

# 海量用户行为数据的存储与分析

黄强@数果智能

# 聊聊当前大数据的技术

**Hadoop**

**Hbase**

**列存储**

**MPP数据库**

**内存查询引擎**

**预聚合**

<https://github.com/onurakpolat/awesome-bigdata>

# 什么是用户行为?

## 一句话

用户在产品上的操作行为的记录

## 要素

时间 地点 人物 做了什么

The screenshot displays a user behavior analysis interface. The top navigation bar includes '数果智能', '图表', '数据分析', '智能分析', '场景应用', '数据管理', '管理中心', '切换项目', '测试数据', and 'admin@广东数果'. The main content area is titled '用户细查' and shows details for a user in the '浙江用户' group. The user's behavior is recorded on '2017-04-03 10:12:28'. The details include: OsScreen: 1280\*720, Operator: 联通, 网络: 3g, 国家: 中国, Media:, IP: 106.91.157.45, Extras:, EventValue:, EventScreen: 选择产品, EventLabel:, EventHour:, EventDateTime: 1491185548336, EventDate:, EventAction: 后台, Creative:, ClientDeviceVersion: 4.4.4, and ClientDeviceModel: 2014813. On the right, a '行为记录' section shows a list of actions with timestamps and descriptions, such as '09:55:31 主题系列测试0510 对焦 播放' and '10:12:28 选择产品 后台'.

# 什么是用户行为?

用户行为通常具有以下特点

- 1、用户基数大 (几十万到上亿)
- 2、高基数维度多 (用户Id、IP、SessionID、IMEI、终端ID等)
- 3、数据量大 (一天几千万到上千亿)
- 4、时序的



# 数果智能如何处理用户行为数据?

## 一、实时接入

全程以**数据流**的形式接入数据

可视化SDK



采集网关服



Kafka



数果TIndex



BinLog采集器

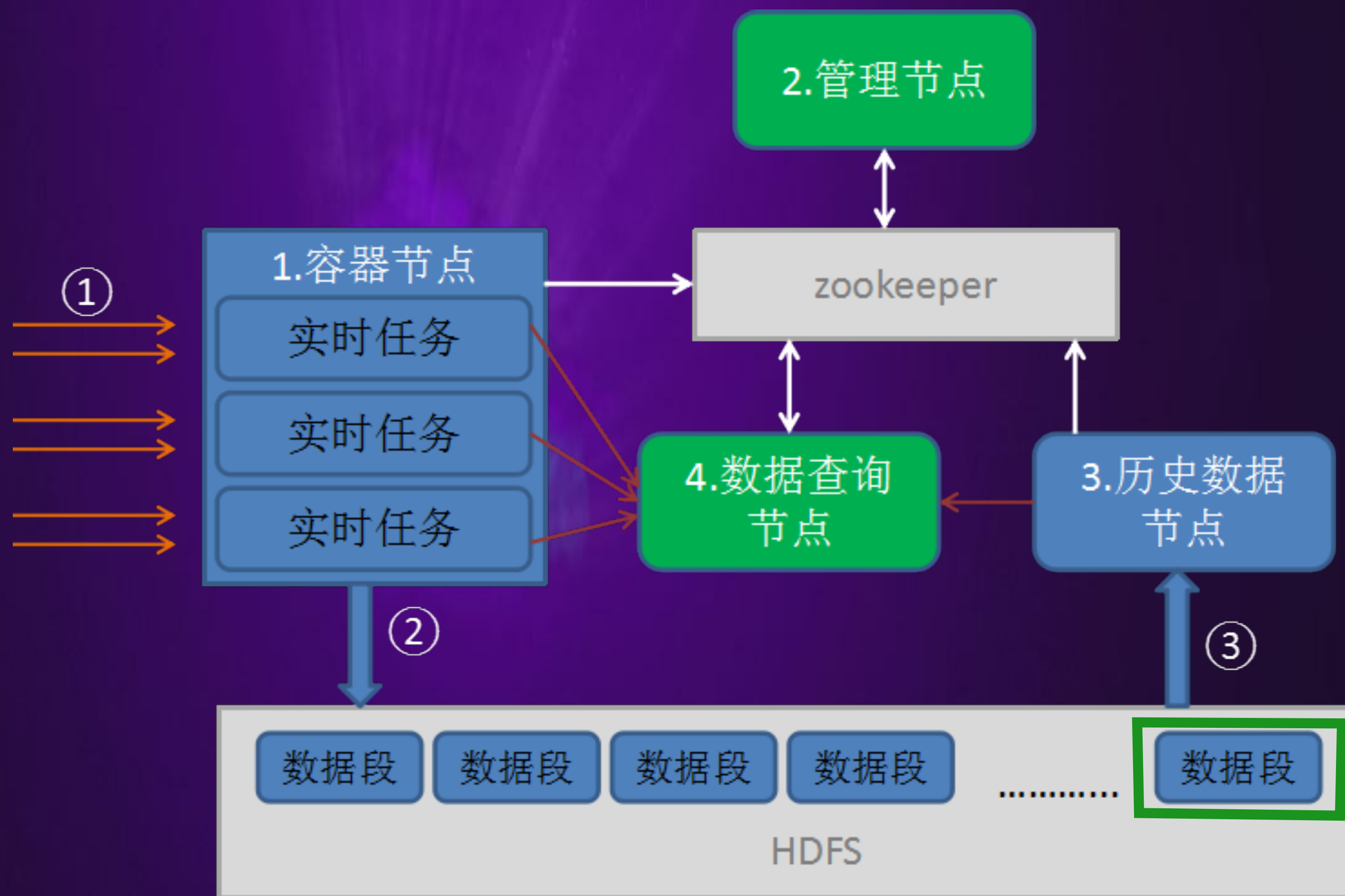


文件采集器



# 数果智能如何处理用户行为数据?

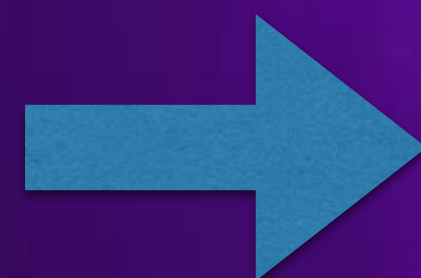
## 二、数据存储





# 数果智能如何处理用户行为数据?

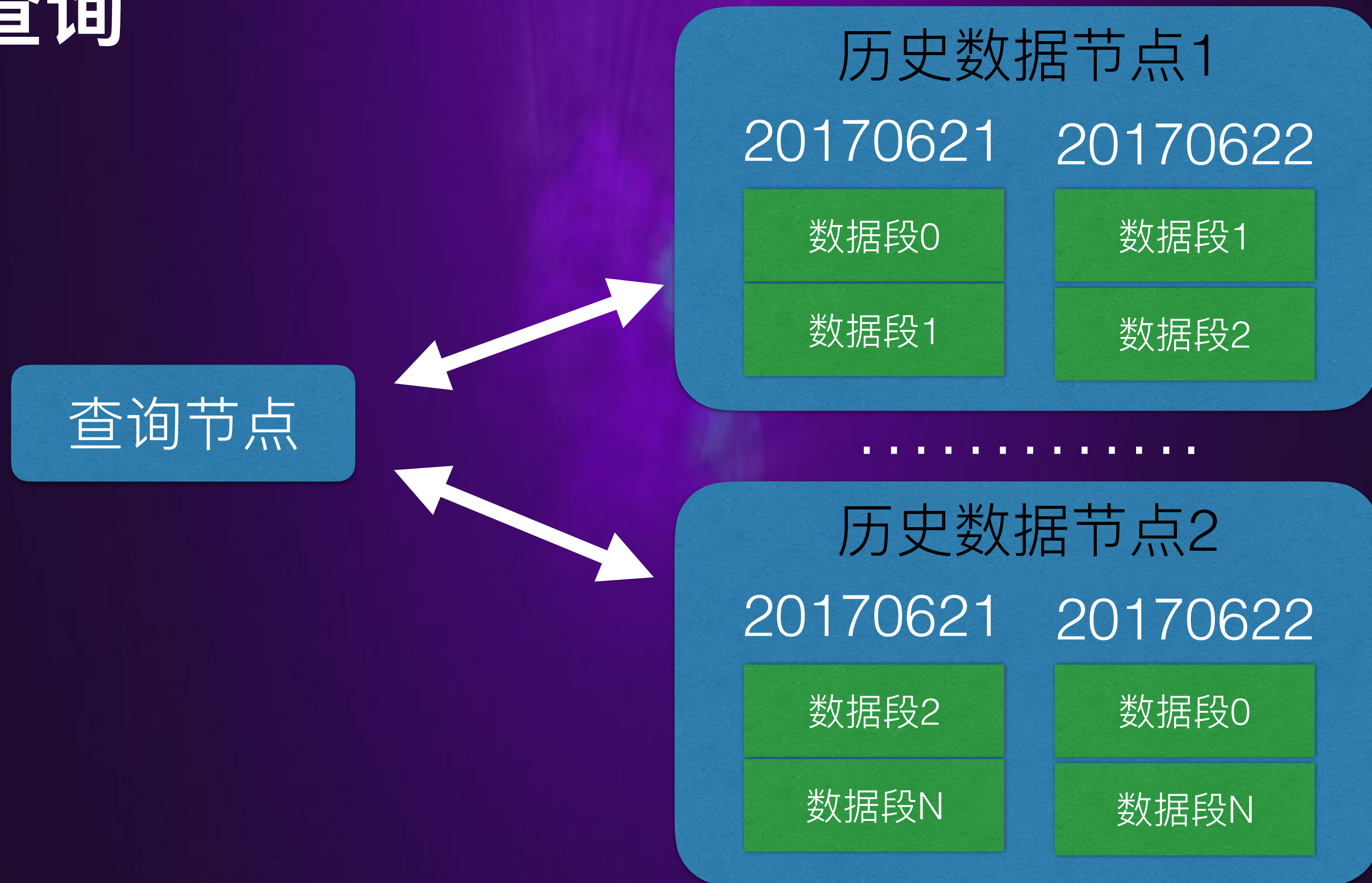
## 二、数据存储





# 数果智能如何处理用户行为数据?

## 三、数据查询





# 数果智能如何处理用户行为数据?



## 四、数据索引

Timestamp	Page	Username	Gender	City
2011-01-01T01:00:00Z	Justin	Boxer	Male	San Fran
2011-01-01T01:00:00Z	Justin	Reach	Male	Waterloo
2011-01-01T02:00:00Z	Ke\$ha	Helz	Male	Calgray
2011-01-01T02:00:00Z	Ke\$ha	Xeno	Male	Taiyuan

# 数果智能如何处理用户行为数据?

## 四、数据索引





# 数果智能如何处理用户行为数据？



## 五、查询多样化

普通查询有timeseries、topN、select、groupBy、firstN、scanQuery等，高级查询包括用户分组、用户漏斗查询、用户留存查询等。

Tindex支持多种条件过滤：日期范围，数字范围，地理坐标范围，字符串的精确匹配、正则匹配、模糊匹配、空值匹配、非空匹配、非等匹配等等。

Tindex支持多种聚合：

### 1. 统计

典型功能：sum、min、max、avg、cardinality、percent、方差、UDF等

### 2. 分组

典型功能：String分组、数字分组、日期分组等

### 3. 聚合再聚合

典型功能：每个地区平均每人的点击数

# 其他大数据方案是否可以处理用户行为



我的看法：可以

但是不够好

- 1、数据的时序性未能充分利用（典型：ELK）
- 2、数据实时性差（典型：内存查询引擎）
- 3、支持维度有限（典型：Hbase、预聚合）
- 4、无法做到查询动态加载数据
- 5、缺少用户行为需要的定制化查询

<https://github.com/Datafruit/gitbook/blob/master/druid/paper.md>



# 基于Tindex我们都做了什么？

## 指标任意定制、维度任意筛选分组

指标管理 ?

项目: 测试数据

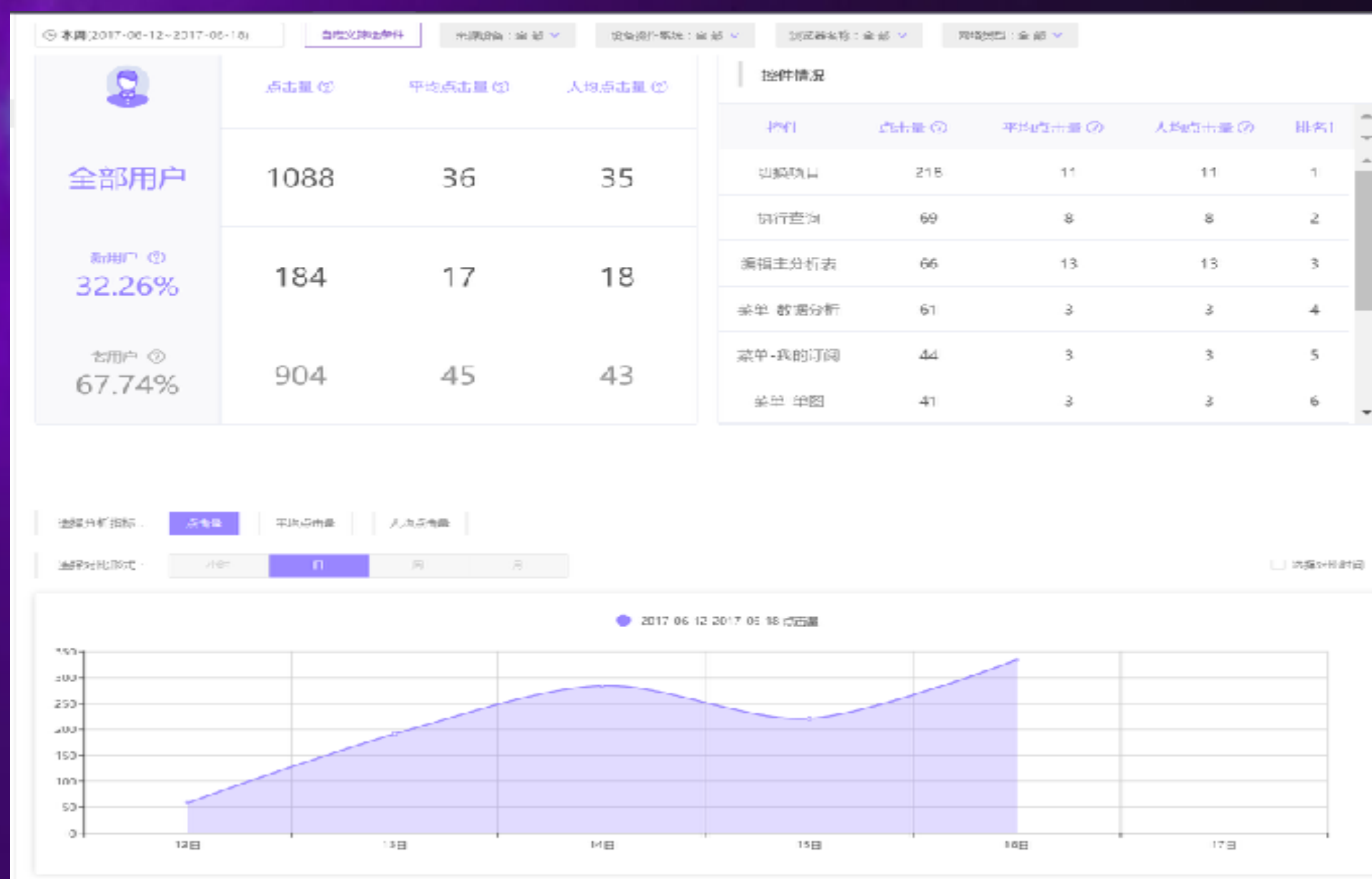
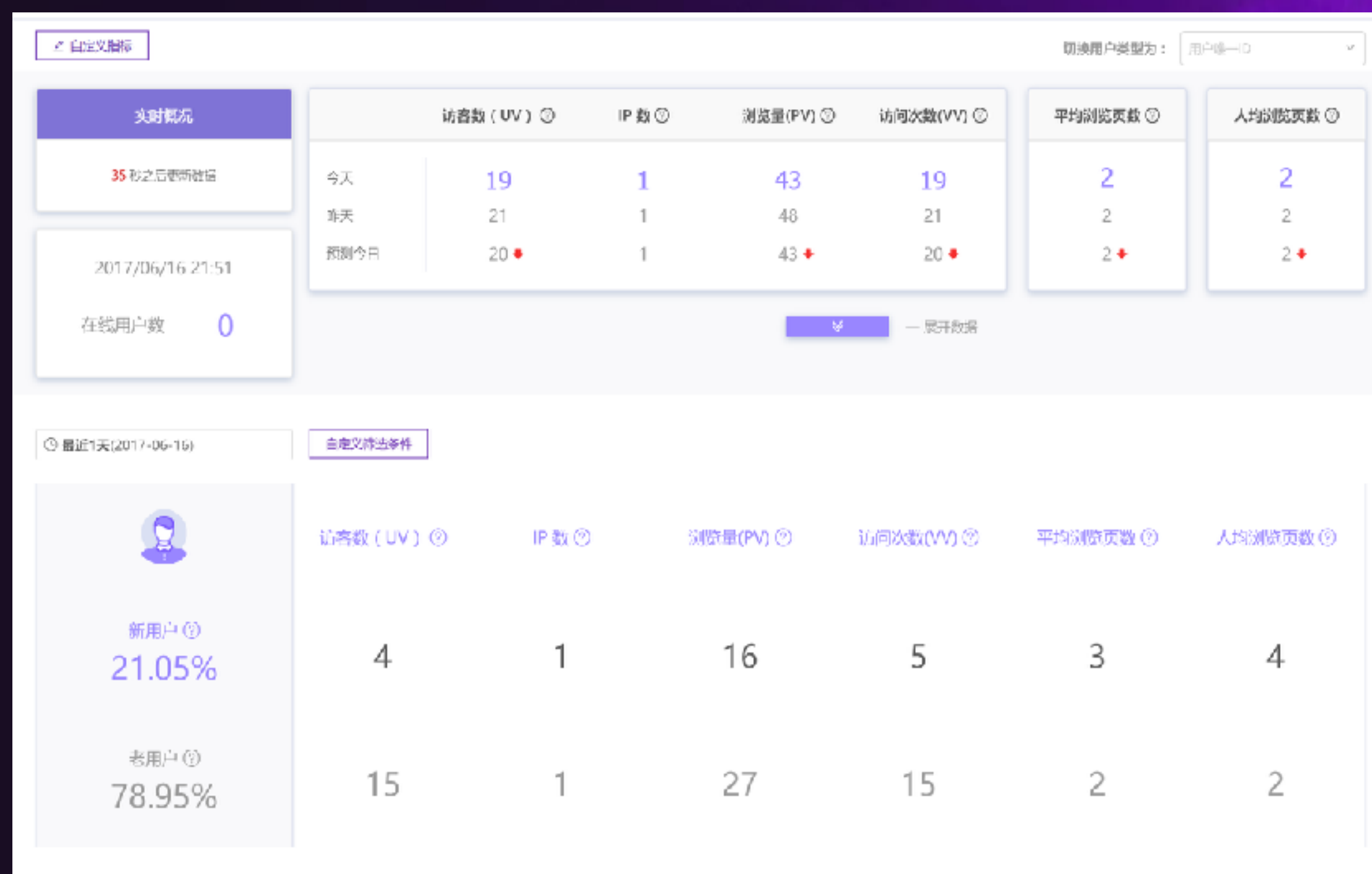
排序和隐藏 全部用户组

<input type="checkbox"/>	名称	公式
<input type="checkbox"/>	菜单数	\$main.countDistinct(\$EventLabel)
<input type="checkbox"/>	用户数(UV)	\$main.countDistinct(\$UserID)
<input type="checkbox"/>	浏览数	\$main.filter(\$EventAction=='浏览').count()
<input type="checkbox"/>	浏览并点击数	\$main.filter(\$EventAction=='浏览').filter(\$EventLabel=='点击').count()
<input type="checkbox"/>	测试点击数	\$main.filter(\$EventAction.in(['浏览','点击'])).count()
<input type="checkbox"/>	启动用户数	\$main.countDistinct(\$ClientDeviceID)
<input type="checkbox"/>	访问用户数	\$main.countDistinct(\$ClientDeviceID)
<input type="checkbox"/>	总记录数	\$main.count()



# 基于Tindex我们都做了什么？

## 用户行为分析





# 基于Tindex我们都做了什么？

## 用户行为分析模型

### 用户分群

用户分群列表 / test-wqc

搜索

- 921 test-wqc
- 1000 taip-1
- 452 test
- 1 一个用户
- 999 浙江用户
- 999 上海用户

用户行为筛选: **并且** 或者

时间范围: 2017-06-14~2017-06-16

指标筛选: **并且** 或者

条件筛选: **并且** 或者

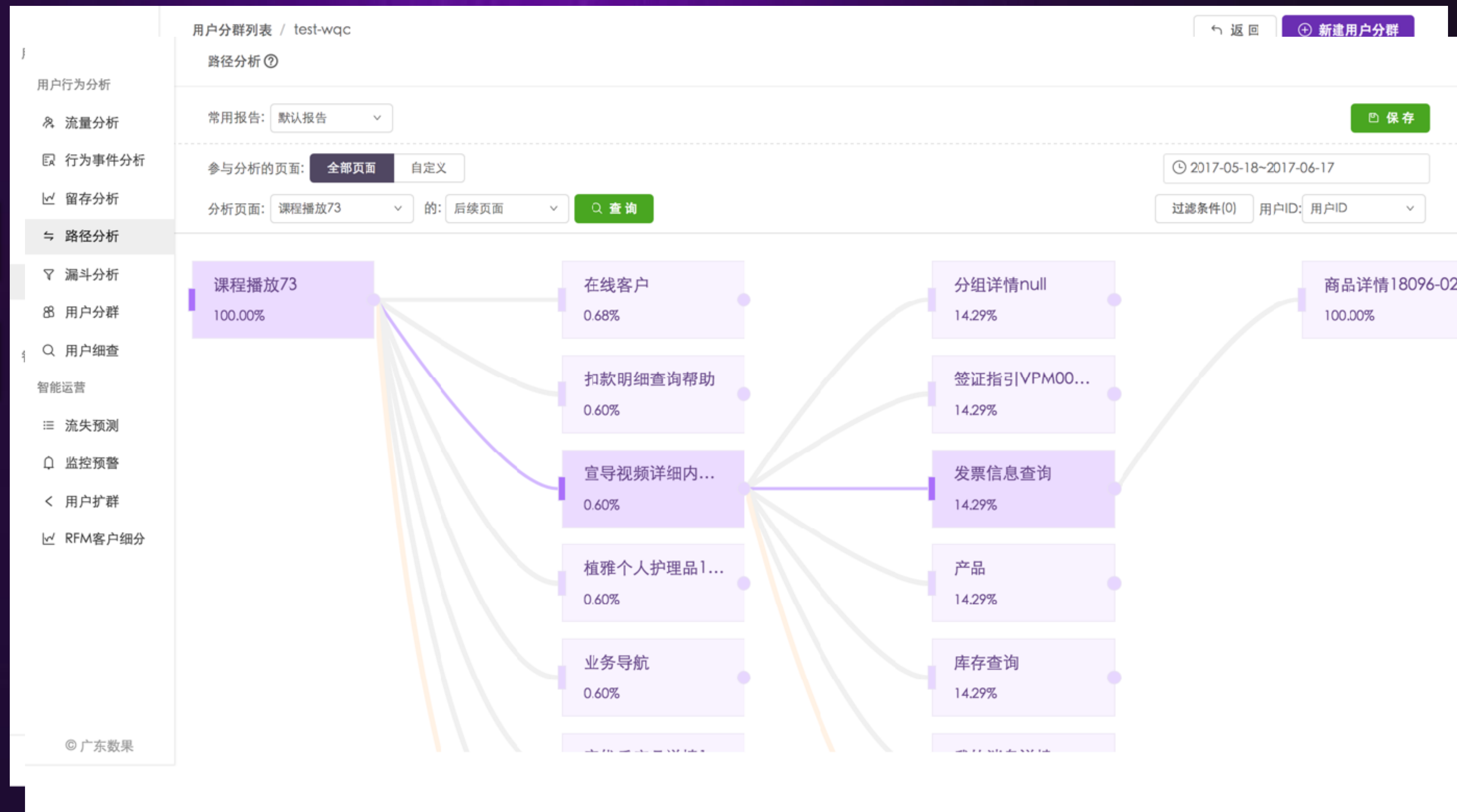
© 广东数果

# 基于Tindex我们都做了什么？

## 用户行为分析模型

用户分群

路径分析





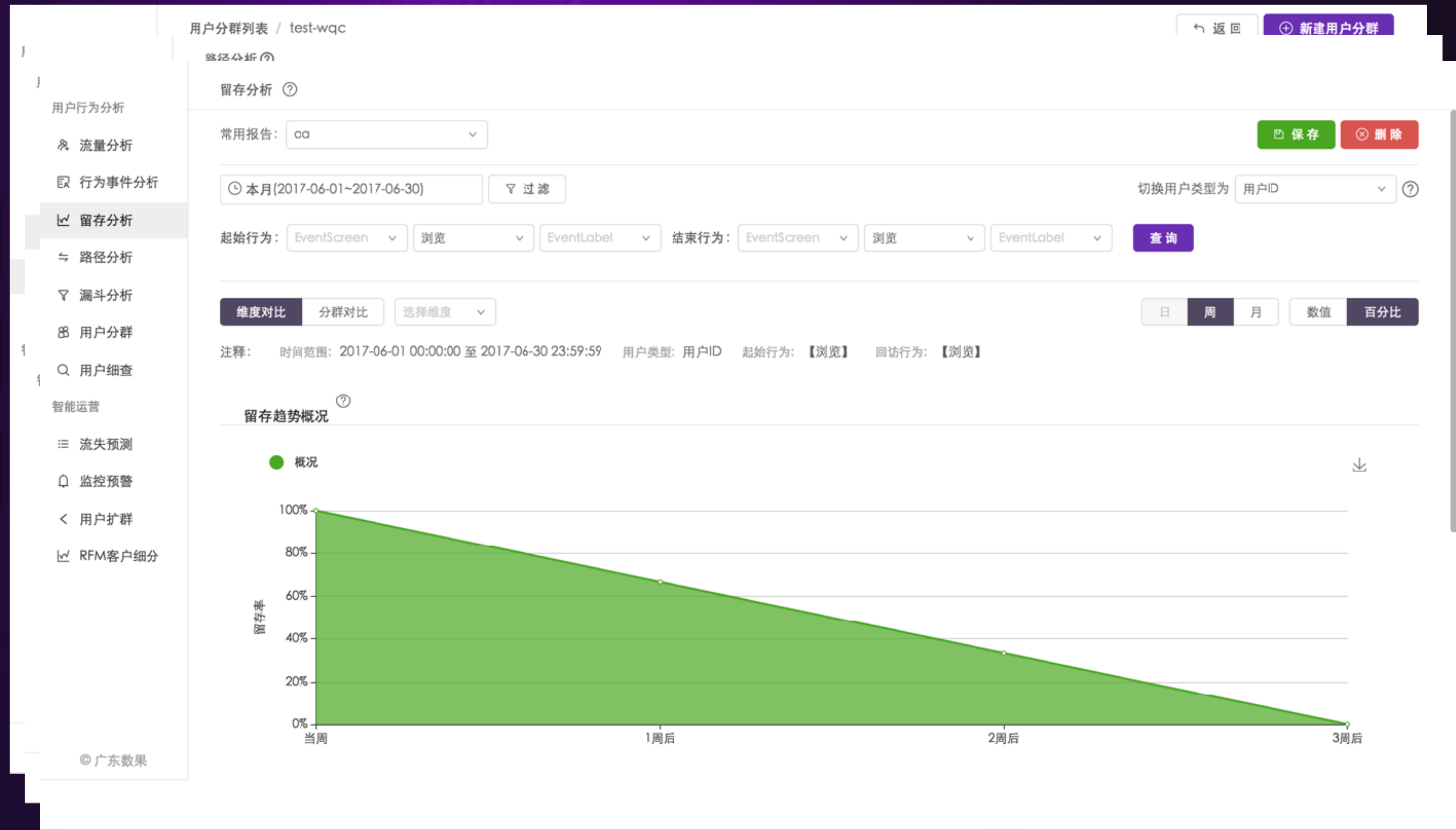
# 基于Tindex我们都做了什么？

## 用户行为分析模型

用户分群

路径分析

留存分析



# 基于Tindex我们都做了什么？

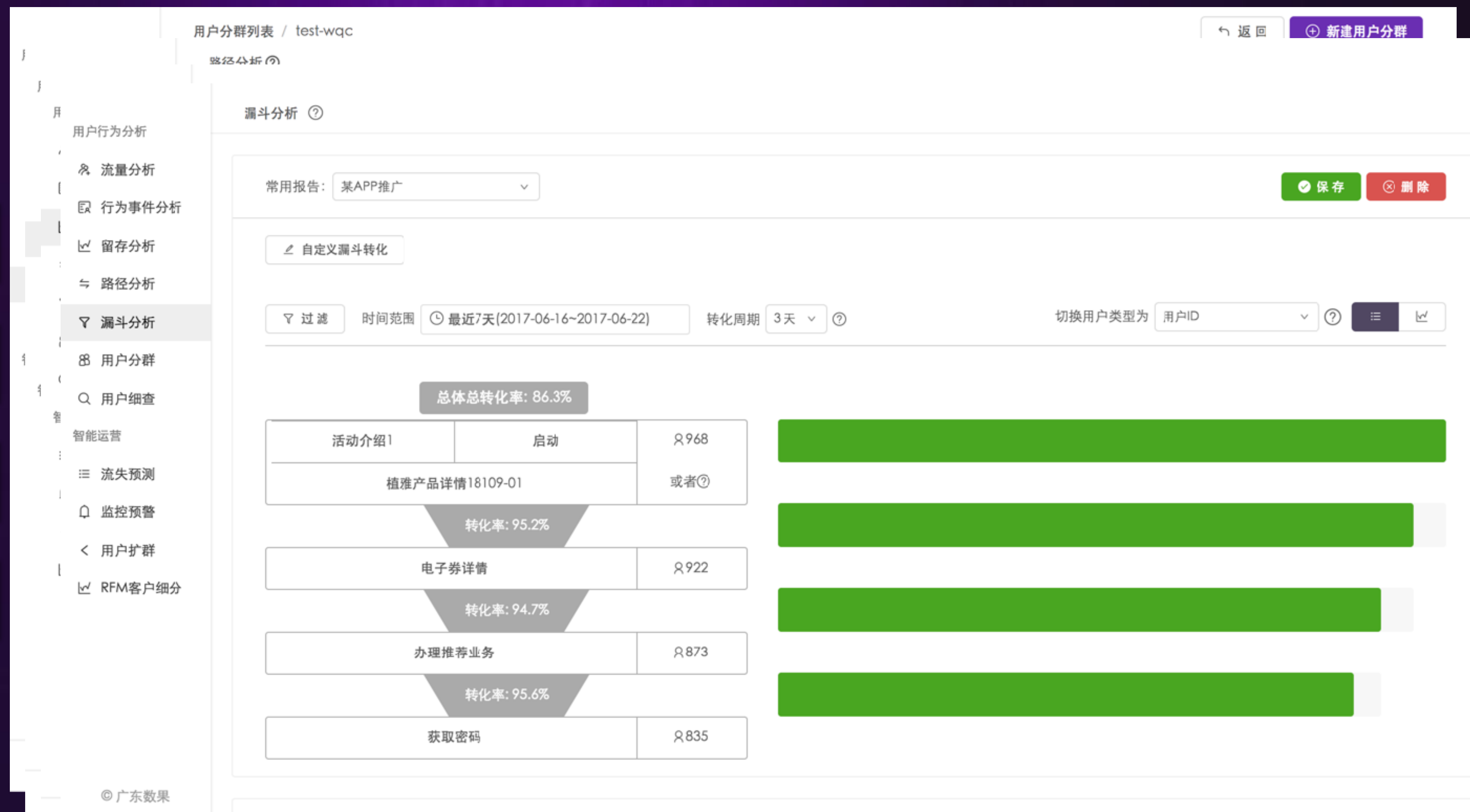
## 用户行为分析模型

用户分群

路径分析

留存分析

漏斗分析



# 基于Tindex我们都做了什么？

## 智能算法模型

## 自定义智能分析

The screenshot displays the Sugon AI analysis interface. The main workspace shows a workflow diagram with the following steps: 读取CSV文件\_1 (Read CSV file\_1), 选择维度\_1 (Select dimension\_1), 设置目标变量\_1 (Set target variable\_1), 随机抽样\_1 (Random sampling\_1), 决策树\_1 (Decision tree\_1), 模型应用\_1 (Model application\_1), and 分类评估\_1 (Classification evaluation\_1). The '决策树\_1' node is highlighted with a purple border. On the right, the '决策树算子设定' (Decision tree operator settings) panel is open, showing various parameters: 选择最佳分割方法 (Select best split method) set to 增益率 (Gain rate), 单棵树最大树状图深度 (Maximum tree diagram depth) set to 20, 剪枝 (Pruning) checked, 置信度 (0~1) (Confidence) set to 0.25, 预剪枝 (Pre-pruning) checked, 最小增益阈值 (Minimum gain threshold) set to 0.1, 叶节点最小个数 (Minimum number of leaf nodes) set to 2, and 最小分割大小 (Minimum split size). The top navigation bar includes options like 图表 (Charts), 数据分析 (Data analysis), 智能分析 (AI analysis), 场景应用 (Scenario application), 数据管理 (Data management), and 管理中心 (Management center). The user is logged in as admin@广东数果.



# 基于Tindex我们都做了什么？



## 智能算法模型

## 自定义智能分析

## 用户扩群

数果智能 SugoiIO 图表 数据分析 智能分析 场景应用 数据管理 管理中心 admin@广东数果

数果智能 SugoiIO 图表 数据分析 智能分析 场景应用 数据管理 管理中心 切换项目 测试数据 admin@广东数果

用户扩群列表 / 根据浙江用户扩群

搜索

状态: 计算中

更新状态 删除

扩群名称: 根据浙江用户扩群

时间范围: 最近7天(2017-06-16~2017-06-22)

参照目标群: 浙江用户 查看分群用户

用户id: 用户ID

特征指标: (选择扩群重要特征指标)

搜索 新建指标 已选择(3)

浏览数 启动用户数 访问用户数 菜单数 用户数(UV)

浏览并点击数 测试点击数 总记录数

用户数量: 1000

扩群算法: 找最相似用户

备注:

© 广东数果

# 基于Tindex我们都做了什么？



智能算法模型

自定义智能分析

用户扩群

RFM用户细分

数果智能 SugoiIO 图表 数据分析 智能分析 场景应用 数据管理 管理中心 admin@广东数果

数果智能 SugoiIO 图表 数据分析 智能分析 场景应用 数据管理 管理中心 切换项目 测试数据 admin@广东数果

数果智能 SugoiIO 图表 数据分析 智能分析 场景应用 数据管理 管理中心 admin@广东数果

智能运营 / RFM客户细分 / RFM客户细分

所属项目: RFM\_data RFM名称: RFM\_12

时间范围: 今年(2017-01-01~2017...)

1. 预设相应的维度 跳转到场景数据设置

购买日期维度: 购买日期

购买金额维度: 购买金额

客户ID维度: 用户ID

2. 配置参数

基础设置 自定义设置

智能划分块数

最近一次消费(R): 2

消费频率(F): 2

累计消费金额(M): 2

查询

© 广东数果



# 基于Tindex我们都做了什么？



智能算法模型

自定义智能分析

用户扩群

RFM用户细分

流失预测

数果智能 SugoIO 图表 数据分析 智能分析 场景应用 数据管理 管理中心 admin@广东数果

数果智能 SugoIO 图表 数据分析 智能分析 场景应用 数据管理 管理中心 切换项目 测试数据 admin@广东数果

数果智能 SugoIO 图表 数据分析 智能分析 场景应用 数据管理 管理中心 admin@广东数果

数果智能 SugoIO 图表 数据分析 智能分析 场景应用 数据管理 管理中心 admin@广东数果

智能运营 / 开始训练

建立训练模型 ?

1 导入文件 2 设置训练字段 ? 3 运行训练数据

温馨提示：请导入用于训练模型的数据，格式为 CSV 文件，并确保你的文件编码方式为 UTF-8 (上传文件说明) 下载导入模版

已上传的数据文件列表

输入要查找的文件名称

- 流失预警训练数据 2017-04-13 11:04 (训练数据样本 - 副本 (4).csv)
- 流失预警训练数据 2017-04-13 11:04 (训练数据样本 - 副本 (4).csv)
- 流失预警训练数据 2017-04-12 17:04 (训练数据样本.csv)
- 流失预警训练数据 2017-04-12 14:04 (流失预测数据170228.csv)
- 流失预警训练数据 2017-04-12 14:04 (训练数据样本.csv)
- 流失预警训练数据 2017-04-12 08:04 (训练数据样本.csv)

点此导入文件

下一步

© 广东数果



# 基于Tindex我们都做了什么？



## 实时监控大屏

**东风日产**

在保车辆: 9,467  
过保车辆: 9,467

各级会员数量

年份	数量
2004	2064
2000	2060
1994	1994
1976	1976
1906	1906

过保留存车辆: 8,385  
过保流失车辆: 1,082

各省消费总额

消费排行

各省客单价

饼图

总计: #fec021

文字大小: 32px

字体: 微软雅黑

Thank you !