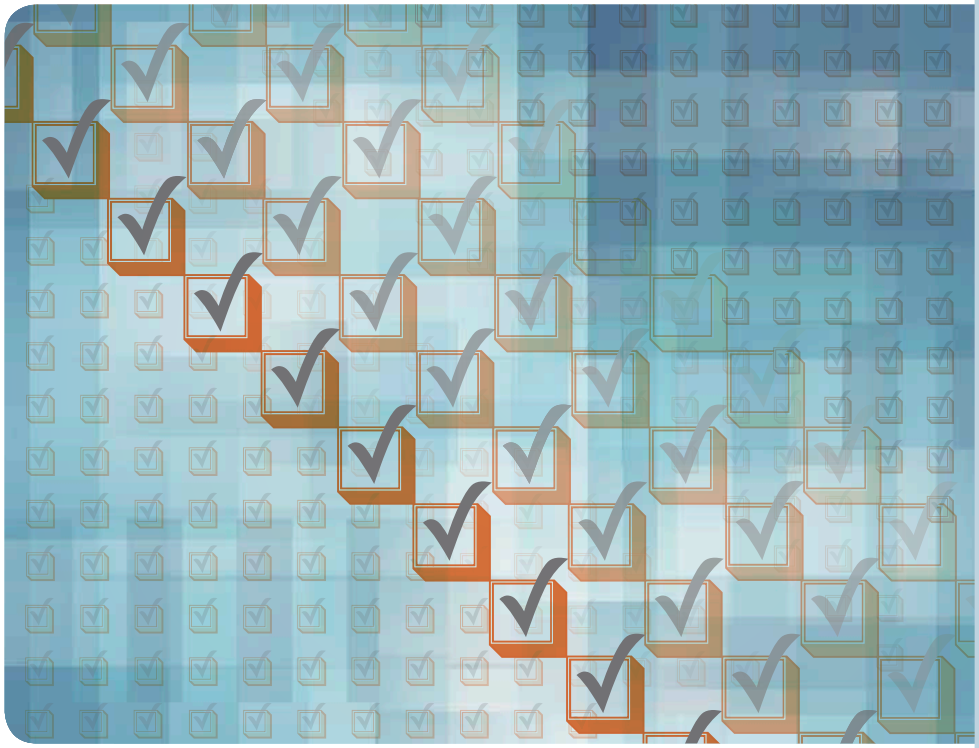
TDWI RESEARCH

TDWI CHECKLIST REPORT

# Enterprise Data Management

By Philip Russom



Sponsored by

DataFlux
A SAS COMPANY

www.tdwi.org

tdwi
THE DATA WAREHOUSING INSTITUTE

OCTOBER 2009

TDWI CHECKLIST REPORT

# Enterprise Data Management

By Philip Russom

## TABLE OF CONTENTS



THE DATA WAREHOUSING INSTITUTE

1201 Monster Road SW, Suite 250
Renton, WA 98057

T 425.277.9126
F 425.687.2842
E info@tdwi.org

www.tdwi.org

## FOREWORD

In most organizations today, data and other information are managed in isolated silos by independent teams using assorted data management tools for data quality, integration, governance, meta- and master data management (MDM), content management, and so on. From a technology viewpoint, the lack of coordination among data management disciplines leads to redundant team staffing and limited developer productivity. Even worse, competing data management solutions can inhibit data's quality, consistency, standards, scalability, architecture, and so on. From a business viewpoint, data-driven business initiatives suffer (including BI, CRM, and business operations) as a result of poor data quality and incomplete information, inconsistent data definitions, noncompliant data, and uncontrolled data usage.

Forward-looking organizations are solving these technology and business problems by adopting enterprise data management (EDM), which TDWI Research defines as:

> A best practice for unifying diverse data management disciplines, so that data and other information are managed according to enterprisewide goals that promote technical efficiencies and support strategic, data-oriented business goals.

EDM focuses on a handful of key tasks:

**Coordinating multiple data management disciplines.** EDM is primarily about unifying teams and tools for data quality, data integration, MDM, and data governance. It also coordinates secondary disciplines, including metadata management, data profiling, and data modeling.

**Expanding the scope of data management.** EDM by nature has an enterprise scope that reaches across operational and analytic IT environments. EDM is accretive over time.

**Aligning data management work with business goals.** Although EDM is initially about coordinating data management functions, it should eventually lead to better alignment between data management work and information-driven business goals of the enterprise.

**Balancing user practices and vendor products.** EDM is both a user-oriented best practice (which users may implement with a variety of vendor products and hand-coded solutions) and a type of unified data management platform (which a few leading software vendors are producing). This TDWI Checklist Report focuses on the former.

## NUMBER ONE

ADOPT EDM FOR GREATER DATA QUALITY, INTEGRATION, CONSISTENCY, AND GOVERNANCE.

Enterprise data management yields several benefits:

**EDM stimulates collaboration for far-flung technical functions.** Specialists from diverse data disciplines can learn from one another, avoid having to re-create the wheel, and create and enforce cross-solution development and data standards.

**EDM supports data-driven corporate objectives.** EDM's information infrastructure enables information-oriented business goals, such as 360-degree CRM, operational business intelligence, and regulatory compliance.

**EDM fosters cross-functional consistency.** Common definitions of business entities expressed through MDM, shared metadata, and cross-system data standards enable consistency. The beneficiaries are business processes that span multiple applications and departments, such as customer identification, financial closings, and insurance claim processing.

**EDM helps the business leverage communal data assets.** EDM's collaborative spirit and common technical infrastructure can turn data into a shared enterprise asset, instead of a departmental property.

**EDM makes data more auditable, which assists with compliance.** A good EDM program will improve the quality of master data and metadata, plus tighten policies for data usage, security, and privacy. Combined, these improvements make data's meaning, lineage, and usage easy to corroborate in audit and compliance situations.

Despite the numerous benefits, there are also barriers to EDM:

**Changes to data management teams can slow down EDM.** Hence, some organizations avoid redistributing data management professionals into new teams and instead coordinate them through a data governance program. Others go so far as to centralize data management disciplines into one or more competency centers.

**Data ownership can inhibit EDM.** EDM should improve data (and data's infrastructure) so it is more easily integrated as an "enterprise asset" across multiple organizations. EDM needs an executive mandate to prevent the owners of data and applications from resisting such sharing. That same mandate should lead stakeholders to adopt common definitions of important business entities, such as customers, products, financials, employees, and so on.

**Existing tool sets may not be conducive to EDM.** An organization may be using several tools for various EDM tasks, which complicates tool integration. EDM works best when diverse tools are integrated at both the development and deployment levels.

## NUMBER TWO

BASE YOUR EDM FRAMEWORK ON PRIMARY DISCIPLINES, SUPPORTED BY SECONDARY ONES.

There are many data management disciplines, but for most organizations, four primary disciplines constitute the core of EDM:

**Data quality (DQ).** An important goal of EDM is to improve data so it can be better leveraged as an enterprise asset. DQ tools and practices satisfy this requirement in spades by cleansing, standardizing, verifying, deduplicating, and enhancing data.

**Data integration (DI).** Another goal of EDM is business integration. This is mostly about sharing data across organizational boundaries so disparate departments and business units have an enterprise view of key business entities (customers, products, financials, and so on). DI tools and practices provide critical infrastructure toward this end.

**Master data management.** EDM also improves and standardizes semantic data so that apples-to-apples associations are made when multiple IT systems manage data about the same entity. MDM has become the preferred method for ensuring consistent associations in data.

**Data governance (DG).** Although DG is not a data management discipline per se, it has become a critical success factor for core EDM disciplines because it supports them with cross-discipline collaboration, change management procedures, and a process for developing and policing data standards and data usage policies.

Besides the primary core disciplines, there are several supporting data management disciplines, which fall into two broad categories:

**Supporting data management practices and their tools.** These secondary disciplines support the primary disciplines and are typically shared across multiple core disciplines. Supporting disciplines include metadata management, data profiling, data monitoring, and data modeling.

**Supporting infrastructure.** This provides interoperability among multiple data management design and deployment environments, plus connectivity with enterprise operational and analytic applications. Note that EDM's infrastructure is not a single monolith; it's an amalgam of interfaces and services supported by its constituent tools.

Put together the primary and secondary disciplines and supporting infrastructure of EDM, and you have an EDM framework, as illustrated in Figure 1. Note that EDM, when organized this way, is a unified hub for data improvement, integration, and collaboration. The hub unifies the multiple data management disciplines involved, as well as disparate enterprise databases and applications and the departments that use them.

Realize that your EDM framework may vary (as compared with Figure 1) based on your organization's needs or maturity. For example, if MDM is not yet a priority, it may not be one of the core disciplines. Many organizations expend much effort on metadata management, such that it's more of a core discipline than a supporting one. In organizations that handle valuable textual information, search, text analytics, or content management may be important supporting or core disciplines.

The rest of this Checklist Report dives into each of the four primary data management disciplines of EDM, as well as EDM's supporting tools and supporting infrastructure.

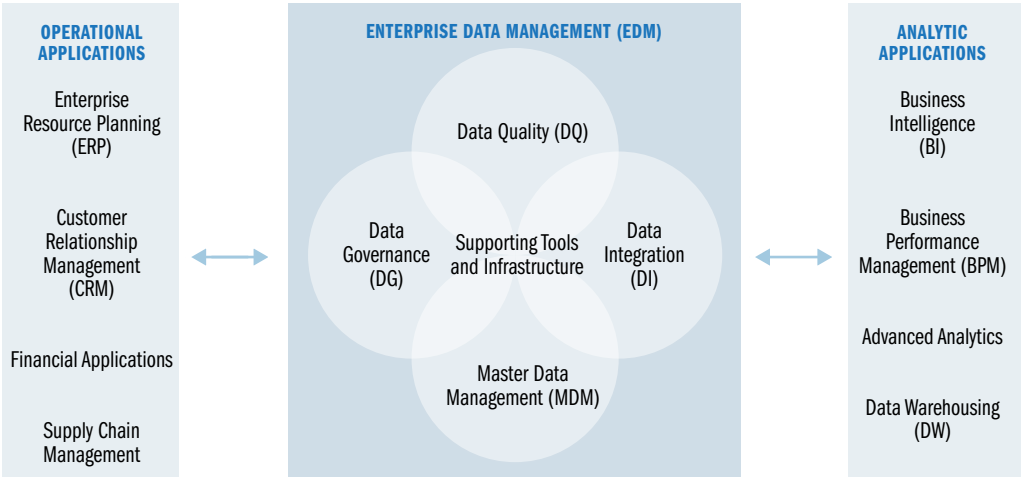



**Figure 1.** Enterprise data management as a unified hub for data improvement, integration, and collaboration.

## NUMBER THREE

APPLY DATA QUALITY FUNCTIONS TO IMPROVE ALL DATA TOUCHED BY EDM.

Data quality (DQ) is a primary data management discipline for enterprise data management (EDM), because improving the content of data makes data far more valuable in general as a shared enterprise asset. DQ also contributes directly to smoother business operations, direct marketing effectiveness, stellar customer service, more accurate decision-making, and so on. The current state of the art in DQ is well suited to EDM's enterprise scope, because DQ applies to both operational and analytic applications, as well as many data domains and business departments.

We say "data quality" as if it were a single monolithic practice. But it's actually a collection of techniques and tool types, including name-and-address cleansing; data standardization, verification, and validation; data enhancement (sometimes called data append); and multiple forms of matching, merging, and deduplication. EDM entails coordinating multiple DQ techniques, as well as coordinating the complete DQ discipline with other data management disciplines.

DQ works well with and complements other data management disciplines in the context of EDM. For example, DI solutions inevitably ferret out data problems that need correction, as well as opportunities for data improvement. DQ solutions can reach more data that needs improvement when assisted by DI. In many firms, MDM is an outgrowth of the DQ program, because MDM improves master, reference, and other semantic data, similar to how DQ improves physical data. The new practice of data governance, in some organizations, grew out of data stewardship, a practice that originated in DQ initiatives.

Many of the secondary support disciplines found in a full-blown EDM program originated and matured within the larger practice of DQ. In particular, without leadership from DQ, we wouldn't have the modern practices of data profiling (to study data in the early phases of solution development) and data monitoring (which re-profiles data as the deployed data quality solution operates on it daily, to ensure that quality metrics are met). Similarly, data stewardship, an important precursor to data governance, was popularized by DQ initiatives as a way of communicating DQ needs and priorities between business and technical people.

## NUMBER FOUR

SHARE ENTERPRISE DATA ASSETS BROADLY VIA DATA INTEGRATION INFRASTRUCTURE.

Data integration (DI) is a primary data management discipline for EDM because it enables one of EDM's top priorities: business integration. Much of EDM's data sharing is accomplished through the infrastructure provided by DI tools, servers, and interfaces. DI divides into two related practices: analytic data integration (largely for data warehousing) and operational data integration (for the migration, consolidation, or synchronization of operational databases). The two combined reach the entire enterprise, which EDM generally demands.

As with DQ, DI is also a collection of related techniques and tool types. These include extract, transform, and load (ETL), its variation ELT, data federation, replication, synchronization, changed data capture, and various hand-coded approaches. Coordinating these techniques with each other, as well as coordinating the greater DI discipline with others, is key to DI's success with EDM.

Increasingly, users implement DI and DQ in tandem. DQ fixes data problems that DI discovers, while DI gives DQ access to more data. In deployment, it's common today that DI and DQ servers call each other as peer components of a single solution. These unified DI/DQ solutions are the very epitome of EDM.

DI depends heavily on secondary EDM disciplines. For example, DI solutions simply aren't possible without hefty doses of metadata management. For this reason, all mature DI tools include sophisticated metadata management functions and a metadata repository. Most tools involved in EDM can manage metadata to some degree, but many EDM programs manage most metadata through a DI tool because it's the most capable. Metadata aside, DI in recent years has come to rely on data profiling and data monitoring, which it inherited from DQ.

In many EDM programs, DI is the bulk of the supporting infrastructure. After all, DI tools and servers are expert at networking with other tools, databases, and applications via a wide array of open interfaces, proprietary application programming interfaces (API), gateways, data standards, Web services, and service-oriented architectures (SOA). However, for broader reach, EDM programs should complement DI with tools for enterprise application integration (EAI).

## NUMBER FIVE

ACHIEVE CONSISTENT DATA DEFINITIONS AND DATA USAGE VIA MDM.

To achieve business integration and other goals of EDM, diverse software tools, applications, and databases share growing volumes of information. In these situations, it's difficult to find the correct or best sources of data and map them accurately between systems. The modern practice of MDM brings accuracy to this task by enabling users to define the meaning of data and how it should be used across various types of applications and their business units, whether in operational processes or analytic decision-making. Because its data definitions bring consistency to data usage, MDM is just as critical to the intense data sharing of EDM as is DI.

Several variations of MDM exist. Some focus on a specific data domain, as with customer data integration, product information management, and financial hierarchy management. Other variations focus on a specific application. For instance, MDM is a critical component of ERP and CRM applications, especially when multiple instances must be synchronized consistently. Similarly, MDM may be applied within a data warehouse for consistency across BI's many reports and analyses. Because of these myriad foci, many organizations deploy multiple MDM solutions, and all or some of these may need coordination through EDM.

Like all primary disciplines for EDM, MDM has synergies with other data management practices, both primary and secondary. Although dedicated MDM applications are available from several vendors, many users prefer to build homegrown MDM solutions using their DI and DQ tools. In many organizations, MDM is an extension of the older practice of metadata management. This makes sense, because both involve the management and improvement of semantic data. Recently, many users have discovered that data profiling is also useful for MDM, not just DQ and DI.

## NUMBER SIX

USE DATA GOVERNANCE LIKE A GLUE THAT HOLDS EDM TOGETHER.

Data governance (DG) is usually manifested as an executive-level data governance board, committee, or other organizational structure that creates and enforces policies and procedures for business use and technical management of data across the entire organization. By its very nature, DG creates a process for coordinating multiple data management disciplines and aligning them with business requirements. This is exactly what EDM needs, so for many organizations, it makes sense to apply DG's coordination and alignment processes to EDM.

Common goals of data governance are to improve data's quality; remediate its inconsistencies; share it broadly; leverage its aggregate for competitive advantage; manage change relative to data usage; and comply with internal and external regulations and standards for data usage. Of course, many of these are also goals of EDM, so again it makes sense to coordinate these data management goals via DG.

DG aside, the other primary data management disciplines of EDM all involve various forms of change. For example, DQ specialists often request changes to operational applications (or to how business people use them) to correct quality issues upstream where they originate. Likewise, MDM specialists may request changes to tools, applications, and databases, so they comply with standardized definitions of business entities established by the MDM team. And DI specialists continually request read and write access to a growing list of enterprise data sources and targets. This is a lot of change to manage, and the processes and procedures of DG effect a kind of change management mechanism for EDM.

This report recommends data governance, although other organizational structures can play an equivalent role in gluing together the pieces of EDM. For example, there are other forms of governance, including corporate governance, IT governance, and BI governance. Many programs for business intelligence, data stewardship, and enterprise data architecture fulfill governance functions, to a degree, as do some data-oriented competency centers and centers of excellence. In fact, many firms begin with one of these preexisting organizational structures and evolve it to become the data governance board. EDM can benefit from any of these, as long as the organizational structure is data-oriented, cross-functional, collaborative, and process-driven.

Many secondary data management techniques and their tools fulfill a support role in EDM, in that they support the primary, core disciplines and are typically shared across multiple core disciplines.

## NUMBER SEVEN

EXTEND EDM'S CORE WITH ADDITIONAL, SUPPORTING DATA MANAGEMENT DISCIPLINES.

For example, metadata is everywhere, so each of the primary disciplines must involve some form of metadata management. Even so, metadata management also needs to be centralized, so that standard metadata can be shared across multiple systems. The point of standardized metadata is to ensure that descriptions of data are complete and meaningful, as well as consistent when multiple applications and databases handle common data. This is very similar to what MDM does for master and reference data, which is why MDM and metadata management should be tightly coordinated. In the context of EDM, the sharing of common MDM data definitions and standard metadata is often done via a central repository.

Data modeling likewise contributes broadly to EDM. Modeling target databases and other data structures is key to data warehousing, customer data integration, and data migrations. Interestingly, many DI tools include data modeling capabilities, so users must decide whether to model targets in a DI tool or with a dedicated modeling tool. When the business entities being defined in an MDM solution are hierarchical (as with a chart of accounts or a bill of materials), modeling the hierarchal definitions may require a data modeling tool.

Data profiling has come a long way this decade. Practiced with mostly manual methods in the 1990s, data profiling today is more often performed with a dedicated tool or with the data profiling functions built into almost all DQ and DI tools. Data profiling and its sibling practice data monitoring are important for all primary EDM practices, because they help you discover data you weren't aware of, as well as its dependencies and changes over time.

The list of secondary, supporting data management disciplines varies greatly. Depending on your organization's requirements and existing systems, additional disciplines may include text analytics and search, operational database administration and architecture, or process management and integration.

## NUMBER EIGHT

ENABLE EDM'S CONNECTIVITY AND INTEROPERABILITY THROUGH A SHARED INFRASTRUCTURE.

EDM's shared infrastructure is not a single monolith; it's an amalgam of interfaces and services supported by its constituent tools. Even so, many users depend heavily on their DI tools for interoperability among multiple data management design and deployment environments, plus connectivity with enterprise operational and analytic applications.

EDM's supporting infrastructure must support a wide range of traditional interfaces, both open standards (such as ODBC, JDBC, ADO.NET, and Java) and proprietary APIs. Web services and SOAs are quickly becoming the preferred media for EDM's connectivity and interoperability.

Organizations with broad EDM ambitions also need support for data exchange standards, integration with leading enterprise applications, and EAI or message-oriented middleware. Even so, just about any tool involved in EDM will have networking and interface capabilities that may apply to EDM. Note that EDM's connectivity and interoperability will have diverse requirements for operating in batch, real time, and many gradations in between.

Acquiring a long list of data management tool types for an organization's EDM portfolio is important, perhaps even a base requirement. Equally important is that each piece of the portfolio integrates and interoperates with others appropriately. To set the proper expectation, don't assume 100% interoperability; design tool and server integration should be selective, based on prominent business and technology needs.

In design and development stages, EDM's shared infrastructure should enable multiple data management tools (both primary and secondary) to share common development artifacts, general business rules, and semantic data (such as MDM data definitions and standardized metadata). In deployment, EDM's infrastructure enables the interoperability of multiple data management servers, which is a hallmark of EDM. Finally, EDM's infrastructure fosters general connectivity to a wide range of enterprise data sources and targets in support of EDM's ultimate goal: the broad sharing of data that is clean, complete, consistent, compliant, and current.

## ABOUT OUR SPONSOR


DataFlux enables organizations to analyze, improve, and control their data through an integrated technology platform. With DataFlux enterprise data quality and data integration products, organizations can more effectively and efficiently build a unified view of customers, products, suppliers, or any other corporate data asset. A wholly owned subsidiary of SAS (www.sas.com), DataFlux helps customers rapidly assess and improve problematic data, and build the foundation for enterprise data governance. Effective data governance delivers high-quality information that can fuel successful enterprise efforts such as risk management, operational efficiency, and master data management.
www.dataflux.com

## ABOUT THE AUTHOR

**Philip Russom** is the senior manager of TDWI Research at The Data Warehousing Institute (TDWI), where he oversees many of TDWI's research-oriented publications, services, and events. He's been an industry analyst at Forrester Research, Giga Information Group, and Hurwitz Group, where he researched, wrote, spoke, and consulted about BI issues. Before that, Russom worked in technical and marketing positions for various database vendors.

## ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data warehousing solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide Membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

## ABOUT THE TDWI CHECKLIST REPORT SERIES

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.