05 Third Quarter

TEN MISTAKES TO AVOID

When Building a Data Quality Program

David Loshin

tdwi
THE DATA WAREHOUSING INSTITUTE
A DECADE OF EXCELLENCE

Ten Mistakes to Avoid When Building a Data Quality Program

David Loshin

FOREWORD

Enterprise quality improvement programs are rapidly becoming more visible as more reports and articles describe the value placed on high-quality information. C-level executives, concerned with regulatory compliance, are finding themselves personally accountable for both the levels and processes associated with data governance and quality assurance. Once a tedious chore relegated to the back office, data quality is now viewed as an organizational necessity.

Data quality improvement involves more than just name and postal address correction. The complexity and impact of the data quality conundrum grows in proportion to the amount of data we capture, store, manage, review, aggregate, summarize, etc. Yet data quality initiatives are frequently doomed when mistakes are ignored.

ABOUT THE AUTHOR

David Loshin is the president of Knowledge Integrity, Inc, a consulting and development company focusing on customized information management solutions, including information quality consulting, information quality training, business intelligence solutions, metadata management, data standards management, and business rules solutions. David is internationally recognized as an expert in Information Management. He serves on the Editorial Board of DM Review *magazine, provides monthly columns for* DM Review, *serves as a featured columnist for* The Data Administration Newsletter *(www.tdan.com), and is the co-editor of the B-EYE-Network's* Business Data Integration *newsletter. David has developed strategies and implementation plans for data quality improvement and business intelligence projects in the financial, insurance, social services, energy, retail, government, pharmaceutical, and non-profit sectors.*

ONE

Using Events and Anecdotes as the Sole Business Drivers

Effectively communicating the value of data quality to senior managers a huge challenge facing potential data quality programs. Demonstrating how specific data quality failures negatively impact the achievement of business objectives is often not enough.

Anecdotes and horror stories are more likely to resonate with upper management because they present an opportunity to (re)act, i.e., correct the data and "be a hero." The problem with such cataclysmic events is that the focus on improved data quality only lasts until the next crisis. While addressing the immediate need solves problems on the surface, it does not help the latent problems, allowing them to fester until the next big crisis.

Anecdotes and crises are good wake-up calls, but the most effective business cases demonstrate a positive return on an enterprise investment in overall data quality improvement. The following list will help you develop a believable value proposition demonstrating how specific data quality improvements lead to more efficient achievement of business objectives:

- Identify the key business impacts associated with poor data quality
- Associate specific costs to specific data flaws and aggregate the costs as a function of how often the flaws occur
- Quantify the impacts in relevant business terms
- Assess the cost of eliminating the sources of those flaws
- Identify key data quality metrics for continuous monitoring and reporting

Take care when evaluating improved data quality ROI because executives are not likely to respond to business cases that rely on vague "industry numbers" relating to operating budgets or revenue percentages. Developing a data quality business case requires serious investigation, discussions with business subject matter experts, and a lot of conservatism.

A clear business case may be less fascinating than a set of horror stories, but a senior executive might be hard pressed to ignore a convincing business case based on supportable numbers.

TWO

Applying Value Judgments to Information

It's easy to develop a habit of referring to data as *good* or *bad*. These terms aren't meant to pass judgment over the data, but rather convey a perception of how well the data meets our business needs. A byproduct of this habit, however, is to unintentionally associate terms meant to describe data with the people managing the data. In other words, when we say that a database record is *bad*, the database doesn't care, but the manager of that database might take offense. Consequently, data quality initiatives are often met with resistance—data owners feel that exposing *bad* data reflects poorly on their personal performance. The result: information hoarding and turf protectiveness.

To avoid this problem, *depersonalize* the characterization of information quality in terms of meeting business data expectations. Assert business expectations in clear and concise statements that can be used to assess business rule compliance thereby removing the value judgments from data. This in turn provides a way to quantify information validity using business-relevant metrics that inspire openness instead of protectiveness.

Recently, our company participated in a data quality assessment evaluating a relatively large contact data set (e.g., address, e-mail, telephone number). Some individuals within the business team accusingly described the data as being *bad*, which prompted the system team to not only claim the opposite, but to also subtly begin restricting access to some of that data—a classic example of turf protectiveness. Our approach objectively profiled the data set using statistical techniques, and then posed non-judgmental questions about potential anomalous behavior within the data.

Identifying clearly defined business rules with specific impact caused members of both teams to relax their positions. The business people conceded that the data flaws were not as critical as previously perceived, while the system team agreed that improvements could be made based on the defined rules. By eliminating the value judgments, the client was able to find common ground for data quality improvement.

THREE

Failing to Evolve from a Reactive to a Proactive Environment

Data crises imbue an organization with a sense of panic. In a *reactive* environment, the offending data is identified, corrected, and interrupted/corrupted processes are rolled back and/or restarted—and everyone sighs with relief until the next crisis erupts. This *reactive* process addresses data crises as they occur, but, in terms of a reasonable process, leaves a lot to be desired. The mistake lies in failing to transition from an environment that *reacts* to problems into a mature, *proactive* one. A *proactive* environment measures conformance with data quality expectations early in the information flow, catching flaws before they become problems.

A structured approach to data quality assessment assumes information consumers and business policies influence rules definitions. That way, when data quality is assessed, we can identify how critical problems impact business objectives. Then we prioritize solution development based on an ROI calculation. The knowledge derived can fix a process and eliminate the problem source.

While this approach is more manageable, the organization is still reactive in respect to data quality events. Even in a well-structured reactive environment, data flaws are addressed *after* the problem has manifested. Yet an organization's current business rules approach can evolve from a reactive to a proactive environment.

Implement business rules earlier in the information processing chain. You can measure data set conformance with business rules long before a corresponding impact materializes. Gauge the compliance of day-to-day information performance metrics fed by the results of proactive monitoring and eliminate problems early on.

Transform into a proactive environment by applying your organization's business rule principles:

- Identify critical business data quality expectations
- Assert those expectations as a set of business rules
- Measure the compliance of data with defined business rules to provide key data quality metrics

Not only will you be able to demonstrate the high quality of your data, but you can also provide auditability of your process. Anyone familiar with regulatory compliance challenges, i.e., Sarbanes-Oxley, will understand how auditability is key to many compliance applications.

FOUR

Buying Software First

One of the first things an organization will do when setting up a data quality program is acquire a data quality tool. Buying a tool too early in the process is indicative of the following:

- **A *reactive* environment.** Once senior management recognizes there is a data quality problem, he/she scrambles to put a solution in place as quickly as possible.
- **Technology-driven data quality.** There is a desire to *fix* noncompliant data instead of eliminating its introduction in the first place.

It is common to purchase a tool and have it sit on the shelf in its shrink wrap for months. Although data quality tools are critical components of a data quality program, you must first question the motivation for purchasing a tool, and then the process itself. What is the improvement potential in terms of contributing to the program's effectiveness?

The impulse to buy a product is driven by the assumption that technology will fix the problem out of the box. You should complete the strategic and tactical groundwork for your data quality program before you purchase a data quality tool. Not only are there different classes of data quality tools, these tools will be most effective after you've established a data quality strategy and developed rules for customizing the tools.

Need more reasons to hold off on tool acquisition?

- Within each class of tools, each product has strong and weak points with respect to addressing any specific set of data quality issues.
- The acquisition process could consume up to six months. For most reasonably sized organizations tool acquisition becomes a project in its own right (i.e., writing a request for proposal, assembling an assessment team, sitting through numerous vendor presentations, installing and evaluating products, etc. until a purchase decision is made—or, in some cases, deferred).
- Even if the acquisition process could be accelerated, there is also a need to train users while simultaneously establishing policies and procedures for general product usage across the enterprise. Focusing attention on bringing a product in without ensuring the proper expertise, operations knowledge, and operating environment are available for using the product can turn the software into shelfware.

Continued on page six.

Buying Software First, continued...

Successful organizations take two important steps before acquiring a data quality tool:

- They perform a business needs assessment to evaluate the kind of data quality problems that exist across the organization. Team members collect the various needs and desires from the organization and prioritize them according to the most critical technology needed to express the business requirements for the acquisition.
- They develop policies and procedures for using the technology in order to deploy resources as soon as the purchase is executed. Not only does this eliminate the shelfware risk, it facilitates the development of an enterprise data quality competency center.

FIVE

Ignoring the Data

There are people who insist issues associated with poor data quality are strictly related to poor processes stemming from problems that lie upstream. As a consequence, the focus on data takes a back seat to assessing processes and workflows, and actual data evaluation is limited to a very small random selection of records from a single table. These "process experts" introduce simplistic metrics as a way to characterize a data quality assessment, often in the absence of any business-oriented context.

Today, it is not unusual for organizations to control gigabyte- and terabyte-sized data systems with an ever-increasing volume of managed data and an ever-accelerating rate of data collection. Automatic analysis exposes some issues but it obscures others for manual reviewers. Therefore, without a comprehensive data analysis, it would be impossible to get a handle on the existence, scope, and scale of the potential data quality issues. You cannot improve the quality of data unless you can understand the kinds of issues that exist.

When coupled with effective analysis and review procedures, data quality techniques implemented with tools (e.g., data profiling or standardization and matching) do much more than frequently distribute column data values. With data profiling, an analyst can discover information flaws that will prevent the achievement of business objectives. Examples of flaws include: embedded relational structure, functional dependencies, and semantic reference differences.

A comprehensive data analysis accelerates the process of identifying data quality problems and defines rules for fixing those problems. Incorporate data quality assertions into automated monitoring and auditing applications to simplify both the discovery and mitigation of introduced data flaws. Organizations should deal with data quality proactively instead of reactively.

SIX

Not Accounting for Organizational Behavior

Without a proper understanding of how people behave within the system, no technology in the world will eliminate data quality problems.

Some common problems that you may encounter include:

- Without the cooperation of upstream system owners, data warehousing managers are often helpless when controlling the quality of incoming data. Stricter data quality needs at the data warehouse level require additional resource allocation from upstream managers. Unfortunately, upstream managers may perceive this as an imposition because their applications may not directly benefit from the desired improvements.
- Finding flaws in the data quality from a set of operational processes is likely to expose inefficiencies with the individuals associated with those processes. For example, finding a large degree of incomplete insurance claims records reflects poorly on the people transcribing those claims. A person's natural reaction to a data quality assessment is to cover up any potential personal performance issues instead of allowing them to be exposed (as well as any consequences that might follow).
- Data collected by inbound call center personnel can reverberate across multiple operational and analytical applications. These individuals are likely to be low paid and rewarded based on volume and not accuracy.

Being aware of these human behaviors can be extremely important when building your data quality program. Be proactive about potential issues before they become full-blown problems. Use strategies that promote positive feedback and incentives. Ultimately, accounting for human behavior will support the technical end of your enterprise data quality program.

SEVEN

Failing to Standardize and Manage Master Reference Data

When someone uses the wrong terms or says one thing while meaning another, humans are quite good at resolving the ambiguity. Unfortunately, data systems don't have the ability to handle ambiguity—they interpret everything literally.

Data quality suffers if we fail to define business terms precisely or accurately. The mistake occurs when we get out of the habit of being both precise and accurate when standardizing how common business terms and their corresponding data element representations are defined and managed.

Frequently, our clients tell us stories about the many hours they have spent trying to agree on definitions of (what are perceived to be simple) concepts such as *customer, product, supplier,* etc. The absence of a process for data standardization and the inability to capture and manage the results of that process allow dissipation of the corporate knowledge that ultimately drives the definition of data quality expectations.

Alternatively, a process that encourages structured collaboration between subject matter experts and information architects provides value in two different ways. It not only establishes a common vocabulary and grammar for clarifying business definitions, but it also directs the framework for centralizing those definitions within a semantic metadata framework. From there, we can express data quality expectations as straightforward assertions against which the actual data sets are assessed when quantifying business rule conformance. Additionally, we can capture data quality assertions as corporate knowledge managed within the enterprise metadata framework.

EIGHT

Isolating Data Quality in the IT Department

Poor data quality is a chimera-like problem because its manifestation relates to a business context, yet the failures occur within a technical context. The result is that the first line of defense is typically the IT department, which essentially means data quality initiatives are spearheaded by technologists. And while technologists are good at developing technical solutions, they are probably less skilled at understanding problem prioritization based on corresponding business impacts. Since data quality is primarily a business issue, IT staff members are the wrong people to place in charge of a data quality program. The result is that the business cases evolved to support a data quality initiative are often technology-heavy, and consequently, focus on the purchase and management of tools instead of the facilitation of measurable business-relevant improvement. Even when the program is approved, a focus on tools and technology does not provide any business impact assessment or solution deployment prioritization. More importantly, it perpetuates the notion that data quality improvement belongs to a *cost center* and not a *profit center*.

Adjust thought processes related to data quality ownership: if the data flaws reflect business impacts related to noncompliance with business expectations, the rules that assert those expectations should be owned by the business client. The IT department can then participate in deploying the tools and methods used to identify nonconformance, and then standardize and remediate problems. The data quality improvement process opens up opportunities for collaboration between IT and business constituents.

NINE

Not Securing the Proper Expertise for Knowledge Transfer

Developing a data quality program is a strategic undertaking—its success depends on having both business and technical expertise. The roles are complicated by the fact that a large part of data quality management, especially at the enterprise level, is advisory. Similarly, there is an expectation that as soon as a data quality program initiates, there should be some visible improvement to the data. Close coupling of process tools and methods adds further complexity. This poses a potential quandary since we assume the data quality manager has responsibility for some action without necessarily having either the knowledge or authority to make it happen. Not only does this contribute to the feeling that the problem is overwhelming, but it also provides no perceivable place to begin.

The mistake is not bringing in the proper agents of change to help the program get off the ground. When establishing a data quality program, implement the following:

- At the program's initiation, hire professionals with experience in managing data quality projects and programs. These individuals will identify opportunities for tactical successes that together contribute to the program's strategic success.
- Engage external experts to help jump-start the improvement process. This reassures your team that your problems are not unique and will allow you to learn from others' best practices.
- Exploit the advisory role and use internal procedures to attach responsibility and accountability for data quality improvement to the existing information management authority.
- Don't forget usage training in policies and procedures—especially in the use of acquired tools.

Data quality improvement is a process that integrates business acumen, high-tech tools, and well-defined processes. While you may think the problems you've encountered are unique, be assured that they are similar to problems encountered in many other organizations in both the concrete and abstract sense.

TEN

Failing to Build an Enterprise Data Quality Center of Excellence

Eventually, money and resources will be invested into purchasing tools and platforms for your data quality improvement program. However, expectations for ROI are not often met when these acquired assets aren't used wisely. In a sea of data, even the best software can turn into an anchor instead of a life preserver when placed in untrained hands.

On the other hand, clever staff members are often able to apply their knowledge to exploit tools and technology in innovative ways. Still, even the most brilliant approach to problem solving becomes useless when the employee who developed the approach walks out the door. The departure of a quality professional often spells doom for his/her innovations.

Being unable to capture best practices and innovations and transfer them across the enterprise is a mistake made by even the most enlightened organizations. For data quality, the impact of this deficiency is multiplied when the knowledge associated with the tools and techniques is not consolidated within a *center of excellence*.

A center of excellence is an enterprise group responsible for deploying corporate data quality strategy. This includes defining guiding principles, helping to assess business needs, recommending tools acquisition, creating the processes to make the best use of those tools, and providing a means for sharing experience in data quality improvement. The major benefits of establishing a center of excellence include:

1. Standardizing the methodology and tools used for addressing particular problems
2. Achieving economies of scale in software and hardware acquisition
3. The provision of a service model for data quality improvement
4. Amortizing investments in program development across multiple divisions or business units
5. The consolidation and documentation of best practices performed across the enterprise, allowing everyone to benefit from common experience
6. Establishing a forum for developing and agreeing to data standards
7. Coordinating and synchronizing professional training in both the use of tools and methods
8. Reducing overall project management costs

ABOUT TDWI

The Data Warehousing Institute™ (TDWI), a division of 101communications, is the premier provider of in-depth, high-quality education and research in the business intelligence and data warehousing industry. TDWI is dedicated to educating business and information technology professionals about the strategies, techniques, and tools required to successfully design, build, and maintain business intelligence and data warehousing solutions. It also fosters the advancement of business intelligence and data warehousing 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 industry leadership, resourceful publications, an in-depth research program, and a comprehensive Web site.

5200 Southcenter Blvd., Suite 250

Seattle, WA 98188

206.246.5059

Fax: 206.246.5952

info@tdwi.org

www.tdwi.org