget.com/tip/0,289483,sid41_gci1169348,00.html

[11] R. Weir, Taoxin Peng, and Kerridge Jon, Best Practice for Implementing a Data warehouse: A Review for Strategic Alignment, DMDW, 2003.