

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 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.

The 6th International Conference on

A Framework for Information Quality in a Data Warehouse

Jonathan Wu
Co-Founder
BASE Consulting Group, Inc.

Version 2.10 29 August 2001

Information Quality

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.

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 2

Presentation Information

- ◆ **Author:** Jonathan Wu
- ◆ **Organization:** BASE Consulting Group
- ◆ **Presentation Title:** A Framework for Information Quality in a Data Warehouse
- ◆ **Contact Information**
 - E-mail: jwu@baseconsulting.com
 - Phone: (510) 628-3300 x224

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 3

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**

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 4

Overview of Data Warehouse Process Flow

The diagram illustrates the data flow from Source Systems (represented by various colored cylinders) to a Data Warehouse (represented by a large blue cylinder). The process involves a Staging Area (2) and Data Warehouse Tables (3). The steps are: 1. Migrating the data, 2. Cleansing the data, 3. Transforming the data, 4. Loading the data warehouse, and 5. Reconciling the data warehouse.

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 5

Data from Source Systems

The diagram shows five types of source systems: Legacy Data Store (purple cylinder), ERP Application (red cylinder), Custom Applications (cyan cylinder), Legacy Systems (light blue cylinder), and Flat Files from External Sources (orange folder icon).

Source Systems

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 6

Data Migration

Data Migration

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 7

Data Cleansing

Customer Information

FIRST NAME	LAST NAME	COMPANY NAME	AREA CODE	PHONE
JIM	Kirk	ibm	212	5551212
Jim	Kirk	ibm	212	5551212
James	Kirk	IBM	212	5551212
JAMES	KIRK	ENTERPRISE	212	5551212
↓	↓	↓	↓	↓
Martin	Zweig	Zweig Funds	415	5551212

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 8

Data Transformation

CUSTOMER_CONTACTS

FIRST_NAME
LAST_NAME
AREA_CODE
PHONE
ADDRESS1
ADDRESS2
ADDRESS3
CITY
STATE
ZIP_CODE

CONTACT_INFORMATION

FULL_NAME
PHONE
ADDRESS
CITY
STATE
ZIP_CODE

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 9

Loading the Data Warehouse

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 10

Reconciliation Process

Company XYZ						
Statement of Comparison between the Data Warehouse (DW) and Source Systems (SS)						
As of August 29, 2001 8:00:58						
TABLE NAME / COLUMN NAME	NUMBER OF ROWS			CONTROL TOTALS		
	DW	SS	DIFFERENCE	DW	SS	DIFFERENCE
GL ACTUAL BALANCES PERIOD_YEAR	980,029	980,029	-	1,856,732,942.00	1,856,732,942.00	-
GL BUDGET BALANCES PERIOD_YEAR	1,053,906.00	1,053,906.00	-	2,104,023,966.00	2,104,023,966.00	-
GL JOURNALS CREDIT	8,350,661.00	8,358,714.00	(8,053.00)	304,016,413,952.71	309,236,482,066.29	(5,220,068,113.58)
GL_CODE_PERIODS CHART_OF_ACCOUNT_NUM	140,271.00	140,271.00	-	14,167,371.00	14,167,371.00	-

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 11

Agenda

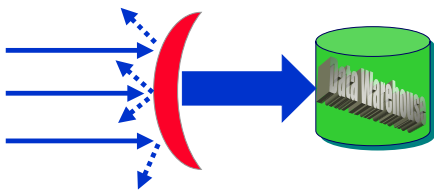
- ◆ Overview of Data Warehouse Process Flow
 - Data from Source Systems
 - Data Migration
 - Data Cleansing
 - Data Transformation
 - Loading the Data Warehouse
 - Reconciling the Data Warehouse
- ◆ Data Control Points as a Framework for Information Quality
 - Prevent Controls
 - Detect Controls
- ◆ Summary

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 12

Data Control Points

◆ **Prevent Controls**

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




Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 13

Data Control Points

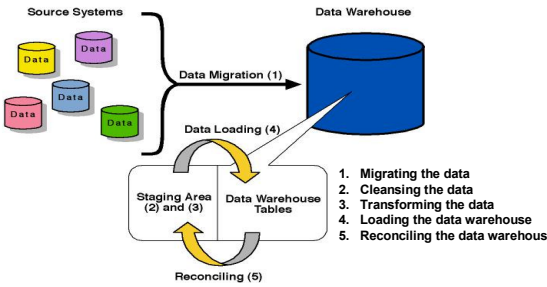
◆ **Detect Controls**

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



Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 14

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

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 15

Data Control Points

1. Migrating the data [Prevent Control]

Goal - Prevent meaningless information by not moving it.

Motto - *"When in doubt, leave it out."*

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 16

Data Control Points

2. Cleansing the data [Prevent Control]

Goal - Prevent unwanted redundant data by comparing and incorrect data by validating.

Motto - *"Cleanliness is next to godliness."*

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 17

Data Control Points

3. Transforming the data [Prevent Control]

Goal - Prevent meaningless data by transforming it.

Motto - *"Business rules."*

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 18

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."*

Data Control Points

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

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.