SEPTEMBER 2014

TDWI E-Book

Data Quality Challenges and Priorities

Sponsored by:

sas

tdwi.org

tdwi

ADDRESSING TODAY'S TOP DATA QUALITY ISSUES

Maintaining data quality has always been a top issue for enterprises, but with changing data needs and business environments—including big data, unstructured data, and data governance—it's never been more challenging. We look at the top issues that enterprises are asking about data quality with Anne Buff, business solutions manager and thought leader for SAS Best Practices.

TDWI: How are industry leaders using data quality to advance business strategy?

Anne Buff: Organizations that design their data management strategy within the context of overarching corporate initiatives are leading their industries, often with large gaps. While there are many great data quality best practices we can learn from these companies, they often share three common elements in their approach:

Designed process. Data quality does not have a one-size-fits-all template—not even within an organization. Designing data quality rules, policies, and procedures around the needs and culture of the business is essential for buy-in and long-term support from the organization.

Business metrics. Metrics-based measurement is an understood management success factor. When it comes to successful data management, though, it is imperative that metrics are business based, not technology based. Data management metrics should have specific, measurable business outcomes and articulate value in at least one of the following areas: increased productivity/efficiency, regulatory compliance, reduced cost/complexity, and decreased risk. Simply put, executives listen when programs make money, save money, or keep them out of jail.

Enterprise view. Although the scope of management matters when governing data, organizations that maintain or are working toward a holistic view of enterprise data rather than maintaining individual data silos are making far greater strides in advancing business strategy. The streamlined, cross-functional capabilities gained from the comprehensive view are fundamental for faster innovation, growth, and development.

Does data quality require data stewardship and data governance?

Data quality initiatives can be successful without data stewardship or data governance, but when completed as ad hoc tasks or projects, they often consume significant resources and time. Data quality programs are most efficient and effective when implemented in a structured, governed environment. Data governance is the business-driven policy making and oversight of corporate data; data management, which includes data stewardship, is the tactical execution of such policies (Dyché, 2010).

Clearly defining roles and outlining the authority, accountability, and responsibility for decisions regarding enterprise data assets provides the necessary framework for resolving conflicts and driving the business forward as the data-driven organization matures.

Consider defining such roles as data stewards, data custodians, subject matter experts, business stakeholders, the data governance council, and executive sponsors/advisors.

As organizations begin to bring big data into their environments, a common question is: "What do we need to add to our data governance program now that we have big data?" The answer is: nothing. Big data is still data—the rules of the game don't change. Big data projects will operate just fine under your existing data governance framework. Not all of the components of the framework will apply to all big data projects. That's okay, just as long as the projects don't run outside the established framework.

When considering data access and availability, is real time realistic?

The need for and definition of real time varies across industries and organization size. Although having access to the most current and accurate data is a reasonable, justifiable expectation (that can require heroic efforts in and of itself in some organizations), real-time access is generally not necessary. There are, of course, use cases in some industries that have little to no tolerance for data latency, such as sensors in life-saving medical devices, data feeds in stock trading, or air traffic control data. Because of the significant investments required to provide and support real-time data, many organizations have weighed business needs against the costs and determined that just-in-time is fast enough.

This will not remain the prevailing answer for long. With the evolution, maturity, and broader adoption of cloud and big data technologies, the expectation of real-time access and availability is increasing rapidly. Realistic or not, organizations must consider new tools and technology solutions to meet these expectations with a very limited budget and resources.

Although business needs and definitions of real time vary across industries, the technology solutions and capabilities to provide and support real time are the same regardless of business or industry. Technologies to explore include event stream processing, data virtualization, in-database embedded processing, cloud computing, and open source big data technologies.

With the evolution, maturity, and broader adoption of cloud and big data technologies, the expectation of real-time access and availability is increasing rapidly.

What is the greatest impact big data will have on the enterprise data environment?

Whether organizations have big data or not, the attention that big data is receiving in mainstream media and across all industries has a powerful direct impact on how they approach and manage data. Executives have tuned in to the big data story and are ready to support enterprise data initiatives and drive organizational change to become data driven. Based on what they have seen and heard, more data means more opportunity, more innovation, more revenue, and better customer experiences—the list of magic that more data brings to the business is ever-growing.

The newfound excitement and support for data is the good news and the bad news. You can't do big data for the sake of the coolness of big data. Although the emerging big data technologies are without a doubt exciting and attractive because of all the possibilities they generate, implementing solutions without a business purpose is doomed to failure. Harnessing the technical "eager beavers" will be a difficult but necessary challenge. Remember, the organizational strategy for managing data, regardless of size, is a business issue. Successful organizations design, manage, and govern their enterprise data programs based on business needs and initiatives.

How will data quality initiatives evolve as organizations add big data to their enterprise environments?

Many early adopters sought to redefine data quality initiatives based on the size or type of data (structured, unstructured, etc.) as they introduced big data to their environments. This approach did not prove to be successful because the business needs had not changed. In the end, big data was still data. The business rules and requirements were still necessary and applicable.

The evolution organizations will see for data quality initiatives as they integrate big data will not be based on the size of the data but rather on context of use. Business rules and quality requirements differ based on the intended use of data.

A data management trend that big data brings to the table is the concept of data lakes (or other large data containment bodies) to hold enormous amounts of unmanaged data. The store-everything approach is not the unique piece of the trend but rather the concept of "manage at consumption" that it brings. Organizations want to take advantage of the significantly lower data storage costs of big data technologies, but applying the requisite policies, standardizations, and transformations to support all business needs to such large data volumes becomes implausible.

To meet the needs of the business and capitalize on the significant data storage cost savings, organizations are starting to employ late-binding processes that apply the data management rules, processes, and policies at the time data is requested within the context of the request.

Should organizations manage and govern all data equally?

The type of data does not determine whether all data should be governed and managed equally—scope does—and the answer is no, organizations should not manage and govern data equally. Management and governance needs will vary as the scope changes. All defined processes, policies, and procedures should comply and adhere to the overarching enterprise data governance program. As the scope narrows from the enterprise level to the business unit, department, and even down to a specific project, the rules and requirements will become more specific. It is critical to apply governance with appropriate scope because the degree to which an organization can use data strategically is the degree to which data is effectively governed.

What is the major differentiator between leaders and laggards in regard to data quality and management?

Leaders consistently treat data as a corporate asset to drive business value. They are keenly aware of the costs and risks that low-quality, incomplete, and inaccurate data present. They understand the implications of not delivering timely, relevant data to the business. In these organizations, executives make available all of the dedicated resources, funding, and technology needed to support a successful enterprise data environment.

These organizations have developed their data management strategies by understanding the needs of the business. Although the business drives how they manage data, they do not get bogged down in whether the business or IT owns data. Instead, business and IT are strategically aligned to support data initiatives as a united front across the enterprise.

References

Dyché, Jill [2010]. "Data Governance Next Practices: The 5 + 2 Model," BeyeNETWORK, December 9. http://www.b-eye-network.com/view/14782

TOP 10 PRIORITIES FOR DATA QUALITY SOLUTIONS

By Philip Russom, TDWI Research

The 10 priorities listed here provide an inventory of techniques, team structures, tool types, methods, mindsets, and other characteristics that are desirable for a fully modern, next-generation data quality (DQ) solution. Few organizations will need or want to embrace all 10 priorities; you should pick and choose according to your organization's business and technology requirements. My intent is to help user organizations prioritize and plan their next-generation data quality program or solution.

Priority #1: Broader Scope for Data Quality

We say *data quality* as if it's a single, solid monolith. In reality, DQ is a family of eight or more related techniques. Data standardization is the most commonly used technique, followed by verification, validation, monitoring, profiling, matching, and so on. TDWI regularly encounters user organizations that apply just one technique, sometimes to just one data set or one data domain. Most DQ solutions need to expand into more DQ techniques, data sets, and data domains.

Priority #2: Real-Time Data Quality

According to a TDWI survey, real-time data quality (RTDQ) is the second-fastest-growing data management discipline, after master data management (MDM) and just before real-time data integration. Make RTDQ a high priority so data can be cleansed and standardized as it's created or updated.

Priority #3: Data Quality Services

DQ techniques need to be generalized so they are available as services that can be called from a wide range of tools, applications, databases, and business processes. Data quality services enable greater interoperability among tools and modern application architectures as well as reuse and consistency in DQ solutions.

Priority #4: Coordination with Other Data Management Disciplines

DQ functions are beneficial to related data management disciplines. For example, DQ functions should be applied to the reference data managed by an MDM solution, and data integration solutions invariably uncover DQ problems and opportunities.

Priority #5: Data Stewardship and Governance

Instead of re-inventing the wheel, user organizations can borrow some of the organizational structures and processes of DQ's stewardship and apply them to data governance. This minimizes the risks and decreases the time-to-use of data governance. Likewise, there are stewardship capabilities built into many DQ tools that can help document, automate, and scale up data governance processes.

Priority #6: Non-traditional Data Types

New types and sources of data are coming from many directions, and all need a DQ strategy. As data is deduced and extracted from Web data, multi-structured data, and social media, it should be subject to DQ functions and quality metrics, as with all data.

Priority #7: Internationalization

This is second-, third-, or later-generation priority for most DQ solutions. Prepare for it by selecting vendor tools that support internationalization functions for national postal standards, Unicode pages, and DQ tool GUI localization.

Priority #8: Value-Add Process

Techniques such as standardization and data append add value by repurposing and augmenting data, respectively. Deduplication adds value to data by reducing its redundancies. Data profiling reveals opportunities for more value-adding actions by DQ techniques. Focus on the value-add process to ensure the continuous improvement expected of a DQ program.

Priority #9: Deeper Profiling

Data profiling is too often shallow, just generating simple statistics for values found in a single database, table, or column. It should be broadened to enable more profound discoveries within data. Profile data repeatedly as a kind of monitoring that tests whether data's quality is truly improving.

Priority #10: Vendor Tools

Many first-generation DQ solutions are homegrown and hand-coded. For example, standardization is the most commonly used DQ technique, and (at the low end) standardization can be hand-coded in SQL or developed using a tool for extract, transform, and load (ETL). Hand-coded DQ solutions can prove the usefulness of software automation for DQ, but you should anticipate life cycle stages that demand functionality that very few organizations can build themselves, such as identity resolution, probabilistic matching, internationalization, real-time operation, DQ services, and hub-based architecture.

For a more detailed discussion, read the article "Ten Goals for Next-Generation Data Quality" in TDWI's *What Works: Case Studies and Solutions*, Volume 33. TDWI members can access the magazine at tdwi.org/whitepapers/2012/05/what-works-volume-33/

Philip Russom is director of TDWI Research for data management and oversees many of TDWI's research-oriented publications, services, and events. He is a well-known figure in data warehousing and business intelligence, having published over 500 research reports, magazine articles, opinion columns, speeches, Webinars, and more. Before joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and BI consultant and was a contributing editor with leading IT magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.

ENGAGING AND EMPOWERING BUSINESS USERS TO IMPROVE DATA QUALITY

Who owns the data has much to do with who is responsible for its quality. Here's how IT and business users can share responsibility.

Who owns the data, really: business or IT? It's a question that's provoked no end of discussion and dissension between the line of business and its IT "custodians." Thanks to a combination of technological, economic, and cultural factors, it's also a mostly moot question.

The simple fact of the matter is that both business and IT own the data; the reach, rights, and responsibilities of both groups can and should be neatly demarcated; and—going forward—wrangling about ownership will prove to be divisive, distractive, and ultimately destructive.

This isn't weak-tea pragmatism, insists Matthew Magne, global product marketing manager for data management (DM) with SAS. A concept of what might be called "shared ownership," based on the insight that IT's data management policies (to say nothing of its portfolio of DM tools and services) can and should be aligned with the needs of the business, is the new normal.

"It's actually very important that we align the creation of data and [the] management of data across its life cycle with business drivers," Magne acknowledges.

"Although IT owns the tactical execution of how [a company] manage[s] data—[e.g.,] what tools do we use to manage the data and what architectural strategy do we use to manage data?—it is critical that IT's priorities are aligned with the business drivers of the organization, too."

It's in this sense, Magne suggests, that the business can be said to "own" the data. Put differently, data must be managed in a way that's transparent or intelligible to the business. The business "owns" the data to the extent that it sets priorities, provides a reference for alignment, and—in the form of data stewards and other IT-to-business liaisons—works with IT to see that this is the case.

"It's no longer a question of IT implementing these business rules the way it sees fit, on its own terms, [albeit] in a way that's consistent with policy or regulatory requirements," Magne continues. "It's now [a question of] proactively tracking business rules in order to try to get ahead of challenges. Before marketing launches a massive direct-mail campaign and spends 20 percent more than it should because its address data is riddled with data quality problems, we're able to measure and detect those [issues] so we can alert the IT team that's responsible for fixing [the data]."

This is a specific example, but it gets at the kind of co-ownership experience Magne has in mind. In the old model, data quality was treated as something that somehow belonged to data—that is, as an inherent characteristic or property of that data, irrespective of how that data was used or what it was used for. In the new model,

quality is contextual: it's a function of how data gets used in the context of particular business processes or by different business domains. This last is actually aligned with the process by which data quality problems typically get redressed, at least in practice. In an ideal world, all data would be consistent and standardized; in the real world, this isn't always the case.

In most cases, in fact, it's practically impossible for an organization to simply "fix" its data quality issues overnight. More commonly, data quality and other governance issues get addressed as part of a phased approach. The larger point is that not all data quality metrics matter equally to all business stakeholders. From marketing's perspective, data must have certain quality characteristics, such as complete ZIP codes and e-mail addresses; from the perspective of other business units (e.g., customer service), other quality characteristics (complete telephone numbers or consistent product numbers) might be as (or even more) important.

An organization could undertake a massive, top-to-bottom effort to cleanse and standardize all of its data—with the result that critical business activities would probably be severely constrained during that process. It makes more sense to empower the business to identify and redress data quality issues, such that governance is aligned with business priorities.

For example, before marketing kicks off that new direct-mail campaign, salient quality issues—missing ZIP codes or incomplete e-mail addresses, for example—must be fixed. "If you look at it this way, the question becomes, 'How do we implement these business rules in such a way that we're proactively tracking levels of data quality in an organization and alerting the right people in the right places when there's an issue?'" Magne points out. Furthermore, we have to learn to avoid "Groundhog Day" data quality—that is, cleaning up the data every month without ever addressing the root causes of poor data quality.

Teaching the Business to Engage

The problem, now as ever, has to do with enabling business people to take an active, actionable role in managing—or, more precisely, in stewarding—their data. The traditional (IT-centric) approach to data management can seem highly technical and even unintelligible to business people. Magne and SAS believe that technologies such as data visualization and self-service access, exposed in new offerings that are designed explicitly for the line of business, can help to meaningfully involve business people in the stewardship of their data.

In a sense, IT should welcome this change. Most IT organizations are already chronically understaffed; they're overwhelmed by legacy baggage, not least of which include chronic governance issues and the persistence of departmental silos. On top of this, IT needs to build up (or train up) new skill bases to deal with big data and data-in-motion, the cloud, and other new modes. To the extent that the business wants to take more responsibility for its data, IT should welcome it. This is what Magne has in mind when he speaks of "business user engagement," which is a SAS-specific term.

"If you look at the technology component [of the problem], most of the data quality tooling is created for technical users. What we [at SAS] focused on is providing user interfaces that reduce or eliminate the need for these vast amounts of training and that don't require kind of a computer science degree to use," he says.

"If you have a wizard-based, browser-based, simple, intuitive way to manage data in order to create business rules, which is what we deliver as part of our SAS Data Quality solution, that's one way [to promote] business user engagement," Magne continues. "Think of how you [use a wizard] to manage your e-mail inbox. If you know how to create rules that assign a high priority to your boss's e-mail address, you can create [actionable] rules in our SAS Data Quality suite."

Data quality is just one piece of a unified information management framework, argues Magne. Think of it as a spectrum: at one end is data integration (DI) activity, in which data is consolidated from different sources. At the other end, there are the front-end tools. In between, there are the governance and policies that determine how data gets managed. These include general-purpose data quality and master data management offerings, along with policy-specific or point products that address requirements or problems in specific verticals.

A credible take on business user engagement must address this spectrum in its entirety, argues Magne: from the point of ingestion (the identification, selection, and loading of data; its preparation; and, if necessary, conformation) to the processes by which data is governed to its ultimate dissemination to information consumers. This, he says, is what he means by a "unified" framework that manages data over its life cycle.

Magne points to SAS's own product portfolio as a case in point: SAS Data Quality is a component of SAS Data Management, which addresses DI and governance, along with vertical- or regulatory-specific DI and governance issues. (SAS also markets a dedicated governance offering, SAS Data Governance, that maintains a

business glossary that relates business and technical metadata and facilitates lineage tracking and impact analysis, among other practices.) In addition, SAS markets front-end business intelligence (BI), analytic discovery (Visual Analytics), statistical analysis (Visual Statistics), and other, vertical- or requirement-specific tools.

"We have this holistic solution that spans data, analytics, and decision management. That's something that most companies can't do well. That's something that I think is really unique," he comments. "We always were known as an analytics company, but people don't think about the data that drives effective decisions from the use of analytics. The data to decision life cycle. You gather data, profile it, clean it up, explore it, create analytical models with it, deploy it as part of your business process, and then drive better business decisions. That's the cycle. It keeps looping. The framework to do this includes being able to track a semantic web, or network of metadata from analytics, reports, the business process that spawned it, the source systems where it resides, the reports that are viewed, and the analytic models that drive it."

Magne returns to the example of the marketing department and its data quality problems: "The unified framework lends itself to a phased approach. We might focus on fixing the tactical data quality issues we talked about and then ramp up the governance initiative as [part of] phase two. For phase three, maybe we could look at deploying an MDM solution on a more enterprisewide basis."

Even Big Data Must Be Governed

Even though it might seem forward or even absurd to talk about governance in the context of big data, it's nonetheless important. Big data platforms such as Hadoop are already being used to inexpensively store, manage, and analyze both non-traditional and structured data types. Hadoop's data management feature set is still comparatively primitive, but it's evolving rapidly thanks chiefly to the applications for which Hadoop is being used:

- As a landing zone for data of all kinds
- As a data preparation and staging area—think of it as ETL at industrial-scale—in which structured data is landed and transformed prior to being extracted to data warehouse systems or other relational consumers
- As a platform for advanced analytics, particularly for analytic workloads that involve non-traditional types of data (machine-generated data, text files, images, audio recordings, and files of every type) in combination with advanced statistical, data mining, and other kinds of algorithms

Organizations are already extensively using Hadoop for the first and second use cases; the goal is to work up to leveraging Hadoop for the third use case, too. To that end, SAS and other vendors are working to make Hadoop more user friendly.

For SAS, Magne says, this means extending the metaphor of business user engagement to the Hadoop platform, too. "Probably the biggest thing I'm excited about is SAS Data Loader for Hadoop because it's designed to get around the challenges that people have with big data. They don't have the right skill sets in house to do this, they haven't aligned IT with their business, there's no governance [in the big data model], there's no security either," he observes.

"With SAS Data Loader for Hadoop, you have this browser-based method that you can expose to business users and which they can use to actually manipulate data on a Hadoop cluster without knowing or caring what Hadoop is." The SAS offering, though new, delivers an experience that's consistent with what business people should expect from business user engagement and helps unshackle IT, says Magne. "The idea is engaging business users so that using this wizard-based tool, they can move [a] data [set], filter it, and prepare it for visualization, all without the assistance of IT, freeing them up to be more productive. A marketing person can prepare data to do segment analysis [on a data set] for an upcoming campaign.

"Under the covers, [SAS Data Loader for Hadoop] translates this into MapReduce jobs that parallelize the work effort. If a user knows SAS code, that's even better. Doing anything on Hadoop requires highly specialized expertise. You have to code for MapReduce, and those [MapReduce jobs] are really difficult to write. What we've done is create a way to leverage your existing skill sets. If you know SAS code, you can write it in such a way that it can ... be deployed [and run in the context of] MapReduce. Even if you don't know anything about coding in Java or coding in SAS, you benefit from that wizard-based business user engagement component."

sas.com

SAS understands that data drives everything. We want to help you make sure it's right.

Is your data easy to access, clean, integrate, and store? Do you know which types of data are used by everyone in the organization? And do you have a system in place for analyzing data as it flows in?

Spend less time maintaining your information and more time running your business with **SAS® Data Management**. It's an industry-leading solution built on a unified platform and designed with IT and business collaboration in mind. It's also the fastest, easiest, and most comprehensive way to get data under control, with in-memory and in-database performance improvements helping to deliver trusted information. When it comes to master data management, data integration, data quality, data governance, and data federation, SAS can help you transform big data into big opportunities.

With SAS, you can create a culture where data quality is truly valued and improve data quality where your data resides. You'll have unprecedented levels of data quality to support business processes, and we'll help you manage your entire data quality life cycle.

Learn more and discover our free white papers, webinars, and videos: sas.com/data.

- sas.com/dataquality
- dataroundtable.com

tdwi.org

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education and research in the business intelligence, data warehousing, and analytics industry. TDWI is dedicated to educating business and information technology professionals about the best practices, strategies, techniques, and tools required to successfully design, build, maintain, and enhance business intelligence and data warehousing solutions. TDWI also fosters the advancement of business intelligence and data warehousing research and contributes to knowledge transfer and the professional development of its members. TDWI offers a worldwide membership program, five major educational conferences, topical educational seminars, role-based training, on-site courses, certification, solution provider partnerships, an awards program for best practices, live Webinars, resourceful publications, an in-depth research program, and a comprehensive website, tdwi.org.