

考试资料职业发展  
技术读书笔记分享

B站/闲鱼：大西洋活跃的锅巴  
公众号：不太甜

# DAMA-DMBOK

## 数据管理知识体系指南CDGA/CDGP认证

第12章 元数据管理（完整课程视频请扫描二维码）





# 第12章 元数据管理

01

引言

02

活动

03

工具和方法

04

实施指南

05

治理



# 01

## 引言

定义、业务驱动因素、目标和原则、基本概念

元数据不仅包括技术和业务流程、数据规则和约束，还包括逻辑数据结构与物理数据结构等。描述了数据本身（如数据库、数据元素、数据模型），数据表示的概念（如业务流程、应用系统、软件代码、技术基础设施），数据与概念之间的联系（关系）。元数据可以帮助组织理解其自身的数据、系统和流程，同时帮助用户评估数据质量，对数据库与其他应用程序的管理是不可或缺的。

### 可靠且良好的元数据管理有助于：

- 1) 通过提供上下文语境和执行数据质量检查提高数据的可信度
- 2) 通过扩展用途增加战略信息（如主数据）的价值
- 3) 通过识别冗余数据和流程提高运营效率
- 4) 防止使用过时或不正确的数据
- 5) 减少数据的研究时间
- 6) 改善数据使用者和IT专业人员之间的沟通
- 7) 创建准确的影响分析，从而降低项目失败的风险
- 8) 通过缩短系统开发生命周期时间缩短产品上市时间
- 9) 通过全面记录数据背景、历史和来源降低培训成本和员工流动的影响
- 10) 满足监管合规

### 管理不善容易导致的问题：

- 1) 冗余的数据和数据管理流程
- 2) 重复和冗余的字典、存储库和其他元数据存储
- 3) 不一致的数据元素定义和与数据滥用的相关风险
- 4) 元数据的不同版本相互矛盾且有冲突，降低了数据使用者的信心
- 5) 怀疑元数据和数据的可靠性

### 目标：

- 1) 记录和管理与数据相关的业务术语的知识体系，以确保人们理解和使用数据内容的一致性
- 2) 收集和整合来自不同来源的元数据，以确保人们了解来自不同组织不同部门的数据之间的相似与差异
- 3) 确保元数据的质量、一致性、及时性和安全
- 4) 提供标准途径，使元数据使用者（人员、系统和流程）可以访问元数据
- 5) 推广或强制使用技术元数据标准，以实现数据交换

### 原则：

- |         |       |         |
|---------|-------|---------|
| 1) 组织承诺 | 2) 战略 | 3) 企业视角 |
| 4) 潜移默化 | 5) 访问 | 6) 质量   |
| 7) 审计   | 8) 改进 |         |

## 1、元数据与数据

元数据也是一种数据。一个人的元数据可能是另一个人的数据

## 2、元数据的类型

信息技术之外的类别：

### 1) 描述元数据

描述资源并支持识别和检索，如作者标题等

### 2) 结构元数据

描述资源及其组件之间的关系，如字数章节

### 3) 管理元数据

描述管理生命周期的元数据，如版本号存档日期

等

(1) 业务元数据

(2) 技术元数据

(3) 操作元数据

### (1) 业务元数据

主要关注数据的内容和条件，另包括与数据治理相关的详细信息。包括主题域、概念、实体、属性的非技术名称和定义、属性的数据类型和其他特征，如范围描述、计算公式、算法和业务规则、有效的域值及其定义。

- 1) 数据集、表和字段的定义和描述
- 2) 业务规则、转换规则、计算公式和推导公式
- 3) 数据模型
- 4) 数据质量规则和检核结果
- 5) 数据的更新计划
- 6) 数据溯源和数据血缘
- 7) 数据标准
- 8) 特定的数据元素记录系统
- 9) 有效值约束
- 10) 利益相关方联系信息（数据所有者、数据管理专员）
- 11) 数据的安全/隐私级别
- 12) 已知的数据问题
- 13) 数据使用说明



### (2) 技术元数据

提供有关数据的技术细节、存储数据的系统以及在系统内和系统之间数据流转过程的信息。

- 1) 物理数据库表名和字段名
- 2) 字段属性
- 3) 数据库对象的属性
- 4) 访问权限
- 5) 数据GRUD增删改查规则
- 6) 物理数据模型，包括数据表名、键和索引
- 7) 记录数据模型与实物资产之间的关系
- 8) ETL作业详细信息
- 9) 文件格式模式定义
- 10) 源到目标的映射文档
- 11) 数据血缘文档，包括上游和下游变更影响的信息
- 12) 程序和应用的名称和描述
- 13) 周期作业（内容更新）的调度计划和依赖
- 14) 恢复和备份规则
- 15) 数据访问的权限、组、角色

### (3) 操作元数据

描述了处理和访问数据的细节

- 1) 批处理程序的作业执行日志
- 2) 抽取历史和结果
- 3) 调度异常处理
- 4) 审计、平衡、控制度量的结果
- 5) 错误日志
- 6) 报表和查询的访问模式、频率和执行时间
- 7) 补丁和版本的维护计划和执行情况，以及当前的补丁级别
- 8) 备份、保留、创建日期、灾备恢复预案
- 9) 服务水平协议（SLA）要求和规定
- 10) 容量和使用模式
- 11) 数据归档、保留规则和相关归档文件
- 12) 清洗标准
- 13) 数据共享规则和协议
- 14) 技术人员的角色、职责和联系信息

### 3、ISO/IEC 11179元数据注册标准

提供了用于定义元数据注册的框架，旨在基于数据的精确定义，从数据元素开始，实现元数据驱动的数据交换。

- 1) 数据元素生成和标准化框架
- 2) 数据元数据分类
- 3) 数据元素的基本属性
- 4) 数据定义的形成规则和指南
- 5) 数据元素的命名和识别原则
- 6) 数据元素的注册

#### 4、非结构化数据的元数据

描述元数据，如目录信息和同义关键字

结构元数据，如标签、字段结构、特定格式

管理元数据，如来源、更新计划、访问权限和导航信息

书目元数据，如图书馆目录条目

记录元数据，如保留策略

保存元数据，如存储、归档条件和保存规则

## 5、元数据来源

(1) 应用程序中元数据存储库

(2) 业务术语表：

1) 业务用户

2) 数据管理专员

业务术语表应包含业务术语属性：

1) 术语名称、定义、缩写或简称，以及任何同义词

2) 负责管理与术语相关的数据的业务部门和/或应用程序

3) 维护术语的人员姓名和更新日期

4) 术语的分类或分类间的关联关系（业务功能关联）

5) 需要解决的冲突定义、问题的性质、行动时间表

6) 常见的误解

7) 支持定义的算法

8) 血缘

9) 支持该术语的官方或权威数据来源

序

### 5、元数据来源

- (3) 商务智能工具
- (4) 配置管理工具
- (5) 数据字典

数据字典定义数据集的结构和内容，通常用于单个数据库、应用程序或数据仓库。可用于管理数据模型中每个元素的名称、描述、结构、特征、存储要求、默认值、关系、唯一性和其他属性。还包含表或文件定义

- (6) 数据集成工具
- (7) 数据库管理和系统目录

数据库目录是元数据的重要来源，它们描述了数据库的内容、信息大小、软件版本、部署状态、网络正常运行时间、基础架构正常运行时间、可用性，以及许多其他操作元数据属性。

- (8) 数据映射管理工具
- (9) 数据质量工具
- (10) 字典和目录
- (11) 事件消息工具

## 5、元数据来源

### （12）建模工具和存储库

生成与应用程序或系统模型设计相关的元数据，如主题域、逻辑实体、逻辑属性、实体和属性关系、父类型和子类型、表、字段、索引、主键和外键、完整性约束以及模型中其他类型的属性。

### （13）参考数据库

记录各种类型的枚举数据的业务价值和描述，在系统的上下文中使用。

### （14）服务注册

是从面向服务的架构（SOA）角度管理和存储有关服务和服务终端的技术信息，如定义、接口、操作、输入和输出参数、制度、版本和示例使用场景。一些与服务相关的最重要的元数据包括服务版本、服务位置、数据中心、可用性、部署日期、服务端口、IP地址、统计端口、连接超时和连接重试超时。

### （15）其他元数据存储

事件注册表、源列表或接口、代码集、词典、时空模式、空间参考、数字地理数据集的分发、存储库的存储库和业务规则

## 6、元数据架构的类型

元数据生命周期：

- 1) 元数据创建和采集
- 2) 元数据在一个或多个存储库中存储
- 3) 元数据集成
- 4) 元数据交付
- 5) 元数据使用
- 6) 元数据控制和管理



## 6、元数据架构的类型

### （1）集中式元数据架构

由单一的元数据存储库组成，包含来自各种不同源的元数据副本。集中式存储库的优点有：

- 1) 高可用性
- 2) 快速的元数据检索，因为存储库和查询功能在一起
- 3) 解决了数据库结构的问题，使其不受第三方或商业系统特有属性的影响
- 4) 抽取元数据时可进行转换，自定义或使用其他源系统中的元数据进行补充，提高了元数据的质量。

缺点有：

- 1) 必须使用复杂的流程确保元数据源头中的更改能够快速同步到存储库中。
- 2) 维护集中式存储库的成本可能很高
- 3) 元数据的抽取可能需要自定义模块或中间件
- 4) 验证和维护自定义代码会增加对内部IT人员和软件供应商要求

## 6、元数据架构的类型

### (2) 分布式元数据架构

维护了一个单一的接入点。分布式元数据架构没有持久化的存储库

优点：

- 1) 元数据总是尽可能保持最新且高效，因为它是从其数据源中直接检索的。
- 2) 查询是分布式的，可能会提高响应和处理的效率
- 3) 来自专有系统的元数据请求仅限于查询处理，而不需要
- 4) 自动化元数据查询处理的开发可能更简单，只需要很少的人工干预
- 5) 减少了批处理，没有元数据复制或同步过程

缺点：

- 1) 无法支持用户定义或手动插入的元数据项，因为没有存储库可以放置这些添加项。
- 2) 需要通过统一的、标准化的展示方式呈现来自不同系统的元数据
- 3) 查询功能受源系统可用性的影响
- 4) 元数据质量完全取决于源系统

## 6、元数据架构的类型

### （3）混合式元数据架构

降低了对专有系统进行手动干预和自定义编码访问功能的工作量。基于用户的优先级和要求，使用时尽可能是最新且有效的。不会提高系统的可用性。

## 6、元数据架构的类型

### （4）双向元数据架构

允许元数据在架构的任何部分（源、数据集成、用户界面）中进行更改，然后将变更从存储库（代理）同步到其原始源以实现反馈。

该设计强制元数据存储库包含最新版本的元数据源，并强制对源的更改管理，必须系统地捕获变更，然后加以解决；必须构建和维护附加的一系列处理接口，以将存储库的内容回写至元数据源。



# 02

## 活动

## 1、定义元数据战略

- 1) 启动元数据战略计划
- 2) 组织关键利益相关方的访谈
- 3) 评估现有的元数据资源和信息架构
- 4) 开发未来的元数据架构
- 5) 制定分阶段的实施计划

## 2、理解元数据需求

- 1) 更新频次
- 2) 同步情况
- 3) 历史信息
- 4) 访问权限
- 5) 存储结构
- 6) 集成要求
- 7) 运维要求
- 8) 管理要求
- 9) 质量要求
- 10) 安全要求

## 3、定义元数据架构

### 1、创建元模型（第一步是这个，考题）

创建一个元数据存储库的数据模型，也叫元模型；是定义元数据战略和理解业务需求后的第一个设计步骤。

### 2、应用元数据标准

对内标准包括命名规范、自定义属性、安全、可见性和处理过程文档，组织对外元数据标准包括交换格式和应用程序接口设计

### 3、管理元数据存储

控制活动包括：

- 1) 作业调度和监控
- 2) 加载统计分析
- 3) 备份、恢复、归档、消除
- 4) 配置修改
- 5) 性能调优
- 6) 查询统计分析分析
- 7) 查询和报表生成
- 8) 安全管理



### 3、管理元数据存储

质量控制活动：

- 1) 质量保证，质量控制
- 2) 数据更新频率——与时间表匹配
- 3) 缺失元数据报告
- 4) 未更新的元数据报告

元数据管理活动包括：

- 1) 加载、探测、导入和标记数据资产
- 2) 记录与源的映射和迁移关系
- 3) 记录版本
- 4) 用户界面管理
- 5) 连接数据集的元数据维护——为NoSQL提供支持
- 6) 数据对内数据采集建立连接——自定义连接和作业元数据
- 7) 外部数据源和订阅源的许可
- 8) 数据增强元数据，如关联GIS

培训活动包括：

- 1) 教育和培训用户和数据专员
- 2) 生成和分析管理指标
- 3) 对控制活动、查询、报告进行培训

## 4、创建和维护元数据

元数据管理的几个一般原则描述了管理元数据质量的方法：

- 1) 责任
- 2) 标准
- 3) 改进

### 1、整合元数据

对元数据存储库的扫描有两种不同的方式：

1) 专用接口：单步方式，扫描程序从来源系统中采集元数据，直接调用特定格式的装载程序，将元数据加载到元数据存储中。不需要输出任何中间元数据文件。

2) 半专用接口：两步方式，扫描程序从来源系统中采集元数据，并输出到特定格式的数据文件中。扫描程序只产生目标存储库能够正确读取和加载的数据文件。可以被多种方式读取，所以这种接口的架构更加开放。

在此过程中，扫描程序产生和使用多种类型的文件：

- 1) 控制文件
- 2) 重用文件
- 3) 日志文件
- 4) 临时和备份文件

### 2、分发和传递元数据

- 1) 元数据内部网站
- 2) 报告、术语表和其他文档
- 3) 数据仓库、数据集市和BI（商务智能）工具
- 4) 建模和软件开发工具
- 5) 消息传送和服务
- 6) web服务和应用程序接口（API）
- 7) 外部组织接口方案（如供应链解决方案）

## 5、查询、报告和分析元数据

- 1、根据需要给用户分组
- 2、将工具与用户要求相匹配



# 03

## 工具和方法

主要工具是元数据存储库，包括整合层和手工更新的接口。处理和使用元数据的工具集成到元数据存储库中作为元数据来源。

元数据管理工具提供了在集中位置（存储库）管理元数据的功能，元数据可以手动输入，也可以通过专门的连接器从其他各种源中提取。，元数据存储库还提供与其他系统交换元数据的功能。

## 1、数据血缘和影响分析

发现和记录数据资产的元数据的一个重要意义在于提供了数据如何在系统间转移信息。

基于程序编码的当前版本的血缘称为“实现态血缘”。映射规范文档中描述的血缘称为“设计态血缘。”

血缘创建的局限性在于元数据管理系统的覆盖范围

拼接的结果是一个表示数据从原始位置（数据源或记录系统）转移到最终位置的全景视图

### （1）业务焦点

根据业务优先级寻找数据源的血缘关系

### （2）技术焦点

从源系统开始识别直接相关的数据使用者，依次识别间接的数据使用者，直到识别出所有系统为止。



## 2、应用于大数据采集的元数据

元数据标签应在采集时应用于元数据，然后元数据可以用来识别可访问的数据湖中的数据内容。数据剖析可以识别出数据域、数据关系和数据质量问题，并打上标签。



# 04

## 实施指南

就绪评估/风险评估、组织和文化变革

## 1、就绪评估/风险评估

评估缺失高质量元数据可能带来的影响如下：

- 1) 因不正确、不完整和不合理的假设或缺乏数据内容的知识导致错误判断
- 2) 暴露敏感数据，使客户和员工面临风险，影响商业信誉和导致法律纠纷
- 3) 如果了解数据的那些领域专家们离开了，那么他们了解的知识也随之被

带走了。

组织准备情况的评估解决方法为：

对元数据相关活动现状进行正式的成熟度评估，评估内容应包括重要的业务数据元、可用的元数据术语表、数据血缘、数据剖析和数据质量管理过程、主数据管理成熟度和其他方面。

风险评估的交付成功包括元数据战略和实施战略

## 2、组织和文化变革



# 05

## 治理

过程控制、解决方案文档、标准和指南

## 1、过程控制

数据管理团队负责定义标准和管理元数据的状态变化，同时可以负责组织内的质量提升活动、培训活动或实际培训活动。

通过多个不同阶段和状态的决策来确定业务术语和定义

## 2、元数据解决方案的文档

- 1) 元数据管理实施状态
- 2) 源和目标元数据存储
- 3) 元数据更新的调度计划信息
- 4) 留存和保持的版本
- 5) 内容
- 6) 质量声明或警告
- 7) 记录系统和其他数据源状态（如数据内容历史加载、删除或更新标志）
- 8) 相关工具、架构和人员
- 9) 敏感信息和数据源的移除或脱敏策略

### 3、元数据标准和指南

XML模式、文档类型定义（DTD）、XML模式定义（XSD）。

### 元数据管理环境的建议指标包括：

1) 元数据存储库完整性：将企业元数据（范围内所有产品和实例）的理想覆盖率与实际覆盖率进行比较

2) 元数据管理成熟度：根据能力成熟度模型（CMM-DMM）的成熟度评估方法

3) 专职人员配备：以及职位描述中的角色定义说明

4) 元数据使用情况：存储库的访问次数

5) 业务术语活动：使用、更新、定义解析、覆盖范围

6) 主数据服务数据遵从性：显示SOA解决方案中数据的重用情况

7) 元数据文档质量

一个指标是通过自动和手动两种方式评估元数据文档的质量

另一个是度量具有定义的属性的百分比，以及随着时间推移而发生变化的趋势

8) 元数据存储库可用性：正常运行时间、处理时间（批处理和查询）



考试资料职业发展  
技术读书笔记分享

B站/闲鱼：大西洋活跃的锅巴  
公众号：不太甜

# 本章完结 感谢观看

完整课程视频请扫描二维码咨询

