



北京邮电大学

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：

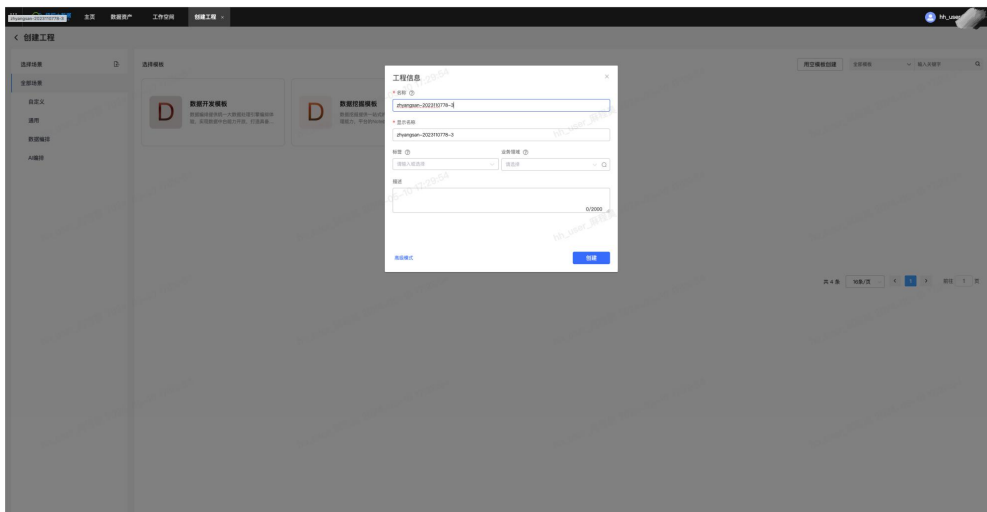
序号	是否为 5G 网络感知质差用户	用户数量
1	不是	4586475
2	是	1022

三、 实验环境：

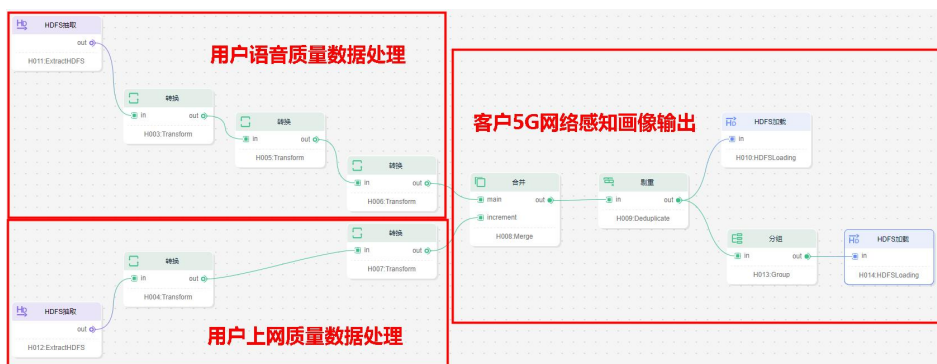
梧桐·鸿鹄大数据实训平台

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：

编辑 HDFS抽取

基础配置

输出列

数据源名称

DATAcube_HADOOP_DS_1

若没有所属集群，请点击这里 创建

物理模型

选择输入或选择物理模型

若没有所属物理模型（表名），请点击这里 创建

业务领域

选择输入

文件路径

/user/popularization/data/

文件名

td_ns_cs_call_d.txt

文件编码

UTF-8

ASCII

ISO-8859-1

GB18030

GBK

文件格式

列分隔符

*名称+值+对

定长字符串

定长字节

选择列分隔符、定长字符串和定长字节文件时，抽取的字段按照输出列顺序依次匹配。删除输出列中的字段可能造成抽取失败，请谨慎操作。

分隔符

|

文件压缩类型

未压缩

.gz

.snappy

取消

保存

结果如下图所示。

编辑 HDFS抽取

基础配置

输出列

增加

1

搜索关键词

Q

<input type="checkbox"/>	输入名...	输出名...	是否输出	数据类型	格式	最小值	最大值	默认值	表达式	长度	允许为...	检查	描述	操作
1	<input type="checkbox"/> FIELD_0	status_d...	是	string						8		关闭	待输入	删除
2	<input type="checkbox"/> FIELD_1	subs_id	是	string						8		关闭	待输入	删除
3	<input type="checkbox"/> FIELD_2	net_type	是	string						8		关闭	待输入	删除
4	<input type="checkbox"/> FIELD_3	mo_req...	是	string						8		关闭	待输入	删除
5	<input type="checkbox"/> FIELD_4	mo_unc...	是	string						8		关闭	待输入	删除
6	<input type="checkbox"/> FIELD_5	mo_con...	是	string						8		关闭	待输入	删除
7	<input type="checkbox"/> FIELD_6	connect...	是	string						8		关闭	待输入	删除
8	<input type="checkbox"/> FIELD_7	dropcall...	是	string						8		关闭	待输入	删除
9	<input type="checkbox"/> FIELD_8	dropcall...	是	string						8		关闭	待输入	删除

[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

输出列

增加

1

搜索关键词

Q

<input type="checkbox"/>	输入名称	输出名称	是否输出	数据类型	格式	表达式类型	表达式	操作
1	<input type="checkbox"/> status_date	status_date	否	string		原有字段输出	status_date	删除
2	<input type="checkbox"/> subs_id	subs_id	是	string		原有字段输出	subs_id	删除
3	<input type="checkbox"/> net_type	net_type	是	string		原有字段输出	net_type	删除
4	<input type="checkbox"/> mo_request_times	mo_request_times	否	string		原有字段输出	mo_request...	删除
5	<input type="checkbox"/> mo_unconnecte...	mo_unconnecte...	否	string		原有字段输出	mo_unconn...	删除
6	<input type="checkbox"/> mo_connected_r...	mo_connected_r...	否	string		原有字段输出	mo_connect...	删除
7	<input type="checkbox"/> connected_times	connected_times	否	string		原有字段输出	connected_ti...	删除
8	<input type="checkbox"/> dropcall_times	dropcall_times	否	string		原有字段输出	dropcall_tim...	删除
9	<input type="checkbox"/> dropcall_rate	dropcall_rate	否	string		原有字段输出	dropcall_rate	删除
10	<input type="checkbox"/>	sh_status	是	string		表达式计算	case when t...	删除

共 11 条 < 1/2 >

别名

Transform

便捷功能，用于快速区分操作节点功能，会显示在画布算子的下方。

描述

请输入描述信息,不能包含字符]]>。

取消

保存

编辑转换算子（2）：

	<input type="checkbox"/>	输入名称	输出名称 ▾	是否输出 ▾	数据类型 ▾	格式 ▾	表达式类型 ▾	表达式	操作
11	<input type="checkbox"/>		dh_status	是	string		表达式计算	case when t...	删除

共 11 条 < 2/2 >

“始呼质差标志”计算公式下图所示。

公式编辑 ×

校验

```

1 case when toint(mo_request_times)>20 and toint(mo_unconnected_times)>1 and tofloat(mo_connected_rate)<99
2 then '1' else '0' end

```

插入 源表 函数 键

“掉话质差标志”计算公式如下图所示。

公式编辑 ×

校验

```

1 case when toint(dropcall_times)>1 and tofloat(dropcall_rate)>0.01 then '1' else '0' end

```

插入 源表 函数 键

[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。

编辑 转换 ×

输入列

	<input type="checkbox"/>	输入名称	输出名称 ▾	是否输出 ▾	数据类型 ▾	格式 ▾	表达式类型 ▾	表达式	操作
1	<input type="checkbox"/>	subs_id	subs_id	是	string		原字段输出	subs_id	删除
2	<input type="checkbox"/>	net_type	net_type	是	string		原字段输出	net_type	删除
3	<input type="checkbox"/>	sh_status	sh_status	否	string		原字段输出	sh_status	删除
4	<input type="checkbox"/>	dh_status	dh_status	否	string		原字段输出	dh_status	删除
5	<input type="checkbox"/>		zc_status	是	string		表达式计算	case when t...	删除

别名 Transform

描述 请输入描述信息,不能包含字符[]>

> 更多配置

取消 保存

“语音通话质差标志”计算公式如下图所示。

公式编辑 ×

校验

```

1 case when sh_status=='1' or dh_status=='1' then '1' else '0' end

```

插入 源表 函数 键

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。

编辑 转换

增加 1

搜索关键词

	输入名称	输出名称	是否输出	数据类型	格式	表达式类型	表达式	操作
1	<input type="checkbox"/>	subs_id	是	string		原有字段输出	subs_id	删除
2	<input type="checkbox"/>	net_type	否	string		原有字段输出	net_type	删除
3	<input type="checkbox"/>	zc_status	否	string		原有字段输出	zc_status	删除
4	<input type="checkbox"/>	ic_zc_user	是	string		表达式计算	case when net_type='5G' and zc_status='1' then '1' else '0' end	删除

别名Transform
便捷功能，用于快速区分操作节点功能，会显示在画布算子的下方。

描述请输入描述信息，不能包含字符[<]。>

更多配置

取消

保存

“5G 语音通话质差标志”计算公式如下图所示。

公式编辑

校验

1 case when net_type='5G' and zc_status='1' then '1' else '0' end

插入

函数

键

[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，

编辑 HDFS抽取

基础配置

输出列

数据源名称

物理模型

业务领域

文件路径

文件名

文件编码

文件格式

分隔符

文件压缩类型

DATACUBE_HADOOP_DS_1

请输入或选择物理模型

请输入

/user/popularization/data/

td_ns_ps_http_d.txt

UTF-8 ASCII ISO-8859-1 GB18030 GBK

列分隔符 名称+值对 定长字符串 定长字节

|

未压缩 .gz .snappy

取消

保存

编辑 HDFS抽取

基础配置 输出

增加 1

搜索关键词

<input type="checkbox"/>	输入名...	输出名...	是否输出	数据类型...	格式	最小值	最大值	默认值	表达式	长度	允许为...	检查	描述	操作
1	<input type="checkbox"/> FIELD_0	statistic_d...	是	string						8		关闭	省输入	删除
2	<input type="checkbox"/> FIELD_1	subs_id	是	string						8		关闭	省输入	删除
3	<input type="checkbox"/> FIELD_2	net_type	是	string						8		关闭	省输入	删除
4	<input type="checkbox"/> FIELD_3	app_type	是	string						8		关闭	省输入	删除
5	<input type="checkbox"/> FIELD_4	page_re...	是	string						8		关闭	省输入	删除
6	<input type="checkbox"/> FIELD_5	page_br...	是	string						8		关闭	省输入	删除
7	<input type="checkbox"/> FIELD_6	page_d...	是	string						8		关闭	省输入	删除

编辑 转换

取消 保存

输出

增加 1

搜索关键词

	<input type="checkbox"/>	输入名称	输出名称 ▾	是否输出 ▾	数据类型 ▾	格式 ▾	表达式类型 ▾	表达式	操作
1	<input type="checkbox"/>	static_date	static_date	否	string		原有字段输出	static_date	删除
2	<input type="checkbox"/>	subs_id	subs_id	是	string		原有字段输出	subs_id	删除
3	<input type="checkbox"/>	net_type	net_type	是	string		原有字段输出	net_type	删除
4	<input type="checkbox"/>	app_type	app_type	是	string		原有字段输出	app_type	删除
5	<input type="checkbox"/>	page_req_times	page_req_times	是	string		原有字段输出	page_req_ti...	删除
6	<input type="checkbox"/>	page_browsing_d...	page_browsing...	否	string		原有字段输出	page_browsi...	删除
7	<input type="checkbox"/>	page_download...	page_download...	否	string		原有字段输出	page_downloa...	删除
8	<input type="checkbox"/>		brow_status	是	string		表达式计算	case when t...	删除

别名

Transform

便捷功能，用于快速区分操作节点功能，会显示在画布节点的下方。

描述

请输入描述信息,不能包含字符[]>。

> 更多配置

公式编辑

校验

1

`case when tofloat(page_browsing_delay) >= 5000 or tofloat(page_download_throughput) < 50 then '1' else '0' end`

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。

编辑 转换 ①

输出列

	输入名称	输出名称	是否输出	数据类型	格式	表达式类型	表达式	操作
1	subs_id	subs_id	是	string		原有字段输出	subs_id	删除
2	net_type	net_type	否	string		原有字段输出	net_type	删除
3	app_type	app_type	否	string		原有字段输出	app_type	删除
4	page_req_times	page_req_times	否	string		原有字段输出	page_req_times	删除
5	brow_status	brow_status	否	string		原有字段输出	brow_status	删除
6		is_zc_user	是	string		表达式计算	case when n...	删除

别名: Transform

描述: 便捷功能, 用于快速区分操作节点功能, 会显示在画布算子的下方。

更多配置

取消 保存

计算公式如下:

公式编辑

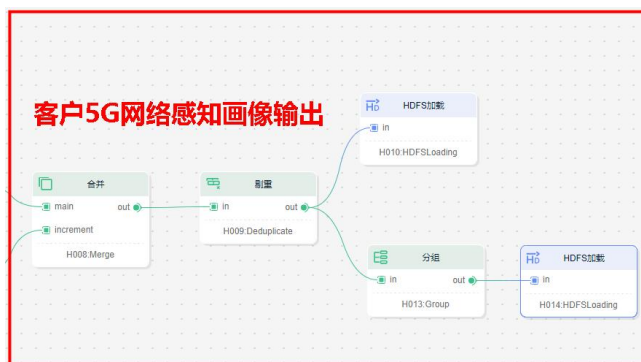
```
1 case when net_type=='5G' and app_type=='HTTP' and toint(page_req_times)>500 and brow_status=='1'
2 then '1' else '0' end
```

插入 函数 键

[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。

编辑 合并

关系键映射

与输入 'increment' 映射

按名称映射

新增映射

1

主输入 'main'		副输入 'increment'		
列名称	数据类型	列名称	数据类型	操作
subs_id	string	subs_id	string	删除
is_zc_user	string	is_zc_user	string	删除

映射关系

与输入 'increment' 映射

按名称映射

新增映射

1

主输入 'main'		副输入 'increment'		
列名称	数据类型	列名称	数据类型	操作
subs_id	string	subs_id	string	删除
is_zc_user	string	is_zc_user	string	删除

别名

Merge

便捷功能，用于快速区分操作节点功能，会显示在画布算子的下方。

描述

取消

保存

然后使用别重算子，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。

编辑 别重

别重类型

第一行

最后一行

任意一行

对于排序后的记录，发现重复时，保留第一行数据。

字段信息配置

搜索关键词

	别重字段	排序字段	输入名称	数据类型	格式	优先级	排序
1	是	否	subs_id	string		0	升序
2	否	是	is_zc_user	string		1	降序

别名

Deduplicate

便捷功能，用于快速区分操作节点功能，会显示在画布算子的下方。

描述

请输入描述信息,不能包含字符[]>。

取消

保存

[给出别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，

编辑 HDFS加载

基础配置

输出列

* 数据源名称

DATAcube_HADOOP_DS_1

若没有所属集群，请点击[在这里创建](#)

物理模型

请输入或选择物理模型

若没有所属物理模型（表名），请点击[在这里创建](#)

* 文件路径

/user/wutong/example_data/

* 文件名

user_5G_zicha_d.csv

业务领域

请输入

* 文件编码

UTF-8

ASCII

ISO-8859-1

GB18030

GBK

* 文件格式

列分隔符

"名称-值"对

是长字符串

* 分隔符

|

生成多文件

是

否

当选择“启用”时,使用多线程加载文件,可以提高运行效率。当配置的加载文件名名为textname.txt,则生成多文件的文件各格式为textname.txt-0,textname.txt-1,textname.txt-2...

结果如下图所示。

编辑 HDFS加载

基础配置 **输出列**

	<input type="checkbox"/>	源字段	目标字段	源字段序号	是否输出	数据类型	格式	对齐方式	精度	变量	描述	操作
1	<input type="checkbox"/>	subs_id	subs_id	1	是	string		left	0		读输入	删除
2	<input type="checkbox"/>	is_zc_user	is_zc_user	2	是	string		left	0		写输入	删除

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 用户画像按照是否质差用户进行分组计算，如下图所示。

编辑 分组

* 算法
自动 端串 预排序 Map-Reduce Hash-Map-Reduce
用于指定计算时使用的算法。

分组字段

<input checked="" type="checkbox"/>	输入名称	数据类型
<input checked="" type="checkbox"/>	is_zc_user	string
<input type="checkbox"/>	subs_id	string

* 汇总计算

	<input type="checkbox"/>	输入名称	输出名称	是否输出	数据类型	格式	表达式类型	表达式	操作
1	<input type="checkbox"/>	is_zc_user	is_zc_user	是	string		原有字段输出	is_zc_user	删除
2	<input type="checkbox"/>		user_num	是	float		统计方法	Count(subs_	删除

* 别名
Group

便捷功能：用于快速区分操作节点功能，会显示在画布算子的下方。
描述：
请输入描述信息，不能包含字符[]<>。

[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,

编辑 HDFS加载

基础配置

输出列

数据源名称

DATAcube_HADOOP_DS_1

若没有所属集群, 请点击这里 创建

物理模型

请输入或选择物理模型

若没有所属物理模型 (表名), 请点击这里 创建

文件路径

/user/wutong/example_data/

文件名

user_5G_zicha_group_d.csv

业务领域

请输入

文件编码

UTF-8

ASCII

ISO-8859-1

GB18030

GBK

文件格式

列分隔符

*名称+值*对

定长字符串

分隔符

,

生成多文件

是

否

当选择 "启用" 时, 使用多线程加载文件, 可以提高运行效率。当配置的加载文件名为textname.txt, 则生成多文件的文件名格式为textname.txt-0, textname.txt-1, textname.txt-2...

取消

保存

结果如下图所示。

编辑 HDFS加载

基础配置

输出列

按名称映射

增加

1

重置

清除

上一步

下一步

打印

删除

复制

Aa

搜索关键词

	源字段	目标字段	源字段序号	是否输出	数据类型	格式	对齐方式	精度	变量	描述	操作
1	is_zc_user	is_zc_user	1	是	string		left	0		请输入	删除
2	user_num	user_num	2	是	float		left	0		请输入	删除

取消

保存

[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。

调试器

基本信息

问题

输入

输出

日志

out

选择输出列

	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。

调试器	基本信息	问题	输入	输出	日志
out	▼	选择输出列			
	is_zc_user			user_num	
1	0			4586475.0	
2	1			1022.0	

[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) 剔除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）