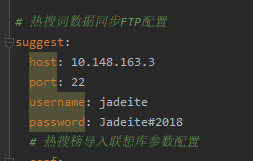
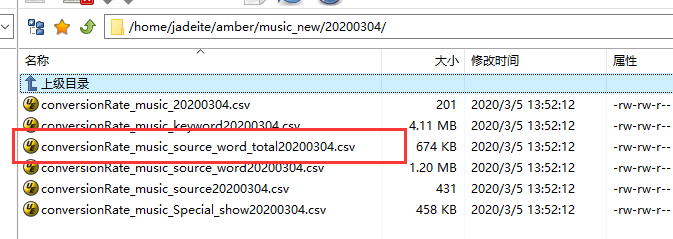
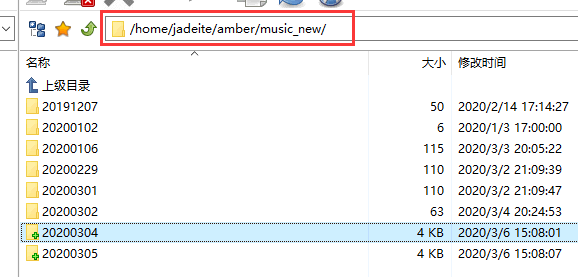
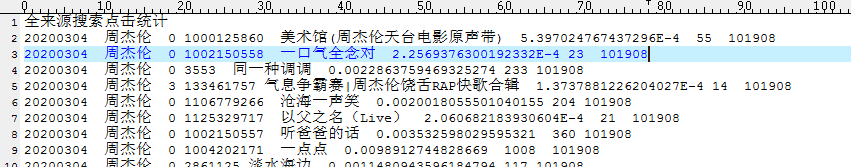
1. Amber文件位置 ，ftp到如下目录





数据含义：搜索词|点击内容类型|点击内容id|点击内容标题|转化率|点击量|搜索量



1. 启动数据处理任务jar包，启动任务参数如下：

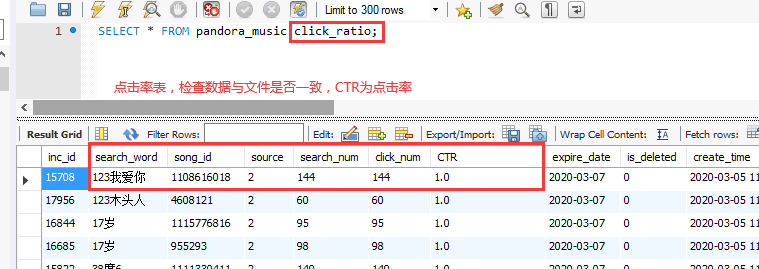


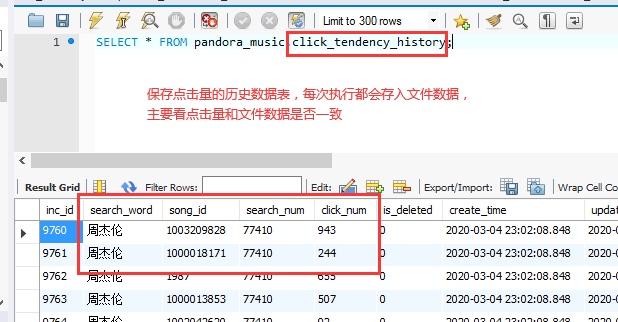
读取不同目录文件更改dt参数即可

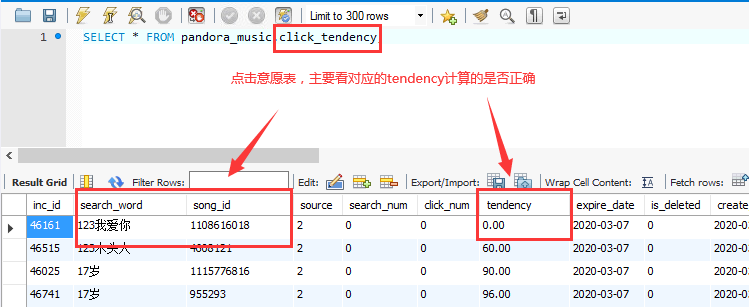
1. 任务跑完后检查表数据是否正确

（文件中有NULL的数据和有科学计数法的行数据会被过滤，不会入库，科学计数法比如1.667767776789E-4这种带E的）

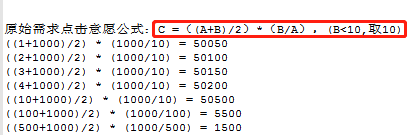
运营平台修改过的数据（即source不等于2的）不会被覆盖，文件中有相同search\_word&&song\_id的数据直接过滤不会入库







点击意愿计算公式如下：A为前日点击量，B为昨日点击量（前日和昨日是对应跑任务的机器时间，并非dt参数的时间。比如今天跑的任务dt 20200301指的就是昨日数据而不是0301日的数据）



点击意愿补充规则：如果只有A没有B则C=0 只有B没有A则C=1

1. 检查redis数据

点击率数据ClickRatio的值为数据库clikc\_ratio表中CTR的值\*1.5之后的百分比数据(ctr=0.5,则入redis数值为75%)；

>99%的按99%算，<60%的直接过滤不入redis

如果同一搜索词前两个song\_id的点击率差值<10%(比如搜索词周杰伦，一路向北点击95%，稻香90%)则整个搜索词(周杰伦)的数据都不入redis



