**kettle 4.2.0探索**

# Kettle介绍

## 什么是 kettle

Kettle 也叫 PDI，在2006年 Kettle 加入了开源的 BI 组织 Pentaho, 正式命名为PDI，英文全称为Pentaho Data Integeration。Kettle 是“Kettle E.T.T.L. Envirnonment”只取首字母的缩写，这意味着它被设计用来帮助你实现你的ETTL需要：抽取、转换、装入和加载数据；翻译成中文名称应该叫水壶，名字的起源正如该项目的主程序员 MATT 在一个论坛里说的哪样：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。

Spoon是一个图形用户界面，它允许你运行转换或者任务，其中转换是用Pan工具来运行，任务是用Kitchen来运行。Pan是一个数据转换引擎，它可以执行很多功能，例如：从不同的数据源读取、操作和写入数据。Kitchen是一个可以运行利用XML或数据资源库描述的任务。通常任务是在规定的时间间隔内用批处理的模式自动运行。

## Kettle的安装

要运行kettle工具必须安装Sun公司的JAVA运行环境，kettle 4.2.0需要运行java 1.6或者更高版本，Kettle的下载可以到<http://kettle.pentaho.org/>取得最新版本。kettle不需要安装，安装好java环境后，在操作系统环境变量path中配置jre路径，把kettle工具压缩包解压后可直接使用。

## 运行Spoon

下面是在不同的平台上运行 Spoon所支持的脚本：

Spoon.bat：在windows 平台运行Spoon。

Spoon.sh： 在 Linux、Apple OSX、Solaris 平台运行 Spoon。

## 资源库

资源库是用来保存转换任务的，用户通过图形界面创建的的转换任务可以保存在资源库中。资源库可以使多用户共享转换任务，转换任务在资源库中是以文件夹形式分组管理的，用户可以自定义文件夹名称。资源库有两种形式：

一、Kettle database repository，即保存在各种常见的数据库资源库类型，用户通过用户名/密码来访问资源库中的资源，默认的用户名/密码是admin/admin和guest/guest。

二、Kettle file repository，保存在服务器硬盘文件夹内的资源库类型，此类型的资源库无需用户进行登录，直接进行操作。

当然，资源库并不是必须的，如果没有资源库，用户还可以把转换任务保存在xml文件中。为了方便管理，建议用户建立并使用数据库类型资源库Kettle database repository。

温馨提示：

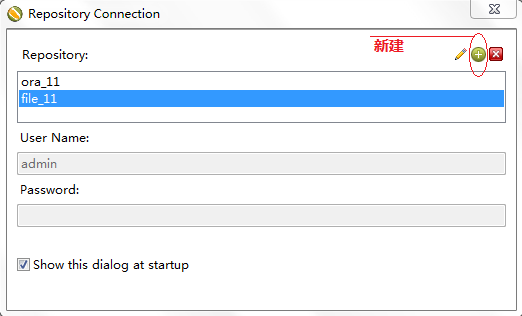
一、在删除资源库中单个内容时，不会提示“是否确定需要删除”，需要特别注意。

二、win7系统下，“Kettle file repository”类型资源库创建文件夹后，无法删除文件夹，不确定是Kettle工具本身的bug或者是在win7操作系统下的bug。

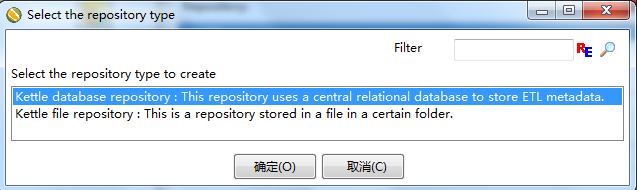
三、“Kettle database repository”类型资源库虽然有需要用户和密码进行登录，但目前还没有相应的权限控制。

## 资源库的创建

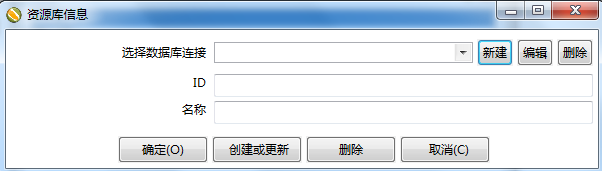
点击新建



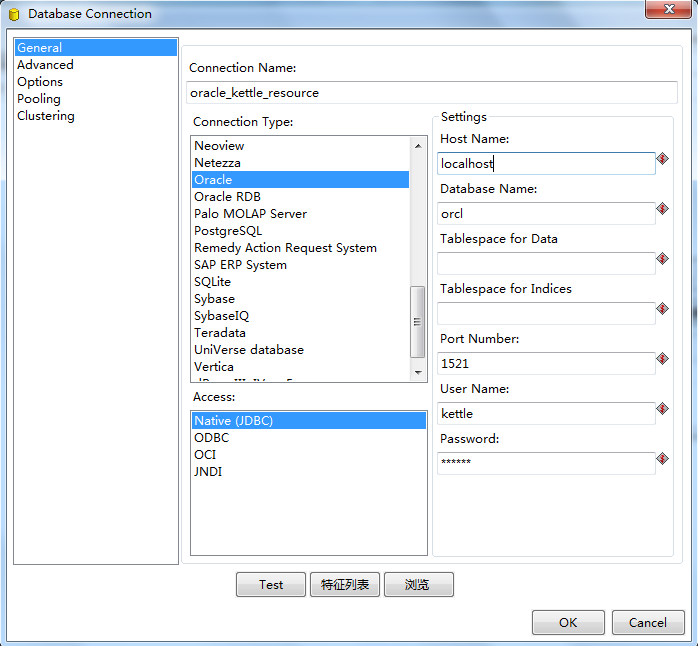
选择数据库类型资源库Kettle database repository，点击“确定”



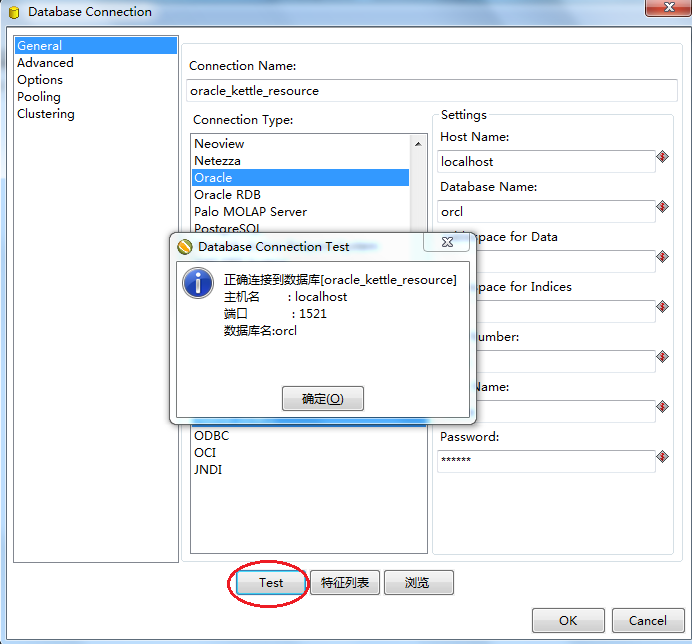
在“资源信息库”窗口中，点击“新建”按钮，新建一个数据库连接



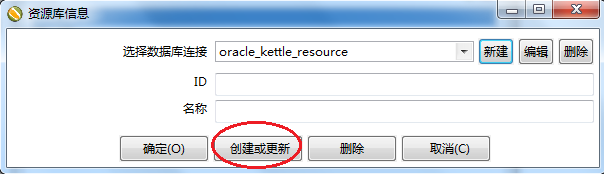
在“Database Connection”窗口，选择数据库类型，并填写数据库连接的相关信息，本实例选用oracle数据库。

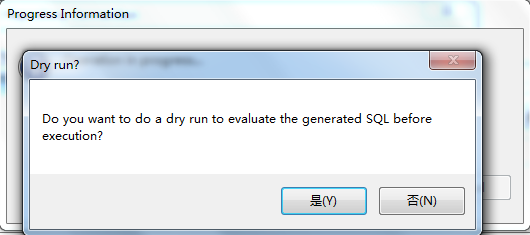


点击test按钮进行数据库连接测试

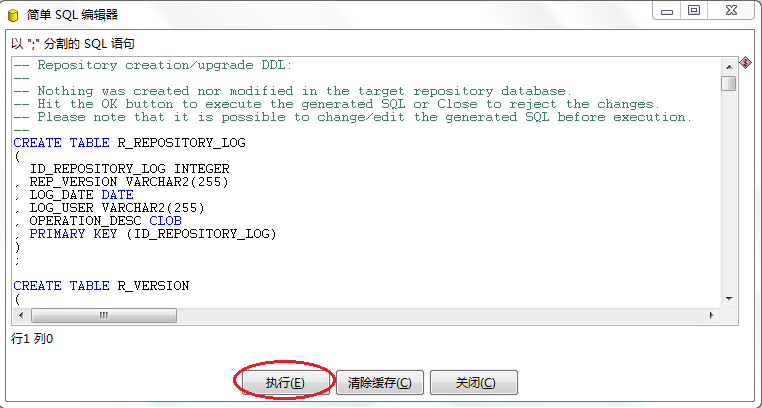


数据库连接正确后，点击“确定”，再点击“OK”，回到“资源库信息窗口”，点击“创建或更新”按钮

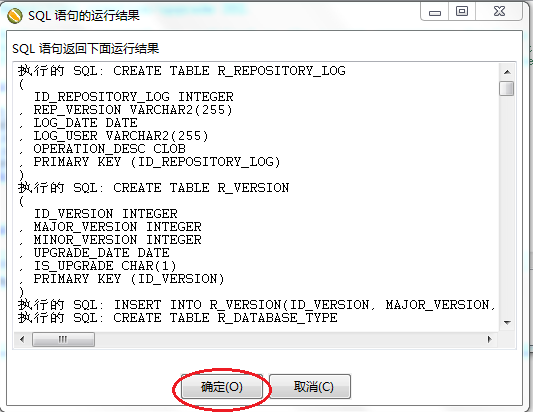




在弹出的“简单SQL编辑器”窗口中，点击“执行”按钮，执行相应的资源库创建脚本。（用户可以查看这些脚本具体是进行了什么操作：创建了42张表，并插入了一些相关信息数据，如R\_USER表里插入了admin和guest两个用户的信息，这两个用户也是该资源库的登录用户，想深入研究资源库的用户可以并一步去了解相关内容）



在弹出“SQL语句运行结果”窗口，点击“确定”按钮



再点击“关闭”按钮，关闭“简单SQL编辑器”窗口

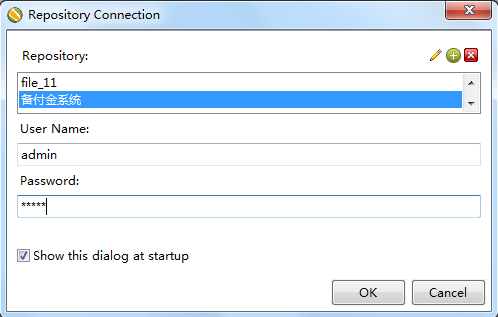


在“资源库信息”窗口，填写资源库的ID和名称，然后点击“确定”按钮。

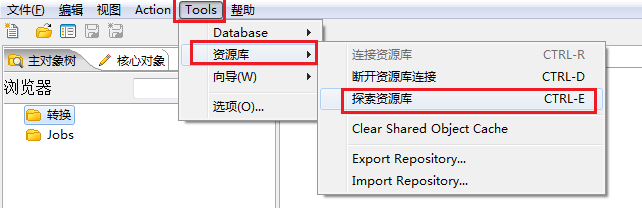
温馨提示：建议使用英名，因为使用中文容易出现乱码，而使得退出后，下次登录资源库时无法找到相应的配置，遇到此类情况，可根据下一章节“资源库名称乱码问题解决方法”进行相关修改设置。



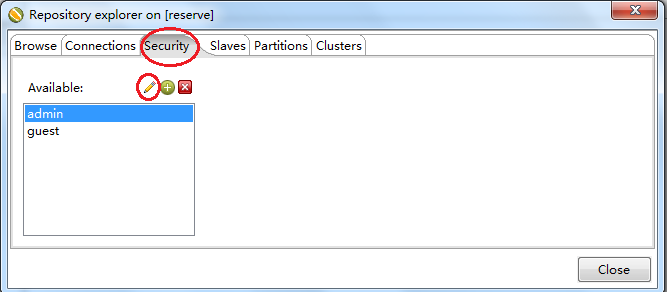
在“资源库连接Repository Connection”窗口，选择刚才创建的资源库“备付金系统”，填写用户名和密码，新建的资源库的默认用户和密码有两个，admin/admin和guest/guest，然后点击“确定OK”按钮，即可登录资源库。



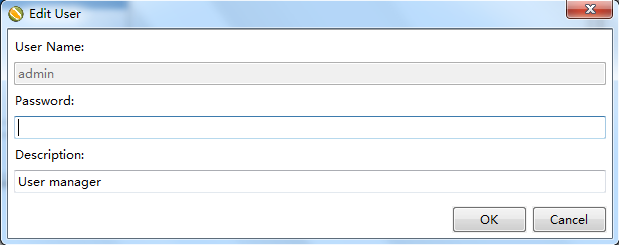
为了资源库的安全，我们需要修改资源库用户的登录密码，“Tools” 🡪“资源库”🡪“探索资源库”，可以使用快捷键CTRL + E。



在资源库探索“Repository explorer”窗口，选择“Security”选项卡，然后选择需要修改的用户再点击编辑按钮



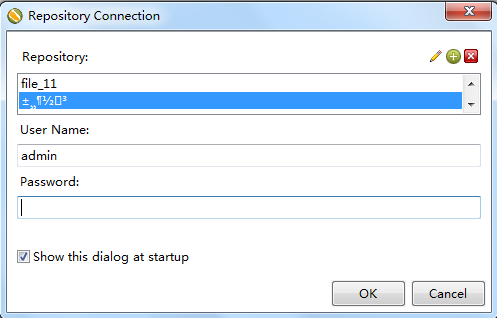
填写新密码，然后点击“OK”按钮即可完成用户密码的修改



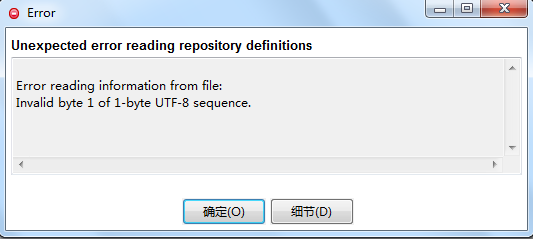
## 资源库名称乱码问题解决方法

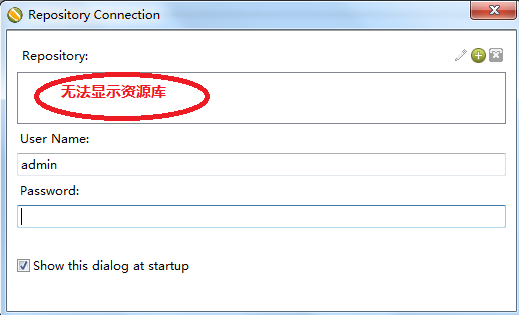
如果在创建资源库时，名称使用的是中文，容易出现乱码或者出现异常报错等情况

情况一、乱码

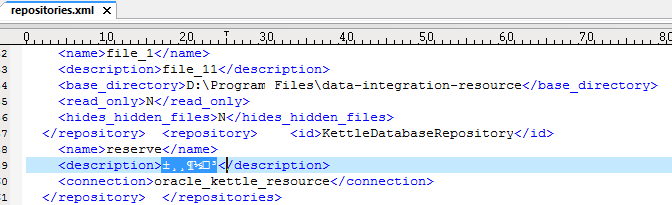


情况二、异常，并且无法显示资源库





解决方法：根据以下路径“c:\Documents and Settings\<username>\.kettle\repositories.xml”文件，打开后找到乱码内容，如下图所示，把乱码删除后修改成中文即可。



## 运行日志

转换运行日志有4张表，JOB日志有3张表，可根据需要设置自己所需日志表和表字段，这些表的表名称和表字段可以根据需要进行修改，为了统一管理，建议把表名称和表字段固定。

|  |  |  |
| --- | --- | --- |
| **日志类型** | **日志表名称** | **日志内容** |
| 转换 | LOG\_TRANS\_STEP | 转换的各个步骤执行结果 |
| 转换 | LOG\_TRANS\_PERFORMANCE | 转换的性能日志 |
| 转换 | LOG\_TRANS\_CHANNEL | 转换的各个渠道执行结果 |
| 转换 | LOG\_TRANS | 转换的执行结果 |
| JOB | LOG\_JOB\_ENTRY | JOB的接口执行结果 |
| JOB | LOG\_JOB\_CHANNEL | JOB的各个渠道执行结果 |
| JOB | LOG\_JOB | JOB的执行结果 |

### 转换运行日志

打开转换，设置转换运行日志，方法有三种

一、编辑 🡪 设置 🡪 日志

二、右键 🡪 转换设置 🡪日志

三、快捷键CTRL+T 🡪 日志

### JOB运行日志

打开JOB，设置JOB运行日志，方法有三种

一、编辑 🡪 设置 🡪 日志

二、右键 🡪 作业设置 🡪日志

三、快捷键CTRL+J 🡪 日志