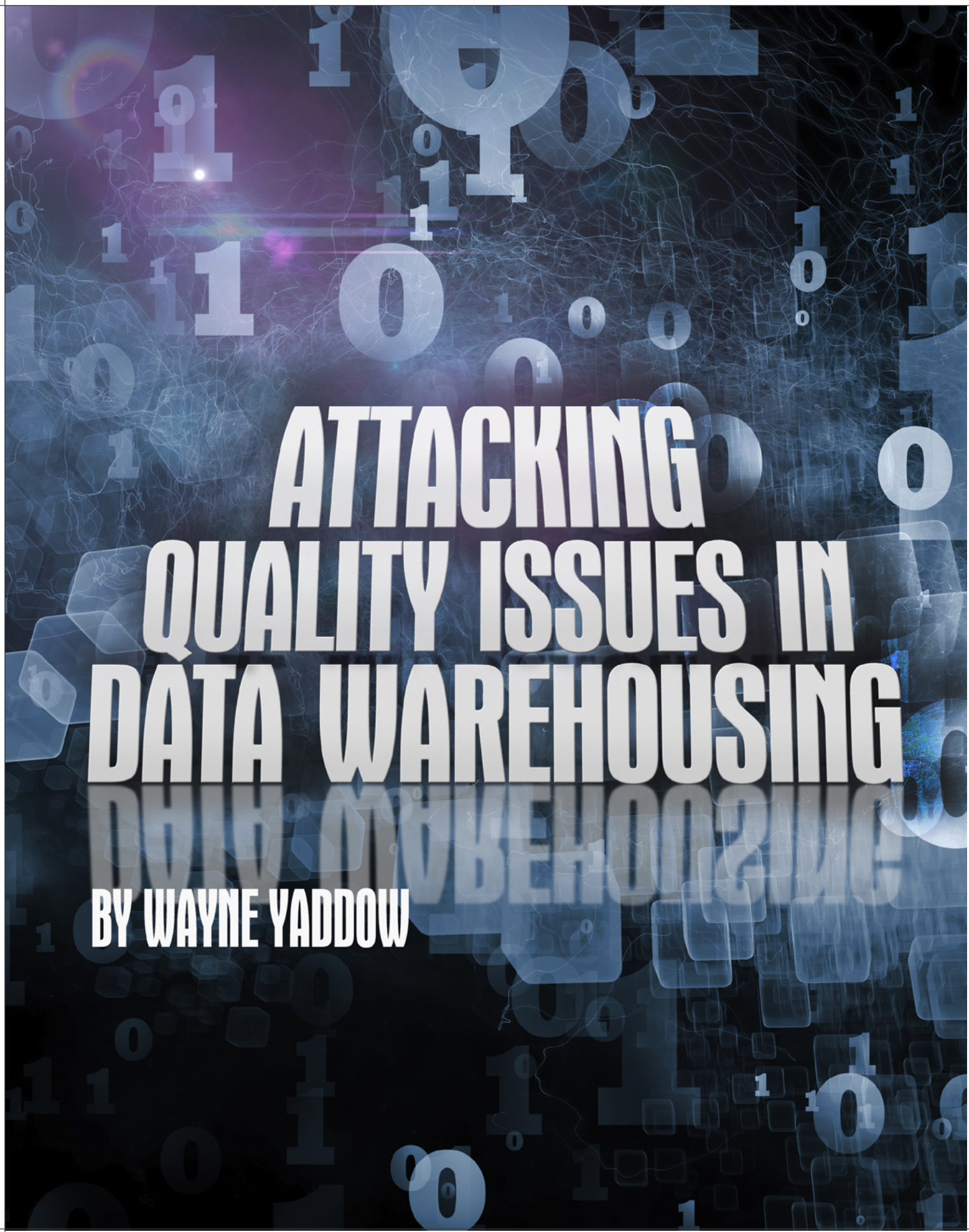

ATTACKING QUALITY ISSUES IN DATA WAREHOUSING

BY WAYNE YADDOW

Implementation and growth of data warehouses continue to gain attraction as organizations become more aware of the benefits of decision and analytic-oriented databases. Nevertheless, there is often one important obstacle to the rapid development of commercial data warehouses: data quality. Serious problems are often discovered when planning and populating a warehouse that, if not resolved, can delay or eventually result in terminating the project.

During the past twenty years, researchers have contributed to the understanding of data quality issues, yet little research has been collectively compiled to identify root causes of data quality problems that occur throughout major phases of data warehousing.

Based on my experience, the following are primary causes of data quality defects in data warehousing [1]:

- Flaws in the data warehouse modeling and schema design
- Defects in data sources used as input to the data warehouse
- Failure to effectively profile source and target data
- Weaknesses in the design and implementation process for data warehouse staging and extract, transform, and load processes

Using Early-Phase Defect Prevention Methods

This article highlights the reasons for data deficiencies related to the root causes listed above together with timely quality assurance efforts that can be implemented for discovery and correction. It is hoped that data warehouse designers, developers, and testers cooperate and benefit by examining these quality issues before moving forward with data integration into the data warehouse.

Figure 1 displays a high-level view of the common data warehousing extract, transform, and load process zwhere data quality and functional testing are recommended.

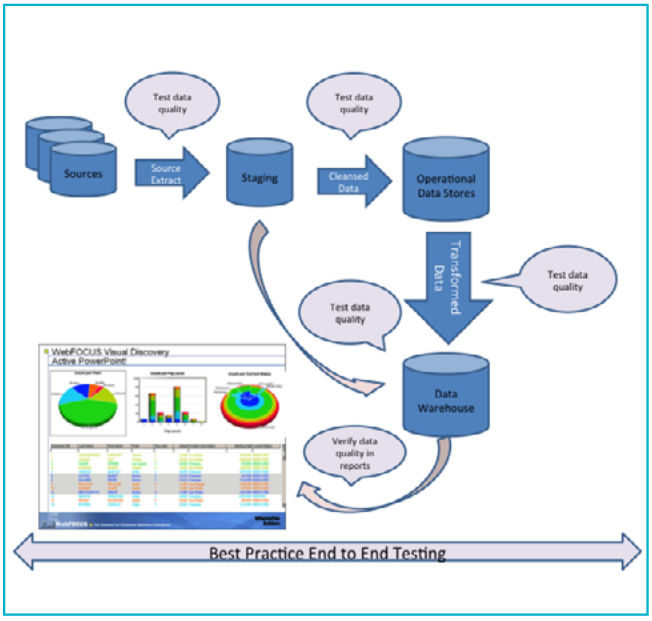

Figure 1: Data warehouse flow and recommended phases for data quality testing

Data Quality Issues Related to Data Model and Schema Design

Design of the data model for the data warehouse greatly influences the quality of the analysis by programs that use the data. A flawed schema will negatively impact information quality.

Data modeling is the process used to define and analyze data requirements needed to support business processes within the scope of application needs. The data modeling process should involve trained data modelers working closely with business stakeholders, developers, quality assurance, and potential users of the information system. Data modeling defines not only data elements, but also their structures and the relationships between them.

Data modeling methodologies should be used to model data in a standard, consistent, and predictable manner in order to manage the data as a resource. The use of modern data modeling standards and tools is strongly recommended for all projects.

Table 1 shows highlights of how a proper data warehouse design review can make or break your data warehouse. [2]

Quality Issues in the Data Warehouse Source Data

A leading cause of data warehousing and business intelligence project failures is finding and then loading incorrect or poor-quality data. The source system often consists of transaction and production raw data, which is where the details are pulled from and made suitable for the data warehouse. Each of these data sources usually has its own diverse methods of storing data, which may contribute to data quality problems if proper care is not taken.

Data warehouse environments provide the source of information used by business units to make strategic decisions. However, much of that data is created outside the warehouse. That means data quality problems can originate at the source and can therefore persist due to faulty data acquisition and delivery processes, or interpretation and transformation glitches.

Data quality problems in source systems need to be recognized as requiring mitigation. This can be accomplished by either addressing these problems as defects or by getting approval by stakeholders that these issues are acceptable. The QA team must then ensure that data warehouse users are aware of these data quality deficiencies in cases where they are not fixed before being loaded into the data warehouse.

Under certain conditions, source files are the product of multiple file consolidations. Consolidated files can, in turn, result in data quality being compromised before being loaded into the data warehouse staging area. Table 2 summarizes a few other possible causes of data quality issues as data sources are staged into the warehouse.

Other reasons for data pollution issues in the data warehouse may be cases where data was never being fully captured by source systems, the use of heterogeneous system integra-

	Questions during data warehouse modeling and schema design	Impacts and data quality risk mitigation issues
1	Have the major data warehouse dimensions been broken down into lower levels of detail? • Have the keys been identified? • Have the attributes been identified? • Have the keys and attributes been grouped together? • Have the relationships between groupings of data been identified? • Have the time variances of each group been identified?	There needs to be a data model that serves as the intellectual heart of the data warehouse environment. The data model normally has three levels: a high-level model where entities and relationships are identified; a mid-level model where keys, attributes, and relationships are identified; and a low-level model where detailed database design can be accomplished. While not all of the data needs to be modeled down to the lowest detail in order for the data warehouse environment to be built, at least the high-level model must be complete.
2	Have levels of data granularity been defined? A high level? Low level? Multiple levels? • Will rolling summarization be done? • Will there be a level of true archival data?	An important design issue in the data warehouse environment is that of granularity of data and the possibility of multiple levels of granularity.

Table 1: A small sample of a data modeling and schema development checklist

	QA checks for source data quality problems	Impacts and risk mitigation efforts
1	Insufficient data source selection and profiling of candidate warehouse data may cause data quality problems (i.e., source data that does not comply to business rules). • An age field contains values greater than 150 years • ZIP codes contain non-numeric values for U.S. addresses	For each required data source field, testers should run queries or profiling tools to verify that the data complies with business rules.
2	Data warehouse business analysts may possess insufficient knowledge of interdependencies among data sources that are used to populate the warehouse.	Review the quality of source data entry processes, loads, and merges used to create each data source. This effort may also include the establishment and monitoring of service level agreements, communication protocols with data suppliers, and data quality assurance policies.

Table 2: Checking for source data quality issues

tions, and a failure to have an adequate policy for data warehouse project planning.

Discovering Data Quality Issues Using Data Profiling Techniques

When potential data sources are identified and then finalized and agreed to, data profiling should be implemented immediately on that source data. Data profiling is the examination and assessment of your source systems' data quality, integrity, and consistency—sometimes known as source systems analysis. [3]

As important as data profiling is, it is often ignored and, as a result, data warehouse quality can be significantly compromised.

The data quality assurance analyst supports an organization's data quality initiatives by analyzing data. Profiling is the primary method for performing a data quality assessment. Data profiling is also used to quantify the extent of problems surfaced by other means and to measure the impact that data quality remedies have had.

Listed below are examples of problems that are easily uncovered through data profiling:

- Data fields used for purposes other than expected
- Fields that contain no data for any record
- Missing values when a field is defined as NOT NULL
- Violations of business rules

Business analysts can determine problem root causes during data analysis that could result in a substantial number of data quality problems that need to be corrected.

At the beginning of a data warehouse project and as soon as potential data sources are identified, data profiling assessments should be conducted to prepare for a go/no-go decision about proceeding with the project.

Table 3 depicts just a few of the possible causes of data

	Data quality issues discovered through data profiling	Impacts and risk mitigation efforts
1	Counts and types of distinct values in each field are not as expected.	Analyzing the number of distinct values within each field will help identify possible unique keys in the source data, which are referred to as natural keys. Identification of natural keys is a fundamental requirement for database and extract, transform, and load architecture, particularly when processing inserts and updates.
2	Counts of zero, blank, and NULL values in each field may reveal defects.	Analyzing each field for missing or unknown data helps you identify potential data issues. This information aids database and extract, transform, and load architects to set up appropriate default values or to allow NULLs on the target database fields where an unknown data element is unacceptable.

Table 3: Discovering data quality issues through data profiling

	Data quality problems discovered during extract, transform, and load phases	Impacts and early risk mitigation efforts
1	Inadequate source-to-target data mapping preparation.	Data mapping documents should be developed and continually maintained throughout the project. Tools designed specifically for data mapping should be considered over the use of manually created spreadsheets.
2	Improper extraction of data to the required target fields.	Design of extract, transform, and load program logic should be reviewed before development has been completed to assure that data will be loaded correctly to each field in target tables.
3	Failure to generate data flow and data lineage documentation that depicts the extract, transform, and load process	Data integration in warehouses is highly complex for most projects. Assure that modern data flow and data model diagrams are used to illustrate the planned flows.

Table 4: Checks for issues in the extract, transform, and load process

quality degradation discovered at the profiling stage of data warehousing.

Data Quality Issues Discovered During Data Loading

An important design consideration is whether data cleansing should be conducted for each source input during the staging phase, during the extract, transform, and load process, or within the data warehouse. The data staging area is where "grooming" is often conducted on data after it is loaded from source systems. [4]

Data staging and the extract, transform, and load phases are considered to be the most crucial phases of data warehousing, where maximum responsibility for data quality efforts occurs. These are prime phases for validating data quality from sources or auditing and discovering data issues. There may be several reasons for data quality problems during the staging and extract, transform, and load phases. A few of those are listed in table 4.

When data quality problems are encountered while importing data into the data warehouse, there are four viable actions that can be taken: exclude the data, accept the data, correct the data, or insert a default value. These are some of the design decisions that must be faced while working to improve data quality in early phases of data warehouse projects.

In Summary

There are many causes of data quality problems that may be found throughout all phases of data warehouse development. Data quality issues have been classified and described in a way that should help data warehouse practitioners, implementers, and tool providers find and resolve these issues as they move forward with each phase of data warehousing. {end}

wyaddow@aol.com

Sticky Notes

Click here to read more at StickyMinds.com.

- References