A Framework for Information Quality in a Data Warehouse: 
IQ in the context of Data Marts and Data Warehouses

Jonathan Wu
BASE Consulting Group, Inc.
475 14th Street, Suite 600
Oakland, California 94612-1900 USA
(510) 628-3300 Ext. 224, (510) 628-3311
jwu@baseconsulting.com

Practice-Oriented Paper

Executive Summary

Data warehousing technology provides integrated data from a multitude of sources that is non-volatile and transformed into meaningful information for decision-making purposes. As organization embrace data warehousing technology as a means of accessing information, the need for quality information within a data warehouse is imperative to the sustained success and use of this technology. There have been several instances where poor quality of the data within a warehouse has led directly to the abandonment of this technology. By understanding the process flow of data from its source of origin through the various stages of manipulation and into the data warehouse, the potential for data errors can be mitigated.

While the quality of information within transactional systems must be addressed because it directly impacts the quality within a data warehouse, the process flow of data from source to target is of greater concern for information quality due to the various stages of data movement and manipulation. By developing a framework of information quality within a data warehouse, issues with data quality can be identified and addressed in a timely manner. The benefits of an established framework include: 1) establishing confidence that the data warehouse contains quality information, 2) identifying data issues from the source systems, 3) discovering changes in business or system processes that have not been reflected in the data transformation process, and 4) providing the group responsible for maintaining the data warehouse with the means of addressing user questions concerning data integrity.
The 6th International Conference on
A Framework for Information Quality in a Data Warehouse
Jonathan Wu
Co-Founder
BASE Consulting Group, Inc.

Presentation Abstract

Data warehousing technology provides integrated data from a multitude of sources that is non-volatile and transformed into meaningful information for decision-making purposes. As organizations embrace data warehousing technology as a means of accessing information, the need for quality information within a data warehouse is imperative to the sustained success and use of this technology. There have been several instances where poor quality of the data within a warehouse has led directly to the abandonment of this technology. By understanding the process flow of data from its source of origin through the various stages of manipulation and into the data warehouse, the potential for data errors can be mitigated.

While the quality of information within transactional systems must be addressed because it directly impacts the quality within a data warehouse, the process flow of data from source to target is of greater concern for information quality due to the various stages of data movement and manipulation. By developing a framework of information quality within a data warehouse, issues with data quality can be identified and addressed in a timely manner. The benefits of an established framework include: 1) establishing confidence that the data warehouse contains quality information, 2) identifying data issues from the source systems, 3) discovering changes in business or system processes that have not been reflected in the data transformation process, and 4) providing the group responsible for maintaining the data warehouse with the means of addressing user questions concerning data integrity.

Overview of Data Warehouse Process Flow

1. Migrating the data
2. Cleansing the data
3. Transforming the data
4. Loading the data warehouse
5. Reconciling the data warehouse

Source Systems
- Legacy Data Store
- ERP Application
- Custom Applications
- Legacy Systems
- Flat Files from External Sources

Data from Source Systems
Data Control Points

**Prevent Controls**
- Controls over the accuracy and completeness of data before it is loaded into the data warehouse.

![Diagram of data flow](Image)

Data Control Points

**Detect Controls**
- Controls over the accuracy and completeness of data at the completion of each stage or after it is loaded into the data warehouse.

Overview of Data Warehouse Process Flow

1. Migrating the data [Prevent Control]
   - Goal: Prevent meaningless information by not moving it.
   - Motto: “When in doubt, leave it out.”

2. Cleansing the data [Prevent Control]
   - Goal: Prevent unwanted redundant data by comparing and incorrect data by validating.
   - Motto: “Cleanliness is next to godliness.”

3. Transforming the data [Prevent Control]
   - Goal: Prevent meaningless data by transforming it.
   - Motto: “Business rules.”
Data Control Points

4. Loading the data warehouse [Prevent Control]
   Goal - Prevent unwanted data through conditions and filters.
   Motto - “If the data does not fit, you must omit.”

5. Reconciling the data warehouse [Detect Control]
   Goal - Detect data quantity and quality exceptions by reconciling.
   Motto - “If at first you don’t prevent, reconcile, reconcile, reconcile.”

Agenda

◆ Overview of Data Warehouse Process Flow
  - Data from Source Systems
  - Data Migration
  - Data Cleansing
  - Data Transformation
  - Loading the Data Warehouse
  - Reconciling the Data Warehouse

◆ Data Control Points as a Framework for Information Quality
  - Prevent Controls
  - Detect Controls

Summary

The success of a data warehouse rests with the users’ perceptions of it.

If the data is incorrect or incomplete, user confidence and use of the data warehouse will diminish.