

## 实验 2-3 kettle 实现源数据的预处理

建议课时：60 分钟

### 一、实验目的

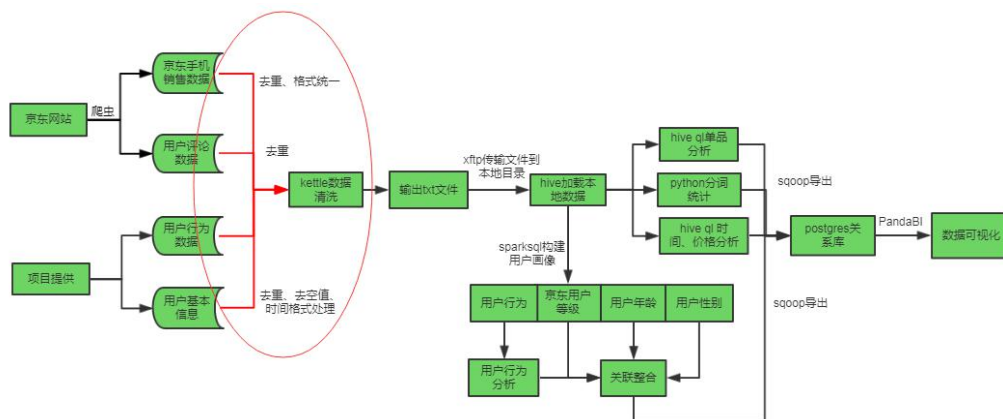
- 熟练使用 kettle 工具做数据预处理；
- 掌握 kettle 中常用步骤的配置；
- 能按清洗目标准确实现数据清洗工作；

### 二、实验环境

Kettle7.0

### 三、实验步骤

本节实验所做内容如下红色标注：



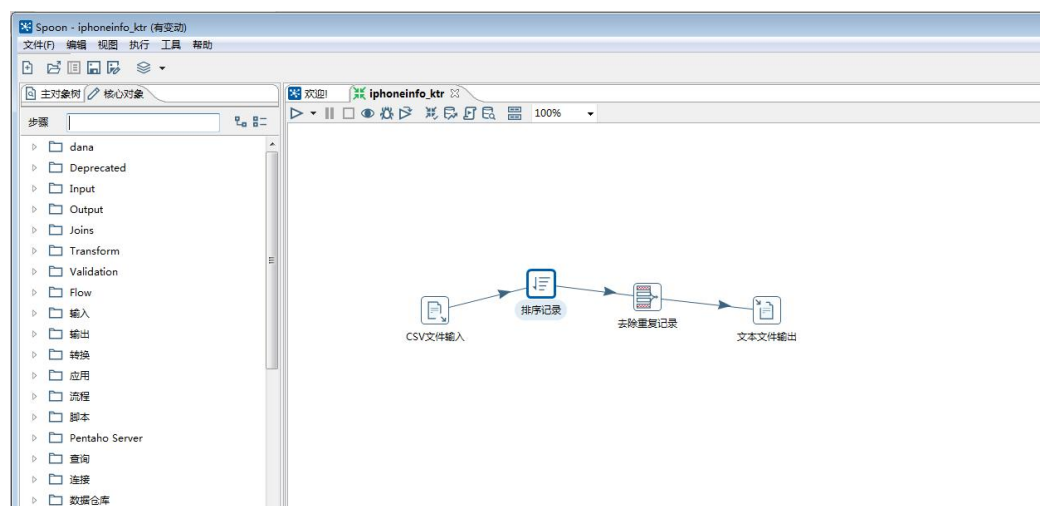
本节实验主要是通过 kettle 实现源数据的预处理。

具体实验步骤如下：

1. 下载安装 kettle7.0 工具
2. 运行 Spoon.bat 文件启动 kettle

file.txt	2019/5/17 16:03	文本文档	1 KB
Import.bat	2018/6/12 14:35	Windows 批处理...	1 KB
import.sh	2018/6/12 14:35	Shell Script	1 KB
import-rules.xml	2018/6/12 14:35	XML 文档	3 KB
Kitchen.bat	2018/6/12 14:35	Windows 批处理...	1 KB
kitchen.sh	2018/6/12 14:35	Shell Script	1 KB
LICENSE.txt	2018/6/12 14:36	文本文档	14 KB
Pan.bat	2018/6/12 14:36	Windows 批处理...	1 KB
pan.sh	2018/6/12 14:36	Shell Script	1 KB
PentahoDataIntegration_OSS_License...	2018/6/12 14:36	HTML 文档	9,114 KB
purge-utility.bat	2018/6/12 14:37	Windows 批处理...	1 KB
purge-utility.sh	2018/6/12 14:37	Shell Script	1 KB
README.txt	2018/6/12 14:37	文本文档	2 KB
runSamples.bat	2018/6/12 14:37	Windows 批处理...	1 KB
runSamples.sh	2018/6/12 14:37	Shell Script	1 KB
set-pentaho-env.bat	2018/6/12 14:37	Windows 批处理...	5 KB
set-pentaho-env.sh	2018/6/12 14:37	Shell Script	4 KB
Spoon.bat	2018/6/12 14:37	Windows 批处理...	4 KB
spoon.command	2018/6/12 14:37	COMMAND 文件	1 KB
spoon.ico	2018/6/12 14:37	WPS看图 ICO 图...	362 KB
spoon.png	2018/6/12 14:37	WPS看图 PNG 图...	2 KB
spoon.sh	2018/6/12 14:37	Shell Script	7 KB
SpoonConsole.bat	2018/6/12 14:37	Windows 批处理...	1 KB
SpoonDebug.bat	2018/6/12 14:37	Windows 批处理...	2 KB
SpoonDebug.sh	2018/6/12 14:37	Shell Script	2 KB
varn.sh	2018/6/12 14:39	Shell Script	2 KB

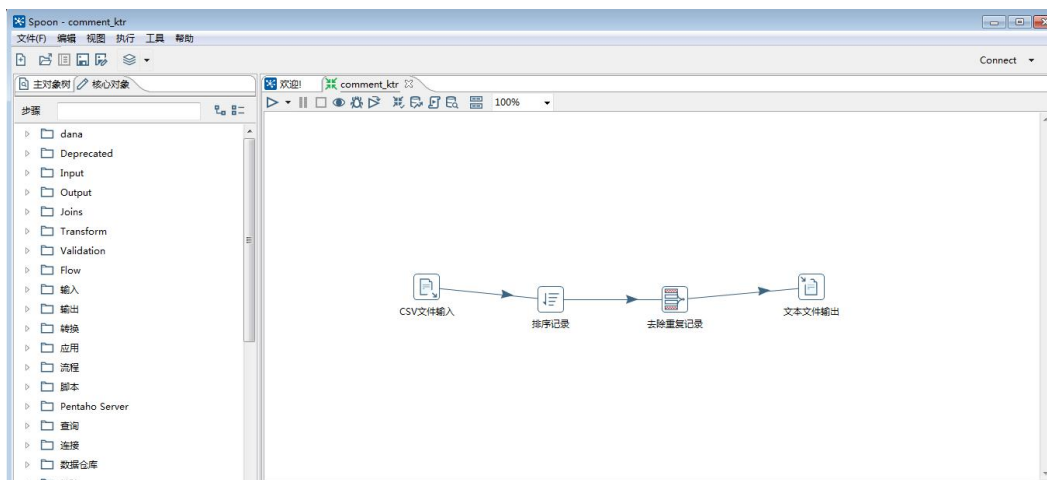
### 3. 新建转换去除手机销售信息表的重复记录



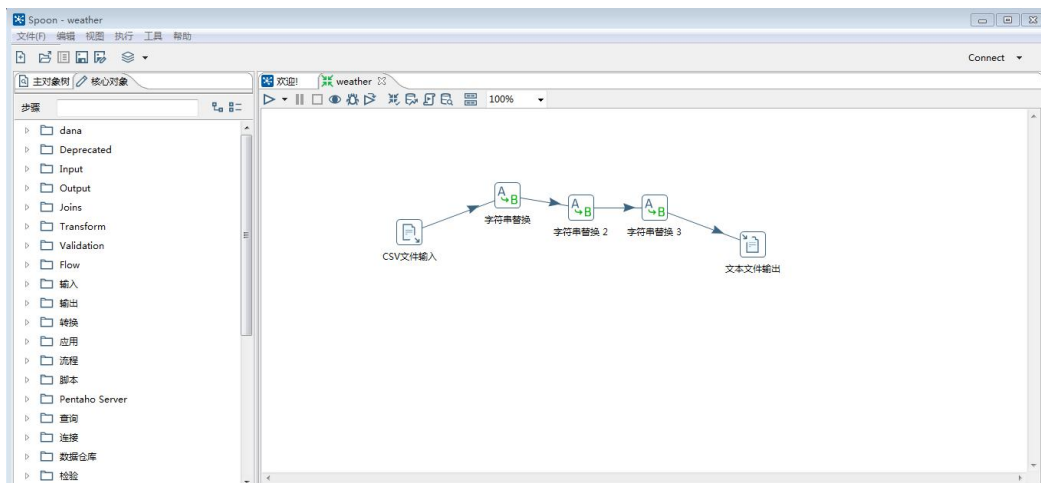
### 4. 在上述转换的基础上清洗手机销售信息表型号字段中的数据

**要求：**去除该字段中的所有空格，方便后续聚合统计，字母统一大小写，去除该字段中的所有特殊字符（各种标点符号）

### 5. 新建转换去除用户评论信息表的重复记录



6. 新建转换处理用户信息表中出生日期字段（将 2019 年 5 月 20 日转换为 2019-5-20）



7. 在每个转换中的文本文件输出步骤中，指定分隔符为英文状态下的逗号，去除输出文件的表头数据，编码格式保持一致，最终的编码格式为 utf-8。



## 四、实验成果

本次实验完成后，需要得到以下结果：

- 去除手机销售信息表的重复记录；
- 去除用户评论表的重复记录；
- 清洗手机销售信息表中型号字段数据（同一型号数据格式保持一致）；
- 对用户信息表中的出生日期字段进行格式处理；
- 清洗完成后保存为文本文件；