第 1 章 数据管理

1.1 引言

数据管理（Data Management）是为了交付、控制、保护并提升数据和信息资产的价值，在其整个生命周期中制订计划、制度、规程和实践活动，并执行和监督的过程。

1.2 基本概念

1.2.1 数据

1、【数据的概念】

长期以来，对数据的定义强调了它在反映客观事实方面的作用。在信息技术中，数据也被理解为**以数字形式存储**的信息（尽管数据不仅限于已数字化的信息，而且与数据库中的数据相同，数据管理的原则也适用于**纸面上的数据**）。

Q：数据是以数字形式存储的信息？

A：错，数据是以数字形式+纸面形式存储的信息。

2、【数据本身也需要被解释，如南京市长江大桥】

数据既是对其所代表对象的解释，也是必须解释的对象（SebastianColeman，2013）。

1.2.2 数据和信息

1、【数据和信息的关系，信息是被处理过的数据】

数据被称为“信息的原材料”，而信息则被称为“在上下文语境中的数据”。

2、组织内部在数据和信息之间画一条线，可能有助于清晰地沟通不同利益相关方对不同用途的需求和期望（如“这是上季度的销售报告”（信息）。它基于数据仓库中的数据（数据）。下一季度，这些结果（数据）将用于生成季度绩效指标（信息）。认识到要为不同的目的准备数据和信息，将使数据管理形成一个核心原则：**数据和信息都需要被管理；**如果再将两者的使用和客户的需求结合在一起进行管理，则两者应具有更高的质量。在本书中，**这些术语可以互换使用。**

Q：数据是原材料，不需要被管理，信息是加工了的数据，所以需要被管理。

A：错，**数据和信息都需要被管理**。

1.2.4 数据管理原则【12 项原则一定会考】

（1）数据是有独特属性的资产

数据是一种资产，但相比其他资产，其在管理方式的某些方面有很大差异。对比金融和实物资产，其中最明显的一个特点是**数据资产在使用过程中不会产生消耗**。

（2）数据的价值可以用【也应该用】经济术语来表示

将数据称为资产意味着它有价值。虽然有技术手段可以测量数据的数量和质量，但还未形成这样做的标准来衡量其价值。想要对其数据做出更好决策的组织，应该开发一致的方法来量化该价值。他们还应该衡量低质量数据的成本和高质量数据的好处。

（3）管理数据意味着对数据的质量管理

确保数据符合应用的要求是数据管理的首要目标。为了管理质量，组织必须了解利益相关方对质量的要求，并根据这些要求度量数据。

数据管理的首要目标：提高数据质量/确保数据符合应用的要求。

数据管理的终极目标：实现数据的价值。

（4）管理数据需要元数据【元数据先行】

管理任何资产都需要首先拥有该项资产的数据（员工人数、账户号码等）。用于管理和如何使用数据的数据都称为元数据。因为数据无法拿在手中或触摸到，要理解它是什么以及如何使用它，需要以元数据的形式定义这些知识。元数据源于与数据创建、处理和使用相关的一系列流程，包括架构、建模、管理、治理、数据质量管理、系统开发、IT 和业务运营以及分析。

（5）数据管理需要规划

即便是小型组织，也可能有复杂的技术和业务流程蓝图。数据在多个地方被创建，且因为使用需要在很多存储位置间移动，因而需要做一些协调工作来保持最终结果的一致，需要从架构和流程的角度进行规划。

（6）数据管理须驱动信息技术决策【业务驱动技术落地】

数据和数据管理与信息技术和信息技术管理紧密结合。管理数据需要一种方法，确保技术服务于而不是驱动组织的战略数据。

（7）数据管理是跨职能的工作

数据管理需要一系列的技能和专业知识，因此**单个团队无法管理组织的所有数据**。数据管理需要技术能力、非技术技能以及协作能力。

（8）数据管理需要企业级视角【整体角度】

虽然数据管理存在很多专用的应用程序，但它必须能够有效地被应用于整个企业。这就是为什么数据管理和数据治理是交织在一起的原因之一。

（9）数据管理需要多角度思考

数据是流动的，数据管理必须不断发展演进，以跟上数据创建的方式、应用的方式和消费者的变化。

（10）数据管理需要全生命周期的管理，不同类型数据有不同的生命周期特征

数据是有生命周期的，因此数据管理需要管理它的生命周期。因为数据又将产生更多的数据，所以数据生命周期本身可能非常复杂。数据管理实践活动需要考虑数据的整个生命周期。不同类型数据有不同的生命周期特征，因此它们有不同的管理需求。数据管理实践需要基于这些差异，保持足够的灵活性，以满足不同类型数据的生命周期需求。

（11）数据管理需要纳入与数据相关的风险

数据除了是一种资产外，还代表着组织的风险。数据可能丢失、被盗或误用。组织必须考虑其使用数据的伦理影响。数据相关风险必须作为数据生命周期的一部分进行管理。

（12）有效的数据管理需要领导层承担责任

数据管理涉及一些复杂的过程，需要协调、协作和承诺。为了达到目标，不仅需要管理技巧，还需要来自领导层的愿景和使命。

1.2.5 数据管理的挑战

1.数据与其他资产的区别

实物资产是看得见、摸得着、可以移动的，在同一时刻只能被放置在一个地方。金融资产必须在资产负债表上记账。然而数据不同，它不是有形的。尽管数据的价值经常随着时间的推移而变化，但它是**持久的、不会磨损**的。数据很容易被复制和传送，但它一旦被丢失或销毁，就不容易重新产生了。因为它在**使用时不会被消耗**，所以它甚至可以在不损耗的情况下被偷走。**数据是动态的，可以被用于多种目的。**同样，数据甚至可以在同时被许多人使用，而对实物资产或金融资产来说，这是不可能的。数据被多次使用产生了更多的数据，大多数组织不得不管理不断提升的数据量和越来越复杂的数据关系。

Q：数据只能用于一种目的。

A：错，数据是动态的，可以被用于多种目的，如营销、风控等。

资产类别	可复制	用后消耗	容易估值	实体还是无形	处理后才有价值
石油	否	是	是	有形	是
金钱	否	是	是	有形	否
血液	否	是	部分	有形	是
人力	否	否	否	有形	是
房产	否	部分	是	有形	否
物料	否	是	是	有形	部分
知识产权	否	否	部分	无形	部分
数据	是	否	否	无形	是

2.数据价值

【DAMA 依赖于成本法、市场法、收益法】

1）获取和存储数据的**成本**。

2）如果数据丢失，更换数据需要的**成本**。

3）数据丢失对组织的影响。

4）风险缓解成本和与数据相关的潜在风险**成本**。

5）改进数据的**成本**。

6）高质量数据的优势。

7）竞争对手为数据付出的费用。

8）数据潜在的销售价格。

9）创新性应用数据的预期收入。

Q：只包括获取数据的成本，不包括存储数据的成本。

A：错，数据价值包含获取和存储数据的成本。

9.数据生命周期【DAMA 认为数据的生命周期起始于计划，和其他体系不同】

1）创建和使用是数据生命周期中的关键点。

2）**数据质量管理**必须贯穿**整个数据生命周期**。

3）**元数据质量管理**必须贯穿**整个数据生命周期**。

4）数据管理还包括确保数据安全，并降低与数据相关的风险。

5）数据管理工作应聚焦于关键数据。组织产生了大量的数据，其 中很大一部分实际上从未被使用过，试图管理每一条数据是不可能 的。生命周期管理要求将重点放在组织关键的数据上，并将数据 ROT（冗余的 Redundant、过时的 Obsolete、碎~~片化的~~不重要的 Trivial）降至最低（Aiken，2014）。

【处置不是销毁】

数据管理对数据生命周期的关注有几个重要影响：

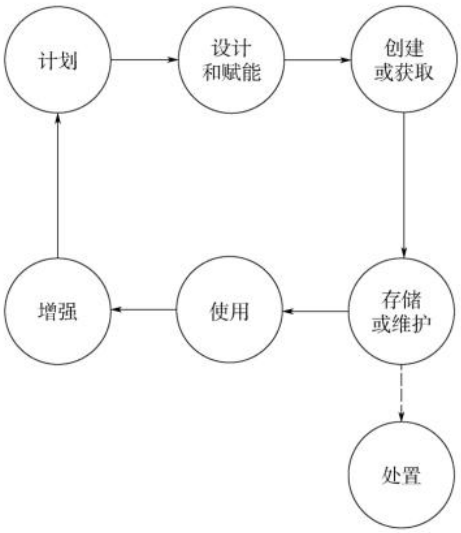

图1-2 数据生命周期中的关键活动

1.2.6 数据管理战略

【非常重要】【数据战略是必须的】

战略是一组选择和决策，它们共同构成了实现高水平目标的高水平行动过程。

在许多组织中，数据管理战略由 CDO 拥有和维护，并由数据治理委员会支持的数据管理团队实施。通常，CDO 会在数据治理委员会成立之前起草一份初步的数据战略和数据管理战略，以获得高级管理层对建立数据管理和治理的支持。

数据管理战略的组成应包括：

1）令人信服的数据管理**愿景**。

2）数据管理的商业案例总结。

3）指导原则、价值观和管理观点。

4）数据管理的使命和长期目标。

5）数据管理成功的建议措施。

6）符合 SMART 原则（具体、可衡量、可操作、现实、有时间限制）的短期（12～24 个月）数据管理计划目标。

【需清除 SMART 各字母含义，在中国国情下，短期目标为 3-6 个月】

7）对数据管理角色和组织的描述，以及对其职责和决策权的总结。

8）数据管理程序组件和初始化任务。

9）具体明确范围的优先工作计划。

10）一份包含项目和行动任务的实施路线图草案。

【非常重要】数据管理战略规划的可交付成果包括：

1）数据管理章程。包括总体愿景、业务案例、目标、指导原则、成功衡量标准、关键成功因素、可识别的风险、

运营模式等。

2）数据管理范围声明。包括规划目的和目标（通常为 3 年），以及负责实现这些目标的角色、组织和领导。

3）数据管理实施路线图。确定特定计划、项目、任务分配和交付里程碑（参见第 15 章）。

Q：数据管理战略规划可交付成果包含数据管理章程、范围声明，实施路线图在后期提交。

A：错，数据管理战略规划的可交付成果包含 3 项：数据管理章程、数据管理范围声明、数据管理实施路线图。

1.3 数据管理框架

1）前两个模型，即**战略一致性模型和阿姆斯特丹（Amsterdam）**信息模型，展示了组织管理数据的高阶关系。

2）DAMA-DMBOK 框架（**DAMA 车轮图【技术层面】、六边形图和语境关系图**）描述了由 DAMA 定义的数据管理知识领域，并解释了它们在 DMBOK 中的视觉表现。

Q：战略一致性模型和阿姆斯特丹（Amsterdam）信息模型讨论了什么之间的关系？A.软件和硬件；B.人和机器；C.业务与技术

A：业务与技术。

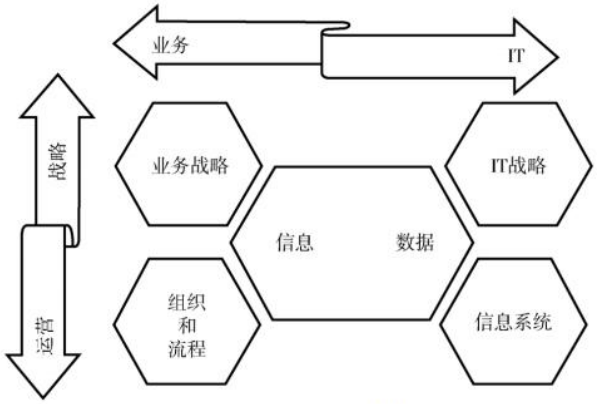

图1-3 战略一致性模型[12]

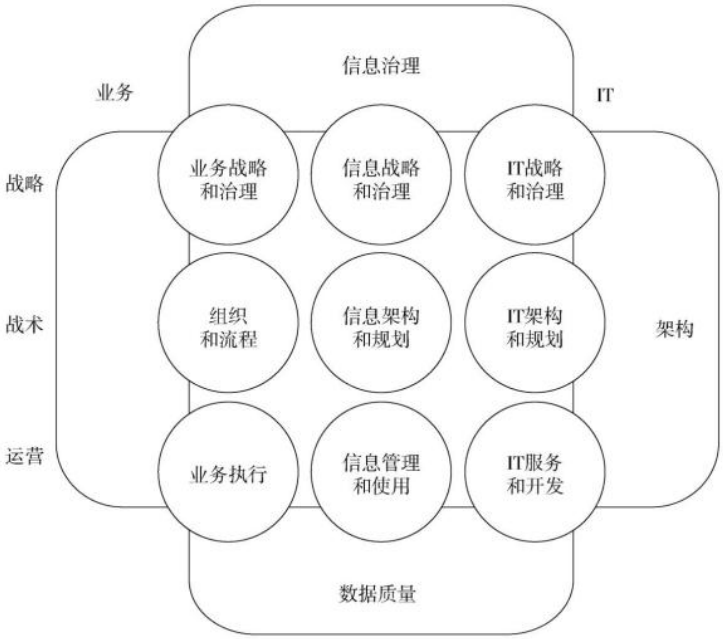

图1-4 阿姆斯特丹信息模型[14]

图1-5 *DAMA-DMBOK2*数据管理框架（DAMA车轮图）

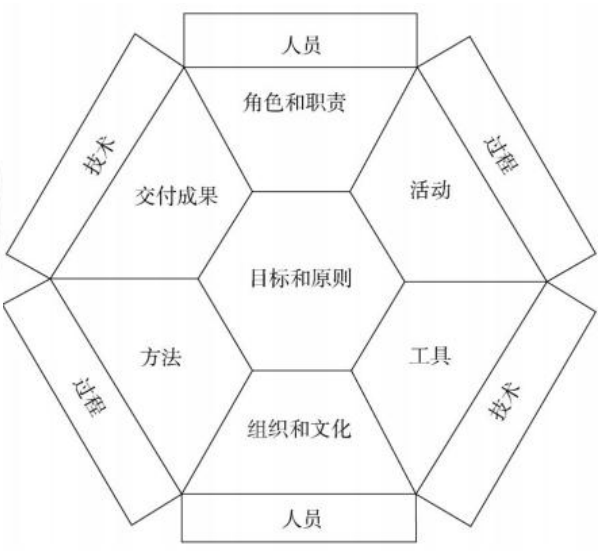

图1-6 DAMA环境因素六边形图

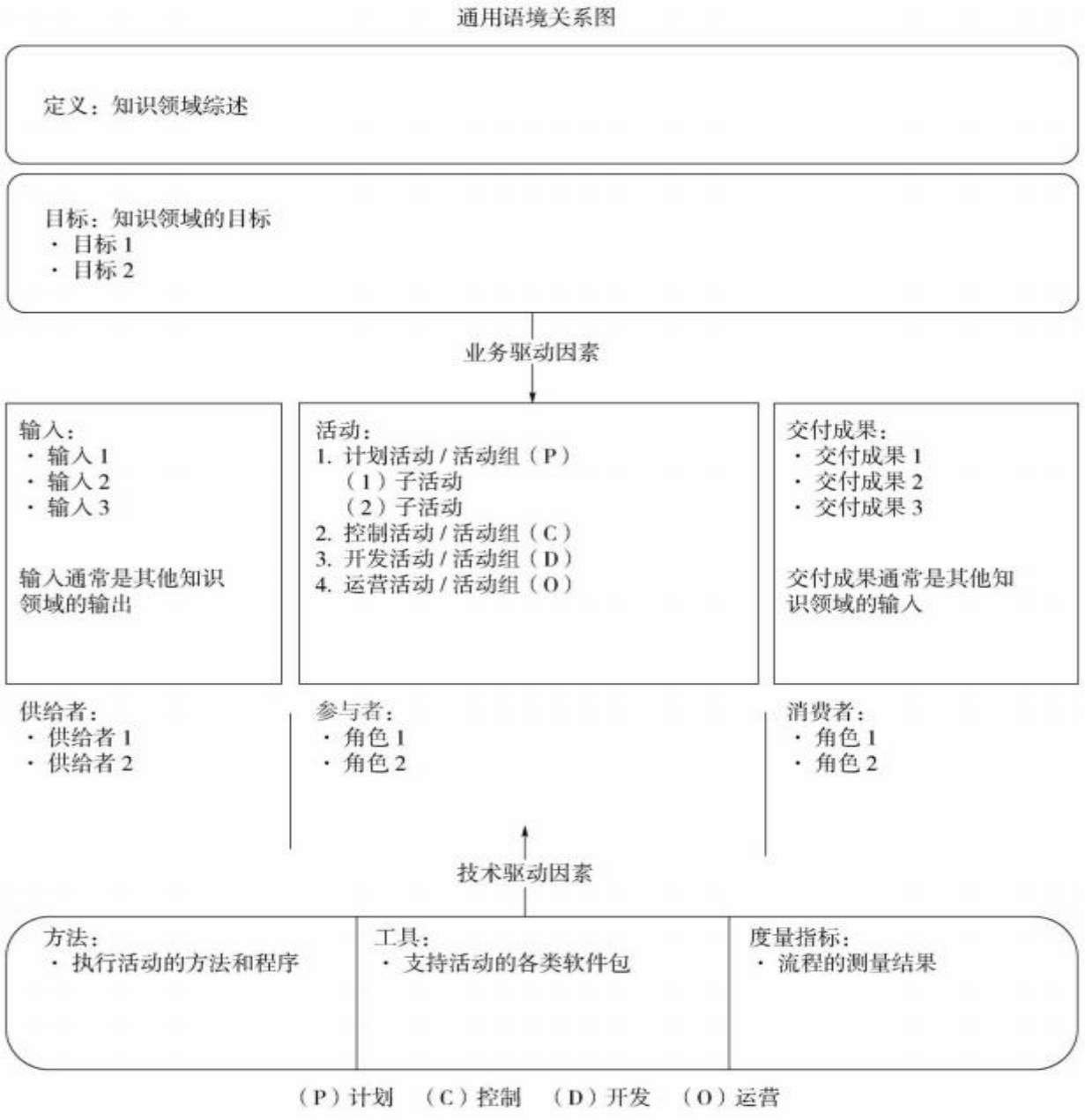

图1-7 知识领域语境关系图

1.3.5 DAMA 数据管理框架的进化

【官方观点：元数据先行】

图1-9 DAMA功能领域依赖关系图

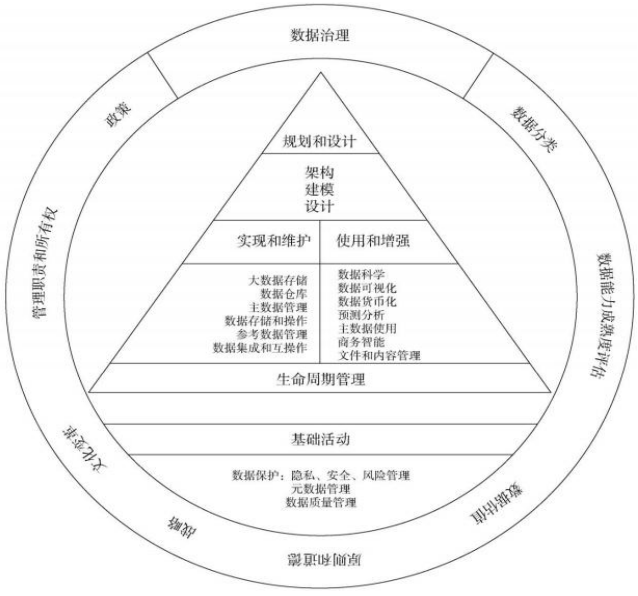

图1-11 DAMA车轮图演变

第 2 章 数据处理伦理

2.1 引言

伦理是建立在是非观念上的行为准则。伦理准则通常侧重于公平、尊重、责任、诚信、质量、可靠性、透明度和信任等方面。

数据伦理是一项社会责任问题。【不是法律问题】

【可能考题】

度量指标：培训员工人数、合规/不合规事件、企业高管参与

2.2 业务驱动因素

正如爱德华·戴明【P357 戴明环 PDCA】关于质量的定义，伦理意味着“在没有人注意的情况下正确做事（Doing it right when no one is looking）”。【慎独】 【提到管理学大师 德鲁克】

2.3 基本概念

2.3.1 数据伦理准则

【4 项内容，注意不是 3 项，团结友爱不是 4 项原则之一】【腾讯 科技向善白皮书】

（1）尊重他人

（2）行善原则

（3）公正

（4）尊重法律和公众利益（US-DHS，2012）

2.3.2 数据隐私法背后的原则

欧盟、加拿大和美国隐私法。

1、GDPR 欧盟通用数据保护条例（GDPR，2016）

【GDPR 针对好公民，不是犯罪分子，是目前数据最严格的法律】

【目的限制，即使是政府要求也不例外】

表 2-1 GDPR 准则

GDPR 准则	描述
公平、合法、透明	数据主题中的个人数据应以合法、公平和透明的方式进行处理
目的限制	必须按照指定、明确、合法的目标去采集个人数据，并且不得将数据用于采集目标之外的方面
数据最小化	采集的个人数据必须足够相关，并且仅限于与处理目的相关的必要信息
准确性	个人数据必须准确，有必要保持最新的数据。必须采取一切合理步骤，确保在完成个人数据处理后能及时删除或更正不准确的个人数据
存储限制	数据必须以可以识别的数据主体（个人）的形式保存，保存时间不得超过处理个人数据所需的时间
诚信和保密	必须确保个人数据得到安全妥善的处理，包括使用适当技术和组织方法防止数据被擅自或非法处理，防止意外丢失、被破坏或摧毁等
问责制度	控制数据的人员应负责并能够证明符合上述这些原则

Q：抓捕犯人，也需遵循公平、合法、透明准则？

A：不需要，特殊情况可以不透明。

Q：疫情期间，因防疫要求，已和居民签约不会给第三方的门禁人脸数据，是否可以给政府使用？

A：不可，签约时未说明会给政府（目的限制），需联合公安、法院章，同时修改居民合约。

修改：Q：政府收集信息，遵循公平、合法、透明原则，有些场合是否可以不透明，如抓捕犯人时。

A：错，不针对抓捕犯人，GDPR 针对好公民，不是犯罪分子，必须透明，抓犯人适用于其他法律。

Q：疫情期间，因防疫要求，已和居民签约，不会给第三方门禁人脸数据，是否可以将数据给政府使用？

A：不可，签约时仅做开门使用，未说明会给政府使用（目的限制，政府的要求也不例外），1 修改合约，让居民同意，2 联合公安、法院盖章确认如发生数据泄露排除责任，3 为防止数据泄露，最好将服务器保存至公安。

2、加拿大 PIPEDA（个人信息保护及电子文件法）

表 2-2　基于 PIPEDA 的法定义务

准则	描述
问责制度	组织有责任对其控制下个人信息负责，并设立专职人员去保证组织遵守这些准则
目的明确	组织在收集个人信息之时或之前必须明确采集的目的

（续）

准则	描述
授权	组织去采集、使用或披露个人信息时需征求当事人的知情和同意，但不适用的情况除外
收集、使用、披露和留存限制	个人信息必须限定于为该组织确定的目标所必需的采集。信息采集应当采取公平、合法的方式。除经个人同意或法律要求外，不得将个人信息用于采集个人信息目的以外的其他用途或披露个人信息。个人信息仅在为实现这些目的所需的时间内保留
准确性	个人信息必须准确、完整、最新，以达到使用目标
保障措施	采集的个人信息必须受到与信息敏感程度相匹配的安全保障措施的保护
透明度	组织必须向个人提供有关其个人信息的信息管理制度和实践相关的具体信息
个人访问	个人应被告知其个人信息的存在、使用和披露情况，并有权访问这些信息。个人应当能够对信息的准确性和完整性提出质疑，并酌情予以修正
合规挑战	个人应能够针对以上原则的遵从性，向负责组织或个人发起合规性质疑

3、2012 年 3 月，美国联邦贸易委员会（FTC）发布了一份报告，建议组织按照报告描述的最佳实践去设计和实施自己的隐私计划。报告中重申了 FTC 对公平信息处理原则的重视（表 2-3）。

表 2-3　美国隐私方案标准

准则	描述
发布/告知	数据采集者在采集消费者个人信息之前，必须披露对这些信息的用途和过程
选择/许可	个人信息是否采集或如何采集，以及会被用于超出采集目标之外的情况，都必须征求被采集者的意见
访问/参与	消费者可以查询，并且质疑其个人数据的准确性和完整性
诚信/安全	数据采集者需要采取合理的步骤，以确保从消费者采集的信息是准确的，并且防止未经授权使用
执行/纠正	使用可靠机制对不遵守这些公平信息实践的行为实施制裁

2.3.4 违背伦理进行数据处理的风险

利用数据歪曲事实是有可能的。达莱尔·哈夫（Darrell Huff，1954）的经典之作《统计数字会撒谎》（How to Lie with Statistics）描述了数据可以被歪曲的事实，同时创造一个事实的虚假表象。

1.时机选择【明星凌晨 2-4 点发布瓜】

2.可视化误导【如股票取一段时间上涨/取一天时间上涨】

3.定义不清晰或无效的比较【如特朗普大选，城市和农村选票】

4.偏见

1）预设结论的数据采集。

2）预感和搜索。

3）片面抽样方法。

4）背景和文化。【如宗教、男女、中美关系】

5.转换和集成数据【一般记前 4 个】
6.数据的混淆和修订【一般记前 4 个】

2.3.6 数据伦理和治理

DAMA 国际数据管理专业人士认证（CDMP）要求被认证人员签署一份正式的伦理准则，其中包括在聘用他们的组织之外进行数据处理时，也要履行处理数据的伦理义务。

第 3 章 数据治理

3.1 引言

【1、数据管理>数据治理，数据治理是数据管理的 1/11，考试时用 DAMA 的概念，而不是用国内泛指的概念】

【2、数据治理是对数据管理的管理】

【3、数据治理并不直接管理数据】

【数据治理包含 2 个核心内容：组织架构、规章制度】

1、【数据治理定义】

数据治理（Data Governance，DG）的定义是在管理数据资产过程中行使权力和管控，包括计划、监控和实施。

2、【数据治理项目范围，8 个内容】

1）战略（Strategy）。定义、交流和驱动数据战略和数据治理战略的执行。

2）制度（Policy）。设置与数据、元数据管理、访问、使用、安全和质量有关的制度。

3）标准和质量（Standards and Quality）。设置和强化数据质量、数据架构标准。

4）监督（Oversight）。在质量、制度和数据管理的关键领域提供观察、审计和纠正等措施（通常称为管理职责 Stewardship）。

5）合规（Compliance）。确保组织可以达到数据相关的监管合规性要求。

6）问题管理（Issue Management）。识别、定义、升级和处理问题，针对如下领域：

数据安全、数据访问、数据质量、合规、数据所有权、制度、标准、术语或者数据治理程序等。

7）数据管理项目（Data Management Projects）。增强提升数据管理实践的努力。

8）数据资产估值（Data Asset Valuation）。设置标准和流程，以一致的方式定义数据资产的业务价值。

3、【数据治理 4 个度量指标】3.5 节漏了遵从法规和内部数据规范

遵从法规和内部数据规范、价值、有效性、持续性。

3.1.1 业务驱动因素

【数据管理短期目标：提高数据质量】【数据管理长期目标：实现数据价值】

1、数据治理最常见的驱动因素是法规遵从性【合规】，特别是重点监控行业。例如，金融服务和医疗健康，需要引入法律所要求的治理程序。

Q：数据治理最常见的驱动因素是法规遵从性，特别是重点监控行业，下列哪些是重点监控行业？A.金融服务 B.医疗健康 C.交通运输

A：金融服务、医疗健康

2、数据治理的驱动因素大多聚焦于**减少风险或者改进流程**。

3、数据治理**不是一次性的行为**。治理数据是一个**持续性**的项目集，【或者是一个过程】以保证组织一直聚焦于能够从数据获得价值和降低有关数据的风险。

4、**数据治理要与 IT 治理区分开。**IT 治理制定关于 IT 投资、IT 应用组合和 IT 项目组合的决策，从另一个角度还包括硬件、软件和总体技术架构。IT 治理的作用是确保 IT 战略、投资与企业目标、战略的一致性。COBIT（Control Objectives for Information and Related Technology）框架提供 IT 治理标准，但是其中仅有很少部分涉及数据和信息管理。其他一些重要法规，如萨班斯法案（Sarbanes-Oxley）则覆盖企业治理、IT 治理和数据治理多个领域。相反，数据治理仅聚焦于管理数据资产和作为资产的数据。

Q：数据治理不能和 IT 治理分开

A：错。**数据治理要与 IT 治理区分开。**数据治理涉及数据资产及数据，IT 治理涉及软件、硬件。

3.1.2 目标和原则

数据治理的目标是使组织能够将数据作为资产进行管理。数据治理提供治理原则、制度、流程、整体框架、管理指标，监督数据资产管理，并指导数据管理过程中各层级的活动。为达到整体目标，数据治理程序必须包括以下几个方面。

（1）可持续发展（Sustainable）

治理程序必须富有吸引力。**它不是以一个项目作为终点，而是一个持续的过程。**需要把它作为整个组织的责任。数据治理必须改变数据的应用和管理方式，但也不代表着组织要作巨大的更新和颠覆。数据治理是超越一次性数据治理组件实施可持续发展路径的管理变革。**可持续的数据治理依靠于业务领导、发起者和所有者的支持。**

（2）嵌入式（Embedded）

数据治理不是一个附加管理流程。【不能违背伦理要求】数据治理活动需要融合软件开发方法、数据分析应用、主数据管理和风险管理。

Q：什么是嵌入式？

（3）可度量（Measured）

数据治理做得好有积极的财务影响，但要证明这一影响，就需要了解起始过程并计划可度量的改进方案。【不能完全定性，需要定量】

3.1.3 基本概念

【非常重要】

正如财务审计人员实际上并不执行财务管理一样，数据治理确保数据被恰当地管理而不是直接管理数据（参见第 15 章）。数据治理相当于将监督和执行的职责分离。

【数据治理是对数据管理的管理，数据治理本身不管理数据】

2.数据治理组织

【有考题】

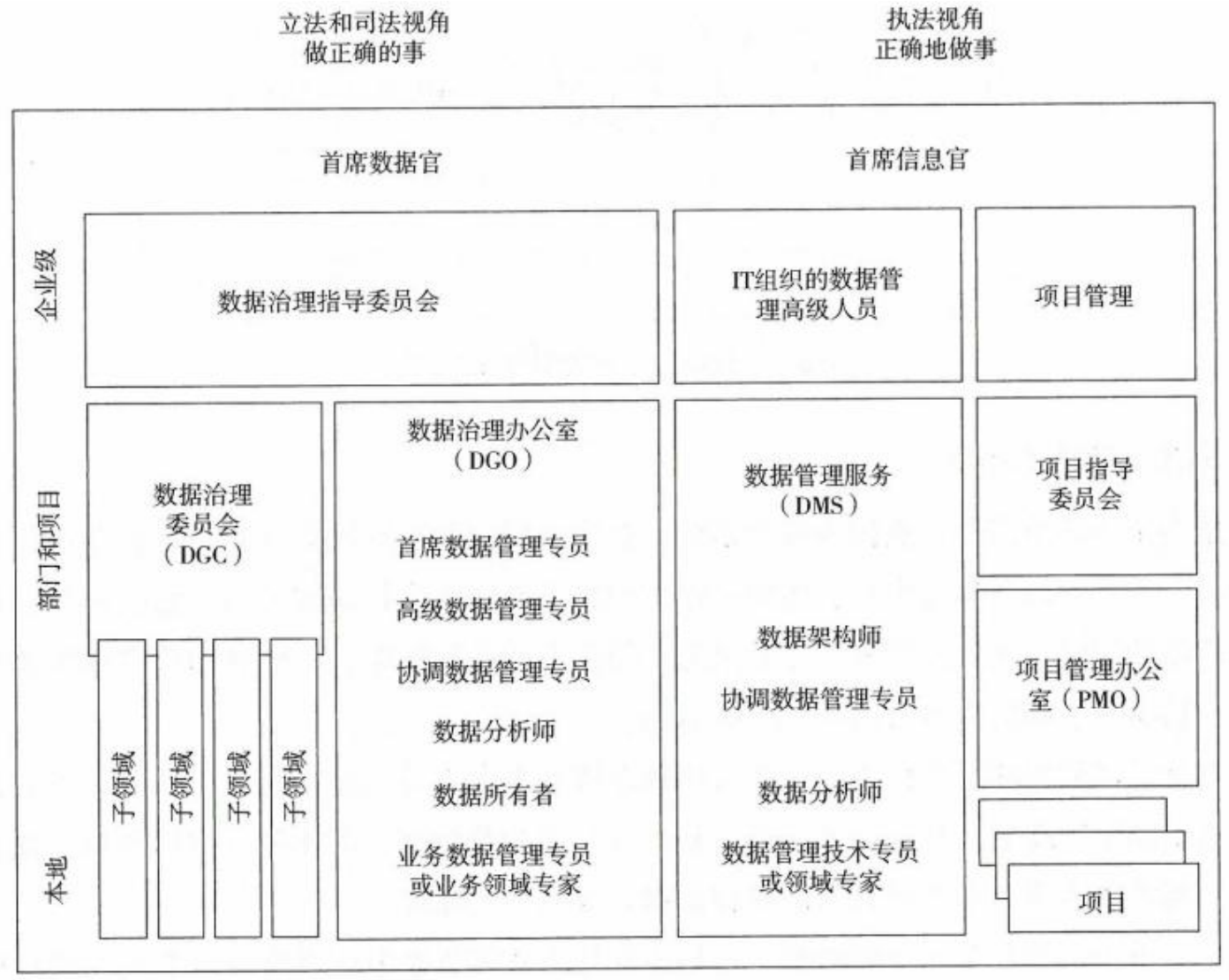

图 3-3　数据治理组织的组成部分

Q1：决定企业数据是否上云，是谁的决策？首席数据官，首席信息官？

A1：首席信息官（技术层面）

Q2：决定使用哪个数据库（oracle 达梦），是谁的决策？首席数据官，首席信息官？

A2：首席信息官（技术层面）

Q3：决定数据价值评估体系，带领企业数字化转型，是谁的决策？首席数据官，首席信息官？

A3：首席数据官

表 3-1　典型数据治理委员会

数据治理机构	说明
数据治理指导委员会	组织中数据治理的主要和最高权威组织，负责监督、支持和资助数据治理活动。由跨职能的高级管理人员组成 通常根据 DGC 和 CDO 的建议，为数据治理发起的活动提供资金。该委员会可能会反过来受到来自更高级别组织或者委员会的监督
数据治理委员会	管理数据治理规划（如制度或指标的制定）、问题和升级处理。根据所采用的运营模型由相关管理层人员组成（参见图 3-4）
数据治理办公室	持续关注所有 DAMA 知识领域的企业级数据定义和数据管理标准，由称为数据管理专员、数据保管人和数据拥有者等协调角色组成
数据管理团队	与项目团队在数据定义和数据管理标准方面进行协作、咨询，由聚焦于一个或者更多领域或项目的成员组成，包括业务数据管理专员、技术数据管理专员或者数据分析师（注：偏重管理职责）
本地数据治理委员会	大型组织可能有部门级或数据治理指导委员会分部，在企业数据治理委员会（DGC）的指导下主持工作。小型组织应该避免这种复杂设置

Q：数据治理委员会起草编写数据的标准、规划？
A：错误，数据治理委员会成员都是部门老大，**审阅、发布标准，而不是起草编写**。

3.数据治理运营模型类型
【3 种类型运营框架：集中式、分布式、联邦式】
在集中式管理模式中，数据治理组织监督所有业务领域中的活动。【阿里巴巴，基本是数据】
在分布式管理模式中，每个业务单元中采用相同的数据治理运营模型和标准。【华为，华为本部、华为欧盟，业务条线多，分公司到处有，无法统一规定】
在联邦式管理模式中，数据治理组织与多个业务单元协同，以维护一致的定义和标准。【央企，各地有省公司、分公司，集团能定集团定，集团定不了分公司定，取最大公约数】

4.数据管理职责
数据管理职责（Data Stewardship）描述了数据管理岗位的责任，以确保数据资产得到有效控制和使用。

5.数据管理岗位的类型
管理专员（Steward，直译为管家，本书译为管理专员）指其职责是为别人管理财产的人。数据管理专员代表他人的利益并为组织的最佳利益来管理数据资产（McGilvray，2008）。数据管理专员代表所有相关方的利益，必须从企业的角度来确保企业数据的高质量和有效使用。有效的数据管理专员对数据治理活动负责，并有部分时间专门从事这些活动。
Q：数据管理专员/数据管家，是 IT 部门还是业务部门？
A：业务团队。

3.2 活动
3.2.2 制定数据治理战略
1.定义数据治理运营框架（3 种运管框架）
构建组织运管框架时考虑以下几方面：
1）数据对组织的价值。如果一个组织出售数据，显然数据治理具有巨大的业务影响力。将数据作为最有价值事物的组织（如 Facebook、亚马逊）将需要一个反映数据角色的运营模式。对于数据是操作润滑剂的组织，数据治理形式就不那么严肃了。
2）业务模式。分散式与集中式、本地化与国际化等是影响业务发生方式以及如何定义数据治理运营模式的因素。

与特定 IT 策略、数据架构和应用程序集成功能的链接，应反映在目标运营框架设计中（图 3-6）。

3）文化因素。就像个人接受行为准则、适应变化的过程一样，一些组织也会抵制制度和原则的实施。开展治理战略需要提倡一种与组织文化相适应的运营模式，同时持续地进行变革。

4）监管影响。与受监管程度较低的组织相比，受监管程度较高的组织具有不同的数据。治理心态和运营模式。可能还与风险管理或法律团队有联系。

6.评估法规遵从性要求

对管理信息资产有重大影响的部分全球性法规：

1）会计准则。政府会计准则委员会（GASB）和财务会计准则委员会（FASB）的会计准则对（在美国）管理信息资产具有重大影响。

2）BCBS 239（巴塞尔银行监管委员会）和巴塞尔 II。这是指有效的风险数据汇总和风险报告原则，是一整套针对银行的法规。自 2006 年以来，在欧盟国家开展业务的金融机构必须报告证明流动性的标准信息。

3）CPG 235。澳大利亚审慎监管局（APRA）负责监督银行和保险实体，公布了一些标准和指南以帮助被监管对象满足这些标准，其中包括 CPG235，一个管理数据风险的标准。制定这个标准的目的是解决数据风险的来源，并在整个生命周期中管理数据。

4）PCI-DSS。支付卡行业数据安全标准（PCI-DSS）。

5）偿付能力标准 II。欧盟法规，类似巴塞尔协议 II，适用于保险行业。

6）隐私法。适用于各地区、各主权实体和国际的法律。

Q：请从下列选项中选择在美国很受重视，但在中国重视程度较低的法规？

A.会计准则　　B.BCBS239（巴塞尔银行监管委员会）和巴塞尔 II

C.CPG235（一个管理数据风险的标准）　　D.PCI-DSS（支付卡行业数据安全标准）

A：D.PCI-DSS，支付卡行业数据安全标准。中美消费差异，美国更多的使用信用卡，中国更多支付宝/微信支付。

3.3 工具和方法

3.3.1 线上应用/网站

数据治理也应该能够线上体现，可以通过中心门户或者协作门户提供核心文档。【数据治理交付物】

3.5 度量指标【数据治理 4 个度量指标】

遵从法规和内部数据规范、价值、有效性、持续性。

（1）价值

1）对业务目标的贡献。

2）风险的降低。

3）运营效率的提高。

（2）有效性

1）目标的实现。

2）扩展数据管理专员正在使用的相关工具。

3）沟通的有效性。

4）培训的有效性。

5）采纳变革的速度。

（3）可持续性

1）制度和流程的执行情况（即它们是否正常工作）。

2）标准和规程的遵从情况（即员工是否在必要时遵守指导和改变行为）

第 4 章 数据架构

4.1 引言

1、【架构定义】

将架构定义为“系统的基本结构，具体体现在架构构成中的**组件、组件之间的相互关系**以及**管理其设计和演变的原则**”。

Q：架构表示什么？

A：1、组件；2、组件之间的相互关系；3、管理其设计和演变的原则。

2、【解决当下的问题，预测未来的问题】数据架构的构件包括当前状态的描述、数据需求的定义、数据整合的指引、数据管控策略中要求的数据资产管理规范。

4.1.1 业务驱动因素

数据架构的目标是在业务战略和技术实现之间建立起一座通畅的桥梁，数据架构是企业架构中的一部分，其主要职责为：

1）利用新兴技术所带来的业务优势，从战略上帮助组织快速改变产品、服务和数据。

2）将业务需求转换为数据和应用需求，以确保能够为业务流程处理提供有效数据。

3）管理复杂数据和信息，并传递至整个企业。

4）确保业务和 IT 技术保持一致。

5）为企业改革、转型和提高适应性提供支撑

4.1.3 基本概念

1.企业架构类型

数据架构的设计及实施与其他架构紧密相关，企业架构包括**业务架构、数据架构、应用架构和技术架构**。

Q：数据架构师要规划业务规则。

A：错，数据架构师理论上负责数据架构规划，业务规划有业务架构师。

3.企业数据架构

企业数据架构描述必须包括企业数据模型（如数据结构和数据规范）和数据流设计。

1）企业数据模型。企业数据模型是一个整体的、企业级的、独立实施的概念或逻辑数据模型，为企业提供通用的、一致的数据视图。通常用于表示高层级简化的数据模型，也表示了不同抽象层级。企业数据模型包括数据实体（如业务概念）、数据实体间关系、关键业务规则和一些关键属性，它为所有数据和数据相关的项目奠定了基础。任何项目级的数据模型必须基于企业数据模型设计。企业数据模型应该由利益相关方审核，以便它能一致有效地代表企业。**【概念模型、逻辑模型】**

【物理模型不是企业数据架构模型】

2）数据流设计**【也叫数据价值链、数据分布、数据流程】**。定义数据库、应用、平台和网络（组件）之间的需求和主蓝图。这些数据流展示了数据在业务流程、不同存储位置、业务角色和技术组件间的流动。

Q：企业数据架构包含什么内容

A：业务数据模型，数据流设计/数据价值链/数据分布/数据流程。

4.2 活动

简化数据和企业架构所面临的复杂问题，基于以下两种方式解决：**面向质量、面向创新**。

4.2.1 建立企业数据架构

项目中的企业数据架构角色依赖软件开发过程。因采用的方法不同，将架构活动嵌入到项目中的过程也不同，具体采用的方式有以下**三种【瀑布方式、迭代方式、敏捷方式】**：

1）瀑布方式。作为整个企业设计的一部分，在连续阶段中理解需求和构建系统。这种方法包括设计用于控制变化的关口。按照这种方式开展数据架构活动通常没有太多问题，但需确保能够从企业视角设计架构和考虑问题，以避免局限性。**【数据架构只能先做瀑布式，建大厦四梁八柱】**

2）迭代方式。逐步学习和构建（如小型瀑布模型）。这种方式适合总体需求模糊的原型。这种方式在启动阶段至关重要，最好是早期迭代中创建一个全面的数据设计。

3）敏捷方式。这种方式是指在离散的交付包中学习，构建并测试（称为“sprints”冲刺）。离散的交付包很小，如需要丢弃，也不会损失太多。敏捷模型（Scrum，快速开发，统一流程）能提高目标导向的模型，强调用户界面设计、

软件设计和系统行为。使用数据模型、数据捕获、数据存储和数据分布规范完成这些方法。当程序员和数据架构师有很强的工作联系，并且他们的标准和指南兼容时，可以采用 DevOps 方法。DevOps 是一种新兴且流行的敏捷方法，它可以帮助改进数据设计，并使得数据设计的选择更有效。

4.3 工具

3 种数据架构工具：**数据建模**工具、**资产管理**软件、**图形设计**应用。

4.6 数据架构治理

4.6.2 度量指标

企业数据架构衡量指标反映了架构目标：架构标准接受度、实施趋势、业务价值度量指标。数据架构衡量工作通常作为项目总体业务客户满意度的一部分，每年开展一次。【请专家评估】

（没有技术价值度量指标、合规价值度量指标）

第 5 章 数据建模和设计

5.1.引言

【数据建模是什么】数据建模→建表结构→存储数据

【建模重要性】：影响存储成本、影响性能体验（8 小时出报告→7 秒出报告）

【6 种模式】【一定会考】

数据可以采用多种不同的模式来表示。其中最为常见的 **6 种模式**分别是：**关系模式、多维模式、面向对象模式、事实模式、时间序列模式和 NoSQL 模式**。

【3 层模型】【一定会考】

按照描述详细程度的不同，每种模式又可以分为 3 层模型：**概念模型、逻辑模型和物理模型**。每种模型都包含一系列组件，如实体、关系、事实、键和属性。

Q1：数据采用哪 6 种模式表示？

A1：关系、多维、面向对象、时间序列、事实、NoSQL。

Q2：数据采用哪 3 种模型表示？

A2：概念、逻辑、物理。

【视频补充概念模型、逻辑模型、物理模型区别】

概念模型（偏业务）：只有实体的名称，没有属性。【实体、属性、关系】

逻辑模型（偏业务）：有实体的名称，也有属性。【实体、属性、关系】

物理模型（偏技术）：【表、字段、外键】

1）标准和规则 a.不能有空格；b.长度不能超过 30byte；c.不能由数字开头，一定要由字母开头，如不能 123catch

2）其他许多工作。

Q1：概念模型、逻辑模型、物理模型区别

A1：a.概念、逻辑模型主要由业务部门主导，物理模型主要由技术部门主导；b.概念、逻辑模型属于数据架构产物（业务），物理模型属于数据建模输出（DBA）。

Q2：下图是什么模型？

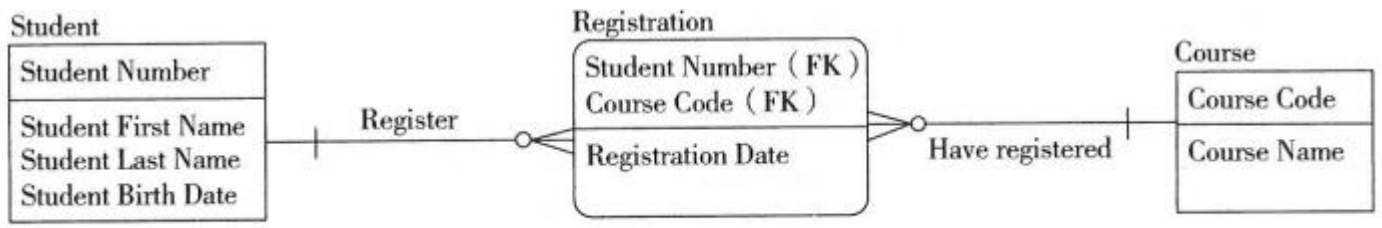

图 5-11 独立和非独立实体

A2：逻辑模型，物理模型中不允许有空格，比概念模型要详细。

【SQL（Structured query language）4 种命令】

1、query，select

2、DDL：创建表结构

3、DML：改变数据

4、DLL：rollback，commit 语句。

5.1.3 基本概念

3.数据模型组件

大多数数据模型都包含基本相同的组件：实体、关系、属性和域。

1、实体：与业务相关的所有内容都是实体（业务部门主导）。

2、关系：4 种，1 对多，多对 1，多对多（不允许此类形式，违反三范式，如老师对学生，多对多关系，需要增加中间表），1 对 1（不允许此类形式，合并）。

Q1：实体与实体之间有几种关系？

A1：4 种，1 对多，多对 1，多对多（不允许此类形式，违反三范式，如老师对学生，多对多关系，需要增加中间表），1 对 1（不允许此类形式，合并）

Q2：下图中，学生和老师是什么关系？

A.1 对多　　B.多对 1　　C.多对多　　D.基于本图没直接关系

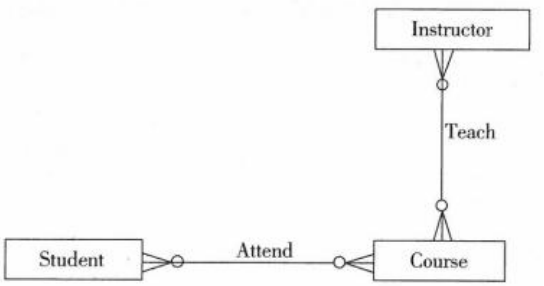

图 5-3　关系

A2：学生、老师基于本图没有直接关系。

3、属性：属性（Attribute）是一种定义、描述或度量实体某方面的性质。属性可能包含域，这将在后面展开讨论。实体中属性的物理展现为表、视图、文档、图形或文件中的列、字段、标记或节点等。【有实体肯定有属性】【老师的属性：姓名、性别、身高、体重】

Q3：一个实体有多少个属性，应该问谁？

A3：业务部门。

4、标识符：标识符（Identifiers）也称为**键**，是唯一标识实体实例的**一个或多个**属性的集合。本节根据键的结构（单一键、组合键、复合键、代理键）和功能（候选键、主键、备用键）进行分类。

Q4：身份证号码能否用为标识符？

A4：可用，但不安全，基本不会使用，存在数据泄露风险。

Q5：姓名是一个属性，还是 2 个属性？

A5：遵从原子性，不可再分，应属于 2 个属性。

5、域：在数据建模中，域（Domain）代表某一属性可被赋予的全部可能取值。域可以用不同的方式来表达（参见本章节末的要点）。域提供了一种将属性特征标准化的方法。【一种便携修改方法】

域可以用多种不同的方式定义：

1）数据类型（Data Type）。域中的某一属性中的数据有特定的标准类型要求。例如，整数、字符（30 字节）和日期都属于数据类型域。

2）数据格式（Data Format）。使用包括模板和掩码等格式的域，如邮政编码和电话号码以及字符的限制（仅用字母数字代码，字母数字代码和某些特殊符号等），用这些格式来定义有效值。

3）列表（List）。含有有限个值的域。很多人都非常熟悉下拉列表就属于此类。例如，订单状态域的值可以限制在订单开立、发货、订单结束、退货等状态。

4）范围（Range）。允许相同数据类型的所有值在一个或多个最小值和/或最大值之间的域。有些范围可以是开放式的。例如，订单送货日期（Order Delivery Date）必须在订单下达日期（Order Date）之后的三个月之内。

5）基于规则（Rule-Based）。域内的值必须符合一定的规则才能够成为有效值。规则包括将关系或组合中的值与计算值或其他属性值进行对比。例如，物品价格必须高于物品成本。

Q6：域是干嘛用的？

A6：把许多类型加载一起，通过改变域改变属性。

4.数据建模方法

常见的 6 种数据建模方法：关系建模、维度建模、面向对象建模、基于事实建模、基于时间建模和非关系型建模。

Q：张三在 2022 年上海卖出了多少辆车，有几个维度几个指标？

A：4 个维度（张三、2022、上海、车），1 个指标（车）。

（2）维度建模

1）事实表，在维度模型中，事实表（Fact Tables）的行对应于特定的数值型度量值。

2）维度表，维度表（Dimension Tables）表示业务的重要对象，并且主要包含文字描述。维度是事实表的入口点或链接，充当“查询”或“报表”约束的主要来源。维度通常是高度反范式的，通常占总数据的 10%左右。**【维度越多越复杂，业务部门主导】**

维度也有一些属性，它们以不同的速率发生变化。渐变类的维度【SCD】根据变化的速率和类型来管理变化。3 种主要的变化类型有时被称为 ORC，具体如下：

①第一类，覆盖（Overwrite）。新值覆盖旧值。

②第二类，新行（New Row）。新值写在新行中，旧行被标记为非当前值。

③第三类，新列（New Column）。一个值的多个实例列在同一行的不同列中，而一个新值意味着将系列中的值向下一点写入，以便在前面为新值留出空间。最后一个值被丢弃。

3）雪花模型。

雪花模型（Snowflaking）的含义是将星型模式中的平面、单表、维度结构规范为相应的组件层次结构或网络结构。

4）粒度。**【粒度越细，灵活性越高，如防疫，从社区到户】**

粒度（Grain）这一概念是指事实表中的单行数据的含义或者描述，这是每行都有的最详细信息。定义一个事实表中的粒度是维度建模的关键步骤之一。例如，如果一个维度模型用于度量学生注册过程，粒度可能为学生、日期和班级。

Q：维度建模由几个表构成？

A：2 个，事实表、维度表。

通常有 4 类 NoSQL 数据库：文档数据库、键值数据库、列数据库和图数据库。

Q：非关系型数据库有哪几种？

A：4 种，文档、列、图、键值。

5.数据模型级别

1975 年，美国国家标准协会的标准规划与需求委员会（SPARC）发布了数据库管理的三重模式，它们分别是：

1）概念模式（Conceptual）。概念模式体现了正在数据库中建模企业的“真实世界”视图，代表了企业当前的“最佳模式”或“经营方式”。

2）外模式（External）。它是数据库管理系统的各个用户操作与特定需求相关企业模型的子集。称为“外模式”。

3）内模式（Internal）。数据的“机器视图”由内模式描述。该模式描述了企业信息的存储表示形式（Hay，2011）。这 3 个层次通常分别在概念层次、逻辑层次和物理层次上进行细节展现。在项目中，概念数据建模和逻辑数据建模是需求规划和分析活动的一部分，而物理数据建模属于设计活动。

概念数据模型（Conceptual Data Model，CDM）

逻辑数据模型（Logical Data Model，LDM）

物理数据模型（Physical Data Model，PDM）

Q：下图一定是物理数据模型？

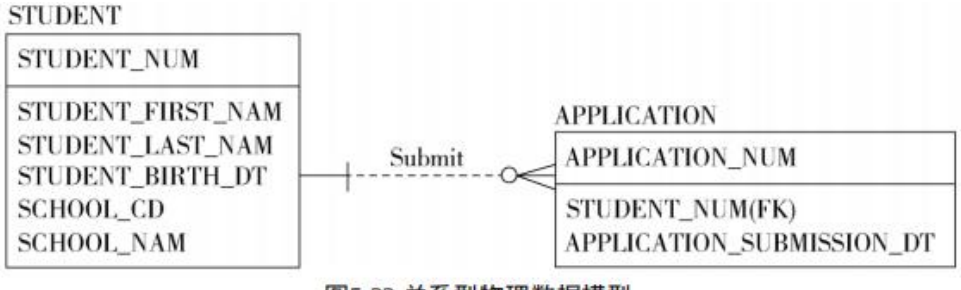

图5-22 关系型物理数据模型

A：错，最有可能是物理模型，也有可能是逻辑模型。

4）逆规范化【牺牲空间，换取时间】【只在 OLAP 中做，OLTP 交易型中不会有】

逆规范化（Denormalization）是将符合范式规则的逻辑数据模型经过慎重考虑后，转换成一些带冗余数据的物理表。换言之，逆规范化有意将一个属性放在多个位置。将数据逆规范化有很多原因，**最重要的是提高性能**，如：
①提前组合来自多个其他表的数据，以避免代价高昂的运行时连接。
②创建更小的、预先过滤的数据副本，以减少昂贵的运行时计算和/或大型表的扫描。
③预先计算和存储昂贵的数据计算结果，以避免运行时系统资源竞争。

5.2 活动

5.2.2 建立数据模型

1.正向工程

正向工程是指从需求开始构建新应用程序的过程。首先需要通过建立概念模型来理解需求的范围和核心的术语；然后建立逻辑模型来详细描述业务过程；最后是通过具体的建表语句来实现物理模型。

2.逆向工程

逆向工程是记录现有数据库的过程。**物理数据建模通常是第一步**，以了解现有系统的技术设计；**逻辑数据建模是第二步**，以记录现有系统满足业务的解决方案；**概念数据建模是第三步**，用于记录现有系统中的范围和关键术语。大多数数据建模工具支持各种数据库的逆向工程。

实际工作中，可能只有物理模型，没有其他模型，直接从物理模型开始。

Q：逆向工程在主数据管理/数据质量中起到重要作用。

A：错，逆向工程在元数据管理中起到非常重要作用。

5.3.6 行业数据模型

行业数据模型是为整个行业预建的数据模型，包括**医疗保健、电信、保险、银行、制造业**等行业。这些模型通常范围广泛且内容详细。一些行业的数据模型包含数千个实体和属性。可以通过供应商购买行业数据模型，也可以通过 ARTS（零售）、SID（通信）或 ACORD（保险）等行业组织获得。

【先看联合国是否有相关标准、再看是否有国标、接着看行业标准、是否有地方性标准、团标、最后没有自建】

5.4.2 数据库设计中的最佳实践【可能会考】

在设计和构建数据库时，DBA 应牢记以下 PRISM 设计原则：

1）性能和易用性（Performance and Ease of Use）。确保用户可快速、轻松地访问数据，从而最大限度地提高应用程序和数据的业务价值。

2）可重用性（Reusability）。应确保数据库结构在适当的情况下，能够被多个应用重复使用，并且可用于多种目的（如业务分析、质量改进、战略规划、客户关系管理和流程改进）。避免将数据库、数据结构或数据对象耦合到单个应用程序中。

3）完整性（Integrity）。无论语境如何，数据应始终具有有效的业务含义和价值，并且应始终反映业务的有效状态。实施尽可能接近数据的数据完整性约束，并立即检测并报告数据完整性约束的违规行为。

4）安全性（Security）。应始终及时向授权用户提供真实准确的数据，且仅限授权用户使用。必须满足所有利益相关方（包括客户、业务合作伙伴和政府监管机构）的隐私要求。强化数据安全性，就像数据完整性检查一样，执行数据的安全性约束检查，尽可能确保数据的安全性。如果检查发现存在违反数据安全性约束的情况，则立刻报告违规行为。

5）可维护性（Maintainability）。确保创建、存储、维护、使用和处置数据的成本不超过其对组织的价值，以能够产生价值的成本方式执行所有数据工作；确保尽可能快速地响应业务流程和新业务需求的变化。

第 6 章 数据存储和操作

6.1 引言

数据库管理员（DBA）在数据存储和操作上述两个方面中都扮演着重要的角色。DBA 这个角色是数据专业中最常见，也是最被广泛接纳的角色。数据库管理实践可能也是数据管理实践领域最成熟的。在数据安全方面，DBA 同样发挥着主导作用。

Q：什么是数据专业中最常见，也是最被广泛接纳的角色？

A：DBA 数据库管理员。

6.1.3 基本概念

1.数据库术语

（4）节点

一台单独的计算机作为分布式数据库处理数据或者存储数据的一个部分。

Q：什么是节点？

A：一台服务器在集群里就是节点。

3.管理员

数据库管理员（DBA）是数据专业中最常见、也是最广泛被接纳的角色。DBA 在数据存储与操作活动中承担着主导角色，在数据安全活动及物理模型建模、数据库设计活动中也是关键的角色。DBA 为开发环境、测试环境、QA 环境及其他特殊数据库环境提供支持。

Q：DBA 为哪些环境提供支持？

A：生产环境、开发环境、测试环境、QA 环境及其他特殊数据库环境。

5.数据处理类型

数据库处理有两种基本类型：ACID 和 BASE。

（1）ACID

缩写词 ACID 是在 20 世纪 80 年代末期出现的一个合成词，含义是保证数据库事务可靠性不可或缺的约束。数十年来，它为事务处理提供了坚实的基础。

1）原子性（Atomicity）。所有操作要么都完成，要么一个也不完成。因此，如果事务中的某部分失败，那么整个事务就都会失败。

2）一致性（Consistency）。事务必须时刻完全符合系统定义的规则，未完成的事务必须回退。

3）隔离性（Isolation）。每个事务都是独立的。

4）持久性（Durability）。事务一旦完成，就不可撤销。

在关系型数据库存储中，ACID 相关技术是最主要的工具，通常采用 SQL 作为接口。

（2）BASE

通常在**大数据环境**里会使用 BASE 类型的系统，如大型互联网公司和社交媒体公司。因为，它们的业务场景任何时候都不需要立即准确地拿到所有数据。

Q：ACID 用于大数据环境？

A：错，ACID 用于关系型数据库，不能有任何出错，在大数据环境里使用 BASE 系统。

（3）CAP

CAP 定理指的是分布式系统不可能同时满足 ACID 的所有要求。系统规模越大，满足的要求点越少。分布式系统必须在各种属性（要求）间进行权衡。

7.数据库环境

（1）生产环境

（2）非生产环境

常见的非生产环境包括**开发环境、测试环境、支持环境和特别用途环境**。

2）测试环境。

测试环境通常用于执行质量保证和用户验收测试，有些情况下，也用于压力测试或性能测试。为了防止测试结果因为环境差异而失真，理想的测试环境应该与生产环境使用完全一样的软硬件配置，这一点对于性能测试来说尤为重要。测试或许可以通过网络连接读取生产数据，但是，测试环境永远不要写数据到生产系统。

测试环境通常用于：

①质量保证测试（QA）。依据需求进行功能测试。

②集成测试。将独立开发或更新的多个模块作为一个整体系统进行测试。

③用户验收测试（UAT）。从用户视角进行系统功能测试。在这个场景下，测试用例是最常见的测试输入。

④性能测试。任何时候都可考虑进行的高复杂度或大容量的测试，而不必等到下班后，或者对生产系统的高峰时间产生不利的影响

Q：有哪些测试环境？

A：质量保证测试（QA）、集成测试、用户验收测试（UAT）、性能测试。

数据沙盒的价值如同进行一场概念验证（Proof-of Concept，POC）。

Q：数据沙盒或实验环境是测试环境？

A：错，不是测试环境，测试环境是 1→end，沙盒是 0→1。

6.4 方法

6.4.2 物理命名规则

Q：物理模型一定能要遵从命名标准？

A：对

Q：实体和属性需要遵从命名标准？

A：错，逻辑概念模型不需要遵从命名规则，但物理模型表、字段需要遵从。

ISO/IEC 11179 元数据注册表（Metadata Registries，MDR）

6.6 数据存储和操作治理

6.6.1 度量指标

4 部分：数据存储、性能度量、操作度量、服务度量。

数据存储的度量指标，包括：

1）数据库类型的数量。

2）汇总交易统计。

3）容量指标。

4）已使用存储的数量。

5）存储容器的数量。

6）数据对象中已提交和未提交块或页的数量。

7）数据队列。

8）存储服务使用情况。

9）对存储服务提出的请求数量。

10）对使用服务的应用程序性能的改进。

性能度量评估指标，包括：

1）事务频率和数量。

2）查询性能。

3）API 服务性能。

操作度量指标，包括：

1）有关数据检索时间的汇总统计。

2）备份的大小。

3）数据质量评估。

4）可用性。

服务度量指标，包括：

1）按类型的问题提交、解决和升级数量。

2）问题解决时间。

DBA 应与数据架构师和数据质量团队一起讨论度量指标的需求。

第 7 章 数据安全

7.1 引言

网络安全不等于数据安全。

7.1.1 业务驱动因素

降低风险和促进业务增长是数据安全活动的主要驱动因素。确保组织数据安全，可降低风险并增加竞争优势。安全本身就是宝贵的资产。

7.1.2 目标和原则

1.目标

数据安全活动目标，包括以下几个方面：

1）支持适当访问并防止对企业数据资产的不当访问。

2）支持对隐私、保护和保密制度、法规的遵从。

3）确保满足利益相关方对隐私和保密的要求。

Q：支持适当访问并防止对企业数据资产的不当访问？防止对企业数据资产的不当访问，并支持适当访问？

A：支持适当访问并防止对企业数据资产的不当访问。【首先共享开放，而后不共享】

6.安全过程

数据安全需求和过程分为 4 个方面，即 4A：访问（Access）、审计（Audit）、验证（Authentication）和授权（Authorization）、权限（Entitlement）。

Q：4A1E 有敏捷嘛？

A：没有，访问、审计、验证、授权、权限。

8.加密

加密（Encryption）是将纯文本转换为复杂代码，以隐藏特权信息、验证传送完整性或验证发送者身份的过程。加密数据不能在没有解密密钥或算法的情况下读取。解密密钥或算法通常单独存储，不能基于同一数据集中的其他数据元素来进行计算。**加密方法主要有 3 种类型，即哈希、对称加密、非对称加密**，其复杂程度和密钥结构各不相同。

9.混淆或脱敏【肯定会考】

数据混淆或脱敏是解决数据使用过程中的一种安全手段。数据脱敏分为两种类型：**静态脱敏和动态脱敏**。

静态脱敏按执行方式又可以分为**不落地脱敏和落地脱敏**。

（1）静态数据脱敏

静态数据脱敏（Persistent Data Masking）永久且不可逆转地更改数据。这种类型的脱敏通常**不会在生产环境中使用**，而是在**生产环境和开发（或测试）环境**之间运用。静态脱敏虽然会更改数据，但数据仍可用于测试、应用程序、报表等。

1）不落地脱敏（In-flight Persistent Masking）。当在数据源（通常是生产环境）和目标（通常是非生产）环境之间移动需要脱敏或混淆处理时，会采用不落地脱敏。由于不会留下中间文件或带有未脱敏数据的数据库，不落地脱敏方式非常安全。另外，如果部分数据在脱敏过程中遇到问题，则可重新运行脱敏过程。

2）落地脱敏（In-place Persistent Masking）。当数据源和目标相同时，可使用落地脱敏。从数据源中读取未脱敏数据，进行脱敏操作后直接覆盖原始数据。假定当前位置不应该保留敏感数据，需要降低风险，或者在安全位置中另有数据副本，在移动至不安全位置之前就应当进行脱敏处理。这个过程存在一定的风险，如果在脱敏过程中进程失败，那么很难将数据还原为可用格式。该技术在一些细分领域中还有些用途，但一般来说，不落地脱敏能更安全地满足项目需求。

（2）动态数据脱敏

动态数据脱敏（Dynamic Data Masking）是在**不更改基础数据**的情况下，在最终用户或系统中改变数据的外观。当用户需要访问某些敏感的生产数据（但不是全部数据）时，这就相当有用。例如，在数据库中，假设社会保障号码存储为 123456789，那么采用此方法后，呼叫中心人员需要验证 通话对 象时， 看到的 该数 据显示的是* * * - * * -6789。

（3）脱敏方法

可以脱敏或混淆数据的方法有以下几种：

1）替换（Substitution）。将字符或整数值替换为查找或标准模式中的字符或整数值。例如，可以用列表中的随机值替换名字。

2）混排（Shuffling）。在一个记录中交换相同类型的数据元素或者在不同行之间交换同一属性的数据元素。例如，在供应商发票中混排供应商名称，以便将发票上的原始供应商替换为其他有效供应商。

3）时空变异（Temporal Variance）。把日期前后移动若干天（小到足以保留趋势）**【大到足以使它无法推测原来数据】**，足以使它无法识别。

4）数值变异（Value Variance）。应用一个随机因素（正负一个百分比，小到足以保持趋势）**【大到足以使它无法推测原来数据】**，重要到足以使它不可识别。

5）取消或删除（Nulling or Deleting）。删除不应出现在测试系统中的数据。

6）随机选择（Randomization）。将部分或全部数据元素替换为随机字符或一系列单个字符。

7）加密技术（Encryption）。通过密码代码将可识别、有意义的字符流转换为不可识别的字符流。**【加密是一种特殊的脱敏方法】**

8）表达式脱敏（Expression Masking）。将所有值更改为一个表达式的结果。例如，用一个简单的表达式将一个大型自由格式数据库字段中的所有值（可能包含机密数据）强制编码为“这是个注释字段”。

9）键值脱敏（Key Masking）。指定的脱敏算法/进程的结果必须是唯一且可重复的，用于数据库键值字段（或类似字段）脱敏。这种类型脱敏对用于测试需要保持数据在组织范围内的完整性极为重要。

10.网络安全术语

在未经测试以确保真正安全时，新建的网络和网站是不完整的。在渗透测试（Penetration Testing）中，来自组织本身或从外部安全公司聘任的“白帽”黑客试图从外部侵入系统，正如恶意黑客一样，试图识别系统漏洞。通过渗透测试发现的漏洞应该在应用程序正式发布之前予以解决。

Q：黑客都是坏蛋嘛？

A：错误，白帽黑客渗透测试。

11.数据安全类型

（1）设施安全（数据中心放在温度较低的地方，科罗拉、贵州，不同国家对待设施安全举措不同）

（2）设备安全（移动设备）

（3）凭据安全（ID、密码、面部识别）

（4）电子通信安全（拦截）

12.数据安全制约因素

数据安全制约因素包括数据的保密等级和监管要求。

（1）机密数据（5 级分类）

1）对普通受众公开（For General Audiences）。可向任何人（包括公众）提供的信息。

2）仅内部使用（Internal Use Only）。仅限员工或成员使用的信息，但信息分享的风险很小。这种信息仅供内部使用、可在组织外部显示或讨论，但不得复制。

3）机密（Confidential）。若无恰当的保密协议或类似内容，不得在组织以外共享。不得与其他客户共享客户机密信息。

4）受限机密（Restricted Confidential）。受限机密要求个人通过许可才能获得资格，仅限于特定“需要知道”的个人。

5）绝密（Registered Confidential）。信息机密程度非常高，任何信息访问者都必须签署一份法律协议才能访问数据，并承担保密责任。

（2）监管限制的数据

①个人身份信息（PII）。个人身份信息也称为个人隐私信息（PPI），包括任何可以识别个人或一组人的信息，如姓名、地址、电话号码、日程安排、政府 ID 号码、账号数据报、年龄、种族、宗教、生日、家庭成员或朋友的姓名、职业和薪酬等数据。高度类似的保护行动可以满足欧盟隐私指令、加拿大隐私法（PIPEDA）、日本 PIP 法案（2003）、PCI 标准、美国 FTC 要求、GLB 与 FTC 标准以及大多数信息安全泄露法案的要求。

②财务敏感数据。所有财务信息，包括可能称为“股东”或“内部人士”的数据以及尚未公开披露的所有当前财务信息。另外，还包括未公布的任何未来业务计划、计划中的并购或分拆、公司重大问题的非公开报告、高级管理层的意外变化、综合的销售以及订单和账单数据。对所有这些信息都可归为此类别并采用相同的保护策略。在美国，这些信息受内幕交易法、SOX（萨班斯-奥克斯利法案）或 GLBA（格兰姆-利奇-布莱利/金融服务现代化法案）的管辖。注意：萨班斯-奥克斯利法案限制和管理谁可以更改财务数据，从而确保数据完整性，而内幕交易法则对所有能够查

看财务数据的人都有影响。
③医疗敏感数据/个人健康信息（PHI）。有关个人健康或医疗的所有信息。在美国，HIPAA（健康信息可移植性和责任法）涵盖了这些信息。其他国家/地区也有关于保护个人信息和医疗信息的限制性法律。因此，要确保公司法律顾问意识到在业务开展或拥有客户的国家/地区，组织需要遵守法律要求的必要性。
④教育记录。有关个人教育的所有信息。在美国，这些信息由 FERPA（家庭教育权利和隐私法）涵盖。
Q：哪些数据需要被保护？
A：个人身份信息 PII、财务敏感数据、医疗敏感数据/个人健康信息 PHI、教育记录。

13.系统安全风险
（2）滥用合法特权
7.4 方法
7.4.1 应用 CRUD 矩阵【一定会考】
创建和使用数据-流程矩阵和数据-角色关系（CRUD—创建、读取、更新、删除）矩阵有助于映射数据访问需求，并指导数据安全角色组、参数和权限的定义。某些版本中添加 E（Execute）执行，以创建 **CRUDE 矩阵**。

7.5 实施指南
7.5.4 外包世界中的数据安全
任何事情皆可外包，但责任除外。
任何形式的外包都增加了组织风险，包括失去对技术环境、对组织数据使用方的控制。数据安全措施和流程必须将外包供应商的风险既视为**外部风险，又视为内部风险**。
负责、批注、咨询、通知**（RACI）矩阵**也有助于明确不同角色的角色、职责分离和职责，包括他们的数据安全义务。

7.6 数据安全治理
7.6.2 度量指标
5 项度量指标：**安全实施指标、安全意识指标、数据保护指标、安全事件指标、机密数据扩散。**
1.安全实施指标
这些常见的安全指标可以设定为正值百分比：
1）安装了最新安全补丁程序的企业计算机百分比。
2）安装并运行最新反恶意软件的计算机百分比。
3）成功通过背景调查的新员工百分比。
4）在年度安全实践测验中得分超过 80%的员工百分比。
5）已完成正式风险评估分析的业务单位的百分比。
6）在发生如火灾、地震、风暴、洪水、爆炸或其他灾难时，成功通过灾难恢复测试的业务流程百分比。
7）已成功解决审计发现的问题百分比。
可以根据列表或统计数据的指标跟踪趋势：
1）所有安全系统的性能指标。
2）背景调查和结果。
3）应急响应计划和业务连续性计划状态。
4）犯罪事件和调查。
5）合规的尽职调查以及需要解决的调查结果数量。
6）执行的信息风险管理分析以及导致可操作变更的分析数量。
7）制度审计的影响和结果，如清洁办公桌制度检查，由夜班安保人员在换班时执行。
8）安全操作、物理安全和场所保护统计信息。
9）记录在案的、可访问的安全标准（制度）。
10）相关方遵守安全制度的动机。
11）业务行为和声誉风险分析，包括员工培训。
12）基于特定类型数据（如财务、医疗、商业机密和内部信息）的业务保健因素和内部风险。
13）管理者和员工的信心和影响指标，作为数据信息安全工作和制度如何被感知的指示。
随着时间的推移，在适当的类别中选择和维护合理数量的可操作指标，以确保合规性；在问题成为危机之前被发现，

并向高级管理层表明保护企业信息的决心。

2.安全意识指标

考虑以下这些常规领域并选择适当的指标：

1）风险评估结果。评估结果提供了定性数据，需要反馈给相关业务单位，以增强其责任意识。

2）风险事件和配置文件。通过这些事件和文件确定需要纠正的未管理风险敞口。在安全意识倡议实施后，通过后续的测试来确定风险敞口以及制度遵从方面的缺失或可衡量改进的程度，以了解这些信息的传达情况。

3）正式的反馈调查和访谈。通过这些调查和访谈确定安全意识水平。此外，还要衡量在目标人群中成功完成安全意识培训的员工数量。

4）事故复盘、经验教训和受害者访谈。为安全意识方面的缺口提供了丰富的信息来源。具体指标可包括已减小了多少漏洞。

5）补丁有效性审计。涉及使用机密和受控信息的计算机，以评估安全补丁的有效性（尽可能推荐自动补丁系统）。

3.数据保护指标

需求决定哪些指标与组织相关：

1）特定数据类型和信息系统的关键性排名。如果无法操作，那么将对企业产生深远影响。

2）与数据丢失、危害或损坏相关的事故、黑客攻击、盗窃或灾难的年损失预期。

3）特定数据丢失的风险与某些类别的受监管信息以及补救优先级排序相关。

4）数据与特定业务流程的风险映射，与销售点设备相关的风险将包含在金融支付系统的风险预测中。

5）对某些具有价值的数据资源及其传播媒介遭受攻击的可能性进行威胁评估。

6）对可能意外或有意泄露敏感信息的业务流程中的特定部分进行漏洞评估。

敏感数据的可审计列表的位置信息，要在整个组织中传播。

4.安全事件指标

安全事件指标包括：

1）检测到并阻止了入侵尝试数量。

2）通过防止入侵节省的安全成本投资回报。

5.机密数据扩散

应衡量机密数据的副本数量，以减少扩散。机密数据存储的位置越多，泄露的风险就越大。

补充内容：18 种数据安全能力

能力一：数据安全综合治理平台

能力二：数据资源梳理及分类分级

能力三： 重要数据识别指南

能力四：权限的控制

能力五：数据加密

能力六：数据静态脱敏

能力七：数据动态脱敏

能力八：数据库防火墙

能力九：数据库漏洞扫描

能力十：数据库审计系统

能力十一：防勒索

能力十二：对数据库运行持续监控以保障其可靠高效运行

能力十三：撞库、拖库攻与防

能力十四：驻场等外来人员危险行为监测及防护

能力十五：利用 AI 技术对异常行为监控和防护

能力十六：数据泄露后的确权和溯源

能力十七：电子取证的能力

能力十八：隐私计算的能力

第 8 章 数据集成和互操作

8.1 引言

8.1.3 基本概念

1.抽取、转换、加载

数据集成和互操作的核心是抽取、转换和加载（ETL）这一基本过程。无论是在物理状态下或虚拟状态下，批量的或实时的执行 ETL 都是在应用程序和组织之间数据流动的必要步骤。

【ETL 目标数仓，BI 商业智能，业务场景是知道的，处理结构化梳理，非结构化无法处理】

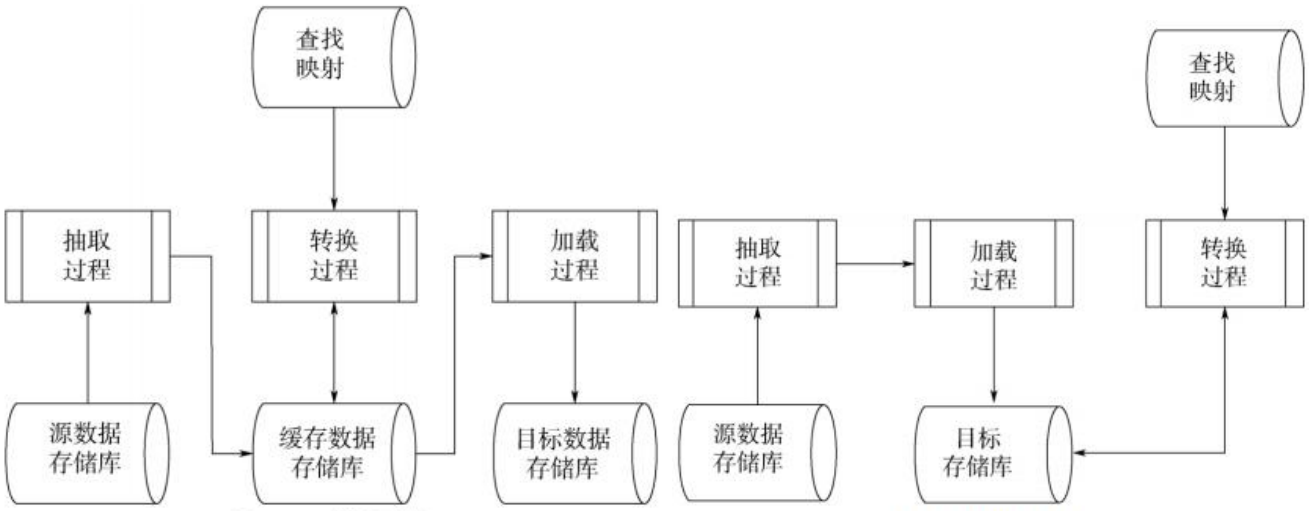

图8-2 ETL处理过程　　图8-3 ELT处理过程

Q：为什么要做转换？

A：生产数据是用三范式设计，大数据可能用逆规范化，所以需要转换。

转换的例子：

1）格式变化。技术上的格式转换，如从 EBCDIC 到 ASCII 的格式转换。

2）结构变化。数据结构的变化，如从非规范化到规范化的记录。

3）语义转换。数据值转换时保持语义的一致化表达，如源性别代码可以包括 0、1、2 和 3，而目标性别代码可以表示为 UNKNOWN、FEMALE、MALE 或 NOT PROVIDED。

4）消除重复。如规则需要唯一的键值或记录，以确保包括扫描目标、检测和删除重复行的方法。

5）重新排序。改变数据元素或记录的顺序以适应已定义的模式。转换可以批量执行，也可以实时执行，或者是将转换结果存储在物理状态下的缓存区域，或者是将转换后的数据存储在虚拟状态下的内存中，直至移动到加载步骤为止。转换阶段所产生的数据应准备好与目标结构中的数据进行集成。

（4）抽取、加载、转换（ELT）

如果目标系统比源系统或中间应用系统具有更强的转换能力，那么数据处理的顺序可以切换为 ELT——抽取、加载、转换（图 8-3）。 ELT 允许在数据加载到目标系统后再进行转换。ELT 允许源数据以原始数据的形式在目标系统上实例化，这对其他进程是有用的。用 ELT 的方式加载至数据湖，这在大数据环境中是很常见的。

【ELT 目标数据湖，不明确业务场景，不知道什么时候要用】

Q：什么数据进数据湖？　A：结构化+非结构化

Q：什么数据进数据仓库？ A：结构化

2.时延

时延（Latency）是指从源系统生成数据到目标系统可用该数据的时间差。不同的数据处理方法会导致不同程度的数据延迟。延迟可以是很高（批处理）或较高（事件驱动），甚至是非常低（实时同步）。

Q：什么是时延？

A：从源系统生成数据到目标系统可用该数据的时间差。

（1）批处理

大多数数据在应用程序和组织之间以一批文件的形式移动，要么是根据数据使用者的人工请求，要么是按周期自动触发。这种类型的交互称为批处理或 ETL。

（2）变更数据捕获【CDC，增量】

变更数据捕获是一种通过增加过滤来减少传送带宽需求的方法，只包含在特定时间范围内更改过的数据。变更数据捕获监视数据集的更改（插入、更改、删除），然后将这些更改（增量）传送给使用这些数据的其他数据集、应用程序和组织。作为变更数据捕获过程的一部分，对数据也可以用标记或时间戳等标识符来标识。变更数据捕获可以是基于数据的，也可以是基于日志的。

（3）准实时和事件驱动

与批处理相比，**准实时（NearReal-Time）**处理具有更低的延迟，而且通常因为工作是随时间分布的，所以系统负载较低。但是，它通常比同步数据集成解决方案要慢一些。准实时数据集成解决方案通常是使用**企业服务总线**（企业服务总线（Enterprise Service Bus，ESB）是用于在多个系统之间接近实时共享数据的数据集成解决方案，其数据中心是一个虚拟概念，代表组织中数据共享的标准和规范格式）来实现。

（4）异步

在异步数据流中，提供数据的系统在继续处理之前不会等待接收系统确认更新。异步意味着发送或接收系统可能会在一段时间内离线，而另一个系统可以正常运行。

（5）实时，同步

（6）低延迟或流处理

【由业务部门确认时延方法】

6.交互模型（描述了在系统之间建立连接以传送数据的方式）

（1）点到点

（2）中心辐射型

（3）发布与订阅

7.数据集成和互操作架构概念

（2）编排和流程控制

编排（Orchestration）是一个术语，用来描述在一个系统中如何组织和执行多个相关流程。所有处理消息或数据报的系统，必须能够管理这些流程的执行顺序，以保持一致性和连续性。

Q：什么是编排？安排人事更有效？

A：错，安排工作，组织和执行多个相关流程。

8.6 数据集成和互操作处理

8.6.3 度量指标

1）数据可用性。请求数据的可获得性。

2）数据量和速度。它包括：传送和转换的数据量，分析数据量，传送速度，数据更新与可用性之间的时延，事件与触发动作之间的时延，新数据源的可用时间。

3）解决方案成本和复杂度。它包括：解决方案开发和管理成本，获取新数据的便利性，解决方案和运营的复杂度，使用数据集成解决方案的系统数量。

第 9 章 文件和内容管理【技术本身不成熟】

【文件管理有成熟的软件系统】

【内容管理有赖于 NLP（自然语言处理），还不成熟】

9.1 引言

在许多组织中，非结构化数据和结构化数据有着直接的关系，有关内容的管理决策应同样适用于非结构化数据的管理要求。【数据都是需要管理的】

9.1.1 业务驱动因素（4 项）

文件和内容管理的主要业务驱动因素包括**法规遵从性要求、诉讼响应能力和电子取证请求能力以及业务连续性要求**。

文档管理 P258【】

ARMA 国际（非营利性的档案和信息管理专业协会）在 2009 年发布了一套被普遍接受的档案保存指导原则®（GARP），它描述了应该如何维护业务档案。

1）问责原则（Accountability）。组织应指派适当的高级管理人员，采用制度和流程来指导员工，并确保计划的可审计性。

2）完整原则（Integrity）。建立信息治理规划，使组织创建或管理的档案和信息具有合理性以及适当的真实性和可靠性保证。

3）保护原则（Protection）。建立信息治理规划，确保对个人信息或其他需要保护的信息提供合理的保护。

4）遵从原则（Compliance）。建立信息治理规划，遵从适用的法律法规和其他有约束力的机构及组织的制度要求。

5）可用原则（Availability）。组织应确保以及时、高效和准确检索其信息的原则来维护其信息。

6）保留原则（Retention）。组织的信息应保留适当的时间，并考虑所有运营、法律、监管和财政以及其他所有相关约束的要求。

7）处置原则（Disposition）。组织应根据其制度、适用的法律法规以及其他有约束力的机构要求，提供安全和适当的信息处置。

8）透明原则（Transparency）。组织应以工作人员和利益相关方可以理解的方式记录其制度、流程和活动，包括其信息治理规划。

9.1.3 基本概念

1.内容

（1）内容管理

内容管理（Content Management）包括用于组织、分类和构造信息资源的流程、方法和技术，以便以多种方式存储、发布和重复使用这些资源。当在整个企业范围内进行内容管理时，称之为企业内容管理（ECM）。

（2）内容元数据

元数据对于管理非结构化数据至关重要，无论是传统上认为的内容和文件，还是现在理解的“大数据”。如果没有元数据，就无法对内容进行编目和组织。

非结构化数据内容的元数据基于：

1）格式。通常数据格式决定了访问数据的方法（如电子非结构化数据的电子索引）。

2）可搜索性。是否已经具备用于搜索相关非结构化数据的工具。

3）自我描述性。元数据是否有自我描述能力（如在文件系统中）。在这种情况下，因为可以简单地采用现有工具，开发的需求是最小的。

4）既有模式。是否可以采用或者适配现有的方法和模式（如在图书馆目录中）。

5）内容主题。人们可能在寻找的东西。

6）需求。需要进行彻底和详细的检索能力（如制药或核工业）。因此，内容级的详细元数据可能是必要的，并且可能需要一个能够进行内容标记的工具。

2.受控词表

受控词表（Controlled Vocabularies）是被明确允许用于通过浏览和搜索对内容进行索引、分类、标引、排序和检索术语的定义列表。系统地组织文件、档案和内容离不开受控词表。词汇表的复杂程度包括从简单的列表或选项列表，到同义词环圈或规范表、分类法以及最复杂的主题词表和本体。**受控词表的一个例子是用于出版物分类的都柏林核心元素集（Dublin Core Element，DC）。**

3.文件和档案

文件（Document）是包含任务说明，对执行任务或功能的方式和时间的要求以及任务执行和决策的日志等的电子或纸质对象。文件可用于交流并分享信息和知识。程序、协议、方法和说明书都属于文件。

只有部分文件才能称为档案（Record）。档案可用于证明所做的决策和所采取的行动是符合程序的；可作为组织业务活动和法规遵从的证据。档案通常是由人来创建的，但仪器和监控设备也可以提供数据来自动生成档案。

Q：文件与档案关系

A：并非所有文件都会成为档案。

精心管理的档案具有以下特点：

1）内容。内容必须准确、完整和真实。

2）背景。关于档案的创建者、创建日期或与其他档案关系的描述性信息（元数据）应该在创建档案时收集、组织并维护。

3）及时性。档案应该在事件、行为或决定发生后立即创建。

4）永久性。一旦成为档案，则在档案的法定**保存期内**不能改变其内容。

5）结构。档案内容的外观和排版需要清晰，它们应被记录在正确的表格或模板上。内容应清晰易读，对术语的使用应始终保持如一。

许多档案同时以电子和纸张两种形式存在。档案管理要求组织知道哪个副本（电子或纸质）是正式的“档案副本”，以履行档案保存义务。一旦档案副本确定下来，其他的副本便可以安全销毁。

Q：在档案的法定保存期外不能改变。

A：错，档案永久性：一旦成为档案，则在档案的法定保存期内不能改变其内容。

5.电子取证

“取证”（Discovery）是一个法律术语，指诉讼的预审阶段，双方当事人互相要求对方提供信息，以查明案件事实，并了解双方的论点有多强。自 1938 年以来，美国联邦民事诉讼规则（FRCP）已经在诉讼和其他民事案件中要求对发现的证据进行管理。几十年来，基于纸质的取证规则被应用到**电子取证（E-discovery）**。2006 年，FRCP 的修订版纳入了 ESI 在诉讼过程中的取证实践和要求。**【微信记录可作为取证】**

9.语义搜索 【舆情分析】

语义搜索（Semantic Search）侧重于语义和语境而非预先设定的关键字。语义搜索引擎可以使用人工智能基于单词及其语境来识别查询匹配。这样的搜索引擎可以根据位置、意图、单词变体、同义词和概念匹配来进行分析。

10.非结构化数据

据估计，多达 80%的数据存储是在关系型数据库之外维护的。这非结构化数据有多种电子格式：文字处理文件、电子邮件、社交媒体、聊天室、平面文件、电子表格、XML 文件、事务性消息、报告、图形、数字图像、缩微胶片、视频和音频。纸质文件中也存在大量非结构化数据。

Q：哪些是非结构化数据？只包含电子格式？

A：纸质文件中也存在大量非结构化数据。

9.3 工具

9.3.4 标准标记和交换格式

4.Schema.org

使用语义标记来给内容打标签（如开源 Schema.org 所定义）使语义搜索引擎更容易索引内容，并使网络爬虫更容易将内容与搜索查询匹配。Schema.org 提供了一组用于页面标记的共享词汇表或模式，以便主流的搜索引擎可以理解它们。它侧重于网页上的文字含义以及术语和关键词。

Schema.org 词汇表集合还可用于结构化数据的互操作（如与 JSON）。

第 10 章参考数据和主数据

Q1：主数据的定义是什么

A1：通过对共享的数据建设标准，提高数据质量

1 共享的数据，2 标准化建设路程，3 落脚点，提高数据质量，不是共享

Q2：主数据难点

A2：如何识别主数据：1 共享的实体（都属于主数据），2 属性（重要、相对稳定的属性）

Q3：主数据是数据之源？

A3：错，元数据是数据之源。

10.1 引言

10.1.2 目标和原则

参考数据和主数据管理规划的**目标**包括：

1）确保组织在各个流程中都拥有**完整、一致、最新且权威**的参考数据和主数据。

2）促使企业在各业务单元和各应用系统之间**共享**参考数据和主数据。

3）通过采用**标准的、通用的**数据模型和整合模式，降低数据使用和数据整合的成本及复杂性。

参考数据和主数据管理遵循以下指导**原则**：

1）共享数据。为了能在组织中实现参考数据和主数据共享，必须把这些数据管理起来。

2）所有权。参考数据和主数据的所有权属于整个组织，而不是属于某个应用系统或部门。因为需要广泛共享，所以需要全局的组织管理。

3）质量。参考数据和主数据需要持续的数据质量监控和治理。【最终落脚点】

4）管理职责。业务数据管理专员要对控制和保证参考数据的质量负责。

5）控制变更。

①在给定的时间点，主数据值应该代表组织对准确和最新内容的最佳理解。改变数据值的匹配规则，应该在有关监督下谨慎地运用。任何合并或拆分参考数据和主数据的操作都应该是可追溯的。

②对参考数据的更改应该遵循一个明确的流程：在实施变更之前应该进行沟通并得到批准。

6）权限。主数据值应仅从记录系统（System of Record）中复制。

为了实现跨组织的主数据共享，可能需要建立一个参考数据管理系统（System of Reference）。

10.1.3 基本概念

1.主数据和参考数据的区别

主数据：需要收集、清洗和解析。

参考数据：直接拿来用就可以。

2.参考数据

表 10-5 通用标准产品与服务分类（UNSPSC）【联合国沿用】

（3）行业参考数据

行业参考数据（Industry Reference Data）是一个宽泛的术语，用于描述由行业协会或政府机构而不是由某个组织创建和维护的数据集，以便为编码重要的概念提供一个通用的标准。这种编码引出了一种常见的理解数据的方式，也是数据共享和互操作性的先决条件。例如，国际疾病分类代码（ICD）提供了一种常见的方法对健康状况（诊断）和治疗（程序）进行分类，从而在卫生保健和治疗结果方面提供了统一的说明方法。如果每个医生和医院都为疾病制定自己的代码集，那么了解疾病的趋势和模式几乎是不可能的事情。

3.主数据

主数据应该代表与关键业务实体有关的权威的、最准确的数据。业务规则通常规定了主数据格式和允许的取值范围。一般组织的主数据包括下列事物的数据：

1）参与方。个人和组织，以及他们扮演的角色，如客户、公民、病人、厂商、供应商、代理商、商业伙伴、竞争

者、雇员或学生等。
2）产品和服务，包括内部和外部的产品及服务。
3）财务体系。如合同、总账、成本中心、利润中心。
4）位置信息。如地址和 GPS 坐标。

（2）可信来源，黄金记录

基于自动规则和数据内容的手动管理的结合，**可信来源（Trusted Source）被认为是"事实的最佳版本"。**可信来源也可以称为一种单一视图、360 度视图。要想让主数据管理系统成为可信来源，就必须有效地管理它们。在可信来源中，表示一个实体、实例的最准确数据的记录，可以被称为**黄金记录（Golden Record）。**

【可能考题】评估一个组织的主数据管理情况，需要识别以下几点：
1）哪些角色、组织、地点和事物被反复引用。
2）哪些数据被用来描述人、组织、地点和事物。
3）数据是如何被定义和设计的，以及数据粒度细化程度如何。
4）数据在哪里被创建或来源于哪里，在哪里被储存、提供和访问。
5）数据通过组织内的系统时是如何变化的。
6）谁使用这些数据，为了什么目的。
7）用什么标准来衡量数据及其来源的质量和可靠性。
④主数据 ID 管理。【数据中台 one ID】

（8）产品主数据

产品主数据（Product Master Data）专注于组织的内部产品和服务，或全行业的产品和服务（包括竞争对手）。不同类型的产品主数据解决方案支持不同的业务功能。
1）产品生命周期管理（PLM）系统侧重于从构想、开发、制造、销售、交付、服务和废弃等方面管理产品或服务的生命周期。组织通过实施产品生命周期管理系统以加快产品的上市。在产品开发周期长的行业（医药行业中长达 8～12 年），产品生命周期管理系统使组织能够跟踪跨过程的成本和法律协议，因为产品的构想从最开始的想法发展到潜在产品的过程会变换名称，还可能会依据不同的许可协议。
2）产品数据管理（PDM）系统通过捕获和实现对设计文档（如 CAD 图样）、配方（制造说明书）、标准操作程序和物料清单（BOM）等产品信息的安全共享，以支持工程和制造功能。产品数据管理功能可以通过专门的系统或 ERP 系统实现。
3）企业资源规划（ERP）系统的产品数据主要关注库存单位，以支持从订单录入到库存阶段，可以通过多种技术识别各种独立的产品。
4）制造执行系统（MES）中的产品数据主要关注原材料库存、半成品和成品，其中成品与可以通过 ERP 系统来存储和订购的产品相关联。这些数据在整个供应链和物流系统中也很重要。
5）客户关系管理（CRM）系统支持营销、销售和交互支持，系统中的产品数据可以包括产品系列和品牌、销售代表协会、客户区域管理以及营销活动等。

实现主数据中心环境的三种基本方法：
1）注册表（Registry）。注册表是指向多种记录系统（System of Record）中主数据记录的索引。记录系统管理应用程序本地的主数据，可以根据主索引访问主数据。注册表相对容易实现，因为它很少需要对记录系统进行更改。但是，要对多个系统中的主数据进行组合时通常需要复杂的查询。此外，还需要实施多个业务规则，以解决跨系统时产生的语义差异。
2）交易中心（Transaction Hub）。在该种方法中，各应用程序与中心系统交互，实现对主数据的访问和更新。主数据存在于交易中心内，而不存在于任何其他的应用程序中。交易中心是主数据的记录系统。交易中心使更好的治理成为可能，并对外提供一致的主数据源。
但是，从现有的记录系统中删除更新主数据功能的成本很高。业务规则仅被实施在单一系统中，即中心系统。
3）混合模式（Consolidated）。混合模式是注册表和交易中心的混合体。记录系统管理应用程序本地的主数据。主数据在一个公共存储库中被合并，并经由数据共享中心实现共享，如此消除了从记录系统直接进行访问的需要。混合法在提供企业视图的同时，能尽量减少对记录系统的影响，但是它需要在系统间进行数据复制，而且数据中心和记录系统之间会有延迟。

10.5 参考数据和主数据治理

10.5.2 度量指标

以下指标可以与参考数据和主数据质量以及支持这些努力的过程结合起来。

1）数据质量和遵从性。数据质量仪表板可以描述参考数据和主数据的质量。这些指标应该说明主题域实体或相关属性的置信度（百分比），以及它在整个组织中符合实际需求的使用价值。

2）数据变更活动。审核可信数据的血缘对于提高数据共享环境中的数据质量是必要的。指标应该展示数据值的变化率，能够帮助人们深入理解为共享环境提供数据的系统，并可被用于调整主数据管理进程中的算法。

3）数据获取和消费。数据由上游系统供应，由下游系统和流程使用。这些指标应该显示和追踪哪些系统在贡献数据，哪些业务区域在共享环境中订阅数据。 【不建议主数据放数据仓库，建议单独放在数仓】

4）服务水平协议（SLA）。应建立 SLA 并传达给贡献者和订阅者，以确保整个数据共享环境的使用和采用。遵循 SLA 可以为支持流程、技术问题和数据问题提供解释，而这些问题都有可能减缓主数据管理应用的速度。

5）数据管理专员覆盖率。【主数据上线后，应该有专人负责、专业团队运营】这些指标应该关注对数据内容负责的个人或团队，并展示覆盖率的评估频率。它们可以用来识别支持方面的差距。

6）拥有总成本。这个指标有多种影响因素、多种表达方式。从解决方案的角度来看，成本可以包括环境基础设施、软件许可证、支持人员、咨询费、培训等。这一指标的有效性主要是基于其在整个组织中的持续应用。

7）数据共享量和使用情况。需要跟踪纳入主数据的数据量和使用情况，以确定数据共享环境的有效性。这些指标应该展示数据共享环境中流入和流出数据的定义、纳入和订阅的数量和速率。

第 11 章 数据仓库和商务智能

【数据仓库-后端，商务智能-前端】

基本算法：关联关系（牵手-谈恋爱），集群关系（杭州人爱吃酸甜口），决策树，线性回归，贝叶斯，神经网络，时间序列

用法：精准营销，客户价值分析，旅客生命周期价值分析，风险，聚类和集群，实施需求和匹配，社会地位参数，忠诚度和客户粘度，时间序列。

11.1 引言

数据仓库（Data Warehouse，DW），商务智能（Business Inteligence，BI）

11.1.2 目标和原则【非常重要】

一个组织建设数据仓库的目标通常有：

1）支持商务智能活动。

2）赋能商业分析和高效决策。

3）基于数据洞察寻找创新方法。

数据仓库建设应遵循如下指导原则：

1）聚焦业务目标。确保数据仓库用于组织最优先级的业务并解决业务问题。

2）以终为始。让业务优先级和最终交付的数据范围驱动数据仓库内容的创建。

3）全局性的思考和设计，局部性的行动和建设。让最终的愿景指导体系架构，通过集中项目快速迭代构建增量交付，从而实现更直接的投资回报。

4）总结并持续优化，而不是一开始就这样做。以原始数据为基础，通过汇总和聚合来满足需求并确保性能，但不替换细节数据。

5）提升透明度和自助服务。上下文（各种元数据）信息越丰富，数据消费者越能从数据中获得更多数据价值。向利益相关方公开集成的数据及其流程信息。

6）与数据仓库一起建立元数据。数据仓库成功的关键是能够准确解释数据。能回答一些基本问题，如“这个数字为什么是 X”“这个怎么计算出来的”“这个数据哪里来的”。元数据的获取应该作为软件开发周期的一部分，元数据的管理也应该作为数据仓库持续运营的一部分。

7）协同。与其他数据活动协作，尤其是数据治理、数据质量和元数据管理活动。

8）不要千篇一律。为每种数据消费者提供正确的工具和产品。

11.1.3 基本概念

1.商务智能商务智能两层含义。

第一层含义，商务智能指的是一种理解组织诉求和寻找机会的**数据分析活动**。数据分析的结果用来提高组织决策的成功率。当人们说数据是竞争优势的关键要素时，他们其实是在说商务智能的内在逻辑：如果一个组织向自己的数据“正确提问”，他就能获得关于产品、服务及客户方面的洞见，为实现自己的战略目标做出更好的决策。

第二层含义，商务智能指的是支持这类数据分析活动的**技术集合**。决策支持工具、商务智能工具的不断进化，促成了数据查询、数据挖掘、统计分析、报表分析、场景建模、数据可视化及仪表板等一系列应用，它们被用于从预算到高级分析的方方面面。

2.数据仓库

数据仓库有两个重要组成部分：一个集成的决策支持数据库和与之相关的用于收集、清理、转换和存储来自各种操作和外部源数据的软件程序。

4.数据仓库建设的方法【一定会考】

大部分关于数据仓库构建的讨论，都受到两位有影响力的思想领袖 Bill Inmon 和 Ralph Kimball 的影响，他们各有不同的数据仓库建模和实施方法。Inmon 把数据仓库定义为“面向主题的、整合的、随时间变化的、相对稳定的支持管理决策的数据集合”，用规范化的关系模型来存储和管理数据。而 Kimball 则把数据仓库定义为“为查询和分析定制的交易数据的副本”，他的方法通常称作多维模型（参见第 5 章）。虽然 Inmon 和 Kimball 提倡的数据仓库建设方法不同，但他们遵循的核心理念相似：

1）数据仓库存储的数据来自其他系统。
2）存储行为包括以提升数据价值的方式整合数据。
3）数据仓库便于数据被访问和分析使用。
4）组织建设数据仓库，因为他们需要让授权的利益相关方访问到可靠的、集成的数据。
5）数据仓库数据建设有很多目的，涵盖工作流支持、运营管理和预测分析。

【OLTP 尽量少用索引】

【数仓侧重点是 BI，但也可以做 AI，数据湖侧重 BI】

Q：数仓目标是 BI

A：错，数仓侧重点是 BI。

多维模型通常称为星型模型，由事实表（包含有关业务流程的定量数据，如销售数据）和维度表（存储与事实表数据相关的描述性属性，为数据消费者解答关于事实表的问题，如这个季度产品 X 卖了多少）组成。

Q：看下图回答问题，数据集市是必须要的产品嘛？

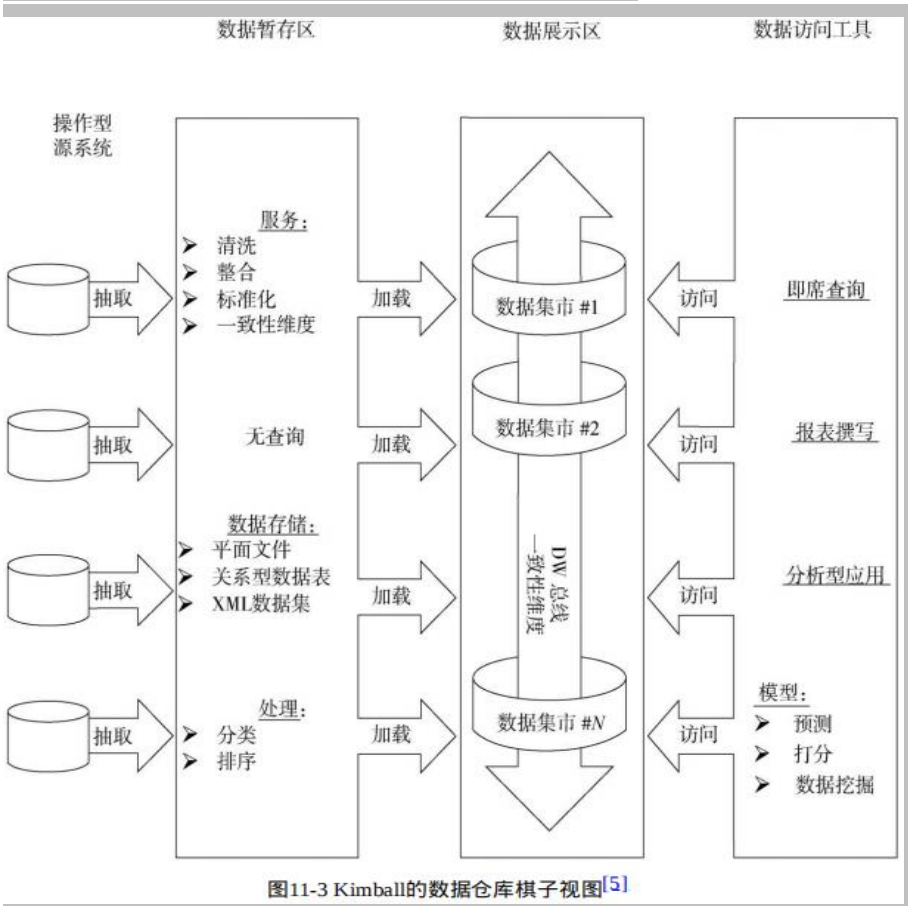

图11-3 Kimball的数据仓库棋子视图[5]

A：错，数据集市是数仓的一小部分，

Q：主数据是必须要的嘛？

A：错，数据源、ETL、核心数据仓库是必须要的。

7.数据仓库架构组件

（1）源系统

（2）数据集成

数据结构的设计元素包括：

①基于性能考虑而设计的业务主键和代理主键之间的关系。

②创建索引和外键以支持维度表。

③用于检测、维护和存储历史记录的**变更数据捕获（Change Data Capture，CDC）**技术。

8.加载处理的方式

数据仓库建设涉及两种主要的数据集成处理类型：历史数据加载和持续不断的数据更新。历史数据通常只需要加载

一次，或者为了处理数据问题加载有限的几次，然后再也不会加载。“持续不断的数据更新”需要始终如一地规划和执行，以保证数据仓库中包含最新的数据。

表11-2 CDC技术比对

方法	对源系统的要求	复杂度	事实表加载	维表加载	重叠	删除
时间戳增量加载	源系统中的变化由系统日期和时间戳标识	低	快	快	是	否
日志表增量加载	捕获源系统中的变化并记录在日志表	中	普通	普通	是	是
数据库交易日志	在交易日志记录数据库变化	高	普通	普通	否	是
消息增量	源系统中的变化发布在实时消息（队列）	极高	慢	慢	否	是
全量加载	没有更改标识符，抽取全表数据并比较判断改动	极低	慢	普通	是	是

Q：有几种方法识别增量？

A：4 种：时间戳增量加载、日志表增量加载、数据库交易日志、消息增量。

Q：处理数据量最大的方法

A：全量加载。

11.3 工具

11.3.3 商务智能工具的类型

常见的 OLAP 操作包括切片和切块、向下钻取、向上钻取、向上卷积和透视。

1）切片（Slice）。切片是多维数组的子集，对应不在子集中的维度的**一个或多个成员**的单个值。

2）切块（Dice）。切块操作是数据立方体上**两个以上**维度的切片，或者是两个以上的连续切片。

3）向下/向上钻取（Drill down/up）。向下钻取或向上钻取是一种特定的分析技术，用户可以在不同数据级别之间导航，范围从最概括（向上）到最详细（向下）。

4）向上卷积（Roll-up）。卷积涉及计算一个或多个维度的所有数据关系。为此，需要先定义计算关系或公式。

5）透视（Pivot）。透视图会更改报表或页面的展示维度。

三种经典的 OLAP 实现方法如下：

1）关系型联机分析处理（ROLAP）。ROLAP 通过在关系数据库（RDBMS）的二维表中使用多维技术来支持 OLAP。星型架构是 ROLAP 环境中常用的数据库设计技术。

2）多维矩阵型联机分析处理（MOLAP）。MOLAP 通过使用专门的多维数据库技术支持 OLAP。

3）混合型联机分析处理（HOLAP）。它是 ROLAP 和 MOLAP 的结合。HOLAP 实现允许部分数据以 MOLAP 形式存储，而另一部分数据存储在 ROLAP 中。控件的实现方式各不相同，设计师对分区的组合也各有不同。

11.4 方法

驱动需求的原型，自助式商务智能，可查询的审计数据。

11.6 数据仓库/商务智能治理

11.6.5 度量指标（3 个）

3 个度量指标：**使用指标、主题域覆盖率、响应时间和性能指标**。

1.使用指标

数据仓库中使用的度量指标通常包括注册用户数、连接用户数或并发用户数。这些度量指标表示组织内有多少人正在使用数据仓库。为每个工具授权多少个用户账户是一个很好的开始，特别是对于审计员而言。但是，实际有多少用户连接到该工具是一个更好的度量指标，并且每个时间段由用户社区申请的查询（或与查询相当）数量对于容量规划是更好的技术指标。允许多个分析指标，如审核用户、已生成的用户查询量和使用用户。

2.主题域覆盖率

主题域覆盖百分比衡量每个部门访问仓库的程度（从数据拓扑的角度来看），还强调哪些数据是跨部门共享的，哪些还不是但也可能是共享的。

将操作源映射到目标是另一种自然的扩展，它强制和验证已经收集的血缘关系和元数据，并可以提供渗透分析，确定哪些部门在使用哪些源系统分析。通过减少对大量使用的源对象的更改，有助于将工作调整集中在那些具有高影响力的分析查询上。

3.响应时间和性能指标

大多数查询工具会测量响应时间。通过工具检索响应或性能指标。此数据指标代表用户的数量和类型。

第 12 章 元数据管理

Q：元数据是数据资产目录

A：错，资源目录。

【元数据管理原则：应规尽规，应收尽收】

12.1 引言

元数据最常见的定义是“关于数据的数据”。这个定义非常简单，但也容易引起误解。可以归类为元数据的信息范围很广，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。它描述了数据本身（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）。【相当于图书馆的目录卡片】

如果没有可靠的元数据，组织就不知道它拥有什么数据、数据表示什么、数据来自何处、它如何在系统中流转，谁有权访问它，或者对于数据保持高质量的意义。如果没有元数据，组织就不能将其数据作为资产进行管理。实际上，如果没有元数据，组织可能根本无法管理其数据。

与其他数据一样，元数据需要管理。

12.1.2 目标和原则

【元数据最终目标：查询、分析】

12.1.3 基本概念

1.元数据与数据

如在简介中所述，元数据也是一种数据，应该用数据管理的方式进行管理。

2.元数据的类型

元数据通常分为三种类型：**业务元数据、技术元数据和操作元数据**。

【不是描述元数据、结构元数据、管理元数据——这是图书馆类别】

（1）业务元数据

业务元数据（Business Metadata）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。业务元数据包括主题域、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。业务元数据的示例包括：

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（如数据所有者、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

技术元数据（Technical Metadata）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。技术元数据示例包括：

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 CRUD（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

操作元数据（Operational Metadata）描述了处理和访问数据的细节，例如：
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

ISO 的元数据注册标准 ISO/IEC 11179 中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。该标准由以下几部分组成：
第 1 部分：数据元素生成和标准化框架。
第 2 部分：数据元数据分类。
第 3 部分：数据元素的基本属性。
第 4 部分：数据定义的形成规则和指南。
第 5 部分：数据元素的命名和识别原则。
第 6 部分：数据元素的注册。

4.非结构化数据的元数据【数据湖】

5.元数据来源

（1）应用程序中元数据存储库

（2）业务术语表

业务术语表（Business Glossary）的作用是记录和存储组织的业务概念、术语、定义以及这些术语之间的关系。
业务词汇表应用程序的构建需满足三个核心用户的功能需求：
1）业务用户（Business users）。数据分析师、研究分析师、管理人员和使用业务术语表来理解术语和数据的其他人员。
2）数据管理专员（Data Stewards）。数据管理专员使用业务术语表管理和定义术语的生命周期，并通过将数据资产与术语表相关联增强企业知识，如将术语与业务指标、报告、数据质量分析或技术组件相关联。数据管理员收集术语和使用中的问题，以帮助解决整个组织的认识差异。

3）技术用户（Technical users）。技术用户使用业务术语表设计架构、设计系统和开发决策，并进行影响分析。
（3）商务智能工具
（4）配置管理工具
（5）数据字典【90%元数据信息来自数据字典，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。】
（6）数据集成工具
（7）数据库管理和系统目录
（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
（12）建模工具和存储库
（13）参考数据库
（14）服务注册
（15）其他元数据存储

12.4 方法

12.4.1 数据血缘和影响分析

【数据血缘：由下到上，影响分析：由上到下】

12.6 元数据治理

12.6.4 度量指标

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（范围内的所有产品和实例）的理想覆盖率与实际覆盖率进行比较。参照元数据管理范围定义的策略。

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，开发用于判断企业元数据成熟度的指标（参见第 15 章）。

3）专职人员配备。通过专职人员的任命情况、整个企业的专职人员覆盖范围，以及职位描述中的角色定义说明，来评估的组织对元数据的承诺。

4）元数据使用情况。可以通过存储库的访问次数衡量用户对元数据存储库的使用情况和接受程度。在业务实践中，用户引用元数据是一个很难跟踪的指标，可能需要定性的调研措施获取评估结果。

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。主数据服务上的元数据帮助开发人员决定新的开发任务可以使用哪些现有服务。

7）元数据文档质量。一个质量指标是通过自动和手动两种方式评估元数据文档的质量。自动评估方式包括对两个源执行冲突逻辑的比对、测量二者匹配的程度以及随时间推移的变化趋势。另一个度量指标是度量具有定义的属性的百分比，以及随着时间的推移而发生变化的趋势。手动评估方式包括基于企业质量定义进行随机或完整的调查。质量度量表明存储库中元数据的完整性、可靠性、通用性等。

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）。

第 13 章 数据质量

原则：重要的数据先开始。

重点：PDCA；评估数据质量维度；根因分析；数据质量报告

13.1 引言

有数据质量团队（Data Quality Program Team）。数据质量团队负责与业务和技术数据管理专业人员协作，并推动将质量管理技能应用于数据工作，以确保数据适用于各种需求。与数据治理和整体数据管理一样，数据质量管理不是一个项目，而是一项持续性工作。它包括项目和维护工作，以及承诺进行沟通和培训。最重要的是，数据质量改进取得长期成功取决于组织文化的改变及质量观念的建立。

13.1.1 业务驱动因素

高质量数据本身并不是目的，它只是组织获取成功的一种手段。

Q：数据管理直接目标？

A：提高数据质量。

Q：数据管理终极目标？

A：实现数据价值。

数据质量管理原则：

1）重要性。数据质量管理应关注对企业及其客户最重要的数据，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，包括其在系统内部和系统之间流转时的数据管理（数据链中的每个环节都应确保数据具有高质量的输出）。

3）预防。数据质量方案的重点应放在预防数据错误和降低数据可用性等情形上，不应放在简单的纠正记录上。

4）根因修正。提高数据质量不只是纠正错误，因为数据质量问题通常与流程或系统设计有关，所以提高数据质量通常需要对流程和支持它们的系统进行更改，而不仅仅是从表象来理解和解决。

5）治理。数据治理活动必须支持高质量数据的开发，数据质量规划活动必须支持和维持受治理的数据环境。

6）标准驱动。数据生命周期中的所有利益相关方都会有数据质量要求。在可能的情况下，对于可量化的数据质量需求应该以可测量的标准和期望的形式来定义。

7）客观测量和透明度。数据质量水平需要得到客观、一致的测量。应该与利益相关方一同讨论与分享测量过程和测量方法，因为他们是质量的裁决者。

8）嵌入业务流程。业务流程所有者对通过其流程生成的数据质量负责，他们必须在其流程中实施数据质量标准。

9）系统强制执行。系统所有者必须让系统强制执行数据质量要求。

10）与服务水平关联。数据质量报告和问题管理应纳入服务水平协议（SLA）。

13.1.3 基本概念

1.数据质量

数据质量如达到数据消费者的期望和需求，也就是说，如果数据满足数据消费者应用需求的目的，就是高质量的；反之，如果不满足数据消费者应用需求的目的，就是低质量的。因此，数据质量取决于使用数据的场景和数据消费者的需求。

2.关键数据

虽然关键的特定驱动因素因行业而异，但组织间存在共同特征，可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，尤其是差异化竞争战略

3.数据质量维度

Q：关于数据质量的大咖

A：Strong-Wang 框架　Thomas Redman《信息时代的数据质量》　Larry English《改善数据仓库和业务信息质量》

2013 年，DAMA UK 发布了一份白皮书，描述了**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。比较事物多种表述与定义的差异。

6.数据质量改进生命周期

戴明环 PDCA 休哈特图

新周期开始于：

①现有测量值低于阈值。

②新数据集正在调查中。

③对现有数据集提出新的数据质量要求。

④业务规则、标准或期望变更。

Q：每天表整合为每月表，是否需要 PDCA？

A：不需要

8.数据质量问题的常见原因【非常重要】

从创建到处置，数据质量问题在数据生命周期的任何节点都可能出现。在调查根本原因时，分析师应该寻找潜在的原因，如数据输入、数据处理、系统设计，以及自动化流程中的手动干预问题。

Q：数据质量最常见问题？

A：缺乏领导力导致。

（1）缺乏领导力【和企业文化】导致的问题

（2）数据输入过程引起的问题

（3）数据处理功能引起的问题

（4）系统设计引起的问题

（5）解决问题引起的问题

9.数据剖析

Q：数据剖析是解决数据质量的方法。

A：错，数据剖析不是解决数据质量的方法。

数据剖析（Data Profiling）是一种用于检查数据和评估质量的数据分析形式。数据剖析使用统计技术来发现数据集合的真实结构、内容和质量（Olson，2003）。剖析引擎生成统计信息，分析人员可以使用这些统计信息识别数据内容和结构中的模式。例如：

1）空值数。标识空值存在，并检查是否允许空值。

2）最大/最小值。识别异常值，如负值。

3）最大/最小长度。确定具有特定长度要求的字段的异常值或无效值。

4）单个列值的频率分布。能够评估合理性（如交易的国家代码分布、频繁或不经常发生的值的检查，以及用默认值填充的记录百分比）。

5）数据类型和格式。识别不符合格式要求的水平，以及意外格式识别（如小数位数、嵌入空格、样本值）。

13.4 方法

13.4.4 有效的数据质量指标

Q：基于 DAMA 理解，数据质量指标可以定性也可以定量。

A：错，必须是可度量的。

1）可度量性。数据质量指标必须是可度量的——它必须是可被量化的东西。例如，数据相关性是不可度量的，除非设置了明确的数据相关性标准。即便是数据完整性这一指标也需要得到客观的定义才能被测量。预期的结果应在离

散范围内可量化。
2）业务相关性。虽然很多东西是可测量的，但并不能全部转化为有用的指标。测量需要与数据消费者相关。如果指标不能与业务操作或性能的某些方面相关，那么它的价值是有限的。每个数据质量指标都应该与数据对关键业务期望的影响相关联。
3）可接受性。数据质量指标构成了数据质量的业务需求，根据已确定的指标进行量化提供了数据质量级别的有力证据。根据指定的可接受性阈值确定数据是否满足业务期望。如果得分等于或超过阈值，则数据质量满足业务期望；如果得分低于阈值，则不满足。
4）问责/管理制度。关键利益相关方（如业务所有者和数据管理专员）应理解和审核指标。当度量的测量结果显示质量不符合预期时，会通知关键利益相关方。业务数据所有者对此负责，并由数据管理专员采取适当的纠正措施。
5）可控制性。指标应反映业务的可控方面。换句话说，如果度量超出范围，它应该触发行动来改进数据。如果没有任何响应，那么这个指标可能没有什么用处。
6）趋势分析。指标使组织能够在一段时间内测量数据质量改进的情况。跟踪有助于数据质量团队成员监控数据质量 SLA 和数据共享协议范围内的活动，并证明改进活动的有效性。一旦信息流程稳定后，就可以应用统计过程控制技术发现改变，从而实现其所研究的度量结果和技术处理过程的可预测性变化。

13.4.6 根本原因分析

导致问题产生的根本原因一旦消失，问题本身也会消失。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。其目的是识别潜在的条件，这些条件一旦消除，问题也将消失。
常见的根因分析技术包括帕累托分析（80/20 规则）、鱼骨图分析、跟踪和追踪、过程分析以及五个为什么等（McGilvray，2008）。

13.6 数据质量和数据治理

13.6.2 度量指标

数据质量团队的大部分工作将集中于质量的度量和报告上。数据质量的高阶指标包括：
1）投资回报。关于改进工作的成本与改进数据质量的好处的声明。
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（趋势），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状态（已解决、未解决、已升级）。
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。
数据质量计划示意图。现状和扩展路线图。

第 14 章 大数据和数据科学

14.1 引言

14.1.3 科学理念

1.数据科学

数据科学将数据挖掘、统计分析和机器学习与数据集成整合，结合数据建模能力，去构建预测模型、探索数据内容模式。数据科学依赖于：

1）丰富的数据源。具有能够展示隐藏在组织或客户行为中不可见模式的潜力。

2）信息组织和分析。用来领会数据内容，结合数据集针对有意义模式进行假设和测试的技术。

3）信息交付。针对数据运行模型和数学算法，进行可视化展示及其他方式输出，以此加强对行为的深入洞察。

4）展示发现和数据洞察。分析和揭示结果，分享洞察观点（表 14-1）对比了传统的数据仓库/商务智能与基于数据科学技术实现的**预测性分析和规范性分析**的作用。

表14-1 分析对比

数据仓库/传统商务智能	数据科学	
描述性分析	预测性分析	规范性分析
事后结论	洞察	预见
基于历史： 过去发生了什么 为什么发生	基于预测模型： 未来可能会发生什么	基于场景： 我们该做什么才能保证事情发生

3.大数据

早期，人们通过3V来定义大数据含义的特征：**数据量大（Volume）、数据更新快（Velocity）、数据类型多样/可变（Variety）**（Laney，2001）。随着越来越多的组织开始深挖大数据的潜力，已经不止于以上三个V。V列表有了更多的扩展：

1）数据量大（Volume）。大数据通常拥有上千个实体或数十亿个记录中的元素。

2）数据更新快（Velocity）。指数据被捕获、生成或共享的速度。大数据通常实时地生成、分发及进行分析。

3）数据类型多样/可变（Variety/Variability）。指抓取或传递数据的形式。大数据需要多种格式储存。通常，数据集内或跨数据集的数据结构是不一致的。

4）数据黏度大（Viscosity）。指数据使用或集成的难度比较高。

5）数据波动性大（Volatility）。指数据更改的频率，以及由此导致的数据有效时间短。

5.大数据来源 结构化数据+非结构化数据

6.数据湖

数据湖是一种可以提取、存储、评估和分析不同类型和结构海量数据的环境，可供多种场景使用。如可以提供：

1）数据科学家可以挖掘和分析数据的环境。

2）原始数据的集中存储区域，只需很少量的转换（如果需要的话）。

3）数据仓库明细历史数据的备用存储区域。

4）信息记录的在线归档。

5）可以通过自动化的模型识别提取流数据的环境。

数据湖的风险在于，它可能很快会变成**数据沼泽——**杂乱、不干净、不一致。为了建立数据湖中的内容清单，在数据被摄取时对元数据进行管理至关重要。

Q：数据湖管理不好会变成？A 池塘 B 沼泽 C 大海

A：不是池塘，是沼泽。

Q：数据湖是否管理好表示什么？A 元数据是否管理好？B 数据质量得到保证

A：元数据是否管理好

7.基于服务的架构

基于服务的体系结构（Services-Based Architecture，SBA）

8.机器学习

机器学习探索了学习算法的构建和研究。这些算法一般分为三种类型：

1）监督学习（Supervised learning）。基于通用规则（如将 SPAM 邮件与非 SPAM 邮件分开）。

2）无监督学习（Unsupervised learning）。基于找到的那些隐藏的规律（数据挖掘）。

3）强化学习（Reinforcement learning）。基于目标的实现（如在国际象棋中击败对手）。

Q：预测明天销售额是多少？

A：有无限可能性，无监督学习

Q：预测明年销售额是否笔今年多？A 多 B 少 C 一样 D 不知道

A：监督学习

12.规范分析

规范分析（Prescriptive Analytics）比预测分析更进一步，它对将会影响结果的动作进行定义，而不仅仅是根据已发生的动作预测结果。规范分析预计将会发生什么，何时会发生，并暗示它将会发生的原因。由于规范分析可以显示各种决策的含义，因此可以建议如何利用机会或避免风险。规范分析可以不断接收新数据以重新预测和重新规定。该过程可以提高预测准确性，并提供更好的方案。

第 15 章 数据管理成熟度评估

15.1 引言

能力成熟度评估（Capability Maturity Assessment，CMA）是一种基于能力成熟度模型（Capability Maturity Model，CMM）框架的能力提升方案，描述了数据管理能力初始状态发展到最优化的过程。

1）0 级。无能力级。

2）1 级。初始级或临时级：成功取决于个人的能力。

3）2 级。可重复级：制定了最初级的流程规则。

4）3 级。已定义级：已建立标准并使用。

5）4 级。已管理级：能力可以被量化和控制。

6）5 级。优化级：能力提升的目标是可量化的。

15.1.3 基本概念

1.评价等级及特点

CMM 通常定义 5～6 个成熟度级别，每个级别有各自的特性，从初始级到优化级，

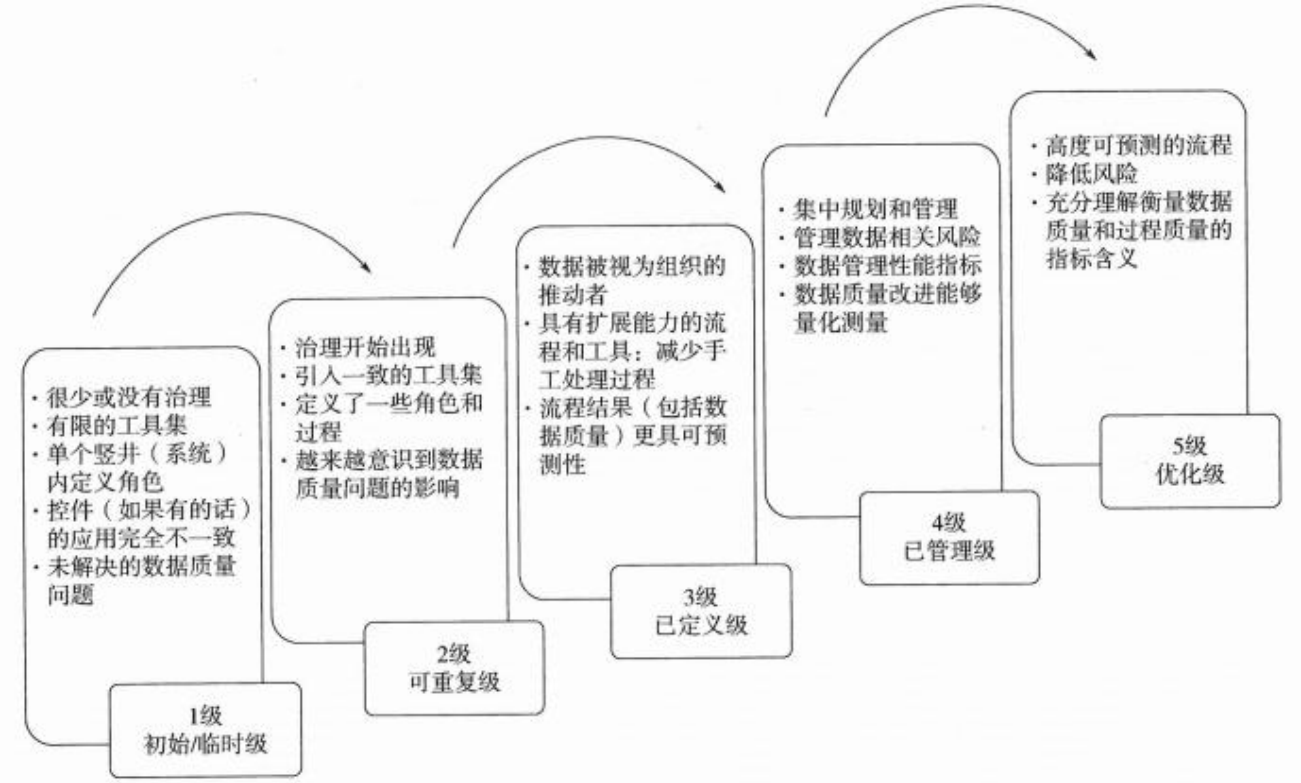

图 15-2 数据管理成熟度模型示例

1）0 级：无能力。在数据管理中，管理活动或正式企业流程处于无组织的状态。很少有组织处在 0 级阶段，这个级别在成熟度模型中是为了定义才被设定的。

2）1 级初始/临时。使用有限的工具集进行通用的数据管理，**很少或根本没有治理活动。**数据处理高度依赖于少数专家，角色和责任在各部门中分开定义。每个数据所有者自主接收、生成和发送数据控件（如果有的话）的应用不一致。管理数据的解决方案是有限的。数据质量问题普遍存在，但无法得到解决，基础设施支持处于业务单元级别。评估标准可能包括对任意一个流程进行控制，如记录数据质量问题。

3）2 级可重复。有一致的工具和角色定义来支持流程执行。在 2 级中，组织开始使用**集中化**的工具，并为数据管理提供更多的监控手段。角色的定义和流程并不完全依赖于特定专家。组织对数据质量问题和概念有认识，开始认识到主数据和参考数据的概念。评估标准可能包括组件中的正式角色定义，如职位描述、流程文档以及利用工具集的能力。

4）3 级已定义：新兴数据管理能力。第 3 级将引入可扩展的数据管理流程将其**制度化**，并将数据管理视为一种组织促成因素。其特点包括在组织中的数据复制受到控制，总体数据质量普遍提高，有协调一致的政策定义和管理。越正式的流程定义越能显著减少人工干预，这样伴随着集中化的设计流程，意味着流程的结果更加可预测。评估标准可能包括制定数据管理政策、可扩展过程的使用以及数据模型和系统控制的一致性。

5）4 级已管理。从 1～3 级增长中获得的经验积累使组织能够在即将开展新项目和任务时预测结果，并开始管理与数据相关的风险，数据管理包括一些绩效指标。4 级的特点包括从桌面到基础设施的数据管理工具标准化，以及结构良好的集中规划和治理功能。此级别的机构在数据质量和全组织数据管理能力（如端到端的数据审核）等方面有显著性提高。评估标准可能包括与项目成功相关的指标、系统的操作指标和数据质量指标。

6）5 级优化。当数据管理实践得到优化时，由于流程自动化和技术变更管理，它们是高度**可预测**的，这个成熟度级别的组织会更关注于持续改进。在第 5 级，工具支持跨流程查看数据。控制数据的扩散防止不必要的复制，使用容易理解的指标来管理和度量数据质量和过程。

2.评估标准

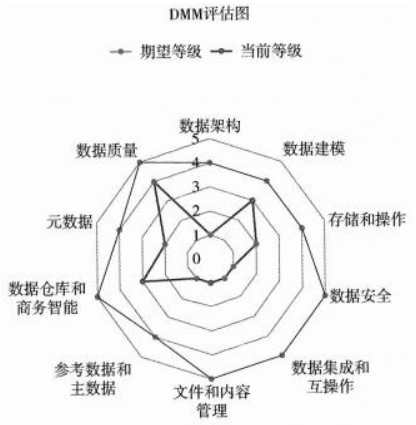

图 15-3　数据管理成熟度评估可视化示例

至少有 3 个地方错误

1、数据仓库和商务智能有赖于数据集成，数据集成只有 1 级，数据仓库和商务智能却有 5 级

2、数据质量依赖于参考数据和主数据，参考数据和主数据只有 1 级，数据质量却有 4 级

3、架构只有 1 级，数据建模却有 3 级。

3.现有 DMMA 框架

（1）CMMI 数据管理成熟度模型（DMM）

（2）EDM 委员会 DCAM

企业数据管理委员会（Enterprise Data Management Council）是总部设在美国的金融服务行业宣传组织，它开发了数据管理能力评估模型（Data management Capability Assessment Model，DCAM）。

（3）IBM 数据治理委员会成熟度模型

数据管理委员会成熟度模型基于 55 个组织委员会组成。

（4）斯坦福数据治理成熟度模型

斯坦福大学的数据治理成熟度模型是为该大学开发的。

（5）Gartner 的企业信息管理成熟度模型

15.2 活动

5 个活动：规划评估活动、执行成熟度评估、解释结果及建议、制定有针对性的改进计划、**重新评估成熟度**。

15.2.1 规划评估活动

15.2.2 执行成熟度评估

15.2.3 解释结果及建议

15.2.4 制定有针对性的改进计划

15.2.5 重新评估成熟度

重新评估也可以重振或重新集中精力。可衡量的进展有助于保持整个组织的认同和热情。监管框架的变动、内外部政策、可治理方法和战略创新的变化是定期重新评估的其他原因。

补充内容：

数据管理能力成熟度评估模型（GB/T 36073-2018）

本标准给出了数据管理能力成熟度评估模型以及相应的成熟度等级，定义了数据战略、数据治理、数据架构、数据应用、数据安全、数据质量、数据标准和数据生存周期等 8 个能力域。

4 个交付物，数字化转型评估不等于 DCMM

评分结果、评估报告、符合性证书、数据管理发展路线图（不够详细，详细的另请公司）

第 16 章 数据管理组织与角色期望

16.3 数据管理组织的结构

16.3.2 网络运营模式

通过 RACI（谁负责，Responsible；谁批准，Accountable；咨询谁，Consulted；通知谁，Informed）责任矩阵，利用一系列的文件记录联系和责任制度，使分散的非正规性组织变得更加正式，称为网络模式。它作为人和角色之间的一系列已知连接运行，可以表示为“网络”。

16.4 关键成功因素

无论数据管理组织的架构如何，有 10 个因素始终被证明对其成功发挥着关键作用：

1）高管层的支持。
2）明确的愿景。
3）主动的变更管理。
4）领导者之间的共识。
5）持续沟通。
6）利益相关方的参与。
7）指导和培训。
8）采用度量策略。
9）坚持指导原则。
10）演进而非革命。

16.6 数据管理组织与其他数据相关机构之间的沟通

16.6.1 首席数据官

虽然大多数公司在某种程度上已认识到数据是有价值的公司资产，但只有少数公司指定了首席数据官（CDO）来帮助弥合技术和业务之间的差距，并在高层建立企业级的高级数据管理战略。然而，CDO 这一角色正在兴起。Gartner 认为，到 2017 年，所有受监管公司中有一半将聘用 CDO（Gartner，2015）。

虽然 CDO 的要求和职能受限于每个组织的文化、组织结构和业务需求，但许多 CDO 往往是业务战略家、顾问、数据质量管理专员和全方位数据管理大使中的一员。

2014 年，Dataversity 发布了概述 CDO 常见任务的研究。【一定会考】

1）建立组织数据战略。
2）使以数据为中心的需求与可用的 IT 和业务资源保持一致。
3）建立数据治理标准、政策和程序。
4）为业务提供建议（以及可能的服务）以实现数据能动性，如业务分析、大数据、数据质量和数据技术。
5）向企业内外部利益相关方宣传良好的信息管理原则的重要性。
6）监督数据在业务分析和商务智能中的使用情况。

16.7 数据管理角色

16.7.2 个人角色

1.执行官角色；2.业务角色；3.IT 角色 ；4.混合角色

混合角色需要同时具备业务和技术知识，根据组织的不同情况确定担任这些角色的人员是汇报给 IT 或业务部门。

1）数据质量分析师（Data Quality Analyst）。负责确定数据的适用性并监控数据的持续状况；进行数据问题的根因分析，并帮助组织识别提高数据质量的业务流程及技术改进。

2）元数据专家（Metadata Specialist）。负责元数据的集成、控制和交付，包括元数据存储库的管理。

3）BI 架构师（Business Intelligence Architect）。负责商务智能用户环境设计的高级商务智能分析师。

4）BI 分析师/管理员（Business Intelligence Analyst/Administrator）。负责支持业务人员有效使用商务智能数据。

5）BI 项目经理（Business Intelligence Program Manager）。负责协调整个公司的 BI 需求和计划，并将它们整合成一个整体的优先计划和路线图。

Q：数据管理专员是什么角色？

A：业务角色。

第 17 章 数据管理和组织变革管理

17.2 变革法则

1）组织不变革，人就变。
2）人们不会抗拒变革，但抵制被改变。
3）事情之所以存在是惯性所致。
4）除非有人推动变革，否则很可能止步不前。
5）如果不考虑人的因素，变革将很容易。

17.3 并非管理变革：而是管理转型过程

变革管理专家威廉·布里奇斯（William Bridges）【管理学大师，不是数据质量的】强调转型过程在变革管理进程中的核心地位。

17.4 科特的变革管理八大误区

约翰·科特（John P.Kotter）【不是数据质量的】是变革管理领域最受尊敬的研究者之一。

17.4.1 误区一：过于自满
17.4.2 误区二：未能建立足够强大的指导联盟
17.4.3 误区三：低估愿景的力量
17.4.4 误区四：10 倍、100 倍或 1000 倍地放大愿景
17.4.5 误区五：允许阻挡愿景的障碍存在
17.4.6 误区六：未能创造短期收益（国内 3-6 个月）
17.4.7 误区七：过早宣布胜利
17.4.8 误区八：忽视将变革融入企业文化

17.5 科特的重大变革八步法

17.5.1 树立紧迫感【可能会考】

在信息管理方面，促使紧迫感产生的因素有如下几种：
1）监管变化。
2）信息安全的潜在威胁。
3）业务连续性风险。
4）商业策略的改变。
5）兼并与收购。
6）监管审计或诉讼风险。
7）技术变革。
8）市场竞争对手的能力变化。
9）媒体对组织或者行业信息管理问题的评论。

17.5.3 发展愿景和战略

1.为何需要愿景

愿景是一幅关于未来的图景，其中隐含着人们为何要努力创造未来的明确或隐含的解释。一个好的愿景有三个重要特指：**明确性、动力性和一致性。**

2.有效愿景的特性【可能会考】

1）充满想象。描绘了一幅未来的图景。
2）吸引力。有利于增加员工、客户、股东和其他利益相关方的长期利益。
3）可行性。目标现实、可实现。
4）重点突出。为决策提供明确指导。
5）灵活性。它足够普适，允许个人采取主动，并在条件或约束发生变化时做出替代计划和响应。
6）可交流性。容易在 5 分钟或者更短时间内分享和清晰交流。

17.9 数据管理价值的沟通

17.9.1 沟通原则

1）有明确的目标和期望的结果。
2）由支持所需结果的关键消息构成。
3）为受众/利益相关方量身定制。
4）通过适合受众/利益相关方的媒介传达。

虽然沟通可能涉及一系列主题，但沟通的总体目标可以归结为：

1）通知。
2）教育。
3）设定目标或愿景。
4）定义问题的解决方案。
5）促进变革。
6）影响或激励行动。
7）获得反馈。
8）获得支持。