Kettle 使用手册（Windows 版）

一、测试对象

Kettle 版本：Pentaho pdi-ce-8.2.0.0-342
Oracle 版本：Oracle 11g EER 11.1.0.6.0-Production
TiDB 版本：TiDB-v4.0.8 TiDB Server
jdk 版本：1.8.0 及以上版本

二、部署搭建

1.1 安装 Java 环境

下载地址：https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html
首先，按照上述地址下载安装包，一路点击，结果如图 1。

图 1

最后，配置 JAVA_HOME 环境变量,Kettle 启动会读此变量,配置方法如图 2。
注意：即使安装 JDK 成功（CMD 验证成功），未配置环境变量也无法启动 Kettle，会报错 found java one folder up/DEBUG: _PENTAHO_JAVA_HOME。

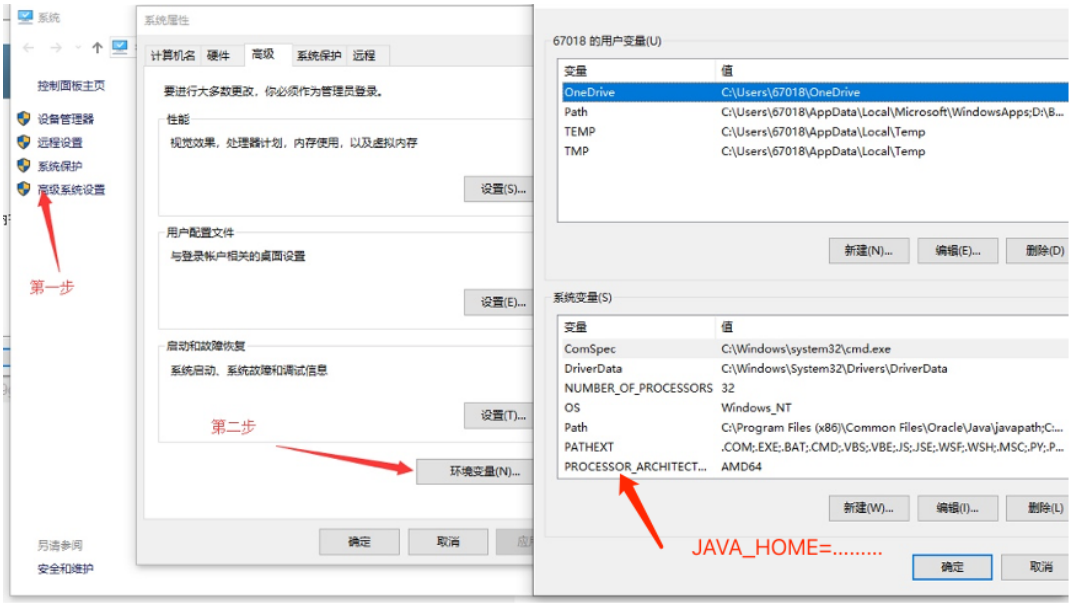


图 2

1.2 下载 kettle 软件

下载地址： http://mirror.bit.edu.cn/pentaho/Pentaho%208.2/client-tools/
下载完毕后，直接解压，双击 Spoon.bat 运行，如图 3。

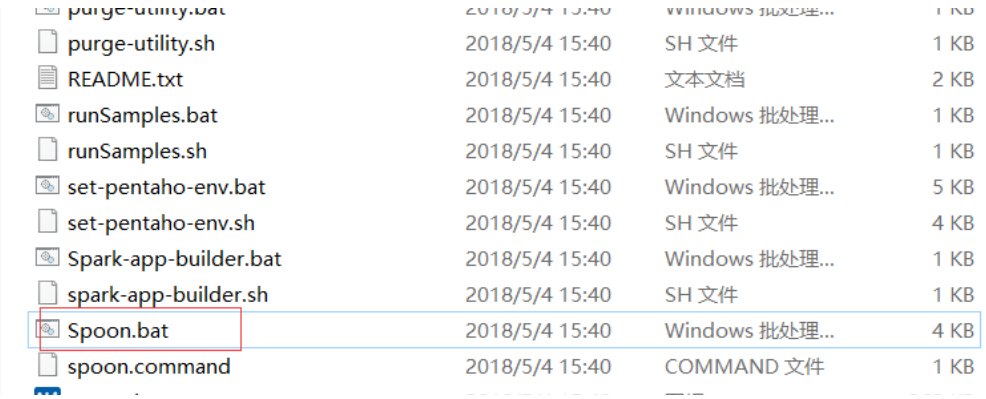


图 3

1.3 添加 Jar 依赖包

Ojdbc5.jar 是从 Oracle 的 $ORACLE_HOME/jdbc/lib 下拷贝至 kettle 的 lib 目录。
否则，kettle 连接 Oracle 数据库时会报错：Driver class ‘oracle.jdbc.driver.OracleDriver’ could not be found, make sure the ‘Oracle’ driver (jar file) is installed。

Mysql-connector-java-5.1.49.jar 从 MySQL 官网下载，拷贝至 lib 目录。

否则，kettle 连接 MySQL 数据库时报错：Driver class 'org.gjt.mm.mysql.Driver' could not be found, make sure the 'MySQL' driver (jar file) is installed.org.gjt.mm.mysql.Driver。

将 jar 包添加至 lib 目录下后，结果如图 4。

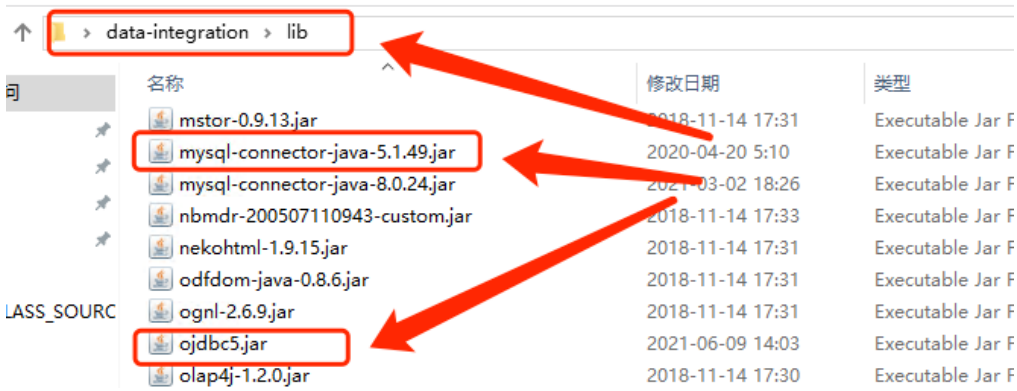


图 4

Oracle solution 参考文章：http://blog.itpub.net/27571661/viewspace-1807978/

MySQL solution 参考文章：https://blog.csdn.net/qq_44895681/article/details/108316605

1.4 修改内存参数

修改内存使用参数，如图 5。常见设置如下：

（1）-Xms、 -Xmx 设置相等的值，以避免在每次 GC 后调整堆的大小；

（2）-Xmn 为 1/4 的-Xmx 值，新生代堆内存解释参考（https://blog.csdn.net/xuheng8600/article/details/81478426）；

注意：配置完参数后，一定要重启 Kettle 生效参数。

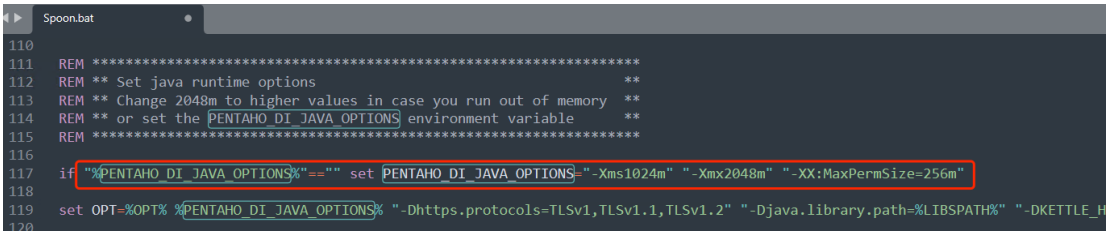


图 5

参数行为：

（1）JVM 初始分配的内存由-Xms 指定，默认是物理内存的 1/64；

（2）JVM 最大分配的内存由-Xmx 指 定，默认是物理内存的 1/4；

（3）默认空余堆内存小于 40%时，JVM 就会增大堆直到-Xmx 的最大限制；

（4）空余堆内存大于 70%时，JVM 会减少堆直到-Xms 的最小限制。

1.5 创建转换（ktr）任务

点击文件→新建→转换，创建转换任务，如图 6。

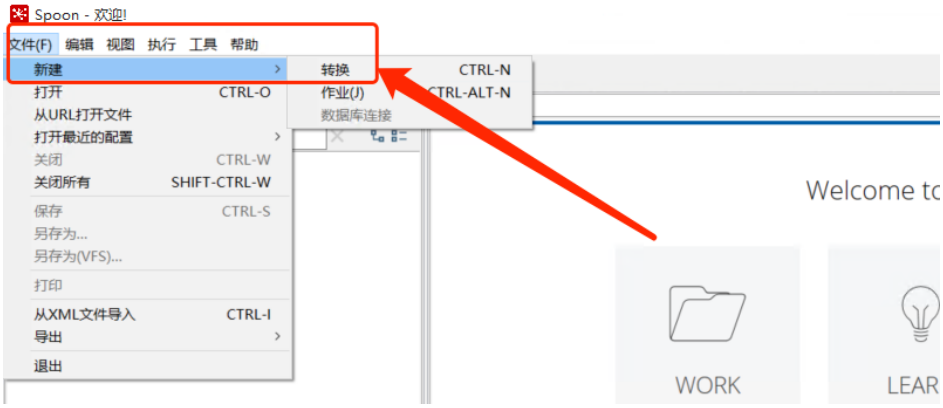


图 6

Ctrl + S 保存转换任务为文件，以便重用，如图 7。

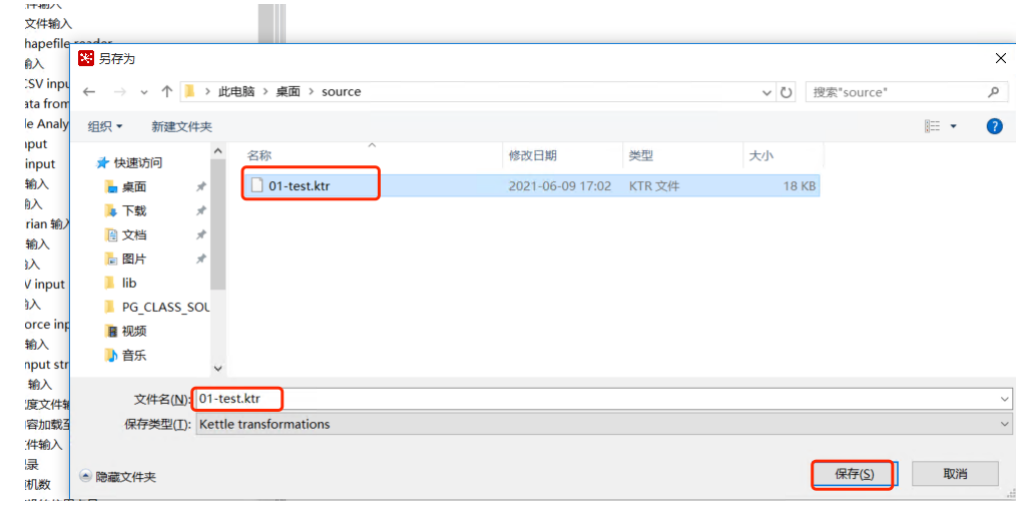


图 7

1.6 初始化 Oracle 连接

第一步：点击【主对象树】→【DB 连接】→【新建】，如图 8。

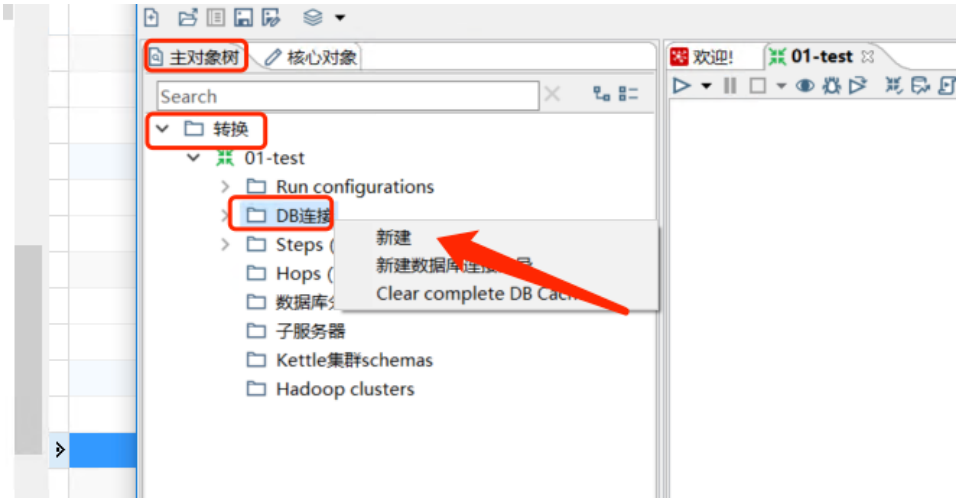


图 8

第二步：依据选框要求，填写连接数据库的必要选项，如图 9。

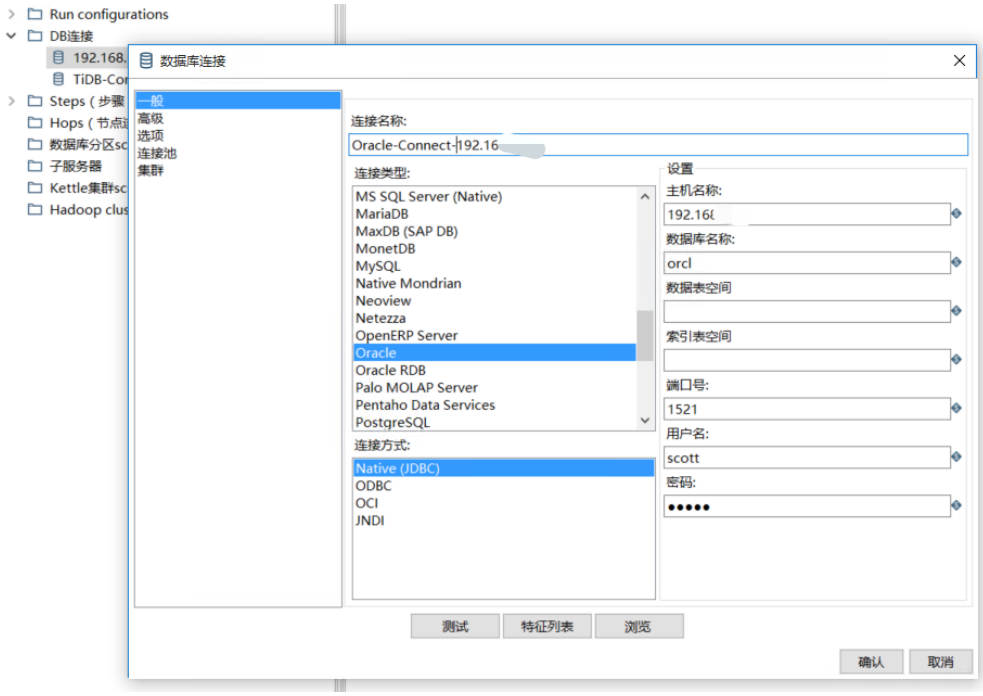


图 9

1.7 创建 steps--表输入

第一步：在【转换】→【核心对象】→双击【表输入】，或选中将【表输入】拖拽到右侧空白区域，如图 10。

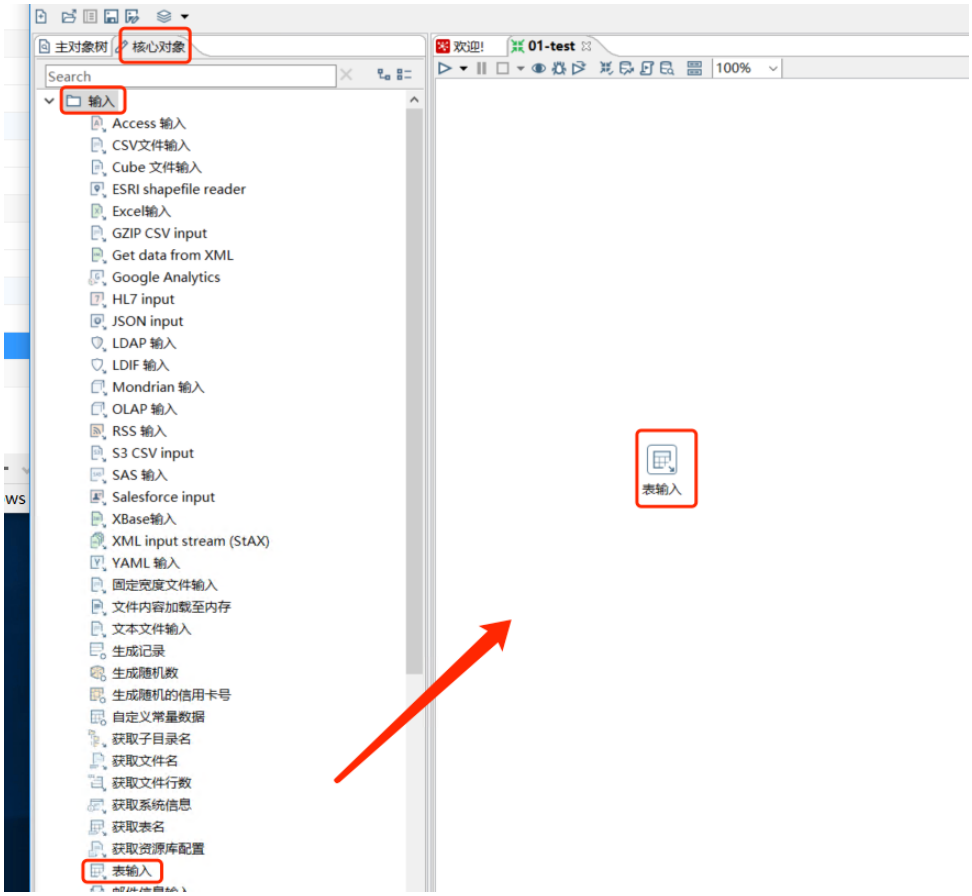


图 10

第二步：双击你拖进来的【表输入】，修改“步骤名称”，选择源数据，点击获取【获取 SQL 查询语句】，选择同步表点击确定即可，也可以自己写 SQL 语句，如图 11。

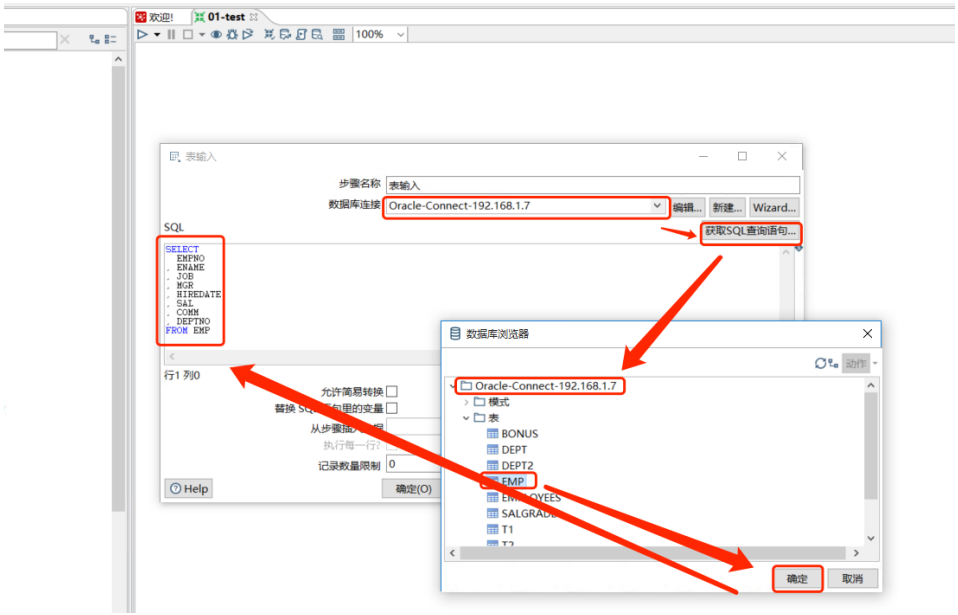


图 11

1.8 初始化 MySQL 连接

操作步骤与初始化 Oracle 步骤相同，如图 12。

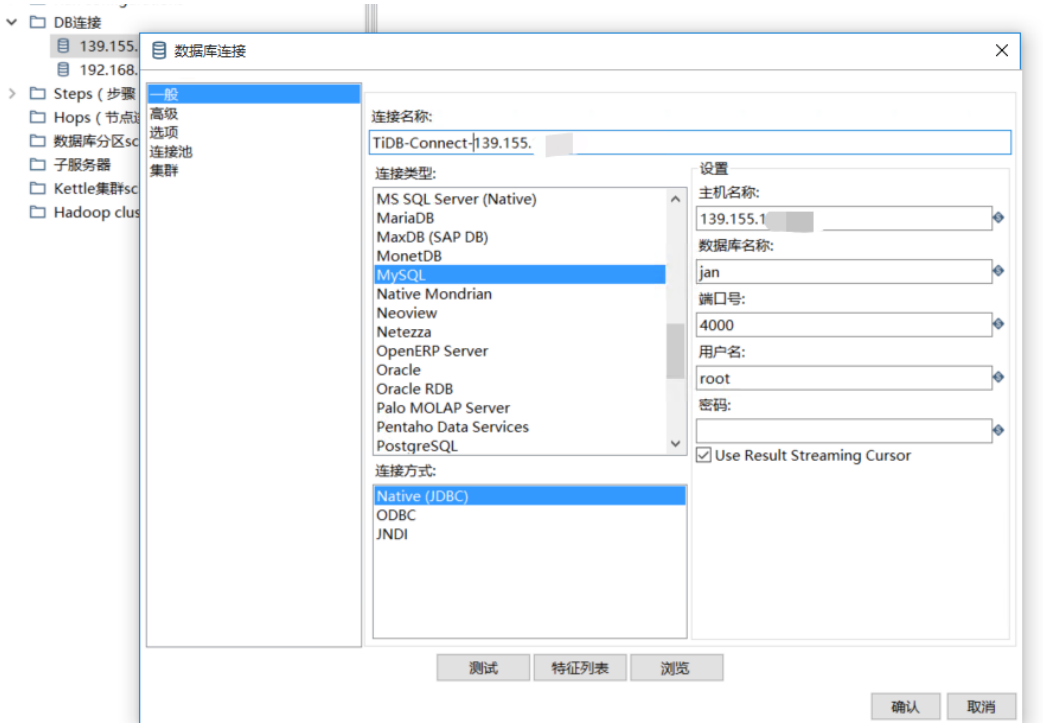


图 12

1.9 创建 steps--表输出

第一步：【转换】→拖拽【表输出】到右侧空白区域，按住 shift 拖动鼠标连接【表输入】和【表输出】，如图 13、图 14；

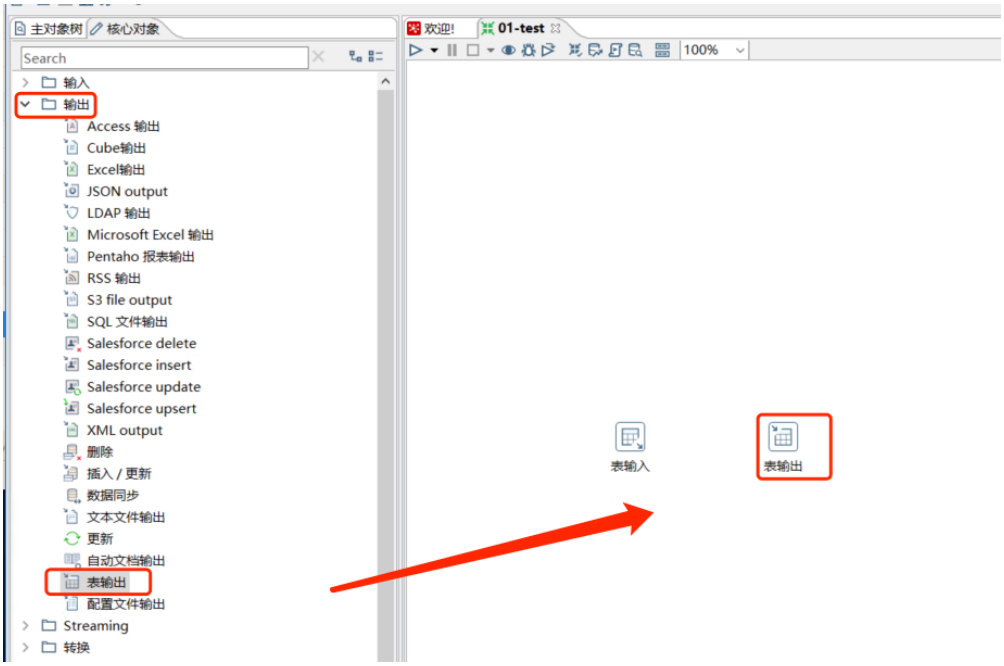


图 13

图 14

第二步：双击【表输出】，修改“步骤名称”，选择“数据库连接”，选择“目标表”，获取字段，如图 15、图 16。

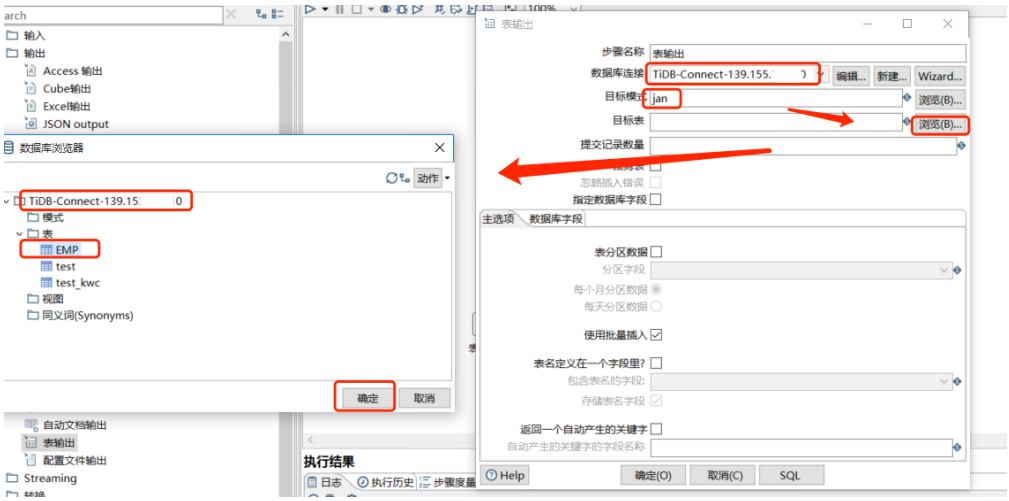


图 15

注意：提交记录数量必须填写数值，如图 16，否则执行时会报错 java.lang.NumberFormatException:null。

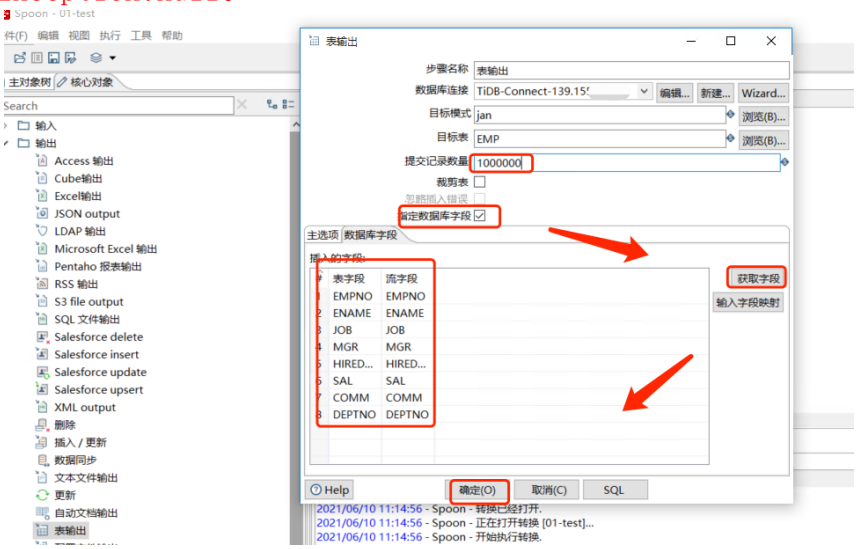


图 16

2.1 验证结果

第一步：点击启动按钮，弹确认窗，点击【启动】确认执行，如图 17；

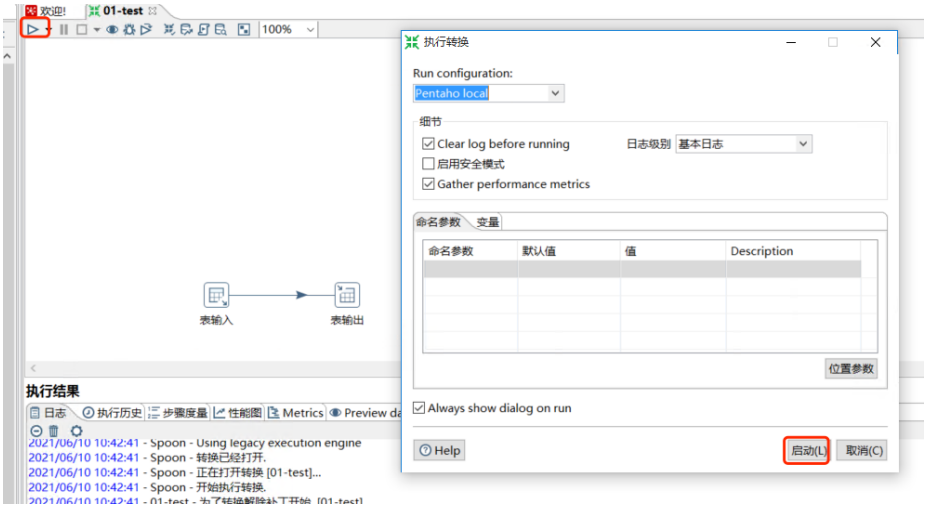


图 17

第二步：点击【是】，保存转转换到之前创建的 ktr 文件中，如图 18；

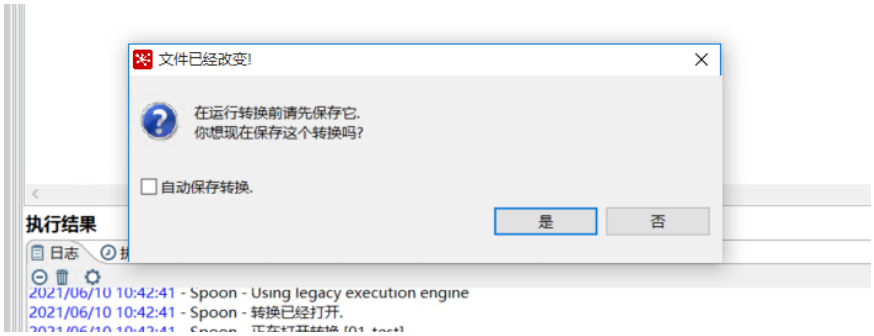


图 18

第三步：查看日志无报错，如图 19；

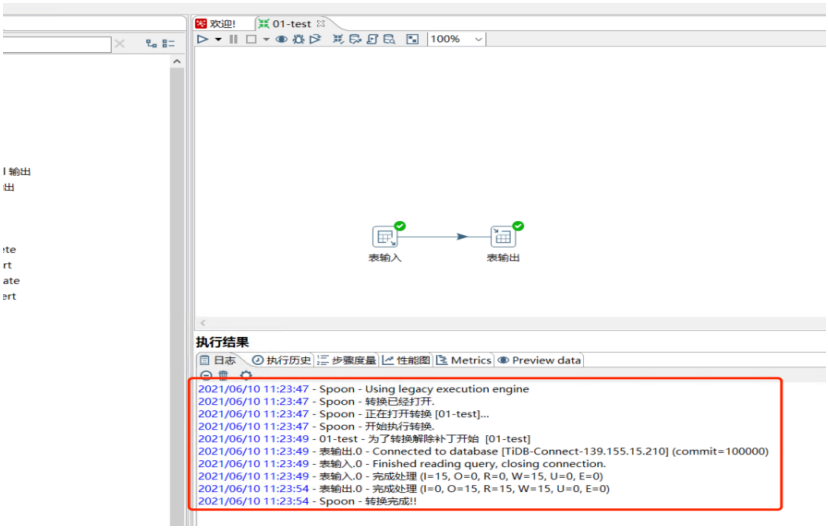


图 19

第四步：结果集验证，数据迁移准确无误，如图 20。

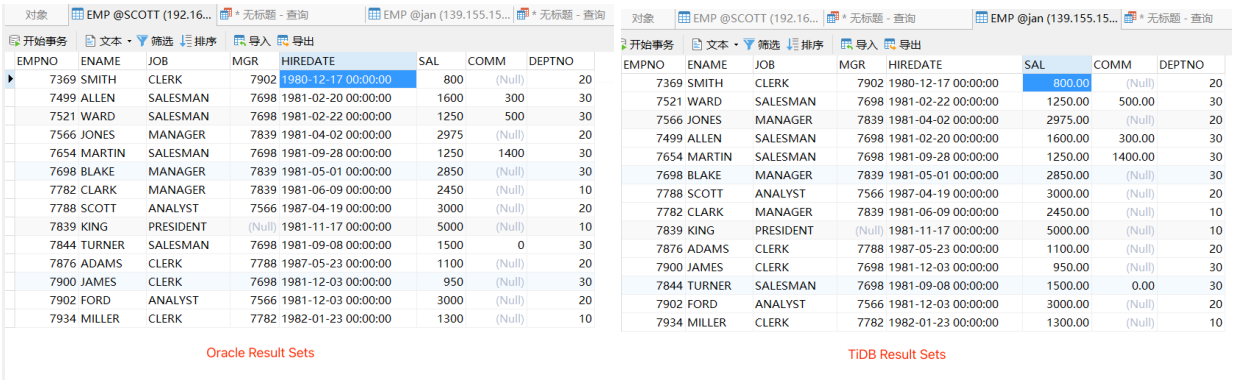


图 20

三、效率对比

1.1 单线程与多线程

（1）单线程执行效率如下，表输出总耗时 2.0s，平均插入 8 条/s，如图 21。

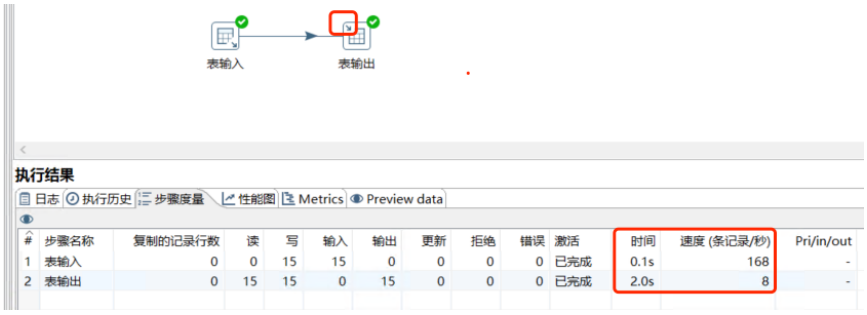


图 21

（2）右键【表输出】改变“更改开始复制的数量”的值，可使用多线程导入加快导入速度，如图 22。

注意：表输入不可以使用多线程，否则会导致重复数据。

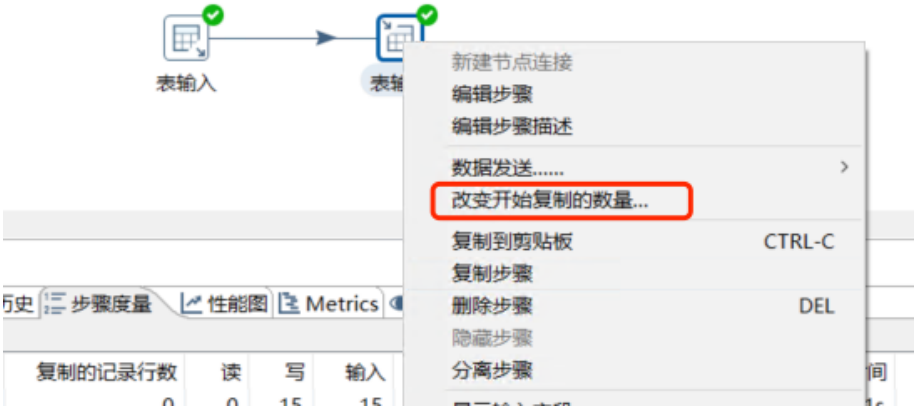
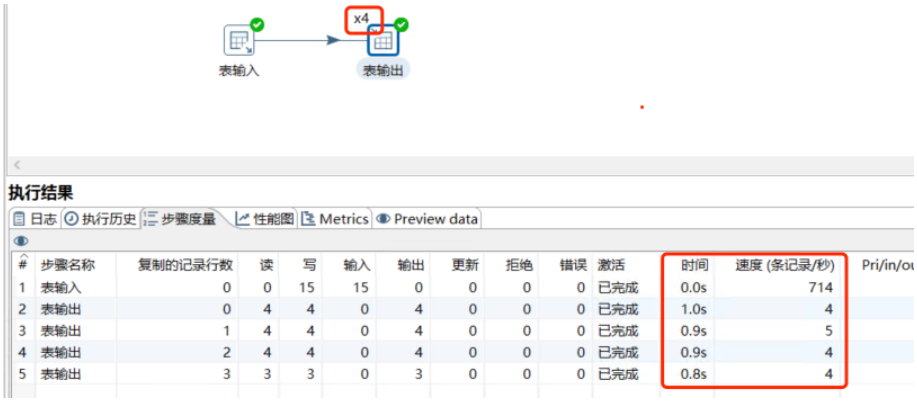


图 22

（3）4 个并发，表输出总耗时约 1.0s，平均每线程插入 4 条/s，如图 23。



#	步骤名称	复制的记录行数	读	写	输入	输出	更新	拒绝	错误	激活	时间	速度 (条记录/秒)	Pri/in/ou
1	表输入	0	0	15	15	0	0	0	0	已完成	0.0s	714	
2	表输出	0	4	4	0	4	0	0	0	已完成	1.0s	4	
3	表输出	1	4	4	0	4	0	0	0	已完成	0.9s	5	
4	表输出	2	4	4	0	4	0	0	0	已完成	0.9s	4	
5	表输出	3	3	3	0	3	0	0	0	已完成	0.8s	4	

图 23

四、Oracle 迁移至 TiDB 问题

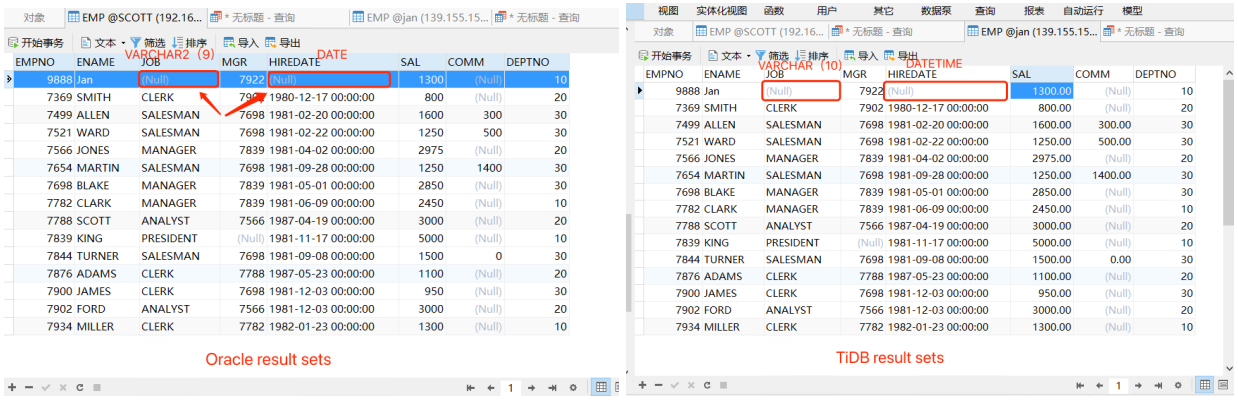
1.1 Oracle 与 TiDB（MySQL）对空串行为判定不一致问题

结论：Kettle 天生自动规避空串行为不一致问题，Kettle 从 Oracle 读取空串时会自动识别成 NULL。

Oracle 源端 EMP 表插入测试数据，测试语句如下：

```
INSERT INTO SCOTT.EMP(empno、ename、job、mgr、hiredate、sal、comm、deptno) VALUES (9888,'jan','',7922,'',1300,NULL,10);
```

迁移后，对比两端结果集，如图 23，Oracle VARCHAR2 类型空串自动转换成 TiDB NULL，Oracle DATE 类型空串自动转换成 TiDB NULL。



EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
9888	Jan	(Null)	7922	(Null)	1300	(Null)	10
7369	SMITH	CLERK	7902	1980-12-17 00:00:00	800	(Null)	20
7499	ALLEN	SALESMAN	7698	1981-02-20 00:00:00	1600	300	30
7521	WARD	SALESMAN	7698	1981-02-22 00:00:00	1250	500	30
7566	JONES	MANAGER	7839	1981-04-02 00:00:00	2975	(Null)	20
7654	MARTIN	SALESMAN	7698	1981-09-28 00:00:00	1250	1400	30
7698	BLAKE	MANAGER	7839	1981-05-01 00:00:00	2850	(Null)	30
7782	CLARK	MANAGER	7839	1981-06-09 00:00:00	2450	(Null)	10
7788	SCOTT	ANALYST	7566	1987-04-19 00:00:00	3000	(Null)	20
7839	KING	PRESIDENT	(Null)	1981-11-17 00:00:00	5000	(Null)	10
7844	TURNER	SALESMAN	7698	1981-09-08 00:00:00	1500	0	30
7876	ADAMS	CLERK	7788	1987-05-23 00:00:00	1100	(Null)	20
7900	JAMES	CLERK	7698	1981-12-03 00:00:00	950	(Null)	30
7902	FORD	ANALYST	7566	1981-12-03 00:00:00	3000	(Null)	20
7934	MILLER	CLERK	7782	1982-01-23 00:00:00	1300	(Null)	10

EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
9888	Jan	(Null)	7922	(Null)	1300.00	(Null)	10
7369	SMITH	CLERK	7902	1980-12-17 00:00:00	800.00	(Null)	20
7499	ALLEN	SALESMAN	7698	1981-02-20 00:00:00	1600.00	300.00	30
7521	WARD	SALESMAN	7698	1981-02-22 00:00:00	1250.00	500.00	30
7566	JONES	MANAGER	7839	1981-04-02 00:00:00	2975.00	(Null)	20
7654	MARTIN	SALESMAN	7698	1981-09-28 00:00:00	1250.00	1400.00	30
7698	BLAKE	MANAGER	7839	1981-05-01 00:00:00	2850.00	(Null)	30
7782	CLARK	MANAGER	7839	1981-06-09 00:00:00	2450.00	(Null)	10
7788	SCOTT	ANALYST	7566	1987-04-19 00:00:00	3000.00	(Null)	20
7839	KING	PRESIDENT	(Null)	1981-11-17 00:00:00	5000.00	(Null)	10
7844	TURNER	SALESMAN	7698	1981-09-08 00:00:00	1500.00	0.00	30
7876	ADAMS	CLERK	7788	1987-05-23 00:00:00	1100.00	(Null)	20
7900	JAMES	CLERK	7698	1981-12-03 00:00:00	950.00	(Null)	30
7902	FORD	ANALYST	7566	1981-12-03 00:00:00	3000.00	(Null)	20
7934	MILLER	CLERK	7782	1982-01-23 00:00:00	1300.00	(Null)	10

图 23