互联网营销精准策略

## 实验2-3 总结

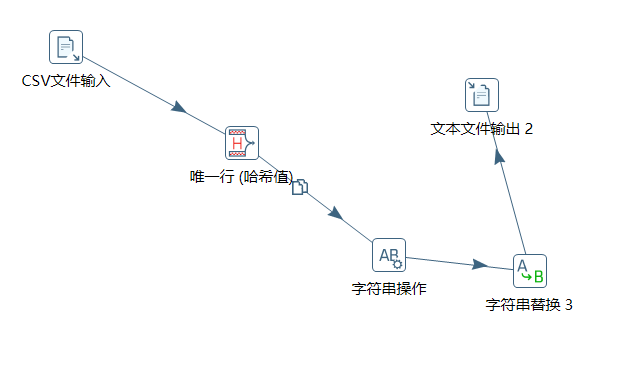
### Kettle数据清洗

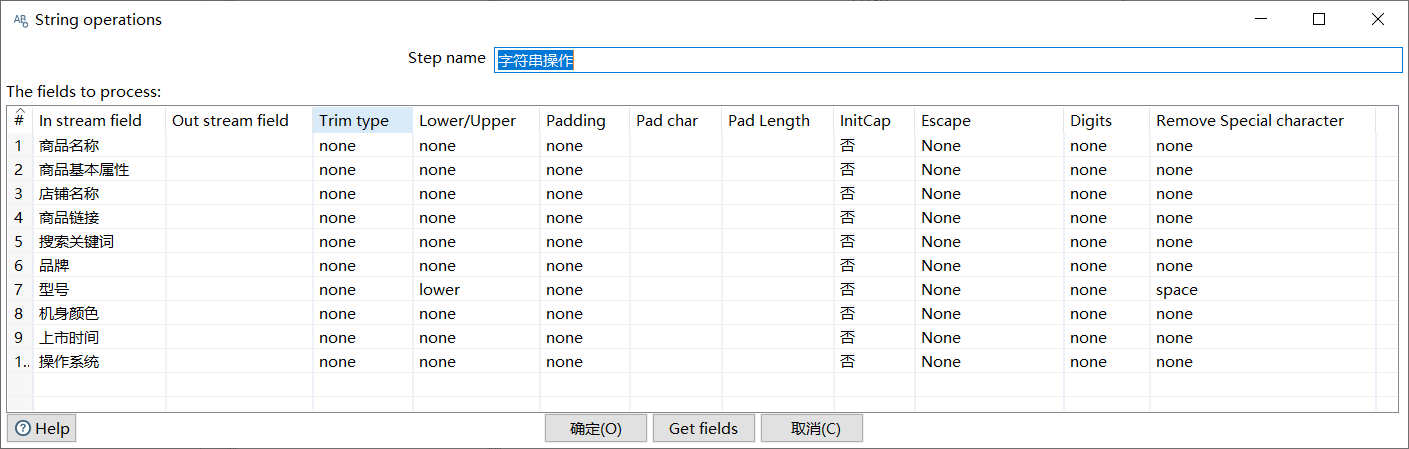
1. 新建转换去除手机销售信息表的重复记录
2. 在上述转换的基础上清洗手机销售信息表中型号字段中的数据

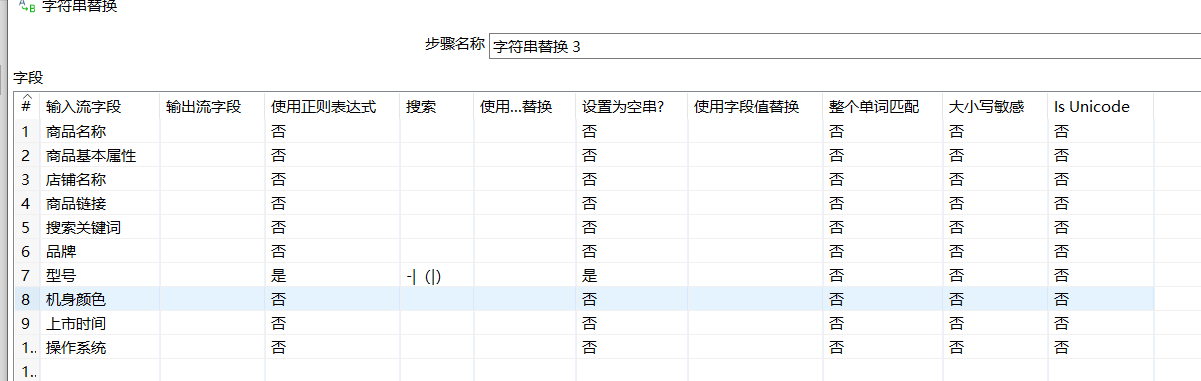
**要求**：去除该字段中的所有空格，方便后续聚合统计，字母统一大小写，

去除该字段中的所有特殊字符（各种标点符号）

一和二一起操作 用哈希去重 再用字符串操作去掉型号里的空格 统一大小写，用字符串替换换掉里面的“（）-”三个字符为空值，展示如下;







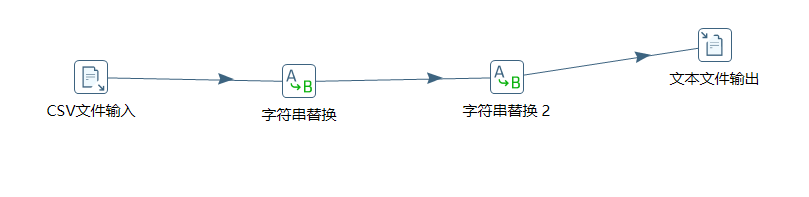
1. 新建转换去除用户评论信息表的重复记录

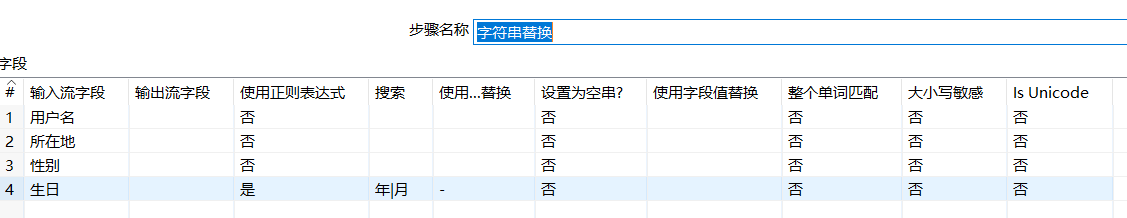
去重环节同上。

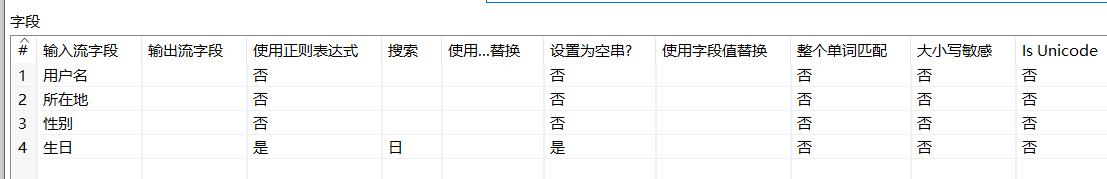


1. 新建转换处理用户信息表中出生日期字段（将 2019 年 5 月 20 日转换为 2019-5-20）

用字符串替换 替换掉年月日







1. 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。

更改编码格式实在txt文档下改的，步骤为：将文本另存为，这个时候选择编码格式为UTF-8.

实验3-1 总结

1. 建表；
2. 将本地文本数据上传至hive数据库，使用代码：

load data local inpath 'sales\_info.txt' overwrite into table sales\_info ;

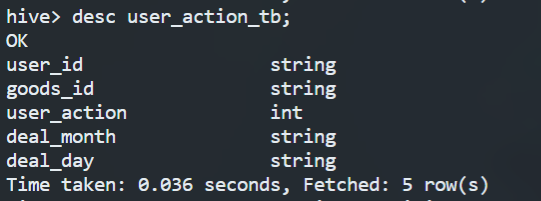
1. 处理手机销售表，删除不是手机的数据：

创建新表goods\_sail\_info：create table goods\_sail\_info like goods\_sail\_tb;

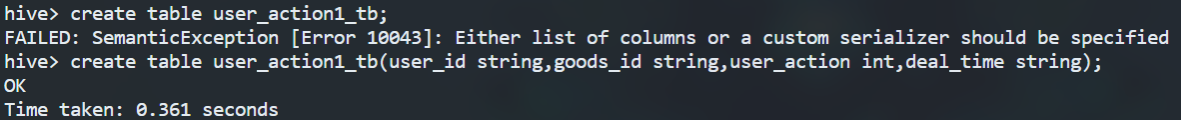
将旧表数据删除没用的再导入新表：insert into goods\_sail\_info select \* from goods\_sail\_tb where operate\_system != "";

1. 处理用户行为表数据，将交易月份与交易日拼接为新字段存储

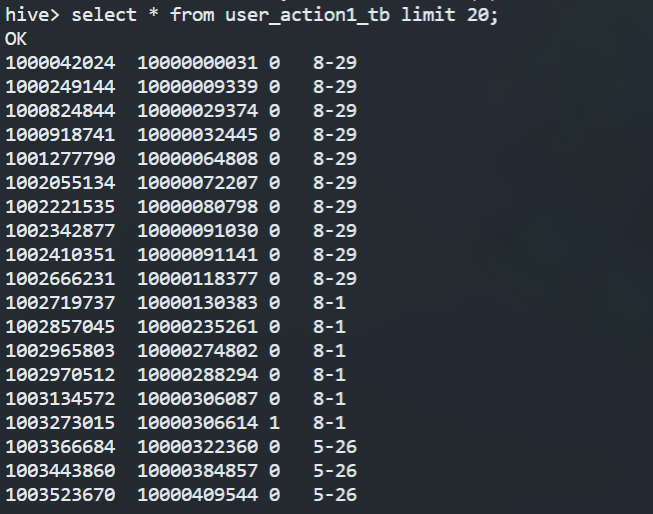
旧表：



创建新表：

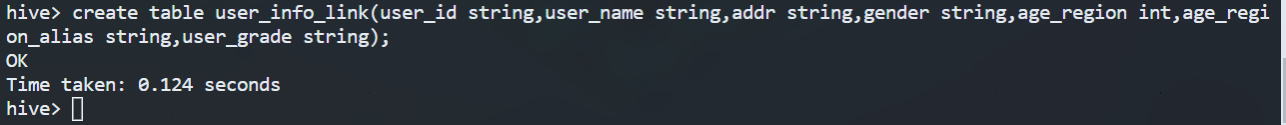






1. 处理用户信息表数据

建立新的信息表：user\_info\_link



建立中间表，表关联user\_info\_new\_tb和user\_comment\_tb

create table userinfo\_mid as select i.user\_id,i.user\_name,i.user\_region,i.gender,

floor(datediff(concat('',from\_unixtime(unix\_timestamp()),''),

concat('',birth,'')) / 365) age,c.user\_level from user\_info\_new\_tb i

join user\_comment\_tb c on i.user\_id=c.user\_id;

分段：

insert into user\_info\_link select user\_id,user\_name,user\_region,gender,

case when age>=0 and age<=17 then 1

when age >= 18 and age <=24 then 2

when age >= 25 and age <= 29 then 3

when age>= 30 and age <= 34 then 4

when age>= 35 and age <= 39 then 5

when age>=40 and age<=49 then 6

when age>=50 then 7 end as age\_region,

case when age>=0 and age<=17 then "18岁以下"

when age>=18 and age<=24 then "[18,24]"

when age>=25 and age<=29 then "[25,29]"

when age>=30 and age<=34 then "[30,34]"

when age>=35 and age<=39 then "[35,39]"

when age>=40 and age<=49 then "[40,49]"

when age>=50 then "50岁以上" end as age\_region\_alias

,user\_level from userinfo\_mid ;

