# 第三章 可视化ETL平台——Kettle

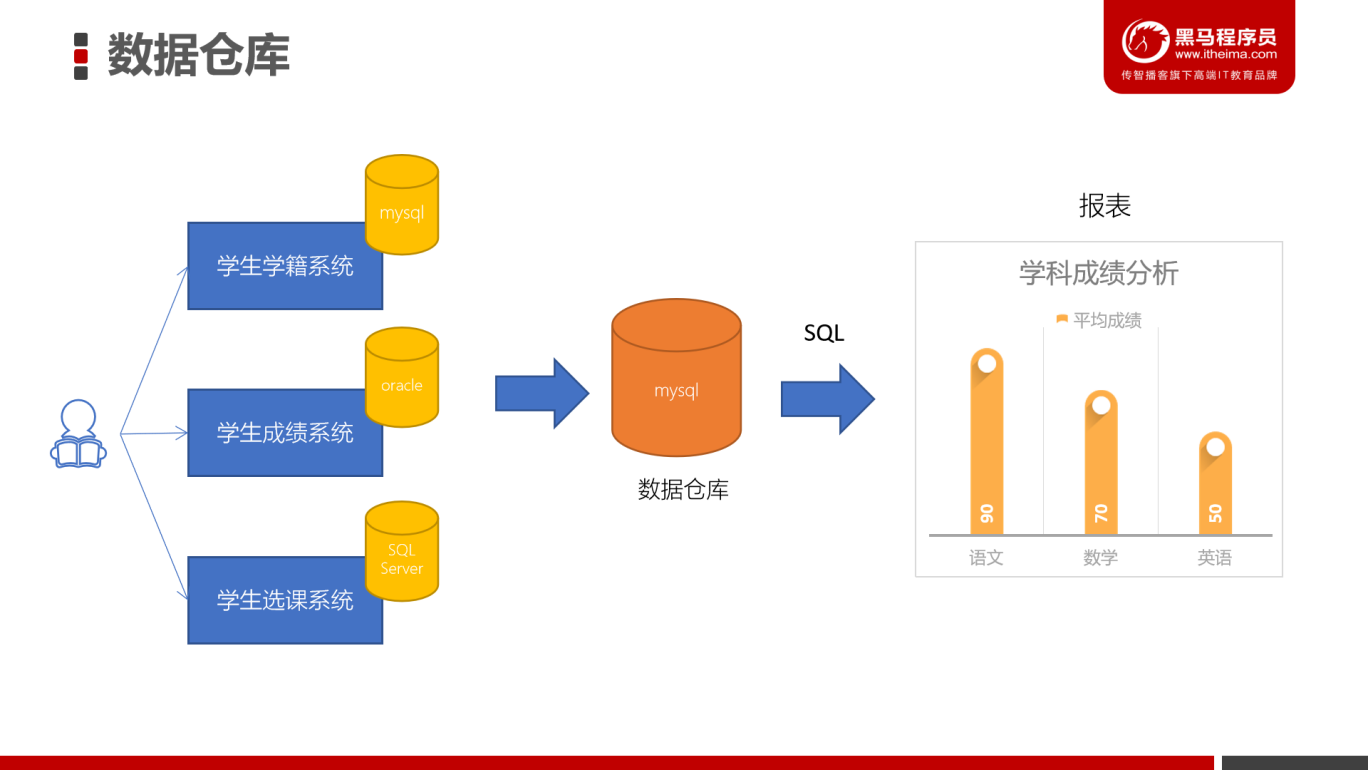
**课程目标**

* 能够理解Kettle的主要用途
* 能够搭建Kettle开发环境
* 能够掌握Kettle输入组件的使用
* 能够掌握Kettle输出组件的使用
* 能够掌握Kettle转换组件的使用

## Kettle介绍

### 数据仓库

数据仓库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的，对多样的业务数据进行筛选与整合。它为企业提供一定的BI能力，指导业务流程改进，监视时间成本、质量以及控制。

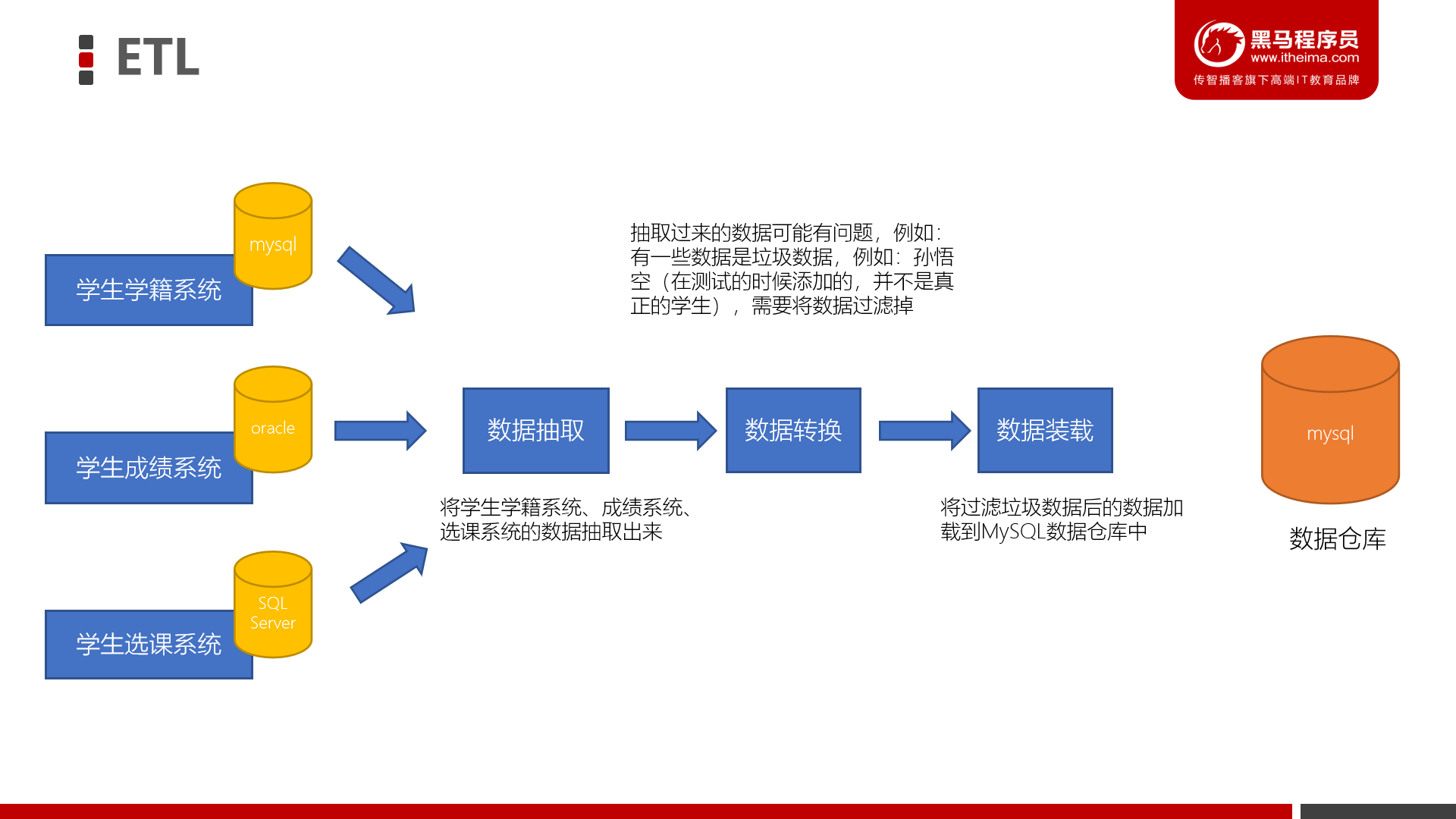


**重点**

数据仓库是专门用来进行数据分析的，它可以是一个MySQL，也可以是Oracle，也可以将来后面大数据会学习到的Hive等。

### ETL

ETL，是英文Extract-Transform-Load的缩写，用来描述将数据从来源端经（extrac过抽取t）、转换（transform）、加载（load）至目的端的过程。ETL是将业务系统的数据经过抽取、清洗、转换之后加载到数据仓库的过程，目的是将企业中分散、零乱、标准不统一的数据整合到一起。



### Kettle介绍

对于企业或行业应用来说，经常会遇到各种数据的处理，转换，迁移，掌握一种etl工具的使用，必不可少，这里要学习的ETL工具是——Kettle，现在已经更名为PDI。

* Kettle是一款国外开源的ETL工具，纯java编写，可以在Window、Linux、Unix上运行，绿色无需

安装

* Kettle中文名称叫水壶，该项目的主程序员MATT希望把各种数据放到一个壶里，然后以一种指

定的格式流出

* Kettle允许管理来自不同数据库的数据，提供一个图形化的用户环境来描述想做什么，无需关心

怎么做

|  |
| --- |
|  |
|  |
|  |
|  |

## windows安装Kettle

### 安装JDK

要在windows系统中使用Kettle，必须先安装好JDK。

#### 安装JDK软件

|  |  |
| --- | --- |
| **截图** | **步骤** |
|  | 1. 找到资料/安装包中的jdk软件安装包，双击打开。 |
|  | 1. 点击 「下一步」按钮 |
|  | 1. 将路径修改为 c:\opt\jdk1.8.0\_241，点击「确定」按钮 |
|  | 1. 点击「下一步」，等待进度条完成。 |
|  | 1. 在新弹出的窗口中，点击「更改」按钮，将JRE的安装路径修改为 c:/opt/jre1.8.0\_241。 |
|  | 注：碰到以下错误可以忽略（是因为机器名中包含了一些JDK不识别的支付） |
|  | 1. 点击「关闭」完成安装。 |

#### 配置环境变量

要让JDK环境生效，需要配置下系统的环境变量。此处大家先无需关心环境变量的意义，先按照下面的步骤配置好。

|  |  |
| --- | --- |
| **截图** | **步骤** |
|  | 1. 打开我的电脑，右键点击「此电脑」。 |
|  | 1. 点击「高级系统设置」。 |
|  | 1. 在弹出窗口中点击「高级」选项卡中的「环境变量」。 |
|  | 1. 点击「新建...」按钮添加环境变量。 |
|  | 1. 变量名写入：JAVA\_HOME，变量值设置为之前安装JDK的路径：c:\opt\jdk1.8.0\_241 |
|  | 1. 点击「确定」按钮保存。 |
|  | 1. 将JAVA\_HOME添加都Path中 |

### 安装Kettle

Kettle的安装非常简单，只需要下载后解压即可。但前提需要安装好JDK。

|  |  |
| --- | --- |
|  | 1. 下载Kettle（可以在资料/安装包/pdi-ce-9.0.0.0-423.zip) |
|  | 1. 解压Kettle（解压到非中文目录） |
|  | 1. 进入解压后的文件夹data-integration,双击Spoon.bat，启动kettle |

注：kettle也可以到可以到<https://sourceforge.net/projects/pentaho/files/>下载

## 快速入门

### 需求

有一个txt文件，内容如下：

|  |
| --- |
| id,name,age,gender,province,city,region,phone,birthday,hobby,register\_date  392456197008193000,张三,20,0,北京市,昌平区,回龙观,18589407692,1970-8-19,美食;篮球;足球,2018-8-6 9:44  267456198006210000,李四,25,1,河南省,郑州市,郑东新区,18681109672,1980-6-21,音乐;阅读;旅游,2017-4-7 9:14  892456199007203000,王五,24,1,湖北省,武汉市,汉阳区,18798009102,1990-7-20,写代码;读代码;算法,2016-6-8 7:34  492456198712198000,赵六,26,2,陕西省,西安市,莲湖区,18189189195,1987-12-19,购物;旅游,2016-1-9 19:15  392456197008193000,张三,20,0,北京市,昌平区,回龙观,18589407692,1970-8-19,美食;篮球;足球,2018-8-6 9:44  392456197008193000,张三,20,0,北京市,昌平区,回龙观,18589407692,1970-8-19,美食;篮球;足球,2018-8-6 9:44 |

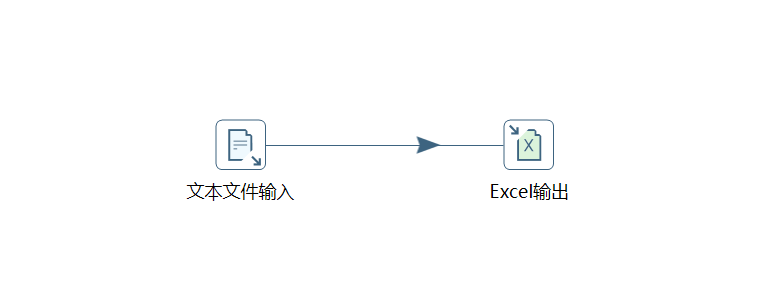
我们需要使用Kettle将这个文件中的数据抽取出来，然后装载到Excel中。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | name | age | gender | province | city | region | phone | birthday | hobby | register\_date |
| 392456197008193000 | 张三 | 20 | 0 | 北京市 | 昌平区 | 回龙观 | 18589407692 | 1970-08-19 | 美食;篮球;足球 | 2018-08-06 09:44:43 |
| 267456198006210000 | 李四 | 25 | 1 | 河南省 | 郑州市 | 郑东新区 | 18681109672 | 1980-06-21 | 音乐;阅读;旅游 | 2017-04-07 09:14:13 |
| 892456199007203000 | 王五 | 24 | 1 | 湖北省 | 武汉市 | 汉阳区 | 18798009102 | 1990-07-20 | 写代码;读代码;算法 | 2016-06-08 07:34:23 |
| 492456198712198000 | 赵六 | 26 | 2 | 陕西省 | 西安市 | 莲湖区 | 18189189195 | 1987-12-19 | 购物;旅游 | 2016-01-09 19:15:53 |
| 392456197008193000 | 张三 | 20 | 0 | 北京市 | 昌平区 | 回龙观 | 18589407692 | 1970-08-19 | 美食;篮球;足球 | 2018-08-06 09:44:43 |
| 392456197008193000 | 张三 | 20 | 0 | 北京市 | 昌平区 | 回龙观 | 18589407692 | 1970-08-19 | 美食;篮球;足球 | 2018-08-06 09:44:43 |



### Kettle的基本开发步骤

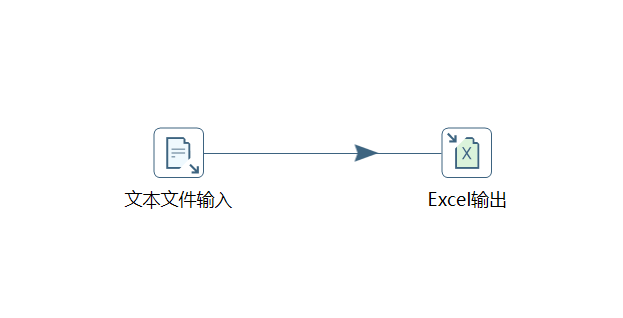
1. 要使用Kettle来进行数据处理，首先要做的是构建Kettle的数据流图，也就是用可视化操作像排积木的方式，把要做的事情排列出来。



1. 配置数据流图中的各个组件
2. 保存并启动执行

### 构建Kettle数据流图

效果图：



**开发步骤**

|  |
| --- |
| 1. 将左边的核心对象中的 **输入文件夹下的文本文件输入** 拖拽到中间空白区域。 |
|  |
| 1. 从**输出文件夹中将Excel输出** 组件拖拽到中间空白区域。 |
|  |
| 1. 按住Shift键点击 文本文件输入组件，移动鼠标到Excel输出，然后释放鼠标，这样就可以将两个组件连接到一起。 |
|  |

#### 配置文件文本输入组件

|  |
| --- |
| 1. 双击文本文件输入组件，在弹出窗口中点击「浏览」按钮。 |
|  |
| 1. 选择 资料/测试数据 中的 user.txt 文件 |
|  |
| 1. 点击「增加」按钮，将文件加入到要抽取的数据中来。注意：不点添加按钮，将不会抽取数据。 |
|  |
| 1. 点击「内容」选项卡，将分隔符修改为逗号（注意是英文状态的逗号），将编码方式修改为：UTF-8。 |
|  |
| 1. 点击「字段」选项卡，再点击「获取字段」按钮，可以读取到txt文件中的所有字段。 |
|  |
| 1. 点击「预览记录」按钮，查看Kettle是否能够读取到 user.txt 中的数据。 |
|  |
|  |
| 1. 点击「确定」按钮保存。 |
|  |

#### 配置Excel输出组件

配置好了抽取 user.txt 组件后，接下来要配置将数据装载到Excel的组件。

|  |
| --- |
| 1. 双击 Excel输出组件，在弹出窗口中点击「浏览」按钮。 |
|  |
| 1. 点击「内容」选项卡，设置要将Excel文件输出到哪儿的路径，然后点击确定，再删除文件后缀xls |
|  |
| 1. 点击「字段」选项卡，再点击「获取字段」，将age的格式设置为0，表示只输出不带小数点的数字。 |
|  |

### 保存并启动执行

|  |
| --- |
| 1. 点击工具条上方的保存按钮，或者按快捷键Ctrl + s。 |
|  |
| 1. Kettle会提示让我们保存该数据流图（转换），我们保存到一个指定位置即可，然后点击「保存」按钮即可。 |
| 1. 点击工具栏上的播放按钮启动执行。 |
|  |
| 1. 执行完后，我们可以到指定的位置，发现Excel文件已经生成，可以看到里面的内容。 |
|  |

## 程序员小姐姐的烦恼

### 问题来了

公司来了有个漂亮的程序员小姐姐叫小花，她刚大学毕业，项目经理安排她这样一项工作：

|  |
| --- |
| 有这样一个Excel文件：user.xls，这个文件内容是这样的。    项目经理想要让她将这些数据导入到MySQL中来。  小花刚来，急得团团转，不知所措，机会来了。 |

### 数据抽取的需求

需要从Excel中将这些用户的数据，使用Kettle抽取到MySQL中

### 准备工作

为了完成本案例，我们需要准备以下几件工作：

#### 找到小姐姐的Excel文件

在资料/测试数据 文件夹中可以找到user.xlsx文件

|  |  |
| --- | --- |
|  |  |

#### 在MySQL数据库中创建数据库

为了方便将Excel文件中的数据抽取到MySQL中，我们必须要创建一个名字叫kettle\_demo的数据库，后续Excel中的数据会装载到该数据库的表中。

|  |  |
| --- | --- |
|  | 1. 在DataGrip中右键点击MySQL连接，选择New/Schema |
|  | 1. 在创建数据库的对话框中输入kettle\_demo，点击Execute |
|  | 1. 我们可以看到kettle\_demo数据库名称就已经创建好了 |

#### 在kettle中加载MySQL驱动

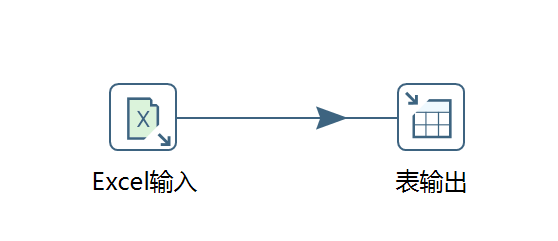
Kettle要想连接到MySQL，必须要安装一个MySQL的驱动，就好比我们装完操作系统要安装显卡驱动一样。加载MySQL驱动只需以下两步：

1. 将资料中的 MySQL jdbc 驱动包mysql-connector-java-5.1.47.jar和mysql-connector-java-8.0.13.jar导入到 **data-integration/lib** 中

3.重启Kettle即可

### 构建Kettle数据流图

效果图：



**开发步骤**

|  |
| --- |
| 1. 在Kettle中创建一个转换（两种方式）    1. 点击菜单：文件/新建/转换    2. 点击按钮：    3. 快捷键：Ctrl + N |
|  |
| 1. 从左边的核心对象中，分别拖入「输入/Excel输入」、「输出/表输出」两个组件到中间区域 |
|  |
| 1. 然后按住Shift键，在 「Excel输入」组件上点击鼠标左键，拖动到「表输出」组件上，连接两个组件，这样数据流图就构建好了 |
|  |

### 配置Kettle数据流图中的组件

刚刚已经把数据流图构建好了，那么Kettle就可以将Excel文件中的数据抽取到MySQL中吗？

显然是不行的。Kettle根本不知道要将哪个Excel文件中的数据，抽取到哪个MySQL中。我们需要配置这两个组件，告诉Kettle从哪个Excel文件中抽取，以及将数据装载到哪个MySQL中。

#### 配置Excel输入组件

|  |
| --- |
| 1. 双击Excel输入组件，会弹出一个对话框，我们可以再该对话框中配置该组件 |
|  |
| 1. 因为此处要抽取的Excel文件为Excel 2007版本，所以指定**表格类型**为Excel 2007 XLSX (Apache POI) |
|  |
| 1. 随后我们需要找到要抽取的那个Excel文件，点击「浏览」按钮，找到 「资料/测试数据/user.xlsx」文件 |
|  |
| 1. 再点击旁边的「增加」按钮，切记：一定要点击增加按钮哦！否则没有效果！ |
|  |
| 1. 在弹出菜单中，点击「Sheet1」工作簿，并点击 「>」 按钮移动到右边。 |
|  |
| 1. 点击「字段」选项卡，点击「获取来自头部数据的字段...」按钮，Kettle会从Excel中读取第一行字段名称。 |
|  |
| 1. 将 age 字段的格式设置为#，register\_date的格式设置为 yyyy-MM-dd HH:mm:ss。 |
|  |
| 1. 点击「预览记录」按钮查看抽取到的数据。 |
|  |
| 1. 点击「确定」按钮保存。 |
|  |

#### 配置MySQL组件

##### 创建数据库连接

要使用Kettle操作MySQL，必须要建立Kettle与MySQL的连接，否则Kettle也不知道操作哪个MySQL库。

|  |
| --- |
| 1. 双击「表输入」组件，会自动弹出配置窗口，点击「新建」按钮 |
|  |
| 1. 配置MySQL连接    1. 输入连接名称，此处用mysql\_开头，数据库名称kettle\_demo为结尾    2. 在连接类型列表中，选择MySQL    3. 输入连接方式:    4. 输入MySQL的连接参数 |
|  |
| 1. 点击测试按钮，测试Kettle是否能够正确连接到MySQL |
| 1. 点击确认保存，到这里数据库连接就应该创建好了。 |
|  |

##### 使用Kettle在MySQL中自动创建表

要保存数据到MySQL，必须先要创建好表。那么，我们是否需要自己手动在MySQL中创建一个表，用来保存Excel中抽取过来的数据呢？

答案是：不需要。Kettke可以自动为我们在MySQL中创建表。

|  |
| --- |
| 1. 输入目标表的名称为：t\_user，后续Kettle将在MySQL中创建一张名为 t\_user 的表格。 |
|  |
| 1. 点击下方的「SQL」按钮，可以看到Kettle会自动帮助我们生成MySQL创建表的SQL语句 |
|  |
| 将age、gender字段类型设置为INT |
|  |
| 1. 点击执行按钮。Kettle将会让MySQL执行该SQL脚本。执行完后，可以在DataGrip中刷新在数据库，可以查看到Kettle帮助我们创建的t\_user表。 |
|  |
|  |
| 1. 点击「确定」按钮，保存配置 |
|  |

### 保存并启动执行Kettle转换

数据流图中的组件都已经准备好了，接下来就可以开始执行Kettle转换了。

|  |
| --- |
| 1. 点击保存按钮保存转换。 |
|  |
| 1. 点击工具栏上的播放按钮启动执行。 |
|  |
|  |
| 1. 执行成功后，可以看到以下界面。组件上都显示了绿色的对号，执行结果中可以看到：转换完成！日志，说明Kettle的转换已经执行成功！！ |
|  |
|  |

### 确认执行结果

Kettle是否已经帮助我们将Excel中的数据抽取并装载到MySQL呢？我们需要到MySQL中看一看，t\_user表中是否有数据呢。

|  |
| --- |
| 在DataGrip中双击 t\_user 表格，可以看到Excel中的数据都已经抽取到了MySQL。 |
|  |

到这里，恭喜你，程序员小姐姐小花的问题你已经帮她解决了，晚上让她请你吃个饭吧。

## Transformation（转换）开发

### 共享数据库连接

在后续的Kettle中，我们需要多次用到上面的数据库连接。那么是不是每一次都要创建数据库连接呢？这样就比较麻烦了。在Kettle中，可以将一个数据库**共享，**这样其他的Kettle转换就都可以复用该数据库连接了。

|  |
| --- |
| 1. 在转换中，点击「主对象树」，点击[DB连接]右键新建数据库连接（如果连接已经存在，则无需创建） |
|  |
| 1. 右键单击需要的数据库连接，选择「共享」 |
|  |
| 1. 发现刚刚选择的数据库连接已经变成黑体字，说明已经共享成功。 |
|  |
| 我们兴奋地发现，在新建的转换中，都可以看到该数据库连接了。这样就避免了每次我们都需要配置该数据库连接了。 |
|  |

### 表输入组件

#### 需求

前面我们已经将Excel中数据抽取到了MySQL的t\_user表中。



现在有了新需求，要将MySQL数据库中的 t\_user 表中的数据抽取出来，装载到另外一张表 t\_user1中。

#### 构建Kettle数据流图

|  |
| --- |
| 1. 从 核心对象的输入组件中，将「表输入」组件拖拽到中间的空白处。 |
|  |
| 1. 从输出中将「 表输出」组件拖拽到中间空白处。 |
|  |
| 1. 安装Shift键，并鼠标左键点击表输入组件，并拉动鼠标，移动到表输出组件，松开鼠标。 |
|  |

#### 配置Kettle数据流图中的组件

接下来，需要指定，数据从哪个数据库中读取数据。

##### 配置表输入组件

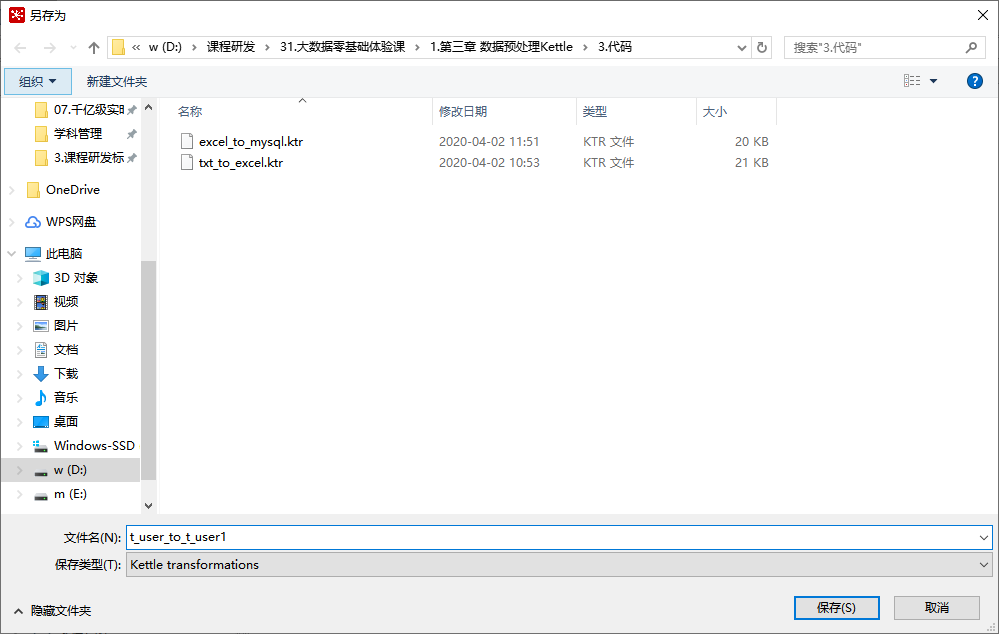
|  |
| --- |
| 1. 双击表输入组件，在弹出对话框中选择「获取SQL查询语句」。 |
|  |
| 1. 选择 t\_user 表，点击确定。 |
|  |
| 1. 在弹出对话框中选择「否」。 |
|  |
| 1. 点击「预览」按钮，查看是否能够从MySQL读取数据。 |
|  |
| 1. 可以看到，Kettle中可以查看到 t\_user 表中的数据。 |
|  |

##### 配置表输出组件

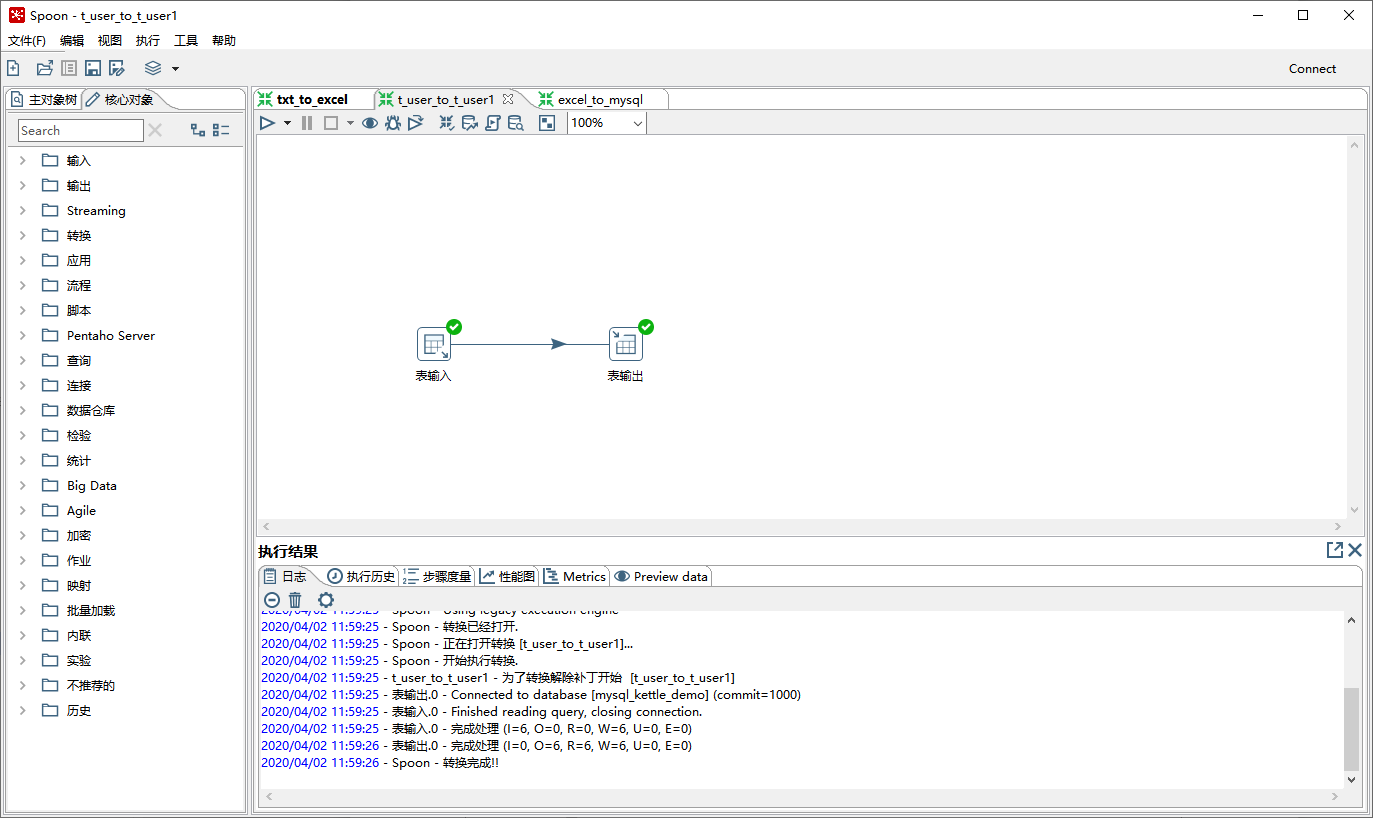
|  |
| --- |
| 1. 双击表输出组件，在目标表中输入 t\_user1 |
|  |
| 1. 点击「SQL」按钮，让Kettle中自动创建表结构。 |
|  |
| 1. 可以看到，Kettle自动为我们生成了创建表的SQL语句。 |
|  |
|  |
| 1. 我们可以到DataGrip中看到 t\_user1 已经创建，但表中没有任何数据。 |
|  |
| 1. 点击「确定」按钮保存。 |
|  |

#### 保存并启动执行

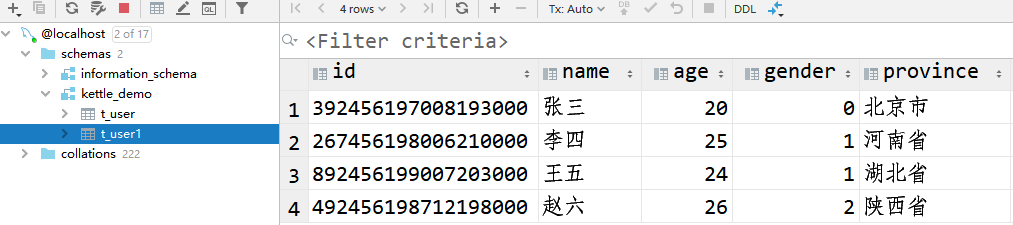
1. 点击保存按钮保存该转换。



1. 点击播放按钮执行，如果看到每个组件上都打上了绿色的对号，说明转换已经执行成功。



1. 再打开DataGrip，刷新下表格，可以看到数据已经加载到 t\_user1 中来了。



### 插入/更新组件

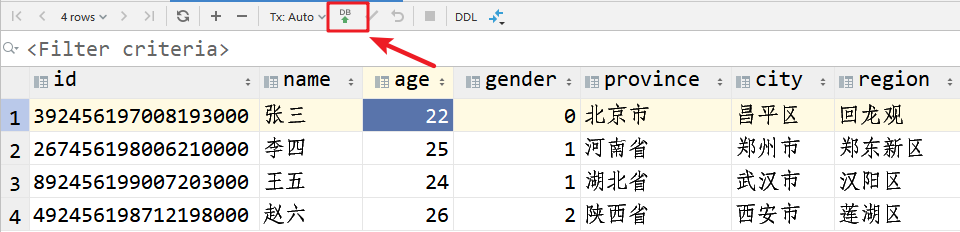
#### 组件介绍

插入/更新组件能够将Kettle抽取的数据，与某个表的数据进行对比，如果数据存在就更新，不存在就插入。



#### 需求

修改 t\_user中的张三这一行数据，修改age为22



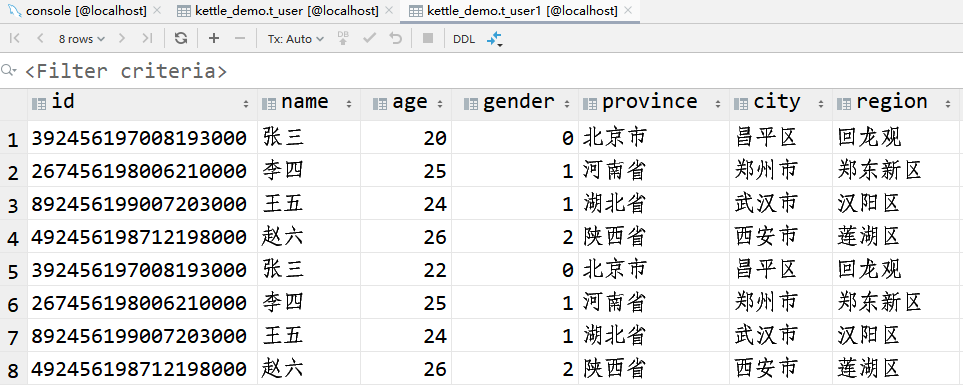
同时，我们想要使用Kettle将 t\_user1 中的张三这一行数据的age也修改为22。



#### 测试之前开发的t\_user\_to\_t\_user1转换

我们是否能够使用 t\_user\_to\_t\_user1.ktr转换来进行数据的同步呢？

直接执行转换，我们发现，Kettle又将t\_user表中的数据新增到了t\_user1表中



说明，表输入组件根本无法同步数据，只是将抽取到的数据，装载到指定的表中。

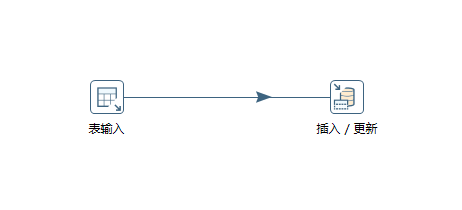
#### 恢复数据

为了方便后续的测试，我们需要恢复 t\_user1中的数据。

|  |
| --- |
| 1. 清空 t\_user1 中的数据 |
|  |
| 1. 点击OK执行清空 |
|  |
| 1. 点击刷新按钮，发现数据已经清空 |
|  |
| 1. 重新运行 t\_user\_to\_t\_user1.ktr 转换 |
| 1. 并将张三对应的年龄恢复为20 |
|  |

#### 构建Kettle数据流图

效果图：



**开发步骤**

|  |
| --- |
| 1. 拖入一个表输入组件，用于读取 t\_user 表中的数据。 |
|  |
| 1. 从 输出中拖入 「插入/更新」组件。 |
|  |
| 1. 将两个组件连接起来。 |
|  |

#### 配置Kettle数据流图中的组件

##### 配置表输入组件

|  |
| --- |
| 1. 双击表输入组件，点击获取SQL查询语句，选择 t\_user1 表。 |
|  |
| 1. 点击 预览按钮，查看Kettle是否能够从MySQL中读取数据。 |
|  |
|  |

##### 配置插入/更新组件

|  |
| --- |
| 1. 双击 插入/更新 组件，点击浏览按钮，找到 t\_user1 表。 |
|  |
| 1. 添加用来查询的关键字，设置表字段为：id，比较符为：=，流里的字段为：id。 |
|  |
| 1. 点击「获取和更新字段」，这样Kettle将会自动更新、或插入所有的字段。 |
|  |

#### 保存并启动执行Kettle转换

|  |
| --- |
| 1. 保存并启动该Kettle转换。 |
|  |
| 1. 执行完后，打开DataGrip刷新 t\_user1 表格，发现张三的年龄已经更新为22。 |
|  |

注：我们可以往 t\_user 中添加一条数据，然后再执行Kettle转换，再用DataGrip查看 t\_user1 表中是否也插入了一条数据。

### 删除组件

删除组件能够按照指定条件，将表中的数据删除。



#### 需求

有以下一个文本文件，文本文件包含了要删除的两个用户id：

|  |
| --- |
| id  392456197008193000  267456198006210000 |

需要使用Kettle将文本文件中两个ID对应的t\_user1表的数据删除。

#### 构建Kettle数据流图

效果图：

|  |
| --- |
| 1. 将文本文件输入组件拖拽到中间的空白区域。 |
|  |
| 1. 再从 输出文件夹中输出一个「删除」组件。 |
|  |
| 1. 按住Shift键，将两个组件连接到一起。 |
|  |

#### 配置Kettle数据流图组件

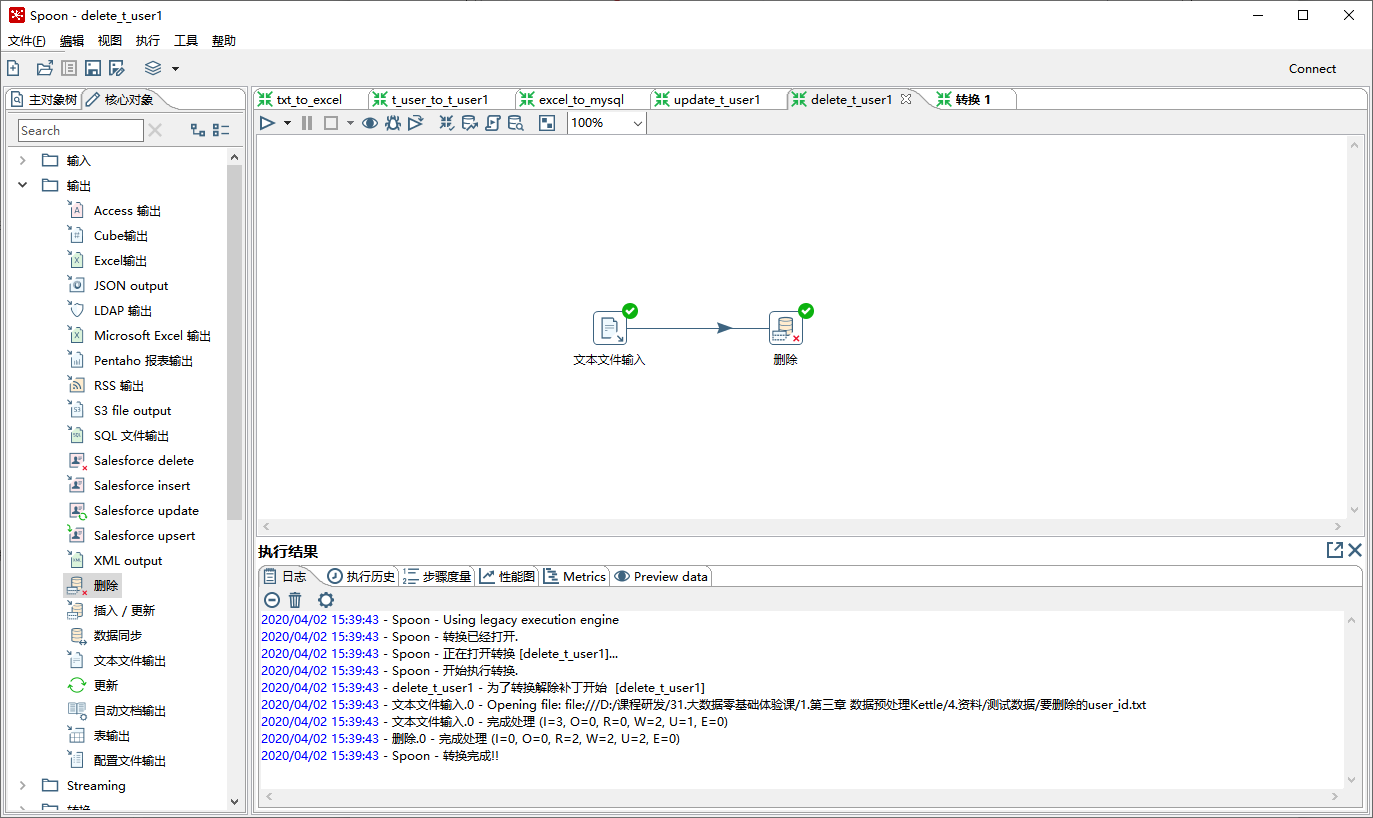
##### 配置文本文件输入组件

|  |
| --- |
| 1. 双击文本文件输入组件，点击浏览，找到 资料/测试数据 中的 **要删除的user\_id.txt** 文件，再点击OK按钮。 |
|  |
| 1. 点击「增加」按钮。 |
|  |
| 1. 再点击「字段」选项卡，点击「获取字段」，将文本文件中的字段读取出来。 |
|  |
| 1. 点击预览记录按钮，可以查看到Kettle能够读取到文本文件中的内容。 |
|  |

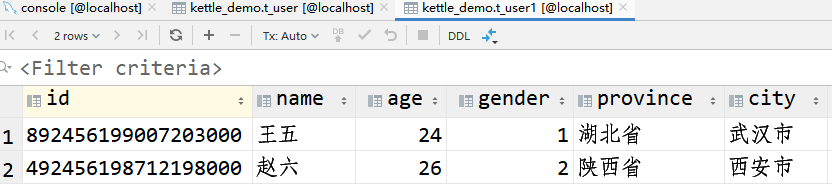
##### 配置删除组件

|  |
| --- |
| 1. 双击删除组件，然后点击浏览按钮，选择 t\_user1 表。 |
|  |
| 1. 在查询值所需的关键字，设置要根据什么条件删除数据。这里表字段设置为：id，比较符设置为：=，流里的字段1设置为：id。 |
|  |

#### 保存并启动运行



我们发现，在DataGrip中，指定ID的行已经被删除了。



### 排序记录组件

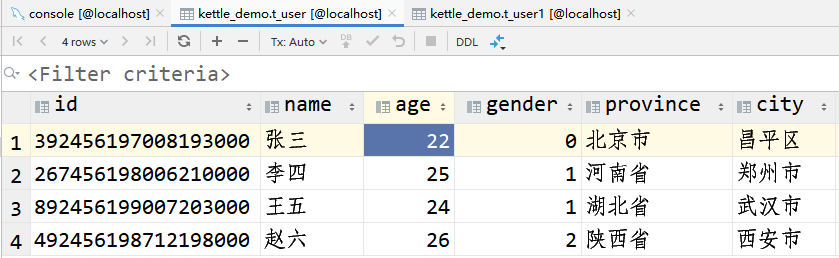
#### 组件介绍

排序组件可以将Kettle数据流中的数据进行排序，可以指定升序、还是降序排列。



#### 需求

使用Kettle将t\_user表中的用户数据，按照年龄升序排序，并将排序后的数据装载到Excel





#### 构建Kettle数据流图

效果图：



**开发步骤**

|  |
| --- |
| 1. 将表输入组件拖拽到中间空白区域。 |
|  |
| 1. 将转换文件夹中的「排序记录」拖拽到中间的空白区域。 |
|  |
| 1. 再将输出文件夹中的Excel输出组件拖拽到中间空白区域。 |
|  |
| 1. 按住Shift键将组件连接起来。 |
|  |

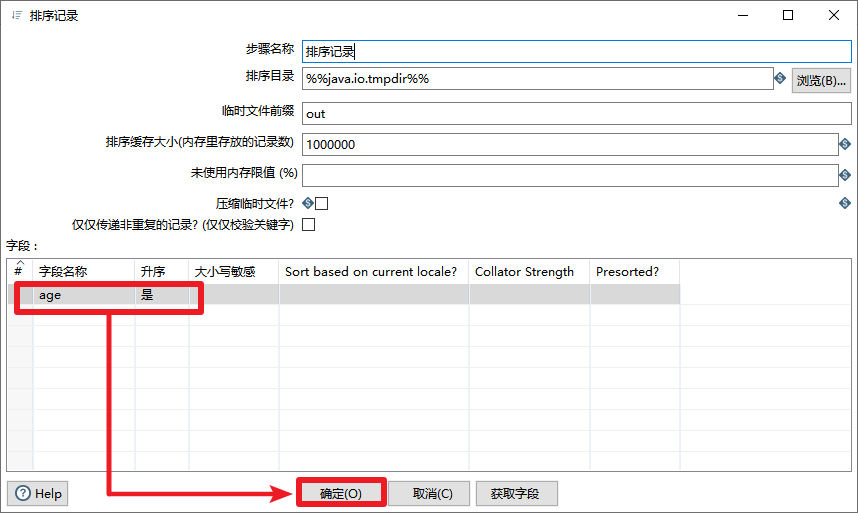
#### 配置Kettle数据流图组件

##### 配置表输入组件

|  |
| --- |
| 1. 双击表输入组件，选择数据库连接，然后点击获取SQL查询语句。再预览数据查看是否能够读取到 t\_user 中的数据。 |
|  |
|  |

##### 配置排序记录组件

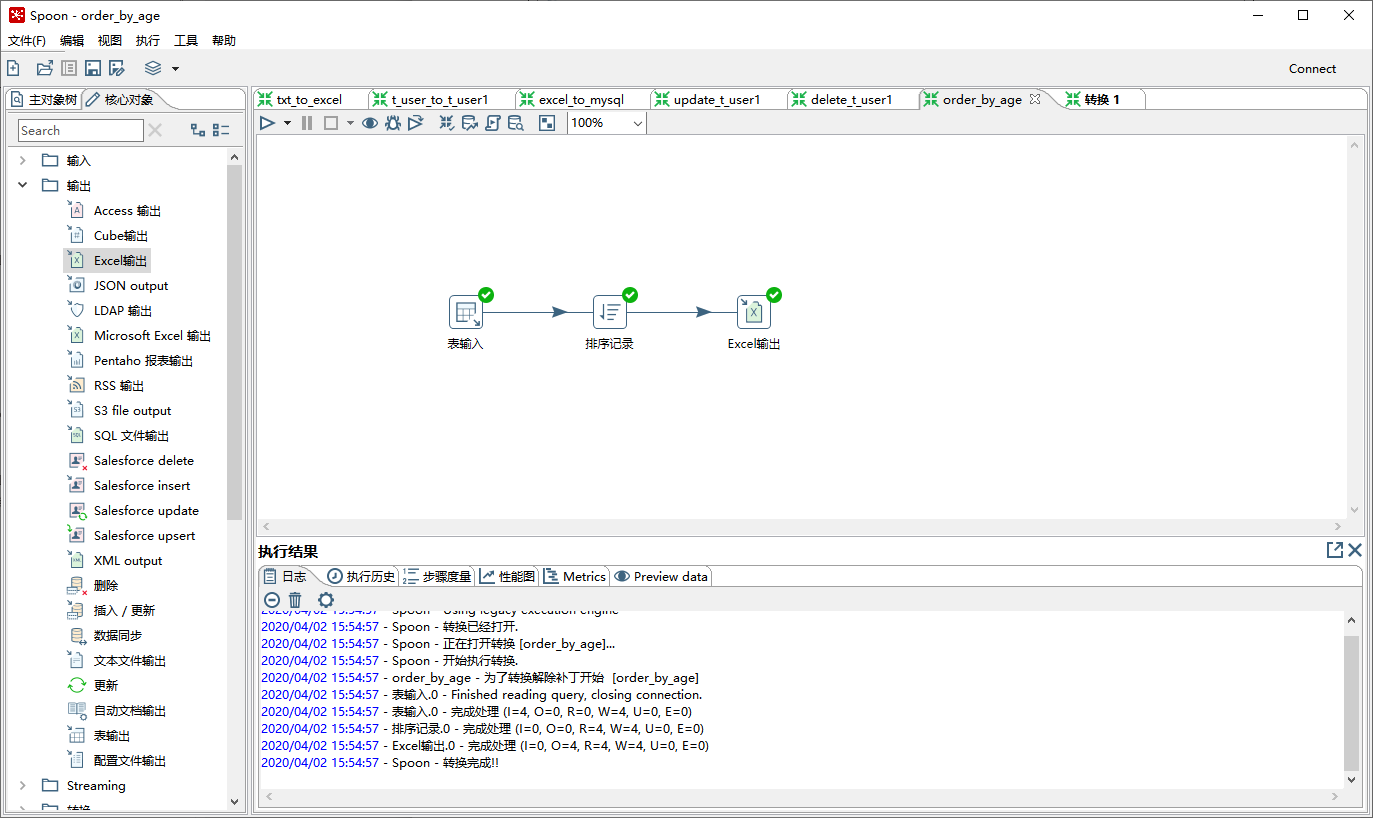
双击 排序记录 组件，设置字段名称为 age，升序选择「是」，这样该组件将会以 age 字段来进行升序排序。



##### 配置Excel输出组件

|  |
| --- |
| 1. 双击Excel输出组件，点击浏览按钮，选择要将Excel文件输出到哪个位置。 |
|  |
| 1. 点击「获取字段」，并将age的格式设置为0。 |
|  |

#### 保存并启动运行



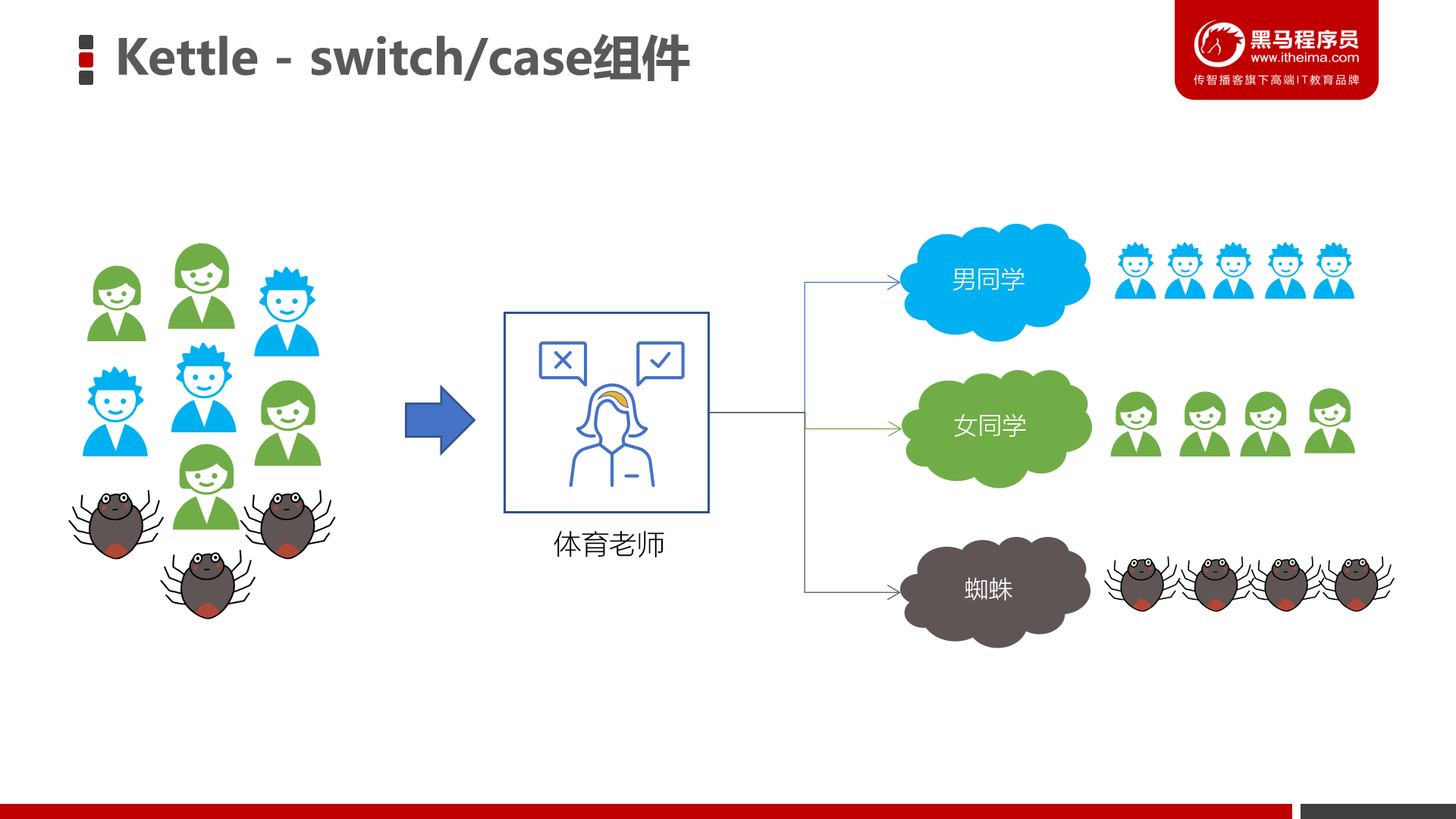
打开Excel，发现所有数据都是按照年龄升序排序的。



### switch/case组件

#### 机智的体育老师——条件判断

有一天，体育老师要让班上的男女同学分别排成两队。但这个班上还有几名同学，很特殊——他们是**蜘蛛**！！所以，机智的体育老师需要把他们排成三队，男同学一队，女同学一队，蜘蛛一队。

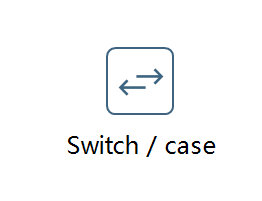


体育老师要做一件非常重要的事情：判断学生是男孩还是女孩、或者是蜘蛛，然后让他们各自到指定的队伍中。

体育老师做的事情，我们同样也会在Kettle中会经常用来。在Kettle中，switch/case组件可以来做类似于体育老师要做的事情——判断。

#### 组件介绍

switch/case提供了一种条件判断的实现。

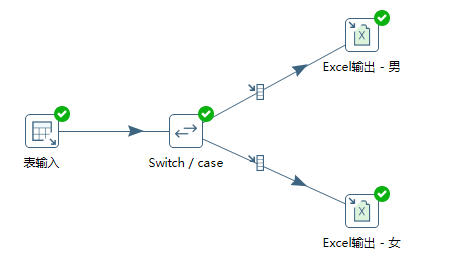


#### 需求

从 t\_user 表中读取所有用户数据，我们需要将性别为男的用户导出到一个Excel、性别为女的导出到另外一个Excel。

#### 构建Kettle数据流图

效果图：



**开发步骤**

|  |
| --- |
| 1. 将表输入组件拖拽到中间的空白区域。 |
|  |
| 1. 从流程文件夹中将 Switch/case 组件拖拽到中间的空白区域。 |
|  |
| 1. 再分别拖入两个Excel输出组件。 |
|  |
| 1. 将组件按照下图方式连接起来。 |
|  |

#### 配置Kettle数据流图组件

##### 配置表输入组件

|  |
| --- |
| 1. 双击表输入组件，指定从 t\_user 表中抽取数据。 |
|  |

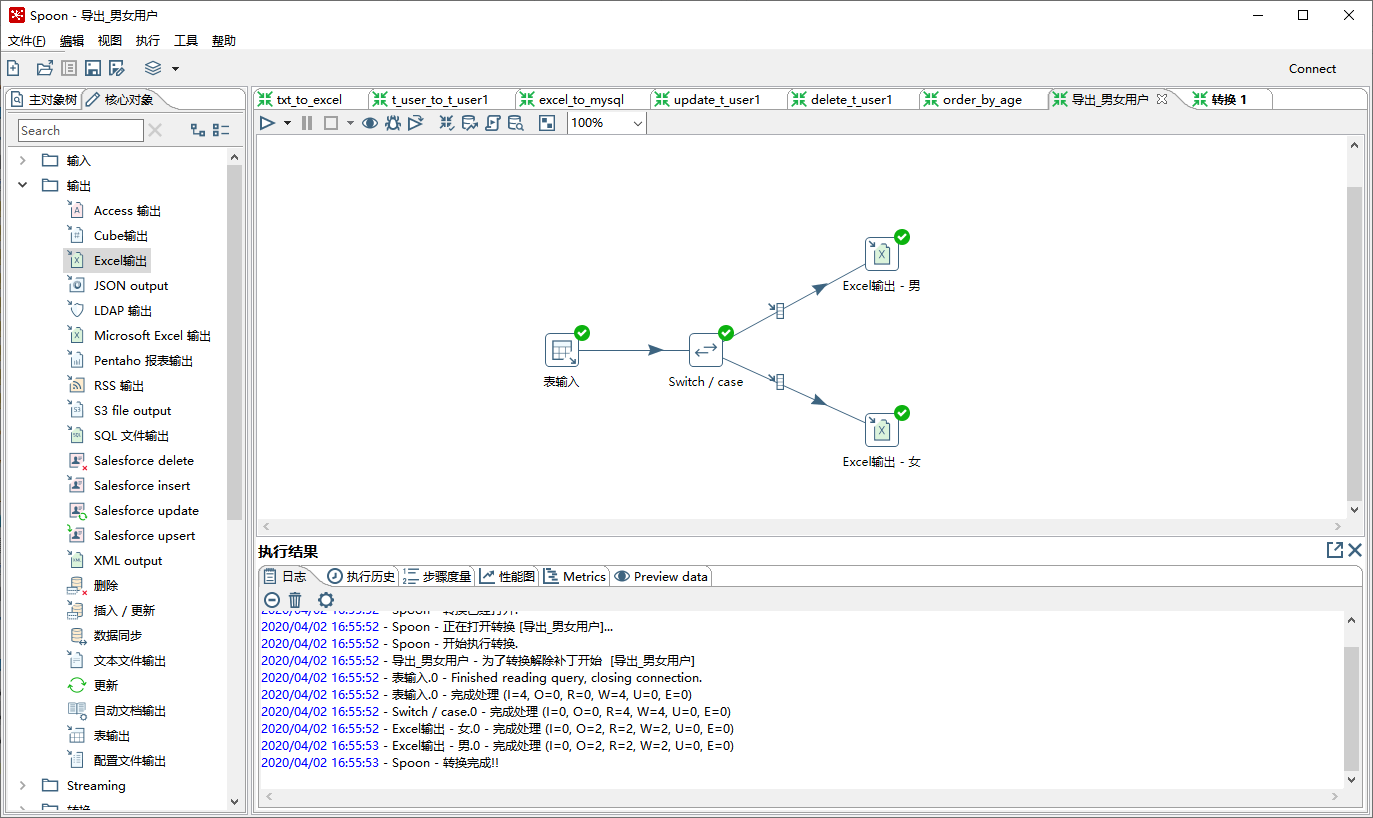
##### 配置switch/case组件

|  |
| --- |
| 1. 此处要指定，按照性别来判断输出到Excel文件。需要设置 Switch字段为gender，在Case值列表中将值分别改为男、女。如果gender是男的话，则将数据装载到 Excel输出 - 男，如果gender是女的话，将数据装载到 Excel输出 - 女。 |
|  |

##### 配置Excel输出组件

|  |
| --- |
| 1. 双击Excel输出组件，分别指定输出到指定的文件夹中。 |
|  |
|  |
|  |
|  |
|  |

#### 启动执行



Kettle会自动生成两个Excel文件，一个文件保存了性别为男的所有用户，另一个文件保存了性别为女的所有用户。

|  |  |
| --- | --- |
|  |  |
|  |  |

### SQL脚本组件

接下来，我们来讲解一个高级用法。在实际开发中，也经常容易使用得到。假设我们有一段SQL脚本，想要用Kettle来执行，此时该使用哪个组件呢？

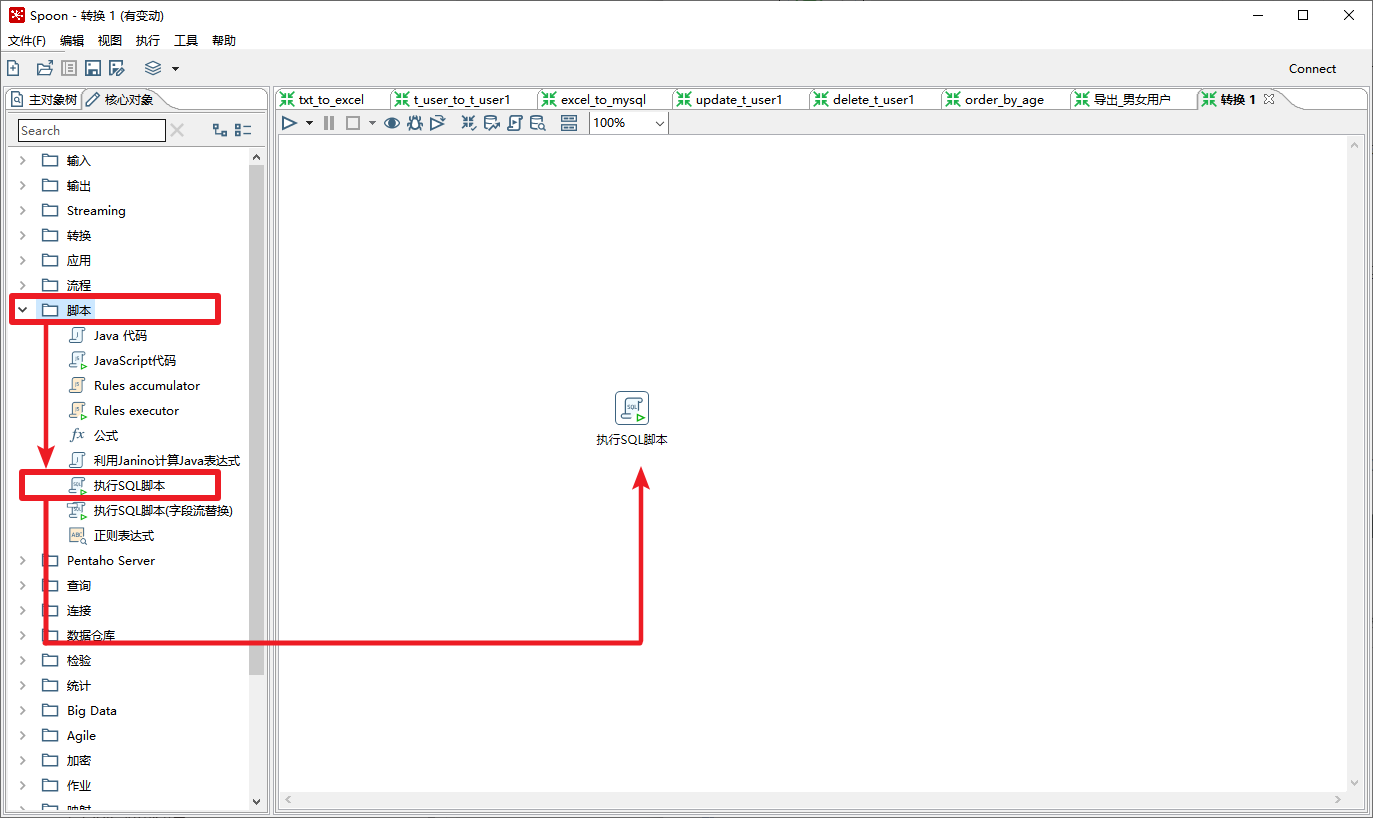
#### 组件介绍

执行SQL脚本组件，可以让Kettle执行一段SQL脚本。我们可以利用它来自动执行某些操作。

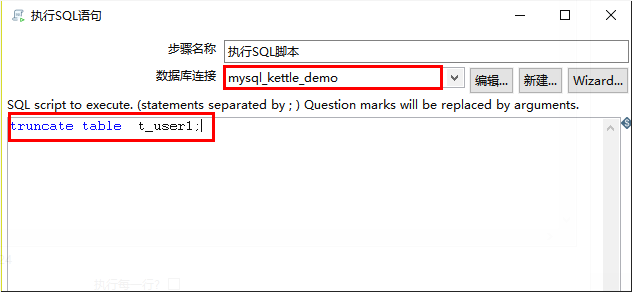
#### 需求

使用Kettle执行SQL脚本，将 t\_user1 表中的数据清空。

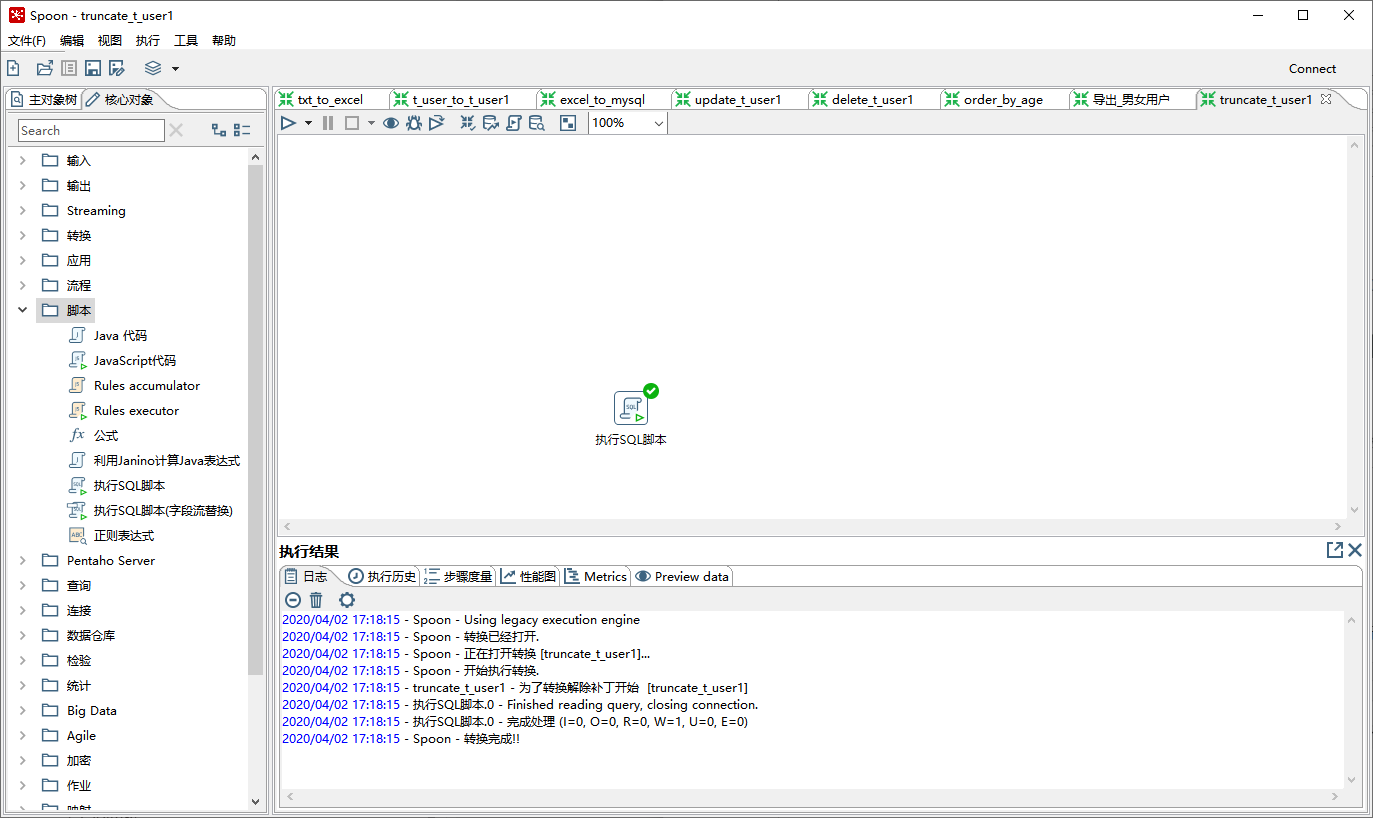
#### 构建Kettle数据流图



#### 配置Kettle数据流图组件参数



#### 启动执行



我们发现，t\_user1中的数据已经被完全清除掉了。说明：Kettle已经帮助我们执行了SQL脚本

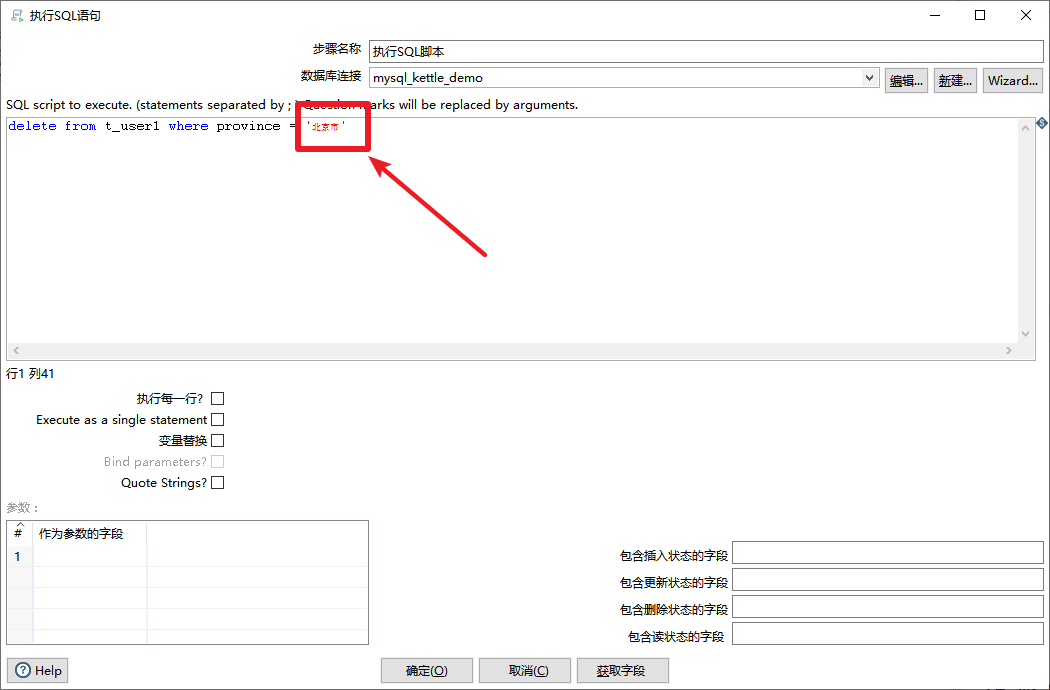


### 设置转换参数

#### 需求

让Kettle帮助我们删除指定省份的用户，具体删除哪个省份不确定，由执行Kettle时指定。

首先我们可以使用执行SQL脚本组件来删除某个省份的用户，类似下图：

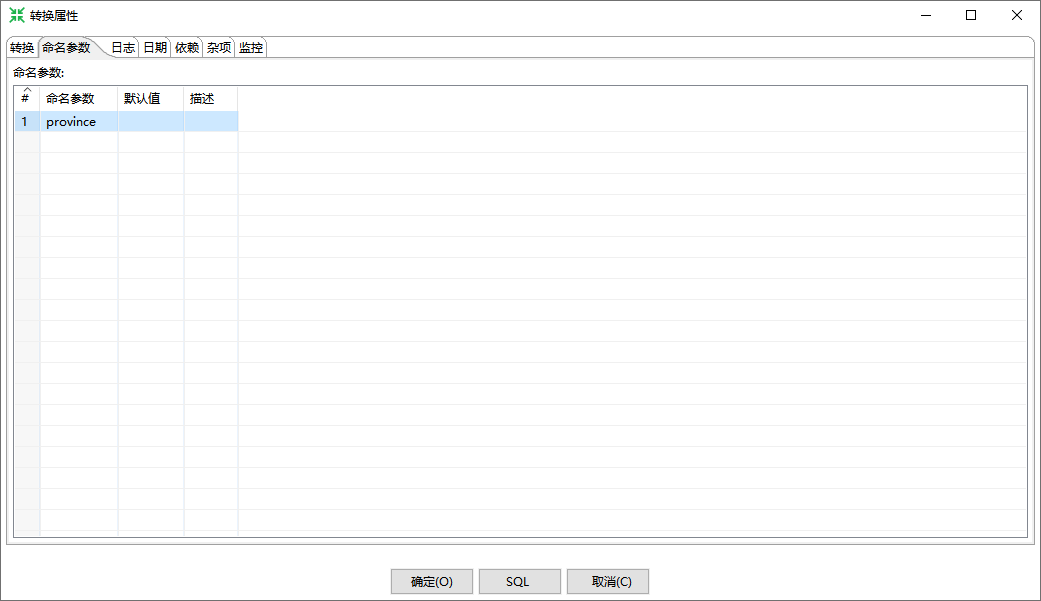


|  |
| --- |
| **delete** **from** t\_user1 **where** province **=** **'北京市'** |

但如果把北京市写在执行SQL脚本组件中，那么每次删除的都是北京市的用户。而此处的需求是，要执行Kettle转换时，我们需要手动来指定要删除哪个省份的用户。此时就需要用到Kettle中的参数了。

#### 参数的使用方法

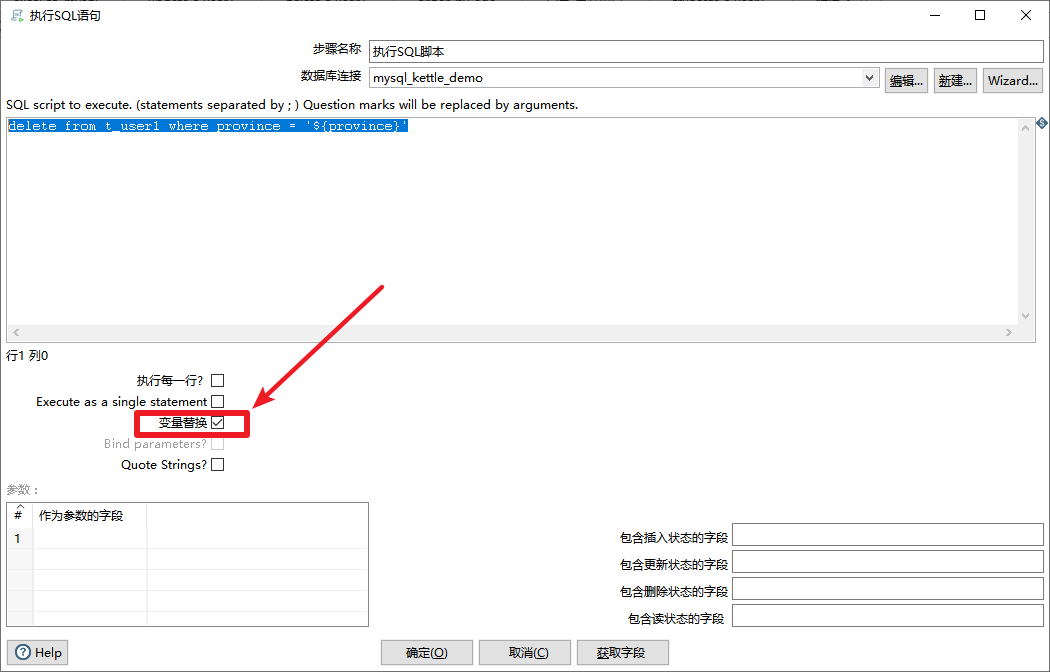
双击Kettle转换的空白处，会弹出转换属性窗口。我们可以在属性窗口中找到命名参数选项卡，在里面配置Kettle转换所需要的参数。例如：下图设置了一个叫做province的参数



#### 在SQL脚本组件中使用Kettle转换参数

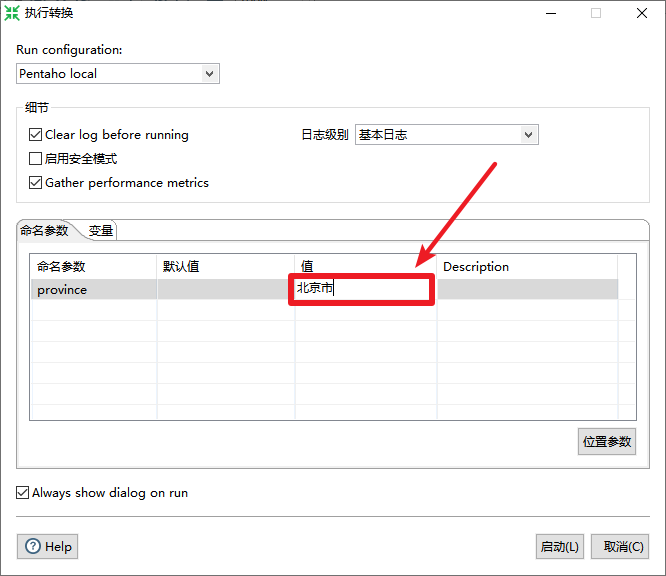
|  |
| --- |
| delete from t\_user1 where province = '${province}' |

通过 ${province} 可以应用Kettle配置的转换参数，而且一定要勾选上：「变量转换」

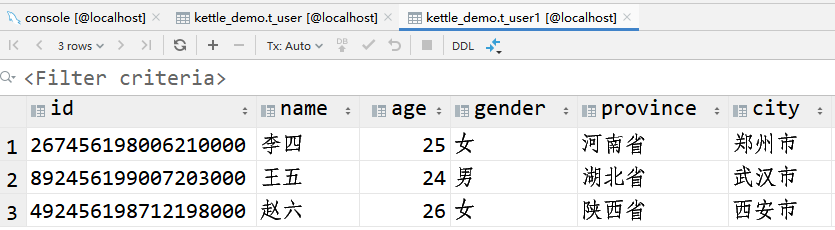


#### 运行带参数的转换

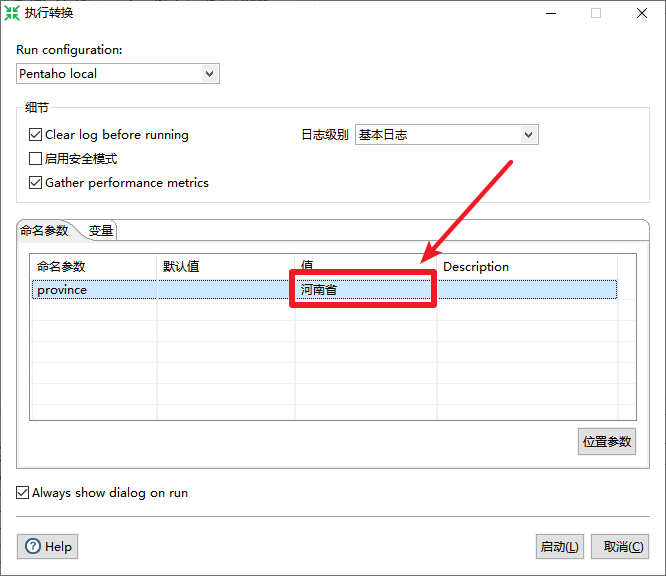
在运行转换时，可以指定参数的值，例如：此处指定要删除的城市为北京市。



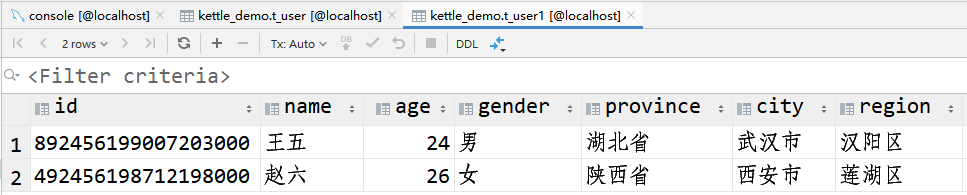
运行完后，我们发现：北京市的用户已经被删除了。



再来测试下删除河南省的用户：



同样，我们发现当执行完Kettle转换后，河南省的用户也被删除了。



## JOB（作业）开发

### 程序员小姐姐的第二次邂逅——JOB

前几天帮助程序员小姐姐小花解决了使用Kettle从Excel中抽取数据到MySQL问题，小姐姐特别高兴，请你吃了一顿饭，好一顿魂牵梦绕。小姐姐好几天都没有联系了，今天小姐姐又抱着电脑来找你了。你知道，又有新的问题来了。

|  |
| --- |
| 项目经理要求小姐姐小花能够每5秒钟执行一次Kettle转换，也就是每5秒钟将Excel中的数据抽取并装载到MySQL中。  怎么实现呢？ |

要实现这个需求，我们需要学习Kettle的JOB，也就是作业。

Kettle中的作业(job）定义了转换应该如何执行，可以配置转换来进行定时执行。

### JOB定时任务开发

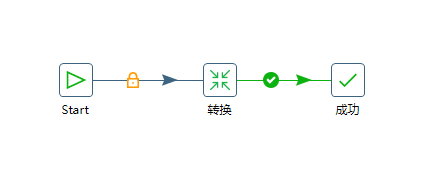
#### 需求

每5秒钟执行一次Kettle转换，也就是每5秒钟将Excel中的数据抽取并装载到MySQL中

#### 创建作业

#### 构建作业流组件图

效果图：



**开发步骤**

|  |
| --- |
| 1. 将核心对象中，通用文件夹下的 Start 组件拖拽到中间的空白区域。 |
|  |
| 1. 将 通用文件夹的「转换」组件拖拽到中间的空白区域。 |
|  |
| 1. 将通用文件夹中的「成功」组件拖拽到中间的空白区域。 |
|  |
| 1. 同样使用Shift键，将组件都连接起来。 |
|  |

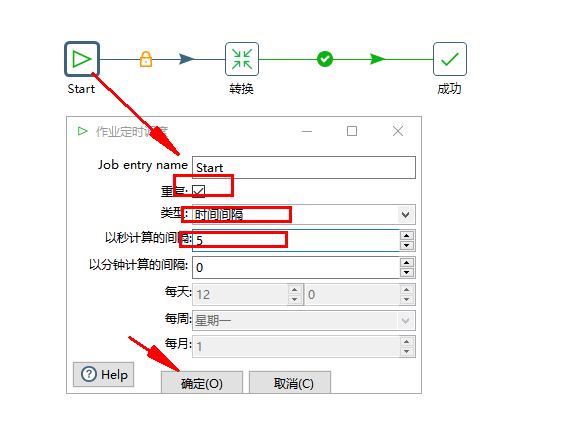
#### 配置作业流图组件

##### 配置转换组件

|  |
| --- |
| 配置转换这里选择作业中要执行的转换，此处选择之前开发好的excel\_to\_mysql.ktr即可 |
|  |
| 注意：此处要先保存作业，然后再配置转换。 |
|  |

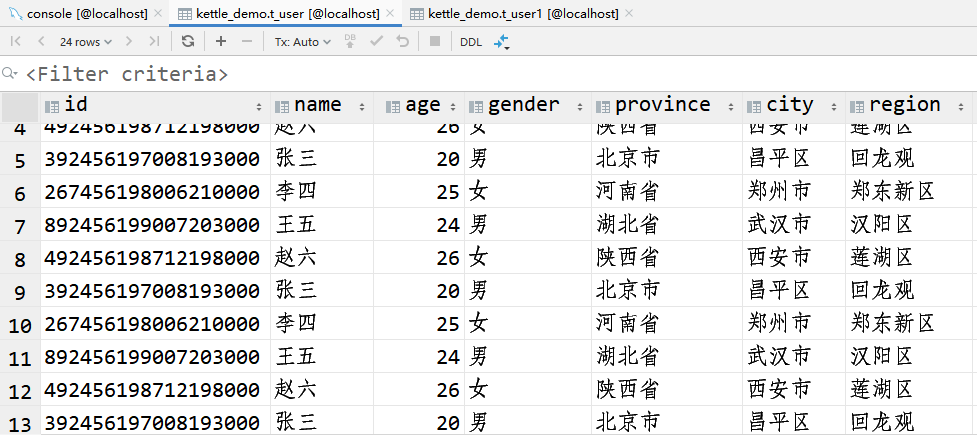
##### 配置Start启动组件

我们需要在Start组件中，配置定时执行。



#### 启动运行作业

点击播放箭头启动作业，并观察数据库中的数据是否会5秒钟增加一次。



我们看到数据每隔5秒钟就会增加一次。

## Kettle的Linux系统部署

### Linux安装Kettle

课程提供的Linux系统中已经安装好了Kettle，可以直接使用，kettle的安装文档请参考文档:06-安装Kettle.doc

### 9Pan转换执行引擎

pan.sh可以用来在服务器中执行一个转换

pan.sh的命令行参数:

-version：显示版本信息

-file: 指定要运行的转换文件（XML文件）

-level: 设置日志级别(Basic,Detailed,Debug,Rowlevel,Error,Nothing)

-log: 指定日志文件

-param:key=value （该参数可以指定多个）

需求：

在Linux中，将 /root/kettle/user.txt数据抽取到 /root/kettle/out\_user.xls 表格中

id,name,age,gender,province,city,region,phone,birthday,hobby,register\_date

392456197008193000,张三,20,0,北京市,昌平区,回龙观,18589407692,1970-8-19,美食;篮球;足球,2018-8-6 9:44

267456198006210000,李四,25,1,河南省,郑州市,郑东新区,18681109672,1980-6-21,音乐;阅读;旅游,2017-4-7 9:14

892456199007203000,王五,24,1,湖北省,武汉市,汉阳区,18798009102,1990-7-20,写代码;读代码;算法,2016-6-8 7:34

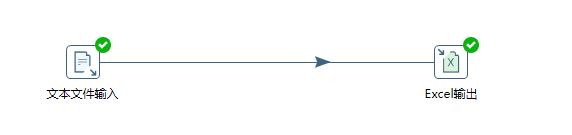
492456198712198000,赵六,26,2,陕西省,西安市,莲湖区,18189189195,1987-12-19,购物;旅游,2016-1-9 19:15

392456197008193000,张三,20,0,北京市,昌平区,回龙观,18589407692,1970-8-19,美食;篮球;足球,2018-8-6 9:44

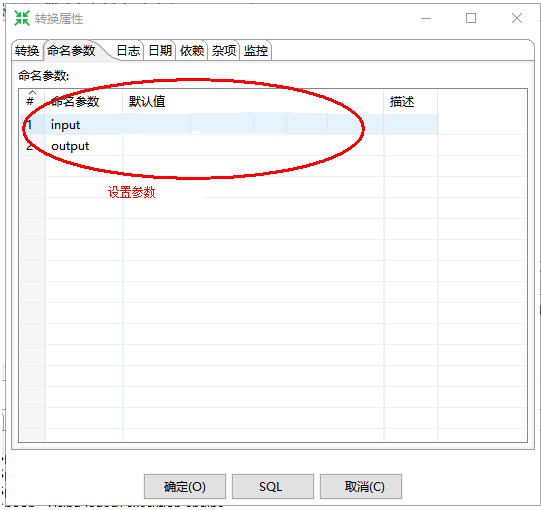
392456197008193000,张三,20,0,北京市,昌平区,回龙观,18589407692,1970-8-19,美食;篮球;足球,2018-8-6 9:44

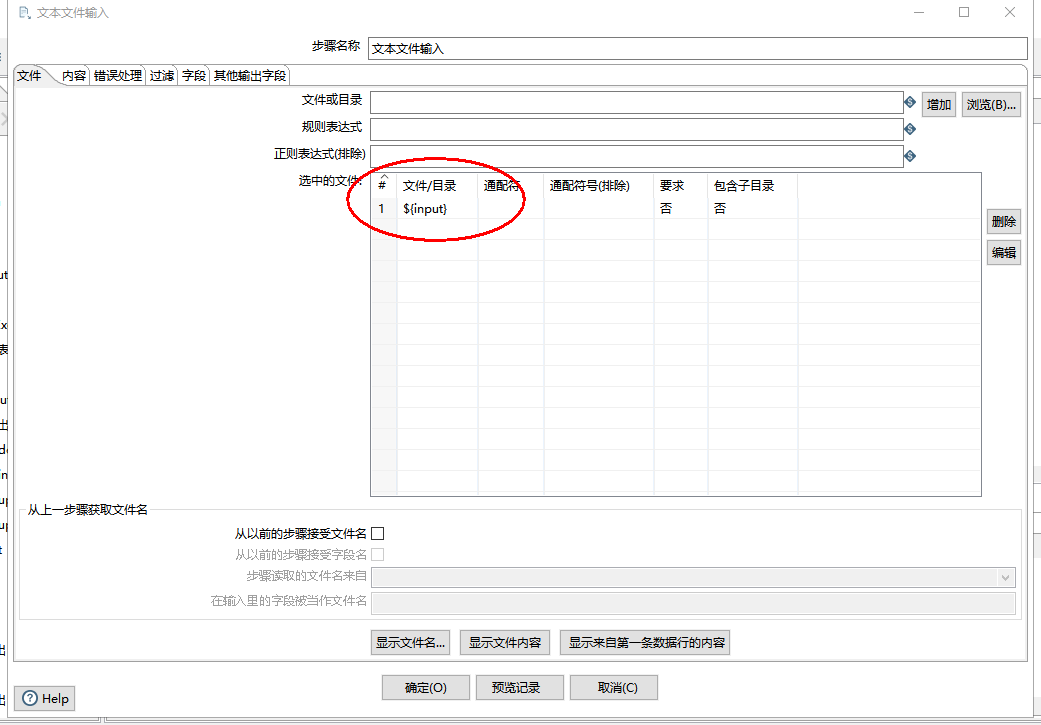
实现步骤：

1. 在 windows 中开发转换，将 文本文件数据抽取装载到Excel文件中

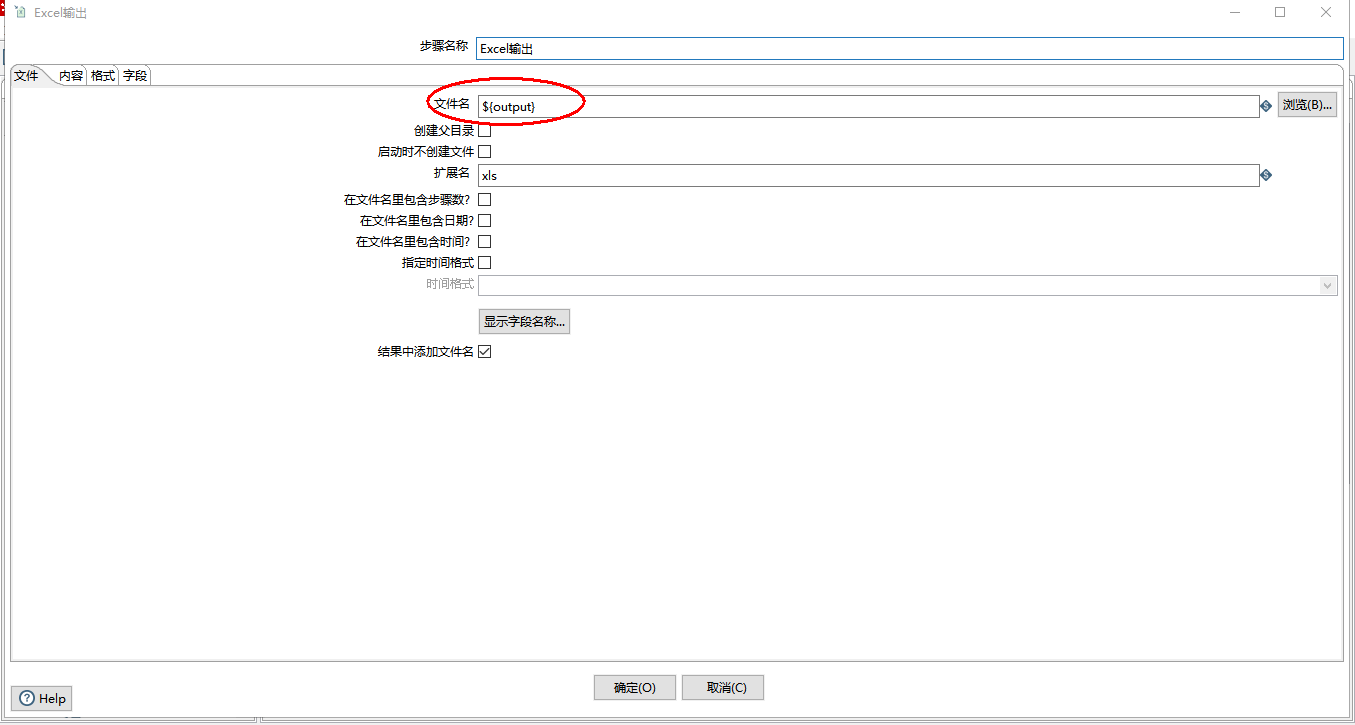


1. 抽取路径参数，通过参数指定输入路径和输出路径



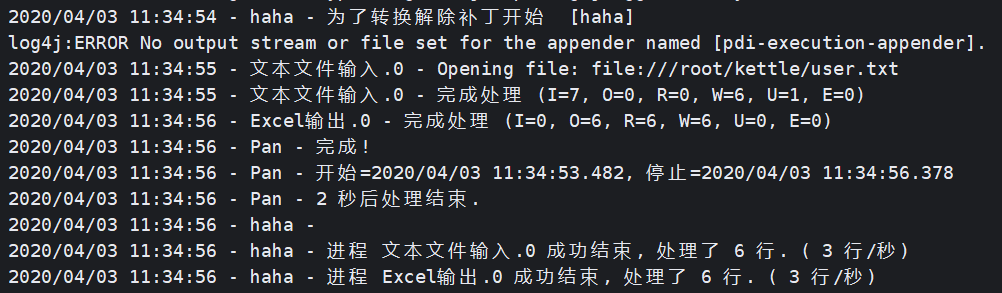






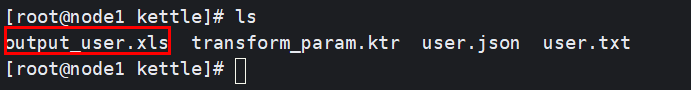
1. 将转换配置保存为transform\_param.ktr文件
2. 将数据文件user.txt和转换文件transform\_param.ktr都上传到 /root/kettle目录
3. 使用 pan.sh 执行转换

|  |
| --- |
| cd /root/kettle  pan.sh -file transform\_param.ktr -level Basic -param:input=/root/kettle/user.txt -param:output=/root/kettle/output\_user |



1. 结果查看

在/root/kettle目录会生成输出文件output\_user.xls



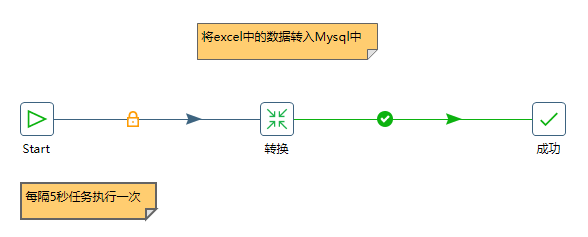
从Linux将该文件下载下来查看内容



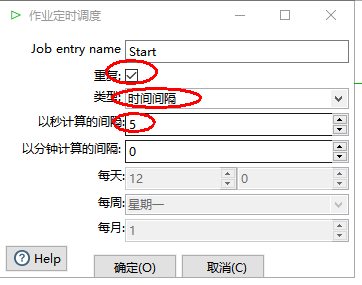
### Kitchen作业执行引擎

在Linux中对Kettle做Linux配置（和Windows相同，添加驱动jar包）

1. 在Windows中开发作业

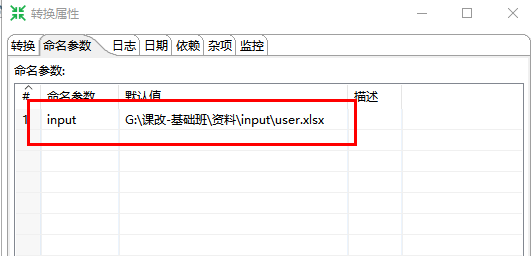


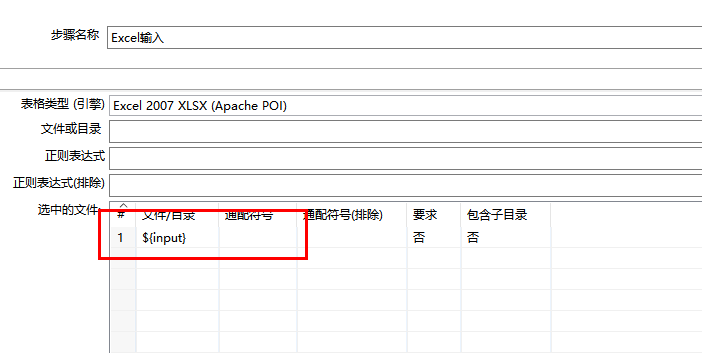
2、配置Start组件



1. 配置转换组件

修改4.5中Excel存入Mysql的转换配置如下





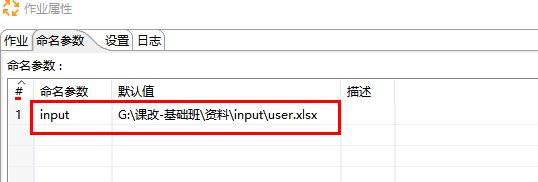
将转换文件保存为2-excel-mysql.ktr中

作业转换配置如下:

${Internal.Entry.Current.Directory}/2-excel-mysql.ktr



1. 配置作业命名参数

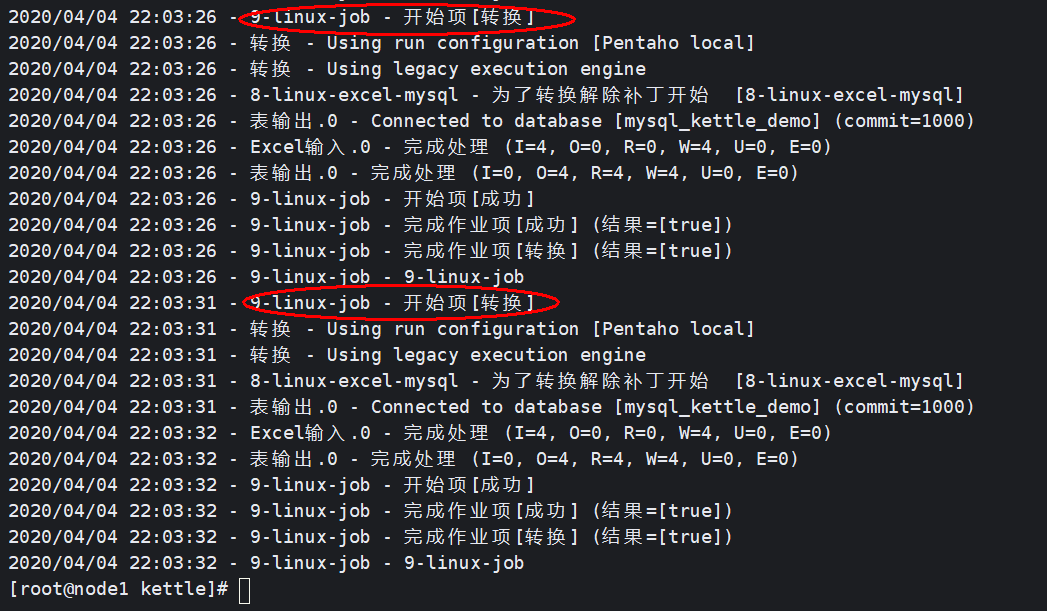


1. 将作业配置到9-linux-job.kjb文件中，并将该job文件和user.xlsx文件上传到Linux服务器的/root/kettle/目录,

7、使用kitchen.sh执行作业

|  |
| --- |
| cd /root/kettle  kitchen.sh -file 9-linux-job.kjb -level Basic -param:input=/root/kettle/user.xlsx |

打开可查看终端信息，发现每隔5秒钟执行一次转换任务



最后查看t\_user表，发现每隔几秒，数据会增多几行数据



## 问题1：错误连接数据库[MySQL]

|  |
| --- |
| 报错：无法连接到数据库。 |
|  |
| 解决办法：  将MySQL的驱动包复制到 pdi-ce-8.2.0.0-342\data-integration\lib 中，然后重启kettle |