# 第5章

# Kettle数据清洗

**本章学习目标**

- 了解 Kettle 数据清洗的概念
- 掌握 Kettle 数据清洗的方法
- 能使用 Kettle 进行数据清洗

本章先介绍 Kettle 数据清洗的概念和基本步骤，再介绍 Kettle 数据清洗的常用方法，最后介绍 Kettle 数据清洗的实例。

视频讲解

## 5.1 Kettle 数据清洗概述

### 1. Kettle 数据清洗介绍

使用 Kettle 可以完成数据仓库中的数据清洗与数据转换工作，常见的操作有数据值的修改与映射、数据排序、重复数据的清洗、超出范围的数据清洗、日志的写入、JavaScript 代码数据清洗、正则表达式数据清洗、数据值的过滤以及随机值的运算等。

在 Kettle 中进行数据清洗的时候基本上没有单一的清洗步骤，很多时候数据清洗工作需要结合多个步骤来完成。例如，数据清洗可以从数据抽取时就开始执行，并在多个步骤中通过设定清洗内容完成操作。

### 2. Kettle 数据清洗基本步骤

如图 5-1 所示，Kettle 在“转换”列表中提供了多种数据清洗步骤，在这里对其中使用频率较高的步骤进行简单的介绍。

（1）计算器：对一个或多个字段进行计算，该步骤提供了很多预定义的函数用于处理输入字段，并且随着版本的更新还在不断增多。

（2）字符串替换：可以理解为对字符串进行查找和替换，该步骤看上去很简单，不过它支持正则表达式，从而可以实现很多复杂的功能。

（3）字符串操作：提供了很多常规的字符串操作，如大小写转换、字符填充、移除空白字符等。

（4）值映射：使用一个标准的值替换字段中的其他值。

（5）字段选择：对字段进行选择、删除、重命名等操作，还可以更改字段的数据类型以及长度等元数据。

（6）去除重复记录：主要通过指定字段去除重复记录，但是一般需要结合其他步骤共同实现其功能。

（7）增加常量：用于增加 $x$ 个字段，每个字段的值都是常量（这里的 $x$ 是一个大于或等于 0 的自然数）。

（8）排序记录：对指定的字段进行排序（升序或降序）。

（9）拆分字段：把字段按照分隔符拆成两个或多个字段。

（10）列拆分为多行：把包含指定的分隔符的字段拆分为多行。


图 5-1　Kettle 中的清洗步骤

（11）将字段值设置为常量：用常量值代替原值，此时无论原值有多少行，该行的所有值都会被一个值所替换。

（12）增加序列：一个序列是在某个起始值和增量的基础之上，经常改变的整数值。可以使用数据库定义好的序列，也可以使用 Kettle 决定的序列。

（13）剪切字符串：对字符串进行剪切。

此外，“应用”列表中的“写日志”步骤、“流程”列表中的“过滤记录”和“识别流的最后一行”步骤、“脚本”列表中的“正则表达式”“公式”和“Java 代码”步骤、“查询”列表中的“检查文件是否存在”和“模糊匹配”步骤、“检验”列表中的“数据检验”步骤、“统计”列表中的“分析查询”“分组”和“单变量统计”步骤也可以进行数据清洗，如图 5-2～图 5-7 所示。


图 5-2 “应用”列表中的数据清洗步骤


图 5-3 “流程”列表中的数据清洗步骤


图 5-4 “脚本”列表中的数据清洗步骤

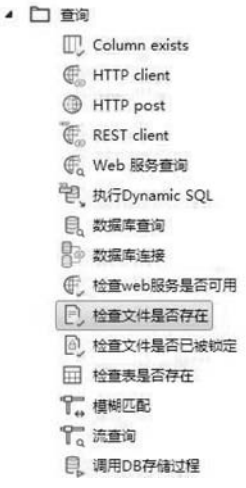



图 5-5 “查询”列表中的数据清洗步骤


图 5-6 “检验”列表中的数据清洗步骤


图 5-7 “统计”列表中的数据清洗步骤

## 5.2 Kettle 数据清洗实现

### 5.2.1 清洗简单数据

#### 1. Kettle 数据清洗介绍


视频讲解

【例 5-1】 值映射。

(1) 启动 Kettle 后,新建转换,在“输入”列表中选择“生成记录”步骤,在“转换”列表中选择“值映射”步骤,拖动到右侧工作区中,其中“值映射”步骤拖动两次,并建立彼此之间的节点连接关系,如图 5-8 所示。

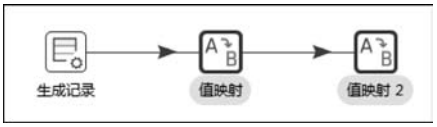



图 5-8 Kettle 值映射工作流程

(2) 双击“生成记录”图标,在“限制”输入框中输入值为 1000,并设置字段内容,生成需要的内容,如图 5-9 所示。


图 5-9 设置生成记录

(3) 单击“预览”按钮,可查看生成记录,如图 5-10 所示。

(4) 双击“值映射”图标,在“使用的字段名”下拉列表中选择 name,并设置字段值内容,从而生成需要的内容,如图 5-11 所示。

(5) 双击“值映射 2”图标,在“使用的字段名”下拉列表中选择 value,并设置字段值内容,从而生成需要的内容,如图 5-12 所示。

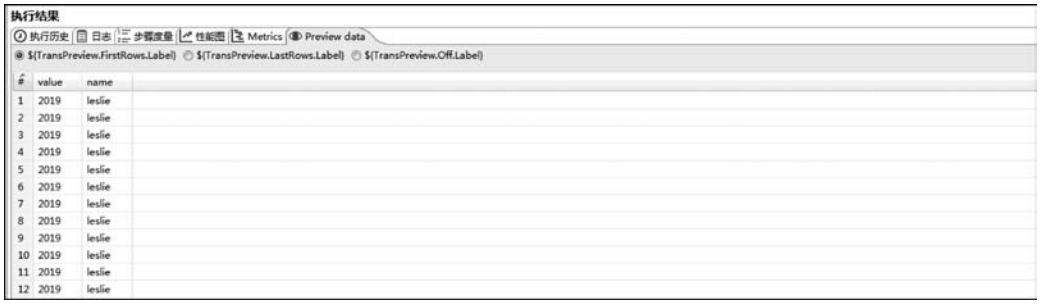

图 5-10　预览生成记录


图 5-11　设置值映射


图 5-12　设置值映射 2

(6) 保存该转换并运行，在执行结果区域的 Metrics 选项卡中可查看数据清洗的过程，在 Preview data 选项卡中查看已经清洗好的数据，如图 5-13 和图 5-14 所示。


图 5-13　查看数据清洗过程




| # | value | name |
|---|---|---|
| 1 | 2019 | leslie |
| 2 | 2019 | leslie |
| 3 | 2019 | leslie |
| 4 | 2019 | leslie |
| 5 | 2019 | leslie |
| 6 | 2019 | leslie |
| 7 | 2019 | leslie |
| 8 | 2019 | leslie |
| 9 | 2019 | leslie |
| 10 | 2019 | leslie |
| 11 | 2019 | leslie |
| 12 | 2019 | leslie |

图 5-14　查看清洗好的数据


视频讲解

**【例 5-2】** 使用 Kettle 实现数据排序。

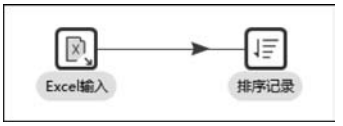



图 5-15　Kettle 数据排序工作流程

(1) 启动 Kettle 后，新建转换，在“输入”列表中选择“Excel 输入”步骤，在“转换”中选择“排序记录”步骤，拖动到右侧工作区中，并建立彼此之间的节点连接关系，如图 5-15 所示。

(2) 双击“Excel 输入”图标，导入 Excel 数据表，如图 5-16 所示，数据表内容如图 5-17 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 5-18 所示。


图 5-16　导入 Excel 数据表

(3) 双击“排序记录”图标，对字段中的“成绩”按照降序排序，如图 5-19 所示。

(4) 保存该文件，运行转换，在执行结果区域的 Preview data 选项卡预览生成的数据，如图 5-20 所示。

| | A | B | C | D |
|---|---|---|---|---|
| 1 | 姓名 | 成绩 | | |
| 2 | 蔡明 | 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 | | | | |
| 15 | | | | |

图 5-17　Excel 数据表内容


图 5-18　获取字段


图 5-19　对字段排序

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 5-20 查看排序结果


视频讲解

**【例 5-3】** 使用 Kettle 去除重复数据。

(1) 在例 5-2 的基础上完成此操作,在“转换”列表中选择“去除重复记录”步骤,拖动到右侧工作区中,并建立彼此之间的节点连接关系,如图 5-21 所示。

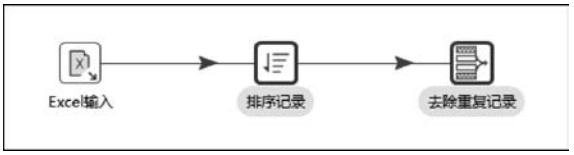



图 5-21 去除重复数据工作流程

(2) 双击“去除重复记录”图标,将“字段名称”设置为“姓名”,如图 5-22 所示。


图 5-22 设置去除重复记录

（3）保存该文件，运行转换，在执行结果区域中的 Preview data 选项卡预览生成的数据，如图 5-23 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 | 宗树生 | 73.0 |
| 5 | 李凡 | 70.0 |
| 6 | 蔡明 | 68.0 |
| 7 | 徐红 | 67.0 |
| 8 | 刘甜 | 63.0 |
| 9 | 周明 | 61.0 |
| 10 | 张敏 | 57.0 |
| 11 | 刘健 | 47.0 |

图 5-23　查看去重结果

需要注意的是，在 Kettle 中使用“去除重复记录”步骤时，首先要对数据进行排序。

### 5.2.2　清洗复杂数据

**【例 5-4】** 使用 Kettle 清洗超出范围的数据。

（1）启动 Kettle 后，新建转换，在“输入”列表中选择“自定义常量数据”步骤，在“检验”列表中选择“数据检验”步骤，在“输出”列表中选择“文本文件输出”步骤，分别拖动到右侧工作区中，其中“文本文件输出”步骤拖动两次，并建立彼此之间的节点连接关系，如图 5-24 所示。值得注意的是，在“数据检验”与“文本文件输出 2”节点的连接中，需要右击“数据检验”，选择“定义错误处理”，并在“步骤错误处理设置”对话框中设置错误处理步骤，如图 5-25 和图 5-26 所示。

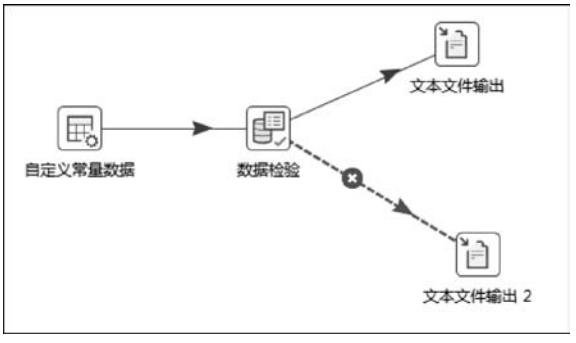



图 5-24　清洗超出范围数据工作流程


图 5-25　定义错误处理


图 5-26　设置错误处理步骤

(2) 双击“自定义常量数据”图标,在“元数据”和“数据”选项卡中设置内容,如图 5-27 和图 5-28 所示。

(3) 双击“数据检验”图标,单击“增加检验”按钮,将新增的检验命名为 sco。选中


图 5-27 设置元数据(1)


图 5-28 设置数据(1)

sco,将“检验描述”设置为 sco,选择“要检验的字段名”为 score,并将 score 类型中的取值范围设置为 0～100,如图 5-29 所示。

(4) 分别双击“文本文件输出”和“文本文件输出 2”图标,设置清洗后将要保存的文件路径和文件名,保留数据为 file6,抛弃数据为 file7。保存该文件,运行转换,并在最终保存的文本文件中查看清洗结果,如图 5-30 所示。

在 Kettle 中使用数据检验可以检查数据是否遵循了预定义的业务规则,从而找出不符合业务规则的数据。在第 1 次编辑“数据检验”步骤时,这个步骤是空的,必须要选择“增加检验”创建一个新检验。除此之外,还可以进行信用卡检验、电子邮箱地址检验以及 XML 检验。例如,“检验邮件地址”步骤,它不仅可以用于验证字符串是否满足电子邮箱的规则,还可以检验电子邮箱的有效性,有兴趣的读者可以自行尝试。

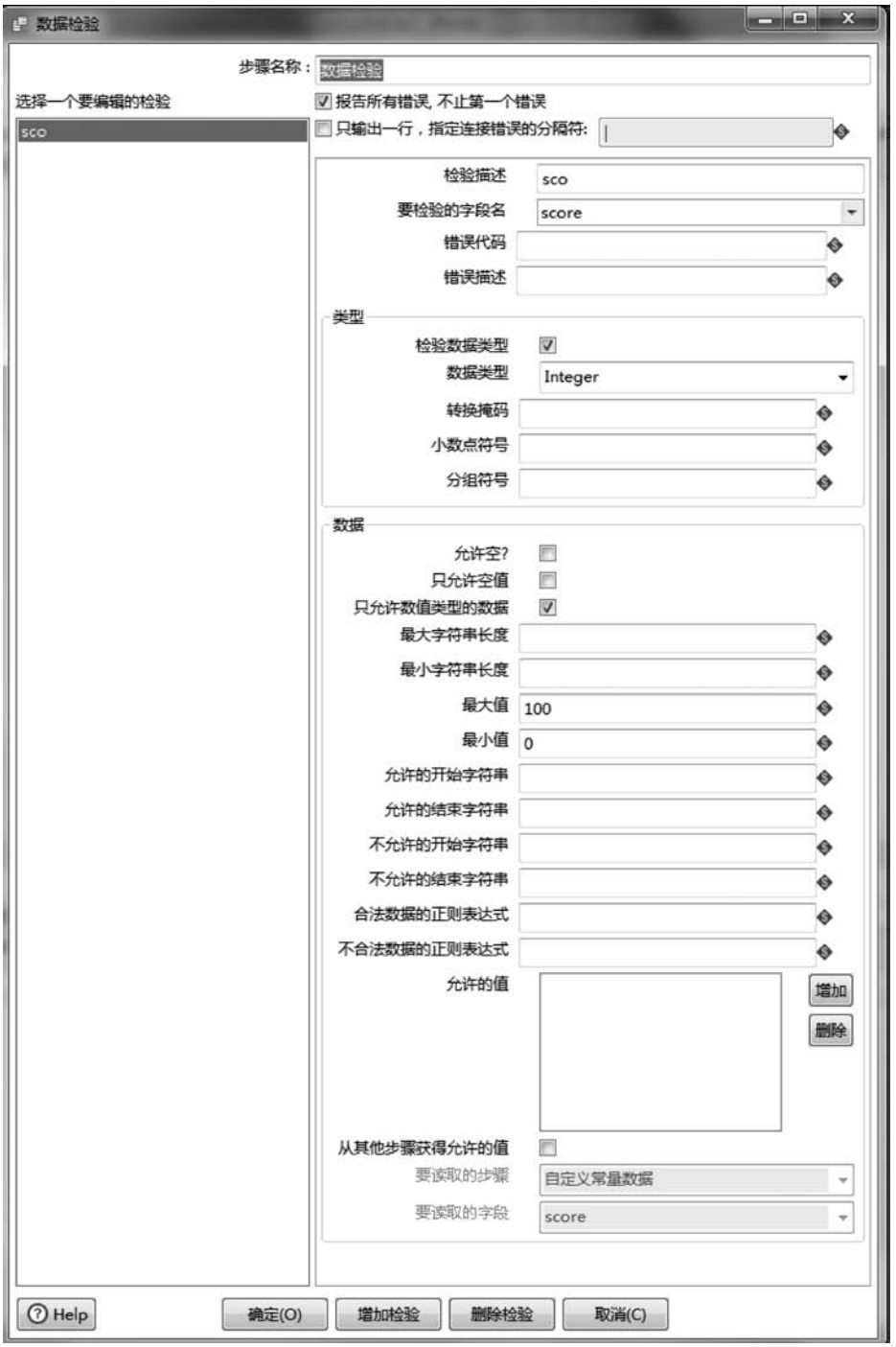

数据检验
步骤名称：
数据检验
选择一个要编辑的检验
sco
报告所有错误，不止第一个错误
只输出一行，指定连接错误的分隔符：
检验描述
sco
要检验的字段名
score
错误代码
错误描述
类型
检验数据类型
数据类型
Integer
转换掩码
小数点符号
分组符号
数据
允许空?
只允许空值
只允许数值类型的数据
最大字符串长度
最小字符串长度
最大值
100
最小值
0
允许的开始字符串
允许的结束字符串
不允许的开始字符串
不允许的结束字符串
合法数据的正则表达式
不合法数据的正则表达式
允许的值
增加
删除
从其他步骤获得允许的值
要读取的步骤
自定义常量数据
要读取的字段
score
Help
确定(O)
增加检验
删除检验
取消(C)

图 5-29 设置清洗规则


图 5-30 查看清洗结果

视频讲解

**【例 5-5】** 使用 Kettle 过滤记录。

(1) 启动 Kettle 后,新建转换,在“输入”列表中选择“自定义常量数据”步骤,在“流程”列表中选择“过滤记录”步骤,在“流程”列表中选择“空操作”步骤,分别拖动到右侧工作区中,其中“空操作”步骤拖动两次,分别重命名为“可以开车”和“不可以开车”,并建立彼此之间的节点连接关系,如图 5-31 所示。值得注意的是,该流程中需要双击“过滤记录”图标,在“发送 true 数据给步骤”下拉列表中选择“可以开车”,在“发送 false 数据给步骤”下拉列表中选择“不可以开车”,如图 5-32 所示。

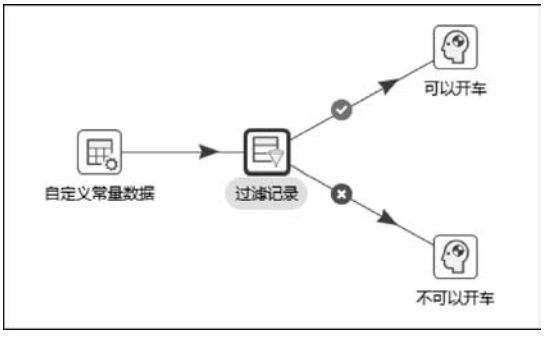



图 5-31 过滤记录工作流程


图 5-32 设置过滤记录

(2) 双击“自定义常量数据”图标,在“元数据”和“数据”选项卡中分别设置内容,如图 5-33 和图 5-34 所示。


图 5-33 设置元数据(2)


图 5-34 设置数据(2)

(3) 双击“过滤记录”图标,将条件设置为 age <= 60,如图 5-35 所示。

(4) 保存该文件,运行转换,单击“可以开车”图标,在执行结果区域中的 Preview data 选项卡查看运行结果;单击“不可以开车”图标,在执行结果区域中的 Preview data 选项卡查看运行结果,如图 5-36 和图 5-37 所示。本例通过过滤记录将年龄大于 60 岁的人设置为“不可以开车”。

在 Kettle 中,“过滤记录”步骤通过条件和比较运算符过滤记录,通常有以下两个步骤。


图 5-35　设置过滤记录条件

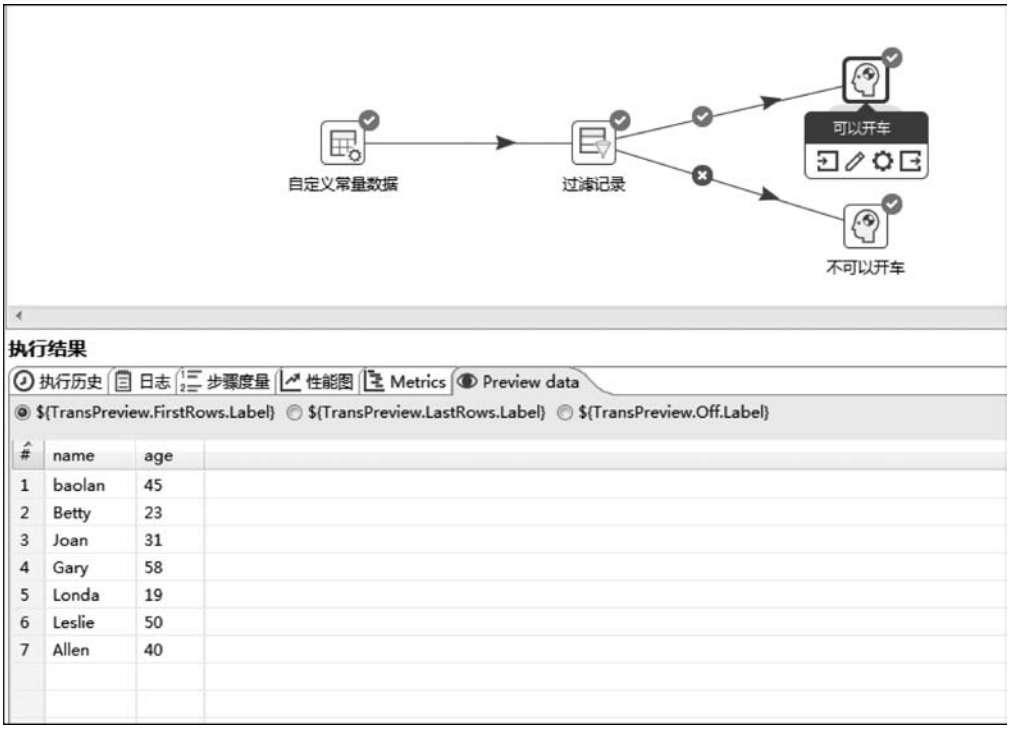



图 5-36　查看结果(可以开车)

(1) 发送 true 数据给步骤：指定条件返回 true 的数据将发送到此步骤。

(2) 发送 false 数据给步骤：指定条件返回 false 的数据将发送到此步骤。

注意：true 和 false 步骤必须指定。

视频讲解

**【例 5-6】** 使用 Kettle 生成多个随机数并相加。

(1) 启动 Kettle 后，新建转换，在“输入”列表中选择“生成随机数”步骤，在“转换”列表中选择“计算器”步骤，分别拖动到右侧工作区中，并建立彼此之间的节点连接关系，如图 5-38 所示。

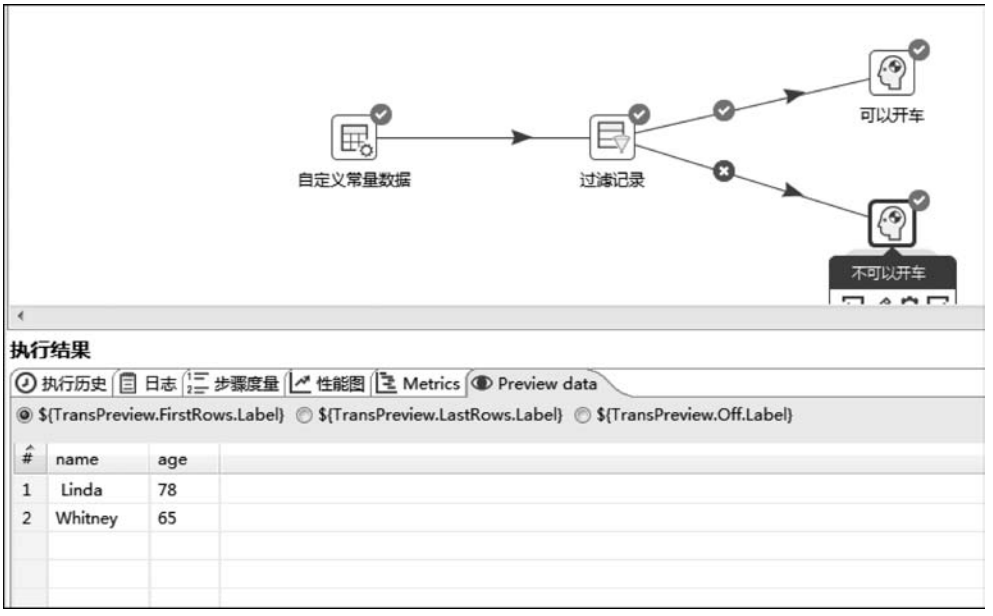



图 5-37　查看结果(不可以开车)

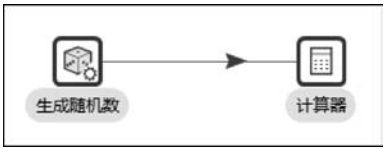



图 5-38　生成多个随机数并相加工作流程

(2) 双击“生成随机数”图标,在弹出的对话框中设置字段,如图 5-39 所示。

(3) 单击“确定”按钮,右击“生成随机数”图标,在弹出的快捷菜单中选择“开始改变复制的数量”,在文本框中输入 10,如图 5-40 所示。


图 5-39　设置随机数


图 5-40　设置复制的数量

（4）双击“计算器”图标，在弹出的对话框中设置字段内容，如图 5-41 所示。

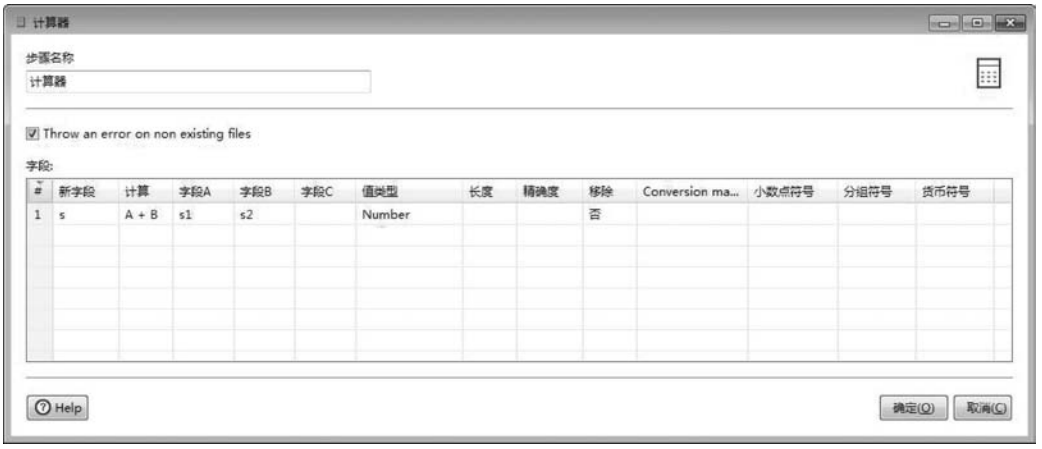
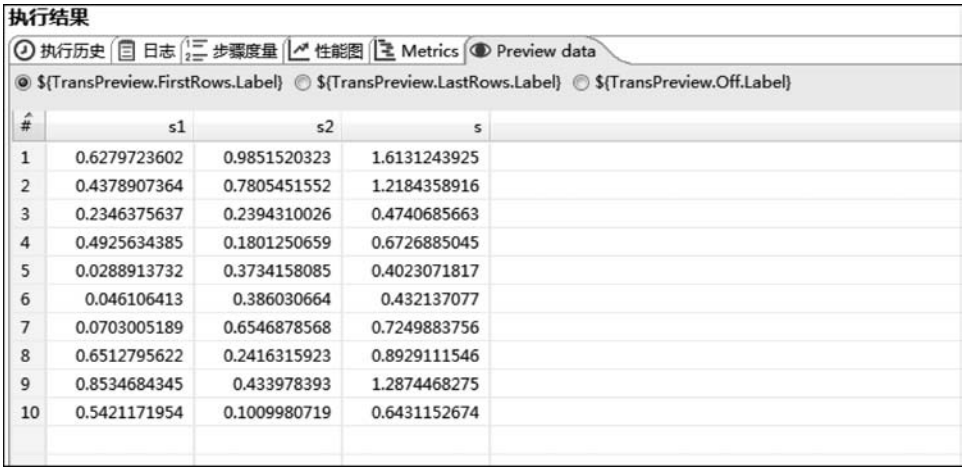



图 5-41　设置计算器字段内容

（5）保存该文件，运行转换，结果如图 5-42 所示。




| # | s1 | s2 | s |
|---|---|---|---|
| 1 | 0.6279723602 | 0.9851520323 | 1.6131243925 |
| 2 | 0.4378907364 | 0.7805451552 | 1.2184358916 |
| 3 | 0.2346375637 | 0.2394310026 | 0.4740685663 |
| 4 | 0.4925634385 | 0.1801250659 | 0.6726885045 |
| 5 | 0.0288913732 | 0.3734158085 | 0.4023071817 |
| 6 | 0.046106413 | 0.386030664 | 0.432137077 |
| 7 | 0.0703005189 | 0.6546878568 | 0.7249883756 |
| 8 | 0.6512795622 | 0.2416315923 | 0.8929111546 |
| 9 | 0.8534684345 | 0.433978393 | 1.2874468275 |
| 10 | 0.5421171954 | 0.1009980719 | 0.6431152674 |

图 5-42　保存并运行转换

视频讲解

**【例 5-7】** 使用 Kettle 对数据进行统计分析。

（1）启动 Kettle 后，新建转换，在“输入”列表中选择“Excel 输入”步骤，在“统计”列表中选择“单变量统计”步骤，分别拖动到右侧工作区中，并建立彼此之间的节点连接关系，如图 5-43 所示。

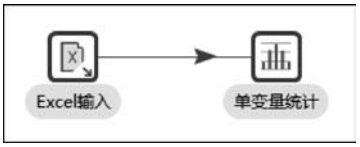



图 5-43　统计分析工作流程

（2）双击“Excel 输入”图标，在“文件”选项卡中添加外部 XLS 文件，数据表内容见例 5-2，如图 5-44 所示；并在“字段”选项卡中进行相关设置，如图 5-45 所示。


图 5-44　添加外部数据表


图 5-45　设置字段

（3）单击“单变量统计”图标，设置统计内容，如图 5-46 所示。

（4）保存该文件，运行转换，结果如图 5-47 所示。

在本例中，成绩(N)表示数据个数；成绩(mean)表示平均成绩；成绩(stdDev)表示成绩的标准差；成绩(min)表示成绩的最小值；成绩(max)表示成绩的最大值；成绩(median)表示成绩的中位数。

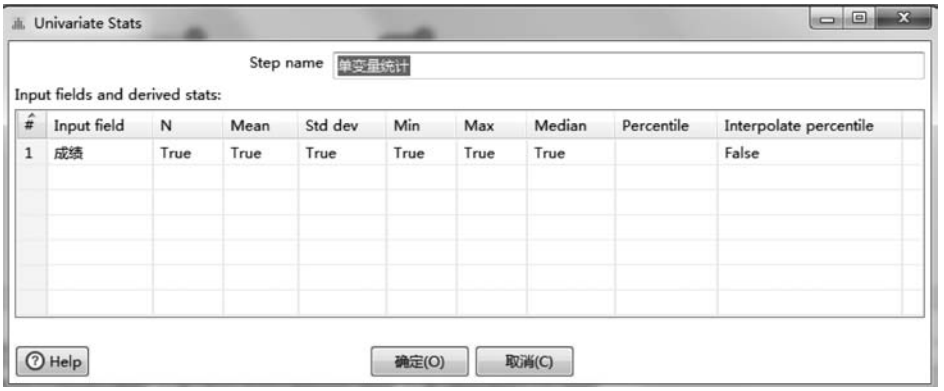



图 5-46　设置统计内容


图 5-47　统计分析结果


视频讲解

**【例 5-8】** 使用 Kettle 书写日志。

（1）启动 Kettle 后，新建转换，在“输入”列表中选择“生成记录”步骤，在“应用”列表中选择“写日志”步骤，分别拖动到右侧工作区中，并建立彼此之间的节点连接关系，如图 5-48 所示。

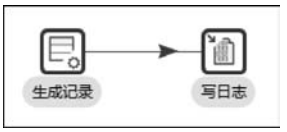



图 5-48　书写日志工作流程

（2）双击“生成记录”图标，在“限制”文本框中输入 40，并分别设置“字段”中的名称、类型和值，如图 5-49 所示。


图 5-49　设置生成记录

（3）双击“写日志”图标，单击“获取字段”按钮，自动获取字段名称，并在“写日志”文本框中输入自定义内容，如图 5-50 所示。


图 5-50　写日志(1)

（4）保存该文件，运行转换，在执行结果区域中的“日志”选项卡中查看日志状态，在 Preview data 选项卡中预览生成的数据，如图 5-51 和图 5-52 所示。

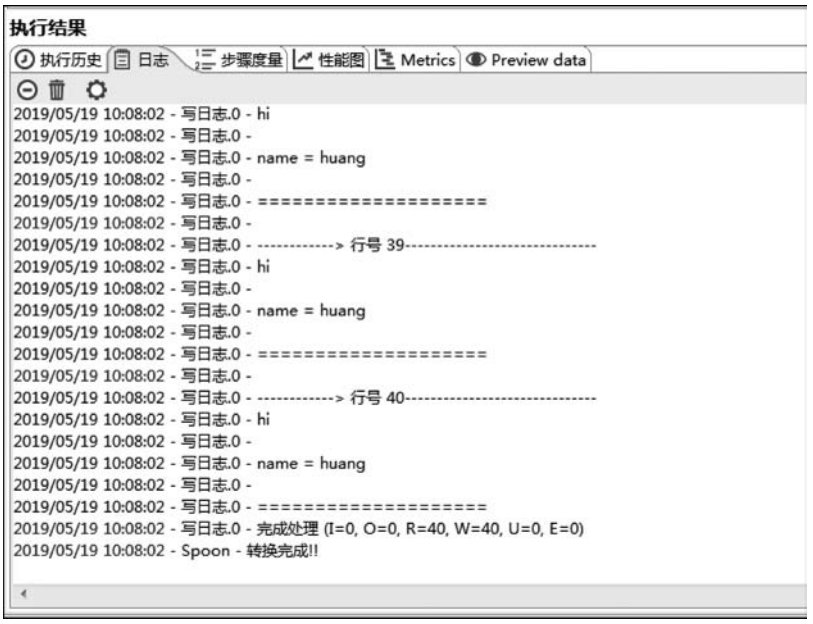



图 5-51　查看日志状态(1)

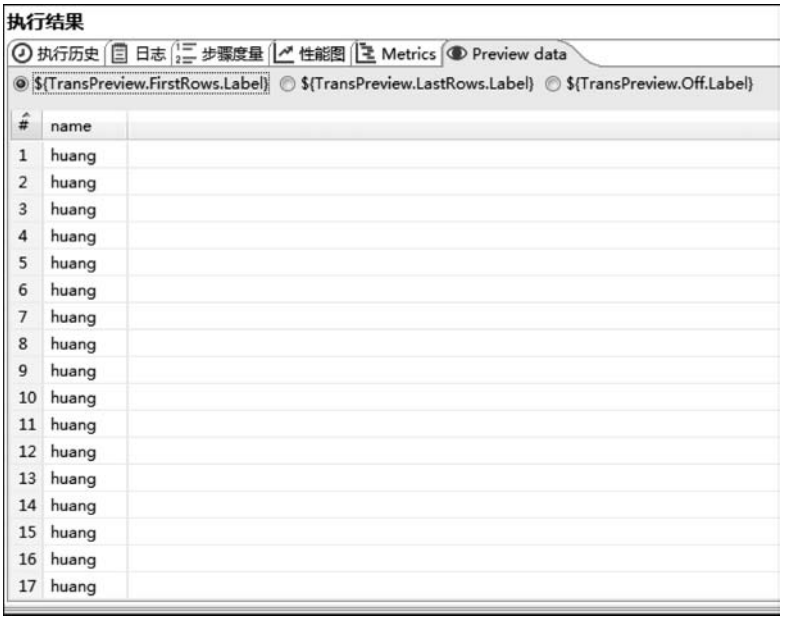



图 5-52　预览生成的数据(1)

日志是针对运行过程的信息反馈,在程序监控和调用中十分有用。此外,在 Kettle 执行结果中通过查看日志可以更好地进行数据仓库的开发与测试。

**【例 5-9】** 使用 Kettle 自定义常量并输入日志。

(1) 启动 Kettle 后,新建转换,在“输入”列表中选择“自定义常量数据”步骤,在“应用”列表中选择“写日志”步骤,分别拖动到右侧工作区中,并建立彼此之间的节点连接关系,如图 5-53 所示。

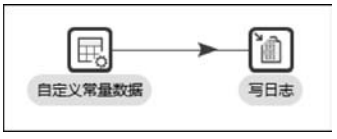



图 5-53　自定义常量并输入日志工作流程

(2) 双击“自定义常量数据”图标,在“元数据”选项卡中设置名称和类型,并将“设为空串?”的值设为否,如图 5-54 所示。


图 5-54　设置元数据(3)

(3) 在“数据”选项卡中输入数据内容,如图 5-55 所示。


图 5-55 设置数据(3)

(4) 双击“写日志”图标,单击“获取字段”按钮,自动获取字段名称,并在“写日志”文本框中输入自定义内容,如图 5-56 所示。


图 5-56 写日志(2)

(5) 保存该文件,运行转换,在执行结果区域的“日志”选项卡中查看日志状态,在 Preview data 选项卡中预览生成的数据,如图 5-57 和图 5-58 所示。

**【例 5-10】** 使用 Kettle 正则表达式清洗数据。

(1) 启动 Kettle 后,新建转换,在“输入”列表中选择“自定义常量数据”步骤,在“检

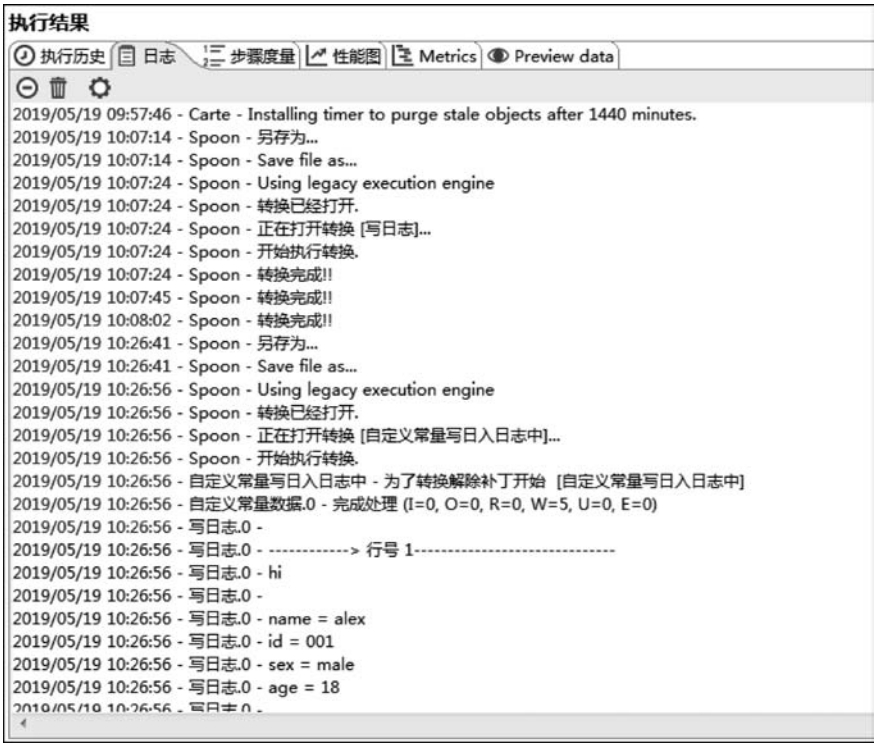



图 5-57　查看日志状态(2)

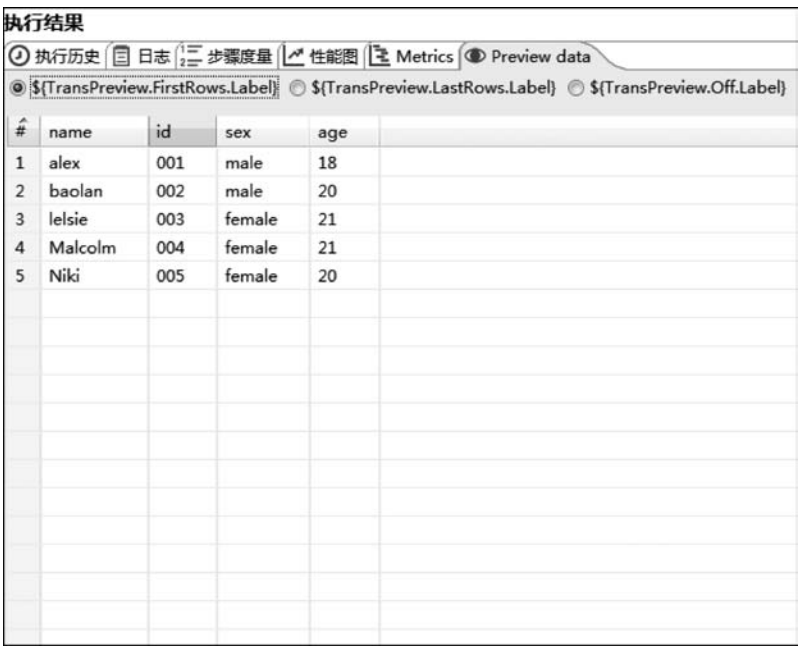



图 5-58　预览生成的数据(2)

验”列表中选择“数据检验”步骤，在“输出”列表中选择“文本文件输出”步骤，分别拖动到右侧工作区中，其中“文本文件输出”步骤拖动两次，并建立彼此之间的节点连接关系，如图 5-59 所示。值得注意的是，在“数据检验”与“文本文件输出 2”节点的连接中，需要在“数据检验”中设置错误处理步骤。

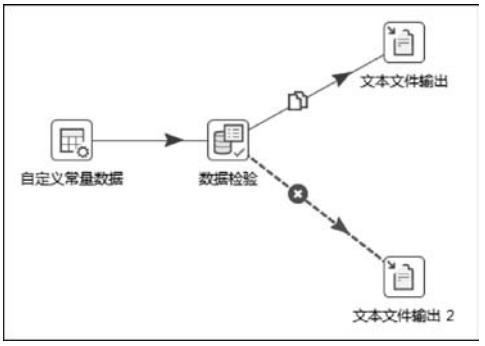



图 5-59　正则表达式清洗数据工作流程

（2）双击“自定义常量数据”图标，在“元数据”和“数据”选项卡中设置内容，如图 5-60 和图 5-61 所示。


图 5-60　设置元数据(4)

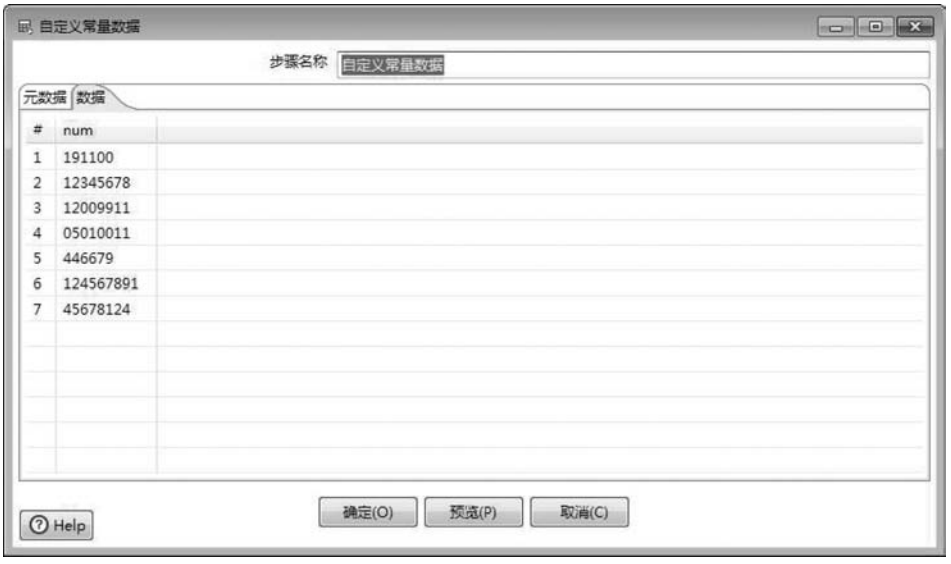



图 5-61　设置数据(4)

（3）双击“数据检验”图标，在“检验描述”文本框中输入 day，选择“要检验的字段名”为 num，并在“合法数据的正则表达式”文本框中填写\d{3,6}，该表达式含义为输出长度为 3～6 位的数据，如图 5-62 所示。


图 5-62　设置正则表达式

（4）保存该文件，运行转换，分别单击“文本文件输出”和“文本文件输出 2”图标，在执行结果区域的 Preview data 选项卡中查看运行结果，如图 5-63 和图 5-64 所示。

本例使用了正则表达式，正则表达式又称为规则表达式，是对字符串操作的一种逻辑公式。其特点是用事先定义好的一些特定字符以及这些特定字符的组合，组成一个“规则字符串”，这个规则字符串用来表达对字符串的一种过滤逻辑，通常被用来检索、替换那些

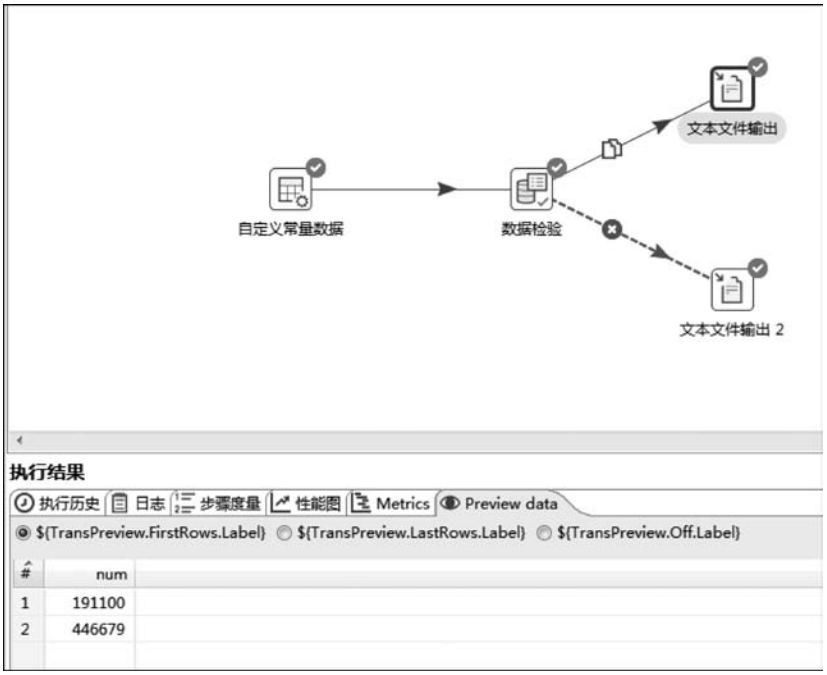



图 5-63　查看结果(文本文件输出)

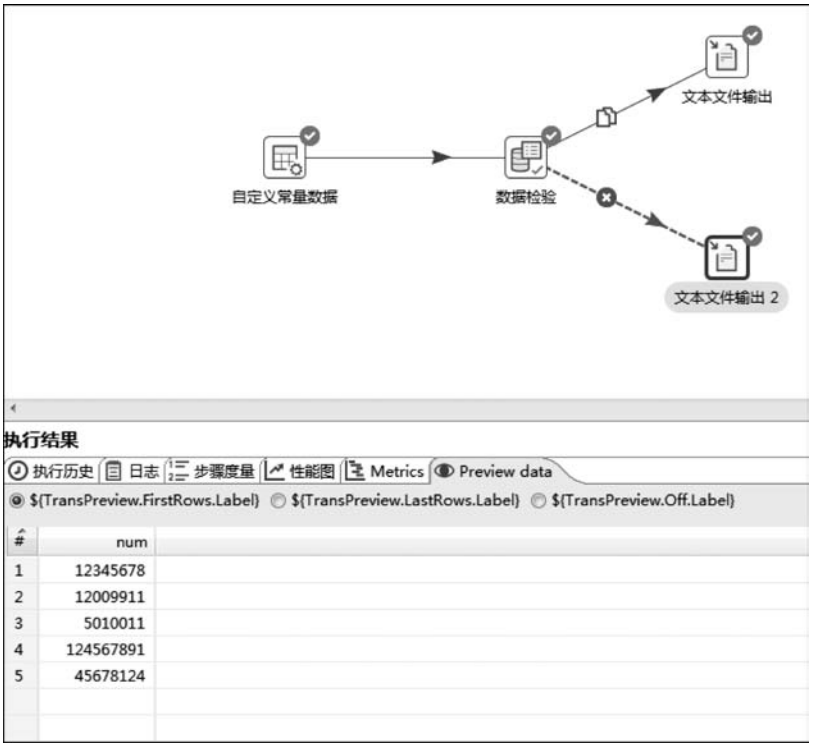



图 5-64　查看结果(文本文件输出 2)

符合某个模式(规则)的文本。构造正则表达式的方法和创建数学表达式的方法一样,也就是用多种元字符与运算符可以将小的表达式结合在一起,创建更大的表达式。正则表达式的组件可以是单个字符、字符集合、字符范围、字符间的选择或所有这些组件的任意组合。表 5-1 给出了常见的正则表达式规则说明。

**表 5-1 常见的正则表达式规则说明**

| 正则表达式 | 说 明 | 正则表达式 | 说 明 |
|---|---|---|---|
| \d | 代表一个数字 | + | 表示前一个字符至少出现一次 |
| * | 代表任意的字符 | — | 表示一个范围 |
| [] | []内的字符只能取其一 | ? | 表示前一个字符可出现 0 次或一次 |
| {} | 指定字符的个数 | | |

例如,声明电话号码。该数据类型由 ***-******** 组成,如 023-67670011,可使用正则表达式表示为"\d{3}-d{8}"。

再比如,声明密码。该数据类型由 ********* 组成,前 3 位是字母,后 6 位是数字,如 abc123456,可使用正则表达式表示为"[a-z]{3}[0-9]{6}"。

正则表达式有以下常见应用:

(1) 替换指定内容;

(2) 数字替换;

(3) 删除指定字符;

(4) 删除空行。

视频讲解

**【例 5-11】** 使用 Kettle 对字段进行拆分。

(1) 启动 Kettle 后,新建转换,在“输入”列表中选择“自定义常量数据”步骤,在“转换”列表中选择“列拆分为多行”步骤,分别拖动到右侧工作区中,并建立彼此之间的节点连接关系,如图 5-65 所示。

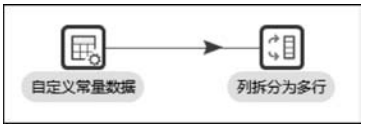



图 5-65 拆分字段工作流程

(2) 双击“自定义常量数据”图标,在“元数据”和“数据”选项卡中设置内容,如图 5-66 和图 5-67 所示。

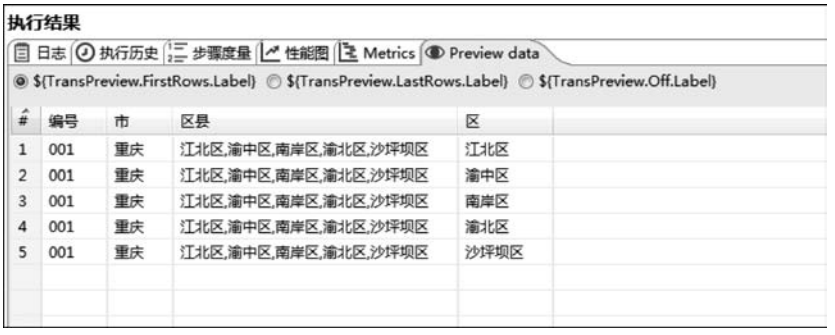

图 5-66 设置元数据(5)


图 5-67 设置数据(5)

(3) 双击“列拆分为多行”图标,在“要拆分的字段”下拉列表中选择“区县”,在“分隔符”文本框中输入[,],勾选“分隔符是一个正则表达”,并将“新字段名”设置为“区”,如图 5-68 所示。


图 5-68 设置列拆分为多行

(4) 保存该文件,运行转换,在执行结果区域的 Preview data 选项卡中查看运行结果,这时可以看到数据增加了一个“区”字段,即将最初的区县字段的内容进行了拆分,如图 5-69 所示。




| # | 编号 | 市 | 区县 | 区 |
|---|---|---|---|---|
| 1 | 001 | 重庆 | 江北区,渝中区,南岸区,渝北区,沙坪坝区 | 江北区 |
| 2 | 001 | 重庆 | 江北区,渝中区,南岸区,渝北区,沙坪坝区 | 渝中区 |
| 3 | 001 | 重庆 | 江北区,渝中区,南岸区,渝北区,沙坪坝区 | 南岸区 |
| 4 | 001 | 重庆 | 江北区,渝中区,南岸区,渝北区,沙坪坝区 | 渝北区 |
| 5 | 001 | 重庆 | 江北区,渝中区,南岸区,渝北区,沙坪坝区 | 沙坪坝区 |

图 5-69 查看拆分结果

**【例 5-12】** 使用哈希值清洗重复数据。

(1) 启动 Kettle 后,新建转换,在“输入”列表中选择“自定义常量数据”步骤,在“转换”列表中选择“唯一行(哈希值)”步骤,分别拖动到右侧工作区中,并建立彼此之间的节点连接关系,如图 5-70 所示。

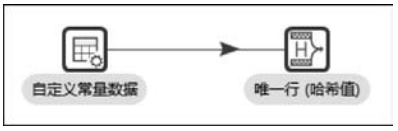



图 5-70 使用哈希值清洗重复数据工作流程

(2) 双击“自定义常量数据”图标,在“元数据”和“数据”选项卡中设置内容,如图 5-71 和图 5-72 所示。


图 5-71 设置元数据(6)


图 5-72 设置数据(6)

(3) 双击“唯一行(哈希值)”图标,在“字段名称”中选择“姓名”,即对“姓名”字段进行比较,如图 5-73 所示。

(4) 保存该文件,运行转换,在执行结果区域的 Preview data 选项卡中查看运行结果,这时可以看到数据中已经去除了重复数据“张鑫”,如图 5-74 所示。


图 5-73 设置用于比较的字段

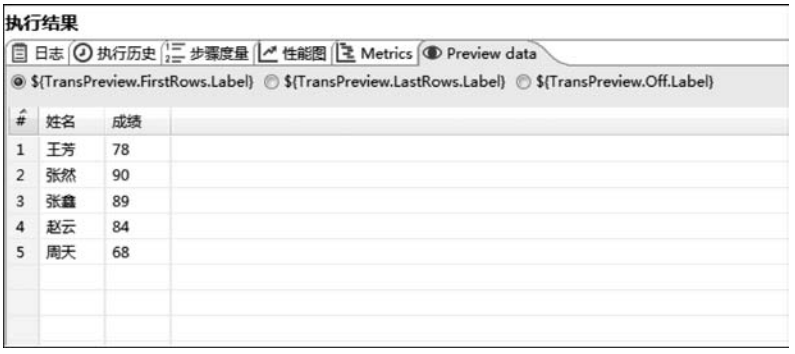



图 5-74 查看去重结果

可以看出,使用“唯一行(哈希值)”步骤可以不事先对数据进行排序,它是在内存中对数据进行去重操作。“唯一行(哈希值)”步骤是根据哈希值进行比较的,而“去除重复记录”步骤是根据相邻两行数据是否一致进行比较的。


视频讲解

**【例 5-13】** 使用 Kettle 对数据进行模糊匹配。

(1) 启动 Kettle 后,新建转换,在“输入”列表中选择“自定义常量数据”步骤,在“查询”列表中选择“模糊匹配”步骤,在“应用”列表中选择“写日志”步骤,分别拖动到右侧工作区中,其中“自定义常量数据”步骤拖动两次,并重命名为 tab_a 和 tab_b,建立彼此之间的节点连接关系,如图 5-75 所示。

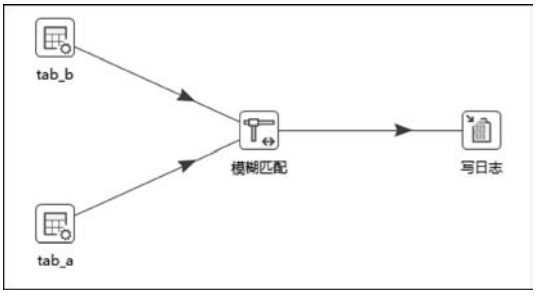



图 5-75 模糊匹配工作流程

（2）双击 tab_b 图标，在“元数据”和“数据”选项卡中分别设置内容，如图 5-76 和图 5-77 所示。


图 5-76　设置 tab_b 元数据


图 5-77　设置 tab_b 数据

（3）双击 tab_a 图标，在“元数据”和“数据”选项卡中分别设置内容，如图 5-78 和图 5-79 所示。


图 5-78　设置 tab_a 元数据


图 5-79 设置 tab_a 数据

(4) 双击“模糊匹配”图标,在“一般”选项卡和“字段”选项卡中分别设置内容,如图 5-80 和图 5-81 所示。其中,在“一般”选项卡中将“匹配步骤”设置为 tab_b,将“匹配字段”设置为 name,将“主要流字段”设置为 name,将“算法”设置为 Jaro Winkler,将“最小值”设置为 0,“最大值”设置为 1;并在“字段”选项卡中将“匹配字段”设置为 match,将“值字段”设置为 measure value。


图 5-80 设置“一般”选项卡


图 5-81 设置“字段”选项卡

本例采用 Jaro Winkler 算法进行字符串之间的模糊匹配,这是计算两个字符串之间相似度的一种算法。算法得分越高,说明相似度越大,如得 0 分,表示没有任何相似度,1 分则代表完全匹配。Jaro Winkler 算法得分公式为

$$d_j=\frac{1}{3}\left(\frac{m}{|s1|}+\frac{m}{|s2|}+\frac{m-t}{m}\right)$$

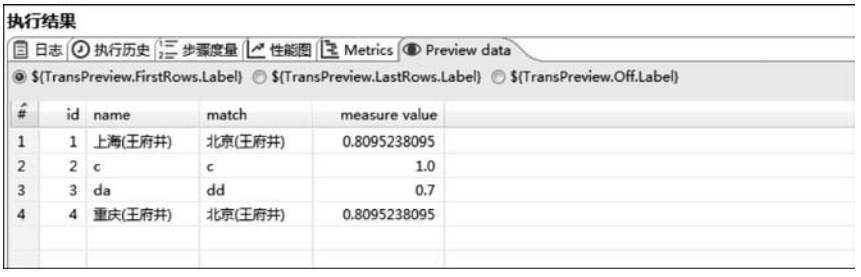
其中,s1 和 s2 表示要比对的两个字符;d_j 表示最后得分;m 表示要匹配的字符数。

此外，还可以根据需要选择其他算法，如 Needleman Wunsch（文本比较算法）、Levenshtein（编辑字符串距离算法）、Damerau Levenshtein（最佳字符串匹配算法）、SoundEx（语音算法）等。

（5）双击“写日志”图标，设置写入的日志字段内容，如图 5-82 所示。


图 5-82　写日志

（6）保存该文件，运行转换。单击“写日志”图标，在执行结果区域的 Preview data 选项卡中查看结果，如图 5-83 所示。




| # | id | name | match | measure value |
|---|---|---|---|---|
| 1 | 1 | 上海(王府井) | 北京(王府井) | 0.8095238095 |
| 2 | 2 | c | c | 1.0 |
| 3 | 3 | da | dd | 0.7 |
| 4 | 4 | 重庆(王府井) | 北京(王府井) | 0.8095238095 |

图 5-83　查看模糊匹配结果

从图 5-83 可以看出，字符串“上海（王府井）”和字符串“北京（王府井）”的匹配度为 0.809 523 809 5，字符串 c 与字符串 c 的匹配度为 1.0（表示完全匹配），字符串 da 与字符串 dd 的匹配度为 0.7。

## 5.3　本章小结

使用 Kettle 可以完成数据仓库中的数据清洗与数据转换工作，常见的有：数据值的修改与映射、数据排序、重复数据的清洗、超出范围数据的清洗、日志的写入、JavaScript

代码数据清洗、正则表达式数据清洗、数据值的过滤以及随机值的运算等。

在 Kettle 中进行数据清洗的时候基本上没有单一的清洗步骤，通常数据清洗工作需要结合多个步骤完成。例如，数据清洗可以从数据抽取时就开始执行，并在多个步骤中通过设定清洗内容完成操作。

## 5.4　实训

### 1. 实训目的

通过本章实训了解数据清洗的特点，能进行简单的与数据清洗有关的操作。

### 2. 实训内容

1）使用 Kettle 查看数据中的空值

（1）启动 Kettle 后，新建转换，在“输入”列表中选择“自定义常量数据”步骤，在“检验”列表中选择“数据检验”步骤，在“输出”列表中选择“文本文件输出”步骤，分别拖动到右侧工作区中，其中“文本文件输出”步骤拖动两次，并建立彼此之间的节点连接关系，如图 5-84 所示。值得注意的是，在“数据检验”与“文本文件输出 2”步骤的节点连接中，需要在“数据检验”步骤中设置错误处理步骤。

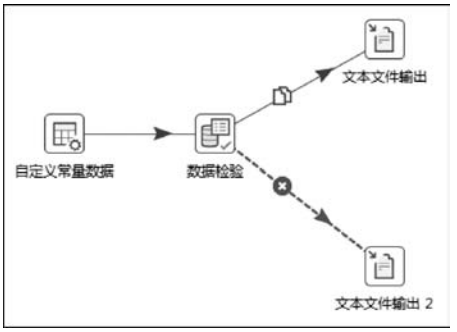



图 5-84　查看空值工作流程

（2）双击“自定义常量数据”图标，在“元数据”和“数据”选项卡中设置内容，如图 5-85 和图 5-86 所示。


图 5-85　设置元数据(7)

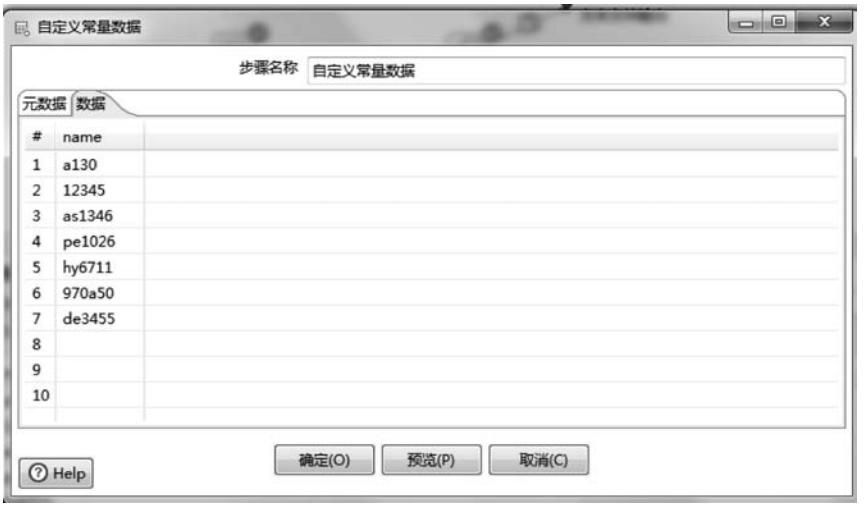



图 5-86　设置数据(7)

(3) 设置完成后,单击“预览”按钮,预览数据,如图 5-87 所示。


图 5-87　预览数据

(4) 双击“数据检验”图标,在“检验描述”文本框中输入 na,选择“要检验的字段名”为 name,在“合法数据的正则表达式”文本框中填写 null,如图 5-88 所示。

(5) 保存该文件,运行转换,分别单击“文本文件输出”和“文本文件输出 2”图标,在执行结果区域的 Preview data 选项卡中查看运行结果,如图 5-89 和图 5-90 所示。


视频讲解

2) 使用 Kettle 采样数据并输出结果

(1) 启动 Kettle 后,新建转换,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,在“统计”列表中选择“数据采样”步骤,在“流程”列表中选择“识别流的最后一行”步骤,分别拖动到右侧工作区中,并建立彼此之间的节点连接关系,如图 5-91 所示。

(2) 双击“Excel 输入”图标,添加需要的数据表,并获取表中字段。

(3) 双击“排序记录”图标,设置字段名称为“成绩”,如图 5-92 所示。


数据检验
步骤名称：
数据检验
选择一个要编辑的检验
na
报告所有错误，不止第一个错误
只输出一行，指定连接错误的分隔符：
检验描述
na
要检验的字段名
name
错误代码
错误描述
类型
检验数据类型
数据类型
String
转换掩码
小数点符号
分组符号
数据
允许空?
只允许空值
只允许数值类型的数据
最大字符串长度
最小字符串长度
最大值
最小值
允许的开始字符串
允许的结束字符串
不允许的开始字符串
不允许的结束字符串
合法数据的正则表达式
null
不合法数据的正则表达式
允许的值
增加
删除
从其他步骤获得允许的值
要读取的步骤
要读取的字段
Help
确定(O)
增加检验
删除检验
取消(C)

图 5-88 设置数据检验

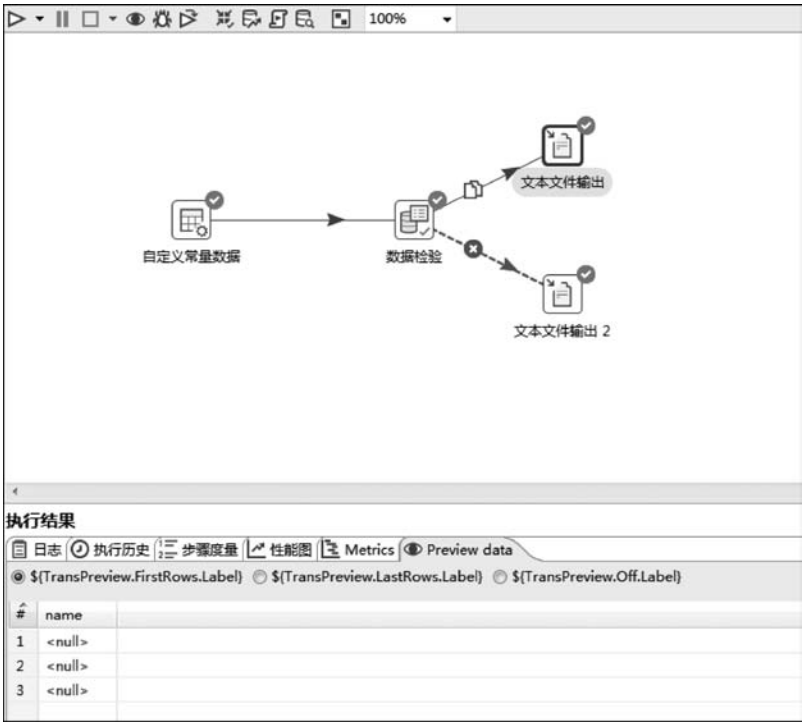



图 5-89　查看空值结果(文本文件输出)

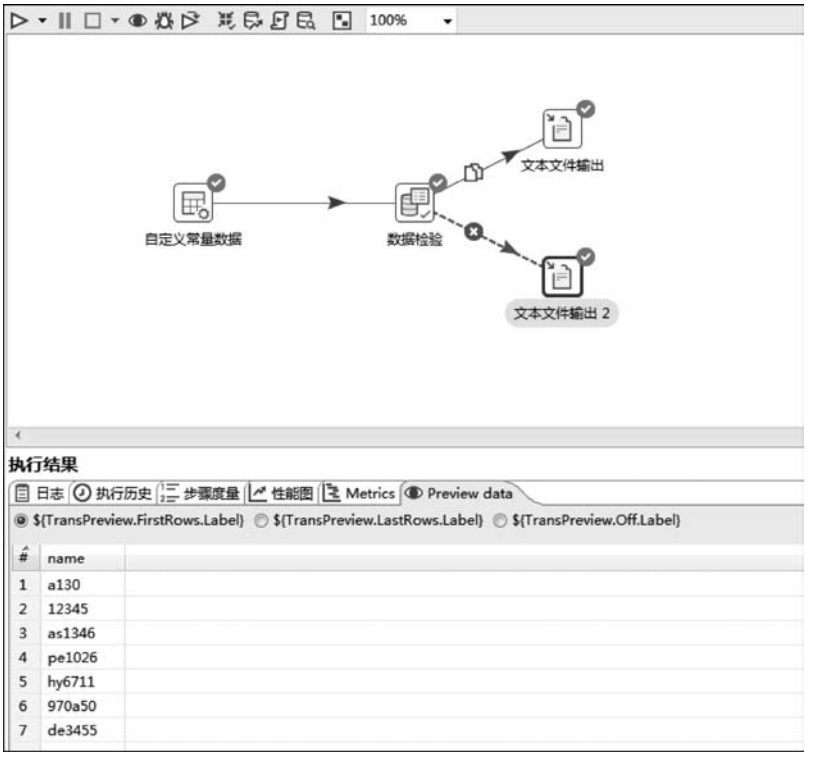



图 5-90　查看空值结果(文本文件输出 2)

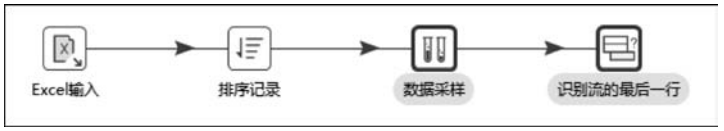



图 5-91 采样数据并输出工作流程


图 5-92 设置排序记录

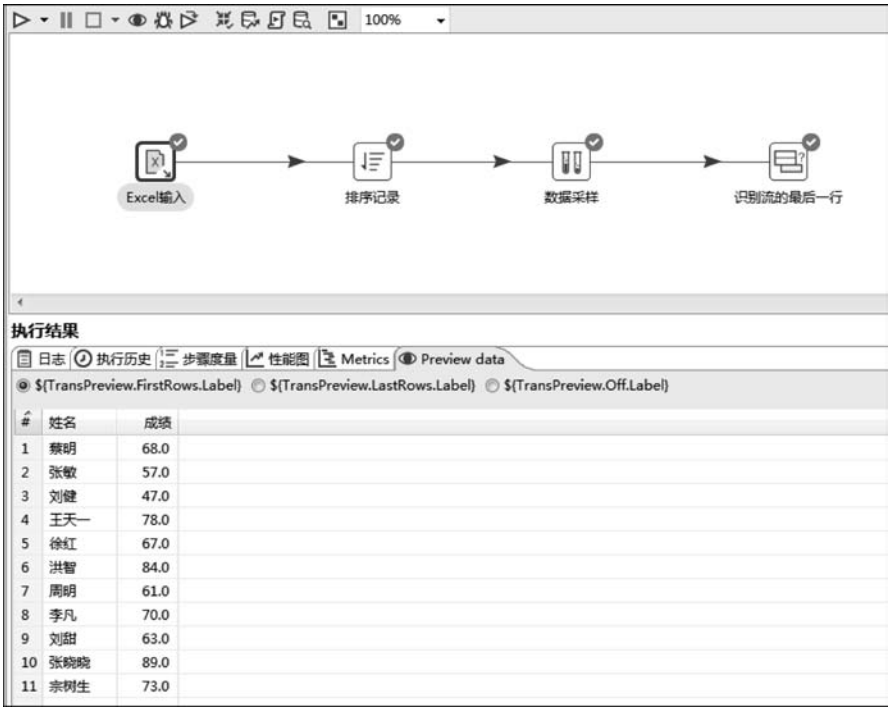
(4) 双击“数据采样”图标，设置 Sample size(样本容量)为 5，Random seed(随机种子数)为 1，如图 5-93 所示。

(5) 双击“识别流的最后一行”图标，设置“结果字段名”为 Last，该字段用于获得最后一行的数据，如图 5-94 所示。


图 5-93 设置采样参数

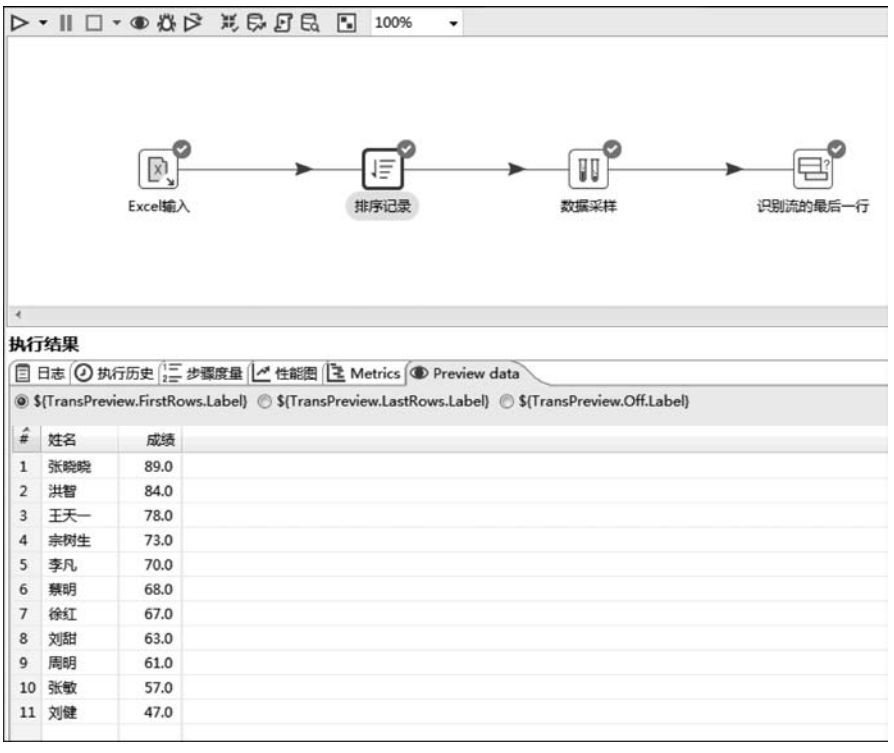

图 5-94 设置识别流的最后一行

(6) 保存该文件，运行转换，依次单击“Excel 输入”“排序记录”“数据采样”“识别流的最后一行”图标，在执行结果区域的 Preview data 选项卡中查看最终的结果，如图 5-95～图 5-98 所示。




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

图 5-95　Excel 输入结果




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

图 5-96　排序结果

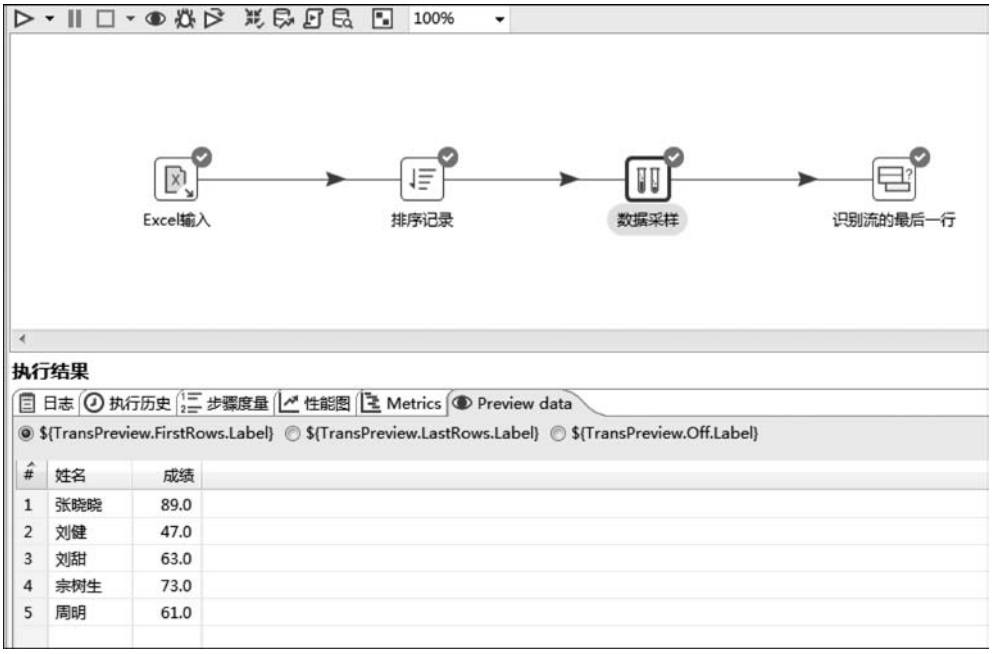



图 5-97 采样结果

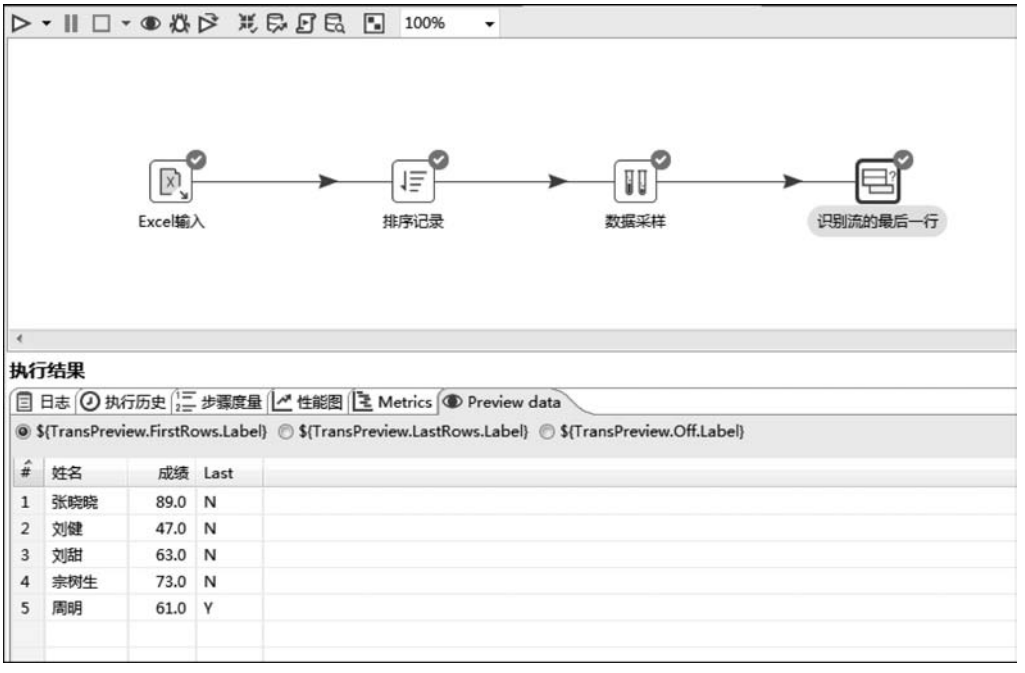



图 5-98 最终输出结果

本例导入的 Excel 表格内容如图 5-99 所示。

3）使用 Kettle 实现字符串替换

（1）启动 Kettle 后，新建转换，在“输入”列表中选择“自定义常量数据”步骤，在“转

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 姓名 | 成绩 | | | |
| 2 | 蔡明 | 68 | | | |
| 3 | 张敬 | 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 | | | | | |
| 14 | | | | | |
| 15 | | | | | |

图 5-99　本例导入的 Excel 表格内容

换”列表中选择“字符串替换”步骤，分别拖动到右侧工作区中，并建立彼此之间的节点连接关系，如图 5-100 所示。

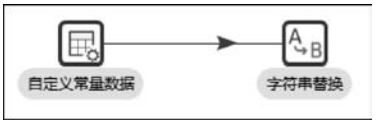



图 5-100　字符串替换工作流程

（2）双击“自定义常量数据”图标，在“元数据”和“数据”选项卡中分别设置内容，如图 5-101 和图 5-102 所示。


图 5-101　设置元数据(8)


图 5-102　设置数据(8)

(3) 双击“字符串替换”图标,设置如图 5-103 所示。


图 5-103 设置字符串替换内容

在本例中,将输入流字段 name 改为输出流字段 new_name,并使用正则表达式将 a～z 的字母用数字 5 进行替换。

(4) 保存该文件,运行转换,单击“字符串替换”图标,在执行结果区域的 Preview data 选项卡中查看运行结果,如图 5-104 所示。


图 5-104 查看字符串替换结果

从图 5-104 可以看出,本例输入的字符串 jj＃kljl 和 jj＃＃kljl 分别被替换成了输出字符串 55＃5555 和 55＃＃5555。

4) 使用 Kettle 拆分字段并保存为日志

(1) 启动 Kettle 后,新建转换,在“输入”列表中选择“自定义常量数据”步骤,在“转换”列表中选择“拆分字段”步骤,在“应用”列表中选择“写日志”步骤,分别拖动到右侧工作区中,并建立彼此之间的节点连接关系,如图 5-105 所示。


图 5-105 拆分字段并写日志工作流程

(2) 双击“自定义常量数据”图标,在“元数据”和“数据”选项卡中分别设置内容,如图 5-106 和图 5-107 所示。

视频讲解


图 5-106　设置元数据(9)


图 5-107　设置数据(9)

(3) 双击“拆分字段”图标,在“需要拆分的字段”下拉列表中选择 name,在“分隔符”文本框中输入“,”,并设置新的字段,如图 5-108 所示。


图 5-108　设置拆分字段

(4) 双击“写日志”图标,在弹出的对话框中设置字段内容,如图 5-109 所示。

(5) 保存该文件,运行转换,依次单击“自定义常量数据”和“写日志”图标,在执行结果区域的 Preview data 选项卡中分别查看结果,如图 5-110 和图 5-111 所示。


图 5-109 设置写日志

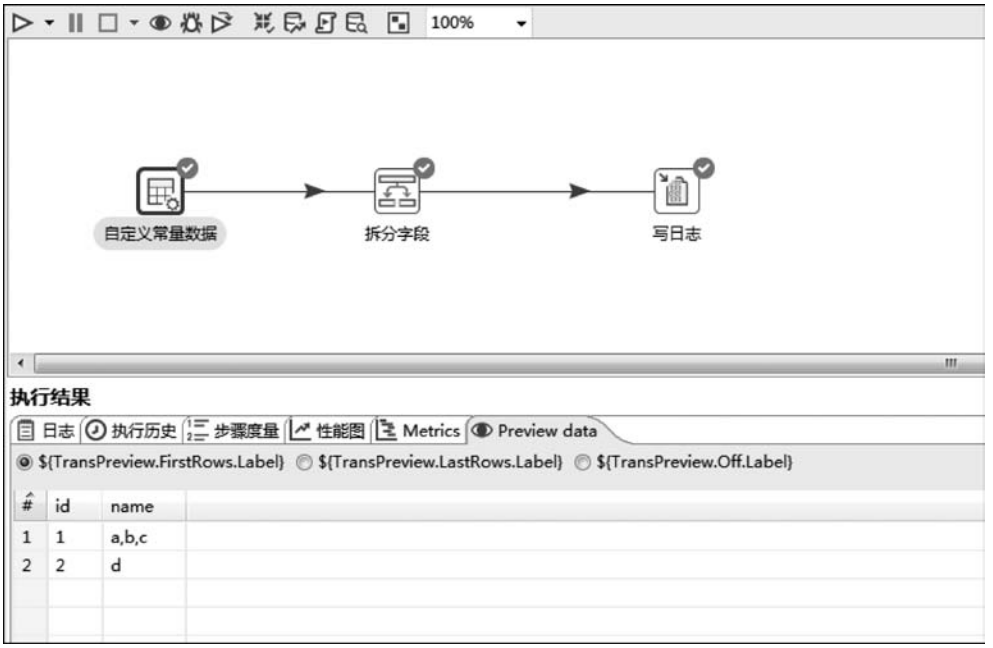



图 5-110 初始设置

从该例可以看出,在 Kettle 中可以使用“拆分字段”步骤将初始字段(如“a,b,c”)拆分为多个字段(如“a”“b”和“c”)。

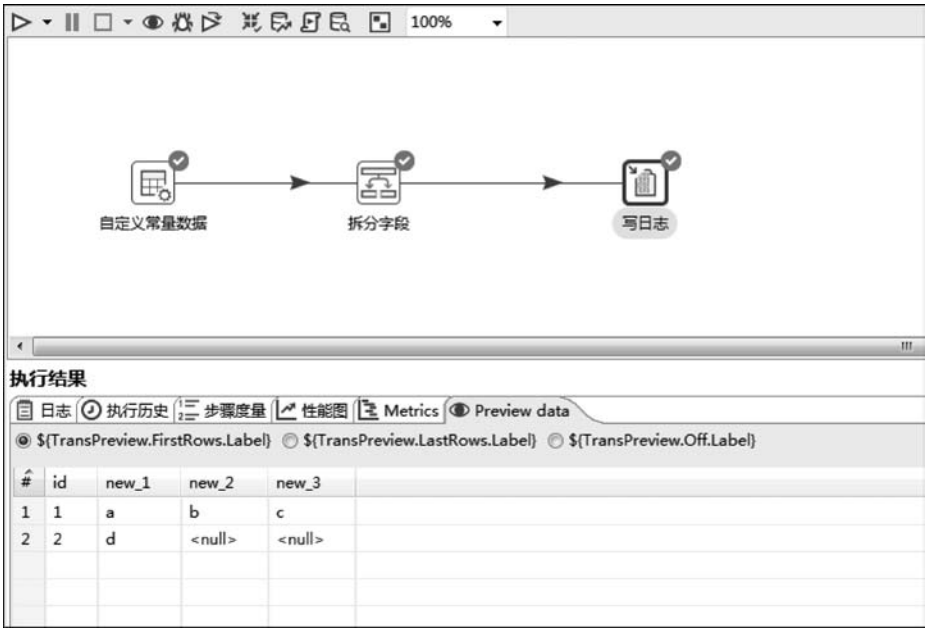



图 5-111　拆分后的结果

## 习题 5

(1) 请阐述 Kettle 数据清洗的常见步骤。
(2) 在 Kettle 中什么是值映射?
(3) 在 Kettle 中什么是采样?
(4) 在 Kettle 中什么是计算器? 如何使用计算器?
(5) 在 Kettle 中如何对重复数据进行清洗?

# 第 6 章

# 数据迁移

**本章学习目标**

- 了解数据迁移的概念
- 了解数据迁移的过程
- 了解数据迁移的方法
- 了解数据迁移的相关技术
- 掌握 Kettle 中的数据迁移实现

本章先介绍数据迁移的概念和数据迁移的过程，再介绍数据迁移的方法，接着介绍数据迁移的相关技术，最后介绍使用 Kettle 实现数据迁移。

视频讲解

## 6.1 数据迁移概述

### 1. 数据迁移简介

在企业的实际应用中，数据通常以不同的方式存储于不同的数据库中。因此，如何采集和获取这些数据量大并且来源复杂的数据，成为现代企业进行大数据分析的一个难题。

数据迁移又称为分级存储管理（Hierarchical Storage Management，HSM），是一种将离线存储与在线存储融合的技术，是数据系统整合中保证系统平滑升级和更新的关键部分。它将高速、高容量的非在线存储设备作为磁盘设备的下一级设备，然后将磁盘中常用的数据按指定的策略自动迁移到磁带库（简称为带库）等二级大容量存储设备上。当需要使用这些数据时，分级存储系统会自动将这些数据从下一级存储设备调回到上一级磁盘上。对于用户，上述数据迁移操作完全是透明的，只是在访问磁盘的速度上略有怠慢，而

在逻辑磁盘的容量上明显感觉大大提高了。图 6-1 所示为数据迁移示意图。

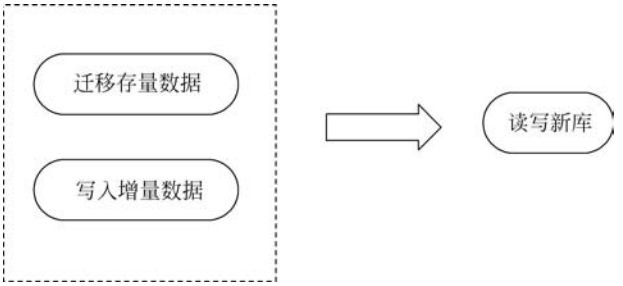



图 6-1　数据迁移

从图 6-1 可以看出，数据迁移既包含了迁移存量数据（老数据），又包含了写入增量数据（新数据）。不过简单来讲，数据迁移就是将数据从一个地方挪到另外一个地方，它是将很少使用或不用的文件移到辅助存储系统（如磁带或光盘）的存档过程。这些文件通常是在未来任何时间可进行方便访问的图像文件或历史信息。数据迁移工作通常与数据备份策略相结合，并且要求定期备份数据。此外，数据迁移还包括计算机数据迁移，如迁移旧计算机（旧系统）中的数据、应用程序、个性化设置等到新计算机（新系统），特别是在系统升级后很有必要。图 6-2 显示了使用 Access 链接到 ODBC 数据库，以实现数据表的导入和导出。


图 6-2　Access 链接到 ODBC 数据库


视频讲解

**【例 6-1】** 使用 Access 迁移数据表中所有记录。

（1）运行 Access 2007，新建表 stu，并添加字段及对应的数据，如图 6-3 所示。

（2）右击表 stu，在弹出的快捷菜单中将该表导出为 Excel，如图 6-4 所示。

stu

| ID | 字段1 | 字段2 | 字段3 | 添加新字段 |
|---|---|---|---|---|
| 1 | id | name | score | |
| 2 | 001 | Lucy | 78 | |
| 3 | 002 | Tony | 89 | |
| 4 | 003 | Tom | 68 | |
| (新建) | | | | |

图 6-3　新建表 stu，并添加字段及对应的数据

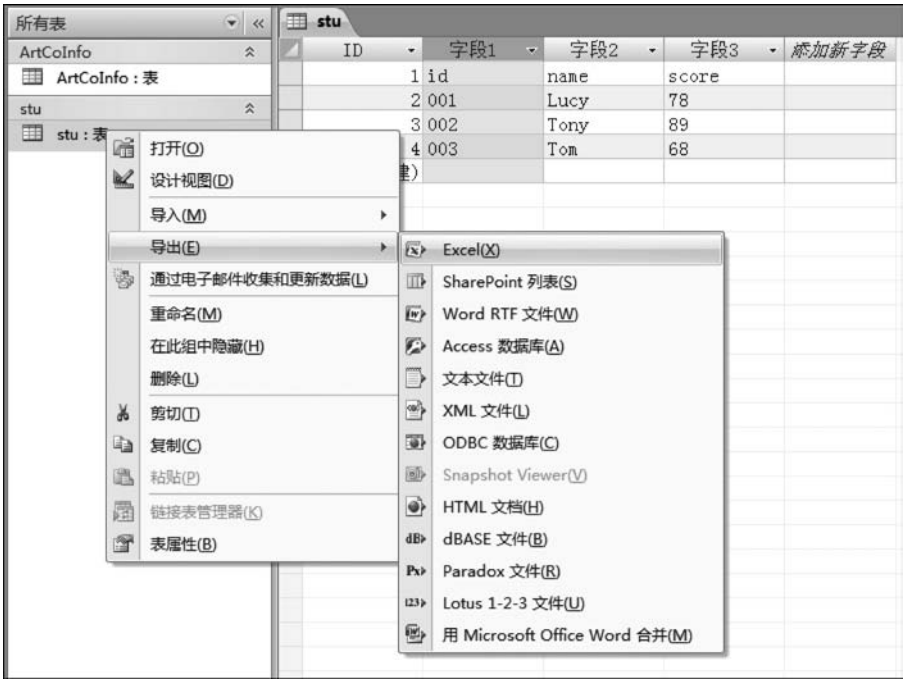



图 6-4　将表 stu 导出为 Excel 表

（3）打开导出的 Excel 表，如图 6-5 所示。

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | ID | 字段1 | 字段2 | 字段3 | | |
| 2 | 1 | id | name | score | | |
| 3 | 2 | 001 | Lucy | 78 | | |
| 4 | 3 | 002 | Tony | 89 | | |
| 5 | 4 | 003 | Tom | 68 | | |
| 6 | | | | | | |
| 7 | | | | | | |
| 8 | | | | | | |

图 6-5　导出的 Excel 表内容

### 2. 数据迁移的过程

数据迁移的实现可以分为 3 个阶段：数据迁移前的准备、数据迁移的实施和数据迁移后的校验。由于数据迁移的特点，大量的工作都需要在准备阶段完成，充分而周到的准备工作是完成数据迁移的主要基础。具体而言，要进行待迁移数据源的详细说明(包括数据的存储方式、数据量、数据的时间跨度)；建立新旧系统数据库的数据字典；对旧系统的历史数据进行质量分析；对新旧系统数据结构进行差异分析；对新旧系统代码数据进行差异分析；建立新旧系统数据库表的映射关系；确定对无法映射字段的处理方法；开发、部属 ETL 工具，编写数据转换的测试计划和校验程序；制订数据转换的应急措施。

其中，数据迁移的实施是实现数据迁移的 3 个阶段中最重要的环节。它要求制订数据转换的详细实施步骤流程；准备数据迁移环境；业务上的准备，结束未处理完的业务事项，或将其告一段落；对数据迁移涉及的技术都得到测试；最后实施数据迁移。

具体来讲，在数据迁移实施中常包含以下 3 个步骤。

(1) 从源数据表查询出要迁移的数据。

(2) 把数据插入新表。

(3) 把旧表的数据删除或更新。

另外，数据迁移后的校验是对迁移工作的检查，数据校验的结果是判断新系统能否正式启用的重要依据。可以通过质量检查工具或编写检查程序进行数据校验，通过试运行新系统的功能模块，特别是查询、报表功能，检查数据的准确性。

### 3. 数据迁移标准

1) 数据一致性

在进行数据迁移时，迁移完成后不能丢失记录，单条记录的数据不能缺失字段，并且在迁移之后需要保证新的库和旧的库的数据是一致的。

2) 业务可用性

数据迁移应该是在线的迁移，也就是在迁移的同时还会有数据的写入，因此需要保证业务写入的可用性。

3) 迁移过程可中断、可回滚

这点要求很高，是确保数据万无一失的策略。在迁移数据的各个阶段发现有问题，都可以回滚到原来的库，不会对系统的可用性造成影响，以此保证业务正常运行。

### 4. 数据迁移的准备

数据转换与迁移通常包括多项工作：旧系统数据字典整理、旧系统数据质量分析、新系统数据字典整理、新旧系统数据差异分析、建立新旧系统数据之间的映射关系、开发部署数据转换与迁移程序、制订数据转换与迁移过程中的应急方案、实施旧系统数据到新系

统的转换与迁移工作、检查转换与迁移后数据的完整性与正确性。

一般来讲，数据转换与迁移的过程大致可以分为抽取、转换、装载 3 个步骤。数据抽取、转换是根据新旧系统数据库的映射关系进行的，而数据差异分析是建立映射关系的前提，这其中还包括对代码数据的差异分析。转换步骤一般还包含数据清洗的过程，数据清洗主要是针对源数据库，对出现二义性、重复、不完整、违反业务或逻辑规则等问题的数据进行相应的清洗操作；在清洗之前需要进行数据质量分析，以找出存在问题的数据，否则数据清洗将无从谈起。数据装载是通过装载工具或自行编写的 SQL 程序将抽取、转换后的结果数据加载到目标数据库中。

**5. 数据迁移后的校验**

在数据迁移完成后，需要对迁移后的数据进行校验。数据迁移后的校验是对迁移质量的检查，同时数据校验的结果也是判断新系统能否正式启用的重要依据。

可以通过以下两种方式对迁移后的数据进行校验。

(1) 新旧系统查询数据对比检查，通过新旧系统各自的查询工具，对相同指标的数据进行查询，并比较最终的查询结果。

(2) 先将新系统的数据恢复到旧系统迁移前一天的状态，然后将最后一天发生在旧系统上的业务全部补录到新系统，检查有无异常，并和旧系统比较最终产生的结果。

对迁移后的数据进行质量分析，可以通过数据质量检查工具或编写有针对性的检查程序进行。对迁移后数据的校验有别于迁移前历史数据的质量分析，主要是检查指标的不同。迁移后数据校验的指标主要包括以下 5 方面。

(1) 完整性检查：引用的外键是否存在。

(2) 一致性检查：相同含义的数据在不同位置的值是否一致。

(3) 总分平衡检查：如欠税指标的总和与分部门、分户不同数据的合计对比。

(4) 记录条数检查：检查新旧数据库对应的记录条数是否一致。

(5) 特殊样本数据的检查：检查同一样本在新旧数据库中是否一致。

## 6.2　数据迁移实现技术

数据迁移实现技术的选择是建立在对系统软硬件以及业务系统的各环节的分析基础之上的。目前开放平台系统中可以采用的数据迁移技术根据发起端的不同，可以分为基于主机的迁移方式、备份恢复的迁移方式、基于存储的迁移方式、基于文件系统的迁移方式以及基于数据库的迁移方式等。

### 6.2.1　基于主机的迁移方式

**1. 利用操作系统命令直接复制**

该方式利用操作系统命令直接复制要迁移的数据，然后复制到要迁移的目的地，优点

是安全可靠，缺点是一般需要脱机迁移。

#### 2. 逻辑卷数据镜像方法

对需要迁移的每个卷都做逻辑卷镜像，适合已经拥有逻辑卷管理器的环境，支持在线迁移。缺点是需要准确获取所有逻辑卷管理(Logical Volume Manager，LVM)配置信息以镜像所有卷，主机层面的相关性强，迁移过程耗用主机的资源多，对业务影响较大。另外，如果同时识别不同厂家的存储，一些系统参数和多路径软件经常不兼容，在线迁移时可能会对生产造成不可知的影响。

### 6.2.2 备份恢复的迁移方式

数据备份是容灾的基础，是指为防止系统出现操作失误或系统故障导致数据丢失，而将全部或部分数据集合从应用主机的硬盘或阵列复制到其他存储介质的过程。备份恢复的迁移方式通常是利用备份管理软件对数据做备份，然后恢复到目的地，对于联机要求高的环境，可以结合在线备份的方法，然后恢复到目的地。此方式可以有效缩短停机时间窗口，一旦备份完成，数据的迁移过程完全不会影响生产系统。

### 6.2.3 基于存储的迁移方式

#### 1. 存储虚拟化

通过存储虚拟化技术将数据从源端迁移到目的地，兼容主流的存储设备，支持不同厂商不同品牌存储设备间的迁移和容灾，适合频繁移动数据的大型企业。

#### 2. 盘阵内复制方法

通过盘阵复制软件对数据进行迁移，这是比较好的迁移方法。通常来讲，该方式对业务影响极小，最大的缺点是一般只支持同一厂商的同类产品间复制。

#### 3. 不同盘阵间复制方法

对于两套同系列的磁盘阵列，可以通过阵列之间的数据复制技术实现数据的迁移，如目前 HDS 的 TureCopy、HUR 复制技术，以及 EMC 的 SRDF 技术，都可以实现在两套磁盘阵列之间的数据迁移，并且此种方法不占用主机资源，对应用透明。但是源磁盘阵列和目标磁盘阵列必须是同一厂家同一系列的产品，而且迁移过程对生产系统有一定的性能影响。

### 6.2.4 基于文件系统的迁移方式

文件系统的复制技术由来已久，常见的 CIFS/NFS 文件共享复制、X-Copy 等都属于文件系统复制技术。这种复制技术简单、快捷，对环境几乎没有复杂要求，可以快速地完成文件系统的整体复制操作。但这种方法存在以下缺点。

(1) 以文件为单位,文件发生一个字节的变化,增量复制时都需要重新传送这个文件。

(2) 文件属性会在复制之后发生变化,如 Owner 的属性、读写、执行权限等。

### 6.2.5 基于数据库的迁移方式

#### 1. 同构数据库数据迁移

同构数据库的数据迁移技术是利用数据库自身的备份和恢复功能实现数据的迁移,可以是整个库或单个表。

同构数据库的数据迁移比较简单,且不限操作系统平台,但是这种方法的缺点是在数据迁移过程中迁移的速度取决于主机的读写速度和网络传输速度。

例如,可以在 Hadoop 系统中实现 Hive 数据迁移,其中 Hive 是基于 Hadoop 的一个数据仓库工具,用来进行数据提取、转化、加载,这是一种可以存储、查询和分析存储在 Hadoop 中的大规模数据的机制。Hive 数据仓库工具能将结构化的数据文件映射为一张数据库表,并提供 SQL 查询功能,能将 SQL 语句转换成 MapReduce 任务来执行。

Hive 数据迁移常见步骤如下。

(1) 迁移表结构:从旧 Hive 中导出表结构并在新 Hive 中导入表结构。

(2) 迁移表数据:首先将 Hive 表数据导出至 HDFS;接着将 HDFS 数据下载至主机中;再将数据压缩并将数据发送至目标 Hive 集群的内网主机中;然后解压数据,将数据上传至 HDFS 中;最后将 HDFS 数据上传至 Hive 表中。

#### 2. 异构数据库数据迁移

异构数据库的数据迁移一般使用第三方软件实现,这种方法适用于纯数据库迁移,并且不需要关注具体的存储过程。如今第三方软件大多提供了不同数据库转换的解决方案,不过无论哪种方式均需对数据库迁移后的数据进行反复的测试。

异构数据库的数据迁移不限操作系统和数据库平台,不过需要大量的时间和费用,特别是专门的定制开发,更需要长时间的测试才能真正投入使用。

1) Navicat Premium

Navicat Premium 是一款数据库管理工具,是一个可多重连线数据库的管理工具,使用 Navicat Premium 可以让操作者快速地在多种数据库系统间传输资料。

图 6-6 显示了使用 Navicat Premium 进行异构数据库的数据迁移。

2) Liquibase

此外,在进行数据库迁移时,也可以使用其他数据库迁移工具,如 Liquibase 或 Flyway 等。Liquibase 是一款开源的数据库迁移工具,是一个用于跟踪、管理和应用数据库变化的数据库重构工具,它将所有数据库的变化(包括结构和数据)都保存在 XML 文件中,便于版本控制。Liquibase 的核心就是用 changeLog 文件记录跟踪数据库更新。Liquibase 支持的 changeLog 格式包括 XML、JSON、YAML 和 SQL。图 6-7 所示为 Liquibase 中的 changeLog 文件,该文件用 databaseChangeLog 标签表示。

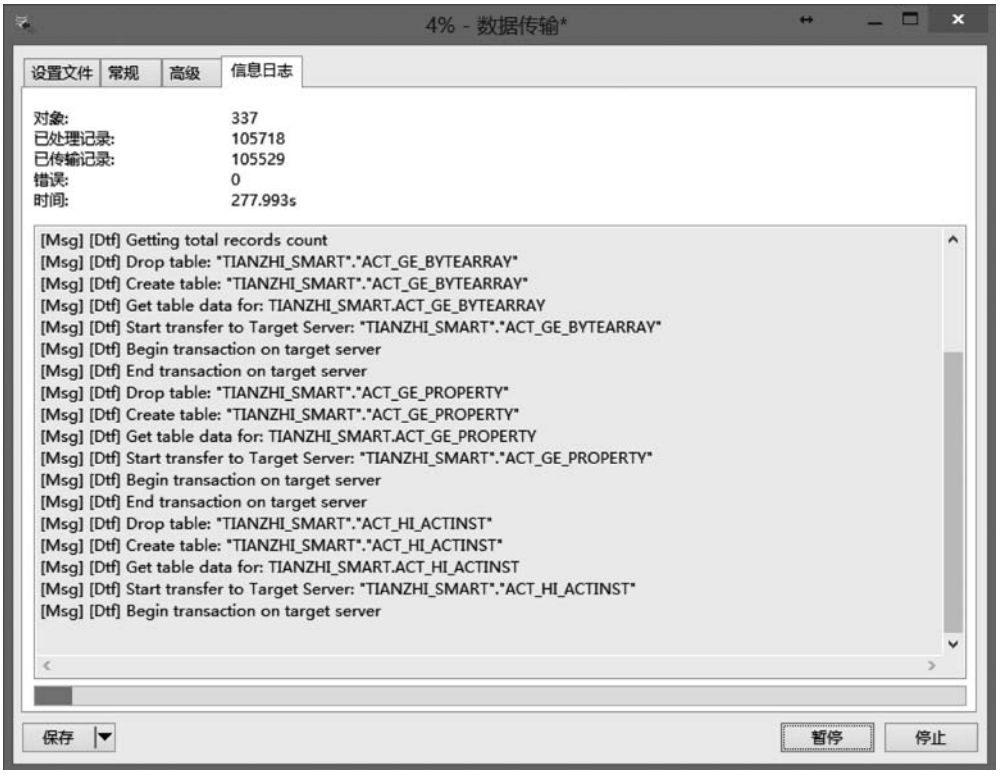



图 6-6　使用 Navicat Premium 进行异构数据库的数据迁移

```
<databaseChangeLog xmlns="http://www.liquibase.org/xml/ns/dbchangelog/1
                   xsi:schemaLocation="http://www.liquibase.org/xml/ns/
    <precondition>
        //
    </precondition>
    <changeSet id="initial" author="garfield" context="initial">
        <comment>create the factory table</comment>
        <preConditions onFail="MARK_RAN">
            <not>
                <tableExists tableName="FACTORY" schemaName="public"/>
            </not>
        </preConditions>
        <sql>
            CREATE TABLE FACTORY (
            ID VARCHAR(50) NOT NULL,
            JSON_CONTENT CLOB NOT NULL,
            PRIMARY KEY (ID)
            );
        </sql>
        <rollback>
            DROP TABLE FACTORY;
        </rollback>
    </changeSet>

</databaseChangeLog>
```

图 6-7　Liquibase 中的 changeLog 文件

Liquibase 的特点如下。

(1) 不依赖于特定的数据库,因此可以支持各种主流数据库,如 MySQL、PostgreSQL、DB2、Oracle、SQL Server、Sybase、Cache 等。

(2) 提供数据库比较功能,比较结果保存在 XML 文件中,基于该 XML 文件可用 Liquibase 轻松部署或升级数据库。

(3) 提供变化应用的回滚功能,可按时间、数量或标签(Tag)回滚已应用的变化。通过这种方式,开发人员可轻易地还原数据库至任何时间点的状态。

(4) 可生成数据库修改文档(HTML 格式)。

(5) 提供数据重构的独立的集成开发环境(Integrated Development Environment,IDE)和 Eclipse 插件。

(6) 支持多种运行方式,如命令行、Spring 集成、Maven 插件、Gradle 插件等。


视频讲解

3) Sqoop

Apache 框架 Hadoop 是一个越来越通用的分布式计算环境,主要用来处理大数据。随着云提供商利用这个框架,更多的用户将数据集在 Hadoop 和传统数据库之间转移,能够帮助数据传输的工具变得更加重要。Apache Sqoop 就是这样一款工具,它可以在 Hadoop 和关系型数据库之间转移大量数据。值得注意的是,在使用 Sqoop 前首先要下载并安装,Sqoop 的下载地址为 http://sqoop.apache.org/。

因此,Sqoop 是一个用来将 Hadoop 和关系型数据库中的数据相互转移的工具,可以将一个关系型数据库(如 MySQL、Oracle、Postgres 等)中的数据导入 Hadoop 的 HDFS 中,执行导入时,Sqoop 可以写入 HDFS、Hive 和 HBase。导入分为两步:连接到数据源以收集统计信息,然后触发执行实际导入的 MapReduce 作业。此外,Sqoop 也可以将数据从 Hadoop 系统中抽取出并导出到关系型数据库。

图 6-8 显示了使用 Sqoop 在关系型数据库和 Hadoop 之间实现数据迁移。

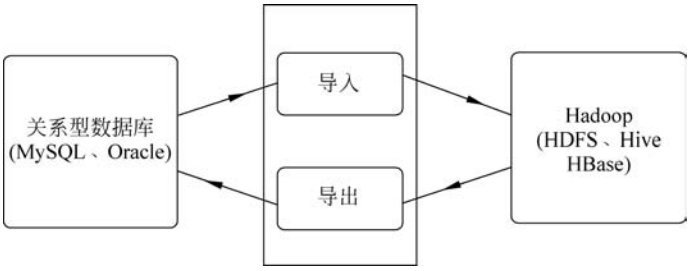



图 6-8　Sqoop 的导入和导出


视频讲解

4) Kettle

Kettle 允许用户管理来自不同数据库的数据,并通过提供一个图形化的用户环境描述用户想做什么,而不是用户想怎么做。图 6-9 显示了 Kettle 连接数据库的界面;图 6-10 显示了 Kettle 对数据库文件的输入和输出的工作流程。


图 6-9　Kettle 连接数据库的界面

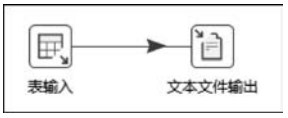



图 6-10　Kettle 对数据库文件的输入和输出的工作流程

## 6.3　数据迁移实现

### 6.3.1　数据库安装与使用

#### 1. MySQL 概述

MySQL 是一个小型的关系型数据库管理系统，由于该软件的体积小、运行速度快、

操作方便等优点，目前被广泛应用于 Web 上的中小企业网站的后台数据库中。

MySQL 数据库的优点如下。

（1）体积小、速度快、成本低。

（2）使用的核心线程是完全多线程的，可以支持多处理器。

（3）提供了多种语言支持，MySQL 为 C、C++、Python、Java、Perl、PHP、Ruby 等多种编程语言提供了 API，方便访问和使用。

（4）MySQL 支持多种操作系统，可以运行在不同的平台上。

（5）支持大量数据查询和存储，可以承受大量的并发访问。

（6）免费开源。

### 2. MySQL 安装与使用

1）MySQL 的下载

登录 MySQL 的官网 www.mysql.com，单击 DOWNLOADS 按钮，进入下载页面，下载对应操作系统的版本，本章下载的 MySQL 版本是 5.6，下载界面如图 6-11 所示。

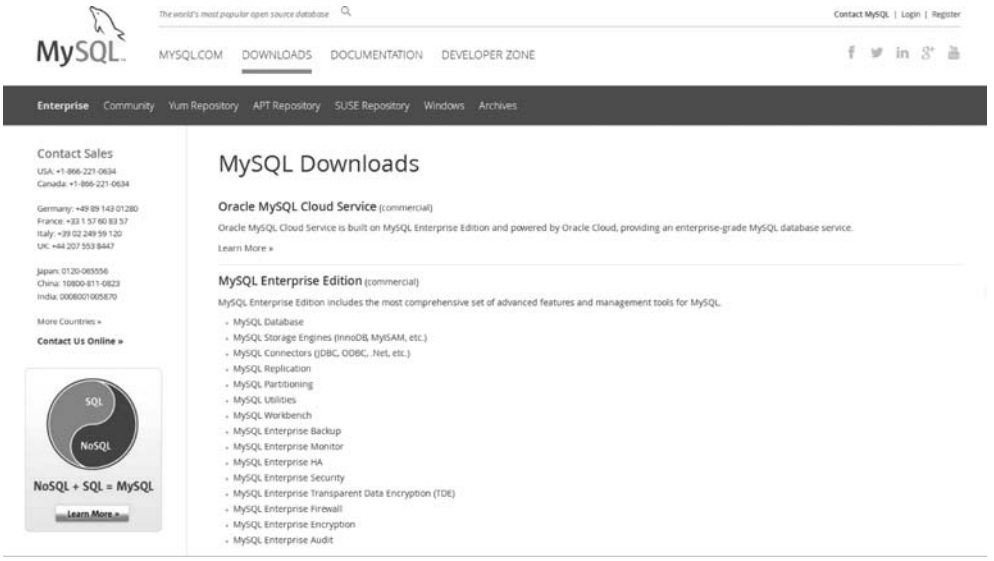



图 6-11　下载 MySQL

2）MySQL 的安装与使用

确保在当前系统中已经安装了 Microsoft.NET Framework 4.0，双击已经下载好的安装文件，即可将 MySQL 安装到本地计算机上。此外，在安装过程中，还需要设置 root 用户的密码，以便今后登录时使用。在本次安装中将该密码设置为空。

在本地计算机安装好 MySQL 后，在 Windows 命令行中输入 net start mysql 命令，即可启动该程序。要进入 MySQL 可执行程序目录，输入 mysql -u root 命令，即可进入 MySQL 的命令行模式，如图 6-12 所示。

要退出命令行模式，在提示符 mysql>后输入 quit 命令即可，如图 6-13 所示。


图 6-12 MySQL 的运行


图 6-13 MySQL 的退出

3）MySQL 中数据表的创建

在 MySQL 中新建数据库 test，在数据库 test 中新建数据表 user，并设置该表的字段名和字段类型，如图 6-14 所示。

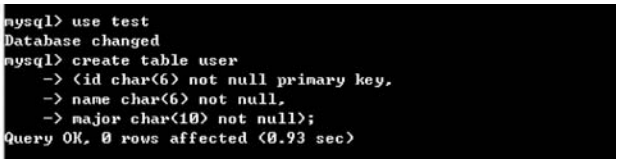



图 6-14 创建数据表 user

查看表结构，如图 6-15 所示。

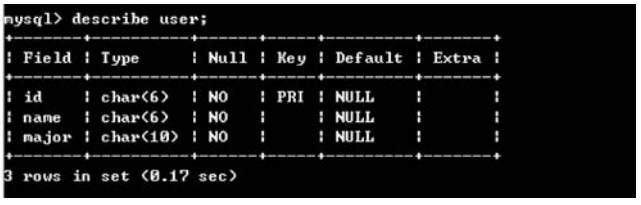



图 6-15 数据表结构

向数据表 user 中添加内容，如图 6-16 所示。

显示数据表 user 中的内容，如图 6-17 所示。

```
mysql> insert into user values('00001','join','computer');
Query OK, 1 row affected (0.23 sec)

mysql> insert into user values('00002','lily','computer');
Query OK, 1 row affected (0.12 sec)

mysql> insert into user values('00003','matt','computer');
Query OK, 1 row affected (0.18 sec)

mysql> insert into user values('00004','ben','computer');
Query OK, 1 row affected (0.09 sec)

mysql> insert into user values('00005','tony','computer');
Query OK, 1 row affected (0.10 sec)
```

图 6-16　向数据表 user 中添加内容

```
mysql> select * from user;
+-------+------+----------+
| id    | name | major    |
+-------+------+----------+
| 00001 | join | computer |
| 00002 | lily | computer |
| 00003 | matt | computer |
| 00004 | ben  | computer |
| 00005 | tony | computer |
+-------+------+----------+
5 rows in set (0.00 sec)
```

图 6-17　显示数据表 user 中的内容

### 6.3.2　Kettle 数据迁移

**【例 6-2】**　使用 Kettle 迁移数据表中所有记录。

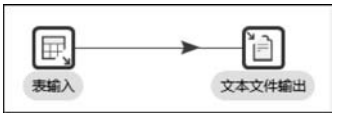



图 6-18　迁移数据表中所有记录工作流程

（1）启动 Kettle 后，新建转换，在“输入”列表中选择“表输入”步骤，在“输出”列表中选择“文本文件输出”步骤，分别拖动到右侧工作区中，并建立彼此之间的节点连接关系，如图 6-18 所示。

（2）双击“表输入”图标，在弹出的“表输入”对话框中单击“新建”按钮，该操作用于连接数据库，如图 6-19 所示。


图 6-19　新建数据库连接

（3）在弹出的“数据库连接”对话框中，输入连接名称为 hy，连接类型选择为 MySQL，连接方式为 Native (JDBC)，在设置中输入主机名称为 localhost，数据库名称为 test，端口号为 3306，用户名为 root，密码为空，如图 6-20 所示。最后单击“确认”按钮，以确保连接成功。


图 6-20　设置数据库连接

(4) 数据库连接成功后,双击“表输入”图标,如图 6-21 所示,输入 SQL 语句:

```
SELECT *
FROM user
```


图 6-21　输入 SQL 语句

该语句获取了 user 数据表中的所有数据。

(5) 双击“文本文件输出”图标,设置输出的文本文件目录和名称,如图 6-22 所示。


图 6-22　设置文本文件输出

(6) 保存该文件,运行转换,在执行结果区域的 Preview data 选项卡中预览迁移结果,如图 6-23 所示。

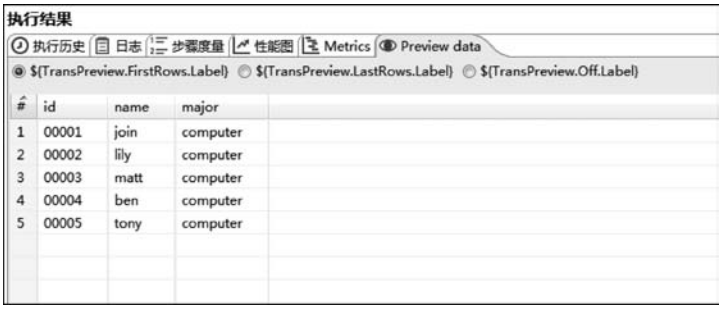



图 6-23　预览迁移结果

(7) 打开生成的 file 文件,可以看到 user 数据表中的数据已经被迁移到了文本文件中,如图 6-24 所示。


图 6-24　查看生成的文本文件

## 6.4 本章小结

(1) 数据迁移就是将数据从一个地方挪到另外一个地方,它是将很少使用或不用的文件移到辅助存储系统(如磁带或光盘)的存档过程。

(2) 数据迁移的实现可以分为 3 个阶段:数据迁移前的准备、数据迁移的实施和数据迁移后的校验。

(3) 数据转换与迁移通常包括多项工作:旧系统数据字典整理、旧系统数据质量分析、新系统数据字典整理、新旧系统数据差异分析、建立新旧系统数据之间的映射关系、开发部署数据转换与迁移程序、制订数据转换与迁移过程中的应急方案、实施旧系统数据到新系统的转换与迁移工作、检查转换与迁移后数据的完整性与正确性。

(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 ETL 工具和分布式并行处理实现。

## 6.5 实训

### 1. 实训目的

通过本章实训了解数据迁移的特点,能进行简单的与数据迁移有关的操作。

### 2. 实训内容

(1) 在 MySQL 中建立数据库 test,新建表 xs,在表 xs 中建立字段 xuehao、xingming、zhuanye、xingbie 和 chengji,将字段 xuehao 设置为主键,并输入数据,如图 6-25 和 6-26 所示。

```
+----------+-------------+------+-----+---------+-------+
| Field    | Type        | Null | Key | Default | Extra |
+----------+-------------+------+-----+---------+-------+
| xuehao   | char(6)     | NO   | PRI | NULL    |       |
| xingming | char(6)     | NO   |     | NULL    |       |
| zhuanye  | char(10)    | YES  |     | NULL    |       |
| xingbie  | tinyint(1)  | NO   |     | 1       |       |
| chengji  | tinyint(1)  | YES  |     | NULL    |       |
+----------+-------------+------+-----+---------+-------+
```

图 6-25 在 MySQL 中新建表并新建字段

```
+--------+----------+----------+----------+----------+
| xuehao | xingming | zhuanye  | xingbie  | chengji  |
+--------+----------+----------+----------+----------+
| 001    | cheng    | jisuanji |        1 |       85 |
| 002    | leslie   | jisuanji |        1 |       99 |
| 003    | tom      | jisuanji |        1 |       71 |
+--------+----------+----------+----------+----------+
```

图 6-26 在 MySQL 中新建表并输入数据

(2) 运行 Kettle,新建转换,将“表输入”和“文本文件输出”步骤拖到工作区中,并建立连接。

（3）双击“表输入”图标，在弹出的对话框中单击“编辑”按钮，建立 Kettle 与 MySQL 数据库的连接，设置完成以后可以单击“测试”按钮，查看连接状况，如图 6-27 所示。


图 6-27　连接数据库并测试

值得注意的是，如果显示无法建立连接，有可能是没有安装对应的数据库链接驱动，需要去官网下载 mysql-connector-java-5. 1. 46-bin. jar 文件（对应不同版本有不同的文件）。

（4）在“表输入”对话框中的 SQL 输入框中输入查询语句：SELECT xingming FROM xs WHERE chengji＞80，查找成绩大于 80 分的学生姓名，并单击“确定”按钮，如图 6-28 所示。


图 6-28　输入 SQL 语句

(5) 保存该文件,运行转换。右击“文本文件输出”图标,在弹出的快捷菜单中选择Preview,在弹出的对话框中选择“文本文件输出”选项,并单击“快读启动”按钮,即可查看运行结果,如图 6-29 所示。

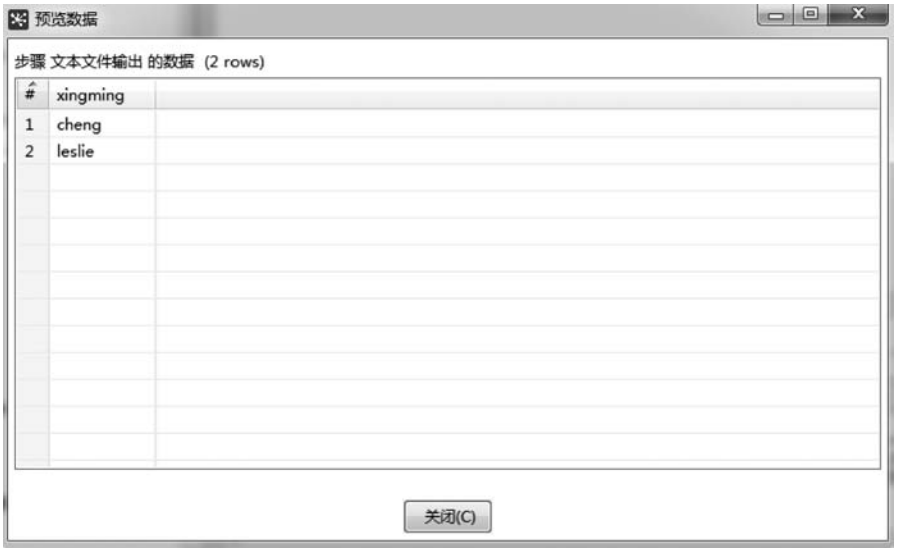



图 6-29 查看输出结果

## 习题 6

(1) 什么是数据迁移?
(2) 数据迁移有哪些过程?
(3) 数据迁移有哪些方法?
(4) 如何使用 Kettle 实现数据迁移?