P158-27

A Scorecard approach to improving Data Quality

Phil Nousak, Rob Phelps, PWC Consulting, Chapel Hill, NC

ABSTRACT

An ever-increasing number of strategic and tactical business decisions are being made from analyzing data gathered in Data Warehouses and Data Marts. Bad decisions, poorly performing predictive models and monetary losses result when data quality is not monitored. What you think you know about your organization, your customers, or your suppliers may be distorted by undependable data.

Data quality cannot be improved independently of the source or the context in which these data are used. Technology-only approaches are not sufficient to provide sustained data quality improvements. The road to data quality improvement involves several factors. These include a technical understanding of the source and data gathering processes. It also includes the establishment of a reporting process to monitor changes in data quality as well as people with well-defined roles, responsibilities, and authority to develop a culture that supports data quality improvement.

This paper will describe a scorecard-based approach to identify, measure and monitor data quality. It will cover the people and processes needed to sustain such an effort, as well as an implementation using SAS software to build the technical infrastructure. Although the focus for the paper is on data quality assessment in a data warehouse, much of the approach can also be implemented outside of a formal system.

INTRODUCTION

DATA QUALITY BUSINESS ISSUES

The PricewaterhouseCoopers LLP Global Risk Management Solutions Data Management Survey 2001 sampled a broad mix of major 'Top 500' corporations, middle-market businesses, and companies primarily engaged in e-business.

Results from this survey show that over 75% of Chief Information Officers, IT directors or equivalent executives at 600 companies across the US, Australia and UK reported having experienced significant problems because of defective data.

The survey also shows that poor data quality causes hard dollar loss, failed securities deliveries, missed corporate actions, or erroneous trading decisions. Data quality issues may result in the following:

- Extra costs to prepare reconciliations
- A delay or scrapping of a new system implementation
- A failure to bill or collect receivables
- Inability to deliver orders or lost sales because of incorrect stock records
- Failure to meet a significant contractual requirement or service level performance

The following are some of the areas related to Human Resource (HR) systems that are adversely affected by poor data quality.

- ERP Conversion
- Outsourced HR programs such as Pension Plan Administration
- Employee Workforce Planning
- Globalization and /or Integration of Business Operations
- Mergers, Acquisitions, Divestitures and Reorganizations
- Government Reporting
- Labor Negotiations

As e-business becomes more pervasive, the rising exposure to poor data quality increases the risk of incurring greater internal costs as well as costs to on-line commercial relationships. Investors are becoming increasingly sensitive to data problems as a sign of a deep malaise at the core of an organization. As reporting methods expand across non-financial areas in support of strategic balanced scorecard management models, the reliability of all kinds of data will come under growing scrutiny.

DATA QUALITY BUSINESS STRATEGY

Companies must take a strategic view of ensuring quality data. This includes a process for monitoring and correction of data quality issues supported by sound and demonstrable data metrics.

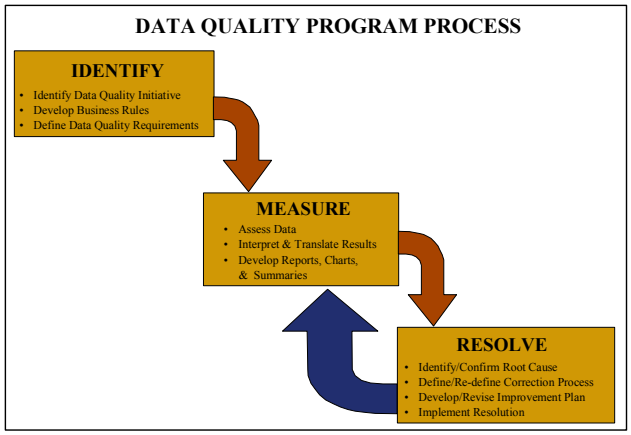
This paper represents a holistic approach to the ongoing issues related to organizational data management. It begins with an overview of a fundamental process as the basis for the establishment of a program for improving data quality over time. Next, an approach for the establishment of a Data Quality Program (DQP) is presented with a discussion of the types of errors found in any collection of information and a primer describing the process for quantifying systematic errors. A description of the metrics and reports needed to systematize the DQP follow, as well as a section describing the roles and responsibilities for people to support a data quality effort. The final section includes a technical framework with consideration for implementation using SAS software.

OVERVIEW OF A DATA QUALITY PROGRAM

The Data Quality Program (DQP) is a single point of reference for addressing issues affecting data quality in an organization or business unit. It provides a forum representing all points of view within the Information

Systems community in defining, identifying, measuring, analyzing, and resolving data quality issues. The DQP provides a foundation for making decisions and provides direction throughout the quality management process. Figure 1 provides an overview of the data quality program process.

FIGURE 1



IDENTIFICATION

The first step is to identify data quality improvement opportunities and define business rules and data quality requirements. It is important to re-visit the identification phase, as often as new data quality measurement needs are determined.

In the Identification phase, the DQP analysis files are built from extracts of source system data or from the Operational Data Store (ODS) in a data warehouse. These data, including tables, records and elements, are extracted in their "native" form and are made available to a reporting tool before any transformation process has occurred. Business rules are used to monitor and report data quality problems. In the DQP, these rules are called filters. Examples of business rules that can be verified with filters include:

- A salary change date that cannot be earlier than the hire date
- A salary that cannot be less than the minimum wage
- An employee who must be 15 years old before his/her date of hire
- Social Security numbers that are not numeric
- A last shipped date that is less that the last ordered date

Suspect records that do not meet these data quality requirements (e.g., consistent, valid and complete) are then identified. The records are presented in detailed reports and used to provide the information to point to the root cause of the DQ issues.

MEASURE

This step involves the application of data quality metrics to data attributes or records from data loaded into a data warehouse or other system. These metrics are translated into business terminology. The measures are communicated by the creation of detailed reports, summaries, trend analyses, and scorecards that portray data quality levels.

RESOLVE

Understanding of the root cause of these quality problems are needed before resolution can occur. After identifying and measuring quality issues the next step is to formulate a correction process with business case justifications. This involves the development of a task schedule and assigning responsibility to execute the correction process.

The correction steps provide the ability to confirm (or modify) the root cause, re-define the correction process, and incorporate modifications into an improvement plan. It will be necessary to repeat the measurement and resolution phases for the data quality improvement process to ensure quality maintenance.

The resolution part of a DQP depends on the coordination of the individuals responsible for the data quality review. Described in a latter section are the roles and responsibilities for staffing such a program.

In the Resolution phase, the DQP Business Analyst collects the DQ detail reports and distributes them to the source data owners as appropriate. The automation of this process happens in the latter stages of deployment.

After reviewing the reports, the source data owner can:

- Correct the data in the source system, or
- Recommend modifications or additions to the filter list, or
- Confirm that the suspect data is acceptable and document its occurrence

.

AN APPROACH TO MANAGING DATA QUALITY

TYPE OF ERRORS

Problems or errors in data can occur either randomly or systematically. Systematic errors often occur as a result of a misapplication or misinterpretation of business rules. Systematic errors can be dealt with in a number of ways, including modifying the data collection process, introducing a systematic correction method, or simply reporting the inconsistencies. Random errors, in contrast, require direct review of input records and are not resolvable by systematic means.

CHECKING FOR ERRORS

One way to check for problems in data is to use filters to identify suspect records. The records are suspect for identification and classification purposes. Later steps are taken to either correct the information identified by filters or to document any discrepancy.

Filters represent simple conditions that identify a single issue with the data analyzed. The types of filters range from the simplest checks of the domain or range of a single variable, through comparisons of one or more fields within or between records in a table, to a complex review of multiple fields across multiple tables and source systems.

The following outline presents types of filters that may be prepared for a database. The outline presents the filters in order of increasing complexity, from simpler within-record checks to the more complex between-table and between-system checks. The key to building a sound DQP is to begin with the simpler filters within a system and build to the more complex filters among systems.

Within variable

- value in domain (e.g., categorical variable matches reference list; numeric variable in range)
- test for missing value when appropriate (e.g., table key, required variable)
- valid date

Between variables

- cross-check in domain (e.g., salary increase matches compensation transaction code)
- test for logical relationships

Between records

- unique key when appropriate
- proper sequence (e.g., employment activity only before termination)
- missing records (e.g., apparent pay change record not in table)
- proper transaction variable assignment (e.g., comparison of selected variables between records meets business rules requirements for transaction)
- comparison of values consistent (e.g., comparison of base rate between records does not match indicated pay change)

Between tables

- check key relationship across tables (e.g., list of expected employ ids matches appropriately across tables with employee data)
- cross-check in domain (similar to within table check)
- test for logical relationships (e.g., observation in separate evaluation table matches employee work history sequence)
- valid date relationships (e.g., separate work history tables join to produce acceptable representation of employee's work history)

Between systems

- check key relationships across tables between systems
- cross-check in domain
- test for logical relationships
- valid date relationships

DATA QUALITY ACTIVITIES AND PROCESSES

The DQP activities identify where quality problems exist, ascertain the magnitude of the problems, and propose solutions. These activities require the DQP to regularly measure data quality levels, provide mechanisms to share data quality information, and maintain organizational accountability for data quality. The activities identified in this section occur throughout the development, enhancement, and maintenance of the DQP life cycle.

PROCESSES

The following is an outline of the process steps necessary for the establishment of a DQP:

- Develop detailed procedures that provide a logical, organized approach to addressing and resolving data quality issues using the high-level process depicted in Figure 1. The DQ process must be scaleable for use on small or large quality problems.
- Develop data quality metrics that measures the level of data quality in information systems monitored.
- Define procedures that apply the quality metrics to the data for each information system in order to monitor its data quality level over time.
- Identify procedures for communicating DQP activities.
- Develop a Data Ownership Policy and define procedures for data owners to perform data quality functions. Clear identification of data and process owners and definition of their responsibilities will facilitate an effective DQP.

ACTIVITIES

The following are the activities needed to establish the DQP processes:

- Identify or verify the authoritative sources of data and assign the necessary data ownership responsibilities.
- Define data quality measurement criteria.
- Propose recommendations to improve the level of data quality based on results of data reviews.
- Develop goals, objectives, plans, and tasks for data quality improvement activities.
- Maintain data quality for current Information Systems by establishing and communicating procedures to personnel whose job function it is to create, update, and delete data .
- Maintain data quality for current Information Systems by developing edit and validation rules that filter the data before storing it in the database, if applicable.
- Report the status and results of data quality improvement activities to the necessary personnel, e.g., Upper Management, Information Systems Management, application support personnel.
- Regularly publish data quality level information for identified Information Systems to audiences at the necessary organizational levels.
- Manage and control DQP work products.
- Coordinate the necessary data quality improvement tasks with the appropriate data and process owners and application support teams.

DATA QUALITY METRICS

Data quality metrics are defined and categorized into two groups. *Generic* metrics apply to all columns and/or all tables (e.g., a record count metric, such as the number of records in the tables). *Specific* metrics apply to specific columns, or combinations of columns, in specific tables (e.g., an accuracy metric, such as the number of valid values in a table).

DEFINING DQ METRICS

The following are principles to consider when defining data quality metrics:

- Metrics should be insensitive to changes in the number of records in the warehouse;
- Metrics should accurately reflect the degree to which the data meets the associated data quality need;
- Metrics should be independent of each other, so that no two metrics are actually measuring the same effect; and
- The number of metrics chosen should be kept to a reasonable number, as too many metrics can often confuse rather than clarify.

DATA QUALITY CATEGORIES

We classify Data quality metrics into categories that describe the methods used to analyze quality of data. A data value is generally accepted as having high quality if it meets the appropriate combination of the categories that are applicable to the element. Following in Figure 2 are the categories used to group the measures of data quality.

The Data Element Quality Scorecard contains the measured level of quality for each data element in the DQP. The scorecard lists data quality categories as columns and data elements as rows. Single summary statistics are created for four types of cells on the worksheet:

- Data element and data quality category combination (e.g., Field1/Valid);
- Quality category across all data elements (e.g., Valid for Field1 –> Field4);
- Data element for all quality categories scored (e.g., Field1: Valid, Unique, Complete etc. - some categories may not be scored); and
- Overall data quality including all data elements and quality categories scored.

FIGURE 2

Name	Description	Example
Valid	Data element passes all edits for acceptability	A Person record has a Name that contains numbers
Complete	Data element is (1) always required or (2) required based on the condition of another data element	A Payroll record misses a value for Person
Consistent	Data element is free from variation and contradiction based on the condition of another data element	A New Hire record has a Hire Date before their birth date Leave of Absence is checked, but the employee is at work
Unique	Data element is unique—there are no duplicate values	Two Person records have the same Social Security Number
Timely	Data element represents the most current information resulting from the output of a business event	A New Hire record references an Organization that has been sold
Accurate	Data element values are properly assigned	An HR Organization record has an inaccurate or invalid hierarchy
Precise	Data element is used only for its intended purpose, i.e., the degree to which the data characteristics are well understood and correctly utilized	HR Organization Department codes are used for different organizational entities between different records

The summary statistics calculation uses mathematical principles and formulas that allow for uniquely categorizing data quality problems. Both row and column totals in all cases are equal to or less than the individual cell totals. This is because any one record may have encountered multiple filters that identified suspect information. Filters must not encounter any suspect fields in order for the record to have complete quality data. The formula for calculating cell total is described in a latter section.

DATA QUALITY REPORTING

The DQP produces a set of reports for different purposes.

Detail Field Suspect Report. The report lists records that are suspect for each filter. It is created to help the Information System representatives or service centers examine the suspect records and make corrections when appropriate.

Summary Filter Suspect Report. The report lists the description and summary of suspect records by filter in the current run of the DQP. It is useful for the application developers to examine the appropriateness of the filters they have implemented.

Summary Field Suspect Report. The Summary Field Report presents the number of suspect records in two dimensions: data elements in row and data quality categories in column. A data element is a collection of one or more source fields in the system. For example, employee's date of birth is a data element, while the employee's name may contain three fields in the source system, the first name, the last name and the middle name initial. Individual cells in the report contain the count of suspect records that fall into the data element and data quality category.

Since a data element may contain more than one source

field, and a data quality category may contain more than one filter, a cell in the report may involve multiple filters. The DQP incorporates an algorithm that prevents a multiple count of the same record for each cell, including the row and column totals. Therefore, a row total is the number of records that do not meet one or more data quality criteria for the data element. On the other hand, the column total is the number of records for which one or more data elements do not meet the data quality criteria in that data quality category.

Finally, the total at the lower right corner is the total number of records that do not meet DQP criteria in at least one of the data elements that is audited. As with the *Trend Analysis,* the total number of a row or a column or the whole table is the number of suspect records, not the number of errors by filter or field. Therefore, the total is not the simple arithmetic sum of its elements. This report is illustrated in Figure 3a.

Figure 3a

DW Data Quality Field Report Records Processed: 9,999								
Element	Data Element Quality Category (# Suspect Records)							Total
Name	Valid	Unique	Complete	Consistent	T..	A..	P..	
Field1	0	0	0	N/A	N/A	N/A	N/A	**0**
Field2	259	N/A	176	N/A				**260**
Field3	0	2	228	N/A				**228**
Field4	720	N/A	604	4				**720**
Total	**720**	**2**	**1000**	**4**				**1000**

Data Quality Scorecard.
The DQP scorecard converts the Summary Field Suspect Report into a percentage of records that pass all the data criteria set in the DQP, i.e., cell percentages are calculated according to the following formula:

Percentage = [1 – (Cell count in the worksheet) / (Total number of records)] × 100%.

An example of a Data Quality Scorecard is shown in Figure 3b.

Figure 3b

HRDW Data Quality Scorecard Records Processed: 9,999								
Element	Data Element Quality Category (% Quality Data)							Total
Name	Valid	Unique	Complete	Consistent	T..	A..	P..	
Field1	100.00%	100.00%	100.00%	N/A	N/A	N/A	N/A	**100.00%**
Field2	97.40%	N/A	98.24%	N/A				**97.40%**
Field3	100.00%	99.99%	97.72%	N/A				**97.72%**
Field4	92.80%	N/A	93.95%	99.96%				**92.80%**
Total	**92.80%**	**99.99%**	**90.00%**	**99.96%**				**90.00%**

Trend Analysis. The report presents the number of suspect records by field for the current and previous runs of the DQP. It presents to the Information System representatives or service centers the progress of the quality of the data for which they are responsible.

DATA QUALITY STAFFING

Collectively, the members of the DQP are responsible for both the data quality management process and data integrity in the core systems. The job descriptions that follow are not necessarily full time positions but represent the roles and responsibilities for various aspects of a DQP. As the number of systems monitored increase and the DQP is better established, the time commitment for these activities will vary.

The members of the DQP must have well-defined roles, responsibilities, and authority to successfully improve the quality of the data in the organization. The DQP may have individuals serving one or many roles. It is important to understand that the creation of a separate DQP group is an option, but not required.

Each member of the DQP is assigned to at least one of the roles listed below:

- Data Warehouse Data Quality Manager
- Information Systems Manager
- Data Warehouse Data Quality Personnel
- Application Developers
- Data Users / Data Producers
- Business Unit Functional Experts

The following section contains a description of each role and defines the function or purpose of the role in the following two scenarios: day-to-day data quality operations and DQP meetings, forums, or activities. The responsibility description attaches accountability and authority to a role. Each role will list the skills required of an individual to perform the job on the DQP.

DATA QUALITY MANAGER

Daily Data Quality Role: Gains economic support for the data quality management process, defines the budget, and monitors the development and maintenance schedules.

DQP Role: Serves as a full-time member by participating in every DQP meeting in the capacity of ensuring adequate resources and funding for performing data quality improvement and maintenance activities.

Skills: Project management, data quality concepts, and knowledge about the Business processes that produce the data.

Responsibility: Keeps the data quality management initiatives on-track by being on schedule and within budget. Data Quality Personnel and Developers should receive direction form the Data Quality Manager as the scope and focus of the DQP evolves.

INFORMATION SYSTEMS MANAGER

Daily Data Quality Role: Gains economic support for system projects, defines the budget, and monitors system development / maintenance schedules. Implements data

administration policies, procedures, criteria and standards for information systems data. Reviews and reconciles Information Systems data models at logical levels of detail, if applicable. Implements procedures for communicating data requirements and resolution activities among all relevant personnel.

DQP Role: Serves as a part-time member by participating in selective DQP meetings in the capacity of ensuring adequate resources and funding are available to support data quality improvement and maintenance activities.

Skills: Project management, systems life-cycle development, systems integration, and knowledge about the Business processes that produce the data.

Responsibility: Ensures data quality improvement and maintenance tasks are included and tracked in the DQP project plans.

DATA QUALITY PERSONNEL

Daily Data Quality Role: Ensures all data quality management tasks are performed in sequence and in an expeditious manner. Identifies or captures data quality problems, defines data quality requirements, assesses the specified data, and develops reports, charts, and summaries that depict the data quality status of the Data Warehouse. Designs automated components and manual processes that will identify, measure and communicate the quality of the data from the source systems in the data warehouse. Supports the analysis of source data, extract/transform. Measures the data quality levels at regular intervals to monitor data improvement activities.

DQP Role: Serve as full-time members by participating in every DQP meeting in the capacity of providing information both on the status of data quality issues and specifics about each data quality issue. Presents the root cause of poor data quality and recommends methods to improve the data quality levels.

One member is responsible for facilitating every DQP meeting to ensure that the meeting objectives are met and manages data quality improvement and maintenance activities.

Searches for the causes of poor data quality and resolves incompatibilities between the systems.

Skills: Business area process knowledge, database management, system life-cycle development, data quality metrics, and basic programming logic.

Responsibility: Institutionalize the data quality management process and knowledgeable about data quality issues.

APPLICATION DEVELOPER

Daily Data Quality Role: The role is typically a role performed by those who develop and maintain the data warehouse and system(s) that capture data that feed the data warehouse. A system analyst or programmer manages the physical database of his/her respective system.

DQP Role: A system analyst or programmer from each project team serves as an invited member to participate in DQP meetings or activities depending on the data quality issue or the type of information that is planned to be improved. Communicates information about the data structure and content in their respective system.

Skills: Business process knowledge, database management, systems life-cycle development, structured programming.

Responsibility: Accountable for the integrity of the physical data structures and the extent that the physical database structure reflects the requirements provided by the business units. Uses standard data definitions when creating or modifying the system's database. Responsible for communicating system modifications to the DQP that affects the format or content of the data.

DATA USERS/DATA PRODUCERS

Daily Data Quality Role: A role played by every information system user who interacts with data in the selected Information Systems as part of their job function. This role includes Managers who are accountable for the business process, but may not directly interact with data in the systems. The personnel identify data quality issues and may participate in the data correction processes.

DQP Role: Serves as an invited member to participate in DQP meetings or activities depending on the data quality issue or the type of information that is to be improved, thereby representing their perspective of data usage. Understands the business process and how data are captured and maintained in the source system. Considered the customer of the data warehouse or source systems. These individuals set quality expectations for the DQP by defining requirements for data quality. They are involved in the data quality review and cleanup activities.

Skills: Knowledgeable about the business function and processes that produce the data.

Responsibility: Responsible for using data as originally intended and for notifying the DQP of data quality problems. Additionally, they are accountable for integrity of data within their job function.

FUNCTIONAL EXPERTS

Daily Data Quality Role: A role played by people who are considered expert in a particular business function. Provides the knowledge and expertise essential to developing and validating data checks, filters and business solutions.

DQP Role: Serves as an invited member to participate in

DQP meetings or activities depending on the data quality issue or the type of information planned for improvement. Understands the Data Users'/ Data Producers' data quality requirements. Also, provides solutions that enable data to be captured in a way that satisfies both the operational data quality requirements and the Data Users'/Data Producers' requirements. Collects and verifies business rules that ensure quality data and resolve issues or conflicts concerning data usage or interpretation.

Skills: Knowledgeable about the business environment, business function, and technology and processes that produce the data.

Responsibility: Responsible for defining and validating the names and definitions of data elements within their subject of expertise to meet the Data Users'/Data Producers' needs. Accountable for the integrity of data definitions and ensuring that the Data Users'/Data Producers' maintain data values that are in accordance with the definitions.

DQP TECHNICAL FRAMEWORK

SYSTEM REQUIREMENTS

A system for reporting Data Quality is incorporated into a production stream or used as a stand alone system to do periodic 'ad-hoc' reporting. The best approach is a modular one that allows for relatively easy modification and expansion to suit the needs of the warehouse developers, data owners, and other members of an organization that need to monitor data quality.

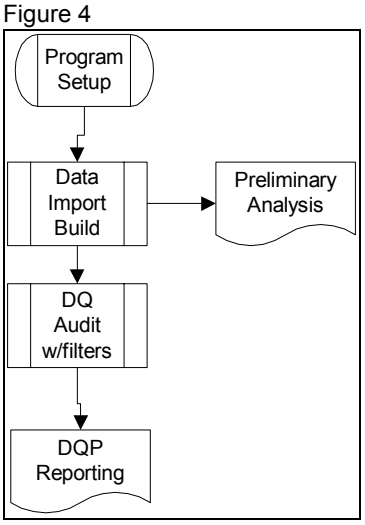
The following are key features of an application for Data Quality Reporting:

- Modular programming that is easy to maintain and is extendable.
- Portability and scalability to other platforms with access to extensive data preparation and analytic tools.
- Easily maintained support files that that can be modified to build the DQP database as well as activate or de-activate new filters.
- Self-documenting support files for identifying data quality issues and preparing the data for analysis.

DQP PROCESS FLOW

Figure 4 is a high-level diagram showing the process needed to evaluate data for Data Quality purposes. The preliminary analysis represents a profile of the characteristics of the data prior to applying 'business rules' for accessing data quality. The DQ Audit with filters represents a detailed reporting process that produces a scorecard and detailed reports for measuring and resolving data quality issues.

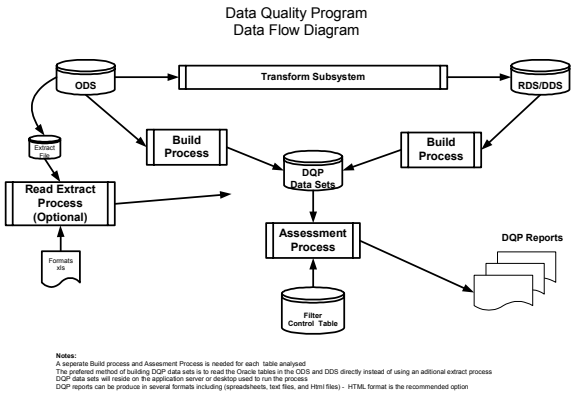
Figure 4



DQP DATA FLOW

Figure 5 is an example diagram showing the data flow in an automated system written using Base SAS® and Macro control language. Customization of this application occurs through entries in tables or spreadsheets. These tables specify the format of the input data and the SAS statements used to establish the filter conditions.

Figure 5



FILTER CREATION/REFINEMENT

Filters represent the 'Business Rules' used to measure the quality of the data in the DQP. These 'rules' form the basis of communication about the details of measuring quality data. A robust solution would include a central repository of these filters and an easy way to document and modify them. It should also provide the following:
Easy to add, activate, or de-activate filters.
Ability to implement a wide range of filters without having to create customized programming logic.
Flexibility to include filters that are more complex if the user is familiar with a programming language.
Ability to point to a diverse set of lookup tables for consistency and completeness checks.
Easy documentation of the filters that are active in the DQP at any time.
Most filters, in our experience, can be implemented with one or two SAS statements. Filters that are more

complex may be implemented using macros. This allows the actual SAS code used in the system to serve as documentation when discussing filter criteria with the business specialist involved in the DQP.

Filters may be thought of as program statements that result in logical binary fields (0,1). These fields are summed in order to determine the number of unique records with suspect data, and allow the creation of a Data Quality Scorecard. There can be multiple filters (with the same DQ category) applied to a single element of data. The presence of any condition that triggers a filter acts as a unique counter for identifying which records contain suspect data. The summation of these fields allows the calculation of overall data quality for the entire table.

Following are excerpts from a set of filters used to measure data quality. Included in Figure 6 is an example of fields contained in a filter control table. It is listed to demonstrate the functionally and flexibility of an automated solution.

Filters for Field=SSN Type=Character

#	Description	Cat	SAS Code
N1	Missing value	Com	If ssn=' '
N1	SSN is Invalid	Val	MACRO SSNCheck(substr(ssn,2,9))

Filters for Field=EMPLID Type=Character

#	Description	Cat	SAS Code
S1	Missing value	Com	if emplid=''
S2	Not unique	Uni	if not (first.emplid and last.emplid)
S3	Invalid - Length is not 6 or not in the range (0 - 999999)	Val	if length(emplid)^=6 or not (0<input(emplid,7.)<=999999)

Filters for Field=Hire Date Type=Date

#	Description	Cat	SAS Code
D1	Invalid - Hire date after May 1, 2001 or before Jan 1, 1930	Val	if hire_dt >mdy(5,1,2001) or hire_dt < mdy(1,1,30)
D2	Missing value	Com	if hire_dt=.
D3	Hire date after termination date, if available	Con	if termination_dt^=. and hire_dt > termination_dt
D4	Employment Hire date not consistent with Job date	Con	if hire_dt ^= PS_JOB.HIR_EFFDT

Filters for Field=Referral Source Type=Character

#	Description	Cat	SAS Code
R1	Value not in lookup table	Val	SD7: sample.LOOKUP
R2	Missing value	Com	if Referral Source=' '

Figure 6

Specifications for FILTERS control	
Field	*Specification*
System	System that the filter audits.
Table	Table that the filter audits.
Field	Field that the filter audits i.e. Character, Numeric, Date.
Type	Type of the field.
Filter CD (#)	Code together with System, Table and Field to uniquely identify the filter.
Filter Description	A text description of the filter
DQ Category	Data quality category of the filter. The categories include: Valid, Complete, Unique, Consistent etc. (See Figure 2)
Active	Indicates whether to include the filter in the current auditing process. Default: No
SAS Code	SAS codes used to define the filter. Code must be provided DQ categories other than Complete or Valid. The code can be one or more lines of SAS statements, or begin with following key words: For DQ category *Valid*, **LIST**: the list of valid codes, or **SD7** : a named SASFILE with valid codes, or **EXCEL**: a named Excel file with valid codes. For any DQ category, **MACRO**: the name of a SAS macro with its parameters.

IMPLEMENTATION PLANNING

Two major streams require coordination in the establishment of a Data Quality Program. These include the establishment of the staffing and process to carry out the DQP, and the technology for the implementation of an automated reporting environment.

A Data Quality Program, like a Data Warehouse, is never completely finished, but changes over time. The best practice for successful implementation includes IT acting as the custodians of the data, with the business units and the organization serving as the owners of quality and accuracy. With upper management attention, this creates a 'hands-on' focus where measuring and improving data quality is a corporate strategy. The DQP then functions to proactively diagnoses and resolves data management issues before they have an adverse effect.

The following are some of the pre-requisite steps needed to establish and sustain a Data Quality Program.

STAFFING AND PROCESS

- Confirm roles and responsibilities for DQP personnel
- Clarify role of DQP Application Support
- Identify personnel who will fill these roles.

- Confirm scope of information (tables and fields) to be included in 1st iteration of DQP
- Establish business rules and define initial filters for major data sources
- Decide on Reporting Frequency
- Perform initial assessment of these elements
- Refine business rules and filters and re-measure
- Establish Reporting Distribution Process

TECHNICAL CONSIDERATIONS

- Confirm Technical Options
- Obtain necessary software components
- Determine Size of Platform for Initial DQP
- Obtain Hardware platform if necessary or identify existing equipment for use
- Configure connectivity for data source access
- Configure programs and options for anticipated environment
- Initiate DQP Processes

CONCLUSION

Many companies are beginning to understand the value of data and the knowledge that it creates as one of their most fundamental assets. The greater prevalence of Data Warehousing is making data available for analysis for more areas of the organization. Few organizations have taken the initiative for making complete data management and quality control of information a strategic initiative.

Organizations need actionable metrics that allow them to quantify improvements for the systems that supply essential information. A Data Quality Scorecard used by all levels of management and the establishment of a Data Quality Program are the best approach to effect change and focus on the issues related to improving the value of information.

Organizations that make data management and improving data quality a strategic initiative benefit in several ways. These include reducing processing costs due to fewer reconciliation activities. Improved data quality provides for increasing sales through better prediction techniques and winning significant contracts through better analysis of data. Greater employee and customer satisfaction through careful and accurate attention to reliable sources of information benefits all levels of an organization.

REFERENCES

PricewaterhouseCoopers LLP, Global Data Management Survey 2001 The new economy is the data economy. 2001.
PricewaterhouseCoopers LLP, Human Resources Data Quality Program – Partner Briefing, March 2001.

ACKNOWLEDGMENTS

Information from PwC internally published documents provided by Mark Miller, Steven Yan, and John Ingersoll was used in the preparation of this paper. Renee Hansen provided additional input and assisted with editing.

TRADEMARK CITATION

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Phil Nousak – Phil.Nousak@us.pwcglobal.com
Rob Phelps – Robert.W.Phelps@us.pwcglobal.com