TEXAS DATA MANAGEMENT FRAMEWORK

Fast Start Learning Guide

OFFICE OF THE CHIEF DATA OFFICER, STATE OF TEXAS

Table of Contents

Introduction

In 2016 the Office of the Statewide Data Coordinator formed the Texas Enterprise Information Management, or TEIM, group. The group was chartered to establish a data and knowledge sharing networking community to help foster collaboration and partnership across state agencies, institutions of higher education, and local government.

A working subgroup of TEIM, composed of multiple state agencies, reviewed and compared several methodologies for implementing a data management program. After careful consideration, the subgroup selected DAMA International's Data Management Body of Knowledge, or DAMA-DMBOK, as the standard that Texas organizations could use to build their individual data programs.

The DAMA-DMBOK describes key principles such as data governance, data architecture, metadata management, reference and master data management, data security, data quality, data ethics, and other important foundational elements that data programs need to incorporate to be successful. The methodology is an overall collaborative creation with input from the leading data management professionals in the industry.

The DAMA-DMBOK is a comprehensive reference manual that provides the reader with extensive detail, examples, and models. To help organizations understand the core components, the Office of Chief Data Officer within the Texas Department of Information Resources analyzed the DAMA-DMBOK (second edition) and created this Fast Start Learning Guide, or FSL.

The intent of the FSL is to capture the critical points of each DAMA-DMBOK chapter so that they can be easily used by members of the Texas data community to help develop their plans and communicate the importance of each key principle to the leadership within their organization. Each of the chapters in the FSL addresses the following:

- What is this key principle?
- Why is this key principle important?
- How does the key principle apply to you?
- When do you use the key principle?
- Who is involved with the key principle?

Each chapter also summarizes key takeaways and, where relevant, guiding principles. Additionally, there is a comprehensive glossary describing significant terms.

Data, if managed appropriately, can be a powerful strategic asset that provides significant insight into an organization. In Texas, we are not short on data and the volume continues to grow each day. To meet this demand, organizations have begun to establish data management offices and hire data professionals to evaluate and assess the maturity of their data assets. Organizations are building data programs that will help them meet their mission goals and objectives. By following the data management best practices outlined in this FSL, agencies, institutions of higher education, and local governments will be able to better serve their customers and use data to drive efficiency and effectiveness for the state of Texas.

Ed Kelly

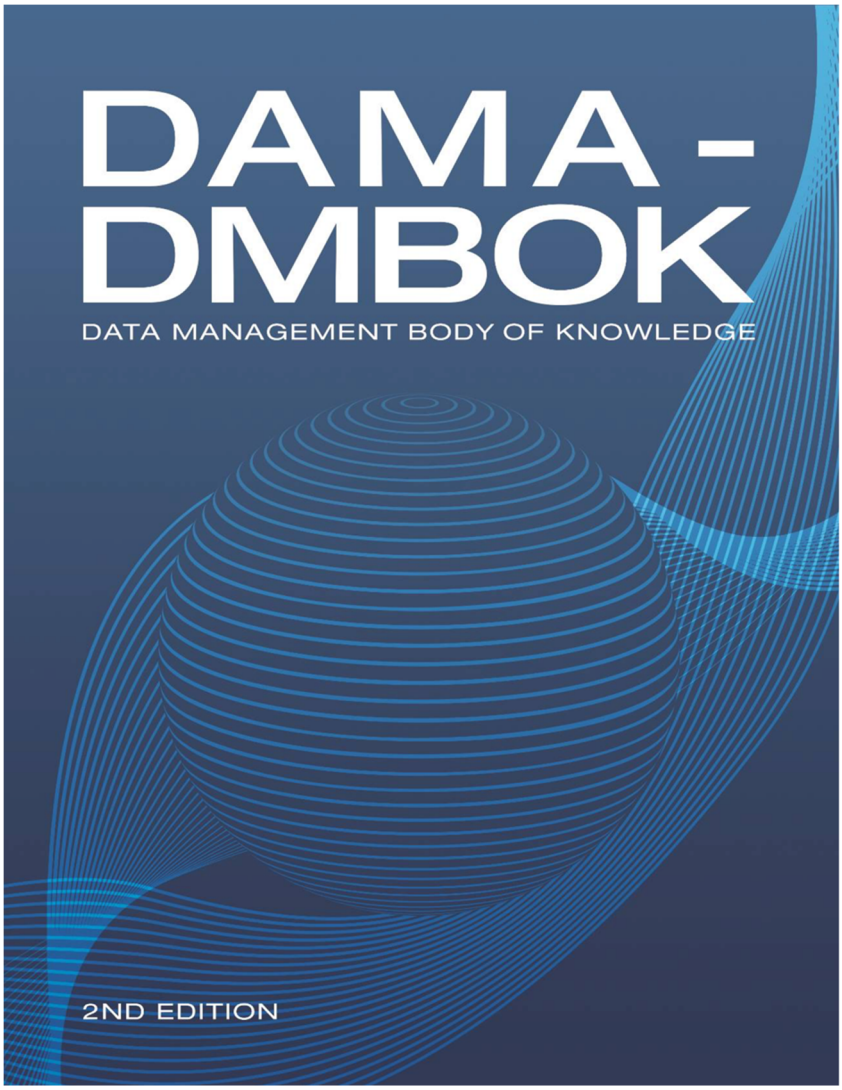
Ed Kelly
Chief Data Officer, State of Texas
October 2019

How to Use

This document is based on DAMA International, *DAMA-DMBOK: Data Management Body of Knowledge*, 2nd ed. (Basking Ridge, NJ: Technics Publications, 2017).

Please note that the page numbers of the printed version of the DAMA-DMBOK differ from the page numbers in the digital version.

The page numbers in the Fast Start Learning Guide are based on the page numbers in the printed version.



Data Management

What Is Data Management?

Data management is the development, execution and supervision of plans, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their life cycles (Figure 1). There are several individual key knowledge areas that make up data management (Figure 2).

Why Is Data Management Important?

Data management enables an organization to:

- Make consistent, informed decisions about how to obtain strategic value from data
- Establish better efficiency, effectiveness, and quality in organizational operations for both business and technical goals
- Ensure the integrity of data assets and the privacy and confidentially of stakeholder data

FIGURE 1: DATA LIFE CYCLE

How Does Data Management Apply to You?

Data management allows you to find the value within the data that your organization possesses and put it to use.

- Applying context to and drawing meaning from data provides information and knowledge that will help you serve your customers.
- Having reliable, high-quality data about your customers, services and operations means you can make better decisions and maintain customer confidence.

When data is not properly managed, there is waste, inefficiency, missed opportunities, and lost constituent and legislative trust (just as when any capital asset is not managed well).

When Do You Use Data Management?

Data management is considered when developing or modifying an organization's strategic goals.

Data management (with its corresponding data strategy), helps establish identifiable links between operational actions and long-term goals. The components of a data management strategy include:

- A compelling vision for data management
- A business case for data management, with examples
- Guiding principles, values, and management perspectives
- The mission and long-term directional goals of data management
- Measures of data management success
- Short-term (twelve to twenty-four months) data management program objectives
- Descriptions of data management roles, with their responsibilities and decision rights
- Descriptions of data management program components and initiatives
- A prioritized program of work with scope boundaries
- A draft implementation roadmap with projects and action items

When the strategic planning process is complete, you should have a:

- **Data management charter.** Overall vision, business case, goals, guiding principles, measures of success, critical success factors, recognized risks, operating model, etc.
- **Data management scope statement.** Goals and objectives for some planning horizon (usually three years) and the roles, organizations, and individual leaders accountable for achieving these objectives
- **Data management implementation roadmap.** Identifying specific programs, projects, task assignments, and delivery milestones

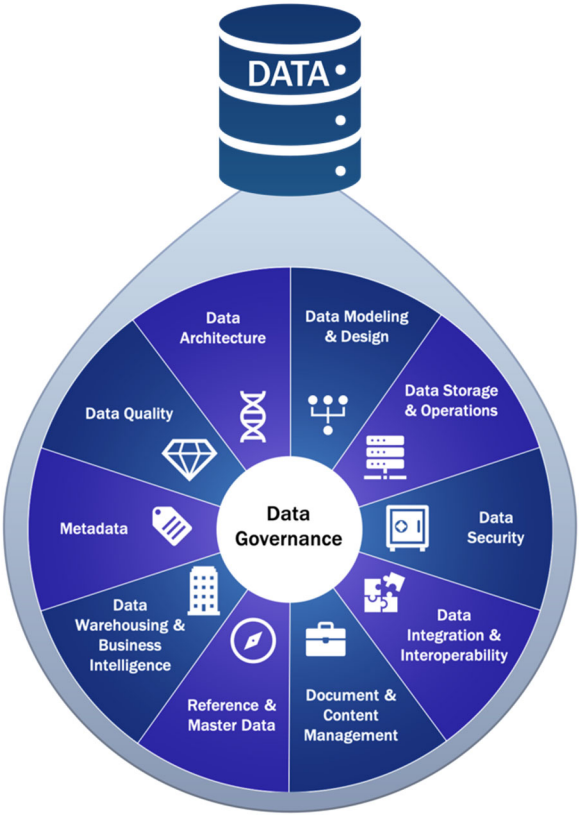


FIGURE 2: DATA MANAGEMENT KNOWLEDGE AREAS

Who Is Involved in Data Management?

Everyone who touches data in an organization is involved in data management. This includes:

- Any person who works in any facet of managing the organization's data and information
- All levels of business management leadership (executive, directors, chief data officers, etc.), data stewards, and data strategists
- All levels of technical management leadership (such as information resources managers, chief information officers, database administrators, network administrators, and programmers)

Key Takeaways

- Following data management best practices is everyone's responsibility.
- Executive leadership, commitment, and support are crucial.
- Collaboration between information technology and the business and program areas is essential.
- If used effectively, data management enables organizations to gain predictive insight about their customers, products, and services.
- Implementing data management is an organizational change, and change management is critical.
- Success requires dedicated time and resources.

Guiding Principles

- **Data is an asset with unique properties.** Unlike other assets, data is not consumed when used.
- **Value of data should be expressed in economic terms.** Low-quality data is a cost to the organization. High-quality data is a benefit.
- **Managing data means managing the quality of data.** The level of data quality should be set by the requirements and expectations of stakeholders.
- **Metadata is critical in understanding data.** Metadata provides the context, description and meaning that makes data useful.
- **Planning is required to manage data.** Data is created and stored in many places. Without planning, there is chaos.

- **Data management is a cross-functional effort.** Managing data requires both technical and non-technical skills and expertise.
- **Data management requires an organizational perspective.** Data is managed best when it's considered from the perspective of the whole organization.
- **Data management must account for a range of perspectives.** Data management must constantly evolve to keep up with change.
- **Data management is life cycle management.** Data must be managed across the entirety of its existence.
- **Different types of data have different life cycle characteristics.** Data management practices must be flexible to accommodate the differing requirements.
- **Managing data includes managing the risks associated with data.** Data represents a risk to the organization. It can be lost, stolen, or misused.
- **Data management requirements must drive technology.** Technology must support, rather than drive, an organization's strategic data needs. Technology and business must partner to successfully meet data management requirements.
- **Effective data management requires leadership commitment.** Data management must have dedicated and supportive leadership vision and purpose to be effective.

DAMA-DMBOK Reference

For more information, see Chapter 1: Data Management, pages 17-48.

Data Handling Ethics

What Are Data Handling Ethics?

Ethics are principles of behavior based on ideas of right and wrong. Ethical principles often focus on ideas such as fairness, respect, responsibility, integrity, quality, reliability, transparency, and trust. Ethics are essential whenever data is handled and must be considered at each stage of the data life cycle, whenever data is **created, stored, used, shared, archived or destroyed**.

Why Are Ethics Important to Data Handling?

While data may seem as if it is merely technical information, its use must be guided by ethical principles or it will present a risk to an organization's continued success. The ethical use of data focuses on several core concepts:

- **Impact on people**. Data represents characteristics of individuals and is used to make decisions that affect people's lives; there is an imperative to manage its quality and reliability
- **Potential for misuse**. Misusing data can negatively affect people and organizations; there is an imperative to prevent the misuse of data
- **Economic value of data**. Data has an economic value; there is an imperative to determine how that value can be accessed and by whom

> *To paraphrase W. Edwards Deming, "Ethics means doing it right when no one is looking."*

How Do Data Handling Ethics Apply to You?

Handling data in an ethical way is a critical part of the duty and responsibilities of those who protect Texas data.

A core function is to manage the sensitive personal data of the public appropriately and to represent it with the highest degree of accuracy. Privacy is of the utmost importance and there is a fundamental expectation that data will be managed, handled, and secured with an appropriate level of transparency (Figure 3).

When Do You Consider Ethical Standards?

The ethical handling of data falls within the areas of both data governance and legal counsel. Together, governance and counsel ensure that employees are kept up to date on legal changes, reducing the risk of ethical impropriety by ensuring employees are aware of their obligations. Data governance programs set the standards and policies for and provide oversight of data handling practices.

FIGURE 3: DATA ETHICS

Who Is Involved in the Ethics of Data Handling?

Everyone who touches data in an organization needs to handle data ethically. This includes:

- Any person who works in any facet of managing the organization's data and information
- All levels of business management leadership (executive, directors, chief data officers, etc.), data stewards, and data strategists
- All levels of technical management leadership (such as information resources managers (IRM), chief information officers (CIO), database administrators, network administrators, and programmers)

Upon starting employment with the state of Texas, each employee must complete a training class on ethics, and then complete a refresher course every two years. These classes are not specific to the handling of data but provide both a foundation and overall guidance on ethical situations.

An organization's data governance program establishes specific standards and policies regarding the ethical handling of data. Other influences include federal guideline requirements for the storage, access, and sharing of specific data types such as HIPAA (Health Insurance Portability and Accountability Act), CJIS (Criminal Justice Information Services), and FERPA (Family Educational Rights and Privacy Act).

Key Takeaways

- Data management includes a duty to handle data according to ethical principles.
- In addition to the state of Texas overall ethics guidelines, an organization must incorporate additional guiding principles for the specific data types that they manage, access, and share.
- Ethics related to the handling of data are a core element of an organization's data governance program.

Guiding Principles

The European Union (EU) General Data Protection Regulation (GDPR) focuses on overall data privacy and the rights of the public. GDPR is currently being used by other states as a baseline for data privacy and ethics statutes and could potentially serve in the future as an effective guideline for data handling ethics in the state of Texas (Table 1).

TABLE 1: GDPR PRINCIPLES

GDPR principle	*Description of principle*
Fairness, lawfulness, transparency	Personal data shall be processed lawfully, fairly, and in a transparent manner in relation to the data subject.
Purpose limitation	Personal data must be collected for specified, explicit, and legitimate purposes, and not processed in a manner that is incompatible with those purposes.
Data minimization	Personal data must be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed.
Accuracy	Personal data must be accurate, and where necessary, kept up-to-date. Every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purpose for which they are processed, are erased or rectified without delay.
Storage limitation	Data must be kept in a form that permits identification of data subjects [individuals] for no longer than is necessary for the purposes for which the personal data are processed.
Integrity and confidentiality	Data must be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical, or organizational measures.
Accountability	Data controllers shall be responsible for and be able to demonstrate compliance with [these

DAMA-DMBOK Reference

For more information, see Chapter 2: Data Handling Ethics, pages 49-66.

Data Governance

What Is Data Governance?

Data governance is the exercise of authority and control over the management of data assets. It guides all other data management functions. Data governance focuses how decisions about data are made, what people are expected to do with data, and the goals that processes related to data management should accomplish.

Similar to the way that auditors set rules for managing financial assets, data governance professionals set rules for managing data assets. Like with auditing, other business areas carry out these rules.

Why Is Data Governance Important?

Data governance is the foundation of an effective data management program. While the scope and focus of the data governance program will depend on the unique needs of an organization (more complex organizations may need more complex data governance structures), most programs include the following components:

- **Strategy.** Defining, communicating, and driving execution of an organization-wide methodology for managing data
- **Policy.** Setting and enforcing policies related to data management, access, usage, security, and quality
- **Standards and quality.** Setting and enforcing standards related to data quality and architecture
- **Oversight.** Providing hands-on observation, guidance, and direction in key areas of quality, policy and data management (this is often called data stewardship)
- **Compliance.** Ensuring the organization can meet regulatory compliance requirements
- **Issue/operational management.** Identifying and resolving issues related to key aspects of data management (including access, sharing, quality, regulatory compliance, ownership, policies, standards, terminology, and procedures)
- **Data management projects.** Sponsoring initiatives to improve data management practices
- **Data asset valuation.** Setting standards and creating processes that define the business value of data

How Does Data Governance Apply to You?

Data governance enables organizations to manage data as an asset. In organizations that do not have a strong data management culture, implementing a data governance program will be a cultural change. Business areas that access, share, use, or create data will need to be brought into alignment with policies and procedures created, implemented, and monitored by the data governance program. These areas may include:

- Procurement and contracts
- Budget and funding
- Regulatory compliance
- Application/system development

When Do You Use Data Governance?

Data governance is essential to effective data management. Whenever data is being managed, data governance must be followed. To ensure that governance is consistently used to guide data management activities at all levels, an effective data governance program is:

- **Sustainable.** Data governance is not a project with a defined end date; it is an ongoing process that requires organizational commitment to change. Sustainable data governance depends on business leadership, sponsorship, and ownership.
- **Embedded.** Data governance is not an add-on process. Data governance activities need to be incorporated into the development (or purchase) of software, the use of data for analytics, the management of data, and risk management.

- **Measured.** Data governance done well results in efficient operations and cost savings. To demonstrate this value, data governance activities and results must be measured and tracked.

Who Is Involved in Data Governance?

Depending on the organization, data governance can include legislative-like functions (defining policies, standards, rules), judicial-like functions (issue management and escalation), and executive functions (administrative responsibilities). Most organizations adopt a representative form of data governance, so that all stakeholders can be heard.

To determine who should be involved in data governance, consider these factors:

- Each organization should adopt a data governance model that supports its business strategy and fits within the organization's cultural context.
- Whichever model is chosen, organizations should be prepared to evolve the data governance model to meet new challenges (formal versus informal or centralized versus distributed).
- Data governance organizations may have multiple layers to address different concerns. Each layer would have a special purpose and level of oversight.

Effective data governance models can give insight to activities at different levels within the organization as well as separation of governance responsibilities (Figure 4). Guidance also exists that shows the typical committees that might be established within a data governance operating framework (Table 2).

TABLE 2: TYPICAL DATA GOVERNANCE COMMITTEES AND BODIES

Data Governance Body	*Description*
Data Governance Steering Committee	The primary and highest authority organization for data governance in an organization, responsible for oversight, support, and funding of data governance activities. Consists of a cross-functional group of senior executives. Typically releases funding for data governance and data governance-sponsored activities as recommended by the DGC and CDO. This committee may in turn have oversight from higher-level funding or initiative-based steering committees.
Data Governance Council (DGC)	Manages data governance initiatives (e.g., development of policies or metrics), issues, and escalations. Consists of executives according to the operating model used.
Data Governance Office (DGO)	Ongoing focus on enterprise-level data definitions and data management standards across all DAMA-DMBOK Knowledge Areas. Consists of coordinating roles that are labelled as data stewards or custodians, and data owners.
Data Stewardship Teams	Communities of interest focused on one or more specific subject-areas or projects, collaborating or consulting with project teams on data definitions and data management standards related to the focus. Consists of business and technical data stewards and data analysts.
Local Data Governance Committee	Large organizations may have divisional or departmental data governance councils working under the auspices of an Enterprise DGC. Smaller organizations should try to avoid such complexity.

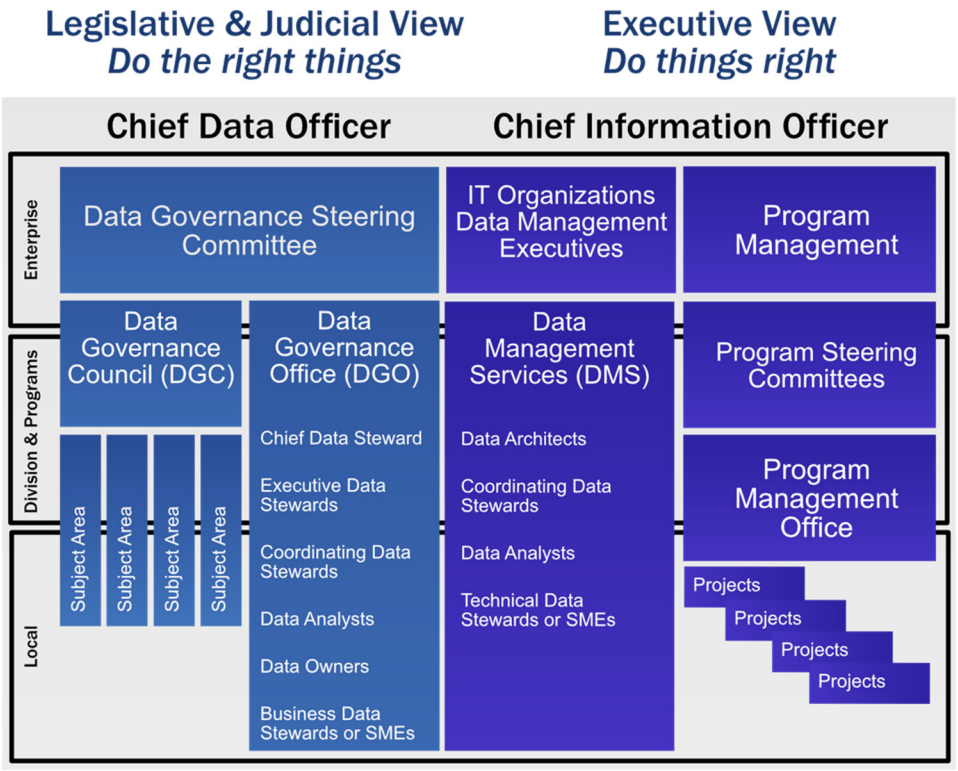


FIGURE 4: GENERIC DATA GOVERNANCE ORGANIZATION MODEL

Key Takeaways

- Data governance is the foundation of a successful data management program.
- Data governance sets the rules for managing data assets as business assets. Other areas carry out the rules.
- Implementing a data governance program will likely be a cultural change and should be treated as such.
- Expect the model of data governance to change as the organization encounters and meets new business challenges.

Guiding Principles

- Successful data governance starts with leadership setting the strategy, vision, and commitment.
- Traditionally, data governance is a business program, and, as such, must govern IT decisions related to data, as much as it governs business interaction with data. (Although note that in Texas there are examples of IT leadership successfully driving data governance.)
- Data governance is a shared responsibility between business data stewards and technical data management professionals.
- Data governance occurs on multiple levels within an organization (including enterprise and program levels).
- Data governance activities require coordination across multiple functional areas.
- For data governance to be successful, a core set of principles and best practices must be articulated as part of the overall policy.

DAMA-DMBOK Reference

For more information, see Chapter 3: Data Governance, pages 67-96.

Data Architecture

What Is Data Architecture?

Data architecture describes how data should be structured, how data should be integrated into business processes, and how data should be controlled so that it can be effectively managed by an organization. A data architecture program serves as a plan for data management in the same way that a builder's blueprint provides a plan for the construction of a building.

A well-developed data architecture is essential if data is to function as part of an organization's business processes.

The most detailed data architecture design document is a formal enterprise data model, containing data names and metadata definitions. It outlines conceptual and logical entities and their relationships and business rules. This model is the fundamental roadmap for an organization's data.

Why Is Data Architecture Important?

Data architecture enables data to be treated as a business asset. As a business asset, data can fully support the needs of the entire organization.

Well-developed data architecture acts as a bridge between business strategy and technology execution. Business drivers and data architecture work together to:

- Translate business needs into data and system requirements so that processes consistently have the data they require
- Manage complex data and information delivery throughout the organization
- Facilitate alignment between business areas and information technology
- Act as agents for change, transformation, and agility

How Does the Concept of Data Architecture Apply to You?

Data architecture interacts with other architectures in an organization (such as business, application, and technical architectures) to create efficiencies throughout an organization. Architects from these different domains must work together to determine how to develop requirements (Figure 5).

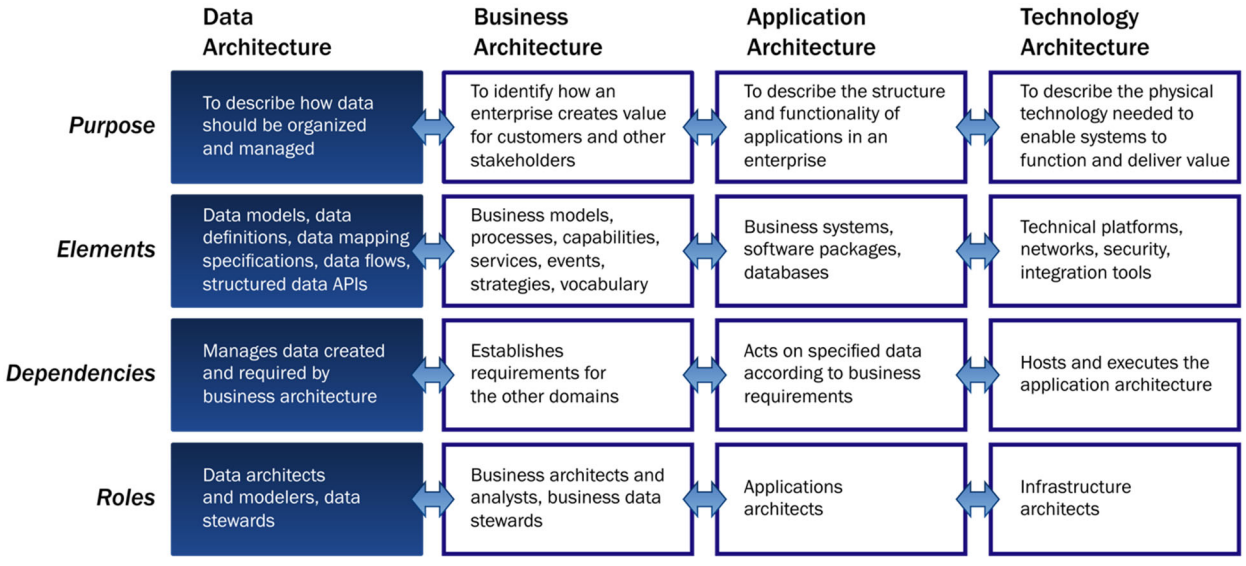


FIGURE 5: DATA ARCHITECTURE INTERACTION

When Do You Use the Concept of Data Architecture?

Data architecture is an integral part of an organization's overall enterprise architecture (as seen in Figure 5). Data architecture is often used when developing projects and system releases that depend on the data that an organization is managing. For example, data architecture can have influence when you are:

- **Defining project data requirements.** Helps guide data requirements for individual projects
- **Reviewing project data designs.** Ensures conceptual, logical, and physical data models are consistent with data architecture and long-term organizational strategy
- **Determining data lineage impact.** Ensures business rules and data flow are consistent and traceable
- **Data replication control.** Ensures sufficient and appropriate replication controls are in place to achieve consistency
- **Enforcing data architecture standards.** Provides for standards in the form of principles, procedures, guidelines and blueprints. Includes expectations regarding compliance
- **Guide data technology and renewal decisions.** Helps manage the versions, patches, and policies each application uses

Who Is Involved in Data Architecture?

Different roles have different relationships with data architecture and provide different perspectives. The perspectives given by roles include:

- **The executive perspective (business context).** Business elements defining scope in identification models
- **The business management perspective (business concepts).** Clarifies relationships between business concepts as defined by the executives in definition models
- **The architect perspective (business logic).** System logical models detailing system requirements in representation models
- **The technician perspective (component assemblies).** Technology-specific view of how components are assembled and operate in configuration models
- **The user perspective (operational and function).** Actual functioning instances used by the users. No models are associated at this level

Key Takeaways

- Data architecture describes how data should be structured, how data should be integrated into business processes, and how data should be controlled so that it can be effectively managed by an organization.
- A well-developed data architecture is essential if data is to serve as an asset to an organization's business processes.
- Data architecture does not stand alone, but rather exists in relationship to other organizational architectures such as business, application, and technology. Each of these is dependent on the others.

DAMA-DMBOK Reference

For more information, see Chapter 4: Data Architecture, pages 97-120.

Data Modeling and Design

What Is Data Modeling and Design?

Data modeling is a critical component of data management—it is the process of discovering, analyzing, and scoping data requirements. The data requirements are then both represented by and communicated through a precise form called a data model. The modeling process, as depicted in an entity relationship diagram (ERD), requires that organizations discover and document how their data fits together (Figure 6).

Data models enable an organization to understand its data assets by tracing the flow of data through a database. Data flow is represented by schemes. Models of these schemes exist at three levels of detail: conceptual (the least complex), logical, and physical (the most complex). Each model contains a set of basic components that includes entities (tables), attributes (columns), and relationships (keys).

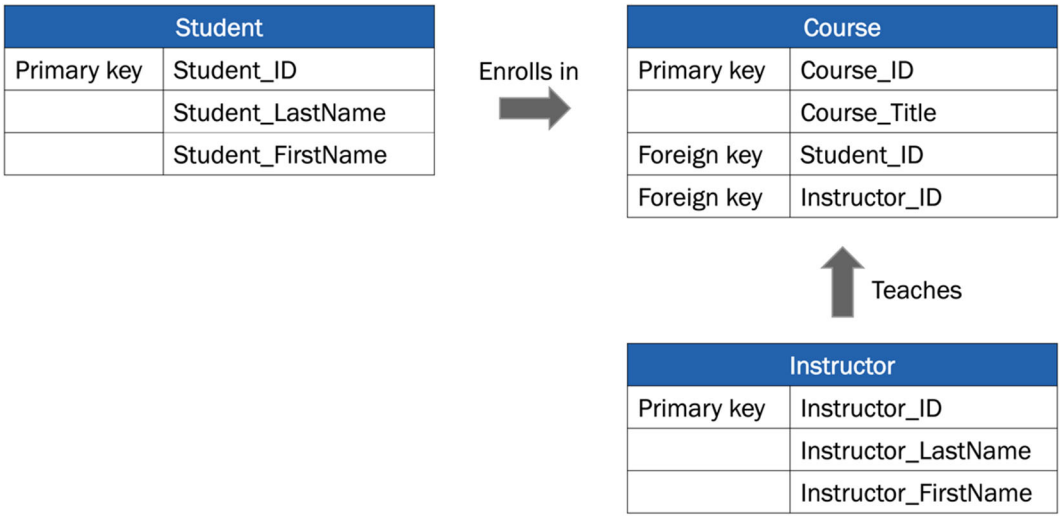


FIGURE 6: SAMPLE ENTITY RELATIONSHIP DIAGRAM (ERD)

Why Is Data Modeling and Design Important?

Data modeling confirms and documents how data is stored in a database. Proper data modeling leads to lower support costs and increases the ability of an organization to reuse data for future initiatives, which can reduce the costs of building new applications.

Data models are a form of metadata. They facilitate:

- **Formalization.** A data model documents a concise definition of data structures and relationships.
- **Scope definition.** A data model can help explain the boundaries for data context and implementation of purchased application packages, projects, initiatives, or existing systems.
- **Knowledge retention/documentation.** A data model can preserve corporate memory regarding a system or project by capturing knowledge in an explicit form. The data model becomes a reusable map to help stakeholders understand data structure within the environment.

How Does Data Modeling and Design Apply to You?

Data models are critical to effective management of data as they:

- Provide a common vocabulary around data used by various stakeholders in an organization
- Capture and document explicit knowledge about an organization's data and systems
- Serve as a primary communications tool during projects
- Provide the starting point for customization, integration, or even replacement of an application

When Do You Use Data Modeling and Design?

Data modeling is most frequently performed in the context of systems development and maintenance efforts, known as the system development life cycle (SDLC). Data modeling can also be performed for broad-scoped initiatives (such as business and data architecture, master data management, and data governance initiatives) where the immediate end result is not a database but an understanding of organizational data.

Who Is Involved in Data Modeling and Design?

Database modeling involves a variety of stakeholders within an organization, including those who supply information about business and technical requirements, participants who create data models, and those who use data models to better understand how data is stored.

- **Suppliers.** Business professionals, business analysts, data architects, database administrators and developers, subject matter experts, data stewards, metadata administrators
- **Participants.** Business and systems analysts, data modelers
- **Consumers.** Business analysts, data modelers, database administrators and developers, software developers, data stewards, data quality analysts

In addition, gathering requirements from business stakeholders and subject matter experts is critical. Understanding the information that stakeholders expect to retrieve from an information system, how they hope to access data, and how they plan to perform queries is critical when deciding what data to include and how data will flow through the system.

Key Takeaways

- Data modeling is the process of discovering, analyzing, and scoping data requirements.
- Proper data modeling leads to lower support costs and increases the ability of an organization to reuse data for future initiatives, which can reduce the costs of building new applications.
- For data modeling to be as effective as possible, business stakeholders should be consulted when determining how data will be retrieved from an information system, how data will be accessed, and how queries will be performed.

Guiding Principles

To assess a data model's quality, the following standards are useful as guiding principles:

- How well does the model capture the requirements?
- How complete is the model in terms of including requirements and thoroughness of metadata?
- How well does the model match its scheme?
- How structurally sound is the model as a basis for building a database?
- How well does the model follow naming standards in terms of structure, term, and style?
- How good are the definitions? Are they clear, complete, and accurate?
- How consistent is the model with other enterprise data documentation?
- How well does the metadata match the data?

DAMA-DMBOK Reference

For more information, see Chapter 5: Data Modeling and Design, pages 121-164.

Data Storage and Operations

What Are Data Storage and Operations?

Data storage and operations include the design, implementation, and support of stored data to maximize its value throughout its life cycle from creation or acquisition to disposal (Figure 7).

Why Are Data Storage and Operations Activities Important?

Organizations rely on their information systems to run their operations. Data storage and operations activities are crucial to organizations that rely on data. Business continuity is a critical driver of these activities. If a system becomes unavailable operations may be impaired or stopped completely. A reliable data storage infrastructure for IT operations provides business with access and availability to data.

The goals of data storage and operations include:

- Managing the availability of data throughout the data life cycle
- Ensuring the integrity of data assets
- Managing the performance of data transactions

FIGURE 7: DATA LIFE CYCLE

How Do Data Storage and Operations Activities Apply to You?

The long-term value of secure, reusable, and high-quality data is not always easily recognized or appreciated. A data storage and operations program maintains and assures the accuracy and consistency of data throughout its entire life cycle, and addresses the design, implementation, and usage of any system that stores, processes, or retrieves data.

Essential data operations management activities include:

- Ensuring the performance and reliability of the database through performance tuning, monitoring, error reporting, and other activities
- Implementing backup and recovery mechanisms to ensure data can be recovered if lost in any circumstance
- Implementing mechanisms for clustering and failover of the database (if continual data availability is a requirement)

When Do You Need to Perform Data Storage and Operations Activities?

The two main activities in data storage and operations are database technology support and database operations support.

- **Database technology support.** Specific to the data and processes that the software manages. Includes defining technical requirements that will meet organizational needs, defining technical architecture, installing and administering technology, and resolving issues related to technology.
- **Database operations support.** Specific to selecting and maintaining the software that stores and manages the data. Focuses on activities related to the data life cycle, from initial implementation of a database environment, through obtaining, backing up, and purging data. It also includes ensuring the database performs well.

Data storage and operations are used when:

- **Configuring the database environment**, including ensuring that instances of the database management system on the supporting server are of a sufficient size and capacity to ensure adequate performance, as well as ensuring the appropriate level of security, reliability, and availability.
- Creating mechanisms and processes for **controlled implementation of changes** to databases.
- Implementing mechanisms for **ensuring the availability, integrity, and recoverability of data** in response to all circumstances that could result in loss or corruption of data.
- Developing mechanisms for **detecting and reporting any error** that occurs in the database or the data server.
- **Monitoring database performance** as workloads and data volumes change.

Who Is Involved in Data Storage and Operations?

Data storage and operations involve working with a variety of stakeholders within an organization. Stakeholders include those who supply data to be stored or ask that it be managed in a particular way, those who manage the data in a database management system, and those who use data from those systems.

- **Suppliers.** Data architect, data modeler, software developer, application testing team
- **Participants.** Database administrator, data architect
- **Consumers.** Data modeler, software developer, application testing team, infrastructure operations

Database administrators (DBA) play the dominant role in data storage and operations and take primary responsibility for data operations management. The DBA is the custodian of all database changes. While many parties may request changes, the DBA defines the precise changes to make to the database, implements the changes, and controls the changes.

Key Takeaways

- Data storage and operations include the design, implementation, and support of stored data to maximize its value throughout its life cycle from creation or acquisition to disposal.
- Data storage and operations activities are crucial to organizations. If a system becomes unavailable, company operations may be impaired or stopped completely.
- A data storage and operations program maintains and assures the accuracy and consistency of data throughout its entire life cycle.

Guiding Principles

- **Identify and act on automation opportunities.** Automate database development processes, which will shorten each development cycle, reduce errors and rework, and minimize impact on the development team.
- **Build with reuse in mind.** Develop and promote the use of abstracted and reusable data objects, which will prevent applications from being tightly coupled to database schemas.
- **Connect database standards to support requirements.** For example, the Service Level Agreement (SLA) can reflect DBA-recommended and developer-accepted methods of ensuring data integrity and data security.
- **Set expectations for the DBA role in project work.** Have a dedicated primary and secondary DBA during analysis and design; and clarify expectations about DBA tasks, standards, work effort, and timeliness.

DAMA-DMBOK Reference

For more information, see Chapter 6: Data Storage and Operations, pages 165-208.

Data Security

What Is Data Security?

Data security is the planning, development, and execution of security policies and procedures to provide the proper protection of and access to data. It includes understanding and classifying data, controlling access to data, and auditing of data. Data security practices function in alignment with privacy and confidentiality regulations, contractual agreements, and business requirements to protect an organization's assets.

Overall requirements for data security come from stakeholders, government regulations, proprietary business concerns, legitimate access needs, and contractual obligations (Figure 8).

Why Is Data Security Important?

Effective data security is essential to reducing the risk of organizational losses and retaining the public's trust. The impact of security breaches and other incidents on well-established organizations in recent years has demonstrated the potential of major financial loss and a significant loss of customer confidence.

How Does Data Security Apply to You?

Ensuring that data remains secure, like other aspects of data management, requires the participation of all of those who touch data. In particular, security professionals are often tasked with managing IT compliance requirements, polices, practices, data classifications, and access authorization rules across the organization.

Data security begins by inventorying and classifying an organization's data to determine the applicable security controls and protections for the data. The overall process includes the following general steps:

- Identify and classify sensitive data assets, e.g. personally identifiable information (PII), protected health information (PHI), etc.
- Locate sensitive data throughout the organization and information systems.
- Determine how each asset needs to be protected.
- Identify how this information interacts with business processes.
- Apply appropriate security controls and routinely monitor the protections to ensure they are compliant.

FIGURE 8: DATA SECURITY REQUIREMENTS

In Texas, data classification follows standards that have been set by the legislature and managed through the Texas Administrative Code (TAC), https://www.sos.texas.gov/tac/index.shtml.

Specific references include:

- Title 1, Part 10, Chapter 202, Information Security Standards
- Title 13, Part 1, Chapter 6, Subchapter C, Standards and Procedures for Management of Electronic Records, §6.93 and §6.94.

When Do You Use Data Security?

Data security is associated with regulatory compliance, fiduciary responsibility for the organization and stakeholders, and an organization's reputation.

Maintaining proper data security is a legal and moral responsibility, and protects the private and sensitive information of employees, customers, business partners, and the public.

Who Is Involved in Data Security?

Data security should be an organization-wide effort. If business areas find individual solutions to security needs, there will be a danger of increasing overall cost while also increasing the risk of a security breach (due to inconsistent protection processes).

An operational security strategy that is properly executed, systems-orientated, and consistent across the organization can reduce these risks.

In Texas, Information Security Officers (ISO) have been trained on the best practices and regulations associated with data security. Data professionals should contact their Information Security Officer to ensure alignment between organizational data protection strategies and their approach to data management.

For example, Information Security Officers note that information security is composed of three key security attributes: confidentiality (the right individuals access the right data), integrity (the data is reliable and protected from unauthorized alteration, deletion, or modification) and availability (the data is accessible to the right people at the right times). These three attributes are often referred to as the "CIA Triad."

ISOs strongly recommend that, when implementing a security program, organizations strive for a defense-in-depth approach, using people, processes, and technology to protect data and assets and to manage risk.

Key Takeaways

- Data security is the planning, development, and execution of security policies, procedures, and controls to provide the proper protections of data.
- While data management professionals have a significant role to play in ensuring data remains secure, effective data security requires the participation of all of those who touch data.
- In Texas, data classification (required for effective data security) follows specific regulations, described in the Texas Administrative Code, TAC 202 and TAC 6.
- Data security will be most effective if it is executed consistently across all business areas.

Guiding Principles

- **Collaboration.** Data security is a collaborative effort involving IT security, data stewards, data governance, internal/external auditors, and legal counsel.
- **Enterprise approach.** Data security standards and policies must be applied consistently across the entire organization to realize the greatest benefits.
- **Proactive management.** Success in data security depends on being proactive and dynamic, engaging all stakeholders, managing change, and overcoming organizational or cultural barriers.
- **Clear accountability.** Roles and responsibilities must be clearly defined, including the "chain of custody" for data across organizational departments and roles. Auditing and accountability measures should be applied based on the needs and sensitivity of the data.
- **Metadata-driven.** Security classifications for data elements is an essential part of data definitions.
- **Reduce risk by reducing exposure.** Minimize sensitive and confidential data proliferation, especially to non-production environments by implementing the use of least privilege and routine data access reviews.

DAMA-DMBOK Reference

For more information, see Data Security, Chapter 7, pages 209-255.

Additional References

Data classification guide:
https://pubext.dir.texas.gov/portal/internal/resources/DocumentLibrary/Data%20Classification%20Guide%20v1.1.docx

Data classification template:
https://pubext.dir.texas.gov/portal/internal/resources/DocumentLibrary/Data%20Classification%20Template.xls

Data Integration and Interoperability

What Are Data Integration and Interoperability?

Data integration and interoperability (DII) ensures that data is located where it is needed, available when it is needed, and in the form in which it is needed. Data integration is the process of transforming data into consistent forms and migrating the transformed data from source to target systems.

Data interoperability ensures that data is uniform and system independent so that it can be shared, migrated, and consumed by disparate systems. Data integration and interoperability describes the processes related to the movement and consolidation of data within and between data stores, applications, and organizations (Figure 9).

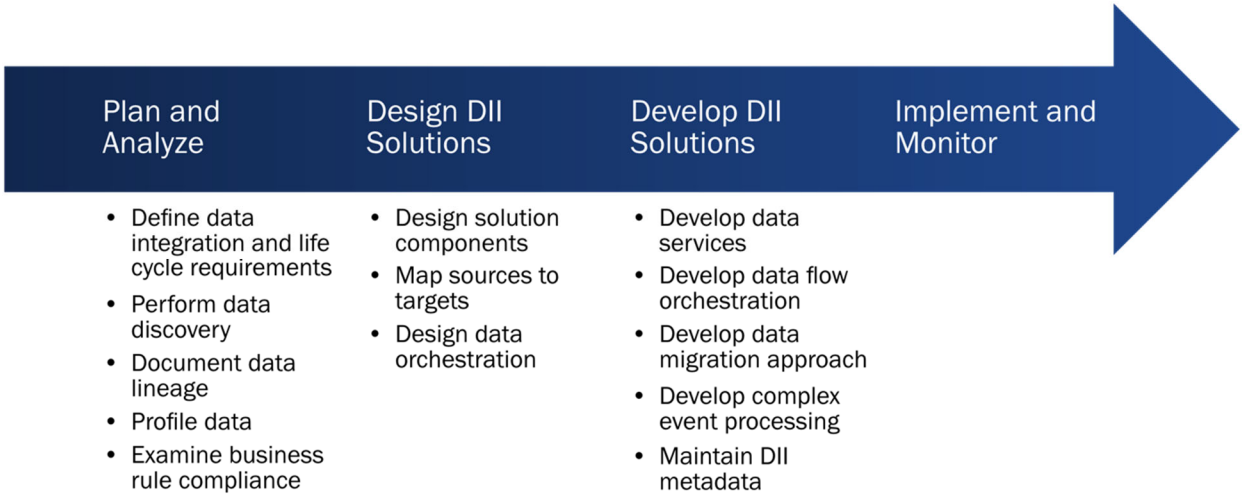


FIGURE 9: DATA INTEGRATION AND INTEROPERABILITY PROCESS

Why Are Data Integration and Interoperability Important?

A primary driver for data integration and interoperability is the need to manage data movement efficiently. As the amount and different types of data collected by organizations increases, so does the number of disparate data storage systems. If data is not integrated properly, the process of moving data between numerous legacy and vendor-specific systems can overwhelm IT resources, drain budgets, and strain support services.

Transforming and integrating data between multiple systems is an essential responsibility of every information technology organization.

How Do Data Integration and Interoperability Apply to You?

Data integration and interoperability are critical to several aspects of data management, especially data warehousing and business intelligence. Data warehousing and business intelligence focus on transforming and integrating data from source systems to consolidated data hubs and from data hubs to target systems, where it can be delivered to data consumers (both system and human).

Data integration and interoperability is also central to the emerging area of big data management, which seeks to integrate various types of data. Once integrated, data can be mined, used to develop predictive models, and deployed in operational intelligence activities.

When Are Data Integration and Interoperability Needed?

The need for data integration and interoperability arises whenever organizations migrate data from one system to another. As systems become obsolete and are no longer supported, data must be pulled into newer data storage systems (or risk the loss of vital information).

Data integration and interoperability are also needed when data is shared across an entire organization. When organizations purchase applications from software vendors rather than develop custom applications, each application comes with its own set of data stores. As the number of different data stores increase, an enterprise model of data integration, such as a data warehouse, is more efficient and cost effective than point-to-point data integration solutions between individual applications (Figure 10).

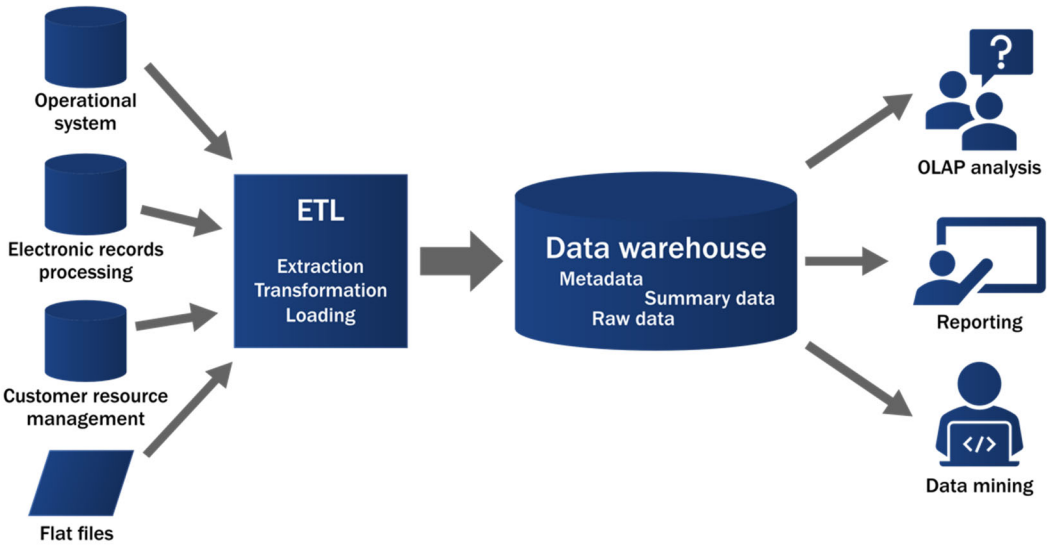


FIGURE 10: DATA WAREHOUSE ARCHITECTURE

Who Is Involved in Data Integration and Interoperability?

Stakeholders include those who supply data from source systems, those who participate in migrating and consolidating data into target systems, and those who consume data for business intelligence and when making decisions.

- **Suppliers.** Data producers, subject matter experts
- **Participants.** Data architects, business and data analysts, data modelers, data stewards, ETL (extract, transform, and load) developers, project and program managers
- **Consumers.** Information consumers, knowledge workers, managers and executives

Data integration solutions are frequently perceived as purely technical; however, to successfully deliver value, they must be developed based on deep business knowledge. Involving those who understand business goals and requirements is key to program success.

Key Takeaways

- Data stores can be present in multiple forms and in multiple places (such as in legacy systems and from software vendors). Data integration and interoperability (DII) ensures that data is located where it is needed, available when it is needed, and in the form in which it is needed.
- If data is not integrated properly, the process of moving data between numerous systems can overwhelm IT resources, drain budgets, and strain support services.
- To successfully deliver value, the development of data integration solutions must be based on deep business knowledge. Involving those who understand business goals and requirements is key to program success.

Guiding Principles

- Make data available in the format and timeframe needed by data consumers.
- Consolidate data physically and virtually into data hubs.
- Lower cost and complexity of managing solutions by developing shared data integration models.
- Support business intelligence, analytics, master data management, and operational efficiency efforts.
- Take an enterprise perspective in design to ensure future extensibility.
- Ensure business accountability for data integration and interoperability design and activity.

DAMA-DMBOK Reference

For more information, see Chapter 8: Data Integration and Interoperability, pages 257-286.

Document and Content Management

What Is Document and Content Management?

Document and content management focuses on controlling the capture, storage, access, and use of data and information stored outside relational databases. Document and content management maintains the integrity of and enables access to documents and other unstructured or semi-structured information.

In many organizations, unstructured data has a direct relationship to the structured data in databases. When that is the case, management decisions about unstructured content should be consistent with decisions made about structured data.

Documents and unstructured content are expected to be secure and of high quality, which requires governance, reliable architecture, and well-managed metadata.

Why Is Document and Content Management Important?

Both business and technical concerns drive the need for effective document and content management. Primary business drivers include:

- **Laws and regulations** that require organizations to maintain records of certain kinds of activities. Organizational policies, standards, and best practices for record keeping help ensure that records (which include paper documents and electronically stored information, or ESI) are properly retained.
- **Litigation and e-discovery,** which often requires that documents be provided.
- **Business continuity**, which requires planning, practice and good records management to reestablish business operations.

A document management program can also improve technical efficiency. Technological advances in document management can help organizations streamline processes, manage workflow, eliminate repetitive manual tasks, and enable collaboration. These technologies make it possible for people to locate, access, and share documents quickly, and also help prevent documents from being lost.

How Does Document and Content Management Apply to You?

Document and content management is essential during e-discovery (the process of finding electronic records that might serve as evidence in a legal action). In Texas, e-discovery is commonly known as a public information request (PIR). Many Texas state agencies have a public information office that coordinates PIR tracking, reporting, and overall response.

The ability of an organization to respond to an e-discovery request or PIR depends on how proactively it has managed records such as email, chats, websites, and other electronic documents, as well as raw application data and metadata. Big data has become a driver for more efficient e-discovery, records retention, and strong information governance.

When Do You Use Document and Content Management?

Document management is used throughout the document life cycle (in general, document management concerns files, with little attention to file content. Figure 11).

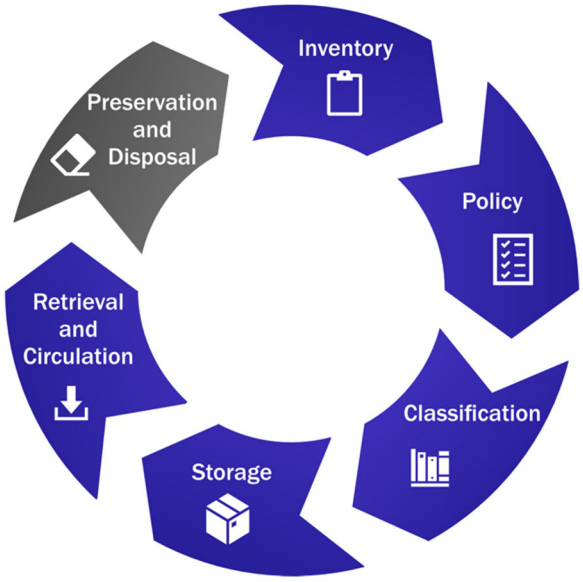


FIGURE 11: DOCUMENT LIFE CYCLE

The life cycle of documents and records includes:

- **Inventory.** Identification of existing and newly created documents and records
- **Policy.** Creation, approval, and enforcement of document and record policies, including a retention policy
- **Classification.** Categorizing documents and records according to a classification scheme
- **Storage.** Short- and long-term storage of physical and electronic documents and records
- **Retrieval and circulation.** Allowing access to and circulation of documents and records in accordance with policies, security and control standards, and legal requirements
- **Preservation and disposal.** Archiving and destroying documents and records according to organizational needs, statutes, and regulations

Who Is Involved in Document and Content Management?

Document and content management involves working with a variety of stakeholders, including those who create and supply documents and records, participants who manage and control access to documents and records, and those who need access to documents and records.

- **Suppliers.** Legal team, business team, IT team, external parties
- **Participants.** Data steward, data management professional, records management staff, content management staff, web development staff, librarians
- **Consumers.** Business user, IT user, government regulatory organization, audit team, external customers

Key Takeaways

- Document and content management focuses on controlling the capture, storage, access, and use of data and information stored outside relational databases.
- Document and content management is essential during e-discovery (in Texas state organizations, commonly known as public information requests).
- Document management is used throughout the document life cycle, which includes inventory, policy, classification, storage, retrieval and circulation, and preservation and disposal.

Guiding Principles

The primary goals of document and content management are to:

- Improve access to information
- Control the growth of materials taking up physical space
- Reduce operating costs
- Minimize litigation risks
- Safeguard vital information
- Support better decision-making

Best practices around document and content management include:

- Ensuring effective and efficient retrieval and use of data and information in unstructured formats
- Ensuring integration capabilities between structured and unstructured data
- Complying with legal obligations and customer expectations

DAMA-DMBOK Reference

For more information, see Chapter 9: Document and Content Management, pages 287-326.

Reference and Master Data

What Are Reference Data and Master Data?

Reference data is data that is used to characterize and define other data (a simple example of reference data is a list that identifies which three-digit code represent which Texas county, Figure 12). The goal of reference data management (RDM) is to ensure an organization has access to a complete set of accurate and current values for each concept represented in the reference data (compare Figure 13 and Figure 14).

Code	Description
001	Anderson
002	Andrews
003	Angelina
004	Archer
006	Armstrong

FIGURE 12: EXAMPLE OF SIMPLE REFERENCE DATA

Master data is data about business entities (real-world objects, like customers, products, employees, or vendors) that provides the context for business transactions and analysis. Master data should represent the authoritative, most accurate data about key business entities. The goal of master data management (MDM) is to ensure that master data values and identifiers are consistent across systems and represent the most accurate and timely data about essential business entities.

ID	Name	Address	Telephone
123	John Smith	123 Main, Dataland, SQ 98765	
234	J. Smith	123 Main, Dataland, DA	2345678900
345	Jane Smith	123 Main, Dataland, DA	234-567-8900

FIGURE 13: DATA REPRESENTING THE SAME ENTITY CAN LOOK DIFFERENT

ID	Name	Address	Telephone	Candidate ID	Party ID
123	John Smith	123 Main, Dataland, SQ 98765		XYZ	1
234	**J. Smith**	**123 Main, Dataland, SQ 98765**	**+1 234-567-8900**	**XYZ, ABC**	**2**
345	**Jane Smith**	**123 Main, Dataland, SQ 98765**	**+1 234-567-8900**	**ABC**	**2**

FIGURE 14: WITH MANAGEMENT, IT BECOMES APPARENT THAT ID 234 AND ID 345 REPRESENT THE SAME ENTITY

Why Are Reference Data and Master Data Important?

Reference data and master data differ in important ways—reference data doesn't change as often as master data; nor does reference data require the matching, merging, and linking that master data does.

They also share some characteristics—for example, they provide context and meaning for other data and need to be managed at the organizational level. They enable data to be meaningfully understood.

Without reference data management and master data management, the same data might be represented in different ways. Not having a consistent understanding and agreement for both reference data and master data will lead to inefficiency, which will lead to inconsistency. Inconsistency leads to ambiguity, and ambiguity introduces risk to an organization.

How Does Reference Data and Master Data Apply to You?

The most basic reference data consists of codes and descriptions, but some reference data can be more complex and incorporated into mappings and hierarchies. Reference data exists in virtually every data store. Some of these stores include code tables in relational databases, RDM systems that maintain business entities values, or object-attribute-specific metadata that defines permissible values. As such, those who use data will use some form of reference data.

Master data provides the organization the most authoritative, accurate data about key business entities. Business rules typically dictate the format and allowable ranges of master data values. Those who create or use data will need to ensure business rules that govern master data are followed to guarantee that data becomes and remains useful in achieving business goals.

When Do You Use Reference Data and Master Data?

Both reference data and master data are important principles of an overall successful data management program. They are linked and integrated into the other principles of data management, such as data governance, data quality, metadata management, and data integration.

Who Uses Reference Data and Master Data?

Many of the same people use both (Figure 15).

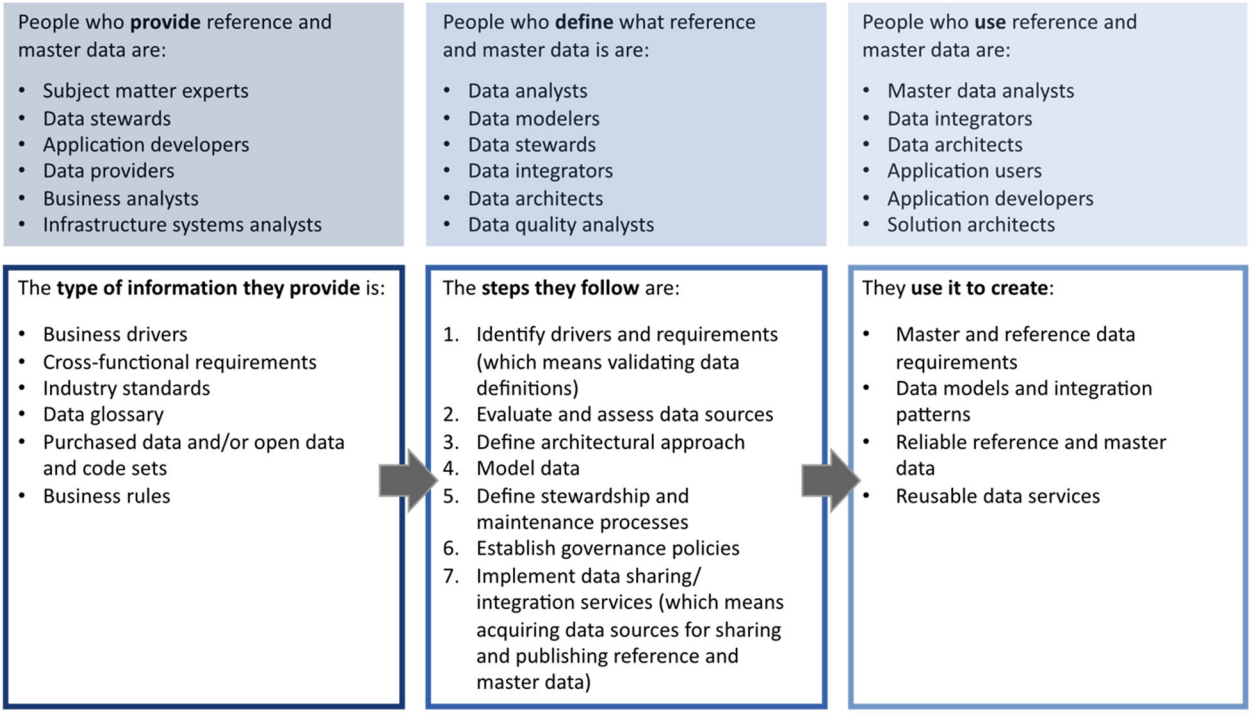


FIGURE 15: PEOPLE INVOLVED IN REFERENCE DATA AND MASTER DATA

Key Takeaways

- Reference data is data that is used to characterize and define other data.
- Master data is data about business entities that provides the context for business transactions and analysis.
- Both reference data and master data must be managed at an organizational level to ensure data is accurate and timely throughout the organization.
- If reference data or master data is not used consistently, different areas can define the same data in different ways, leading to inconsistency, ambiguity, and risk.

Guiding Principles

- **Shared data.** Reference data and master data must be managed so that they are shareable across the organization.
- **Ownership.** Reference data and master data belong to the organization, not to a particular application or department. Because they are widely shared, they require a high degree of stewardship.
- **Quality.** Reference and master data management require ongoing governance and monitoring for data quality.
- **Stewardship.** Business data stewards are accountable for controlling and ensuring the quality of reference data.
- **Controlled change.** At any given point of time, master data values should represent the organization's best understanding of what is accurate and current. Changes to reference data values should follow a defined approval process and should be well communicated prior to implementation.
- **Authority.** Master data values should be replicated only from the system of record. A system of reference may be required to enable sharing of master data across an organization.

DAMA-DMBOK Reference

For more information, see Chapter 10: Reference and Master Data, pages 327-358.

Data Warehousing and Business Intelligence

What Are Data Warehousing and Business Intelligence?

Data warehouses integrate and store data from a range of sources into a common data model. The process of extracting cleansing, transforming, controlling, and loading data into the warehouse is called data warehousing.

Business intelligence is a type of data analysis that is performed on the data stored in data warehouses. The goal of business intelligence is to improve an organization's success (business intelligence also refers to the tools that support this analysis).

Why Are Data Warehousing and Business Intelligence Important?

Organizations build data warehouses because they need to make reliable, integrated data available to a variety of authorized stakeholders (Figure 16). Stakeholders will often best understand data when it is organized by business process areas and structured as data marts (copies of sections of data from the warehouse). Data marts reduce the chance of confusion and erroneous interpretation of the data by aligning the data content with the analyst's mental model.

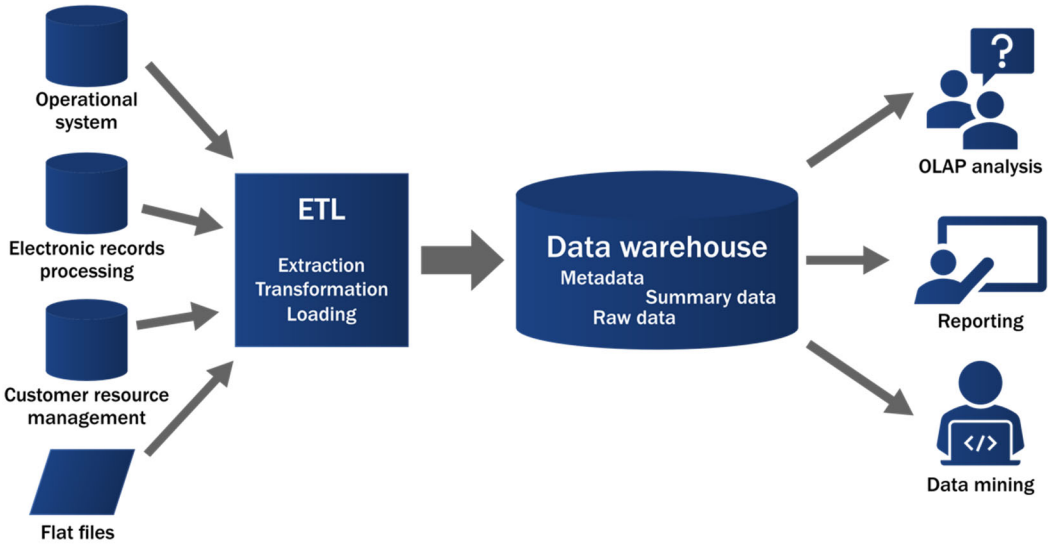


FIGURE 16: DATA WAREHOUSE ARCHITECTURE

Supporting business intelligence is the primary reason for a data warehouse, and business intelligence depends on extracting information from data warehouses through analytical processing. Any type of data warehouse is an OLAP, or online analytical processing system; an OLAP contains historical data used to follow trends and support business intelligence. An OLAP is often compared to an OLTP, or online transaction processing system, which holds the operational data used to record transactions and lacks the historical data present in a data warehouse (Figure 17).

How Should Data Warehouses and Business Intelligence Be Organized?

OLTP models are very good at containing a set of data related to a particular subject area, such as an accounting or human resources database.

When analysts need to draw on data from disparate data sources, however, an OLAP (or data warehouse) model provides the ability to retrieve data from multiple data sources in a way that enables the data to be understood and used.

Data warehouses can be organized in several ways. One effective design is the dimensional data warehouse model. Often referred to as a star schema, a dimensional data warehouse model is composed of "facts," which contain quantitative data about business processes, such as sales numbers; and "dimensions," which store descriptive attributes related to fact data. (A fact table joins with many dimension tables, and when viewed as a diagram, appears as a star. Figure 18.)

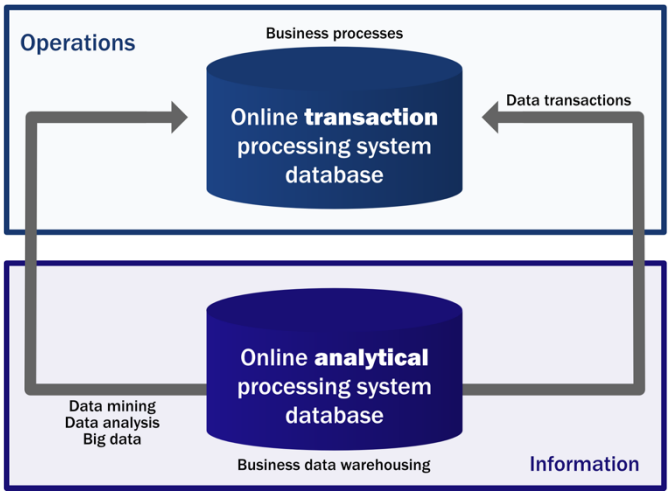


FIGURE 17: OLTP COMPARED WITH OLAP

Analysts and other data consumers can query the data warehouse to answer questions about the facts, for example questions like "how many units of product X were sold this quarter from region Y?"

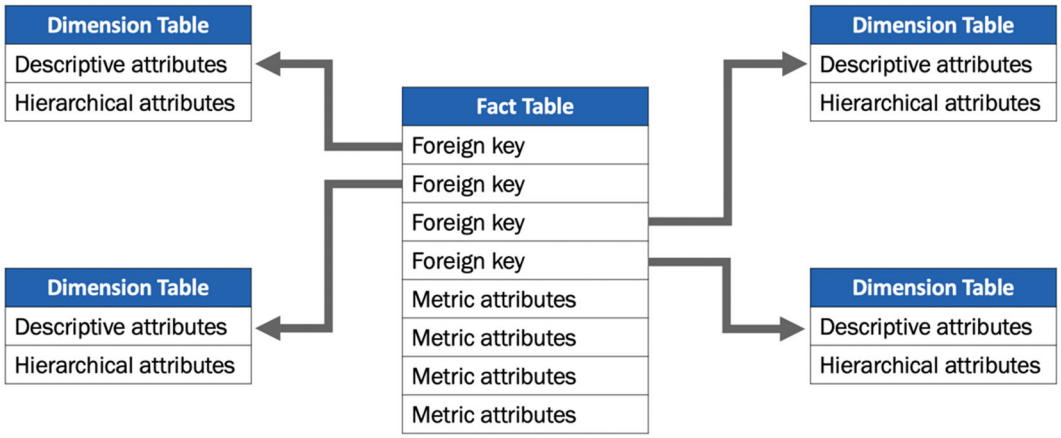


FIGURE 18: EXAMPLE OF A STAR SCHEMA

Common data warehouse operations employed by analysts include:

- **Slice.** A slice is a subset of data corresponding to a single value for one or more attributes in a dimension.
- **Dice.** A dice is a slice on more than two dimensions in a cube, or more than two consecutive slices.
- **Drill down/up.** Drilling down or up is where people navigate among levels of data, ranging from the most detailed (down) to the most summarized (up).
- **Roll-up.** A roll-up involves computing all of the data relationships for one or more dimensions.

When Do You Use Data Warehousing? When Do You Use Business Intelligence?

Organizations should build data warehouses when data from a variety of business areas need to be integrated into a single repository that is accessible to stakeholders.

Analysts often need to repeatedly traverse known hierarchies to analyze data, such as by date or product. Enabling people to examine data themselves allows them to select and analyze data in the way they understand it best.

By implementing data warehouses, organizations can unleash the power of their data by making it more readily available to a wide variety of data consumers. In doing so, organizations improve their ability to deliver more insights with greater business intelligence.

Who Is Involved in Data Warehousing and Business Intelligence?

Those involved in building data warehouses and business intelligence activities include:

- **Suppliers.** Data producers and subject matter experts
- **Participants.** Architects and analysts, data warehouse specialists in data storage and information management, project managers, and change management specialists
- **Consumers.** Analysts, customers, managers, and executives

While building a data warehouse requires technical expertise, it is essential that business stakeholders are included in all stages of a data warehouse project. The purpose of a data warehouse is to support business intelligence, and only those who are business subject matter experts can best determine which data aligns with business goals and how business analysts might be expected to interact with that data.

Key Takeaways

- Data warehouses integrate and store data from a range of sources into a common data model. Organizations build data warehouses because they need to make reliable, integrated data available to a variety of authorized stakeholders.
- Business intelligence depends on extracting information from data warehouses through analytical processing.
- The more that a data warehouse can be aligned with the needs of business intelligence, the more that business analysts will be able to use data to improve business operations.
- Business stakeholders must be included in all stages of a data warehouse project, because they are the ones who know what data needs to be included for effective business intelligence.

Guiding Principles

- **Focus on business goals.** Make sure the data warehouse serves organizational priorities and solves business problems.
- **Start with the end in mind.** Let the business priority and scope of end-data-delivery in the business intelligence space drive the creation of the data warehouse content.
- **Think and design globally; act and build locally.** Let end-vision guide the architecture, but build and deliver incrementally, through focused projects or sprints that enable more immediate return on investment.
- **Summarize and optimize last, not first.** Build on the atomic data. Aggregate and summarize to meet requirements and ensure performance, not to replace the detail.

DAMA-DMBOK Reference

For more information, see Chapter 11: Data Warehousing and Business Intelligence, pages 359-391.

Metadata Management

What Is Metadata Management?

The most common definition of metadata is “data and content about data.” While accurate, the definition is misleadingly simple. Metadata can include information about technical and business processes, data rules and constraints, and logical and physical data structures.

Like other data, metadata requires management. As the capacity of organizations to collect and store data increases, the role of metadata in data management grows in importance.

Metadata is typically categorized into three types: descriptive (describes the content of the data), structural or technical (information about the technical details and systems that store the data), and administrative or operational (details about the processing and accessing of data) (Figure 19).

Why Is Metadata Management Important?

Metadata helps an organization understand its data, its systems, and its workflows.

Across an organization, different individuals will have different levels of knowledge, but no individual will know everything about all data. Information must be documented or the organization risks losing valuable knowledge about itself.

Metadata provides the primary means of capturing and managing organizational knowledge about data. Reliable, well-managed metadata helps:

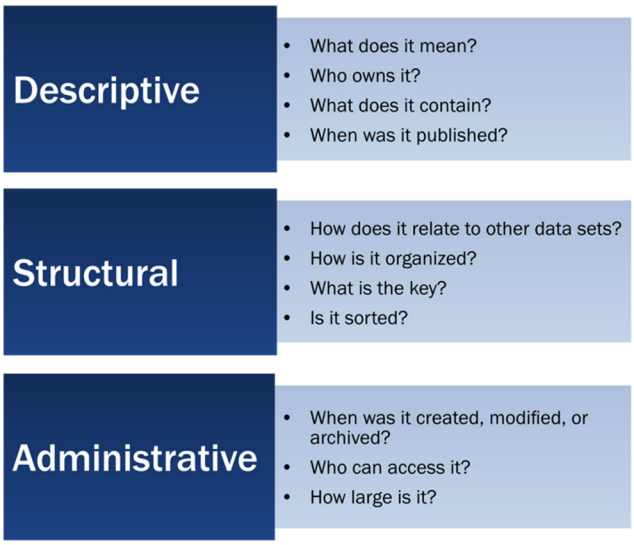


FIGURE 19: METADATA TYPES

- Increase confidence in data by providing context and enabling the measurement of data quality
- Increase the value of strategic information by enabling multiple uses
- Improve operational efficiency by identifying redundant data and processes
- Prevent the use of out-of-date and incorrect data
- Create accurate impact analysis, thus reducing the risk of project failure

How Does Metadata Management Apply to You?

Metadata enables an organization to know what data it has, what the data represents, where it originates, how it moves through systems, who has access to it, and what it means for the data to be of high quality.

Without metadata, an organization cannot manage its data as an asset. Risks include:

- Errors in judgement due to incorrect, incomplete, or invalid assumptions
- Errors due to lack of knowledge about the context of the data
- Exposure of sensitive data, which may put customers or employees at risk, reduce the credibility of the organization, or lead to legal expenses
- That the small set of subject matter experts who know the data will leave and take their knowledge with them

When Do You Use Metadata Management?

Metadata is often a by-product of data modeling and application processing rather than an end product itself. The majority of administrative metadata, for example, is generated as data is processed. While it is possible to

reverse engineer knowledge about data from existing systems and harvest metadata from existing data dictionaries, models, and process documentation, it is better to be intentional about developing definitions.

In this respect, metadata is like other data: it should be created as the product of a well-defined process, using tools that will support its overall quality.

Who Is Involved in Metadata Management?

Those involved in metadata management activities include those who supply or create metadata, participants who manage metadata, and consumers who rely on metadata to better understand and analyze data.

- **Suppliers.** Business data stewards, data managers, data governance bodies, data modelers, database administrators
- **Participants.** Data stewards, project managers, data architects, business analysts, system analysts
- **Consumers.** Application developers, data integrators, business users, knowledge workers, customers and collaborators, data scientists, data journalists

Key Takeaways

- Metadata is data and content about data. It can include information about technical and business processes, data rules and constraints, and logical and physical data structures.
- Metadata enables an organization to know what data it has, what the data represents, where it originates, how it moves through systems, who has access to it, and what it means for the data to be of high quality.
- Metadata should be created as the product of a well-defined process, using tools that will support its overall quality.

Guiding Principles

- **Organizational commitment.** Secure organizational commitment to metadata management as part of an overall strategy to manage data as an asset for the organization.
- **Strategy.** Develop a metadata strategy that accounts for how metadata will be created, maintained, integrated, and accessed. The metadata strategy must align with business priorities.
- **Organizational perspective.** Take an organizational perspective to ensure future extensibility, but implement through iterative and incremental delivery to bring value.
- **Quality.** Recognize that metadata is often produced through existing processes (such as data modeling, SDLC (system development life cycle), business process definition) and hold process owners accountable for the quality of metadata.

DAMA-DMBOK Reference

For more information, see Chapter 12: Metadata Management, pages 393-422.

Data Quality

What Is Data Quality?

Effective data management programs enable an organization to use data to achieve strategic goals. For data to be useful, it must be reliable, trustworthy and of high quality.

Data quality depends on the context and the needs of those who use it. It is high quality to the extent it effectively serves those purposes, and low quality when it doesn't. Data quality depends on the context; the same data that is high quality in one situation (that is, it meets the need) might be low quality in a different situation (if it doesn't meet the need).

While there is no single, agreed-upon set of requirements that data must meet to be considered high quality, these six core dimensions can be considered indicative of high-quality data:

- **Completeness.** The degree to which all requisite information is included and data values have no missing elements
- **Uniqueness.** No entity instance (thing) will be recorded more than once based upon how that thing is identified
- **Timeliness.** The degree to which the currency of data aligns with business needs
- **Validity.** Data is valid if it conforms to the syntax (format, type, range) of its definition
- **Accuracy.** The degree to which data correctly describes the real object or event being described
- **Consistency.** The absence of difference, when comparing two or more representations of a thing against a definition

In addition, other characteristics that impact data quality are:

- **Usability**. Is it understandable, simple, relevant, accessible, maintainable?
- **Timing**. Is it stable yet responsive to legitimate change requests?
- **Flexibility**. Is comparable and compatible with other data? Does it have useful groupings and classifications? Can it be repurposed and is it easy to manipulate?
- **Confidence**. Are the data governance, protection, and security processes in place? Is the data verified and verifiable?
- **Value**. Is there a good cost/benefit case for the data?

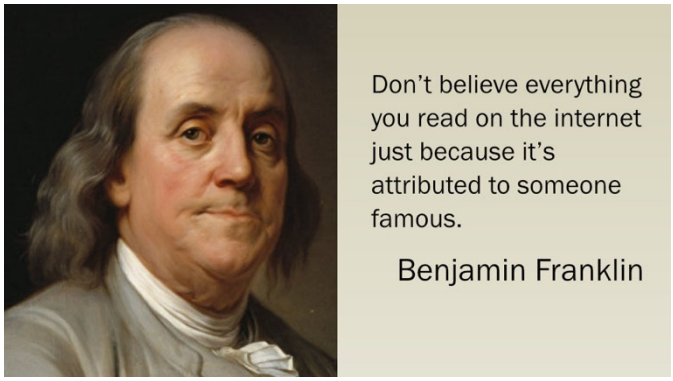


FIGURE 20: HIGH-QUALITY DATA

Why Is Data Quality Important?

Data quality is foundational to ensuring that the output of any analysis based on that data is valid. If data fails to meet quality standards, the output of an analysis that uses that data would be suspect and invalid.

How Does the Concept of Data Quality Apply to You?

Many people recognize poor-quality data when they see it; however, fewer can define what they mean by high-quality data. To help assess and improve an organization's state of data quality, ask these types of questions:

- What do stakeholders mean by "high-quality data"?
- What is the impact of low-quality data on business operations and strategy?
- How will higher quality data enable business strategy?
- What priorities drive the need for data quality improvement?
- What is the tolerance for poor-quality data?

- What governance is in place to support data quality improvement?
- What additional governance structures will be needed?

When Do You Consider the Quality of Data?

Data quality is a concern throughout the data life cycle. Data quality includes setting standards; building quality into the processes that create, transform and store data; and measuring data against requirements. (In this, formal data quality management is similar to continuous quality management for other processes or functions.)

Note that managing data to this level requires a data quality program team. This team is responsible for engaging both the business and the technical areas to apply quality management techniques and ensure that the data is fit for consumption for its intended purpose.

> *If you move data quality upstream and embed it in the [business process], it's much better than trying to catch [flawed data] downstream and then fixing it in all the different applications that used it.*
>
> *Aaron Zornes*

Who Is Involved in Ensuring the Quality of Data?

Data quality depends on all who interact with data.

Not only do all data management disciplines contribute to the quality of data, anyone who interacts with data can affect its quality. Producing high-quality data requires cross-functional commitment and coordination.

Key Takeaways

- Data quality management is a program, not a project. It includes both project and maintenance work across the organization, along with a commitment to communications and training.
- Data quality can be best represented and measured in six core dimensions: completeness, uniqueness, timeliness, validity, accuracy, and consistency.
- Data quality is everyone's business. All who come into contact with data can affect the quality of data.

Guiding Principles

- Data quality should focus on the data most critical to the enterprise and its customers.
- The quality of data should be managed across the data life cycle, from creation or procurement through disposal.
- Data quality should focus on preventing data errors and conditions that reduce the usability of data.
- Problems with the quality of data should be understood and addressed at their root causes.
- Data governance must support the development of high-quality data.
- All stakeholders in the data life cycle should have data quality requirements, and those requirements should be measurable. Measurements and methodology for measuring should be consistent.

DAMA-DMBOK Reference

For more information, see Chapter 13: Data Quality, pages 423-467.

Big Data and Data Science

What Are Big Data and Data Science?

Big data and data science result from significant technological changes that have allowed people to generate, store, and analyze larger and larger amounts of data. Big data refers to several aspects of data: the volume of data collected, its variety (structured and unstructured data, documents, files, streaming data, etc.), and the speed at which it is produced.

Data science refers to the analysis of big data. While traditional business intelligence provides analysis of structured data to describe past trends, data science uses a variety of methods to gain insight from big data and predict future behaviors.

Why Are Big Data and Data Science Important?

Big data and data science provide organizations the ability to find and act on business opportunities that can be discovered through data sets generated through a diversified range of processes.

Big data can stimulate innovation by making more and larger data sets available for exploration. This data can be used to define predictive models that anticipate customer needs and enable personalized presentation of products and services.

Using big data, data science can improve operations. Machine learning algorithms, for example, can automate complex time-consuming activities, thus improving organizational efficiency, reducing costs, and mitigating risks.

How Do Big Data and Data Science Apply to You?

The massive amounts of data that accumulates every day as we move through the world, interact with each other, and transact business creates a new data landscape. Big data is produced through email, social media, online orders, and even online video games. Data is generated by phones, point-of-sale devices, surveillance systems, sensors in transportation systems, medical monitoring systems, and military equipment.

The result is big data volumes that are exceptionally large. In traditional warehousing and analytic solutions, very large volumes of data pose challenges to data loading, modeling, cleansing, and analytics. The size and variety of data sets changes the overall way that data is stored and accessed, how data is understood, and how data is managed. Big data and data science programs address these issues in the new data landscape.

When Do You Use Big Data? When Do You Use Data Science?

The most common characteristics of big data include:

- **Volume.** Big data often has thousands of entities or elements in billions of records.
- **Velocity.** The large volume of big data is often generated in real time.
- **Variety.** Big data is generated in multiple forms, which requires storage of multiple formats.

When these characteristics are present, at an organization needs to redefine how its data is stored and managed.

Just as storing and managing large data sets pose challenges to an organization, so too does using traditional business intelligence tools to perform analysis (Figure 21). To provide business insights to big data sets, data science employs new methods of analysis, including:

- **Machine learning.** Programming machines to quickly learn from queries and adapt to changing data sets
- **Sentiment analysis.** Used to understand what people say and feel about brands, products, or services captured in surveys and other unstructured data sets

- **Data and text mining.** Help discover unknown relationships by revealing patterns. The key to data exploration and classification.
- **Predictive analytics.** Attempts to model data and predict future outcomes through evaluation of probability estimates

Who Is Involved in Big Data and Data Science?

As with data warehouse and business intelligence, a big data implementation will bring together a number of key cross-functional roles, including:

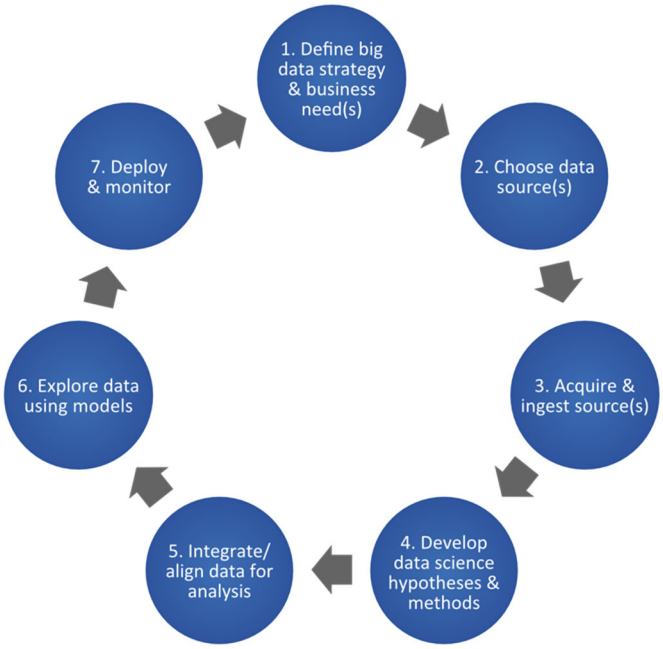


FIGURE 21: DATA SCIENCE PROCESS

- **Big data platform architects.** Hardware, operating systems, filesystems, and services.
- **Ingestion architects.** Data analysis, system of record, data modeling, and data mapping.
- **Metadata specialists.** Metadata interfaces, metadata architecture, and contents.
- **Analytic design leads.** End user analytic design, best practice guidance implementation in related toolsets, and end user result set facilitation.
- **Data scientists.** Provides architecture and model design consultation based on theoretical knowledge of statistics and computability, delivery on appropriate tools, and technical application to functional requirements.

Key Takeaways

- Big data refers to the extremely large volumes of data collected, often in real time, from a variety of sources.
- Data science is the analysis of big data, which can provide insight that is different from traditional business intelligence and has the capability of predicting future behaviors.
- To provide business insights to big data sets, data science requires new methods of analysis, such as machine learning, sentiment analysis, data and text mining, and predictive analytics.

Guiding Principles

A big data strategy must include criteria to evaluate:

- What problems the organization is trying to solve and why it needs analytics.
- What data sources to use or acquire.
- The timeliness and scope of the data to provision.
- The impact on and relation to other data structures.
- Influences on existing modeled data.

DAMA-DMBOK Reference

For more information, see Chapter 14: Big Data and Data Science, pages 469-500.

Data Management Maturity Assessment

What Is a Data Management Maturity Assessment?

A Data Management Maturity Assessment (DMMA) is an approach to process improvement that assesses where an organization stands on a scale from least mature (possessing no processes) to most mature (possessing fully optimal processes). Once the organization's current status is determined, a plan to improve and optimize an organization's data management can be developed and followed.

A DMMA can evaluate data management overall, or it can focus on a single DAMA knowledge area, or even examine a single process. It can help bridge the gap between the business and technology perspectives on the health and effectiveness of data management practices. A DMMA provides a common language for depicting what progress looks like regarding data management knowledge priorities and helps set goals for improvement across all areas of data management practices.

Why Is a Data Management Maturity Assessment Important?

The data management capability and maturity assessment process places the organization on a maturity scale by clarifying specific strengths and weaknesses. It helps organizations identify, prioritize, and implement improvement opportunities.

Organizations conduct maturity assessments for several reasons, including:

- **Regulation.** To ensure minimum levels of maturity are met according to the regulation of specific types of data
- **Data governance.** For planning and compliance validation of data management governance
- **Organizational readiness for process improvement.** For implementing new data management practices
- **Organizational change.** To determine an organization's readiness to address change related to data management overall
- **New technology.** To understand advancements in data management technology and how they apply to the likelihood of successful adoption
- **Data management issues.** To understand and address data issues to make better decisions about how to implement change

How Do You Use a Data Management Maturity Assessment?

When an organization gains an understanding of the characteristics of each state, it can determine its current level of maturity. Once it knows the level of maturity, it can put in place a plan to improve its capabilities.

Maturity assessments define five or six levels of maturity, each with its own characteristics that span from non-existent (or ad hoc), to optimized (or high performance) (Figure 22).

Items that can comprise a scale include:

- 0 – Absence of capability
- 1 – Initial or Ad Hoc: Success depends on the competence of individuals
- 2 – Repeatable: Minimum process discipline is in place
- 3 – Defined: Standards are set and used
- 4 – Managed: Processes are quantified and controlled
- 5 – Optimized: Process improvement goals are quantified

FIGURE 22: DATA MANAGEMENT MATURITY ASSESSMENT STATES

When Do You Use a Data Management Maturity Assessment?

A DMMA can be used whenever an organization desires to improve its data management program. A DMMA starts by establishing the state at which an organization's program exists.

This can be done by measuring how well data management capabilities meets criteria such as:

- **Activity.** To what degree is the activity in place?
- **Tools.** To what degree is the activity automated and supported by a common set of tools?
- **Standards.** To what degree is the activity supported by a common set of standards?
- **People and resources.** To what degree is the organization staffed to carry out the activity?

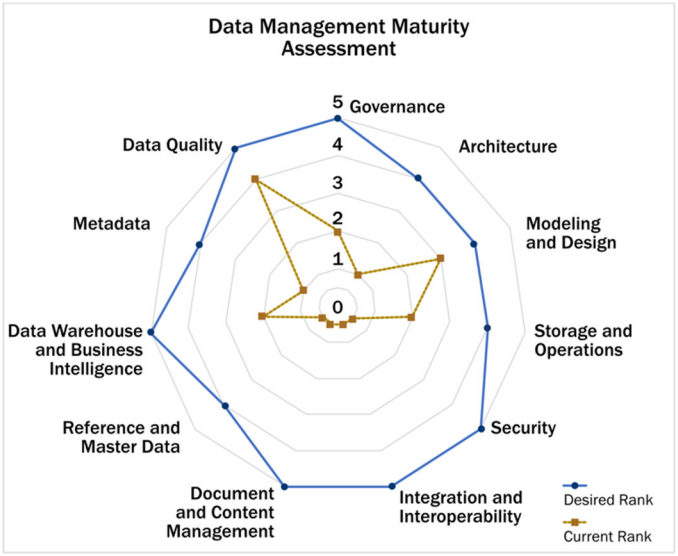


FIGURE 23: GRAPHICAL REPRESENTATION OF DMMA

Findings from a DMMA can be presented in several ways, including visually (Figure 23).

For each of the capabilities (governance, architecture, etc.), the outer ring of the display shows the level of capability the organization has determined it needs to achieve. The inner ring displays the level of capability as determined through the assessment. Areas where the distance between the two rings is largest represent the greatest risks.

Visual presentations of DMMA results can help set priorities and measure progress toward the goals.

Who Is Involved in Creating a Data Management Maturity Assessment?

An accurate data management maturity assessment is reached by creating a consensus view of current capabilities. Evidence comes from an examination of artifacts and through interviews of stakeholders in data management, including business areas, data management personnel, information technology participants, subject matter experts, and leadership.

Key Takeaways

- The primary goal of a data management capability and maturity assessment is to evaluate the current state of critical data management activities in order to plan for improvement.
- A DMMA can evaluate data management overall, or it can focus on a single DAMA knowledge area, or even a single process.
- Understanding the current level of data management maturity comes from an examination of artifacts and through interviews of stakeholders in data management.
- Once the maturity level of the focus area of the DMMA is established, a target maturity level can be identified, and a plan put in place to achieve the goal.

Guiding Principles

- Educate stakeholders about data management concepts, values, and practices.
- Clarify stakeholder roles and responsibilities in relation to organization data.
- Highlight the need to manage data as a critical asset.
- Broaden recognition of data management activities across the organization.
- Contribute to improving the collaboration necessary for effective data governance.

DAMA-DMBOK Reference

For more information, see Chapter 15: Data Management Maturity Assessment, pages 501-518.

Data Management Organization and Role Expectations

What Is the Data Management Organizational Structure?

While every organization is unique, there are ways to incorporate data management that have been tested and proven over time. Using one of these models (Figure 24) can help you successfully integrate data management into your organization. Each model has benefits and drawbacks (Table 3).

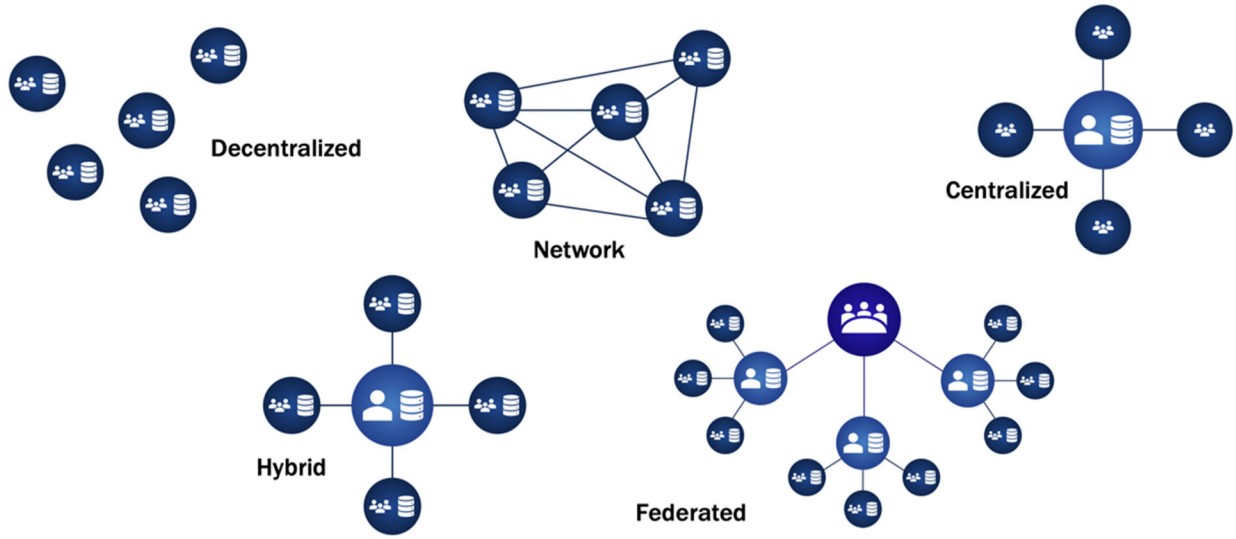


FIGURE 24: DATA MANAGEMENT ORGANIZATION MODELS

Why Is the Data Management Organizational Structure Important?

The way in which data management is organized and incorporated into the business processes of your organization is an essential factor in the success of the entire data management program. Effective incorporation will mean that data management processes (and knowledge areas) are evangelized and used throughout the organization.

How Does the Data Management Organizational Structure Apply to You?

Start by understanding the current status of data management. Ask the relevant parties about the following:

- **The role of data in the organization.** What role does data play in day-to-day business processes? What role does it play in organizational strategy?
- **Cultural norms about data.** What obstacles might exist to implementing or improving management and governance structures?
- **Data management and data governance practices.** Who currently makes decisions about data?
- **How work is organized and executed.** What business processes and structures are currently in place that can support formal data management?
- **How reporting relationships are organized.** What's the current data management organizational structure?
- **Skill levels.** What abilities do people have to manage data?

TABLE 3: DATA MANAGEMENT ORGANIZATIONAL STRUCTURE

	Decentralized	*Network*	*Centralized*	*Hybrid*	*Federated*
Characteristics	Responsibilities distributed across different business and IT areas.	Distributed, but with connections and accountabilities documented in a RACI (Responsible, Accountable, Consulted, and Informed) matrix.	Formal and mature. Everything is owned by data management organization. Led by a data management leader.	Centralized data management works with decentralized business groups, usually through executive steering committee, and tactical working groups. Organizational culture determines how responsibilities are divided.	Centralized strategy with decentralized execution; similar to a multiple-hybrid model.
Benefits	Flat structure. Each area has a clear understanding of its own data requirements. Relatively easy to implement or improve.	Flat structure. Each area has a clear understanding of its own data requirements. Easy to implement or improve. RACI matrix helps create accountability.	Formal executive position, one person at the top. Easier decision-making. Data can be managed by type of subject area.	Establishes direction from the top.	For large enterprises, can be only model that works. Different lines of business are empowered to meet data requirements. Easier to prioritize efforts across the organization.
Drawbacks	Many participants involved in decision-making. Hard to implement collaborative decisions. Hard to sustain over time. Difficult to enforce consistency.	Many participants involved in decision-making. Hard to implement collaborative decisions. Hard to sustain over time. Difficult to enforce consistency. Need to maintain and enforce RACI matrix.	Requires significant organizational change. Separation of data from core business processes can result in knowledge being lost over time.	Requires additional staff for "Center of Excellence." Different business unit priorities need to be managed from central perspective. Can be conflicts in priorities between central and decentralized units.	Significantly more complex than other models. Need to maintain a balance between lines of business and overall needs of organization.

What Factors Do You Need to Consider for Success?

These factors should be considered to help ensure a successful transition to a fully integrated data management process:

- **Executive sponsorship.** Need to ensure processes are effectively implemented and sustained for the long term
- **Clear vision.** Need to make sure that all who use data know what data management is, why it's important, and how their work will be affected by it
- **Proactive change management.** Data management is a change in business processes and should be managed as such
- **Leadership alignment.** Leaders need to agree on how success will be defined and be consistent in communication
- **Communication.** Communicate what data management is and why it is important. Communicate the vision
- **Stakeholder engagement.** Working with stakeholders will help create buy-in and support for the new approach
- **Orientation and training.** Train on new policies, processes, techniques, procedures, and tools. Tailor training so it is appropriate to the need
- **Adoption measurement.** Establish metrics that will measure the success of incorporating data management, track the metrics, and use the metrics to determine whether the change is working
- **Adherence to guiding principles.** Establish principles that articulate shared values to serve as reference points from which decisions can be made
- **Evolution, not revolution.** A guided, measured organizational change process will help ensure behavioral change is sustained

Who Is Involved in Creating an Organizational Structure?

Central roles include the following:

- **Executive**. Executives who support and lead the data management program for the organization on the IT side (Chief Information Officer, Information Resource Manager, Chief Technology Officer) and the business side (Chief Operating Officer, Chief Financial Officer, executive directors)
- **Chief Data Officer or Data Director.** Leads the overall efforts for the organization and helps to bridge the gap between technology and business areas
- **Business and program**. Focus on data governance functions, especially data stewardship. Data stewards are subject matter experts who understand the data at the most fundamental level and how it applies within the overall organization. They define business terms and valid values for the data and help resolve issues concerning data. Data stewards also help establish standards, policies, and procedures around the data.
- **IT data management.** Provides leadership and collaboration on a wide range of services, including data storage and security, application and technical architecture, and database administration
- **Specific IT roles.** Includes different types of architects, developers at different levels, database administrators, and a range of supporting functions. Examples include:
    - Data architect
    - Data modeler
    - Data model administrator
    - Database administrator
    - Data security administrator
    - Data integration architect
    - Data integration specialist
    - Analytics/report developer
    - Application architect
    - Technical architect
    - Technical engineer
    - Help desk administrator
    - IT auditor

Key Takeaways

- Data management can be incorporated into an existing organizational structure in multiple ways. Several models exist; an organization can determine which of the models will be the most effective.
- Data management involves all areas within an organization, including IT and business areas.
- For an organization's incorporation of data management to be successful, it's essential to consider these factors: sponsorship and leadership, clear vision, communication, and proactive change management. "Evolution not revolution!"

DAMA-DMBOK Reference

For more information, see Chapter 16: Data Management Organization and Role Expectations, pages 519-538.

What Is Change Management?

For data management to be effective, it must be incorporated into an organization. For many organizations, this can be a significant structural and cultural change.

Why Is Change Management Important?

Without an organizational change management strategy that contains a clear and compelling vision and a well-implemented, consistent communication plan, incorporating data management into an existing organization will not succeed.

To ensure the organizational change process is successful, these key factors need to be considered (Figure 25):

- **Vision.** Create a clear, feasible, and focused picture of the future that identifies what the end point of the change process will be. For data management initiatives, the vision must articulate the challenges with existing data management practices, the benefits of improvement, and the path to get to a better future state.
- **Communication.** Clearly share the vision. Use metaphor, analogy, and examples; consistently communicate the vision and explain any seeming inconsistencies. Create a formal communication plan, implement the plan, and ensure communication continues through and after the closing of the change management process.
- **People.** Change happens when people behave differently, not when a new policy is approved. Change management is effective when people believe that systems must be changed, when they are engaged in defining the change, and when they are involved in how and when the change will take place.

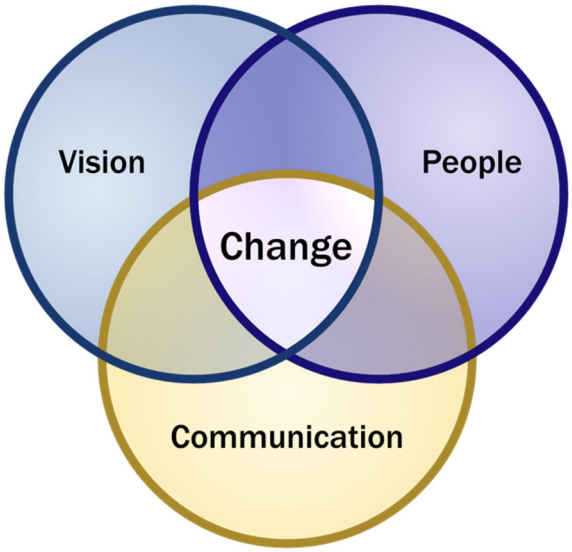


FIGURE 25: KEY FACTORS FOR ORGANIZATIONAL CHANGE

How Does Change Management Apply to You?

If data management is not currently fully incorporated into your organization, some level of change management will need to be used to ensure that data management is effectively deployed.

John P. Kotter, one of the most respective researchers in change management, notes common obstacles to change:

- Inward-focused cultures
- Paralyzing bureaucracy
- Parochial politics
- Low levels of trust
- Lack of teamwork
- Arrogance
- Lack of or failure of leadership
- Fear of the unknown

If some or all of these elements are present (in a greater or lesser degree) in your organization, you will likely need to account for them when incorporating data management.

What Are the Key Steps for Change Management?

Kotter recognizes eight steps for effective organizational change. The four foundational steps are:

1. **Establish a sense of urgency.** Without urgency, there is no motivation to change.
2. **Create the guiding coalition.** Change results from having effective leadership (to drive the change) and management (to keep the process under control). Members of a successful guiding coalition have influence over their peers, either through formal authority or as a result of their status or experience.
3. **Develop a vision and strategy.** Neither an authoritarian decree nor micromanagement will result in long-lasting, effective change. Only motivating people through a clear vision and strategy will result in success.
4. **Communicate the change vision.** The change vision needs to be communicated for the process to be effective. If you think you're communicating enough, you likely still need to communicate more.

Following those steps, Kotter notes four additional steps:

5. Empower broad-based actions.
6. Create short-term wins.
7. Consolidate gains and produce more change.
8. Anchor new approaches in the culture.

Who Is Involved in Change Management?

Key roles that are needed for successfully incorporating data management into an organization include:

- **Change agents.** Help the change proceed as smoothly as possible. Change agents actively listen to employees, customers, and other stakeholders to identify problems before they arise.
- **Managers in charge of the affected units.** Need to be able to reduce complacency in teams under their direct control.

Key Takeaways

- A clear and compelling vision is essential for the data management program. The vision must clarify the direction, motivate people to take the right steps, and align the organization in an efficient way.
- Data management programs are an ongoing program effort, not a one-time project. Communication that supports the program needs to be measured and sustained for ongoing success.
- Data management programs need to cultivate ongoing support. Change management practices need to be applied in a continual approach as new employees and leaders come into the organization so the data management program can continue to grow and be supported.

DAMA-DMBOK Reference

For more information, see Chapter 17: Data Management and Organizational Change Management, pages 539-574.

administrative metadata. Data that is used to manage resources over their life cycle (examples include version numbers and archive dates).

attribute. Property that identifies, describes, or measures an entity.

big data. Characterized by high volume (elements in billions of records), intense velocity (often collected in real time), and variety (requiring storage of multiple formats with data structure that is inconsistent within or across data sets).

Business Data Steward. Business professional, most often a recognized subject matter expert, accountable for a subset of data. They work with stakeholders to define and control data.

business glossary. System of record for business terms related to data. Many organizations develop their own internal vocabulary.

business intelligence. Data analysis that is performed on the data stored in data warehouses in order to improve an organization's success. Business intelligence can also refer to the tools that support this analysis.

candidate key. Minimal set of one or more attributes (such as a simple or compound key) that identifies the entity instance to which it belongs. Minimal means that no subset of the candidate key uniquely identifies the entity instance.

centralized operating model. The most formal and mature data management operating model. Everything is owned by the data management organization.

change management. Process and procedures to identify, propose, document, review, evaluate, authorize, and track any changes to project baselines such as project scope and budget changes.

Chief Data Officer. An organization officer responsible for organization-wide governance and use of data as an asset.

Chief Data Steward. Data steward who manages data assets on behalf of an entire organization.

CJIS. Criminal Justice Information Services.

classification scheme. Codes that represent controlled vocabulary.

communication plan. In the context of change management, a communication plan provides a roadmap to guide the work toward the goal of the change. It includes elements such as the message, goal, audience, channel, timing, frequency, materials, communicators, expected response, metrics, and a budget and resource plan.

conceptual model. Identifies the different entities in data and how they relate to each other without referencing technology.

confidential data. Generally, data that cannot be shared outside the organization without a properly executed non-disclosure agreement or similar in place.

controlled vocabulary. Defined list of explicitly allowed terms used to index, categorize, tag, sort, and retrieve content through browsing and searching.

Coordinating Data Steward. Leads and represents teams of business and technical data stewards in discussions across teams and with executive data stewards. Coordinating data stewards are particularly important in large organizations.

data cube. A representation of data that is a three-dimensional matrix, or cube. Subject areas form two dimensions (the rows and columns). Factors, or measures, of the subject areas are represented in a third dimension, which creates a data cube.

data dice. Slice on more than two dimensions in a cube, or more than two consecutive slices.

data dictionary. Defines the structure and contents of data sets, often for a single database, application, or warehouse.

data enrichment. Adding attributes that can improve entity resolution services.

data lake. Environment where a vast amount of data of various types and structures can be ingested, stored, assessed, and analyzed.

data latency. Time difference between when data is generated in the source system and when the data is available for use in the target system.

data lineage. Metadata that describes where data came from and where it moves over time.

data management framework. The defined approach of aligning and applying the key

independent knowledge areas and principles to manage an organization's data. The DAMA-DMBOK framework has been adopted by the state of Texas for the implementation of enterprise data management practices.

data map. Inventory of all electronically stored information data sources, applications, and IT environments that includes the owners of the applications, custodians, relevant geographical locations, and data types.

data mart. Provide data prepared for analysis. This data is often a sub-set of warehouse data designed to support particular kinds of analysis or a specific group of data consumers.

data mining. Analysis that helps discover unknown relationships by revealing patterns in data using various algorithms.

data model. Describes an organization's data as the organization understands it, or as the organization wants it to be. Data models are the main medium used to communicate data requirements from business to IT and within IT from analysts, modelers, and architects, to database designers and developers.

Data Owner. Business data steward who has approval authority for decisions about data within their domain.

data quality dimensions. Dimensions of quality help people understand what is being measured. Consistent application of dimensions will help with measurement and issue management processes.

data science. Field of analysis that integrates methods from mathematics, statistics, computer science, signal processing, probability modeling, pattern recognition, machine learning, uncertainty modeling, and data visualization in order to gain insight and predict behaviors based on big data sets.

data slice. Subset of data corresponding to a single value for one or more attributes in a dimension.

data standardization. Ensuring data content conforms to standard reference data values (e.g., country codes), formats (e.g., telephone numbers) or fields (e.g., addresses).

Data Steward. Person who manages data assets within an organization. Data stewards are generally responsible for data categorization.

data stewardship. See Data Steward.

data store. Repository for storing, managing, and distributing data sets on an enterprise level.

data validation. Identifying data prove-ably erroneous or likely incorrect or defaulted (for example, removal of clearly fake email addresses).

data warehouse. Combination of two primary components: An integrated decision support database and the related software programs used to collect, cleanse, transform, and store data from a variety of operational and external sources. In its broadest context, a data warehouse includes any data stores or extracts used to support the delivery of data for business intelligence purposes.

data-at-rest. Situation where data resides in a system without moving between systems. Data-at-rest can be protected by a firewall. Compare with **data-in-motion**.

data-in-motion. Situation where data requires a network in order to move between systems.

database. Any collection of stored data, regardless of structure or content. Some large databases refer to instances and schema.

decentralized operating model. An organization-level model for data management in which responsibilities are distributed across different lines of business and information technology. Collaboration is committee-based; there is no single owner.

denormalization. Deliberate transformation of normalized logical data model entities into physical tables with redundant or duplicate data structures.

descriptive metadata. Data that describes a resource and enables identification and retrieval (examples include title, author, and subject).

development environment. Database environment used to create and test developer changes that will be implemented in a production environment.

discovery. Legal term that refers to pre-trial phase of a lawsuit where both parties request information from each other to find facts for the case and to see how strong the arguments are on either side.

document. Electronic or paper object that contains instructions for tasks, requirements for how and when to perform a task or function, and logs of task execution and decisions.

drill down/up. Navigating among levels of data, ranging from the most detailed (down) to the most summarized (up).

encryption. Process of translating plain text into complex codes to hide privileged information, verify complete transmission, or verify the sender's identity.

Enterprise Data Steward. Data steward who has oversight of a data domain across business functions.

entity. Within data modeling, an entity is a thing about which an organization collects information.

entity resolution. Process of determining whether two references to real world objects refer to the same object or to different objects.

ESI. Electronically stored information.

ETL. Extract, transform, and load. Essential steps in moving data around and between applications and organizations.

Executive Data Steward. Senior manager who serves on a Data Governance Council.

federated operating model. An organizational-level model for data management. The federated model provides layers of centralization/decentralization, which are often required in large global enterprises. A variation on the hybrid operating model.

FERPA. Family Educational Rights and Privacy Act. United States federal law that protects the privacy of student education records.

fit for purpose. the quality of data for a particular purpose in terms of its completeness, timeliness, conformity, uniqueness, integrity, consistency, accuracy, as well as how well the data meets the expectations of how users define useful information.

foreign key. Used in physical and sometimes logical relational data modeling schemes to represent a relationship.

GDPR. General Data Protection Regulation. Legal framework that sets guidelines for the collection and processing of personal information from individuals who live in the European Union (EU).

golden record. Record that represent the most accurate data about entity instances within a trusted source. See also **trusted source**.

guiding coalition. In the context of change management, a guiding coalition is the powerful and enthusiastic team of volunteers from across the organization that helps to put new strategies into effect and transform the organization.

hash encryption. Encryption that uses algorithms to convert data into a mathematical representation.

hierarchical database. Database in which data is organized into a tree-like structure with mandatory parent/child relationships: each parent can have many children, but each child has only one parent (also known as a 1-to-many relationship).

HIPAA. Health Insurance Portability and Accountability Act. United States federal law protecting the privacy and security of certain health information.

hub-and-spoke model. Data interaction model where data is consolidated (either physically or virtually) in a central data hub that many applications use.

hybrid operating model. An organizational-level model for data management where a centralized data management Center of Excellence works with decentralized business unit groups, usually through both an executive steering committee representing key lines of business and a set of tactical working groups addressing specific problems. The hybrid operating model encompasses benefits of both the decentralized and centralized models.

information architecture. Process of creating structure for a body of information or content.

instance. An execution of database software controlling access to a certain area of storage. An organization will usually have multiple instances executing concurrently, using different areas of storage. Each instance is independent of all other instances.

logical model. Explains the data in as much detail as possible without regard to how they will be physically implemented.

machine learning. Explores the construction and study of learning algorithms.

malware. Any malicious software created to damage, change, or improperly access a computer or network.

masking. Security measure in which data is made less available by process that removes, shuffles, or otherwise changes the appearance of the data, without losing the meaning of the data or the relationships the data has to other data sets, such as foreign key relationships to other objects or systems. Also called **obfuscation**.

metadata. Data that provides information and context about other data.

network operating model. An organization-level model for data management in which data management operates as a series of known connections between people and roles. The model can be diagrammed as a "network" and is made formal through a documented series of connections and accountabilities via a RACI (Responsible, Accountable, Consulted, and Informed) matrix.

normalization. Process of applying rules in order to organize business complexity into stable data structures.

obfuscation. Security measure in which data is made less available by process that removes, shuffles, or otherwise changes the appearance of the data, without losing the meaning of the data or the relationships the data has to other data sets, such as foreign key relationships to other objects or systems. Also called **masking**.

OLAP. Online analytical processing. Method of modeling data in a database using normalized transactional or operational data.

OLTP. Online transactional processing. Method of modeling data in a database using denormalized historical data.

ontology. Type of taxonomy that represents a set of concepts and their relationships within a domain.

organization. Administrative and functional entity that applies data management practices. Can be used to refer to state agencies, institutions of higher education, and public and private government bodies.

phishing. A phone call, instant message, or email meant to lure recipients into giving out valuable or private information without realizing they are doing so.

physical model. Represents exactly how data will be created within a physical location such as a database.

PII. See personal identification information. Any information that can personally identify the individual (individually or as a set), such as name, address, phone numbers, schedule, government ID number, account numbers, age, race, religion, ethnicity, birthday, family members' names or friends' names, employment information (HR data), and in many cases, remuneration. Also known as Personally Private Information.

point-to-point model. Data interaction model where systems share data by directly passing the data to each other.

predictive analytics. Sub-field of supervised learning (a form of machine learning) that attempts to model data elements and predict future outcomes through evaluation of probability estimates.

primary key. The candidate key that is chosen to be the unique identifier for an entity. Even though an entity may contain more than one candidate key, only one candidate key can serve as the primary key for an entity.

private-key encryption. Encryption that uses one key to encrypt the data. Both the sender and the recipient must have the key to read the original data.

production environment. Technical database environment where all business processes occur.

public-key encryption. Encryption where the sender and the receiver have different keys. The sender uses a public key that is freely available, and the receiver uses a private key to reveal the original data.

record. A subset of documents that provide evidence that actions were taken and decisions were made in keeping with procedures; records can serve as evidence of the organization's business activities and regulatory compliance.

regulated data. Data that is regulated by external laws, industry standards, or contracts that influence how data can be used, as well as who can access it and for what purposes.

relational database. Database where data elements or attributes (columns) are related into rows. Tables in relational databases are sets of relations with identical structure.

relationship. An association between entities. A relationship captures the high-level interactions between conceptual entities, the detailed interactions between logical entities, and the constraints between physical entities.

replication. Creating exact copies of data sets in multiple physical locations.

risk. Refers both to the possibility of loss and to the thing or condition that poses the potential loss.

roll-up. Computing all of the data relationships for one or more dimensions.

schema. A subset of a database objects contained within the database or an instance. Schemas are used to organize objects into more manageable parts.

sentiment analysis. Field dedicated to the exploration of subjective opinions or feelings collected from various sources about a particular subject.

spam. Unsolicited, commercial email messages sent out in bulk, usually to tens of millions of users in hopes that a few may reply.

stakeholder. Group or individual who can affect or who is affected by the success of a project.

structural metadata. Data that describes relationships within and among resources and their component parts (examples include number of pages and number of chapters).

structured data. Data that is organized and formatted according to a model.

taxonomy. Umbrella term referring to any classification or controlled vocabulary.

Technical Data Stewards. Information technology professionals operating within one of the knowledge areas, such as data integration specialists, database administrators, business intelligence specialists, data quality analysts, or metadata administrators.

test environment. Database environment that serves several purposes: quality assurance, integration testing, user acceptance testing, and performance testing.

text mining. Analyzes documents with text analysis and data mining techniques to automatically classify content into different ontologies.

thesaurus. Type of controlled vocabulary used for content retrieval.

threat. Potential offensive action that could be taken against an organization.

trusted source. Data that is recognized as the "best version of the truth" based on a combination of automated rules and manual stewardship of data content.

unstructured data. Stored data that is maintained outside of relational databases. Unstructured data does not have a data model that enables the understanding of its content or how it is organized.

virus. Program that attaches itself to an executable file or vulnerable application and delivers a payload that ranges from annoying to extremely destructive.

vision. In a business or organizational context, a vision is a statement that focuses on what the organization will become. In change management, the vision provides the context and meaning of the change effort.

vulnerability. Weaknesses or defect in a system that allows it to be successfully attacked and compromised—essentially a hole in an organization's defenses. Some vulnerabilities are called exploits.

worm. Program built to reproduce and spread across a network by itself.

Texas Department of Information Resources
300 West 15th St., Suite 1300
Austin, TX 78701
dir.texas.gov