DAMA-DMBOK

数据管理知识体系指南CDGA/CDGP认证

第13章 数据质量（完整课程视频请扫描二维码）

第13章 数据质量/Contents

01 引言

定义、业务驱动因素、目标和原则、基本概念

定义

定义：为确保满足数据消费者的需求，应用数据管理技术进行规划，实施和控制等管理活动。

导致低质量数据产生的因素：

组织缺乏对低质量数据影响的理解等、缺乏规划、孤岛式系统设计、不一致的开发过程、不完整的文档、缺乏标准或缺乏治理等。

业务驱动因素

建议正式数据质量管理的业务驱动因素包括：

1）提高组织数据价值和数据利用的机会
2）降低低质量数据导致的风险和成本
3）提高组织效率和生产力
4）保护和提高组织的声誉

许多直接成本均与低质量数据有关：

1）无法正确开具发票
2）增加客服电话质量，降低解决问题的能力
3）因错失商业机会造成收入损失
4）影响并购后的整合进展
5）增加受欺诈的风险
6）由错误数据驱动的错误业务决策造成损失
7）因缺乏良好信誉而导致业务损失

目标和原则

目标：

1）根据数据消费者的需求，开发一种受管理的方法，使数据适合要求
2）定义数据质量控制的标准和规范，并作为整个数据生命周期的一部分
3）定义和实施测量、监控和报告数据质量水平的过程
4）通过过程和系统改进，识别和提倡提高数据质量的机会

原则：

1）重要性：改进的优先顺序应根据数据的重要性以及数据不正确时的风险水平来判定
2）全生命周期管理：数据质量管理应覆盖从创建或采购直至处置的数据全生命周期，包括其系统内部和系统之间流转时的数据管理。
3）预防：预防数据错误和降低数据可用性
4）根因修正：对流程和支持它们的系统进行更改
5）治理：必须支持高质量数据的开发
6）标准驱动：所有利益相关方都会有数据质量要求
7）客观测量和透明度
8）嵌入业务流程：流程中实施数据质量标准
9）系统强制执行：强制执行数据质量要求
10）与服务水平关联：纳入管理水平协议（SLA）

1、数据质量

既指高质量数据的相关特征，也指用于衡量或改进数据质量的过程。挑战之一是，与质量相关的期望并不总是已知的。

2、关键数据

根据以下要求评估关键数据：

1）监管报告
2）财务报告
3）商业政策
4）持续经营
5）商业战略

3、数据质量维度（包括哪些？）

是数据的某个可测量的特性。

数据质量维度提供了可定义数据质量要求的一组词汇，通过这些维度定义可以评估初始数据质量和持续改进的成效。为了衡量数据质量，组织需要针对重要业务流程和可以测量的参数建立特征。

3、数据质量维度

Strong-Wang 数据质量的四个大类和15个指标

（1）内在数据质量：

1）准确性　2）客观性　3）可信度　4）信誉度

（2）场景数据质量：

1）增值性　2）关联性　3）及时性　4）完整性

5）适量性

（3）表达数据质量：

1）可解释性　2）易理解性　3）表达一致性　4）简洁性

（4）访问数据质量：

1）可访问性　2）访问安全性

3、数据质量维度——Thomas《信息时代的数据质量》

（1）数据模型

1）内容：①数据关联性 ②获取价值的能力 ③定义清晰性

2）详细程度：①特征描述颗粒度

（2）属性域的精准度

1）构成：①自然性 ②可识别性③同一性④最小必要冗余性

2）一致性：①模型各组成部分的语义一致性
②跨实体类型属性的结构一致性

3）应变性：①健壮性 ②灵活性

4）数据值：①准确性 ②完备性 ③时效性 ④一致性

5）数据表达：①适当性②可解释性③可移植性④格式精确性
⑤格式灵活性 ⑥表达空值的能力
⑦有效利用存储 ⑧数据的物理实例与其格式一致

3、数据质量维度——Larry English 《改善数据仓库和业务信息质量》

（1）固有质量特征

1）定义的一致性
2）值域的完备性
3）有效性或业务规则一致性
4）数据源的准确性
5）反映现实的准确性
6）精确性
7）非冗余性
8）冗余或分布数据的等效性
9）冗余或分布数据的并发性

（2）实用质量特征

1）可访问性　　2）及时性
3）语境清晰性　　4）可用性
5）多源数据的可整合性　　6）适当性或事实完整性

3、数据质量维度——DAMA UK白皮书 数据质量的6个核心维度

1）完备性

存储数据量与潜在数据量的百分比

2）唯一性

在满足对象识别的基础上不应多次记录实体实例（事物）

3）及时性

数据从要求的时间点起代表现实的程度

4）有效性

如数据符合其定义的语法（格式、类型、范围），则数据有效

5）准确性

数据正确描述所描述的“真实世界”对象或事件的程度

6）一致性

比较事物多种描述与定义的差异

3、数据质量维度——DAMA UK白皮书的其他特征

1）可用性

数据是否可理解、简单、相关、可访问、可维护，且达到正确的精度水平

2）时间问题

是否稳定，是否对合法的变更请求作出及时响应

3）灵活性

是否具有可比性，是否与其他数据有很好的兼容性？是否具备可用的分组和分类？是否能被重用？是否易于操作？

4）置信度

数据治理、数据保护和数据安全等管控是否到位？数据的可信性如何

5）价值

数据是否有良好的成本/收益实例？是否得到了最佳应用？是否危及人们的安全、隐私或企业的法律责任

3、数据质量维度——一组普遍的数据质量维度定义

准确性：数据正确表示真实实体的程度

完备性：是指是否存在所有必要的数据；完备性可以在数据集、记录或列级进行测量

一致性：可以指确保数据值在数据集内和数据集之间表达的相符程度。它也可以表示系统之间或不同时间的数据集大小和组成的一致程度。

完整性：包括与完备性、准确性和一致性相关的想法。在数据中，完整性通常指的是引用完整性或数据集内部的一致性

合理性：合理性是指数据模式符合预期的程度。

及时性：及时性的概念与数据的几个特性有关

唯一性/数据去重：唯一性是指数据集内的任何实体不会重复出现

有效性：是指数据值与定义的值域一致。

4、数据质量和元数据

5、数据质量ISO标准

ISO将质量数据定义为：“符合规定要求的可移植数据”

ISO 8000定义了数据供应链中任何组织都可以测试的一些特性，从而可以客观地确定数据与ISO8000之间是否具有一致性

6、数据质量改进生命周期

计划Plan执行Do检查Check行动Act

7、数据质量业务规则

1）定义一致性　2）数值存在和记录完备性
3）格式符合性　4）值域匹配性
5）范围一致性　6）映射一致性
7）一致性规则：指根据这些属性的实际值，在两个或多个属性之间关系的条件判定。
8）准确性验证：将数据值与记录系统或其他验证来源中的相应值进行比较
9）唯一性验证
10）及时性验证：表明与数据可访问性和可用性预期相关特征的规则

聚合检查的示例包括：

1）验证文件中记录数量的合理性
2）验证从一组交易中计算出的平均金额的合理性
3）验证指定时间段内交易数量的预期差异

8、数据质量问题的常见原因

（1）缺乏领导力导致的问题

有效管理数据质量的障碍包括：

1）领导和员工缺乏意识
2）缺乏治理
3）缺乏领导力和管理能力
4）难以证明改进的合理性
5）测量价值的工具不合适或不起作用

8、数据质量问题的常见原因

（2）数据输入过程中引起的问题

1）数据输入接口问题
2）列表条目放置
3）字段重载
4）培训问题
5）业务流程的变更
6）业务流程执行混乱

8、数据质量问题的常见原因

（3）数据处理功能引起的问题

1）有关数据源的错误假设

2）过时的业务规则

3）变更的数据结构

8、数据质量问题的常见原因

（4）系统设计引起的问题 （5）解决问题引起的问题

1）未能执行参照完整性。

参照完整性对于确保应用程序或系统级别的高质量数据是必要的。如果没有强制执行参照完整性，或者关闭了验证，则有可能出现各种数据质量问题：

①产生破坏唯一性约束的重复数据

②既可以包含，又可以排除在某些报表中的孤儿数据，导致同样的计算生成多个值

③由于参照完整性要求已还原或更改，无法升级

④由于丢失的数据被分配为默认值而导致数据准确性

2）未执行唯一性约束：表或文件中的多个实例副本预期包含唯一实例

3）编码不准确和分歧：数据映射或格式不正确，或处理数据的规则不准确，处理过的数据就会出现质量问题

4）数据模型不准确：如果数据模型内的假设没有实际数据的支持，则会出现数据质量问题，包括实际数据超出字段长度导致数据丢失、分配不正确ID或键值等

5）字段重载：随着时间的推移，为了其他目的重用字段，而不是更改数据模型或代码，可能会导致混淆的值集、不明确的含义。

6）时间数据不匹配：采用不同的日期格式

7）主数据管理薄弱：不成熟的主数据管理可能为数据选择不可靠的数据源

8）数据复制： ①单源-多个本地实例 ②多源-单一本地实例

9、数据剖析

是一种用于检查数据和评估质量的数据分析形式。使用统计技术来发现数据集合的真实结构、内容和质量。剖析引擎生成统计信息，可以识别数据内容和结构中的模式：

1）空值数
2）最大/最小值
3）最大/最小长度
4）单个列值的频率分布
5）数据类型和格式

还包括跨列分析，识别不符合格式要求的水平，以及意外格式识别

基本概念——10、数据质量和数据处理

（1）数据清理

1）实施控制以防止数据输入错误
2）纠正源系统中的数据
3）改进数据录入的业务流程

（2）数据增强

是给数据集添加属性以提高其质量和可用性的过程。通过集成组织内部的数据集可以获得，也可以通过购买外部数据

1）时间戳
2）审计数据
3）参考词汇表
4）语境信息
5）地理信息
6）人口统计信息
7）心理信息：如偏好、成员资格、休闲活动、交通方式
8）评估信息：针对资产评估、库存和销售数据

（3）数据解析和格式化

数据解析时使用预先确定的规则来解释其内容或值的分析过程

（4）数据转换和标准化

数据转换建立在标准化技术的基础之上。通过将原始格式和模式中的数据值映射到目标表述形式来指导基于规则的转换。

标准化是分析人员经过反复分析语境、语言学，以及公认的最常见的惯用语等，为获取规则而进行的一种特殊的格式转换。

02 活动

1、定义高质量数据

从不同角度探讨这个问题：

1）了解业务战略和目标
2）与利益相关方面谈，以识别痛点、风险和业务驱动因素
3）通过资料收集和其他剖析形式直接评估数据
4）记录业务流程中的数据依赖关系
5）记录业务流程的技术架构和系统支持

2、定义数据质量战略

1）了解并优先考虑业务需求
2）确定满足业务需求的关键数据
3）根据业务需求定义业务规则和数据质量标准
4）根据预期评估数据
5）分享调查结果，并从利益相关方哪里获得反馈
6）优先处理和管理问题
7）确定并优先考虑改进机会
8）测量、监控和报告数据质量
9）管理通过数据质量流程生成的元数据
10）将数据质量控制集成到业务和技术流程中

3、识别关键数据和业务规则

识别能描述或暗示有关数据质量特征要求的业务规则。
完整性规则反映了字段是强制的还是可选的。

4、执行初始数据质量评估

POC的步骤包括：

1）定义评估的目标
2）确定要评估的数据，
3）识别数据的用途和数据的使用者
4）利用待评估的数据识别已知风险，包括数据问题对组织的潜在影响。
5）根据已知和建议的规则检查数据
6）记录不一致的级别和问题类型
7）根据初步发现进行额外的深入分析，以便：
    ①量化结果
    ②根据业务影响优化问题
    ③提出关于数据问题根本原因的假设
8）与数据管理专员、领域专家和数据消费者会面，确认问题和优先级
9）使用调查结果作为规划的基础
    ①解决问题，找到根本原因
    ②控制和改进处理流程，以防止问题重复发生
    ③持续控制和汇报

5、识别改进方向并确定优先排序

数据剖析和分析的步骤：

定义目标、了解数据使用和风险，根据规则衡量，记录并与领域专家确认结果，利用这些信息确定补救和改进工作的优先级。

剖析是分析数据质量的第一步。

6、定义数据质量改进目标

确定改进的投资回报率：

1）受影响数据的关键性
2）受影响的数据量
3）数据的龄期
4）受问题影响的业务流程数量和类型
5）受问题影响的消费者、客户、供应商或员工数量
6）与问题相关的风险
7）纠正根本原因的成本
8）潜在的工作成本

7、开发和部署数据质量操作

1、管理数据质量规则

预先定义规则将：

1）对数据质量特征设定明确的期望
2）提供防止引入数据问题的系统编辑和控制要求
3）向供应商和其他外部方提供数据质量要求
4）为正在进行的数据质量测量和报告创建基础

规则应该是：

1）记录的一致性
2）根据数据质量维度定义
3）与业务影响挂钩
4）数据分析支持
5）由领域专家确认
6）所有数据消费者都可以访问

2、测量和监控数据质量

进行业务数据质量的原因：

1）向数据消费者通报质量水平

2）管理业务或技术流程，改变引入的变更风险

ValidDQI(r)=TestExecution(r)-ExceptionsFound(r)/TestExecution(r)

I nvalidDQI(r)=ExceptionFound(r)/TestExecutions(r)

r为正在测试的规则

3、制定管理数据问题的操作过程

（1）诊断问题

1）在适当的信息处理流程下查看数据问题，并隔离出现缺陷过程的位置
2）评估是否存在任何可能导致错误的环境变化
3）评估是否有其他过程问题导致了数据质量事件
4）确定外部数据是否存在影响数据质量的问题

（2）制定补救方案

1）纠正非技术性根本原因，如缺乏培训、缺乏领导支持、责任和所有权不明确等
2）修改系统以消除技术类的根本原因
3）制定控制措施以防止问题发生
4）引入额外的检查和监测
5）直接修正有缺陷的数据
6）基于变更的成本和影响对比更正后的数据的价值分析，不采取任何操作

（3）解决问题

1）评估替代方案的相对成本和优点　　2）推荐计划中的一个备选方案
3）提供开发和实施该解决方案的计划　　4）实施该解决方案

事件跟踪系统将收集与解决问题、分配工作、问题数量、发生频率，以及做出响应、给出诊断、计划解决方案和解决问题所需时间相关的性能数据。这些指标可以为当前工作流的有效性、系统和资源利用率提供有价值的洞察，它们是重要的管理数据点。

进行有效的跟踪需要做到以下几点：

1）标准化数据质量问题和活动　　2）提供数据问题的分配过程
3）管理问题升级过程　　4）管理数据质量解决方案工作流

7、开发和部署数据质量操作

4、制定数据质量服务水平协议

规定了组织对每个系统中数据质量问题进行响应和补救的期望。

数据质量SLA中定义的数据质量控制操作包括：

1）协议涵盖的数据元素

2）与数据缺陷相关的业务影响

3）与每个数据元素相关的数据质量指标

4）从每个已确定指标的数据元素出发，识别数据价值链上每个应用程序系统中的质量期望

5）测量这些期望的方法

6）每次测量的可接受性阈值

7）如果不满足可接受性阈值，应通知数据管理专员

8）预期解决或补救问题的时间和截止日期

9）升级策略，以及可能的奖励和惩罚

5、编写数据质量报告

1）数据质量评分卡
2）数据质量趋势
3）服务水平协议（SLA）指标
4）数据质量问题管理
5）数据质量问题管理。
6）IT和业务团队对数据质量政策的一致性

03 工具和方法

工具

1、数据剖析工具
2、数据查询工具
3、建模和ETL工具
4、数据质量规则模板
5、元数据存储库

1、预防措施

预防方法包括：

1）建立数据输入控制：创建数据输入规则，防止无效或不准确的数据进入系统

2）培训数据生产者：确保上游系统的员工了解其数据对下游用户的影响，对数据的准确性和完整性进行激励或基础评估，让其不仅仅追求录入速度

3）定义和执行规则

4）要求数据供应商提供高质量数据：检查外部供应商的流程，以检查其结构、定义、数据源和数据出处

5）实施数据治理和管理制度：确保定义并执行以下内容的角色和责任：参与规则、决策权和有效管理数据和信息资产的责任。

6）制定正式的变更控制：确保在实施之前对存储数据的所有变更进行定义和测试

2、纠正措施

1）自动修正

自动更正技术包括基于规则的标准化、规范化和更正。修改后的值是在没有人工干预的情况下获取或生成和提交的。

2）人工检查修正

使用自动工具矫正和纠正数据，并在纠正提交到持久存储之前进行人工检查。

3）人工修正

在缺乏工具、自动化程度不足或者确定人工监督能更好地处理变更的情况下，人工更正是唯一的选择。

3、质量检查和审核代码模块

创建可共享、可连接、可重用的代码模块，开发人员可以从存储库中拿到它们，重复执行数据质量检查和审计过程

4、有效的数据质量指标

1）可度量性：数据质量指标可被量化
2）业务相关性
3）可接受性
4）问责/管理制度：关键利益相关方
5）可控制性：应触发行动来改进数据
6）趋势分析：一段时间内测量数据质量改进的情况

5、统计过程控制

SPC是一种通过分析过程输入、输出或步骤的变化测量值来管理过程的方法。SPC基于假设：当一个具有一致输入的过程被一致执行时，它将产生一致的输出。它使用集中趋势（变量的值接近其中心值的趋势，如平均值、中值或模式）和围绕中心值可变性（如范围、方差、标准偏差）的度量来确定过程中的偏差公差。

主要工具是控制图，是一个时间序列图，包括平均值的中心线，以及描述测算的上下控制界限。

SPC通过识别过程中的变化来衡量过程结果的可预测性。

第一步是对过程进行度量，以识别和消除特殊原因

第二步是尽可能早地发现异常变化，因为早期发现问题简化了对问题根源的调查过程

6、根本原因分析

是一个理解导致问题发生的因素及其作用原理的过程。目的是识别潜在的条件，这些条件一旦消除，问题也将消失。

04 实施指南

就绪评估/风险评估、组织和文化变革

数据质量项目的实施计划：

1）有关数据价值和低质量数据成本的指标
2）IT/业务交互的操作模型
3）项目执行方式的变化
4）对业务流程的更改
5）为补救和改进项目提供资金
6）为数据质量运营提供资金

1、就绪评估/风险评估

1）管理层承诺将数据作为战略资产进行管理
2）组织对数据质量的当前理解
3）数据的实际情况
4）与数据创建、处理或使用相关的风险
5）可扩展数据质量监控的文化和技术就绪。

2、组织与文化变革

首先是提高数据对组织作用和重要性的认识。

培训应着重于：

1）导致数据问题的常见原因

2）组织数据生态系统中的关系以及为什么提高数据质量需要全局方法

3）糟糕数据造成的后果

4）持续改进的必要性

5）要“数据语言化”，阐述数据对组织战略与成功、监管报告和客户满意度的影响。

05

治理

过程控制、解决方案文档、标准和指南

数据质量团队的利益相关方合作：

1）风险与安全人员可以帮助识别与数据相关的组织弱点

2）业务流程工程和培训人员，可以帮助团队实施流程改进

3）业务和运营数据专员以及数据所有者，他们可以识别关键数据、定义标准和质量期望，并优先处理数据问题

治理组织可以通过以下方式加快数据质量方案的工作：

1）设定优先级
2）确定和协调有权参与各种数据质量相关决定和相关活动的人
3）制定和维护数据质量标准
4）报告企业范围内数据质量的相关测量
5）提供有助于员工参与的指导
6）建立知识共享的沟通机制
7）制定和应用数据质量和合规政策
8）监控和报告绩效
9）共享数据质量检查结果，以提高认识，确定改进机会，并就改进达成共识
10）解决变化和冲突，提供方向性指导

1、数据质量制度

1）制度的目的、范围和适用性

2）术语定义

3）数据质量团队的职责

4）其他利益相关方的责任

5）报告

6）策略的实施，包括与之相关的风险、预防措施、合规性、数据保护和数据安全性等

度量指标

1）投资回报

2）质量水平

测量一个数据集内或多个数据集之间的错误或不满足甚至违反需求情况的数量和比率

3）数据质量趋势

随着时间的推移，针对阈值和目标的质量改进，或各阶段的质量事件

4）数据问题管理指标

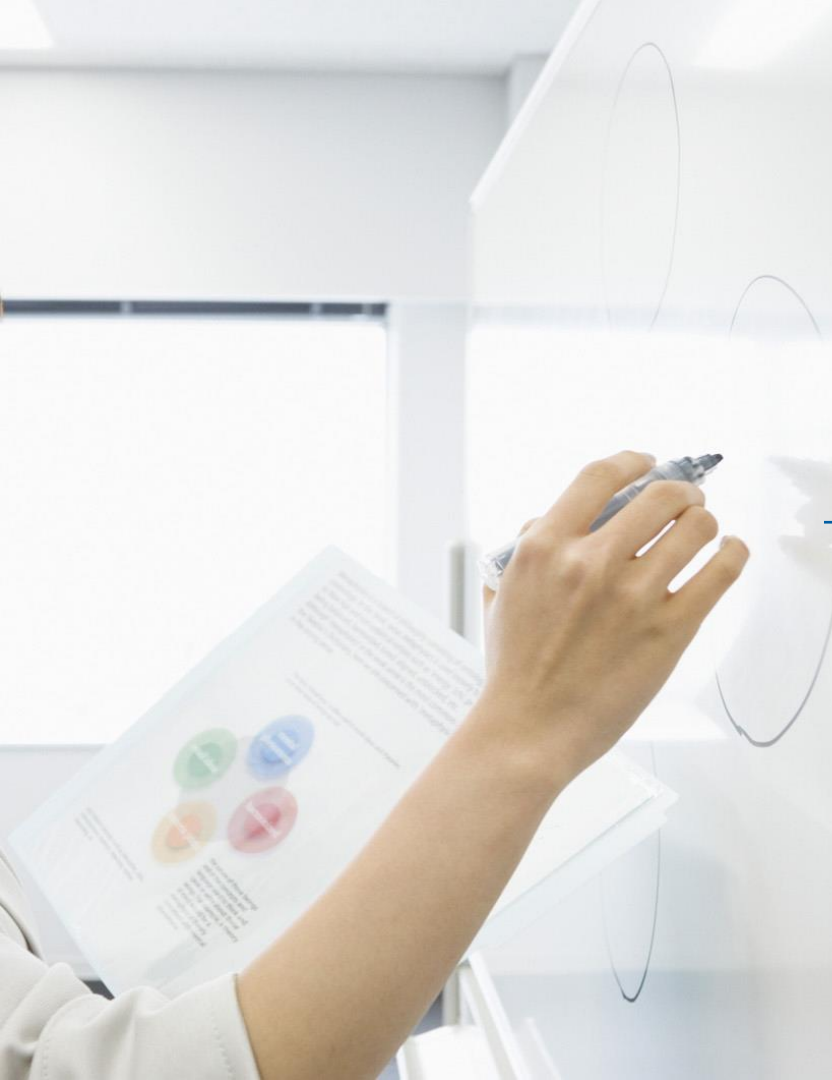
①按数据质量指标对问题分类与计数
②各业务职能部门及其问题的状态
③按优先级和严重程度对问题排序
④解决问题的时间

5）服务水平的一致性

包括负责人员在内的组织单位对数据质量评估项目干预过程的一致性。

6）数据质量计划示意图

现状和扩展路线图



本章完结 感谢观看

完整课程视频请扫描二维码咨询