CDMP Fundamentals Notes

Data Strategy Professionals

Contents

Introduction to the CDMP Fundamentals Exam

Thank you for your purchase of the CDMP Fundamentals Notes. We hope you find this guide useful in your Data Strategy journey.

All CDMP exams are based on the ***Data Management Body of Knowledge*** **(*DMBOK*)**. The CDMP Fundamentals Notes walks you through the first 14 chapters of the *DMBOK* at the level of detail required to serve as an aid to your studying, helping you become familiar with these concepts more quickly.

The **Fundamentals Exam** is required for all CDMP certification levels. It consists of **100 questions** that you will have **90 minutes**[1] to answer. In addition to these Notes, Data Strategy Professionals offers a **Study Plan** that may assist with your preparation. It's available as an email series that's sent over the course of 90-days or all at once through the 'immediate access' option. You may also be interested in the **Guided Study Sessions** on each *DMBOK* chapter offered as part of the **Community & Events Membership**.

After completing the CDMP Fundamentals Exam, you'll receive your score immediately. If you scored **80% or above**, congrats, you're done! If you scored 60-69%, you're set for the **Associate certification**, but you'll need to retake the exam if you want to proceed to the **Practitioner** (at least 70%) and/or **Master** level (at least 80%).

Beyond helping you ace the Fundamentals Exam, the CDMP Fundamentals Notes are also useful for the **Specialist Exams**. Specialist Exams are a deep dive on a specific chapter of the *DMBOK*. After the Fundamentals Exam, you must take **two Specialist Exams** in order to gain recognition at the Practitioner or Master level. There are seven options for the Specialist Exams, and Data Strategy Professionals offers a guide to each one.

In addition to reading (or thoroughly skimming) the *DMBOK* before you take the Fundamentals and/or Specialist Exams, you may choose to study **additional reading materials**. We have a list of recommended reading on our website.

This study guide is **not a replacement for the *DMBOK***. We still recommend the purchase of this book and its use on the CDMP Fundamentals Exam and CDMP Specialist Exams. Note that only **one book** can be used on the CDMP Exam (we strongly recommend the *DMBOK*). You can use either the hardcopy or electronic version, but not both.

In terms of choosing which version to use, some members of the **CDMP Study Group** have enjoyed the ability to use ctrl + f to find information in the ebook during the exam. Either way, you're encouraged to take notes, highlight, and put sticky notes in your copy of the book.

Because CDMP exams are now either **open book or open notes**, if you chose to use the *DMBOK* as your one book, you cannot use this document as a reference during the test.

[1] If you purchase the English as a Second Language (ESL) version of the exam, you'll have 110 minutes to complete it. There's no downside to taking the ESL version, so you definitely should if English is not your native language.

Data Management

chapter 1 | page 17 | 2% of exam

Summary: An organization controls how it obtains and creates data. If data is viewed as an asset as well as a potential source of risk, then better decisions will be made throughout the data lifecycle. Data Management requires a collaborative approach to governance, architecture, modeling, and other functions.

Notes:
More data exists today than at any time in history, and understanding how to use data is key to an organization's success. Data Management allows an organization to capitalize on its competitive advantage.

Data Management refers to the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout the data lifecycle

Data lifecycle activities:
- Plan
- Design and Enable
- Create and Obtain
- Store and Maintain
- Use
- Enhance
- Dispose of

Here are some paradoxical, quasi-mystical statements from the *DMBOK* about the nature of data:
- Data is both an interpretation of the objects it represents and an object that must be interpreted
- Data is not consumed when it is used
- Data begets more data
- Data is easy to copy and transport, but it is not easy to reproduce if it is lost or destroyed
- Data can be stolen without being gone

Data Management is made more complicated by the fact that different types of data have different lifecycle management requirements

Types of data include:
- Transactional data
- Reference data
- Master data
- Metadata
- Category data
- Resource data
- Event data

Different departments may have different ways of representing the same concept – subtle or blatant differences can create significant challenges in managing data. This challenge is covered in Data Integration & Interoperability (ch. 8).

Reliable metadata (i.e., data about data) is required to manage the organization's data assets; types of metadata include:

- Business metadata
- Technical metadata
- Operational metadata
- Data Architecture metadata
- Data models
- Data security requirements
- Data integration standards
- Data operations processes

Data not only represents value; it also represents risk:

- Low quality data contributes to poor organizational decision-making
- Misunderstandings can have devastating consequences
- Regulatory processes govern all aspects of the data lifecycle
- Privacy is paramount for maintaining ethical use of data

Data Ethics

chapter 2 | page 49 | 2% of exam

Summary: Data Ethics refers to a set of standards to manage data assets properly. It includes data integrity standards and data privacy practices.

Notes:

Ethical principles for Data Management stem from the Belmont Principles (1979):

1. Respect for Persons
2. Beneficence
3. Justice

In 2015, the **E**uropean **D**ata **P**rotection **S**upervisor (EDPS) set out an opinion on digital ethics, specifically focused on Big Data. It called for:

- Future-oriented regulation of data processing and respect for the rights to privacy and to data protection
- Accountable controllers who determine personal information processing
- Privacy conscious engineering and design of data processing products and services
- Empowered individuals

The European Union set forth the **G**eneral **D**ata **P**rotection **R**egulation (GDPR) principles in 2016 and implemented the groundbreaking regulation in 2018. The principles of GDPR are as follows:

- Fairness, Lawfulness, Transparency
- Purpose Limitation
- Data Minimization
- Accuracy
- Storage Limitation
- Integrity and Confidentiality
- Accountability

Other Data Privacy legislation includes:

- Canada's **P**ersonal **I**nformation **P**rotection and **E**lectronic **D**ocuments **A**ct (PIPEDA) in 2000
- US's **F**ederal **T**rade **C**ommission (FTC) Privacy Program Criteria in 2012

Data Governance

chapter 3 | page 67 | 11% of exam

Summary: Data Governance refers to the exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets. It provides direction and oversight for Data Management by establishing a system of decision rights over data that accounts for the needs of the organization.

Notes:
Data Governance should include strategy, policy, standards and quality, oversight, compliance, issue management, Data Management projects, and data asset valuation. Change management is required for success.

Higher volumes of data are generated in the modern economy through:
- Increased proliferation of data-centric systems such as **c**ustomer **r**elationship **m**anagement (CRM)
- Widespread adoption of **i**nternet **o**f **t**hings (IOT) devices
- More time spent on digital platforms
- More value placed on digital platforms
- Information created by machine learning and generative AI

Data Governance is critical to an organization's success; it helps to…
- Produce trust in information through improved data quality
- Enhance understanding of the organization's data assets
- Create opportunities to generate insights through analytics and data science
- Promote regulatory compliance (i.e., keeping the CEO out of jail)
- Enable the organization to act on its competitive advantage
- Promote agility and ability to respond to change

Motivations for setting up Data Governance:
- Reactive (e.g., response to a data breach)
- Preemptive (e.g., in the face of major change or threats)
- Proactive (e.g., to improve capability to resolve risk and fix data issues)

Organizational structures for Data Governance:
- Steering Committee composed of executives and responsible for releasing funding
- Data Governance Council that manages initiative
- Data Governance Office that provides ongoing focus on enterprise-level data definitions and data management standards, and also functions as a **P**roject **M**anagement **O**ffice (PMO)
- Data Stewardship teams
- Communities of Interest focused on specific subject-areas or projects
- Local Data Governance Working Groups may be needed in a large organization; this refers to additional distributed teams at the department-level that may be set up on a short-term basis as needed

Data Governance process:
- Start with Data Governance Maturity Assessment that defines current state
- Move to Gap Assessment
- Build Data Governance policies
- Create a Roadmap with transition steps
- Work to maintain the target state

What makes for an effective Data Governance process?
- Sustainable: given that governance is a process and not a project with a defined end, it must be "sticky"
- Embedded: efforts need to be incorporated into existing aspects of the organization, including software development methods, data ownership, master data management, and risk management
- Measured: employing governance has a positive financial impact, but articulating this benefit requires understanding the baseline and capturing measurable improvements

Principles for getting more value from your organization's data:
- Data should be managed as a corporate asset
- Data management best practices should be incentivized across the organization
- Enterprise data strategy must be directly aligned with overall business strategy
- Data management processes should be continuously improved

Generally Accepted Information Principles can be used to calculate the value of information as an asset
1. Accountability
2. Asset
3. Audit
4. Due Diligence
5. Going Concern
6. Level of Valuation
7. Liability
8. Quality
9. Risk
10. Value

Key elements of Data Governance strategy:
- Charter
- Operating framework that lists accountabilities
- Implementation roadmap
- Vision of what constitutes operational success

Data Stewards should create a business glossary that associates terms (such as synonyms, metrics, lineage, business rules, owner, etc.) with important metadata

Given that it is a process and not a project with a defined end date, implementing Data Governance requires flexibility, such as…
- Updates to the mapping between Data Governance outcomes and business needs
- Continual adjustments to roadmap for creating Data Governance
- Ongoing enhancement of the business case for Data Governance
- Frequent assessment of Data Governance metrics

Data Governance is also a regulatory compliance project; it requires working with business and technical leadership to find the best answers to a standard set of regulatory compliance questions (such as how, why, when, etc.)

Primary focus of Data Governance must be on improving data (quality, accessibility, security, privacy, retention, etc.) over time, not simply monitoring for issues

Data Architecture

chapter 4 | page 97 | 6% of exam

Summary: Data Architecture presents data assets in a structured and easy-to-understand format. It defines the blueprint for managing data assets in alignment with organizational strategy through the establishment of strategic data requirements and the development of architectural designs to meet these requirements.

Notes:

Most organizations have more data than a single individual can comprehend — therefore, it's necessary to be able to represent data at various levels of abstraction so that it can be understood for decision-making purposes

Enterprise **D**ata **M**odel (EDM) is a holistic, enterprise-level, implementation-independent conceptual or logical data model providing a common consistent view of data across the enterprise. It is comprised of:
- Data names
- Data and metadata definitions
- Conceptual and logical entities and relationships
- Business rules

Data flow design is a diagram that defines requirements and master blueprint for storage and processing across databases, applications, platforms, and networks; it illustrates where data originated, where it's stored and used, and how it is transformed as it moves inside and between diverse processes and systems

CRUD stands for **c**reate, **r**ead, **u**pdate, **d**elete; referring to the four basic operations of persistent storage

Objectives of Data Architect:
- Enable consistent data standardization and integration across the organization
- Serve as a bridge between business strategy and technical execution
- Create and maintain organizational knowledge about data and systems
- Identify opportunities for data usage, cost reduction, and risk mitigation
- Iimplement and enforce semantics via common business vocabulary

Operational data enables up-time services for **i**nternet **o**f **t**hings (IOT) objects like manufacturing equipment, healthcare equipment, and consumer goods

Zachman Framework is an enterprise architectural framework from 1980s:
- Shows what models should exist, where each cell is a unique design artifact
- Reification transformation refers to the concept of changing an abstract idea into concrete instance (i.e., instantiation)

Architecture initiative requires: strategy, culture, organization, working methods, results

Outcomes:
- More accurate project data requirements
- Review project data designs
- Determine data lineage impact
- Data replication control
- Enforce data architecture standards
- Guide data technology and renewal decisions

Roadmap covers 3-5 years of the development path; move from least dependent at top of diagram to most dependent at bottom

Agile methodology requires learning, constructing, and testing in discrete delivery packages (called “sprints”) that are small enough that if work needs to be discarded, not much is lost

Lifecycle stages:
- Current
- Deployment period
- Strategic period
- Retirement
- Preferred
- Containment
- Emerging
- Reviewed

Data Architecture is required to…
- Oversee projects
- Manage architectural designs, lifecycles, and tools
- Define standards
- Create data-related artifacts

Metrics:
- Architecture standard compliance rate
- Implementation trends:
  - Use / reuse / replace / retire measurements
  - Project execution efficiency measurements
- Business value measurements
- Business agility improvements
- Business quality
- Business operation quality
- Business environment improvements

Data Modeling & Design

chapter 5 | page 121 | 11% of exam

Summary: Data Modeling is the process of taking unstructured information and turning it into structured information through the creation of Conceptual, Logical, and Physical Data Models.

Notes:
Move from Conceptual (taxonomy) to Logical (entity-relationship diagram) to Physical Model (plan for storage and operations)

Conceptual Model may start off looking like a business glossary:
- Should incorporate enterprise terminology
- Build by connecting nouns (entities) and verbs (relationships)

Logical Model captures detailed data requirements:
- Analyze information requirements and existing documentation
- Add associative entities
- Add attributes
- Assign domains
- Assign keys

Physical Model outlines how data will be stored in enterprise systems:
- Resolve logical abstractions
- Add attributes
- Add reference data objects
- Assign surrogate keys
- Denormalize for performance
- Index for performance
- Partition for performance
- Create views

Once database architecture is complete, compare a reverse-engineered version of the Physical Model to the Conceptual Model to ensure initial business requirements have been met

PRISM represents best practices in database design:
- **P**erformance and ease of use
- **R**eusability
- **I**ntegrity
- **S**ecurity
- **M**aintainability

Data Modeling schemes:
- Relational (Information Engineering diagram uses 'crow's feet' to depict cardinality): capture business rules
- Dimensional (Fact and Dimension tables): capture navigation paths required to answer business questions
- Object Oriented (**U**nified **M**odeling **L**anguage [UML]): classes, attributes, operations
- Fact Based
- Time Based

— NoSQL

Mid to large organizations usually have an application landscape with multiple schemes and models evolved over time

Cardinality refers to the relationships between the data in two database tables; defines how many instances of one entity are related to instances of another entity (e.g., zero, one, many)

Unary relationship involves only one entity (i.e., multiple instances of the same type); for example, the relationship between a pre-requisite and an academic course (both are courses); it is also known as a recursive or self-referencing relationship

When diagramming, rectangular boxes are used to represent primary key entities, and rounded boxes are used for foreign key entities

Construction keys:
— Surrogate: simple counter that provides unique id within a table
— Compound: 2+ attributes to uniquely id instance (e.g., phone number is composed of area code + exchange + local number)
— Composite: compound key + simple or compound key

Function keys:
— Super key: any set of attributes that uniquely id an entity instance
— Candidate key: a minimal set of one or more attributes that id entity instance
— Natural key: business key
— Primary key: candidate key chosen as unique id (versus alternate keys)

Often, primary key is surrogate key, and alternate keys are business keys

When primary key of parent is migrated as foreign key to child's primary key, this is known as the identifying relationship

Domain refers to the complete set of possible values an attribute can be assigned; it can be restrict with constraints (i.e., rules on format and/or logic), and is typically characterized in the following ways:
— Data type (e.g., Character(30))
— Data format (e.g., phone number template)
— List
— Range
— Rule-based (e.g., ItemPrice > ItemCost)

Dimensional modeling fact table:
— Rows correspond to particular measurements and are numeric
— 90% of contents of database
— Many rows

Dimensional modeling dimensions table:
— Mostly textual descriptions
— Must have a unique identifier for each row (surrogate or natural key)

Options to change dimensions:
1. Overwrite

2. New row – old row marked as not current
3. New column

Star schema dimensional model is denormalized ("collapsed")

Denormalize to…
- Combine data and avoid run-time joins
- Create smaller, pre-filtered copies of data to reduce table scans of large tables
- Pre-calculate and store expensive calcs

Denormalization introduces risk of errors due to duplication

Snowflaking refers to the normalization of a dimensional model; this is not recommended because it degrades performance

Grain refers to the level of detail in a dataset

In Kimball data model, conformed facts / dimensions are built with entire organization's needs in mind; they are standardized definitions, and can be used across data marts

Unified **M**odeling **L**anguage (UML) is a graphical language for modeling software

Fact-based modeling is based on forming plausible sentences business person might use; it's also referred to as **O**bject-**R**ole **M**odeling (ORM)

Data vault is a type of time based model where a normalized data store is composed of hubs, links, and satellites

Partitioning refers to splitting a table to facilitate archiving or to improve retrieval performance – can be either vertical or horizontal

Create data lineage through source-target mapping

Canonical model used for data in motion between systems; it describes structure sending and receiving services should use, a process described in more detail in Data Integration and Interoperability (ch. 8)

Steps to build a data model:
- Select scheme and notation
- Gather entities and relationships
- Utilize business specific terminology (i.e., from data glossary)
- Obtain signoff

Data Storage & Operations

chapter 6 | page 165 | 6% of exam

Summary: The focus of Data Storage & Operations is to maintain data integrity and ensure availability throughout the operational lifecycle. Data Storage encapsulates the design, implementation, and support of stored data to maximize its value. Data Operations provides support throughout the data lifecycle from planning for collection to designing appropriate strategies for the disposal of data.

Notes:

Goals for Data Storage & Operations:
- Manage availability of data through lifecycle
- Ensure integrity
- Manage performance of data transactions

Best practices:
- Automation opportunities
- Build with reuse in mind
- Connect database standards to support requirements
- Set expectations for **d**ata**b**ase **a**dministrators (DBAs) in project work

Information lifecycle:
- Plan: governance, policies, procedures
- Specify: architecture (conceptual, logical, physical modeling)
- Enable: install / provision servers, networks, storage, databases; put access controls into place
- Create & Acquire
- Maintain & Use: validate, edit, cleanse, transform, review, report, analyze
- Archive & Retrieve
- Purge

Activities:
- Database support
  - Performance tuning, monitoring, error reporting
  - Failover for 24/7 data availability
  - Backup and recovery
  - Archiving
- Data technology management

Database support:
- Implement and control database environment
- Acquire externally sourced data (optional); metadata very important
  - Marketing and demographics
  - Industry standards
  - Elections data
  - Geographic / geospatial data (e.g., images, infrared, etc.)
  - Dunning & Bradstreet company hierarchies
    - Linkage refers to the relationship between different active business entities or specific sites within a corporate family
    - Linkage occurs in Dunning & Bradstreet's database when one business location has financial and legal responsibility for another business location

    - Percentage of financial and legal responsibility determines the type of linkage relationship
    - 19.7M active records in Dun & Bradstreet's global database
- Plan for data recovery: backup and recover data
    - Database backup schedule
    - Maintain logs
    - Provide continuity of data to the organization
- Set database performance service levels
- Monitor and tune database performance
- Archive, retrieve, and purge data
    - Test that archives can be retrieved
    - Just clicking delete isn't sufficient; need to follow processes to ensure data is actually removed
- Manage specialized databases
    - Geospatial, graph, **c**omputer-**a**ided **d**esign (CAD), E**x**tensible **M**arkup **L**anguage (XML), object etc.

Data technology management:
- Understand requirements
- Define database architecture
- Evaluate data technology
- Install and administer data tech
- Inventory and track tech licenses
- Support data tech usage and issues
- Set up and maintain **I**nformation **T**echnology **I**nfrastructure **L**ibrary (ITIL)

Options for selection of tools:
- Enterprise: swiss army knife, multiple capabilities,
- Best of breed: best fit per discipline after thoroughly evaluating tools
- Tactical: best quick fit for the job (quick and dirty) – just in time – meets immediate needs – use what you know

ACID for achieving reliability within database transactions:
- **A**tomicity: all operations are performed, or none of them are, so that if one part of the transaction fails, then the entire transaction fails
- **C**onsistency
- **I**solation: each transaction is independent
- **D**urability

BASE as another approach to reliability:
- **B**asically **A**vailable
- **S**oft state: data in a constant state of flux
- **E**ventual consistency

CAP theorem (i.e., Brewer's theorem) – data practitioners are forced to pick two:
- **C**onsistency
- **A**vailability
- **P**artition tolerance

Lambda architecture uses two paths for data:
- Speed path where availability and partition tolerance are most important
- Batch path where consistency and availability are most important

Open **D**ata**b**ase **C**onnectivity (ODBC) is an **a**pplication **p**rogramming **i**nterface (API) that enables database abstraction

Clustering refers to the practice of combining more than one servers or instances connecting a single database

Columnar storage reduces **i**nput / **o**utput (I/O) bandwidth by storing column data using compression

Environments:
- Prod
- Pre-Prod
  - Test
    - **Q**uality **A**ssurance (QA)
    - Integration
    - **U**ser **A**cceptance **T**esting (UAT) with realistic use cases
    - Performance: high volume / high complexity
  - Development

Sandbox allows only read-only connection to production; it is used for experiments by users (not DBAs)

Replication through:
- Mirroring: updates to the primary database are replicated immediately to the secondary database as part of a two-phase commit process
- Log shipping: a secondary server receives and applies copies of the primary database transaction logs at regular intervals

Sharding refers to the process by which small chunks of the database are isolated so that they can be can be updated independently of other shards; replication is merely a file copy

Data**b**ase **A**dministator (DBA) role:
- Level 2 technical support: working with help desks and tech vendor support
- Working knowledge of data modeling, use case analysis, analysis
- Backups and recovery tests
- Data integration
- Test deployment into pre-prod database
- Document processes and procedures
- Predict ebbs and flows of usage patterns and put into place processes to handle high traffic and take advantage of low traffic (e.g., for complex / high resource activities)
- Define recovery plan
- Manage the physical storage environment using **s**oftware **c**onfiguration **m**anagement (SCM) processes to record config status
  - Configure identification
  - Configure change control
  - Configure status accounting
  - Configure audits
- Manage access controls:
  - Environment (e.g., firewalls, network health, patch management, etc.)
  - Physical security (e.g., data audit logging, disaster management, database backup planning)
  - Monitoring: continuous hardware and software monitoring of critical servers
  - Controls: access, auditing, intrusion detection, and vulnerability assessment tools
- Implement physical data model

- Load data
- Set up replication
- Manage performance: availability and speed
  - Set **s**ervice **l**evel **a**greements (SLAs) with IT and business users
  - Availability
    - Manageability
    - Recoverability
    - Reliability
    - Serviceability
- Issue remediation
- Maintain dev and test environments, sandboxes, offline backups, failover, resiliency support systems
- Manage datasets for testing software applications
- Manage data migration

Hot backup is a backup taken while database is running

Physical Data Model includes storage objects, indexing objects, and any encapsulated code objects

required to enforce data quality rules, connect database objects, and achieve performance

Data Security

chapter 7 | page 209 | 6% of exam

Summary: Data Security refers to the set of policies and procedures designed to reduce legal and/or financial risks and to grow and protect the business. It ensures that data privacy and confidentiality are maintained, that data is not breached, and that data is accessed appropriately.

Notes:

Security motivations:
- Protect stakeholders (e.g., clients, patients, employees, suppliers, partners, etc.)
- Comply with government regulations
- Protect proprietary business concerns (to protect business competitive advantage)
- Provide legitimate data access
- Meet contractual obligations (e.g., **P**ayment **C**ard **I**ndustry [PCI] Standard mandates encryption of user passwords, etc.)

Good security responsible for risk reduction and business growth

Steps:
- Identify and classify sensitive data assets depending on industry and organization
- Locate sensitive data throughout the enterprise
- Determine how each asset needs to be protected
- Identify how information interacts with business processes

Security-related metadata increases quality of transactions, reporting, and business analysis

Prioritize risks based on combination of:
- Potential severity of damage to the organization
- Likelihood of occurrence

Classify risk level associated with data as critical, high, or moderate

Large businesses may have a **C**hief **I**nformation **S**ecurity **O**fficer (CISO) who reports to CIO or CEO

National **I**nstitute of **S**tandards and **T**echnology (NIST) provides a Risk Management Framework that categorizes all enterprise information to locate sensitive info

Four As of data security:
1. Access – actively connect to an information system
2. Audit – review of security actions and user activity to ensure compliance with regulations and conformance with company policy and standards
3. Authentication – validate user access
4. Authorization – grant individual privileges to access specific views of data as appropriate to role

Sarbanes-Oxley regulations are mostly concerned with protecting financial information integrity by identifying rules for how financial information can be created and edited.

Methods of encryption:
- Hash: algorithm that converts data into math
- Symmetric / private-key: sender and recipient have key to read original data

- Public-key: sender and recipient have different keys; sender uses public key that is freely available and receiver uses a private key to reveal original data (e.g., for a clearinghouse)
- Obfuscation or masking

Data-in-motion requires network protection:
- Firewall is insufficient
- Each machine on network requires line of defense
- Webservers require particularly sophisticated protection

Backdoor refers to an overlooked or hidden entry into a computer system or application (e.g., accidentally keeping default password)

Bot / zombie is a workstation that's been taken over by trojan, virus, phish

Cookie refers to small data file that website installs on a computer's hard drive to id returning visitors and provides their preferences; often used for Internet commerce

Firewall is software or hardware that filters network traffic to protect an individual computer or an entire network from authorized attempts to access or attack the system; may scan both incoming and outgoing communications for restricted or regulated info to prevent it from passing without permission (i.e., data loss prevention)

De**m**ilitarized **Z**one (DMZ) is an area between two firewalls that is used to pass or temporarily store data between organizations

Super User Account has administrator or root access to a system to be used only in an emergency

Virtual **p**rivate **n**etwork (VPN) connection creates an encrypted tunnel into organization's environment, allowing communication between users and internal network

Confidentiality levels:
- For general audiences
- Internal use only
- Confidential
- Restricted confidential
- Registered confidential

Family **E**ducational **R**ights and **P**rivacy **A**ct (FERPA) protects educational records

Implement principle of least privilege

Vulnerabilities:
- Abuse of excessive privilege: user with privileges that exceed the requirements of their job
  - Query-level access control restricts database privileges to minimum-required SQL operations and data (triggers, row-level security, table security, views)
- Abuse of legitimate privilege (e.g., healthcare worker prying into patient records)
- Typically, apps restrict viewer to accessing one record at a time
- Unauthorized privilege elevation (i.e., taking on privileges of administrator)
  - Vulnerabilities may occur in stored procedures, built-in functions, protocol implementations, and SQL statements
  - **I**ntrusion **P**revention **S**ystems (IPS)

    - Query-level access control intrusion prevention
    - Inspect database traffic to id patterns that correspond to known vulnerabilities
- Service accounts (i.e., batch IDs for specific processes) and shared accounts (i.e., generic IDs created when an app can't handle total user accounts) create risk of data security breach, complicating ability to trace breach to source
- Platform intrusion attacks
    - Software updates (i.e., patches)
    - Implementation of **I**ntrusion **P**revention **S**ystem (IPS) and **I**ntrusion **D**etection **S**ystem (IDS)
- SQL injection attack: attacker inserts unauthorized statements into vulnerable SQL data channel (e.g., stored procedures or web application input spaces); execution in database provides attacker unrestricted access to database
    - To prevent, sanitize all inputs before passing them to server
- Change default passwords
- Encrypt backup data
- Social engineering / phishing
- Malware: malicious software (including viruses, worms, spyware, key loggers, adware)
    - Adware: from download, captures buying behaviors to sell to marketing firms or for id theft
    - Spyware: can store credit card info, etc.
    - Trojan horse: destructive program
    - Virus: attaches to an executable or vulnerable app
    - Worm: built to reproduce and spread across network to send out a continuous stream of infected messages

Track metadata to avoid security risks

Sanitize documents before sharing to avoid sharing confidential information

Data Integration & Interoperability

chapter 8 | page 257 | 6% of exam

Summary: Data Integration refers to the process of merging data from various datasets into unified data using both technical and business processes. This process improves communication and efficiency in an organization. Data Interoperability refers to the process of designing data systems so that data will be easy to integrate. These fields involve processes related to the movement and consolidation of data within and between data stores, applications, and organizations.

Notes:

Fundamental concepts:
- Integration: data exchange; process of sending and receiving data
- Interoperability: data sharing; includes metadata
- Hub: system of record

The Mars Climate Orbiter (1989) represents an example of Data Integration gone wrong. The mission failed due to a navigation error caused by a failure to translate English units to metric. Commands from Earth were sent in English units (in this case, pound-seconds) without being converted into the metric standard (Newton-seconds).

Point-to-point approach to Data Integration:
- Build an interface between system A and B, system A and C, system B and C etc.
- Number of interfaces = (n-1) !
- Will provide highest potential performance
- Good for getting things done quickly without procurement
- Short life expectancy
- Doesn't scale because custom coding is required
- Support is difficult

Hub approach to Data Integration:
- Data updates into a central repo which distributes data into authorized applications
- Commonly used for Master & Reference Data Management
- Keeping a copy in the repo is a potential security or regulatory risk
- Latency and slow performance

Bus Distribution approach to Data Integration:
- Enterprise Service Bus: cannot be used as a record of reference because data is not stored, rather this is a data transportation system
- **S**ervices **o**riented **a**rchitecture (SOA)
- One app "pushes" data into central service which then pushes it to authorized applications that can "pull" the data
- Publish-subscribe model
- Routing model "broker"
- Scalable

Extract **T**ransform **L**oad (ETL), **E**xtract **L**oad **T**ransform (ELT), and **C**hange **D**ata **C**apture (CDC) approach to Data Integration:
- ETL and ELT about batch distribution: scheduling, parallel processing, complex data transformation, cross reference, and data mapping

    - Typically run overnight
- CDC is event driven and delivers real time incremental replication
- Data moves from database to database

Message Synchronization and Propagation approach to Data Integration:
- **E**nterprise **A**pplication **I**ntegration (EAI): involves the use of hub
- **E**nterprise **S**ervice **B**us (ESB): involves the use of bus
- Event-driven business process automation
- May be loosely or tightly coupled
    - Tightly coupled: processes talk directly to each other
    - Loosely coupled: processes talk through **a**pplication **p**rogramming **i**nterface (API) services
- Data moves from application to application

Abstraction / Virtual Consolidation approach to Data Integration:
- Similar to database views: consuming applications see data as though in their own systems
- **E**nterprise **I**nformation **I**ntegration (EII)
- No need to touch underlying source data
- Data federation from database to application

Document & Content Management

chapter 9 | page 287 | 6% of exam

Summary: Document & Content Management promotes efficient asset retrieval from various platforms and systems, ensures the availability of semi-structured and unstructured data assets, and aids in compliance and audit practices. This field includes planning, implementation, and control activities used to manage the lifecycle of data and information in a range of unstructured media, especially documents needed to support legal and regulatory compliance requirements.

Notes:
Controlled vocabularies are a type of Reference Data (ch. 10) and records are a subset of documents

Record referns to evidence that actions were taken and decisions were made in keeping with procedures

Folksonomy is a classification scheme obtained through social tagging

Information architecture:
- Controlled vocabularies
- Taxonomies and ontologies
- Navigation maps
- Metadata maps
- Search functionality and specifications
- Use cases
- User flows

Semantic modeling is a type of knowledge modeling that describes a network of concepts and their relationships

Policies:
- Scope and compliance with audits
- Identification and protection of records
- Purpose and schedule for retaining records
- How to respond to information hold orders
- Requirements for onsite and offsite storage
- Use and maintenance of hard drive and shared network drives
- Email management, addressed from content management perspective
- Proper destruction methods for records

E**x**tensible **M**arkup **L**anguage (XML) represents both structured and unstructured data
- **R**esource **D**escription **F**ramework (RDF): standard model for data interchange on the web
  - SPARQL: used for semantic querying
  - **S**imple **K**nowledge **O**rganization **S**ystem (SKOS)
  - OWL (W3C Web Ontology Language): vocabulary extension of RDF; used when information contained in documents needs to be processed by application

Organization should set up an **E**nterprise **C**ontent **M**odel (ECM)

E-discovery is the process of finding electronic records that might serve as evidence in a legal action

Master & Reference Data Management

chapter 10 | page 327 | 10% of exam

Summary: Master & Reference Data Management supports the organization of enterprise-level data through ongoing reconciliation and maintenance of core critical shared data that is used to enable consistency across systems. Master & Reference Data should represent the most accurate, timely, and relevant information about essential business entities. As such, Master & Reference Data should be considered infrastructure for the organization.

Notes:

Master **D**ata M**a**nagement (MDM) provides control over master values and identifiers that enable consistent use across systems; it provides a single version of customers, accounts, materials, products, etc.

Master Data provides control over domain values and definitions, which may include:
- Codes and descriptions
- Classifications
- Mappings
- Hierarchies

Master Data provides a unified view of important entities such as
- Products
- Vendors / suppliers
- Business units
- Aspects of the legal structure
- Aspects of the financial Structure
- Locations

Terms associated with Master Data Management: golden record, system of truth, master values

Terms associated with Reference Data: list of values, taxonomy, cross reference

Both Master Data and Reference Data are forms of Data Integration (ch. 8)

Some general facts about Master & Reference Data:
- Reference Data can be a subset of Master Data
- Both Reference and Master Data can provide context for transaction data
- Reference Data is typically smaller than Master Data
- Master Data reduces risk that might be associated with ambiguous identifiers
- Master Data requires a trusted version of truth for each instance of conceptual entities
- Both Master and Reference Data should be shared at the enterprise level
- Reference Data typically comes from outside the organization

Golden Record encompasses data from multiple source systems, matching and merging processes to formulate the final “record”

MDM key processing steps:
- Data model management
- Data acquisition
- Data validation, standardization, and enrichment
- Entity resolution and stakeholder management
- Data sharing and stewardship

Master Data hub manages interaction with spokes such as source systems, data stores, etc.

Approaches to data sharing:
- Registry: index that points to Master Data in systems of record, where Master Data is managed
- Transaction Hub: application interface with hub to access and update Master Data; in this system Master Data exists only in Transaction Hub, which is the system of record
- Consolidated: hybrid where systems of record manage Master Data local to their applications and then Master Data is consolidated and made available from a data sharing hub, which forms the system of reference for Master Data

Metrics:
- Data Quality and compliance
- Data change activity
- Data ingestion and consumption
- **S**ervice **l**evel **a**greements (SLAs)
- Data Steward coverage
- **T**otal **c**ost of **o**wnership (TCO)
- Data sharing volume and usage

Data Warehousing & Business Intelligence

chapter 11 | page 359 | 10% of exam

Summary: Data Warehousing & Business Intelligence involves the planning, implementation, and control processes to manage decision support and to enable knowledge workers to get value from data through analysis and reporting. The Data Warehouse stores data from various databases and supports strategic decisions.

Notes:

Warehousing:
- Stores data from other systems
- Storage includes organization that increases value
- Makes data accessible and usable for analysis

Two main Data Warehousing methodologies:
- Inmon: single data warehouse layer with atomic level data
- Kimball: departmental data marts informing conformed dimensions and facts

Inmon's **C**orporate **I**nformation **F**actory (CIF) is a normalized relational model:
- Subject oriented: based on business entities
- Integrated: unified and cohesive, consistently structured
- Time variant: records are like snapshots
- Non-volatile: new data appended to existing data
- Includes atomized and aggregated data
- Historical

Kimball's Dimensional Model (i.e., star schema):
- Composed of facts (quantitative) and dimensions (descriptive)
- Fact table joins with many dimension tables, forming a star

Components of Kimball Data Warehouse:
- Operational source systems: operational / transactional applications within the enterprise create the data that is integrated into the **O**perational **D**ata **S**tore (ODS) and then into data warehouse
- Data staging area: clean, combine, standardize, conform dimensions, sort, and sequence
- Data presentation area: datamarts linked by Data Warehouse bus of conformed dimensions
- Data access tools that focus on end users' data requirements

Kimball's Data Warehouse Bus represents shared or conformed dimensions unifying multiple datamarts

Bus-matrix is a table of business processes against subject areas

Data Storage areas:
- Staging area: intermediate data store between original data source and centralized data repository; data is staged before transformation, integration, and prep for loading into warehouse
- Reference and Master Data conformed dimensions
- Central Warehouse: maintains historical atomic data as well as latest instance of batch run; considerations include:
  - Relationship between business key and surrogate keys for performance
  - Creation of indices and foreign keys to support dimensions
  - **C**hange **d**ata **c**apture (CDC) techniques used to detect, maintain, and store history

- **O**perational **D**ata **S**tore (ODS): lower latency for operational use; single time window
- Data mart: presents a departmental or functional subset of data warehouse
- Cubes: support **O**n**l**ine **A**nalytical **P**rocessing (OLAP); can be relational, multi-dimensional, or hybrid

OLAP is composed of a server component and client-facing component on desktop or web

Update methods:
- Trickle feeds: source accumulation
- Messaging: bus accumulation
- Streaming: target accumulation

Implementation considerations:
- Conceptual data model
- Data Quality feedback loop
- End-to-end metadata
- End-to-end verifiable data lineage

Metrics:
- Usage
- Subject area coverage percentages
- Response and performance

Metadata Management

chapter 12 | page 393 | 11% of exam

Summary: Metadata Management refers to the process of ensuring the quality of metadata (i.e., data about data). This work involves planning, implementation, and control activities to enable access to high quality, integrated metadata, including definitions, models, data flows, and other information critical to understanding data and the systems through which it is created, maintained, and accessed.

Notes:

Metadata is data about data:
- Info about technical and business processes
- Data rules and constraints
- Logical and physical data structures

Metadata describes the data itself as well as concepts the data represents, and connections between data and concepts; it's important for the organization to standardize access to metadata

Types of metadata:
- Business / Descriptive (e.g., title, owner, business area)
- Technical / Structural (e.g., number of rows / columns)
- Operational / Administrative (e.g., version number, archive date, **s**ervice **l**evel **a**greement [SLA] requirements)

Storing metadata:
- Metadata repo
- Business glossary
- Business intelligence
- **C**onfiguration **m**anagement **d**ata**b**ase (CMDB) for IT assets
- Data dictionary
- Data integration tools
- Database management / system catalogs
- Data mapping mgmt tools
- Data Quality tools
- Event messaging tools: move data between diverse systems
- Modeling tools
- Master & Reference Data repos
- Service registries: **s**ervice-**o**riented **a**rchitecture (SOA) perspective enables reuse of services

Metadata storage architecture:
- Centralized
- Distributed
- Bi-directional

Activities:
- Readiness / risk assessment
- Cultural analysis and change management
- Create metadata governance
- Data lineage
- Impact analysis
- Apply tags when ingesting data into a data lake

Some less obvious business drivers of Metadata Management:
- Provide context to increase confidence
- Make it easier to identify redundant data and processes
- Prevent the use of out of date or improper data
- Reduce data-oriented research time
- Improve communications between data consumers and IT
- Create accurate impact analysis
- Reduce training costs associated with data use by improving documentation of data context, history, and origin
- Support regulatory compliance

Activities related to Metadata Management:
- Collect and integrate from diverse sources
- Provide standard way to access
- Ensure metadata quality and security

Other key terminology:
- Application metadata repo: where data is physically stored in the organization
- Business glossary: documents with concepts, definitions, relationships, and terminology
- Data dictionary (or catalog): structure and contents of datasets
- Data Integration tools: move data from one module / system to another with executables / **a**pplication **p**rogramming **i**nterfaces (APIs)
- Database management: content of databases, plus size, software version, deployment status, network uptime requirements, availability requirements, etc.
- Data mapping management tools: used during analysis and design phase of project to transform requirements into mapping specifications
- Data Quality tools: use of validation rules to assess quality
- Directories and catalogs: systems, sources, and location of data
- Event messaging tools: data movement
- Modeling tools and repos
- Reference data repositories
- Service registries: technical info about services and end-points

Architecture can be centralized, distributed, hybrid, or bi-directional

“Metadata guides the use of data assets. It supports business intelligence, business decisions, and business semantics.”

Dealing with data lakes:
- Metadata tags should be applied upon ingestion into a data lake

Data profiling to identify domains, relationships, and data quality issues

Data Quality

chapter 13 | page 423 | 11% of exam

Summary: Data Quality assures that data is fit for consumption and meets the organization's needs. Quality management techniques should be applied in the planning, implementing, and controlling stages to measure, assess, and improve the degree to which data is fit for use within an organization.

Notes:

Fundamental frameworks:
- Strong-Wang framework – Intrinsic, Contextual, Representational, Accessibility
- Shewhart / Deming cycle – "plan, do, check, act"

Data Quality issues caused by system design:
- Failure to enforce referential integrity
- Failure to enforce uniqueness constraints
- Coding inaccuracies and gaps
- Data model inaccuracies
- Field overloading
- Temporal data mismatches
- Weak **M**aster **D**ata **M**anagement (MDM)
- Data duplication

Dimensions of Data Quality:
- Completeness
- Uniqueness
- Timeliness
- Validity
- Accuracy
- Consistency

Activities:
- Maturity assessment
- Profiling

Data profiling is a form of data analysis used to inspect data and assess quality using statistical techniques:
- Count of nulls
- Min / max value
- Min / max length
- Frequency distribution
- Data type and format

Profiling also includes cross-column analysis to identify overlapping or duplicate columns and expose embedded value dependencies
- Inter-table analysis explores overlapping value sets and helps identify foreign key relationships

Data enhancement:
- Date / time stamps
- Audit data (e.g., data lineage)
- Reference vocabularies (i.e., business specific terminology, ontologies, and glossaries)

- Contextual information (i.e., adding context such as location, environment, or access methods, and tagging data for review and analysis)
- Geographic information (e.g., geocoding)
- Demographic information
- Psychographic information (i.e., customer segmentation based on behaviors and preferences)
- Valuation information (e.g., asset valuation, inventory, and sale)

Data parsing used to analyze data using predetermined rules to define content or value

Data Quality initiative requires identifying and prioritizing potential improvements

Provide continuous monitoring by incorporating control and measurement processes into information processing flow

Data Quality incident tracking requires staff to be trained on how issues should be classified, logged, and tracked for root cause remediation

Data Quality reporting:
- Data Quality scorecard
- Data Quality trends across the organization
- **S**ervice **L**evel **A**greement (SLA) metrics
- Data Quality issue management
- Conformance to Data Governance policies
- Impact of improvement projects

Preventative actions:
- Establish data entry controls
- Train data producers
- Define and enforce rules
- Demand high quality data from data suppliers
- Implement governance and stewardship
- Institute formal change control

Corrective actions:
- Automated
- Manually-directed
- Manual

Big Data

chapter 14 | page 469 | 2% of exam

Summary: Big Data refers to advanced analytics, data mining, and data science.

Notes:
Extract **L**oad **T**ransform (ELT) is typically used for data lakes; metadata is particularly valuable

Abate Information Triangle shows context added to data and distinction between Business Intelligence and Data Science

The Vs of Big Data: volume, velocity, variety, viscosity (i.e., how difficult to integrate), volatility, veracity

Services-**b**ased **a**rchitecture (SBA) to provide immediate data:
- Batch layer: data lake that contains both recent and historical data
- Speed layer: contains only real time data; **o**perational **d**ata **s**tore (ODS)
- Serving layer: interface to join data from batch and speed layers

Data mining reveals patterns in data through algorithms:
- Feature selection
- Correlation analysis
- Clustering
- Dimensionality reduction

Data Science:
- Predictive analytics based in probability estimates
- Smooth data with a moving average after regression analysis
- Use unsupervised learning to tag unstructured data

Operational analytics:
- Segmentation
- Sentiment analysis
- Geocoding
- Psychological profiling

Technology enabling data science: Moore's law, **G**raphical **P**rocessing **U**nits (GPUs), hand-held devices, **i**nternet **o**f **t**hings (IOT)

Massively **P**arallel **P**rocessing (MPP) for analyzing huge volumes of information
- Data is partitioned (logically distributed) across multiple processing servers (computational nodes) with each server having its own dedicated memory to process data locally

Hadoop as low cost storage platform for a variety of data types

Considerations:
- Relevance
- Readiness
- Economic viability
- Prototype

Next Steps

Congrats on finishing your review of the **CDMP Fundamentals Notes**! As a next step, we suggest purchasing the **CDMP Fundamentals Exam** if you've not done so already. When you do, you'll get access to the **official CDMP test bank** of 200 practice questions. You'll also receive a **free three year membership to DAMA International**.

If you'd like additional structure for your studies, Data Strategy Professionals offers a **CDMP Study Plan** that can be sent to you as emails over the course of 90 days or all at once through the 'immediate access' option. If you'd like **additional practice questions**, you can purchase those here. You may also be interested in the **Guided Study Sessions** on each *DMBOK* chapter offered as part of the **Community & Events Membership.**

You have an unlimited amount of time between purchasing a CDMP exam and actually taking the test. When you feel ready, you can take the CDMP Fundamentals Exam using **Google Chrome**. Your test will be proctored via the **Honorlock browser system**. Make sure to have your copy of the *DMBOK* close at hand given that the exam is **open book**.

You'll receive your score immediately after completing the exam. If you scored a 60-69%, you're set for the Associate certification, but you'll need to retake the exam if you want to proceed to the Practitioner (70%+) and/or Masters (80%+) levels.

The CDMP awards badges through the **openbadges standard** at badgr.com. You can share this credential through LinkedIn and other social platforms of your choosing.

If you do choose to proceed with the Specialist Exams, make sure you **sign up using the same email** you used for the Fundamentals Exam. You won't automatically receive your certification unless you took the three required exams in the same Canvas account.

To activate your free three year membership to DAMA International, contact their team at **cdmp@dama.org** at the **end of the month** following your purchase of the CDMP Fundamentals Exam. Provide your order number and date, and they will activate your membership for you.

If you have any issues or remaining questions, you can contact DAMA (the organization that runs the CDMP exams) here.