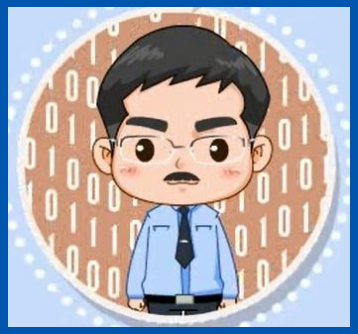


《数据采集与预处理》

教材官网： http://dblab.xmu.edu.cn/post/data-collection/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第7章 ETL工具Kettle

（PPT版本号：2022年1月版本）

林子雨 副教授

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页：**http://dblab.xmu.edu.cn/linziyu**

提纲

7.1 Kettle的基本概念
7.2 Kettle的基本功能
7.3 安装Kettle
7.4 数据抽取
7.5 数据清洗与转换
7.6 数据加载

本PPT是以下教材的配套讲义
林子雨编著《数据采集与预处理》
人民邮电出版社

教材官网：
http://dblab.xmu.edu.cn/post/data-collection

7.1 Kettle的基本概念

一个数据抽取过程（如图7-1所示）主要包括创建一个作业（Job），每个作业由一个或多个作业项（Job Entry）和连接作业项的作业跳（Job Hop）组成。每个作业项可以是一个转换（Transformation）或是另一个作业。一个转换由一个或多个步骤（Step）和连接步骤的跳（Hop）组成。

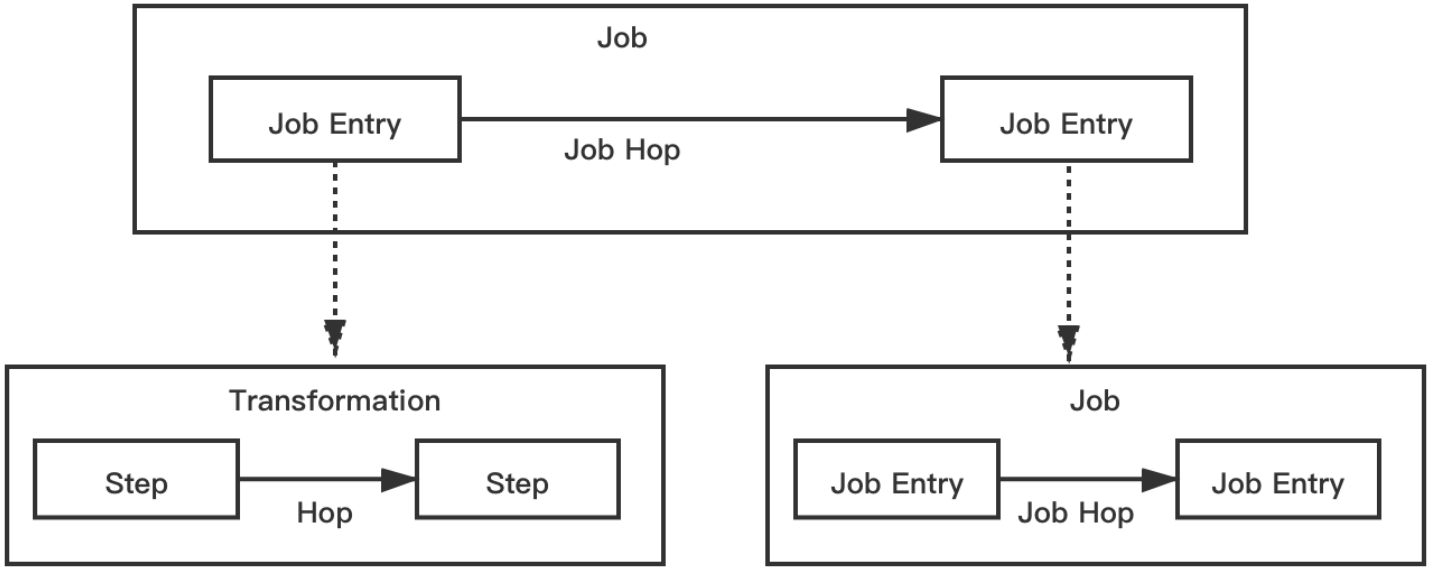


图7-1 一个数据抽取过程的构成要素

7.1 Kettle的基本概念

转换主要用于数据的抽取（Extraction）、转换（Transformation）以及加载（Load），比如读取文件、过滤输出行、数据清洗或加载到数据库等步骤。一个转换包含一个或多个步骤，每个步骤都是单独的线程，当启动转换时，所有步骤的线程几乎并行执行。步骤之间的数据以数据流方式传递。所有的步骤都会从它们的输入跳中读取数据，并把处理过的数据写到输出跳，直到输入跳里不再有数据就终止步骤的运行；当所有步骤都终止了，整个转换就终止了。由于转换里的步骤依赖前一个步骤获取数据，因此转换里不能有循环。

7.1 Kettle的基本概念

相较于转换，作业是更加高级的操作。作业由一个或多个作业项（作业或转换）组成。所有的作业项是以某种自定义的顺序串行执行的。作业项之间可以传递一个包含了数据行的结果对象。当一个作业项执行完成后，再传递结果对象给下一个作业项。作业里可以有循环。

跳是步骤之间带箭头的连接线，它定义了一个单向通道，用于连接两个步骤，实现将数据从一个步骤（写入数据到行集）流向另一个步骤（从行集中读取数据）。跳是两个步骤之间的被称为“行集”（Row Set）的数据行缓存（可以在转换设置中定义行集大小）。若行集满了，则向行集写数据的步骤将停止写入，直到行集里又有空间。若行集空了，则从行集读取数据的步骤就会停止读取，直到行集里又有可读取的数据行。跳对于向行集写入数据的步骤来说是输出跳，一个步骤可以拥有多个输出跳；跳对于从行集中读取数据的步骤来说是输入跳。

作业跳是作业项之间带箭头的连接线，它定义了作业的执行路径。

7.2 Kettle的基本功能

Kettle的基本功能包括转换管理和作业管理。转换管理主要包括输入、输出、转换、应用、流程、脚本、查询、检验、作业、映射和批量加载等功能。作业管理主要包括通用、邮件、文件管理、条件、脚本、批量加载等功能。

7.3 安装Kettle

在Windows系统中打开浏览器，访问Kettle官网（https://sourceforge.net/projects/pentaho/），下载Kettle安装文件pdi-ce-9.1.0.0-324.zip。或者，也可以直接到教材官网的“下载专区”的“软件”目录中下载pdi-ce-9.1.0.0-324.zip文件。

把pdi-ce-9.1.0.0-324.zip解压缩到“D:\”目录下（或者也可以选择一个其他目录，比如“C:\”），会生成一个“data-integration”目录，该目录下就包含了Kettle。在data-integration目录里包含了Spoon的启动文件，即spoon.bat，双击该文件就可以启动Spoon，启动界面如图7-2所示。

7.3 安装Kettle

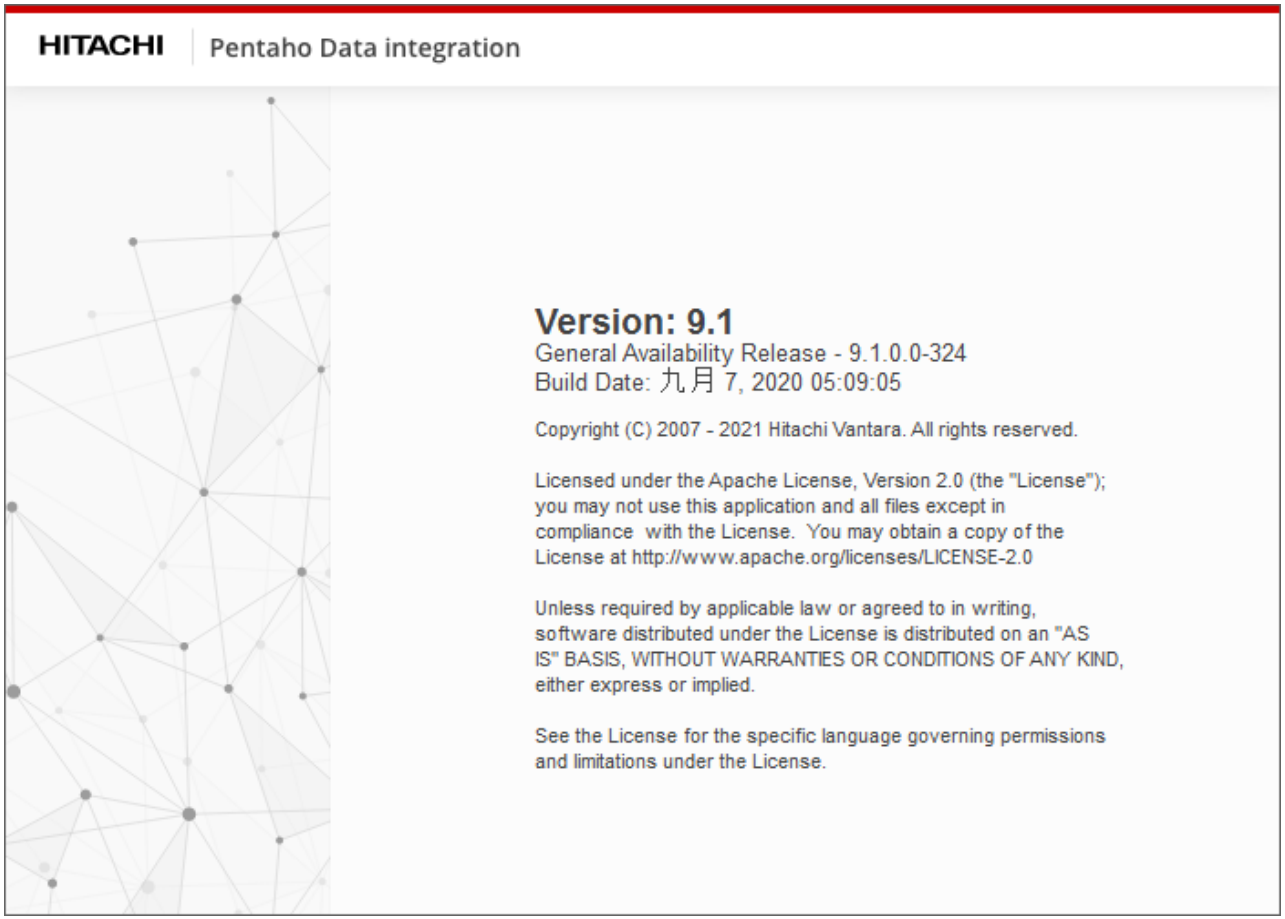


图7-2 Spoon启动界面

7.3 安装Kettle

启动成功以后的界面如图7-3所示。

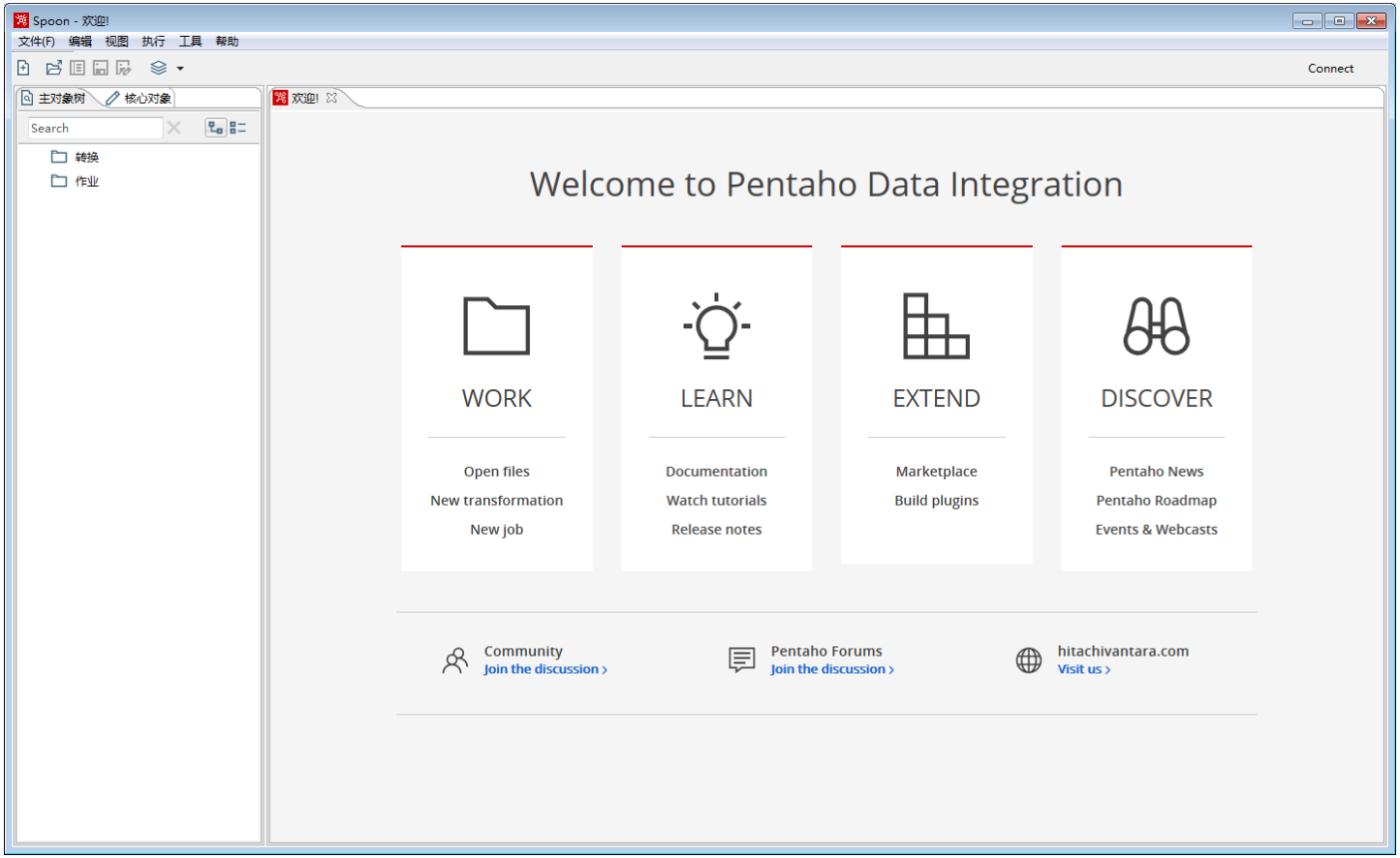


图7-3 Spoon启动以后的欢迎界面

7.4 数据抽取

7.4.1 把文本文件导入到Excel文件中
7.4.2 把文本文件导入MySQL数据库中（请直接参考教材）
7.4.3 把Excel文件导入到MySQL数据库中（请直接参考教材）

7.4.1 把文本文件导入到Excel文件中

这里给出一个实例，演示如何使用Kettle把文本文件导入到Excel文件中，具体包括如下步骤：

- 创建文本文件；
- 建立转换；
- 设计转换；
- 执行转换。

7.4.1 把文本文件导入到Excel文件中

1.创建文本文件

在“D:\”目录下新建一个文本文件studentinfo.txt，其内容如图7-4所示，文件的第1行是字段名称，包括sno、name、sex和age，字段之间用“|”隔开，其余行都是记录，字段之间也是用“|”隔开。

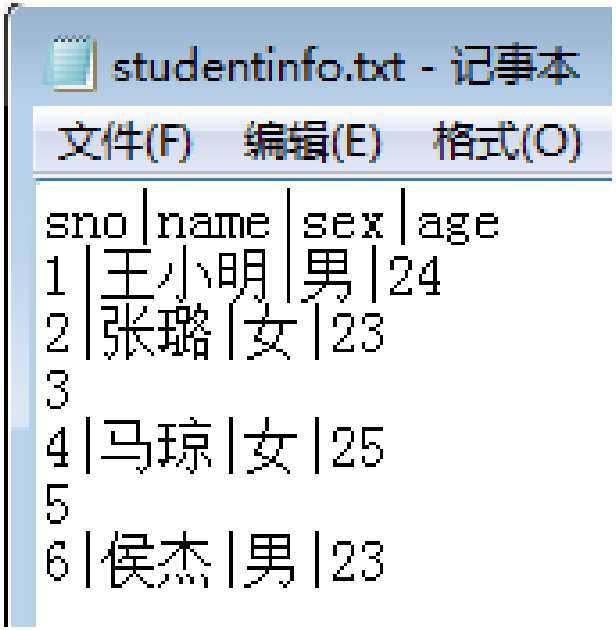


图7-4 studentinfo.txt文件内容

7.4.1 把文本文件导入到Excel文件中

2.建立转换

在Spoon主界面的“主对象树”栏目中，在“转换”上面（如图7-5所示）单击鼠标右键，在弹出的菜单中点击“新建”。点击Spoon主界面左上角的“保存”图标，把这个转换保存到某个路径下并且名称为“text_to_excel”。

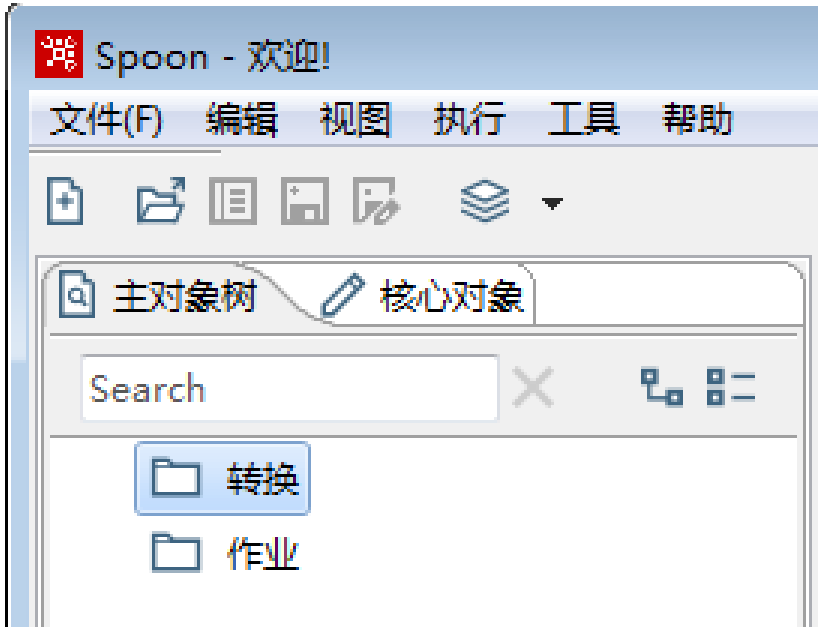


图7-5 新建“转换”

7.4.1 把文本文件导入到Excel文件中

3.设计转换

在“核心对象”栏目中，在“输入”控件里把“文本文件输入”拖到右侧设计区域，然后在“输出”控件里把“Excel输出”拖到右侧设计区域，然后为这两个控件建立连线（如图7-6所示），这里的连线就是前文介绍过的“跳”。为这两个控件建立连线的方法是，按住键盘上的Shift键，然后用鼠标左键单击“文本文件输入”控件图标，再用鼠标左键单击“Excel输出”控件图标，最后在其他空白区域单击鼠标左键，这样就建立了一条从“文本文件输入”到“Excel输出”的连线。

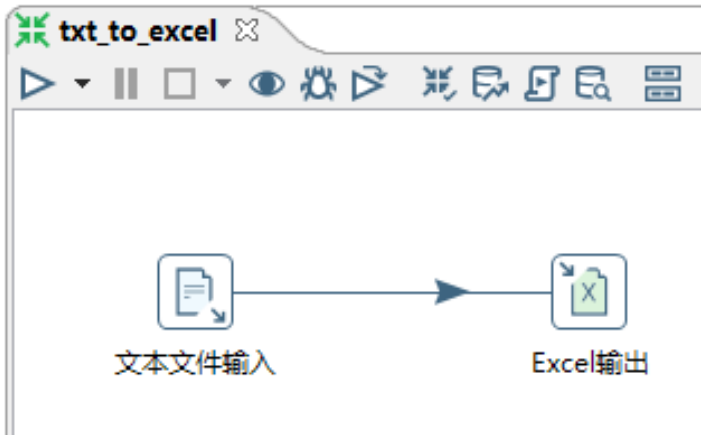


图7-6 放置文本文件输入和Excel输出两个控件

7.4.1 把文本文件导入到Excel文件中

双击设计区域的“文本文件输入”控件，打开设置界面，点击“文件”选项卡，点击“文件或目录”右侧的“浏览”按钮（如图7-7所示），把studentinfo.txt文件添加进来，然后点击“增加”按钮，studentinfo.txt文件就会被增加到“选中的文件”中，增加后的效果如图7-8所示。

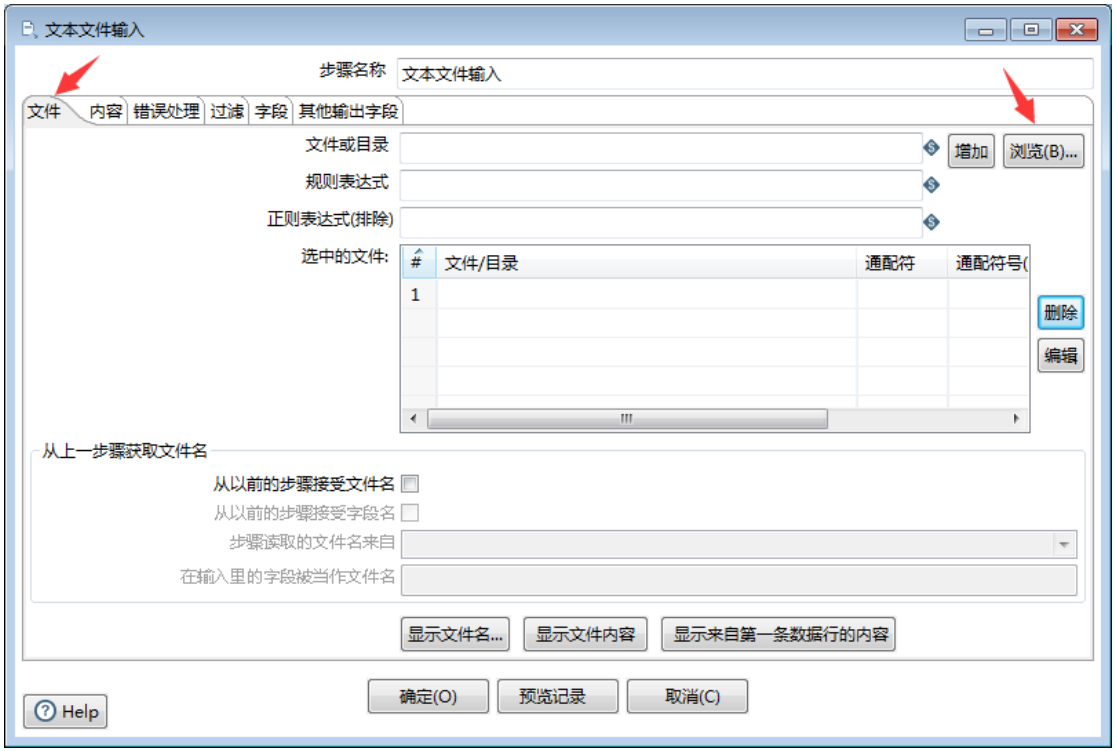


图7-7 添加文件

7.4.1 把文本文件导入到Excel文件中

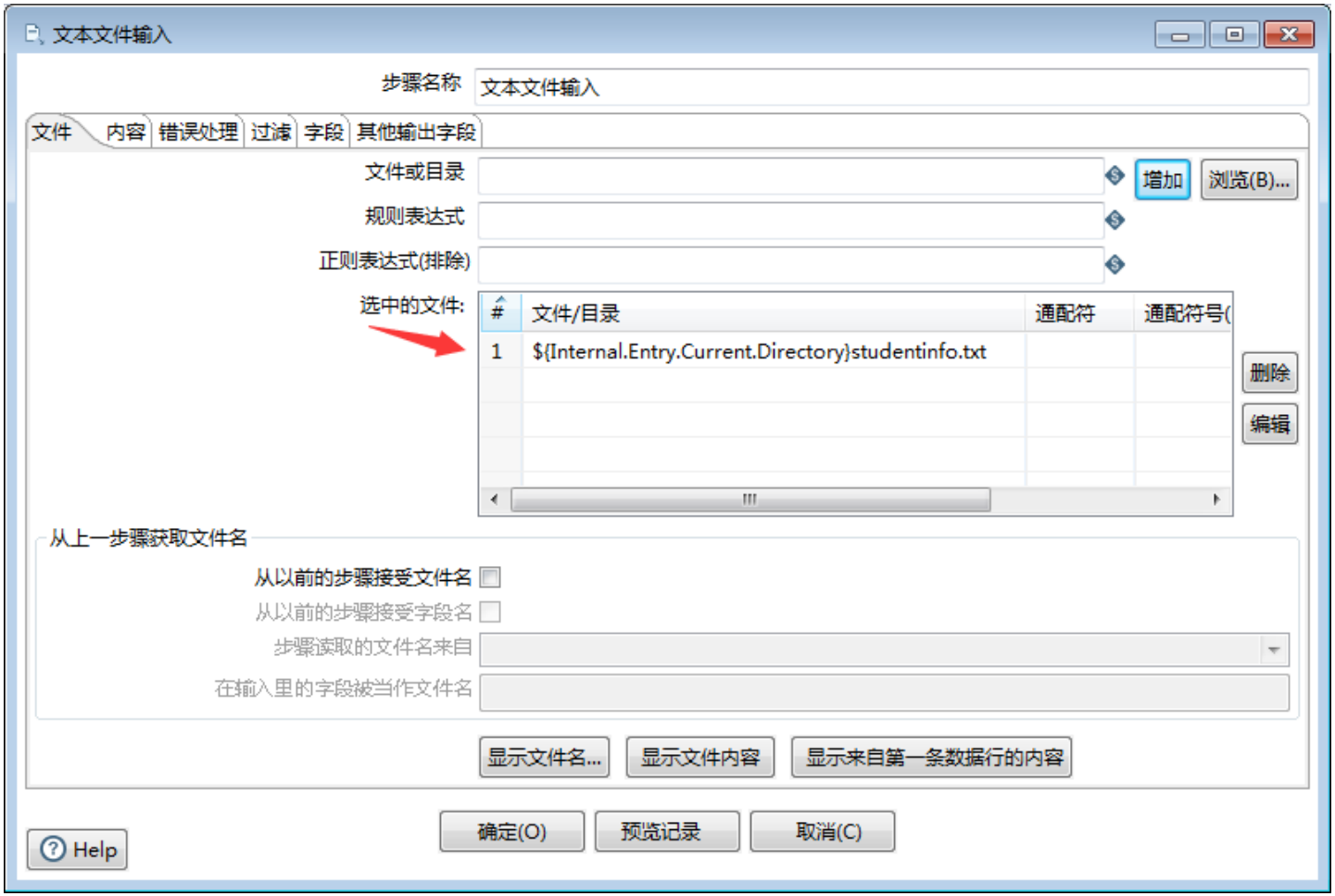


图7-8 添加文件以后的效果

7.4.1 把文本文件导入到Excel文件中

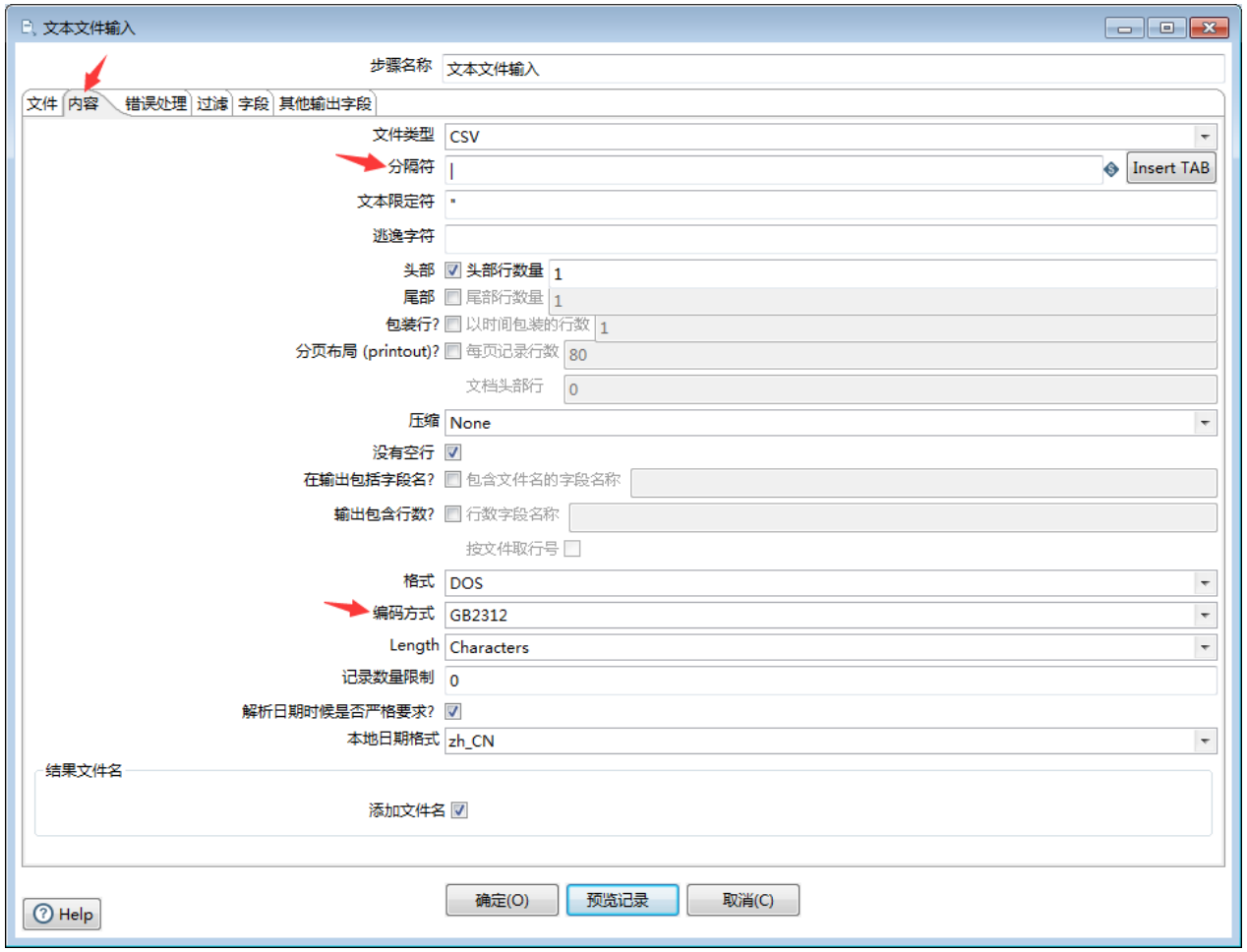


图7-9 设置“内容”选项卡

7.4.1 把文本文件导入到Excel文件中

在“字段”选项卡中（如图7-10所示），点击“获取字段”按钮，会弹出如图7-11所示的样本数据行数设置界面，直接点击“确定”按钮，会得到如图7-12所示结果。这时，点击界面底部的“预览记录”，就可以看到如图7-13所示的数据。最后，点击界面底部的“确定”按钮，完成文本文件输入控件的设置。

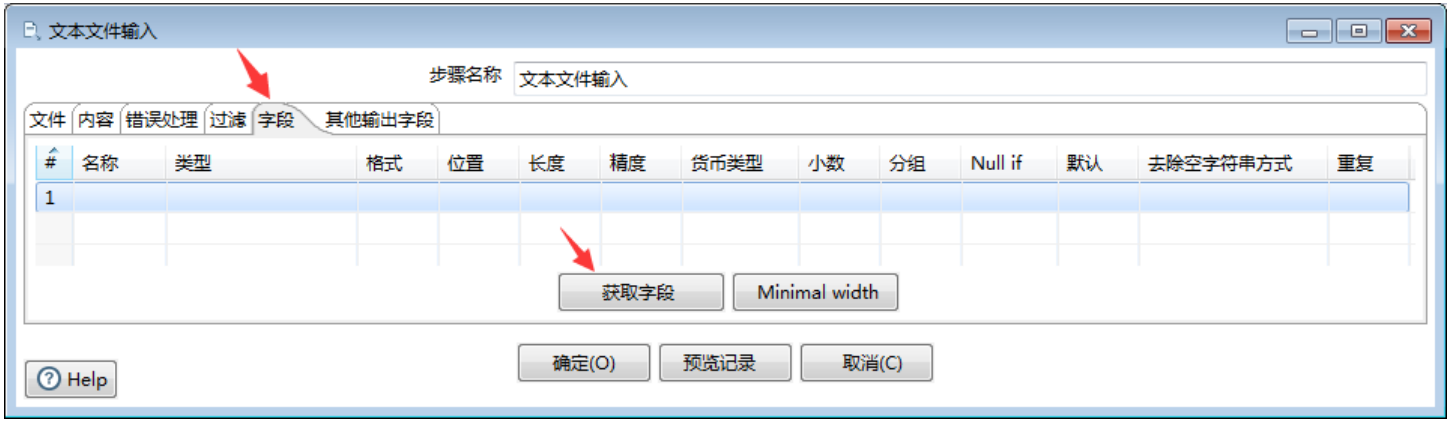


图7-10 设置“字段”选项卡

7.4.1 把文本文件导入到Excel文件中

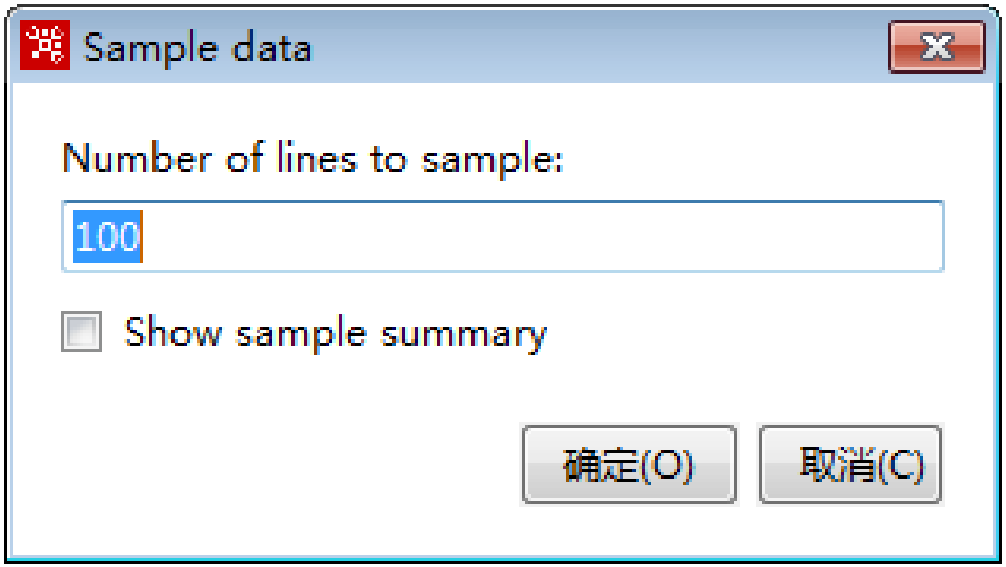


图7-11 设置样本数据行数

7.4.1 把文本文件导入到Excel文件中

图7-12 获取字段以后的效果

7.4.1 把文本文件导入到Excel文件中

预览数据

步骤 文本文件输入 的数据 (6 rows)

#	sno	name	sex	age
1	1	王小明	男	24
2	2	张璐	女	23
3	3	<null>	<null>	<null>
4	4	马琼	女	25
5	5	<null>	<null>	<null>
6	6	侯杰	男	23

图7-13 预览记录

7.4.1 把文本文件导入到Excel文件中

双击设计区域的“Excel输出”控件图标，打开设置界面（如图7-14所示），在“文件”选项卡中，设置“文件名”为“D:\file”。

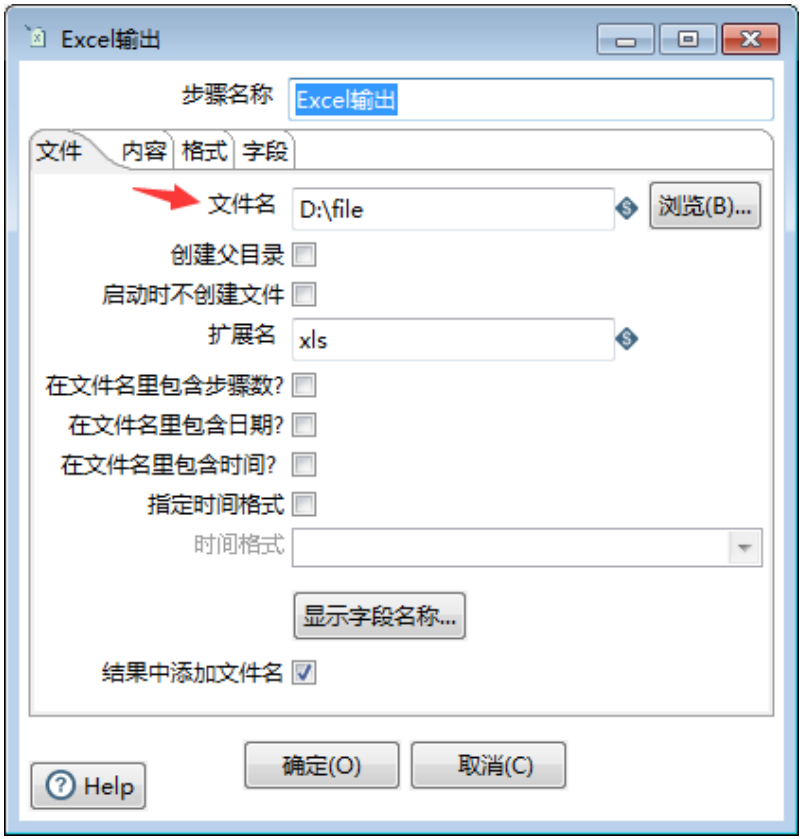


图7-14 设置文件名

7.4.1 把文本文件导入到Excel文件中

在“字段”选项卡中（如图7-15所示），点击界面底部的“获取字段”按钮，成功获取字段以后的效果如图7-16所示，然后把“sno”和“age”字段的“格式”设置为“#”。最后，点击“确定”按钮完成“Excel输出”控件的设置。全部设置完成以后，需要保存设计文件。

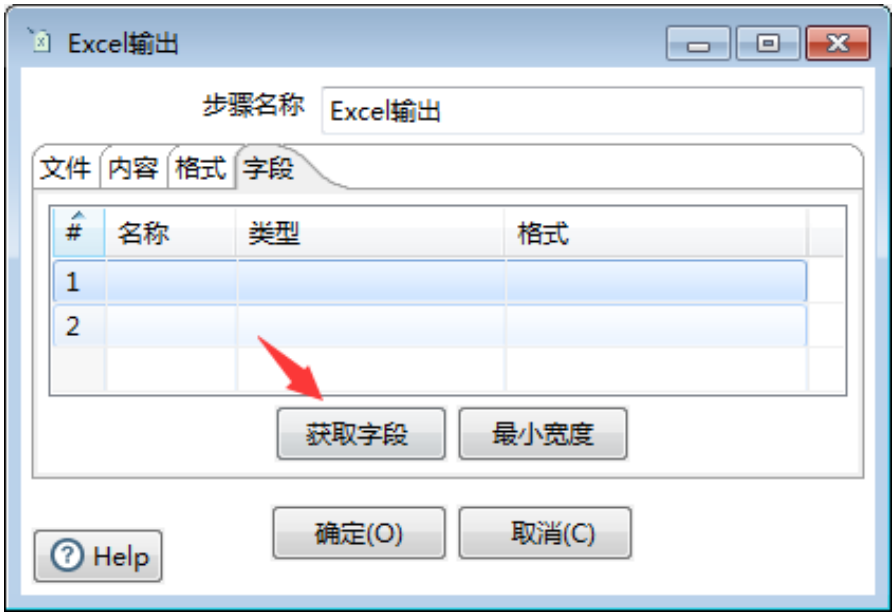


图7-15 “字段”选项卡

7.4.1 把文本文件导入到Excel文件中

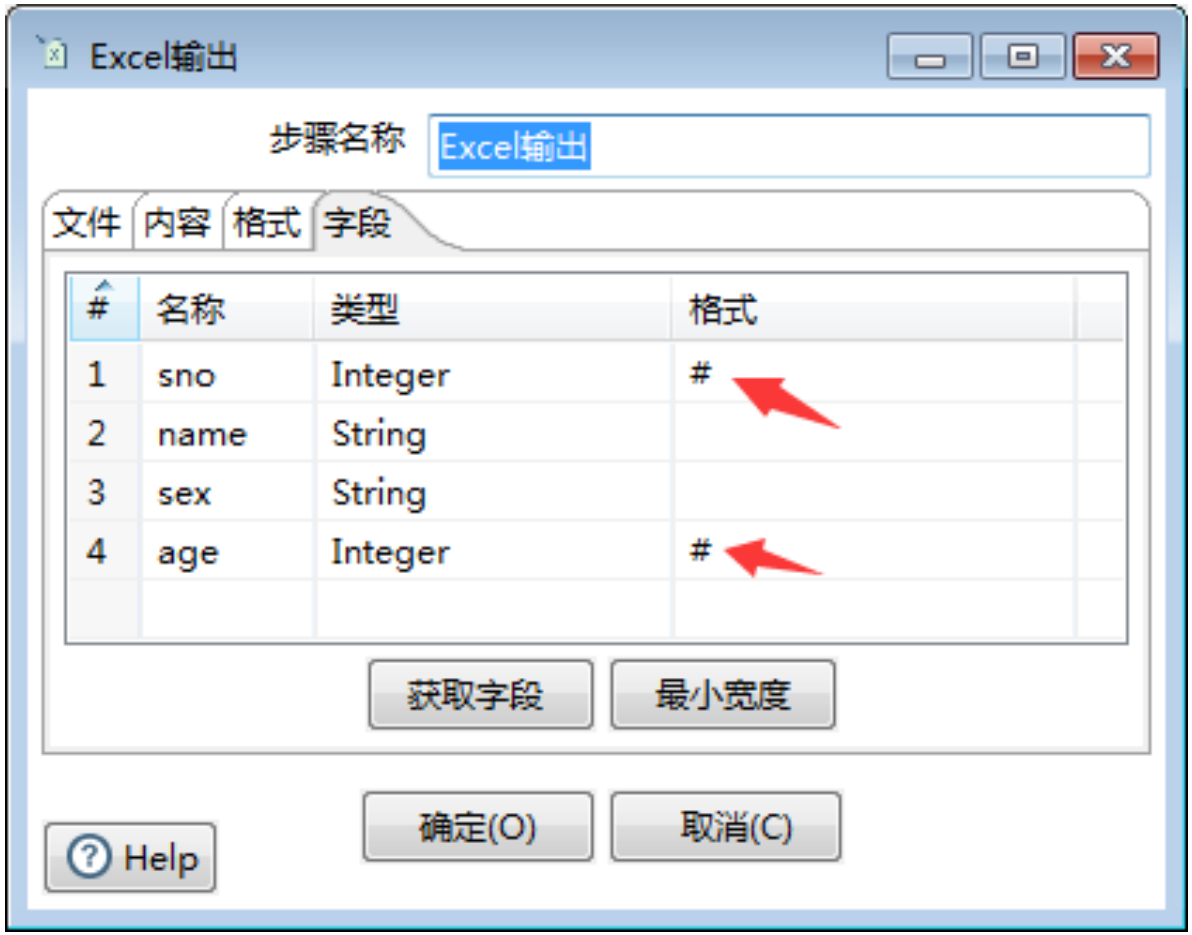


图7-16 获取字段后的效果

7.4.1 把文本文件导入到Excel文件中

4. 执行转换

在转换设计界面中（如图7-17所示），点击三角形按钮开始执行转换，会弹出如图7-18所示界面，在界面中点击“启动”，如果转换执行成功，会显示如图7-19所示的效果，在两个控件图标上都会显示绿色的勾号。这时，到D盘根目录下就可以看到新生成的文件file.xls，可以使用Excel软件打开file.xls查看内容（如图7-20所示）。

图7-17 运行转换

7.4.1 把文本文件导入到Excel文件中

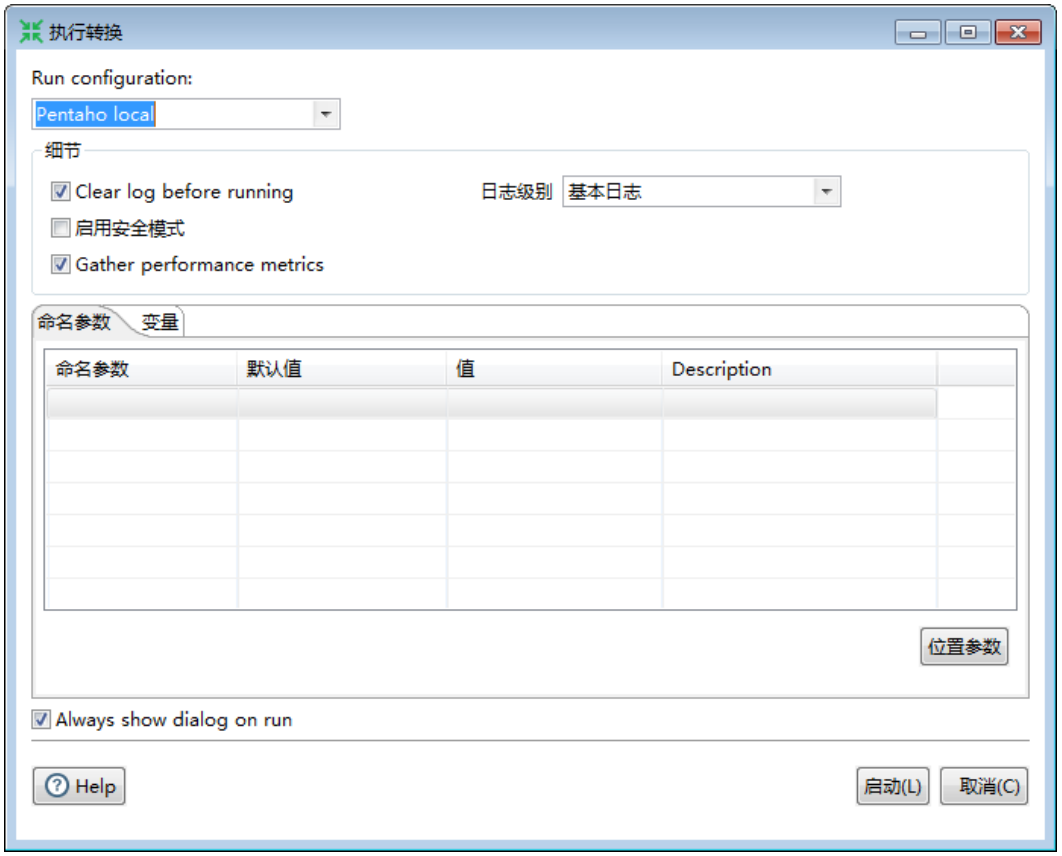


图7-18 转换启动界面

7.4.1 把文本文件导入到Excel文件中

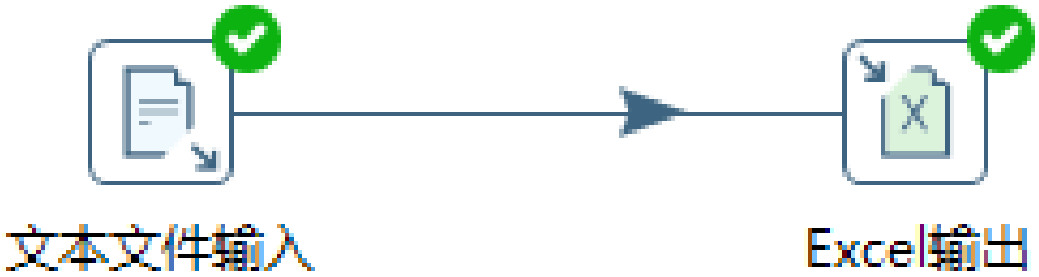
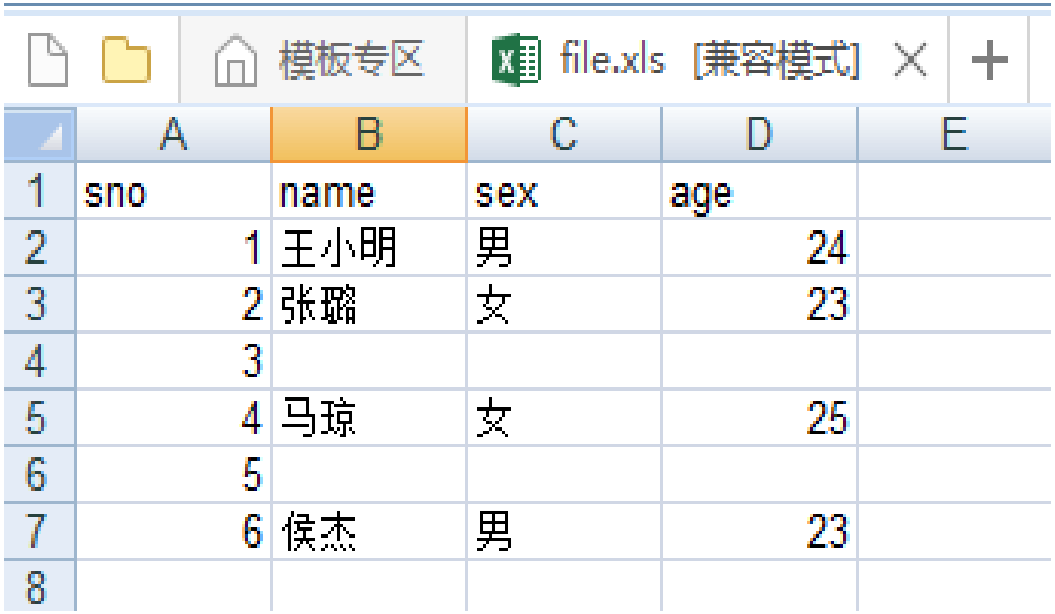


图7-19 转换执行成功的效果

7.4.1 把文本文件导入到Excel文件中



	A	B	C	D	E
1	sno	name	sex	age	
2	1	王小明	男	24	
3	2	张璐	女	23	
4	3				
5	4	马琼	女	25	
6	5				
7	6	侯杰	男	23	
8					

图7-20 file.xls文件内容

7.5 数据清洗与转换

7.5.1 使用Kettle实现数据排序
7.5.2 在Kettle中用正则表达式清洗数据（请直接参考教材）
7.5.3 使用Kettle去除缺失值（请直接参考教材）
7.5.4 使用Kettle转化MySQL数据库中的数据（请直接参考教材）

7.5.1 使用Kettle实现数据排序

这里给出一个实例，演示如何使用Kettle实现数据排序，具体包括如下步骤：

- 创建文本文件；
- 建立转换；
- 设计转换；
- 执行转换。

7.5.1 使用Kettle实现数据排序

1.创建文本文件

在“D:\”目录下新建一个文本文件score.txt，其内容如图7-51所示，文件的第1行是字段名称，包括name和score，字段之间用分号隔开，其余行都是记录，字段之间也是用分号隔开。

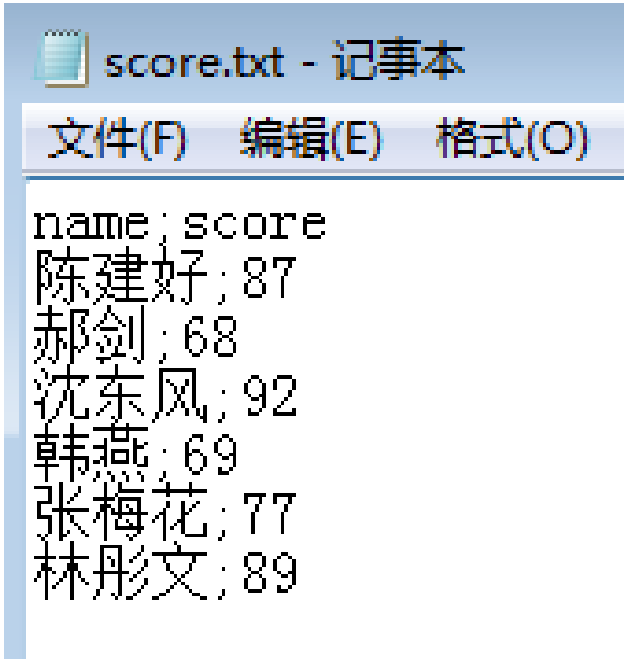


图7-51 score.txt文件内容

7.5.1 使用Kettle实现数据排序

2.建立转换

在Spoon主界面的“主对象树”栏目中，在“转换”上面（如图7-52所示）单击鼠标右键，在弹出的菜单中点击“新建”。点击Spoon主界面左上角的“保存”图标，把这个转换保存到某个路径下并且名称为“sort_data”。

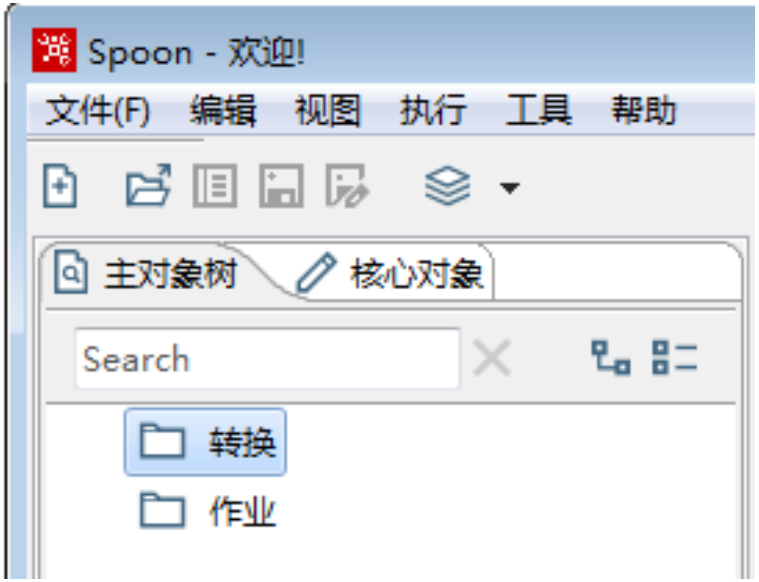


图7-52 新建“转换”

7.5.1 使用Kettle实现数据排序

3.设计转换

在“核心对象”栏目中，在“输入”控件里把“文本文件输入”拖到右侧设计区域，然后在“转换”控件里把“排序记录”拖到右侧设计区域，然后为这两个控件建立连线（如图7-53所示）。

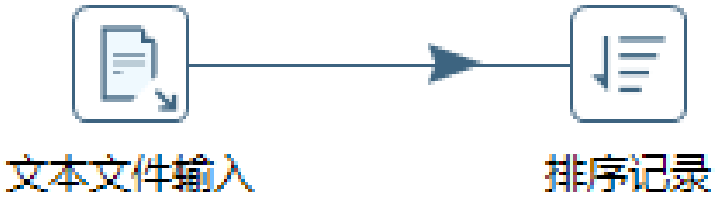


图7-53 放置文本文件输入和排序记录两个控件

7.5.1 使用Kettle实现数据排序

双击设计区域的“文本文件输入”控件图标，打开设置界面（如图7-54所示），点击“文件或目录”右侧的“浏览”按钮，添加文件“D:\score.txt”，然后，点击“增加”按钮，执行效果如图7-55所示。

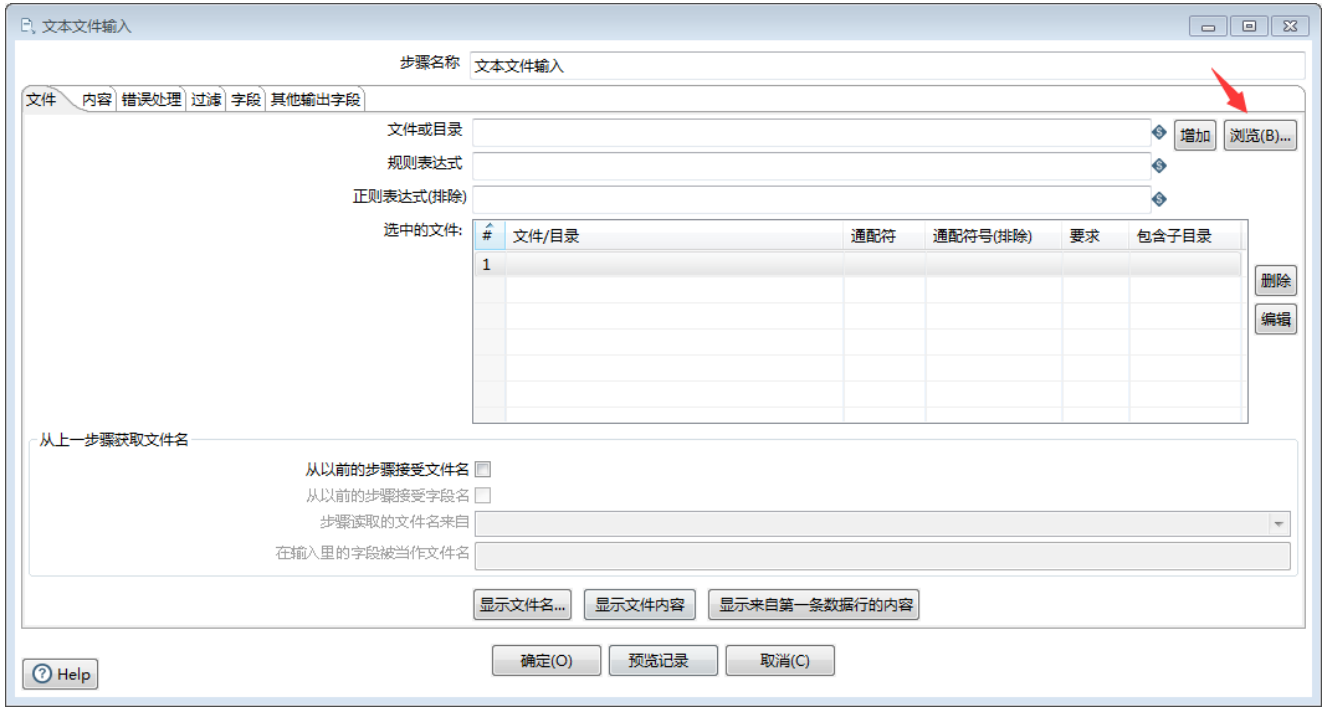


图7-54 添加文件

7.5.1 使用Kettle实现数据排序

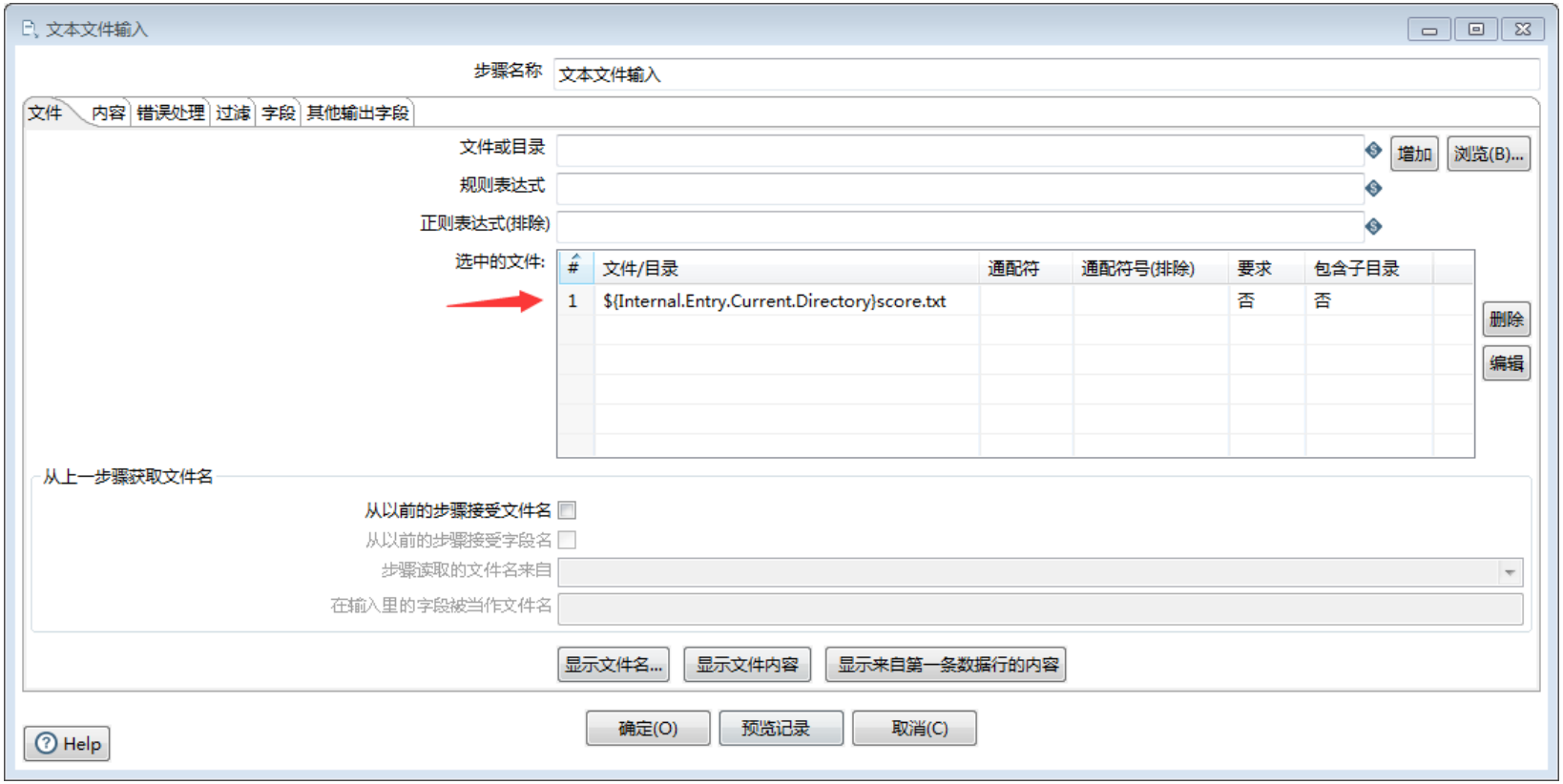


图7-55 添加文件以后的效果

7.5.1 使用Kettle实现数据排序

在“内容”选项卡中，设置分隔符为分号“;”（如图7-56所示）。

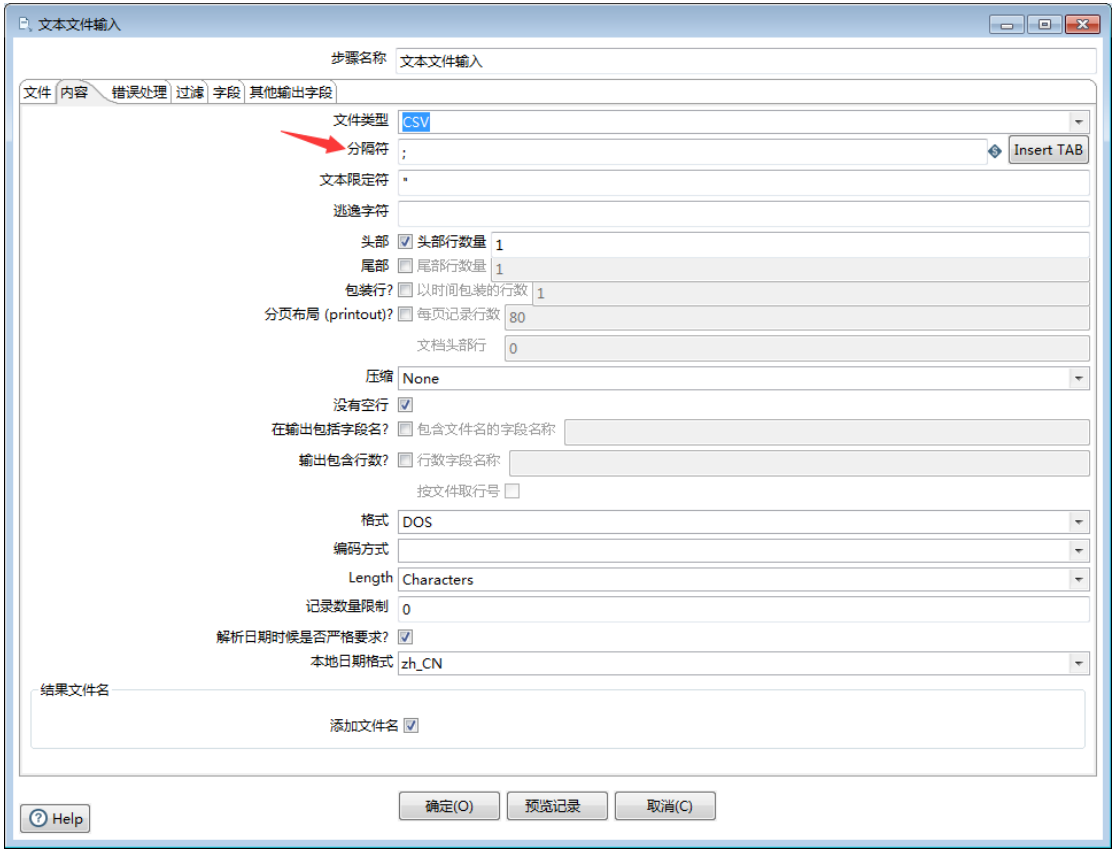


图7-56 设置“内容”选项卡

7.5.1 使用Kettle实现数据排序

在“字段”选项卡中（如图7-57所示），点击“获取字段”按钮，成功获取字段以后的效果如图7-58所示。

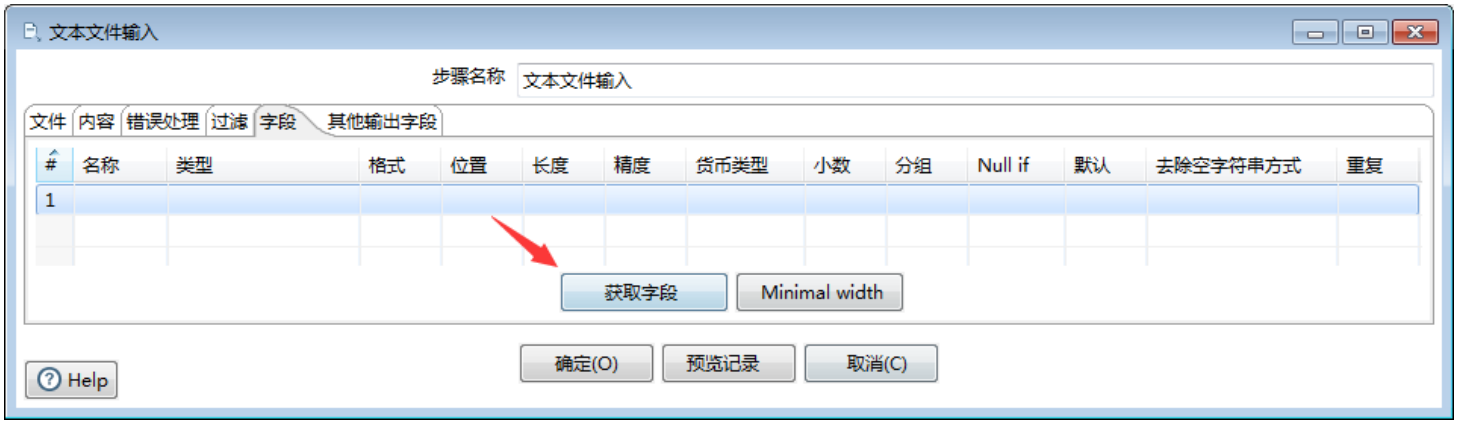


图7-57 获取字段

图7-58 获取字段成功以后的效果

7.5.1 使用Kettle实现数据排序

这时，点击界面（如图7-58所示）底部的“预览记录”按钮，就可以预览数据（如图7-59所示）。最后，点击界面底部的“确定”按钮，完成“文本文件输入”控件的设置。

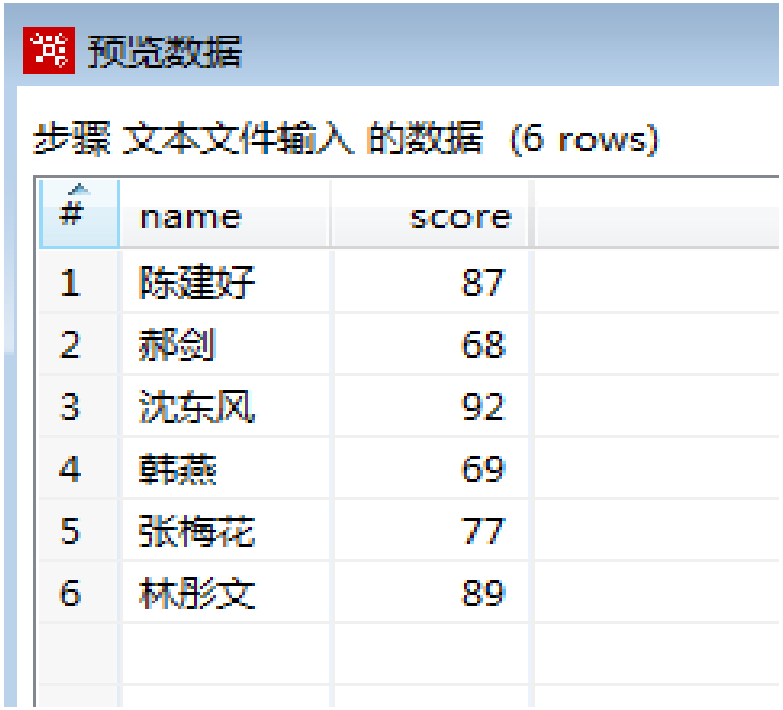


图7-59 预览数据

7.5.1 使用Kettle实现数据排序

双击设计区域的“排序记录”控件图标，打开设置界面（如图7-60所示），在“字段名称”下拉列表中选择“score”，在“升序”下拉列表中选择“是”，然后点击“确定”按钮完成设置。全部设置完成以后，需要保存设计文件。

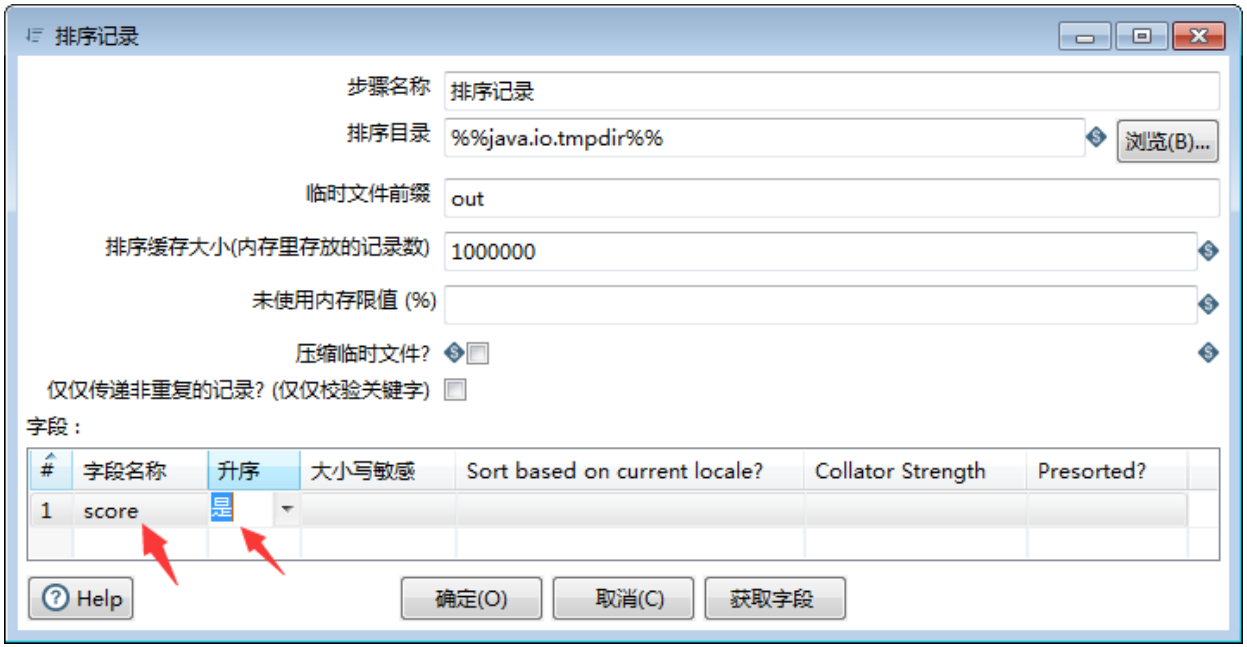


图7-60 排序记录设置界面

7.5.1 使用Kettle实现数据排序

4.执行转换

在转换设计界面中（如图7-61所示），点击三角形按钮开始执行转换，在弹出的界面中点击“启动”，如果转换执行成功，会显示如图7-62所示的效果，在两个控件图标上都会显示绿色的勾号。这时，在“执行结果”的“Preview data”选项卡中就可以预览排序后的数据（如图7-63所示）。

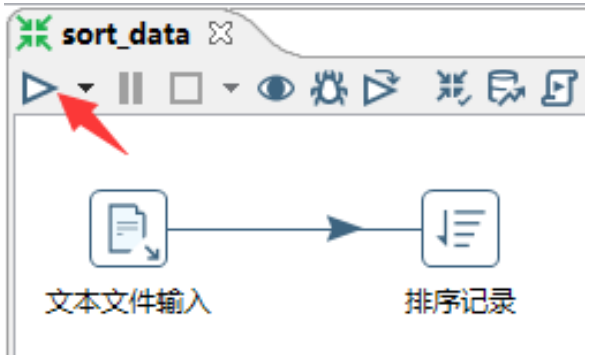
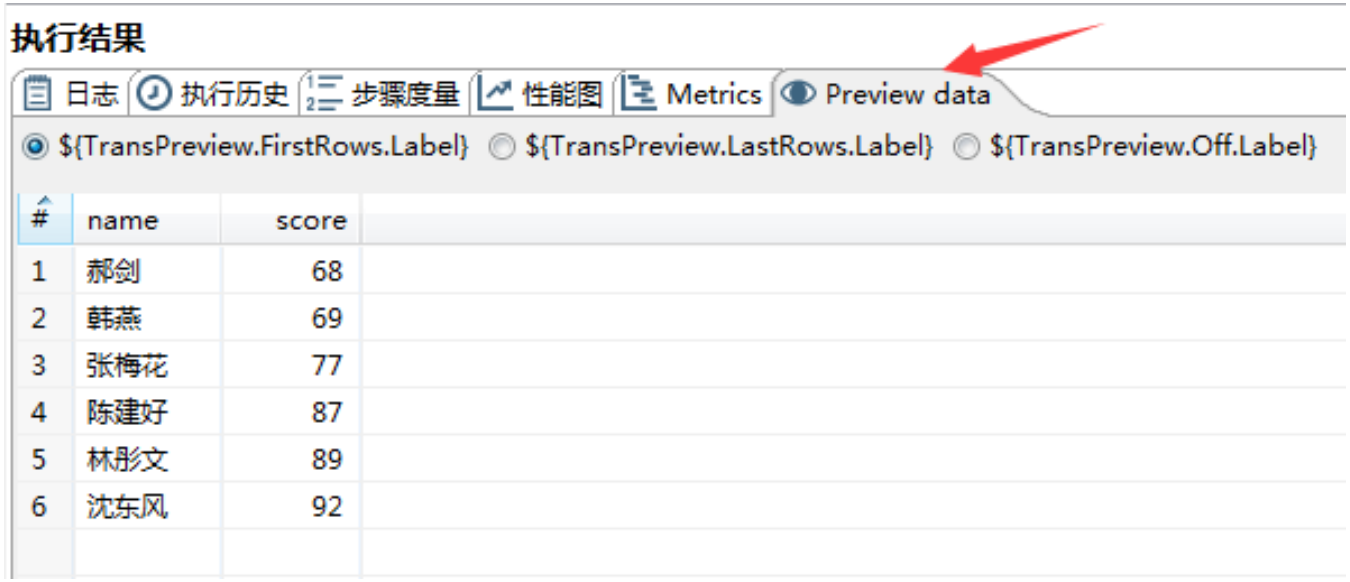


图7-61 运行转换

图7-62 转换执行成功

7.5.1 使用Kettle实现数据排序



#	name	score
1	郝剑	68
2	韩燕	69
3	张梅花	77
4	陈建好	87
5	林彤文	89
6	沈东风	92

图7-63 排序后的数据

7.6 数据加载

7.6.1把本地文件加载到HDFS中（请直接参考教材）
7.6.2把HDFS文件加载到MySQL数据库中

7.6.2把HDFS文件加载到MySQL数据库中

这里给出一个实例，演示如何使用Kettle把HDFS文件导入到MySQL数据库中，具体包括如下步骤：

- 新建HDFS文件；
- 创建数据库；
- 建立转换；
- 创建MySQL连接和Hadoop连接；
- 设计转换；
- 执行转换。

7.6.2把HDFS文件加载到MySQL数据库中

1. 新建HDFS文件

在Windows系统中打开一个cmd窗口，启动Hadoop。在“D:\”目录下新建一个文本文件student.txt，其内容如图7-140所示，文件的第1行是字段名称，包括no、name、sex和age，字段之间用“|”隔开，其余行都是记录，字段之间也是用“|”隔开。

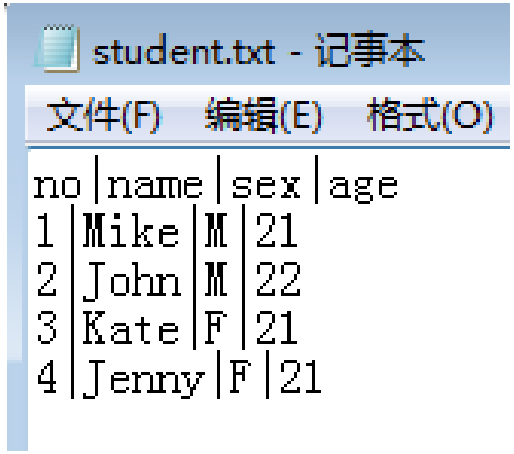


图7-140 student.txt文件内容

7.6.2把HDFS文件加载到MySQL数据库中

在cmd窗口中执行如下命令，把本地文件student.txt上传到HDFS系统的根目录下：

> cd c:\hadoop-3.1.3\bin

> hadoop fs -put D:\book.txt hdfs://localhost:9000/

可以继续执行如下命令查看HDFS中student.txt的内容：

> hadoop fs -cat hdfs://localhost:9000/student.txt

或者，也可以打开浏览器，访问“http://localhost:9870”，使用HDFS的WEB管理界面查看文件内容。

7.6.2把HDFS文件加载到MySQL数据库中

2.创建数据库

在Windows系统中启动MySQL服务，打开MySQL命令行客户端，执行如下SQL语句创建数据库：

```
CREATE DATABASE kettle;
```

继续执行如下SQL语句创建student_table表：

```
USE kettle;
#------------创建表student_table
DROP TABLE IF EXISTS student_table;
CREATE TABLE student_table (
no int,
name VARCHAR(10),
sex VARCHAR(2),
age int
);
```

7.6.2把HDFS文件加载到MySQL数据库中

3.建立转换

在Spoon主界面的“主对象树”栏目中，在“转换”上面（如图7-141所示）单击鼠标右键，在弹出的菜单中点击“新建”。点击Spoon主界面左上角的“保存”图标，把这个转换保存到某个路径下并且名称为“hdfs_to_mysql”。

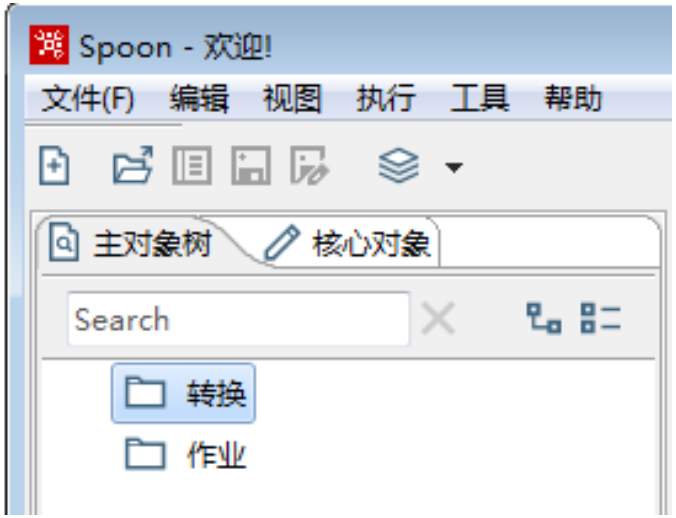


图7-141 新建“转换”

7.6.2把HDFS文件加载到MySQL数据库中

3. 创建MySQL连接和Hadoop连接

参照本章7.4.2节的内容，建立一个名称为“mysql”的数据库连接（如图7-142所示）。

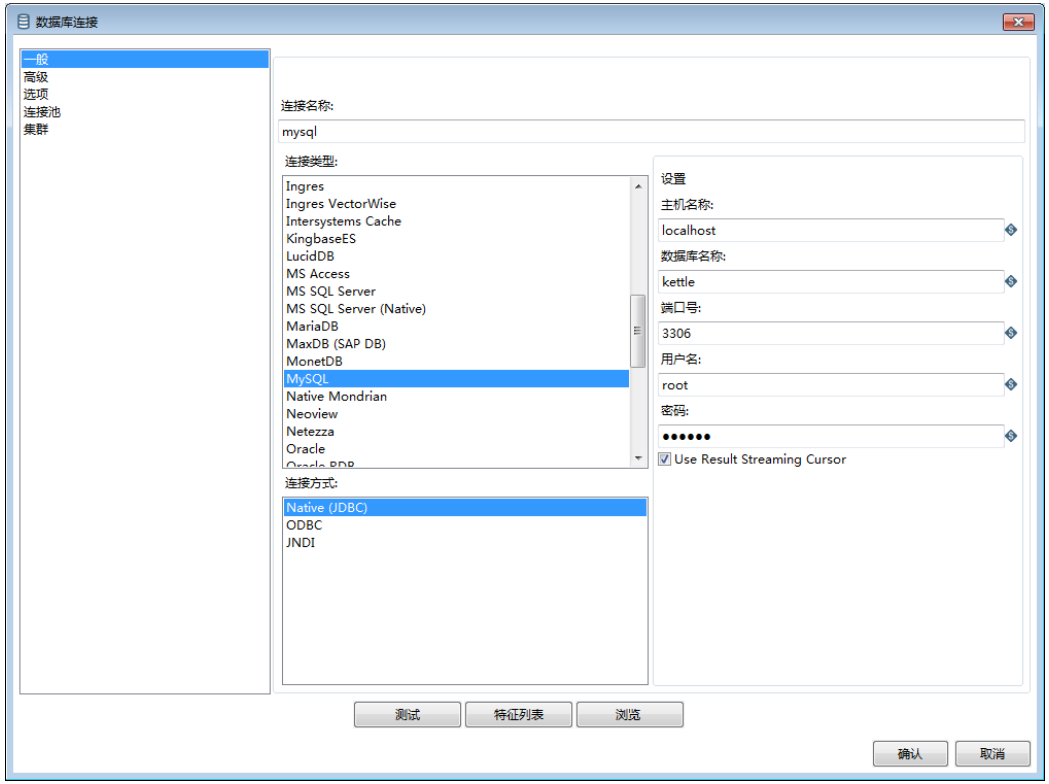


图7-142 建立数据库连接

7.6.2把HDFS文件加载到MySQL数据库中

参照本章7.6.1节的内容，建立一个名称为“Hadoop3”的Hadoop连接（如图7-143所示）。

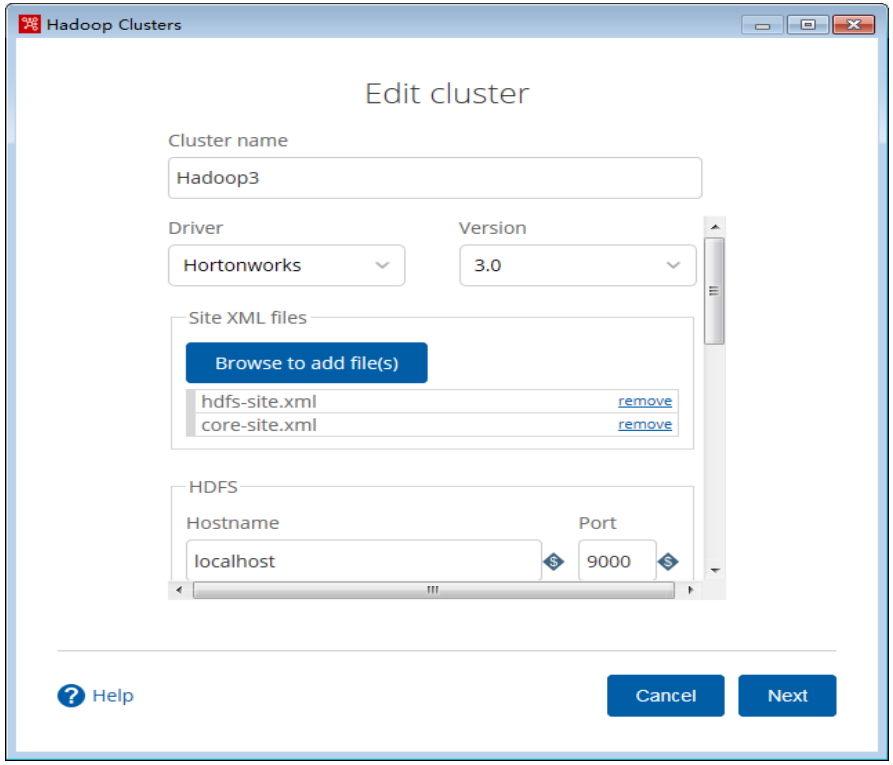


图7-143 建立Hadoop连接

7.6.2把HDFS文件加载到MySQL数据库中

4. 设计转换

在Spoon主界面的“核心对象”的“Big Data”里面，找到“Hadoop file input”控件，放置到设计区域，在“核心对象”的“输出”里面，找到“表输出”控件，放置到设计区域，为两个控件建立连线（如图7-144所示）。

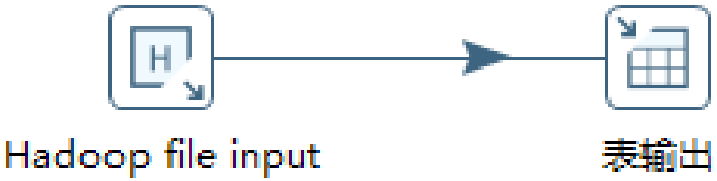


图7-144 放置“Hadoop file input”和“表输出”控件

7.6.2把HDFS文件加载到MySQL数据库中

在设计区域双击“Hadoop file input”控件图标，打开设置界面（如图7-145所示），用鼠标点击“Environment”下面的空白单元格，会出现如图7-146所示的下拉列表，选中“Hadoop3”。

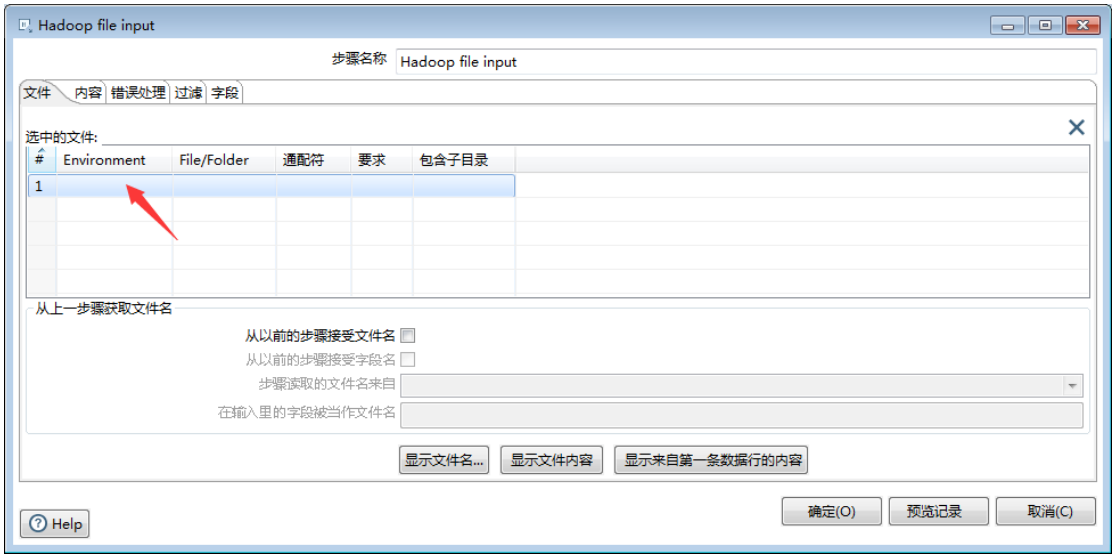


图7-145 “Hadoop file input”设置界面

7.6.2把HDFS文件加载到MySQL数据库中

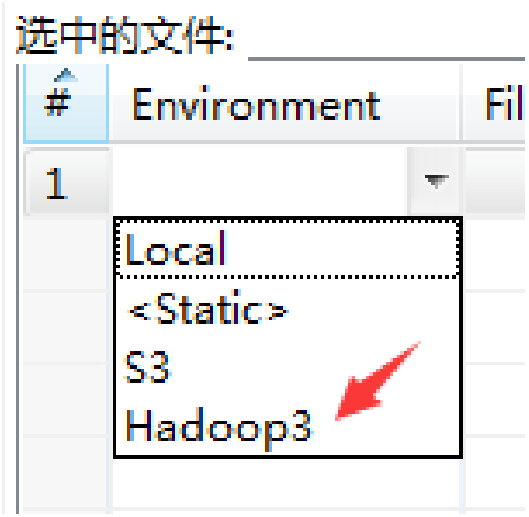


图7-146 设置Environment

7.6.2把HDFS文件加载到MySQL数据库中

点击“File/Folder”下面的空白单元格，会出现如图7-147所示的效果，点击省略号按钮，会弹出如图7-148所示的界面，选中HDFS中的student.txt文件，点击“OK”按钮，返回到“Hadoop file input”设置界面。

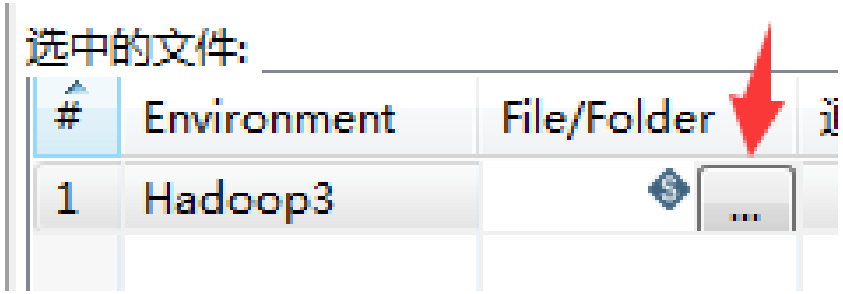


图7-147 设置File/Folder

7.6.2把HDFS文件加载到MySQL数据库中

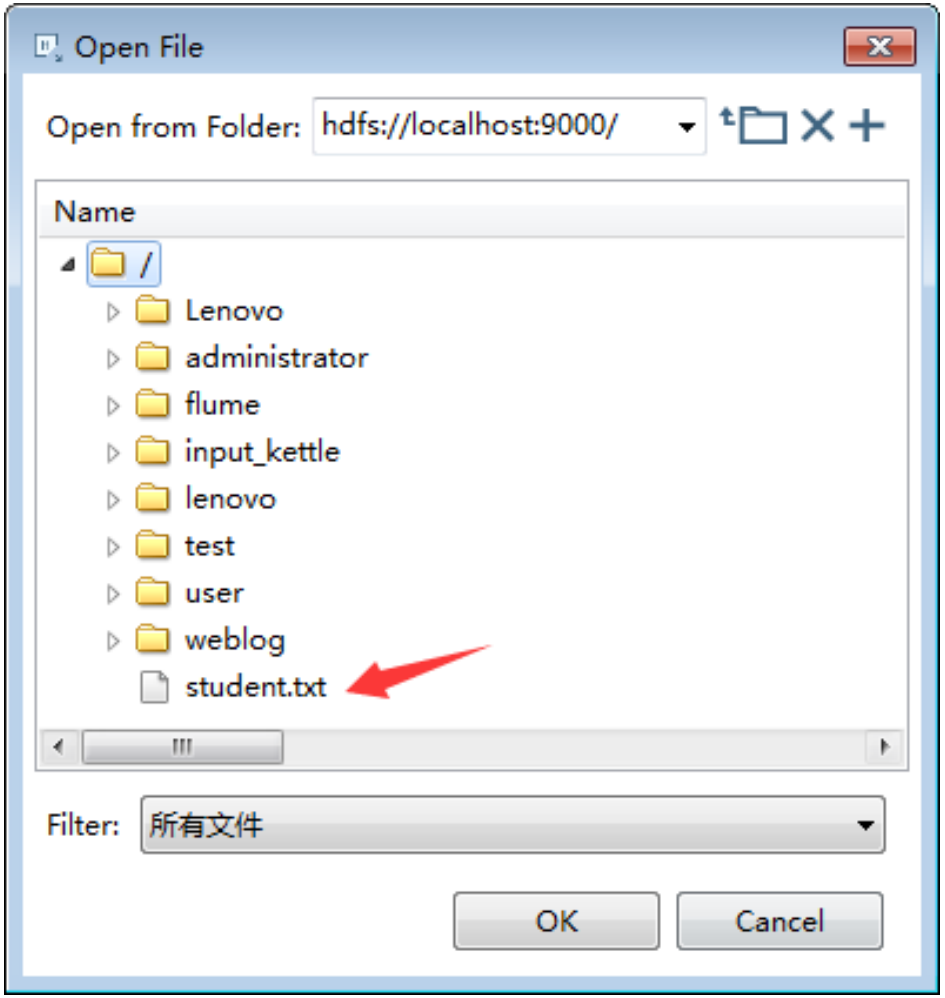


图7-148 选中HDFS中的student.txt文件

7.6.2把HDFS文件加载到MySQL数据库中

点击“Hadoop file input”设置界面的“内容”选项卡，会出现如图7-149所示的界面，“文件类型”选择“CSV”，把“分隔符”设置为“|”，“头部”后面的勾号选中，设置“头部行数量”为1。

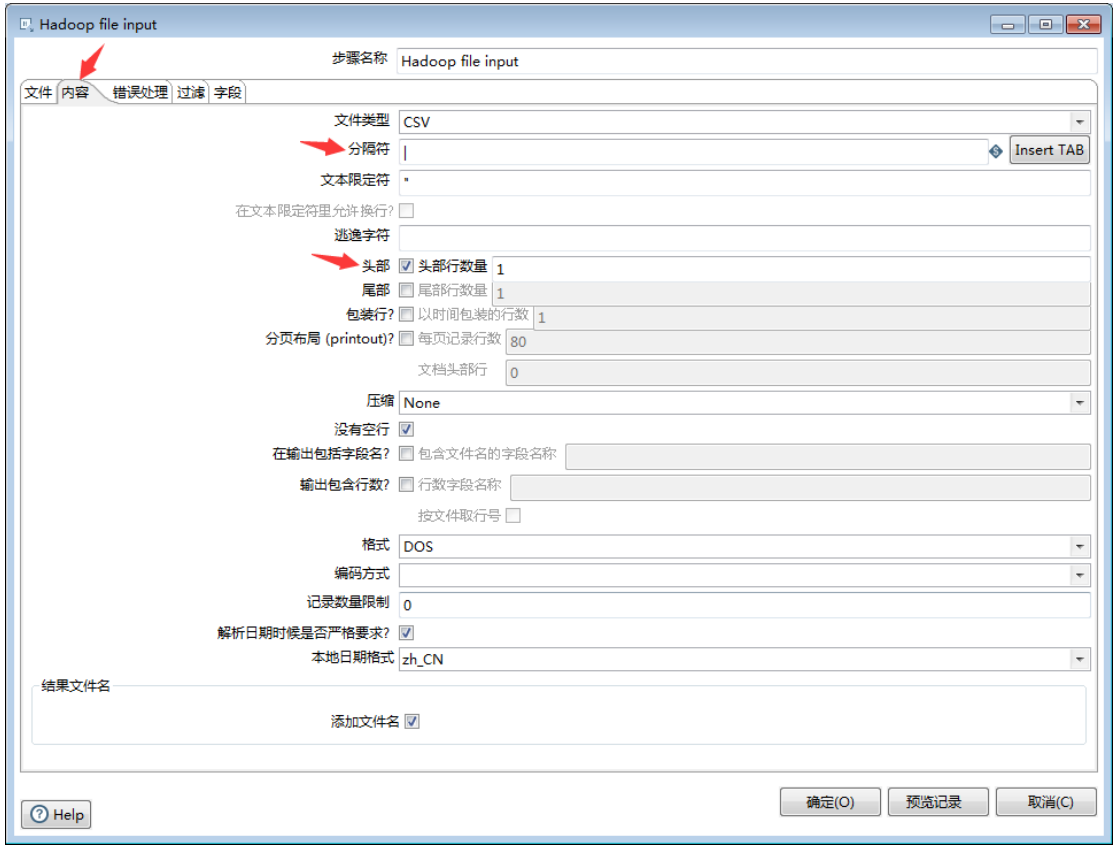
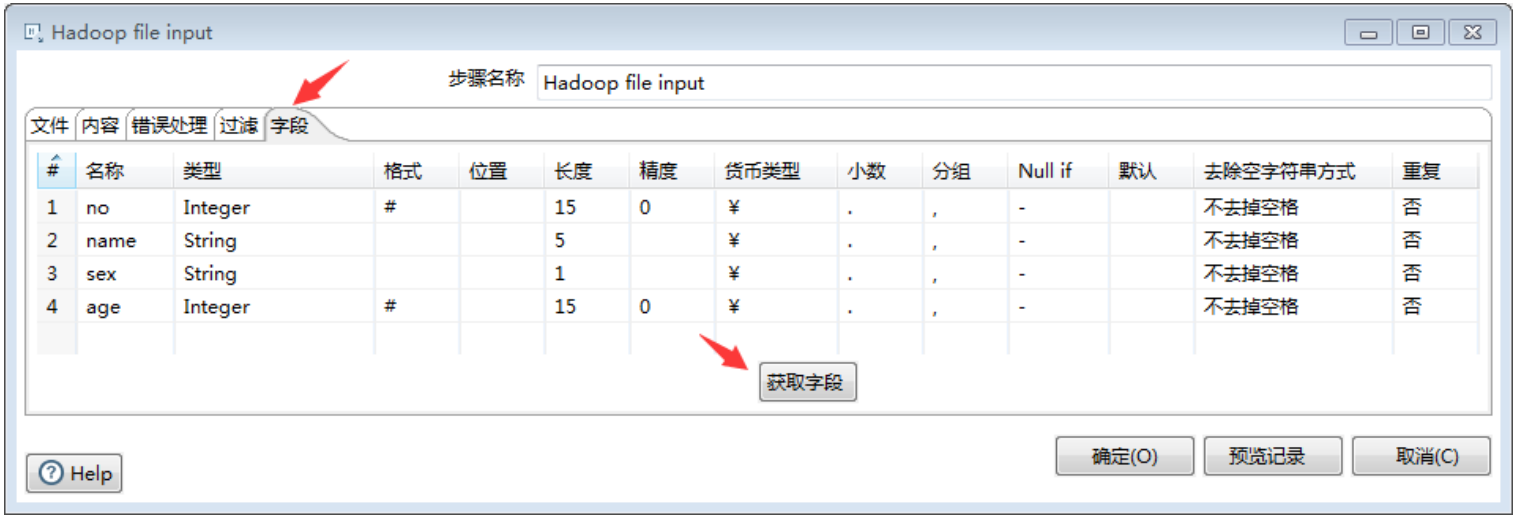


图7-149 设置分隔符

7.6.2把HDFS文件加载到MySQL数据库中

点击“Hadoop file input”设置界面的“字段”选项卡，会出现如图7-150所示的界面，点击界面底部的“获取字段”按钮，会弹出如图7-151所示的界面，直接点击“确定”按钮返回字段设置界面，最后，点击该界面上的“确定”按钮。



#	名称	类型	格式	位置	长度	精度	货币类型	小数	分组	Null if	默认	去除空字符串方式	重复
1	no	Integer	#		15	0	¥	.	,	-		不去掉空格	否
2	name	String			5		¥	.	,	-		不去掉空格	否
3	sex	String			1		¥	.	,	-		不去掉空格	否
4	age	Integer	#		15	0	¥	.	,	-		不去掉空格	否

图7-150 设置“字段”选项卡

7.6.2把HDFS文件加载到MySQL数据库中

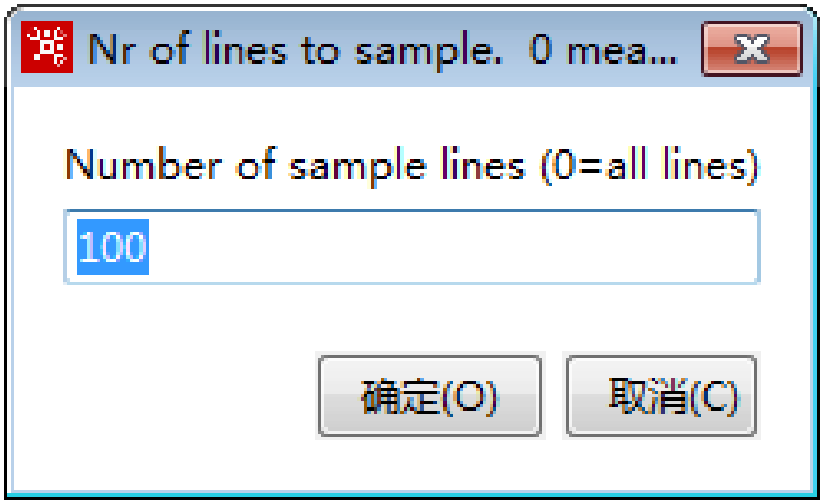


图7-151 设置取样行数

7.6.2把HDFS文件加载到MySQL数据库中

双击设计区域的“表输出”控件图标，打开“表输出”设置界面（如图7-152所示），在“数据库连接”右边的下拉列表中选择“mysql”，点击“目标表”右侧的“浏览”按钮，会弹出如图7-153所示界面，在界面中选中“student_table”表，点击“确定”按钮，返回到“表输出”设置界面，再点击“确定”按钮，完成设置。全部设置完成以后，需要保存设计文件。

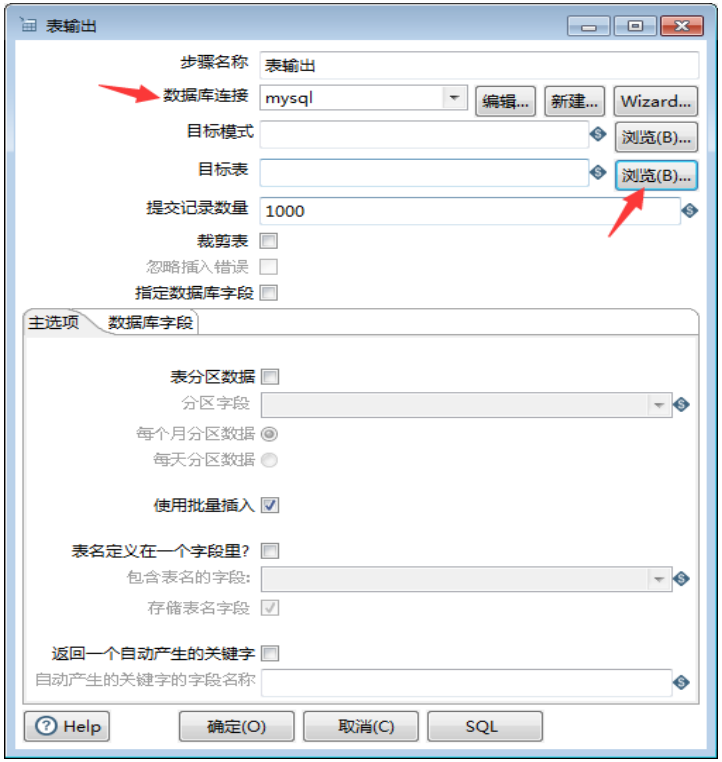


图7-152 表输出设置界面

7.6.2把HDFS文件加载到MySQL数据库中

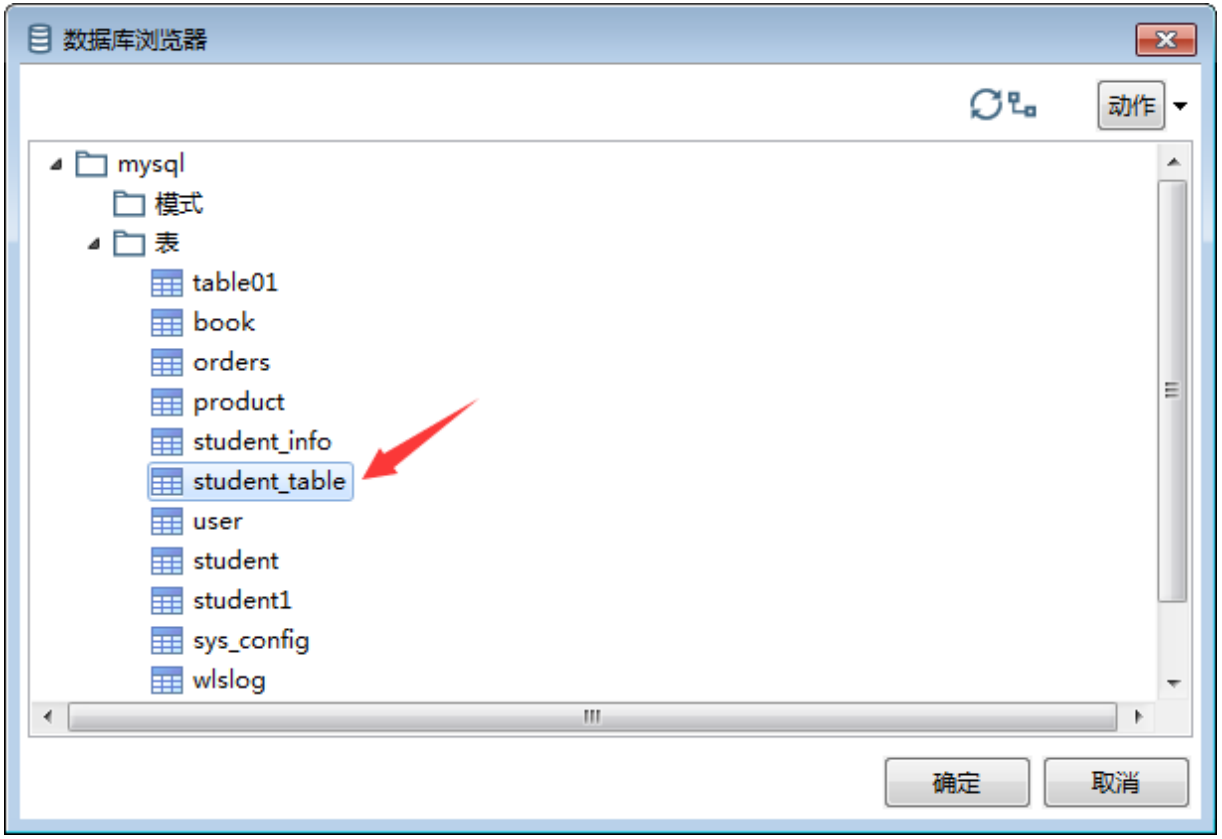


图7-153 选择student_table表

7.6.2把HDFS文件加载到MySQL数据库中

5. 执行转换

在转换设计界面中（如图7-154所示），点击三角形按钮开始执行转换，在弹出的界面中点击“启动”，如果转换执行成功，会显示如图7-155所示的效果，在两个控件图标上都会显示绿色的勾号。

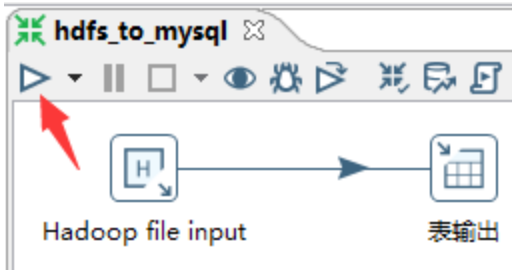


图7-154 转换设计界面

图7-155 执行成功的效果

7.6.2把HDFS文件加载到MySQL数据库中

这时，到MySQL数据库命令客户端执行如下SQL语句查看数据库中的数据：

```
mysql> USE kettle;
mysql> SELECT * FROM student_table;
```

上面SQL语句的执行结果如图7-156所示。

```
mysql> select * from student_table;
+------+-------+------+------+
| no   | name  | sex  | age  |
+------+-------+------+------+
|    1 | Mike  | M    |   21 |
|    2 | John  | M    |   22 |
|    3 | Kate  | F    |   21 |
|    4 | Jenny | F    |   21 |
+------+-------+------+------+
4 rows in set (0.00 sec)
```

图7-156 数据库查询执行结果

7.7 本章小结

大数据应用系统的构建过程中，数据清洗是一个非常重要的环节。通过使用ETL工具，可以大幅提高数据清洗的效率，本案例采用开源工具Kettle实现数据的ETL操作。本章介绍了Kettle的基本概念、基本功能和安装方法，并通过实例演示了使用Kettle进行数据抽取、数据清洗与转换、数据加载的具体方法。本章介绍的内容属于比较基础的Kettle使用方法，如果要学习更加高级复杂的Kettle使用方法，可以参考相关书籍或网络资料。

附录A：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学与技术系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://dblab.xmu.edu.cn/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），全国高校知名大数据教师，现为厦门大学计算机科学系副教授，厦门大学信息学院实验教学中心主任，曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度、2017年度和2020年度厦门大学教学类奖教金获得者，荣获2019年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过1000万字高价值的研究和教学资料，累计网络访问量超过1000万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过400万次，累计访问量超过1500万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：林子雨大数据系列教材

了解全部教材信息：http://dblab.xmu.edu.cn/post/bigdatabook/

附录D：《大数据导论（通识课版）》教材

开设全校公共选修课的优质教材

本课程旨在实现以下几个培养目标：

- 引导学生步入大数据时代，积极投身大数据的变革浪潮之中
- 了解大数据概念，培养大数据思维，养成数据安全意识
- 认识大数据伦理，努力使自己的行为符合大数据伦理规范要求
- 熟悉大数据应用，探寻大数据与自己专业的应用结合点
- 激发学生基于大数据的创新创业热情

高等教育出版社 ISBN:978-7-04-053577-8 定价：32元 版次：2020年2月第1版
教材官网：http://dblab.xmu.edu.cn/post/bigdataintroduction/

附录E：《大数据导论》教材

- 林子雨 编著《大数据导论》
- 人民邮电出版社，2020年9月第1版
- ISBN:978-7-115-54446-9 定价：49.80元

教材官网：http://dblab.xmu.edu.cn/post/bigdata-introduction/

开设大数据专业导论课的优质教材

扫一扫访问教材官网

附录F：《大数据技术原理与应用（第3版）》教材

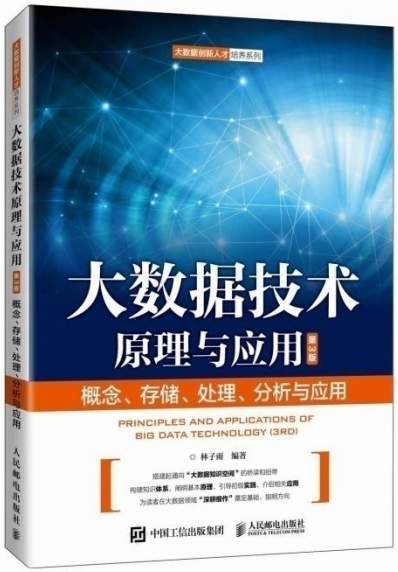
《大数据技术原理与应用——概念、存储、处理、分析与应用（第3版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社　ISBN:978-7-115-54405-6　定价：59.80元

全书共有17章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、Flink、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase、MapReduce、Spark和Flink等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata3

扫一扫访问教材官网



附录G：《大数据基础编程、实验和案例教程（第2版）》

本书是与《大数据技术原理与应用（第3版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

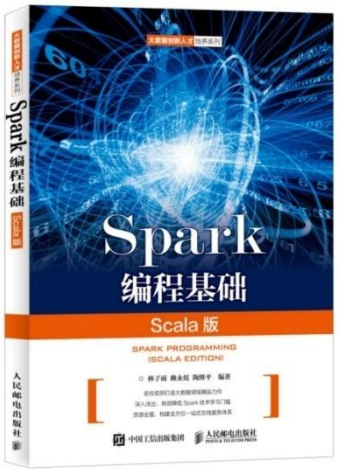
厦门大学林子雨编著

配套实验指导书

- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，八套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

林子雨编著《大数据基础编程、实验和案例教程（第2版）》
清华大学出版社 ISBN:978-7-302-55977-1 定价：69元 2020年10月第2版

附录H：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

厦门大学 林子雨，赖永炫，陶继平 编著

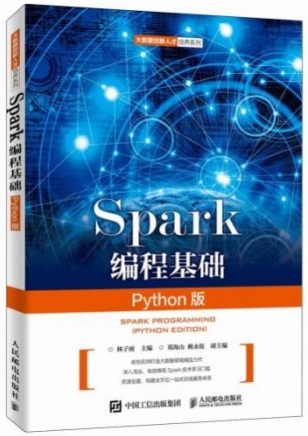
披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-48816-9

教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录I：《Spark编程基础（Python版）》



《Spark编程基础（Python版）》

厦门大学 林子雨，郑海山，赖永炫 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-52439-3

教材官网： http://dblab.xmu.edu.cn/post/spark-python/

本书以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、上机实验指南等。

附录J：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

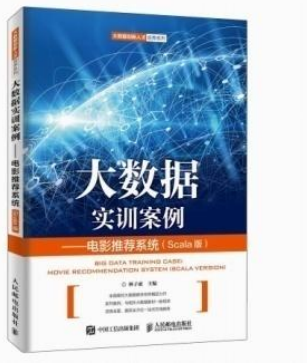
扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

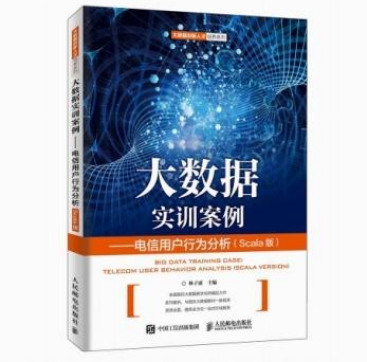
附录K：高校大数据实训课程系列案例教材

为了更好满足高校开设大数据实训课程的教材需求，厦门大学数据库实验室林子雨老师团队联合企业共同开发了《高校大数据实训课程系列案例》，目前已经完成开发的系列案例包括：

《电影推荐系统》（已经于2019年5月出版）
《电信用户行为分析》（已经于2019年5月出版）
《实时日志流处理分析》
《微博用户情感分析》
《互联网广告预测分析》
《网站日志处理分析》



系列案例教材将于2019年陆续出版发行，教材相关信息，敬请关注网页后续更新！
http://dblab.xmu.edu.cn/post/shixunkecheng/



扫一扫访问大数据实训课程系列案例教材主页


Thank You!