【31】

上节视频讲了如何对脏数据进行数据清理。这节视频给大家介绍预处理的中数据集成。

【32】

【32.1】先看下数据集成，什么是数据集成，数据集成是指将互相关联的分布式异构数据源**集成**到一起，使用户能够以透明的方式访问这些数据源。集成是指维护数据源整体上的数据一致性、提高信息共享利用的效率；**透明的方式**是指用户无需关心如何实现对异构数据源数据的访问，只关心以何种方式访问何种数据。

【32.2】为什么要进行数据集成呢，一般信息系统建设通常具有阶段性和分布性的特点，它们的数据分散的存储在各自数据库系统中，这就导致“信息孤岛”现象的存在。

“信息孤岛”是指不同软件间，尤其是不同部门间的数据信息不能共享，造成系统中存在大量冗余数据、垃圾数据，无法保证数据的一致性。

数据集成的目的就是要打破孤岛间的壁垒，最终在获取全面数据的基础上完成数据分析。

【33】

数据集成的方法有多种，包括联邦数据库、中间件集成和数据复制等方法

【33.1】联邦数据库是早期采用的一种模式集成方法。其基本思想是，在构建集成系统时将各数据源的数据视图集成为全局模式，使用户能够按照全局模式透明地访问各数据源的数据。

【33.2】中间件集成方法通过使用统一的全局数据模型来访问异构的数据库或Web资源等。中间件位于异构数据源系统和应用程序之间，向下协调各数据源系统，向上为访问集成数据的应用提供统一数据模式和数据访问的通用接口。

【33.3】数据复制最常用的方法是数据仓库，它将各个数据源的数据复制到同一处，用户则像访问普通数据库一样直接访问数据。

【34】

在实际应用中，数据仓库是数据集成最常用的方法，接下来我们将介绍使用数据仓库的方法来集成数据时需要处理的问题。包括实体识别、冗余和相关分析以及数据冲突和检测

【35】

首先看下实体识别，它所解决的问题是如何匹配多个信息源在现实世界中的实体事物

【35.1】通常有两种实体识别问题需要解决，一个是同义不同名，比如一个数据库中的*customer\_id*和另一个数据库中的*cust\_number*指的是同一实体，

【35.2】还有一个是同名不同义，比如dicount在一个数据库表中是指对某件商品进行打折，而在另一个数据库中是对满额订单进行打折，如果处理不对会企业照成损失。

通常，这两种情况就需要根据数据库或者数据仓库中的元数据来区分这些实体的真正含义。

【36】

冗余是数据集成的另一个重要问题。集成多个数据源时，经常会出现冗余数据，常见的有属性重复、属性相关冗余和元组重复。

【36.1】属性的重复包括同一属性在一个表中出现多次，只是命名不一致而已，比如性别，在一个表中同时有female和性别两个属性字段，这两个属性肯定是重复的。

【36.2】属性相关冗余是指一个属性可以由另外一个属性导出，如：一个顾客数据表中的平均年收入属性，可以根据月收入属性计算出来。另外两个属性之间的相关性有多大，可根据相关系数和协方差来评估一个属性的值如何随另一个变化。

【36.3】还有一种冗余是元组重复，对于给定的唯一数据实体，存在两个或多个相同的元组。

【37】

对于现实世界的同一实体，来自不同数据源的属性值可能不同。这可能是因为表示、尺度、编码或比例的差异。

例如，重量属性可能在一个系统中以公制单位公斤存储，而在另一个系统中以英制单位磅存储。

再比如，某一表示长度的属性在一个数据库中用“厘米”表示，在另一个数据库中却使用“分米”表示。所以检测到这类数据值冲突后，可以根据需要修改某一数据库的属性值，以使不同数据库中同一实体的属性值保持一致。

【38】

本节视频介绍了数据预处理的数据集成。

数据集成是将互相关联的分布式异构数据源**集成**到一起，使用户能够以透明的方式访问这些数据源的一种方法，在集成过程中需要解决实体识别、冗余与相关分析和数据冲突检测这三种问题。