

---

# Kettle初探

Start : 2011.03.01

Updated : 2011.03.01

王凡(wangfan)

[wf141732@sohu.com](mailto:wf141732@sohu.com)

[woshiwangfan@gmail.com](mailto:woshiwangfan@gmail.com)

<http://t.qq.com/lingmengfei>

版本	修改日期	内容
1.0	2011.03.01	创建

---

## 1. 简介

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，数据抽取高效稳定。

在这里主要对比Oracle Data Integrater作对比

### 1.1. 环境信息

---

Kettle:4.10ce

OS: Red Hat Enterprise Linux 5 64-bit

### 1.2. 相关文档

---

---

## 2. 软件准备

## 2.1. 软件下载

---

官网需要验证信息，到如下地址可以下载 Kettle：

<http://nchc.dl.sourceforge.net/project/pentaho/Data%20Integration/4.1.0-stable/pdi-ce-4.1.0-stable.tar.gz>

JDK: [http://cds.sun.com/is-bin/INTERSHOP.enfinity/WFS/CDS-CDS\\_Developer-Site/en\\_US/-/USD/VerifyItem-Start/jdk-6u24-linux-x64.bin?BundledLineItemUUID=mCiJ\\_hCwiC4AAAEuob4AGW8J&OrderID=uPiJ\\_hCwgMYAAAEuiL4AGW8J&ProductID=oSKJ\\_hCwOIYAAAEtBcoADqmS&FileName=/jdk-6u24-linux-x64.bin](http://cds.sun.com/is-bin/INTERSHOP.enfinity/WFS/CDS-CDS_Developer-Site/en_US/-/USD/VerifyItem-Start/jdk-6u24-linux-x64.bin?BundledLineItemUUID=mCiJ_hCwiC4AAAEuob4AGW8J&OrderID=uPiJ_hCwgMYAAAEuiL4AGW8J&ProductID=oSKJ_hCwOIYAAAEtBcoADqmS&FileName=/jdk-6u24-linux-x64.bin)

安装JRE也可

## 2.2. 软件安装

---

无需安装，解压出来

需要配置PENTAHO\_JAVA\_HOME

```
export PENTAHO_JAVA_HOME=/usr/java/jdk1.6.0_23
```

或者修改set-pentaho-env.sh

```
if [ -n "$PENTAHO_JAVA_HOME" ]; then
    # echo "DEBUG: Using PENTAHO_JAVA_HOME"
    _PENTAHO_JAVA_HOME="$PENTAHO_JAVA_HOME"
    _PENTAHO_JAVA="$_PENTAHO_JAVA_HOME"/bin/$_LAUNCHER
```

将PENTAHO\_JAVA\_HOME修改为JAVA\_HOME，前提是JAVA\_HOME已经配置过

## 3. 一个简单的传输

### 3.1. 体积

---

Kettle:118M

ODI:432M(仅包括oracledi)

### 3.2. 建立数据库连接用户

---

在数据库中建立kettle的用户

```
-- Create the user
create user kettle
  identified by wangfan
  default tablespace USERS
  temporary tablespace TEMP;
-- Grant/Revoke role privileges
grant connect to kettle;
grant resource to kettle;
```

### 3.3. 建立资料库

---

解压后到data-integration目录运行./spoon.sh



**Repository Connection**

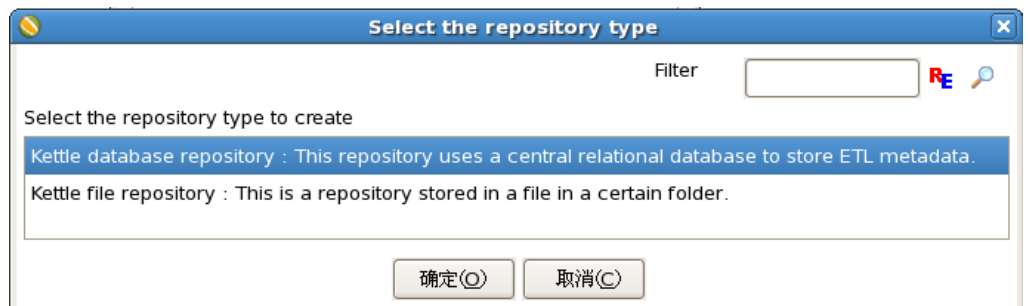
Repository: ✎ + ✕

User Name:

Password:

☒ Show this dialog at startup

点击+，新建资料库，kettle提供两种形式的资料库存储，数据库存储和文件存储

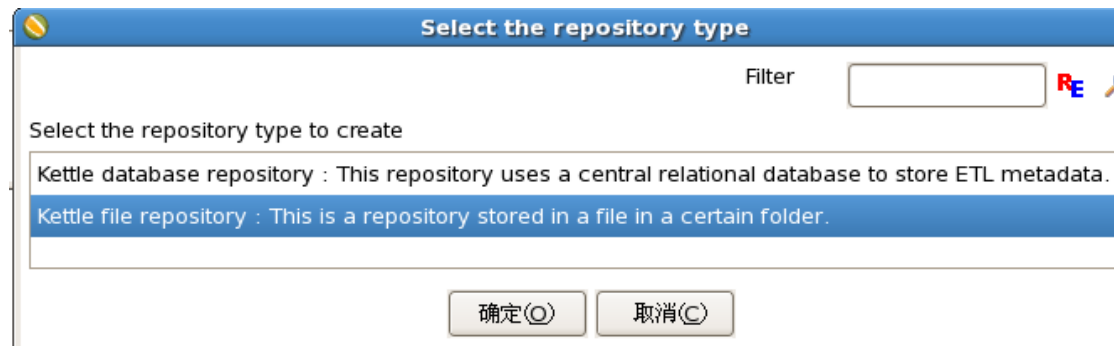


**Select the repository type**

Filter

Select the repository type to create

- Kettle database repository : This repository uses a central relational database to store ETL metadata.
- Kettle file repository : This is a repository stored in a file in a certain folder.



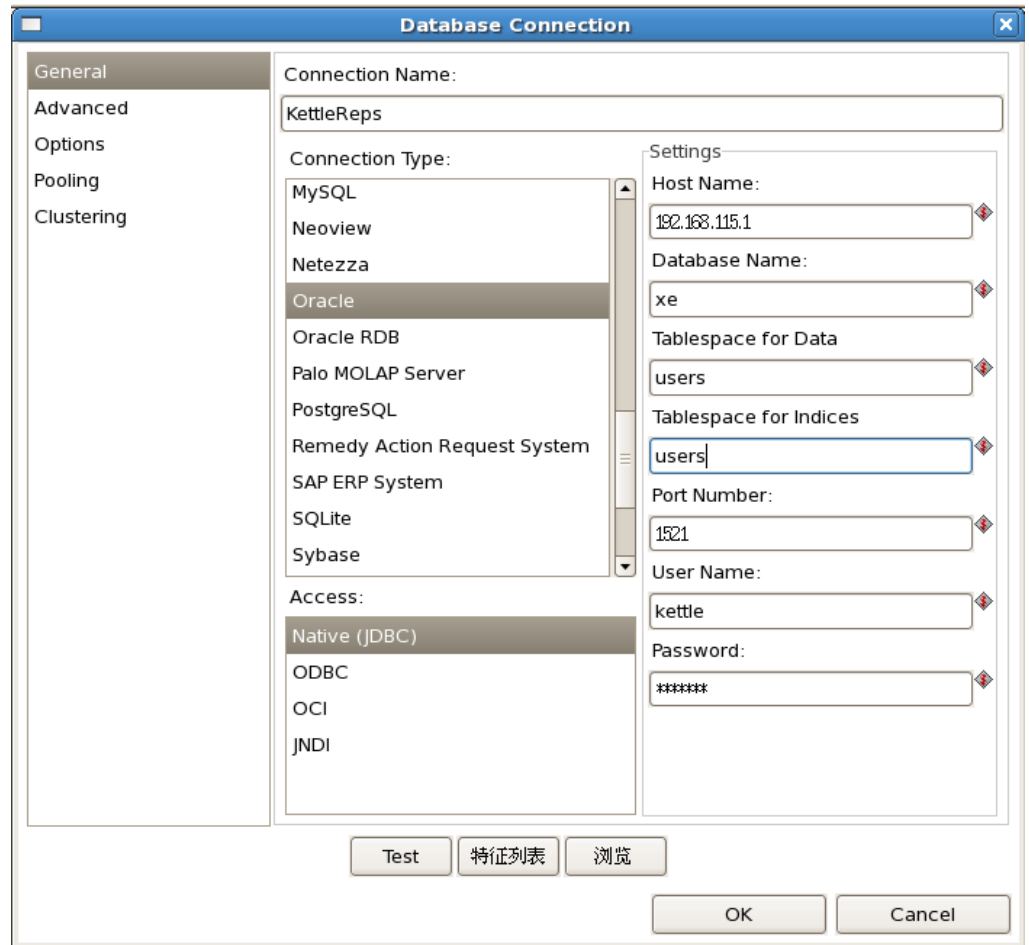
**Select the repository type**

Filter

Select the repository type to create

- Kettle database repository : This repository uses a central relational database to store ETL metadata.
- Kettle file repository : This is a repository stored in a file in a certain folder.

建立名为KettleReps的资料库



The 'Database Connection' dialog box is shown with the 'General' tab selected. The 'Connection Name' is 'KettleReps'. The 'Connection Type' is 'Oracle'. The 'Access' is 'Native (JDBC)'. The 'Settings' section includes: 'Host Name' (192.168.115.1), 'Database Name' (xe), 'Tablespace for Data' (users), 'Tablespace for Indices' (users), 'Port Number' (1521), 'User Name' (kettle), and 'Password' (\*\*\*\*\*). Buttons at the bottom include 'Test', '特征列表', '浏览', 'OK', and 'Cancel'.

Database Connection

General

Advanced

Options

Pooling

Clustering

Connection Name: KettleReps

Connection Type: MySQL, Neoview, Netezza, Oracle, Oracle RDB, Palo MOLAP Server, PostgreSQL, Remedy Action Request System, SAP ERP System, SQLite, Sybase

Access: Native (JDBC), ODBC, OCI, JNDI

Settings

Host Name: 192.168.115.1

Database Name: xe

Tablespace for Data: users

Tablespace for Indices: users

Port Number: 1521

User Name: kettle

Password: \*\*\*\*\*

Test 特征列表 浏览 OK Cancel



The 'File repository settings' dialog box is shown. The 'Base directory' is '/mnt/hgfs/development-soft/kettle'. The 'Read-only repository?' checkbox is unchecked. The 'ID' is 'kettle1'. The 'Name' is 'KettleResp'. Buttons at the bottom include '确定(O)' and '取消(C)'.

File repository settings

Base directory: /mnt/hgfs/development-soft/kettle 浏览(B)...

Read-only repository? ☐

ID: kettle1

Name: KettleResp

确定(O) 取消(C)

在资源库这里点击创建或更新开始创建

**资源库信息**

选择数据库连接: KettleReps [新建] [编辑] [删除]

ID: [ ]

名称: [ ]

[确定(O)] [创建或更新] [删除] [取消(C)]

点击确定

**确定**

你确信要 创建 这个资源库在这个指定数据库连接?

[是(Y)] [否(N)]

执行完毕后会显示执行SQL的详细信息

SQL 语句的运行结果

```

SQL 语句返回下面运行结果
执行的 SQL: CREATE TABLE R_TRANSFORM_LOCK
(
  ID_TRANSFORM_LOCK INT(10)
, ID_TRANSFORM_LOCK INT(10)
, ID_USER INT(10)
, LOCK_MESSAGE CLOB
, LOCK_DATE DATE
, PRIMARY KEY (ID_TRANSFORM_LOCK)
)
TABLESPACE users
执行的 SQL: CREATE TABLE R_JOB_LOCK
(
  ID_JOB_LOCK INT(10)
, ID_JOB INT(10)
, ID_USER INT(10)
, LOCK_MESSAGE CLOB
, LOCK_DATE DATE
, PRIMARY KEY (ID_JOB_LOCK)
)
TABLESPACE users
执行的 SQL: CREATE TABLE R_USER
(
  ID_USER INT(10)
, LOGIN VARCHAR2(255)
, PASSWORD VARCHAR2(255)
, NAME VARCHAR2(255)
, DESCRIPTION VARCHAR2(255)
, ENABLED CHAR(1)
, PRIMARY KEY (ID_USER)
)
TABLESPACE users
执行的 SQL: INSERT INTO R_USER(ID_USER, LOGIN, PASSWORD, NAME, DESCRIPTION, ENABLED) VALUES (1,'admin','2be98afc86aa7f2e4cb79ce71da9fa6d4','Administrator','User manager','Y')
执行的 SQL: INSERT INTO R_USER(ID_USER, LOGIN, PASSWORD, NAME, DESCRIPTION, ENABLED) VALUES (2,'guest','2be98afc86aa7f2e4cb79ce71da9fa6d4','Guest account','Read-only guest account','Y')
执行了 343 个 SQL 语句
  
```

[确定(O)] [取消(C)]

然后给资源库一个唯一ID

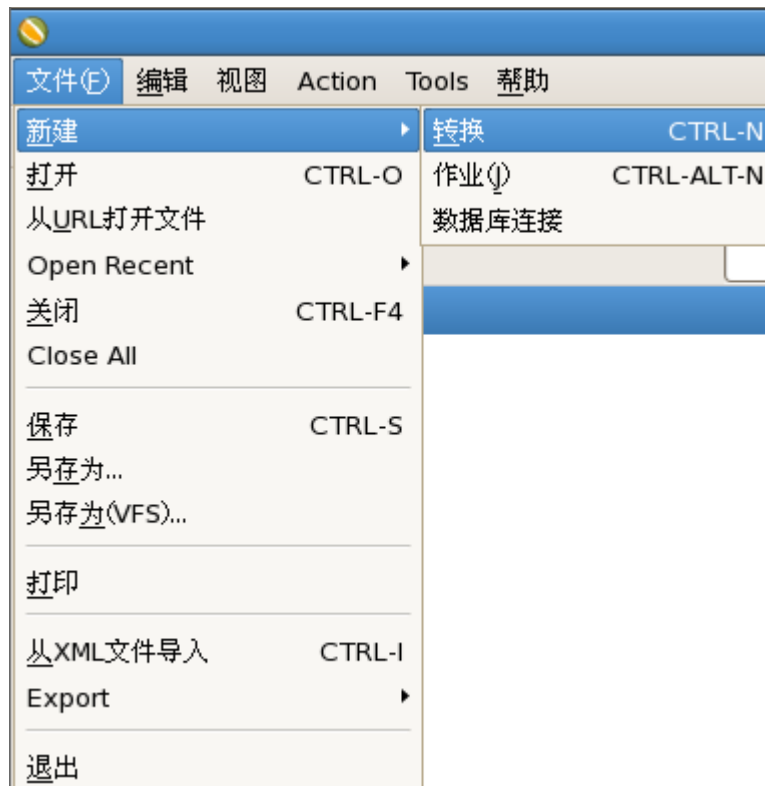


用户名密码admin/admin,进入Kettle

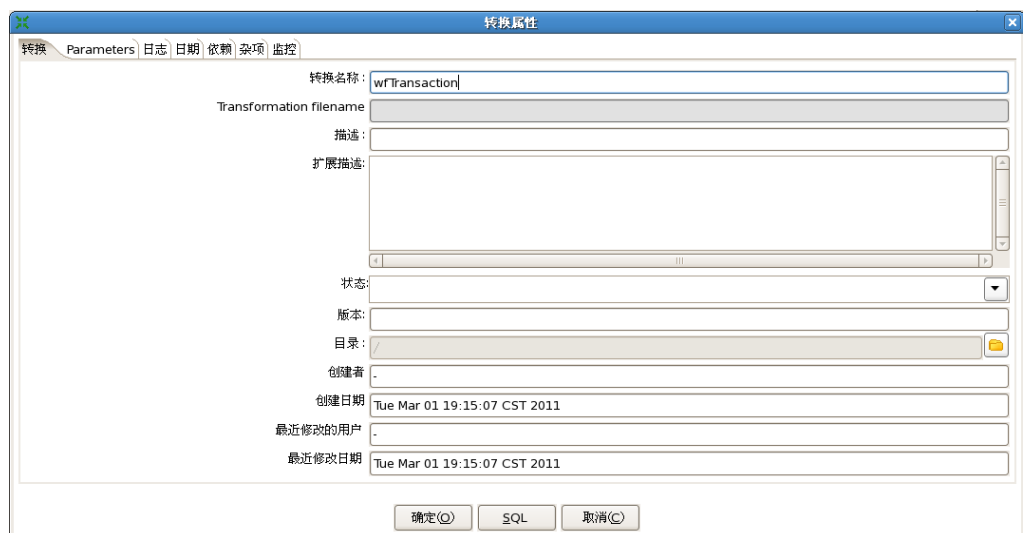


### 3.4. 建立数据库连接

点击文件->新建->转换，建议一个转换

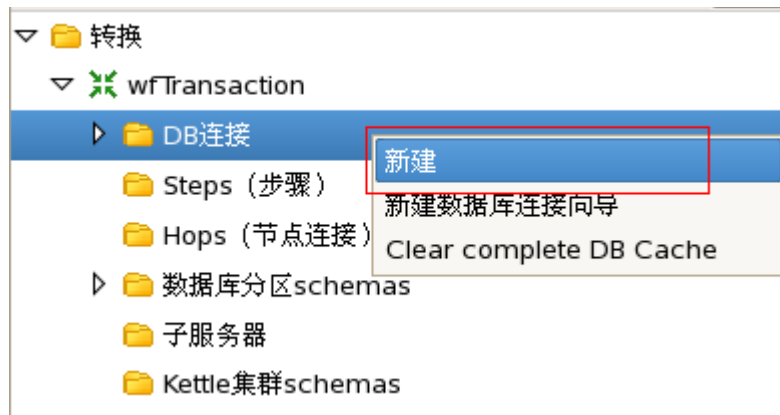


在新建的转换上右键->编辑，可以修改转换属性

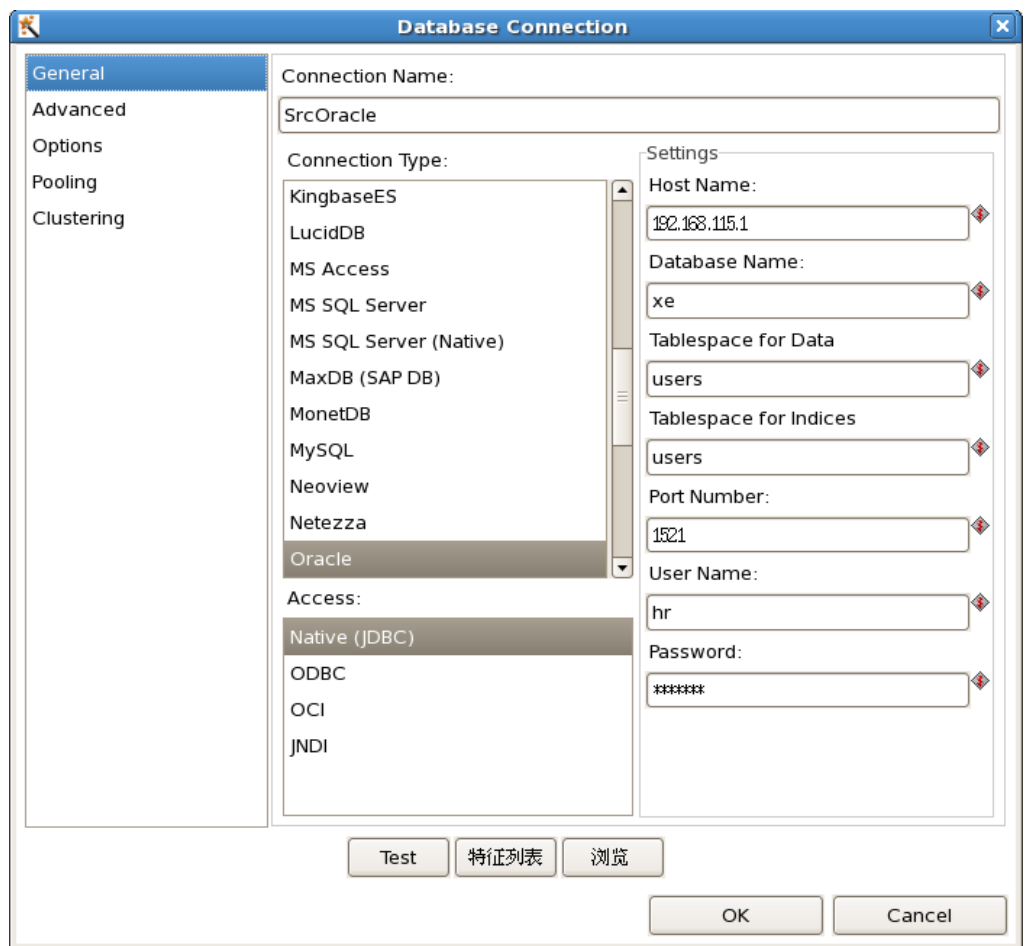


在下面转换下面DB连接右键->新建



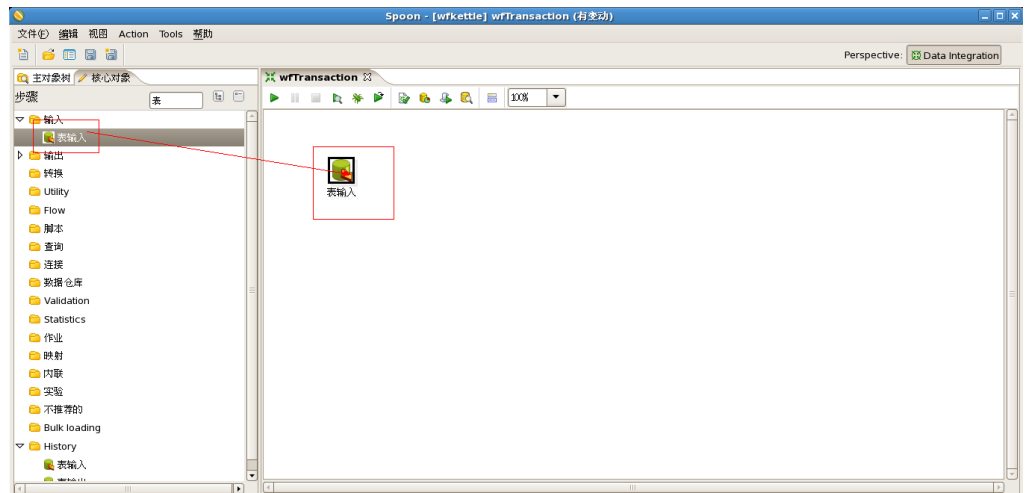


输入连接信息

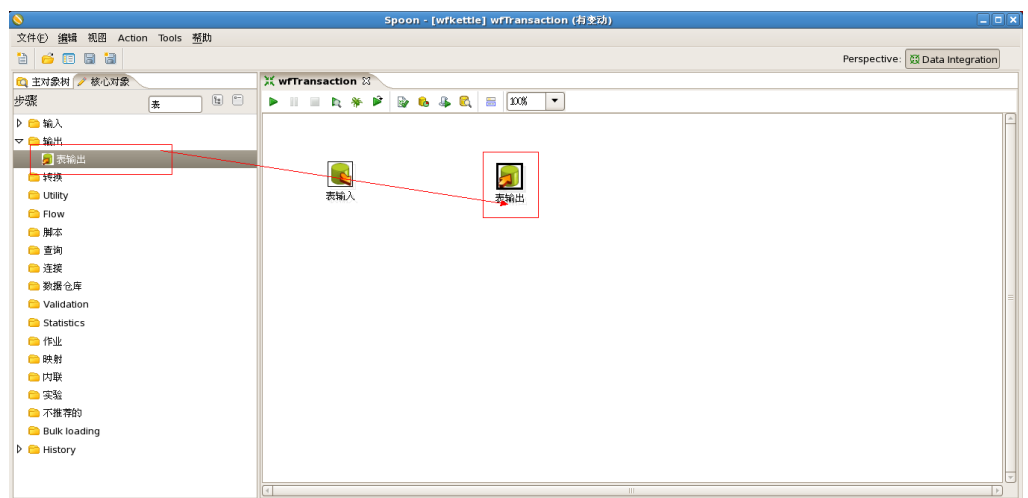


### 3.5. 数据传输

在核心对象里面拖动表输入到wfTransaction



在拖动一个表输入到wfTransaction



双击表输出，编辑表输出属性

表输出

步骤名称: 表输出

数据库连接: SrcOracle [编辑...] [新建...]

目标模式: HR [浏览(B)...]

目标表: HTG\_JOBS [浏览(B)...]

提交记录数量: 1000

裁剪表: ☐

忽略插入错误: ☐

Specify database fields: ☐ [图标]

Main options Database fields

表分区数据: ☐

分区字段: [选择框]

每个月分区数据: ☒

每天分区数据: ☐

使用批量插入: ☒

表名定义在一个字段里?: ☐

包含表名的字段: [选择框]

存储表名字段: ☒

返回一个自动产生的关键字: ☐

自动产生的关键字的字段名称: [选择框]

[确定(O)] [取消(C)] [SQL]

双击表输入，进入表输入属性对话框，选择“获取SQL查询语句”

**表输入**

步骤名称: 表输入

数据库连接: SrcOracle 编辑... 新建...

SQL 获取SQL查询语句...

```
SELECT
  JOB_ID
, JOB_TITLE
, MIN_SALARY
, MAX_SALARY
FROM JOBS
```

行1 列0

允许延迟转换 ☐

替换 SQL 语句里的变量 ☐

从步骤插入数据

执行每一行? ☐

记录数量限制: 0

确定(O) 预览(P) 取消(C)

在Hops右键新建

**Hop: From --> To**

起始步骤: 表输入

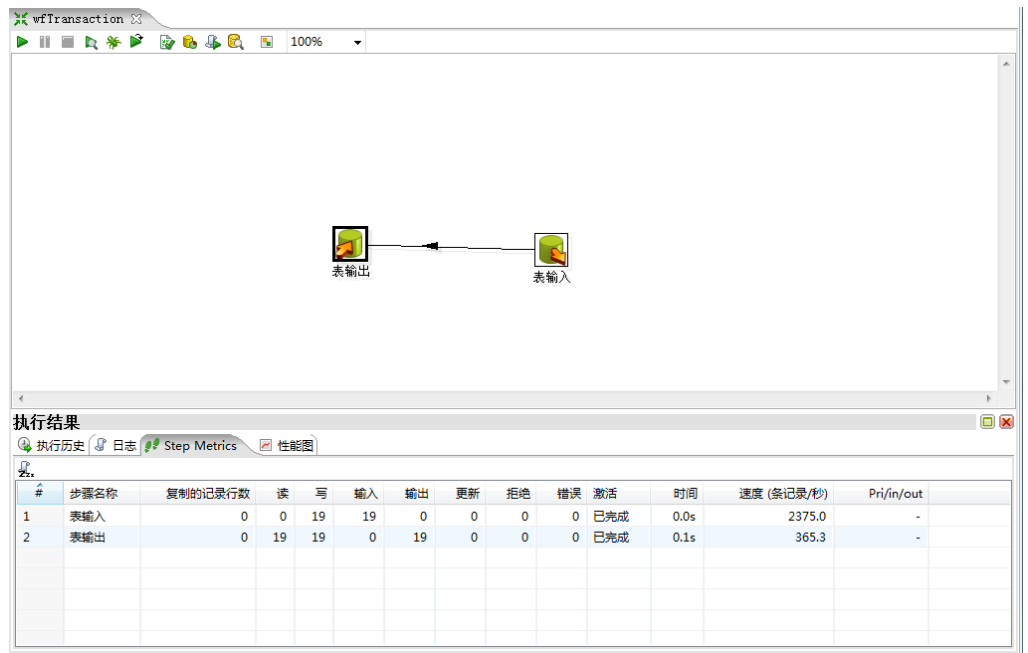
目标步骤: 表输出

使连接生效? ☒

From <-> To

确定(O) 取消(C)

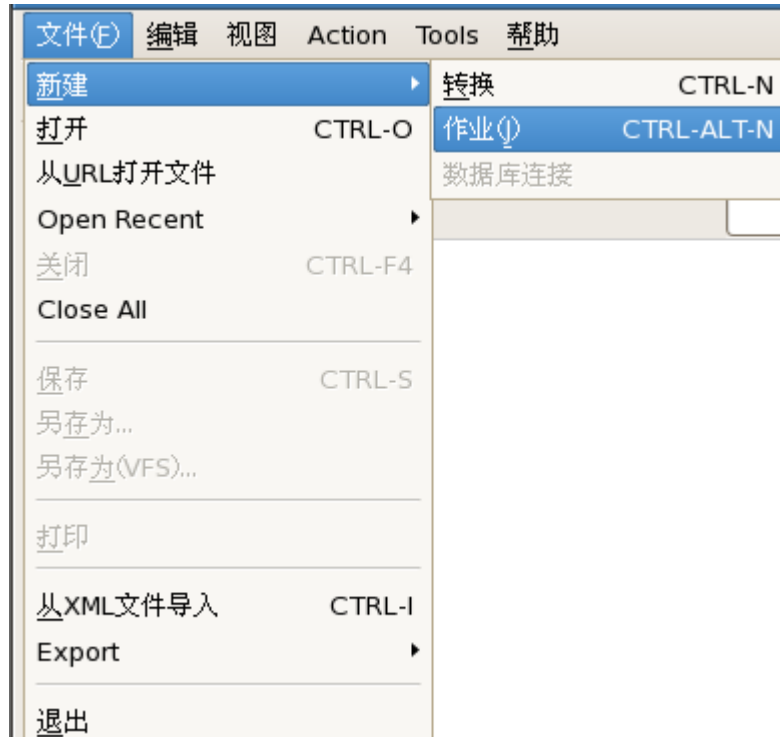
在wfTransaction上点击执行



可以看到有19条记录已经传输完毕

### 3.6. 调度执行

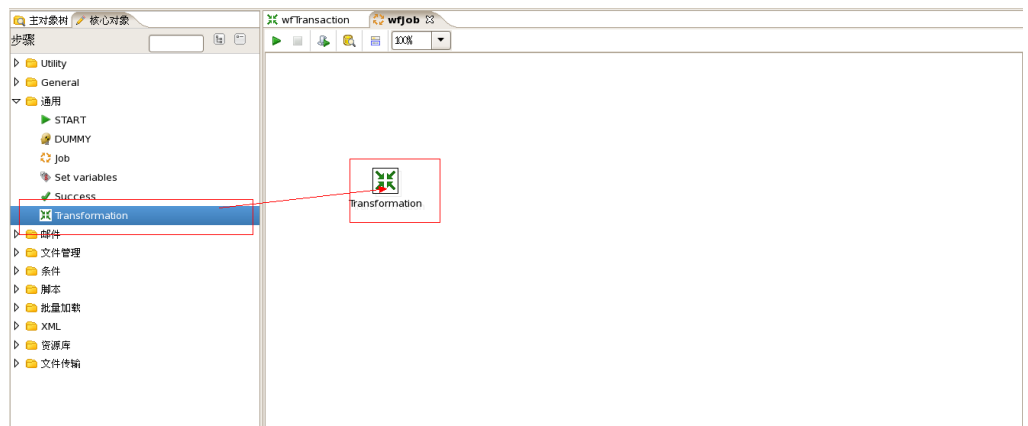
在文件->新建->作业，新建一个作业



在新建的job上右键，编辑，修改job属性



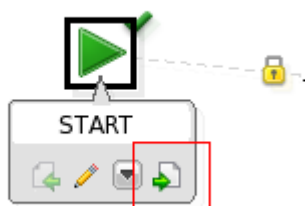
选择wfJob，在核心对象->通用下面选择Transformation



双击Transformation，选择转换名称



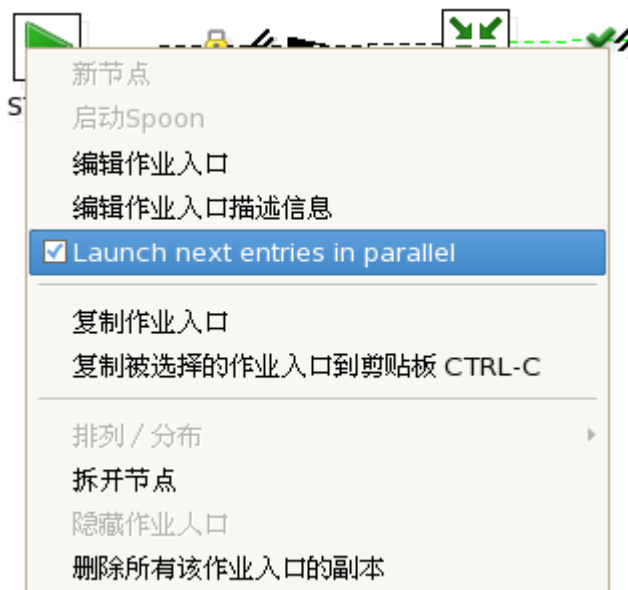
在选择的时候选择右键头 ( shift+ 鼠标拖动 )



可以修改指示的条件



需要设置如下：



执行调度的脚步为：

```
./kitchen.sh -rep=kettle1 -user=admin -pass=admin -level=Basic -job=wfJob
```

## 4. 常见问题

### 4.1. No repository defined

---

No repository defined

Unexpected error during transformation metadata load

No repository defined!

注意rep=id