

# Data Quality Management

**Author:** BALWANT RAI  
**Organization:** Evaltech, Inc.  
**Evaltech Research Group,  
Data Warehousing Practice.**  
**Date:** 07/17/2004  
**Email:** erg@evaltech.com

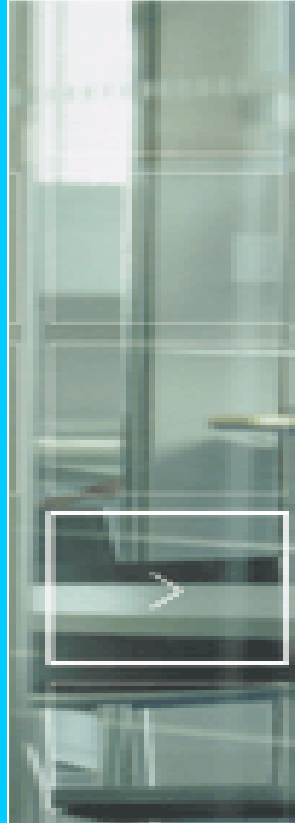


## **Abstract:**

The difficulties associated with dealing with the mountains of data produced in businesses brought about the concept of information architecture which has spawned projects such as Operational Data Stores (ODS), Data Warehousing and Data Marts. Along with these came a set of associated complementary technologies which help companies collect, massage, process, analyze and deliver useful information from this mass of raw, unconnected data. The growth of Data Warehousing into \$ billions market demonstrates the degree to which organizations have taken a pro-active role in managing their data.

## **Intellectual Property / Copyright Material**

All text and graphics found in this article are the property of the Evaltech, Inc. and cannot be used or duplicated without the express written permission of the corporation through the Office of Evaltech, Inc.



## **INTRODUCTION**

The difficulties associated with dealing with the mountains of data produced in businesses brought about the concept of information architecture which has spawned projects such as Operational Data Stores (ODS), Data Warehousing and Data Marts. Along with these came a set of associated complementary technologies which help companies collect, massage, process, analyze and deliver useful information from this mass of raw, unconnected data. The growth of Data Warehousing into \$ billions market demonstrates the degree to which organizations have taken a pro-active role in managing their data. In a few short years, data warehousing has passed from theory to conventional wisdom. In the explosive growth that has transpired, a body of thought has developed surrounding it. From the beginning, data warehousing was never a theoretical exercise, but has always been rooted in pragmatism. But as is inevitable given the breathtaking growth that has been the lot of data warehousing, an organized, thorough intellectual framework has begun to grow around both its infrastructure and rationale. Here are man aspects to this intellectual framework. One of the important considerations, critical to the infrastructure, is the quality of data that courses through the veins of components of the warehouse. Indeed, quality in many different forms is one the cornerstones of data warehousing. If the data warehouse is ever to achieve the loft goal of becoming a foundation for enterprise intelligence, data quality must become a reality. It is simply unthinkable that analysis for important corporate decisions should proceed on the basis of incorrect and incomplete data. Here fore, a de facto prerequisite for enterprise intelligence is quality throughout the data warehouse environment.

### **Enterprise Data Quality Management**

After some years of attempting to deal with the issue of data quality, a new discipline has emerged within information architecture development to address the need for appropriately managing data quality. This discipline, known as Enterprise Data Quality Management, (EDQM) is intended to ensure the accuracy, timeliness, relevance and consistency of data throughout an organization, or multiple business units within an organization, and therefore to ensure that decisions are made on consistent and accurate information.

The single most important reason for improving the quality of data within organizations and the processes that drive EDQM is to improve customer relationships that ultimately add to the bottom line. It represents the added revenues that are realized when businesses correctly model and track their customer relationships, product or service preferences. With reliable data quality processes;

One of the largest software companies in the world was able to decrease their IT personnel costs devoted to data quality by 60%, revenues and sales reports were generated in two days, instead of five. The IS department now provides a value-added corporate-wide service by ensuring quality data for channel analysis, revenue forecasts, inventory re-stocking, general ledger postings and other functions. Within the first year of implementation they experienced a ten-fold ROI from our initial investment.

Similarly, an insurance company was able to cleanse and standardize the names and addresses from its customer information files, resulting in a 62% reduction in names and an 80% reduction in addresses from duplications. This translated into huge savings in processing time, storage and mailing costs, in the confidence users have in their own data, analysis and conclusions, but most importantly in the cost of contacting customers and managing ongoing customer relationships. Clearly, information is of value only if it is accurate, and in today's more complex information technology, when internal and external data are blended together in data warehouses and more advanced OLAP (on-line analytical processing) applications, new technology processes to ensure

the accuracy of information are required. Today, more than ever, it is imperative to tackle the data quality issue from a point of prevention as well as cleansing existing data stores.

### **Issues of Data Quality**

The heart of the corporate information factor is the data warehouse. The first major issue of data quality in the corporate information factor is how to ensure the data arrives in the data warehouse with the highest degree of quality. There are three opportunities for ensuring data quality as data is prepared for loading into the data warehouse. All of these opportunities for data quality have their own considerations. In fact, it is recommended that all three be used together for maximum effectiveness.

#### **Data Cleansing at the Application Level**

At first glance, it appears that the most natural place for assuring data quality is in the application. Data first enters the corporate information factor and is captured in the application. Indeed, the cleaner the data at the point of entry, the better off the corporate information factor. One theory says that if the data is perfectly cleaned at the application level, it need not be cleaned elsewhere. Unfortunately, this is not the case at all.

Several mitigating factors prevent the application from being the panacea for data quality. The first difficulty is the state of the application itself. In many cases, the application is old and undocumented. Applications programmers are legitimately scared to go back into old application code and alter it in a significant way. The fear is that one problem may be fixed, but two others may arise. Fixing one problem then might set off a cascade of other problems. The result is that the application is worse off than it was before the application was maintained. The second reason why applications developers are loathe going back into old code is that they see no benefit in doing so. Application developers focus on immediate requirements, and they see no urgency, or for that matter, a motivation, in going back into old code and modifying it to solve someone else's problems. But even if you could magically and easily do anything you wished at the application level, you would still need to cleanse data elsewhere in the corporate information factor. The reason why integration and transformation cleansing is still necessary elsewhere, even when application data is perfect, is that application data is not integrated. Data may be just fine in the eyes of a single application developer or user. But the data residing in the application still needs to be integrated across the corporate information factor. There is a big difference between cleaning application data and integrating application data. Only after the data comes out of the application is there a need and opportunity for integrating the data. The first opportunity for integration arises as data passes into the integration and transformation layer.

#### **Data Cleansing in the Integration and Transformation Layer**

Multiple applications pass data into the integration and transformation layer. Each application has its own interpretation of data, as originally specified by the application designer. A key, attributes, structures, encoding conventions is all different across the many applications. But in order for the data warehouse to contain integrated data, the many application structures and conventions must be integrated into a single, cohesive set of structures and conventions. There is then a complex task in store for the integration and transformation processing. Not only are keys, structures, and encoding conventions different across the many applications, but, in many cases, relationships between data within systems, as well as across systems, are undetected. Legacy information is often buried and floating within free-form text fields such as name and address lines, comment fields, and other data fields that have become a storage closet for meanings and relationships not accounted for in the original system. Data relationships may be hidden because initial systems did not provide a key structure that linked all relevant records, e.g., multiple account numbers might block the fact that all the records are from subsidiaries of the same company. Data anomalies in names, addresses, part descriptions, account codes are another area to rectify. And inconsistencies between Meta field definitions and the applications tend to surface over time as the application systems become part of the operational fabric of an organization, e.g., commercial

names mixed with personal names, addresses with missing information, truncated information, use of special characters as separators, missing values, abbreviations, etc. These quality issues can be found in one set of application data, can be multiplied when integrated within multiple applications, and can put the effectiveness of the resulting data warehouse at risk for delivering enterprise intelligence. The result of the tedious and difficult integration and transformation processing is integrated data. And the process of integrating the many applications together is certainly one form of cleansing data. It is noteworthy that this form of cleansing data is not possible until the data has passed out of the application. Therefore, there is another separate opportunity for data quality other than cleansing data in the application. But there is also third place where data quality needs to be addressed: after the data has been loaded into the data Warehouse.

#### **Data Quality inside the Data Warehouse**

Suppose that you could create perfect application and perfect integration and transformation programs. Would you still need a data quality facility within the data warehouse itself? The answer is "Yes." First of all, as new application data is added to the data warehouse environment, all the integration and transformation layer issues will be readdressed, and the new data may also uncover more hidden anomalies and relationships even in the warehouse itself. However, another key reason is that the data warehouse contains data collected over a spectrum of time. In some cases, the spectrum is as long as ten years. The problem with data collected over time is that data itself changes over time. In some cases, the changes are slow and subtle. In other cases, the changes are fast and radical. In any case, it is simple a fact of life that data changes over time. And with these changes comes the need to integrate data over time within the data warehouse after it has already been loaded. Even if data is entered perfectly from the applications and integration and transformation programs, there will still be a need to examine data quality inside the data warehouse over time. But has the data remained constant over those years? Hardly.

Even if data quality has been perfected elsewhere, it remains to be perfected one more time after the data enters the warehouse simply because data ages inside the warehouse.