Information and Data Management

inShare

SHARE

Print

Definition: *Information and Data Management (IDM) is the practice of putting into place policies, procedures, and best practices that ensure that data is understandable, trusted, visible, accessible, optimized for use, and interoperable. IDM functions include processes and procedures that cover strategy, planning, modeling, security, access control, visualization, data analytics, and quality among others. Outcomes encompass the improvement of data quality and assurance, enablement of information sharing, and the fostering of data reuse by minimizing data redundancy.*

Keywords: *business intelligence, data, data analysis, data governance, data management, data mart, data mining, data modeling, data quality, data warehouse, database, database management system (DBMS), information management, master data management, metadata, data migration*

MITRE SE Roles & Expectations: MITRE systems engineers (SEs) will encounter IDM-related activities on most government programs. They are expected to understand the customer organization's data requirements and help develop concepts for how to use and manage data, as well as how to apply appropriate IDM mechanisms in the organization's system environment. The IDM SE role may start before system acquisition, when only general requirements are known. Typically it encompasses planning, training, and operational support for the awareness, coordination, and integration of data and information management activities. MITRE SEs are expected to be able to determine the size of data, data security and privacy requirements, and data sharing requirements. This may include specifying the information needs, data, software, and hardware, as well as the skills and staffing required to support the system's operational IDM needs. At the end of a system life cycle, the SE may need to consider where and how data is stored or disposed.

Discussion

Data is the"life blood" of an organization, for as it flows between systems, databases, processes, and departments, it carries with it the ability to make the organization smarter and more effective. The highest performing organizations pay close attention to the data asset, not as an afterthought but rather as a core part of defining, designing, and constructing their systems and databases. Data is essential to making well-informed decisions that guide and measure achievement of organizational strategy. For example, an organization may analyze data to determine the optimal enforcement actions that reduce non-compliant behavior. Similarly, data is also at the heart of the business processes. An organization may enhance a process to catch fraudulent activities by including historical risk-related data. Over time, this type of process improvement can result in material savings. Even a single execution of a business process can translate into substantial benefits, such as using data patterns to stop a terrorist at a border or filtering a cyber-attack.

How an organization uses and manages the data is just as important as the mechanisms used to bring it into the environment. Having the right data of appropriate quality enables the organization to perform processes well and to determine which processes have the greatest impact. These fundamental objectives leverage data by transforming it into useful information. The highest performing organizations ensure that their data assets are accessible to the processes and individuals who need it, are of sufficient quality and timeliness, and are protected against misuse and abuse. Successfully leveraging data and information assets does not happen by itself; it requires proactive data management by applying specific disciplines, policies, and competencies throughout the life of the data.

Similar to systems, data goes through a life cycle. Figure 1 presents the key phases of the data life cycle.

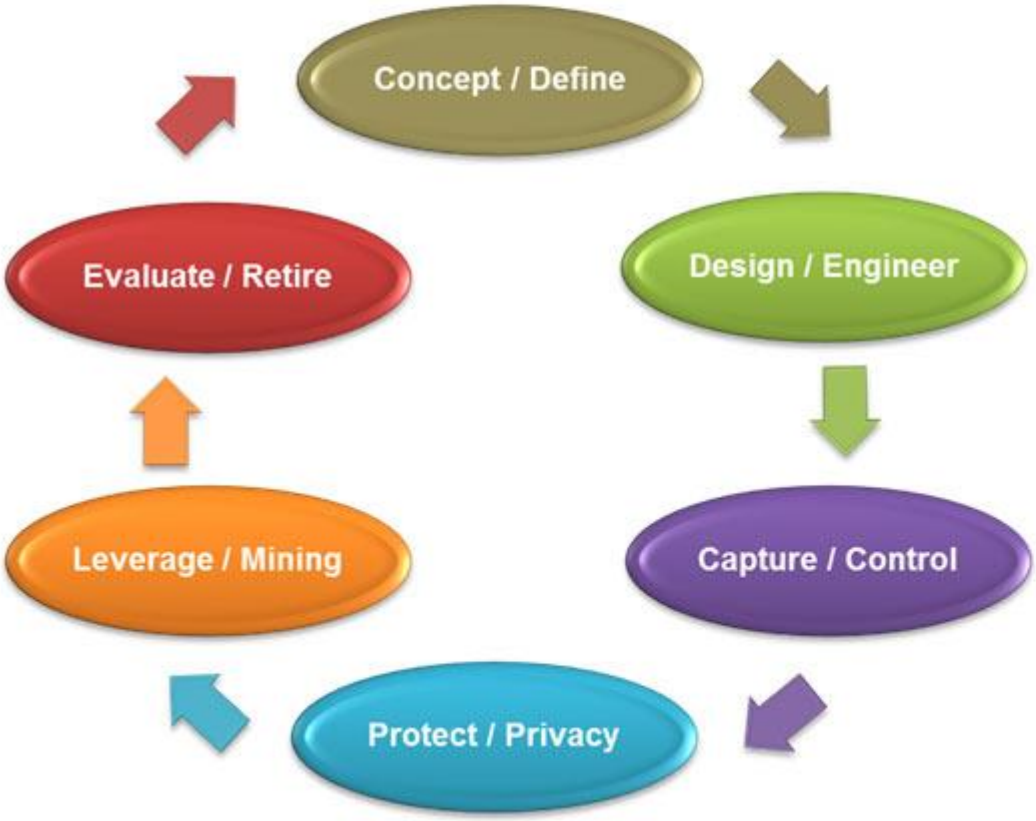

Figure 1. Data Life Cycle

Effective data management through all of the data life-cycle phases is the foundation for reliable information. Data may have different uses at different times and requires different management handling in the life-cycle phases. For instance, an organization may consider critical data required for discovery as very valuable during a key event, but when the event is over, the information diminishes in value quickly (e.g., data collected for predicting the weather).

Data may typically have a longer lifespan than the project that creates it. Though the funding period formally defines the lifespan of most projects, the resultant data may be available for many years afterwards. If an organization manages and preserves the data properly, the data is available for use well into the future, increasing the investment made in generating it by increasing visibility and usefulness. The time spent in planning and implementing effective data management pays dividends far in excess of its investment costs.

IDM is the set of related disciplines that aims to manage the data asset fully, from conception to retirement. Figure 2 presents a high-level view of data management disciplines.

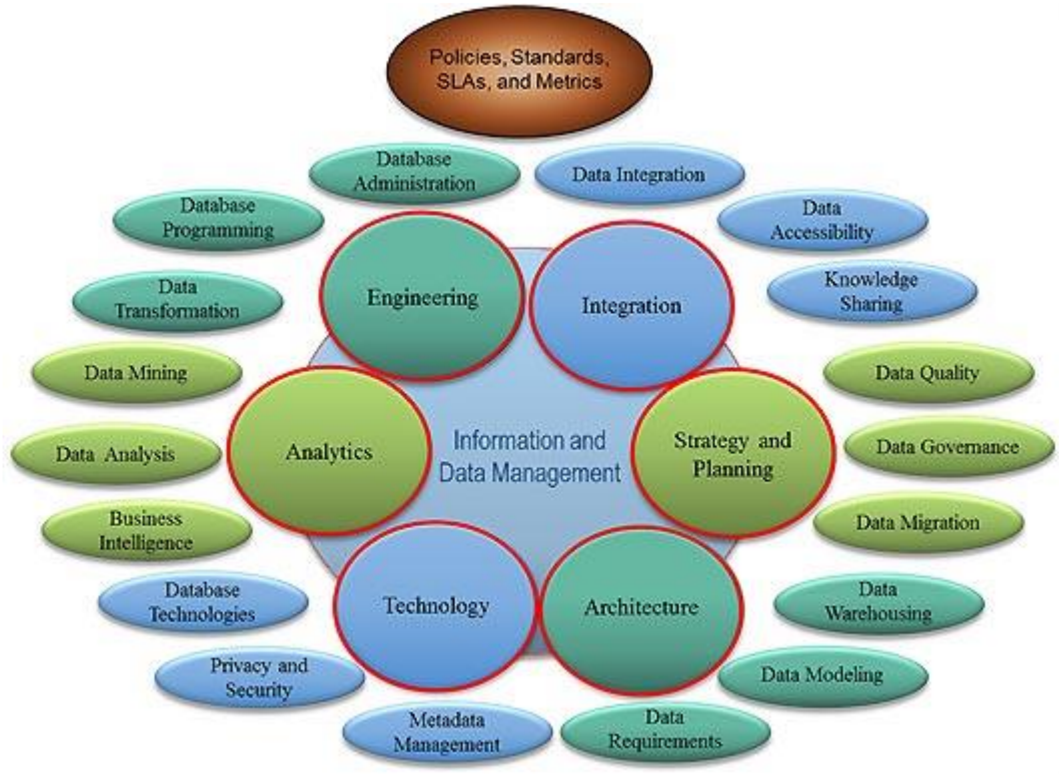

Figure 2. Data Management Disciplines

Data without context has no value; data that consumers never use is worthless, too. The value of data is in the information it contains and uses. The extraction of information and providing it in an appropriate format may be summarized as data analysis and reporting. However, data analysis and reporting encompasses several overlapping disciplines, among them statistical analysis, data mining, predictive analysis, artificial intelligence, and business intelligence. IDM has an appreciation for these disciplines and may use the same tools and incorporate some of these disciplines. The common ground among all of these disciplines and IDM is making good use of data.

Knowledge Required

An SE dealing with data should be knowledgeable in at least one of the following environments or disciplines:

- **Operational data:** Operational environments provide core transactional capabilities (i.e., processing applications, claims, payments, etc.) that typically work with database management systems.
- **Data exchange:** Organizations use data exchanges and data exchange standards to share information with internal or external parties.

Standardizing exchange formats and metadata minimizes impacts to both the sending and receiving systems and reduces cost and delivery time. A related discipline is master data management (MDM). An example is a vendor list. The U.S. Treasury requires specific information identifying contractors before the federal government reimburses them. Most federal agencies use this centrally collected list. Exchange, transform, and load (ETL) tools typically support these types of data exchange activities. ETL tools manipulate data and move it from one database environment to another.

- **Data warehouses [1]:** The integration of similar and disparate data from across organizational, functional, and system boundaries can create new data assets. The organizations can use the new data to ensure consistent analysis and reporting and to enhance the information needed for decision making. Data may be structured, unstructured, or both. Business intelligence (BI) has become a recognized discipline. It takes advantage of data warehouses (or similar large data consolidation) to generate business performance management, and reporting.
- **Data mining and knowledge discovery:** Mining applications explore the patterns within data to discover new insight and predictive models. An organization may use specialized software that applies advanced statistics, neural net processing, graphical visualization, and other advanced analytical techniques against targeted extracts of data. In addition, tools may evaluate continuously streaming data within operational sources.
- **Database administration [2]:** Knowledge in this discipline requires specific training related to a specific DBMS and being certified. A certified database administrator (DBA) is responsible for the installation, configuration, and maintenance of a DBMS (e.g., storage requirements, backup and recovery), as well as database design, implementation, monitoring, integrity, performance, and security of the data in the DBMS.
- **Data architecture:** A data architect is responsible for the overall data requirements of an organization, its data architecture and data models, and the design of the databases and data integration solutions that support the organization. The data structure must meet business requirements and regulations. Good communication and knowledge of the business must be part of the data architect's arsenal. A specialized area in data architecture is the role of the data steward. The data steward is usually responsible for a specific area of data such as one or more master data.

Best Practices and Lessons Learned

What is in it for me? Conveying the importance of information and data management to federal executives is the most common challenge that an SE will encounter. Most organizations focus their time and energy on application development and the technical infrastructure. For information systems, at best this approach leads to delays in implementation and, at worst, data is not trusted and system failures occur. The organization needs to coordinate data and IT staff with the business staff to align strategy and improvement initiatives. The best approach for long-term success is to initiate a program that gradually addresses the multifaceted challenges of data management.

An effective data management program begins with identifying core principles and collaborative activities that form the foundation for providing efficient, effective, and sustainable data. The organization should interweave the following core principles throughout all of the data management activities:

- Data collected is timely, accurate, relevant, and cost effective.
- Data efforts are cost efficient and purposeful, and they minimize redundancy and respondent burden.
- Data is used to inform, monitor, and continuously improve policies and programs.
- Data activities seek the highest quality of data and data collection methodologies and use.
- Data activities are coordinated within the organization, maximizing the standardization of data and sharing across programs.
- Partnerships and collaboration with all stakeholders are cultivated to support common goals and objectives around data activities.
- Activities related to the collection and use of data is consistent with applicable confidentiality, privacy, and other laws, regulations, and relevant authorities.
- Data activities adhere to appropriate guidance issued by the organization, its advisory bodies, and other relevant authorities.

The data management program supports the framework that facilitates relationships among the organization's staff, stakeholders, communities of interest (COIs), and users. It also provides a plan and approach to accomplish the next level of work needed to implement the technical architecture. The ultimate goal of the program is to define a data-sharing environment to provide a single, accurate, and consistent source of data for the organization.

Design for use. A simple analogy is to view data as a set of books. With a small number of books and only one individual interacts with them, organizing the books is a matter of preference. Further, finding sections of interest in the books is manageable. However, as the number of books increases and the number of individuals interacting with them also increases, additional resources are required to acquire, organize, and make the books available when requested.

In the discipline of data management, acquiring, managing, and extracting information are also true for data, but at a more intricate level. The complexity of the tasks related to database 1 design grows as requirements, number of users, and data relationships increases. The most common approach to deal with large amount of data with multiple users is to store data in a Database Management System 2 (DBMS). In many cases, the DBMS is a relational DBMS (RDBMS), which reduces reliance on software developers and provides an environment to establish data standards. However, working with any DBMS requires knowledge of the specific DBMS. In addition, an SE would have to be proficient in specific tools such as data modeling, query language or others. A DBMS designer also would take into considerations:

- Business requirements
- Operational requirements (is it mainly an interactive system for data collection or is it for querying?)
- Access and usage requirements
- Performance
- Data structure and replications requirements
- Interfaces and data-sharing requirements
- Reporting and analytical requirements
- Data volume
- Privacy and security

The complexity of data may require both a data architect and a certified DBA. A MITRE SE may play these roles or advise someone playing these roles. A data architect is usually associated with data strategy and data modeling. The data architect may propose a physical data model, but it is in coordination with the DBA. Though the DBA's responsibilities usually start with the physical database model, their responsibilities span into all physical data responsibilities while data is in the DBMS.

Fit for consumption. The Federal Data Architecture Subcommittee (DAS) Data Quality Framework [3] defines data quality as"the state of excellence that exists when data is relevant to its intended uses, and is of sufficient detail and quantity, with a high degree of accuracy and completeness, consistent with other sources, and presented in appropriate ways." A simpler definition is "data fit for its intended use." A set of characteristics provides the best definition for data quality. These are data accessibility, data completeness, data consistency, data definition, data accuracy, data relevancy, data timeliness, and data validity. Emphasis on one characteristic over another depends on the environment. The following environments introduce key considerations:

- Standalone: Usually data from a single application with limited or no interfaces
- Enterprise-wide: Data of relevance to the enterprise with no interfaces to the external world
- Multi-enterprise: Data shared outside the enterprise with the need to meet external regulations

In a standalone environment, obtaining an acceptable level of data quality is relatively simple. The organization can meet most of the characteristics because they are part of the application requirements and design. In such a case, data quality usually means data accuracy and data validity. The organization manages the data quality by ensuring that data collection meets requirements and there are tools (automated or otherwise) to control and monitor data validity and accuracy.

The picture changes in an enterprise environment because there are competing needs for the same sets of data. For example, an accounting department must account for every penny to avoid legal consequences, whereas budgeting operations are typically not concerned with small dollar variations. In this environment, all the data quality characteristics are important, but usage determines what is acceptable and what is not. Another common factor is the variation in terminology, such as using the same word to mean two different things or using different coding lists for equivalent attributes. A recommended solution to eliminate or reduce miscommunications is to establish data stewardships and data governance to facilitate mediation and conflict management. In addition, as in most large endeavors, documentation and standards are critical for success.

The multi-enterprise environment adds complexity (i.e., data sharing). An organization may use the data in the manner originally intended. Documentation of data content is important, and control of data use is more limited, so standards are harder to enforce. As an example, the unique identification of an individual varies from state to state. A federal agency integrating data from states that do not share unique identifiers may introduce data incompatibility issues (e.g., fraud may go on unnoticed). This issue is not easily resolved because one state may mandate the use of social security number as an identifier, whereas another state may forbid it. In such a case, compromised data quality will occur until the organization implements an innovative solution that ensures uniqueness.

'Cause they said so. Data governance encompasses roles, responsibilities, accountability, policy enforcement, processes, and procedures that ensure data value, quality improvement, and standard definitions. It also entails the overall management of the availability, usability, integrity, and security of the data employed in the enterprise. A sound governance program includes a governing council, an accountability structure, a defined set of procedures, and a plan to execute those procedures. The Data Governance Framework presented in Figure 3 provides an overview of the expected governance roles and responsibilities, accountability, and authority for the strategic, collaborative, and operational levels and the IT subject matter experts.

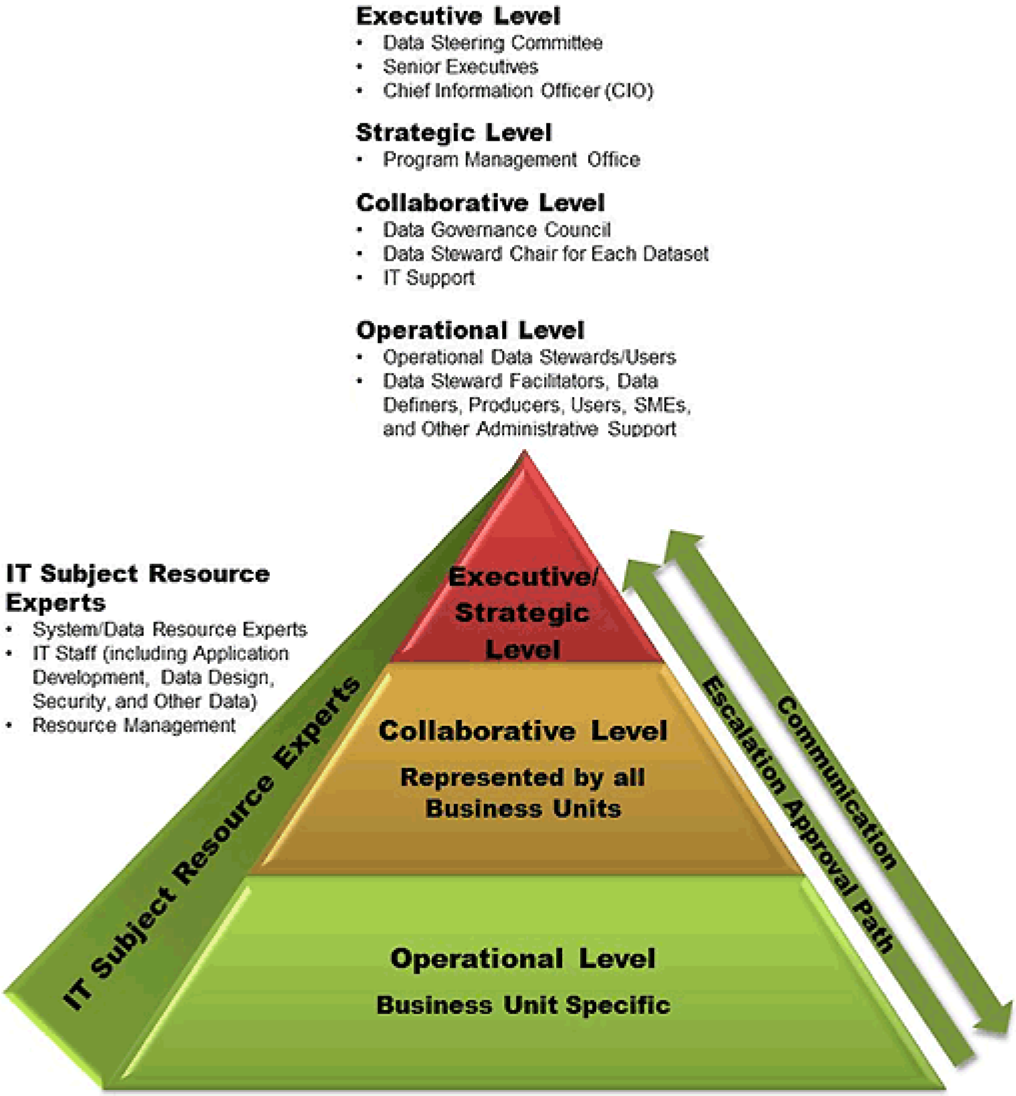
Executive Level
• Data Steering Committee
• Senior Executives
• Chief Information Officer (CIO)
Strategic Level
• Program Management Office
Collaborative Level
• Data Governance Council
• Data Steward Chair for Each Dataset
• IT Support
Operational Level
• Operational Data Stewards/Users
• Data Steward Facilitators, Data Definers, Producers, Users, SMEs, and Other Administrative Support
IT Subject Resource Experts
• System/Data Resource Experts
• IT Staff (including Application Development, Data Design, Security, and Other Data)
• Resource Management
IT Subject Resource Experts
Executive/ Strategic Level
Collaborative Level
Represented by all Business Units
Operational Level
Business Unit Specific
Escalation Approval Path
Communication

Figure 3. Data Governance Framework

The line of business (LOB) chief has a clear responsibility over the business. In addition, the staff at the operational level (i.e., data stewards, SMEs, etc.) receive direction from the LOB chief. Operational data stewards are responsible for managing data in the best interest of the LOB. However, when several LOBs are dealing with the same set of data, conflicts may arise because of their varying needs. Resolution of these issues requires collaboration among the LOBs. The most important role of the data governance council (or equivalent) is conflict resolution. Business and technical staffs, specifically the collaborative data stewards should define the composition of the data governance council. The collaborative data stewards should be knowledgeable in more than one LOB as part of proposing solutions that are best for the enterprise. By promoting accountability for data as an enterprise asset and providing for efficient collaboration among stakeholders, the data governance council fosters an environment that ensures optimal mission performance. Even with the best of intentions, the data governance council may deadlock. In such cases, the collaborative steward must escalate the issues to the executive/strategic level.

Establishing a data governance council may be easy, but an effective council must be committed to collaboration. The role and responsibilities should be clear and focused to accomplish what is best for the enterprise. In some organizations, the council is composed of individuals from the LOBs, whereas in others, a separate independent group is established. Success with either approach depends on the organization.

Secure your belongings. Data security [4] protects data from unauthorized access, use, disclosure, and destruction, as well as the prevention of unwanted changes that can affect the integrity of data. Ensuring data security requires paying attention to physical security, network security, and security of computer systems and files. Data security is required to protect intellectual property rights, commercial interests, or to keep sensitive information safe. Security defines the methods of protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability, whether in storage or in transit. Confidentiality will prevent the disclosure of information to unauthorized individuals or systems. Integrity means that the data cannot be modified without authorization (i.e., integrity is violated when an individual accidentally or with malicious intent deletes important data files). Availability means that the information must be obtainable when a user requests the data. These three concepts are core principles of information security.

Data about data. Informative and relevant metadata (i.e., data about data) supports your organization and helps everyone that uses your data. A data steward, working under the direction of a data architect and a DBA, is usually responsible for managing a portion of the metadata. Metadata describes the definition, structure, and administration of information with all contents in context to ease the use of the captured and archived data for further use. The traditional data administration approach uses metadata to define data structures and relationships (e.g., data models) to support the development of databases and software applications. In addition to supporting systems development, metadata may be associated with all data in the enterprise for the purposes of"advertising" data assets for discovery. Organizations have to identify and document all data to facilitate its subsequent identification, proper management, and effective use, and to avoid collecting or purchasing the same data multiple times. There are many types of metadata, including vocabularies, taxonomic structures used for organizing data assets, interface specifications, and mapping tables.

Metadata management not only encapsulates basic data dictionary content but also ensures data's ongoing integrity. Metadata aids in the comprehension of the data to avoid making incorrect decisions based on their interpretations. Data lineage, the understanding of data from its inception to its current state, is a foundation capability of metadata management. As users reuse data from an original source system to the downstream support systems, they need to understand the lineage of that data. Data longevity is roughly proportional to the comprehensiveness of the metadata. For example, during an emergency event, it can be difficult to know where data is in order to assemble it expeditiously. Access to the data is critical when saving time means saving lives. Good metadata can help overcome the obstacles and get the right information into the hands of the right people as fast as possible.

Going to a better place. Data migration is the process of transferring data from one system to another. Migration includes the following steps:

- Identify the migrating legacy data and associated business rules.
- Map and match the legacy data to the target system.
- Aggregate, cleanse, and convert legacy data, as needed, to fit appropriately in the target system.
- Migrate the data in an appropriate sequence to the target system.

The most frequent challenges a data migration effort may face are an underestimation of the task and a postponement until the target system is almost ready to go operational. The complexity of a migration effort is in the implementation, and challenges exist at every step of the process. It is easy to reach Step 4 and discover that Step 1 is not complete. In some instances, legacy data cannot be migrated because it does not meet business rules in the target system and there may be a cascading effect on the cleansed data. Data cleansing is the process of detecting and correcting or removing corrupt or inaccurate records from a record set, table, or database.

Defining data elements and associated business rules can be a daunting exercise but a necessary task to ensure a successful migration effort. Legacy systems may not always document the data well, and business rules may not be fully enforced.

For example, the definition of an existing data element could change mid-stream and affect associated business rules. Mapping may be possible, but the business rules may differ significantly to render legacy data useless. A detailed data cleansing routine will ease the pain during the tedious process of weeding out duplicates and obsolete data, as well as correcting any errors in the data.

Finally—and this is a common mistake—never assume that the previous steps worked perfectly. Routines to cleanse, transform, and migrate the data have to be run several times and at times modified to ensure completeness. The best advice for data migration is to start early in the system migration process. Be prepared. Understand as much as possible what data is available (i.e., legacy system) and where data is moving (i.e., target system). Be patient, be flexible, and expect the unexpected.

Play nice in the sandbox. Information sharing is the exchange among individuals, organizations, systems, and databases across domains and organizational boundaries. The goal of information sharing is to provide the right data at the right place in order to support timely and effective decision making. Information-sharing solutions support the collection of data from enterprise systems and their assembly into concise, understandable, actionable, and when possible, unclassified formats. An organization can have an information-sharing culture that embraces the exchange of information and an information-sharing environment that includes policies, governance, procedures, and technologies that link resources (people, process, and technology) of stakeholders to facilitate information sharing, access, and collaboration. A mature organization will exhibit continual information sharing in a standardized manner with guaranteed data quality.

References & Resources

1. The Data Warehousing Institute (TDWI).
2. DAMA's Data Management Body of Knowledge (DMBOK).
3. Federal Data Architecture Subcommittee (DAS) Data Quality Framework, v1.0, October 2008.
4. National Institute of Standards and Technology (NIST), Information Security Handbook: A Guide for Managers Information Security, Special Publication (SP) 800-100, Revision 3 (March 2007).

[1] A database is a collection of related data. It may be stored in a single or several files. It may be structured or unstructured.

[2] A DBMS is software that controls the organization, creation, maintenance, retrieval, storage, and security of data in a database. Applications make requests to the DBMS, but they do not manipulate the data directly.