数据清洗

过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。 不符合要求的数据主要是有不完整的数据、错误的数据、重复的数据三大类。 是一个反复的过程，不可能在几天内完成，只有不断的发现问题，解决问题。

- 待清理数据的主要类型

  残缺数据

  错误数据 EXP 数值数据输成全角数字字符 字符串数据后面有回车 日期格式 日期越界 对于类似于全角字符、数据前后有不可见字符的问题，只能通过写SQL语句的方式找出来，然后要求客户在业务系统修正之后抽取。 日期格式不正确的或者是日期越界的这一类错误会导致ETL运行失败，这一类错误需要去业务系统数据库用SQL的方式挑出来，交给业务主管部门要求限期修正，修正之后再抽取。

  重复数据 将重复数据记录的所有字段导出来，让客户确认并整理。

- 数据清洗的内容

  一致性检查 无效值和缺失值的处理

- 数据清洗的实现方式

  1 - 专门编写应用程序，解决特定的问题 2 - 解决某类特定应用域的问题，根据概率统计学原理查找数值异常 对姓名 地址 邮政编码等清理 3 -

- 数据清洗的方法

> 数据清理是将数据库精简以除去重复记录，并使剩余部分转换成标准可接收格式的过程。 1 解决不完整数据（即值缺失）的方法\ 某些缺失值可以从本数据源或其它数据源推导出来，这就可以用平均值、最大值、最小值或更为复杂的概率估计代替缺失的值，从而达到清理的目的。 2 错误值的检测及解决方法 用统计分析的方法识别可能的错误值或异常值，如偏差分析、识别不遵守分布或回归方程的值，也可以用简单规则库（ 常识性规则、业务特定规则等）检查数据值，或使用不同属性间的约束、外部的数据来检测和清理数据。 3 重复记录的检测及消除方法 数据库中属性值相同的记录被认为是重复记录，通过判断记录间的属性值是否相等来检测记录是否相等，相等的记录合并为一条记录（即合并/清除)。合并/清除是消重的基本方法。 4 不一致性（ 数据源内部及数据源之间）的检测及解决方法 数据迁移工具允许指定简单的转换规则，如：将字符串gender替换成sex。sex公司的PrismWarehouse是一个流行的工具，就属于这类。 数据清洗工具使用领域特有的知识（ 如，邮政地址）对数据作清洗。它们通常采用语法分析和模糊匹配技术完成对多数据源数据的清理。某些工具可以指明源的“ 相对清洁程度”。工具Integrity和Trillum属于这一类。 数据审计工具可以通过扫描数据发现规律和联系。因此，这类工具可以看作是数据挖掘工具的变形。

- 数据清洗的步骤

- 定义和确定错误类型

> 1.1 数据分析 检测数据中的错误或不一致情况。 手动检查+分析程序获得关于数据属性的元数据 1.2 定义清洗转换规则 数据分析得到的结果来定义清洗转换规则与工作流 执行大量的数据转换和清洗步骤 尽可能为模式相关的数据清洗和转换指定一种查询和匹配语言，从而使转换代码的自动生成变成可能。

- 搜寻并识别错误实例

> 2.1 自动检测属性错误 利用高的方法自动检测数据集中的属性错误 主要的方法有：基于统计的方法 聚类方法 关联规则的方法 2.2 检测重复记录的算法 针对两个数据集或者一个合并后的数据集 检测出标识同一个现实实体的重复记录，即匹配过程。 检测重复记录的算法主要有：基本的字段匹配算法 递归的字段匹配算法 Smith-Waterman算法 Cosin相似度算法e

- 纠正所发现的错误

> 直接在源数据进行清洗时，要备份源数据 根据脏数据存在形式的不同，执行一系列的转换步骤来解决模式层和实例层的数据质量问题。 3.1 从自由格式的属性字段中抽取值 3.2 确认和改正 （处理输入和拼写错误），尽可能使其自动化。基于字典查询的拼写检查对于发现拼写错误是很有用的 3.3 标准化 把属性值转换成一个一致和统一的格式。

- 干净数据回流

> 干净的数据替换数据源中原来的脏数据。 提高原系统的数据质量，还可避免将来再次抽取数据后进行重复的清洗工作。

- 数据清洗的评价标准
- 数据可信性

  精确性（与对应客观实体的特征相一致） 完整性 一致性（同一实体同一属性在不同系统中是一致的） 有效性（满足用户定义或在一定的阈值范围内） 唯一性（没有重复记录）
- 数据的可用性

  时间性 是当前数据还是历史数据 稳定性 数据是否稳定 是否在有效期内
- 数据清洗的代价

  考虑物质和时间开销，是否会超过组织的承受能力 系统性工程，需要多方配合以及大量人员参与，需要多种资源的支持。成本效益估算
- 常见的数据清洗算法
- 空缺值的清洗

  忽略元祖，人工填写空缺值，使用全局变量填充空缺值，使用属性的平均值等或更为复杂的概率统计函数值
- 噪声数据的清洗

  分箱(binning) 等深或等宽的箱中，用属性值的平均值或中值代替 计算机和人工检查结合，计算机检查可疑数据，然后进行人工判断 使用简单规则库检测和修正错误 使用不同属性间的约束检测和修正错误 使用外部数据源检测和修正错误
- 不一致数据清洗

  知识工程工具可以用来检测违反限制的数据。例如，知道属性间的函数依赖，可以查找违反函数依赖的值。此外，数据集成也可能产生数据不一致。
- 重复数据清洗

  基本思想：排序和合并 先将数据库中的记录排序，然后通过比较邻近记录是否相似来检测记录是否重复。 消除重复记录的算法主要有：优先队列算法，近邻排序算法(Sorted—Neighborhood Method)，多趟近邻排序(Multi—Pass Sorted—Neighborhood)。
- 异常点检测算法
- 偏差检测

  例如聚类，最近邻等。
- 基于统计的异常点检测算法

例如极差，四分位数间距，均差，标准差等，这种方法适合于挖掘单变量的数值型数据。全距(Range)，又称极差，是用来表示统计资料中的变异量数(measures of variation) ，其最大值与最小值之间的差距；四分位距通常是用来构建箱形图，以及对概率分布的简要图表概述。

- 基于距离的异常点检测算法

  主要通过距离方法来检测异常点，将数据集中与大多数点之间距离大于某个阈值的点视为异常点，主要使用的距离度量方法有绝对距离（曼哈顿距离）、欧氏距离和马氏距离等方法。

- 基于密度的异常点检测算法

  考察当前点周围密度，可以发现局部异常点，例如LOF算法