META DATA QUALITY CONTROL ARCHITECTURE IN DATA WAREHOUSING

Ramesh Babu Palepu[1], Dr K V Sambasiva Rao[2]

Dept of IT, Amrita Sai Institute of Science & Technology[1]
MVR College of Engineering[2]
asistithod@gmail.com

Abstract:

The quality is the key concept in each and every analysis as well as in computing applications. Today we gather large volumes of information and store them in multidimensional mode that is in data warehouses then analyze the data to be used in exact decision making in various fields. The studies proved that most of the volumes of data are not useful for analysis due to lack of quality caused by improper data handling techniques. This paper try to find out a solution to achieve the quality of data from the foundation of data repositories and try to avoid quality anomalies at meta data level. This paper also proposes the new model of Meta data architecture.

Keywords:

Data warehouse, Meta data, ETL, Total Quality Management

INTRODUCTION:

Implementation of Data Warehouse is right solution for complex business intelligent applications. The Data Warehouses fail to meet the expectations because of lack of data quality. Once the data warehouse is built, the issue of data quality does not go away. The data quality is a serious issue in implementation and management of data warehouses. Before using the data within data warehouses effectively, the data is to be analyzed and cleaned. In order to reap the benefits of Business Intelligence, it is necessary to apply ongoing data cleaning processes and procedures and track data quality levels over time.

The key area where the data warehouse fails is, moving data from various sources into easily accessible single repository, that is, integration of the data. The data quality begins with understanding the entire data within the data warehouse. The development of data warehouse for business application for coordinated development of information systems is preferred to which process different applications for the same purpose.

When the data warehouse is defined, the main task is to define granularity of data in the data warehouse and different level of abstraction which will support the best decision making process. A recent study by Standish Group states that 83% of data warehouse projects overrun their budgets primarily as the result of misunderstanding about the source data and data definitions. A similar survey conducted by Gartner Group point to data quality as a leading reason for overrun and failed projects.

DOI : 10.5121/ijcseit.2012.2402

To avoid such a pitfall, a thorough assessment of the quality of data is to be performed when data warehouse is build and the data warehouse is populated with data gradually.

1. Quality Assessment Criteria:

The following is the assessment criteria defined according to English's Total Quality data Management (TQdM) Methodology and Forino's recommendations.

Data quality assurance is a complex problem that requires a systematic approach. English proposes a comprehensive Total Quality data Management Methodology which consists of six processes of measuring and improving information quality and also about cultural and environmental changes to sustain information quality improvement. The following are the six processes,

1. Assess data definition and information architecture quality.
2. Assess information quality.
3. Measure non quality information costs.
4. Reengineer and cleanse data.
5. Assess Information process quality.
6. Establish information quality environment.

The above steps are further divided into sub processes to achieve the desired data quality.
To determine the quality of data a field-by-field assessment is required. Simply possessing data is not enough, but the context for which the data is to exist must also be known, that is, Meta data. Now, the assessment of the quality can be extended depending on availability of Meta data. Now the quality criteria is defined as,

1. Data type integrity.
2. Business rule integrity.
3. Name and address integrity.

On the other hand, English defines the following inherent information quality characteristics and measures;

1. Definition conformance
2. Completeness of values
3. Validity or business rule conformance
4. Accuracy to surrogate source
5. Accuracy to reality
6. Precision
7. Non duplication of occurrence
8. Equivalence of redundant or distributed data
9. Concurrency of redundant or distributed data
10. Accessibility

2. Classification of Data Quality Issues:

In order to analyze and to determine the scope of the underlying root causes of data quality issues and to plan the design, it is valuable to understand these common data quality issues. This classification is very helpful to the data warehouse and data quality community.

2.1 Data Quality Issues at Data Sources:

Different data sources have different kind of problems associated with it such as data from legacy data sources do not even have metadata that describe them. The sources of dirty data include data entry error by a human or computer system, data update error by a human or computer system. Part of the data comes from text files, part from MS Excel files and some of the data is direct Open Data Base Connectivity (ODBC) connection to the source database. Some files are result of manual consolidation of multiple files as a result of which data quality might be compromised at any step. The following are the some of the causes of data quality problems at data sources.

1. Inadequate selection of candidate data sources.
2. Inadequate knowledge of inter dependencies among data sources.
3. Lack of validation routines at sources.
4. Unexpected changes in source systems.
5. Multiple data sources generate semantic heterogeneity which leads to quality issues.
6. Contradictory information presents in data sources.
7. Having inconsistent/Incorrect data formatting.
8. Multiple sources for the same data.
9. Inconsistent use of special characters of data.

2.2 Causes of Data Quality Issues at Data Profiling Stage:

When possible candidate data sources are identified and finalized data profiling comes in play immediately. Data profiling is the examination and assessment of our source systems' data quality, integrity and consistency sometimes also called as source systems analysis. Data profiling is a fundamental, yet often ignored or given less attention as result of which data quality of the data warehouse is compromised. The following are the causes of data quality issues at data profiling stage. The following are the causes for data quality at data profiling stage.

1. Insufficient data profiling of data sources.
2. Inappropriate selection of automated tools.
3. Insufficient structural analysis of data sources at profiling stage.
4. Unreliable and in complete Meta data causes data quality problems.

2.3 Data quality issues at data staging ETL (Extraction, Transformation and Loading):

The extraction, transformation and loading phase is very crucial where maximum responsibilities of data quality efforts reside in data warehousing. The data cleaning process is executed in data staging in order to improve the accuracy of data warehouse. The following are data quality issues at ETL.

1. The data warehouse architecture affects the data quality.
2. The relational and non relational effects in data warehouse affect the data quality.
3. Some times the business rules applied on data sources may cause for quality problems.
4. Lack of periodical refreshment of integrated data cause for data quality problems.

2.3 Data quality problems at schema design stage:

It is well known that the quality of data is depending on three things.

1. The quality of data itself.

2. The quality of application program.
3. The quality of data base schema.

A special attention during the schema design by considering some issues such as slowly and rapidly changing dimensions, multi valued dimensions etc in data warehouse. A flaw schema has negative effect on data quality. The following are data quality issues at schema modeling phase.

1. In complete or wrong requirement analysis for schema design.
2. The selection of dimensional modeling such as star, snowflake, and fact constellation may affect the data quality.
3. Multi valued dimensions, late identification of slowly changing dimensions and late arrival dimensions may cause data quality problems.
4. In complete and wrong identification of facts may cause the data quality problems.

3. Data Quality problems in Data Warehouses:

The one definition of data quality is that it's about bad data - data that is missing or incorrect or invalid in some context. A broader definition is that data quality is achieved when organization uses data that is comprehensive, understandable, consistent, relevant and timely. Understanding the key data quality dimensions is the first step to data quality improvement. The dimensions of data quality typically include accuracy, reliability, importance, consistency, precision, timeliness, fineness, understandability, conciseness and usefulness. Here the quality criteria is undertaken by taking six key dimensions as depicted below.

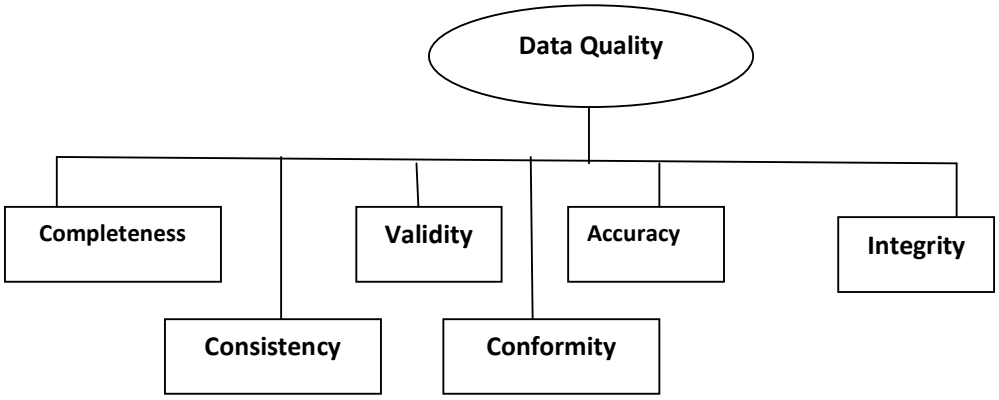

Figure 1: Data Quality Criteria

4. Data Warehouse Structure to achieve the Quality of Data:

A data Warehouse is a collection of technologies aimed at enabling the knowledge worker to make better and faster decisions. It is expected to present the right information in the right place at the right time with in the right cost in order to support the right decision. The practice proved that the traditional online transaction processing (OLTP) systems are not appropriate for decision support and the high speed networks cannot by themselves solve the information accessibility problem. Data Warehouse has become an important strategy to integrate heterogeneous information sources in organization, and to enable online Analytical Processing. The quality of the data within the data warehouse is also depending on the semantic models of data warehouse architecture.

The Data Warehouse stores the data which is used for analytical processing, but when frame the Data Warehouse it face two essential questions:

1. How to reconcile the stream of incoming data from multiple heterogeneous sources?
2. How to customize the derived data storage to specific applications?

The design decision of Data Warehouse to solve the above two problems is depending on the business needs, therefore in Data Warehouse design, the change management and design supports play an important role. The following are the objectives of Data Warehouse Quality.

1. The Meta databases with formal models of information quality to enable adaptive and quantitative design optimization of data warehouses.
2. The information resource models to enable more incremental change propagation and conflict resolution.
3. The Data Warehouse schema models to enable designers and query optimizers to take explicit advantage of the temporal, spatial and aggregate nature of DW data.

During the whole spectrum of the Data Warehouse modeling, design and development will be focused on:

1. The Data Warehouse Quality framework and system architectures.
2. Data Warehouse Meta modeling and designing methods and languages.
3. Optimization.

To achieve the above objectives, a quality model can be designed for data. This quality model has become specialized for the utility of Data Warehousing and has great emphasis on historical as well as aggregated data.

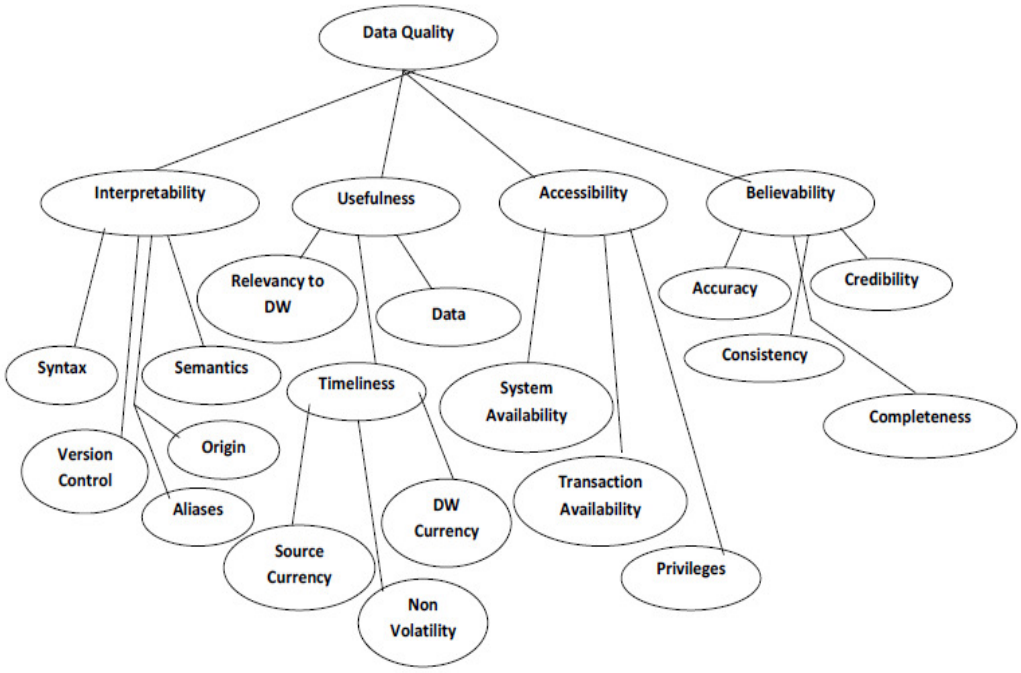

Figure 2: Quality Factors In Data Warehousing

5. Data Warehouse Quality Architecture:

While interpreting the Data Warehousing in more detail; any Data Warehouse component can be analyzed in the **conceptual perspective, logical perspective,** and **physical perspective.** Moreover in the design, operation, and in evolution of Data Warehouse, it is important that these three perspectives are maintained consistent with each other. Finally quality factors are associated with specific perspective or specific relationship between perspectives. The following diagram represents how to link quality factors to Data Warehouse tasks.

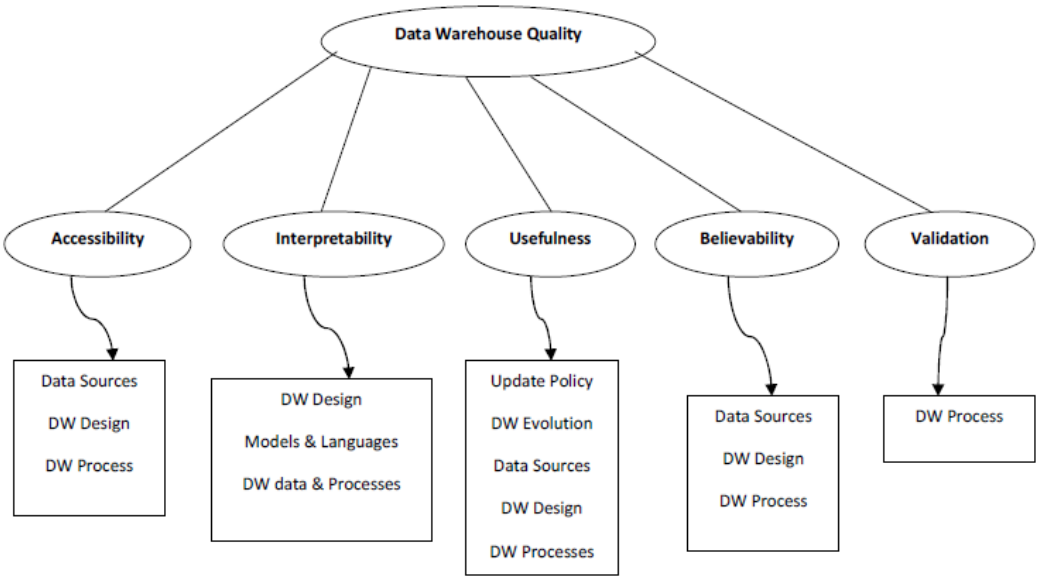

Figure 3: Linking Quality Factors to Data Warehousing Tasks

5.1 Conceptual Perspective: In terms of Data Warehouse quality, the conceptual model defines a theory of the organization. Actual observation can be evaluated for quality factors such as accuracy, timeliness, completeness with respect to this theory. Moreover, the fact that data warehouse views intervene between client views and source views can have both a positive and a negative impact on information quality.

5.2 Logical Perspective: The logical perspective conceives a Data Warehouse from the view point of the actual data models involved. Researchers and practitioners following this perspective are the ones that consider a Data Warehouse simply a collection of materialized views on top of each other, based on existing information sources.

5.3 Physical Perspective: The physical perspective interprets the data warehouse architecture as a network of data stores, data transformers and communication channels, aiming at the quality factors of reliability and performance in the presence of very large amounts of slowly changing data.

The following diagram shows DW architecture within three perspectives.

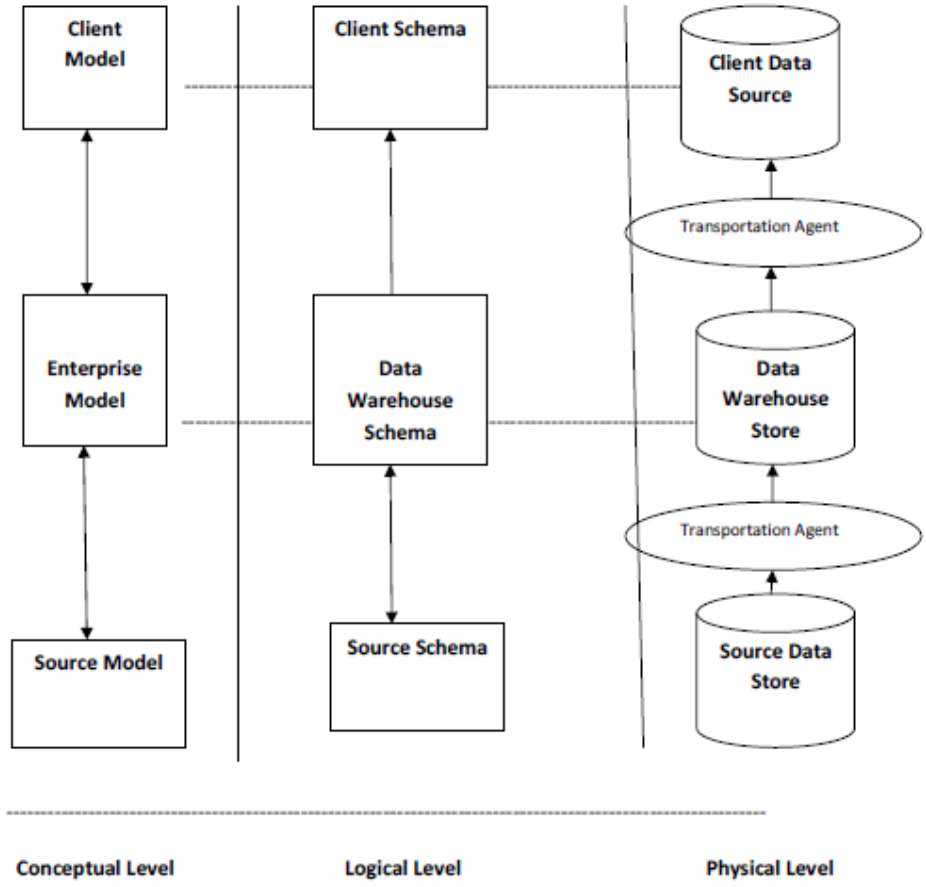

Figure 4: Three Aspects of Data Warehouse Architecture

6. Data Quality Tools:

It is known that 75% of the effort spent on data warehouse is attributed to back end issues, such as readying the data and transporting it into the data warehouse. Now a days, hundreds of tools are available to automate the tasks associated with auditing, cleansing, extracting and loading data into data warehouses. The quality tools are emerging as a way to correct and clean data at many stages in building and maintaining a data warehouse. These tools are used to audit the data at the source, transform the data so that it is consistent throughout the warehouse, segment the data into atomic units and ensure that the data matches the business rules. The tools can be stand alone packages, or can be integrated with data warehouse packages.

The data quality tools generally fall into one of three categories: auditing, cleansing and migration. The main focus of these tools is on cleaning and auditing and the data and extraction cum migration data tools will have a limited look.

Data auditing tools enhance the accuracy and correctness of the data at the source. These tools generally compare the data in the source database to a set of business rules. When using a source external to the organization, business rules can be determined by using a source external to the organization, business rules can be determined by using data mining techniques to uncover

patterns in the data. Business rules that are internal to the organization should be entered in the early stages of evaluating data sources. Lexical analysis may be used to discover the business sense of words within the data. The data that does not adhere to the business rules could then be modified as necessary.

Data cleansing tools are used in the intermediate staging area. The tools in this category have been around for number of years. A data cleansing tool cleans names, addresses, and other data that can be compared to an independent source. These tools are responsible for parsing, standardizing and verifying data against known lists.

The third type of tool, the data migration tool, is used in extracting data from a source database, and migrating the data into an intermediate storage area. The migration tools also transfer data from the staging area into the data warehouse. The data migration tool is responsible for converting the data from one platform to another. A migration tool will map the data from the source to the data warehouse. It also checks for Y2K compliance and other simple cleansing activities.

7. Data quality Standards

The data quality standards are useful to assess whether the requirements specified by the customers are achieved or not. The following are data quality standard objectives.

2.4 Accessibility
2.5 Accuracy.
2.6 Timeliness.
2.7 Integrity.
2.8 Validity.
2.9 Consistency.
2.10 Relevance.

The data quality standards are meant for;

1. Making the decisions based on facts.
2. Useful to take corrective actions.
3. Useful in access source of quality problems.
4. Useful to estimate losses due to lack of quality.
5. Estimate or calculate the solutions which below and or above the intended goals.
6. To provide quality in solutions.

Let us consider the measurements for a specific problem or a product and then frame the quality metrics, before actual data quality standards are defined. The data quality metrics should always met the legal regulations, business rules, and urgent, special cases. The metrics defined to access the information quality should satisfy the following criteria.

1. All metrics should frequently report about the quality control factors with in an organization.
2. Define methods, which are used to controlling the metrics.
3. Each and every metric should consist of certain attributes such as metric name, metric definition, calculation, data elements, and data source.

8. Meta data based quality control system

To achieve high level data quality in data warehouse, let us consider the customer requirements, participation of stack holders, and all most all aspects with in the organization. This is known as the Total Quality Management (TQM). But to control the activities that are performed in the assessment of quality. The organizational structure, functionalities and responsibilities with continuous quality improvement are considered. To control the quality assessment activities consider two things are considered.

1. Quality planning - Here select the criteria for quality assessment, classify it and assign priorities to the controlling activities.
2. To measure the quality quantitatively.

In Meta data quality control system, first the information requirements or demands for quality from the users are collected and then transform these requirements into specifications. The Meta data Quality Control system (MQC) comprises the whole data warehouse architecture and the quality will be measured along the data flow. The following is the architecture of MQC system.

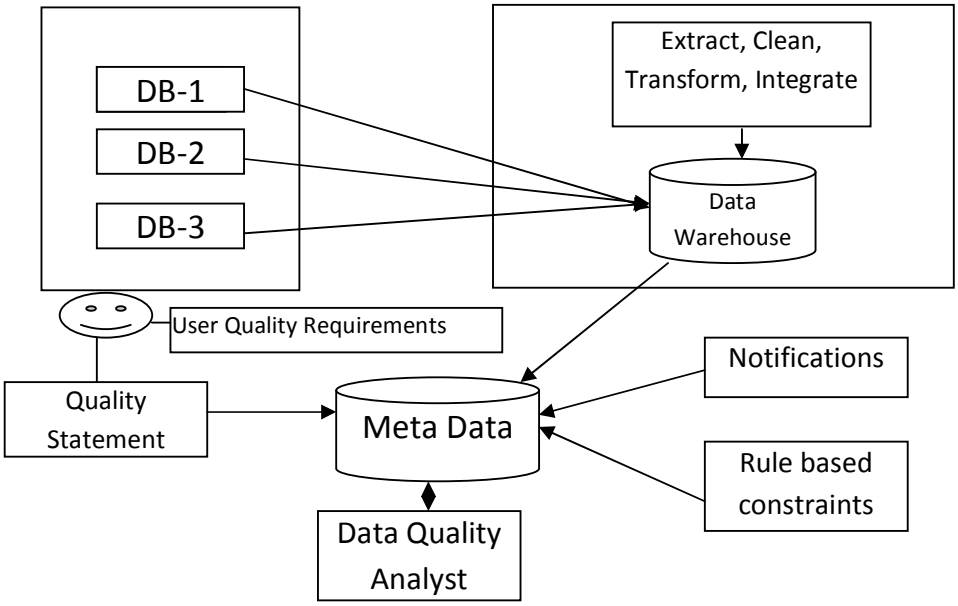

Figure 5: Meta data Quality Architecture.

Conclusion:

It is well known that Extraction, Transformation and Loading (ETL) operations are important as well as crucial in forming of Data warehouse. Let us consider maintaining the quality of data in Data Warehouse, first of all extracting the information from various small operational databases and then transforming the data into the schema model of Data Warehouse and finally load the data into the Data Warehouse. During this entire process one of the key components is Meta data repository which stores each and every characteristic of the data items loaded in the Data Warehouse. Hence screening the Meta data, identifying the quality control problems, and then applying quality standards on it leading to rectification of quality problems in Data Warehouse frame work. In addition to this, the changes are proposed in the Data Warehouse architecture

according to the quality control standards within Meta database. It leads to synchronization of the operations of Meta database with in Data Warehouse architecture. In this paper, a new Meta data quality architecture is proposed to overcome the data quality problems in Data Warehouse at the bottom level of Data Warehouse formation.

The proposed data quality issues in data warehouse provides planning, data quality assessment criteria, Total data Quality management, Classification data quality issues, data quality problems in data warehouse, data warehouse structures how they effect the data quality, different data quality tools, Quality control standards and Meta data quality control system.

References:

1. "Data quality tools for data warehousing – A small sample survey" available at www.ctg.albany.edu.
2. www.informatica.com/INFA_Resources/brief_dq _dw_bi_6702.pdf
3. Ranjit Singh and Dr. Kawaljeet Singh "A descriptive classification of causes of data quality problems in data warehouse" International Journal of Computer Science Issues, Vol 7, Issue 3, No 2 May 2010.
4. Valjan Mahnic and Igor Rozanc "Data Quality: A prerequisite for successful data warehouse implementation".
5. Markus Helfert and Clemens Herrmann "Proactive data quality management fir data warehouse systems" available at http://datawarehouse.iwi.unisg.ch.
6. Philip Crosby "The high costs of low quality data" available at http://pdf4me.net/pdf-data/data-warehouse-free-book.php.
7. Matthias Jarke and Yannis vassiliou "Data warehouse Quality: A review of the DWQ Project".
8. http://iaidq.org/main
9. www.thearling.com/index.htm