101

TDWI
REPORT SERIES

DATA QUALITY AND THE BOTTOM LINE

Achieving Business Success through a Commitment to High Quality Data

by Wayne W. Eckerson

THE DATA WAREHOUSING INSTITUTE™

Research Sponsors

Arkidata Corporation

DataFlux Corporation/SAS Institute

DataMentors, Inc.

Sagent Technology, Inc.

Vality Technology, Inc.

Acknowledgements

TDWI would like to thank many people who contributed to this report. First, our report sponsors diligently reviewed outlines, survey questions, and report drafts. Second, several individuals besides sponsors reviewed and commented on the manuscript: Larry English, David Wells, Sid Richardson, Pieter Mimno, and Louis Barton. Also, we would like to thank the many individuals—most of whom are quoted in the report—who talked with us about their approaches to data quality. Finally, we would like to recognize TDWI's production team: Denelle Hanlon, Marie McFarland, Donna Padian, and Dale Chinn.

DATA QUALITY AND THE BOTTOM LINE

Achieving Business Success through a Commitment to High Quality Data

by Wayne W. Eckerson

Table of Contents

Illustrations

About the Author

WAYNE W. ECKERSON is the Director of Education and Research for The Data Warehousing Institute, the leading provider of high-quality, in-depth education and research services to data warehousing and business intelligence professionals worldwide. Eckerson oversees TDWI's educational curriculum, member publications, and various research and consulting services.

Eckerson has written and spoken extensively on data warehousing and business intelligence since 1994. He has published in-depth reports about data marts, databases, online analytical processing (OLAP), meta data management, Web-based query tools, enterprise information portals, and customer relationship management applications and architectures.

Eckerson has also written about business intelligence for *Data Management Review*, *Computerworld*, the *Journal of Data Warehousing*, *DB2 Magazine*, *Enterprise Systems Journal*, *Application Development Trends*, and *Oracle Magazine*, and has been quoted extensively in a variety of business and industry magazines.

In addition, Eckerson has delivered presentations at industry conferences, users group meetings, and vendor seminars. He has also consulted with many large vendor and user firms.

Prior to joining TDWI, Eckerson was a senior consultant at the Patricia Seybold Group, and Director of the Group's *Business Intelligence & Data Warehouse Service*, which he launched in 1996.

Eckerson has a B.A. in American Studies from Williams College and an M.A.L.S. in literature from Wesleyan University. Eckerson lives and works in the coastal town of Hingham, MA, with his wife and two children.

About the TDWI Report Series

This series is designed to educate technical and business professionals about new business intelligence technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and leading-edge user companies, and is supplemented by a survey of business intelligence professionals.

To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving business intelligence problems or to an emerging technology discipline. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical business intelligence issues. *Please contact Wayne Eckerson at weckerson@dw-institute.com if you would like to suggest a topic that meets these requirements.*

For more information about this report or its sponsors, please visit www.dw-institute.com/dqreport/.

Executive Summary

As we enter the 21st century, we are at the dawn of the Information Age. Data and information are now as vital to an organization's well being and future success as oxygen is to humans. And without a fresh supply of clean, unpolluted data, companies will struggle to survive.

The Data Warehousing Institute estimates that data quality problems cost U.S. businesses more than $600 billion a year. Yet, most executives are oblivious to the data quality lacerations that are slowly bleeding their companies to death. More injurious than the unnecessary printing, postage, and staffing costs is the slow but steady erosion of an organization's credibility among customers and suppliers, as well as its inability to make sound decisions based on accurate information.

High quality data is critical to success in the Information Age.

The problem with data is that its quality quickly degenerates over time. Experts say 2 percent of records in a customer file become obsolete in one month because customers die, divorce, marry, and move. In addition, data entry errors, systems migrations, and changes to source systems, among other things, generate bucket loads of errors. More perniciously, as organizations fragment into different divisions and units, interpretations of data elements mutate to meet the local business needs. A data element that one individual finds valuable may be nonsense to an individual in a different group.

Fortunately, new strategic initiatives, such as CRM, business intelligence, and supply chain management are sounding a wake-up call to top executives. Many are learning the hard way that data quality problems can sabotage the best laid strategies and expose errors to a much broader, and critical, external audience.

The Goal Is Achievable. The good news is that achieving high quality data is not beyond the means of any company. The keys are to treat data as a strategic corporate resource; develop a program for managing data quality with a commitment from the top; and hire, train, or outsource experienced data quality professionals to oversee and carry out the program. Then, it is critical for organizations to sustain a commitment to managing data quality over time and adjust monitoring and cleansing processes to changes in the business and underlying systems.

Achieving high quality data is not beyond the means of any company.

Commercial data quality tools and service bureaus automate the process of auditing, cleaning, and monitoring data quality. They can play a significant role in data quality efforts and be well worth the investment. Most commercial tools are now moving beyond auditing and scrubbing name and address data to tackle other data types. They are also beginning to step up to the challenge of validating company-specific business rules, and augmenting addresses with geospatial and demographic data, among other things.

Data is a vital resource. Companies that invest proportionally to manage this resource will stand a stronger chance of succeeding in today's competitive global economy than those that squander this critical resource by neglecting to ensure adequate levels of quality.

Report Methodology

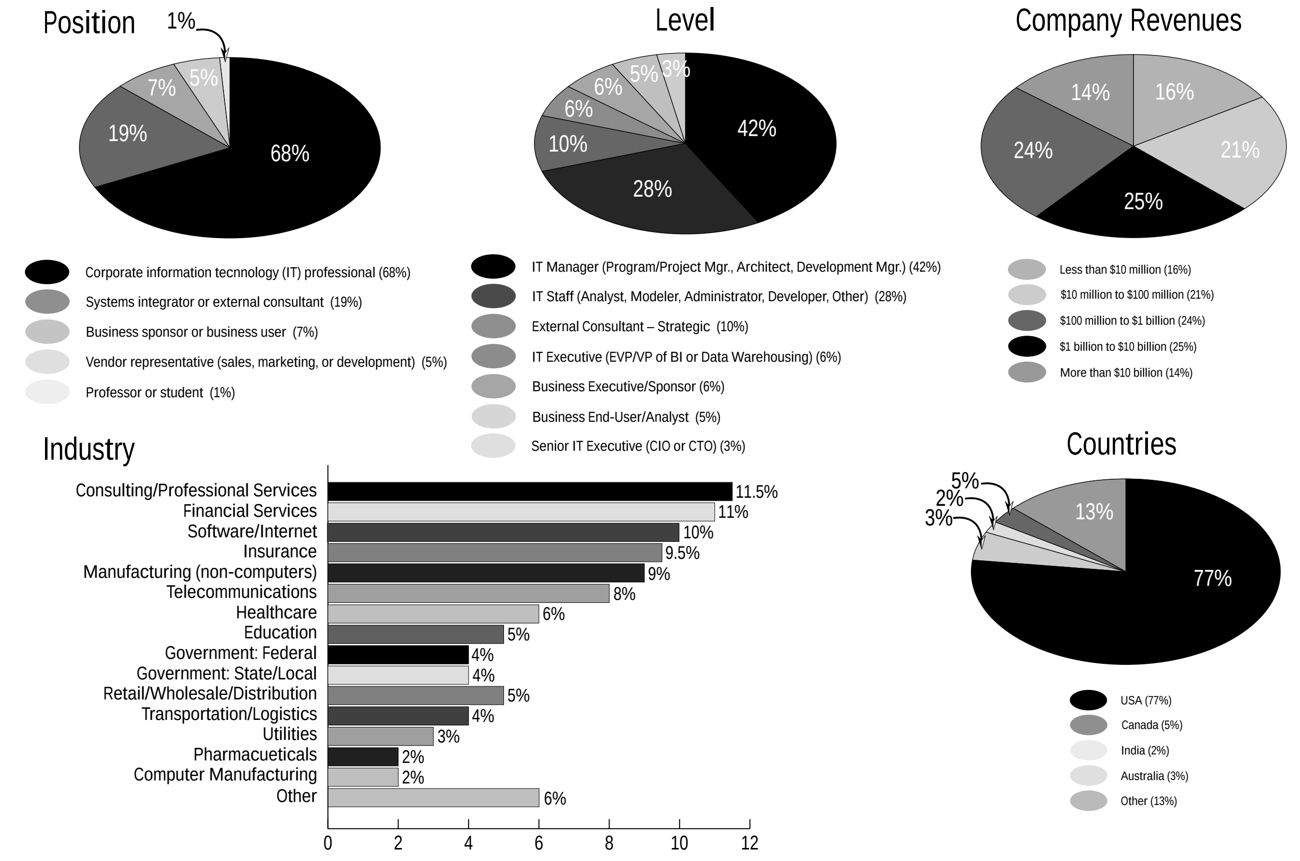
The research for this report was conducted during the fall of 2001, and is based on:

- Interviews with industry experts
- Interviews with business and technical users that have implemented data quality programs
- Books and white papers on data quality
- A comprehensive survey of data warehousing and business intelligence professionals. The survey was conducted during November and December 2001. TDWI contacted all data warehousing and business intelligence professionals in its database. It received 647 valid responses that were included in the survey results. (See below for survey demographics.)

Survey Questions. The report survey, called the Data Quality Survey in the body of the report text, consisted of 36 questions in total. Due to branching logic, most respondents answered between 30 and 32 questions. Respondents were directed to different questions based on their responses to previous questions. Many questions allowed respondents to select more than one answer, causing total results to exceed 100 percent.

Demographics. The survey drew responses from 647 individuals in a range of positions, industries, and countries. The charts below describe the makeup of survey respondents.

Demographics



Data as a Critical Resource

The Business Impact of Poor Quality Data

During the past 50 years, the developed world has moved from an industrial economy to an information economy. Companies now compete on the ability to absorb and respond to information, not just manufacture and distribute products. Intellectual capital and know-how are more important assets than physical infrastructure and equipment. Knowledge workers, not factory hands, dominate the workforce.

Downstream Costs. If information is the currency of the new economy, then data is a critical raw material needed for success. Just as a refinery takes crude oil and transforms it into numerous petroleum products, organizations use data to generate a multiplicity of information assets. These assets form the basis of strategic plans and actions that determine an organization's success. (See Illustration 1.)

Organizations use data to generate a multiplicity of information assets.

Consequently, poor quality data can have a deleterious impact on the health of a company. If not identified and corrected early on, defective data can contaminate all downstream systems and information assets, jacking up costs, jeopardizing customer relationships, and causing imprecise forecasts and poor decisions.

Data Refinery

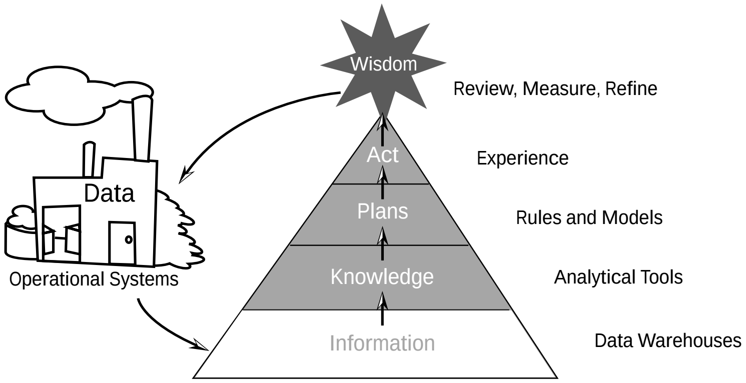


Illustration 1. A data refinery transforms data into information via a data warehouse. Knowledge workers equipped with analytical tools identify patterns in the information and create rules and models, to be used to develop business plans. Companies gain wisdom by reviewing the impact of their plans and repeating the cycle.

RISK EXPOSURE. The Data Warehousing Institute (TDWI) estimates that poor quality customer data costs U.S. businesses a staggering $611 billion a year in postage, printing, and staff overhead.[1] Frighteningly, the real cost of poor quality data is much higher. Organizations can frustrate and alienate loyal customers by incorrectly addressing letters or failing to recognize them when they call or visit a store or Web site. Once a company loses its loyal customers, it loses its base of sales and referrals, and future revenue potential.

And these estimates don't even account for the money organizations are losing due to problems with non-name-and-address data. Although customer contact data is notoriously volatile and difficult to maintain at high accuracy levels (~ 97 to 99 percent), it represents a small fraction of the data entities at most companies that must be monitored for quality.

Poor quality customer data costs U.S. businesses $611 billion a year.

Insurance Example. Consider the following real-life example: an insurance company receives 2 million claims per

[1] TDWI estimate based on cost-savings cited by survey respondents and others who have cleaned up name and address data, combined with Dunn & Bradstreet counts of U.S. businesses by number of employees.

month with 377 data elements per claim. Even at an error rate of .001, the claims data contains more than 754,000 errors per month and more than 9.04 million errors per year![2] If the insurance company determines that 10 percent of the data elements are critical to its business decisions and processes, the firm still must fix almost 1 million errors each year that could damage its ability to do business.

Companies risk losing $10 million a year from poor quality data.

What is the insurance company's exposure to these errors? Let's say the firm estimates its risk at $10 per error. This covers staff time required to fix the error downstream after a customer discovers it, the subsequent loss of customer trust and loyalty, and erroneous payouts (both high and low.) Even at $10 per error, a conservative estimate, the company's risk exposure to poor quality claims data is $10 million a year! And this doesn't include the firm's exposure to poor quality data in its financial, sales, human resources, decision support, and other applications.

Perception versus Reality: Paying Lip Service to Data Quality

Given the business impact of poor quality data, it is bewildering the casual way in which most companies manage this critical resource. Most organizations do not fund programs designed to build quality into data in a proactive, systematic, and sustained manner. According to TDWI's Data Quality Survey, almost half of all companies have no plan for managing data quality. (See Illustration 2.)

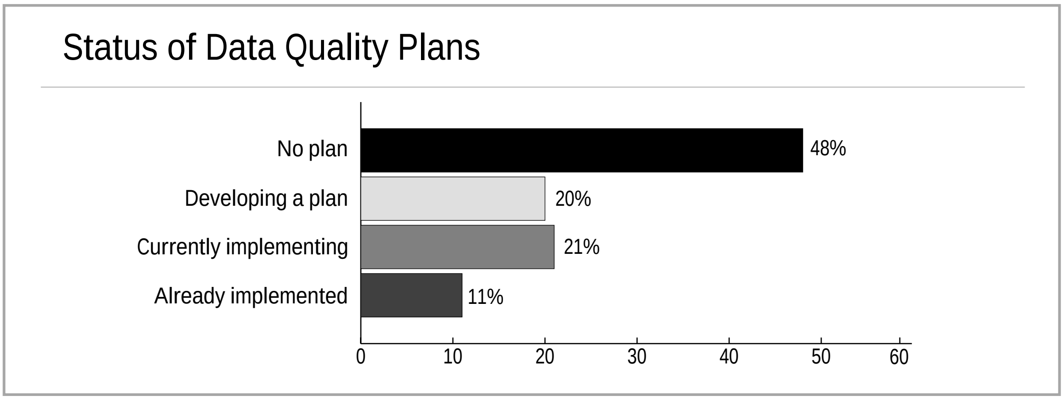


Illustration 2. Almost half of companies (48 percent) do not have a plan for managing or improving data quality.

Organizations overestimate data quality, underestimate costs of errors.

Part of the problem is that most organizations overestimate the quality of their data and underestimate the impact that errors and inconsistencies can have on their bottom line. On one hand, almost half of companies believe the quality of their data is "excellent" or "good." Yet, almost half of respondents also said the quality of their data is "worse than everyone thinks." (See Illustrations 3 and 4.)

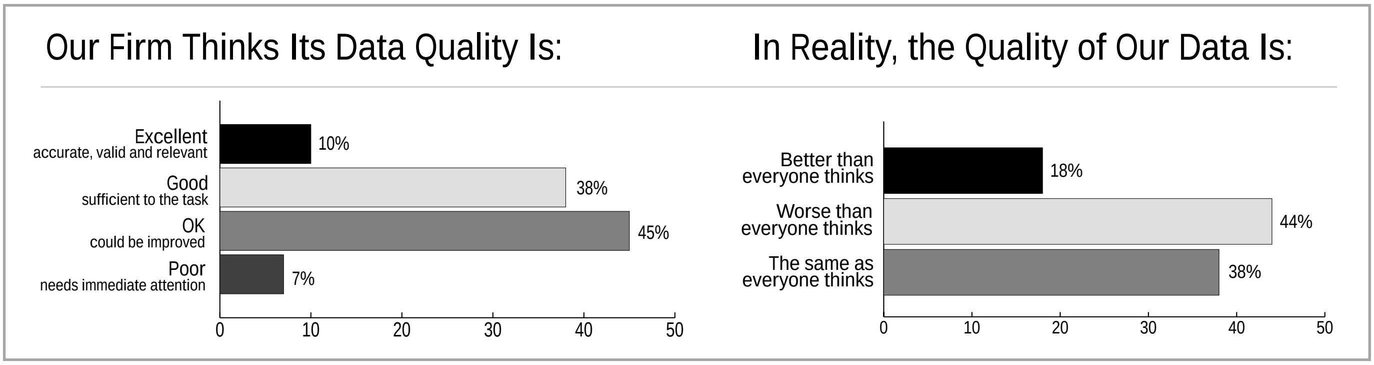


Illustration 3. Almost half of companies (48 percent) think the quality of their data is excellent or good….

Illustration 4. … But almost half of respondents think the quality of their data is worse than everyone thinks.

[2] From *TDWI Data Cleansing: Delivering High Quality Warehouse Data*, TDWI course book, Second Edition, November 2001, page 4-38.

Clearly, these survey results indicate a significant gap between perception and reality regarding the quality of data in most organizations. It is not surprising, then, that most individuals think their organization needs more education on the importance of data quality and the methods for improving and managing it. (See Illustration 5.)

Gap between perception and reality exists on data quality.

Most organizations seem to display an almost laissez-faire attitude about ensuring high-quality data. It appears either that executives are oblivious to the problems of defective data, or that they accept these problems as a normal cost of doing business. Few executives seem to understand that poor quality data puts them at a competitive disadvantage. And even data that today is deemed "sufficient to the task" or "good" may not be adequate to support future information needs.

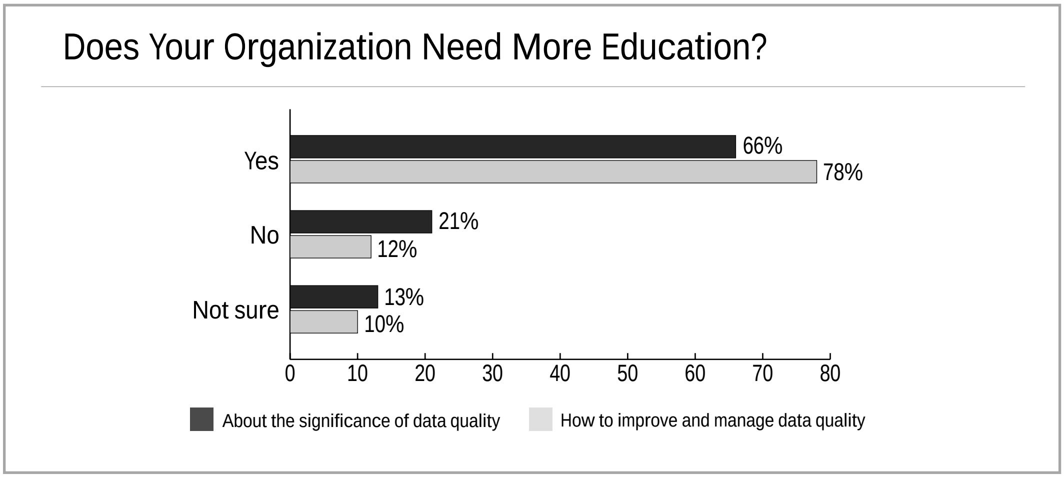


Illustration 5. There is a desperate need for additional education about the importance of data quality and methods to maintain and improve it.

Quality Ratio. One way to raise the issue of data quality with executives is to present them with a chart that compares the firm's investments in maintaining the quality of its top resources. The chart depicts a Quality Ratio, which correlates a firm's expenditures to operate a critical resource with its expenditures to maintain and improve that resource.

A Quality Ratio compares investments in critical resources.

Although simplistic, this metric can help executives begin to see and evaluate the manner in which they are investing in key resources, including data. It also helps elevate data to the upper echelon of resources.

Expenditures on Critical Resources

A	*B*	*C*	*D*
Critical Resource	Annual $$ to Operate	Annual $$ to Improve	Quality Ratio (Column C/B)
Employees	Salary, benefits, bonuses, etc.	Training	%
Products	Manufacturing, Distribution	Research and Development	%
Customers	Sales, Marketing, Advertising	Customer service	%
Equipment/ Systems	Capital equipment budget	Capital maintenance budget	%
Applications	Software programs, development staff	QA staff, test software	%
Data	Information systems less capital expenses	Data quality programs	%

Illustration 6. The Quality Ratio makes it easy for executives to compare investments in maintaining and improving various critical resources, including data.

The goal is to manage the quality of data with the same diligence that companies use in managing their money.

Ultimately, the goal for companies is to manage the quality of data with the same diligence and attention to detail that they devote to managing money. This is not a pie-in-the-sky nicety; it's just good business practice. Once a company values data as a critical resource, it is not that big a leap to make a corporate commitment to manage data quality. This commitment naturally leads to establishing a program that marshals people, processes, systems, and data quality tools to achieve a common goal of high quality data.

The following mission statement for a data quality program initiated at Cullen/Frost Bankers, Inc., an $8 billion financial holding company based in San Antonio, TX, articulates the rationale for investing in high quality data:

More than 98 percent of our company's assets and those of our customers are managed by data and information—and less than 2 percent are in the form of cold, hard cash. Just as we are careful and meticulous in managing cash and negotiables, we have a duty and obligation to exercise a high degree of care with the data that is the basis for customer relations and decision making.[3]

The bottom line is that data is a critical asset in the information economy, and the quality of a company's data is a good predictor of its future success.

The Value Proposition for High Quality Data

A Litany of Problems

Although some companies understand the importance of high quality data, most are oblivious to the true business impact of defective or substandard data.

The cost of poor data may be 10–25 percent of total revenues.

Larry English, a leading authority on data quality issues, writes, "...the business costs of nonquality data, including irrecoverable costs, rework of products and services, workarounds, and lost and missed revenue may be as high as 10 to 25 percent of revenue or total budget of an organization." [4]

High-Profile Failures. Thanks to a raft of new information-intensive strategic business initiatives, executives are beginning to wake up to the real cost of poor quality data. Many have bankrolled high-profile IT projects in recent years—data warehousing, customer relationship management (CRM), and e-business projects—that have failed or been delayed due to unanticipated data quality problems.

Data quality issues torpedoed a $38 million CRM project.

For example, in 1996, FleetBoston Financial Corp. (then Fleet Bank) in New England undertook a much publicized $38 million CRM project to pull together customer information from 66 source systems. Within three years, the project was drastically downsized and the lead sponsors and technical staff were let go. A major reason the project came unraveled was the team's failure to anticipate how difficult and time consuming it would be to understand, reconcile, and integrate data from 66 different systems.

According to TDWI's 2000 industry study, the top two technical challenges facing companies implementing CRM solutions are "managing data quality and consistency" (46 percent) and "reconciling customer records" (40 percent.) Considering that 41 percent of CRM projects were "experiencing difficulties" or "a potential flop," according to the same study, it's clear that the impacts of poor data quality in CRM are far reaching.[5]

Summarization and Integration Problems. Data warehousing, CRM, and e-business projects often expose poor quality data because they require companies to extract and integrate data from multiple operational systems. Data that is

[3] *Inside Data Quality*, Company Newsletter, Cullen/Frost Bankers, Inc., 2001.

[4] L. English. *Improving Data Warehouse and Business Information Quality*. New York: John Wiley & Sons, 1999, p. 12.

[5] *Harnessing Customer Information for Strategic Advantage: Technical Challenges and Business Solutions*. A summary can be found at www.dw-institute.com/download/2000_Industry_Study.pdf.

sufficient to run payroll, shipping, or accounts receivable is often peppered with errors, missing values, and integrity problems that don't show up until someone tries to summarize or aggregate the data.

Also, since operating groups often use different rules to define and calculate identical elements, reconciling data from diverse systems can be a huge, sometimes insurmountable, obstacle. Sometimes the direct intervention of the CEO is the only way to resolve conflicting business practices or political and cultural differences.

Political and cultural differences create data quality problems.

All Too Common Scenarios? Every organization, if it looks hard enough, can uncover a host of costs and missed opportunities caused by inaccurate or incomplete data. Consider the following examples:

- A telecommunications firm lost $8 million a month because data entry errors incorrectly coded accounts, preventing bills from being sent out.
- An insurance company lost hundreds of thousands of dollars annually in mailing costs (postage, returns, collateral, and staff to process returns) due to duplicate customer records.
- An information services firm lost $500,000 annually and alienated customers because it repeatedly recalled reports sent to subscribers due to inaccurate data.
- A large bank discovered that 62 percent of its home equity loans were being calculated incorrectly, with the principal getting larger each month.
- A health insurance company in the Midwest delayed a decision support system for two years because the quality of its data was "suspect."
- A global chemical company discovered it was losing millions of dollars in volume discounts in procuring supplies because it could not correctly identify and reconcile suppliers on a global basis.
- A regional bank could not calculate customer and product profitability due to missing and inaccurate cost data.

In addition, new industry and government regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and the Bank Secrecy Act, are upping the ante. Organizations are now required to carefully manage customer data and privacy or face penalties, unfavorable publicity and loss of credibility.

Most Suffering Losses. Almost half of companies (40 percent) have suffered "losses, problems, or costs" due to poor quality data, according to TDWI's Data Quality Survey. And about the same percentage (43 percent) have yet to study the issue. We suspect they are experiencing similar problems but do not know it!

Forty percent of firms have suffered losses due to poor data quality.

The two most common problems caused by poor quality data are (1) extra time required to reconcile data and (2) loss of credibility in the system or application. (See Illustration 7.) These two problems are related since an inability to reconcile data between the data warehouse and the source systems causes end users to lose confidence in the data warehouse. This is true even if the data in the warehouse is more accurate. Without meta data to track the origin and transformation of data in the warehouse, users typically trust source systems before a data warehouse.

Companies also cite extra costs due to duplicate mailings, excess inventory, inaccurate billing, and lost discounts as well as customer dissatisfaction, delays in deploying new systems, and lost revenue.

Impact on Strategic Planning and Programs. Several survey respondents also noted an extremely serious problem: poor data quality has undermined strategic plans or projects:

Without good data, companies are running blind.

"We had a misdirected major action, based on misunderstanding a specific situation."

"We are unable to track company performance because our data is so suspect."

"It is impossible for us to develop business and market strategies."

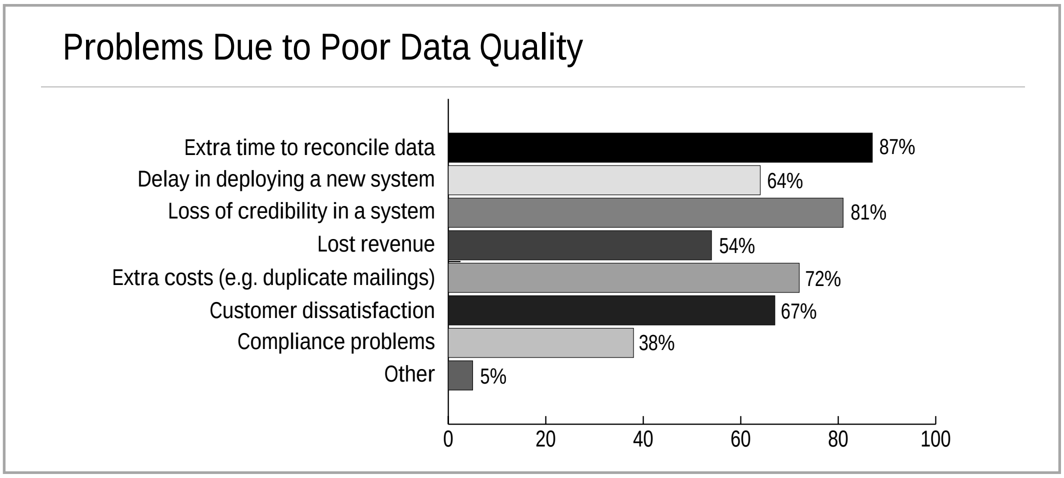


Illustration 7. Defective data causes a litany of problems. Based on 286 respondents who could select multiple answers. TDWI Data Quality Survey, December 2001.

Without good data, organizations are running blind. They can't make good decisions because they have no accurate understanding of what is happening within their company or the marketplace. They rely on intuition, which is dangerous in a fast-moving market with nimble competitors and finicky customers.

Benefits and Improvements

Tangible Benefits. On the other hand, companies that have invested in managing and improving data quality can cite tangible and intangible benefits, often the inverse of the problems mentioned above.

One company is generating $130,000 annually in cost savings from a data quality project.

For example, a data quality project at a medium-sized financial institution is generating cost-savings of $130,000 annually on an outlay of $70,000 ($40,000 for software and $30,000 for data cleansing services). This project's internal rate of return is 188 percent and the net present value is $278,000 with a several month payback.

Almost half of our respondents (47 percent) said their companies have derived benefits from better quality data. Topping the list were customer satisfaction, creating a "single version of the truth," and "greater confidence in analytical systems." (See Illustration 8.)

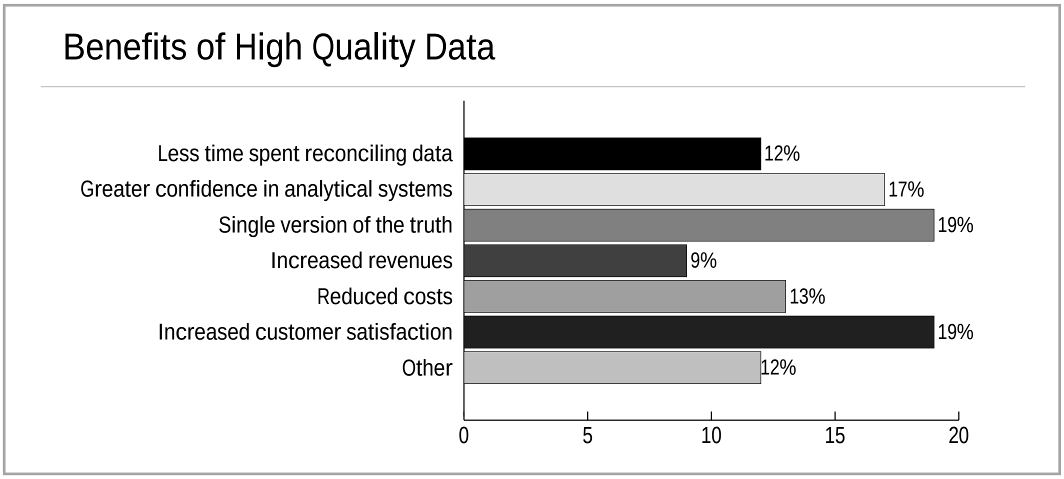


Illustration 8. There are many benefits to high quality data. Based on 304 responses, with respondents selecting one answer. TDWI survey on data quality, December 2001.

A recent report from PricewaterhouseCoopers entitled *Global Data Management Survey 2001* also validates these findings. Based on a survey of top managers at 600 firms, the report found that almost 60 percent had cut their processing costs, more than 40 percent boosted sales through better analysis of customer data, and more than 30 percent had won a significant contract through better analysis of data.

Companies Investing in data quality are "pulling ahead."

The PricewaterhouseCoopers report paints the value of high quality data in stark business terms:

Companies that manage their data as a strategic resource and invest in its quality are already pulling ahead in terms of reputation and profitability from those that fail to do so.[6]

The quality of a company's data generates both tangible and intangible costs and benefits. Clearly, the further we move into the information economy, the more important it will be for companies to invest in maintaining good quality data.

Understanding Data Quality

What Is Data Quality?

Data Quality Attributes. We have talked a lot about the importance of data quality and its business ramifications, but we still need to define what it is. Data quality is not necessarily data that is devoid of errors. Incorrect data is only one part of the data quality equation.

Most experts take a broader perspective. Larry English says data quality involves "consistently meeting knowledge worker and end-customer expectations."[7] Others say data quality is the fitness or suitability of data to meet business requirements. In any case, most cite several attributes that collectively characterize the quality of data:

1. Accuracy: Does the data accurately represent reality or a verifiable source?
2. Integrity: Is the structure of data and relationships among entities and attributes maintained consistently?
3. Consistency: Are data elements consistently defined and understood?
4. Completeness: Is all necessary data present?
5. Validity: Do data values fall within acceptable ranges defined by the business?
6. Timeliness: Is data available when needed?
7. Accessibility: Is the data easily accessible, understandable, and usable?

The first five attributes generally pertain to the content and structure of data, and cover a multitude of sins that we most commonly associate with poor quality data: data entry errors, misapplied business rules, duplicate records, and missing or incorrect data values.

But defect-free data is worthless if knowledge workers cannot understand or access the data in a timely manner. The last two attributes above address usability and usefulness, and they are best evaluated by interviewing and surveying business users of the data.

Defect-Free Data Is Not Required. It is nearly impossible to ensure that all data meet the above criteria 100 percent. In fact, it may not be necessary to attempt this Herculean feat. Data does not need to be perfect. It simply needs to meet the requirements of the people or applications that use it. And different types of workers and applications require different levels of data quality.

Data quality is in the eye of the beholder.

[6] *Global Data Management Survey 2001*, PriceWaterhouseCoopers. Download the report at http://www.pwcglobal.com.

[7] English, op. cit., p. 24.

For example, salespeople may need only a rough description of the type and amount of purchases made by specific customers in the past two years to make an effective sales call. Marketing analysts need details about customer transactions and demographic attributes to create accurate propensity models, but they can work around missing data using sampling and extrapolation. Financial analysts, on the other hand, need to track customer purchases down to the penny to deliver accurate forecasts and be able to reconcile their analyses with operational systems.

Your most stringent users determine level of quality.

Each of these knowledge workers requires a different level of accuracy, completeness, and consistency to make effective use of the data. If the data doesn't meet their expectations, they will lose confidence in the system and look elsewhere to meet their needs.

If you are designing a data warehouse to support multiple groups, you need to meet the needs of knowledge workers with the most stringent data quality requirements. And since these requirements may shift over time, it's best to build in the highest level of data quality possible to meet all potential future needs.

What Can Go Wrong?

The Web is increasing data entry errors.

The sources of poor quality data are myriad. Leading the pack are data entry processes, which produce the most frequent data quality problems, and systems interfaces.

Not surprisingly, survey respondents cite data entry errors by employees as the most common source of data defects. (See Illustration 9.) Examples include misspellings, transpositions of numerals, incorrect or missing codes, data placed in the wrong fields, and unrecognizable names, nicknames, abbreviations, or acronyms. These types of errors are increasing as companies move their businesses to the Web and allow customers and suppliers to enter data about themselves directly into operational systems.

Lack of Validation Routines. Interestingly, many data entry errors can be prevented by using validation routines that check data as it is entered into Web, client/server, or terminal-host systems. Respondents mentioned a "lack of adequate validation" as a source of data defects, noting this grievance in the "Other" category in Illustration 9.

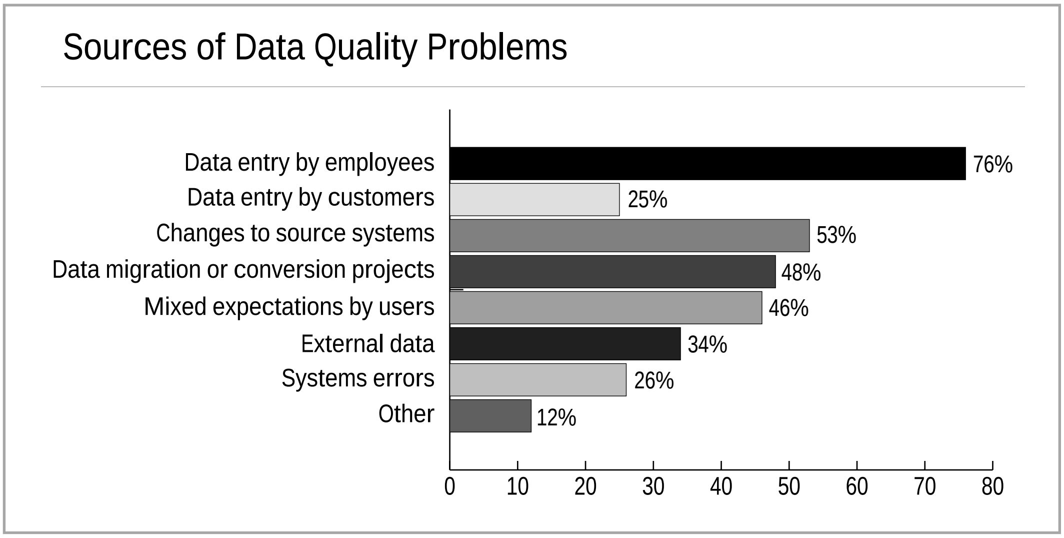


Illustration 9. Data entry and changes to source systems are the biggest causes of data quality problems for data warehousing managers.

A typo can create a valid phone number that is incorrect.

Valid, But Not Correct. But even validation routines cannot catch typos where the data represents a valid value. Although a person may mistype his telephone number, the number recorded is still valid—it's just not the right one! The same holds true for social security numbers, vehicle identification numbers, part numbers, and last names. Database integrity rules can catch some of these errors, but companies need to create complex business rules to catch the rest.

Mismatched Syntax, Formats, and Structures. Data entry errors are compounded when organizations try to integrate data from multiple systems. For example, corresponding fields in each system may use different syntax (e.g. first-middle-last name vs. last-first-middle name), data formats (6 byte date field vs. 4 byte date field), or code structures (e.g. male-female vs. m-f vs. 1-2). Either the data cleansing or the ETL tool needs to map these differences to a standard format before serious data cleanup can begin.

Mapping is required to standardize field formats among systems.

Unexpected Changes in Source Systems. Perhaps a more pernicious problem is structural changes that occur in source systems. Sometimes these changes are deliberate, such as when an administrator adds a new field or code value and neglects to notify the managers of connecting systems about the changes. In other cases, front-line people reuse existing fields to capture new types of information that were not anticipated by the application designers.

Spiderweb of Interfaces. Because of the complex of systems architectures today, changes to source systems are easily and quickly replicated to other systems, both internal and external. Most systems are connected by a spiderweb of interfaces to other systems. Updating these interfaces is time consuming and expensive, and many changes slip through the cracks "infecting" other systems. Thus, changes in source systems can wreak havoc on downstream systems if companies do not have adequate change management processes in place.

System interfaces quickly spread defects.

Lack of Referential Integrity Checks. It is also true that target systems don't adequately check the integrity of data they are loading. For example, data warehouse administrators often turn off referential integrity when loading the data warehouse for performance reasons. If source administrators change or update tables, they can create integrity problems that are not detected.

Poor System Design. Of course, source or target systems that are poorly designed can create data errors. As companies rush to deploy new systems, developers often skirt fundamental design and modeling principles, which leads to data integrity problems down the road.

Data Conversion Errors. In the same vein, data migration or conversion projects can generate defects, as well as ETL tools that pull data from one system and load it into another. Although systems integrators may convert databases, they often fail to migrate business processes that govern the use of data. Also, programmers may not take the time to understand source or target data models and therefore easily write code that introduces errors. One change in a data migration program or system interface can generate errors in tens of thousands of records.

Interpretation and Perception Problems

A Diaspora of Definitions and Rules. But a much bigger problem comes from the fragmentation of organizations into a multitude of departments, divisions, and operating groups, each with its own business processes supported by distinct data management systems. Slowly and inexorably, each group begins to use slightly different definitions for common data entities—such as "customer" or "supplier"—and apply different rules for calculating values, such as "net sales" and "gross profits." Add mergers and acquisitions and global expansion into countries with different languages and customs, and you have a recipe for a data quality nightmare.

The problems that occur in this scenario have less to do with accuracy, completeness, validity, or consistency, than with interpretation and protecting one's "turf." That is, people or groups often have vested interests in preserving data in a certain way even though it is inconsistent with the way the rest of the company defines data.

Some data problems have less to do with accuracy than interpretation.

For example, many global companies squabble over a standard for currency conversions. Each division in a different part of the world wants the best conversion rate possible. And even when a standard is established, many groups will skirt the spirit of the standard by converting their currencies at the most opportune times, such as when a sale was posted versus when the money was received. This type of maneuvering wreaks havoc on a data warehouse that tries to accurately measure values over time.

Slowly Changing Dimensions. Similarly, slowly changing dimensions can result in data quality issues depending on the expectations of the user viewing the data. For example, an analyst at a chemical company wants to calculate the

How you rewrite history affects perceptions of data quality.

Competition in subject areas can raise and lower the perception of data quality.

total value of goods purchased from Dow Chemical for the past year. But Dow recently merged with Union Carbide, which the chemical company also purchases materials from.

In this situation, the data warehousing manager needs to decide whether to roll up purchases made to Dow and Union Carbide separately, combine the purchases from both companies throughout the entire database, or combine them only after the date the two companies merged. Whatever approach the manager takes, it will work for some business analysts and alienate others who will find the data inaccurate and unusable.

In these cases, data quality is a subjective issue. Users' perception of data quality is often colored by the range of available data resources they can access to perform their work. Where there is "competition"—another data warehouse or data mart that covers the same subject area—knowledge workers tend to be pickier about data quality, says Michael Masciandaro, director of decision support at Rohm & Haas. "If there is no competition, it is easier to satisfy user requirements for data quality."

Delivering High Quality Data

"Data quality is not a project, it's a lifestyle."

Given the ease with which data defects can creep into systems, especially data warehouses, maintaining data quality at acceptable levels takes considerable effort and coordination throughout an organization. "Data quality is not a project, it's a lifestyle," says David Wells, enterprise systems manager at the University of Washington and the developer of TDWI's full-day course on data cleansing ("TDWI Data Cleansing: Delivering High Quality Warehouse Data.")

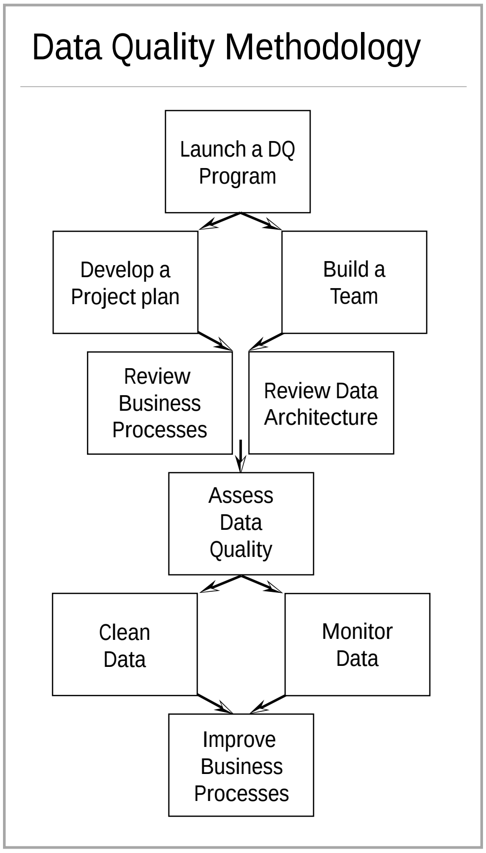


Illustration 10. An eight-step methodology for maintaining data quality.

And progress is not always steady or easy. Improving data quality often involves exposing shoddy processes, changing business practices, gaining support for common data definitions and business rules, and delivering lots of education and training. In short, fixing data quality often touches a tender nerve on the underbelly of an organization. It brings up nettlesome political and cultural issues that are not easy or quick to settle.

One top executive leading a data quality initiative says, "Improving data quality and consistency involves change, pain, and compromise. Expect a lot of frustration. The key is to be persistent and get buy in from the top. Tackle high ROI projects first, and use them as leverage to bring along other groups that may be resistant to change."

Wells emphasizes that managing data quality is a never-ending process. Even if a company gets all the pieces in place to handle today's data quality problems, there will be new and different challenges tomorrow. That's because business processes, customer expectations, source systems, and business rules all change continuously.

To ensure high quality data, companies need to gain broad commitment to data quality management principles and develop processes and programs that reduce data defects over time. To lay the foundation for high quality data, companies need to adhere to a methodology depicted in Illustration 10 and described in the following sections.

1. Launch a Data Quality Program

Getting Executives on Board. The first step to delivering high quality data is to get top managers to admit there is a problem and take responsibility for it. This is not always easy.

"A lot of executives talk about data quality, but few do anything about it," says Jim Lair, chairman of the Center for Data Quality, a data quality services and consulting firm in Reston, VA. "Only a disaster gets people to address data quality issues."

Once disaster strikes—complaints from irate customers, excessive cost outlays, or fruitless arguments among top analysts about whose data is correct—executives are more amenable to taking responsibility.

Responsibility in the Wrong Place. Today, most companies delegate authority for managing data quality to the IT department. (See Illustration 11.) Although IT must be involved in the process, it doesn't have the clout to change business processes or behavior that can substantially improve data quality.

"Improving data quality and consistency involves change, pain, and compromise. Expect a lot of frustration."

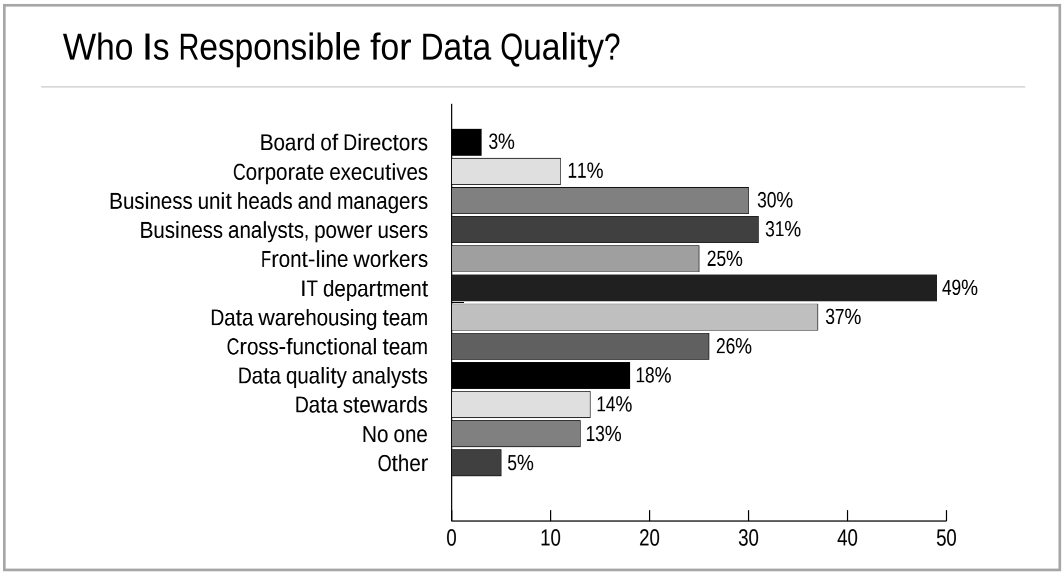


Illustration 11. IT is responsible for managing data quality in most organizations, followed by the data warehousing team. Based on 646 respondents.

To succeed, a data quality program must be initiated by the CEO, overseen by the board of directors, and managed either by a chief data quality officer or senior-level business managers in each area of the business. This notion is reinforced by a recent PricewaterhouseCoopers study on data management.

First step: get executives to admit there is a problem.

> *"For a director in charge of marketing, production or CRM to fail to take an interest in data management, or any responsibility for its quality, is a clear abdication of duty. The companies which have established clear management control over data management are acknowledging the fact that it is a core competency like managing people or customer relationships—and that, as a key foundation of the business, it should be handled at the board level alongside other business-critical issues."*
>
> *Data Management Survey*, 2001, PricewaterhouseCoopers, p. 9.

Data Stewardship Program. The best way to kickstart a data quality initiative is to fold it into a corporate data stewardship or data administration program. These programs are typically chartered to establish and maintain consistent data definitions and business rules so the company can achieve a "single version of the truth" and save time on developing new applications and looking for data. Managing data quality is a natural extension of these activities.

The best way to kickstart a data quality initiative is to fold it into a corporate data stewardship program.

A corporate stewardship committee needs to develop a master plan for data quality that contains a mission statement, objectives, and goals. It then needs to educate all employees about the plan and their roles in achieving the goals.

Top-Down Approach. Cullen/Frost Bankers, Inc., a large financial services firm in San Antonio, TX, recently established a data stewardship and data quality program to fix data quality problems that were adversely affecting some client relationships and impacting its CRM process. With guidance from an outside consultancy, chairman Richard Evans kicked off a joint data stewardship and data quality program by inviting senior managers from all lines of business to serve as chief data stewards in their areas.

State the program's objectives.

Objectives. The bank laid out its objective for the Data Quality Initiative in a company newsletter:

Our key objective ... is to materially improve the profitability of the company by (1) [e]nsuring decisions are based upon fresh and accurate information, and (2) reducing expenses associated with the inefficiencies incurred from rework, complaint handling, elongated processing time, unplanned service, and redundant efforts.

Set goals and expectations.

Goals. According to one executive, the company's goal is to achieve "zero defects" and continuously measure its progress toward that goal, along with the costs of variance. The bank also set forth expectations to the stewards and employees: "The journey to quality information does not happen overnight. The process to correct and prevent recurrences can be lengthy."

Oversee data quality groups and functions throughout the organization.

Oversight Tasks. The corporate stewardship committee also needs to oversee and provide direction to all data quality teams or functions scattered throughout the company. Specifically, the committee should:

- Provide education on the importance of data quality to the company
- Communicate data quality improvements to all employees
- Define mandatory data elements that need to be measured
- Review metrics for measuring the quality of these elements
- Define methods of reporting on data quality levels
- Set precedence for establishing service level agreements for data quality
- Establish the "owners" and "custodians" of each major subject area and data store
- Resolve cross-divisional enterprise data quality issues
- Ensure code sets are updated regularly

In essence, the corporate stewardship committee needs to be a clearinghouse for ideas and information pertaining to data quality, and work diligently with all teams and individuals to sustain momentum to achieve the program's goals.

2. Develop a Project Plan

Prioritize projects by their upside potential.

The next step is to develop a data quality project plan, or series of plans. You do this by prioritizing projects that have the greatest upside for the company, and tackle them one by one.

Plan Composition. A project plan should define the scope of activity, set goals, estimate ROI, perform a gap analysis, identify actions, and measure and monitor success. To perform these tasks, the team will need to dig into the data to assess its current state, define corrective actions, and establish metrics for monitoring conformance to goals.

- Define Scope. After a preliminary data assessment performed during the launch phase (see step 1), a company should have a decent understanding of the high priority areas it needs to tackle. A scope document identifies the key data elements that will be analyzed, measured for validity, and then cleaned and monitored. For example, Alliance Blue Cross Blue Shield identified 61 key data elements out of 588 that it would focus on.
- Set Goals. A plan should have a well defined but realistic goal that gives the project clarity and momentum to succeed. A goal might be to reduce duplicate customer records in the data warehouse to less than one percent in three months.

CFB estimated $100,000 annual savings and a quick payback period.

- Define ROI. The plan should also estimate the ROI and payback period for each project. For example, Cullen/Frost Bank decided to first tackle incorrect customer addresses. The team determined that it would save more than $100,000 annually by reducing the cost of postage, printing, and staff needed to handle returned mail, and earn a payback in less than a year. The project will also yield less tangible results, such as improved customer service and satisfaction.

- Identify Actions. A plan defines how data will be corrected (i.e. prevention, detection, correction, filtering, verification) and where (source, staging, area, target.) (See steps 6 and 8.)
- Measure and monitor success. This involves building programs to monitor data quality of key data elements on an ongoing basis. (See step 7.)

3. Build a Data Quality Team

Positions. To implement a data quality plan, organizations must assign or hire individuals to create the plan, perform initial assessment, scrub the data, and set up monitoring systems to maintain adequate levels of data quality.

Industry experts outline numerous data quality positions. (See "Data Quality Roles.") Although your organization may not be large enough to dedicate individuals to each of these positions, it's important that someone assumes the responsibilities described. In small or mid-sized organizations or departments, a single individual may assume responsibility for multiple roles. (TDWI's *2001 Data Warehousing Salary, Roles, and Responsibilities Report*, for example, shows that data warehousing managers assume an average of 4.8 roles!)

Data Quality Roles

The following data quality positions may not exist in your organization, but it's important that someone assumes the roles described below. For a more complete description of these and other data quality positions, see Larry English's *Improving Data Warehouse and Business Information Quality* (John Wiley & Sons, 1999) p. 450-453.

> Chief Quality Officer—A business executive who oversees the organization's data stewardship, data administration, and data quality programs.
>
> Data Steward—A business person who is accountable for the quality of data in a given subject area.
>
> Subject Matter Expert—A business analyst whose knowledge of the business and systems is critical to understand data, define rules, identify errors, and set thresholds for acceptable levels of data quality.
>
> Data Quality Leader—Oversees a data quality program that involves building awareness, developing assessments, establishing service level agreements, cleaning and monitoring data, and training technical staff.
>
> Data Quality Analyst—Responsible for auditing, monitoring, and measuring data quality on a daily basis, and recommending actions for correcting and preventing errors and defects.
>
> Tools Specialists—Individuals who understand either ETL or data quality tools or both and can translate business requirements into rules that these systems implement.
>
> Process Improvement Facilitator—Coordinates efforts to analyze and reengineer business processes to streamline data collection, exchange, and management, and improve data quality.
>
> Data Quality Trainer—Develops and delivers data quality education, training, and awareness programs.

Data quality analysts monitor quality levels, analyze the source of problems, and recommend corrective actions.

Data Quality Analysts. Most successful data warehousing programs hire one or two full-time data quality analysts who serve on the front lines of data quality management. These analysts monitor data quality levels, analyze the source of problems, and recommend corrective actions to the oversight committee.

These analysts may report to a data or information quality leader or, more likely, the head of data warehousing or data administration, but they communicate consistently with custodians of source systems and subject-area data stewards.

Data quality analysts should be recruited from an operational business role.

Analyst Background and Skills. Ideally, data quality analysts should be recruited from an operational role on the business side of the organization. They should be familiar with how transactions are executed and understand the data fields and business rules driving the operational applications.

BlueCross BlueShield of North Carolina. For example, BlueCross BlueShield of North Carolina typically recruits data quality analysts who have spent several years in operational roles such as member enrollment, suspense, or claims processing, says Celia Fuller, director of corporate data warehousing. These individuals receive training in SQL and Unix so they can become proficient at navigating databases and manipulating data.

Analysts write business rules and data integrity scripts, set tolerances, and measure progress.

The analysts implement business rules defined by business users and data stewards. They also write data integrity scripts under the guidance of the chief architect. These scripts identify defects among approximately 150 key data elements (KDE) as they are loaded into the data warehouse and measure the level of quality for each field based on predefined measures, such as number of null values or missing fields or constraint violations. The analysts also set tolerances that define acceptable quality levels for each KDE and measure progress against those tolerances.

For example, the analysts oversee the "balancing" process, which reconciles the KDEs in the data warehouse with corresponding fields in the source systems. If there is a significant deviation from defined tolerances, the analysts report the errors to a requirements committee along with recommendations for corrective action.

4. Review Business Processes and Data Architecture

Once there is corporate backing for a data quality plan, the stewardship committee—or a representative group of senior managers throughout the firm—needs to review the company's business processes for collecting, recording, and using data in the subject areas defined by the scope document. With help from outside consultants, the team also needs to evaluate the underlying systems architecture that supports the business practices and information flows.

The business and technical reviews can take several days or weeks.

The business and technical reviews can take several days or weeks depending on the scale and scope of the data quality plan and the number of stakeholders involved.

BUSINESS PROCESS REVIEW. On the business side, the team needs to document how data elements in each subject area are collected and where they are stored. The team also needs to identify who owns the data, who uses the data, what kind of reports they need, and what quality checks exist, if any.

Cullen/Frost Banks discovered that customer records were stored in a half dozen systems.

Cullen/Frost Banks, for example, discovered that customer records were stored in a half dozen systems due to a spate of acquisitions. This made it difficult for the bank to create a common, accurate view of all its customers. The bank decided to consolidate customer data from diverse systems into an operational data store (ODS) that, once deployed, will support various operational applications in marketing, sales, and customer service, and funnel customer data into the warehouse for analysis and modeling.

Deliverables. The business process review should generate a document that exposes business processes that need to be modified and suggest metrics for measuring data quality, says Elizabeth Vannan, project manager at the Centre for Education Information in Victoria, British Columbia. Vannan is helping the BC Ministry of Advanced Education establish data quality practices that will enable 22 data warehouses to collect standard data for reporting and analysis.

TECHNICAL REVIEW. At the same time, the technical review can reveal whether there are opportunities to re-architect or replace systems to improve data quality and optimize data collection. These changes may also be needed to pave the way for new strategic applications such as CRM, sales force automation, e-commerce, personalization, or supply chain analytics.

A good technical architecture will do a number of things to enhance data quality. Illustration 12 shows an example of the possible results when companies transform their data architectures from data quality quagmires to data quality greenfields.

From data quality quagmires to data quality greenfields.

- Implement Validation Routines. Implement robust validation routines at data collection points.
- Implement ETL and Data Quality Tools. Use data quality and extraction, transformation, and load (ETL) tools to automate the continuous detection, cleansing, and monitoring of key files and data flows.
- Implement Data Quality Checks. Implement data quality checks or audits at reception points or within ETL processes. Stringent checks should be done at source systems and a data integration hub.
- Consolidate Data Collection Points. Consolidate source systems or data collection points to minimize divergent data entry practices.
- Consolidate Shared Data. Model subject areas so data can be integrated physically or logically. Use a data warehouse or ODS to physically consolidate data used by multiple applications. Or use reference tables and keys to logically integrate data that must remain distributed across systems.
- Minimize System Interfaces. Minimize system interfaces by (1) backfilling a data warehouse behind multiple independent data marts, (2) merging multiple operational systems or data warehouses, (3) consolidating multiple non-integrated legacy systems by implementing packaged enterprise application software, and/or (4) implementing a data integration hub (see next).
- Implement a Data Integration Hub. A hub can minimize systems interfaces and provide a single source of clean, integrated data for multiple applications. This hub uses a variety of middleware (e.g. message queues, object request brokers) and transformation processes (ETL, data quality audits) to prepare and distribute data for use by multiple applications.
- Implement a Meta Data Repository. Create a repository for managing meta data gleaned from all enterprise systems. The repository should provide a single place for systems analysts and business users to look up definitions of data elements, reports, and business views; trace the lineage of data elements from source to targets; identify data owners and custodians; and examine data quality reports. In addition, enterprise applications, such as a data integration hub or ETL tools, can use this meta data to determine how to clean, transform, or process data in its workflow.

5. Assess Data Quality

DATA AUDITING. After reviewing information processes and architectures, an organization needs to undertake a thorough assessment of data quality in key subject areas. This process is also known as data auditing or profiling.

The purpose of the assessment is to (1) identify common data defects (2) create metrics to detect defects as they enter the data warehouse or other systems, and (3) create rules or recommend actions for fixing the data. This can be long, arduous, and labor-intensive work depending on the scale and scope of the project, and the age and cleanliness of source files.

Data warehousing consultants have seen too many companies fall victim to the “code, load, and explode” phenomenon. That is, they bypass doing a rigorous source code analysis in order to code extract and transform programs. Then, when they load data into the warehouse, it spits out huge numbers of errors, making the warehouse virtually unusable and forcing developers to start the process anew.

“Code, load, and explode”—the result of an insufficient data audit.

Consequently, many consultants recommend that companies minimize the number of source systems they extract data from when building their first data warehouse. This technique minimizes project delays due to poor quality data and shortens the laborious process of auditing and fixing source data.

Systematic Review. The assessment should systematically review all data elements. It should identify problems such as missing data, incorrect values, duplicate records, and business rule violations.

Architecting for Data Quality

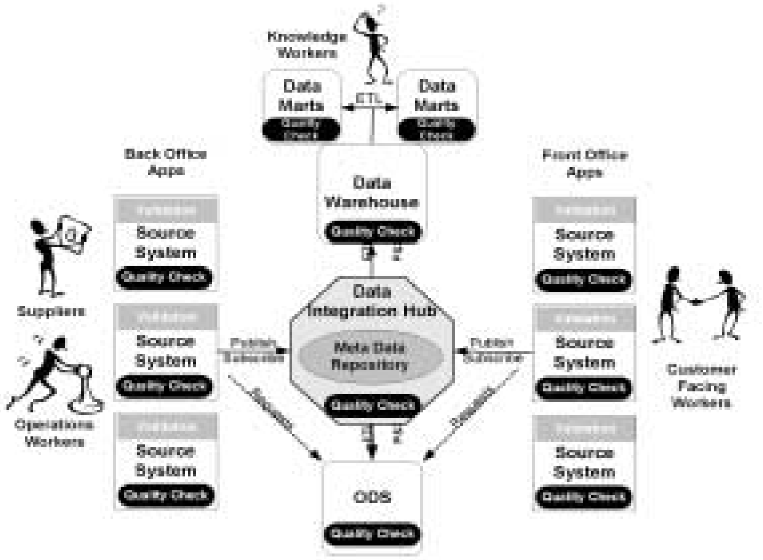


Illustration 12. An example of a cleaner architecture that optimizes data quality, in contrast to the spider web of interfaces that exists within a typical organization.

The result of data quality audits can be shocking.

The result of such assessments can be shocking, if not depressing for many companies. Consider the following data anomalies discovered by the Center for Data Quality in audits it has performed for clients:

- Required social security number missing from 82 percent of claims
- Blank fields in 30 percent of a million records from a securities firm
- Active customer status missing in 47 percent of the database
- A database contains 1,100 organization codes but only 18 are valid
- 82 percent of loans are calculated on the low side
- Duplicate payroll numbers

Or consider these defects that the British Columbia Ministry of Advanced Education discovered during its assessment phase:[8]

- Students more than 2,000 years old and students not yet born
- Course sections that occurred before the college was established
- Course sections that ended before they started
- Students registered in the same course section multiple times
- Invalid program, course section, and course codes

METHODS FOR AUDITING DATA. There are several ways to audit existing data files. Companies can (1) issue SQL queries against a sample of data, (2) use a commercial data profiling tool, or (3) send a data sample to a service bureau for evaluation.

[8] "Quality Data—An Improbable Dream?" Educause Quarterly, Number 1, 2001.

SQL. Of the three options above, most organizations prefer to perform the work in house using SQL, according to our survey. (See Illustration 13.) One drawback to this approach is that it limits queries to known or existing conditions, such as those that users have been complaining about.

Most firms use SQL to audit data files.

Profiling Tools. On the other hand, commercial data auditing tools use pattern recognition and classification techniques to dig deeper into the tables. They often analyze dependencies among fields in one or more tables, allowing them to uncover more anomalies in the data. Dependency profiling also enables them to reverse engineer the source system data models, which can facilitate data warehouse design.

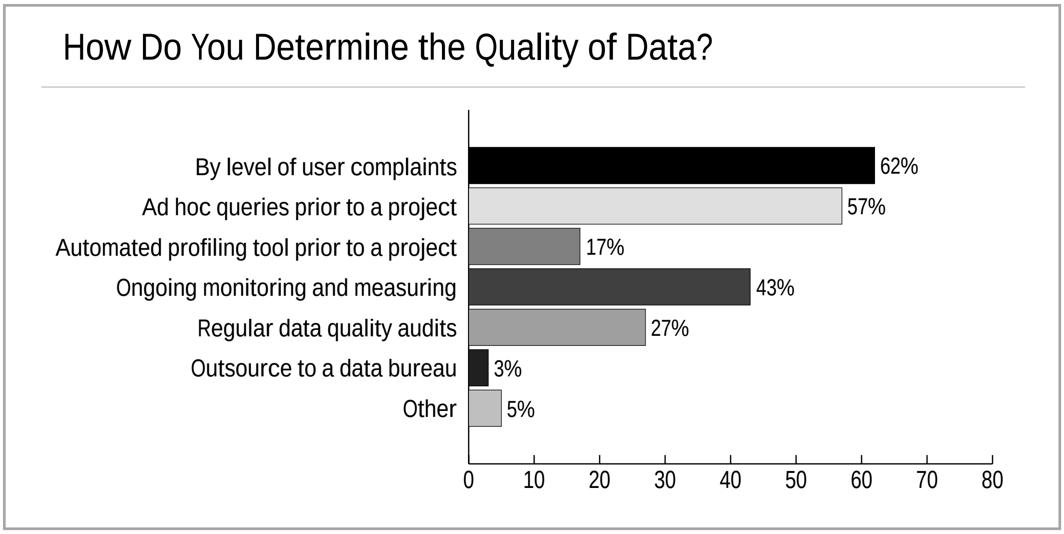


Illustration 13. Other than user complaints, a majority of organizations use SQL queries to audit data files prior to a project. Based on 640 respondents who could each select multiple answers.

Service Bureaus. The last option—outsourcing sample data to a service bureau—is not popular yet, according to our survey. But it can be a good alternative if you need to:

Service bureaus provide an outside, independent review of data quality levels.

- Kickstart a data quality initiative by getting an independent assessment of your organization's data quality.
- Perform an audit but can't spare in-house resources.
- Verify against third-party databases, such as the United States Postal Service database of valid U.S. addresses, or customer tracking databases, such as Acxiom's Abilitec or Experian's TruVue.
- Adhere to a company policy of obtaining regular outside audits.
- Augment existing cleansing processes and push accuracy rates to 96 or 98 percent.
- Rely on people with expertise and experience in data quality management that do not exist in house.

Peter Harvey, CEO of Intellidyn, a marketing analytics firm in New York, says that when his firm audits recently "cleaned" customer files from clients, it finds that 5 percent of the file contains duplicate records. The duplication rate for untouched customer files can be 20 percent or more.

Duplication rates in customer files range from 5 to 20 percent.

Harvey says service bureaus should be able to demonstrate a huge ROI and immediate (three-to-six-month) payback on the fees they charge. In addition, service bureaus should provide a free audit on a data sample to demonstrate their capabilities and value proposition, he adds.

Most service bureaus parse and match data, but they also need to correct errors and fill in missing data, says Harvey. This is vital if the client (or service bureau) needs to model customer behavior using data mining or other analytical tools. In addition, if your customer files contain millions of records, you should check that a service bureau has high-performance systems to process files in a timely manner.

Subject matter experts, who know both the business and the data, define rules and metrics.

BUILD RULES AND METRICS. The key to any data auditing approach is working closely with subject matter experts who understand both the business and the data. These experts need to study audit reports and determine which "anomalies" represent defects and which are valid data elements.

Data Cleansing Rules. The subject matter experts then need to define business rules for cleaning defects in the file. These rules might include mapping codes, assigning default values to missing fields, or standardizing fields against a reference library. (See "Clean the Data" below) The experts also need to recommend whether to fix data at the source, the staging area, or in the data warehouse, or whether to change business processes to prevent defects from occurring in the first place.

Monitor on a continuous basis.

Data Quality Metrics. The experts then should suggest metrics for measuring the quality of the data file on a continuous basis. (See "Monitor Data Continuously," page 24) Potential metrics are:

- Acceptable counts on the number of invoices or claims issued per month
- Proper syntax on various fields, such as phone number, names, dates
- Counts on unique identifiers, such as customer IDs and vehicle IDs
- Acceptable date ranges for data elements such as birth date, age, orders, and shipments
- Formulas for calculating loan amounts
- Rules that define constraints between fields, such as item counts and dollar totals in master-detail records.

These data quality metrics can be applied in the source system, staging area, or data warehouse.

6. Clean the Data

Prevention is more cost effective over the long haul.

THE ECONOMICS OF PREVENTION. Once the audit is complete, the job of cleaning the data begins. A fundamental principle of quality management is to detect and fix defects as close as possible to the source to minimize costs.

Prevention is the least costly response to defects, followed by correction and repair. Correction involves fixing defects in house, while repair involves fixing defects that affect customers directly. (See Illustration 14.) Examples of repair are direct mail pieces that are delivered to a deceased spouse, or software bugs in a commercially available product. Defect prevention programs may cost significant money to implement but pay bigger dividends in the long run.

CLEANSING METHODS. There are four basic methods for "cleaning" data:

You will need to fix defective data elements and records.

- Correct. Most cleansing operations involve fixing both defective data elements and records.

  Correcting data elements typically requires you to (1) modify an existing incorrect value (e.g. fix a misspelling or transposition), (2) modify a correct value to make it conform to a corporate or industry standard (e.g. substitute "Mr." for "Mister"), or (3) replace a missing value. You can replace missing values by either inserting a default value (e.g. "unknown") or a correct value from another database, or by asking someone who knows the correct value.

  Correcting records typically requires you to (1) match and merge duplicate records that exist in the same file or multiple files, and (2) decouple incorrectly merged records. Decoupling is required when a single record contains data describing two or more entities, such as individuals, products, or companies. (See "Role of Data Quality Tools," p. 26, for more information on matching and consolidation techniques.)

  To correct data in a relational database, analysts use SQL or a commercial data quality tool with built-in SQL support. To correct defects in non-SQL databases, you must use the native data manipulation language. To cor-

Data Quality Economics

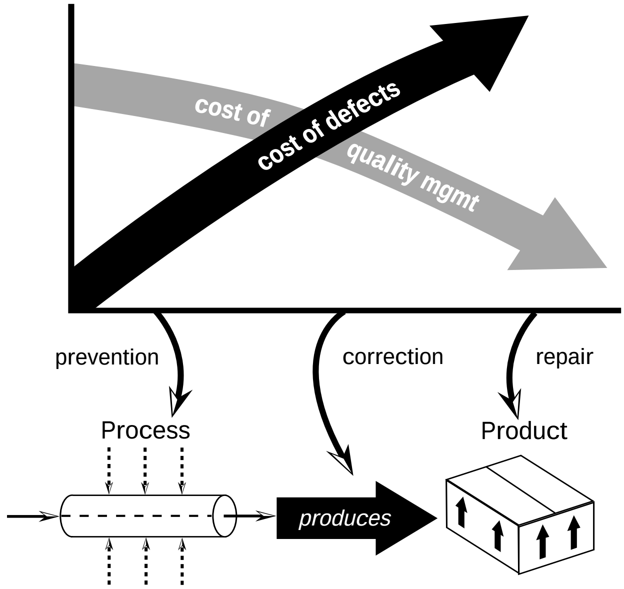


Illustration 14. Defects increase in cost the longer they go undetected, while the cost of quality management decreases as defects are prevented. From TDWI's course, "TDWI Data Cleansing: Delivering High-Quality Warehouse Data."

rect data on the fly during extraction or loading processes, you can use ETL tools, which offer a visual interface for defining transformations.

- **Filter.** Filtering involves deleting duplicate, missing, or nonsensical data elements, such as when an ETL process loads the wrong file or the source system corrupts a field. Caution must be taken when filtering data because it may create data integrity problems.
- **Detect and Report.** In some cases, you may not want to change defective data because it is not cost-effective or possible to do so. For example, if 20 percent of your customer records are missing social security numbers, but the records are more than 30 years old, there may be no business value to cleaning this data. In these cases, analysts need to notify users and document the condition in meta data.
- **Prevent.** Prevention involves educating data entry people, changing or applying new validations to operational systems, updating outdated codes, redesigning systems and models, or changing business rules and processes.

In some cases, you may not want to change defective data because it is not cost-effective or possible to do so.

WHERE TO CLEAN? If defects can't be prevented, the best place to clean data is in the source system so the defects cannot spread to the data warehouse and other downstream systems. Most of our survey respondents fix data at the source. (See Illustration 15.)

"Our goal is to fix at the source," says Jim Funk, IS Manager of Global Information Architecture at SC Johnson. "We're not 100 percent successful, but we don't want to introduce errors into the data warehouse if we can help it."

Fixing errors at the source is very important when totals in the data warehouse must reconcile exactly with data in the source systems. This is typically the case with financial applications. Fixing data at the source means both systems are operating off the same data—for better or worse.

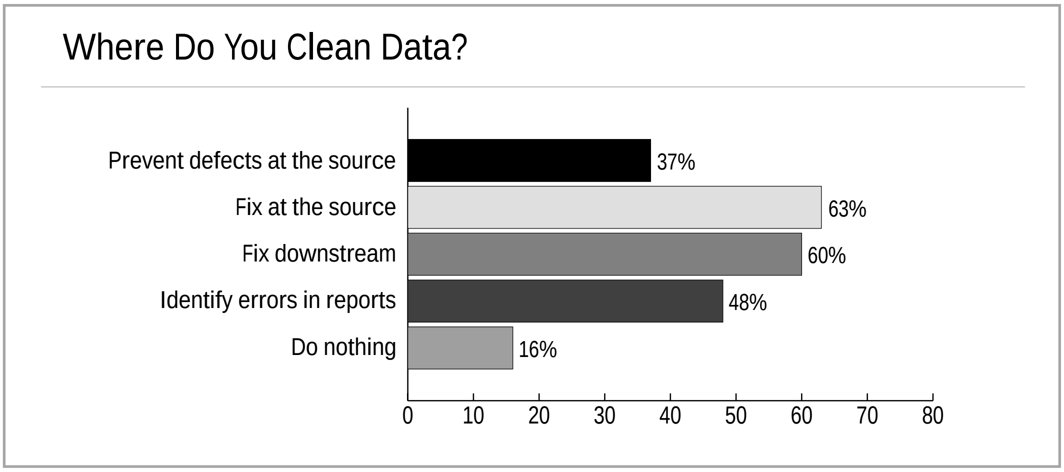


Illustration 15. Most companies clean defects at the source.

"There is a real problem with having too much data quality."

—Darrell Piatt

"If you can't reconcile totals, the data warehouse will face a credibility problem, even if its data is more accurate," says Darrell Piatt, former data warehousing manager at CompuCom and now a consultant at Kaloke Technologies. "There is a real problem with having too much quality."

Other Places to Clean Data. Besides cleaning data at the source, organizations can fix data in other places:

- Staging area—Fixing defective data in a staging area prevents errors from source systems, external data files, or the extraction process itself from entering the data warehouse. It also reduces the workload on ETL processes, especially if there are many complex cleansing operations.
- ETL process—ETL tools can perform some, but not all, data cleansing operations. (See "The Role of Data Quality Tools," p. 26.) However, they can be combined or, in some cases, integrated with data cleansing tools to provide a more complete solution. ETL tools will become a more important data cleansing option as companies update data warehouses in near real time using trickle feed connections to transaction-based middleware.

Cleaning data in the warehouse shifts the burden of defective data from the data creators to the data users.

- Data Warehouse—Since ETL processes can introduce errors, you will need to audit and clean data in the data warehouse before allowing users to access it. However, cleaning data in the data warehouse shifts the burden of dealing with defective data from the data creators to the data users. "If an error exists in the data warehouse, it has already impacted the business," says Celia Fuller, director of corporate data warehousing at BlueCross BlueShield of North Carolina.

No matter where companies decide to clean defective data, the most important thing is to have a strategy and a plan for detecting and cleaning defects.

7. Monitor Data

Firms can quickly lose the benefits of data cleansing efforts if they fail to monitor data quality continuously.

It is time consuming to prepare data files when loading a database for the first time. But organizations can quickly lose the benefits of these data preparation efforts if they fail to monitor data quality continuously.

To monitor data quality, companies need to build a program that audits data at regular intervals, or just before or after data is loaded into another system such as a data warehouse. Companies then use the audit reports to measure their progress in achieving data quality goals and complying with service level agreements negotiated with business groups.

Service Level Agreements. Service level agreements should specify tolerances for critical data elements and penalties for exceeding those tolerances. Service level agreements can help boost confidence in the data warehouse or source systems. For example, some organizations don't allow knowledge workers to access the data warehouse until the team meets data quality levels specified in a service level agreement.

If thresholds are exceeded, however, data quality analysts need to examine the data in question and develop a plan for remedying the situation. In some cases, this may mean fixing data in the source and reloading the data warehouse. At other times, it may mean tweaking thresholds to accommodate business expansion or taking no action at all.

For example, a flu epidemic in December may cause the number of insurance claims loaded into a data warehouse to spike and exceed a predefined threshold. But by talking with business users, the data warehousing team quickly discovers they do not need to take action since the "abnormal" count is accurate.

Service level agreements can help boost confidence in the data warehouse.

Alliance Blue Cross Blue Shield. Missouri-based Alliance Blue Cross Blue Shield built a data warehouse in 1994. Due to defective data, the data warehouse suffered credibility problems and didn't get much use, according to Dalton Holt, data warehouse development manager. To reinvigorate the data warehouse, Holt and his team put together a data quality plan and spent three years fixing a long list of data quality problems.

Once Alliance cleaned up the most egregious problems, the team built a simple program to monitor data quality. The team identified 61 key data elements (KDEs) out of 588 to monitor. They selected the KDEs based on their value in supporting critical business decisions and processes, Holt says. The team defined metrics to assess the quality of each KDE. The monitoring program audits the values of KDEs as they are loaded into the data warehouse and generates a report that measures data quality against limits or tolerances set in conjunction with business users.

The team identified 61 KDEs out of 588 to monitor.

In the first six months of the data monitoring program, Alliance decreased errors by 65.8 percent. The improvements in data quality increased the credibility and usage of the data warehouse, according to Holt. This gave the data warehousing team confidence to build an executive information system that had long been delayed by data quality problems.

Errors decreased by 65.8 percent.

8. Improve Business Processes

As mentioned earlier, preventing data defects involves changing attitudes and optimizing business processes. "A data quality problem is a symptom of the need for change in the current process," says Brad Bergh, a veteran database designer with Double Star, Inc. Improving established processes often stokes political and cultural fires, but the payoff for overcoming these challenges is great.

"A data quality problem is a symptom of the need for change in the current process."
—Brad Bergh

Having a corporate data stewardship program and an enterprisewide commitment to data quality is critical to making progress here. Under the auspices of the CEO and the direction of corporate data stewards, a company can begin to make fundamental changes in the way it does business to improve data quality.

There are several ways to improve business processes and practices:

- Educate. Use newsletters, corporate memos, the news media, and employee and shareholder meetings to communicate to employees the importance of data to the company.
- Train and Reward. Continuously train new and existing data entry people about data standards. More importantly, reward them for improving data quality through bonuses or other incentives. And make sure existing incentives—such as rewards for quantity of calls taken per hour—don't undermine the quest to improve quality.
- Drive Business Impact Close to Home. Source system owners aren't motivated to make changes until data defects materially affect their business. Show business units how their unique data entry practices cost them money and lost business.
- Apply Validations. Prevent errors by building stringent validation checks on data entry fields. Use real-time verification software to prevent CSRs or Web-based customers from entering incorrect addresses.
- Standardize Codes, Rules, and Definitions. Bring together senior managers to agree on standard codes, data definitions, and business rules. Lock the door and don't let them out until they have resolved their differences.

- **Leverage Successes.** Some groups and business units will resist standardization, in some cases for legitimate business reasons. Tackle easy projects first, and leverage these successes to put pressure on holdouts to change.
- **Redesign Systems.** Redesign data models to more accurately reflect the business and better enforce referential integrity.

The above techniques, although not easy to implement in all cases, can help bring a company closer to achieving a strong foundation on which to build an information-based business. The key is to recognize that managing data quality is a perpetual endeavor. Companies must make a commitment to build data quality into all information management processes if they are going to reap the rewards of high quality data—and avoid the pitfalls caused by data defects.

The Role of Data Quality Tools

The Market for Data Quality Tools

A tool is not a silver bullet.

Good data quality is primarily the result of managing people and processes in an effective manner. Technology alone cannot solve a company's data quality problems, but it plays an important role.

For example, after implementing a data quality tool, an online education firm now saves $500,000 annually in postage on 20 million direct mail pieces it sends out each year to customers and prospects. The tool validates addresses in real time for telemarketers and customer service representatives, and it appends extra digits to zip codes so the company qualifies for postal service discounts.

Outside of scrubbing name and address data, however, many organizations today use homegrown tools to automate data quality activities. Typically, companies use SQL to audit data files and custom developed programs to clean and monitor data quality. Many companies have not yet felt the need to purchase a packaged data quality tool to clean or monitor non-name-and-address data. However, this may change as data quality vendors expand their offerings. (See "Emerging Functionality," p. 29.)

"We get customers when they've tried to do it themselves and failed."

Data Quality Vendors. Although the market for data quality tools is small, it is growing. Companies are investing millions of dollars in CRM initiatives, Web extranets, and business intelligence systems, which rely heavily on high quality customer data. These systems expose bad data to a much broader audience. As a result, more companies than ever are beginning to investigate commercial data quality solutions.

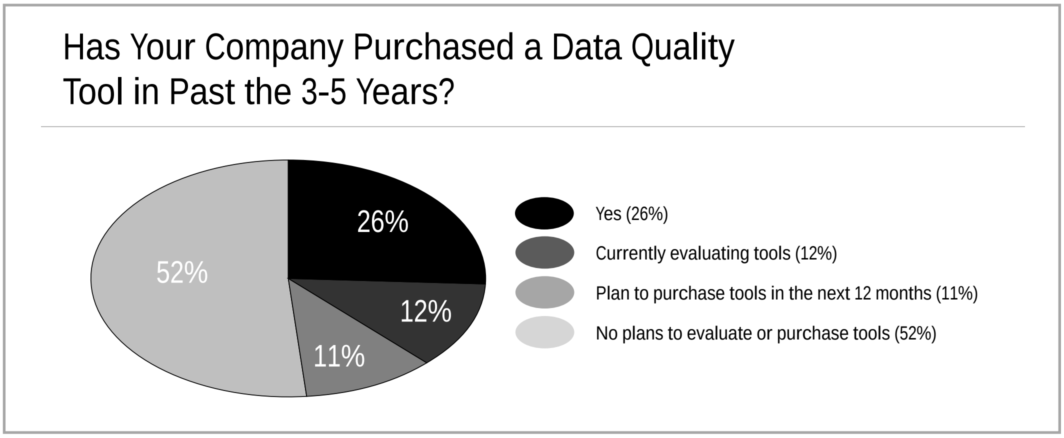


Illustration 16. Only one-quarter of companies have purchased a data quality tool in the past 3-5 years. Based on 647 responses.

"We get customers when they've tried to do it themselves and failed," says a representative from one data quality vendor. "That's when they realize that our tools are really worth the investment."

Market Penetration. According to our survey, slightly more than a quarter of organizations have purchased a data quality tool in the past three to five years. However, almost another 25 percent are evaluating or planning to purchase commercial tools in the next 12 months. (See Illustration 16.)

There are currently about a dozen vendors in the market, mostly privately-owned firms or subsidiaries of larger public companies. Several have sponsored the research for this report. (See page 33 for descriptions of these vendors and their products.)

Customer-Centric Tools. Since their debut in the 1990s, most commercial data quality tools have focused on cleaning up only name and address data. That's because many were spun out of direct mail service bureaus, and were developed to parse, standardize, and de-dupe huge customer lists prior to a direct mail campaign.

Most tools have a mail house ancestry.

Today, most data quality vendors still focus on name and address data, although this is changing. Several vendors—especially those that did not originate in mailing houses or marketing service bureaus—audit and scrub other types of data.

Traditionally, vendors have focused on name and address elements because they are the most volatile fields in corporate databases. "Two percent of a perfectly clean customer database will contain errors each month because of people who have deceased, divorced, or moved," says Peter Harvey, Intellidyn CEO. Rarely do people inform their suppliers of these life changes, he added.

Customer data is very volatile, ever changing.

Consequently, tools vendors view customer data as the sweet spot of the data quality market. Over the years, they have developed robust parsing engines and extensive reference libraries to aid in standardizing data, and built sophisticated algorithms for matching and householding customer records. But many data quality vendors have recently extended their parsing and matching algorithms to handle non-name-and-address data. (See "Emerging Functionality," p.29.)

Data Quality Tool Functionality

CORE CAPABILITIES. Customers today are looking for a range of features in data quality tools. Standardization and verification, available in most commercial tools today, are at the top of the list. Not far behind, customers are looking for tools that define and validate general business rules, which is not something most tools support today, although some are moving in this direction along with ETL tools. Other important features are matching, consolidation, and integration with other enterprise applications, such as ETL tools. (See Illustration 17.)

Although not all vendors offer all features listed above, most offer the following standard features:

- Data Auditing. Also called data profiling or data discovery, these tools or modules automate source data analysis. They generate statistics about the content of data fields. Typical outputs include counts and frequencies of values in each field; unique values, missing values, maximum and minimum values; and data types and formats. Some of these tools identify dependencies between elements in one or more fields or tables, while others let users drill down from the report to individual records.
- Parsing. Parsing locates and identifies individual data elements in customer files and separates them into unique fields. For example, parsers identify "floating fields"—data elements that have been inserted into inappropriate fields—and separate them. For example, a parser will transform a field containing "John Doe, age 18" into a first name field ("John"), last name field ("Doe"), and age field ("18"). Most parsers handle standard name and address elements: first name, last name, street address, city, state, and zip code. More sophisticated parsers identify complex name and address elements, such as DBA (doing business as) or FBO (for the benefit of). Newer parsers identify products, email addresses, and so on.
- Standardization. Once files have been parsed, the elements are standardized to a common format defined by the customer. For example, the record "John Doe, 19 S. Denver Dr." might be changed to "Mr. John Doe, 19

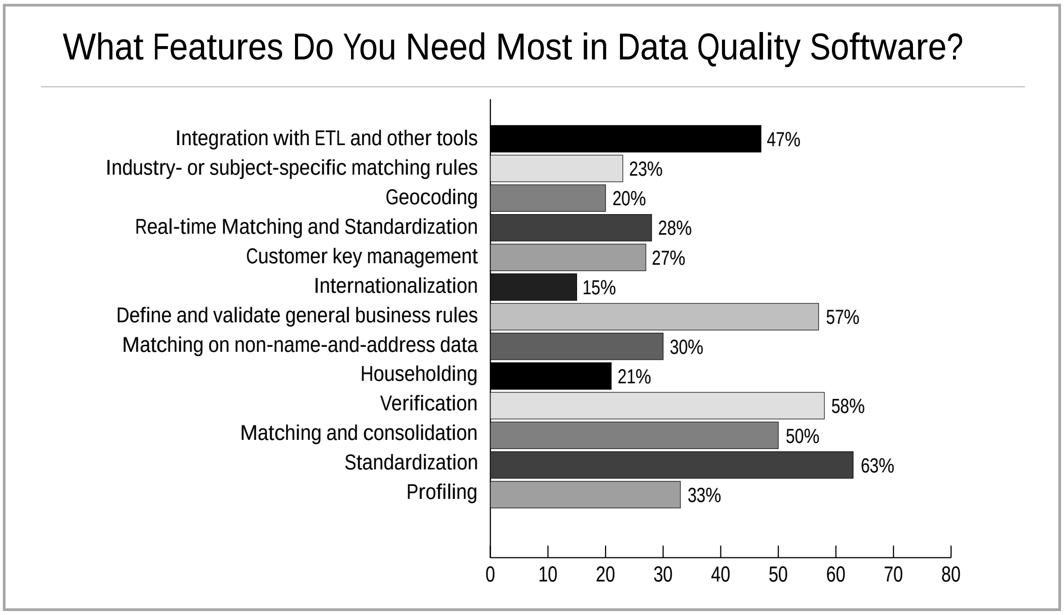


Illustration 17. Based on 632 respondents.

South Denver Drive." Standardization makes it easier to match records. To facilitate standardization, vendors provide extensive reference libraries, which customers can tailor to their needs. Common libraries include lists of names, nicknames, cardinal and ordinal numbers, cities, states, abbreviations, and spellings.

- Verification. Verification authenticates, corrects, standardizes, and augments records against an external standard, most often a database. For example, most companies standardize customer files against the United States Postal Service database.
- Matching. Matching identifies records that represent the same individual, company, or entity. Vendors offer multiple matching algorithms and allow users to select which algorithms to use on each field. There are several common algorithms: (1) key-code matching examines the first few characters in one or more fields; (2) soundexing matches words by their pronunciation; (3) fuzzy matching computes a degree of likeness among data elements; and (4) weighted matching lets users indicate which fields should be given more weight.
- Consolidation/Householding. Consolidation combines the elements of matching records into one complete record. Consolidation also is used to identify links between customers, such as individuals who live in the same household, or companies that belong to the same parent.

The above capabilities can help an organization boost the accuracy levels of its customer files into the high 90 percent range, especially if these processes are automated to run against both source files and downstream systems on a regular basis.

Companies need tools that handle a broader range of issues.

DOWNSIDES. Unfortunately, vendors' historical focus on name and address data has pigeonholed them in the eyes of customers. Names and addresses represent a small fraction of the total number of fields in a corporate database. To justify their investments, companies need data quality tools to provide broader functionality and handle more complex problems. (See Illustration 18.)

Customers are dismayed by the high cost of data quality tools.

In addition, many potential customers have been dismayed by the high cost of some tools ($100,000+) and sluggish performance when run on Windows servers. The pricetag for some tools causes many organizations to spend their precious capital dollars on databases, ETL tools, and analytical applications, and limp along with home-grown data quality software. Customers also want easier-to-use tools that integrate better with other products.

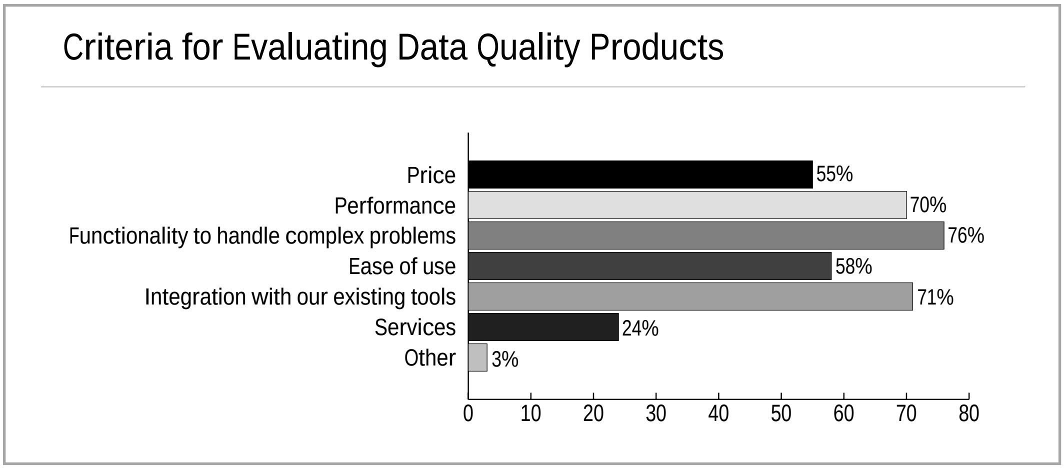


Illustration 18. Based on 143 respondents who are evaluating data quality software.

VERSUS ETL. Some companies mistakenly believe that ETL tools offer much the same functionality as data quality tools. Unfortunately, ETL vendors have not gone out of their way to squash this myth.

Although ETL tools can perform some cleansing operations, such as mapping between code sets, and can apply procedural logic to catch critical data integrity problems, the products don't provide many basic data quality capabilities, such as auditing, parsing, standardization via extensive reference libraries, matching, verification, and householding. In fact, ETL tools assume the data they receive is clean and correct! Otherwise, their mapping and transformation programs do not output the correct data. Hence, the "code, load, and explode" phenomenon.

ETL tools don't provide many basic data quality capabilities.

As a result, some ETL vendors recently have partnered with or acquired data quality vendors. By combining ETL and data quality toolsets, vendors can offer customers a more comprehensive data integration solution. From a customer perspective, it is difficult to understand why ETL and data quality tools have existed in separate but parallel markets. Fortunately, customer requirements are now bringing these two toolsets closer together.

However, there is a slight tradeoff when combining ETL and data quality tools. Second-generation tools, such as Informatica's PowerMart or Ascential Software's DataStage, work on one record at a time. This limits the accuracy of some data cleansing operations, such as matching and householding, which achieve the highest match rates by comparing large numbers of records in one or more databases.

Therefore, except for simple validation and standardization processes, it is still wise to perform data quality operations in batch prior to kicking off an ETL job. In many cases, the ETL tool can be used to kick off and monitor these batch operations. (See Illustration 19.)

However, ETL tools will become a more important option as more companies update data warehouses in near real time using trickle feed connections to transaction-based middleware. Since these processes feed data to ETL engines on a transaction basis, ETL tools (either alone or with data quality tools) will need to assume greater responsibility for checking the validity of incoming data.

Emerging Functionality

Data Integration. Not surprisingly, data quality vendors have been listening to customers and adding functionality to address significant issues. Some vendors now offer lower-priced tools that are easy enough for a business person to learn and use. Also, vendors are improving performance and throughput by running on higher-end systems and improving the efficiency of matching algorithms.

ETL-Data Quality Integration

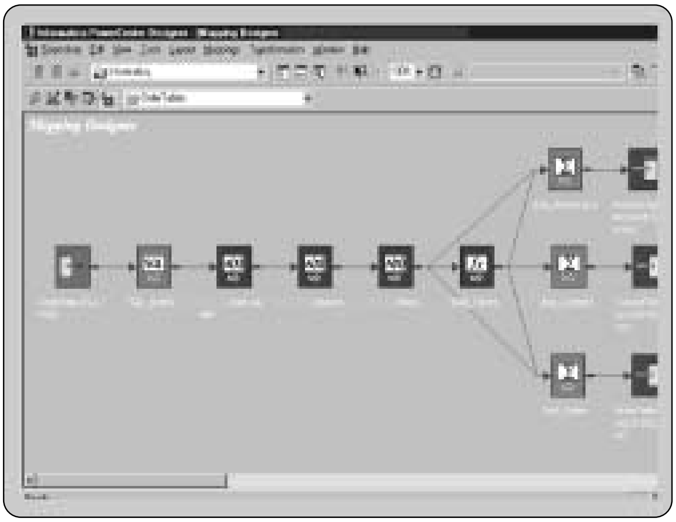


Illustration 19. The screen shot above depicts a data flow diagram from Informatica's PowerMart ETL tool. The steps before the branch represent data cleansing operations.

Vendors are viewing their technology as a foundation for data integration.

More interestingly, vendors are beginning to view their technology from a broader perspective, as a foundation for enterprise data integration. For example, some are leveraging their fuzzy matching technology to create customer key codes for tracking customers over time and across systems.

Others are building data integration hubs that serve as a central repository of clean, standardized records and through which all shared system data flows, getting validated, cleaned, and matched in the process. Still others are developing tools for searching and assembling product catalogs, facial recognition systems, and even terrorist tracking applications.

With all the innovation in the market, a number of features will emerge to become standard fare, if they are not already:

- Non-Name-and-Address-Data. Vendors are developing parsing algorithms to identify new data types, such as emails, documents, and product numbers and descriptions. They are also leveraging standardization and matching algorithms to work with other data types besides names and addresses.
- Internationalization. To meet the needs of global customers, vendors are adding support for multi-byte and unicode character strings. They are also earning postal certifications from the U.S., Canada, Australia, and Great Britain, and adapting to address reference files in other countries.
- Data Augmentation. While the USPS database can add zip+4 and other fields to a record, some vendors now can augment addresses with geocode data (i.e. latitude/longitude, census tracts, and census blocks) and demographic, credit history, and psychographic data from large information service providers such as Polk, Equifax, and Claritas.
- Real-Time Cleaning. Traditionally, data quality tools clean up flat files in batch on the same platform as the tool. Most vendors now offer tools with a client/server architecture so that validation, standardization, and matching can happen in real time across a local-area network or the Web.

- **Customer Key Managers.** Some vendors are marketing internal match keys as a convenient way to associate and track customers across time and systems. If desired, companies can use the keys to leave customer records in the source systems in their original (unclean) state. This is important if the data can't be moved or integrated or if it is owned by another company. Some companies create and manage the customer keys at the vendor site (Axciom and Experian), while others let users create and manage their own keys. Outsourced keys track customer activity across a multitude of third-party databases and applications, providing a rich, historical record of a customer and demographic profile. Insourced key applications track far fewer customer data points, but are less expensive and let companies manage customer keys in house.
- **Integration With Other Tools.** Many vendors offer a software developer's kit (SDK) which makes it easy for ETL and application vendors to embed data cleansing routines into their applications. These embeddable routines will grow in importance as companies begin trickle feeding data warehouses and ODSs that support analytic and operational applications. Tools that are truly integrated will exchange rich sets of meta data. For example, a data quality tool should be able to exchange or map its customer keys with the surrogate keys that many ETL tools assign to customer records.
- **Data Integration Hubs.** Most companies have a spider web of interfaces that propagate data and errors among disparate systems. Data integration hubs channel these interfaces into a central repository that maps incoming data against a clean set of standardized records. Source systems are updated as incoming data is parsed, standardized, and matched against repository data in a two-way exchange of information.

Tool ROI. As data quality becomes a more significant obstacle in establishing a single view of customers and gaining accurate, reliable data for decision making, companies will begin to see the wisdom of investing in data quality tools. Almost two-thirds of survey respondents who have purchased data quality tools have currently broken even or better on their investments. (See Illustration 20.) We suspect the ROI for these tools will only improve as the toolsets become more robust and high-performance, and offer expanded functionality.

A majority of users have gotten a break-even or better payback.

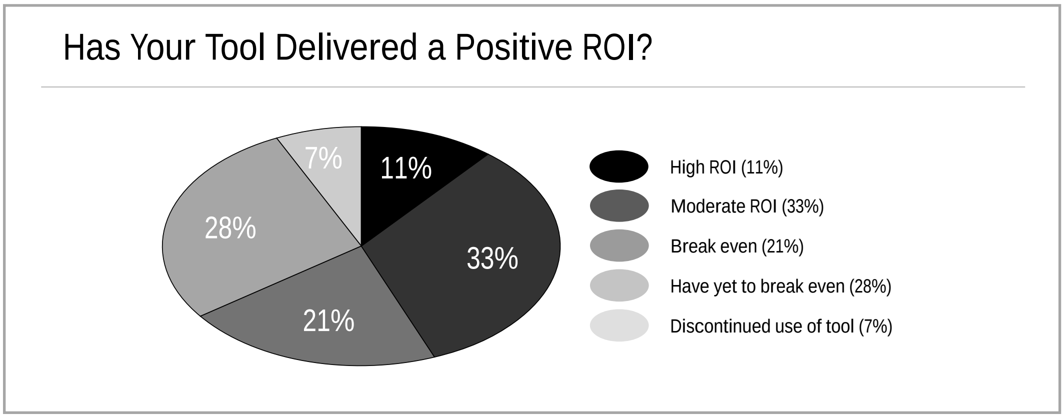


Illustration 20. Based on 148 respondents who have purchased data quality tools.

Overall, data quality tools are worth evaluating, especially if your company is embarking on a strategic project involving customer data. In the next few years, companies will be turning to commercial data quality tools to handle a broader array of data quality issues besides reconciling name and address data. And tools will make it easier for organizations to justify their investments in data quality processes and procedures.

Conclusion

The Vital Role of Data. As we enter the 21st century, we are still at the dawn of the Information Age. Data and information are now as vital to an organization's well being and future success as oxygen is to humans. Without a fresh supply of clean, unpolluted data, companies will struggle to survive and thrive.

Although executives say they view data as a critical resource, few have provided critical leadership or funding.

Although executives say they view data as a critical resource, few provided leadership or funding to establish an enterprisewide program to preserve the value of the data under their management. However, new strategic initiatives, such as CRM, business intelligence, supply chain management, and e-business, require the integration of data from diverse systems and external sources. These initiatives place a premium on high quality data and quickly expose problems to a much broader audience including customers, suppliers, and investors. They are sounding a wake-up call to top executives who are learning the hard way that data quality problems can sabotage the best laid strategies

The problem with data is that its quality quickly degenerates over time. Experts say 2 percent of records in a customer file become obsolete in one month because customers die, divorce, marry, and move. In addition, data entry errors, systems migrations, and changes to source systems generate bucket loads of errors. More perniciously, as organizations fragment into different divisions and units, interpretations of data elements mutate to meet local business needs. A data element that one individual finds valuable may be nonsense to an individual in a different group.

Maintaining high quality data is not beyond the means of any company.

The Goal is Achievable. The good news is that achieving high quality data is not beyond the means of any company. The keys are to treat data as a strategic corporate resource, develop a program for managing data quality with a commitment from the top, and hire, train, or outsource experienced data quality professionals to oversee and carry out the program. Then, it is critical for organizations to sustain a commitment to managing data quality over time and adjust monitoring and cleansing processes to changes in the business and underlying systems.

The only truly effective data quality programs must be initiated by the CEO or board and promulgated throughout the organization via a group of senior-level managers acting as data stewards. These stewards define the mission, goals, and objectives of the program and then monitor compliance and troubleshoot problems.

Technology can help automate processes that teams put in place to clean and then monitor data quality. Today, most companies use manual methods or homegrown programs to manage data quality. Commercial data quality tools, if used, are employed primarily to clean name and address data for direct mail and CRM applications. Eventually, however, more companies will implement commercial data quality products as those tools broaden their focus and address the full range of an organization's data quality requirements.

The time to start is now!

Now, the important step is to get started. Data quality problems cost your organization perhaps hundreds of thousands of dollars each year and, more significantly, slowly undermine customer confidence and loyalty. Data is a vital resource and it's time your organization starts treating it that way, if it doesn't already.

SPONSORS

Arkidata Corporation
1100 W. 31st Street, Suite 520
Downers Grove, IL 60515
630.795.2600 or 877.947.3282
Fax: 630.963.9653
Email: mail@arkidata.com
Web: www.arkidata.com

Arkidata Corporation is a full-service information integration company that delivers data quality solutions through services and software, utilizing their unique business-rule driven data cleansing methodology and proprietary technology, Arkistra™. The Arkistra application, used in all projects, provides a comprehensive solution designed to solve complex data quality problems. Founded in 1997, Arkidata has formed alliances with major consulting firms and systems integrators, and has served a diverse group of Fortune 500 companies. Successful engagements include projects for Kimberly-Clark, Raytheon, Dun & Bradstreet, MediaOne, Sprint, Verizon, and Continental Airlines. These projects have included historic data cleansing and ongoing information integration, ensuring information quality and the integration of information across multiple disparate sources into a single useable format.

DataFlux Corporation
4001 Weston Parkway, Suite 300
Cary, NC 27513
919.674.2153 or 877.846.3589
Fax: 888.769.3589
Email: info@dataflux.com
Web: www.dataflux.com

Incorporated in 1997, DataFlux is a North Carolina-based software company that provides leading-edge technology for data cleansing, data augmentation, data consolidation, and data integration. The versatile DataFlux products are designed to significantly improve the accuracy, consistency, and usability of an organization’s critical data, enhancing the effectiveness of data-driven applications and initiatives such as data warehousing, e-commerce systems, data mining, customer information systems, sales force automation, marketing databases, customer/prospect profiling, and list management. DataFlux is a wholly owned subsidiary of SAS Institute, the world’s largest privately held software company. For more on DataFlux, visit: www.dataflux.com.

DataMentors, Inc.
13153 N. Dale Mabry Hwy
Suite 100
Tampa, FL 33618
813.960.7800
Fax: 813.960.7811
Email: contact@datamentors.com
Web: www.datamentors.com

DataMentors is a full-service data quality solutions company, providing a comprehensive suite of completely customizable data validation, transformation, and database building products. The company offers a fully modular relationship matching and linking system that cleanses, organizes, standardizes, and households data, offering superior efficiency and unprecedented accuracy. DataMentors also provides data profiling for non-name-and-address analysis. Clients choose on-site installation or service bureau processing, or have the opportunity for in-house migration for any solution. Rapid implementation increases customer profitability and ROI while decreasing marketing costs. For more information, call 813.960.7800 or explore the possibilities at www.datamentors.com.
Data Solutions...Clean and Simple.

Sagent Technology, Inc.
800 West El Camino Real, Suite 300
Mountain View, CA 94040
650.493.7100
Fax: 650.815.3500
Email: info@sagent.com
Web: www.sagent.com

Sagent’s suite of enterprise business intelligence solutions enables companies to measurably impact their business by implementing highly successful customer relationship and financial management initiatives. Through Sagent’s powerful enabling technologies, organizations can easily and rapidly turn company data into relevant information that can be used for effective decision making, analysis, and reporting. Information can be extracted from multiple sources (internal and external), optimized for decision support, and delivered in a customized format for Web-based or client applications. Even the most complex analytic application can be developed in weeks, not months.

SAS Institute
SAS Campus Drive
Cary, NC 27513
919.677.8000
Fax: 919.677.4444
Email: software@sas.com
Web: www.sas.com

SAS is the market leader in business intelligence, offering software and services spanning the full data warehousing process and high-end analytics. SAS drives the intelligent enterprise, bringing greater effectiveness and efficiency to critical operations such as CRM, supplier relationship management, and strategic performance management. SAS solutions are built on a proven Intelligence Architecture that is open and scalable, allowing seamless integration of processes and platforms. Software from SAS, the world’s largest privately held software company, is used at more than 37,000 business, government, and university sites. Customers include 90 percent of the Fortune 500 (98 of the top 100 companies). For 25 years, SAS has given customers The Power to Know™. To learn more, visit www.sas.com or call your local SAS office.

Vality Technology, Inc.
100 Summer Street, 15th Floor
Boston, MA 02110
617.338.0300
Fax: 617.338.0368
Email: info@vality.com
Web: www.vality.com

Vality Technology is the recognized leader in enterprise data quality and integration and provides customers and partners with data standardization and matching software and consulting services. Its customers and partners are Global 5000 corporations in finance, healthcare, insurance, manufacturing, pharmaceuticals, retail, telecommunications, energy, and utilities. They include Aetna U.S. Healthcare®, IBM®, Marks & Spencer®, NatWest®, www.onlinesuppliers.com corp.™, Telstra®, and UPS®. These companies rely on Vality’s patent-pending technology to develop and deploy enterprise data quality management solutions that ensure the ROI of Customer Relationship Management (CRM), business intelligence, e-Commerce, and Supply Chain Management (SCM) initiatives. Vality was recently listed in *Boston Magazine’s* Best Places to Work; selected one of the 15 Stars of e-Commerce by eCom Advisory; and named a Top 500 software company by *Software Magazine.*

Mission

The Data Warehousing Institute™ (TDWI), a division of 101communications, is the premier provider of in-depth, high quality education and training in the data warehousing and business intelligence (BI) industry. TDWI is dedicated to educating business and information technology professionals about the strategies, techniques, and tools required to successfully design, execute, and maintain data warehousing and business intelligence projects. It also fosters the advancement of research and contributes to knowledge transfer and professional development of its Members. TDWI sponsors and promotes a worldwide membership program; annual educational conferences; regional educational seminars; onsite courses; solution provider partnerships; awards programs for the best practices and leadership in data warehousing, business intelligence, and other innovative technologies; resourceful publications; an in-depth research program; and a comprehensive Web site.

Membership

As the data warehousing and business intelligence field continues to evolve and develop, it is necessary for information technology professionals to connect and interact with one another. TDWI provides these professionals with the opportunity to learn from each other, network, share ideas, and respond as a collective whole to the challenges and opportunities in the data warehousing and BI industry.

Through Membership with TDWI, these professionals make positive contributions to the industry and advance their professional development. TDWI Members benefit through increased knowledge of all the hottest trends in data warehousing and BI, which makes TDWI Members some of the most valuable professionals in the industry. TDWI Members are able to avoid common pitfalls, quickly learn data warehousing and BI fundamentals, and network with peers and industry experts to give their projects and companies a competitive edge in deploying data warehousing and BI solutions.

TDWI Membership includes more than 4,000 Members who are data warehousing and information technology (IT) professionals from Fortune 1000 corporations, consulting organizations, and governments in 45 countries. Benefits to Members from TDWI include:

- *Quarterly Journal of Data Warehousing*
- *Biweekly FlashPoint electronic bulletin*
- *Quarterly Member Newsletter*
- *Annual Data Warehousing Salaries, Roles, and Responsibilities Report*
- *Quarterly Ten Mistakes to Avoid series*
- *Annual Best Practices in Data Warehousing Awards summaries*
- *Semiannual What Works: Best Practices in Data Warehousing and Business Intelligence corporate case study compendium*
- *TDWI Marketplace Online comprehensive product and service guide*
- *Annual technology poster*
- *Periodic Executive Summary of the Industry Study*
- *Periodic research report summaries*
- *Special discounts on all conferences and seminars*
- *Fifteen-percent discount on all publications and merchandise*

Membership with TDWI is available to all data warehousing, BI, and IT professionals for an annual fee of $245 ($295 outside the U.S.). TDWI also offers a Corporate Membership for organizations that register 5 or more individuals as TDWI Members.

General Membership Inquiries:

Membership
The Data Warehousing Institute
5200 Southcenter Blvd., Suite 250
Seattle, WA 98188
Local: 206.246.5059, ext. 113
Fax: 206.246.5952
Email: membership@dw-institute.com
Web: www.dw-institute.com

Corporate Membership Inquiries:

Local: 206.246.5059, ext. 108
Email: dsmith@dw-institute.com

THE DATA
WAREHOUSING
INSTITUTE™