3-2-cn

1

本节我们讨论数据质量

2

关于数据质量，我们要讨论，单数据资源的数据质量问题和多数据资源的数据质量问题

3

数据质量问题的分类有两个维度，

1）按照数据源的数量进行分类， 数据质量问题可分为单数据源和多数据源两种类型

2）按照数据问题出现的阶段分类。数据质量问题可以分为模式层问题和实例层问题。

模式层问题能够通过改善模式设计、模式转换和模式集成加以解决。

而实例层数据质量问题是指在实际的数据内容中存在错误和不一致，这些问题往往在模式层是不可见的需要在数据清洗过程中解决。

4

我们先来看看 单一数据来源 模型级别和 实例级别的数据质量问题

5

对于单个数据源而言，模式层数据质量问题很大程度上依赖于设计模式对数据的完整性约束。

数据库的完整性约束决定了哪些数据值是可以被接受的。

例如某个数据表示日期时，需要约束日期的格式和类型，确保数据库所有日期数据的格式统一。

然而对于文件，web 数据这些没有统一约束的数据源来说，造成数据值错误和不一致的可能性就大大增加。

6

对于单个数据源而言，实例层面的数据质量问题是模式设计层面无法避免的，例如数据输入的错误等

7

1）属性内部：仅限于单个属性值，例如年龄：2000

2）记录内部（属性之间）：同一条记录中不同属性值不一致，例如年龄和生日无法对应

3） 数据源内部（记录之间）：同一数据源不同记录之间的不一致，例如同一个 ID 的姓名不一致

4） 数据源之间：数据源中的某些属性和其他数据源中的相关值得不一致关系，例如同一个 ID 对应的年龄不一致

不同层级范围的数据质量问题，对应的数据清洗方法也不同，明确数据层次范围，是找到合理的数据清洗方法的基础。

8 现在让我们看看多数据源模型级别和实例级别的数据质量问题

9

对于多数据源的情况，需要对不同数据源的数据进行集成。每个数据源往往由特定的应用程序创建，以满足特定的用户的需求，每个数据源的额数据模型设计会存在很大的差异。此外，每个数据源都可能包含脏数据，且不同数据源对同一数据可能存在不同的表示形式、数据重复或者数据冲突，因此在单一数据源情况下存在的数据质量问题在多数据源的情况下依然存在。此外在多数据源情况下，数据清洗面临许多新问题，比如结构冲突，命名冲突，重复记录等。

多数据源模式层面的主要问题是命名冲突和结构冲突：

命名冲突是指对不同的数据对象采用相同的名字命名或者对同一数据对象采用不同的名字命名，

结构冲突存在很多不同的情况，通常指采用不同的方式表示不同数据源中的同一个数据对象，比如同一个对象在不同数据集中有不同属性的粒度，不同的组成结构，不同的数据类型、不同的完整性约束等

10

数据实例层面的冲突是指：具体数据的冲突。在单数据源中存在的数据质量问题，在不同数据源中可能表现为不同形式，比如记录重复，记录冲突等问题。即使不同数据源之间具有相同属性的名字和数据类

型，也可能存在不同的数据值表示，比如对性别的描述，可以表示为男，女，也可以表示为 M，F
或者对数据值得不同解释，比如美元，欧元等不同货币衡量单位
此外，不同数据源提供的信息可能聚合在不同层次，比如，某个数据源中单条记录描述的是某个产品的销售信息，而另一个数据源中的一条记录描述的是一组同类产品的销售信息
11
本节课我们概要的介绍了数据的质量，今天的课就到这里，谢谢大家