Applied A.I. Solutions

Foundations of Data Management

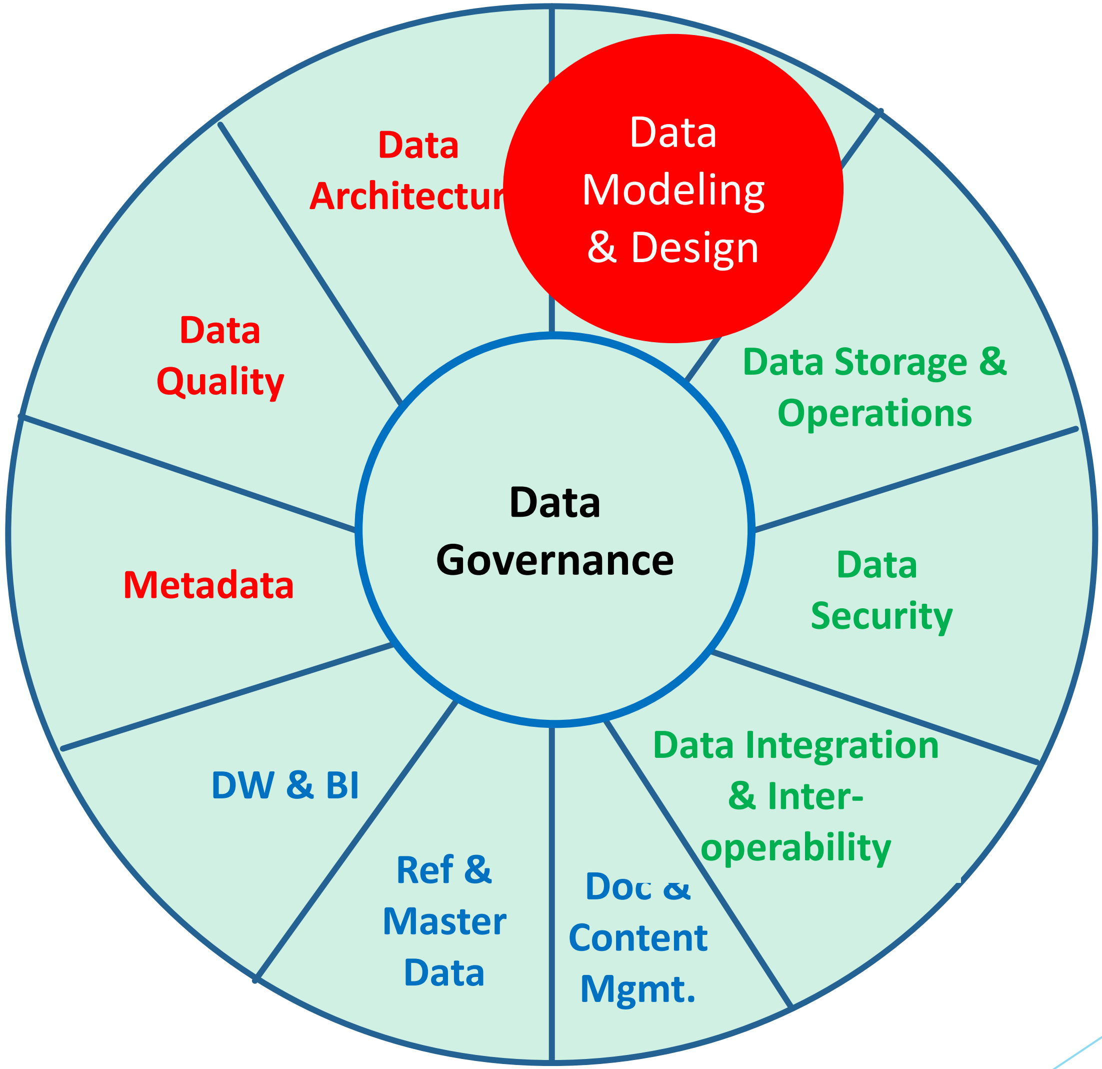
Professor

Daniel Vitaver-Bronstein, B.Sc., EMBA

daniel.vitaver-bronstein@georgebrown.ca

DATA MODELING AND DESIGN

The DAMA Wheel



[1] Main source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

1. INTRODUCTION

- Data Models are a critical component of DM and enable organizations to understand its data assets
- Common schemes are: Relational, Dimensional, Object-Oriented, Fact-Based, Time-Based, and NoSQL
- Model exist at a conceptual, logical, and physical level. Each model contains a set of components
- Components are entities, relationships, facts, keys, and attributes

[1] Main source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

DM&D Framework

Definition

Data Modeling is the process of discovering, analyzing, and scoping data requirements, and then representing and communicating these data requirements in a precise form called the Data Model.

[1] Main source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Goals

- To confirm and documents an understanding of different perspectives, which leads to applications that align with business requirements
- To create a foundation to initiatives such as MDM, DG programs

Business Drivers

[1] Main source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

The framework is guided by the following principles

- Formalization
- Scope Definition
- Knowledge retention / documentation

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Inputs

- Existing data models and databases
- Data standards
- Data sets
- Initial data requirements
- Original data requirements
- Data architecture

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Activities

1. Plan for Data Modeling
2. Build the Data Models (conceptual, logical, physical)
3. Review the Data Models
4. Manage the Data Model

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Deliverables

- Data Models
  - conceptual
  - logical
  - physical

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Suppliers

- Business Professionals
- Business Analysts
- Data Architects
- DBAs, Developers
- SME
- Data Stewards
- Metadata Admin

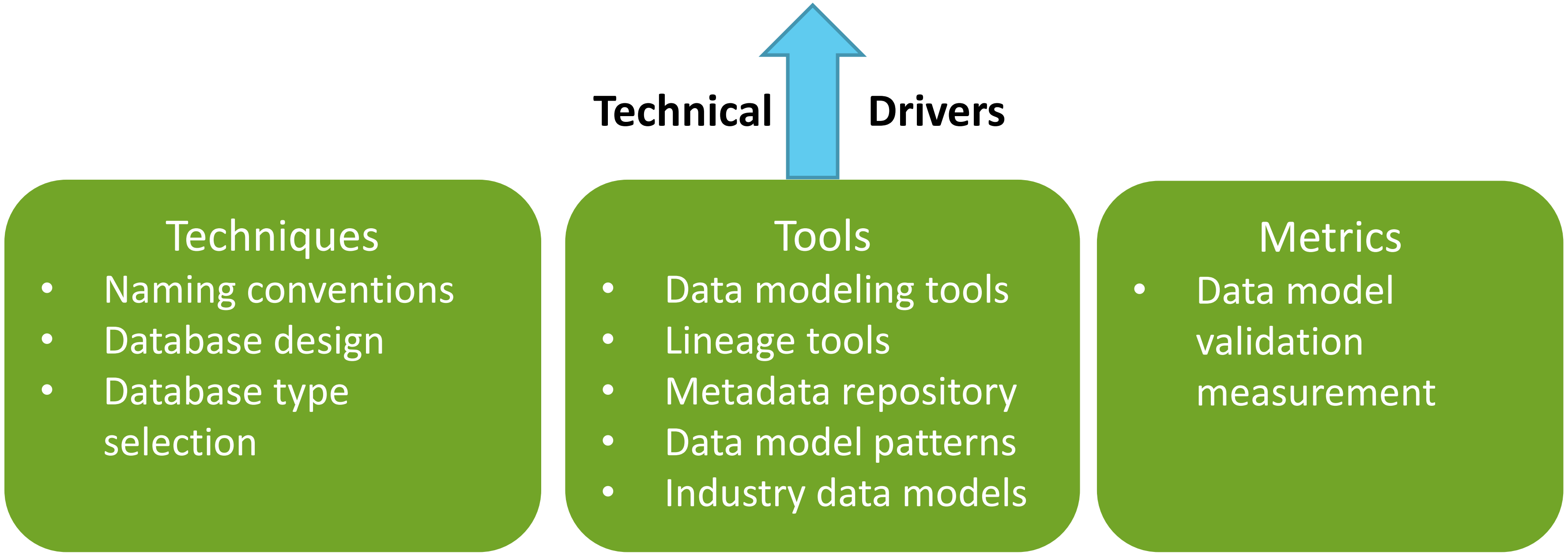
Participants

- Business Analysts
- Data Modelers

Consumers

- Business Analysts
- Data Modelers
- DBAs, and Developers
- Software Developers
- Data Stewards
- Data Quality Analysts
- Data Consumers

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA



Technical **Drivers**

Techniques

- Naming conventions
- Database design
- Database type selection

Tools

- Data modeling tools
- Lineage tools
- Metadata repository
- Data model patterns
- Industry data models

Metrics

- Data model validation measurement

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Drivers

- Provide a **common vocabulary** around data
- Capture and document explicit **knowledge** about an organization’s data and systems
- Serve as primary **communication tool** during projects

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Data Models

- Provide a **common vocabulary** around data
- Capture and document explicit **knowledge** about an organization’s data and systems
- Serve as primary **communication tool** during projects
- Data Models contains 4 main building blocks:
  - Entities
  - Relationships
  - Attributes
  - Domains

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Data Types – 4 main types (Edvinsson, 2013)

- Category information
- Resource information
- Business event information
- Detail transaction information

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Data Model Level of Detail

- Conceptual
- Logical
- Physical

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Entities

- Entity is a thing that exists "per se", separate from other things
- It is a thing about which the organization collects information
- They are referred to as nouns
- The entity represents the answer to:
    - Who
    - What
    - When, Where, Why
    - How
    - Measurement

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

- Entity definitions are core Metadata
- High quality data definition has three main characteristics:
    - Clarity
    - Accuracy
    - Completeness

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

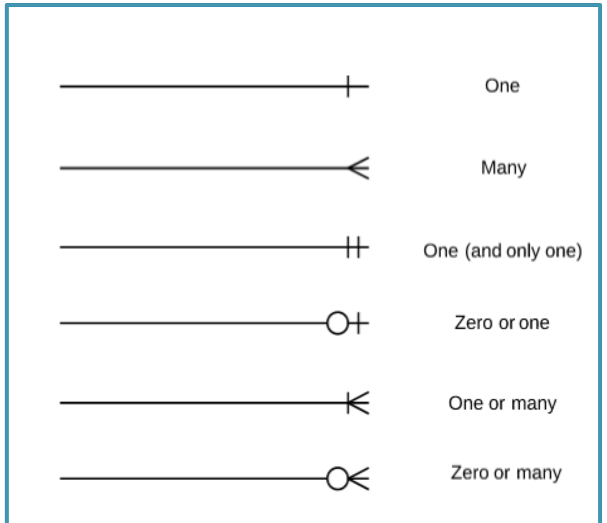
Relationship

- Association between entities (Chen, 1976)
  - It captures the high-level interactions between conceptual entities
  - The detailed interactions between logical entities
  - The constraints between physical entities

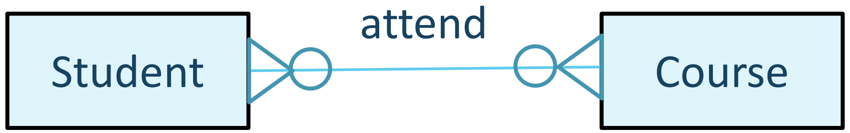
[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Relationship Cardinality

- It captures how many entity-instances participate in a relationship with how many of the other entity



Crow’s foot notation

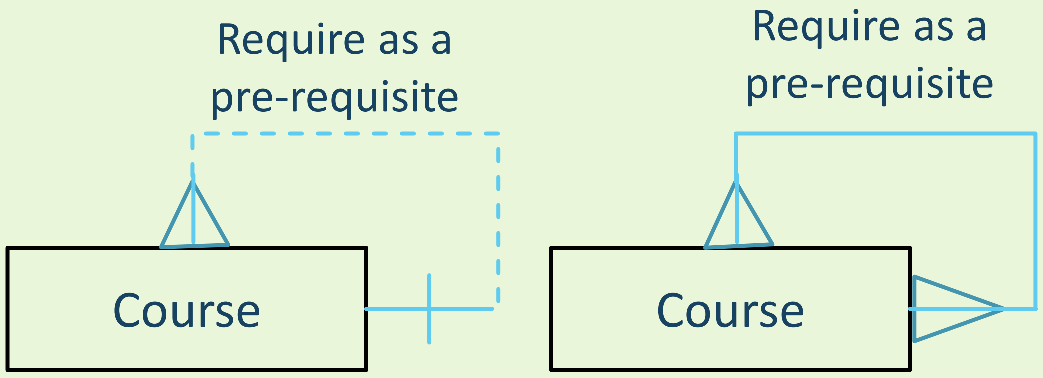


Business Rules

- Each student may attend one or many courses
- Each course may be attended by one or many students

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Examples of relationship



Require as a pre-requisite

Course

Require as a pre-requisite

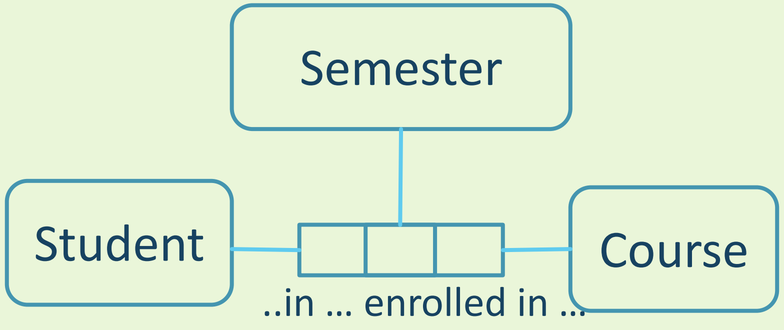
Course

Hierarchy

Network or Graph

Unary relationship

Recursive or self-referencing



Semester

Student

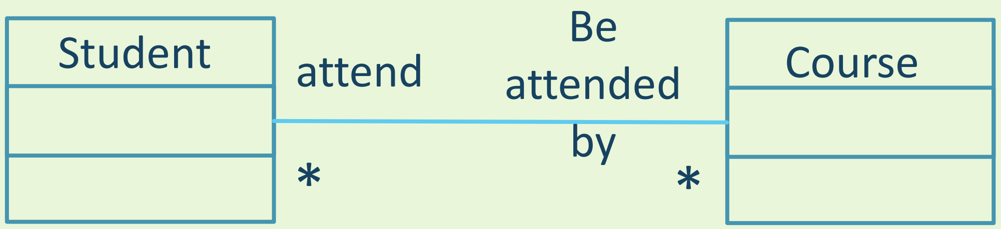
Course

..in … enrolled in …

Ternary relationship

(object-role notation)

Fact-based modeling



Student

attend

Be attended by

Course

*

*

Binary relationship (UML notation)

Attributes

- An attribute is a property that identifies, describes, or measures an entity.
- The physical correspondent of an attribute in an entity is a column, field, tag, or node in a table, view, document, graph or file

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Keys

- Are a set of one or more attributes that uniquely defined an instance of an entity

Foreign Key

- A foreign key is used in physical and sometimes logical relational data modeling schemes to represent relationships.
- A foreign key is created implicitly when a relationship is defined between entities, depending on the database technology or the data modeling tools, and whether the two entities have mutual dependencies

Type of Keys: by construction

- Simple Key: one attribute that uniquely identifies an entity-instance
- Surrogate Key: is a simple key, unique identifier for a table, system-generated, without intelligence, integer, whose meaning is unrelated to its face value
- Compound Key: set of two or more attributes that together uniquely identify an entity-instance
- Composite Key: contains one compound key and at least one other simple or compound key or non-key attribute

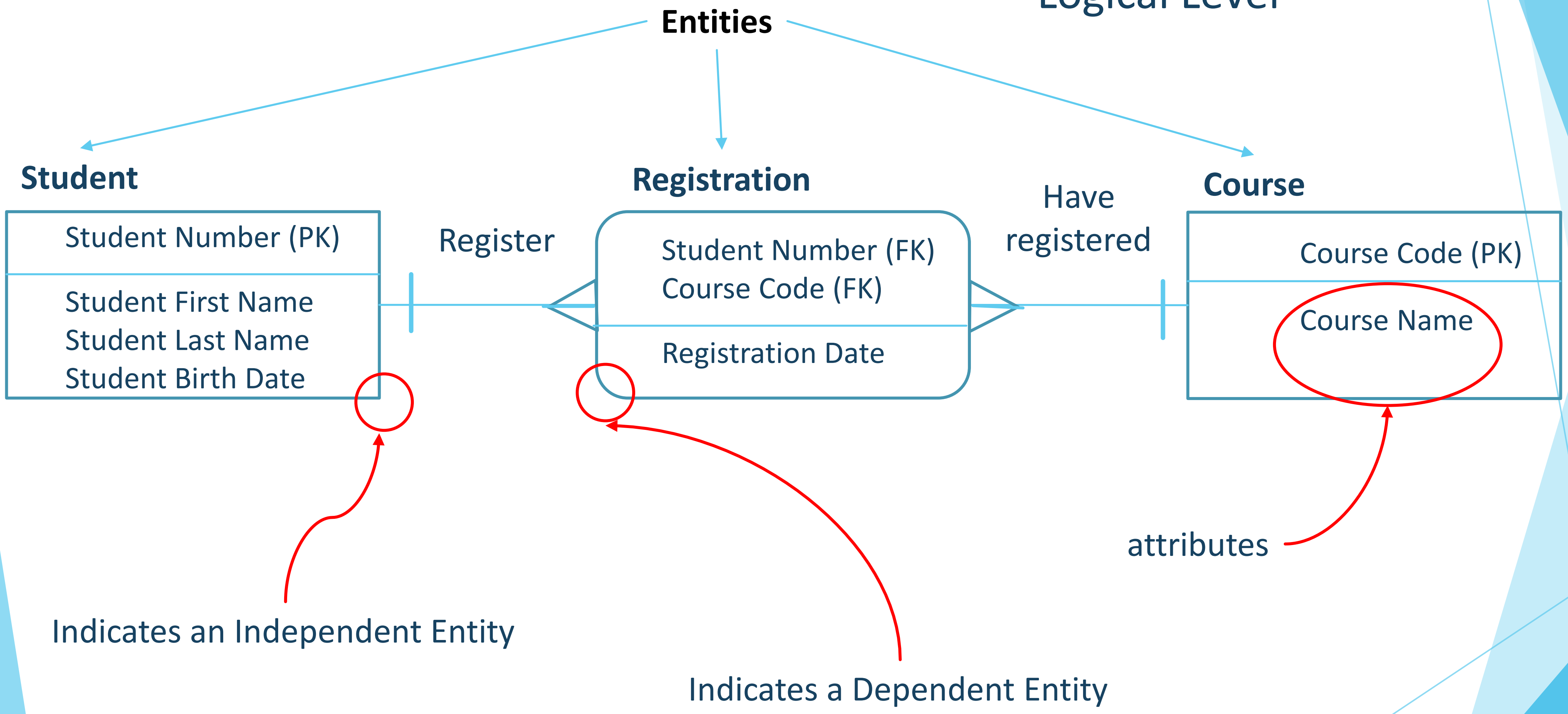
[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Type of Keys: by function

- Super key: set of attributes that uniquely identify and entity-instance
- Candidate key: is a minimal set of one or more attributes that identifies the entity-instance (i.e., a simple or compound key)
- Primary key**:** is the candidate key that is chosen to be the unique identifier for an entity
- Alternate key: are used to find specific entity instances

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Logical Level



[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Domain

- A complete set of possible values that an attribute can be assigned.
- Provides a means of standardizing the characteristics of the attributes
- All values inside a domain are valid values, outside are invalid
- Attributes should not contain values outside its domain
- Domain constraints are rules that restrict a domain with specific rules
- Domains can be defined by data type, format, list, range, rule-based

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

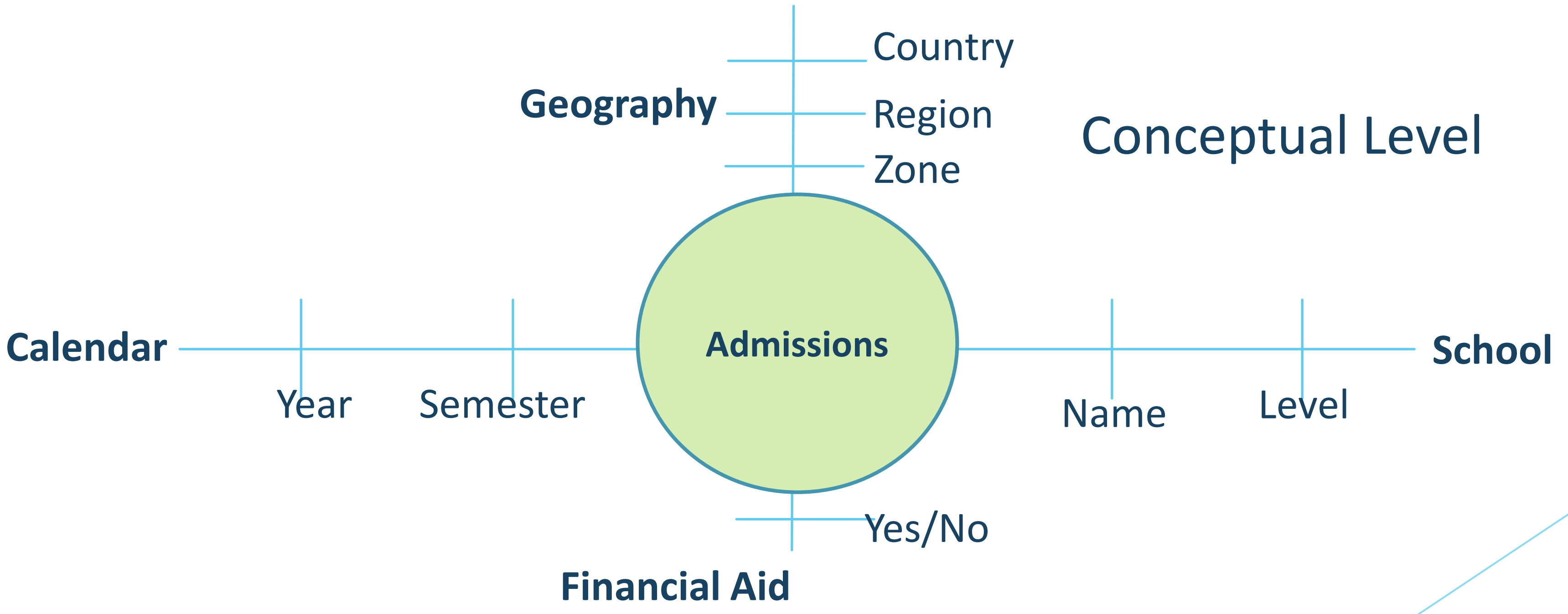
Relational Scheme

- Relational Theory provides a systemic way to organize data so that they reflected their meaning (Codd, 1970)
- The design objectives are to have an exact expression of business data and to have one fact in one place (no redundancies)
- It is ideal for the design of operational systems (transactional database)
- The most common form of notation is Information Engineering (IE) syntax (crow's foot)

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Dimensional Scheme

- Data is structured to optimize the query and analysis of large amount of data
- Dimensional models focus on a particular business process
- The model capture the navigation paths needed to answer questions



[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Object-Oriented Scheme

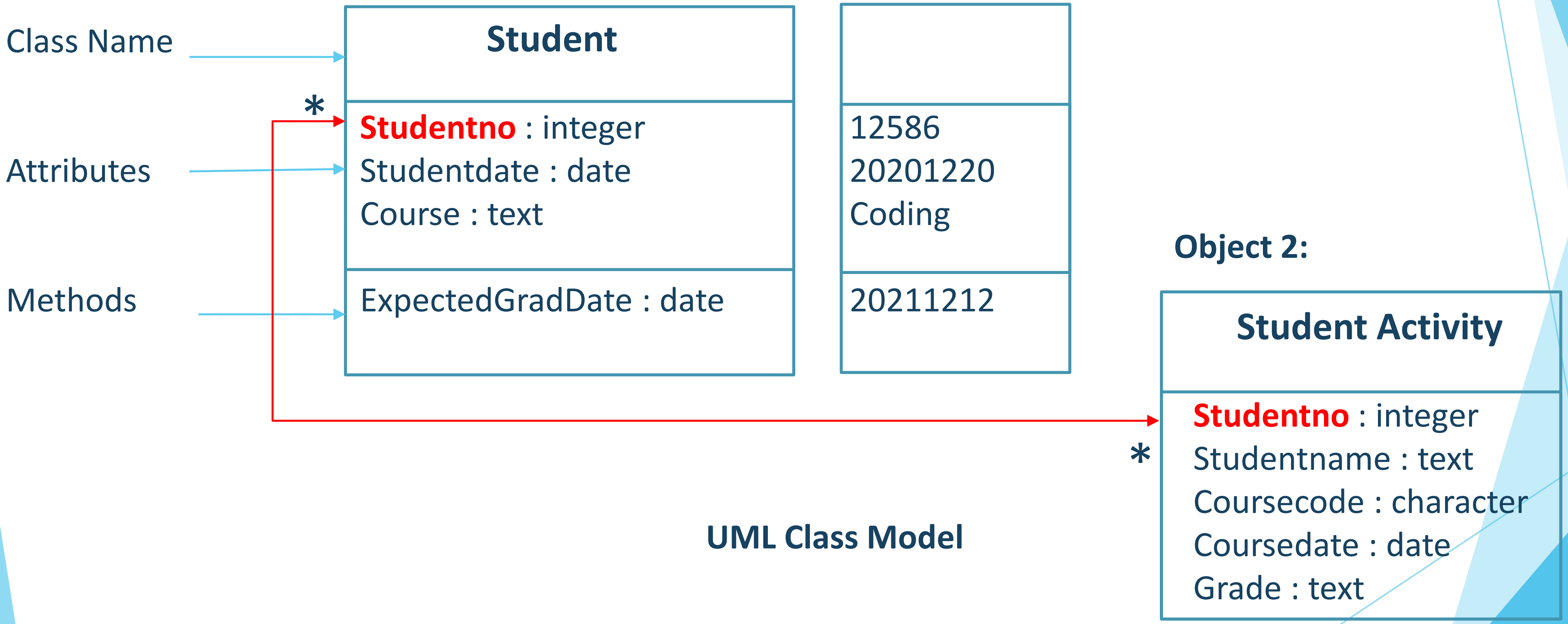
- Object-oriented programming (OOP) is a programming paradigm based on the concept of "**objects**", which can contain data and code: data in the form of fields (attributes or properties), and code, in the form of procedures (methods). [2]

- Many programming languages (such as C++, Java, Python) are multi-paradigm and they support object-oriented [2]

- Unified Modeling Language (UML) is a graphical modeling language

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA
[2] Source: Wikipedia.org

Object-Oriented Scheme



Object 1:

Object 1: instance

Class Name

Attributes

Methods

Student

*

Studentno : integer
Studentdate : date
Course : text

ExpectedGradDate : date

12586
20201220
Coding

20211212

Object 2:

Student Activity

*

Studentno : integer
Studentname : text
Coursecode : character
Coursedate : date
Grade : text

UML Class Model

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

NoSQL Scheme

- Non-relational database technology provides a mechanism for storage and faster retrieval of data that is modeled in means other than the tabular relations used in relational databases [2]
- Types:
  - Document
  - Graph
  - Key-value
  - Column-oriented

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA
[2] Source: Wikipedia.org

1. **Document**: it stores the business subject in a structure called a Document
2. **Graph**: designed for data that is well represented as a set of nodes with a finite number of connections between them
3. **Key-value**: it allows an application to store simple and complex data only in two columns (Key and Value)
4. **Column-oriented**: like RDBMS, it looks data as rows and values, however, it can work with complex data types including unformatted text and multimedia.

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA
[2] Source: Wikipedia.org

Normalization

- Rules to organize business complexity into stable data structures
- Keeps attributes in only one place to eliminate redundancy, inconsistencies
- Rules sort attributes according to the PK and FKs
- Each level corresponds to a separate normal form and do not necessarily include the previous one

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Pros and Cons

- Increases data consistency as it avoids the duplicity
- Helps in grouping related data under the same schema
- Improves searching faster as indexes can be created faster (OLTP)
- Delays the retrieving of data as more table joins are needed
- Normalization is not a good option in OLAP transactions

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

- **1NF**: each entity has a valid PK, every attribute depends on the PK
- **2NF**: each entity has the minimal PK, every attribute depends on the complete PK
- **3NF**: each entity has no hidden PKs, each attribute depends on no attributes outside the whole key

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Normalization – cont’d

1NF

EmployeeNumber	LastName	FirstName	AreaName	AreaCity	AreaCountry
1111	Andrews	Jack	Accounting	Toronto	Canada
1115	Smith	Mike	Technology	Toronto	Canada
1220	Jones	Harry	HR	New York	USA
1250	Harvey	John	Admin	London	UK
1250	Harvey	John	HR	London	UK

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

2NF

EmployeeNumber	LastName	FirstName
1111	Andrews	Jack
1115	Smith	Mike
1220	Jones	Harry
1250	Harvey	John
1250	Harvey	John

Table A

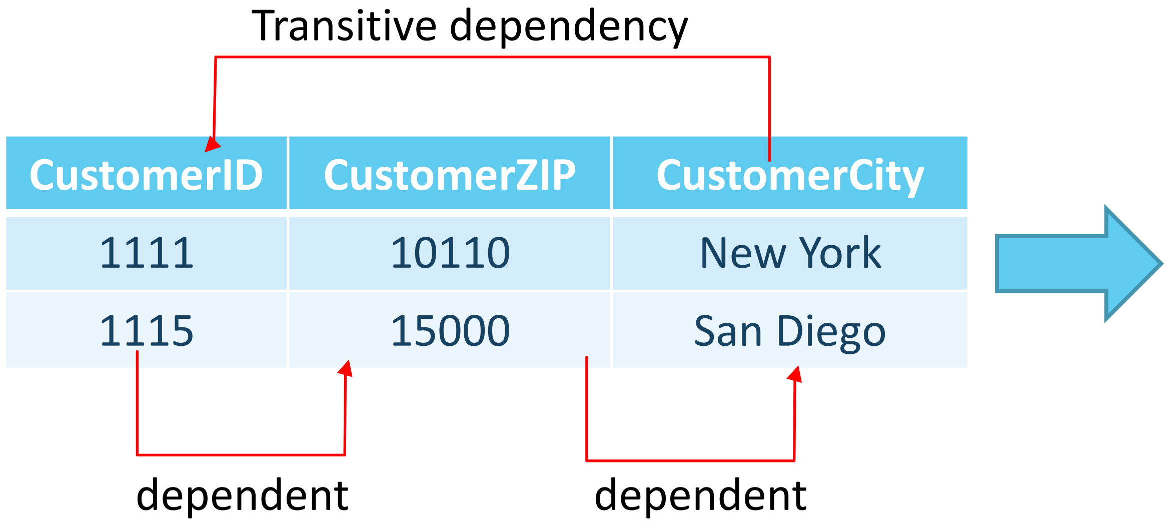
EmpAreaID	EmployeeNumber	AreaNumber
1	1111	10
2	1115	20
3	1220	30
4	1250	40
5	1250	30

Table C

AreaNumber	AreaName	AreaCity	AreaCountry
10	Accounting	Toronto	Canada
20	Technology	Toronto	Canada
30	HR	New York	USA
40	Admin	London	UK

Table B

3NF



Transitive dependency

CustomerID	CustomerZIP	CustomerCity
1111	10110	New York
1115	15000	San Diego

dependent

dependent

Customer Table

Customer Table

CustomerID	CustomerZIP
1111	10110
1115	15000

CustZIP Table

CustomerZIP	CustomerCity
10110	New York
15000	San Diego

[1] Source: Software Testing Help https://www.softwaretestinghelp.com/

- **Boyce / Codd normal form** (BCNF): resolves overlapping composite candidate keys
- **4NF**: resolves many-to-many-to-many relationships (and beyond) in pairs until entities cannot be broken down into smaller pieces
- **5NF**: resolves inter-entity dependencies into basic pairs, and all join dependencies use parts of the PK

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

2. ACTIVITIES

a) Planning

- Diagram
- Definitions
- Issues and outstanding questions
- Lineage

b) Build the Data Model

- Forward Engineering (CDM, LDM, PDM) (see tasks on next page)
- Reverse Engineering

c) Review the Data Model

d) Manage and Maintain the Data Model

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Build the Data Model / Forward Engineering

CDM	LDM	PDM
Select Scheme	Analyze Information Requirements	Resolve Logical Abstractions
Select Notation	Analyze existing Documentation	Add Attribute Details
Complete Initial CDM	Add Associative Entities	Add Reference Data Objects
Incorporate Enterprise Terminology	Add Attributes	Assign Surrogate Keys
Obtain Sign-off	Assign Domains	Renormalize for Performance
	Assign Keys	Index, Partition for Performance
		Create Views

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

3. TOOLS

- Data Modeling tools
- Lineage tools
- Data Profiling tools
- Metadata Repositories
- Data Model Patterns
- Industry Data Models

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

4. **Best Practices in:**

   a) Naming Conventions

   b) Data Design

      - Performance and ease of use
      - Reusability
      - Integrity
      - Security
      - Maintainability

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

5. DATA GOVERNANCE MODEL

a) Data Model and Design Quality Management

- Develop Data Modeling and Design Standards
- Review Data Model and Database Design Quality
- Manage Data Model Versioning and Integration

b) Data Modeling Metrics

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA