

# Kettle

Kettle 是一款纯 Java 编写的开源 ETL 工具。Kettle 是 PDI（Pentaho Data Integration）的前身。表示抽取、转换、装入和加载数据，翻译成中文是水壶的意思，希望把各种数据放到一个壶里，像水一样，以一种指定的格式流出，表达数据流的含义。

Kettle 共有两类设计模式，转换（Transformation）和作业（Job）；转换用于控制数据处理和流转，作业则用于转换集成和流程控制。

## ETL

ETL 的 3 个字母分别代表 Extract（抽取）、Transform（转换）和 Load（装载）。

数据抽取：从数据源端的系统中，抽取目标端系统需要的数据。

数据转换：从数据源端获取的数据按照业务需求，转换成目标端要求的数据形式，并对错误、不规范、不一致的数据（俗称“脏数据”）进行清洗和加工。

数据装载：将转换后的数据装载到指定数据库或文件中。

## 作业练习

1.处理等于 3000 抽取到 emp3000 , 不等于抽取到 empnot3000

欢迎! k06131

100%

```
graph LR; A[表输入] --> B[表输出]; A --> C[过滤记录]; C --> D[表输出 2]; C --> E[表输出 3];
```

### 执行结果

日志 | 执行历史 | 步骤度量 | 性能图 | Metrics | Preview data

2025/06/13 11:29:44 - Spoon - Running transformation using the Kettle execution engine

2025/06/13 11:29:44 - Spoon - 转换已经打开.

2025/06/13 11:29:44 - Spoon - 正在打开转换 [k06131]...

2025/06/13 11:29:44 - Spoon - 开始执行转换.

2025/06/13 11:29:44 - k06131 - 为了转换解除补丁开始 [k06131]

2025/06/13 11:29:44 - 表输出 3.0 - Connected to database [TAG] (commit=1000)

2025/06/13 11:29:44 - 表输出 2.0 - Connected to database [TAG] (commit=1000)

2025/06/13 11:29:44 - 表输入.0 - Finished reading query, closing connection.

2025/06/13 11:29:44 - 表输入.0 - 完成处理 (I=14, O=0, R=0, W=14, U=0, E=0)

2025/06/13 11:29:44 - 过滤记录.0 - 完成处理 (I=0, O=0, R=14, W=14, U=0, E=0)

2025/06/13 11:29:44 - 表输出 3.0 - 完成处理 (I=0, O=12, R=12, W=12, U=0, E=0)

2025/06/13 11:29:44 - 表输出 2.0 - 完成处理 (I=0, O=2, R=2, W=2, U=0, E=0)

2025/06/13 11:29:44 - Spoon - 转换完成!!

```
select * from emp3000
```

	EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
1	7788	SCOTT	ANALYST	7566	1987/4/19	3000.00		20
2	7902	FORD	ANALYST	7566	1981/12/3	3000.00		20

```
select * from empnot3000
```

	EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
1	7369	SMITH	CLERK	7902	1980/12/17	800.00		20
2	7499	ALLEN	SALESMAN	7698	1981/2/20	1600.00	300.00	30
3	7521	WARD	SALESMAN	7698	1981/2/22	1250.00	500.00	30
4	7566	JONES	MANAGER	7839	1981/4/2	2975.00		20
5	7654	MARTIN	SALESMAN	7698	1981/9/28	1250.00	1400.00	30
6	7698	BLAKE	MANAGER	7839	1981/5/1	2850.00		30
7	7782	CLARK	MANAGER	7839	1981/6/9	2450.00		10
8	7839	KING	PRESIDENT		1981/11/17	5000.00		10
9	7844	TURNER	SALESMAN	7698	1981/9/8	1500.00	0.00	30
10	7876	ADAMS	CLERK	7788	1987/5/23	1100.00		20
11	7900	JAMES	CLERK	7698	1981/12/3	950.00		30
12	7934	MILLER	CLERK	7782	1982/1/23	1300.00		10

计算每个员工的年薪

表输入

增加常量

计算器

执行结果

日志 执行历史 步骤度量 性能图 Metrics Preview data

● \${TransPreview.FirstRows.Label} ○ \${TransPreview.LastRows.Label} ○ \${TransPreview.Off.Label}

#	EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO	month	ling	newComm	nian
1	7369	SMITH	CLERK	7902	1980/12/17 00:00:00.000000000	800.0	<null>	20	12	0	0	9600.0
2	7499	ALLEN	SALESMAN	7698	1981/02/20 00:00:00.000000000	1600.0	300.0	30	12	0	300	19500.0
3	7521	WARD	SALESMAN	7698	1981/02/22 00:00:00.000000000	1250.0	500.0	30	12	0	500	15500.0
4	7566	JONES	MANAGER	7839	1981/04/02 00:00:00.000000000	2975.0	<null>	20	12	0	0	35700.0
5	7654	MARTIN	SALESMAN	7698	1981/09/28 00:00:00.000000000	1250.0	1400.0	30	12	0	1400	16400.0
6	7698	BLAKE	MANAGER	7839	1981/05/01 00:00:00.000000000	2850.0	<null>	30	12	0	0	34200.0
7	7782	CLARK	MANAGER	7839	1981/06/09 00:00:00.000000000	2450.0	<null>	10	12	0	0	29400.0
8	7788	SCOTT	ANALYST	7566	1987/04/19 00:00:00.000000000	3000.0	<null>	20	12	0	0	36000.0
9	7839	KING	PRESIDENT	<null>	1981/11/17 00:00:00.000000000	5000.0	<null>	10	12	0	0	60000.0
1.	7844	TURNER	SALESMAN	7698	1981/09/08 00:00:00.000000000	1500.0	0.0	30	12	0	0	18000.0
1.	7876	ADAMS	CLERK	7788	1987/05/23 00:00:00.000000000	1100.0	<null>	20	12	0	0	13200.0
1.	7900	JAMES	CLERK	7698	1981/12/03 00:00:00.000000000	950.0	<null>	30	12	0	0	11400.0
1.	7902	FORD	ANALYST	7566	1981/12/03 00:00:00.000000000	3000.0	<null>	20	12	0	0	36000.0
1.	7934	MILLER	CLERK	7782	1982/01/23 00:00:00.000000000	1300.0	<null>	10	12	0	0	15600.0

2.查询所有大于平均工资的员工信息

欢迎! k06131 k06132 k06133 k06134

100%

表输入

分组

设置变量

执行结果

日志 执行历史 步骤度量 性能图 Metrics Preview data

● \${TransPreview.FirstRows.Label} ○ \${TransPreview.LastRows.Label} ○ \${TransPreview.Off.Label}

#	avgsal
1	2073.2142857143

获取变量

表输入

### 执行结果

日志 执行历史 步骤度量 性能图 Metrics Preview data

● \${TransPreview.FirstRows.Label} ○ \${TransPreview.LastRows.Label} ○ \${TransPreview.Off.Label}

#	EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
1	7566	JONES	MANAGER	7839	1981/04/02 00:00:00.000000000	2975.0	<null>	20
2	7698	BLAKE	MANAGER	7839	1981/05/01 00:00:00.000000000	2850.0	<null>	30
3	7782	CLARK	MANAGER	7839	1981/06/09 00:00:00.000000000	2450.0	<null>	10
4	7788	SCOTT	ANALYST	7566	1987/04/19 00:00:00.000000000	3000.0	<null>	20
5	7839	KING	PRESIDENT	<null>	1981/11/17 00:00:00.000000000	5000.0	<null>	10
6	7902	FORD	ANALYST	7566	1981/12/03 00:00:00.000000000	3000.0	<null>	20

3.表 A 和表 B 数据结构完全相同，表自建，数据自定义 表 A 数据存在新增 & 修改 & 删除 操作，实现表 A 和表 B 的数据保持一致

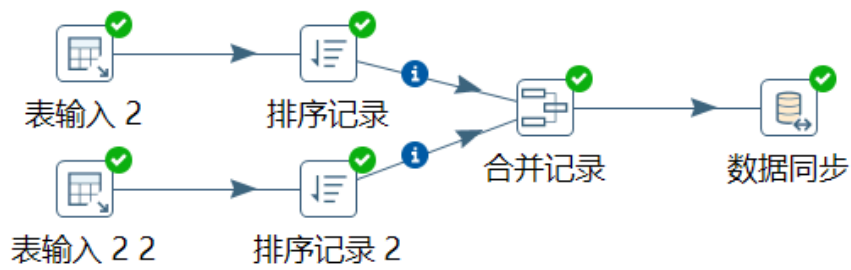
```
create table empA as select * from emp
create table empB as select * from emp
```

```
insert into empA(empno,ename) values(1111,'Tetuka');
update empA set ename='SSSSS' where empno=7369
delete from empA where ename='KING'
```

```
select * from empA
```

	EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
1	7369	SSSSS	CLERK	7902	1980/12/17	800.00		20
2	7499	ALLEN	SALESMAN	7698	1981/2/20	1600.00	300.00	30
3	7521	WARD	SALESMAN	7698	1981/2/22	1250.00	500.00	30
4	7566	JONES	MANAGER	7839	1981/4/2	2975.00		20
5	7654	MARTIN	SALESMAN	7698	1981/9/28	1250.00	1400.00	30
6	7698	BLAKE	MANAGER	7839	1981/5/1	2850.00		30
7	7782	CLARK	MANAGER	7839	1981/6/9	2450.00		10
8	7788	SCOTT	ANALYST	7566	1987/4/19	3000.00		20
9	7844	TURNER	SALESMAN	7698	1981/9/8	1500.00	0.00	30
10	7876	ADAMS	CLERK	7788	1987/5/23	1100.00		20
11	7900	JAMES	CLERK	7698	1981/12/3	950.00		30
12	7902	FORD	ANALYST	7566	1981/12/3	3000.00		20
13	7934	MILLER	CLERK	7782	1982/1/23	1300.00		10
14	1111	Tetuka						

(表A修改后记得提交! )



## 执行结果

日志 执行历史 步骤度量 性能图 Metrics Preview data



2025/06/13 16:43:29 - Spoon - Running transformation using the Kettle execution engine  
2025/06/13 16:43:29 - Spoon - 转换已经打开.  
2025/06/13 16:43:29 - Spoon - 正在打开转换 [k06136]...  
2025/06/13 16:43:29 - Spoon - 开始执行转换.  
2025/06/13 16:43:29 - k06136 - 为了转换解除补丁开始 [k06136]  
2025/06/13 16:43:29 - 表输入 2 2.0 - Finished reading query, closing connection.  
2025/06/13 16:43:29 - 表输入 2.0 - Finished reading query, closing connection.  
2025/06/13 16:43:29 - 表输入 2 2.0 - 完成处理 (I=14, O=0, R=0, W=14, U=0, E=0)  
2025/06/13 16:43:29 - 表输入 2.0 - 完成处理 (I=14, O=0, R=0, W=14, U=0, E=0)  
2025/06/13 16:43:29 - 排序记录.0 - 完成处理 (I=0, O=0, R=14, W=14, U=0, E=0)  
2025/06/13 16:43:29 - 排序记录 2.0 - 完成处理 (I=0, O=0, R=14, W=14, U=0, E=0)  
2025/06/13 16:43:29 - 合并记录.0 - 完成处理 (I=0, O=0, R=28, W=15, U=0, E=0)  
2025/06/13 16:43:29 - 数据同步.0 - 完成处理 (I=0, O=1, R=15, W=15, U=2, E=0)  
2025/06/13 16:43:29 - Spoon - 转换完成!!

执行后表B与表A数据一致（将表A更新到表B）

```
select * from empB
```

	EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
1	7369	SSSSS	CLERK	7902	1980/12/17	800.00		20
2	7499	ALLEN	SALESMAN	7698	1981/2/20	1600.00	300.00	30
3	7521	WARD	SALESMAN	7698	1981/2/22	1250.00	500.00	30
4	7566	JONES	MANAGER	7839	1981/4/2	2975.00		20
5	7654	MARTIN	SALESMAN	7698	1981/9/28	1250.00	1400.00	30
6	7698	BLAKE	MANAGER	7839	1981/5/1	2850.00		30
7	7782	CLARK	MANAGER	7839	1981/6/9	2450.00		10
8	7788	SCOTT	ANALYST	7566	1987/4/19	3000.00		20
9	7844	TURNER	SALESMAN	7698	1981/9/8	1500.00	0.00	30
10	7876	ADAMS	CLERK	7788	1987/5/23	1100.00		20
11	7900	JAMES	CLERK	7698	1981/12/3	950.00		30
12	7902	FORD	ANALYST	7566	1981/12/3	3000.00		20
13	7934	MILLER	CLERK	7782	1982/1/23	1300.00		10
14	1111	Tetuka						

4.数据来自 excel ，利用 PRODUCTID 和 CUSTOMID 进行关联后，插入到本地数据库中的一个表中

--建表

```
create table pcid(  
  ORDERID varchar2(100),  ZTIME varchar2(100),  
  PRODUCTID varchar2(100),  SALE_AMOUNT number(5),  
  CUSTOMID varchar2(100),  PRICE number(10,2),  
  CUSTOMNAME varchar2(100),  COUNTRY varchar2(100)  
);
```



2025/06/13 16:56:16 - Spoon - 转换完成!!

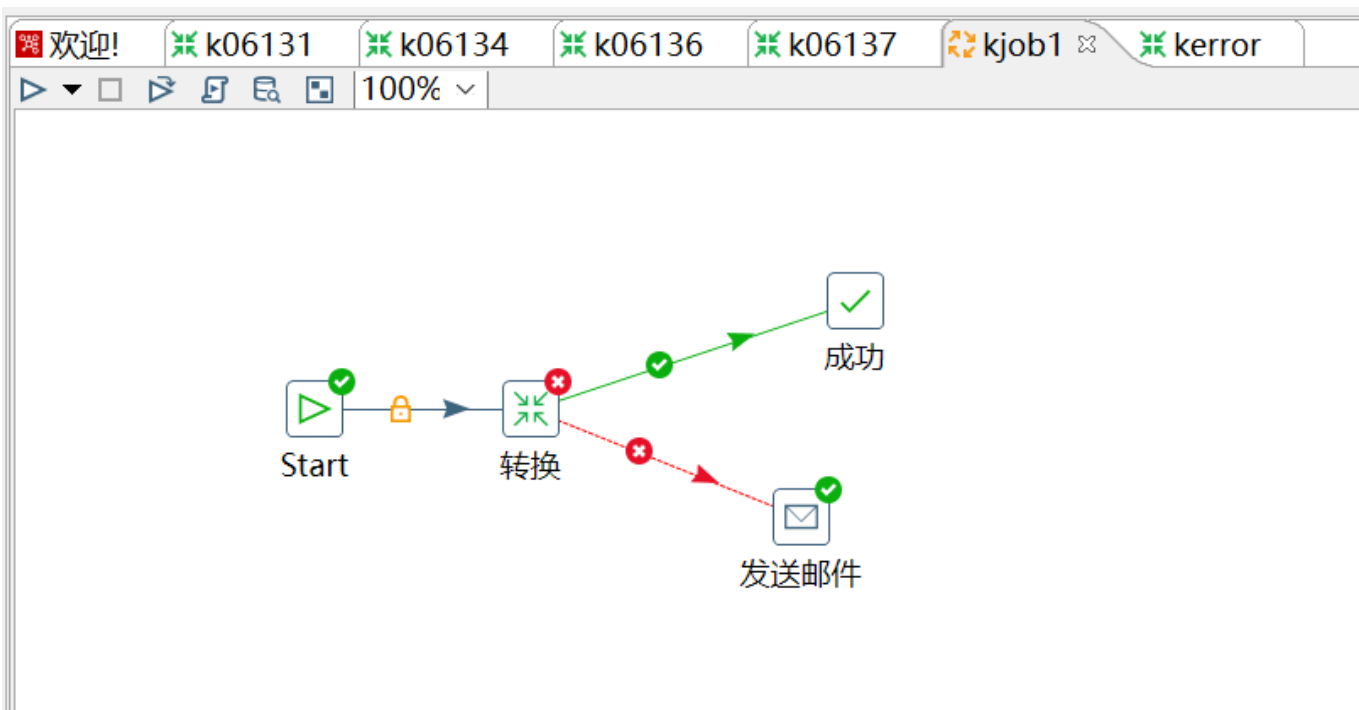
执行后查看表pcid

select \* from pcid

	ORDERID	ZTIME	PRODUCTID	SALE_AMOUNT	CUSTOMID	PRICE	CUSTOMNAME	COUNTRY
1	0001	8/8/2020	P005	11	C002	49.37	客户B	法国
2	0005	8/13/2020	P006	15	C002	10.20	客户B	法国
3	0006	9/4/2020	P009	99	C002	50.28	客户B	法国
4	0009	8/23/2020	P002	73	C003	49.84	客户C	中国
5	0007	8/12/2020	P007	86	C003	22.87	客户C	中国
6	0005	9/12/2020	P001	10	C004	29.18	客户D	英国
7	0004	9/9/2020	P003	99	C004	41.68	客户D	英国
8	0003	9/17/2020	P003	57	C004	41.68	客户D	英国
9	0009	8/15/2020	P004	9	C004	58.88	客户D	英国
10	0007	9/22/2020	P001	73	C005	29.18	客户E	美国
11	0009	8/23/2020	P003	3	C005	41.68	客户E	美国
12	0002	9/29/2020	P005	59	C005	49.37	客户E	美国
13	0005	9/6/2020	P006	3	C005	10.20	客户E	美国
14	0004	8/15/2020	P003	45	C006	41.68	客户F	中国
15	0002	9/27/2020	P005	62	C006	49.37	客户F	中国
16	00010	8/22/2020	P007	54	C006	22.87	客户F	中国
17	0002	8/3/2020	P009	22	C006	50.28	客户F	中国
18	0009	9/4/2020	P001	34	C007	29.18	客户G	法国
19	0007	9/22/2020	P003	70	C007	41.68	客户G	法国
20	0003	8/29/2020	P008	38	C007	23.78	客户G	法国
21	0002	8/30/2020	P009	86	C007	50.28	客户G	法国
22	0001	9/20/2020	P002	3	C008	49.84	客户H	英国
23	0007	9/22/2020	P003	50	C008	41.68	客户H	英国
24	0005	8/23/2020	P004	56	C008	58.88	客户H	英国
25	0004	8/8/2020	P006	51	C008	10.20	客户H	英国
26	0001	8/10/2020	P007	97	C008	22.87	客户H	英国

新建作业，发送邮件

SMTP 服务需要自己登录 qq 邮箱开通：账号与安全-安全设置-根据提示开通服务



转换错误执行发送邮件，之后成功接收邮件



使用 windows 自带任务计划管理程序定时调用 Kettle 作业，同时输出 BASIC 级别日志  
将以下代码写入到winjob.bat

```
D:
cd D:\data-integration
kitchen /rep huangwenzhe /user admin /pass admin /file
D:\kettle\huangwenzhe\kjob1.kjb /level basic>D:\kettle\huangwenzhe\kettle.log
```

Windows管理工具-任务计划程序-操作-创建基本任务



到点任务执行，发送邮件并生成log



```
kettle.log - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
18:43:53,890 INFO [KarafInstance]
*****
*** Karaf Instance Number: 2 at D:\data-integration\.\system\karaf\caches\k ***
*** itchen\data-1 ***
*** Karaf Port:8803 ***
*** OSGI Service Port:9052 ***
*****
2025/06/13 18:43:55 - Kitchen - Logging is at level : 基本日志
2025/06/13 18:43:55 - Kitchen - Start of run.
2025/06/13 18:43:55 - RepositoriesMeta - Reading repositories XML file: C:\Users\HSanae'
ERROR: Kitchen can't continue because the job couldn't be loaded.
18:44:17,164 INFO [DriverManager] Installing driver kars.
18:44:17,165 INFO [DriverManager] 0 drivers will be installed.
18:44:17,165 INFO [DriverManager] Finished installing drivers kars.
2025/06/13 18:44:17 - kjob1 - 开始执行任务
2025/06/13 18:44:17 - Carte - Installing timer to purge stale objects after 1440 minutes.
2025/06/13 18:44:17 - kjob1 - 开始项[转换]
2025/06/13 18:44:17 - 转换 - Using run configuration [Pentaho local]
2025/06/13 18:44:17 - 转换 - Running transformation using the Kettle execution engine
2025/06/13 18:44:17 - kerror - 为了转换解除补丁开始 [kerror]
2025/06/13 18:44:17 - 获取变量.0 - 完成处理 (I=0, O=0, R=1, W=0, U=0, E=1)
2025/06/13 18:44:17 - kerror - 转换被检测
2025/06/13 18:44:17 - kerror - 转换正在杀死其他步骤!
2025/06/13 18:44:17 - 表输入.0 - Finished reading query, closing connection.
2025/06/13 18:44:17 - 表输入.0 - 完成处理 (I=0, O=0, R=0, W=0, U=0, E=1)
2025/06/13 18:44:17 - kerror - 转换被检测
2025/06/13 18:44:17 - kerror - 转换正在杀死其他步骤!
2025/06/13 18:44:17 - kjob1 - 开始项[发送邮件]
2025/06/13 18:44:27 - kjob1 - 完成作业项[发送邮件] (结果=[true])
2025/06/13 18:44:27 - kjob1 - 完成作业项[转换] (结果=[true])
2025/06/13 18:44:27 - kjob1 - 任务执行完毕
2025/06/13 18:44:27 - Kitchen - Finished!
2025/06/13 18:44:27 - Kitchen - Start=2025/06/13 18:43:55.430, Stop=2025/06/13 18:44:2
2025/06/13 18:44:27 - Kitchen - Processing ended after 32 seconds.

< 第 1 行, 第 1 列 100% Windows (CRLF) ANSI
```