WHITE PAPER

Build a Data Warehouse That People Actually Use — and Trust

JIM HARRIS

2349 (02/2013)

Table of Contents

Introduction

Business users and executives require a holistic view of the enterprise to better understand business performance, improve customer experience, and uncover hidden opportunities for revenue generation and competitive differentiation. These business requirements reflect the need to make trusted decisions in less time than ever before, using trusted data and more varied sources and types of data, with more transparency in decision making and its business results.

But most of the data they need is fragmented across many applications, which are hard to integrate. The data can be incomplete and inconsistent, which becomes clear as they try to piece data together across applications to get a single view. Therefore, a data warehouse is built, bringing together disparate operational data sources in an attempt to integrate and transform all of this data into a single system of record. The data warehouse's goal is to provide timely delivery of trusted data to support the current and evolving business intelligence and decision-making needs of the enterprise.

Paraphrasing the famous line from the movie *Field of Dreams*, many of these data warehousing projects are initiated with a "if you build it, they will come" mentality. However, many projects become a data warehouse of broken dreams when, after they go live, few business users come to play — a frustrating, if not infuriating, result after considerable time, effort, and money were expended.

To understand why the data warehouses of today don't get used, we need to revisit the best practices of yesterday. We need to get back to the basics of data warehousing, which are essential for building a next-generation data warehouse, although their key concepts are often overlooked or oversimplified.

This white paper will focus on a high-level overview of three categories of key concepts:

- **Data Quality Management** – When people don't use a data warehouse because its data isn't trusted, we have to examine if overemphasizing data delivery is causing an inattention to data quality best practices, as well as verify that trusted data is not misunderstood as perfect data.
- **Decision Management** – When people don't use a data warehouse because of bad business results, we need to consider not only the data management practices that provide the data to support business decisions but also the decision management practices to evaluate what data was used, and why, before we blame the data warehouse for being untrustworthy.
- **Metadata Management** – When people don't use a data warehouse because communication breakdowns preventing collaboration are exacerbated by a complex data warehouse environment, we must look at the metadata management practices necessary to create a common language that will make the data warehouse much more understandable, more usable, and therefore more trustworthy.

Data Quality Management

When people don't use a data warehouse because they don't trust its data, we have to examine if overemphasizing data delivery is causing an inattention to data quality best practices, as well as verify that trusted data is not misunderstood as perfect data.

When Delivery Is Job One, Quality Is Job None

The reality in data warehousing is that the primary focus is on delivery. The data warehouse team is tasked with extracting, transforming, integrating, and loading data into the warehouse within increasingly tight time frames. Twenty years ago, monthly data warehouse loads were common. Ten years ago, weekly loads became the norm. Five years ago, daily loads were called for. Nowadays, near-real-time analytics demands the data warehouse be loaded more frequently than once a day.

The quality of the data in the warehouse determines whether it's considered a trusted source, but it faces a paradox similar to "which came first, the chicken or the egg?" For the data warehouse, the question is "which comes first, delivery or quality?" However, because users can't complain about the quality of data that hasn't been delivered yet, delivery always comes first in data warehousing.

And with transaction volumes reportedly growing 50 to 60 percent each year, data warehouse teams are struggling to keep up with the flow of data, which, in the era of big data, also includes fast-moving, large volumes of variously structured data (e.g., social interactions and sensor readings). Although new prefixes for bytes (giga, tera, peta, exa, zetta, yotta) measure an increase in space, new prefixes for seconds (milli, micro, nano, pico, femto, atto) measure a decrease in time. More space is being created to deliver more data within the same, or smaller, time frames. Space isn't the final frontier, time is.

To deliver more data in less time, something has to give, and often that something is data quality. When delivery is job one, quality is job none. This is an ineffective strategy: Either data quality issues are discovered upon delivery, reducing business users' trust in the data warehouse, or, even worse, no one notices and poor data quality becomes a ticking time bomb that will eventually explode and wreak havoc on daily business activities.

Reactive Versus Proactive Data Quality

Overemphasizing data delivery sets the data warehouse team up for being reactive, not proactive, regarding data quality. Let's briefly examine the characteristics of these two different approaches.

Reactive data quality focuses entirely on finding and fixing the problems with existing data after it has been extracted from its operational sources and loaded into the data warehouse.

Characteristics of a reactive approach to data quality management include:

- **You Talk About, but Don't Look at, the Data** – Although business requirements and technical specifications are reviewed, no data analysis supplements the planning process, leading to development and delivery delays caused by unexpected, and undocumented, data issues.
- **When in Doubt, You NULL it out** – The dirty little secret of ETL is that its standard practice for "resolving" a data quality issue is often to substitute a bad data value with either a missing or default value. For example, a date stored in a text field in the source that can't be converted into a valid date value is loaded with either a NULL value or the processing date.
- **Code, Load, and Explode** – When data quality management suffers from inattention, or the intentional non-attention of "just get the data loaded," the organization is often blindsided by the negative business impacts of poor data quality in the warehouse. Examples include a customer service nightmare, a financial reporting scandal, or a regulatory compliance failure.
- **You Rinse, Maybe Repeat, but Never Root out** – Data quality issues are treated with data cleansing, which may be repeated periodically, but the root cause of poor data quality is left unresolved and often not even identified. This approach simply hits the snooze button on the ticking time bomb, allowing the same data quality issues to eventually explode again.

Proactive data quality advocates quality control and monitoring. It's impossible to prevent every data defect, but the more control enforced wherever data originates, and the more monitoring performed wherever data flows, the better overall data quality will be in the warehouse.

A proactive approach to data quality management includes the following six steps:

- **Look Before You Leap to the Next Record** – Besides supplementing planning with data analysis, you should also profile during data entry, or wherever data enters the data warehouse environment. Doing so enables you to identify data quality issues as early as possible and provides an opportunity to address them before they flow further downstream.
- **Data Quality Monitoring** – Every time data is moved from point to point throughout the data warehouse environment, track what changes were made to the data along the way. Even when data quality issues can't be resolved, they should always be reported; all resolutions should be logged for verification and communicated to show the progress of improving data quality in the warehouse. Continuous monitoring ensures that subsequent updates don't undo fixes.
- **Address Data Quality as Close to the Source as Possible** – In many data warehouse environments, it's not feasible to push defect prevention back to the source because it would disrupt operational systems. Therefore, the staging area of the data warehouse often becomes, in effect, the first line of defense against poor data quality. Whenever possible, this should also be the last line of defense, so that from this point forward data quality is consistent.
- **Build Reusable Rules** – The data warehouse often processes the same data domains from different sources. Reusing data quality rules helps you avoid re-inventing the wheel, especially when sources are processed separately. Not only will you eliminate redundant data cleansing downstream but you also might be able to pass those reusable rules upstream to prevent (or at least minimize) similar data quality issues.
- **If You Can't Beat 'em, Join 'em** – Sometimes data quality can be improved by switching to an alternative data source. You need to assess data quality first, but don't assume that fixing the data quality issues of an existing source is your only option.
- **Be Reactive in a Proactive Way** – Reactive data cleansing will occasionally still be necessary for critical data quality issues causing immediate business problems. This is why one of the most proactive things you can do is design a rapid remediation process to address these issues when they arise. Planning for unexpected, and inevitable, data quality issues makes you react in a proactive way and reduces the business impacts of poor-quality data.

Trusted Data Is Not Perfect Data

Before we move on to the next topic, there's one more important point to make about data quality. Delivering data with data quality issues is acceptable — as long as it's made clear that data is arriving with a known need for improvement.

Trusted data is not perfect data. Trusted data is transparent data, honest about its imperfections, and realistic about the practical trade-offs between delivery and quality. You achieve trusted data through the transparency that pervasive data quality monitoring provides. You can't fix what you can't see, but even more important, concealing or ignoring known data quality issues is only going to decrease business users' trust of the data warehouse.

Decision Management

When people don't use a data warehouse because of bad business results, we need to consider not only the data management practices that provide the data to support business decisions but also the decision management practices to evaluate what data was used, and why, before we start blaming the data warehouse.

Decision Bias

In his book *The House Advantage: Playing the Odds to Win Big in Business*, Jeffrey Ma explained that "judging the merit of a decision can never be done simply by looking at the outcome. A poor result does not necessarily mean a poor decision. Likewise a good result does not necessarily mean a good decision. This is something that few people intuitively understand. In fact most people believe that if the result turned out positively, then the decision must have been the right one."

In psychology, this is known as outcome bias, which could serve as the framework for jumping to the wrong conclusion when we have an unsatisfactory experience with the data warehouse. When business results are positive, no one questions the decision. However, when the business results are negative, often instead of questioning how the decision was made, what's questioned is the decision-supporting data and its process. We have to identify where the problems are before we implement solutions. Although the emphasis is usually on bad data management, sometimes bad decision management is just as much to blame for bad business results.

In his book *The Half-life of Facts: Why Everything We Know Has an Expiration Date*, Samuel Arbesman discussed three additional biases that can affect how we use data to make decisions:

- **Factual Inertia** – Tendency to adhere to out-of-date data well after it has lost its applicability.
- **Semmelweis Reflex** – Tendency to ignore data because it doesn't fit within our worldview. Named after 19th-century doctor Ignaz Semmelweis, who proved that so-called childbed fever, which resulted in the deaths of many women after childbirth, was actually caused by doctors not washing their hands before delivering babies. However, his data was rejected by the medical community because the germ theory of disease hadn't become widely accepted yet.
- **Confirmation Bias** – Tendency to only seek out data that adheres to our worldview.

Too often, the human biases of decision makers are not taken in consideration when evaluating the effectiveness of the organization's decision-making process.

Trusted Decisions Are Transparent Decisions

If decision makers claim that they're not using the data warehouse because it doesn't provide 100 percent visibility into all data and complete trust in the origin, accuracy, and completeness of the data needed to make decisions, then does that mean they're not making any decisions? Or that what they're using instead of the data warehouse meets all of their data needs? Of course not. Trusted decisions are transparent decisions, which reveal exactly what data was used and why.

Good-Enough Data for Fast-Enough Decisions

As was discussed earlier, overemphasizing data delivery often means data quality suffers. But time constraints also provide a frame for business decisions that cannot be overstated. For example, a decision that must be made within 30 seconds has very different data requirements than a decision that should be made within 30 minutes or a decision that could be made within 30 days.

The growing demand for more near-real-time business decisions sometimes makes decision speed more important than data quality. Although high-quality data is obviously preferable to poor-quality data, less than perfect data quality cannot be used as an excuse to delay making a critical decision. Nor can the fear of making a mistake be used as an excuse to delay making a critical decision, which sometimes also makes decision speed more important than decision quality.

In a constantly changing business world, we often need good-enough data for fast-enough decisions. But what exactly constitutes good-enough and fast-enough isn't something that can be articulated with general rules for data quality and decision speed. Instead, what's necessary is that decision makers are engaged in defining specific data requirements for each business decision.

Key considerations of decision management include:

- **Document the Decision-Making Process** – Most decisions are either undocumented or only document the preparation for but not the execution of them. Document both for every decision.
- **Decision-Specific Data Requirements** – With a clearer understanding of decision needs, efforts can be prioritized to deliver better data quality to the decisions that need it most.
- **Eliminate Unnecessary Data and Reports** – Use the documented decision-making process to eliminate unnecessary data and reports. Reduce the delivery logjam by eliminating what's being delivered without being used, which also streamlines making decisions.
- **Data Usage Monitoring** – Ongoing monitoring of data usage can identify changes in the decision-making process, as well as identify data for archival status after its current usage has lapsed. Such monitoring also reduces data storage and maintenance costs associated with decision support.

Metadata Management

When people don't use a data warehouse because communication breakdowns preventing collaboration are exacerbated by a complex data warehouse environment, we must look at the metadata management practices necessary to create a common language that will make the data warehouse much more understandable, more usable, and therefore more trustworthy.

The Broken Telephone of the Data Warehouse

In many ways, a data warehouse resembles the children's game broken telephone, where a message is passed through a group by being whispered into the ear of one player after another, until the last player announces the message to the entire group. Because errors typically accumulate in the retellings, the final version of the message often differs significantly from its source. Some players appear to deliberately alter what's being said, guaranteeing a garbled message by the end of the game.

As data journeys from operational sources through the staging area, the data warehouse, the data marts, and finally into dashboards and reports, a lot could be lost in translation. As it is processed, data is often deliberately altered to make it accommodate the structure of its next target. Every time data moves from point to point, semantic inconsistencies may be introduced.

Much more than "data about data," metadata can be thought of as a translator providing a definition and context for data. Therefore, metadata plays an integral role in determining data usage. There's also a strong relationship among metadata, data quality, and decision management. Metadata provides a context for evaluating the quality of data, and metadata supplies a frame for interpreting the contents of the dashboards and reports involved in the decision-making process.

Commonly used terms such as revenue and customer often complicate what on the surface seem like straightforward discussions about, for example, how much revenue was generated during a particular fiscal quarter or how many customers the organization has. These discussions often turn into heated debates over how such terms as revenue and customer should be defined and how their data should be integrated and aggregated to support the views displayed in dashboards and reports.

Trusted Metadata Is a Lingua Franca

Although we might be tempted to enforce a single definition for commonly used terms, we must acknowledge that the players in the data warehouse's game of broken telephone are not trying to undermine communication. Instead, they're using terms based on what's often a valid alternative business perspective, reflecting the context of their business needs and job responsibilities.

Both the business and IT bring their own vernacular language to the data warehouse, but a vehicular language is needed to traverse its complex communication pathways. Instead of forcing one group's native language to be the standard, a common language must be crowd sourced, which must support multiple perspectives, multidimensional definitions, and multidirectional translations.

Trusted metadata is a lingua franca, a common language capable of replacing the broken telephone of the data warehouse with the clear communication network needed for business-IT collaboration.

Turn the Data Warehouse into a Glass House

The data warehouse's goal is timely delivery of trusted data to support decision-enabling insights. However, it's difficult to get insights out of an environment that's hard to see inside of.

This is why, as much as is possible given the necessities of data privacy, a data warehouse should be turned into a glass house, allowing us to see data quality and decision management challenges as they truly are. Without this visibility, dangerous assumptions can be made about business problems and related data challenges being well-understood by the collaborative team trying to solve them. The data warehouse needs to provide a clear view of the terminology (both business and technical) surrounding its data and its processes (again, both business and technical) in order to make the data warehouse much more understandable, more usable, and therefore more trustworthy.

Key considerations of metadata management include these five factors:

- **Business Glossary** – Enhance business-IT collaboration with a glossary furnishing the underlying detail about how business terms were derived and where they're used. Such a glossary increases IT productivity by ensuring clear communication with business users.
- **End-to-End Data Lineage** – Documenting the data warehouse's trail of digital breadcrumbs allows tracing data from a report all the way back to the source, giving an overview of any data transformations or data quality rules applied along the way. This history is essential for troubleshooting existing issues and performing impact analysis on proposed changes.
- **You Can't Reuse What You Don't Understand** – Metadata enables reuse because business rules, data services, and prebuilt reports can't be reused if they're not understood.
- **Roll Call with Role Tags** – Some of the most important metadata simply tags who's it, by identifying the business process owners, data stewards, and other subject matter experts who can clarify any confusion about complex concepts or processes.
- **Auditing for Regulatory Compliance** – Comprehensive business and technical metadata provides the audit trail necessary to ensure that regulatory compliance requirements are met.

Conclusion

This white paper opened by explaining that a "if you build it, they will come" mentality doesn't work in data warehousing. What does work is laying the foundation of trust that the business users need to use the data warehouse that is being built to support them. If you want to build a data warehouse that people actually use, then they have to view it as a trusted source of data and metadata and as a trusted advisor for making business decisions. If they trust it, they will use it.

Although beyond the scope of this white paper, data governance provides the framework for the proactive and comprehensive approaches to data quality management, decision management, and metadata management outlined above, which are necessary to instill trust through transparency.

Often the root cause of why a data warehouse does not get used can be traced to the lack of a shared understanding of the roles and responsibilities involved (e.g., business process ownership and data stewardship). Data governance policies can help establish accountability for those roles and furnish the framework for establishing a pervasive program for ensuring that data is of sufficient quality so that it can be trusted to meet the organization's current and evolving business needs.

Furthermore, just like building any well-made product, there is a standardized set of practices and procedures that should be followed. When repetition and experience are applied to this process, best practices can be developed. By eliminating unnecessary steps and focusing on the critical elements of the process, you can accelerate the design, development, and deployment of solutions that transform data chaos into breakthrough business results. Therefore, it's important to remember that you are not just building a better data warehouse, but you also are building a better building process.

The value proposition of the data warehouse goes beyond the final data delivered via data marts, dashboards, and reports. A lot of the technical architecture, metadata infrastructure, and reusable components needed to build a next-generation data warehouse will also furnish your organization with the next generation of data integration and data quality capabilities.

By building a better building process, if you build it, they will come — they will see that you built not a field of dreams but a breeding ground for best practices. They will see that you built not just a next-generation data warehouse but also a factory that is manufacturing reusable data integration and data quality services that can be deployed all across your organization.

In closing, here's a summary of the key points made throughout this white paper:

- **Trusted Data Is Not Perfect Data** – Trusted data is transparent, honest about its imperfections, and realistic about the practical trade-offs between delivery and quality. Perfect data is impossible, but the more control enforced wherever data originates, and the more monitoring performed wherever data flows, the better overall data quality will be in the warehouse.
- **Trusted Decisions Are Transparent Decisions** – Trusted decisions reveal exactly what data was used and why. In a constantly changing business world, we often need good-enough data for fast-enough decisions, which is why decision makers must define specific data requirements for each business decision and document the preparation for, and execution of, every decision.
- **Trusted Metadata Is a Lingua Franca** – Metadata creates the crowd-sourced common language needed to traverse the complex communication pathways of the data warehouse, enabling the business problems and related data challenges to be well-understood by the business and IT teams that must collaborate to solve them.
- **A Trusted Data Warehouse Is a Glass House** – The data warehouse's goal is to provide timely delivery of trusted data to support decision-enabling insights. However, it's difficult to get insights out of an environment that's hard to see inside of. As much as is possible given the necessities of data privacy, a data warehouse should be a glass house supplying a clear view of the business and technical processes surrounding its data and decisions, in order to make the data warehouse much more understandable, more usable, and therefore more trustworthy.

About Jim Harris

Jim Harris is a recognized industry thought leader with more than 20 years of enterprise data management experience, specializing in data quality, data integration, data warehousing, business intelligence, master data management, and data governance.

As Blogger-in-Chief at Obsessive Compulsive Data Quality, Jim Harris offers an independent, vendor-neutral perspective and hosts the popular audio podcast OCDQ Radio, syndicated on iTunes and Stitcher SmartRadio. He is an independent consultant and freelance writer, as well as a regular contributor to Information-Management.com and DataRoundtable.com.

More information about Jim Harris can be found at: www.ocdqblog.com

About Informatica

Informatica provides data integration software and services that enable organizations to gain a competitive advantage in today's global information economy by empowering them with timely, relevant, and trustworthy data for their top business imperatives.

By using the Informatica® Platform for data integration in data warehousing and analytics, companies have been able to deliver current, actionable, and trusted information to their business users in half the time and cost compared to the alternatives. They've ensured the completeness, accuracy and timeliness of the data to make users confident in basing decisions on the data. They've also removed the maintenance nightmare and proliferation of spreadsheets, data marts, and one-off reports that all too often raises costs and leads to inconsistent or inaccurate decision making.

More information about Informatica is available at: www.informatica.com

Additional Resources

Adventures in Data Profiling

For a vendor-neutral blog series (and presentation download — no registration required) about the functionality provided by data profiling tools, visit: ocdqblog.com/adventures-in-data-profiling

Identifying Duplicate Customers

For a vendor-neutral, free (no registration required) presentation and study guide about data matching methodology, visit: ocdqblog.com/identify-duplicate-customers

The Collaborative Culture of Data Governance

An article by Jim Harris on information management: information-management.com/issues/21_1/the-collaborative-culture-of-data-governance-10019477-1.html

The Role of Data Quality Monitoring in Data Governance

An article by Jim Harris on dashboard insight: dashboardinsight.com/articles/business-verticals/the-role-of-data-quality-monitoring-in-data-governance.aspx

Informatica Perspectives

For posts about data integration, data warehousing, data governance, data quality, and more, from a variety of bloggers, including John Schmidt, David Loshin, and Ravi Shankar, visit the Informatica Perspectives blog at: blogs.informatica.com/perspectives

Govern Your Data

For an open peer-to-peer community of data governance practitioners, evangelists, thought leaders, bloggers, analysts, and vendors sharing vendor and product-neutral best practices, methodologies, frameworks, education, and other tools, check out: GovernYourData.com (sponsored by Informatica)