第5章

数据清洗

本章学习目标

- 了解数据清洗的概念
- 了解数据清洗的原理
- 了解数据清洗的常用方法
- 了解数据质量和数据仓库的定义
- 掌握 Kettle 的使用方法

本章先介绍数据清洗的概念，再介绍数据清洗的原理，接着介绍数据清洗的常用方法以及数据质量和数据仓库的定义，最后介绍 Kettle 的使用方法。

视频讲解

5.1 数据清洗概述

1. 数据清洗的定义

数据量的不断剧增是大数据时代的显著特征，大数据必须经过清洗、分析、建模、可视化才能体现其潜在的价值。然而，在众多数据中总是存在着许多“脏数据”，即不完整、不规范、不准确的数据，因此，数据清洗就是指把“脏数据”彻底洗掉，包括检查数据一致性、处理无效值和缺失值等，从而提高数据质量。在实际的工作中，数据清洗通常占开发过程中约 50%～70%的时间。

数据清洗可以有多种表述方式，其定义取决于具体的应用。一般认为，数据清洗的含义是检测和去除数据集中的噪声数据和无关数据，处理遗漏数据，去除空白数据域和知识背景下的白噪声。

2. 数据清洗的应用领域

目前，数据清洗主要应用于 3 个领域：数据仓库、数据挖掘和数据质量管理。

1）数据清洗在数据仓库中的应用

在数据仓库领域，数据清洗一般是应用在几个数据库合并时或多个数据源进行集成时。例如，指代同一个实体的记录，合并后的数据库中就会出现重复的记录。数据清洗过程就是要把这些重复的记录识别出来并消除它们，也就是所谓的合并清洗（Merge/Purge）问题。这个问题的实例字面上可称作记录链接、语义集成、实例识别或对象标识问题。不过值得注意的是，数据清洗在数据仓库中的应用并不是简单地合并清洗记录，它还涉及数据的分解与重组。

2）数据清洗在数据挖掘中的应用

在数据挖掘领域，经常会遇到的情况是挖掘出来的特征数据存在各种异常情况，如数据缺失、数据值异常等。对于这些情况，如果不加以处理，那么会直接影响到最终挖掘模型建立后的使用效果，甚至使最终的模型失效，任务失败。因此，在数据挖掘（早期又称为数据库的知识发现）过程中，数据清洗是第 1 个步骤，即对数据进行预处理的过程。不过值得注意的是，各种不同的知识发现和数据仓库系统都是针对特定的应用领域进行数据清洗的，因此采用的方法和手段都各不相同。

3）数据清洗在数据质量管理中的应用

数据质量管理是贯穿数据生命周期的全过程，在数据生命周期中，数据的获取和使用周期包括一系列活动，如评估、分析、调整、丢弃数据等。因此，数据质量管理覆盖了质量评估、数据去噪、数据监控、数据探查、数据清洗、数据诊断等方面。在此过程中，数据清洗为衡量数据质量的好坏提供了重要的保障。

3. 数据清洗的原理

数据清洗的原理为利用有关技术（如统计方法、数据挖掘方法、模式规则方法等）将脏数据转换为满足数据质量要求的数据。数据清洗按照实现方式与范围，可分为手工清洗和自动清洗。

1）手工清洗

手工清洗是通过人工对录入的数据进行检查。这种方法较为简单，只要投入足够的人力、物力和财力，就能发现所有错误，但效率低下。在大数据量的情况下，手工清洗数据的操作几乎是不可能的。

2）自动清洗

自动清洗是由机器进行相应的数据清洗。这种方法能解决某个特定的问题，但不够灵活，特别是在清洗过程需要反复进行（一般来说，数据清洗一遍就达到要求的很少）时，导致程序复杂，清洗过程变化时，工作量大，而且这种方法也没有充分利用目前数据库提供的强大数据处理能力。

此外，随着数据挖掘技术的不断提升，在自动清洗中常常使用清洗算法与清洗规则帮助完成。清洗算法与清洗规则是根据相关的业务知识，应用相应的技术（如统计学、数据

挖掘的方法)分析出数据源中数据的特点,并且进行相应的数据清洗。常见的清洗方式主要有两种:一种是发掘数据中存在的模式,然后利用这些模式清理数据;另一种是基于数据的清洗模式,即根据预定义的清理规则,查找不匹配的记录,并清洗这些记录。值得注意的是,数据清洗规则已经在工业界被广泛利用,常见的数据清洗规则有编辑规则、修复规则、Sherlock 规则和探测规则等。

例如,编辑规则在关系表和主数据之间建立匹配关系,若关系表中的属性值和与其匹配到的主数据中的属性值不相等,就可以判断关系表中的数据存在错误。

4. 数据清洗的行业发展

在大数据时代,数据正在成为一种生产资料、稀有资产和新兴产业。大数据产业已提升到国家战略的高度,随着创新驱动发展战略的实施,逐步带动产业链上下游,形成万众创新的大数据产业生态环境。数据清洗属于大数据产业链中关键的一环,可以从文本、语音、视频和地理信息等多个领域对数据清洗产业进行细分。

1) 文本清洗领域

文本清洗领域主要基于自然语言处理技术,通过分词、语料标注、字典构建等技术,从结构化、非结构化数据中提取有效信息,提高数据加工的效率。除去国内传统的搜索引擎公司,如百度、搜狗、360 等,该领域的代表公司有拓尔思、中科点击、任子行、海量等。

2) 语音数据加工领域

语音数据加工领域主要是基于语音信号的特征提取,利用隐马尔可夫模型等算法进行模式匹配,对音频进行加工处理。国内该领域的代表公司有科大讯飞、中科信利、云知声、捷通华声等。

3) 视频图像处理领域

视频图像处理领域主要是基于图像获取、边缘识别、图像分割、特征提取等环节,实现人脸识别、车牌标注、医学分析等实际应用。国内该领域的代表公司有 Face++ 、五谷图像、亮风台等。

4) 地理信息处理领域

地理信息处理领域主要是基于栅格图像和矢量图像,对地理信息数据进行加工,实现可视化展现、区域识别、地点标注等应用。国内该领域的代表公司有高德、四维图新、天下图等。

5.2 数据清洗的常用方法

5.2.1 缺失值的处理方法

在数据集中,若某记录的属性值被标记为空白或—,则认为该记录存在缺失值(空值),它也常指不完整的数据。缺失值的产生原因多种多样,主要分为机械原因和人为原因。机械原因导致的数据收集或保存的失败造成数据缺失,如数据存储的失败、存储器损坏、机械故障导致某段时间数据未能收集。人为原因是人的主观失误、历史局限或有意隐瞒造成数据缺失。例如,在市场调查中被访人拒绝透露相关问题的答案,或者回答的问题

是无效的，或者在数据录入时由于操作人员失误漏录了数据等。又如，在一张人员表中，每个实体应当有 5 个属性，分别是姓名、年龄、性别、籍贯和学历，而某些记录只有 4 个属性，因而该记录就存在缺失值。

对于缺失值的清洗方法较多，如将存在遗漏信息属性值的对象（元组、记录）删除；或者将数据过滤出来，按缺失的内容分别写入不同数据库文件并要求客户或厂商重新提交新数据，要求在规定的时间内补全，补全后才继续写入数据仓库；有时也可以用一定的值去填充空值，从而使信息表完备化。填充缺失值通常基于统计学原理，根据初始数据集中其余对象取值的分布情况对一个缺失值进行填充。

处理缺失值按照以下 4 个步骤进行。

（1）确定缺失值范围，对每个字段都计算其缺失值比例，然后按照缺失比例和字段重要性，分别制定策略。

（2）对于一些重要性高，缺失率较低的缺失值数据，可根据经验或业务知识估计，也可通过计算进行填充。

（3）对于指标重要性高，缺失率也高的缺失值数据，需要和取数人员或业务人员了解，是否有其他渠道可以取到相关数据，必要时进行重新采集。若无法取得相关数据，则需要对缺失值进行填充。

（4）对于指标重要性低，缺失率也低的缺失值数据，可只进行简单填充或不作处理；对于指标重要性低，缺失率高的缺失值数据，可备份当前数据，然后直接删除不需要的字段。

值得注意的是，对缺失值进行填充后，填入的值可能不正确，数据可能存在偏置，并不十分可靠。因此，在估计缺失值时，通过考虑该属性的值的整体分布和频率，保持该属性的整体分布状态。

5.2.2 噪声数据的处理方法

噪声数据是指数据中存在着错误或异常（偏离期望值）的数据，这些错误或异常数据对大数据分析造成了干扰。常见的噪声数据包括错误数据、假数据和异常数据。本节主要介绍异常数据的处理。

1. 异常数据的处理

表 5-1 所示为异常数据的常见处理方法。

表 5-1 异常数据的常见处理方法

处 理 方 法	描　　述
删除含有异常数据的记录	直接删除含有异常数据的记录
视为缺失值	将异常数据视为缺失值，利用缺失数据处理的方法进行处理
平均值修正	可用前后两个观测值的平均值修正该异常数据
不处理	直接在有异常数据的数据集上进行数据挖掘建模

值得注意的是，在数据预处理时，异常数据是否删除需视具体情况而定，因为有些异常数据可能蕴含着有用的信息。

2. 异常数据的检测

在数据清洗中，异常数据的检测十分重要。通常对于异常数据的观察和检测，常使用以下几种方法。

1）分箱法

分箱法是通过考查某一数据周围数据的值（即“近邻”）光滑有序数据的值。在分箱法中用“箱的深度”表示不同的箱中有相同个数的数据，用“箱的宽度”表示每个箱值的取值区间。在采用分箱技术时，需要确定的两个主要问题就是如何分箱以及如何对每个箱子中的数据进行平滑处理。常见的分箱法有如下几种。

（1）等深分箱法：每箱具有相同的记录数，每个箱子的记录数称为箱的深度。

（2）等宽分箱法：在整个数据值的区间上进行平均分割，使每个箱子的区间相等，该区间称为箱的宽度。

（3）用户自定义分箱法：根据用户自定义的规则进行分箱处理，当用户明确希望观察某些区间范围内的数据分布时，使用这种方法可以方便地帮助用户达到目的。

2）平滑处理

在分箱之后，需要对每个箱子中的数据进行平滑处理。平滑处理方法主要有按平均值平滑、按边界值平滑和按中值平滑。

（1）按平均值平滑：对同一箱子中的数据求平均值，用平均值代替该箱子中的所有数据。

（2）按边界值平滑：用距离较小的边界值代替箱中每个数据。

（3）按中值平滑：取箱中数据的中值，代替箱中的所有数据。

图 5-1 所示为箱线图。

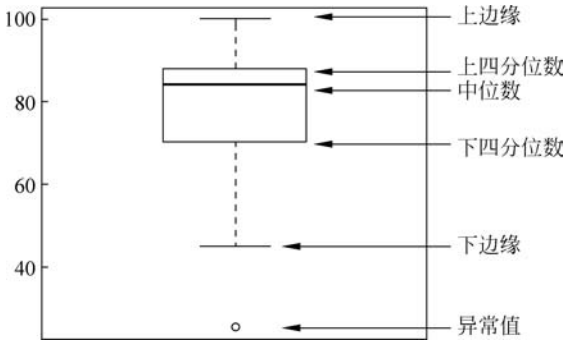

图 5-1　箱线图

3）回归法

回归法是试图发现两个相关的变量之间的变化模式，通过使数据适合一个函数平滑数据，即通过建立数学模型预测下一个数值，包括线性回归和非线性回归。线性回归涉及

找出拟合两个属性(或变量)的“最佳”直线,使一个属性可以用来预测另一个属性。非线性回归是线性回归的扩充,其中涉及的属性多于两个,并且数据拟合到一个多维曲面。图 5-2 所示为回归法。

4）聚类分析

聚类是将数据集合分组为若干个簇,在簇外的值即为孤立点,这些孤立点就是噪声数据,应当对这些孤立点进行删除或替换,如图 5-3 所示。

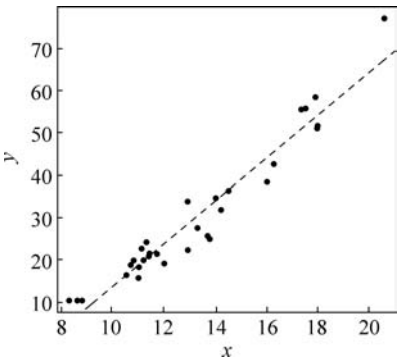

图 5-2　回归法

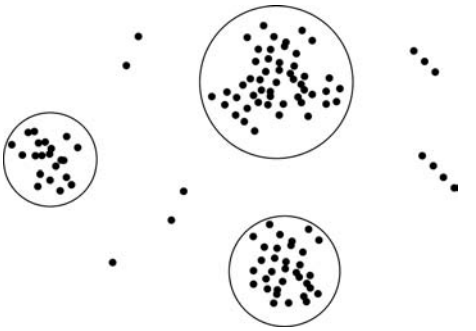

图 5-3　聚类分析

从图 5-3 可以看出,在圆外的点即为噪声数据。

5）估算分析法

对于极个别的异常数据,还可以采取估算分析法,如可以使用均值、中值、众数估算方法等实现。此外,在估算之前,首先应该分析该异常值是自然异常值还是人为的。如果是人为的,则可以用估算值来估算。除此之外,还可以使用统计模型预测异常数据观测值,然后用预测值估算它。

6）3σ 原则

3σ 原则是指如果数据服从正态分布,那么在 3σ 原则下,异常数据为一组测定值中与均值的偏差超过 3 倍标准差的值。因此,如果数据服从正态分布,那么距离均值 3σ 之外的值出现的概率为 $P(|x-\mu|>3\sigma)\leqslant 0.003$(属于极个别的小概率事件),即可认为是异常数据。如果数据不服从正态分布,也可以用远离均值的若干倍标准差来描述。图 5-4 所示为用 3σ 原则检测异常数据。

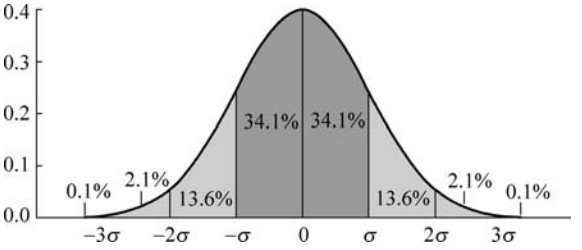

图 5-4　3σ 原则检测异常数据

5.3 数据质量与数据仓库

5.3.1 数据质量的定义

数据无处不在，企业的数据质量与业务绩效之间存在着直接关系。随着企业数据规模的不断扩大，数据数量的不断增加以及数据来源复杂性的不断变化，企业正在努力解决如何处理所有这些问题。

在大数据时代，数据资产及其价值利用能力逐渐成为构成企业核心竞争力的关键要素。然而，大数据应用必须建立在质量可靠的数据之上才有意义，建立在低质量甚至错误数据之上的应用有可能与其初衷南辕北辙，背道而驰。因此，数据质量正是企业应用数据的瓶颈，高质量的数据可以决定数据应用的上限，而低质量的数据则必然拉低数据应用的下限。因此，数据清洗的目的就是真正提高数据质量。

数据质量一般指数据能够真实、完整反映经营管理实际情况的程度，通常可通过以下几方面衡量和评价。

1. 准确性

准确性是指数据在系统中的值与真实值的符合情况。一般而言，数据应符合业务规则和统计口径。常见数据准确性问题如下。

(1) 与实际情况不符：数据来源存在错误，难以通过规范进行判断与约束。

(2) 与业务规范不符：在数据的采集、使用、管理、维护过程中，业务规范缺乏或执行不力，导致数据缺乏准确性。

2. 完整性

完整性是指数据的完备程度。常见数据完整性问题如下。

(1) 系统已设定字段，但在实际业务操作中并未完整采集该字段数据，导致数据缺失或不完整。

(2) 系统未设定字段，或存在数据需求但未在系统中设定对应的取数据字段。

3. 一致性

一致性是指系统内外部数据源之间的数据一致程度，数据是否遵循了统一的规范，数据集合是否保持了统一的格式。常见数据一致性问题如下。

(1) 缺乏系统联动：系统间应该相同的数据却不一致。

(2) 联动出错：在系统中缺乏必要的联动和核对。

4. 可用性

可用性一般用来衡量数据项整合和应用的可用程度。常见数据可用性问题如下。

(1) 缺乏应用功能：没有相关的数据处理、加工规则或数据模型的应用功能以获取目标数据。

（2）缺乏整合共享：数据分散，不易有效整合和共享。

其他衡量标准，如有效性可考虑对数据格式、类型、标准的遵从程度；合理性考虑数据符合逻辑约束的程度。例如，对国内某企业数据质量问题进行的调研显示如下：常见数据质量问题中准确性问题占 33%，完整性问题占 28%，可用性问题占 24%，一致性问题占 8%，这在一定程度上代表了国内企业面临的数据问题。

5.3.2 数据质量的常见问题

常见的数据质量问题可以根据数据源的多少和所属层次分为 4 类。

1. 单数据源定义层

违背字段约束条件（如日期出现 1 月 0 日）、字段属性依赖冲突（如两条记录描述同一个人的某个属性，但数值不一致）、违反唯一性（同一个主键 ID 出现了多次）。

2. 单数据源实例层

单个属性值含有过多信息、拼写错误、空白值、噪声数据、数据重复、过时数据等。

3. 多数据源的定义层

同一个实体的不同称呼（如冰心和谢婉莹，用笔名还是用真名）、同一种属性的不同定义（如字段长度定义不一致、字段类型不一致等）。

4. 多数据源的实例层

数据的维度、粒度不一致（如有的按 GB 记录存储量，有的按 TB 记录存储量；有的按照年度统计，有的按照月份统计）、数据重复、拼写错误。

除此之外，还有在数据处理过程中产生的“二次数据”，其中也会有噪声、重复或错误的情况。数据的调整和清洗也会涉及格式、测量单位和数据标准化与归一化的相关问题，以致对实验结果产生比较大的影响。通常这类问题可以归为不确定性。不确定性有两方面内涵，包括各数据点自身存在的不确定性，以及数据点属性值的不确定性。前者可用概率描述，后者有多重描述方式，如描述属性值的概率密度函数、以方差为代表的统计值等。

5.3.3 数据质量与数据清洗

在不同的学科背景下，数据质量需求的表示方式各不相同。在数据内容层面，要求数据具有正确性、完整性、一致性、可靠性等。在数据效用层面，要求数据具有及时性、相关性、背景性等。学术界的研究多聚焦于数据内容质量的提升，特别是缺失数据补全、数据去重和错误数据纠正，此时指导数据清洗的指标可以具体表示如下。

1. 完整性约束

实体完整性约束规定数据表的主键不能为空和重复；域完整性约束要求表中的列必须满足某种特定的数据类型约束；参照完整性约束规定数据表主键和外键的一致性；此

外，还有用户定义的完整性约束。这些约束大多可以反映属性或属性组之间相互依存和制约的关系，干净的数据需要满足这些约束条件。例如，函数依赖 $X \rightarrow Y$，其中 X 和 Y 均为属性集 $R=\{A_1, A_2, \cdots, A_m\}$ 的子集。给定实例 I 中的两个元组 t_1 和 t_2，如果 $t_1[X]=t_2[X]$，那么必须有 $t_1[Y]=t_2[Y]$。

2. 数据清洗规则

数据清洗规则可以指明数据噪声及其对应的正确值，当数据表中的属性值与规则指出的真值匹配时，该数据满足数据质量需求。有的清洗规则直接把正确值编码在规则中，如修复规则；而有的清洗规则需要通过建立外部数据源与数据库实例之间的对应关系获取正确的数据，如编辑规则、Sherlock 规则和探测规则。常用的外部数据源有主数据和知识库。数据清洗规则含义广泛，数据质量标准和规范大多可以形式化为数据清洗规则。

3. 用户需求

这里的用户需求是指由用户直接指明数据库实例中的错误数据和修复措施。用户标注部分数据后，可以通过监督式学习方法（如支持向量机和随机森林）模拟用户行为。

5.3.4 数据仓库与 ETL


视频讲解

1. 数据仓库

顾名思义，数据仓库（Data Warehouse，DW）是一个很大的数据存储集合，出于企业的分析性报告和决策支持目的而创建，并对多样的业务数据进行筛选与整合。数据仓库是决策支持系统和联机分析应用数据源的结构化数据环境，它研究和解决从数据库中获取信息的问题，并为企业所有级别的决策制定过程，提供所有类型数据支持的战略集合。图 5-5 所示为数据仓库在大数据系统中的地位。

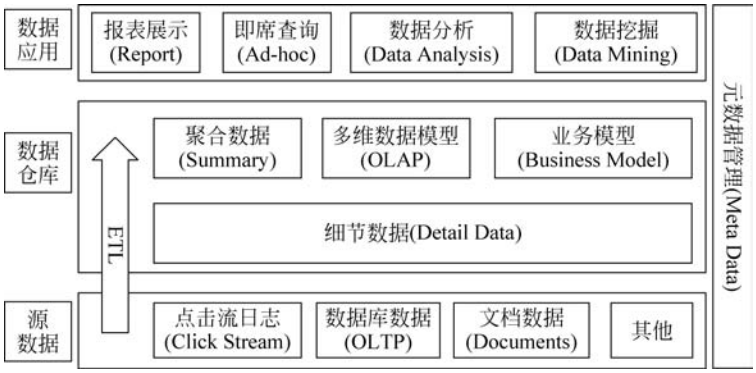

图 5-5 数据仓库在大数据系统中的地位

从图 5-5 可以看出，数据仓库在大数据系统中起着承上启下的作用。一方面，它从各种数据源中提取所需的数据；另一方面，对这些数据集合进行存储、整合与挖掘，从而最终帮助企业的高层管理者或业务分析人员做出商业战略决策或商业报表。

数据仓库的特点如下。

（1）数据仓库是集成的。数据仓库的数据是从原来分散的数据库中的数据抽取而来的。例如，数据仓库的数据有的来自分散的操作型数据，这就需要将所需数据从原来的数据中抽取出来，进行加工与集成，统一与综合之后才能进入数据仓库。

（2）数据仓库是面向主题的。主题是指用户使用数据仓库进行决策时所关心的重点方面，一个主题通常与多个操作型信息系统相关。由于数据仓库都是基于某个明确主题，因此只需要与该主题相关的数据，其他的无关细节数据将被排除掉。

（3）数据仓库是在数据库已经大量存在的情况下，为了进一步挖掘数据资源和决策需要而产生的，数据仓库的方案建设是为前端查询和分析作为基础。

2. ETL

数据仓库中的数据来源十分复杂，既有可能位于不同的平台上，又有可能位于不同的操作系统中，同时数据模型也相差较大。因此，为了获取并向数据仓库中加载这些数据量大同时种类较多的数据，一般要使用专业的工具完成这一操作。

ETL，英文 Extract-Transform-Load 的缩写，用来描述将数据从源端经过抽取、转换、加载至目的端的过程。在数据仓库的语境下，ETL 基本上就是数据采集的代表，包括数据的抽取（Extract）、转换（Transform）和加载（Load）。在转换的过程中，需要针对具体的业务场景对数据进行治理，如对非法数据进行监测与过滤、对数据进行格式转换与规范化、对数据进行替换和保证数据完整性等。

图 5-6 所示为 ETL 在数据仓库中的作用。

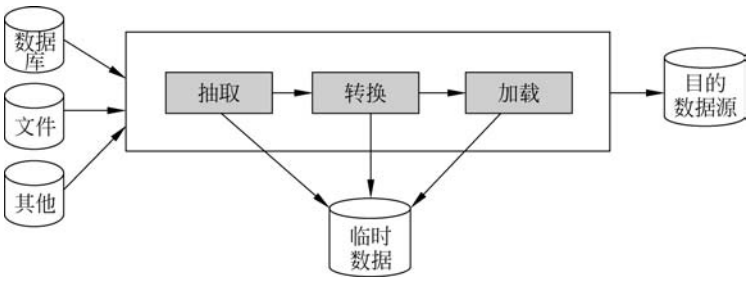

图 5-6　ETL 在数据仓库中的作用

ETL 的流程主要包含数据抽取、数据转换和数据加载，下面分别介绍。

1）数据抽取

数据抽取是指把数据从数据源读出来，一般用于从源文件和源数据库中获取相关的数据，目前在实际应用中，数据源较多采用的是关系数据库。

（1）全量抽取。全量抽取类似于数据迁移或数据复制，它将数据源中的表或视图的数据原封不动地从数据库中抽取出来，并转换成自己的 ETL 工具可以识别的格式，通常来讲全量抽取比较简单。

（2）增量抽取。增量抽取只抽取自上次抽取以来数据库中要抽取的表中新增或修改的数据。如何捕获变化的数据是增量抽取的关键。对捕获方法一般有两点要求：①准确

性，能够将业务系统中的变化数据按一定的频率准确地捕获到；②性能，不能对业务系统造成太大的压力，影响现有业务。在 ETL 使用过程中，增量抽取较全量抽取应用更广。

值得注意的是，数据抽取并不仅仅是根据业务确定公共需求字段，更涉及从不同类型的数据库（Oracle、MySQL、DB2、Vertica 等）、不同类型的文件系统（Linux、Windows、HDFS），以何种方式（数据库抽取、文件传输、流式）、何种频率（分钟、小时、天、周、月）、何种抽取方式（全量抽取、增量抽取）获取数据。所以，具体的实现也包含了大量的工作和技术难点。

2）数据转换

数据转换在 ETL 中常处于中心位置，它把原始数据转换成期望的格式和维度。如果用在数据仓库的场景中，数据转换也包含数据清洗，如根据业务规则对异常数据进行清洗，保证后续分析结果的准确性。值得注意的是，数据转换既可以包含简单的数据格式的转换，也可以包含复杂的数据组合的转换。此外，数据转换还包括许多功能，如常见的记录级功能和字段级功能。

3）数据加载

数据加载是指把处理后的数据加载到目标处，如数据仓库或数据集市中。加载数据到目标处的基本方式是刷新加载和更新加载。其中，刷新加载常用于数据仓库首次被创建时的填充；更新加载则用于目标数据仓库的维护。值得注意的是，加载数据到数据仓库中通常意味着向数据仓库中的表添加新行，或者在数据仓库中清洗被识别为无效的或不正确的数据。此外，在实际的工作中，数据加载需要结合使用的数据库系统（Oracle、MySQL、Spark、Impala 等），确定最优的数据加载方案，以节约 CPU、硬盘 I/O 和网络传输资源。

常见的 ETL 处理方式有以下 3 种。

（1）数据库外部的 ETL 处理。大多数转换工作都在数据库之外，在独立的 ETL 过程中进行。这些独立的 ETL 过程与多种数据源协同工作，并将这些数据源集成。数据库外部 ETL 处理的优点是执行速度比较快；缺点是大多数 ETL 步骤中的可扩展性必须由数据库的外部机制提供，如果外部机制不具备扩展性，那么此 ETL 处理就不能扩展。

（2）数据库段区域中的 ETL 处理。不使用外部引擎而是使用数据库作为唯一的控制点。多种数据源的所有原始数据大部分未作修改就被载入中立的段结构中。如果源系统是关系数据库，段表将是典型的关系型表。如果源系统是非关系型的，数据将被分段置于包含列 VARCHAR2(4000)的表中，以便在数据库内进一步转换。成功地将外部未修改数据载入数据库后，再在数据库内部进行转换，这就是系列方法载入然后转换。数据库段区域中的 ETL 处理方式执行的步骤是抽取、加载、转换，即通常所说的 ELT。在实际数据仓库系统中经常使用这种方式。这种方式的优点是为抽取出的数据首先提供一个缓冲以便进行复杂的转换，降低了 ETL 进程的复杂度。但是这种 ETL 处理方式的缺点有：①在段表中存储中间结果和来自数据库中源系统的原始数据时，转换过程将被中断；②大多数转换可以使用类 SQL 的数据库功能解决，但它们可能不是处理所有的 ETL 问题的最优语言。

（3）数据库中的 ETL 处理。使用数据库作为完整的数据转换引擎，在转换过程中也

不使用段。数据库中的 ETL 处理具有数据库段区域中的 ETL 处理的优点，同时又充分利用了数据库的数据转换引擎功能，但是要求数据库必须完全具有这种转换引擎功能。目前的主流数据库产品 Oracle 9i 等可以提供这种功能。

以上 3 种 ETL 处理方式，数据库外部的 ETL 处理可扩展性差，不适合复杂的数据清洗处理；数据库段区域中的 ETL 处理可以进行复杂的数据清洗；而数据库中的 ETL 处理具有数据库段区域 ETL 处理的优点，又利用了数据库的转换引擎功能。所以，为了进行有效的数据清洗，应该使用数据库中的 ETL 处理方式。

5.3.5 主数据与元数据

1. 主数据

主数据是用来描述企业核心业务实体的数据，如客户、合作伙伴、员工、产品、物料单、账户等。它是具有高业务价值的可以在企业内跨越各个业务部门被重复使用的数据，并且存在于多个异构的应用系统中。

一般来讲，主数据可以包括很多方面，除了常见的客户主数据之外，不同行业的客户还可能拥有其他各种类型的主数据。例如，对于电信行业客户，电信运营商提供的各种服务可以形成其产品主数据。对于航空业客户，航线、航班是其企业主数据的一种。对于某个企业的不同业务部门，其主数据也不同。例如，市场销售部门关心客户信息；产品研发部门关心产品编号、产品分类等产品信息；人事部门关心员工机构、部门层次关系等信息。不过，在企业生产中主数据可以随着企业的经营活动而改变，如客户的增加、组织架构的调整、产品下线等。但是，主数据的变化频率应该是较低的。所以，企业运营过程产生过程数据，如生产过程产生的各种订购记录、消费记录等，一般不会纳入主数据的范围。

值得注意的是，集成、共享、数据质量、数据治理是主数据管理的四大要素。主数据管理要做的就是从企业的多个业务系统中整合最核心的、最需要共享的数据（主数据），集中进行数据的清洗和丰富，并且以服务的方式把统一的、完整的、准确的、具有权威性的主数据分发给全企业范围内需要使用这些数据的操作型应用和分析型应用，包括各个业务系统、业务流程和决策支持系统等。

2. 元数据

元数据又称为中介数据、中继数据，是描述数据的数据，是数据仓库的重要构件，是数据仓库的导航图，在数据源抽取、数据仓库应用开发、业务分析和数据仓库服务等过程中都发挥着重要的作用。

一般来讲，元数据主要用来描述数据属性的信息，如记录数据仓库中模型的定义、各层级间的映射关系、监控数据仓库的数据状态和 ETL 的任务运行状态等。因此，元数据是对数据本身进行描述的数据，或者说，它不是对象本身，它只描述对象的属性，就是一个对数据自身进行描绘的数据。例如，人们想要网购一件衣服，那么衣服就是数据，而挑选的衣服的色彩、尺寸、做工、样式等属性就是它的元数据。

元数据一般分为技术元数据、业务元数据和管理元数据。技术元数据为开发和管理数据仓库的信息技术人员使用，它描述了与数据仓库开发、管理和维护相关的数据，包括

数据源信息、数据转换描述、数据仓库模型、数据清洗与更新规则、数据映射和访问权限等。业务元数据为管理层和业务分析人员服务，从业务角度描述数据，包括商务术语、数据仓库中有什么数据、数据的位置和数据的可用性等，帮助业务人员更好地理解数据仓库中哪些数据是可用的以及如何使用。管理元数据是描述数据系统中管理领域相关概念、关系和规则的数据，主要包括人员角色、岗位职责和管理流程等信息。

元数据的主要作用如下。

(1) 数据描述：对信息对象的内容属性等的描述是元数据最基本的功能。

(2) 数据检索：支持用户发现资源的能力，即利用元数据更好地组织信息对象，建立它们之间的关系，为用户提供多层次多途径的检索体系，从而有利于用户便捷、快速地发现其真正需要的信息资源。

(3) 数据选择：支持用户在不必浏览信息对象本身的情况下能够对信息对象有基础的了解和认识从而决定对检出信息的取舍。

(4) 数据管理：保存信息资源的加工、存档、结构、使用、管理等方面的相关信息权限，如管理权、所有权、使用权、防伪措施、电子水印、电子签名等。

(5) 数据评估：保存资源被使用和被评价的相关信息，通过对这些信息的使用分析，方便资源的建立以及管理者更好地组织资源，并在一定程度上帮助用户确定该信息资源在同类资源中的重要性。

从上面的描述可以看出，元数据最大的好处是使信息的描述和分类可以实现结构化，从而为机器处理创造了可能。

5.4 数据清洗环境与常用工具

5.4.1 数据清洗环境

目前的数据清洗主要是将数据划分为结构化数据和非结构化数据，分别采用传统的数据抽取、转换、加载(ETL)工具和分布式并行处理实现。

具体来讲，结构化数据可以存储在传统的关系型数据库中。关系型数据库在处理事务、及时响应、保证数据的一致性方面有天然的优势。

非结构化数据可以存储在新型的分布式存储中，如 Hadoop 的 HDFS。分布式存储在系统的横向扩展性、降低存储成本、提高文件读取速度方面有着独特的优势。例如，要将传统结构化数据(如关系型数据库中的数据)导入分布式存储中，可以利用 Sqoop 等工具，先将关系型数据库(MySQL、PostgreSQL 等)的表结构导入分布式数据库(Hive)，然后再向分布式数据库的表中导入结构化数据。

图 5-7 所示为在 MySQL 中的数据清洗；图 5-8 所示为在 Hadoop 环境中的数据清洗。

```
select * from mess where content is not null

select * from mess where content <>''

select * from mess where name <>'' and content is not null;
```

图 5-7 在 MySQL 中的数据清洗

```
[root@itcast03 bin]# ./sqoop import --connect jdbc:mysql://169.254.254.1:3306/sqoop --
username root --password root --table Student --target-dir /sqoop/td3 -m 2 --fields-
terminated-by '\t' --columns 'ID,Name,Age'--where 'ID>=3 and ID<=8'
```

图 5-8　在 Hadoop 环境中的数据清洗

5.4.2　数据清洗常用工具

1. Excel

Excel 是人们熟悉的电子表格软件，自 1993 年被微软公司作为 Office 组件发布后，已被广泛使用了近 30 年。虽然在之后的岁月中，各大软件公司也开发出了各类数据处理软件，但至今还有很多数据只能以 Excel 表格的形式获取到。Excel 的主要功能是处理各种数据，它就像一本智能的簿子，不仅可以对记录在案的数据进行排序、筛选，还可以整列整行地进行自动计算；通过转换，它的图表功能可以使数据更加简洁明了地呈现出来。

但 Excel 的局限在于一次所能处理的数据量有限，若要针对不同的数据集实现数据清洗，非常麻烦，这就需要用到 VBA 和 Excel 内置编程语言。因此，Excel 一般用于处理小批量的数据清洗。

Excel 数据清洗和转换的基本步骤如下。

（1）从外部数据源导入数据。

（2）在单独的工作簿中创建原始数据的副本。

（3）确保以行和列的表格形式显示数据，并且每列中的数据都相似；所有的列和行都可见；范围内没有空白行。

（4）先执行不需要对列进行操作的任务，如拼写检查或使用“查找和替换”功能。

（5）再执行需要对列进行操作的任务。对列进行操作的一般步骤如下。

① 在需要清理的原始列（A）旁边插入新列（B）。

② 在新列（B）的顶部添加将要转换数据的公式。

③ 在新列（B）中向下填充公式。在 Excel 表中，将使用向下填充的值自动创建计算列。

④ 选择并复制新列（B），然后将其作为值粘贴到新列（B）中。

⑤ 删除原始列（A），这样，新列 B 将转换为 A。

图 5-9 所示为在 Excel 中使用函数进行数据的字符截取；图 5-10 所示为在 Excel 中使用函数进行数据转换。

	A	B	C	D
1	字符串	公式	返回值	
2	我的名字	=LEFT(A2)	我	
3	我的名字	=LEFT(A2,)		
4	我的名字	=LEFT(A3, 2)	我的	
5	我的名字	=LEFT(A4, 5)	我的名字	
6				

图 5-9　在 Excel 中使用函数进行数据的字符截取

	A	B	C	D	E	F
1	转换前	转换后				
2	99887766554	(0998)8776-6554				
3	88776655443	(0887)7665-5443				
4	77665544332	(0776)6554-4332				
5	66554433221	(0665)5443-3221				
6	55443322110	(0554)4332-2110				
7						

图 5-10　在 Excel 中使用函数进行数据转换

2. Kettle

Kettle 是一款国外开源的 ETL 工具，纯 Java 语言编写，可以在 Windows、Linux、UNIX 上运行，数据抽取高效稳定。Kettle 中有两种脚本文件：转换(Transformation)和工作(Job)，转换完成针对数据的基础转换，工作则完成整个工作流的控制。

使用 Kettle 可以完成数据仓库中的数据清洗与数据转换工作，常见的操作有数据类型的转换、数据值的修改与映射、数据排序、空值的填充、重复数据的清洗、超出范围的数据清洗、日志的写入、数据值的过滤和随机值的运算等。

图 5-11 所示为使用 Kettle 完成数据排序；图 5-12 所示为使用 Kettle 去除重复记录。

图 5-11　Kettle 数据排序

图 5-12　Kettle 去除重复记录

Kettle 生成的文件扩展名为 ktr，可用 Windows 中的记事本打开，并查看其中保存的数据内容。

3. DataCleaner

DataCleaner 是一个简单、易于使用的数据质量的应用工具，旨在分析、比较、验证和监控数据。DataCleaner 包括一个独立的图形用户界面，它能够将凌乱的半结构化数据集转换为所有可视化数据，并可以读取的干净可读的数据集。此外，DataCleaner 还提供数据仓库和数据管理服务。

DataCleaner 的特点有：可以访问多种不同类型的数据存储，如 Oracle、MySQL、MS CSV 文件等；还可以作为引擎清理、转换和统一来自多个数据存储的数据，并将其统一到主数据的单一视图中。

值得注意的是，DataCleaner 提供了一种和 Kettle 类似的运行模式。用户在图形界面通过数据源选择、组件拖动、参数配置、结果输出等一系列操作过程，最终将程序运行的结果保存为一个任务文件(*.xml)。

图 5-13 所示为 DataCleaner 的启动界面；图 5-14 所示为 DataCleaner 的运行界面；图 5-15 所示为 DataCleaner 的数据清洗界面。

图 5-13　DataCleaner 的启动界面

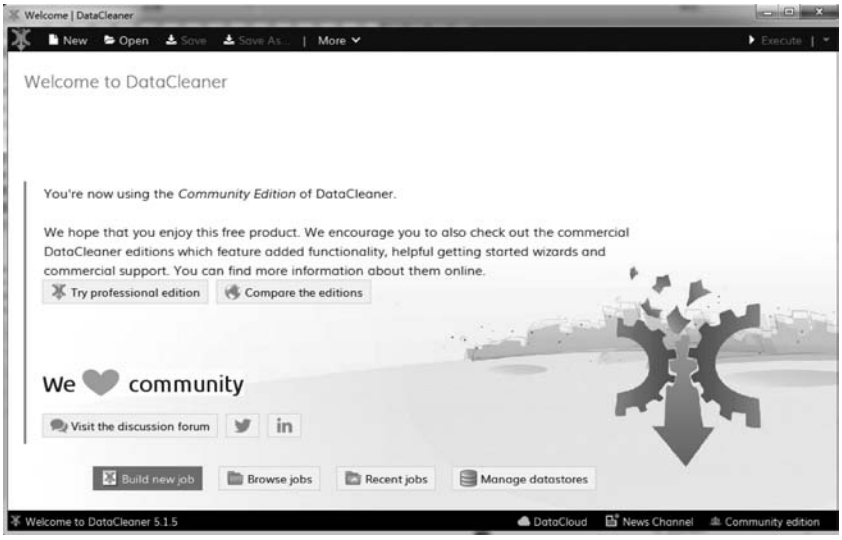

图 5-14　DataCleaner 的运行界面

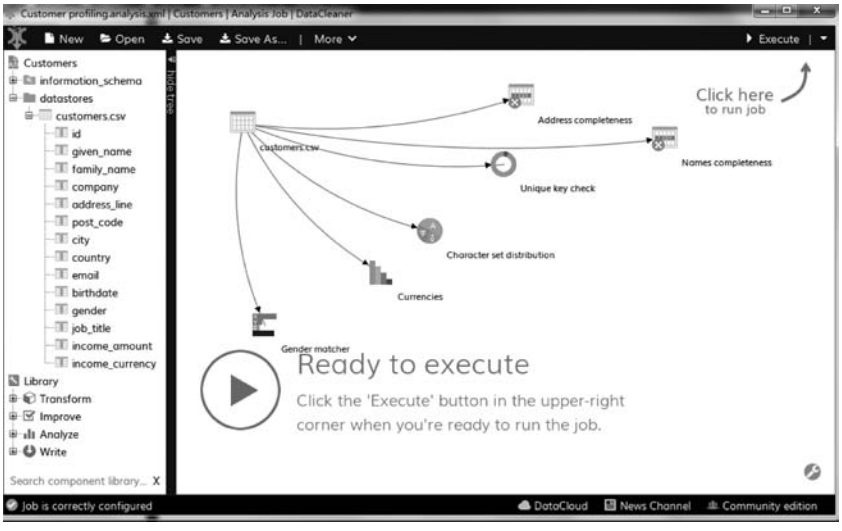

图 5-15　DataCleaner 的数据清洗界面

4. OpenRefine

OpenRefine 又称为 GoogleRefine，是一个新的具有数据画像、清洗、转换等功能的工具，它可以观察和操纵数据。OpenRefine 类似于传统 Excel 表格处理软件，但是工作方式更像是数据库，以列和字段的方式工作，而不是以单元格的方式工作。因此，OpenRefine 不仅适合对新的行数据进行编码，而且功能还极为强大。

OpenRefine 的特点有：在数据导入时，可以根据数据类型将数据转换为对应的数值和日期型等；相似单元格聚类，可以根据单元格字符串的相似性聚类，并且支持关键词碰撞和近邻匹配算法等。

图 5-16 所示为 OpenRefine 读取数据表的界面。

图 5-16 OpenRefine 读取数据表的界面

5.5 本章小结

（1）在众多数据中总是存在着许多“脏数据”，即不完整、不规范、不准确的数据。数据清洗就是指把“脏数据”彻底洗掉，包括检查数据一致性、处理无效值和缺失值等，从而提高数据质量。

（2）在数据清洗中，原始数据源是数据清洗的基础，数据分析是数据清洗的前提，而定义数据清洗转换规则是关键。

（3）数据质量正是企业应用数据的瓶颈，高质量的数据可以决定数据应用的上限，而低质量的数据则必然拉低数据应用的下限。因此，数据清洗的目的就是真正提高数据质量。

（4）目前的数据清洗主要是将数据划分为结构化数据和非结构化数据，分别采用传统的数据抽取、转换、加载（ETL）工具和分布式并行处理来实现。

5.6 实训

视频讲解

1. 实训目的

通过本章实训了解数据清洗的特点，能够进行简单的与数据清洗有关的操作。

2. 实训内容

1）下载并安装数据清洗的常用工具

（1）以 Kettle 为例，下载安装 Kettle 前需要首先从官网上下载 jdk。

（2）配置 Path 变量。下载完之后进行安装，安装完毕后要进行环境配置。在“我的电脑”→“高级”→“环境变量”中找到 Path 变量，并把 Java 的 bin 路径添加进去（多个路径用分号隔开）。注意，要找到自己安装的对应路径，如 D:\Program Files\Java\jdk1.8.0_181\bin。

（3）配置 Classpath 变量，在环境变量中新建一个 Classpath 变量，填入 Java 文件夹中 lib 文件夹下 dt.jar 和 tools.jar 的路径，如 D:\Program Files\Java\jdk1.8.0_181\lib\dt.jar 和 D:\Program Files\Java\jdk1.8.0_181\lib\tools.jar。

（4）配置完后运行 cmd 命令，输入命令 java，如果配置成功，会出现如图 5-17 所示的界面。

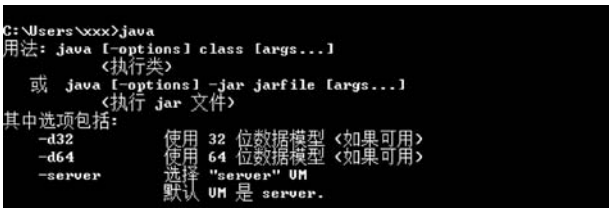

图 5-17 jdk 配置成功

（5）从官网（http://kettle.pentaho.org）下载 Kettle 软件，由于 Kettle 是绿色软件，因此下载后可以解压到任意目录。本书下载 8.2 版本，本书中的 Kettle 程序也可使用 7.1 版本运行。

（6）运行 Kettle。解压后，双击目录下面的 spoon.bat 批处理程序，即可启动 Kettle，如图 5-18 所示。

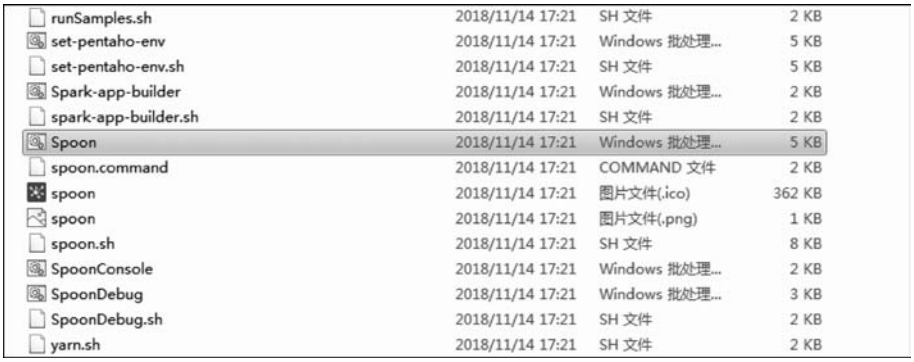

图 5-18 启动 Kettle

（7）Kettle 8.2 运行界面如图 5-19 所示。

2）使用 Kettle 实现数据排序

（1）启动 Kettle 后，执行“文件”→“新建”→“转换”菜单命令，新建一个转换，在“输入”列表中选择“Excel 输入”步骤，在“转换”列表中选择“排序记录”步骤，分别拖动到右

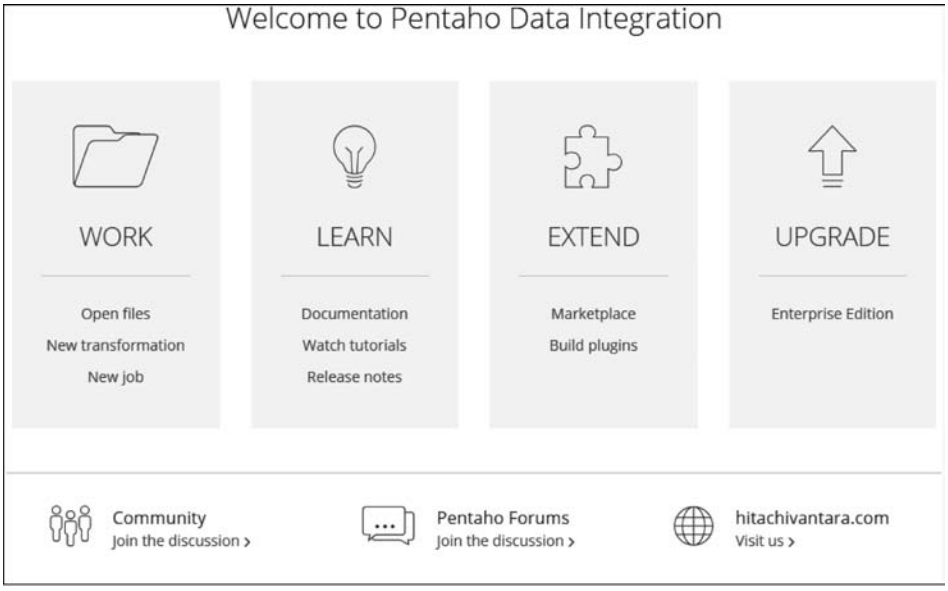

图 5-19 Kettle 8.2 运行界面

侧工作区中，并建立彼此之间的节点连接关系，如图 5-20 所示。

（2）双击“Excel 输入”图标，导入 Excel 数据表，如图 5-21 所示，数据表内容如图 5-22 所示；切换至“字段”选项卡，单击“获取来自头部数据的字段”按钮，如图 5-23 所示。

图 5-20 Kettle 数据排序工作流程

图 5-21 导入 Excel 数据表

	A	B	C
1	姓名	成绩	
2	禁明	68	
3	张敏	57	
4	刘健	47	
5	王天一	78	
6	徐红	67	
7	洪智	84	
8	周明	61	
9	李凡	70	
10	刘甜	63	
11	张晓晓	89	
12	宗树生	73	
13	王天一	78	
14			

图 5-22　Excel 数据表内容

图 5-23　获取字段

（3）双击“排序记录”图标，对“成绩”字段按照降序排序，如图 5-24 所示。

图 5-24　对“成绩”字段排序

（4）保存该文件，单击“运行这个转换”按钮，在“执行结果”区域中的 Preview data 选项卡预览生成的数据，如图 5-25 所示。


执行结果

执行历史 | 日志 | 步骤度量 | 性能图 | Metrics | Preview data

${TransPreview.FirstRows.Label} ${TransPreview.LastRows.Label} ${TransPreview.Off.Label}

#	姓名	成绩
1	张晓晓	89.0
2	洪智	84.0
3	王天一	78.0
4	王天一	78.0
5	宗树生	73.0
6	李凡	70.0
7	蔡明	68.0
8	徐红	67.0
9	刘甜	63.0
10	周明	61.0
11	张敏	57.0
12	刘健	47.0

图 5-25 查看排序结果

习题 5

（1）请阐述数据清洗的应用领域。

（2）请阐述处理缺失值的步骤。

（3）请阐述数据仓库的定义。

（4）请阐述噪声数据的处理方法。