



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

实验二（三选一）

大数据交互式 OLAP 多维分析数据方案  
-经营分析

# 实验二：大数据交互式 OLAP 多维分析数据方案-经营分析

## 一、实验描述

针对关键指标数据进行经营分析统计，呈现企业当前关于客户规模、市场发展、收入指标等各项维度数据，给决策者提供各项数据支撑，其中客户规模、客户价值波动、客户活跃情况等对于运营商是非常重要的运营指标。

## 二、实验目的

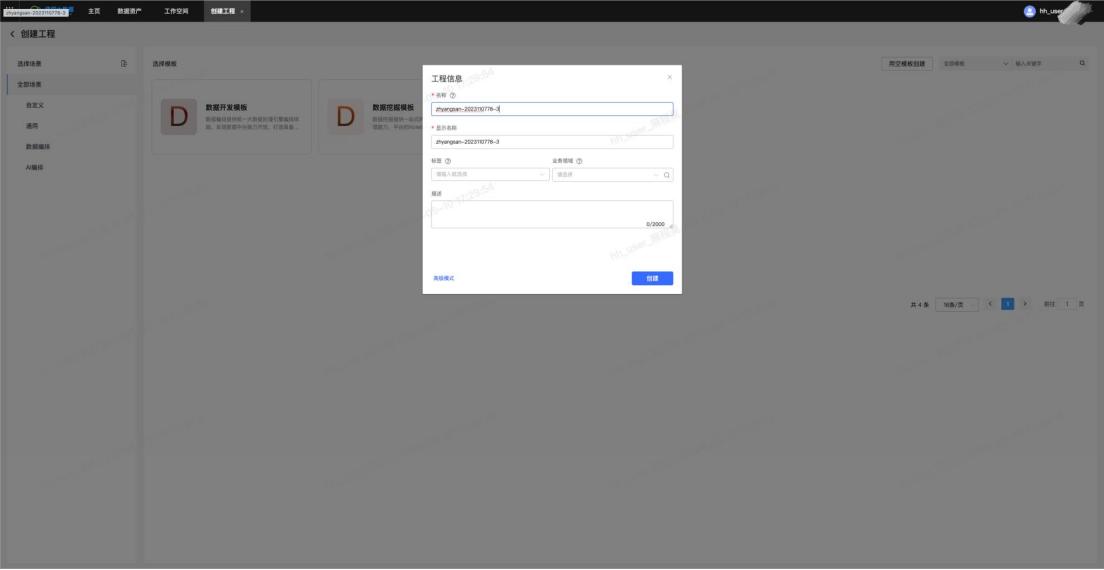
该案例以客户活跃情况为例，监控用户的 DOU（单位为 MB）和 MOU（单位为 min）波动情况。

## 三、实验环境

梧桐·鸿鹄大数据实训 DPaaS 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。



### 4.1 相关表格信息解释

单一用户业务量汇总表”：该表记录了每个用户在不同月份的消费情况：

序号	字段属性	字段名称	数据类型
1	RECORD_NUM	记录行号	STRING
2	STATIS_DATE	数据日期	STRING
3	MSISDN	手机号	STRING
4	PROV_NO	省份标识	STRING
5	CITY_CODE	地市编码	STRING
6	THIS_ACCT_FEE_TAX	本期出账金额（税前）	DECIMAL
7	ACCT_BAL_FEE	账户余额	DECIMAL

8	GPRS_TOTAL_FLUX	GPRS 总流量	DECIMAL
9	VOICE_DURA	通话时长	DECIMAL
10	VOICE_DAYS	本期通话天数	DECIMAL
11	GPRS_2G_FLUX_UP	GPRS 上行流量 (2G)	DECIMAL
12	GPRS_2G_FLUX_DOWN	GPRS 下行流量 (2G)	DECIMAL
13	GPRS_3G_FLUX_UP	GPRS 上行流量 (3G)	DECIMAL
14	GPRS_3G_FLUX_DOWN	GPRS 下行流量 (3G)	DECIMAL
15	GPRS_4G_FLUX_UP	GPRS 上行流量 (4G)	DECIMAL
16	GPRS_4G_FLUX_DOWN	GPRS 下行流量 (4G)	DECIMAL
17	ROAM_SJ_JF_TIMES	省际漫游计费时长	DECIMAL
18	ROAM_GJ_JF_TIMES	国际漫游计费时长	DECIMAL
19	STATIS_YM	统计月份	STRING
20	PROV_ID	省份编码	STRING

按月统计 DOU 和 MOU，体现 MOU 和 DOU 的波动趋势，样例数据统计结果如下表所示。

月份	DOU	MOU
202204	14619.15	19.96
202205	14706.02	20.61
202206	14196.98	19.94

按地市进行细分，体现不同地市的 MOU 和 DOU 的月波动趋势，样例数据统计结果如下表所示。

月份	地市	MOU	DOU
202204	10601	18.92	13913.23
202205	10601	19.39	13791.84
202206	10601	18.87	13512.74
202204	10602	21.02	15556.18
202205	10602	22.17	16123.45
202206	10602	21.14	15246.85
202204	10603	19.76	14548.75
202205	10603	21.23	14751.32
202206	10603	20.32	13961.66
202204	10604	21.01	13374.58
202205	10604	21.71	13541.86
202206	10604	20.84	12737.49
202204	10605	19.72	13053.23
202205	10605	20.59	13600.15
202206	10605	19.71	12571.37
202204	10606	20.26	13639.21
202205	10606	21.02	13943.66
202206	10606	20.21	13377.40
202204	10607	20.33	12562.56
202205	10607	21.05	12882.05
202206	10607	20.38	12484.34
202204	10608	21.57	15019.78
202205	10608	22.33	15410.27
202206	10608	21.51	14698.06
202204	10609	22.02	18261.16
202205	10609	22.79	18305.60

依据客户规模，对具体月份、MOU 和 DOU 进行分层统计，了解 DOU 和 MOU 的客户分层

及环比波动情况，样例数据统计结果如下表所示。

2022 年 5 月 DOU 客户分层：

DOU 分层	客户数	环比
0GB	15804	0.43%
0GB~5GB	58235	-1.03%
5GB~10GB	27623	-0.90%
10GB~15GB	17085	2.56%
15GB~20GB	11507	2.69%
20GB 以上	37072	0.05%

2022 年 6 月 DOU 客户分层：

DOU 分层	客户数	环比
0GB	16732	5.87%
0GB~5GB	59516	2.20%
5GB~10GB	27546	-0.28%
10GB~15GB	16447	-3.73%
15GB~20GB	11002	-4.39%
20GB 以上	36065	-2.72%

2022 年 5 月 MOU 客户分层：

MOU 分层	客户数	环比
0min	15804	0.43%
0min~20min	45337	-5.26%
20min~30min	56221	-15.66%
30min 以上	49964	34.61%

2022 年 6 月 MOU 客户分层：

MOU 分层	客户数	环比
0min	16731	5.87%
0min~20min	46823	3.28%
20min~30min	66784	18.79%
30min 以上	36970	-26.01%

## 4.2 数据准备：

创建工程，工程作为基本管理单元可进行编排开发和数据模型管理。在数据编排工具首页，单击创建工程按钮，如下图所示，然后在弹出的“工程信息”对话框中输入工程相关信息，完成工程创建。（工程名称为姓名+学号+实验序号，详见开头截图示例）

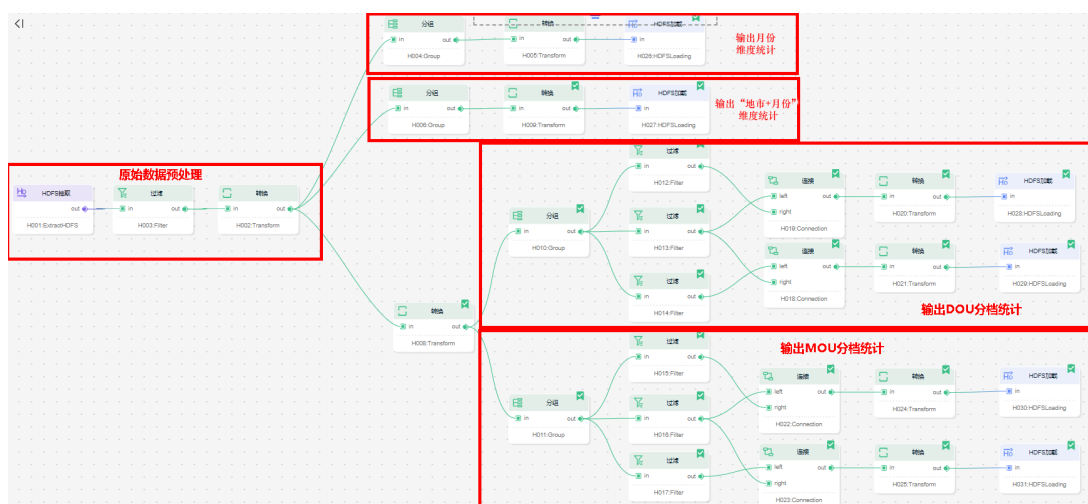




[给出工程创建后的截图，要求能看到实验名称]

### 4.3 数据流程编排：

通过图形化界面按钮进行数据加工。打开创建的工程，在导航栏单击“数据处理”进入数据处理页面，单击批处理下的“数据流”，然后在右侧单击“新建流程”，输入名称完成数据流的创建。在数据流画布中进行算子的编排，编排包括 3 个阶段：第一个阶段是抽取数据，即从 HDFS 中抽取本案例需要的数据到编排的数据流中；第二个阶段是进行数据的处理，即根据实际需要进行表关联或字段计算统计等；第三个阶段是将处理后的数据加载成文件并存放 HDFS 中。实践案例的详细编排结果如下图所示。



说明：本案例输入数据的 HDFS 路径为 /tmp/wutong/example\_data/，该路径下的 TO\_M\_CUST\_86005 表示单一用户业务量汇总月表。

完成数据处理后，将月份维度统计结果、“月份+地市”维度统计结果、DOU 客户分档统

计结果和 MOU 客户分档统计结果加载为 CSV 文件，分别命名为姓名缩写\_mou\_dou\_data\_m.csv、姓名缩写\_mou\_dou\_city\_data\_m.csv、姓名缩写\_dou\_level\_data\_05\_m.csv、姓名缩写\_dou\_level\_data\_06\_m.csv 和姓名缩写\_mou\_level\_data\_05\_m.csv、姓名缩写\_mou\_level\_data\_06\_m.csv，最后将其存储到 HDFS 中，存储路径为实训活动中各自对应的编排输出路径。

以“单一用户业务量汇总表”为基础，通过数据计算，可以得到 6 组不同的统计结果，如上图所示。

### 步骤一：

编制公共使用的“原始数据预处理”数据流程，对源数据进行初步加工，得到后续分析需要使用的数据指标，如下图所示。



第一个算子是 HDFS 抽取算子，负责把源数据抽取到数据流程中，用于后续计算。这里需要配置物理模型，从“物理模型”下拉列表框中选择 TO\_M\_CUST\_86005，如下图所示。

编辑 HDFS抽取

基础配置 输出列

\* 数据源名称: datacube\_source

物理模型: TO\_M\_CUST\_86005

业务领域: 请输入

\* 文件路径: /tmp/wutong/example\_data/

\* 文件名: TO\_M\_CUST\_86005.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

\* 分隔符: €€

文件压缩类型: 未压缩 .gz .snappy

自动适配列名文件: 是 否

取消 保存

完成数据源加载后，使用过滤算子选择后续需要用到的数据。

编辑 HDFS抽取

基础配置 **输出列**

增加 1

	输入名称	输出名称	是否输出	数据类型	格式	最小值	最大值	默认值	表达式	长度	允许为空	检查	逻辑策略	密文长度
1	RECORD_N...	RECORD_N...	<input type="checkbox"/>	string	请选择						拒绝	关闭		
2	STATIS_DATE	STATIS_DATE	<input checked="" type="checkbox"/>	string	请选择					请输入	拒绝	关闭		
3	MSISDN	MSISDN	<input checked="" type="checkbox"/>	string	请选择					请输入	拒绝	关闭		
4	PROV_NO	PROV_NO	<input checked="" type="checkbox"/>	string	请选择					请输入	拒绝	关闭		
5	CITY_CODE	CITY_CODE	<input type="checkbox"/>	string	请选择						拒绝	关闭		
6	THIS_ACCT...	THIS_ACCT...	<input type="checkbox"/>	double	请选择						拒绝	关闭		
7	ACCT_BAL...	ACCT_BAL...	<input type="checkbox"/>	double	请选择						拒绝	关闭		
8	GPRS_TOTA...	GPRS_TOTA...	<input checked="" type="checkbox"/>	double	请选择						拒绝	关闭		
9	VOICE_DURA...	VOICE_DURA...	<input checked="" type="checkbox"/>	double	请选择						拒绝	关闭		
10	VOICE_DAYS	VOICE_DAYS	<input type="checkbox"/>	double	请选择						拒绝	关闭		
11	GPRS_2G_F...	GPRS_2G_F...	<input type="checkbox"/>	double	请选择						拒绝	关闭		

共 20 条 20条/页 < < 1/1 > >

取消 保存

该案例以记录量适中的省份“106”为例，选取 2022 年 4~6 月的数据。数据源中 PROV\_NO 字段表示地市级编码，前三位代表所属省份。过滤公式如下。

`substr(PROV_NO,0,3)== 106 and (STATIS_DATE ==202204 or STATIS_DATE == 202205 or STATIS_DATE ==202206)`

最后使用转换算子，将从源数据加载的日周期转换成月周期，处理空值数据，便于后续计算，如下图所示。

编辑 转换

输出列

增加 1

	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	STATIS_DATE	STATIS_DATE	否	string		原有字段输出	STATIS_DATE	删除
2	MSISDN	MSISDN	是	string		原有字段输出	MSISDN	删除
3	PROV_NO	PROV_NO	是	string		原有字段输出	PROV_NO	删除
4	GPRS_TOTAL_FLUX	GPRS_TOTAL_FL...	否	float		原有字段输出	GPRS_TOTAL...	删除
5	VOICE_DURA...	VOICE_DURA...	否	float		原有字段输出	VOICE_DURA...	删除
6		STATIS_MONTH	是	string		表达式计算	substr( STAT...	删除
7		GPRS_TOTAL_FL...	是	float		表达式计算	nvl(GPRS_TO...	删除
8		VOICE_DURA	是	float		表达式计算	nvl( VOICE...	删除

[给出处理的结果截图，要求能够看到用户个人信息以及时间]

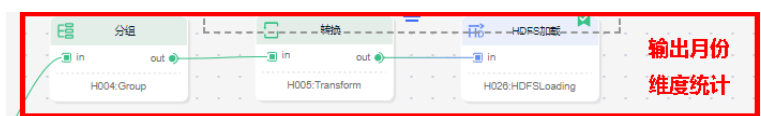
“日期转换”表达式：substr(STATIS\_DATE,0,6)，将不确定格式的日周期转换为月周期格式。

“流量空值转换”表达式：nvl(GPRS\_TOTAL\_FLUX,0)，将流量空值记录转换为 0。

“语音空值转换”表达式：nvl(VOICE\_DURA,0)，将语音空值记录转换为 0。

## 步骤二：

以月份为维度分析用户 DOU 和 MOU 波动趋势，通过数据流程转换输出统计结果，如下图所示。



首先使用分组算子按照月份进行分组，统计“当月用户量”“流量使用总值”和“语音使用

总值”3 个指标，如下图所示。

编辑 分组

算法

自动 哈希 预排序 Map-Reduce Hash-Map-Reduce

用于指定计算时使用的算法。

分组合段

输入名称

MSISDN PROV\_NO **STATIS\_MONTH** GPRS\_TOTAL\_FLUX VOICE\_DURA

数据类型 string string string float float

汇总计算

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	STATIS_MONTH	STATIS_MONTH	是	string		原有字段输出	STATIS_MO...	删除
2		USER_NUM	是	float		统计方法	Count( MSIS...	删除
3		GPRS_TOTAL_FL...	是	float		统计方法	Sum( GPRS_...	删除
4		VOICE_DURA	是	float		统计方法	Sum( VOICE...	删除

“当月用户量”表达式：Count(MISSDN)。  
“流量使用总值”表达式：SUM(GPRS\_TOTAL\_FLUX)。  
“语音使用总值”表达式：SUM(VOICE\_DURA)。  
然后使用转换算子，计算当月 DOU 和 MOU，如下图所示。

编辑 转换

输出列

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	STATIS_MONTH	STATIS_MONTH	是	string		原有字段输出	STATIS_MO...	删除
2	USER_NUM	USER_NUM	否	float		原有字段输出	USER_NUM	删除
3	GPRS_TOTAL_FLUX	GPRS_TOTAL_FL...	否	float		原有字段输出	GPRS_TOTAL...	删除
4	VOICE_DURA	VOICE_DURA	否	float		原有字段输出	VOICE_DURA	删除
5		DOU	是	float		表达式计算	GPRS_TOTAL...	删除
6		MOU	是	float		表达式计算	VOICE_DUR...	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

DOU 表达式：GPRS\_TOTAL\_FLUX/USER\_NUM。  
MOU 表达式：VOICE\_DURA/USER\_NUM。  
最后，使用 HDFS 加载算子，将计算结果加载到 HDFS 的实训中各自的编排输出路径中（图中文件路径仅为示例），如下图所示。





编辑 分组

✕

算法

自动

哈希

预排序

Map-Reduce

Hash-Map-Reduce

用于指定计算时使用的算法。

分組字段

输入名称

数据类型

<input checked="" type="checkbox"/>	PROV_NO	string
<input type="checkbox"/>	MSISDN	string
<input checked="" type="checkbox"/>	STATIS_MONTH	string
<input type="checkbox"/>	GPRS_TOTAL_FLUX	float
<input type="checkbox"/>	VOICE_DURA	float

汇总计算

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"/>	PROV_NO	PROV_NO	是	string		原有字段输出	PROV_NO	<a href="#">删除</a>
2	<input type="checkbox"/>	STATIS_MONTH	STATIS_MONTH	是	string		原有字段输出	STATIS_MO...	<a href="#">删除</a>
3	<input type="checkbox"/>		USER_NUM	是	float		统计方法	Count( MSIS...	<a href="#">删除</a>
4	<input type="checkbox"/>		GPRS_TOTAL_FL...	是	float		统计方法	Sum( GPRS ...	<a href="#">删除</a>
5	<input type="checkbox"/>		VOICE_DURA	是	float		统计方法	Sum( VOICE...	<a href="#">删除</a>

“分地市分月用户量”表达式：Count(MSISDN)。

“分地市分月流量使用总值”表达式：Sum(GPRS\_TOTAL\_FLUX)。

“分地市分月语音使用总值”表达式：Sum(VOICE\_DURA)。

然后使用转换算子，计算分月分地市 DOU 和分月分地市 MOU 数值，如下图所示。

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"/>	PROV_NO	PROV_NO	是	string		原有字段输出	PROV_NO	<a href="#">删除</a>
2	<input type="checkbox"/>	STATIS_MONTH	STATIS_MONTH	是	string		原有字段输出	STATIS_MO...	<a href="#">删除</a>
3	<input type="checkbox"/>	USER_NUM	USER_NUM	否	float		原有字段输出	USER_NUM	<a href="#">删除</a>
4	<input type="checkbox"/>	GPRS_TOTAL_FLUX	GPRS_TOTAL_FL...	否	float		原有字段输出	GPRS_TOTAL...	<a href="#">删除</a>
5	<input type="checkbox"/>	VOICE_DURA	VOICE_DURA	否	float		原有字段输出	VOICE_DURA	<a href="#">删除</a>
6	<input type="checkbox"/>		DOU	是	float		表达式计算	GPRS_TOTAL...	<a href="#">删除</a>
7	<input type="checkbox"/>		MOU	是	float		表达式计算	VOICE_DUR...	<a href="#">删除</a>

分月分地市 DOU 表达式：GPRS\_TOTAL\_FLUX/USER\_NUM。

分月分地市 MOU 表达式：VOICE\_DURA/USER\_NUM。

最后，使用 HDFS 加载算子，将计算结果加载到 HDFS 的实训中各自的编排输出路径中（图中文件路径仅为示例），如下图所示，

编辑 HDFS加载

✕

基础配置

输出列

数据源名称

DATA CUBE\_HADOOP\_DS\_1

若没有所属集群，请点击[这里](#) 创建

物理模型

请选择物理模型

若没有所属物理模型（表名），请点击[这里](#) 创建

文件路径

/user/vutong/example\_data/

文件名

mou\_dou\_city\_data\_m.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

加载结果如下图所示。

编辑 HDFS加载 ②

基础配置 输出列

按名称映射 增加 1

搜索关键词

	<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"/>	PROV_NO	PROV_NO	1	是	string		left	0		请输入	删除
2	<input type="checkbox"/>	STATIS_MO...	STATIS_MO...	2	是	string		left	0		请输入	删除
3	<input type="checkbox"/>	DOU	DOU	3	是	float		left	0		请输入	删除
4	<input type="checkbox"/>	MOU	MOU	4	是	float		left	0		请输入	删除

[给出配置截图以及加载结果截图，要求截图可以显示出个人信息以及时间]

步骤四：

使用转换算子对公共部分的计算结果再次加工，以便用于后续计算，如下图所示。



在转换算子中增加 4 个指标，使用表达式计算输出结果，如下图所示。

编辑 转换 ②

输出列

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"/>	MSISDN	MSISDN	是	string		原有字段输出	MSISDN	删除
2	<input type="checkbox"/>	PROV_NO	PROV_NO	是	string		原有字段输出	PROV_NO	删除
3	<input type="checkbox"/>	GPRS_TOTAL_FLUX	GPRS_TOTAL_FL...	否	float		原有字段输出	GPRS_TOTAL...	删除
4	<input type="checkbox"/>	VOICE_DURA	VOICE_DURA	否	float		原有字段输出	VOICE_DURA	删除
5	<input type="checkbox"/>	STATIS_MONTH	STATIS_MONTH	是	string		原有字段输出	STATIS_MO...	删除
6	<input type="checkbox"/>		DOU_LEVEL	是	string		表达式计算	case when G...	删除
7	<input type="checkbox"/>		MOU_LEVLE	是	string		表达式计算	case when V...	删除

各用户 DOU 分档编码和名称计算公式如下图所示。

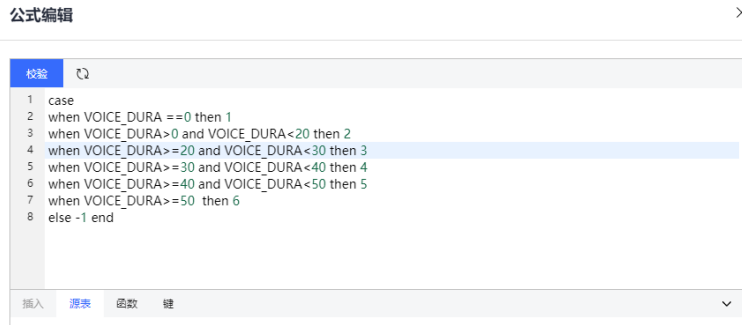
公式编辑

校验

```
1 case
2 when GPRS_TOTAL_FLUX>=0 then 1
3 when GPRS_TOTAL_FLUX>0 and GPRS_TOTAL_FLUX<5000 then 2
4 when GPRS_TOTAL_FLUX>=5000 and GPRS_TOTAL_FLUX<10000 then 3
5 when GPRS_TOTAL_FLUX>=10000 and GPRS_TOTAL_FLUX<15000 then 4
6 when GPRS_TOTAL_FLUX>=15000 and GPRS_TOTAL_FLUX<20000 then 5
7 when GPRS_TOTAL_FLUX>=20000 then 6
8 else -1 end
```

插入 函数 键

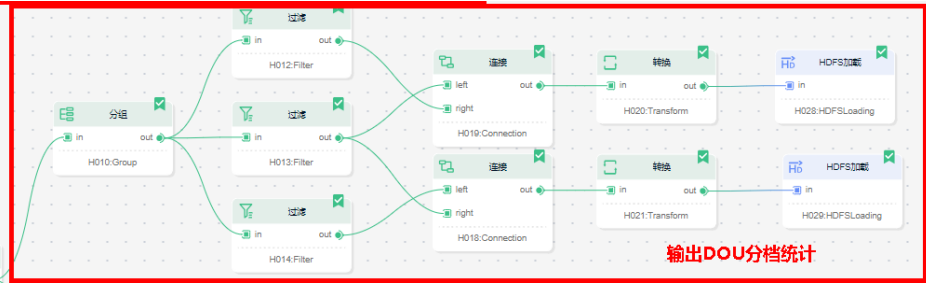
各用户 MOU 分档编码和名称计算公式如下图所示。



[给出 DOU 和 MOU 公式编辑截图]

步骤五：

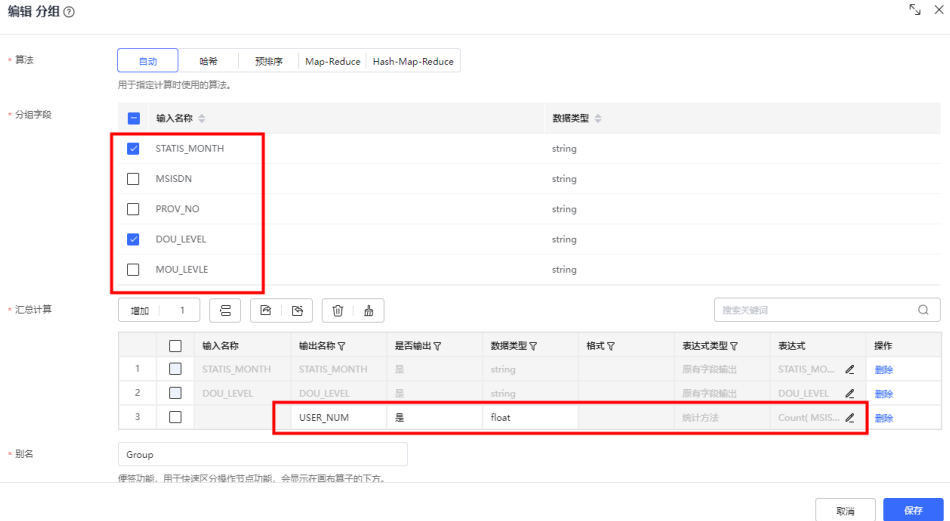
以公共转换算子的结果为输入源，计算 2022 年 5 月和 6 月 DOU 分档相关统计数据，整体流程如下图所示。



首先使用分组算子，以“月份+DOU”分档编码、名称为维度进行分组，计算“分月各分档用户数量”，如图 4-6-22 所示。

“分月各分档用户数量”表达式：Count(MSISON)。

通过 3 个过滤算子分别过滤输出 202204 周期、202205 周期和 202206 周期的各月数据，202204 周期的表达式为 STATIS\_MONTH==202204，同理 202205 周期的表达式为 STATIS\_MONTH==202205 和 202206 周期的表达式为 STATIS\_MONTH==202206。



然后编辑连接算子，将前后两个月的数据以 DOU 分档关联合并，将本月和上月的统计结果同时展现，如下图所示。

编辑连接-基础配置（1）：

编辑连接

基础配置

输出列

主数据源

left

right

选择为主数据源的数据集字段会显示在映射关系左侧。

算法

自动

哈希

预排序

Map-Reduce

Hash-Map-Reduce

用于指定计算时使用的算法。

映射关系

与输入'right'映射

按名称映射

新增映射

1

主输入'left'		副输入'right'		
列名称	数据类型	列名称	数据类型	操作
DOU_LEVEL	string	DOU_LEVEL	string	删除

连接类型

内连接

左外连接

右外连接

外连接

选择需要输出字段的连接条件。

别名

Connection

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。

取消

保存

编辑连接-输出列（1）：

编辑连接

基础配置

输出列

搜索关键词

	是否输出	输出名称	组名	输入名称	数据类型	格式
1	是	DOU_LEVEL	left	DOU_LEVEL	string	
2	是	USER_NUM_05	left	USER_NUM	float	
3	否	DOU_LEVEL	right	DOU_LEVEL	string	
4	是	USER_NUM_04	right	USER_NUM	float	

编辑连接

基础配置

输出列

主数据源

left

right

选择为主数据源的数据集字段会显示在映射关系左侧。

算法

自动

哈希

预排序

Map-Reduce

Hash-Map-Reduce

用于指定计算时使用的算法。

映射关系

与输入'right'映射

按名称映射

新增映射

1

主输入'left'		副输入'right'		
列名称	数据类型	列名称	数据类型	操作
DOU_LEVEL	string	DOU_LEVEL	string	删除

连接类型

内连接

左外连接

右外连接

外连接

选择需要输出字段的连接条件。

别名

Connection

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。

取消

保存

编辑连接-输出列（2）

编辑连接

基础配置

输出列

搜索关键词

	是否输出	输出名称	组名	输入名称	数据类型	格式
1	是	DOU_LEVEL	left	DOU_LEVEL	string	
2	是	USER_NUM_06	left	USER_NUM	float	
3	否	DOU_LEVEL	right	DOU_LEVEL	string	
4	是	USER_NUM_05	right	USER_NUM	float	

关联合并后，使用转换算子，输出“DOU 分档用户量环比”指标。  
输出“DOU 分档用户量环比”（1）：

编辑 转换

输出列

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"/>	DOU_LEVEL	DOU_LEVEL	是	string		原有字段输出	DOU_LEVEL	删除
2	<input type="checkbox"/>	USER_NUM_05	USER_NUM_05	是	float		原有字段输出	USER_NUM_...	删除
3	<input type="checkbox"/>	USER_NUM_04	USER_NUM_04	否	float		原有字段输出	USER_NUM_...	删除
4	<input type="checkbox"/>		USER_NUM_RATE	是	float		表达式计算	USER_NUM_...	删除

输出“DOU 分档用户量环比”（2）：

编辑 转换

输出列

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"/>	DOU_LEVEL	DOU_LEVEL	是	string		原有字段输出	DOU_LEVEL	删除
2	<input type="checkbox"/>	USER_NUM_06	USER_NUM_06	是	float		原有字段输出	USER_NUM_...	删除
3	<input type="checkbox"/>	USER_NUM_05	USER_NUM_05	否	float		原有字段输出	USER_NUM_...	删除
4	<input type="checkbox"/>		USER_NUM_RATE	是	float		表达式计算	USER_NUM_...	删除

2022 年 5 月和 6 月的“DOU 分档用户量环比”表达式分别为

USER\_NUM\_05/USER\_NUM\_04-1  
USER\_NUM\_06/USER\_NUM\_05-1

最后，编辑 HDFS 加载算子，将计算结果加载到 HDFS 的实训中各自的编排输出路径中（图中文件路径仅为示例），如下图所示。

编辑 HDFS 加载算子（1）：

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

dou\_level\_data\_05\_m.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

编辑 HDFS 加载算子（2）：

编辑 HDFS加载

基础配置 输出列

按名称映射

增加 1

搜索关键词

	<input type="checkbox"/>	源字段	目标字段	源字段...	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"/>	DOU_LEVEL	DOU_LEVEL	1	是	string		left	0		源输入	删除
2	<input type="checkbox"/>	USER_NUM_05	USER_NUM_...	2	是	float		left	0		源输入	删除
3	<input type="checkbox"/>	USER_NUM_RATE	USER_NUM_...	3	是	float		left	0		源输入	删除

编辑 HDFS 加载算子（3）

编辑 HDFS加载

基础配置

输出列

数据源名称

DATAcube\_HADOOP\_DS\_1

若没有所需集群，请点击[这里](#) 创建

物理模型

请输入或选择物理模型

若没有所需物理模型（表名），请点击[这里](#) 创建

文件路径

/user/vutong/example\_data/

文件名

dou\_level\_data\_06\_m.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不为textname.txt,则生成多文件的文件格式为textname.txt-0,textname.txt-1,textname.txt-2...

取消

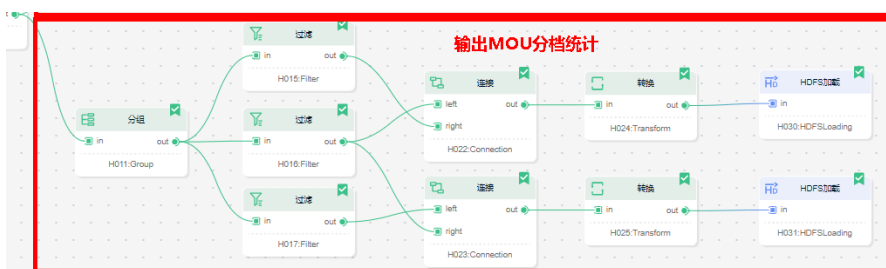
保存

**编辑 HDFS加载**

**基础配置    输出列**

名称映射	增加	1								Aa	搜索关键词	Q
<input type="checkbox"/>	源字段	目标字段	源字段...	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"/> DOU_LEVEL	DOU_LEVEL	1	是	string		left	0		读输入	<a href="#">删除</a>	
2	<input type="checkbox"/> USER_NUM_06	USER_NUM_...	2	是	float		left	0		读输入	<a href="#">删除</a>	
3	<input type="checkbox"/> USER_NUM_RATE	USER_NUM_...	3	是	float		left	0		读输入	<a href="#">删除</a>	

### 步骤六：



编辑 分组

算法

自动 哈希 预排序 Map-Reduce Hash-Map-Reduce

用于指定计算时使用的算法。

分组长段

输入名称

STATIS\_MONTH

MSISDN

PROV\_NO

DOU\_LEVEL

MOU\_LEVEL

string

string

string

string

string

汇总计算

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	STATIS_MONTH	STATIS_MONTH	是	string		原存字段输出	STATIS_MO...	删除
2	MOU_LEVEL	MOU_LEVEL	是	string		原存字段输出	MOU_LEVEL	删除
3		USER_NUM	是	float		统计方法	Count(MSIS...	删除

别名

Group

便捷功能。用于快速区分操作节点功能。会显示在操作节点的下方。

取消

保存

“分月各分档用户数量”表达式：Count(MSISDN)。

通过 3 个过滤算子分别过滤输出 202204 周期、202205 周期和 202206 周期的各月数据，例如 202204 周期的表达式为 STATIS\_MONTH==202204。

然后编辑连接算子，将前后两个月的数据以 MOU 分档关联合并，将本月和上月的统计结果同时展现，如下图所示。

编辑连接算子（1）：

编辑 连接

基础配置

输出列

主数据源

left right

选择为主数据源的数据表字段会显示在映射关系左侧。

算法

自动 哈希 预排序 Map-Reduce Hash-Map-Reduce

用于指定计算时使用的算法。

映射关系

与输入'right'映射

按名称映射 新增映射 1

主输入'left'	数据类型	副输入'right'	数据类型	操作
MOU_LEVEL	string	MOU_LEVEL	string	删除

连接类型

内连接 左外连接 右外连接 外连接

选择需要输出字段的连接条件。

别名

Connection

便捷功能。用于快速区分操作节点功能。会显示在操作节点的下方。

取消

保存

编辑连接算子（2）

编辑 连接

基础配置

输出列

是否输出

输出名称

别名

输入名称

数据类型

格式

1	是	MOU_LEVEL	left	MOU_LEVEL	string	
2	是	USER_NUM_05	left	USER_NUM	float	
3	否	MOU_LEVEL	right	MOU_LEVEL	string	
4	是	USER_NUM_04	right	USER_NUM	float	

编辑连接算子（3）



编辑连接

基础配置 输出列

主数据源

left right

选择为主数据源的数据源字段会显示在映射关系左侧。

算法

自动 哈希 预排序 Map-Reduce Hash-Map-Reduce

用于指定计算时使用的算法。

映射关系

与输入'right'映射

添加映射 新增映射 1

主输入'left'	数据源名称	数据类型	列名称	数据类型	操作
MOU_LEVLE	string	MOU_LEVLE	string	删除	

连接类型

内连接 左外连接 右外连接 外连接

选择需要输出字段的连接条件。

别名

Connection

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。

取消

保存

## 编辑连接算子（4）

编辑连接

基础配置 输出列

是否输出	输出名称	别名	输入名称	数据类型	格式
是	MOU_LEVLE	left	MOU_LEVLE	string	
是	USER_NUM_05	left	USER_NUM	float	
否	MOU_LEVLE	right	MOU_LEVLE	string	
是	USER_NUM_05	right	USER_NUM	float	

关联合并后，编辑转换算子，输出“MOU 分档用户量环比”指标，如下图所示：

编辑转换

输出列

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	MOU_LEVLE	MOU_LEVLE	是	string		原有字段输出	MOU_LEVLE	删除
2	USER_NUM_05	USER_NUM_05	是	float		原有字段输出	USER_NUM_...	删除
3	USER_NUM_04	USER_NUM_04	否	string		原有字段输出	USER_NUM_...	删除
4		USER_NUM_RATE	是	float		表达式计算	USER_NUM_...	删除

别名 Transform

结果如下：

编辑转换

输出列

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	MOU_LEVLE	MOU_LEVLE	是	string		原有字段输出	MOU_LEVLE	删除
2	USER_NUM_06	USER_NUM_06	是	float		原有字段输出	USER_NUM_...	删除
3	USER_NUM_05	USER_NUM_05	否	string		原有字段输出	USER_NUM_...	删除
4		USER_NUM_RA...	是	float		表达式计算	USER_NUM_...	删除

别名 Transform

描述 请输入描述信息,不能包含字母[]>

2022 年 5 月和 6 月的“MOU 分档用户量环比”表达式分别为

USER\_NUM\_05/USER\_NUM\_04-1  
USER\_NUM\_06/USER\_NUM\_05-1

最后，编辑 HDFS 加载算子，将计算结果加载到 HDFS 的实训中各自的编排输出路径中（图中文件路径仅为示例），如下图所示。

编辑 HDFS 加载算子（1）：

编辑 HDFS加载

基础配置

输出列

数据源名称

DATAcube\_HADOOP\_DS\_1

还没有所属集群，请点击这里 创建

物理模型

请输入或选择物理模型

还没有所属物理模型（表名），请点击这里 创建

文件路径

/user/wutong/example\_data/

文件名

mou\_level\_data\_05\_m.csv

业务领域

请输入

文件编码

UTF-8

ASCII

ISO-8859-1

GB18030

GBK

文件模式

列分隔符

名称+值对

增长字符串

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texname.txt，则生成多文件的文件名格式为texname.txt-0,texname.txt-1,texname.txt-2...

编辑 HDFS加载

基础配置

输出列

按名称添加

增加1

Aa

搜索关键字

	<input type="checkbox"/>	源字段	目标字段	源字段...	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"/>	MOU_LEVLE	MOU_LEVLE	1	是	string		left	0		读输入	删除
2	<input type="checkbox"/>	USER_NUM_05	USER_NUM...	2	是	float		left	0		读输入	删除
3	<input type="checkbox"/>	USER_NUM_RATE	USER_NUM...	3	是	float		left	0		读输入	删除

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

mov\_level\_data\_06\_m.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

固定字符串

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。当配置加载文件名时,则生成多文件的文件名格式为:texname.txt,0,texname.txt-1,texname.txt-2,...

取消

保存

编辑 HDFS加载 









基础配置

输出列

综合名称映射


增加





1



Aa

搜索关键词



	<input type="checkbox"/>	源字段	目标字段 	源字段...	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"/>	MOU_LEVLE	MOU_LEVLE	1	是	string		left	0		该输入	<a href="#">删除</a>
2	<input type="checkbox"/>	USER_NUM_06	USER_NUM...	2	是	float		left	0		该输入	<a href="#">删除</a>
3	<input type="checkbox"/>	USER_NUM_RATE	USER_NUM...	3	是	float		left	0		该输入	<a href="#">删除</a>

## 4.4 数据输出

数据编排完成后，点击“保存”进行编辑参数，队列名选择 compute：

编辑参数 ② 请谨慎填入合理范围的值，避免引入安全风险

Hadoop集群

datacube\_source

队列名

midteant02\_dev\_compute

SparkSQL 模式

是

否

默认

如果选择“是”按钮,那么将启用SparkSQL模式;如果选择“否”按钮,那么将禁用SparkSQL模式;如果选择“默认”按钮,那么该配置项将由全局参数控制

增加

1

删除

全删

<input type="checkbox"/>	名称	描述	默认值
<input type="checkbox"/>	runType		Spark
<input type="checkbox"/>	bdi_yarnjob_hadoopclusterId		datacube_source datacube_source
<input type="checkbox"/>	mapreduce_job_queueName		midteant02_dev_compute
<input type="checkbox"/>	is_sql_supported		default

取消

保存

然后进入在线调试，如下图所示：



依次查看各个 HDFS 加载算子的输出结果。

步骤一：

以月为维度统计 DOU 和 MOU，结果如下图所示。

调试器 基本信息 问题 输入 输出 日志			
out		选择输出列	
STATIS_MONTH		DOU	MOU
1	202206	14196.984	19.943148
2	202204	14619.145	19.957817
3	202205	14706.0205	20.605644

步骤二：

以“地市+月份”维度统计 MOU 和 DOU，结果如下图所示。

调试器					基本信息					问题					输入					输出					日志				
out					选择输出列																				搜索				
PROV_NO					STATIS_MONTH					DOU					MOU														
1	10608				202205					15410.274					22.33382														
2	10611				202206					17300.48					22.006243														
3	10601				202204					13913.229					18.918911														
4	10602				202204					15556.177					21.017832														
5	10606				202206					13377.396					20.205036														
6	10615				202205					12557.642					19.323383														
7	10606				202204					13639.214					20.262686														
8	10613				202204					19082.93					22.338427														
9	10613				202206					18491.95					22.38001														
10	10616				202205					13108.355					21.529284														
11	10605				202205					13600.154					20.593449														
12	10607				202205					12882.054					21.048367														
13	10611				202204					17882.44					22.01105														
14	10612				202206					15224.626					21.408491														
15	10615				202206					11520.092					18.630596														
16	10614				202206					12803.001					20.570715														
17	10612				202204					16223.888					21.312284														
18	10601				202205					13791.837					19.390757														
19	10603				202205					14751.316					21.23116														
20	10614				202204					13572.729					20.471264														
21	10604				202205					13541.855					21.711472														
22	10615				202204					11982.392					18.559006														
23	10608				202206					14698.061					21.509699														
24	10602				202205					16123.446					22.17257														
25	10610				202205					16480.17					21.823334														
26	10606				202205					13943.661					21.019321														
27	10607				202206					12484.335					20.38242														
28	10609				202205					18305.596					22.790956														
29	10608				202204					15019.777					21.565228														
30	10611				202205					18289.258					22.867283														
31	10604				202204					13374.581					21.013334														
32	10613				202205					19277.229					23.237474														
33	10601				202206					13512.736					18.874327														
34	10603				202206					13961.655					20.321926														
35	10604				202206					12737.485					20.843773														
36	10603				202204					14548.747					19.759975														
37	10610				202206					16024.877					21.010294														
38	10614				202205					13772.192					20.8905														
39	10602				202206					15246.846					21.13669														
40	10605				202206					12571.367					19.710405														
41	10609				202204					18261.162					22.0225														
42	10605				202204					13053.233					19.721571														
43	10610				202204					16082.763					20.911406														
44	10616				202204					12741.641					20.464912														
45	10616				202206					12065.548					20.720675														
46	10609				202206					17688.033					22.051424														
47	10607				202204					12562.559					20.326656														
48	10612				202205					16508.883					22.35721														

步骤三：

对具体月份 MOU 和 DOU 客户数进行分档统计，了解 MOU 和 DOU 客户分档情况及环比波动，如下图所示。

2022 年 5 月 DOU 客户分档统计：

调试器 基本信息 问题 输入 输出 日志			
out		选择输出列	
DOU_LEVEL		USER_NUM_05	USER_NUM_RATE
1	6	37072.0	4.5883656E-4
2	3	27623.0	-0.009040356
3	4	17085.0	0.025633335
4	2	58235.0	-0.010282099
5	5	11507.0	0.026860595
6	1	15804.0	0.0043213367

2022 年 6 月 DOU 客户分档统计：

调试器 基本信息 问题 输入 输出 日志			
out		选择输出列	
DOU_LEVEL		USER_NUM_06	USER_NUM_RATE
1	5	11002.0	-0.043886304
2	1	16732.0	0.058719277
3	6	36065.0	-0.027163386
4	3	27546.0	-0.0027875304
5	4	16447.0	-0.037342727
6	2	59516.0	0.021997094

2022 年 5 月 MOU 客户分档统计：

调试器 基本信息 问题 输入 输出 日志			
out		选择输出列	
MOU_LEVLE		USER_NUM_05	USER_NUM_RATE
1	1	15804.0	0.0043213367
2	3	56221.0	-0.15662599
3	4	49964.0	0.34608543
4	2	45337.0	-0.052597463

2022 年 6 月 MOU 客户分档统计：

MOU_LEVLE		USER_NUM_06	USER_NUM_RATE
1	3	66784.0	0.1878835
2	4	36970.0	-0.26006722
3	2	46823.0	0.032776713
4	1	16731.0	0.058655977

[给出结果部分上述提到的截图，要求能够看到个人用户信息以及时间]

## 五、 实验结果

1、 提供上述截图，包括：（每张图 1 分，具体要求见指导书）

- （1）工程创建的截图
- （2）源数据处理截图
- （3）以月份为维度分析用户 DOU 和 MOU 波动趋势时 HDFS 算子配置截图
- （4）以月份为维度分析用户 DOU 和 MOU 波动趋势时结果截图
- （5）以“地市+月份”维度分析用户 DOU 和 MOU 波动趋势 HDFS 算子配置截图
- （6）以“地市+月份”维度分析用户 DOU 和 MOU 波动趋势输出结果截图
- （7）DOU 公式编辑截图

- (8) MOU 公式编辑截图
  - (9) 2022 年 5 月和 6 月 DOU 分档相关统计数据 HDFS 算子编辑截图
  - (10) 结果部分截图
- 2、 相关代码以及截图语言描述以及解释（5 分）
  - 3、 实验总结以及自己的思考（5 分）