

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 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', 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, Network: 3g, Country: 中国, Media: (empty), IP: 106.91.157.45, Extras: (empty), EventValue: (empty), EventScreen: 选择产品, EventLabel: (empty), EventHour: (empty), EventDateTime: 1491185548336, EventDate: (empty), EventAction: 后台, Creative: (empty), ClientDeviceVersion: 4.4.4, and ClientDeviceModel: 2014813. A '行为记录' section on the right lists various actions with timestamps, 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'. The interface also includes a sidebar with navigation options like '流量分析', '行为事件分析', '留存分析', '路径分析', '漏斗分析', '用户分群', '用户细查', '智能运营', '流失预测', '监控预警', '用户扩群', and 'RFM客户细分'.

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



BinLog采集器



文件采集器



采集网关服



Kafka

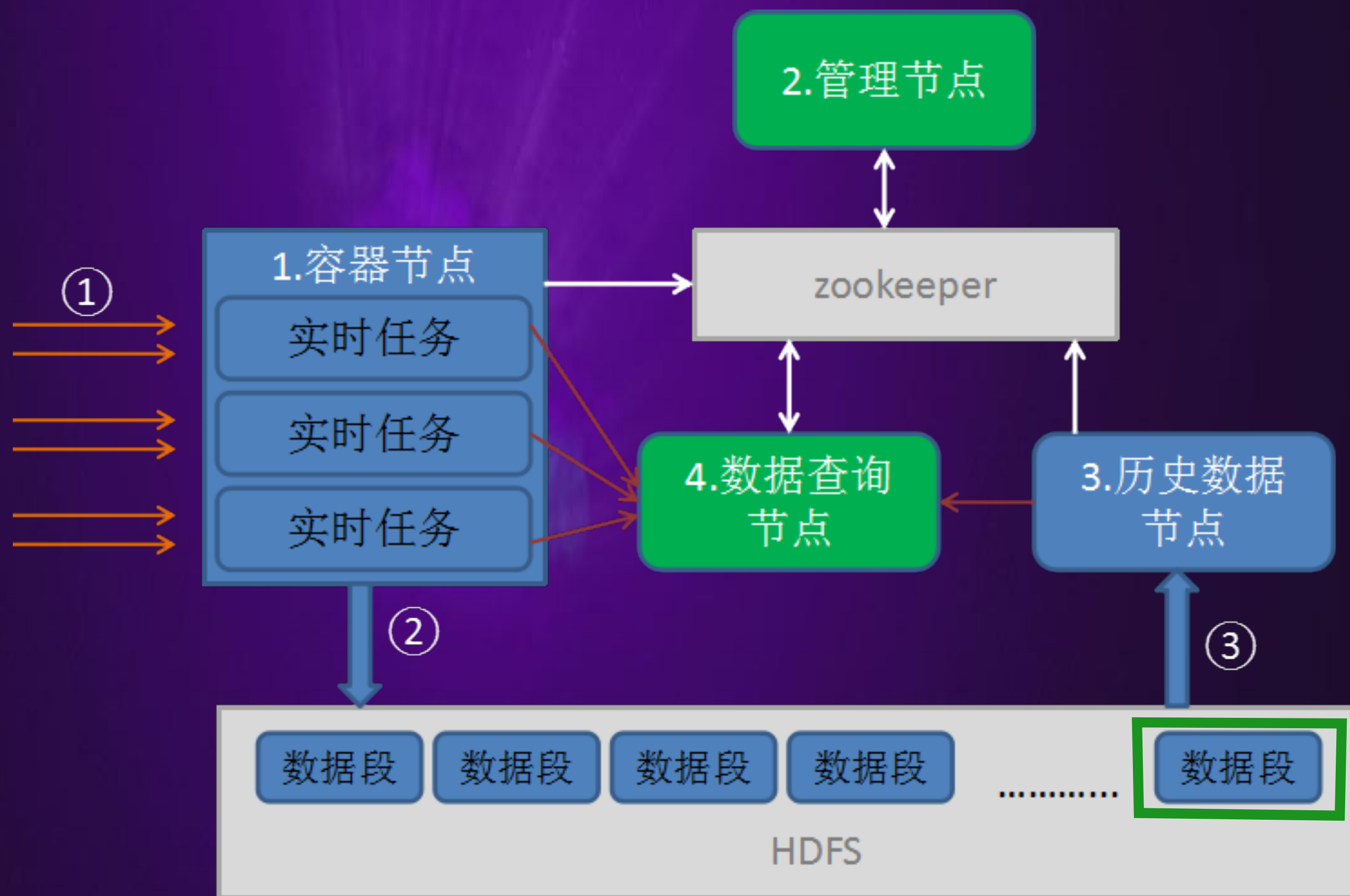


数果TIndex



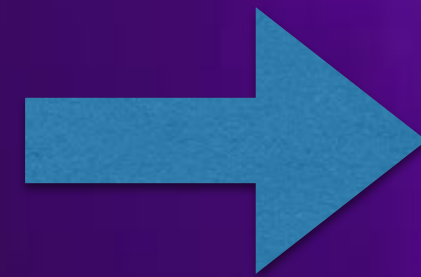
数果智能如何处理用户行为数据？

二、数据存储



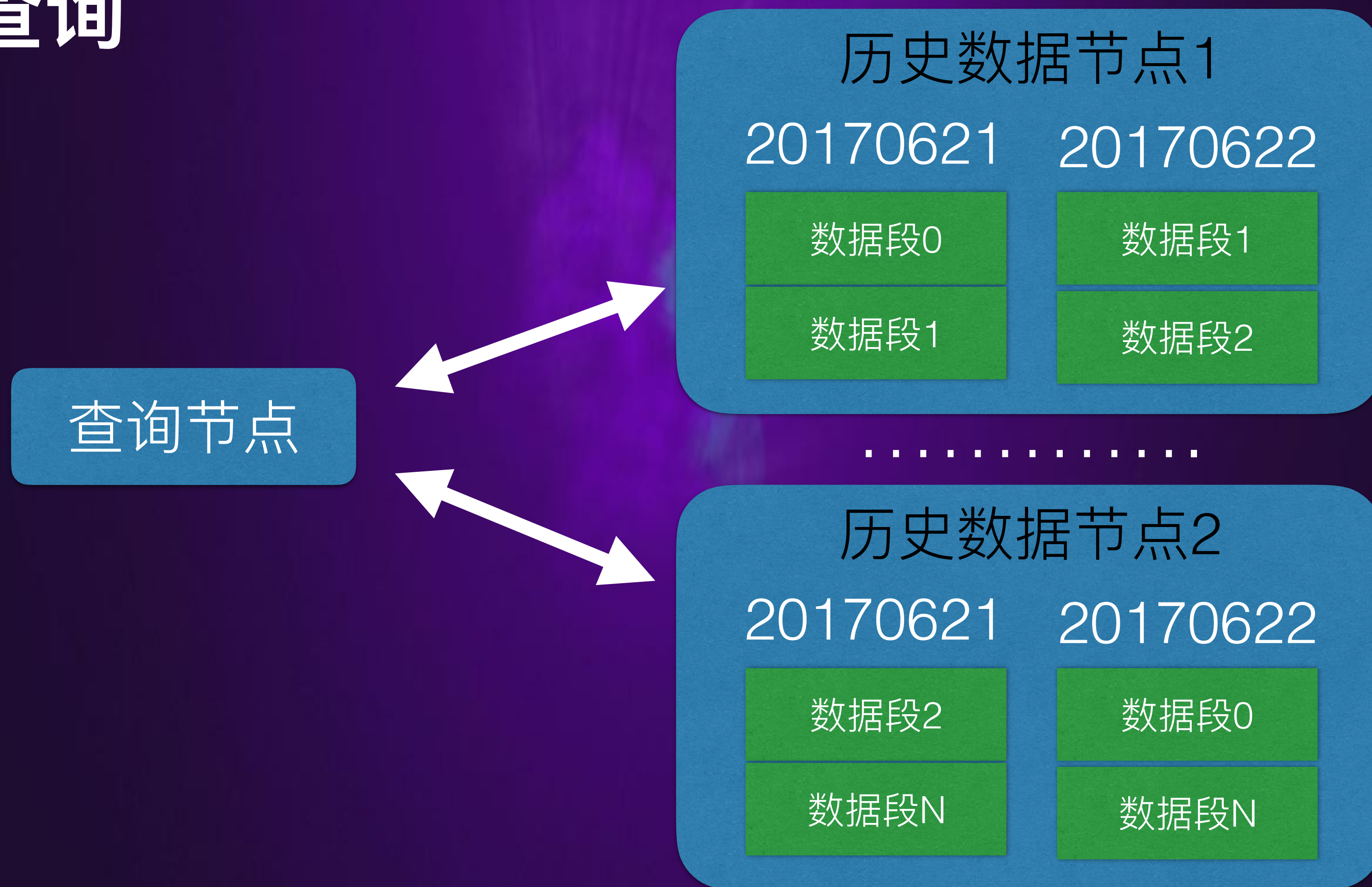
数果智能如何处理用户行为数据？

二、数据存储



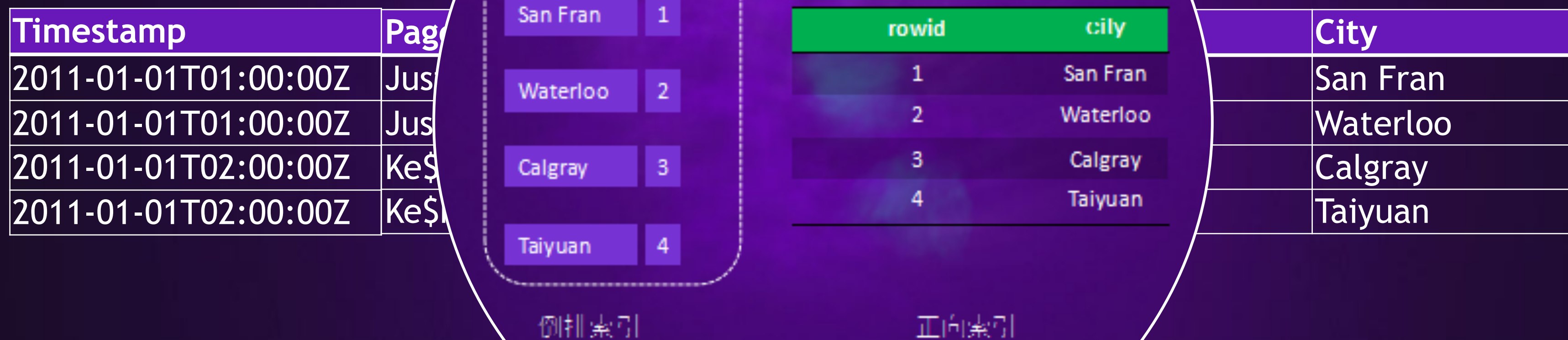
数果智能如何处理用户行为数据？

三、数据查询



数果智能如何处理用户行为数据?

四、数据索引



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

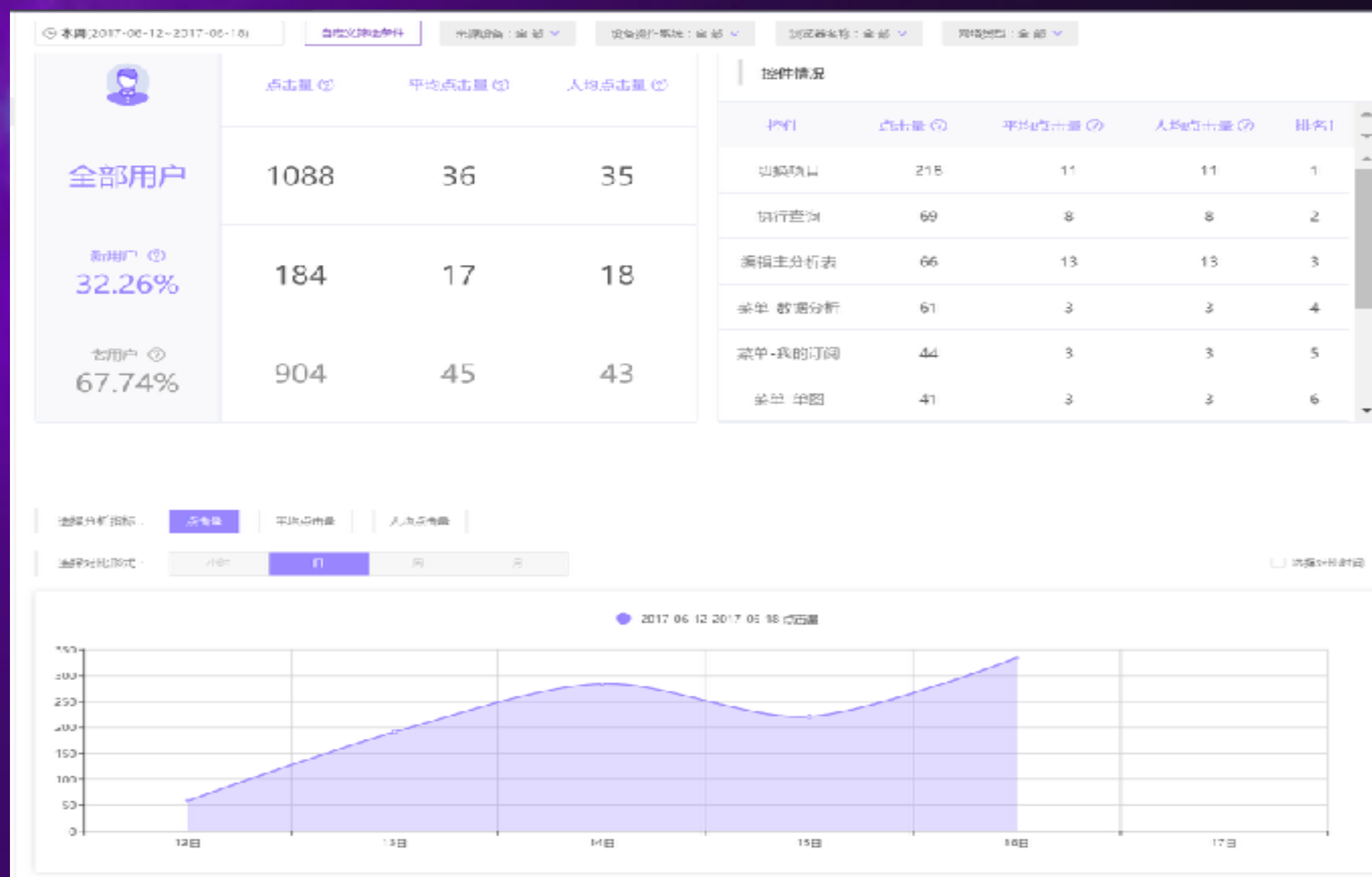
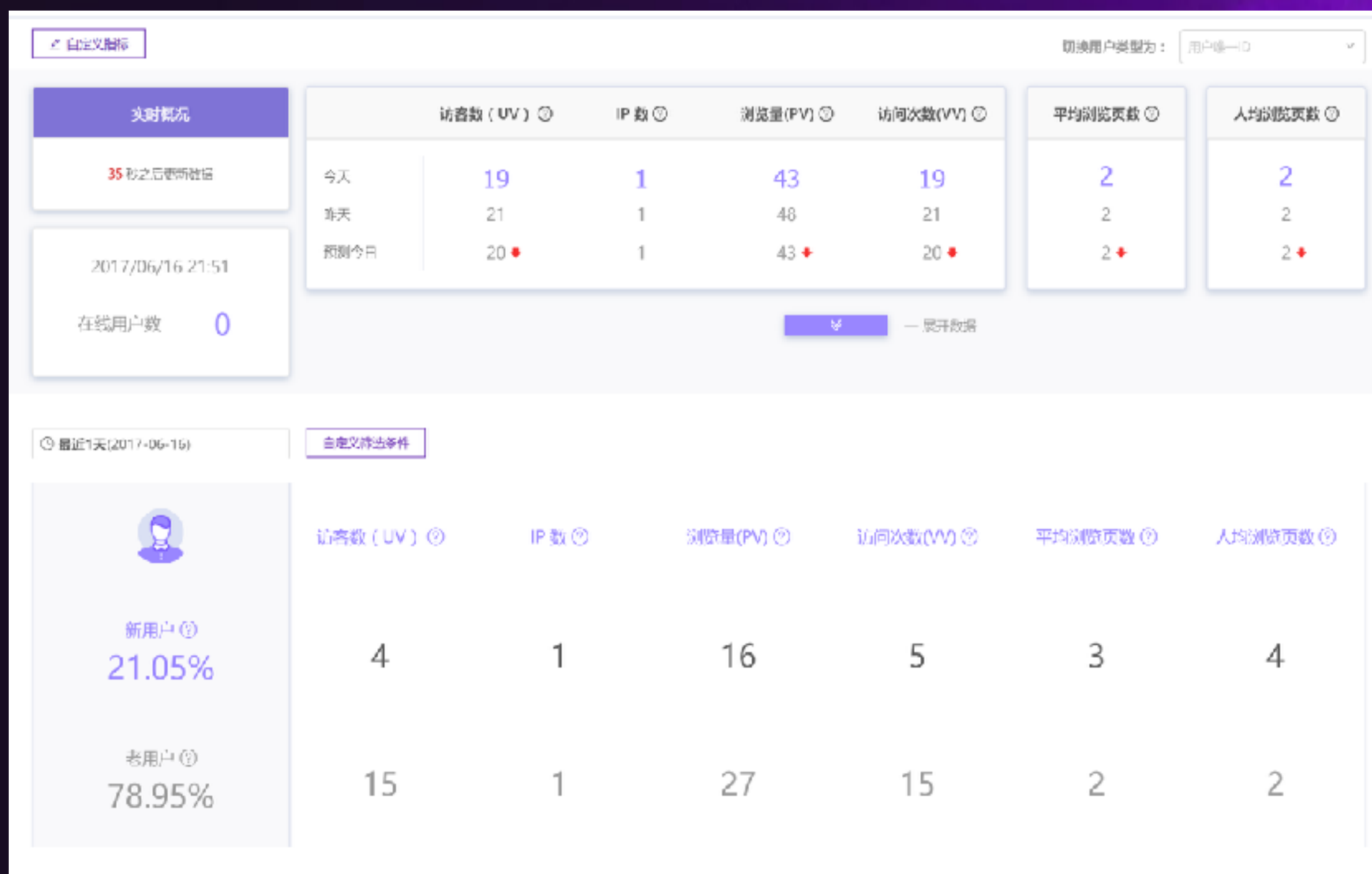
排序和隐藏 全部用户组

<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？

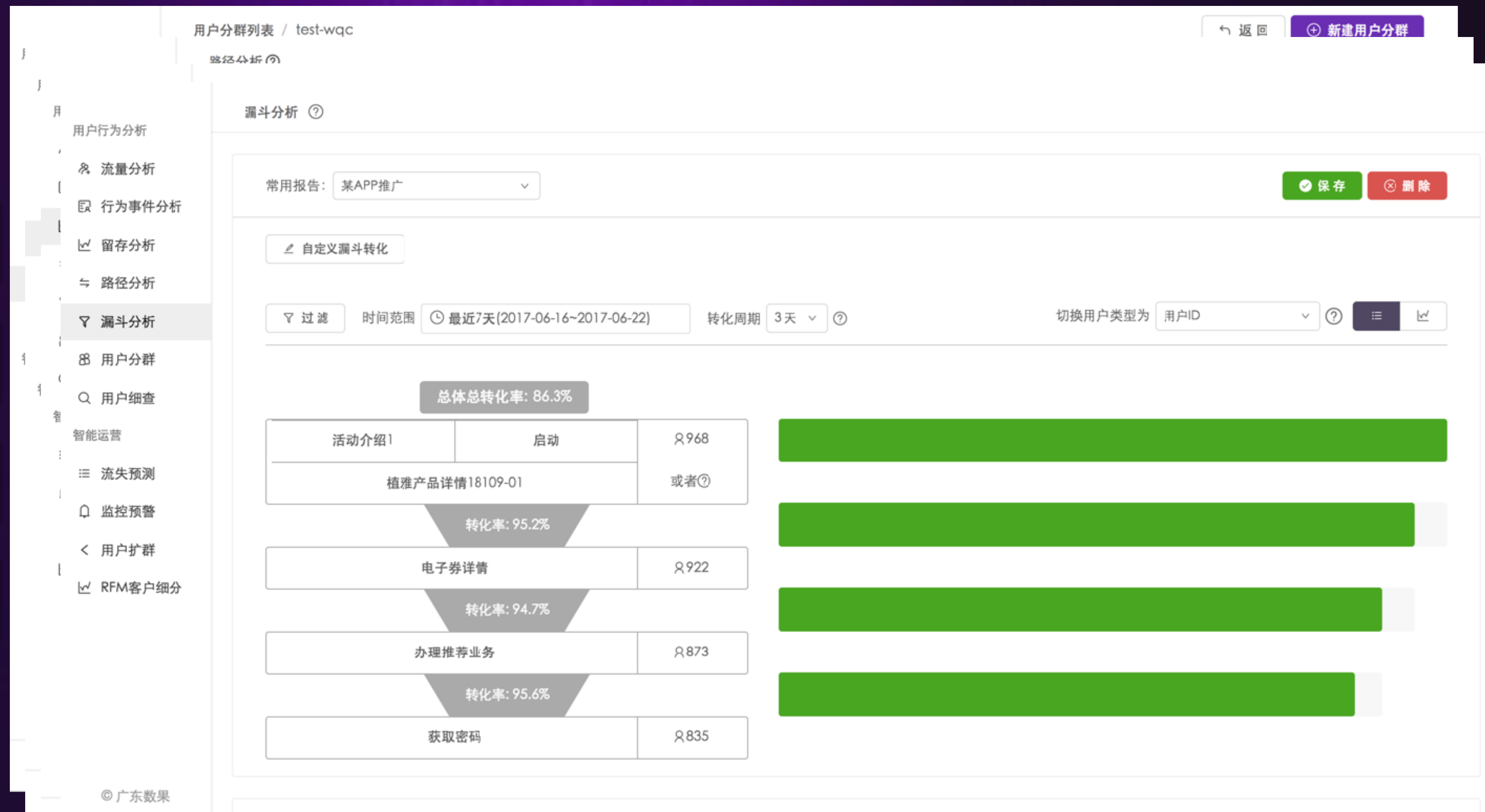
用户行为分析模型

用户分群

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

用户扩群

RFM用户细分

流失预测

The screenshot shows the SugoIO web interface for building a training model. The top navigation bar includes '数果智能' (SugoIO), '图表' (Charts), '数据分析' (Data Analysis), '智能分析' (Smart Analysis), '场景应用' (Scenario Applications), '数据管理' (Data Management), and '管理中心' (Management Center). The user is logged in as 'admin@广东数果'. The main content area is titled '智能运营 / oa / 开始训练' (Smart Operation / oa / Start Training) and '建立训练模型' (Build Training Model). A progress indicator shows three steps: 1. 导入文件 (Import File), 2. 设置训练字段 (Set Training Fields), and 3. 运行训练数据 (Run Training Data). A warning message states: '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)' (Warm tip: Please import data for training the model, in CSV format, and ensure your file encoding is UTF-8 (Upload File Instructions)). A large blue plus sign icon is centered with the text '点此导入文件' (Click here to import file). On the right, a search bar is labeled '已上传的数据文件列表' (List of uploaded data files) and contains a search input field. Below the search bar, a list of files is shown, each with a radio button and a filename: '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)', and '流失预警训练数据 2017-04-12 08:04 (训练数据样本.csv)'. A '下一步' (Next Step) button is at the bottom right. The footer shows '© 广东数果'.

基于Tindex我们都做了什么？

实时监控大屏



Thank you !