

Paper 146-2011

Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution

Pat Herbert, DataFlux, a SAS company, Cary, NC USA

ABSTRACT

Master data management has become one of the fastest-growing application areas in recent times because of the need for organizations to leverage the valuable assets hidden within disparate enterprise data stores. DataFlux, a SAS® company, has created the DataFlux qMDM Solution in order to allow you to connect the dots between these previously disconnected systems without needing to make changes that can negatively impact operational systems. This paper provides an overview of the qMDM Solution and provides real-world examples of how one can achieve the goal of mastering data to connect the information across multiple disparate systems... and the techniques for accomplishing that task are changing. In this presentation, you see how the DataFlux qMDM Solution can put you on the path to developing a complete view of your data.

INTRODUCTION

Of all the factors that drive an organization's success, customers and products are at the top of the list. To increase revenue, it's critical to know who your customers are, what they purchase and how to sell more to them. And you need to know what you're spending on materials across all business units to control costs. Master data management (MDM) is the creation of a single, accurate and unified view of corporate data, integrating information from various data sources into one master record. This master data is then used to feed information back to the applications, creating a consistent view of information across the enterprise.

The insights offered by MDM allow companies to:

- Improve revenue generation
- Reduce risk
- Enhance customer relationships
- Uncover opportunities to maximize spend

THE THREE LEGS OF DATA MANAGEMENT

In a 2009 Gartner survey, the research firm found that the average organization loses \$8.2 million annually from poor-quality data. Further, of the 140 companies surveyed, 22% estimated their annual losses resulting from bad data at \$20 million – and 4% put that figure at a staggering \$100+ million.¹

The reason for much of this lost money is a fundamental misunderstanding of the value of data – or the inability to properly assign a monetary value on data. Enterprise information is the foundation for every business decision and corporate strategy, yet it is not historically treated as a corporate asset on par with a building, vehicle or piece of production machinery. However, as the data on customers, products, employees, finances and assets continues to grow – both in quantity and complexity – an organization's ability to manage this tidal wave of information will ultimately determine a company's ability to react to market trends and create a more competitive product.

There are three primary types of data management technologies and strategies that focus on creating more sustainable, high-value data:

- Data quality
- Data integration
- Master data management

DATA QUALITY

Data quality technology allows organizations to cleanse, correct and enhance any type of enterprise data – and create an accurate view of corporate data. Once called “data cleansing” or “data standardization” tools, modern data quality tools encompass a host of capabilities, from initial data discovery through ongoing monitoring of business rules on a real-time basis.

Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution, continued

While data quality technology has become broader in scope, it has also expanded its presence within the organization. Once the domain of the IT department who were tasked with cleaning up the data within enterprise data warehouses or business applications, current data quality tools are also used by business analysts and data stewards – business-side professionals who have some experience with IT infrastructures.

Data quality tools typically perform these tasks:

- Profile data to discover errors, inconsistencies, redundancies and incomplete information
- Correct, standardize and verify information across the enterprise from a single platform
- Match, merge or link data from a variety of disparate sources (see Figure 1)
- Enrich data using information from internal and external data sources
- Check and control data integrity over time with real-time data monitoring, dashboards and scorecards

The screenshot shows the SAS Scheme Builder window with two main panes: 'Report' and 'Scheme'.

Report Pane: Total count: 1293. It displays a table with columns 'Gro...', 'Value', and 'Count'. The data shows various company names and their counts, with 'First Merit Bank' having the highest count of 48.

Gro...	Value	Count
10	First Merit Bank	48
10	>first Merit Corp	2
10	1st Merit Bank Corp	1
10	First Marret Bank	1
10	1st Merit Bank	1
11	First Insurance Co	13
12	First Interstate Services Co	4
12	First Interstate Bank Corp	1
12	First Interstate Bank	16
13	First Interstate Bancorp	1
14	First Interstate Bank-Project Mgmt	1
15	First Interstate Bank-Operations	1
16	First Approach	2
17	Frst Oil Corp	2
18	First Trust Corp	4
19	First Security Info	1
20	First Century Co	1
21	First Source International	2
22	First City Savings	2
23	First Total Systems Inc	1
24	Three Five Systems	1
25	Fifth Third Bank	13
25	5th Third Bank	10
25	Fifth 3rd Bank	6
25	The 5th 3rd Bank	2
25	Fifth Third Banc	2
25	5th Thrid Bank	2
25	5th 3rd Bank	1
26	Aae Co	1
27	ABC Unified School District	3
28	Adia	2
29	Admar Corp	1
29	Admar Group Inc	4
30	Advo Systems	1
31	Advanced Computer Systems	1
32	Advanta Mortgage	4

Scheme Pane: Entries: 185. It displays a table with columns 'Data' and 'Standard'. The data shows various company names and their standardized equivalents, such as 'Fid' being standardized to 'Fht'.

Data	Standard
Fid	Fht
Fta International	Fht
5th 3rd Bank	Fifth Third Bank
5th Third Bank	Fifth Third Bank
5th Thrid Bank	Fifth Third Bank
Fifth 3rd Bank	Fifth Third Bank
Fifth Third Banc	Fifth Third Bank
Fifth Third Bank	Fifth Third Bank
The 5th 3rd Bank	Fifth Third Bank
Filenet	Filenet Corp
Filenet Corp	Filenet Corp
Filinet Corp	Filenet Corp
Financial Indemnity	Financial Indemnity
Financial Indemnity Co	Financial Indemnity
Fin Plan Strat	Financial Planning Strat
Financial Planning Strat	Financial Planning Strat
Financial Technology	Financial Technology
Financial Technology & ...	Financial Technology
Firemans Fund INS Co	Firemans Fund Insurance Co
Firemans Fund Insurance ...	Firemans Fund Insurance Co
First Interstate Bank	First Interstate Bank
First Interstate Bank Corp	First Interstate Bank
1st Merit Bank	First Merit Bank
1st Merit Bank Corp	First Merit Bank
>first Merit Corp	First Merit Bank
First Marret Bank	First Merit Bank
First Merit Bank	First Merit Bank
Fleming Foods	Fleming Foods
Fleming Foods-Phoenix ...	Fleming Foods
Flour Daniel Inc	Fluor Daniel
Fluor Daniel	Fluor Daniel
Fluor Daniel Inc	Fluor Daniel
Ford	Ford Motor
Ford Motor	Ford Motor
Frd Motor Co	Ford Motor

At the bottom of the Scheme pane, there are buttons for 'Add...', 'Edit...', 'Delete', and 'Options...'. The status bar at the bottom of the window shows 'Ready', 'QKB Locale: English (United States)', and 'NUM'.

Figure 1 – A sample data quality report showing clusters of potentially duplicate data.

Following a data quality implementation, an organization often has more than “clean” data. The process of building rules that guide the delivery of acceptable data creates a unique side-product: a data governance framework. The people and processes behind a data quality implementation must work in concert to agree on definitions of data elements and what constitutes “good” or “bad” data. This governance framework can be a valuable tool for any data management project.

Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution, continued

DATA INTEGRATION

Data integration often builds from the work done in the data quality world. The rules designed to manage the quality are perfectly suited as transformations during a data integration project. A classic example for this has been loading data into a data warehouse. An extract-transform-load (ETL) technology provides a method to upload data to a warehouse on a regular basis. In this instance, the “T” can be a set of data quality rules that can ensure that the data entering a data warehouse has a consistent level of quality.

Data integration can be as straightforward – or as complex – as an organization demands. It can be defined as moving data from source to target (single scope), or as a multi-layered process incorporating a variety of techniques, approaches and methodologies, all in an effort to deliver customized solutions for varying enterprise requirements. Once the process is complete, data can be accessed, consolidated, profiled, improved and made available to a variety of users as a single view for operational needs and analysis.

One of the most important steps to successful data integration is choosing a strategy. Many organizations rarely look beyond ETL – historically the most common method, but not necessarily the best choice for every type of business. Other options for data integration include:

- Data federation - Data federation allows for a virtual view of data across multiple data stores, without the requirements of actually moving or copying the data. While ETL moves data into a single repository, a federated view enables the data to remain in the source repository where it can remain physically unchanged. When an organization needs to access the data for business use, it uses a query processing capability to create a snapshot of the information required. In other words, a user simply specifies the information he or she needs to know, and it's delivered immediately in one neat, clean package without ever moving the original data.
- ELT (in-database/SQL “pushdown”) - An alternative to traditional ETL, some organizations might choose the extract, load and transform (ELT) method, also known as in-database transformation, where data integration occurs after the data has been loaded into a target database or repository. This speeds the time required to load a repository and allows for the transformations to occur within a high-capacity infrastructure.
- Real-time data integration employing Enterprise Application Integration (EAI) - Real-time data integration is about timeliness and accuracy; it's where data integration meets business and operations applications. Whether an organization needs to monitor up-to-the-minute purchase orders, oversee compliance regulations or comply with its customer service model by catching data mistakes before they reach the customer, real-time data integration is a key requirement.

MASTER DATA MANAGEMENT

MDM is the acknowledgement that the spread of information can have serious business impacts, particularly on an organization's daily operations. MDM provides a single view of a data element, such as customer, product or asset, regardless of the originating application.

MDM builds on the work done in data quality and data integration technologies. The rules created in a data quality program serve as the data integrity backbone of the MDM implementation (often known as data governance). Data integration technology is vital, as an MDM repository requires a way to control and manage the movement of data as it flows into and out of the hub repository. MDM adds to these technologies the ability to create a workflow for a new data element as it enters the enterprise. It also helps determine how that new record is processed within the master hub, and how changes are communicated back to enterprise applications.

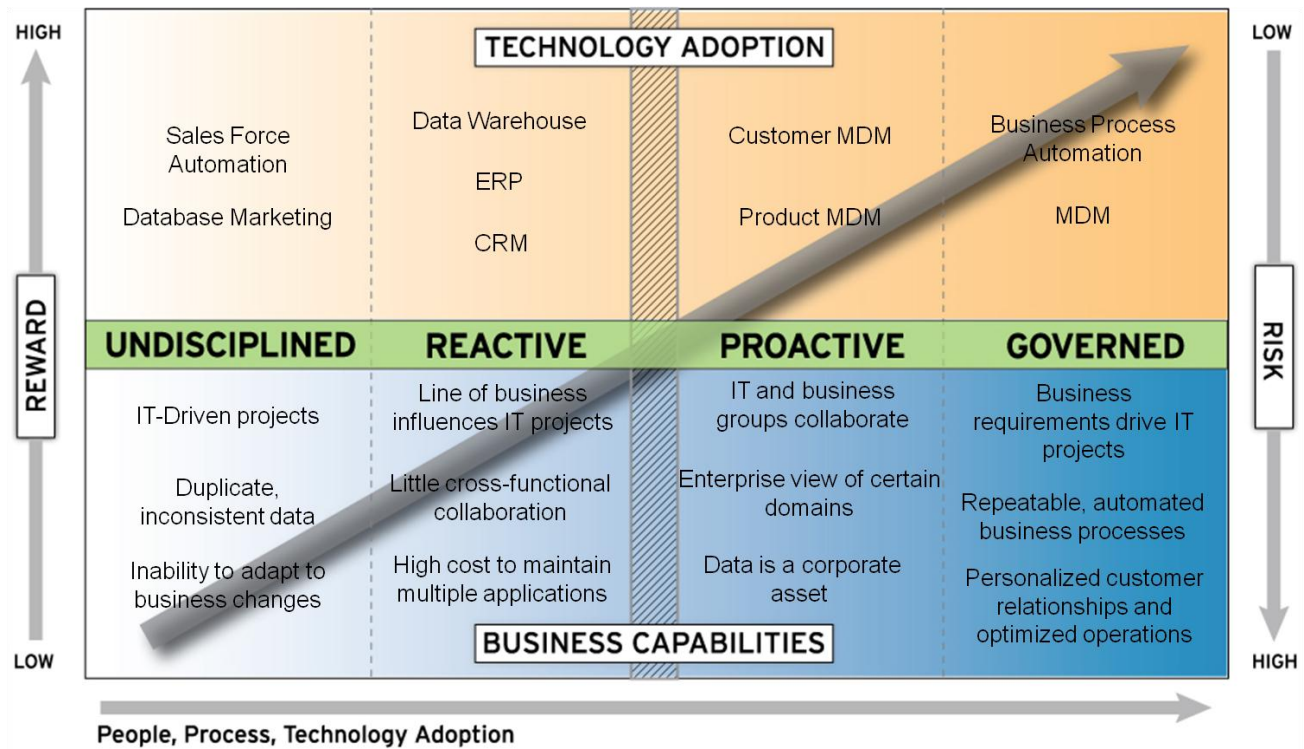
THE ROAD TO MASTER DATA MANAGEMENT

Like many enterprise projects, data management programs often start small before finding the sponsorship and support needed to transcend organizational boundaries. For most companies, data management or data governance takes on a slow but steady evolution as the company matures in its management and control of enterprise data.

Through an established Enterprise Data Maturity Model, organizations can identify and quantify precisely where they are – and where they can go – to create an environment that delivers and sustains high-quality information. An organization's growth toward this ultimate goal invariably follows an understood and established path. There are four stages, which are:

1. Undisciplined
2. Reactive
3. Proactive
4. Governed

Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution, continued



© Copyright DataFlux Corporation, LLC. All Rights Reserved.

Figure 2 – Enterprise data maturity model

Within the Maturity Model, each stage requires certain investments, both in terms of internal resources and third-party technologies. These investments lead to future rewards based on improved data that drives more effective, intelligent processes.

PROGRESSING THROUGH THE ENTERPRISE DATA MATURITY MODEL

The Enterprise Data Maturity Model examines the technology being utilized, along with the people and policies associated with the governance initiative, to ascertain the level of data governance sophistication within that enterprise. In the first stage – the Undisciplined phase, an organization has few defined rules and policies regarding data quality and data integration. The same data may exist in multiple applications and redundant data is often found in different sources, formats and records.

The danger for Undisciplined companies is the real and constant threat that the underlying data will lead to bad business decisions that may, in turn, result in missed business opportunities and decreased customer satisfaction. Often times, it takes a cataclysmic failure to shake the organization out of complacency.

At the next stage – the Reactive phase, a company begins to organize a data governance program, either through grassroots efforts or, more likely, through an executive-driven effort fueled by an earlier failure. At the Reactive stage, organizations try to reconcile the effects of inconsistent, inaccurate or unreliable data as bad records are identified. Here, the gains are often seen on a departmental or divisional level, but the company is starting to establish some best practices for data governance.

The move to the next stage – the Proactive phase, is not an easy one. After years of investing time and resources in complex enterprise applications (such as customer relationship management or CRM systems), a Proactive company understands that a more unified view is necessary if it wants to derive any real value from its information. Applications like CRM often become silos of data and to progress to a unified view and workable format, the organization needs to extend the reach of that data through the checks and balances of the maturity model technology to clearly manage the data and achieve master data management (MDM).

At the Proactive stage, the data governance program becomes cross-functional and has explicit executive support. To build a single view of a customer, for example, every part of the organization – sales, marketing, shipping, finance – has to agree on what attributes make up a customer record.

The final stage – the Governed phase, is where data is unified across data sources according to business rules

Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution, continued

established by an enterprise data governance team. At this final stage of the maturity model, a company has achieved a sophisticated data strategy and framework, and a major culture shift has occurred. Instead of treating issues of data quality and data integration as a series of tactical projects, these companies have a comprehensive program that elevates the process of managing business-critical data.

Although individual applications are still in use by a Governed company, the data that they access comes from a single repository that is propagated across the IT infrastructure. This provides the ultimate in control for the enterprise, as all reports and dashboards pull from the same pool of information.

UTILIZING THE MATURITY MODEL TO CREATE A REASONABLE IT APPROACH

The Enterprise Data Maturity Model helps organizations understand that they will not reach the highest levels of data management overnight. Rather, they should view the process as a journey, with a host of challenges and significant milestones along the way.

One customer that embraced the Enterprise Data Maturity Model is a large non-profit organization headquartered in the U.S. The organization wanted to build a more unified view of its donor network, whose data was splintered in dozens of applications across the country. The IT department decided on an MDM program and began to seek technology vendors to support the effort.

The organization found that although its intentions were good, the people and processes were not ready for that move. Political battles broke out to determine who “owned” the data. No single group was responsible for determining what constituted a good record. The bare necessities of data governance were not in place and initial trials were gridlocked.

Once the organization understood that it was an Undisciplined company according to the maturity model, it had a plan to progress to the next level. Business analysts within the company used data profiling technology to identify bad or incomplete data. The management team established a data stewardship group, which started to codify business rules for data quality and data integration efforts. Within months, the organization had progressed to the Reactive stage – and the goals of MDM were now within reach.

ACHIEVING DATA MATURITY

The phrase “Rome was not built in a day” does not only apply to the construction of capital cities. The same is true for any IT strategy, and particularly for an MDM initiative. Since the advent of MDM, companies have often rushed headlong to install the software to receive the anticipated benefit (“a single view of the customer,” for example) without considering the impacts – on both the IT and business staffs – of this move.

From an IT perspective, MDM requires connecting some – or all – business applications to a new repository of master data. This is a drain on manpower and may necessitate an external consultant to guide the process, particularly in more complex or disparate environments.

On the business side, the problems are less obvious but still quite compelling. The business side must start to come to some agreement on who owns the data, how data elements are defined, and how the organization will represent those items in the master repository. Often, the business units have not explored these aspects in the past, particularly across business units or divisions. And without guidance or consultation, this process can devolve into political battles and “turf wars” about data ownership across the enterprise.

The smart move? Realize that a master view of customers, products, inventory or any other data element is a journey, not a destination. MDM requires a monumental effort across the enterprise, and although the results look great on paper, the risk may not equal the reward. So, organizations are looking at ways to craft a smarter strategy for MDM deployments. And that often starts with a phased approach.

In a phased deployment, an organization can start on the path to MDM but view it as number of separate steps, each with a different reward. For example, as mentioned above, data quality is often viewed as a foundation component of MDM deployments. Starting out with a departmental or cross-functional data quality engagement will not only start to build the underpinnings of later MDM work, but these initiatives may yield initial benefits in the form of improved customer data (leading to more cross-sell or up-sell opportunities) or more valid product data (helping to lower inventory costs and reduce production time).

For a phased approach, an organization must consider a number of areas that are part of the MDM program – and how to work with each of these in concert to achieve a final goal of a master repository. The four key enterprise components for MDM are:

- **People** – Who is involved and what contributions must they make? The best functioning organizations understand that business issues should drive data governance but that IT plays a critical role.

Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution, continued

- Process – What activities must be performed? What business rules must be available to properly govern data? In undisciplined and reactive organizations, data is “owned” by a group. It is not governed by the organization at large and the concept of “sharing” data is a sign of weakness. Proactive and governed organizations believe power derives from sharing information.
- Technology – What investments in technology are necessary? How should these investments be staged? Reaching a mature data state is not about ripping and replacing current systems and solutions or buying one solution that will “solve” the problem. It’s figuring out where you are and choosing solutions that close critical gaps based on sound policies. In some cases, an organization can’t get its data in order because too much of the wrong kind of data is being gathered. To succeed, some solutions need to be pared back, rather than new ones purchased.
- Risks and rewards – What risks does the organization face at the current stage? And what could it gain from progressing forward? This may be the most important consideration for each part of a company’s progression through the maturity model – how much is the next step worth to the organization?

As more of these components begin to mature, the tenets of MDM will begin to solidify, leading to a more cohesive MDM strategy. By concentrating on these four areas, a company can build a blueprint of the people, processes and technologies that need to move forward to achieve MDM success.

DATAFLUX qMDM

DataFlux qMDM incorporates a deep background in data quality and data integration engagements. The DataFlux core data quality engine can be configured to parse, standardize, validate and verify virtually all data types in an MDM environment. DataFlux qMDM can streamline and standardize data, forming a central data repository that enables consistent, uniform control of data across the enterprise. With MDM, organizations can enhance and link enterprise information to create an accurate, unified view.

DataFlux qMDM is a combination of software, templates, documentation, data model and services that together provide the functionality and processes necessary to build and maintain a master entity database. An entity in this case could be a customer, a product, a patient, a site or any other business data object you might define. You can define the attributes for each entity and define transformations that support data quality and identity management functionality.

The DataFlux qMDM database is a set of tables and table relationships that contain entity information and required entity keys and linking information to give end-users or other applications the ability to see a “complete” view of the entity and its relationships with other entities. In addition, the DataFlux qMDM database can be used as the sole source for creating and maintaining “global” entity IDs that can be used for reference in other applications or processes. Source system IDs, important to other enterprise applications or data sources, will also be kept in the DataFlux qMDM database to facilitate linking activities with other data providers and/or consumers.

TECHNOLOGY ARCHITECTURE

The basic architecture for DataFlux qMDM uses DataFlux Data Management Studio with the DataFlux Data Management Server working together to build and manage the MDM database. Data Management Studio can be used alone to build and maintain DataFlux qMDM in batch mode in a Windows environment while the Data Management Server and the associated service-oriented architecture (SOA) is needed for non-Windows environments and for real-time/transactional usage.

Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution, continued

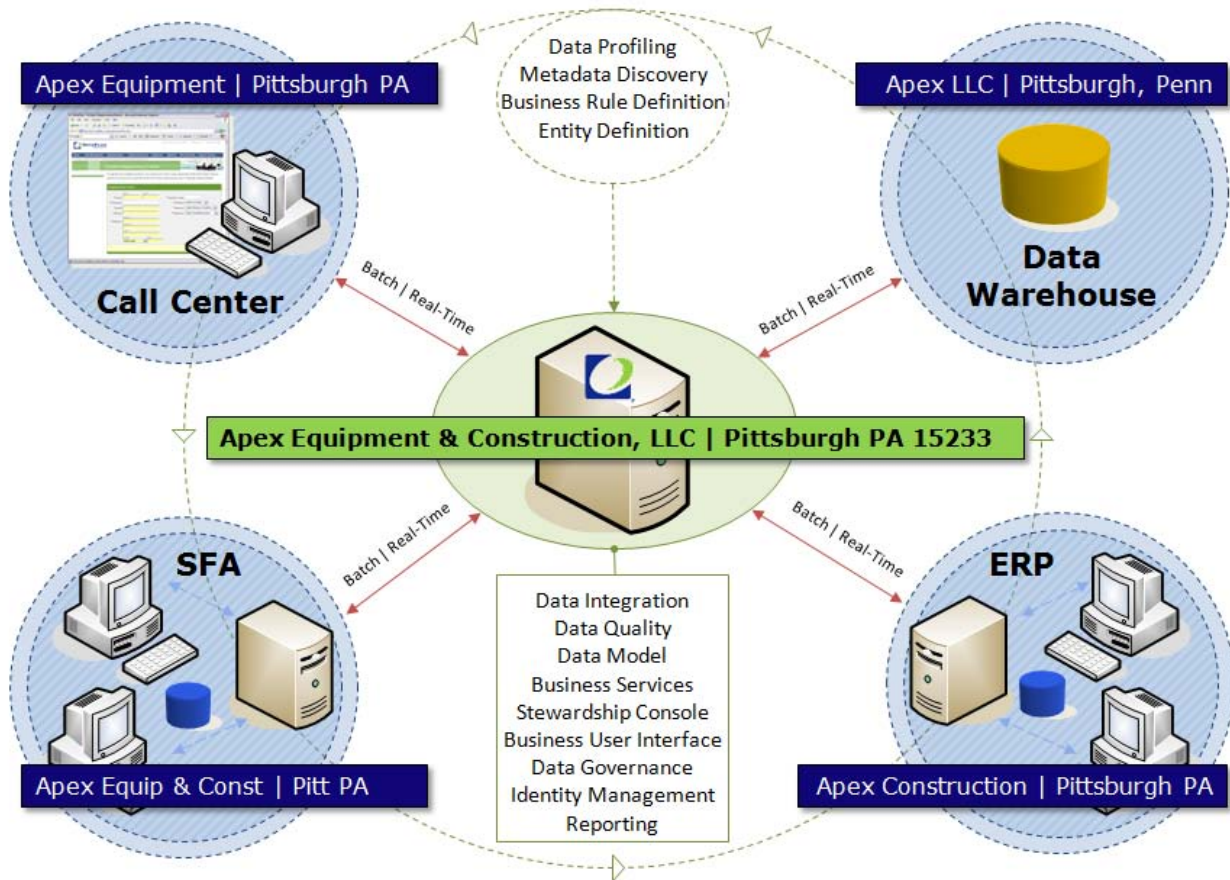


Figure 3 - DataFlux qMDM Solution overview

Data Management Server

Data Management Server is a technology that enables data management and integration processes to be deployed across an IT environment. Data Management Server mainly has the job of performing tasks as indicated in batch jobs or real-time web services. These jobs and services specify activities like database connectivity, data transformation, querying, and so on. Data Management Studio is used to edit workflow templates, and Data Management Server is responsible for processing the work instructions in the templates. Data Management Server can optionally use DataFlux Authentication Server to provide a security layer for job/service access.

Data Management Studio

Data Management Studio is a Windows application that provides the functionality needed to view, build and edit DataFlux qMDM job templates and business services. The job templates are starting points for business services that bring data from many different databases or on a per-record basis into a DataFlux qMDM hub and apply specified transformations to the data such as address verification or record merging. In addition, Data Management Studio direct editing of the Quality Knowledge Base (QKB) contains all of the metadata and algorithmic information needed to apply data quality processes like name standardization, address verification, and so on.

Architect Job Templates/Business Services

As previously mentioned, Data Management Studio can be used to modify and build work flows called jobs. These jobs are delivered as templates, which can be customized by consultants or other IT professionals. These jobs can be used in a batch mode and jobs used in a real-time mode are also known as business services. Many job templates are delivered with the solution to accommodate such things as address verification, merging, assigning IDs, standardizing data, querying data, etc.

Quality Knowledge Base Locales

The QKB contains the files, file relationships and metadata needed to correctly parse, match, standardize and

Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution, continued

otherwise process data as required for DataFlux qMDM. The QKB can be customized to meet unique client needs and new locales can be utilized to internationalize DataFlux qMDM.

Master Repository Manager

The Master Repository Manager is a web-based application for direct interaction with the DataFlux qMDM database. Services can be accessed by an administrator or data steward who can work through the Data Management Server to query the database or make changes directly to it. Business users can also generate ad hoc reports based on information in the DataFlux qMDM database. Figure 4 shows how a user can enter master data directly into the hub repository via the Master Repository Manager.

The screenshot displays the Master Repository Manager web application. The top navigation bar includes 'Dashboard', 'Entities', 'Hierarchies', 'Reports', 'Administration', 'Services', and 'Accelerators'. The main content area is divided into several sections:

- Entity Search:** Shows 'Entity Type: Person' and search options like 'DataFlux Match', 'Search', 'Clear', and 'Show Retired Entities'.
- Person Attribute Browser:** A table listing attributes such as 'Database Name', 'Display Name', 'MDM_SRC_SYS_REC_ID', 'MDM Source System Recor', 'MDM_QKB_LOCALE', 'MDM QKB Locale', 'FULL_NAME', 'Full Name', and 'SURNAME1', 'Surname'.
- Entity Form:** A form for editing a 'Person' entity. It includes sections for 'MDM Identifiers' (MDM QKB Locale: ENUSA) and 'Personal Information' (Full Name: KELLY P SMITH, Name Title: Mrs, Given Name: KELLY, Middle Name: P, Surname: SMITH, Birth Date: 04/11/1994).
- Search Results:** A table showing search results for 'Person (3)'. The table has columns for MDM Entity ID, MDM Best Record, MDM Source System, MDM Source System ID, MDM QKB Locale, MDM Last Modified, * Full Name, Name Title, Given Name, and Middle Name. Two rows are shown for 'KESHA K JONES'.

MDM Entity ID	MDM Best Record	MDM Source System	MDM Source System ...	MDM QKB Locale	MDM Last Modified	* Full Name	Name Title	Given Name	Middle Name
986	SURVIVOR	qMDM	986	ENUSA	08/17/2010 10:23:47 ...	KESHA K JONES	Mrs	KESHA	K
496	NONE	ERP	6	ENUSA	08/17/2010 10:23:47 ...	KESHA K JONES	Mrs	KESHA	K

Figure 4 – The Master Repository Manager provides a data steward interface for the master hub.

DataFlux qMDM Database

An integral part of DataFlux qMDM is the database. At a minimum it supports essential entity attribute storage and important metadata and linkage information for each entity. Users may choose to keep most of their entity data in this database but DataFlux qMDM can also support a “registry”-style deployment where only critical pieces of information are stored in the database and references back to original source systems are used to build virtual views of the data when needed.

Web Services and the SOA Environment

The real-time/transactional piece of the solution is surfaced to other applications through an SOA environment. Specifically, the ability to call business services hosted on the Data Management Server will be done through web services. Batch jobs derived from base DataFlux qMDM templates can be repurposed as business services that accept data on a record-by-record basis. This aspect of the solution allows enforcement of business logic, exception handling, identity management and data quality across the various modes of data processing.

Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution, continued

REFERENCES

¹ Kelly, Jeff. "Poor data quality costing companies millions of dollars annually." SearchDataManagement.com. Aug. 25, 2009. <http://searchdatamanagement.techtarget.com/news/1365965/Poor-data-quality-costing-companies-millions-of-dollars-annually>

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Pat Herbert
Enterprise: DataFlux
Address: 940 NW Cary Pkwy, Suite 201
City, State ZIP: Cary, NC 27513 USA
Work Phone: 919-447-3000
Fax: 919-447-3100
E-mail: pat.herbert@dataflux.com
Web: www.dataflux.com

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.