

Master Data Management: What It Is and Why You Should Care

John E. Bentley, Wachovia Bank, Charlotte NC

ABSTRACT

By early 2006, Master Data Management (MDM) emerged as one of the next big things in enterprise information integration and early-adopter organizations had already successfully implemented it. Companies who like 'proven technology' are not moving as fast but by mid-2006 MDM had certainly moved into the IT mainstream. The concept of system of record has been around for years in the data warehousing field, and data cleaning is an integral part of extract-transform-load processing. So how is MDM different? Is it really something new or is it an old concept marketed in a new package? This paper will explain what MDM is, suggest some approaches and challenges to implementing it, and suggest how SAS[®] can help make it happen.

INTRODUCTION

Master Data Management (MDM), also known as Reference Data Management, is a discipline in Information Technology that focuses on the management of reference or master data that is shared by several disparate IT systems and groups. MDM is required to warrant consistent computing between diverse system architectures and business functions.
Wikipedia, the free encyclopedia

Hot topics in data management change every couple years. Many IT managers have been burned by expensive solutions that never quite deliver on their promise, so a cynic can be forgiven when she says that when data management consultants can't make enough money selling existing services, they invent a new one--data warehousing, for example, morphed into the Corporate Information Factory; a data dictionary evolved into metadata management; and exploratory data analysis became data profiling. Many of these concepts turn out to be quite valuable to the business users even though they take lots of up-front investment and years to recoup that investment. Others never quite deliver on their promise, admittedly often due to a botched implementation rather than conceptual failure.

As Phillip Russom at The Data Warehousing Institute recently put it, data management today is all about integration. It's about integrating multiple sources of customer, supplier, and vendor data, integrating data stored in multiple application silos, integrating multi-source data for BI purposes, integrating our data with our partners' data, integrating with government standards and formats, and accomplishing this integration through new interfaces and channels like Web services. The cynic might say that Master Data Management (MDM) is intended to extend and expand all this integration work. That cynic might also say that MDM looks suspiciously like a new twist on the old "system of record" concept.

Regardless of its provenance, MDM has proven itself to be valuable and so is gaining acceptance in as a mainstream technological discipline. It's widely recognized that a solid foundation of Master Data is needed to support the processes of integrating similar or even supposedly identical data elements from numerous disparate source systems.

Based on an admittedly cursory examination of the professional literature, it appears that in early 2005 Master Data Management began to be implemented by early-adopter organizations, and by 2006 was being recognized by even late majority organizations. (See Everett Rogers, *Diffusion of Innovations* for a theory relevant to technology adoption.) As an indicator of how it's becoming a mainstream technology, a few months ago the DM Review web site started a "Master Data Management" Portal. As further evidence, even some technologically-conservative banks are starting MDM initiatives.

This paper will provide some basic information about Master Data Management, just in case you gentle reader are assigned to a project with a goal of implementing it. "What is master data management?" and "How will it benefit us?" are two questions this paper will answer.

WHY MASTER DATA? WHAT'S THE PROBLEM?

At its most basic level any business object, entity, or concept that is subject to multiple definitions needs Master Data. These multiple definitions arise in part from certain pathologies that evolve as organizations grow in size and materialize in the areas of communication and systems complexity. Generally speaking, the larger the organization, the more autonomy the internal divisions have and the less cross-divisional communication and coordination takes place. Larger organizations also generally have a larger number of more complex information systems and more of them that don't play well with each other.

Most IT systems are basically transaction processing application, and they're usually very good at it. But each of these systems cares only about the data it needs to process its own transactions and maintains (usually) only the minimal amount of metadata needed to get the job done. Even in a data warehouse, metadata describes the data provided by the source system and is often the same metadata used by the source systems. It is important to recognize the stovepipe nature of typical metadata management applications--most enterprises don't have a system that takes a holistic look at metadata.

The 'system of record' concept is well understood—one transactional system is highly trusted (for various reasons) and is designated as the source of authoritative data for some specific purpose(s). The problem with the system of record concept is that it is system specific and consequently is often inadequate for filling information needs. For example, if a bank has a regulatory requirement to report counts of mortgage customers by city it might start with a list of mortgage customers from the mortgage system and then match that data to address data sourced from the designated customer data system of record—the Customer Information System. But what if the report needs to be summarized by city where the property is located? Now the Customer Information System isn't appropriate and the mortgage system, which is not designated a system of record, must be used. Is this an extreme example? Perhaps, but it shows the impact of a system of record approach to having authoritative data.

In a Business Intelligence (BI) environment, even with a robust analysis and reporting system, the facts that describe data sourced from multiple systems are inconsistent, can be inaccurate, and are buried in data transaction structures, copybooks and data dictionaries, databases, content management systems, data marts, spreadsheets, or in email. Even with a solid data warehouse and data mart system, instant access to accurate information is probably still a gleam in the CIO's eye.

Consider a critical concept—"Customer". In different industries, a "customer" is defined differently and that makes sense. Within an industry, a "customer" can be defined differently and that's OK too. But within an organization, different business units can and often do define a "customer" differently and that's probably not a good thing. Within a Bank, a "customer" might be defined as a business that applied for a loan (loan a thousand dollars and you have a customer; loan a million dollars and you have a partner); a person who co-signed for a car loan; the spouse of a person who opened a brokerage account; or another bank that purchased a bundle of mortgages. The Accounting Department might say that there are fourteen million "customers" but the Customer Information System holds the names of twenty-six million "customers". Which is right?

Product is another area that has the potential for multiple definitions, especially products that are services as opposed to tangible objects. Again using a Bank to illustrate the problem, consider a Safe Deposit Box. In a retail branch, a box in the vault might be product SDR3x6 if rented to a person but SDC3x6 if rented to a business. If the customer rents the box through their personal banker who doesn't work in the branch and uses a different transaction system, the same box might be product SDBS.

The point here is that data values that should uniquely describe business entities often differ between business units and information systems. Relationships between the different data values (via information map) are often either non-existent, exist in hard-to-get-to personal databases, or exist in multiple places and differ between information systems.

Compounding the problem of defining an entity, where the same set of base codes or identifiers are used throughout the organization different hierarchies can exist that result in different roll-ups. Consider the geographic identifiers "city", "area", and "region". It's entirely possible different divisions within an organization use different hierarchies to group cities into areas and areas into regions. The result is that Region 1 in the Retail Division is different from Region 1 in the Commercial Division. How does one get an enterprise-level regional report?

Finally, identifiers that one could suppose are unique to each entity are sometimes either used multiple times or used incorrectly. The example here is that a Bank organized into four divisions named demand deposits, consumer loans, mortgages, and safe deposit boxes each with their own transactional systems can allow each of the divisions to use the same 12-digit number for the customer account number. Account number is unique within each division but what happens when those divisions' data is combined in a data warehouse?

All of these issues and others make enterprise-level data integration, management, analysis, modeling, and reporting much more difficult than it should be. Even if the data is cleaned as part of a data warehouse's ETL process, it will still need a massaging and recoding prior to use. Even after the analyst's best efforts the data might still be a mix of apples and not-quite apples, and the result is information that is a bit off but nobody can say by how much. Management, looking at the \$8 million annual data warehouse and its \$2 million annual budget, wonders why they can't get the information they need.

Driven in part by the need to comply with increasingly stringent federally-mandated oversight regulations such as Sarbanes-Oxley, Master Data Management has emerged as a model for how to create and maintain a source of standardized data values and business rules used to describe or define core business elements.

A DEFINITION OF MASTER DATA MANAGEMENT

Master Data is also known as Master Reference Data and can be defined as the authoritative information required for creating and maintaining an enterprise-wide "system of record" for core business entities to allow standardized analysis and reporting.

In a nutshell, Master Data provides an enterprise-wide system of record for core business entities, objects, and concepts. This system of record consists of definitive facts that are (a) obtained from a variety of sources, (b) accepted throughout the organization, and (c) describe and provide context to the core business entities, objects, and concepts. Items needing description and context include products, locations, customers, suppliers, and so on down to concepts like business rules used in data cleaning.

It's important to recognize here that ETL data cleansing is critical but is most often focused on standardizing data formats (yes=1, 2=no) and improves the quality of data by identifying obviously redundant or anomalous records. Implementing Master Data Management requires that "enhanced data cleansing" be implemented, a step that includes rules that recode data formats from disparate sources to a single format that enables enterprise-level or enterprise-wide analysis and reporting.

Once in place, Master Data provides an authoritative reference point to effectively integrate, analyze, and report data regardless of its source. Master Data Management provides a system of record for each business object so that users across the organization have a single complete, accurate, and standardized view of it. Master Data provides, for example, a lookup table for mapping every corporate division's geographic region codes to a single master enterprise region code. Metadata is specific to each separate data management system but Master Data spans the enterprise. In this sense, Master Data might be thought of as 'super metadata'

Master Data is critical when multiple IT systems across a company identify a business object differently. Within a Bank, the retail banking division can continue to define geography differently than the commercial organization for divisional purposes, and both of their geographies may be different from those the securities division uses. It's entirely permissible that in their transactional systems each division might use its own metadata hierarchies for city, area, region, and state. In an enterprise-level database or for an enterprise-wide BI system, however, Master Data is needed to authoritatively define a standard geography so that all users have the same single view of the data. This single view enables accurate data synchronization, integrity, quality, and analysis that are not subject to misinterpretation or confusion.

HOW DO YOU DO MASTER DATA MANAGEMENT?

Haven't IT professionals been trying for years to provide consistent, comprehensive, easily accessible core information across the enterprise? Yes, they have and a lot of progress has been made—think about the evolution of data warehouses, data marts, and BI systems over the past ten years. But even with these successes the problem of heterogeneous metadata still exists. Very few organizations have succeeded in implementing a single system of record for data that spans the enterprise.

Master Data Management is not an end-state. It is an ongoing set of processes that identifies, acquires, and stores master reference data and then uses that data to insure the provision of accurate, consistent, standardized information. In a data warehousing environment, this requires integrating Master Data into the ETL process to insure standardized formatting of data values, even if it means recoding (which is anathema to those who subscribe to the "if it's in the legacy system then load it" approach to ETL). Alternatively, a BI solution can leverage Master Data in real-time to insure reporting accuracy and consistency.

Master Data Management should be a part of any data integration solution. Having a data warehouse that is the mother of all databases, having an ETL tool that processes hundreds of gigabytes of data in time windows unimaginable only a couple years ago, or having the most expensive BI tool set that you can find will not deliver the promised results without the single view of the data that Master Data provides.

There are three major stages to a Master Data Management initiative. The first is to select a manageable (that means small) set of business entities that you want to address. Working with the business users and perhaps the BI tools team, you can quickly identify a set of data elements used in an important analysis, model, or report—focus on those. Very possibly the senior manager who most needs a report will agree to be the business sponsor if you can convince her it's worth the time and effort to make his information more accurate.

With the data elements identified, next understand the extent of the problem. Use data profiling to analyze each data element across source systems for accuracy, completeness, structure, business rules compliance and uniformity of coding. Data accuracy as used here means "conforms to expected contents". Let's say that the 'account status code' field sourced from all sixteen source systems should contain only letters, but in some instances contains a number. In these cases we obviously have a data accuracy problem. If we used SAS to combine all the source systems instances of 'account status code' and assigned an identifier to each source system then we could easily produce a contingency table of status code by source system. We've very quickly see how accurate our data is and which systems are problematic. We could easily assign both system-specific scores and an enterprise-wide score based on the percentage of conforming status codes.

Data completeness is an interesting concept often ignored in data warehousing when the approach is "if it's in the legacy system then load it." Data completeness means simply that a field is populated more often than not. If a field is populated from three source systems but we find that only 40 percent of the records contain any data for the field, we can reasonably ask why the field is being loaded. Do the business users know that 60 percent of the data is missing? Maybe, but maybe not.

Data structure refers to the safe deposit box problem mentioned earlier where three different product codes were used for the same product. Data structure also includes naming conventions—if twelve data mart tables contain a field with a business name of 'product identifier' because that's the name in the source systems and three tables have a field named 'product code', because that's what their source systems use, then it makes a whole lot of sense to standardize the field name as 'product identifier'. A MDM initiative is an opportunity to clean up some of the little mistakes that might have been made years ago but never fixed.

When analyzing a data element for business rules compliance, we have to find out if data is entered into the source and downstream systems is in accordance with the business rules that apply to the field. Many data elements use business rules that reference data elements transformed by their own business rules, and it could be that the inputs to your business rule are assigned by another business rule that has an element that uses a third business rule that was changed a few months ago but nobody heard about it. (In many organizations, impact analysis is a concept that has yet to be fully implemented.) So we might find that our business rule is "working" but the results now are not what we saw when we wrote our business rule and verified and validated it two years ago. It could be that anomalies that would have been caught two years ago are slipping through now.

Finally, uniformity analysis focuses on determining whether or not the same data item has the same form from system to system. Status code could be O (open) and C (closed) in System A; O, P (pending), and C in System B; and A (approved), R (rejected), P (paid), and D (defaulted) in a third.

A Master Data Management initiative is different from other data quality initiatives because all this work is done on the combined instances of the particular data element across all the organization's IT systems, not just as it exists in one system. With cross-system data profiling you can report the status of the data element as it is across the entire enterprise. Using this as a starting point you will ultimately be able to get "a single version of the truth" from your information systems.

Once you have profiled the data elements selected for Master Data status, a Master Data repository should be set up. A dimensional data model may be most appropriate for storing Master Data, but regardless of the data model or architecture the MDM Repository is a database that contains the 'super metadata' that you've collected.. The MDM Repository will supplement the existing system and table-specific metadata in a data warehouse or BI environment and will be incorporated as the system of record for ETL business rules when synchronizing or creating cross-enterprise data elements, mapping tables for data recoding during cleansing, and defining hierarchies for report summarizations.

MASTER DATA MANAGEMENT AND SAS

SAS has a number of products ideally suited for implementing a Master Data Management initiative. First, of course, is the SAS Foundation software that is so incredibly versatile at accessing various and sundry data sources and running data profiling routines. The SAS9 Base, Macro, and SAS/Access products combine to make powerful, robust, and efficient custom-written data profiling and monitoring applications possible.

With either the SAS Enterprise Data Integration Server or the SAS Warehouse Administrator as the point-of-control, you can easily build and manage the ETL processes needed to build the Master Data Repository and insure that it remains

accurate and up-to-date. These same products are also ideal for leveraging the contents of the MDM Repository in warehouse or mart ETL processes.

Finally, SAS and DataFlux combine to offer the SAS Data Quality Solution for an almost off-the-shelf platform for enterprise-level data profiling, monitoring, cleaning, augmentation, and integration. With the Data Quality Solution you can create and manage Master Data as well as incorporate Master Data in ETL processes.

SUMMARY

Master Data Management is much more than an IT consultant's latest silver bullet. Properly implemented it can accomplish great things and deliver on its promise of improving data accuracy, consistency, and reliability. Actually, in a lot of organizations even a haphazard implementation of Master Data Management would provide improvement. The end result is that you have more consistent, more reliable data than you have now. And we all know that you need to use reliable data to get information that can be turned into knowledge.

As with all IT initiatives, beware of those who suggest a big bang approach. (In a room full of IT consultants, how do you spot the expert? She's the one who says the project will take longer and cost more than the project plan calls for.) Even IBM agrees that Master Data Management isn't something that can be accomplished with a big bang project. A realistic MDM initiative will quite likely be a progressive multiyear process of unraveling a data mess that might have been created over decades. Using SAS, though, at least it doesn't have to be a labor-intensive process. An incremental approach using the right tools is what's needed.

REFERENCES AND RESOURCES

Druker Daniel, and Robert Rich. Master Data Management. DB2 Magazine, 3d Quarter 2005
<http://www.db2mag.com/showArticle.jhtml?articleID=167100925>

Griffin, Jane. Information Strategy: The Master Data Management Challenge. DM Review, May 2006.
http://www.dmreview.com/article_sub.cfm?articleID=1026072

Smith, Mark. Master Data Management for Information Management. Intelligent Enterprise, December 15, 2004.
<http://www.intelligententerprise.com/showArticle.jhtml?articleID=55800289>

Waddington, David. Master Data management—What is it? Business Intelligence.Com, June 16, 2005.
http://www.businessintelligence.com/print_research.asp?id=97

Yang, S. Jae. Primer: Master Data Management. Baseline Project Management Center, June 10, 2005.
<http://www.baselinemag.com/article2/0,1540,1826593,00.asp>

AUTHOR BIOGRAPHY AND CONTACT INFORMATION

Since 1987 John Bentley has used SAS in the healthcare, insurance, and banking industries. He is currently with Wachovia Bank's Business Integration Services Group where he supports the Bank's data warehouse and data marts. John holds four SAS certifications, has been a Section Chair at both SUGI and SESUG, and is past-Chair of the Charlotte Area Wachovia In-House SAS Users Group. He has a Masters degree in Political Science with a Concentration in Southeast Asian Politics and is intermittently enrolled in a Masters of Information Systems program. Driven from Chicago by the weather, he hopes to never again live north of the Mason-Dixon Line.

John E. Bentley
Business Integration Services
Wachovia Bank
201 S. College Street, NC-1025
Charlotte NC 28210
john.bentley@wachovia.com

DISCLAIMER

The views and opinions expressed here are those of the author and not those of Wachovia Bank. Wachovia Bank does not necessarily subscribe to, support, or make use of any of the concepts presented here.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.