### spoon 用户指南

#### 1.1 spoon简介

Kettle是“Kettle ETTL Environment”的首字母缩写。Kettle旨在帮助您满足您的ETTL需求，包括数据的提取，转换，运输和加载。

Spoon是一个图形用户界面，允许您设计可以使用Kettle工具--Pan和Kitchen运行的转换和作业。Pan是一种数据转换引擎，可执行多种功能，例如从各种数据源读取，操作和写入数据。Kitchen是一个程序，用于执行由Spoon在XML或数据库存储库中设计的作业。作业通常以批处理模式安排，以定期自动运行。

转换和作业可以使用XML文件描述自己，也可以放在Kettle数据库存储库中。然后，Pan或Kitchen可以读取数据以执行转换中描述的步骤或运行作业。

##### 存储库

Spoon允许您将转换和作业文件存储到本地文件系统或Kettle存储库中。Kettle存储库可以存放在任何常见的关系数据库中。要从数据库存储库加载转换，您必须连接到此存储库。启动Spoon时定义与存储库的数据库连接。

与存储库关联的信息存储在“repositories.xml”中。此文件位于默认主目录中的隐藏目录“.kettle”中。

##### Transformation、Job定义

值：行的一部分

行：一行由0个或多个值组成，值作为单独条目进行处理

输入流：进入步骤的一堆行

跳(hops):两个步骤或作业的一个或多个数据流的表示，一跳始终表示一步的输出流和另一部的输入流。

注释：对转换或作业的描述

##### kettle选项

这里只说明 “使用数据库缓存？”这个选项。

Spoon缓存存储在源数据库和目标数据库中的信息。在某些情况下，当您进行数据库更改时，缓存会导致错误的结果。为防止出错，您可以完全禁用缓存，而不是每次都清除缓存。

注意：当您向数据库连接启动DDL（数据定义语言）语句时，Spoon会自动清除数据库缓存; 但是，使用第三方工具时，可能需要手动清除数据库缓存。

##### 搜索元数据

此选项将搜索所有已加载作业的任何可用字段，连接器或注释以及“过滤器”字段中指定的字符串的转换。元数据搜索返回详细的结果集，显示任何搜索命中的位置。通过从菜单栏中选择“编辑”|“搜索元数据”可访问此功能。

##### 设置环境变量

“设置环境变量”功能允许您显式创建和设置当前用户会话的环境变量。在设计用于测试通常由另一个作业或转换动态设置的变量替换的转换时，这是一个有用的功能。

通过从菜单栏中选择“编辑”|“设置环境变量”，可以访问此功能。

### 1.2 数据库连接

数据库连接描述了Kettle连接到数据库的方法。您可以创建特定于作业或转换的连接，也可以将它们存储在Kettle存储库中，以便在多个转换或作业中重复使用。

需要注意的：

连接池：一些数据库的 配置选项。

集群：此选项卡允许您为数据库连接启用群集并创建与数据分区的连接。要为连接启用群集，请启用“使用群集？” 选项。

要创建新数据分区，请输入分区ID以及用于连接到分区的主机名，端口，数据库，用户名和密码。

高级 | 连接成功需要执行的SQL：此选项卡允许您在连接到数据库后立即输入许多SQL命令。出于许可，配置，日志记录，跟踪等原因，有时需要这样做。

清除数据库缓存：为了加速连接，Spoon使用数据库缓存。当缓存中的信息不再代表数据库的布局时，右键单击主树中的连接并选择“清除数据库缓存...”选项。更改，创建或删除数据库表时，通常使用此命令。

共享：使连接可以在其他转换或作业中使用。

SQL编辑器：见下一小节

### 1.3 SQL编辑器

当您必须为诸如创建表，删除索引和修改字段等任务执行标准SQL命令时，Simple SQL Editor是一个很好的工具。SQL编辑器用于预览和执行Spoon生成的DDL（数据定义语言），如“创建/更改表”，创建索引，“和”创建序列“SQL命令。例如，如果您将表输出步骤添加到转换并单击Table Input对话框底部的SQL按钮，Spoon自动为输出步骤生成必要的DDL以使其正常运行，并通过SQL编辑器将其呈现给最终用户。

注意：

* 多个SQL语句必须用分号分隔。
* 在将这些SQL语句发送到要执行的数据库之前，Spoon会删除返回，换行和分隔分号。
* Kettle清除启动DDL语句的数据库连接的数据库缓存。

限制：

Simple SQL Editor无法识别所有受支持数据库的方言。这意味着创建存储过程，触发器和其他特定于数据库的对象可能会带来问题。考虑在这些实例中使用数据库附带的工具。

### 1.4 hops(跳)

转换：注意方向(代表数据流向)和是否生效(灰色还是蓝色)以及错误处理(红色)

作业：

有🔒的：“无条件”指定无论原始作业条目的结果如何，都将执行下一个作业条目。

有√的：“当结果为真时跟随”指定只有当原始作业条目的结果为真时才执行下一个作业条目，这意味着成功执行，找到文件，找到表格，没有错误，评估结果为假，...

有×的：“当结果为假时跟随”指定只有当原始作业条目的结果为假时才执行下一个作业条目，这意味着执行失败，找不到文件，找不到表，发生错误，评估为假， ...

### 1.5 变量和参数

变量引用：${var}或%%var%%,也可以递归使用变量，例：${%%inner\_var%%}

变量(variables):

用法；${variables}或%%variables%%

全局环境变量：即可作用于全局的变量，通过-D选项传给JVM虚拟机的

kettle局部变量：作用于作业或转换中的动态输入，限定变量作用范围，通过设置变量步骤或修改 kettle.properties设置

内部变量：类似 Internal.Transformation.Filename.Directory 的变量

命名参数(parameters):

参数就像局部变量; 它们是可重用的输入，仅适用于它们所定义的特定转换。定义参数时，可以为其指定一个默认值，

以便在未为其提取一个参数时使用。此功能使其在PDI中的动态输入类型中独一无二。

注意：如果参数和变量之间存在名称冲突，则该参数优先

定义：在转换中右键，选择转换设置，单击命名参数选项卡，或在作业中右击转换、作业组件，单击命名参数选项

位置参数(arguments):是作为命令行参数给出的，每个参数都声明为Pan或Kitchen行的其余部分后给出的以空格分隔的值(即下面的argOne,argTwo,argThree)

sh pan.sh -file：/example\_transformations/example.ktr argOne argTwo argThree

通常在作业中的作业和转换组件对话框中的arguments中定义，以便用来测试。

### 1.6 转换设置和步骤

设置：略

步骤：

并行，即“改变开始复制的数量“

分发还是复制：分发，轮流处理数据，复制，处理同样的数据

错误处理：发生错误时的下一步动作

可以支持VFS:略

步骤详解见第2章。

### 1.7 作业设置和条目

设置：略

条目见第3章。

### Pentaho数据集成步骤

### 第3章 Pentaho数据集成条目

### 第4章 Pan、Kitchen、Carte用户文档