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

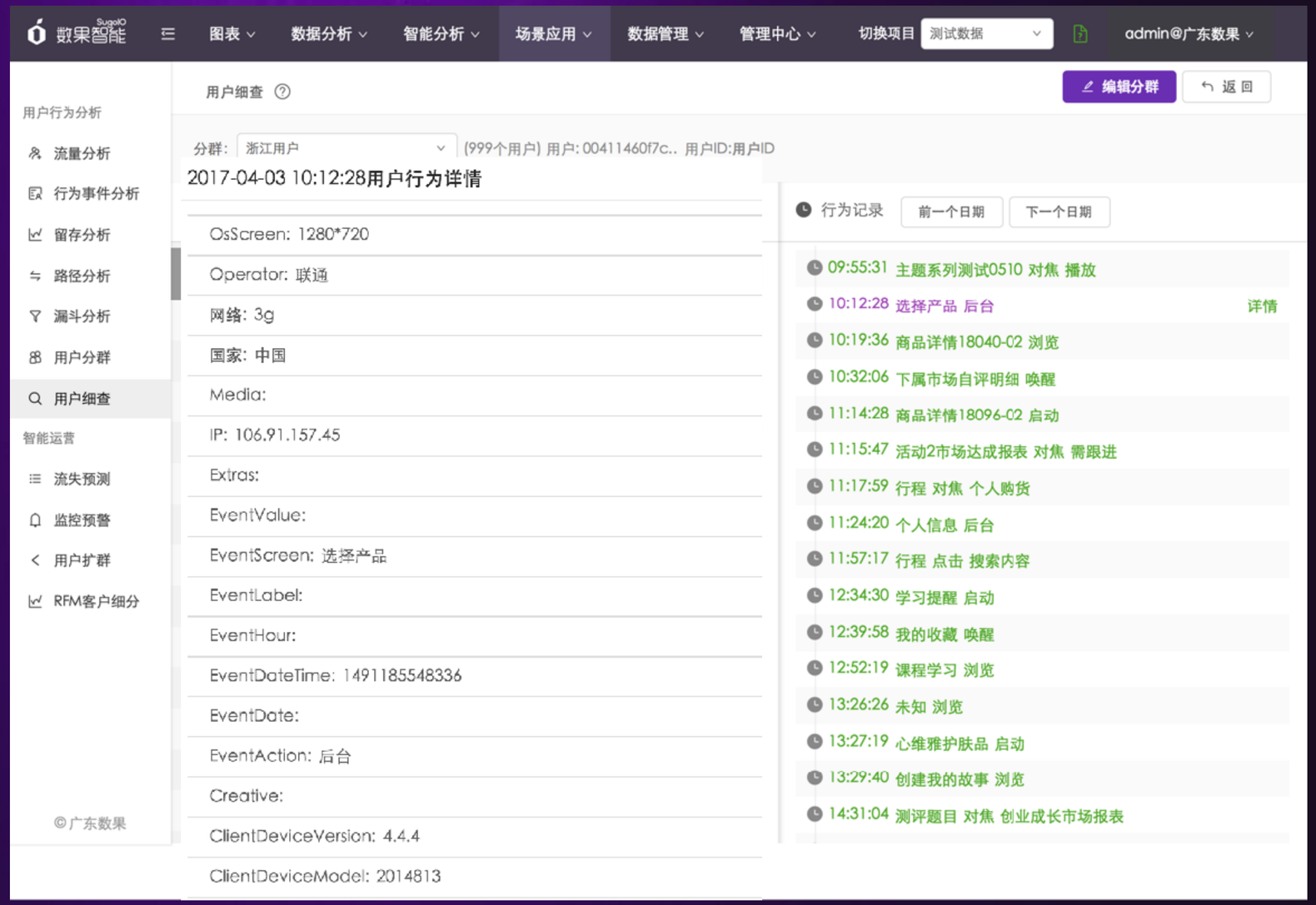
什么是用户行为?

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



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

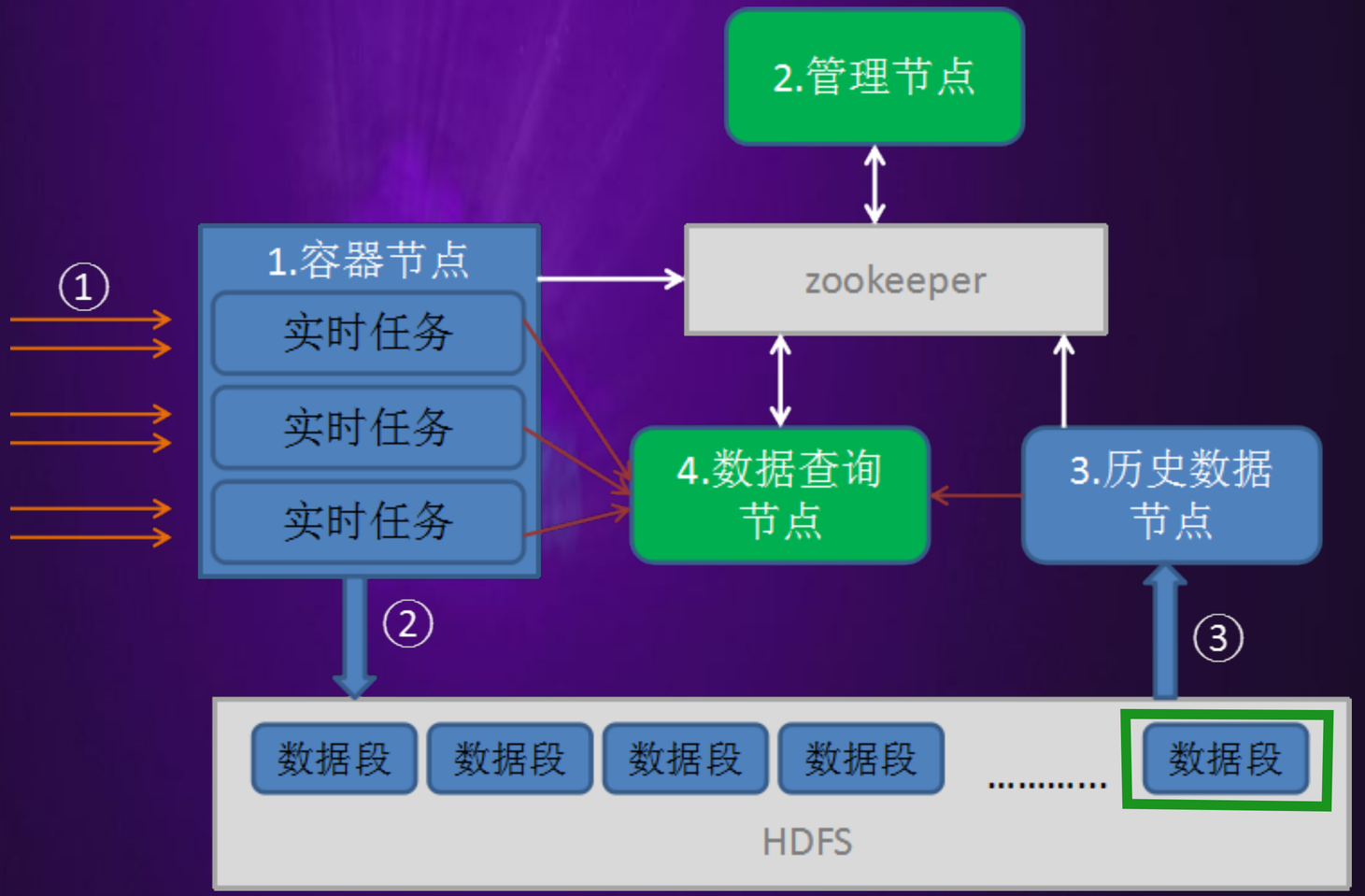
采集网关服

Kafka

数果TIndex

数果智能如何处理用户行为数据?

二、数据存储

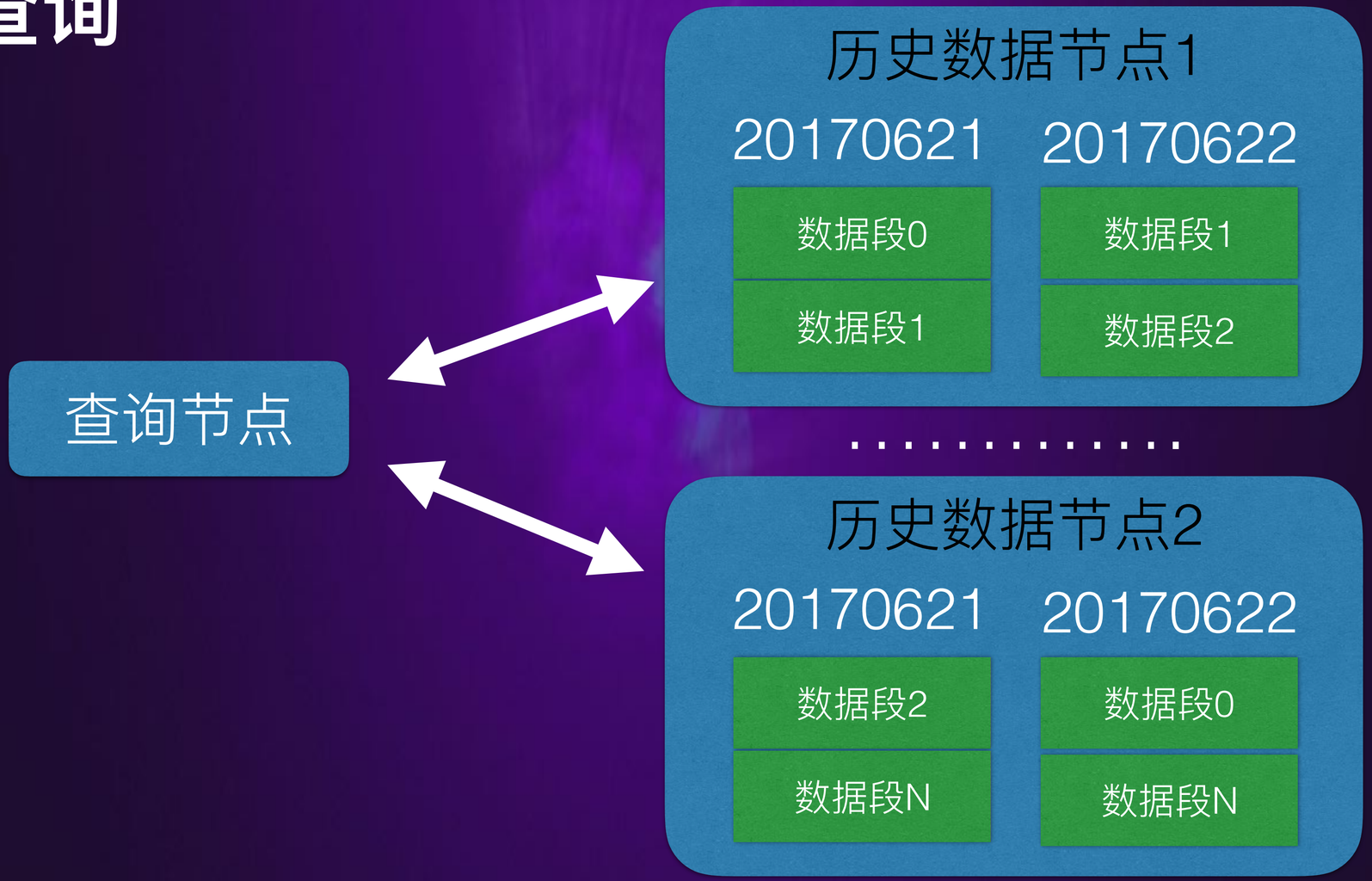


数果智能如何处理用户行为数据?

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

Timestamp	Pag	City
2011-01-01T01:00:00Z	Jus	San Fran
2011-01-01T01:00:00Z	Jus	Waterloo
2011-01-01T02:00:00Z	Ke$	Calgray
2011-01-01T02:00:00Z	Ke$	Taiyuan

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

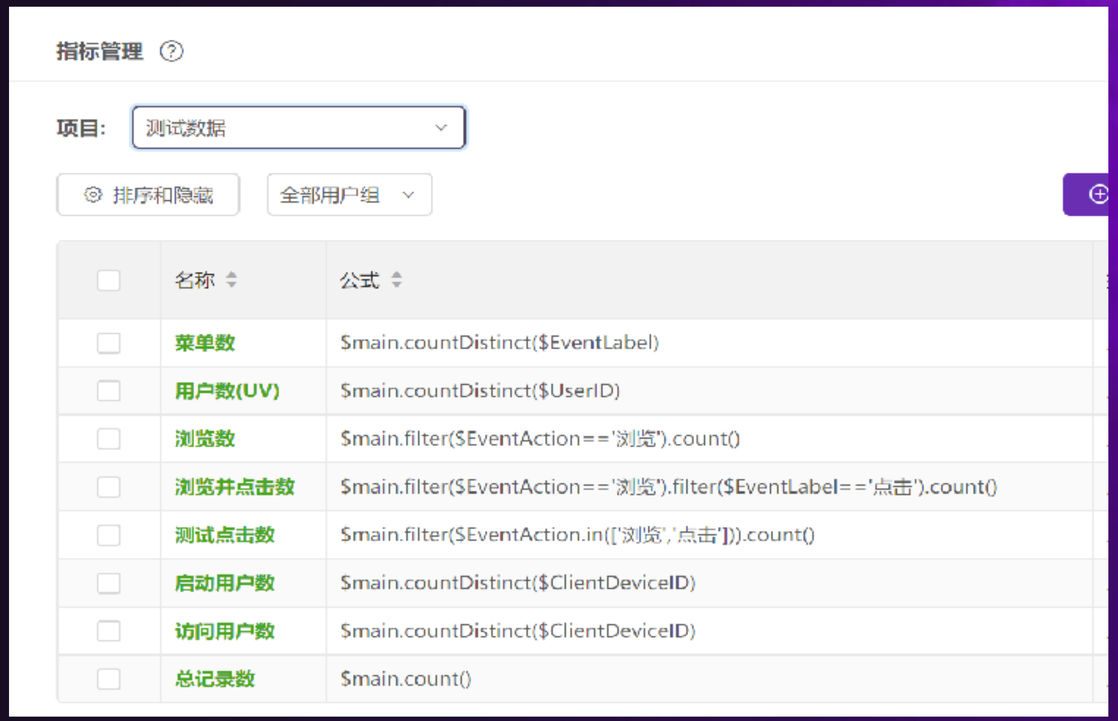
但是不够好

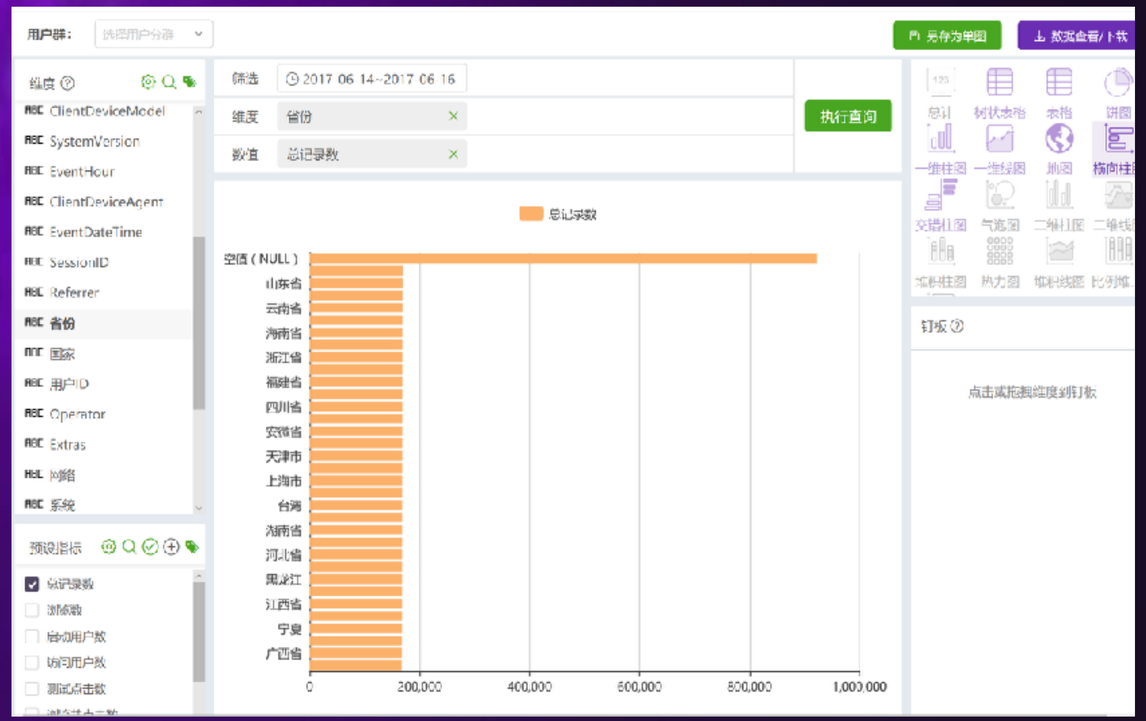
1、数据的时序性未能充分利用（典型：ELK）
2、数据实时性差（典型：内存查询引擎）
3、支持维度有限（典型：Hbase、预聚合）
4、无法做到查询动态加载数据
5、缺少用户行为需要的定制化查询

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

基于Tindex我们都做了什么?

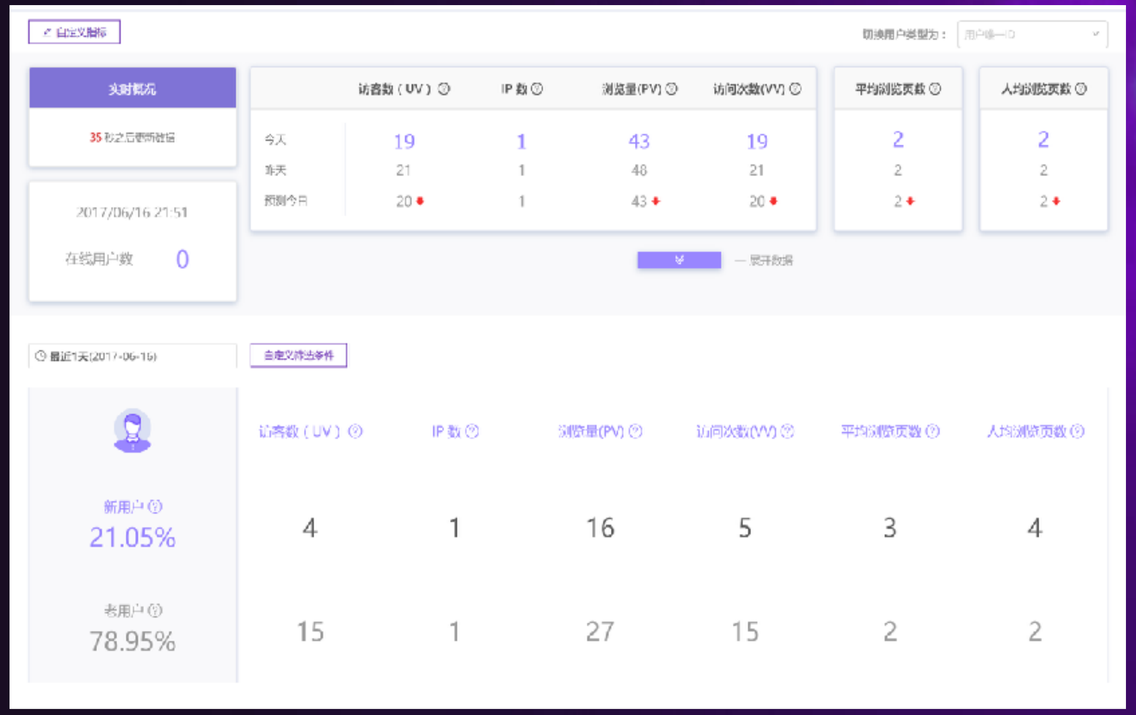
指标任意定制、维度任意筛选分组

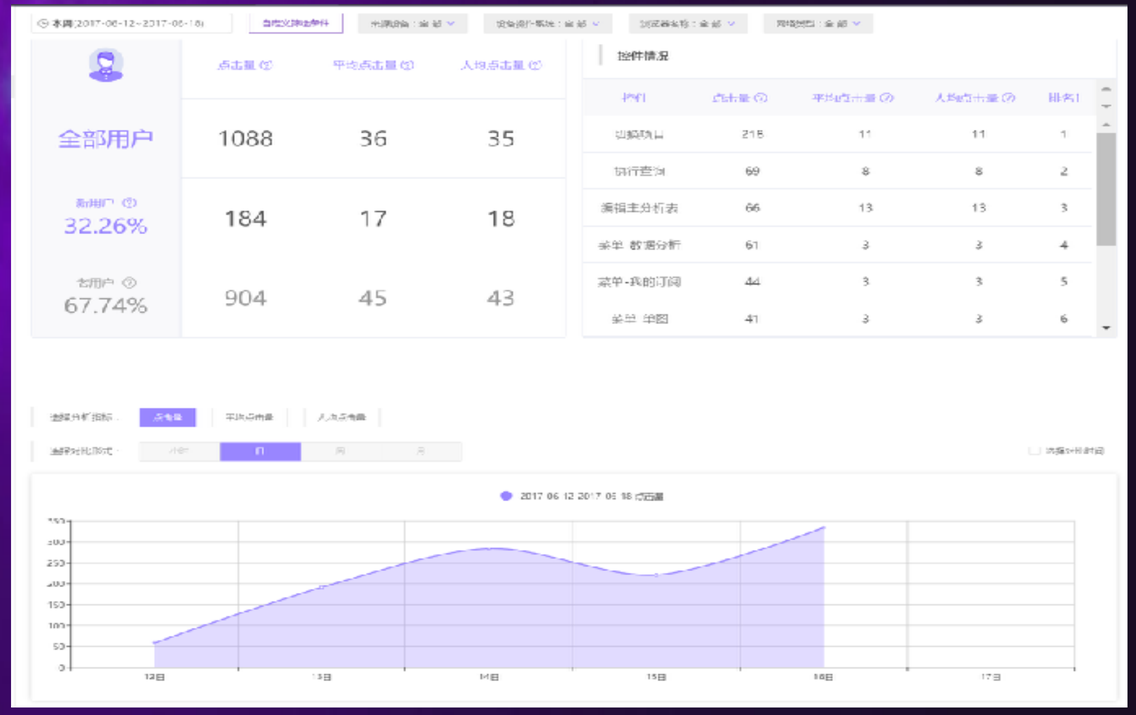




基于Tindex我们都做了什么?

用户行为分析

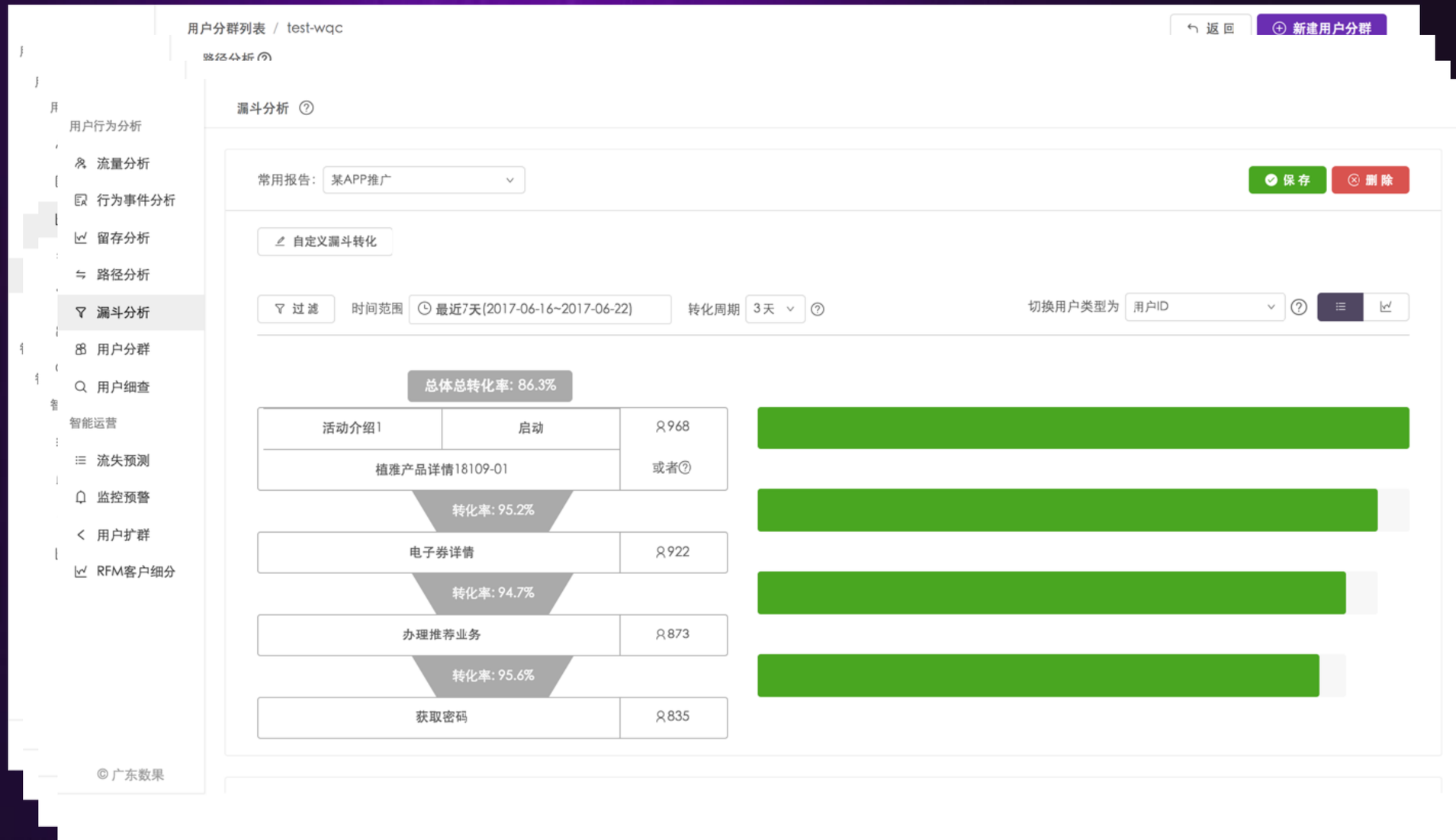




基于Tindex我们都做了什么?

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?

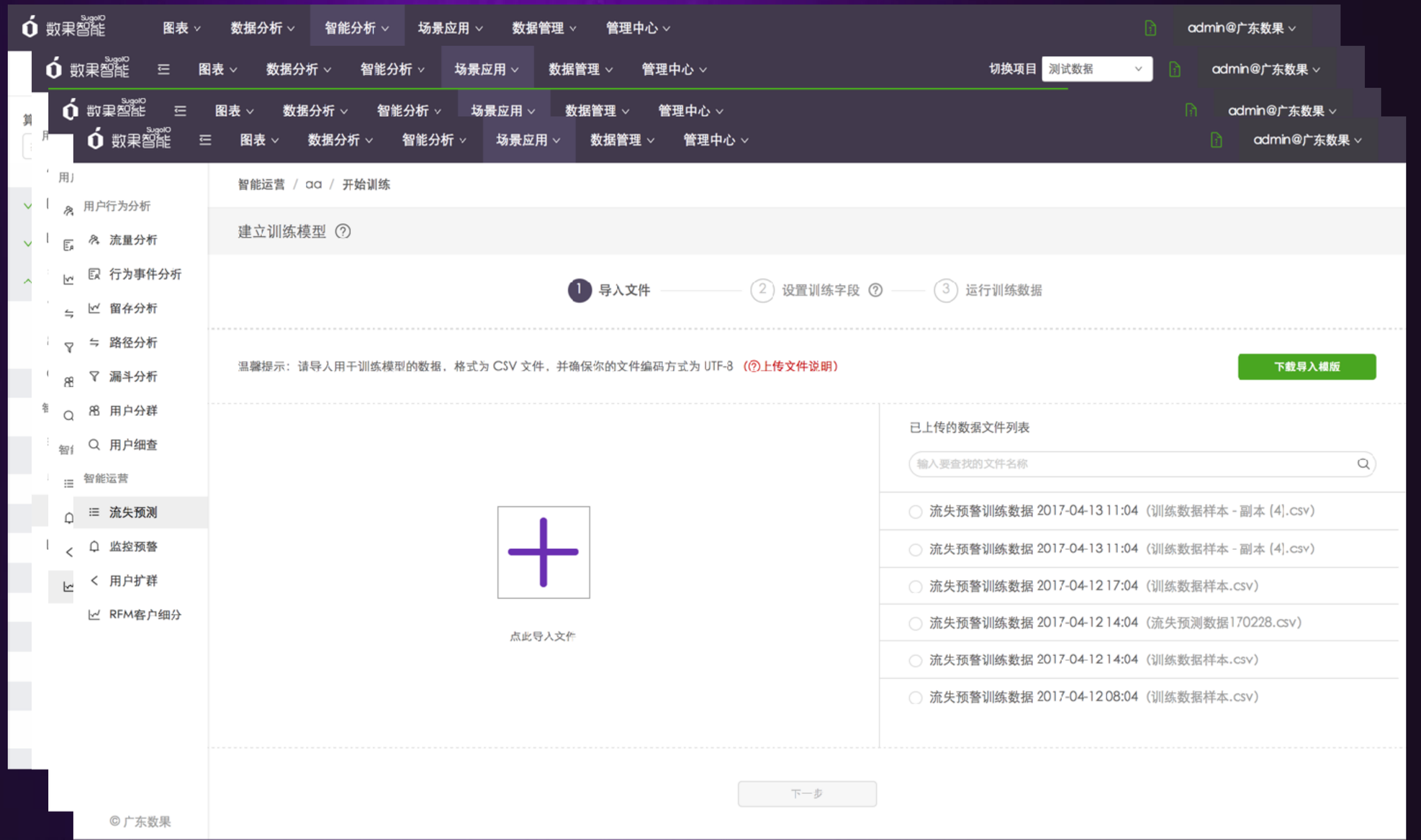
智能算法模型

自定义智能分析

用户扩群

RFM用户细分

流失预测



基于Tindex我们都做了什么?

实时监控大屏

Thank you !