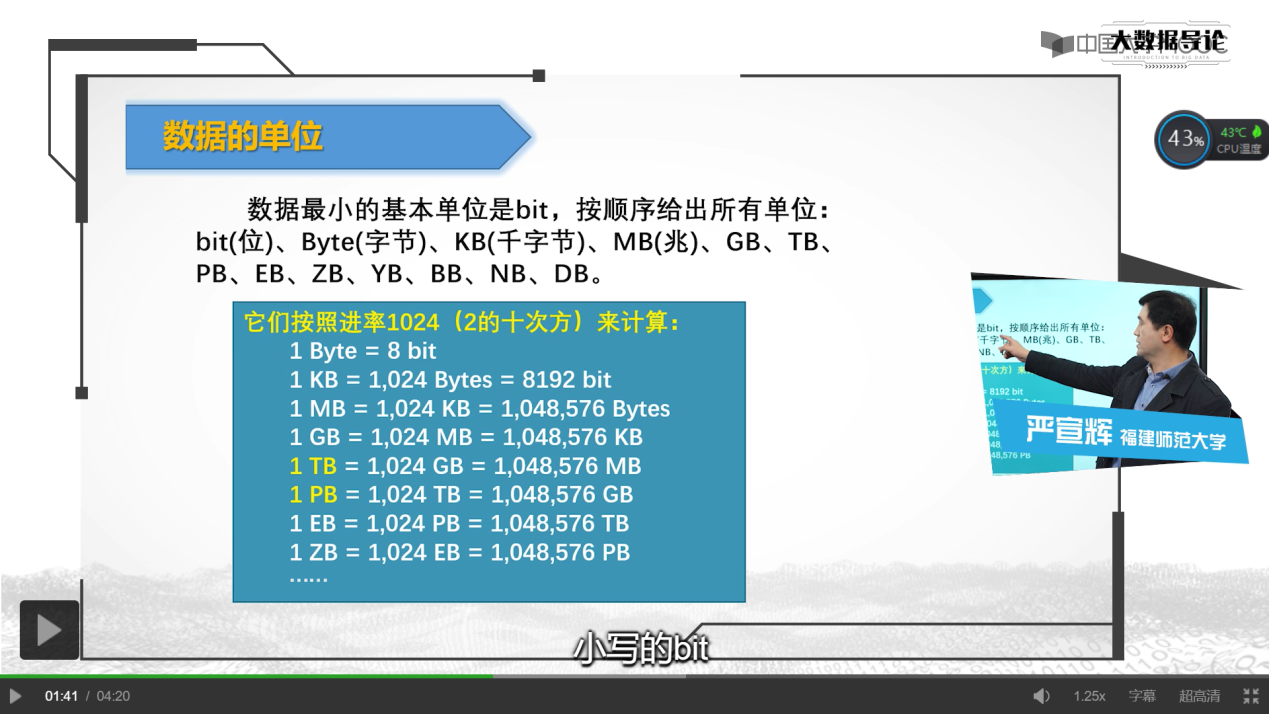
**第三章 数据采集和预处理**

**3.1大数据的来源**

1)大数据的来源非常多，如信息管理系统、网络信息系统、物联网系统、科学实验系统等

2)数据的单位：



3) 对现实世界的测量

按产生数据的主体划分 人类的记录

计算机产生的数据

类型

结构化信息：关系型数据库

按数据格式划分 半结构化信息：XML

非结构化信息：文档、音频

结构化数据简单来说就是可以用关系型数据库表示和存储，表现为二维形式的数据；半结构化数据适于描述两个或多个数据库中的数据；非结构化数据是数据结构不规则，不完整，不方便用数据库二维逻辑表来表现的数据

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据 | 举例 | 特点 | 优点 | 缺点 | 使用场景 |
| 结构化数据 | Excel，mysql | 二维形式的数据 | 方便查询和修改 | 不易扩展 | 字段特征较固定 |
| 半结构化数据 | XML，JSON，html | 包含元数据信息 | 扩展性很好 |  | 易于归档 |
| 非结构化数据 | word,txt,img,video | 不规范的数据 | 格式多样 |  | 大量存储与共享 |

**3.2大数据采集**

1)数据采集：是指从真实世界中获得原始数据的过程

2)数据采集的方法：

无线传感器网络

用于采集物理信息的传感器

有线传感器网络

用于采集数字设备运行状态的日志文件

用于采集互联网信息的网络爬虫: 是一种按照一定的规则，自动的抓取[万维网](https://baike.so.com/doc/5341011-5576454.html" \t "_blank)信息的程序或者脚本。另外一些不常使用的名字还有蚂蚁，自动索引，模拟程序或者蠕虫。

外包：把一个任务外包给其他人来完成

众包：不固定的一群人用开源的思想来共同完成一群任务，如维基百科

3) Chukwa

Flume

Scrible

Kafka

4)Apache Kafka起源于Linkedle ,2011年成为开源Apache的项目，在2012年成为Apache的一流项目。

目前Apache已发展成功能完善的基于分布式的消息发布-订阅系统。

Topics:消息的分类名。

Producers:能够发布消息到Topics的进程。

Consumers:可以订阅一个或多个话题，并从Broker拉数据，从而消费这些已发布的消息。

Broker:组成Kafka集群的单个节点

Apache Flume:

Flume是一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统。

flume的核心是把数据从数据源收集起来，经过传送通道将收集到的数据送到指定的目的地，event是flume传输数据的基本单元。

**3.3大数据预处理**

1、影响数据质量的因素

1. 大数据处理环节对数据质量的影响

采集阶段——数据来源、数据录入

整合阶段——数据集成

分析阶段——数据建模方式、数据间约束条件

可视化阶段——数据表达方式

1. 数据质量的五个维度：数据一致性、精确性、完整性、时效性和实体同一性。
2. 评估数据质量的标准：衡量数据某一方面的性质，如准确性、完整性、一致性、及时性、可信性、可解释性、重复性、关联性等。它们反映了数据质量的特征和用户的需求。

2、数据预处理

1. 数据预处理的含义

数据预处理Data Preprocessing

预处理是指在数据进行分析挖掘之前，对原始数据进行变换、清洗与集成等一系列操作。数据预处理工作能有效提高数据质量。

1. 数据预处理的流程

数据归纳

数据变换

数据集成

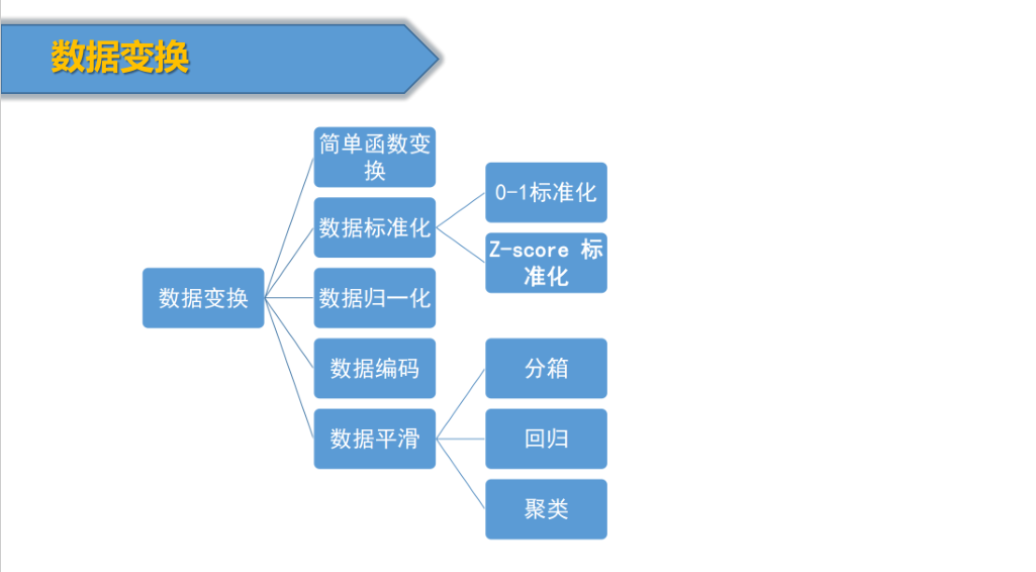
数据清洗

初始数据

的获取

1. 数据变换：将数据转换成适当的形式以便更好地理解和处理。

数据变换的几种类型



* 数据平滑：去除数据中的噪声波动，使数据平滑。常用方法：分箱、回归和聚集。
* 简单函数变换：如平方、开方、对数变换和差分运算等；
* 数据归一化：是把数据变换为[0,1]之间的小数，这样可以把有量纲表示为无量纲表示，成为纯量。如最小最大归一化：f(x) = (x - min)/(max - min)，零-均值规范化，小数定标规范化。
* 数据标准化：是将数据把数据转换成标准正态分布，如Z-score标准化：f(x) = (x - 均值)/标准差。
* 数据编码：用统一的编码标准对信息记录进行编码。

1. 数据归约：归约后的数据集比原数据集小得多，但仍近似地保持原数据的完整性。

数据归约的策略包括数据立方体聚集、属性子集选择、数据压缩、数值归约、离散化和概念分层。

5、数据清洗——提升数据质量

数据清洗通过填写空缺值，消除噪音数据，识别或删除孤立点，并解决不一致性来“清洗”数据，从而改善数据质量，提高数据挖掘的效率和精度。

* 处理空缺值：均值填充法

如果缺失值是数值型的，就该数据变量在其他所有对象的取值的平均值来填充该确实的变量值；

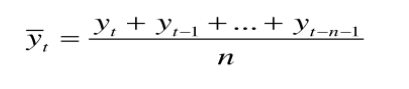
如果缺失值是非数值型的，则使用众数来补齐该缺失的变量值。

其他填充方法：回归填充法、热卡填充填补法、就近补齐法、极大似然估计、期望最大化法、K最近距离邻法等。

* 消除噪音数据

噪声数据是一个测量变量中的随机错误或偏差，包括错误的值或偏离预期的孤立点值。

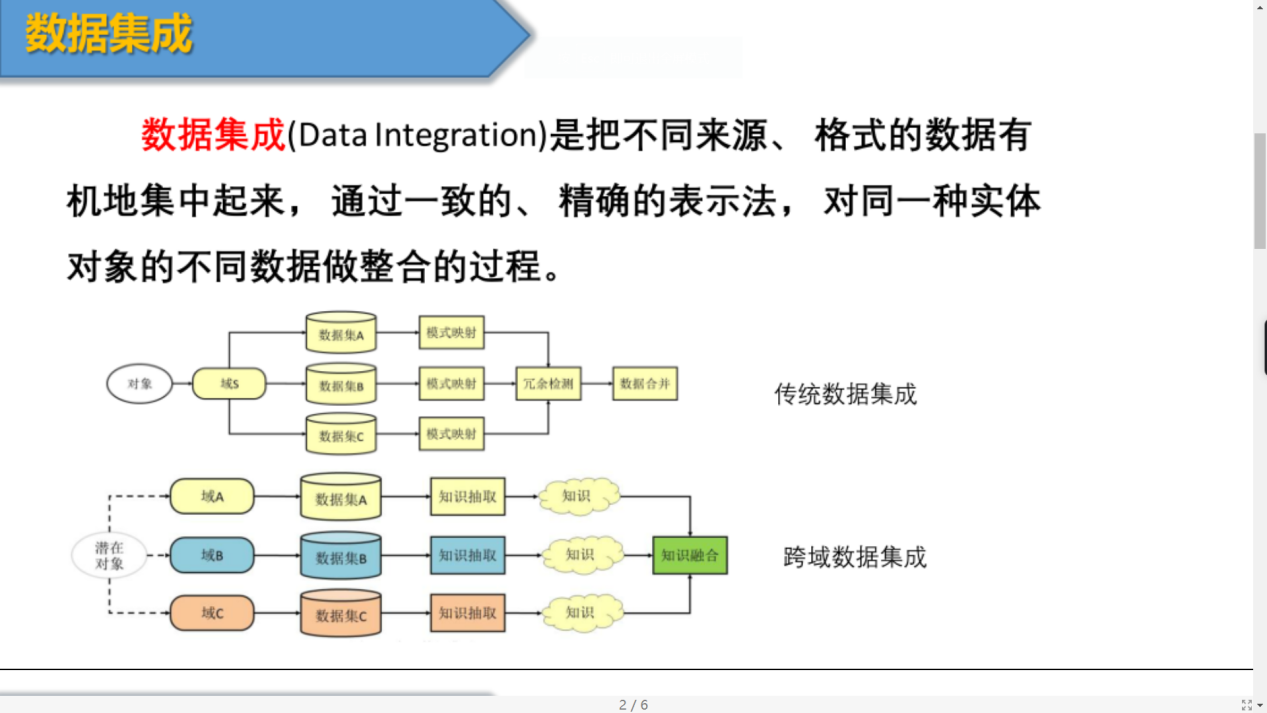
* 分箱法（Binning）：局部的平滑
* 平均数平滑：箱中每一个值被箱的平均数替换
* 中位数平滑：箱中每一个值被箱的中位数数替换
* 箱边界平滑：箱中每一个值被离它最近的箱边界值替换
* 回归法：移动平均

移动平均是一种用以滤除噪声的简单数据处理方法。当接收到输入数据后，就将本次输入数据与其前若干次的输入数据进行平均。

* 聚类法：用聚类算法把相似的数据聚成一个“类簇”，落在各个类簇之外的数据（一般为异常数据或孤立点）可以被看为“噪声”。

**3.4大数据集成**

数据集成(Data Integration)是把不同来源、格式的数据有机地集中起来，通过一致的、精确的表示法，对同一种实体对象的不同数据做整合的过程。



数据集成解决的主要问题：

* 实体识别问题

通常可以根据数据库或数据仓库中的元数据来区分模式集成中的错误。

* 冗余问题

集成多个数据源时，冗余数据经常会出现，常见的是冗余属性。处理检测属性间的冗余外，还应当检测元组之间的重复。

* 数据冲突的检测与处理

*概念提示*

众数：是一组数据中出现次数最多的数值，有时众数在一组数中有好几个。

中位数*:*是按顺序排列的一组数据中居于中间位置的数,对于有限的[数集](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=8312689&ss_c=ssc.citiao.link)，可以通过把所有观察值高低排序后找出正中间的一个作为中位数。如果观察值有偶数个，通常取最中间的两个数值的[平均数](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=795982&ss_c=ssc.citiao.link)作为中位数。中位数只能有一个。

标准差：是[离均差](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=70565987&ss_c=ssc.citiao.link)平方的[算术平均数](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=796022&ss_c=ssc.citiao.link)的[平方根](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=101572146&ss_c=ssc.citiao.link)，用σ表示。标准差是方差的[算术平方根](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=7562263&ss_c=ssc.citiao.link)。所有数减去其平均值的[平方和](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=305544&ss_c=ssc.citiao.link)，所得结果除以该组数之个数（或个数减一，即变异数），再把所得值开根号，所得之数就是这组数据的标准差。

孤立点：在数学上是指坐标满足[曲线方程](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=7611557&ss_c=ssc.citiao.link)，但并不落在曲线上的点，是一种奇点。也可以指是在数据集合中与大多数数据的特征或不一致的数据。

3.1、3.2 潘念然 19760117

3.3、3.4 徐雅婕 19760119

语音汇总 赵亚然 19760120