北京郵電大學

Beijing University of Posts and Telecommunications

期末综合

基于中国移动梧桐大数据平台
行业应用实训实验

实验三（三选一）

大数据交互式 OLAP 多维分析数据方案
-客户标签画像分析

实验三: 大数据交互式 OLAP 多维分析数据方案-客户标签画像分析

一、 实验描述:

针对个体客户, 结合历史数据对客户进行消费习惯、行为偏好、兴趣偏好等维度的分析, 形成多维客户画像, 帮助企业更清楚地了解客户, 结合客户特征开展个性化的产品营销及客户关怀, 提高客户满意度和忠诚度。

客户画像是企业在大数据背景下洞察客户的重要手段。客户画像即结合客户多维历史数据对客户设置标签, 标签包括个人基础信息、消费行为、兴趣偏好等, 如客户年龄、近 3 个月的月均消费、5G 网络感知质差、是否爱好体育等。

二、 实验目的:

以客户服务中的“网络质量感知质差画像”为例进行说明。通过客户 5G 上网话单中的页面访问时长、页面下载速度等, 以及 5G 语音话单中的接通次数、掉话次数等指标, 进行客户 5G 网络质量感知质差分析, 形成客户 5G 网络质量感知质差画像。通过该画像, 运营商可以找出 5G 网络质量感知质差客户, 通过针对这些客户开展个性化关怀活动, 提升客户满意度。通过本次实验达到下面两个目的:

1、 根据客户上网数据和语音通话数据开展 5G 网络质量分析, 确定客户 5G 网络网络感知质差画像。案例共需要使用两个原始数据模型表: 用户语音通话质量表和用户上网质量表, 如下:

用户语音通话质量表:

字段属性	字段名称
START_DATE	时间
SUBS_ID	用户编码
NET_TYPE	网络类型
MO_REQUEST_TIMES	始呼请求次数
MO_UNCONNECTED_TIMES	始呼未接通次数
MO_CONNECTED_RATE	始呼接通率
CONNECTED_TIMES	接通次数
DROPCALL_TIMES	掉话次数
DROPCALL_RATE	掉话率

用户上网质量表:

字段属性	字段名称
START_DATE	时间
SUBS_ID	手机号
NET_TYPE	网络类型
APP_TYPE	业务类型
PAGE_REQ_TIMES	页面访问次数
PAGE_BROWSING_DELAY	页面显示时长

PAGE_DOWNLOAD_THROUGHPUT	页面下载速率

经过大量数据分析，确定 5G 网络环境下各字段质差阈值。为方便说明，假设该案例 5G 网络的感知质差包括上网感知质差和语音通话感知质差，如下表所示。

质差类型	口径说明
始呼质差	始呼请求次数大于 20、始呼未接通次数大于 1 且始呼接通率小于 99%
掉话质差	掉话次数大于 1 且掉话率大于 1%
5G 语音通话质差口径	网络类型为 5G，存在始呼质差或者掉话质差
页面加载质差	页面显示时长大于或等于 5000ms 或页面下载速率小于 50kB/s
5G 上网质差口径	网络类型为 5G、上网类型为 HTTP、页面访问次数大于 500 且存在页面加载质差
5G 网络感知质差口径	存在 5G 语音通话感知质差或 5G 上网感知质差

2、 输出 5G 网络感知质差客户画像和 5G 网络感知质差用户分布。如下表所示：

5G 网络感知质差客户画像：

序号	用户标识	是否为 5G 网络感知质差用户
1	311****5044	不是
2	311****5037	是

5G 网络感知质差用户分布：

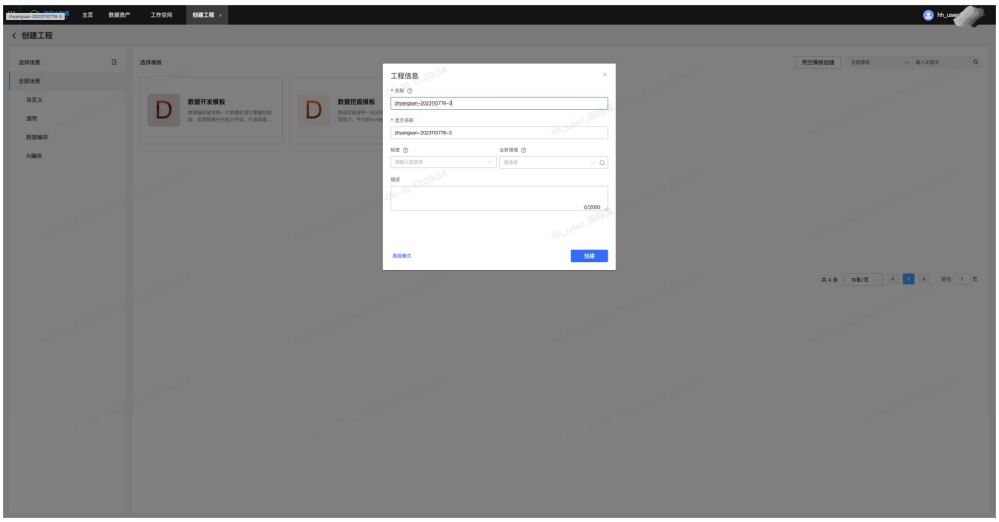
序号	是否为 5G 网络感知质差用户	用户数量
1	不是	4586475
2	是	1022

三、 实验环境：

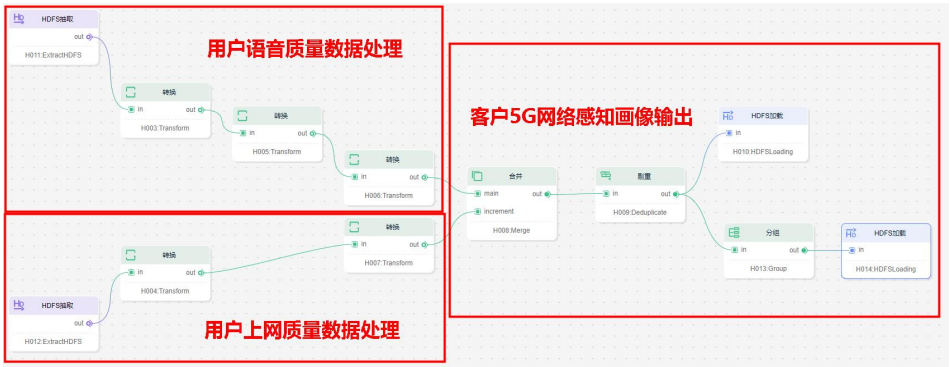
梧桐·鸿鹄大数据实训平台

四、 实验步骤：

附：截图要求：能看到右上角用户信息，右下角若有时间需要将时间也截进来。



场景二的实践案例详细编排流程如下图所示。



说明：本案例输入数据的 HDFS 路径为/user/popularization/data/，路径下的 td_ns_cs_call_d.txt 表示用户语音通话质量表、td_ns_ps_http_d.txt 表示用户上网质量表。

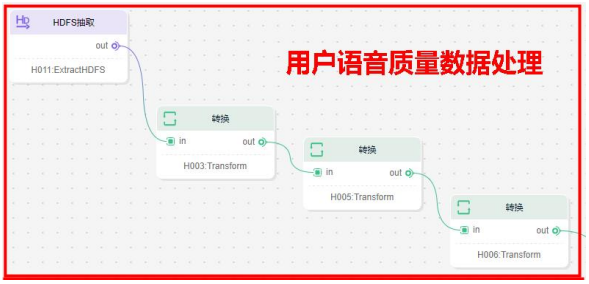
数据处理后将用户 5G 网络感知画像和用户 5G 网络感知汇总结果加载为 CSV 文件，分别命名为 user_5G_zicha_d.csv 和 user_5G_zicha_group_d.csv，最后将其存储到 HDFS 中，存储路径为/user/wutong/example_data/。

以“用户语音通话质量表”和“用户上网质量表”为数据源，分别分析出语音通话质差用户画像和上网质差用户画像，最后组合成用户 5G 网络感知画像以及用户 5G 网络感知汇总结果，整体流程如上图所示。

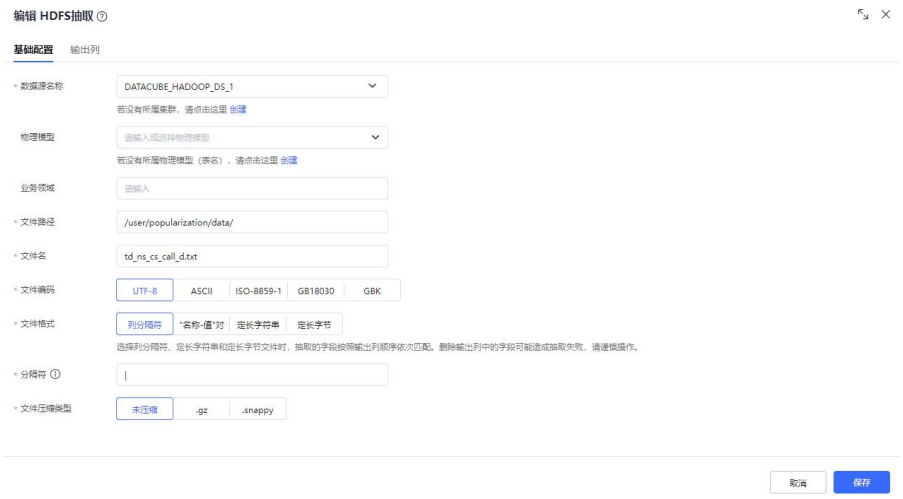
4.1 数据处理

步骤一：

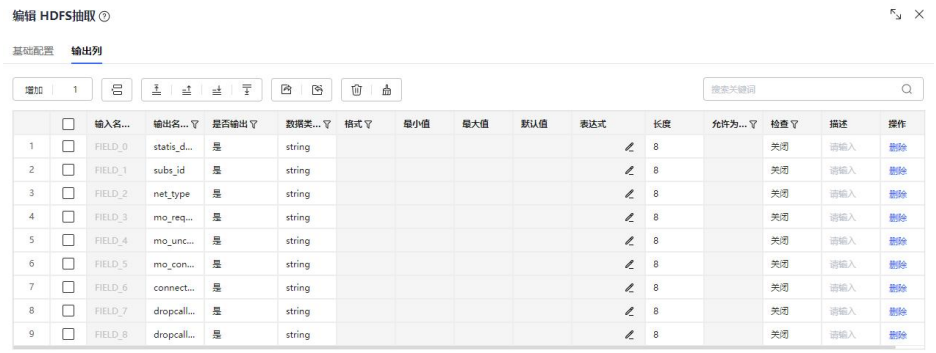
用户 5G 语音通话质量画像计算流程为对用户语音通话质量数据源进行加工，获取用户语音质量的数据集合，如下图所示。



首先使用 HDFS 抽取算子将数据源中的数据抽取到数据流中，如下图所示，编辑 HDFS 算子：



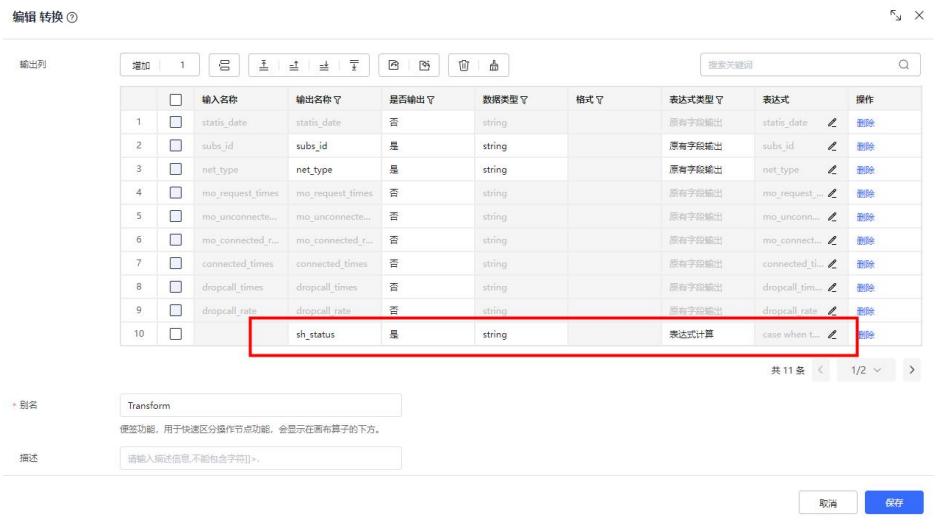
结果如下图所示。



[给出 HDFS 编辑截图以及结果截图，要求能看到用户个人信息以及时间信息]

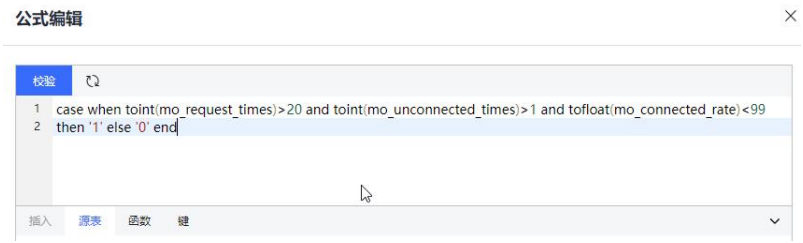
然后使用转换算子，分别计算用户“始呼质差标志”和“掉话质差”标志，如下图所示。

编辑转换算子（1）：

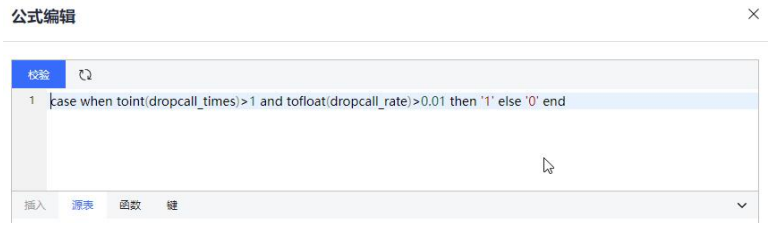


编辑转换算子（2）：

“始呼质差标志”计算公式下图所示。

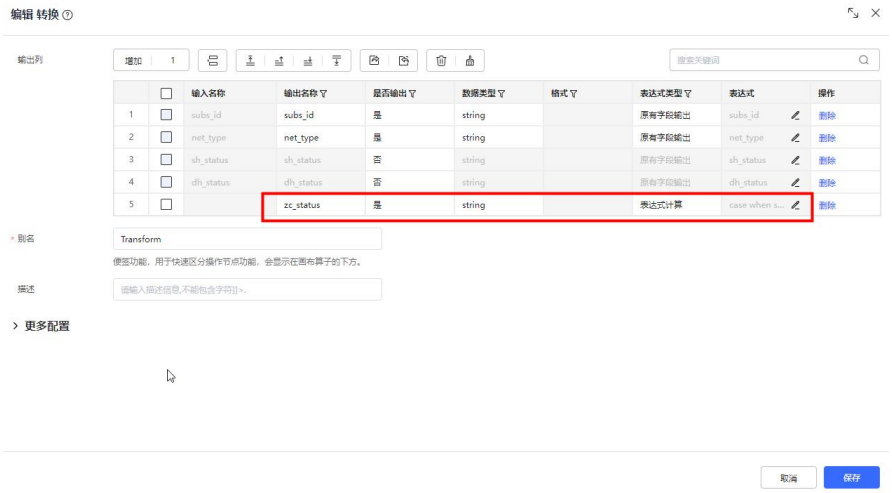


“掉话质差标志”计算公式如下图所示。

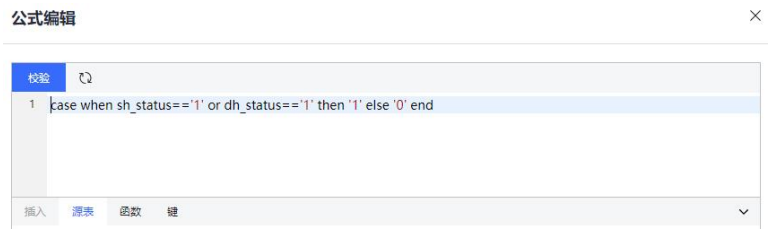


[给出计算用户“始呼质差标志”和“掉话质差”标志算子以及公式截图，要求能看到用户信息以及时间信息]

接下来，再次使用转换算子，将两种质差标志用户组合，不论哪一种质差情况，都应该属于语音通话质差用户，设置为“语音通话质差标志”，如下图所示。

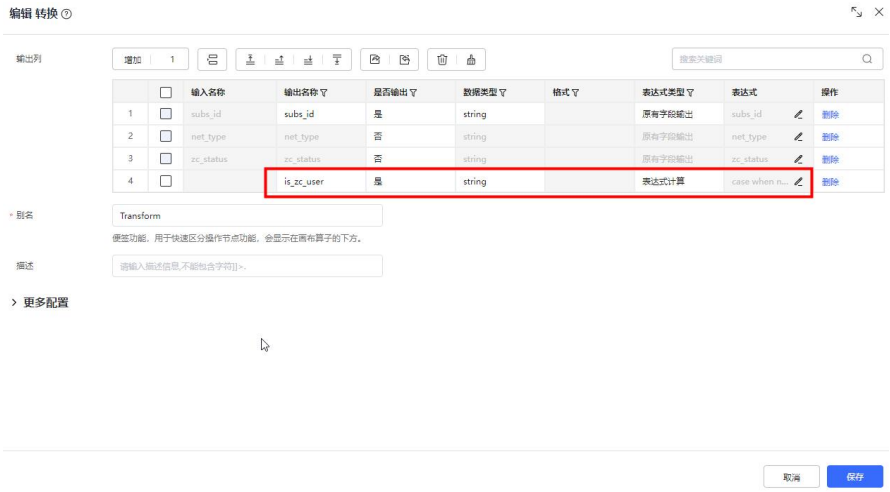


“语音通话质差标志”计算公式如下图所示。

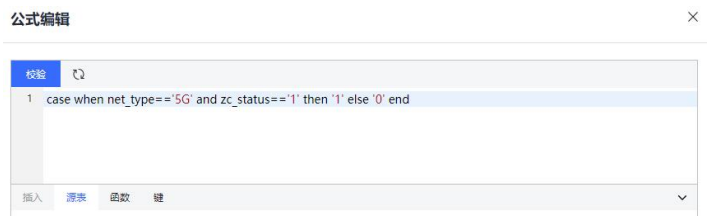


最后，再次使用转换算子，判断网络类型，计算用户的“5G 语音通话质差标志”，如下

图所示。



“5G 语音通话质差标志”计算公式如下图所示。

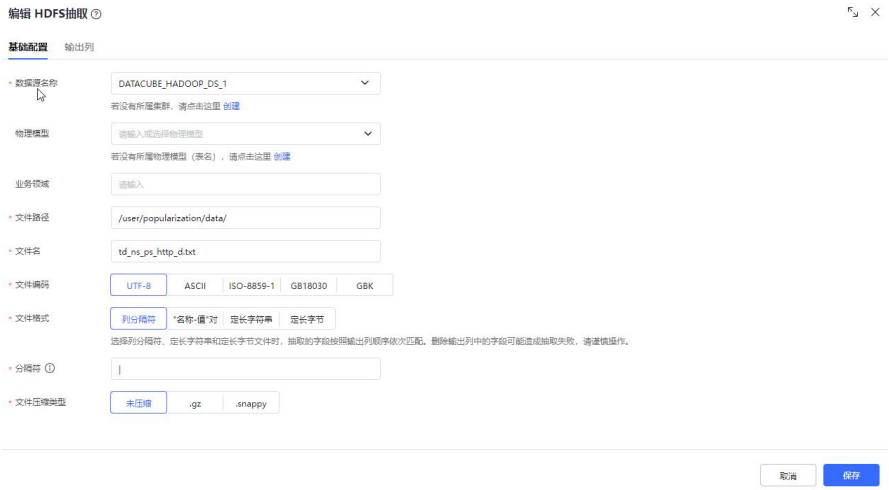


[给出计算用户的“5G 语音通话质差标志”转换算子以及公式截图，要求能看到用户信息以及时间信息]

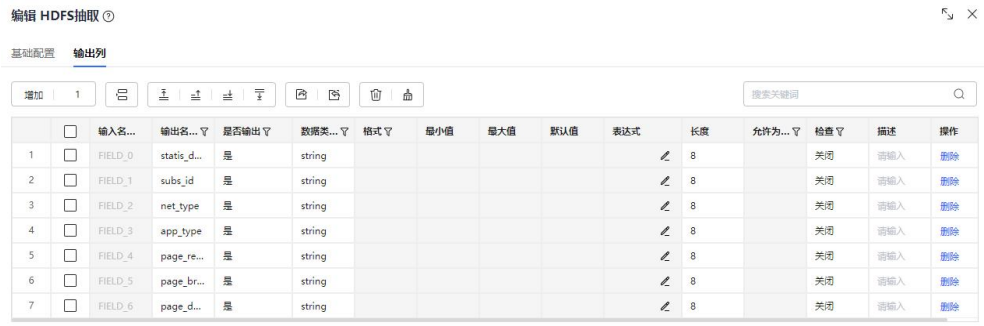
步骤二：

用户上网质量画像计算流程为对用户上网质量数据源进行加工，获取用户 5G 上网质量的数据集合，如下图所示。

首先使用 HDFS 抽取算子将数据源中的数据抽取到数据流中，如下图所示，

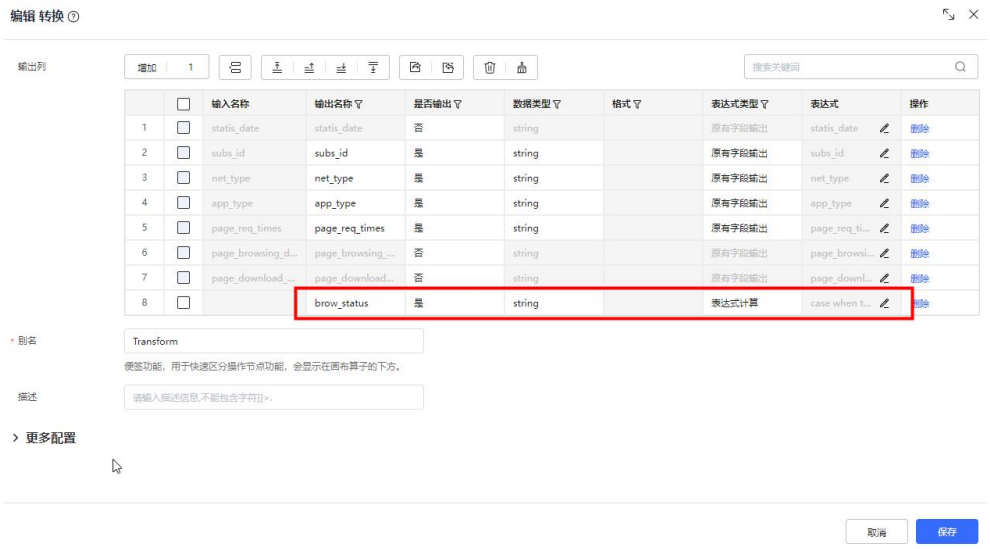


结果如下图所示。



[给出 HDFS 算子截图以及输出列截图，要求能够看到用户的信息以及时间信息]

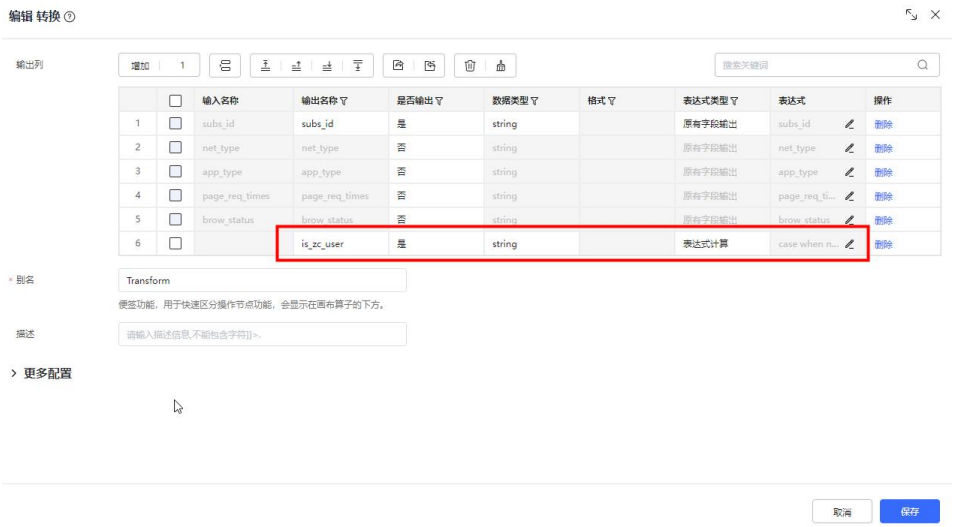
然后使用转换算子，计算用户“页面加载质差标志”，如下图所示。



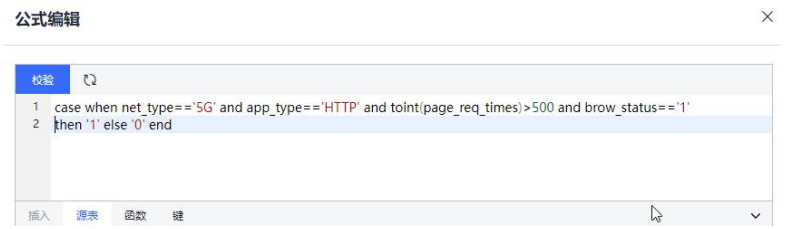
“页面加载质差标志”计算公式如下图所示。

[要求给出算子及公式截图，要求能看到用户信息以及时间信息]

再次使用转换算子，判断网络类型，计算用户的“5G 上网质差标志”，如下图所示。



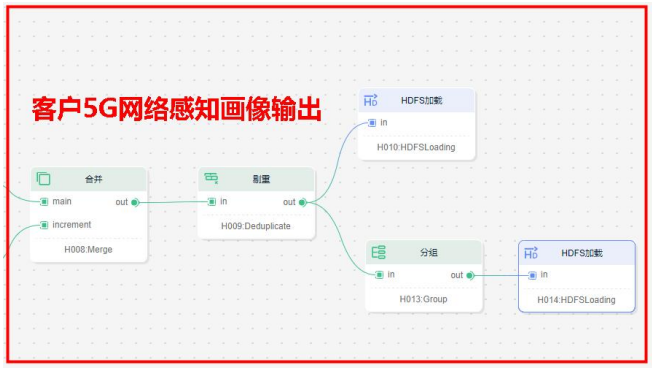
计算公式如下：



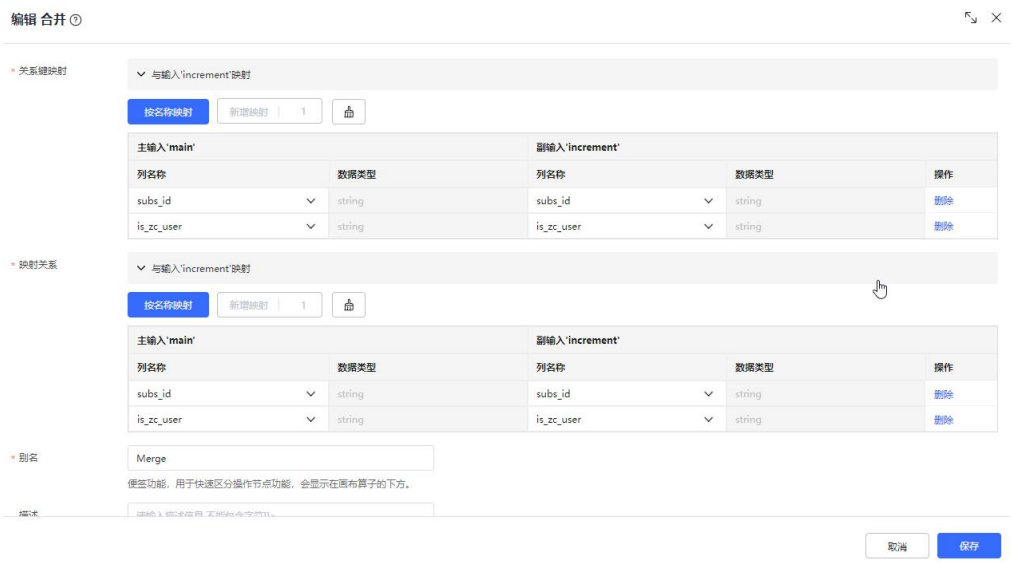
[要求给出算子及公式截图，要求能看到用户信息以及时间信息]

4.2 数据输出：

5G 网络感知画像计算数据流，如下图所示，将计算出的“5G 语音通话质差标志”和“5G 上网质差标志”合并，得到客户 5G 网络感知画像，并统计 5G 网络感知用户分布情况。



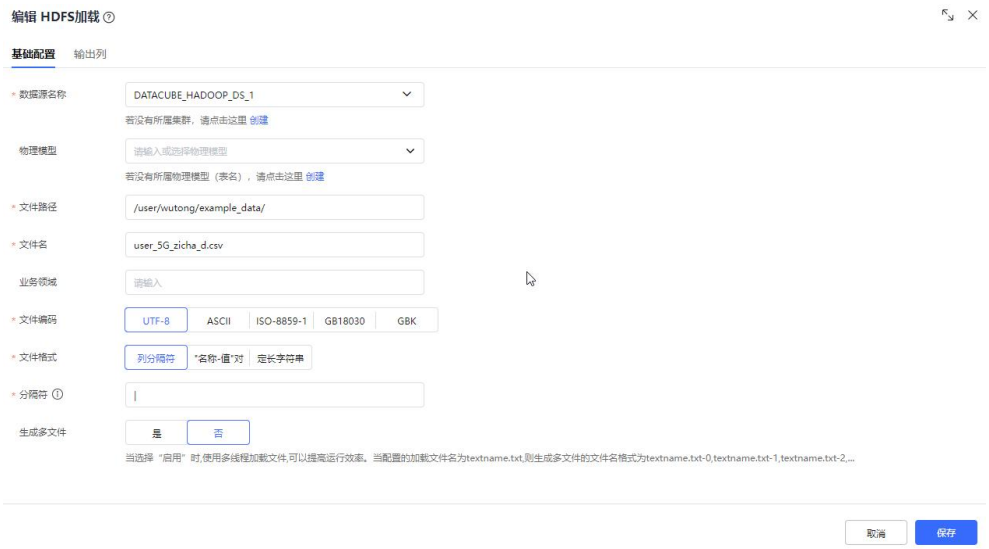
首先使用合并算子将计算出的“5G 语音通话质差标志”和“5G 上网质差标志”合并，如下图所示。



然后使用剔重算子，如下图所示，判断用户是否出现质差标志，只要有一个用户出现质差标志，就被标记为 5G 网络感知质差用户，同一个用户只标记一次。

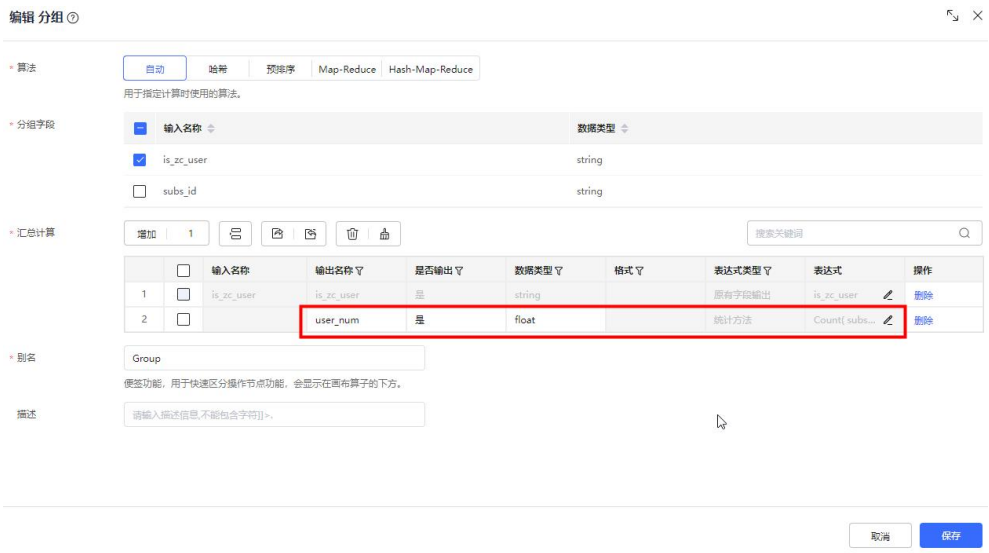
[给出剔重算子以及合并后的算子截图，要求能够看到个人信息以及时间信息]

接下来，使用 HDFS 加载算子将 5G 网络感知用户画像加载到 HDFS 中，如下图所示，



结果如下图所示。

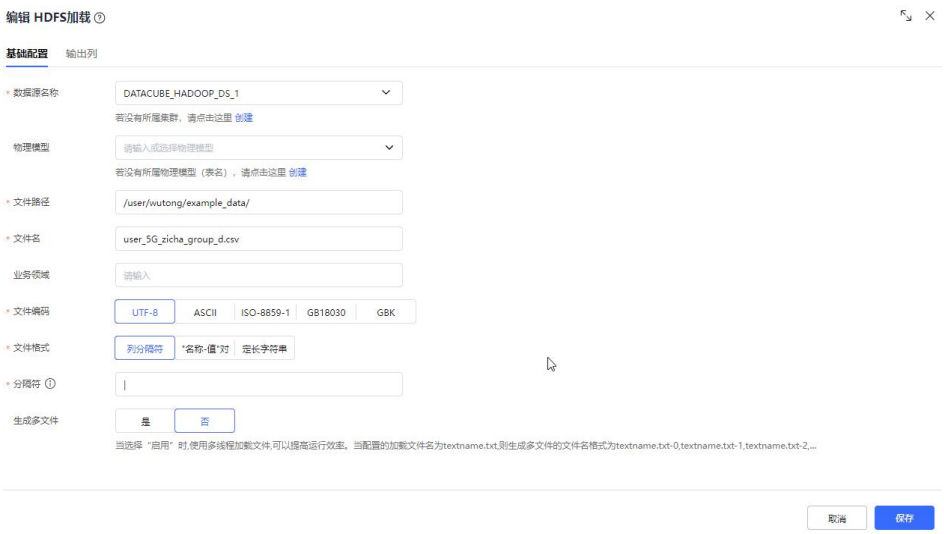
为了得到 5G 网络质差用户和非 5G 网络质差用户的数量，使用分组算子将 5G 网络感知用户画像按照是否质差用户进行分组计算，如下图所示。



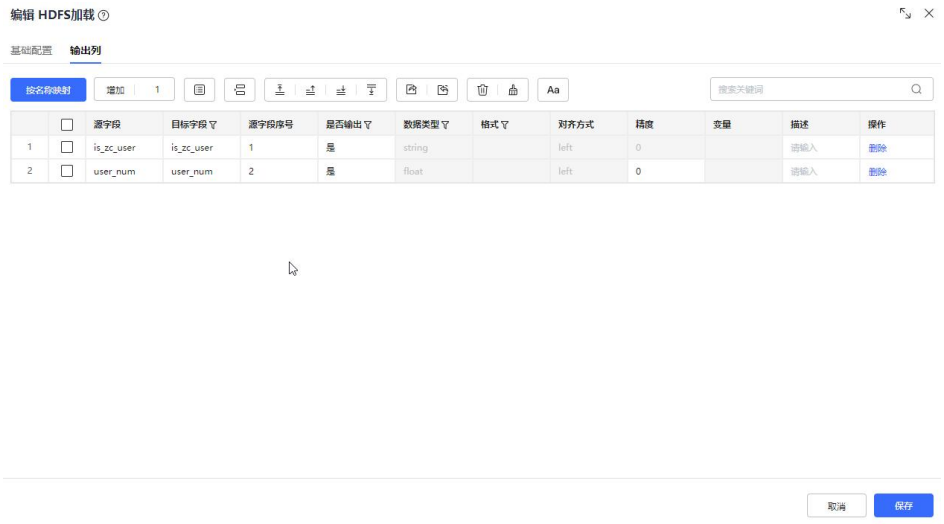
[给出 HDFS 加载算子编辑以及输出截图和分组算子截图，要求能够看到个人信息以及时间信息]

分组用户数量表达式：Count(subs_id)。

最后，使用 HDFS 加载算子将用户 5G 网络感知统计结果加载到 HDFS 上的指定文件中，如下图所示，



结果如下图所示。



[给出加载算子以及结果截图，要求能够看到个人信息以及时间信息]

数据编排完成后，进入在线调测，分别查看 HDFS 加载算子的调测结果。5G 网络感知用户画像如下图所示。

	subs_id	is_zc_user
1	3110110925044	0
2	3112384647848	0
3	3110580025433	0
4	3112008175325	0
5	3110580340028	0
6	3112247149741	0

5G 网络感知用户分布如下图所示。

[给出结果截图，要求能够看到个人信息以及时间信息]

五、 实验结果：

1、 提供上述截图，包括：（每个截图 1 分）

（1） 抽取算子将数据源中的数据抽取到数据流中时 HDFS 编辑截图以及结果截图

（2） 计算用户“始呼质差标志”和“掉话质差”算子以及公式截图

（3） 计算用户的“5G 语音通话质差标志”算子以及公式截图

（4） 对用户上网质量数据源进行加工时的 HDFS 算子及输出列截图

（5） 计算用户“页面加载质差标志”算子及公式截图

（6） 计算用户的“5G 上网质差标志”算子及公式截图

（7） 剔重算子以及合并后的算子截图

（8） HDFS 加载算子编辑以及输出截图和分组算子截图

（9） 将用户 5G 网络感知统计结果加载到 HDFS 上的指定文件中的加载算子以及结果截图

（10） 5G 网络感知用户画像以及 5G 网络感知用户分布截图

2、 截图以及代码的相关语言描述（5 分）

3、 实验总结以及自己的思考（5 分）