Kettle对接FusionInsight

适用场景

> Kettle6.1 ↔ FusionInsight V100R002C60U10

环境准备

###Linux平台

- 安装操作系统
  - 安装CentOS6.5 Desktop
  - 禁用防火墙，SELinux
  - 添加本地主机名解析
  - 使用`vi /etc/hosts`添加本地主机名解析

    ```
    162.1.115.89 kettle
    ```

- 安装FusionInsight HD客户端
  - 下载完整客户端，安装至目录`/opt/hadoopclient`
  - 使用`vi /etc/profile`编辑以下内容插入到文件末尾

    ```
    source /opt/hadoopclient/bigdata_env
    ```

  - 将krb5.conf放在/etc目录下

    ```
    cp /opt/hadoopclient/KrbClient/kerberos/var/krb5kdc/krb5.conf /etc/
    ```

Windows平台

- 安装JDK8

```
C:\Users\Administrator>java -version
java version "1.8.0_112"
Java(TM) SE Runtime Environment (build 1.8.0_112-b15)
Java HotSpot(TM) 64-Bit Server VM (build 25.112-b15, mixed mode)
```

- 配置系统环境变量

```
JAVA_HOME= C:\\Program Files\\Java\\jdk1.8.0_112
```

- 在PATH环境变量添加`%JAVA_HOME%\bin;%JAVA_HOME%\jre\bin;`
- 获取Kerberos配置文件

  在FI管理界面下载用户的认证凭据

  解压后得到Kerberos配置文件krb5.conf和用户密钥文件user.keytab

- 将krb5.conf文件复制`C:\Windows`目录下，重命名为krb5.ini

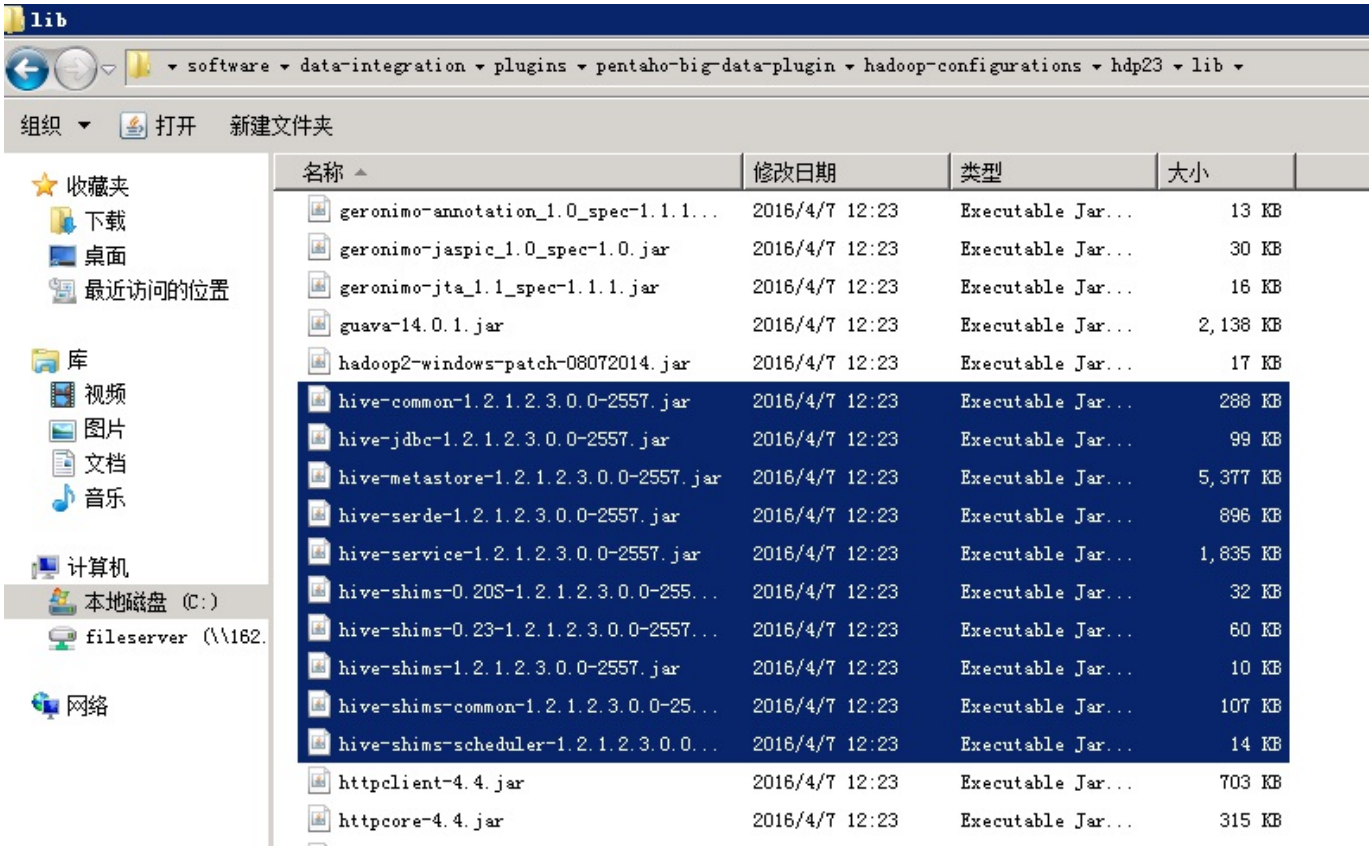
- 添加系统环境变量KRB5_CONFIG（可选步骤）

```
KRB5_CONFIG=C:\Windows
```

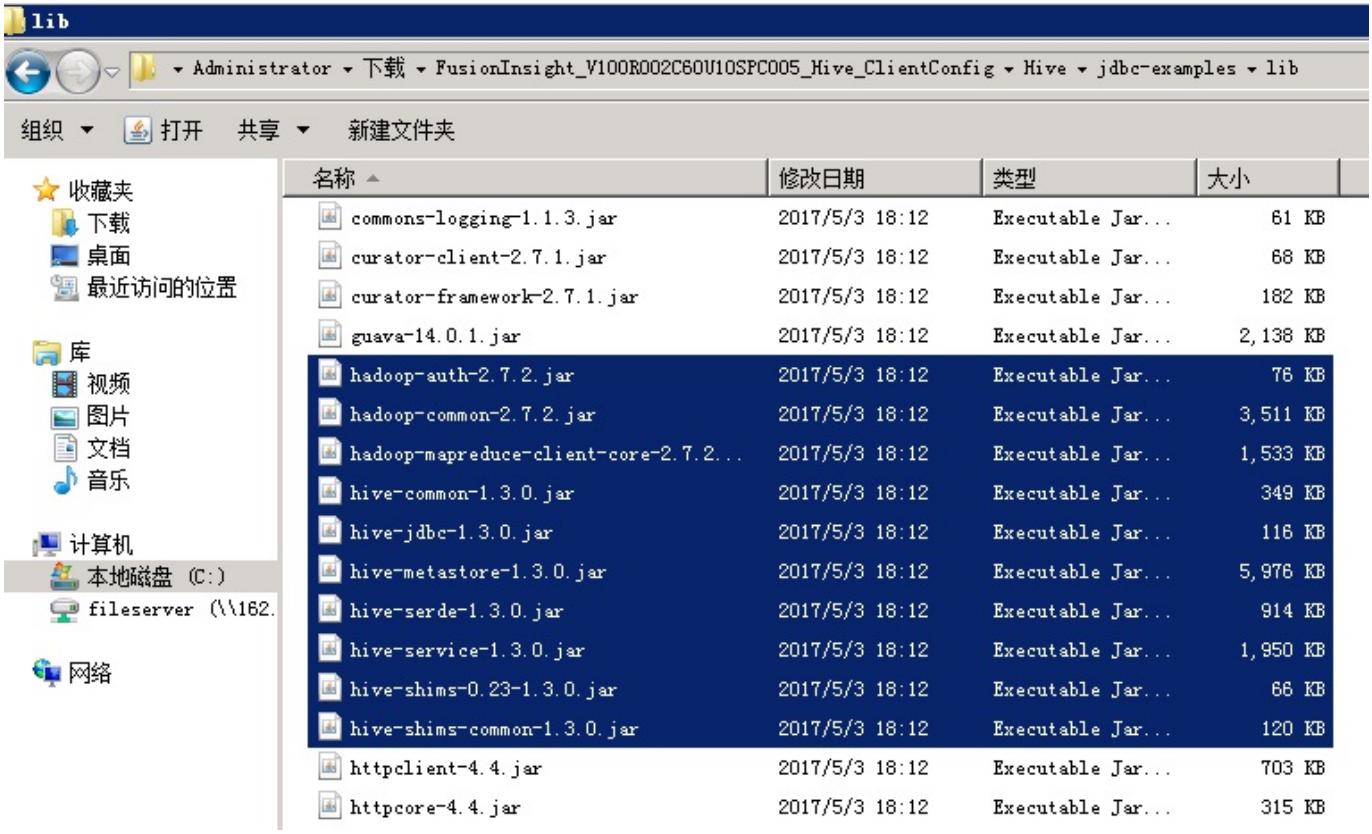
配置并启动Kettle

- 从以下地址 https://sourceforge.net/projects/pentaho/files/Data%20Integration/ 下载Kettle6.1版本
- 解压得到data-integration目录
- 替换pentaho-big-data-plugin下的配置文件
  - 下载FusionInsightHD客户端并解压
  - 用解压目录下 `Hive/jdbc-examples/conf/core-site.xml` 文件
  - 替换 `data-integration/plugins/pentaho-big-data-plugin/hadoop-configurations/hdp23` 目录下的core-site.xml文件
- 替换Hive相关jar包

  将 `data-integration/plugins/pentaho-big-data-plugin/hadoop-configurations/hdp23/lib` 下的hive相关的jar包



替换成Hive客户端下jdbc-examples/lib目录下的以下jar包



- 获取用户keytab文件
  - 在FI管理界面下载用户的keytab文件到本地
- Kerberos认证（可选步骤）

  在对接Hive时，可以使用本地缓存的认证票据，或者在jdbc URL中指定principal和keytab文件进行认证（对接HDFS时，只能使用本地缓存的票据）

  如果使用本地缓存的票据，需要在启动kettle前先完成认证。

```
C:\Users\Administrator>kinit -k -t C:\user.keytab test
New ticket is stored in cache file C:\temp\krb5cache

C:\Users\Administrator>klist

Credentials cache: C:\temp\krb5cache

Default principal: test@HADOOP.COM, 1 entry found.

[1]  Service Principal:  krbtgt/HADOOP.COM@HADOOP.COM
     Valid starting:     May 11, 2017 10:37:54
     Expires:            May 12, 2017 10:37:54
```

> 使用本地缓存票据存在以下问题：kettle只在启动时读取一次票据，而不是连接时实时读取当前票据信息，所以当kettle启动时获取的票据过期以后，连接Hive会失败，必须重启kettle。

- 启动kettle
  - Linux平台

    VNC登录CentOS桌面，打开Terminal

```
cd /opt/data-integration/
./spoon.sh
```

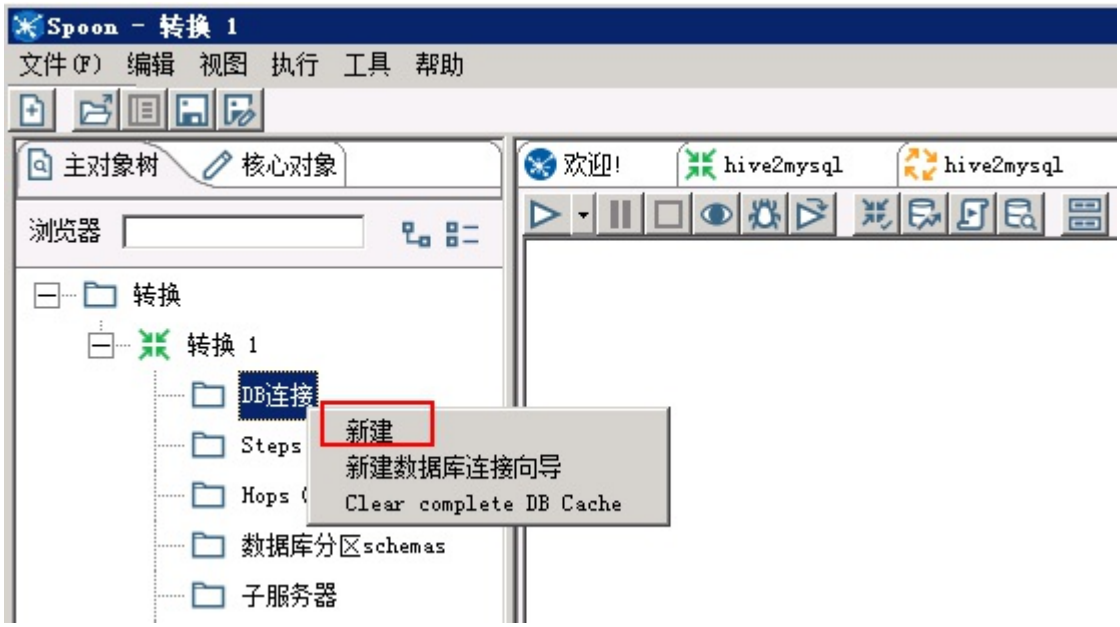
  - Windows平台

    双击data-integration目录下的Spoon.bat

对接Hive

创建Hive连接

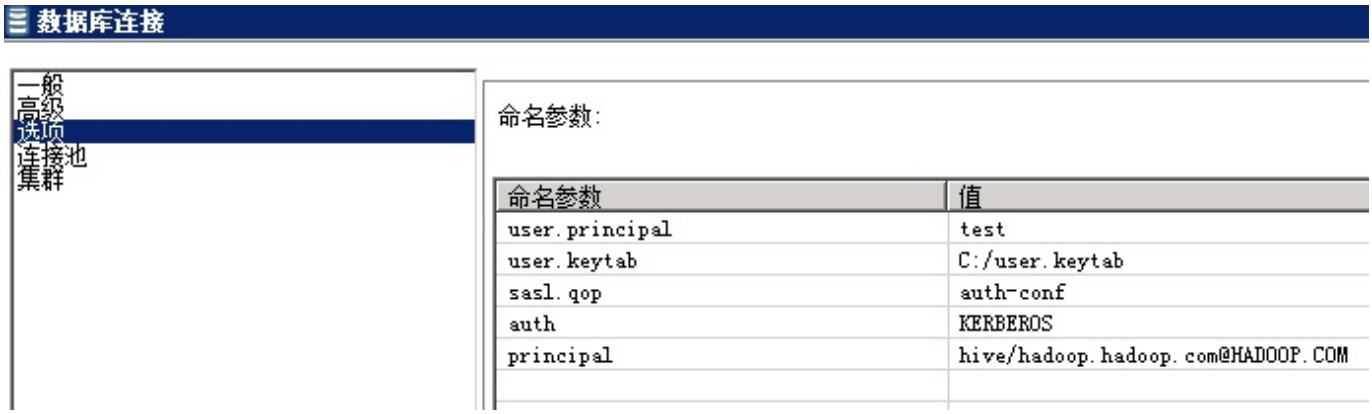
- 选择 **文件** -> **新建** -> **转换**
- 点击 **主对象树** 页签，在页签中选择 **转换** -> **DB连接**，右键选择 **新建**



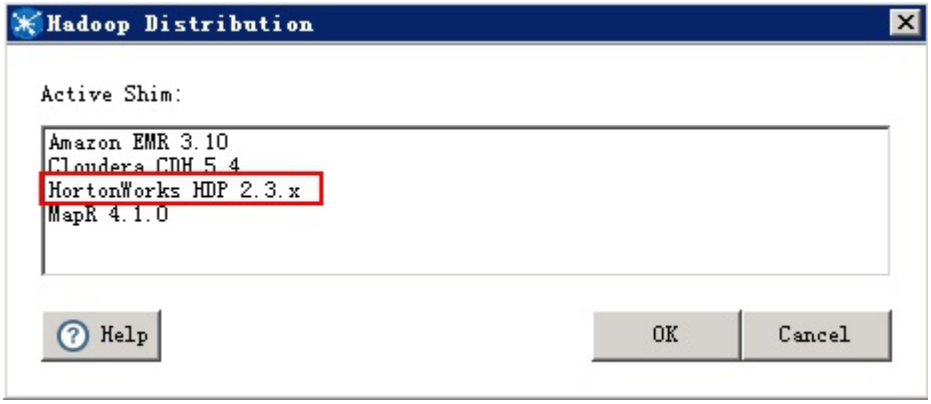
- 连接类型选择Hive 2，填写主机名、端口号、数据库名

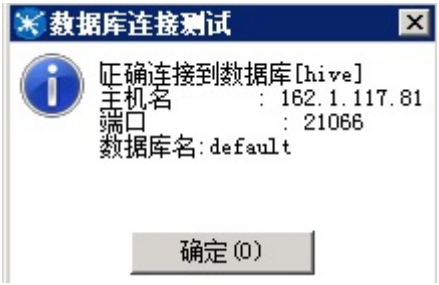
- 点击左侧 **选项**，如果使用本地缓存票据，填写以下参数：

- 如果要在连接Hive时使用keytab文件认证，增加user.principal和user.keytab两个参数：



- 测试连接时，Hadoop版本选用HDP2.3



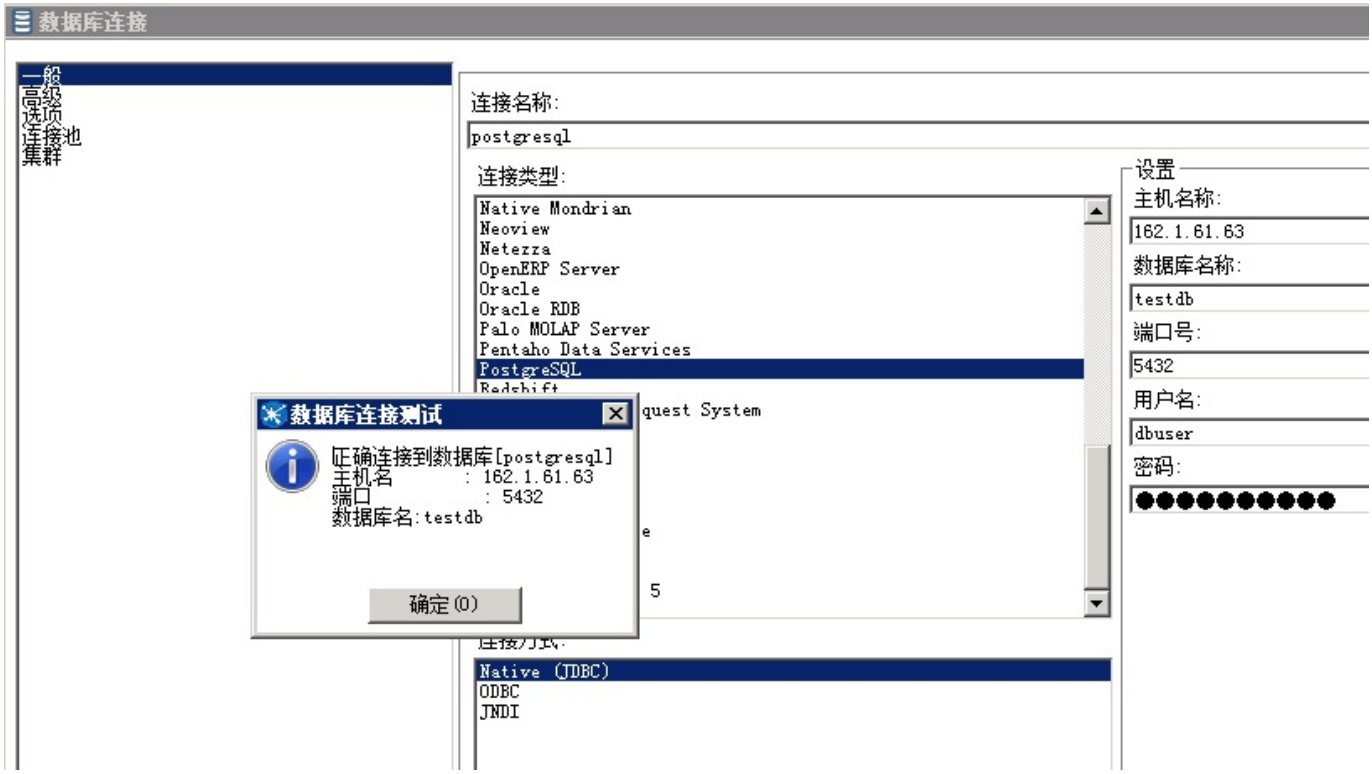


- 连接测试成功后，点击 **确认** 保存连接

读取Hive数据

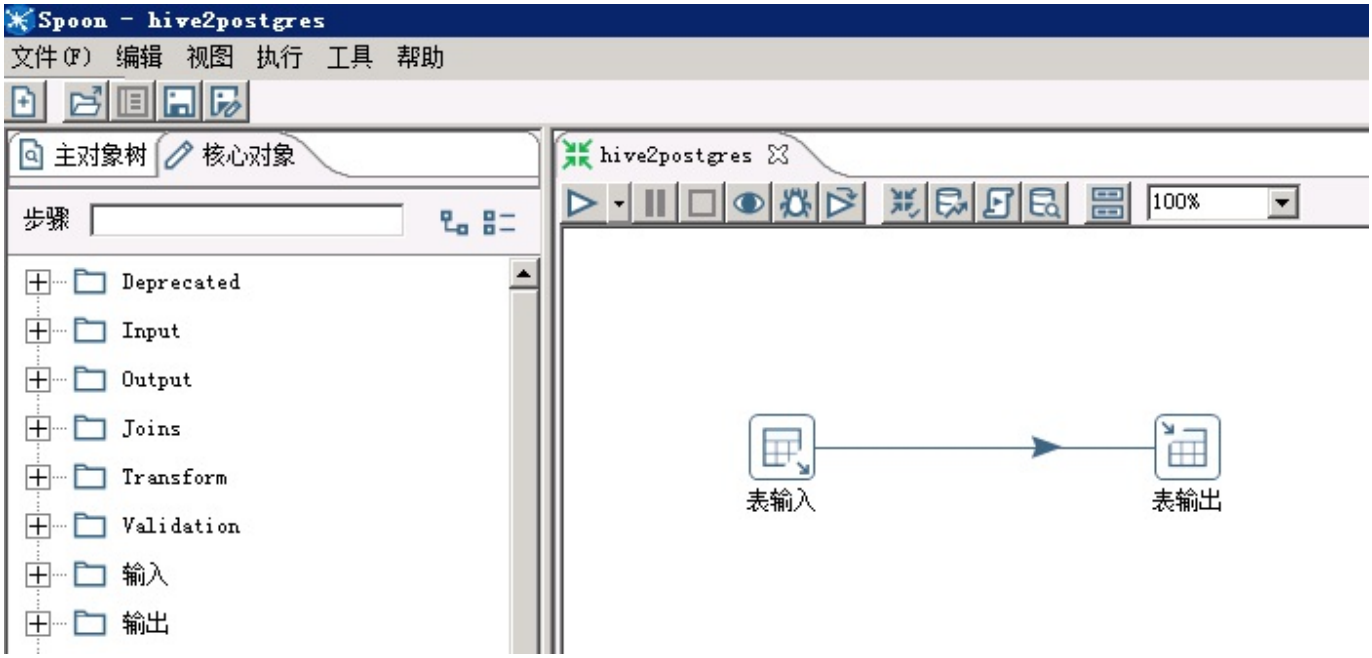
以hive -> postgresql为例

- 将上面创建的转换保存为hive2postgres.ktr
- 创建postgresql连接



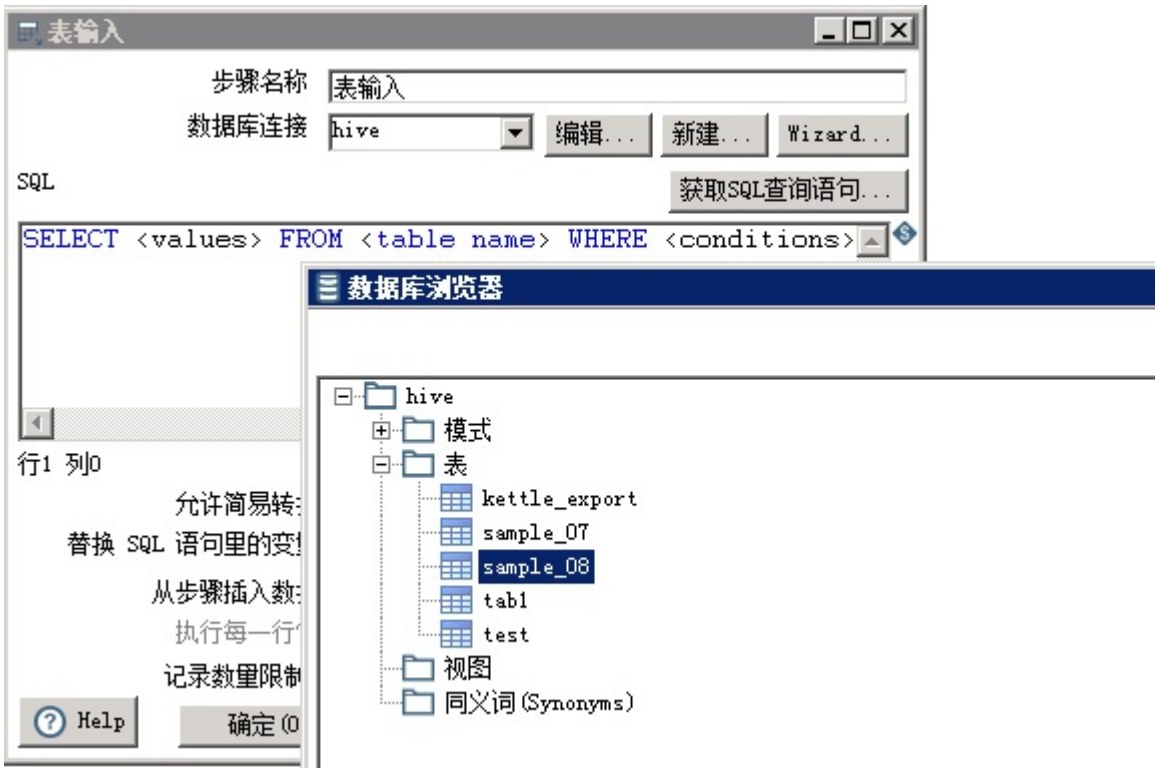
- 添加转换步骤

  在 **核心对象** 页签下，拖动 **输入 -> 表输入**，和 **输出 -> 表输出** 两个步骤到工作区，并连接这两个步骤。



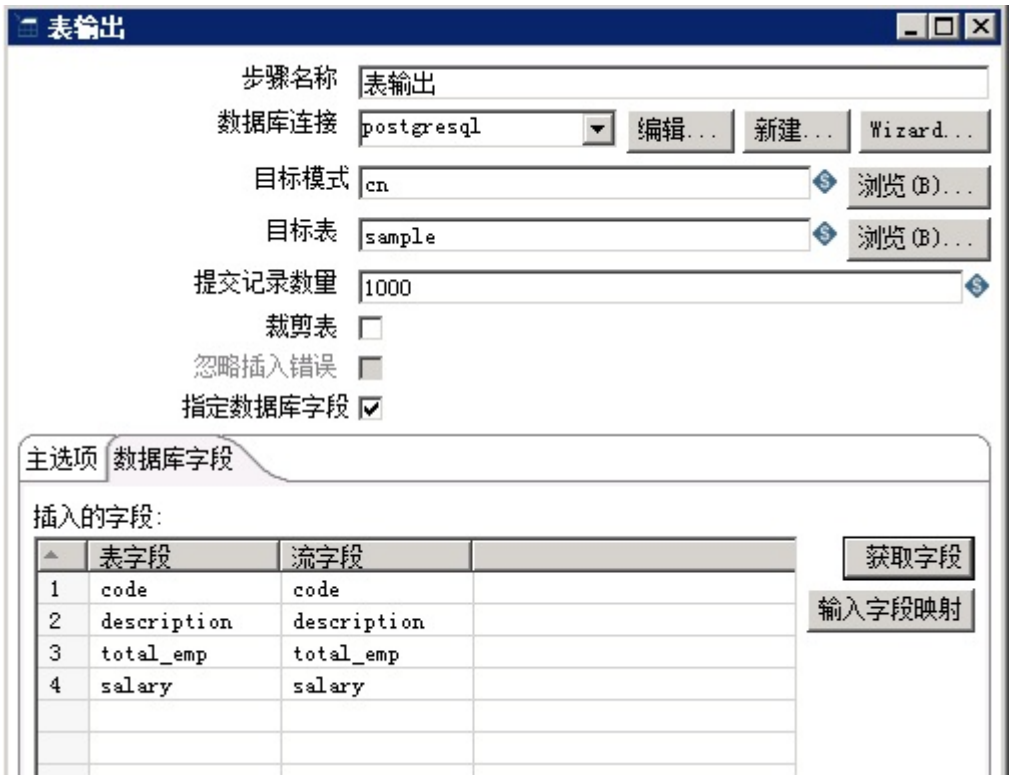
- 修改Hive表输入配置

  双击 **表输入** 步骤， **数据库连接** 选择前面创建的hive连接，点击 **获取SQL查询语句** ，选择需要导入的hive表

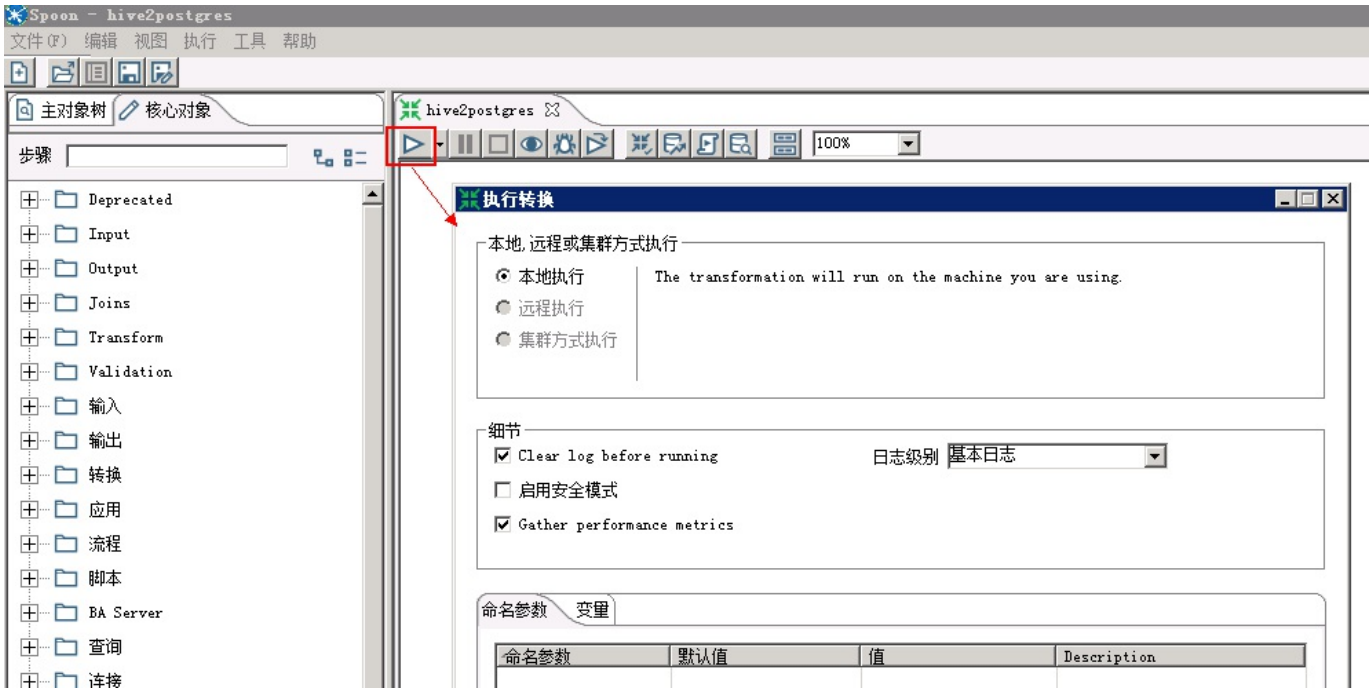


- 修改postgresql表输出配置

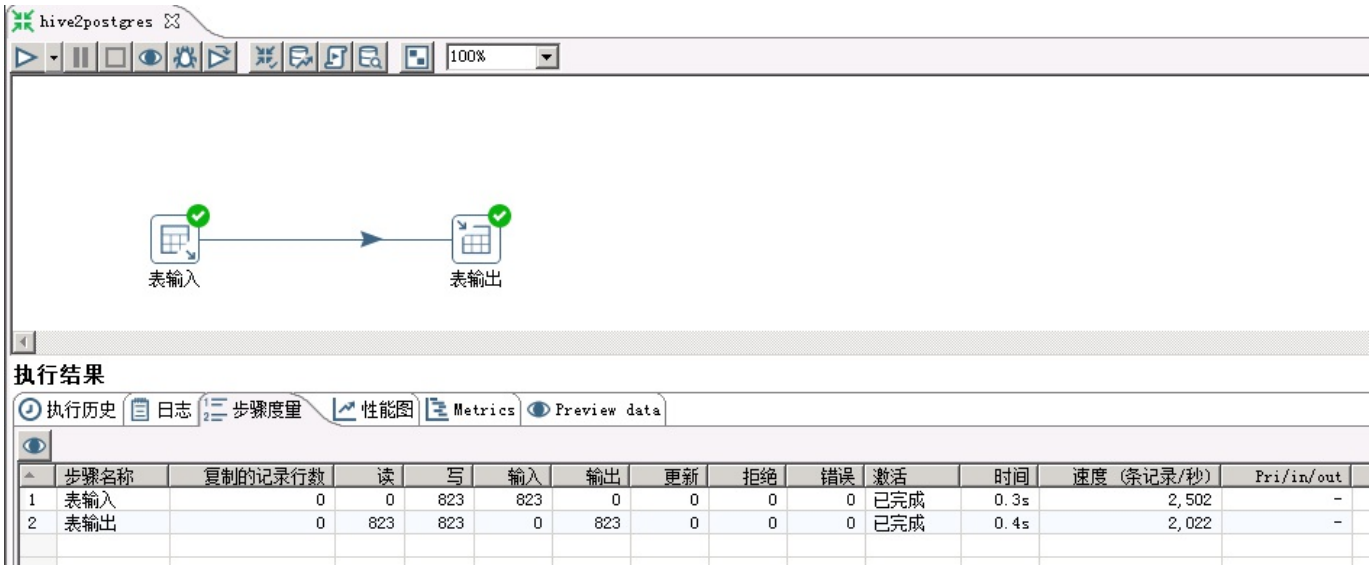
  双击 **表输出** 步骤，**数据库连接中** 选择前面创建的postgresql连接，点击 **获取目标表配置** 如下（需要先在postgresql数据库创建目标表）



- 运行转换
- 保存配置，点击 **执行** 按钮，选择 **本地执行**



执行结果：



	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/out
1	表输入	0	0	823	823	0	0	0	0	已完成	0.3s	2,502	-
2	表输出	0	823	823	0	823	0	0	0	已完成	0.4s	2,022	-

postgresql数据库查看：

```
testdb=> select count(*) from cn.sample;
 count
-------
   823
(1 row)

testdb=> select * from cn.sample limit 10;
  code   |             description             | total_emp | salary
---------+-------------------------------------+-----------+--------
 00-0000 | All Occupations                     | 135185230 |  42270
 11-0000 | Management occupations              |   6152650 | 100310
 11-1011 | Chief executives                    |    301930 | 160440
 11-1021 | General and operations managers     |   1697690 | 107970
 11-1031 | Legislators                         |     64650 |  37980
 11-2011 | Advertising and promotions managers |     36100 |  94720
 11-2021 | Marketing managers                  |    166790 | 118160
 11-2022 | Sales managers                      |    333910 | 110390
 11-2031 | Public relations managers           |     51730 | 101220
 11-3011 | Administrative services managers    |    246930 |  79500
(10 rows)
```

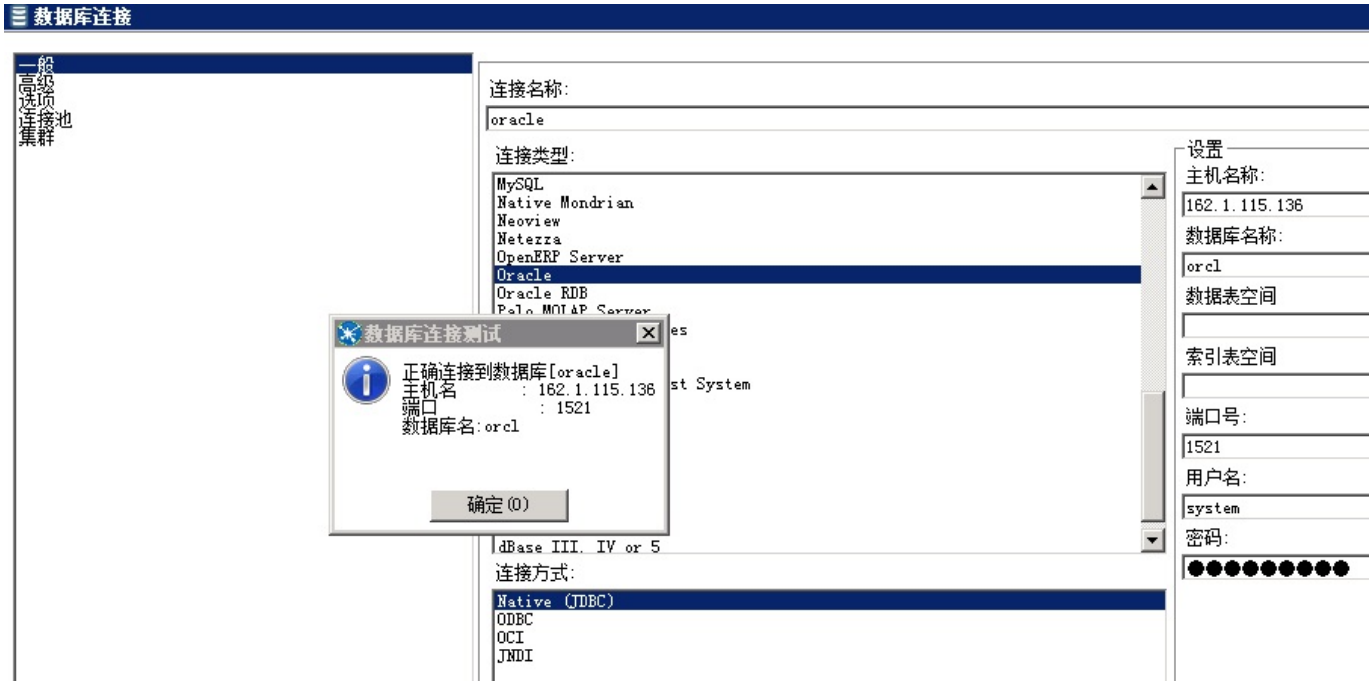
写入Hive数据

以oracle -> hive为例

- 添加Oracle JDBC Driver

  从http://www.oracle.com/technetwork/database/features/jdbc/index-091264.html 下载对应版本的jdbc Driver，放到 `data-integration/lib` 目录下，重启kettle

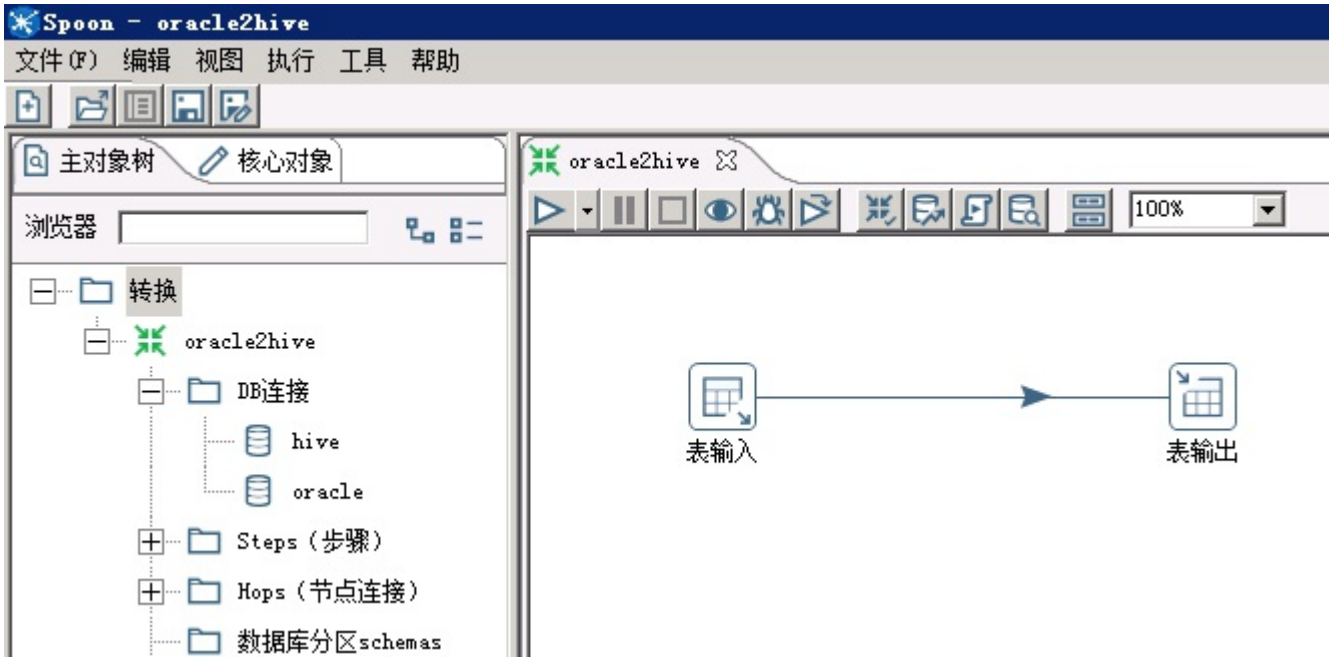
- 新建转换，保存为oracle2hive.ktr
- 创建Oracle连接



- 参考上面章节创建hive连接
- 创建待导入的Hive表

```
CREATE TABLE IF NOT EXISTS kettle_export (
  id int,
  name string
);
```

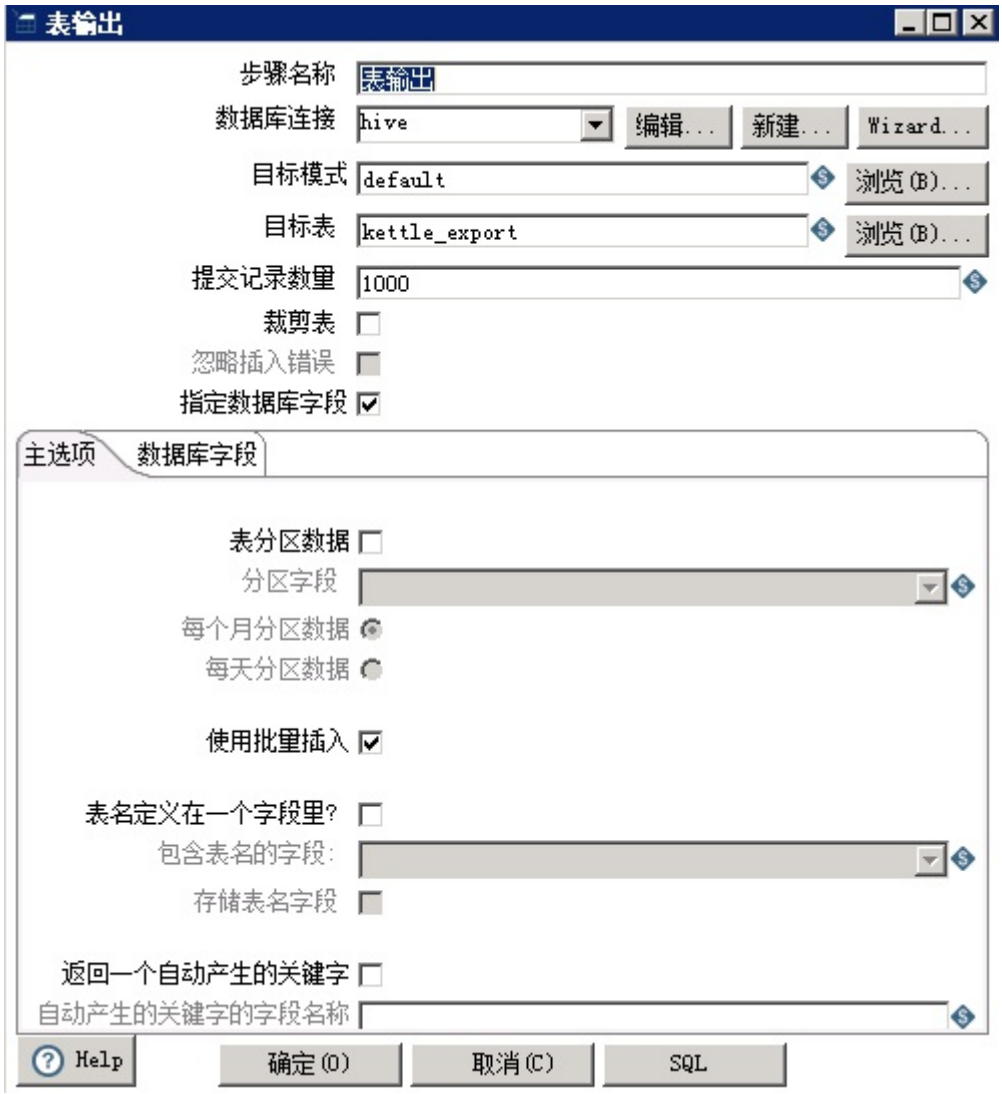
- 添加转换步骤



- 修改步骤配置
- Oracle表输入配置


表输入
步骤名称
表输入
数据库连接
oracle
编辑...
新建...
Wizard...
SQL
获取SQL查询语句...
SELECT
ID
, NAME
FROM TEST.TEST1
行1 列0
允许简易转换
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
Help
确定(O)
预览(P)
取消(C)

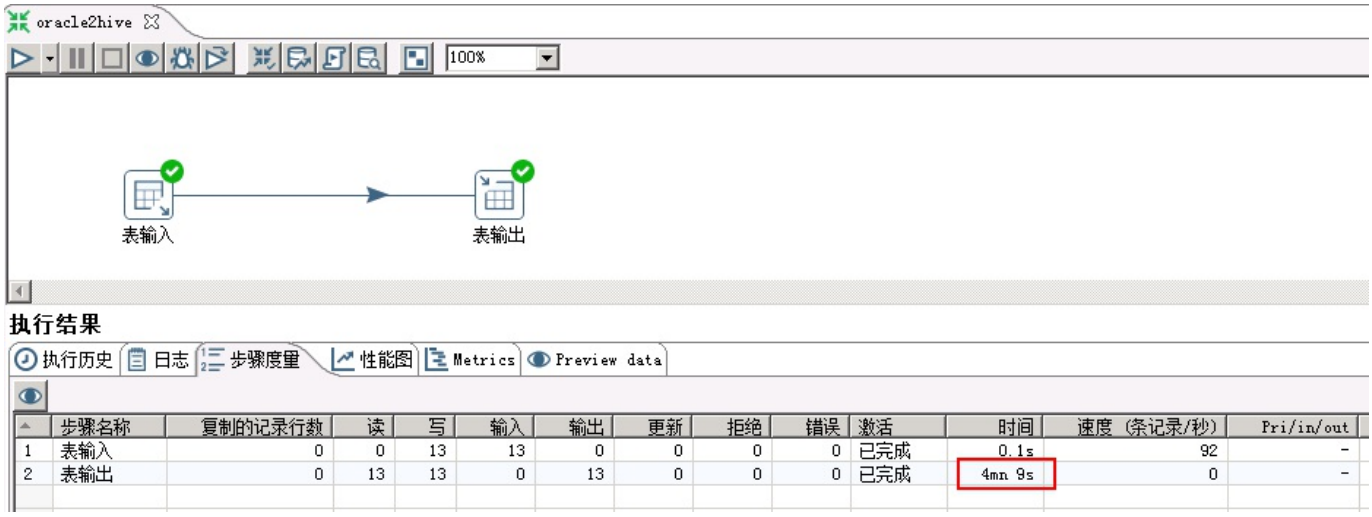
- Hive表输出配置


表输出
步骤名称
表输出
数据库连接
hive
编辑...
新建...
Wizard...
目标模式
default
浏览(B)...
目标表
kettle_export
浏览(B)...
提交记录数量
1000
裁剪表
忽略插入错误
指定数据库字段
主选项
数据库字段
表分区数据
分区字段
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:
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
Help
确定(O)
取消(C)
SQL

- 运行转换

  保存配置，点击 **执行** 按钮，选择 **本地执行**

  执行结果：向Hive表写入13条数据，用时4min+

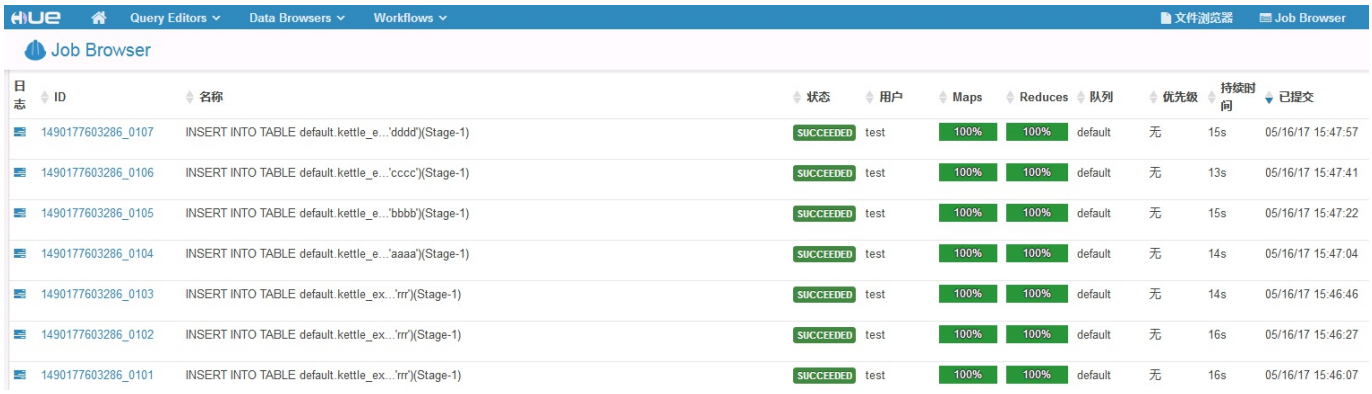


执行结果

	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/out
1	表输入	0	0	13	13	0	0	0	0	已完成	0.1s	92	-
2	表输出	0	13	13	0	13	0	0	0	已完成	4mn 9s	0	-

查看Hive表数据：

```
0: jdbc:hive2://162.1.117.82:21066/> select * from kettle_export;
+-------------------+---------------------+--+
| kettle_export.id  | kettle_export.name  |
+-------------------+---------------------+--+
| 3                 | jjj                 |
| 4                 | jjj                 |
| 5                 | jjj                 |
| 6                 | jjj                 |
| 1                 | nnn                 |
| 2                 | nnn                 |
| 9                 | rrr                 |
| 10                | rrr                 |
| 11                | rrr                 |
| 20                | aaaa                |
| 21                | bbbb                |
| 22                | cccc                |
| 23                | dddd                |
+-------------------+---------------------+--+
13 rows selected (0.193 seconds)
```



> 说明：向Hive表中写入数据，每插入一条数据会起一个MR任务，所以效率特别低，不推荐用这种方式，可以将数据写入HDFS文件

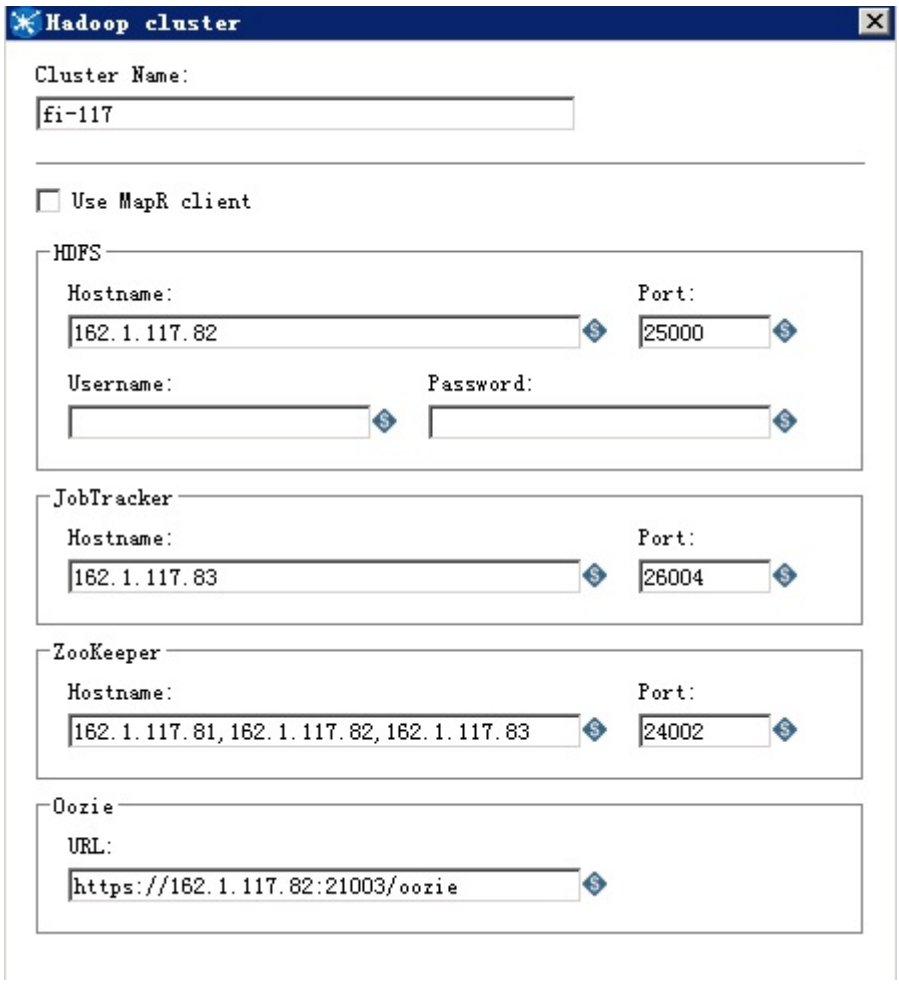
对接HDFS

创建Hadoop Cluster

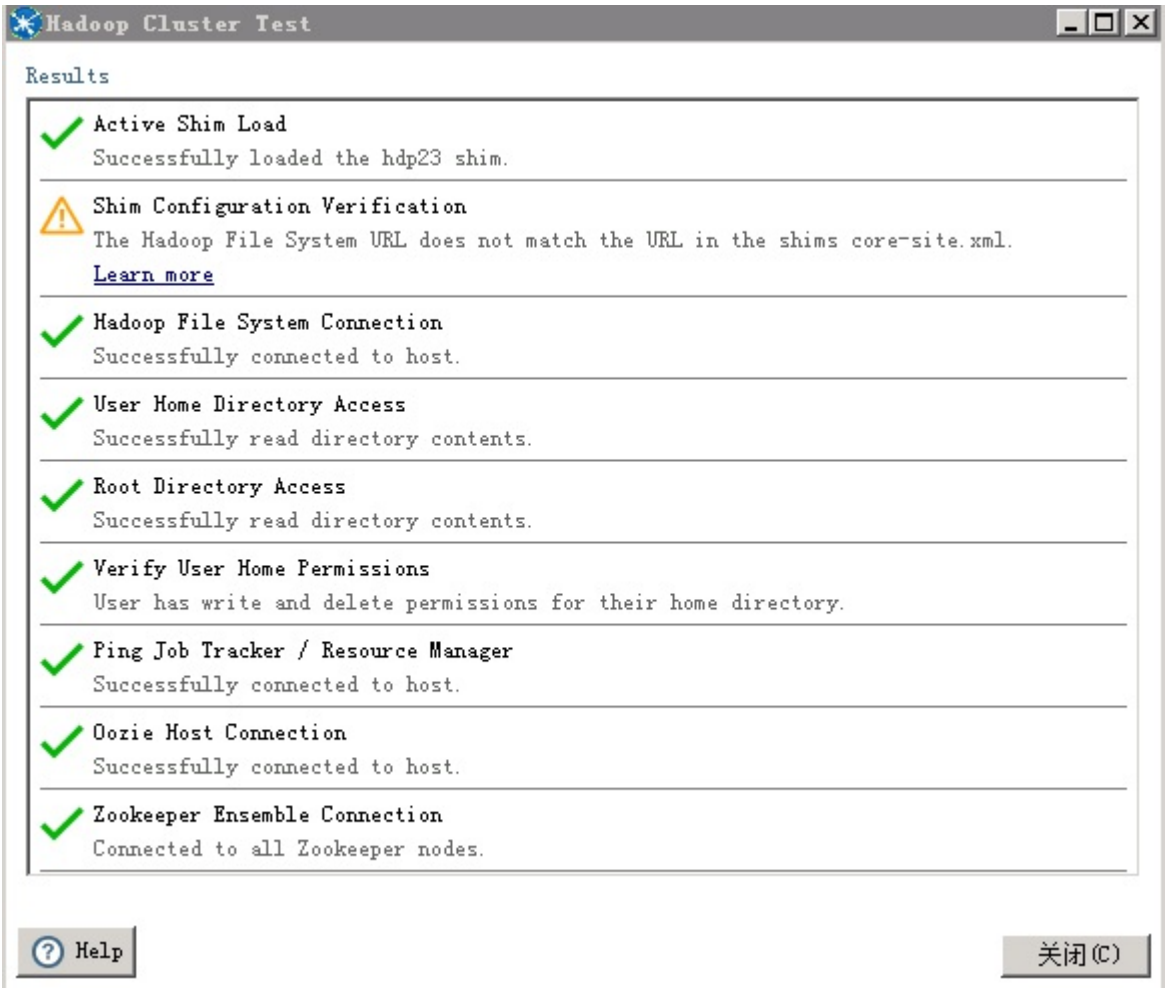
- 选择 **文件** -> **新建** -> **转换**，点击 **主对象树** 页签，在 **Hadoop Clusters** 右键选择 **New Cluster**

  HDFS的Hostname填写NameNode主节点的IP，端口号是25000，如果NaneNode发生主备切换，需要修改IP

  JobTracker的Hostname 填写 Yarn ResourceManager主节点的IP，端口号是26004，如果ResourceManager发生主备切换，需要修改IP。



点击 **测试**



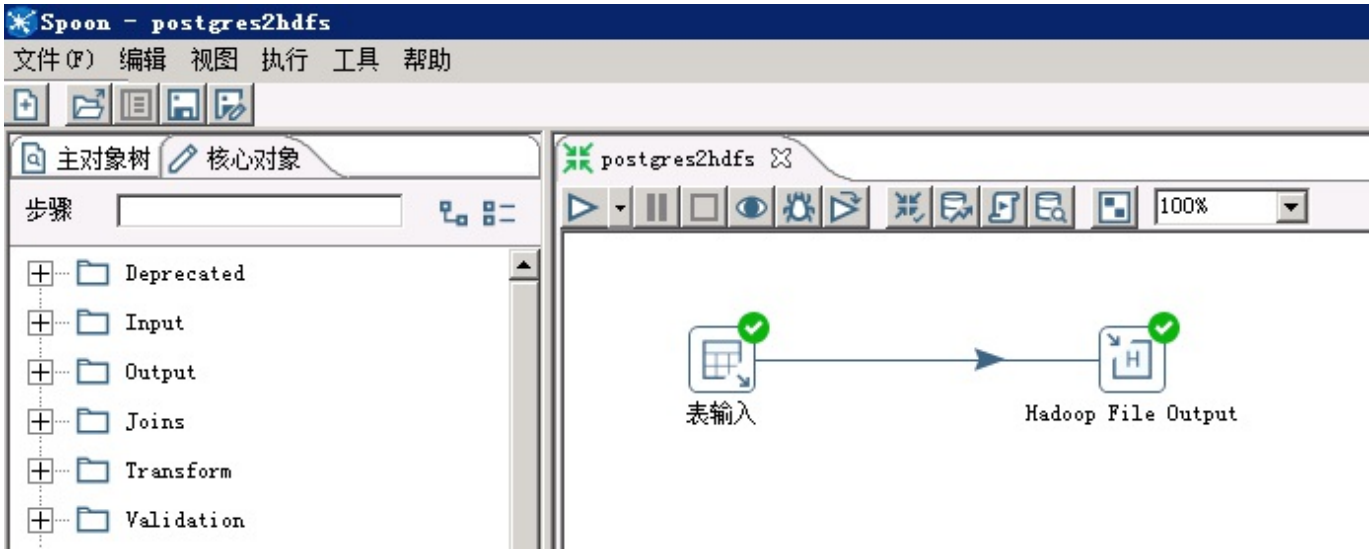
> kettle6.1不支持HDFS NameNode和Yarn ResourceManager的HA配置

导入HDFS文件

以postgresql -> HDFS为例

- 将上面创建的转换保存为postgres2hdfs.ktr
- 参考前面章节创建postgresql连接
- 添加转换步骤

  在 核心对象 页签下，拖动 **输入** -> **表输入** ，和 **Big Data** -> **Hadoop File Output** 两个步骤到工作区，并连接这两个步骤。



- 创建待导入的Hive表

```
CREATE TABLE IF NOT EXISTS sample_kettle_hdfs_test (
code string,
description string,
```

```
  total_emp int,
  salary int
)
ROW FORMAT SERDE
'org.apache.hadoop.hive.contrib.serde2.MultiDelimitSerDe' WITH
SERDEPROPERTIES ("field.delim"="[,]")
STORED AS TEXTFILE;
```

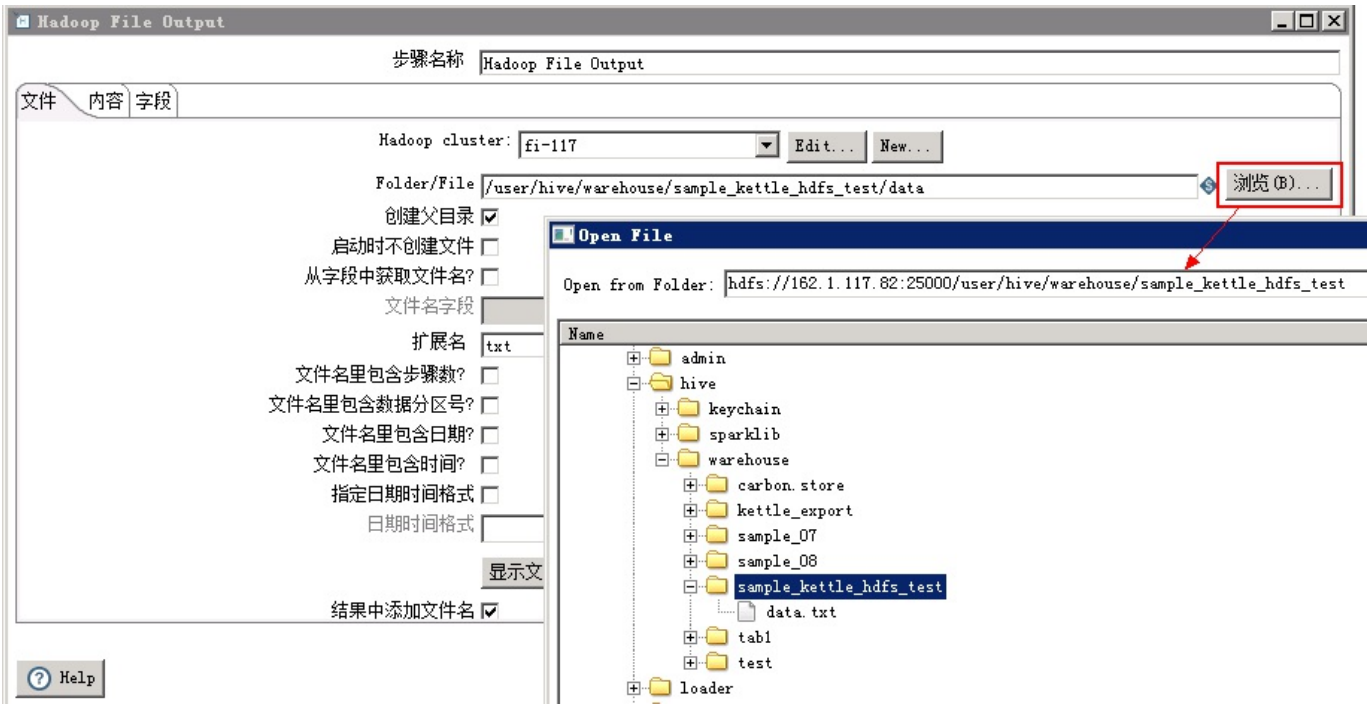
> 如果数据中含有","，列分隔符不可以使用默认的","，本样例使用多字节分隔符"[,]"

- 修改postgresql表输入配置

  双击 **表输入** 步骤，**数据库连接** 选择前面创建的postgresql连接，点击 **获取SQL查询语句**，选择需要导入的表

- 修改Hadoop File Output配置

  双击 **Hadoop File Output** 步骤，在 **文件** 页签下，**Hadoop Cluster** 选择前面创建的集群，**Folder/File** 选择到hive表对应的hdfs目录，文件名可以任意指定



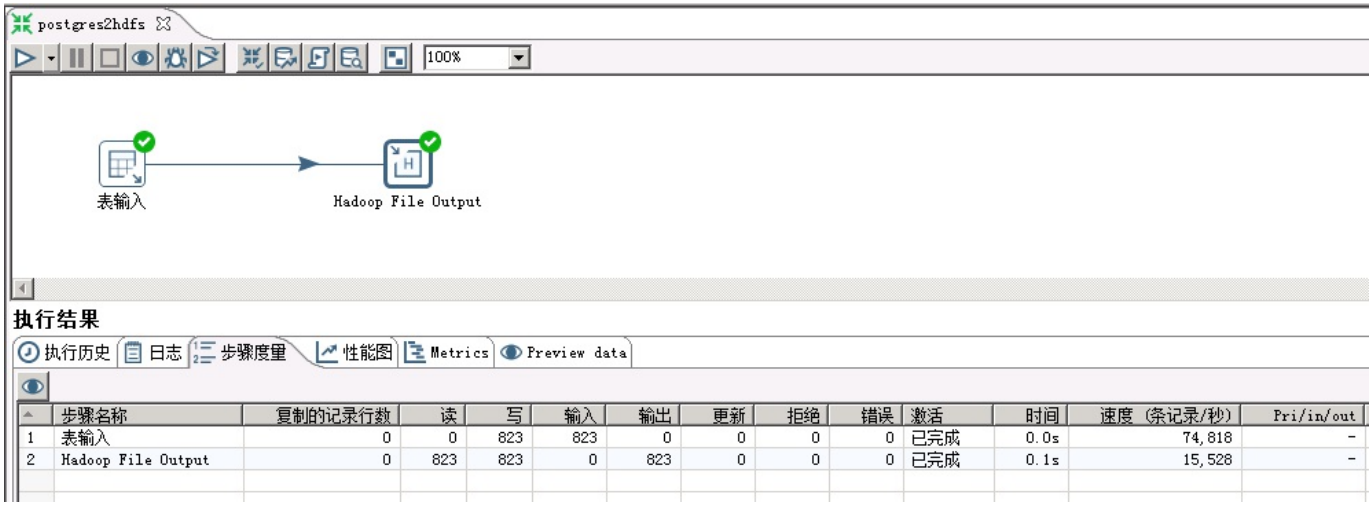
- 点击 **内容** 页签，分隔符设置与前面创建的Hive表相同，勾选 **快速数据存储（无格式）**（否则保存的文件中会按字段长度填充空格）

- 点击 **字段** 页签，获取字段

- 运行转换

  保存配置，点击 **执行** 按钮，选择 **本地执行** 。

- 执行结果：



	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/out
1	表输入	0	0	823	823	0	0	0	0	已完成	0.0s	74,818	-
2	Hadoop File Output	0	823	823	0	823	0	0	0	已完成	0.1s	15,528	-

- 查看导入的HDFS文件：

```
[root@kettle ~]# hdfs dfs -cat /user/hive/warehouse/sample_kettle_hdfs_test/data.txt
00-0000[,]All Occupations[,]135185230[,]42270
11-0000[,]Management occupations[,]6152650[,]100310
11-1011[,]Chief executives[,]301930[,]160440
11-1021[,]General and operations managers[,]1697690[,]107970
11-1031[,]Legislators[,]64650[,]37980
11-2011[,]Advertising and promotions managers[,]36100[,]94720
11-2021[,]Marketing managers[,]166790[,]118160
11-2022[,]Sales managers[,]333910[,]110390
11-2031[,]Public relations managers[,]51730[,]101220
11-3011[,]Administrative services managers[,]246930[,]79500
11-3021[,]Computer and information systems managers[,]276820[,]118710
11-3031[,]Financial managers[,]500590[,]110640
11-3041[,]Compensation and benefits managers[,]38810[,]93410
11-3042[,]Training and development managers[,]29350[,]93830
11-3049[,]Human resources managers, all other[,]60980[,]103920
11-3051[,]Industrial production managers[,]154030[,]91200
11-3061[,]Purchasing managers[,]67150[,]94300
11-3071[,]Transportation, storage, and distribution managers[,]96300[,]84520
11-9011[,]Farm, ranch, and other agricultural managers[,]3410[,]62400
11-9012[,]Farmers and ranchers[,]490[,]49140
11-9021[,]Construction managers[,]220550[,]89770
```

- 查看Hive表数据：

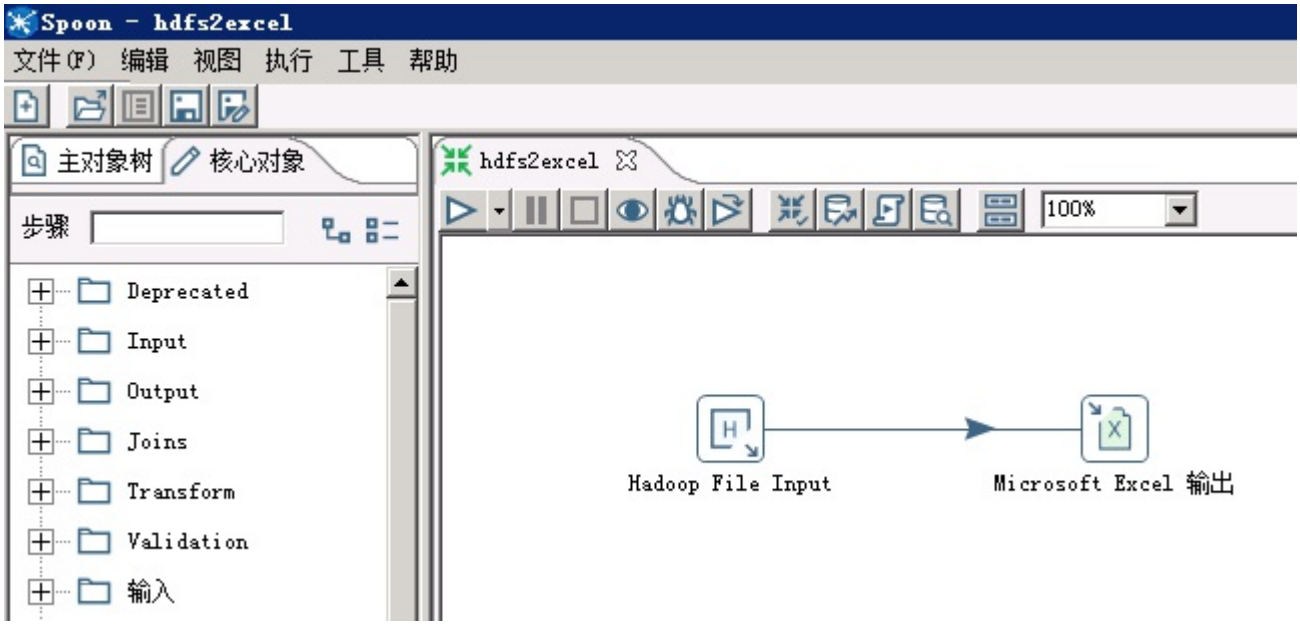
```
0: jdbc:hive2://162.1.117.82:21066/>  select count(*) from sample_kettle_hdfs_test;
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1490177603286_0094
INFO  : Kind: HDFS_DELEGATION_TOKEN, Service: ha-hdfs:hacluster, Ident: (HDFS_DELEGATION_TOKEN token 143959 for test)
INFO  : Kind: HIVE_DELEGATION_TOKEN, Service: HiveServer2ImpersonationToken, Ident: 00 04 74 65 73 74 04 74 65 73 74 21 68 69 76 65 2f 68 61 64 6f 6f 70 2e 68 61 64 6f 6f 7
 43 8a 01 5c 2f 20 54 43 8d 04 ab 0a 78
INFO  : The url to track the job: https://162-1-117-83:26001/proxy/application_1490177603286_0094/
INFO  : Starting Job = job_1490177603286_0094, Tracking URL = https://162-1-117-83:26001/proxy/application_1490177603286_0094/
INFO  : Kill Command = /opt/huawei/Bigdata/FusionInsight_V100R002C60U10/FusionInsight-Hive-1.3.0/hive-1.3.0/bin/..//../hadoop/bin/hadoop job  -kill job_1490177603286_0094
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2017-05-15 18:39:13,030 Stage-1 map = 0%,  reduce = 0%
INFO  : 2017-05-15 18:39:20,644 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.05 sec
INFO  : 2017-05-15 18:39:37,660 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.34 sec
INFO  : MapReduce Total cumulative CPU time: 5 seconds 340 msec
INFO  : Ended Job = job_1490177603286_0094
+------+--+
| _c0  |
+------+--+
| 823  |
+------+--+
1 row selected (38.323 seconds)
0: jdbc:hive2://162.1.117.82:21066/> select * from sample_kettle_hdfs_test limit 10;
+-----------------------------+-----------------------------------+-----------------------------+-----------------------------+--+
| sample_kettle_hdfs_test.code | sample_kettle_hdfs_test.description | sample_kettle_hdfs_test.total_emp | sample_kettle_hdfs_test.salary |
+-----------------------------+-----------------------------------+-----------------------------+-----------------------------+--+
| 00-0000                     | All Occupations                   | 135185230                   | 42270                       |
| 11-0000                     | Management occupations            | 6152650                     | 100310                      |
| 11-1011                     | Chief executives                  | 301930                      | 160440                      |
| 11-1021                     | General and operations managers   | 1697690                     | 107970                      |
| 11-1031                     | Legislators                       | 64650                       | 37980                       |
| 11-2011                     | Advertising and promotions managers | 36100                     | 94720                       |
| 11-2021                     | Marketing managers                | 166790                      | 118160                      |
| 11-2022                     | Sales managers                    | 333910                      | 110390                      |
| 11-2031                     | Public relations managers         | 51730                       | 101220                      |
| 11-3011                     | Administrative services managers  | 246930                      | 79500                       |
+-----------------------------+-----------------------------------+-----------------------------+-----------------------------+--+
```

读取HDFS文件

以HDFS -> Excel为例

- 新建转换，保存为hdfs2excel.ktr
- 添加转换步骤

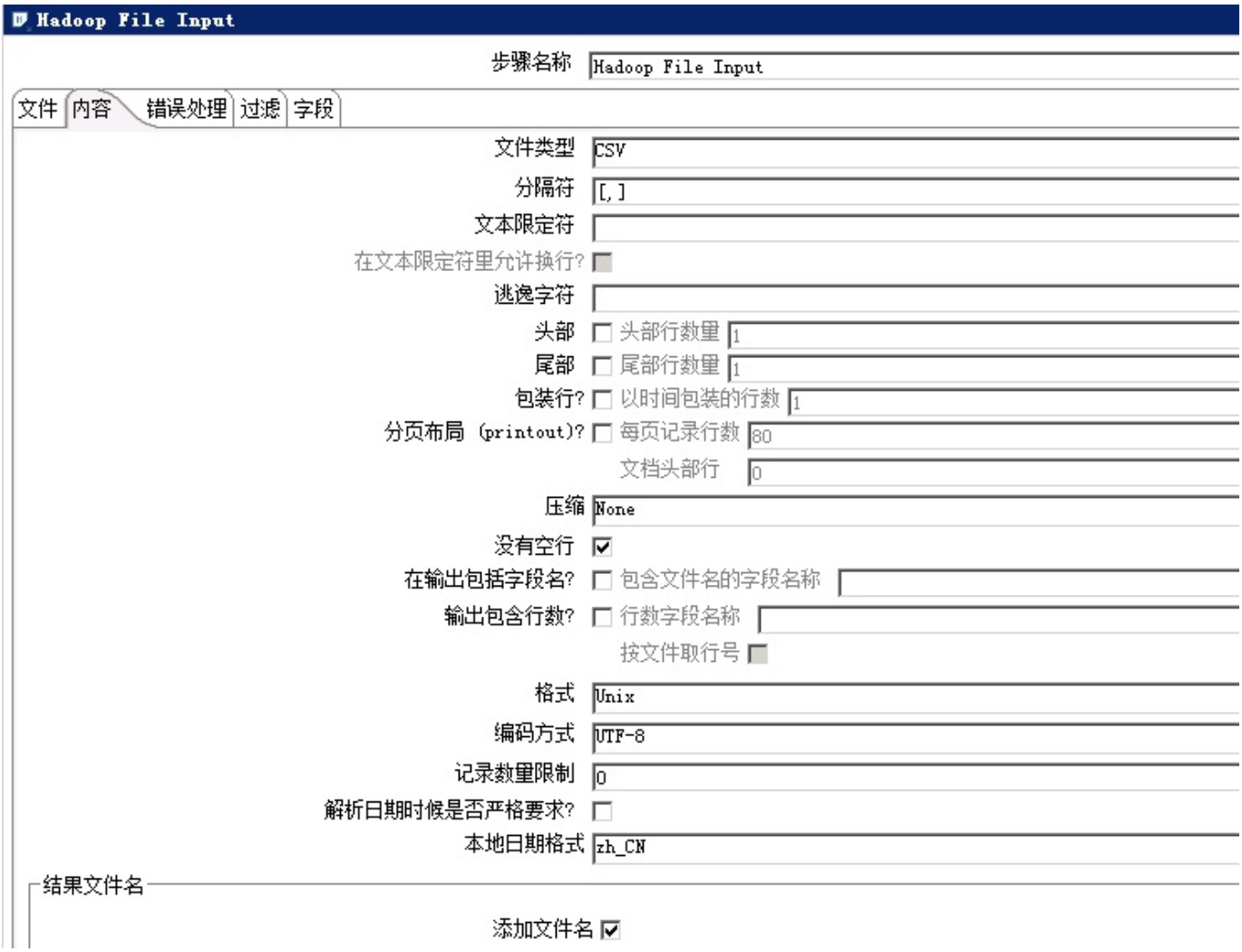
  在 **核心对象** 页签下，拖动 **Big Data** -> **Hadoop File Input** 和 **输出** -> **Microsoft Excel 输出**，两个步骤到工作区，并连接这两个步骤。



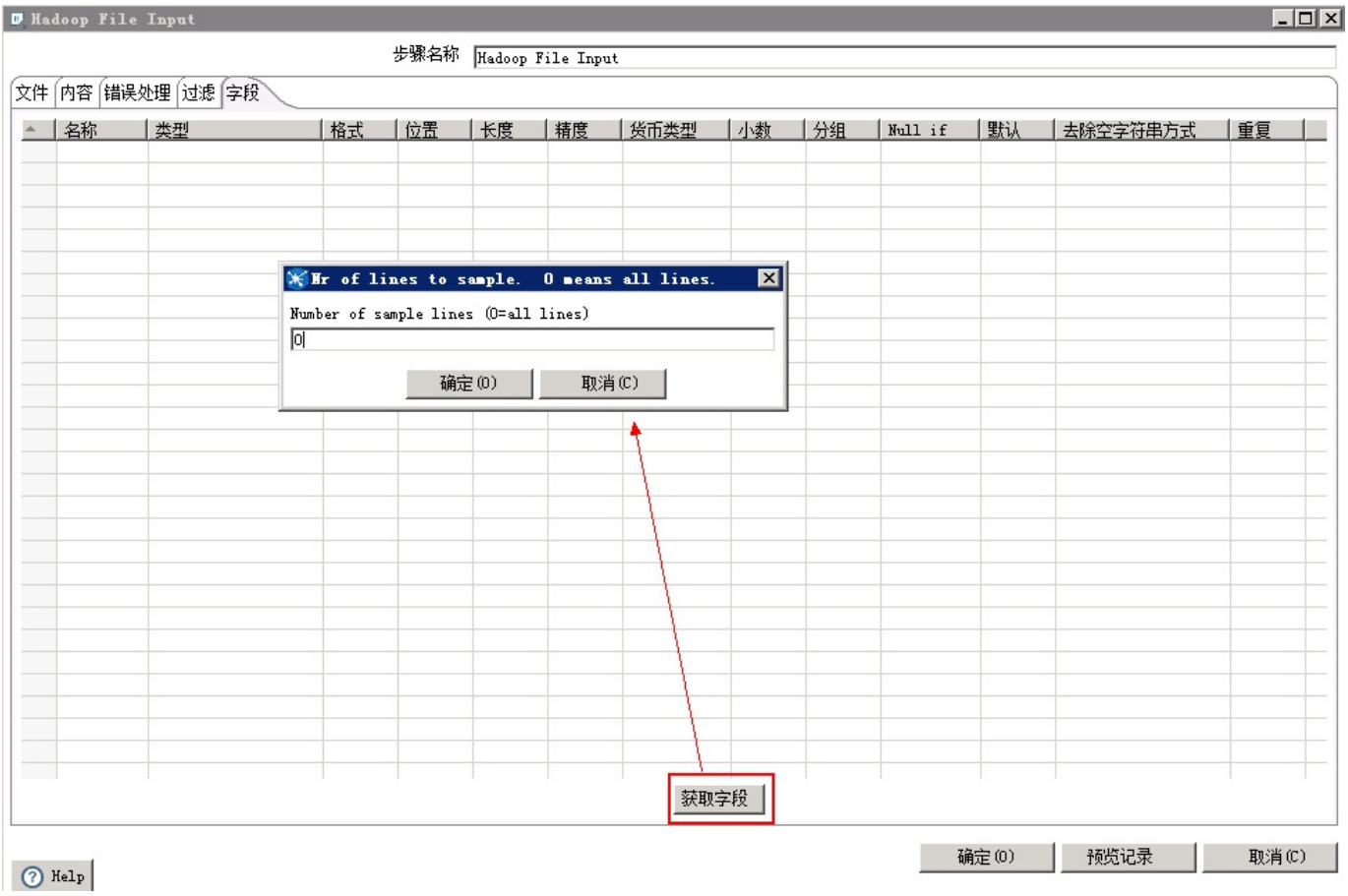
- 修改 Hadoop File Input配置

  双击 **Hadoop File Input** 步骤，**文件** 页签，选择待导出的文件，文件类型支持CSV（txt也可以）和Fixed（固定列宽）

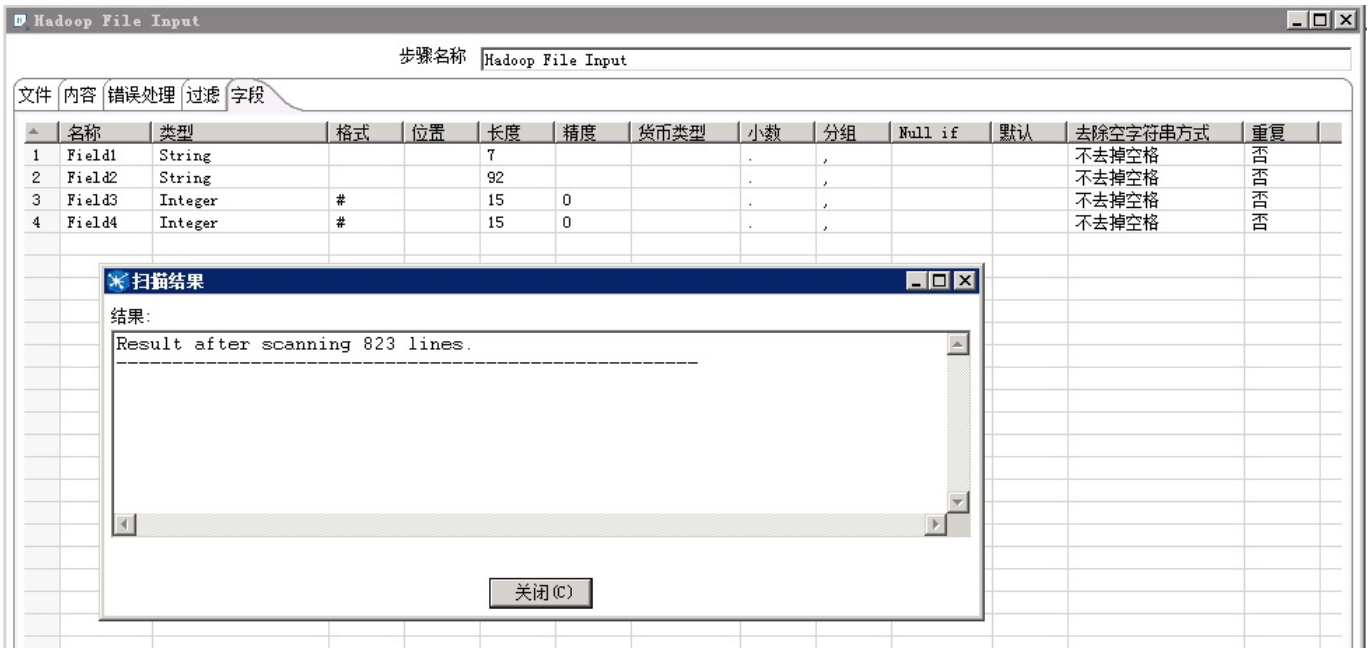
  点击 **内容** 页签，选择文件类型、分隔符、编码方式等



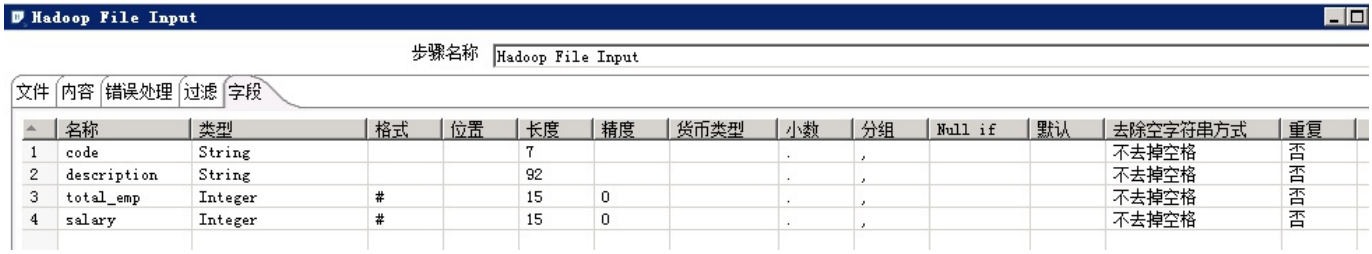
点击 **字段** 页签，获取字段



kettle会自动扫描文件中的字段类型和长度



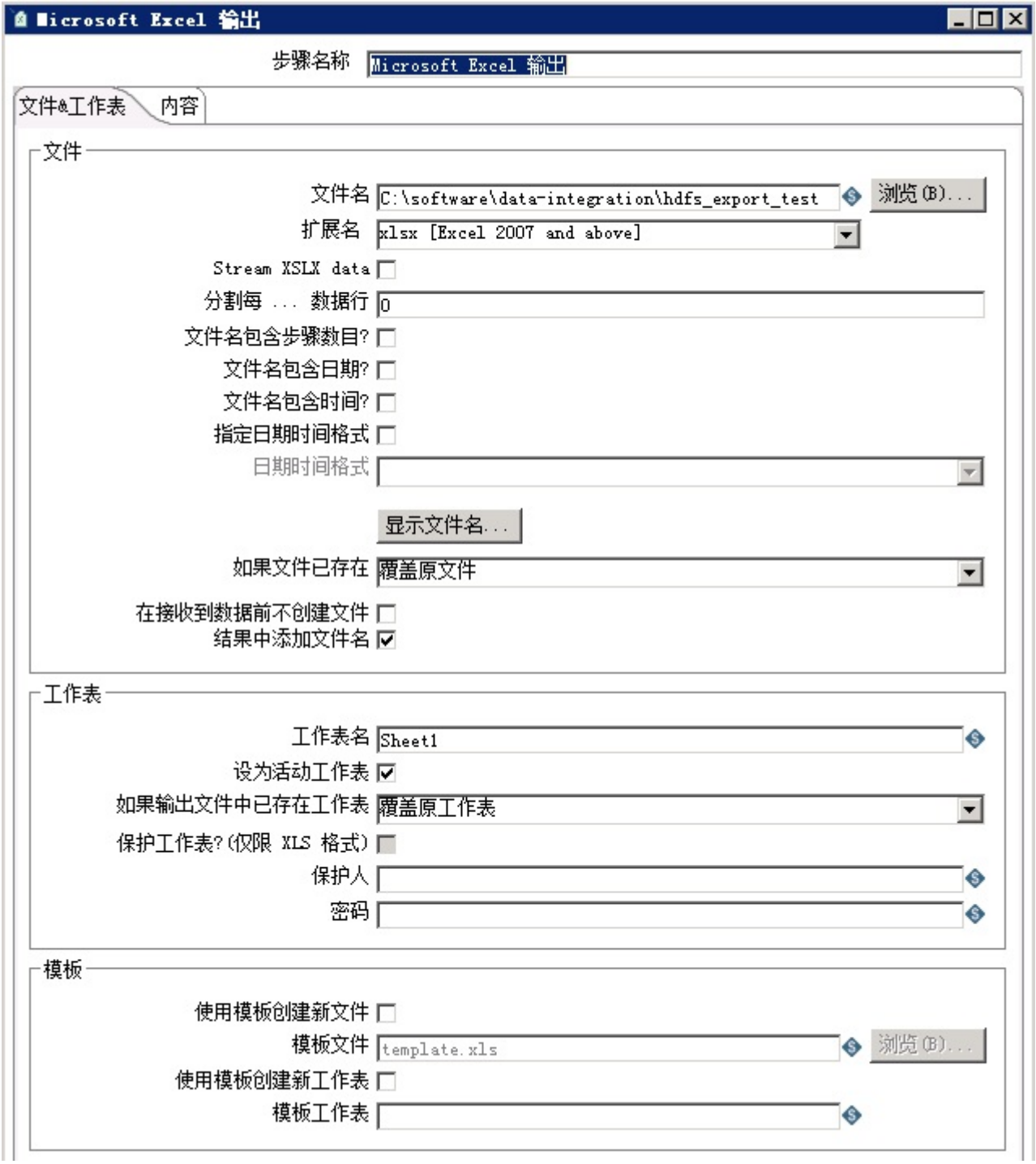
可以手动修改字段名称和长度



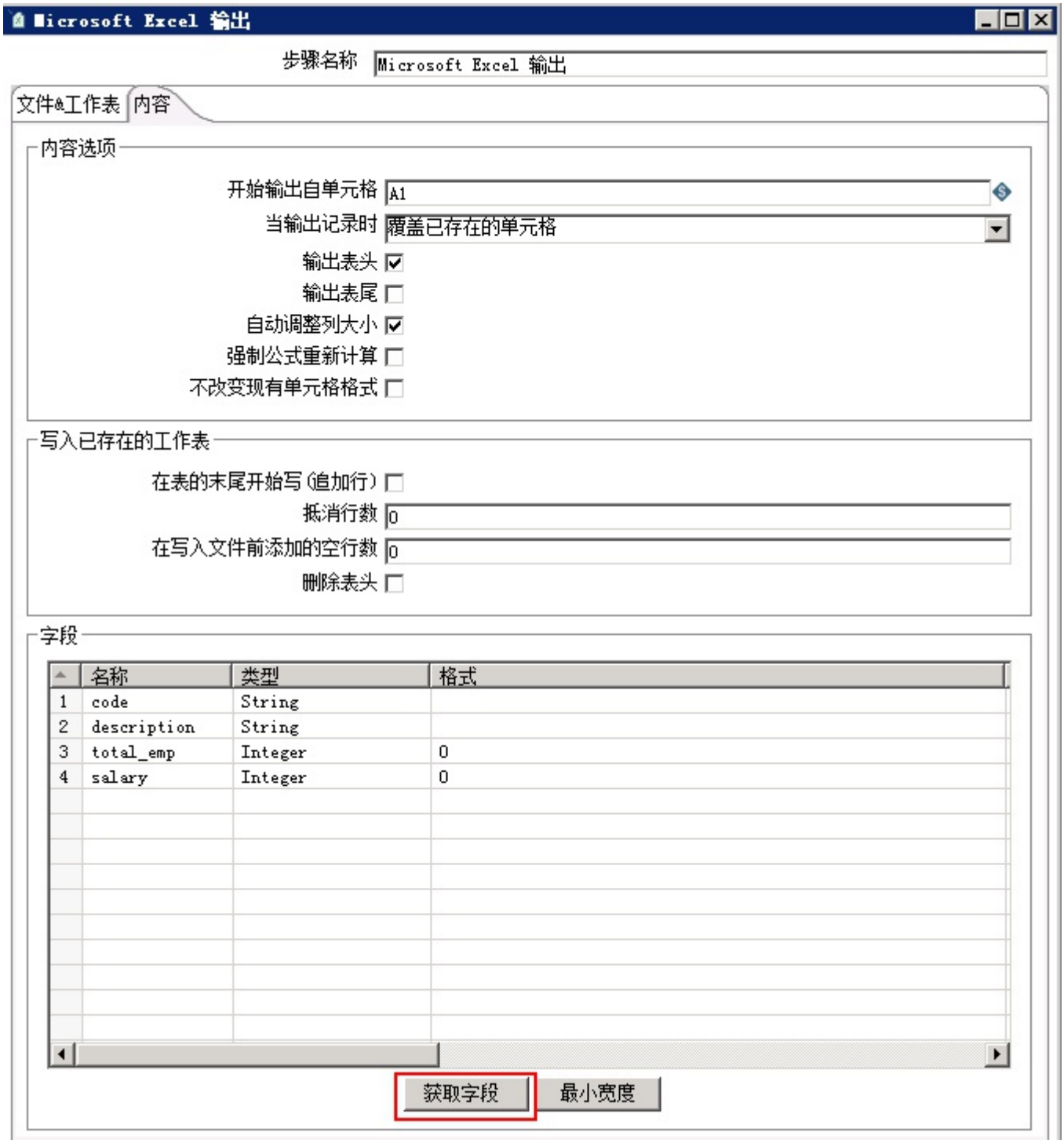
点击 **确定** 按钮，保存配置

- 修改Microsoft Excel输出配置

  双击 **Microsoft Excel 输出** 步骤，选择文件保存位置和文件名


Microsoft Excel 输出
步骤名称
Microsoft Excel 输出
文件&工作表
内容
文件
文件名
C:\software\data-integration\hdfs_export_test
浏览(B)...
扩展名
xlsx [Excel 2007 and above]
Stream XSLX data
分割每 ... 数据行
0
文件名包含步骤数目?
文件名包含日期?
文件名包含时间?
指定日期时间格式
日期时间格式
显示文件名...
如果文件已存在
覆盖原文件
在接收到数据前不创建文件
结果中添加文件名
工作表
工作表名
Sheet1
设为活动工作表
如果输出文件中已存在工作表
覆盖原工作表
保护工作表?(仅限 XLS 格式)
保护人
密码
模板
使用模板创建新文件
模板文件
template.xls
浏览(B)...
使用模板创建新工作表
模板工作表

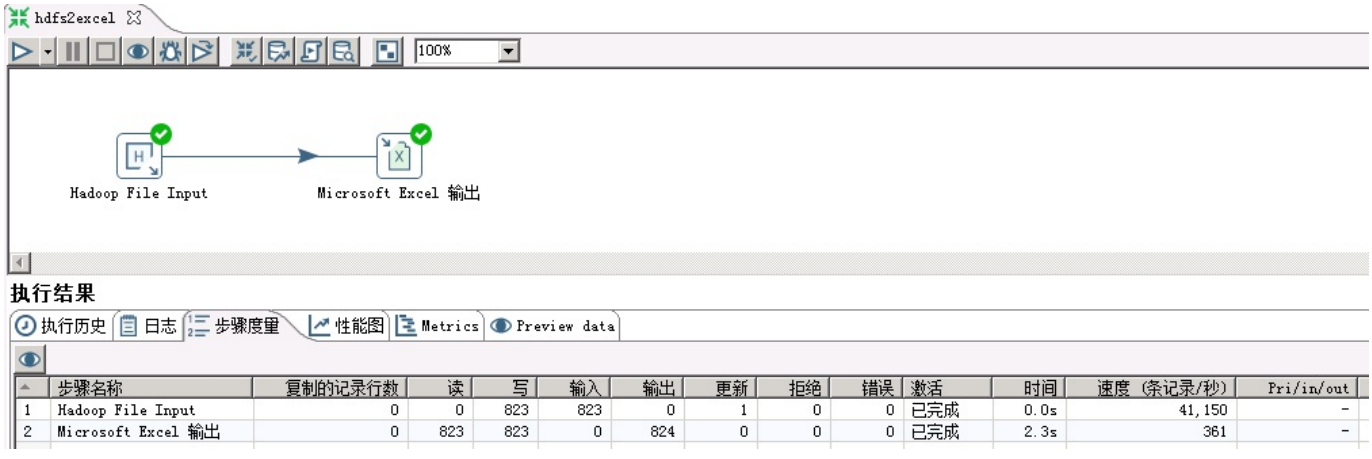
点击 **内容** 页签，获取字段



- 运行转换

  保存配置，点击 **执行** 按钮，选择 **本地执行**

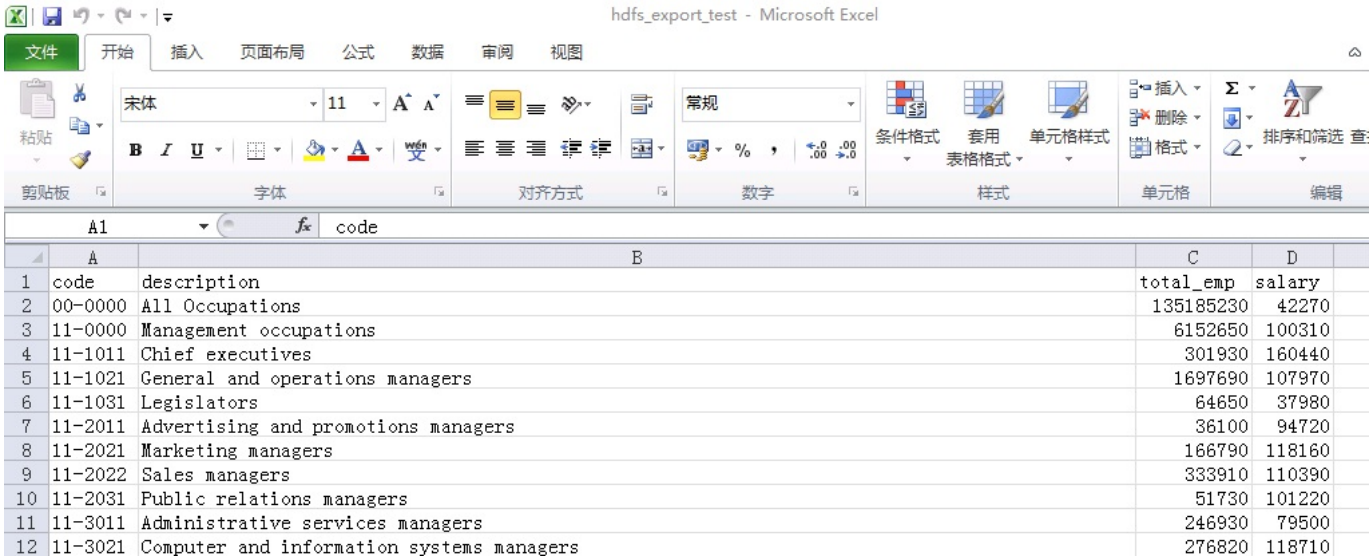
- 执行结果



执行结果

	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/out
1	Hadoop File Input	0	0	823	823	0	1	0	0	已完成	0.0s	41,150	-
2	Microsoft Excel 输出	0	823	823	0	824	0	0	0	已完成	2.3s	361	-

- 查看导出的excel文件



	A	B	C	D
1	code	description	total_emp	salary
2	00-0000	All Occupations	135185230	42270
3	11-0000	Management occupations	6152650	100310
4	11-1011	Chief executives	301930	160440
5	11-1021	General and operations managers	1697690	107970
6	11-1031	Legislators	64650	37980
7	11-2011	Advertising and promotions managers	36100	94720
8	11-2021	Marketing managers	166790	118160
9	11-2022	Sales managers	333910	110390
10	11-2031	Public relations managers	51730	101220
11	11-3011	Administrative services managers	246930	79500
12	11-3021	Computer and information systems managers	276820	118710