



西安电子科技大学
XIDIAN UNIVERSITY

第五章 数据质量管理



马 晶

经济与管理学院 信息管理系
Email: majing@xidian.edu.cn

课前讨论：数据质量引发的现实问题

信用卡欺诈



低质量数据增加企业成本

Gartner 2016 年的一项研究发现，
由于数据质量差，受访组织平均
每年将损失 960 万美元。



课前讨论：数据质量引发的现实问题

公告

中国南方航空 2023-11-09 12:46 发表于广东

11月8日晚间在系统异常期间南航售出的所有机票（支付成功并已出票）全部有效，旅客可正常使用。

公司在双十一期间推出更多优惠活动，邀请您一起参与。感谢您选择南航，南航将一路伴您美好出行！

中国南方航空

2023年11月9日

22时，记者查询南航的票价时发现，售价已全部恢复正常。

针对为何出现超低价机票，多家媒体致电南航客服。客服回应称，当晚20时左右出现的超低价机票是系统Bug导致，“10”“30”是因为系统故障出现的错误代码，并非票价。

数据质量的定义

数据质量

- 基于用户需求定义数据质量：
满足特定用户预期需要的程度；
- 基于数据本身定义数据质量：
包括真实性、完备性、自治性等；
- 基于约束关系定义数据质量：
包括数据的原子性、关联性、
满足约束的程度等；
- 基于生产和使用过程定义数
据质量：包括数据能被正确使
用、存储、传输、共享的程度。

数据一致性

数据集中每个数据在语义表达上是一致的，即不存在语义错误或矛盾。

数据精确性

数据集中的每个数据都能准确地描述现实世界中的实体，即不存在模糊或近似。

数据完整性

数据集中包含足够的数据来支持各种查询和计算，即不存在数据或记录的缺失。

数据时效性

数据集中的每个数据都与时俱进，正确描述当前时刻的数据对象，即不存在过时、失效的数据。

实体同一性

同一实体的标识在所有数据集中必须相同，语义表达必须一致。

表 6-1 数据质量维度（拓展）

数据质量维度拓展

序号	维 度	描 述
1	时效性	数据的老化度对当前工作的影响程度 信息系统对现实世界变化做出反应的及时程度
2	流通性	数据的实时程度 描述信息何时进入数据仓库
3	一致性	当前数据与历史数据有相同的形式且兼容的程度 对数据集定义的语义规则的违反项个数
4	精确性	数据库中描述现实世界实体的数据的准确程度 数据正确、可靠、被认证的程度 一个数据 v , 相对于另一个数据 v' 来说更正确的程度 在有可参考认证的权威来源情况下, 数据的正确程度
5	完整性	数据集能够表示现实世界系统中每个有意义状态的能力 对于当前工作, 数据的范围、广泛性、深刻性对任务的贡献程度 数据仓库中现实世界信息所占的百分比 非空数据的比率
6	可访问性	数据是否可用, 或者是否可被快捷地检索到
7	重复率	(不必要的) 重复数据的比率
8	规范性	衡量数据标准、数据模型、商业数据和参考数据的存在性、完整性、质量及文档记录是否合规
9	表现质量	数据的表示和收集、使用状态
10	声誉度	在存在对数据源和数据本身的额外描述时, 数据在来源和内容方面被重视的程度
11	无害性	数据产生的效果对人、过程或者环境的风险等级
12	适量性	现有数据量适用于当前任务的程度
13	安全性	数据获取权限的限定范围可确保其安全的程度
14	可信性	数据真实可信的程度
15	易懂性	数据明确、无歧义且易理解的程度
16	客观性	数据是否是公正的、公开的、无偏见的
17	切题性 & 有用性	数据对当前工作的适用性和贡献度
18	有效性	用户在指定的上下文中能使用数据准确完整地实现目标功能的程度

第五

序号	维 度	描 述
19	解释性	数据有适当且定义明确的语言、符号和单位的程度
20	易操作性	数据易被处理且能适用于其他相似格式的程度
21	易用度 & 可维护性	数据易于被获取, 以及数据易于被更新、维护、管理的程度
22	可靠性	在特定条件下, 能保持某个性能等级的程度
23	新鲜度	包含一系列的质量因素, 其中每个因素代表某方面的鲜度
24	附加价值性	在使用数据的过程中获取收益的程度
25	易学习性	数据易于被用户学习的程度
26	衰败率	数据发生消极变化的程度
27	简洁度	数据被简洁表示的程度 (不过分简略, 完整且切题)
28	同步性	在不同应用或系统中数据的等价程度, 以及使数据等价化的便捷程度
29	完备性	衡量数据的存在性、有效性、结构良好性及其他基本特性的指标
30	导航度	数据易于被发现及连接的程度
31	高效性	数据供给当前工作应用的便捷程度
32	可用性	在物理上数据可利用的程度
33	覆盖率	有效且完整的数据占总体数据的比率
34	交易度	数据能够产生商业交易或结果的程度
35	时效度 & 可用度	数据的实时程度及在特定条件下的可用程度
36	易变性	数据在现实世界有效的时间段长度



数据质量问题

问题的起源

- **来源不一致：**同一个现实世界中的对象可能在不同的场景中出现，每个场景对应的数据源提供的数据未必及时、准确、一致。
- **背景知识不一致：**参与采集和录入的人的背景知识强弱会极大影响采集和录入的数据质量。
- **计算和存储资源受限：**计算和存储资源不足，数据的精确性、完整性、时效性都会受影响。
- **安全性要求较高：**出于数据安全方面的考虑，精确度的牺牲经常是不可避免的。
- **数据本身的二义性：**描述不同的语义或者同一种语义有不同表达是普遍存在的。
- **数据语义的复杂性：**很多非结构化数据本身包含复杂的语义，进行转换和映射是很困难的。
- **管理手段不适当：**对数据的管理过于严苛或者过于疏松都可能导致数据质量下降。

第五章 数据质量管理

数据质量管理

数据质量管理是指对数据进行全面质量管理，涉及到技术、规则、组织、流程、评价考核等多个方面，目标是针对一个或多个所需要的维度，及时发现并解决数据质量问题，提升数据对应维度的数据质量。

数据质量管理的可分为人工管理、半自动化管理、自动化管理三个层次。

- **人工管理：**完全手工实现，管理的效果取决于数据质量管理者的时间、工作态度、对于数据的熟悉程度，以及背景知识的充分程度等。
- **半自动化管理：**通过制定规则、程序比对、统计分析等方法，先完成基本的劣质数据检测、筛选、处理，将结果转给人工管理者，由其决定最终要采取的措施并反馈给自动化管理模块。
- **自动化管理：**完全依靠自动化和智能化系统完成数据质量的评估和管理。



第五章 数据质量管理

数据质量评估

数据一致性评估

对数据一致性的研究是为了保证数据不违背特定场景下的语义约束。评估数据一致性，直观上就是要了解数据集中究竟有多少数据的表述是不存在矛盾的。

数据库中最常见的**函数依赖**（**Functional Dependency**）就是典型的可用于数据一致性评估的规则。函数依赖仅能检测到数据中的部分不一致错误，很多数据一致性问题无法通过函数依赖表达和检测。

条件函数依赖（**Conditional Functional Dependency**）可以表示部分数据上的数据依赖关系，从而可被用来评估数据一致性以及检测修复不一致数据。



函数依赖：“邮编→城市”

数据库中只要两条记录的邮编相同，那么城市就必然相同。

条件函数依赖： $(X \rightarrow Y, T_p)$

其中， X 和 Y 是两个属性子集， $X \rightarrow Y$ 是标准的函数依赖， T_p 是关于 X 和 Y 的模式表，用于约束 X 和 Y 的取值。

$$\varphi: ([\text{员工}, \text{部门}] \rightarrow [\text{项目}], T_p)$$

$T_p:$	员工	部门	项目
	A		

第五章 数据质量管理

数据完整性评估

对数据完整性的研究是为了保证数据中不存在缺失值。在最简单的情况下，可以对直接对数据集进行扫描，检查数据缺失的比例，完成数据完整性评估。此评估方法虽然简单，但过于粗糙，因为有些NULL值可能是“伪缺失值”。可能有下述两种情况：

- (1) 数据值本身就应该为空；
- (2) 数据虽然缺失，但可以通过其他手段正确填补。

ID	姓名	性别	城市	邮编	婚姻状态	配偶
001	张三			510000	离异	
002	李四	男	广州	510000	未婚	无

示例数据表中共包含14个数据项，其中5个为空（NULL），若直接使用 $1 - \frac{\text{count(NULL)}}{|D|}$ 计算，则数据完整性评估结果为 $1 - \frac{5}{14} \approx 0.643$ 。

在满足数据一致性的条件下，婚姻状态“离异”和“未婚”均为无配偶，若函数依赖“邮编→城市”

成立，则可以将表中的城市补充为“广州”。此时，数据表中仅剩1个缺失数据项，其完整度评估结果为 $1 - \frac{1}{14} \approx 0.929$ 。

第五章 数据质量管理

数据完整性评估

为了处理上述两种情况，需要向数据中引入语义规则。是否适用语义规则可能会很大程度影响数据完整性的评估结果，在使用了恰当的语义规则后，完整性评估结果的准确度会提升。语义规则通常有两种获取途径：

- (1) 由领域专家给出；
- (2) 采用类似于挖掘关联规则的方法从高质量的数据集中挖掘得出。

很多时候上述方法评估得到的结果还未必准确，因为数据未必能够满足封闭世界假设，即“非已知的事物都为假”（所有应该存在的记录都包含于当前数据集）的假设。

包含依赖指明了多个数据表之间的关系，因此可以被用来度量数据完整性。

包含依赖通常表示关系之间的关系（也可以理解为不同的数据集之间的关系）。针对图书采购和展览相关的数据集写出包含依赖，其语义是，图书展览表的 ISBN 列的所有值都应当出现在图书采购表的 ISBN 列中。

$\varphi: \text{图书展览} . ISBN \sqsubseteq \text{图书采购} . ISBN$

图书采购表：

ID	ISBN	书名	出版社	作者	价格

图书展览表：

ID	ISBN	书名	陈列楼层	陈列展厅	行列

数据时效性评估

- 对数据时效性的研究是为了保证数据不陈旧过时。评估数据时效性时，**时间戳**显得尤为重要。
- 理想情况下，数据库中对每个值都存有完整可用的时间戳，不仅包括数据的创建时间，也包括数据的截止时间、有效期、生命周期等。可以基于时间戳来统计数据库中有多少值是已经过时失效的，进而通过过时失效数据的比例来评估整个数据集的时效性。
- 然而在很多情况下，**完整、可用、精确的时间戳**可能会无法获得，或者获取代价极高。
- 为此，应当有不完全依赖于时间戳的数据时效性评估方法，来判定数据时效性。可以将数据时效性划分为两类：
 - (1) **绝对时效性**：可以对给定数据库，形式化地评估单个数据项、元组及数据库整体的时效性；
 - (2) **相对时效性**：针对数据库上的特定查询或分析需求，度量数据库相对于查询或特定分析需求的时效性。

数据精确性评估

- 对数据精确性的研究是为了保证数据能够准确地描述对应实体。
- 数据精确性评估主要考察数据相对于某个标准是否能足够准确地描述对象。
- 数据精确性需要有明确的评价标准，可以扫描数据集中所有存在精确度要求的数据，统计不符合精确度要求的数据量，并以符合精确度要求的数据的百分比作为最后的精确度评估结果。
- 很多情况下，导致数据不精确的未必是质量问题。因此，精确性的评估要放在具体应用值来看。如果数据本身不太精确，而查询要求返回非常精准的结果，就会导致查询结果不对。

城市	人口（万）
A	1531

城市	人口（万）
A	1530.59

如果数据中存在多个精确度不同的副本，那么在查询精确度要求较高时，我们可以尝试在数据集中先寻找满足精确度要求的数据副本，再基于这些数据副本计算查询结果，并度量查询结果的精确性。



第五章 数据质量管理

实体同一性评估

对实体同一性的研究是为了保证描述同一实体的数据是一致的。

两条记录的身份证号相同（可以肯定是同一个人），但性别不同，这就出现了实体同一性问题。

实体同一性问题看起来与数据一致性问题非常类似，但二者的关注点不同。

数据一致性评估关注数据集在整体上是否存在不一致数据，而实体同一性则专门关注同一实体的描述信息中是否存在矛盾。

序号	姓名	身份证号	性别	生日
001	张三	123456200101017890	男	2001-01-01
.....
102	张三	123456200101017890	女	2001-01-10

其他数据质量维度的评估或修复或许会影响实体同一性的评估结果。

实体同一性的评估效果有很大一部分依赖于我们是否能够准确发现哪些记录描述了同一实体。

缺失值填充

什么是缺失值

在大多数情况下，信息系统中的数据都存在某种程度的缺失。造成缺失值的原因是多方面的，主要可能有以下几种：

- **数据难以获取：** 存在各种原因导致时延，或出于权限、隐私保护等原因，数据不能被采集。
- **数据采集时有侧重：** 对数据的重要性有所权衡，代价高昂但作用较小的数据可能会被忽略。
- **数据在流动或使用时有丢失：** 出于技术、制度、流程、操作等原因，一些数据可能丢失。

出于上述原因，数据缺失的方式可以大致分为三类：

- **完全随机缺失：** 缺失值与当前信息系统中的其他值无关。
- **完全依赖缺失：** 缺失值与信息系统中的某些值有完全依赖关系。
- **部分依赖缺失：** 缺失值与系统中的某些值有依赖关系，但是这种依赖是部分的或者有概率的。

缺失值填充

缺失值处理方法

大致说来，处理缺失值的操作主要有三类：删除、填充、忽略。

1. 删除

- 需要讨论是删除记录，还是删除属性。
- 对于关系型数据库，记录和属性分别对应着数据表的行和列。判断如何删除的方法比较简单，基本根据记录或者属性的完整程度来确定。
- 如果某个属性的大部分值都缺失，那么该属性就可以删除，提高数据集的完整性。如果某条记录的大部分属性都缺失，那么这条记录可以删除。
- 删除数据集中的数据时往往需要特别谨慎。有时删除数据可能会丢失数据的隐藏语义，进而使得某些数据分析结果不准确。
- **适用于：数据关键性较低、关联关系不大且数据缺失较为集中的简单情况。**

2. 填充

(1) 重新采集

通过之前使用过的数据采集手段重新采集数据，适合于数据采集代价较低且可以复现的场景。

(2) 默认值填充

一种较为常见的自动填充方法。将缺失的部分填为“Null”、“0”、“ $+\infty$ ”、“ $-\infty$ ”等。

(3) 统计填充

通过统计的方法来填充数据。简单的填充方法包括使用平均数、中位数、最大值、最小值等。

(4) 热卡填充（或就近补齐）

对于一个包含空值的对象 o ，热卡填充法在数据库中找到一个与它最相似的对象 o' ，然后用 o' 的值来填充 o 的对应位置的值。具体的相似度根据实际应用场景有不同的定义。

(5) 期望最大化（Expectation maximization, EM）

EM算法是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。在每一次迭代循环过程中交替执行两个步骤——E步和M步，算法在 E步和 M步之间不断迭代直至收敛。

(6) 预测填充

通过对现有数据建模，预测缺失位置可能的值，用预测结果来填充缺失值。这种方法比较复杂，但取得的效果也比较好。

第五章 数据质量管理

表 6-3 缺失属性和简单的属性分析及处理方法举例

序号	属性	属性分析及处理方法
1	年收入	<p>属性分析：数值型，涉及个人隐私，可基于职业、年龄等信息推断</p> <p>处理方法：有个人历史信息的场景可使用预测填充（回归模型）；有大量相似记录时可以使用热卡填充或预测填充；商品推荐场景可使用平均值填充；借贷场景可使用最小值填充；如非必须填充，出于隐私保护的角度考虑，可不处理（忽略）</p>
2	运动轨迹坐标	<p>属性分析：可能是数值型的二元组/多元组，如经纬度坐标 (x, y) 或三维地图坐标 (x, y, h) 等；也可能是脱敏后的字符串，例如，一些场景下，运动轨迹可以通过用户接入的移动基站来表示轨迹坐标</p> <p>处理方法：可以在数据库中寻找相似轨迹从而使用热卡填充；也可以对坐标划分类别，使用预测填充</p>
3	人体寿命	<p>属性分析：数值型，进行预测评估需要“病史”等隐私数据，进行预测建模相对困难</p> <p>处理方法：保险费用估计场景下使用最大值填充；人口估计场景下使用平均值填充；非必须填充的场景可以不处理</p>
4	邮编	<p>属性分析：字符串型，与“城市”“街道”属性之间存在依赖</p> <p>处理方法：以中国地区的邮编为例，在“城市”“街道”已知的场景下，可基于依赖关系补全详细的邮编；在只知道“城市”的场景下，可以基于属性依赖关系，使用城市默认的邮编进行填充；在不存在相关属性的情况下，可使用默认值填充</p>
5	地址	<p>属性分析：字符串型，可以基于“工作单位”“邮编”等相关属性推断，但由于属性间不存在绝对的依赖关系，所以只能近似推断</p> <p>处理方式：在存在“工作单位”属性的场景下，可以使用工作单位的地址填充；在存在“邮编”时，可以使用邮编对应的城市和街道填充。需要注意的是，以上填充结果均是近似的，实际的真实值可能和填充结果相差甚远</p>



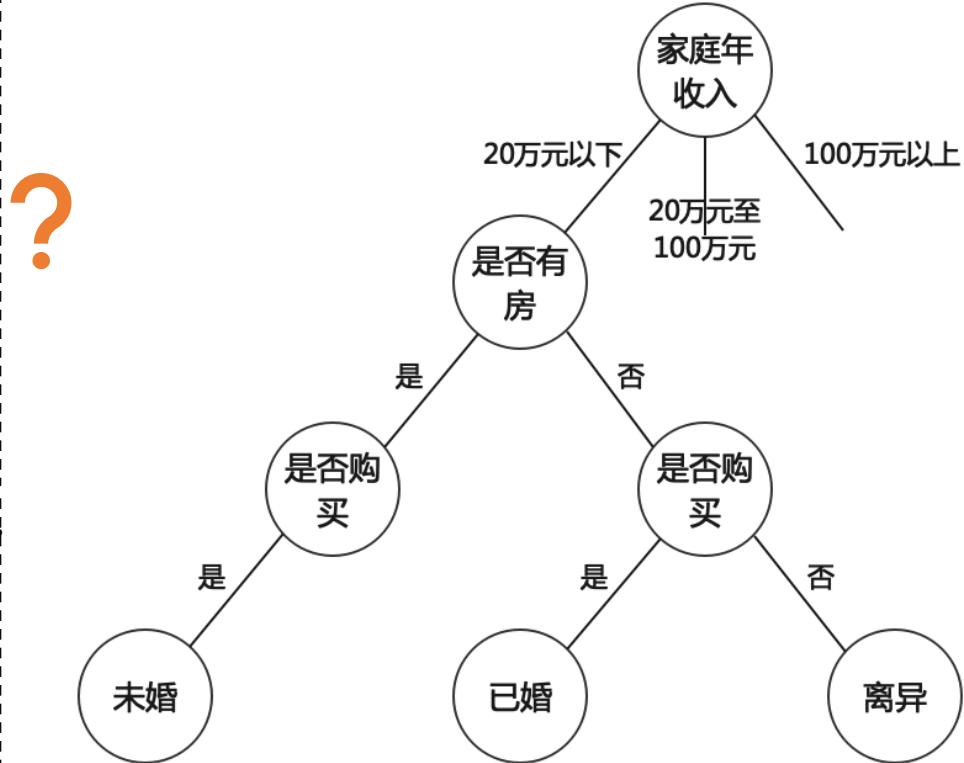
第五章 数据质量管理

例：某公司有一份数据记录了客户是否最终购买了他们的产品及客户的一些个人信息，其中属性包括“家庭年收入”“是否有房”“婚姻状况”，最后的类别为“是否购买”。不过这份数据在“婚姻状况”的属性上存在缺失值，但是公司需要使用“婚姻状况”来进行后续分析，因此必须对缺失值进行填充。

一种可行的思路是，我们将“婚姻状况”（未婚、已婚、离异）作为分类标签，而将“是否购买”挪作一个已知的属性来构建一棵决策树。

这种方法可能要比我们随机填上“未婚”“已婚”“离异”或者填上默认的“NULL”效果好。

家庭年收入 是否有房 婚姻状况 ... 是否购买



实体识别与真值发现

实体识别

实体识别是指，在给定的实体对象（包括实体名和各项属性）集合中，正确发现不同的实体对象，并将其聚类，使得每个经过实体识别后得到的对象簇在现实世界中指代的是同一个实体。

实体识别要解决的问题主要包括以下两类：

- (1) **冗余问题**: 同一类实体可能由不同的名字指代;
- (2) **重名问题**: 不同类的实体可能由相同的名字指代。

针对不同类型冲突的处理，实体识别中主要有两类技术：

- (1) **冗余发现**: 用于处理冗余问题，构造对象名称的相似性函数，并与阈值进行比较，从而判定对象是否属于同一实体簇；
- (2) **重名检测**: 用于处理重名问题，利用基于聚类的技术，通过考察实体属性间的关联程度判定相同名称的对象是否属于同一实体簇。

第五章 数据质量管理

基于规则的实体识别方法

实体识别的目标就是要识别数据集中指代同一实体的元组。传统的识别方法通过比较元组对的相似性来识别实体。此类方法非常直观，但在实际应用中却可能存在一定的问题。

ID	name	coauthors	title	class
o_{11}	Wei Wang	zhang	inferring...	e_1
o_{12}	Wei Wang	duncan, kum, pei	social...	e_1
o_{13}	Wei Wang	cheng, li, kum	measuring...	e_1
o_{21}	Wei Wang	lin, pei	threshold...	e_2
o_{22}	Wei Wang	lin, hua, pei	ranking...	e_2
o_{31}	Wei Wang	shi, zhang	picturebook ...	e_3
o_{32}	Wei Wang	pei, shi, xu	utility...	e_3

在这个例子中共有7个元组，包含3个作者实体。由于他们的名字相同，因此不能用名字来区分不同实体；论文标题的相似度较低，也不能用于识别相同实体。可以考虑使用合作者（coauthors）信息。由于Jaccard相似性常被用于测量集合的相似度，因此可以进一步考虑。

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

第五章 数据质量管理

ID	name	coauthors	title	class
o_{11}	Wei Wang	zhang	inferring...	e_1
o_{12}	Wei Wang	duncan, kum, pei	social...	e_1
o_{13}	Wei Wang	cheng, li, kum	measuring...	e_1
o_{21}	Wei Wang	lin, pei	threshold...	e_2
o_{22}	Wei Wang	lin, hua, pei	ranking...	e_2
o_{31}	Wei Wang	shi, zhang	picturebook	e_3
...				
o_{32}	Wei Wang	pei, shi, xu	utility...	e_3

$$sim(o_{11}, o_{12}) = 0$$

$$sim(o_{11}, o_{31}) = 1/2$$

$$sim(o_{12}, o_{13}) = 1/5$$

$$sim(o_{12}, o_{21}) = 1/4$$

可以看出如果根据相似性比较，我们很难得到正确的实体识别结果。采用其他的相似性度量方式也都有同样的问题。

第五章 数据质量管理

ID	name	coauthors	title	class
o_{11}	Wei Wang	zhang	inferring...	e_1
o_{12}	Wei Wang	duncan, kum, pei	social...	e_1
o_{13}	Wei Wang	cheng, li, kum	measuring	e_1
			...	
o_{21}	Wei Wang	lin, pei	threshold...	e_2
o_{22}	Wei Wang	lin, hua, pei	ranking...	e_2
o_{31}	Wei Wang	shi, zhang	picturebook	e_3
			...	
o_{32}	Wei Wang	pei, shi, xu	utility...	e_3

基于上面的例子，我们可以得到一些观察结果：

- 观察1：某些属性值对的存在对识别元组很有用。例如，“lin”只出现在指代实体 e_2 的元组中。
- 观察2：某些属性值对的不存在也能帮助识别元组。例如，元组 o_{11} 和 o_{31} 中都有（coauthors，“zhang”）的存在，但却并不指代同一实体，但由于指代 e_3 的元组中都包含（coauthors，“shi”），因此（coauthors，“shi”）的不存在可以帮助我们排除 o_{11} 指代 e_3 的可能性。

基于以上观察，我们可以整理出类似下面的规则来识别表中的元组。

R1: $\forall o_i$, 如果 $o_i[\text{name}]$ 是“Wei Wang”且 $o_i[\text{coauthors}]$ 包含“pei”，那么 o_i 指代实体 e_2 。

实体识别规则

可以用 $A \Rightarrow B$ 来表示实体识别规则“ $\forall o$ ，如果 o 满足约束 A ，那么 o 指代实体 B ”。

为了方便，在下面的定义中，我们令 o 表示一个元组， U 表示一个数据集， r 表示一个实体识别规则， R 表示一个实体识别规则集， $LHS(r)$ 和 $RHS(r)$ 分别表示规则 r 的左部和右部。

基于规则的实体识别算法 有了有效的规则，基于规则的实体识别就变得非常直接，也就是通过匹配规则的左部即可以实现有效实体识别。

右侧实体规则识别的伪代码中，第7行表示计算 o_i 指代实体 e_j 的置信度，第8行表示选择置信度最大的实体。

实体识别规则可以由人手工撰写，也可以由程序从数据中自动学习得到。

```

Input: 数据集 U, 规则集 R, 阈值  $\theta_c$ 
Output: U 的划分  $\mathbb{U}$ 
1   for E 中的每个实体  $e_j$  do
2        $U_j \leftarrow \emptyset$ ;
3   for U 中的每个  $o_i$  do
4        $R(o_i) \leftarrow \text{FindRules}(o_i)$  ;
5       for E 中的每个实体  $e_j$  do
6            $R(e_j) \leftarrow \{r \mid RHS(r) = e_j\}$  ;
7            $C(o_i, e_j) \leftarrow \text{CompConf}(R(o_i) \cap R(e_j))$  ;
8       SelEntity( $o_i, \theta_c$ ) ;
9   return  $\mathbb{U} \leftarrow \{U_1, U_2, \dots, U_m\}$  ;
  
```

真值发现

在经过实体识别之后，描述同一个现实世界实体的不同元组被聚到了一起，这些对象的相同属性可能包含冲突值。在描述同一实体同一属性冲突值中发现真实值的操作称为**真值发现**。

目标：在多个对同一实体的描述信息中，找出最为准确的描述。

形式化地，**真值发现问题**定义如下：

我们考虑一组数据源 \mathcal{V} 和一组对象 \mathcal{O} 。一个对象代表了一个现实世界实体的某个特定方面（在关系数据库中，一个对象对应一个表中的一个单元格）。对于每个对象 $o \in \mathcal{O}$ ，数据源 $v \in \mathcal{V}$ ，可以（但不一定）提供真值。提供的这些值被称为“事实”。在为一个对象提供的不同事实中，一个事实正确地描述了真实世界，所以它是真实的，其余的是假的。给定一组数据源 \mathcal{V} ，我们要为每个满足 $o \in \mathcal{O}$ 的对象 o 判断其真相。

真值发现方法与技术

真值发现中最为直观的方法是**投票法**，即对同一个对象的不同取值进行投票，将得票多的描述作为真实值。投票时不同数据源对应的描述的权值应有所不同，这取决于**数据源的精度**。此外，由于现实生活中不同数据源之间存在复制现象，错误数据可能在多个数据源间传播，称为**数据源之间的依赖**。

1. 迭代地计算数据源的精度

数据源的精度将决定投票时相应描述对应的权值。

计算数据源精度的依据是，一个数据源为其他对象提供的真相越多，越有可能为当前的对象提供真相，其精度越高。

2. 数据源之间的依赖关系

如果两个数据源之间存在数据复制，则称它们之间存在依赖关系。

数据源之间的依赖会导致实体的某些描述得票偏高，于是我们可以通过减小它们的权值来减少依赖带来的影响。我们可以通过统计两个数据源对同一实体所提供的描述相同的比例计算两个数据源之间的依赖程度，但要注意，对于某个实体，当两个数据源提供的都是真值时，它们在这个实体上不应被视为复制关系。

数据冲突

由于信息错误、记录丢失、输入错误、数据过时等情况可能存在于不同的数据源中，来自多个数据源的信息常常相互矛盾冲突。三种典型的多源数据冲突场景：

(1) 同一实体信息冲突

- 在互联网上可以通过多种不同途径获取多个数据源对同一实体的描述信息。然而，得到的信息并不能保证是一致的。
- 例如，在搜索引擎中搜索“珠穆朗玛峰的高度”，返回结果包括“29035英尺”、“29002英尺”和“29029英尺”三个不同网站的结果。

(2) 同类实体信息冲突

- 在现实世界中，可以收集到多个数据源对同类实体的描述信息。然而，不同数据源对相同类别的实体描述也可能是不一致的。



(3) 同类实体多语义表达

- 当多源数据以文本形式存在时，描述相同或相似实体的信息可能以不同词汇和句子结构表达出来。
- 例如，来自三个数据源的有关实体China和属性president的文本分别为：
 - The U.S. President Donald Trump visited China weeks ago.
 - Europe is chosen as first overseas trip of China's president Jinpin Xi.
 - There was a card printed 'China, Jintao Hu' on the table indicated it was the position of Chinese president.

上述三个句子中哪句表达了China现阶段的president的真实信息呢？

以上三个场景简要说明了数值型数据、文本型数据等不同类型的多源数据组成都可能在同一实体、同类实体等不同层次上产生冲突。多源数据的真实信息挖掘，也就是多源数据的真值发现，正是面向多类型数据、多层次冲突找到粒度更细、内容更准、价值更高的实体信息的重要技术。

当前，面向多源数据的真值发现技术主要面临**结构和算法**两方面的挑战。

错误检测与修复

格式内容清洗

如果数据的来源是系统日志，那么通常在格式和内容方面会与元数据的描述一致，而如果数据是由人工收集或用户填写而来的，则有很大可能在格式和内容上存在一些问题。简单来说，格式内容问题有以下几类：

(1) 显示格式不一致

这种问题通常与输入端有关。

(2) 有非法字符

最典型的情况就是头、尾、中间包括空格，或姓名中存在数字符号、身份证号中出现汉字等。

(3) 内容与该字段应有内容不符

在一些情况下，用户误将本来属于一个属性的数据填写到了另一个属性中。

逻辑错误清洗

逻辑错误清洗的工作是去掉一些通过逻辑推理就可以发现问题的数据，防止分析结果的偏差。

主要包含如下几个步骤：

1. 去重

去重就是去掉数据中的重复信息。由于数据存在的同名（不同的事物具有相同的名字）和异名（相同的事物具有不同的名字）情况，去重通常要通过上文介绍的实体识别技术来实现。

2. 去除不合理值

有时候用户会填入一些不合理值，需要有效检测和修复这种不合理值。这类不合理值的检测主要依靠属性值上的约束。由于这类不合理值提供的有用信息非常少，因此其修复需要按照缺失值处理。

3. 修正矛盾内容

有些字段是可以互相验证的。这种错误的检测可以通过规则来实现，这就经常用到函数依赖和条件函数依赖。

面向图数据的错误检测

知识图谱的构建有两种手段：手动构建的知识图谱和自动构建的知识图谱。

(1) 简单方法

- ◆ **使用频率修正错误：**目前较多的针对知识图谱的错误检测算法都着重于利用知识图谱本身所具备的频率。存在的问题：无法处理频率较为集中的知识图谱关系。
- ◆ **引入新知识关系：**引入新的知识对现有知识进行修正也是很多目前知识图谱使用的方法。引入的新知识包括：另一个知识图谱中所包含的知识关系、新的网络语料库关系等。这种方法利用外部知识库来消除冲突并提高分类的质量。存在的问题：忽略了不同知识图谱的独特性。

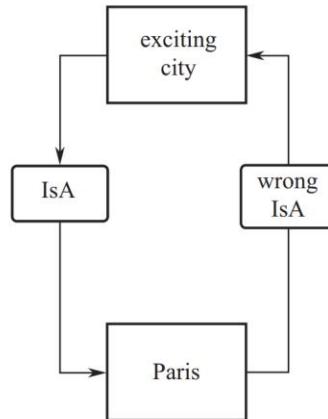


面向图数据的错误检测

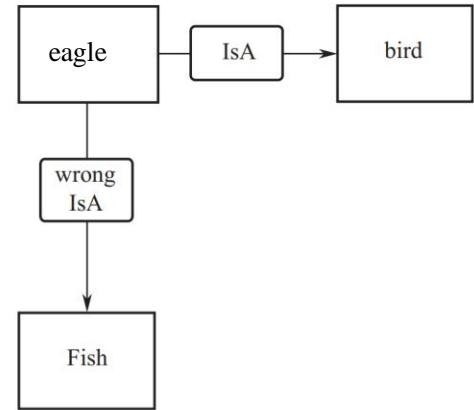
(2) 高级方法

◆ 统计学回归方法：是一种数学方法，使用统计技术用于数据整合。主要针对于处理知识图谱当中的线性关系，通过处理数字属性之间的定量关系，从而得到对于线性关系的回归方法。存在的问题：对于非线性关系无法得出有效的解决方法。

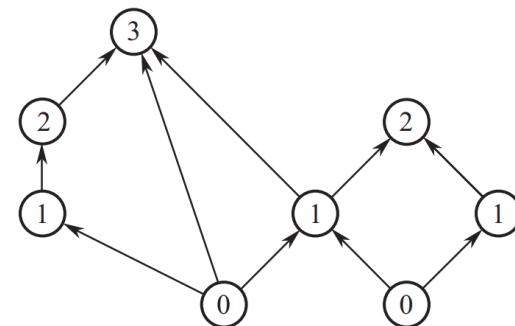
◆ 查找圈关系：最新提出的一种方法，着重点是将知识图谱中间存在的关系视为一种有向图关系，并使用相关算法在该大图上查找圈关系。若存在圈关系，则判定该圈中包含的关系至少有一个是错误的。存在的问题：较低的错误召回率。



(a) 存在圈结构的错误



(b) 不存在圈结构的错误



IsA 关系有向无环图

数据问题根因分析

根因分析

所谓根因分析，就是分析导致数据质量问题的最基本原因。根因分析是一个系统化的问题处理过程，包括确定和分析问题原因，找出适当的问题解决方案，并制定问题的预防措施。

对于企业而言，数据质量问题通常只是一个现象，人们往往只看到数据不准确、不一致、不完整，却没有细致地剖析这些问题发生的原因。尽管技术手段可以改善数据质量，但在此之前，我们应当进行根因分析，以防止“治标不治本”。

进行数据质量问题的根因分析，不仅在于解决业务部门和技术部门的矛盾，更重要的是能够帮助企业利益相关者发现数据质量问题的症结所在，从而找到适当的解决方案。



问题产生阶段

数据的“一生”要经历规划设计、数据创建、数据使用、数据老化、数据消亡五个阶段，每个阶段都可能产生数据质量问题。

- 程序员小K为某程序创建了一个手机号码表，并对其设置了约束条件—11位数字，然而这个程序是跨境使用的。
- 业务员在录入数据时误将“客户地址”填入了“客户名称”输入框。
- 在将ERP与CRM系统中的客户数据进行集成时，没有导出完整的客户信息表。
- 企业员工信息表中的联系地址与联系方式已经超过2年没有更新。
- 3年前的财务账目数据在归档迁移时出现了遗漏，导致系统更新时某业务部门的财务数据无据可查。



根因分析工具

鱼骨图

- **特征**就是“问题的结果”，例如同一客户不能唯一标识。
- **主骨**用来引出问题，问题写在右端。
- **大骨**用来表示问题的直接原因，例如人员因素、系统因素、技术因素等。
- **中骨**用来描述事实，例如操作失误等。
- **小骨**用来描述为什么会这样，例如系统操作不熟悉、随意输入等。
- **主因**用椭圆圈定，在大骨、中骨、小骨每一级均有可能发生。

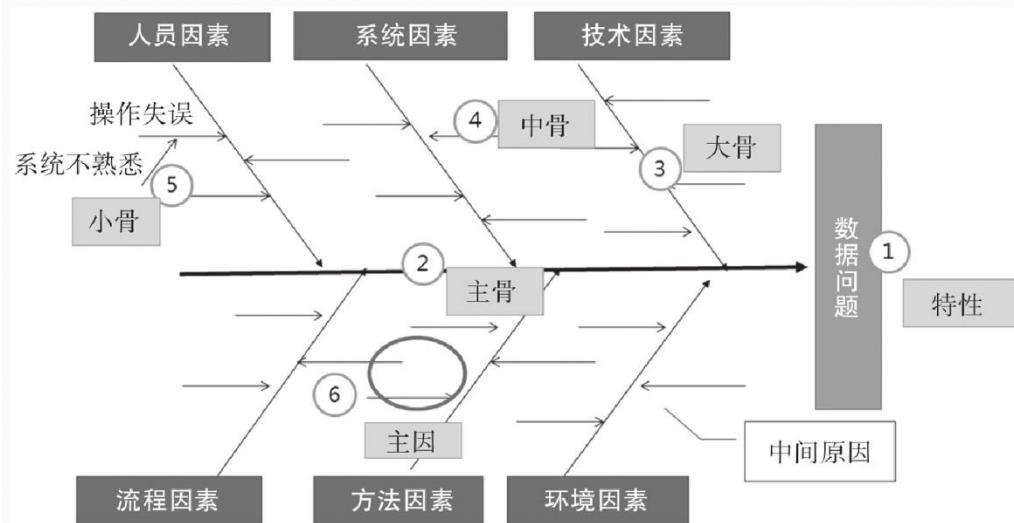


图19-5 鱼骨图分析法

第五章 数据质量管理

5Why图

- 5Why图也称5Why分析法或丰田5问法，首创是丰田公司大野耐一。5Why的精髓就是多问几个为什么，顺藤摸瓜，穿越不同的抽象层面，直至找出原有问题的根本原因。

举例：分析为什么同一客户在数据中不能唯一识别？

➤ 为什么不能识别？

数据中至少有两条重复记录，这是现象。

➤ 为什么会有重复记录？

数据源系统中的客户数据出现重复，这是直接原因。

➤ 为什么数据源系统中客户数据会重复？

业务员输入客户数据时重复了，这是进一步原因。

➤ 为什么业务员会重复输入？

新来的业务员对系统操作不熟悉，这是深入原因。

➤ 业务员不熟悉系统就会重复输入吗？

系统缺乏对客户唯一ID的校验，这是根本原因。



图19-6 5Why分析法

故障树图

- 故障树图是一种逻辑因果关系图，是故障事件在一定条件下的**逻辑推理**方法，可针对某一故障事件进行层层追踪分析。
- 使用故障树图来确定数据还来那个问题的可能原因时，可采用自上而下的推演方法。首先分析问题的直接原因，将顶问题作为逻辑输出的事件，将所有引起顶问题的直接原因作为输入事件，将它们之间的逻辑关系适当的连接起来。然后每一个中间问题用同样的方法逐级向下分析，直到所有输入问题都不需要再分解（找到根本原因）为止。

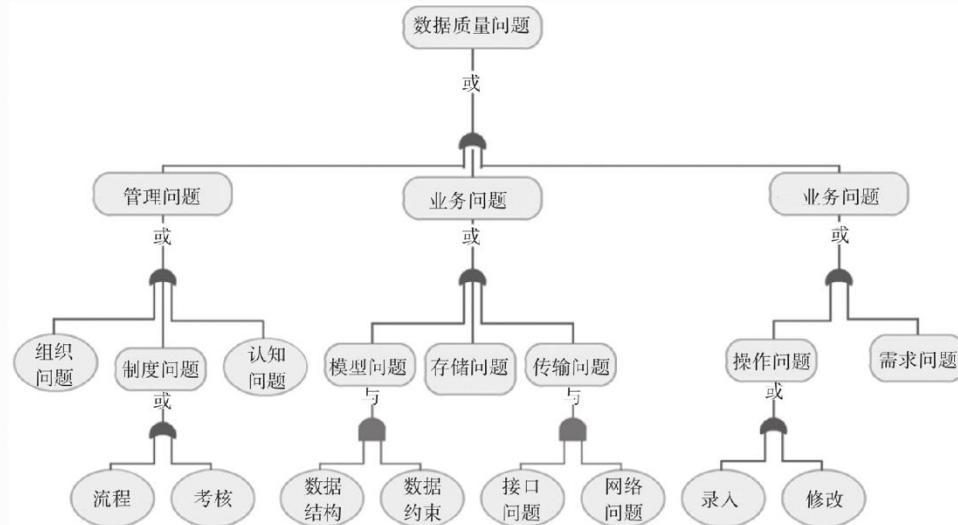


图19-7 故障树分析法

帕累托图

- 帕累托图是柱状图和折线图的结合，柱状图反映问题发生的频率，折线表示累积频率，横坐标表示影响质量的各项因素，按出现频数从左向右排列。
- 通过对排列图的观察分析可以抓住影响质量的主要因素，进而确定问题的优先级。
- 帕累托图是根据80/20法则的分析，即80%发生的问题是由20%的原因引起的。这意味着，有针对性的对主要问题制定解决方案，可以解决大部分的数据质量问题。

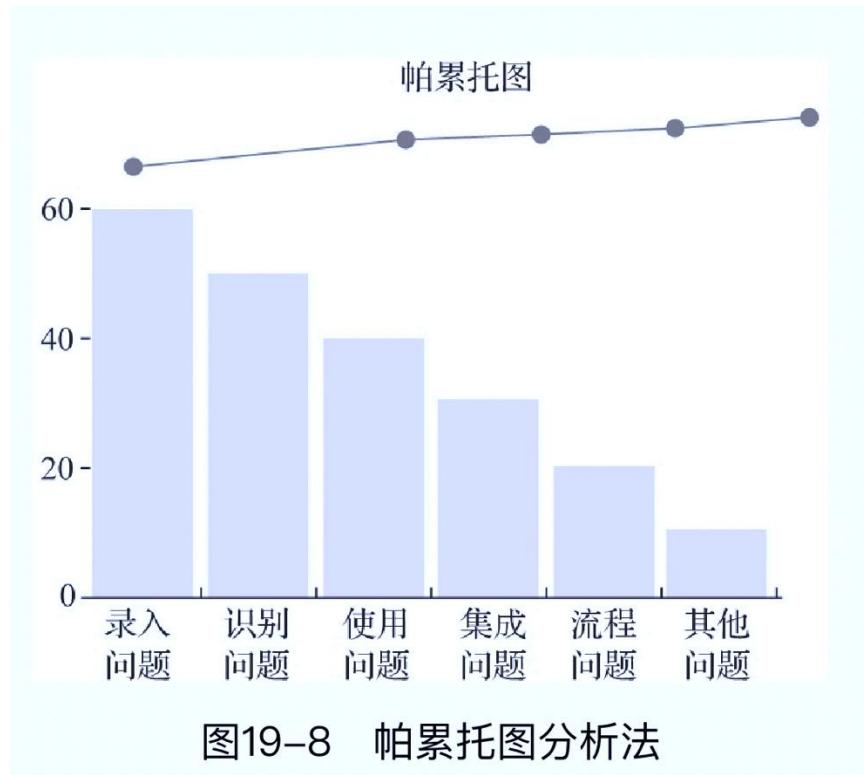


图19-8 帕累托图分析法

数据质量评估框架

DQAF

- **DQAF (Data Quality Assessment Framework, 数据质量评估框架)** 是国际货币基金组织 (IMF) 以联合国政府统计基本原则为基础构建的数据质量评估框架体系，于 2003年7月正式发布。
- DQAF最初目的是建立一种测量数据质量的方法，为数据消费者提供有意义的数据测量结果，并帮助提高数据质量。DQAF对数据质量内涵的界定比较完整，归纳性也比较强，同时提供了句的数据质量测量类型和数据质量指标，并给出了相应的详细解释，这些因素使该框架的可操作性较强。
- 作为一个权威性的国际规范，DQAF所采用的标准定义、概念和良好的统计实践可以用于全面、客观地评估统计数据的质量，为企业的数据质量管理提供可借鉴的范本。

DQAF

- **测量的原因（Why）**：数据测量维度，用来确定测量数据的哪些方面。
- **测量的方法（How）**：分为持续测量和定期测量两种。

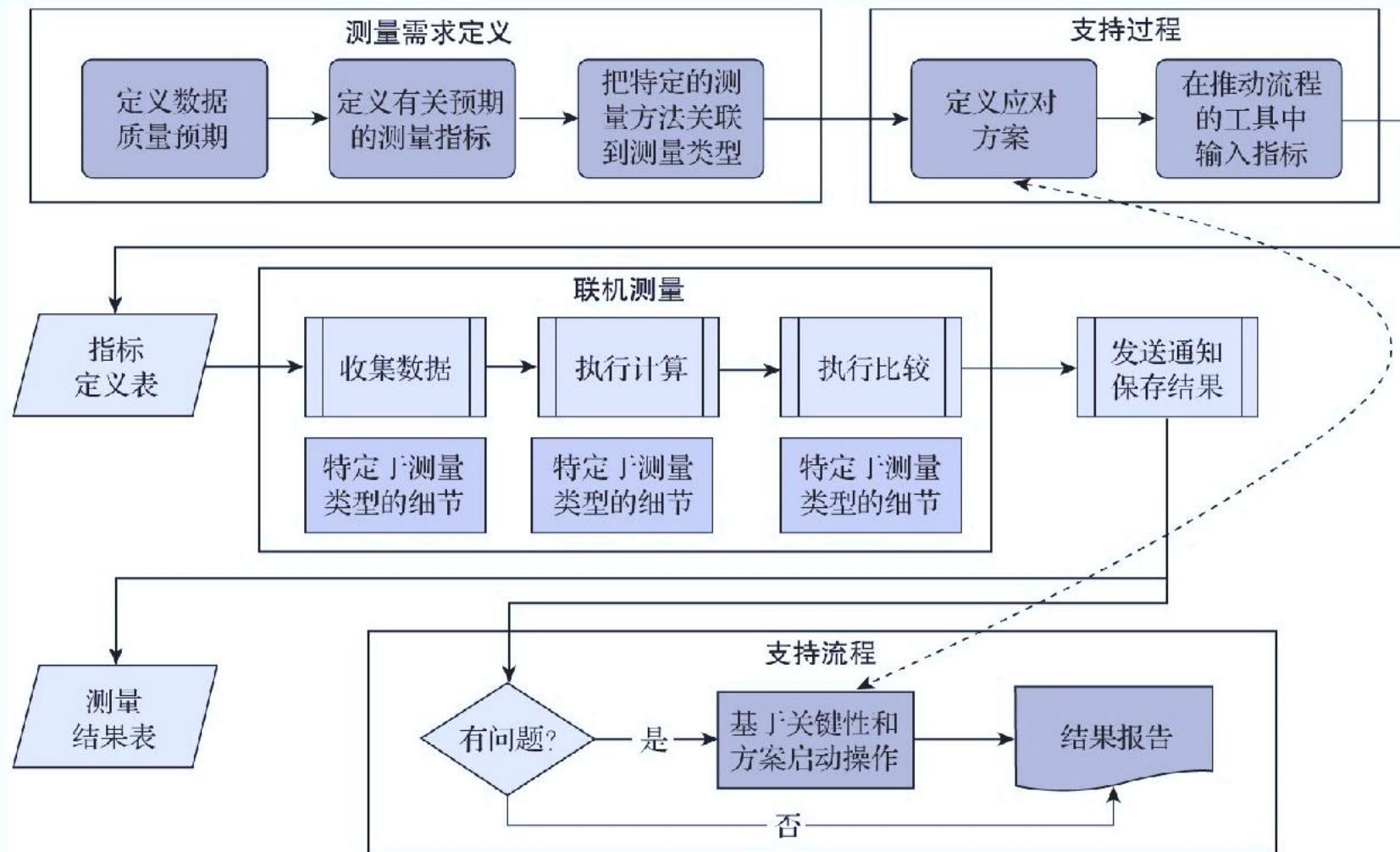
持续测量即对关键的有风险的数据源实施联机持续测量，目的是维持数据质量，它有三个任务：一是监控数据的状况并为数据在某种程度上符合预期提供保障；二是监控和发现数据质量问题；三是确定改进的机会。

定期测量即对非关键性数据和不适合持续测量的数据进行定期重新评估，为数据所处状态符合预期提供一定程度的保障。定期评估可以确保参考数据保持最新，预防业务和技术演进导致意外的数据更改。

维度 测量的原因 Why	完整性	唯一性	一致性	有效性	及时性	降低抽象（增加特定性、 具体性），更接近数据
测量类型 测量的方法 How	确认表之间参 照数据的完整 性，找出没有父 记录的数据	按照一定 的规 则对某一数 据集进 行查看	对某一数据在 不同数据源中 的数据元的值 进行比较	对输入数据的 值与已定义的 有效值域进 行比较	对数据传输的 实际时间与计 划时间进 行比较	
数据质量指标 测量的内容 What	找到未分配 角色的用户	CRM系统中 同名的 客户数量	数仓中销售 额与CRM系 统中销售 额的差异	职工分析中 职工的年龄 的有效性	记录客户下 单到发货的 时间范围	

- **测量的内容（What）**：数据质量测量的内容通常称为数据质量的指标，即衡量数据质量目标的参数，于其中要达到的指数、规格、标准，一般用数据表示。

DQAF的应用



DQAF的应用

1. 测量需求定义：即收集业务需求，定义数据质量测量的维度和指标，将特定的测量方法关联到测量类型，并将测量方法和规则落实到工具中。

注意：数据质量的规则定义更多的是由业务人员而不是IT人员负责。业务人员既是数据的消费者，也是数据质量的定义者。

2. 联机数据质量测量：

数据模型：用于评估数据模型与元数据标准是否一致的活动。

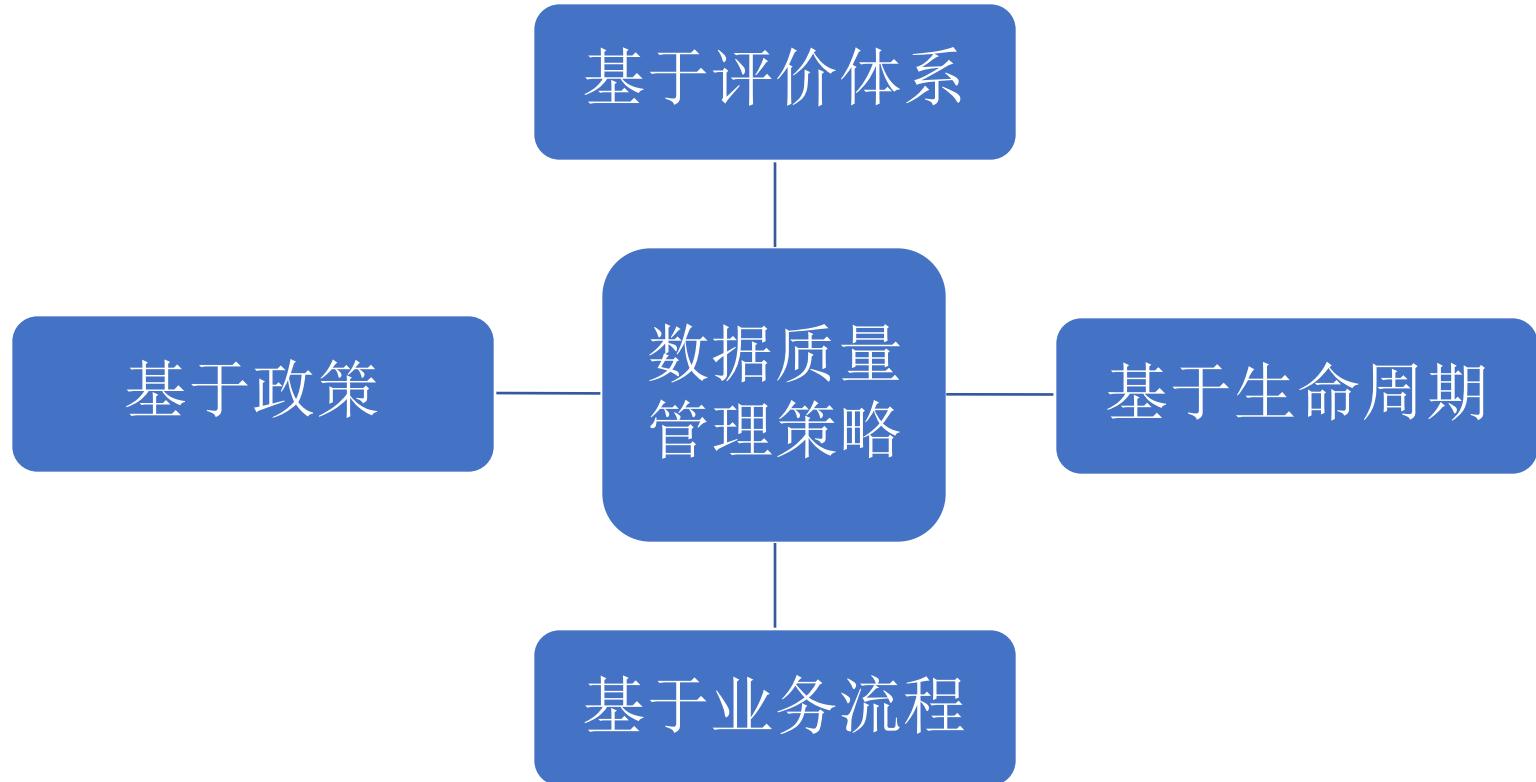
接收数据：用于确保正确接收数据的活动。

数据处理：用于评估数据处理流程质量和数据处理结果的活动。

数据内容：用于评估数据内容的各种活动，包括数据的有效性、特定数据类型的一致性等。

3. 数据质量的初步评估：初步评估的目的是了解企业数据状况，发现数据质量问题的“重灾区”。

4. 启动数据质量改进项目：使数据更加符合业务需求，以支撑数据消费者做出更明智的决策，改进项目可大可小。



根因分析练习

某服装公司正在进行一项客户满意度调查，公司在对调查数据进行分析时发现有一些客户的评分和评论不一致，例如有些客户给出了很高的评分，但是评论中却表达了不满或抱怨。负责该项目的主管认为，有必要对这一问题进行根因分析以进一步获取客户满意度调查的深刻洞察。

请使用鱼骨图法，列举可能影响上述数据质量问题的原因，并逐一分析每个类别下的具体因素。

本章要点

- 掌握数据质量的定义与评估维度
- 熟悉常用的数据质量控制方法
- 熟悉数据质量问题根因分析的工具方法
- 了解数据质量评估框架及数据质量标准