THE QUALITY OF DATA AND METADATA IN A DATAWAREHOUSE

Carmen Răduț[1]

Summary: *Data quality is an important concept for the economic applications used in the process of analysis. Databases were revolutionized when they first started being used with large amounts of data. From this point on, an important process is represented by storing multidimensional data in data warehouses, in order to be processed and analyzed with the purpose of obtaining information which can be used for decision making in various activities. Specialty studies show that most data is not useful for the purpose it has been collected because of both the lack of quality and incorrect techniques of manipulating this data. This study will try to offer a process of obtaining quality data in data archives and how to avoid quality anomalies inside metadata.*

Keywords: *metadata, data warehouse, quality, architecture*

JEL Classification: C81, C89

1. Introduction

The ideal solution for complex **B**usiness **I**ntelligence applications is represented by the introduction of a data warehouse. In most of the cases, the data warehouses do not meet the requirements of the beneficiary because of the low quality of the data. This means that the quality of the data is an important problem for both implementing and managing data warehouses. In order to use efficiently the information from a data warehouse, the data should be put through an analysis and cleaning process. This way, in BI applications, there is the need of continuous cleaning of data for the results to meet the needed standards.

The area where the data warehouse starts to fail is the data transfer from different data sources, into one data warehouse through an integrating process. The development of for business applications, is an important process in the development of informatics systems.

In order to define a data warehouse, first we need to set the **data granularity** and the **abstracting level** needed to sustain the process of decision making in the management activity. Recent studies show that more than 80% of the data warehouse projects go beyond the allocated budgets because of *not understanding data* or *incorrectly defining* them. A similar study shows that the *data quality* is also and important reason which leads to the failure of these projects. This is why the need of proper evaluation of data quality both at the moment of the data warehouse setup and during the gradual process of data introduction (Singh R. i Singh K., 2010).

2. Data quality in data warehouses

Data quality in data warehouses means that the following procedures must be completed: data quality evaluation, data quality related problem classification, data warehouse structure, data warehouse architecture/quality model, data quality tools, data quality standards, metadata quality control system (Palepu and Rao, 2012).

2.1. Information quality evaluation is a defined part by the total quality data management (TQdM). The TQdM methodology of quality management consists of going through six processes of measurement and improvement of the quality of data resulted from the modification of the environment. The six processes regarding *measurement and improvement of information quality*: data definition and quality architecture of data,

[1] Conf.univ.dr., Universitatea “Constantin Brâncoveanu” Pitești, Facultatea Management Marketing în Afaceri economice Râmnicu Vâlcea, c_radut@yahoo.com

information quality evaluation, cost measurement of low-quality information, data restructuring and cleaning, information process quality evaluation, information quality frame setup.

These 6 steps are divided into sub-stages in order to achieve the required quality of data. In order to determine the data quality, an evaluation of all the fields is required. Simply having the data is not enough. The context in which the data is going to be used must be known, in order to obtain **metadata.** Metadata is information attached to the data in order to provide additional features to the data existing in the data warehouse regarding the content and purpose. *Quality criteria* is: data type integrity, business rules integrity, name and address integrity. Also, the *characteristics and measures regarding information quality* are: definition consistency, value integrality, availability and compliance with the business rules, non-duplication of the data, accessibility.

2.2. Classifying data quality related problems. The analysis of the causes which lead to problems related to data quality, data design planning, means, first of all, identifying the problem related to **data quality**. Identifying and classifying these problems is necessary both for the data warehouse and data quality. This way problems regarding data quality can be identified on different levels**: from data sources, during the data profile setup stage, during the ETL stage, during the scheme design period, in the data warehouse obtaining period** (Palepu and Rao, 2012)**.**

a. Problems related to the quality of data received from the sources. Data sources have some problems such as the fact that old data sources do not have metadata, to provide a short description. The main reasons for impure data are data errors which appear during input because of human error or computational system and data updating errors. A part of the information comes from text files or MS Excel files but it also can be obtained by connecting directly through ODBC to data sources. Some files are the result of manual consolidation of files, which leads to the lack of quality of the data. The main reasons which lead to low quality data are: wrong selection of data sources, not knowing the internal dependencies amongst data sources, the lack of validation procedures for data sources, unexpected modifications in source systems, multiple data sources generate semantic heterogeneity, leading to quality problems; creating inconsistent/inexact data, multiple sources for the same data.

b. Problems related to the data quality during the data profile setup stage. After the possible sources of data are identified, the setup of the data profile is set. Setting up the data profile consists of examining and evaluating the quality, integrity and coherence of the data from the source systems. This is also known as ***source systems analysis***. Profiling data is fundamental although many times it is ignored or it is not given the proper attention, which leads to low quality data inside the data warehouses. This way, the reasons for quality issues during the data profile setup are: the data profile from the data sources is insufficient; wrong selection of processing tools; insufficient structural analysis of the data sources during the data profile setup stage; incorrect or incomplete metadata.

c. Data quality issues during the ETL (extraction, transformation, loading) **stage.** The ***extraction, transformation*** and ***loading*** stage is a very important one because the most important aspect regarding data quality starts with proper data storing capacity. The cleaning process of data is executed during the data collection step in order to improve the quality of the data warehouse. The problems which can affect the data quality during the ETL stage are the architecture of the data warehouse; the relational or non-relational connection of the data warehouse; business rules applied to the data sources; lack of periodical update of the data.

d. Data quality problems during the projecting scheme stage. The data quality is dependant on three things: actual data quality, application quality and database scheme quality. Special attention during the projecting scheme stage must be given to some aspects such as dimensions which modify fast/slow such as multiple value dimensions, etc., from data warehouses. A bad designed data scheme has a negative impact on the quality of the information. Problems regarding data quality in the scheme modeling stage are given by: incomplete or erroneous requirements for the scheme; dimensional modeling selection for the star or snow flake type schemes, multiple value dimensions; identification for dimensions with slow change; incomplete and erroneous identification of the events. During the conception period of the data warehouse, dimensional models are used which group data into relational tables in star or snowflake type schemes. In these schemes, quantitative data can be found or grouped by certain criteria. The quantitative data from the dimensional data bases has different types such as transaction numbers, centralization by certain characteristics, totals and activity measures.

e. Data quality problems in data warehouses. A *definition of data quality* refers to incorrect data or missing data which is incorrect or invalid in a certain context. Data quality is reached at the moment when the organization uses complete data, easy to understand, coherent and relevant. Understanding the dimensions regarding data quality is the first step forward to improving it. The data quality dimensions can include: correctness, reliability, importance, coherence, precision, promptitude, finesse, clarity and utility. This way, data quality is given by the following key dimensions: consistency, conformity, validity, accuracy and integrity.

2.3 Data warehouse structure for obtaining data quality. We will start off with the definition of a data warehouse written by W.H. Inmon (2002, p.8) "A **data warehouse** is a collection of data sorted by subject, integrated, historical and nonvolatile destined to supporting the managerial decision making process". A data warehouse is a collection of technologies which allow the ones working with the information to be able to make better and faster decisions. Practice has shown that online transaction processing systems are not the best for supporting decisions and high speed networks and cannot handle the problem regarding information accessibility. The data warehouse has become an important strategy for integrating many heterogeneous information sources in an organization and for allowing Online Analytical Processing of Transactions. Data quality from the data warehouse is dependent on the semantic models or the warehouse's architecture. Because of the stored data in the data warehouse, the information is analytically processed and the warehouse's framework must handle two aspects:

–How can we harmonize the entry flow of data which comes from many heterogeneous sources?

–How can we customize the data storage for different applications?

The data warehouse's designing decision depends on the needs of the business and that will play a major role. The objectives regarding the data warehouse's quality are: meta-databases which contain formal models regarding the information quality for the optimization of the design of the database; modeling the informational resources; modeling the data warehouse scheme so that the designers and the ones who optimize the interrogations to be able to benefit explicitly of the temporal, spatial and aggregated dimension of the data from the data warehouse. During the whole designing stage of the data warehouse, the designing is centered on: the quality of the data warehouse and the system architecture; designing methods, language and designing the data warehouse; optimization. In order to complete these objectives a model is designed for the data.

2.4. Data warehouse quality architecture. During the interpretation of the data these are analyzed from *a conceptual, logical and physical* perspective. These three perspectives need to be maintained because the quality factors are associated with the certain perspective or with the relation between these perspectives. **The conceptual perspective –**

in terms of data warehouse quality, the conceptual model defines the organization's theory. Regarding this theory, real observation can be evaluated from a quality factor point of view such as ***accuracy, opportunity, exhaustiveness.*** After this stage, the ***source model*** is obtained. **The logical perspective** – designs the data warehouse from a real data point of view. Researchers that follow this perspective take into consideration the fact that the data warehouse is a collection of materialized visualizations, extracted from information sources. After this stage ***the source scheme*** is obtained. **The physical perspective** – interprets the data warehouse architecture as a network of smaller data warehouses which generate data and communication channels, targeting quality factors such as ***reliability and performance*** in the presence of large amounts of information with low modification rate. After this stage the ***storage of data sources*** is obtained (Palepu and Rao, 2012).

2.5. Instruments regarding Data Quality. Studies show that 75% of the effort needed with the data warehouse is allocated to problems related to: data readiness and transport to the data warehouse. At the moment there are tools associated with ***auditing, cleaning, extracting*** and ***loading*** data into data warehouses. These tools are used to audit data at the source and transform it so that they can be consistent in the data warehouse and make sure that the data corresponds to the business rules. These tools can be independent packets or can be integrated with the existing data warehouse packets.

The tools which target data quality usually fall into one of the three categories: **auditing, cleaning** or **extracting/migrating**. The main objective is to clean and audit data, extraction and migration tools having a limited share.

a. Data audit tools increase accuracy and precision of the data sources. These instruments compare information from the source database with the set of business rules. When an external source is used, the business rules are determined by the external information source. In this case, it is established using the **data mining** technique and allows to obtain the data model. The internal business rules are introduced in the evaluation stage of data sources, and the lexical analysis is used for discovering the meaning of this information.

b. Data cleaning tools are used in the intermediate stage. These instruments guarantee the analysis, standardization and verification of data based on already known lists.

c. Data migration tools are used for extracting data from a source database and for migrating it in an area for temporary storage. The migration instruments transfer data from the temporary storage area in the data warehouse. The tools provide data conversion from one platform to the other and the migration utility maps the data from the source to the data warehouse, checking the compliance and doing the basic data cleaning activity.

2.6. Data quality standards. The data quality standards are useful for evaluating if the required specifications by the clients are met. The objective quality standards refer to: *accessibility, precision, opportunity, integrity, availability, coherence, relevance*. The data quality standards are meant to: help decision making based on facts; estimate losses because of the lack of quality. Data quality measurements must always respect the legal regulations, business rules and special cases. The measurements must meet the following criteria in order to guarantee the high quality data: all measurements must frequently report the quality control factors from the organization; defining methods used in order to control measurements; every measurement unit has certain attributes: name, definition, calculation method, data elements and data source.

2.7. Metadata quality control system. In order to obtain high quality data for the data warehouse, the starting point is represented by the client's requirements, different aspects present in the organization, the organizational structure, features and responsibilities regarding the continuous quality improvement. This is also known as Total

Quality Management (TQM). The quality control evaluation accounts for: ***planning quality*** **–** the quality evaluation criteria are selected, classified and given priorities for control activities and ***quantitative quality measurement***. **Quality control system metadata** will be collected and then transformed to meet the required specifications such as: information requirements, user quality requests, etc. The metadata quality control system (MQC) includes: data warehouse architecture and the overall quality which is measured along the data flow. MQC system architecture is shown in Figure 1. Operations of extraction, transformation and loading (ETL) data are important / essential in making data warehouse. This way, the data quality contained in the data warehouse, along with the information extraction from different smaller databases and transforming them to meet the requirements of the scheme model of the data warehouse, is a priority, followed by loading data in the data warehouse.

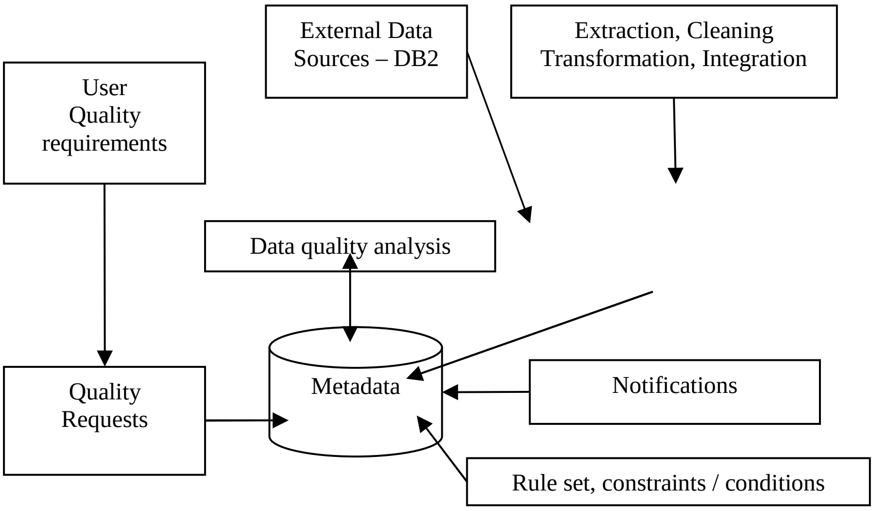
External Data Sources – DB2
Extraction, Cleaning Transformation, Integration
User Quality requirements
Data quality analysis
Quality Requests
Metadata
Notifications
Rule set, constraints / conditions

Figure 1. MQC system architecture

During this process, one of the key components refers to the metadata. Metadata stores features for the data introduced in the data warehouse. Metadata screening, quality standards enforcement, identifying quality control problems, lead to fixing these issues. The modifications suggested in the data warehouse architecture according to the quality control standards can be found in meta-databases. This leads to an operation synchronization with the data warehouse database.

Conclusions

The metadata quality control system guarantees overcoming the problems related to the quality of the data in the data warehouse. **Problems regarding data quality** in the data warehouse are solved through: planning, using evaluation criteria for the data quality, through total data quality management and classifying problems related to data quality in the data warehouse, evaluating the data warehouse structure regarding quality, and using instruments for data quality, quality control standards and metadata quality control system.

Bibliography

1. W. H.Inmon, (2002), ”Building the Data Warehouse”, John Wiley & Sons, San Francisco
2. Ramesh Babu Palepu, Dr K V Sambasiva Rao, (2012), ”Meta data quality control architecture in data warehousing”, *International Journal of Computer Science, Engineering and Information Technology* (IJCSEIT), Vol.2, No.4, pp. 15-24
3. Ranjit Singh, Kawaljeet Singh, (2010), ”A descriptive classification of causes of data quality problems in data warehouse”, *International Journal of Computer Science Issues,* Vol 7, pp. 23-29