# 实验1.1：大数据采集——清洗技术Kettle使用

## **1实验目的**

1. 了解Kettle的基本作用，下载并配置工具；
2. 使用Kettle对数据进导入到指定数据库数据表中；
3. 使用Kettle对数据进行简单清洗操作，并输出。

## **2实验内容**

1. 操作演示Kettle的下载和配置；
2. 操作演示了使用Kettle对数据进行一系列的操作；
3. 操作演示了使用Kettle对数据进行清洗。

## **3实验过程**

**任务一:安装配置Kettle**

打开浏览器，在浏览器地址栏中输入地址:“kettle.pentaho.org”，找到Downloads，在Older versions中找到我们需要的7.0版本。如图所示

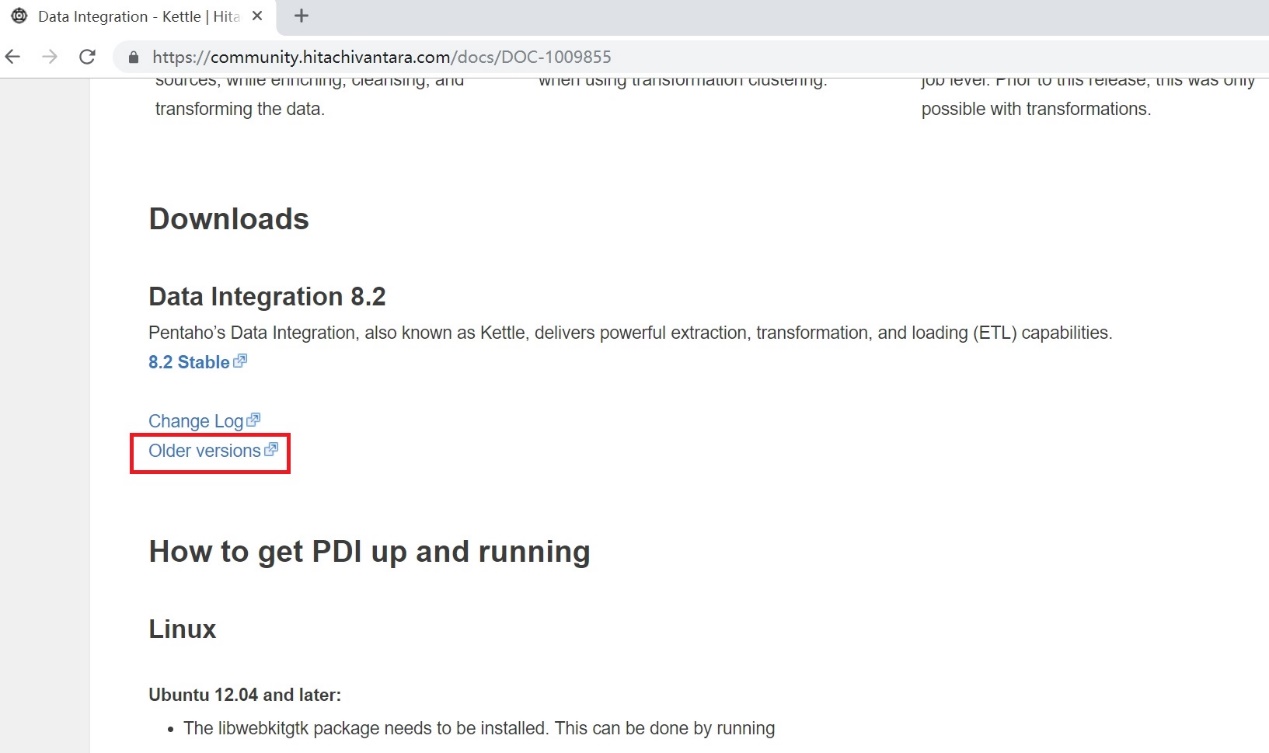


图2-1 Kettle下载页面

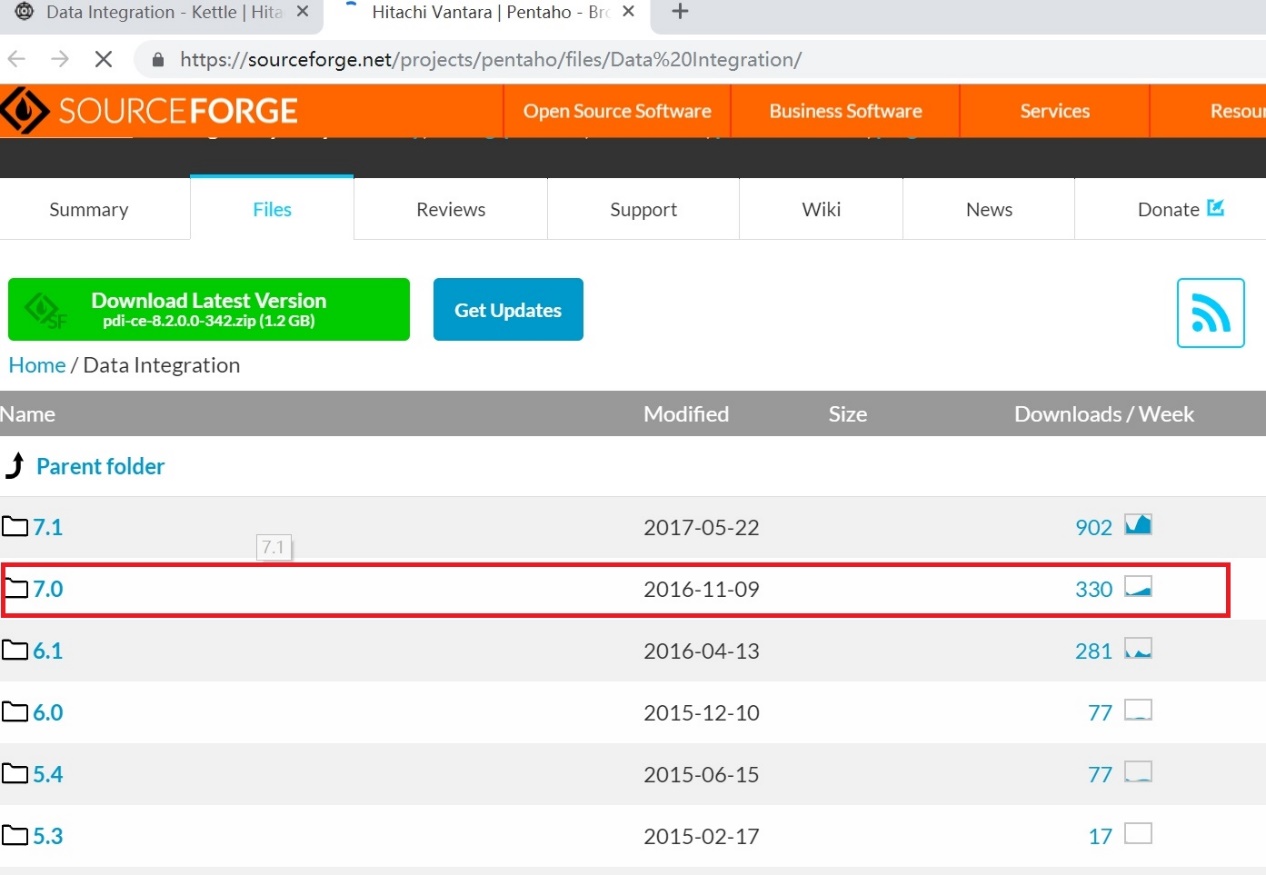


图2-2 Kettle下载页面



图2-3 Kettle下载页面

找到下载好的Kettle文件，并在kettle文件中找到“Spoon.bat”文件。如图所示

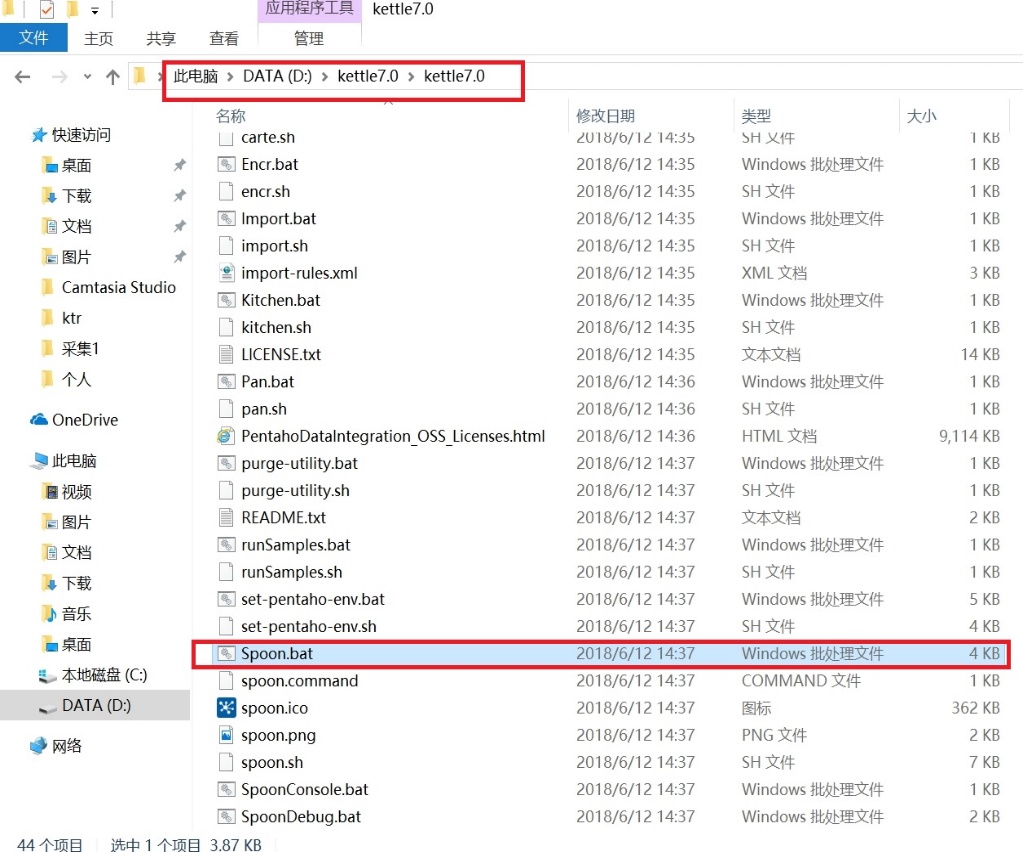


图2-4 Kettle文件

**任务二:使用Kettle对数据操作**

双击“Spoon.bat”文件运行，进入Kettle初始界面，如图所示

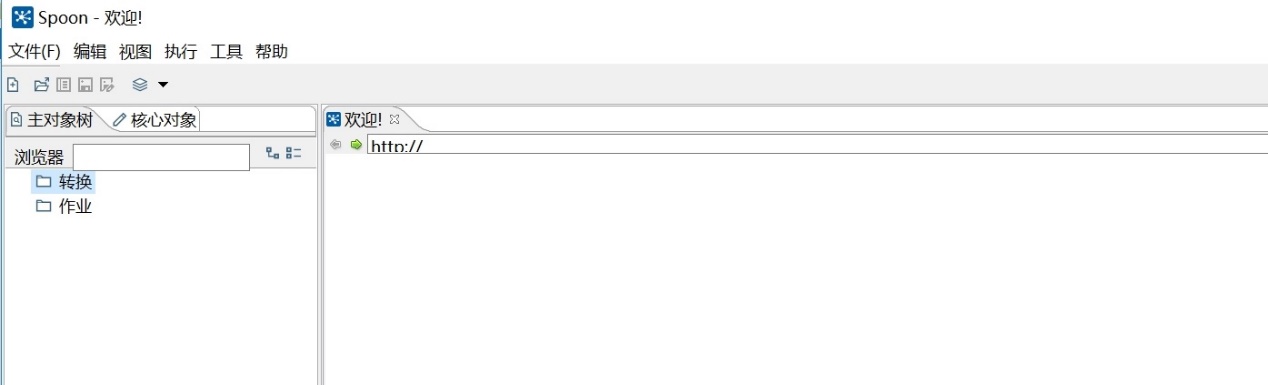


图2-5 Kettle初始界面

打开Navicat中实现建立demo数据库中的ods\_carinfo数据表，选中右键清空表，刷新后数据表为空，如图所示

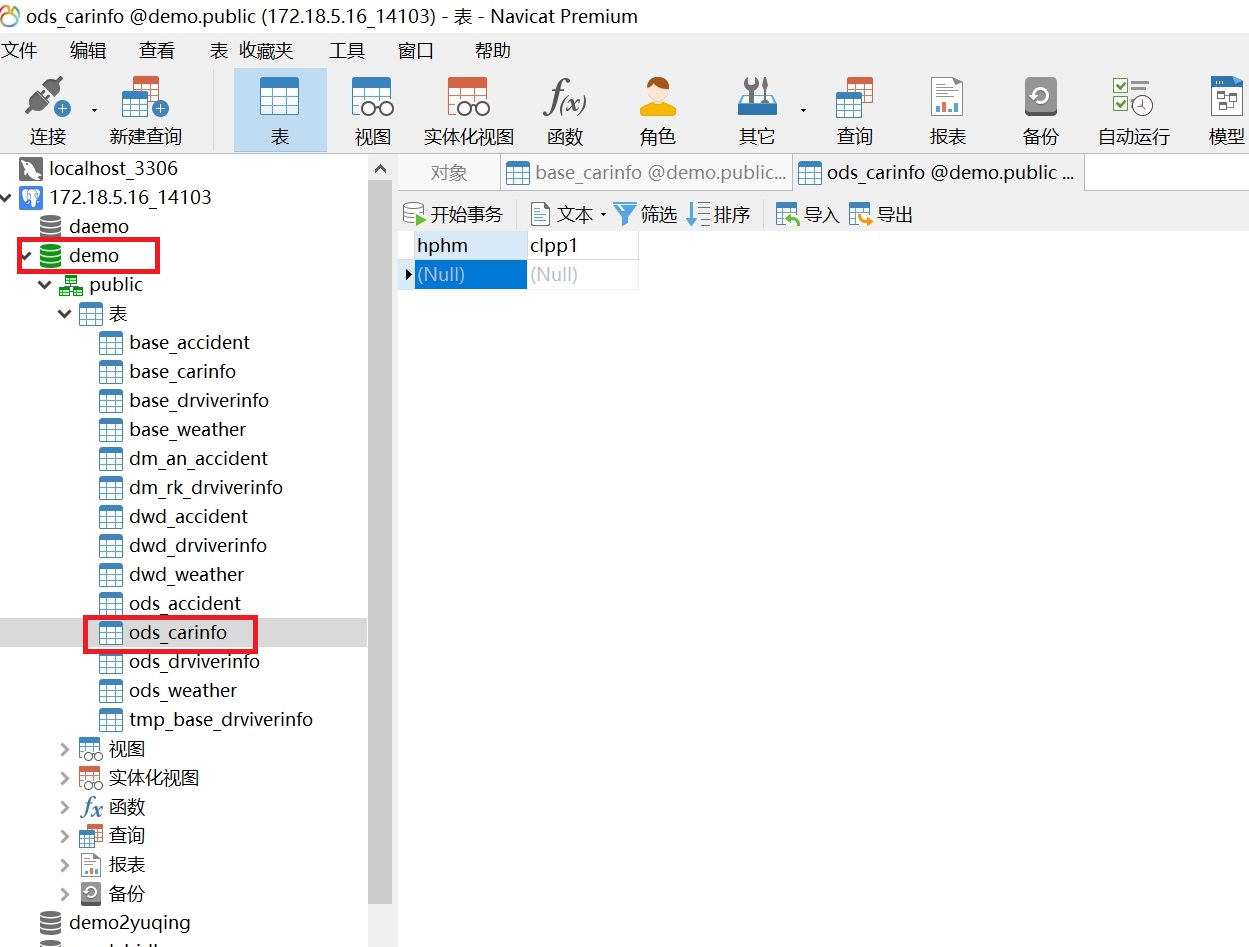


图2-6 清除数据表内容

打开Kettle，点击“文件”再点击“新建”、“作业”。首先添加核心对象，点击“通用”，选择“START”、 “成功”、“转换1”、“转换2”四个对象拖动至右边区域，在脚本目录中选择“SQL”脚本，拖动至工作区域。然后我们建立执行顺序，按住“shift”键不放，鼠标左键从“START”拖动至“SQL”上，同理可以建立完执行顺序，按住鼠标左键不放，选中所有核心对象，点击右键选择排列中的“顶端排列”如图所示

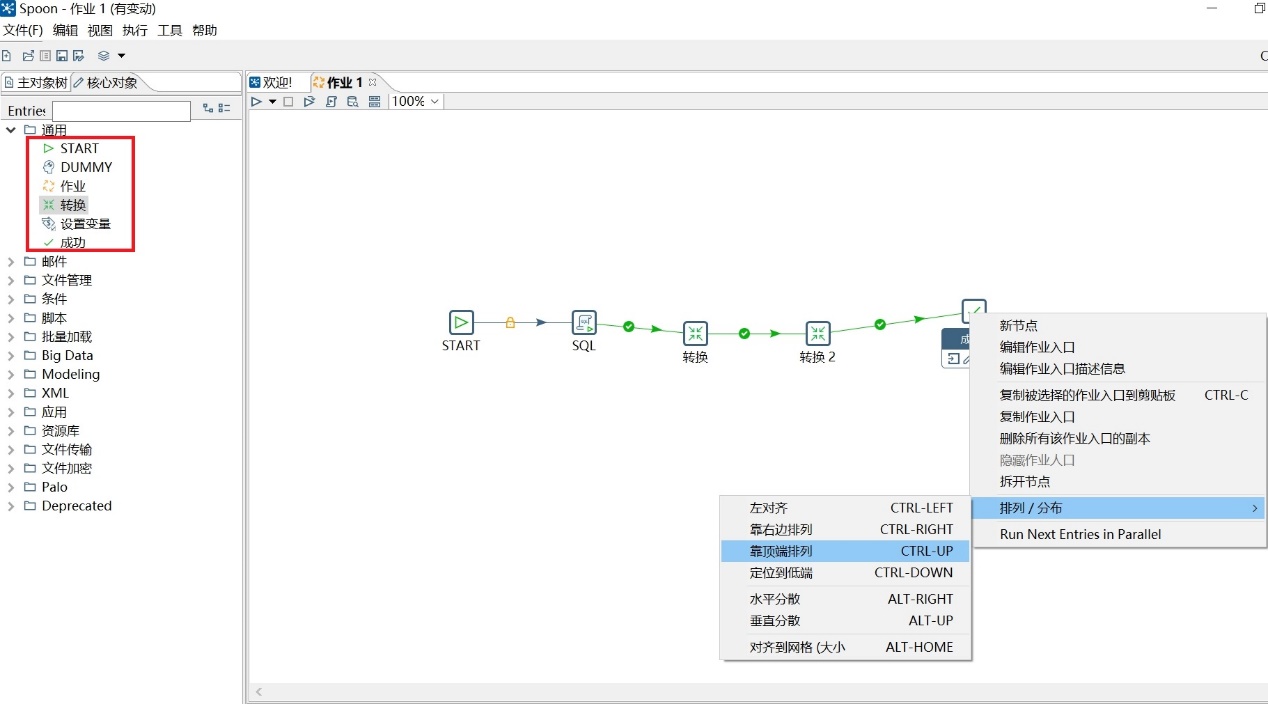


图2-7 作业流程图

选择主对象树，双击“DB连接”，选择连接类型“PostgreSQL”，选择连接方式“Native（JDBC）”,输入主机名称“192.18.5.16（请更改为自己的主机，下同）”，输入数据库名称“demo”、端口号“14103”、用户名及密码“stork”，连接名称“172.18.5.16”，点击测试，测试结果连接正确，如图所示。【备注：这里主机名称，数据库名称，端口，用户名及密码等根据自己配置进行相应设置，下列涉及的这些内容跟你所配置的资源相一致】

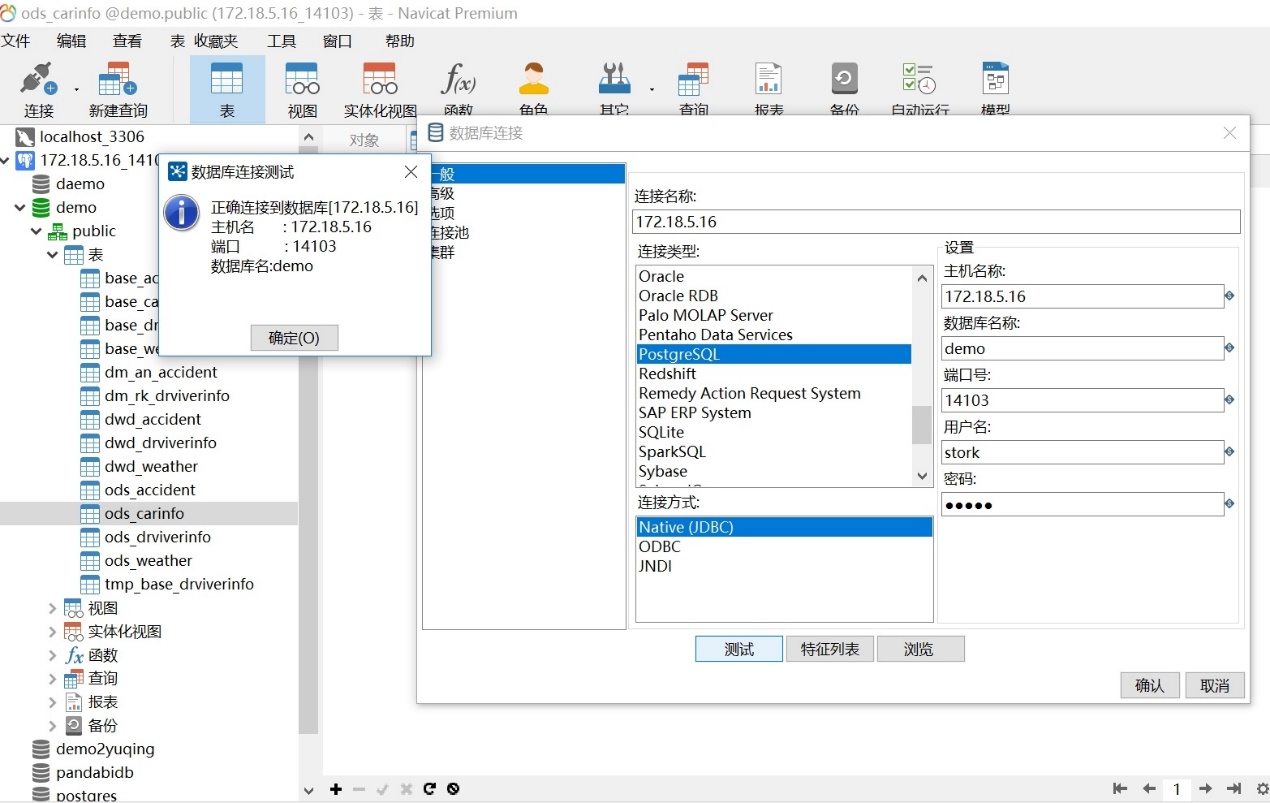
****

图2-8 建立数据库连接

双击作业1中的“SQL脚本”，输入作业项名称“删除表，新建表”，输入数据库连接“172.18.5.16”，输入事先写好的SQL脚本，点击“确定”。如图所示

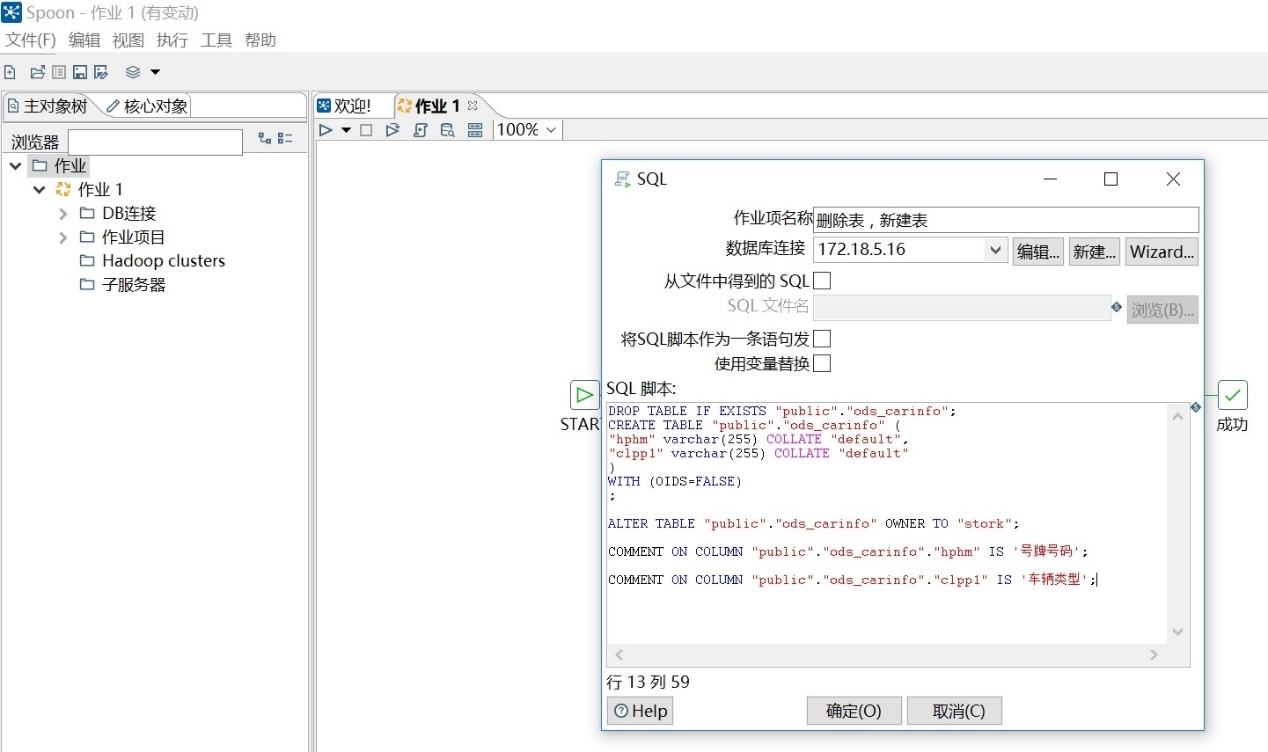


图2-9 定义删除表和新建表

选择“转换”，再点击“文件”、“新建”、“转换”，拖动输入目录下的“CSV文件输入”、转换目录下的“拆分字段”、输出目录下的“表输出”到右边工作页面。重复上文建立执行顺序操作，如图所示

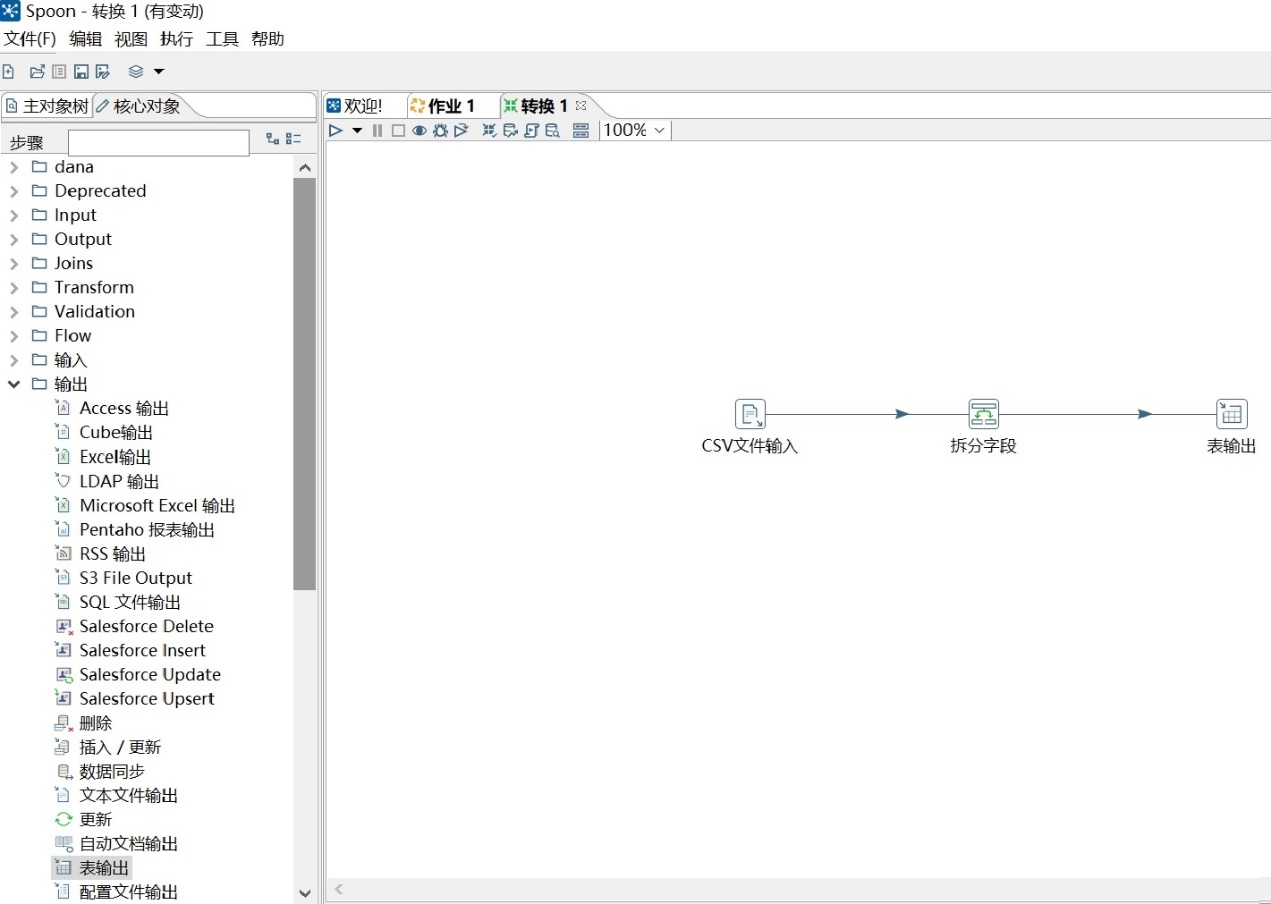


图2-10 定义转换执行流程

双击“DB连接”重复上文建立数据库连接步骤，建立数据库连接如图所示

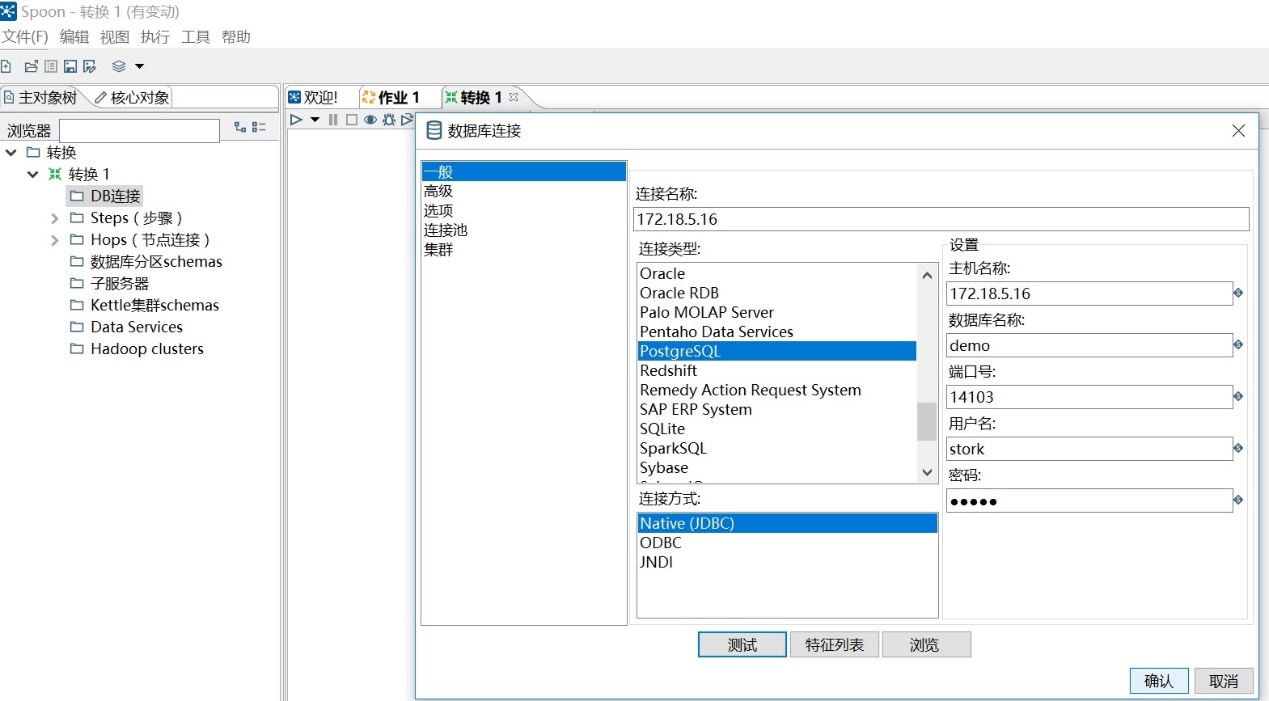
****

图2-11 建立转换1数据库连接

双击“CSV文件输入”对象，定义对象内容，输入步骤名称“carinfo文件输入”，选择已准备好的文件名，使用“GBK”文件编码，输入抽取对象名称“hphm；clpp1”，抽取类型“String”，如图所示



图2-12 定义“CSV文件输入”对象内容

点击“预览”，成功后点击“保存，如图所示



图2-13 输入文件预览

双击拆分字段进行定义，输入步骤名称“拆分字段”，输入分隔符“；”输入新字段1“hphm”新字段2“clpp1”，选择类型“String”选择去除空格类型“去除两端空格，保留存数据信息”。如图所示



图2-14 定义“拆分字段”对象内容

双击“表输出”对象，选择数据库连接“172.18.5.16”，点击目标表后面的“浏览”按钮，选择数据表“ods\_carinfo”，点击“确定”，选择目标模式“public”，勾选“指定数据库字段”，点击“数据库字段”，点击“获取字段” 点击“确定”，如图所示



图2-15 定义“表输出”对象内容

点击“文件”，保存命名为“carinfo\_提取”，点击“保存”。打开“作业1”双击“转换1”，输入作业项名称“carinfo\_提取”，选择转换文件名“carinfo\_提取”，点击“确定”。如图所示

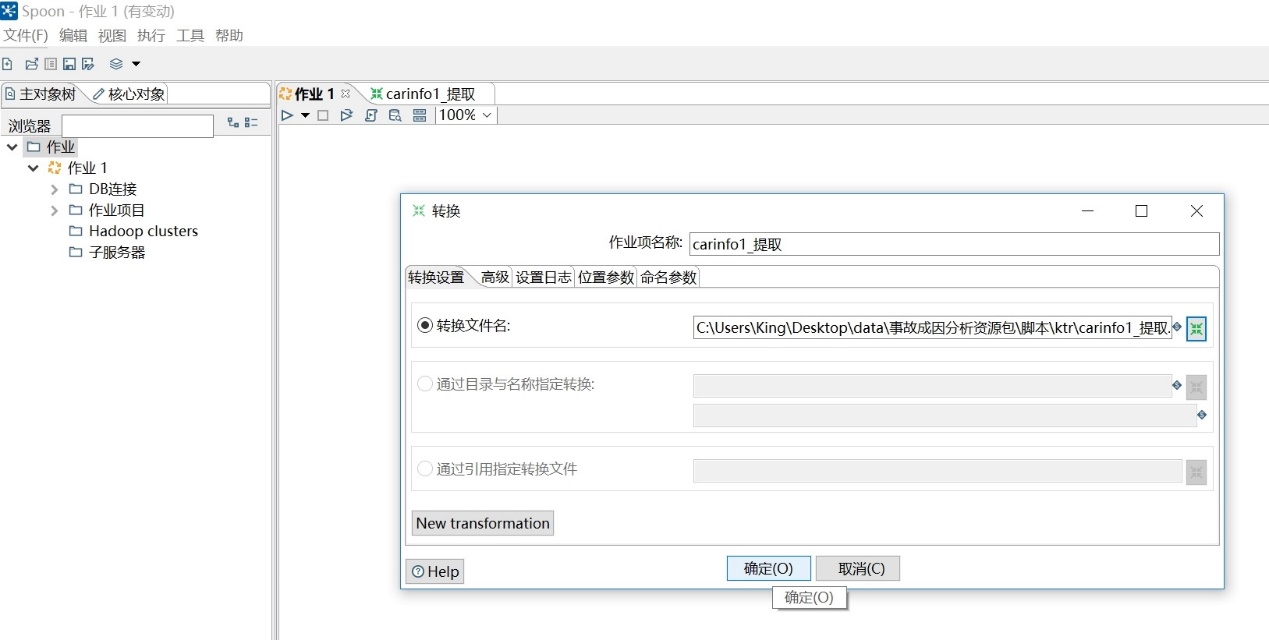


图2-16 定义“carinfo\_提取”对象内容

同理进行转换2的定义，定义方法与上文一致。点击“文件”、“新建”、“转换”。选着对象“获取文件名”、“CSV文件输入”、“拆分字段”、“表输出”，并将其拖至右边，按住“shift”键不放，建立执行顺序，并排列。如图所示

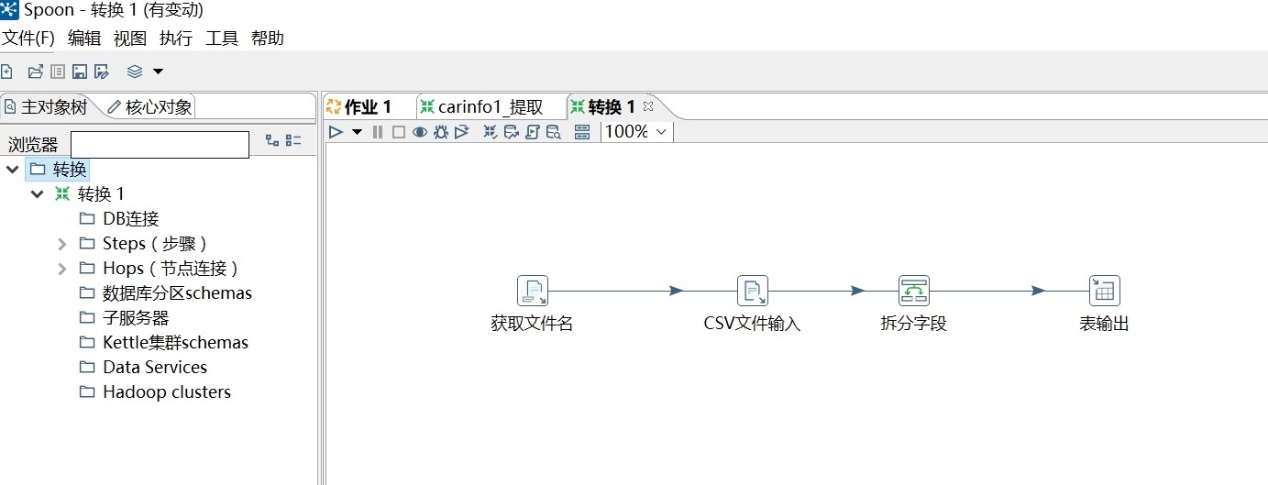


图2-17 定义转换2执行流程

双击“DB连接”，选择连接类型“PostgreSQL”、连接方式“Native（JDBC）”,输入主机名及用户名“172.18.5.16”，数据库名称“demo”，端口号“14103”，用户名及密码“stork”，点击测试。如图所示

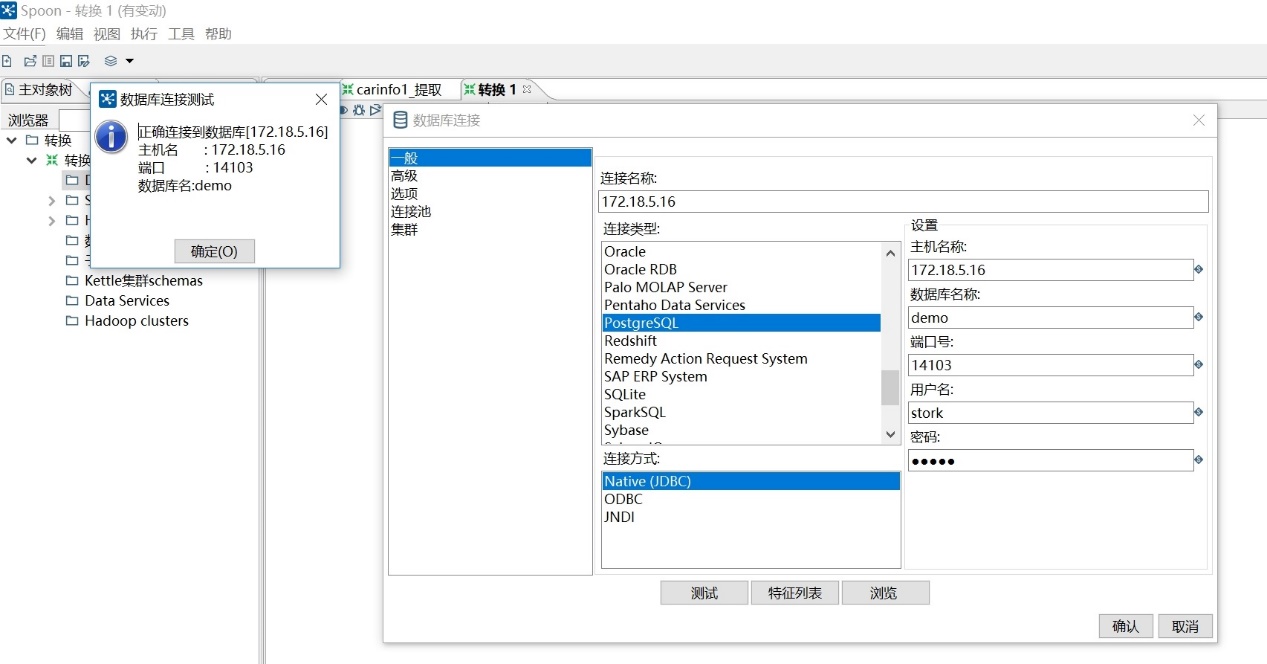


图2-18 建立转换2数据库连接

双击“获取文件名”，将文件所在目录粘贴到“文件/目录”，输入通配符“^carinfo.\*”、通配符（排除）“carinfo1.csv”,点击”显示文件名称”,如图所示

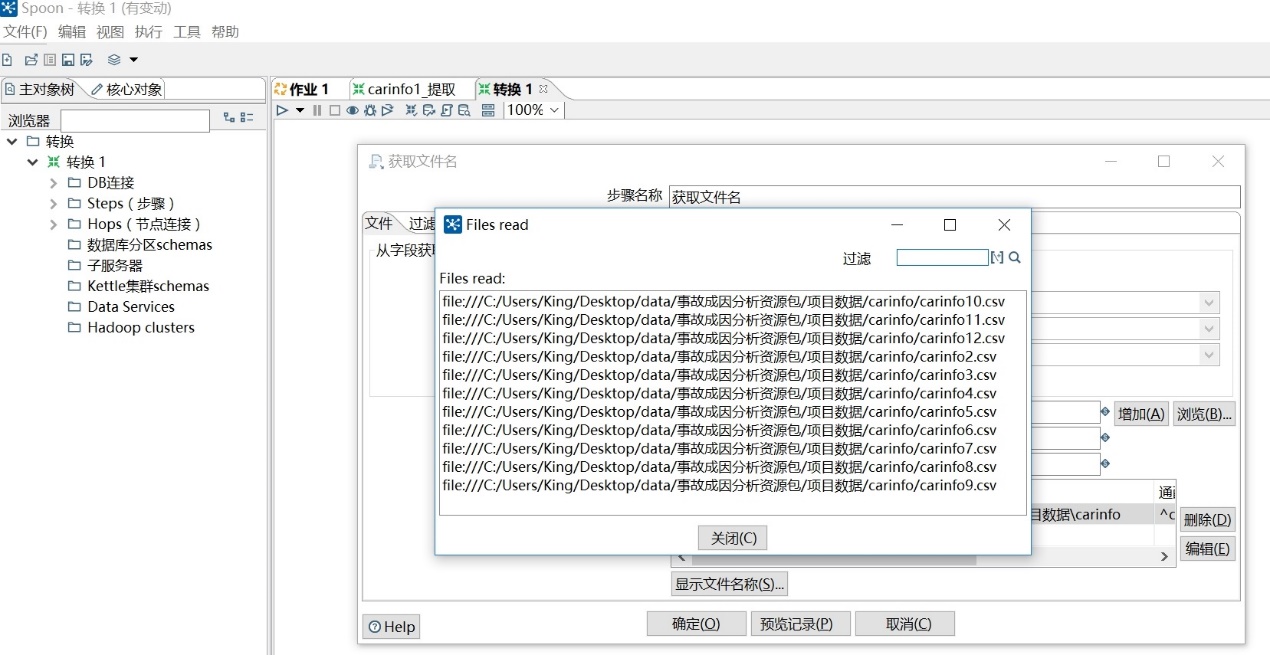


图2-19 获取文件名

双击“CSV文件输入”,选择文件名字段“filename”,文件编码“GBK”,输入名称“hphm；clpp1”、选择类型“String”，选择去掉空格类型“不去掉空格”。点击确定。如图所示



图2-20 定义“CSV文件输入”对象

双击“拆分字段”，选择“需要拆分的字段“hphm；clpp1”、输入分隔符“，”输入新的字段1“hphm”、字段2“clpp1”，类型为“String”，去除空格类型为“去掉左右两端空格”，点击“确定”。如图所示



图2-21 定义“拆分资源”对象

双击“表输出”，选择目标表“ods\_carinfo”，输入目标模式“public”，勾选“指定数据库字段”，并在数据库字段目录下点击“获取字段”，点击“确定”，点击“文件”，选择“保存”，命名为“carinfo\_提取2”。如图所示

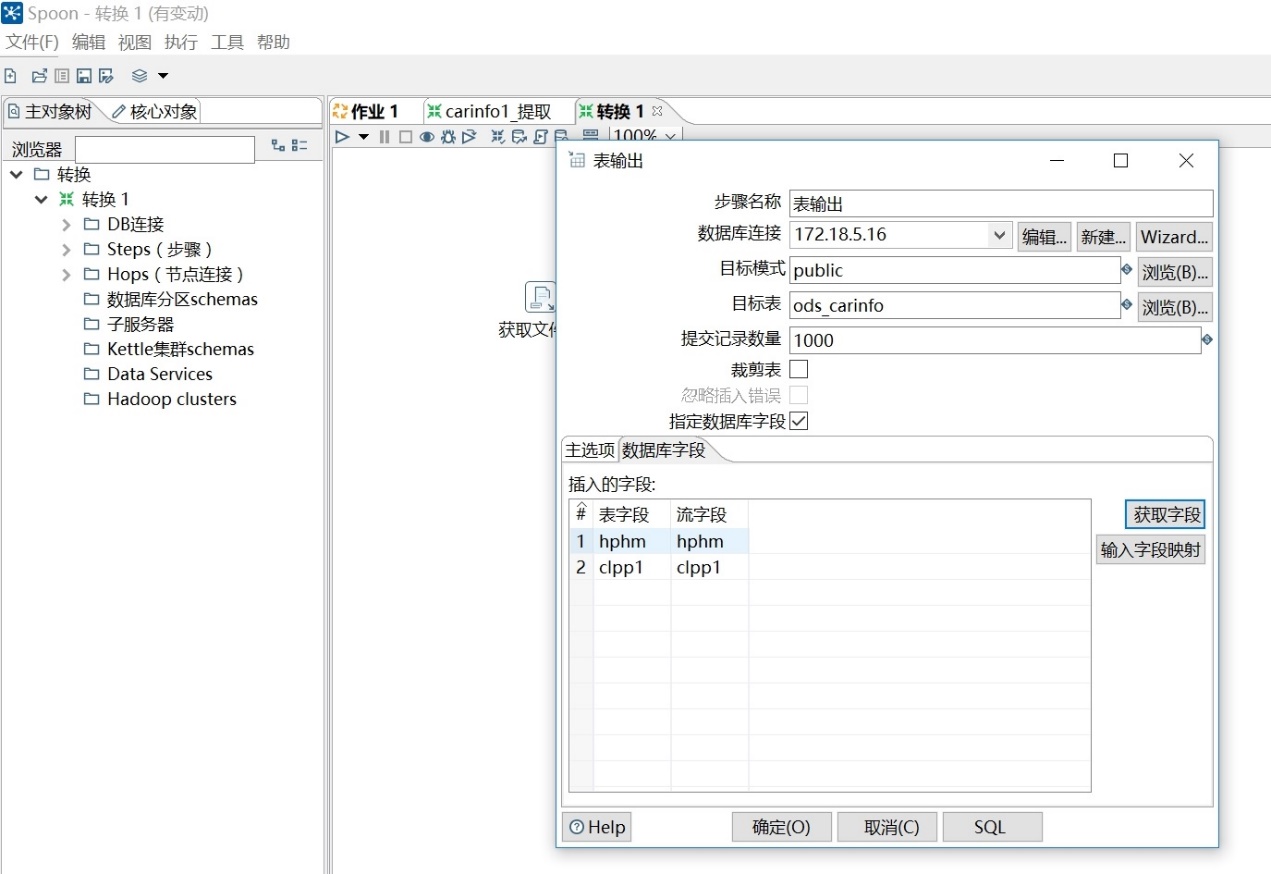


图2-22 定义“表输出”对象

双击 “转换2”对其命名。输入作业项名称“carinfo2提取”，选择转换文件名“carinfo\_提取2”，点击“确定”，点击“文件”，选择“保存”，命名为“ods\_carinfo提取”。如图所示

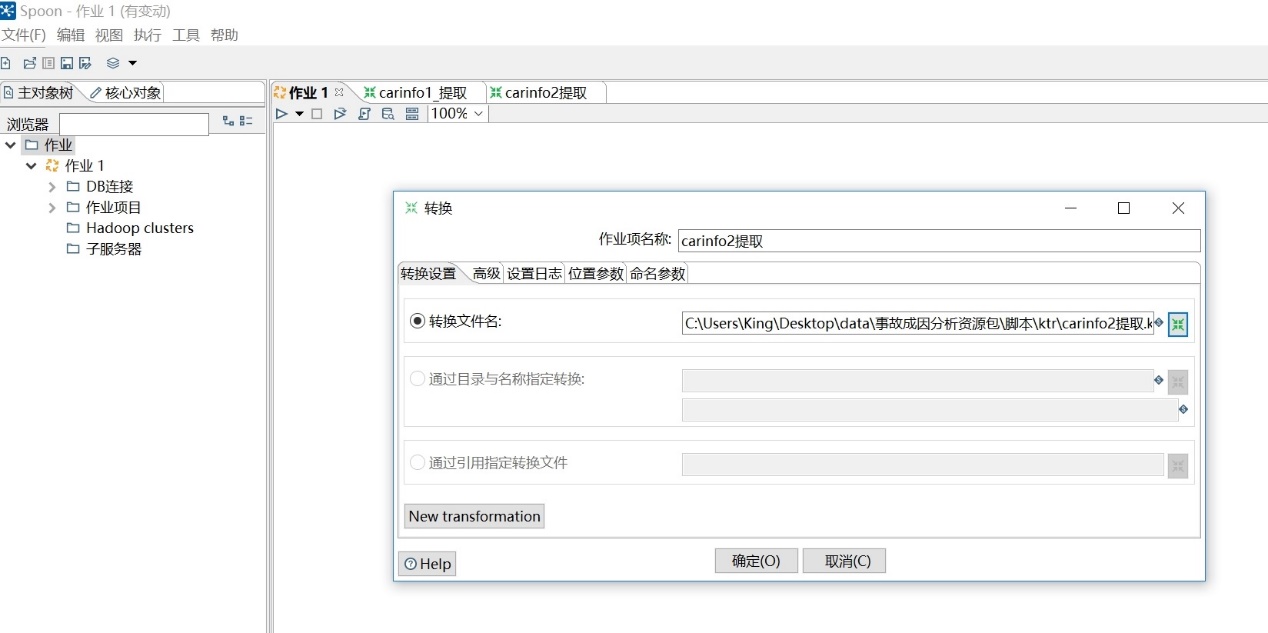


图2-23 对“转换2”命名

点击“执行”选项，点击“运行”，查看试验结果

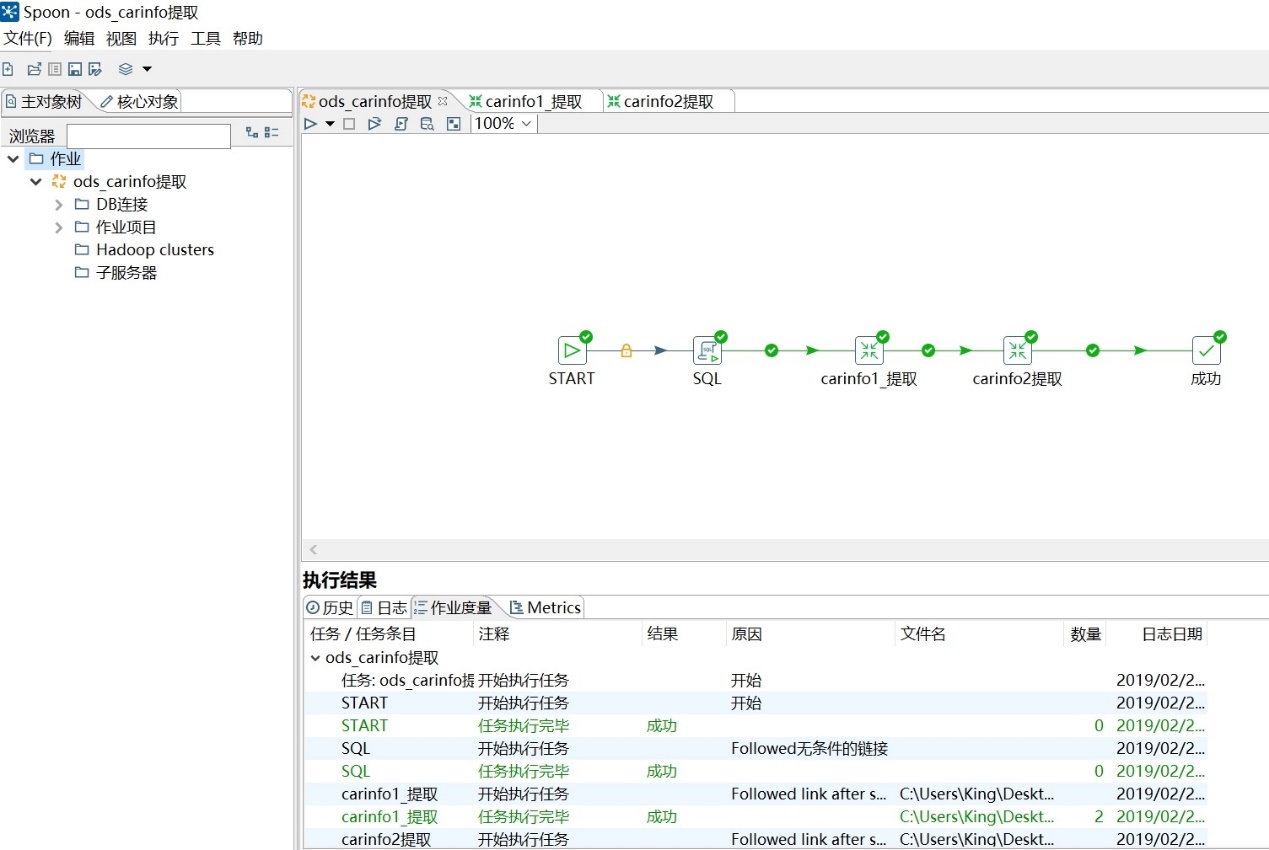


图2-24 Kettle试验执行结果

点击“Navicat”，刷新“ods\_carinfo”数据表，试验结果如图所示

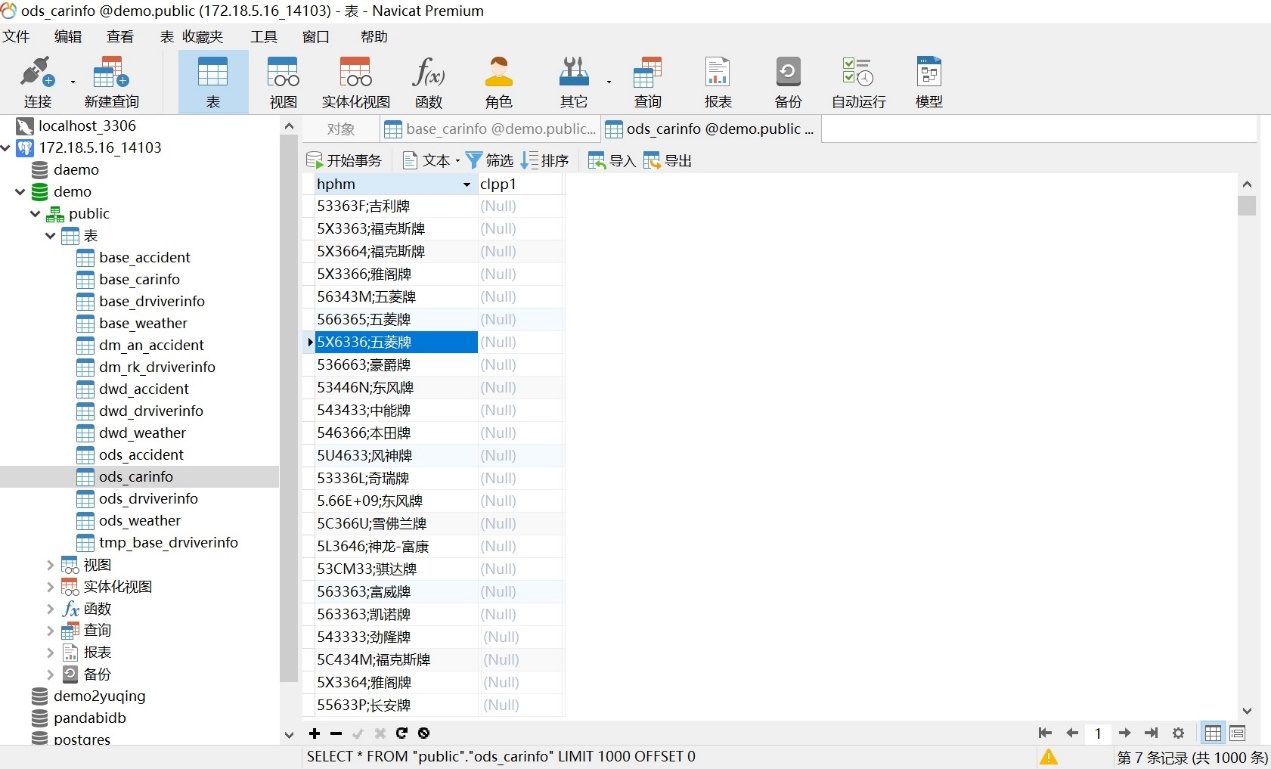


图2-25 Navicat查看实验执行结果

**任务三:使用Kettle对数据进行清洗操作**

点击打开“Kettle”，点击“文件”，选择“新建转换”，点击“主对象树”，双击“DB连接”，输入连接名称“172.18.5.16”，选择连接类型“PostgreSQL”,连接方式“Native（JDBC）”,输入主机名称“172.18.5.16”、数据库名称“demo”、端口号“14103”、用户名及密码“stork”，点击测试。如图所示

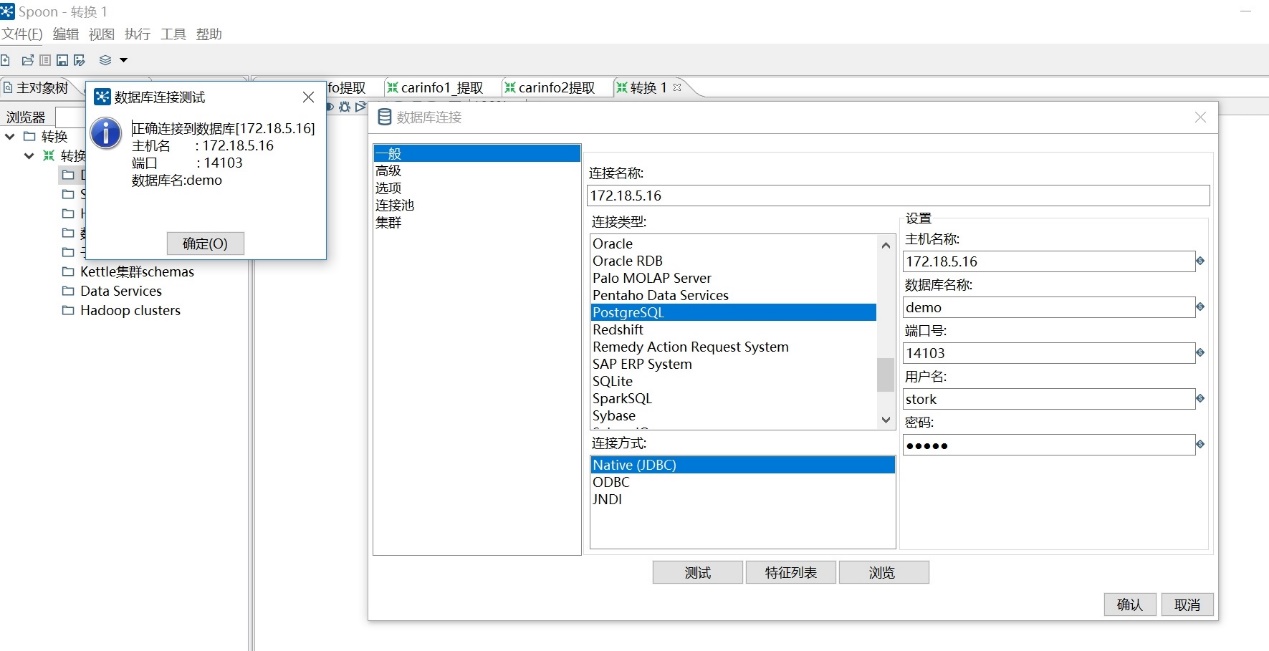


图2-26 新建数据库连接

点击“核心对象”，选择“表输入”、“增加校验列”、“排序对象”、“去重复记录”、“获取系统信息”、“表输出”，拖至右边的工作区。并建立执行顺序。如图所示



图2-27 新建转换

双击“表输入”，选择数据库连接“172.18.5.16”，将准备好的SQL语句粘贴到SQL中，点击预览，输入预览词条“10”。如图所示

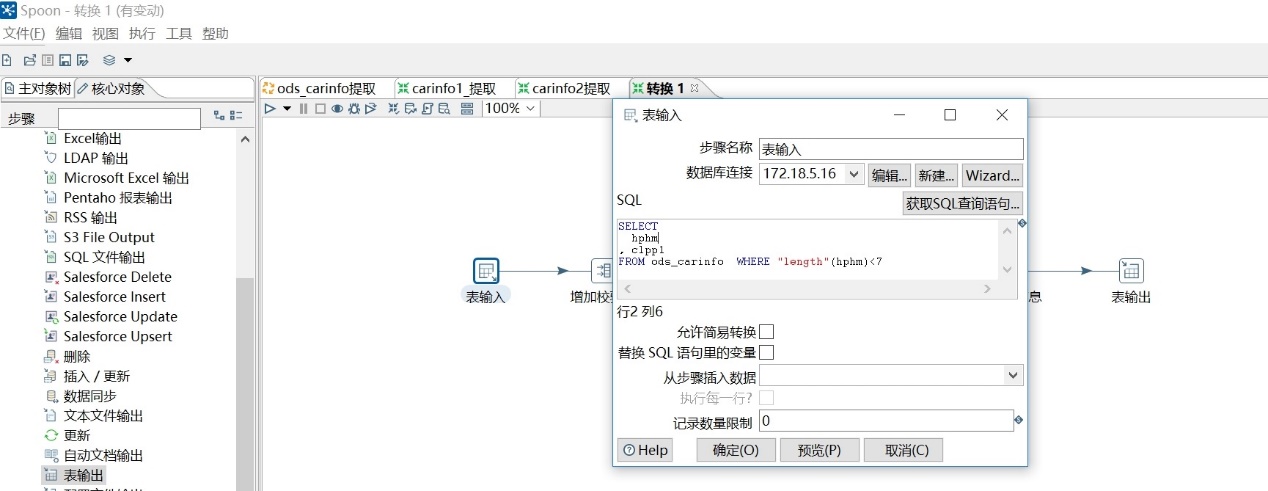


图2-28 定义“表输入”

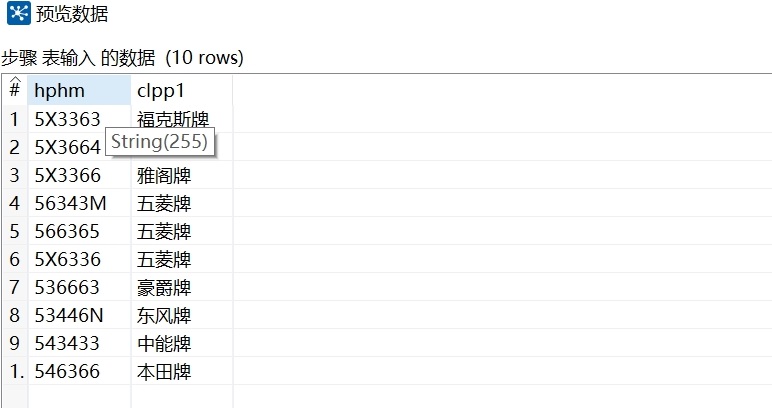


图2-29 预览数据

双击“增加校验列”，选择类型“MD5”,输入结果字段“basez\_zj”，选择字段1“clpp1”点击“确定”。如图所示

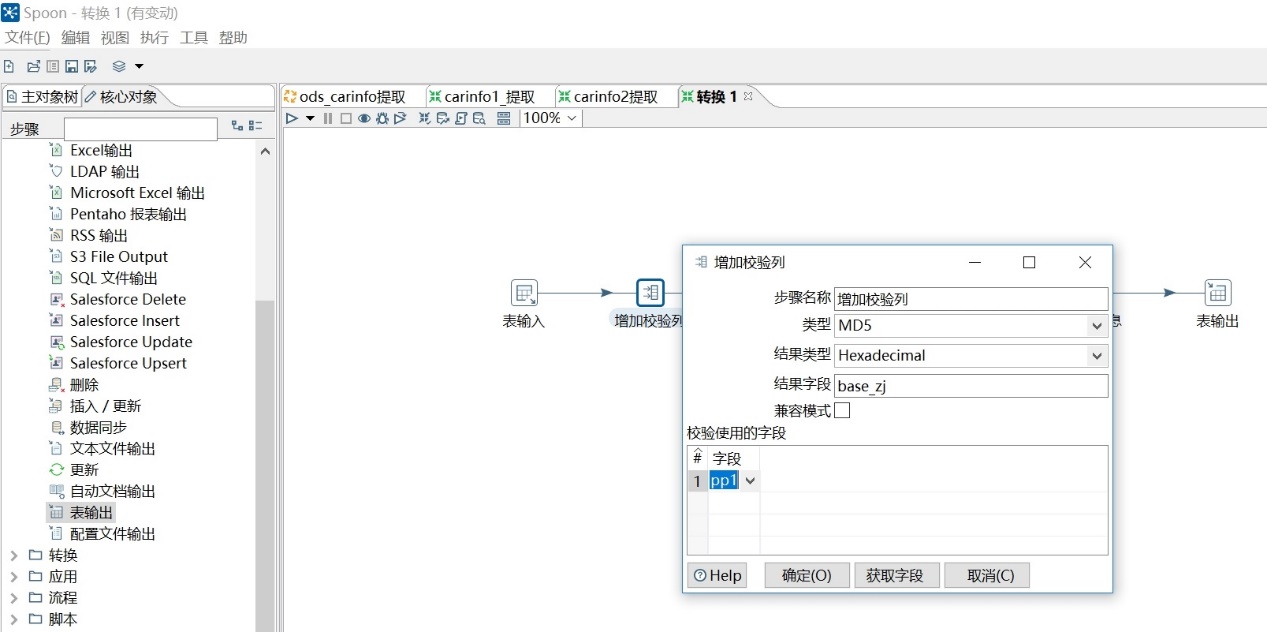


图2-30 定义“增加校验列”

双击“排序记录”，选择字段名称“base\_zj”，选择升序“是”，选择大小写敏感“否”，点击确定。如图所示

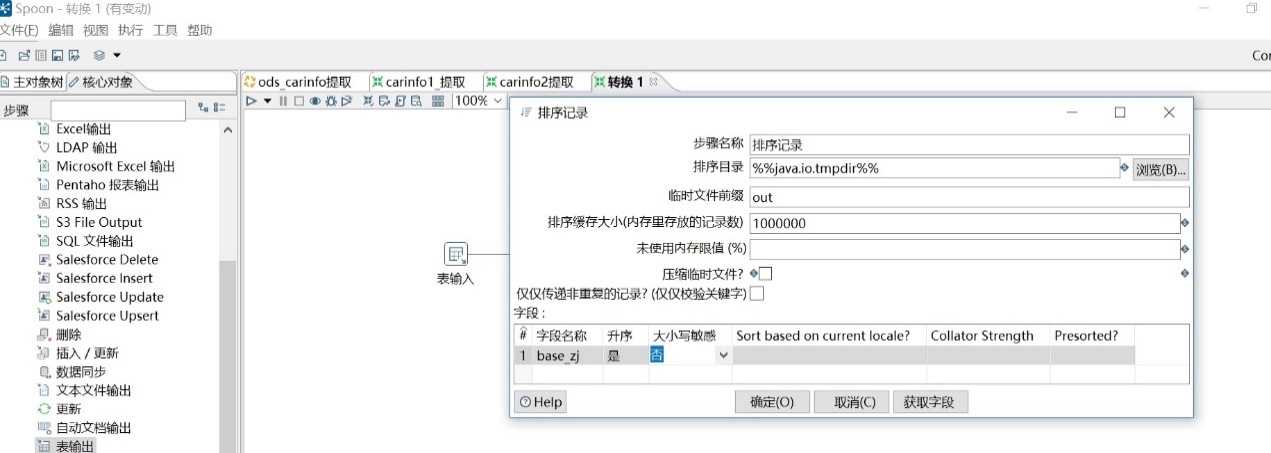


图2-31 定义“排序记录”

双击“去重复记录”，选择字段名称“base\_zj”，选择忽略大小写“N”,点击确定。如图所示



图2-32 定义“去除重复记录”

双击“获取系统信息”，输入名称“base\_rksj”,选择类型“系统日期（可变）”，点击“确定”。如图所示

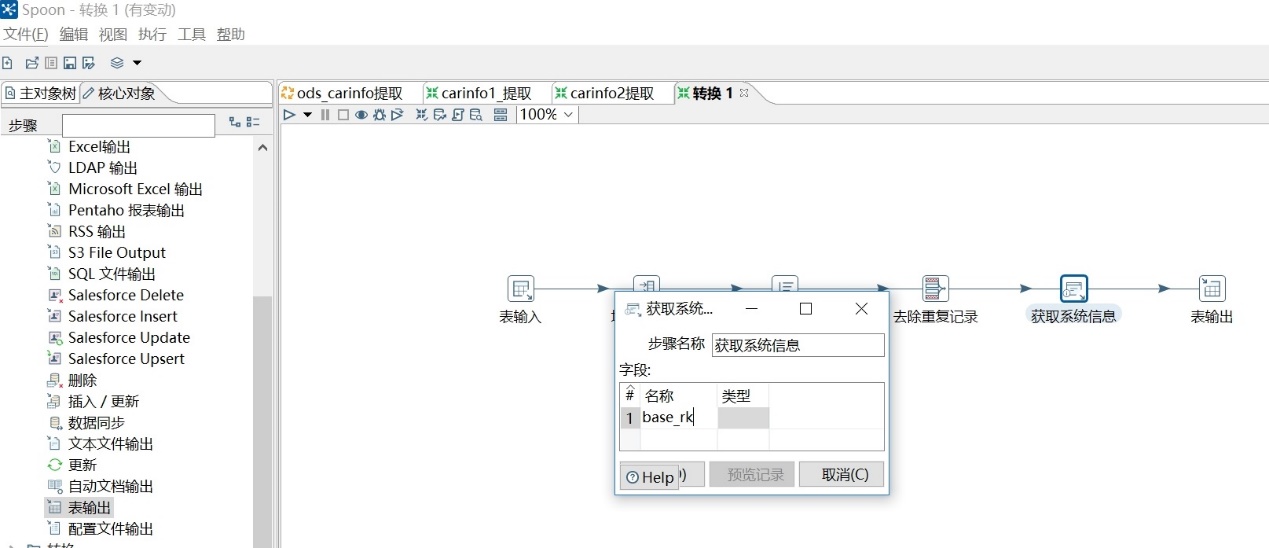


图2-33 定义“获取系统信息”

双击“表输出”，选择目标表“base\_carinfo”，输入目标模式“public”，勾选“指定数据库字段”，点击“获取字段”，点击确实。如图所示

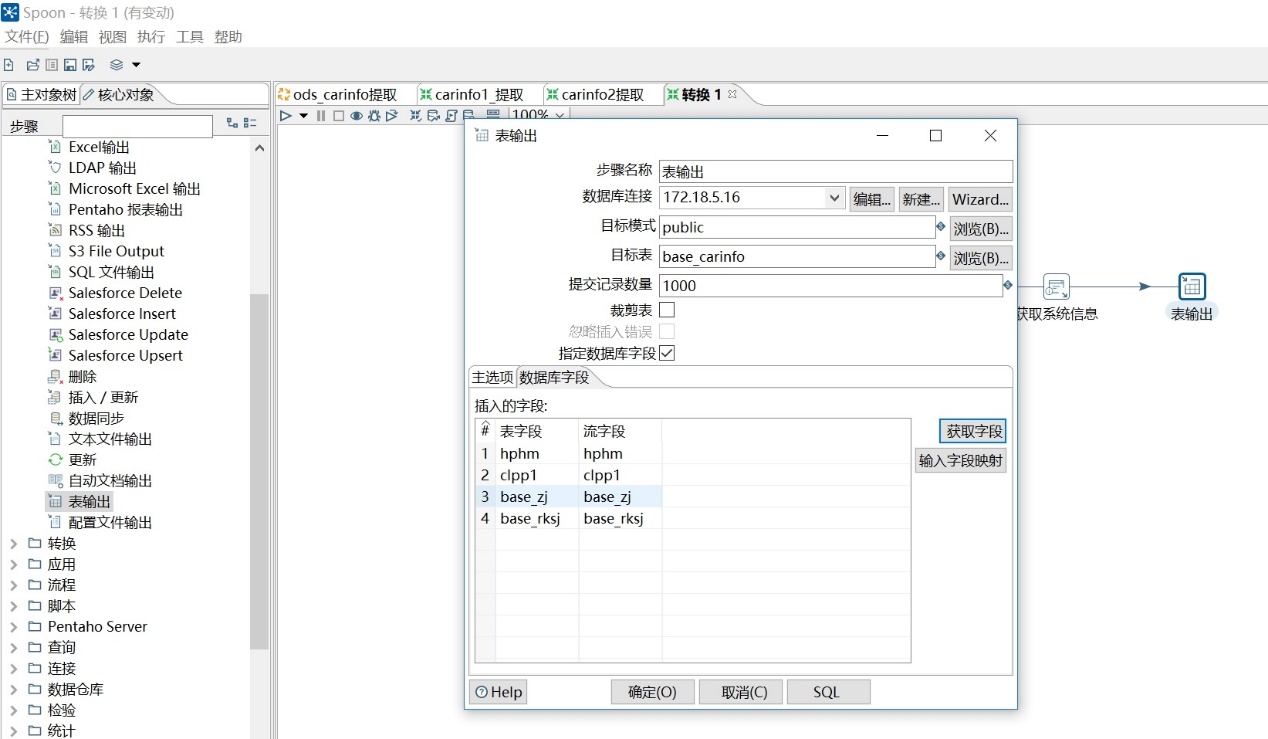


图2-34 定义“表输出”

点击“文件”，选择“保存”，命名为“ods\_carinfo\_qinxi”。点击“运行转换”，点击“Navicat”，查看实验结果。

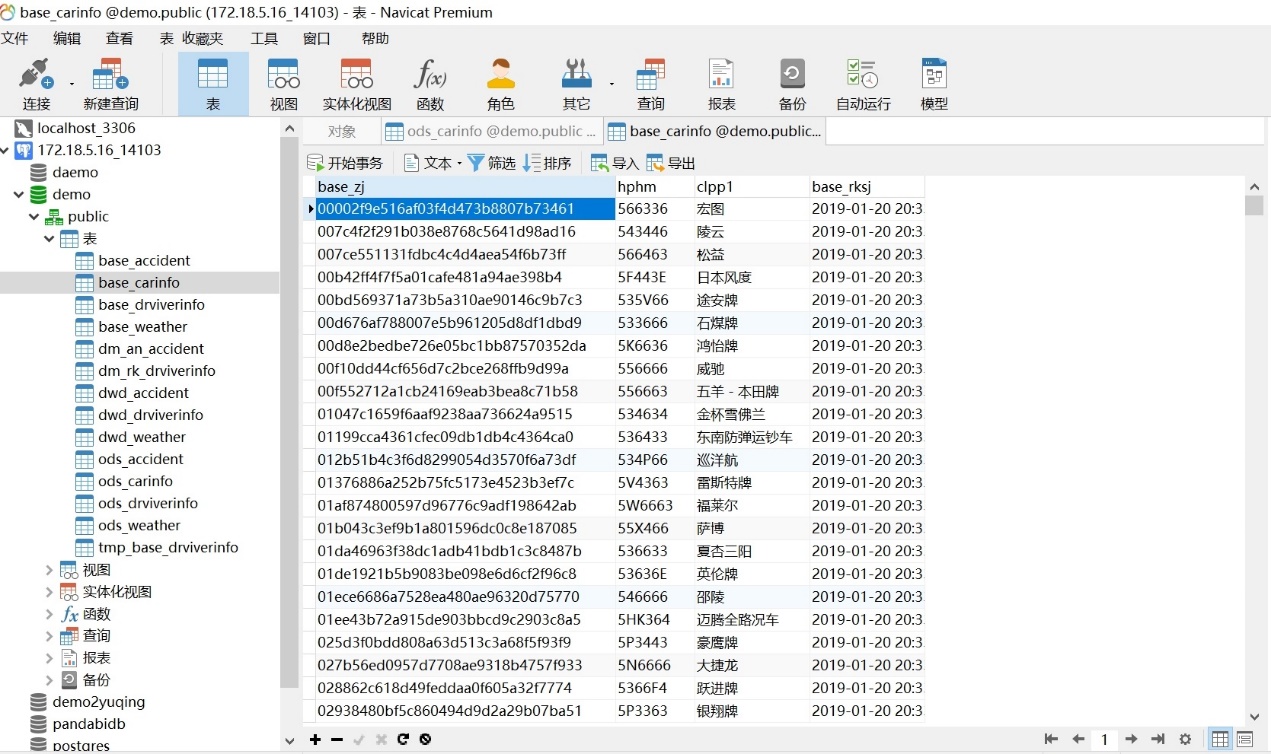


图2-35 Navicat中的实验结果

## **4实验小结**

通过本次实验学习了大数据采集技术Kettle的使用。学习了Kettle的下载及环境的搭建。使用Kettle对数据进行抽取到指定数据库数据表，并使用Kettle对数据进行了简单的清洗工作。