



Data Warehouses and Quality: Not Just for IT Anymore

Thomas C. Redman, Ph.D.

A Navesink Consulting Group White Paper

- January, 2008 -

Sponsored by Teradata Corporation

Correspondence to:

Thomas C. Redman, Ph.D.

President

Navesink Consulting Group

732-933-4669

tomredman@dataqualitysolutions.com

www.dataqualitysolutions.com

Data Warehouses and Quality: Not Just for IT Anymore

Thomas C. Redman, Ph. D.

ABSTRACT

This paper has four interlocking goals: First, is to emphasize (hopefully re-emphasize!) the critical importance of high-quality data if decision makers are to trust their data warehouses and fully utilize them. The second is to contrast the competing approaches for managing data quality. Respectively, these aim to find and fix (clean-up) errors and prevent errors at their sources. The second approach, preventing errors, is clearly superior, yielding orders of magnitude better data that can be trusted for even the most critical decisions.

Preventing errors at their sources requires a single minded focus on the business processes that create data. Therefore, it is primarily a business, not technology, endeavor. Unfortunately, too many enterprises elect the find and fix errors approach, perhaps reasoning that “since the data is in the warehouse, it must be the responsibility of the warehouse manager.” So, the third goal of this paper is to clarify for business managers that it is they, not IT, who must lead the data quality program. Finally, the fourth goal is to help business managers understand what they must do to improve data quality.

INTRODUCTION AND SUMMARY

Over the last dozen years, enterprises have implemented data warehouses at a furious pace. The allure to “get all the data in one place” to improve decision making, unleash data mining, and “get everyone on the same page” has proven intoxicating. And with good reason! In many organizations, data warehouses have helped decision makers see customers and markets in new ways, create new products and services, and improve operations and, in so doing, distance themselves from competitors.

But alone, even the largest, best designed, and most technically sophisticated data warehouse does not enable even the boldest, most aggressive decision maker to fashion these results. For decisions depend critically on the data on which they are based. Bad data leads to bad results and average data to average results. Superior decisions, the kind that “change everything,” transform companies, and create new industries, are only possible when the data are of superior quality.

As this paper describes, any enterprise can dramatically improve its data. Doing so doesn't require special skills, tools, or even a lot of money. Rather, it requires proper orientation, leadership, discipline, and a burning desire to improve. Enterprises with superior-quality data realize that data quality is largely determined the moment the data is created, so they focus on preventing errors. They develop a keen sense of which data is most important and aggressively manage both their business processes that create data and outside suppliers that provide it. They measure quality levels in business terms, and they identify and eliminate root causes of error. Once they've removed a cause, they build controls into their processes and applications so it doesn't return. In so doing, they eliminate entire categories of future errors, improving quality by one to two orders of magnitude, and creating a data warehouse that decision makers can trust.

Note that the business processes that create all this data are, by and large, conducted within the business community. It follows that business departments must lead data quality efforts and do most of the heavy-lifting in preventing errors.

Unfortunately, far too many enterprises settle for average data quality. They assign data quality to warehouse managers who address the issue by trying to find and fix errors using a variety of

data profiling, editing, standardization, and de-duplication tools. It is demanding, time-consuming, and expensive work. Done well, it eliminates many of the most egregious errors, bringing “bad” data up to “average” standards. But too many errors escape technological capture, and the data gets no better.

Said differently, superior-quality data results not from technologically-driven efforts to find and fix errors, but from well-defined and – managed business processes that don’t make errors in the first place. And these must be the purview of the business community. Therefore, it is time for the enterprises that want to make and execute the really bold, industry-changing decisions to task their business departments, not IT, with leading their data quality efforts.

DECISION MAKERS, DATA WAREHOUSES, AND DATA QUALITY

Most decision makers are smart. They instinctively know that their decisions are no better than the data on which they are based. Decision makers function best when they fully trust their input data and have plenty of evidence that this trust is well-placed. So much trust that they can concentrate on the decision at hand and not worry about the data. Only then can they make bold, confident decisions. And only then can they align the enterprise to execute them with power and authority.

When populated with high-quality, trusted data, warehouses, including the associated decision support, business intelligence, and data mining tools, can be powerful enablers. Many enterprises use them to complete 360-degree views of their customers, gain deep insights into market performance, and develop more effective and efficient ways to conduct their operations. Results can be stunning, as the numerous examples in the recent book by Tom Davenport and Jeanne Harris, *Competing on Analytics*,¹ attest. Personally, I think that organizations investing in data warehouses should expect no less.

Unfortunately, too many enterprises are plagued by data quality issues that shatter any hope of reaping these benefits. The issues include:

- data values that are simply wrong, missing, or out-of-date,
- weak data definitions that simply don’t capture reality, and
- inconsistencies in data that should agree, and so forth.

While these issues may pervade transactional systems, they only affect one transaction, one customer, or one product at a time and often get lost in the press of day-in, day-out work. Data warehouses, on the other hand, seem to have a way of bringing data quality issues into the open, often magnifying the problems. Decision makers see examples every day, when they peruse the data, review simple reports, and try to sort out inconsistencies between the data warehouse and other systems. And, quite naturally, they are loathe to base decisions on data they mistrust.

Importantly, few decision makers are capable of making the granular distinction between a data warehouse and the data it provides. So, if the data isn’t up to snuff, they fault the entire warehouse! The net result is that some data warehouse projects fail.² But the more common situation is that the data warehouse suffers “chronic under-use.” Business people may use it for relatively minor tactical decisions, but it never gains full acceptance. So it is used neither to suggest transformative opportunities nor for the really important decisions.

For data warehouses, the stakes in data quality are indeed high!

¹ Davenport, Thomas H., and Harris, Jeanne G., *Competing on Analytics: The New Science of Winning*, Harvard Business School Press, Boston, MA., 2007.

² For example, the following statement “According to a senior research analyst at a Gartner Group Symposium, by 2005, more than 50% of data warehouse and CRM projects will fail, with one of the points of business failure including denial about data quality issues” appeared in *Customer Data Management: Providing a Sound Foundation for CRM*, DataFlux Corporation, 2003.

THINKING ABOUT AND MANAGING DATA QUALITY

Approach Matters

The most important decision an enterprise makes about data quality is the approach it selects for addressing the issue. We'll discuss the candidate approaches using the so-called "data quality lake" of Figure 1. In the analogy, the *lake* is like a database, and the *water* in the lake is like the data it contains. Further, the *stream* feeding new water into the lake is like a business process that creates new data, and the *factories* along the stream that cause the water entering the lake to be less than pure are like the business process issues that cause the new data to be erroneous, late, improperly formatted, or otherwise unfit for use. Finally, and not pictured, are those who drink the lake water, boat on the lake, and fish from it. They are like the business people who use data to serve customers, process transactions, and make decisions.



Figure 1. The data quality lake.

Isolation Approach

As noted above, the data quality lake aims to clarify the enterprise's choices regarding its approach to data quality. There are three choices. The first focuses not on the water itself, but on preventing people from drinking from it. It is usually put into action after several people get sick. An especially observant emergency room doctor or epidemiologist traces the sicknesses back to a common source. A fence and *No Trespassing* signs are put up around the dirty lake. The approach may also involve training doctors to recognize the symptoms, developing better treatments, and settling the lawsuits that are sure to arise.

The equivalent approach in the data world entails taking necessary steps to make good for people impacted by bad data. This may mean giving customers their money back, restating financial statements, or backing out of bad decisions. For large technology projects, it may mean scrapping the project altogether. Finally, managers who've been misled, advise others not to trust that system, the equivalent of the *No Trespassing* sign.

This approach is clearly not smart, and no one would even think of adopting it as the primary approach for maintaining an important water supply. The approach is just as wrong-headed for data quality. Unfortunately, far too many enterprises have given data quality far too little attention, in effect, unconsciously selecting this approach. We will not discuss this approach further.

Cleanup Approach

The second approach is to clean up the water before it does any harm. This is exactly what water purification plants do. They spray the water in the air, run it through settling tanks, and use chemicals to kill bacteria and adjust the pH. They test the water continually and don't allow it to enter the drinking supply until it passes on all scores.

Many enterprises strive to clean up their data before it (the erred data) does any harm. There are two steps: *finding* the errors and *fixing* them. Finding errors involves combing through the data using business rules. The simplest business rules check for missing values—required data that simply is not there. More complex business rules check whether a data value is in its prescribed domain. For example, in the United States, 'STATE' must be one of fifty recognized states. Business rules can be quite complex, involving more than one attribute or field. A simple two-field example involving 'AREA CODE' and 'POSTAL CODE' helps illustrate the point. For a given household, at least one of the following is almost certainly incorrect:

AREA CODE = 212 (New York City) and POSTAL CODE = 90210 (Beverly Hills)

The second step, fixing the errors, can be more problematic. Some errors can be corrected automatically. For example, it is often possible to eliminate duplicate records. Other corrections can be made using sophisticated probabilistic methods. But all too often, employees must go to original sources to make the corrections. As an example, the only way to resolve the inconsistency between the area code and postal code in the example above may be to ask a member of the household.

Finding and fixing errors is popular, primarily because it has two advantages: First, it can be quite effective at eliminating obvious errors and second, much of the work, especially error detection, can be assigned to IT. But relying on this approach has several disadvantages. First, it is time-consuming and expensive. Writing good business rules is hard work, and erroneous data can't be put to use until it is corrected. Second, it adds no true value: no informed customer would ever pay extra for an enterprise to make an error, and then correct it. Third, far too many errors avoid detection by business rules (e.g., transposed digits in a TELEPHONE NUMBER) or prove too difficult to correct. These issues further erode trust in the data.

But the biggest objection to finding and fixing errors is that it never ends! The flawed business processes keep churning out new errors every day. As the quantities of new data grow exponentially, even more errors are created. In some enterprises whole departments must be formed to keep up with the work of finding and fixing errors.

Prevention Approach

Going back to our lake analogy, the third approach is to identify the sources of pollution (the upstream factories) and eliminate or mitigate them, thus allowing clean water to flow into the lake and, over time, improving the overall quality of lake water. This is the approach favored by environmentalists, and is the proven winner.

For data, the equivalent approach involves identifying and eliminating the root causes of error in the business process that creates the data. A good example of a root cause is that one department doesn't understand a particular field that it creates, but does not use to complete its work. People collecting and entering the data have learned, through trial and error, that inputting 'NA' is accepted by the system. Eliminating this root cause involves training current employees and incorporating this training into the new employee training. Enterprises often find that there are a relatively few (say two to four) similar issues, and eliminating them improves quality by a factor of ten. Thereafter, clean data flows into the database and, as decision makers see the improvements, they come to trust the data. Further improvements require a bit more work, but most enterprises, applying diligent effort, can improve by another order of magnitude. This approach is called "preventing errors at their sources" or simply prevention. It is the clear winner for data quality.

The differences in these approaches could not be more stark (per our promise, we will not discuss the isolation approach further). One focuses on the errors that have already been made, while the other focuses on errors that haven't been made yet. One focuses on clean up; the other on prevention. One grows as the rate of data creation grows; the other gets smaller as there are fewer errors to prevent. One is typically conducted by IT; the other must be led by the business community. Each leads to predictable results, as summarized in Table 1. Note especially that the prevention approach leads to far higher quality at far less total cost.

Table 1: Summary Comparisons of Approaches to Managing and Improving Data Quality

	Led By	Field-Level Error Rate ¹	Record-Level Error Rate ²	Overall Cost ³
Find and Fix Errors	IT	1% to 5%	20% to 50% or more	Rule of ten: It costs ten times as much to complete a simple operation when the data is erroneous as when it is perfect.
Prevent Errors at Sources	Business Community	.05% to .5%	.5% to 1%	reduced by two-thirds

Notes: 1. Field-level error rate divides the total number of erred fields by the total number of fields.
 2. Record-level error rate counts a record as erred if one or more fields are erred. We often find that simple operations (accepting a customer order, completing a security transaction, etc.) depend upon and/or create ten to twenty-five important fields. The stats reported here reflect our experience with such processes.
 3. It is nearly impossible to pin down the costs associated with lost opportunity and/or poor decisions due to poor quality data. It is even more difficult to estimate the lost opportunity costs due to data that are not trusted. So here we've provided reasonably reliable estimates of the cost of poor data quality in simple operations.

Preventing errors is not perfect. It does not promise that errors will never leak through (though it does promise that decreasingly fewer will). Nor does it address the erroneous data already created.

Therefore, the smart data quality strategy involves:

1. Focusing first on preventing new errors, especially errors in the most important data.
2. Initiating a cleanup project narrowly focused on the most important data only after the processes for creating high quality data have been established.
3. Understanding and scoping cleanup activities. A very few issues, such as failing to pay employees correctly, may demand immediate cleanup. In these circumstances, it is imperative that a parallel effort to eliminate the root causes of error must be undertaken. Many enterprises make two mistakes on this strategic point:
 - a. they conclude that all data issues demand immediate cleanup.
 - b. they decide to hold off on eliminating the root causes of error and never get around to doing so.
4. Staying in front of new issues and opportunities by following steps 1-3 above as customer requirements evolve.

A Data Warehouse is Like a Sea or an Ocean

Data warehouse managers may well point out that their warehouses and the processes that feed them cannot be likened to a single lake and stream. Even a moderate-sized warehouse is like a vast sea, maybe even an ocean, fed by literally hundreds of streams, tributaries, and rivers of enormous size. How well, they may wonder, does the lake/stream analogy, and especially the approach based on prevention, hold up under these circumstances?

Importantly, the bigger the database and the higher the rate of data creation, the more apt the analogy. Bigger databases, especially data warehouses, have more diverse customers, more demands placed on them, tougher quality requirements, and greater consequences for being wrong. Cleanup costs too much, takes too long, and, except in rare cases, doesn't scale. Prevention, on the other hand, is ideally suited to such environments. One may not be able to take on all streams, rivers, and tributaries at once. But one can move from process to process, confident that sources of error, once eliminated and controlled, will not return.

THE TEN HABITS OF THOSE WITH THE BEST DATA

The sidebar lists the ten habits of those enterprises with the best data. Certainly not all enterprises that enjoy high-quality data follow all ten in equal measure or in the same ways. Instead, they emphasize those that work best for their data, and tailor the workings of each into their cultures. And, perhaps, most importantly, they employ the habits in ways that reinforce one another as part of an overall data quality program that diffuses, and later embeds itself across the enterprise.

Customer Focus

The first habit is that of *customer focus*. It is essential because quality is in the eyes of the customer³, and without understanding customer needs, it is impossible to satisfy them. One aspect of following this habit involves sitting down with customers, learning what decisions they make, how they make those decisions, and working backwards to determine what data they need and the quality levels they will accept. It is usually a good idea to document customer needs in the form of "Customer Data Requirements," which proves useful input to later habits.

This seems straightforward enough. However, the habit of customer focus is more involved.

There are simply too many customers, too many requests, and too much change to develop and keep one grandiose "Customer Data Requirements" document current. The more difficult part of the habit is determining which customers are most important. It's demanding work, because if some customers are more important, then others are less so. And who wants to tell a colleague in product management that he is less important than a marketing director? Those with the best data realize they have no real choice. Failing to focus on the most important customers almost always means that efforts to improve data quality will become diluted.

Because it has these two components, this habit is more correctly stated as *focusing on the most important needs of the most important customers*.

Process Management

With requirements in hand, those with the best data work backwards to the business processes that create the needed data and actively *manage those processes end-to-end* (the second habit). As noted above, individuals and departments that create and input data, say, in customer service or production, often have no idea why they are asked to do so. Good process managers address that gap by sharing the aforementioned requirements document. And immediate improvements

The Ten Habits of Those with the Best Data

The habits of:

1. Customer focus
2. Process management
3. Supplier management
4. Measurement
5. Continuous improvement
6. Control
7. Targets for improvement
8. Clear management accountabilities
9. Managing soft issues
10. Broad, senior group leadership

³ A "data customer" is anyone or thing who uses the data in any way. Customers can be people or departments/enterprises. They may be inside the enterprise or out. They may use the data in the most mundane way, or for strategic decisions. And, anything in between.

often result, as individuals and groups figure out how to improve their work. This is one example of the power of process management.

Supplier Management

Sometimes the business process that creates the needed data lies outside the enterprise, in the supplier base. Enterprises with the best data are not deterred by this. *Manage data suppliers* is their third habit. It seems odd that companies will issue detailed specifications and employ service level agreements for the physical goods they purchase and completely ignore the data. But, the methods of supplier management for physical goods are just as effective for data. Indeed, supplier management can be an effective paradigm for improving quality as data flows from one department to the next inside an enterprise.

Measurement

The fourth habit is *measurement at the source*. There is no mystery here. The old management saw that *you can't manage what you don't measure* applies to data as well as anything else. Good measurements are especially useful in identifying areas for improvement. In addition, many enterprises publish their data quality statistics, at least for key customers. The goal is twofold. If the data is good and getting better, publishing the statistics builds decision makers' confidence. If the data is not so good or not getting better, publishing the statistics increases the urgency for improvement.

Continuous Improvement

The fifth habit is *continuous improvement*, and it is the means by which much of the heavy lifting is accomplished. An improvement project, in and of itself, involves nothing more than investigating a pattern of errors, selecting a category of error to work on, identifying the root cause, and changing the business process to eliminate that cause. Frequently, managers jump to blame inattentive or lazy low-level employees as the most important root cause. This is rarely the case. More often, the root causes are lack of requirements, poorly defined processes, lack of training, unfathomable applications, and illegible inputs.

Enterprises with the best data have a knack for starting and completing improvement projects. They do so in ways that engage the knowledge and intellect of everyone involved. They work on a manageable number of projects and, once one is done, they celebrate success, and start the next one.

Control

The sixth habit is *control*, and it is a bit more involved. Control is defined as "the managerial act of comparing actual performance against requirements and acting on the difference."⁴ A good business rule can form the *comparing performance against requirements* portion of a control. But the definition isn't met unless the enterprise *acts on the difference* by ensuring that the data is corrected.

Enterprises with the best data employ controls on many levels:

- First, to help people enter data correctly.
- Second, to prevent errors from leaking downstream.
- Third, to ensure that systems and other changes work properly.
- Fourth, to distinguish between special and common causes of process variation. The controls needed to do so are called statistical controls, and they are indeed powerful.

⁴ This definition is after the great quality guru, Dr. Joseph Juran. See, for example, *Juran's Quality Handbook, Fifth Edition*, McGraw-Hill, New York, 1999.

Targets for Improvement

Intimately related to the habits of measurement, improvement, and control is the seventh habit, which involves *setting and achieving aggressive targets for improvement*. For example, the head of a business unit remarked that her goal was to cut in half the error rate in all key processes every year. This is impressive, but typical of those with the best data! The habit reflects much more — namely that data has made it into the planning cycle.

Clear Management Accountability

In a perfect world, middle managers could develop all of these habits by themselves. After all, almost all managers buy in, conceptually at least, to the notion of preventing errors at their sources. But the world is far from perfect, and enterprises with the best data recognize the importance of leadership, as described in the last, but far from least, three habits. The first of these three (and eighth overall) is *clear management accountabilities*. Specifically, those who create data are accountable for its quality. The maxim applies to all data—operational data, financial data, reporting data, metadata — all data.

It is easy to dismiss this habit. Once one buys into preventing errors at their sources, it seems obvious. Unfortunately, inertia may take an enterprise in a different direction. Implicitly perhaps, people conclude, “If it’s in the computer, it must be the responsibility of the CIO (or data warehouse manager).” This is exactly counter to the proper responsibilities. So, in many organizations, it is important that managerial accountabilities for data be formalized via a data policy.

Managing Soft Issues and Senior Leadership

The ninth habit calls for *managing the soft issues* and the tenth for *senior leadership*. Most data people are surprisingly naïve about organizational politics. For example, many data people think it obvious that developing a 360-degree view of customers is a good thing. But a senior salesperson, with deep knowledge of a big customer, may well feel threatened. This knowledge is a source of respect, power, and commissions! To this salesperson, a 360-degree view is threatening and may cause him or her to be a reluctant contributor.

There are no easy to resolve political issues. Enterprises with the best data embrace political realities. Leaders of their data quality programs gain and use political capital effectively. They don’t go off “tilting at windmills,” attempting to do more than they are capable of doing. At the very least, they are astute, and don’t get tripped up by the more obvious issues within their enterprises.

The success of most data quality programs extends just as far as the influence of the most senior manager(s) perceived to be leading the effort. Those with the best data recognize this and secure broad, senior support. The higher and broader the better!

DATA QUALITY AND THE BUSINESS COMMUNITY

The previous sections should remove any doubt that data quality must be the responsibility of the business community, not IT. The only proven way to enjoy the benefits of high-quality data is to create it correctly the first time. The heavy lifting to achieve high quality, as evidenced by the ten habits, takes place in the business departments of an enterprise.

Two further observations reinforce this point. First, it is the business community that has the most to gain or lose from data quality. As we’ve noted, poor quality data lead to mistrust of a data warehouse and may harm IT’s reputation. But costs to the business community are more direct, involving both initial investment and ongoing operational dollars. Conversely, if the data is of high quality and the data warehouse trusted, the business community may reap millions from a single better conceived, better executed strategic decision. For IT, the goal is to be much like the referee at a soccer match, invisible to the crowd and the outcome.

Second, more and more managers are coming to realize that data is a business asset, worthy of the same level of attention as capital and people.

So what must the business community do? First, it must make a conscious decision that it is responsible for data. It must explicitly define a data quality program focused on the philosophy of preventing future errors. It must put someone in charge who is senior and well-positioned enough to lead the transition to the new philosophy. It must assess its current data quality program, the current levels of quality for important data, and the business impacts in terms of cost and lost opportunity. And it must craft a data quality program that addresses the shortcomings and advances its interests.

In doing so, it must create a sense of urgency around data quality. It is amazing how tolerant many enterprises are of poor quality, and the high costs that result from it. This tolerance must stop!

Second, the business community must recognize data quality as an essential component of new technology initiatives. More specifically, it must embrace the four points of the smart data quality strategy and the ten data quality habits as early as possible in a new technology initiative. In the worst case, failing to do so may scuttle an otherwise well-planned and executed technology initiative. And waiting too long may limit its options, in effect forcing the enterprise into an enormous and costly cleanup effort during the latter stages of the initiative.

Embracing data quality early thus reduces risk and cost. Even more importantly, it helps prepare the enterprise to fully utilize the data warehouse. It does so because the work to improve data quality forms a powerful lens into the entire enterprise. This lens helps bring into sharp focus the questions that an enterprise must ask if it hopes to transform itself. Questions about everything from product mix, to customer segmentation, to the conduct of operations. Answering these questions is precisely what an enterprise seeks from its data warehouse.

Third, the business community must simultaneously narrow and clarify its expectations of IT with respect to data. Given the current state of data quality, cleanups are inevitable. But *ongoing cleanups* are not. The business community must task IT with the following:

- Carefully evaluating cleanup requests and rejecting them when adequate preventative measures are not in place.
- Using applications and technology effectively to lock in gains when processes are improved.
- Developing and implementing technological infrastructure.

Importantly, while we explicitly reject the view that “data is in the computer and so the purview of the CIO,” IT must step up as well. Too many IT departments have allowed this view to fester, even proliferate. They have been seduced by the power of cleanup tools and vendors’ promises. They have either stuck their heads in the sand regarding the importance of the highest possible quality data, or accepted cleanup assignments that they can only do moderately well. And in so doing, they have not served the needs of their business counterparts. This must change.

Ideally, of course, the business community and IT can work together on data quality. While the business community should lead, both bring skills and perspectives that can accelerate the effort. And, the sooner the enterprise enjoys the benefits of a high-quality, trusted data warehouse, the better!

FINAL REMARKS

A study of some of the world’s best companies and their data quality practices reveals data quality to be a subject of both great mystery and incredible simplicity. This paper reflects much of that simplicity — a simple philosophy of preventing errors at their sources and a number of habits to do so. Neither the philosophy nor the habits described herein are particularly challenging from

a technical perspective. Their implementation requires consistent management effort, leadership, political will, and persistence (persistence above all!).

In some respects, this paper is mistitled. The “Not Just for IT Anymore” is misleading, for data quality has always been predominately a business issue. More and more enterprises are hearing the message, making dramatic improvements, and enjoying the benefits that result. Not yet a groundswell perhaps, but movement is unmistakable. No enterprise that depends on its data can ignore it.

About the Author

Thomas C. Redman, “The Data Doc,” is president of Navesink Consulting Group, Little Silver, NJ. Dr. Redman advises leading corporations on data quality. His fourth book, *Data Driven: Profiting from Your Most Important Business Asset*, is slated for publication by Harvard Business School Press in 2008.

About Teradata Corporation

Teradata Corporation, the global leader in data warehousing, provides solutions that give companies the people, technology, and innovation to make smarter, faster decisions that help them achieve competitive advantage to master their markets. Learn more at teradata.com.

EB-5499 > 0108