同程艺龙

先知-智能营销平台

周辉

数据中台研发部

自我介绍

周辉

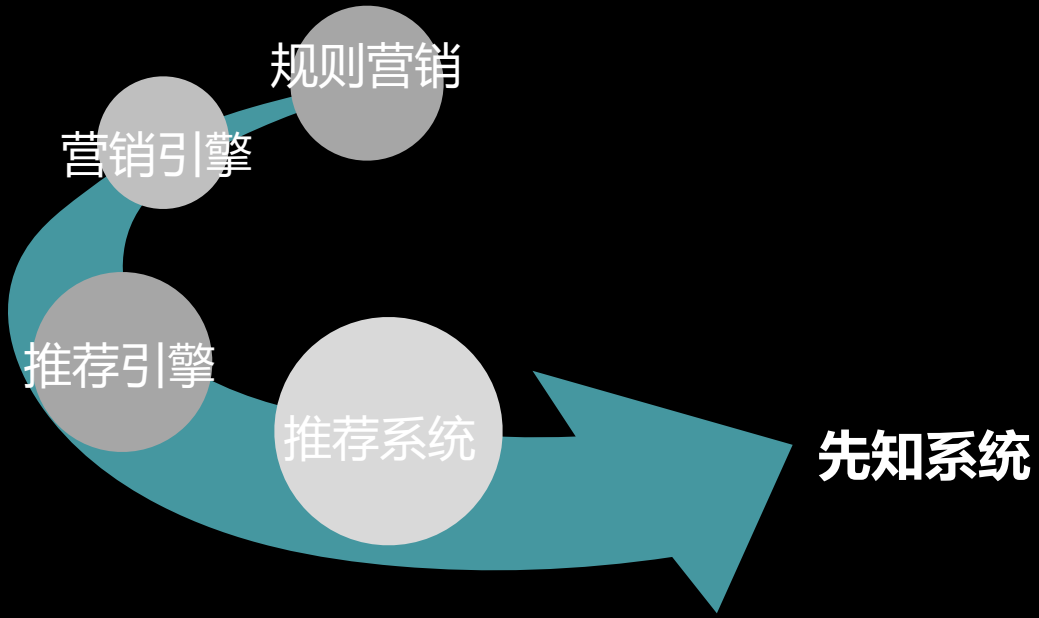
2017年加入同程，现任职于数据中台研发部，先后参与过大数据一站式平台建设，统一轨迹采集，现在主要负责先知系统等数据中台产品开发及实时应用研发。

同程艺龙

目录

先知-业务代号

同程艺龙



基于用户画像智能分配人群的营销平台

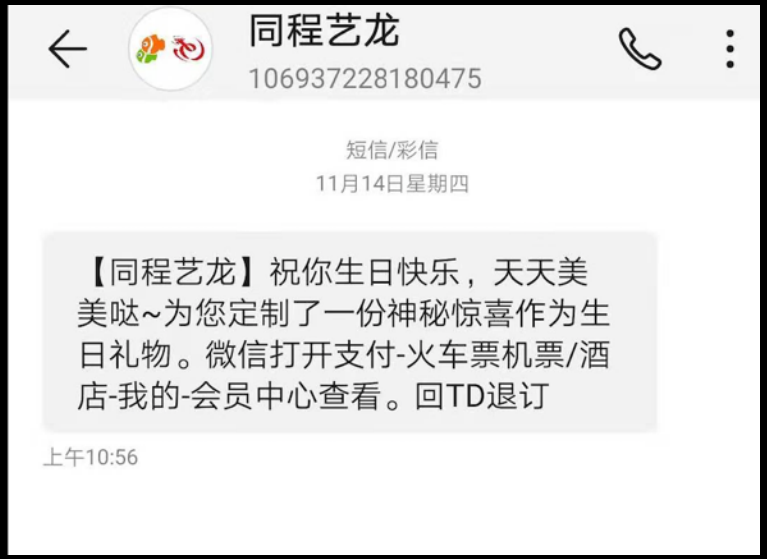
同程艺龙

目录

举例说明-case1

- 【实时】我的页面今日未签到弹屏

举例说明-case2



- 【离线】当天生日祝福短信

举例说明-case3

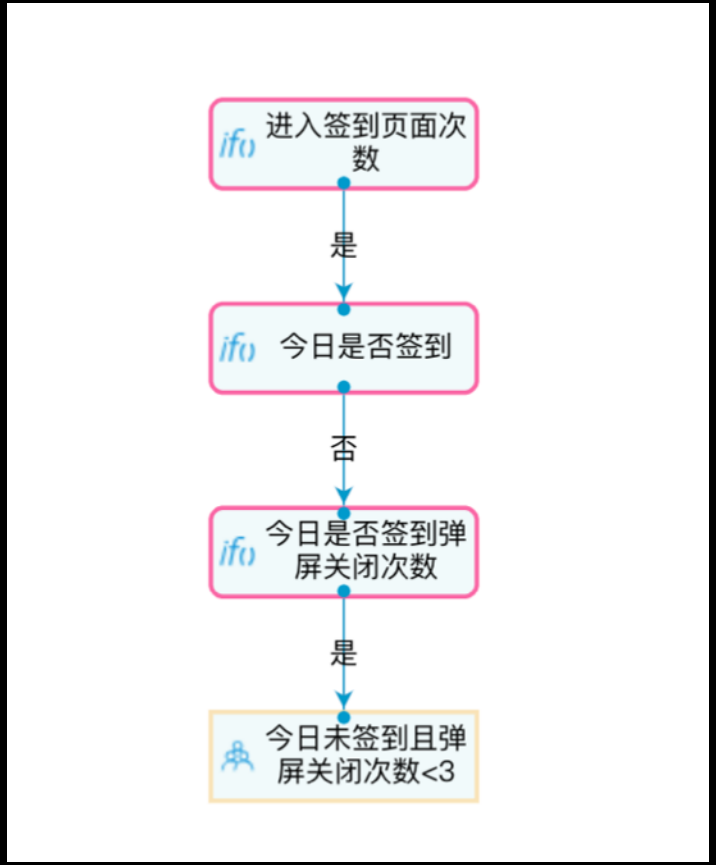
- 【延迟】里程抵现权益使用提醒短信

一般流程

业务痛点：

1. 流程长，修改维护困难
2. Case by case，硬编码
3. 数据分散，ETL口径不统一

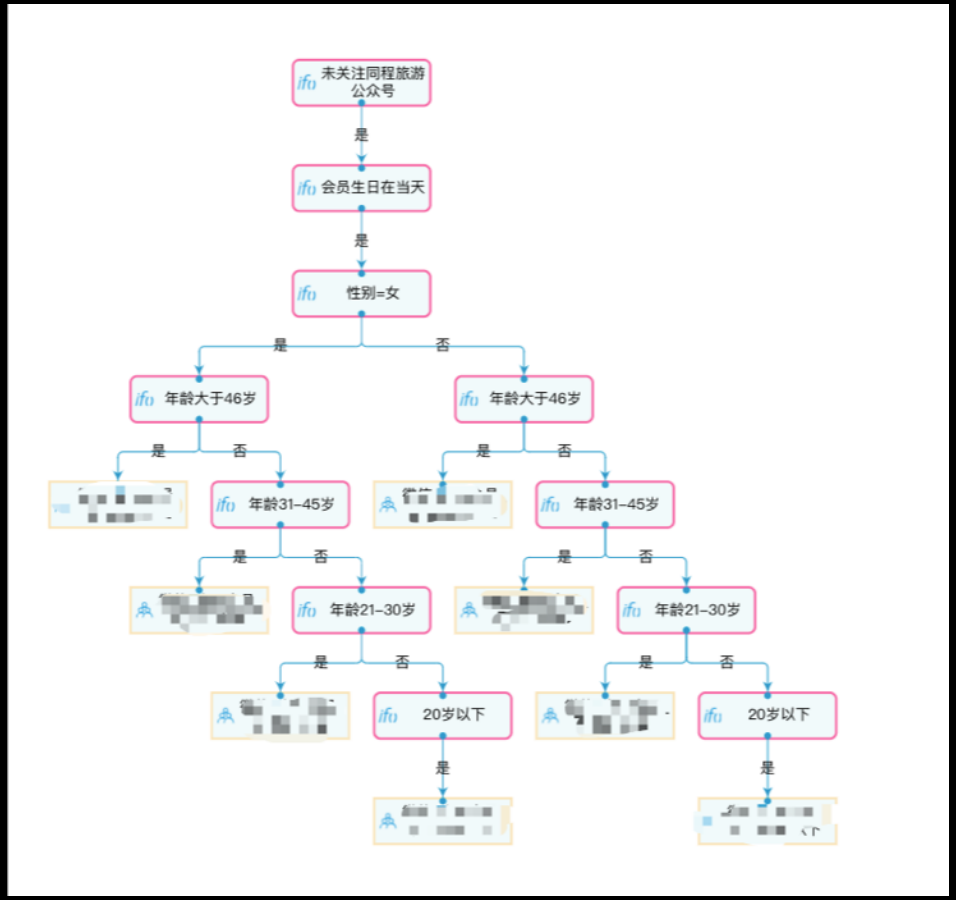
先知流程-【实时】我的页面今日未签到弹屏



Case1:

1. 进入【我的】页面
2. 今日未签到
3. 今日关闭弹屏次数小于3

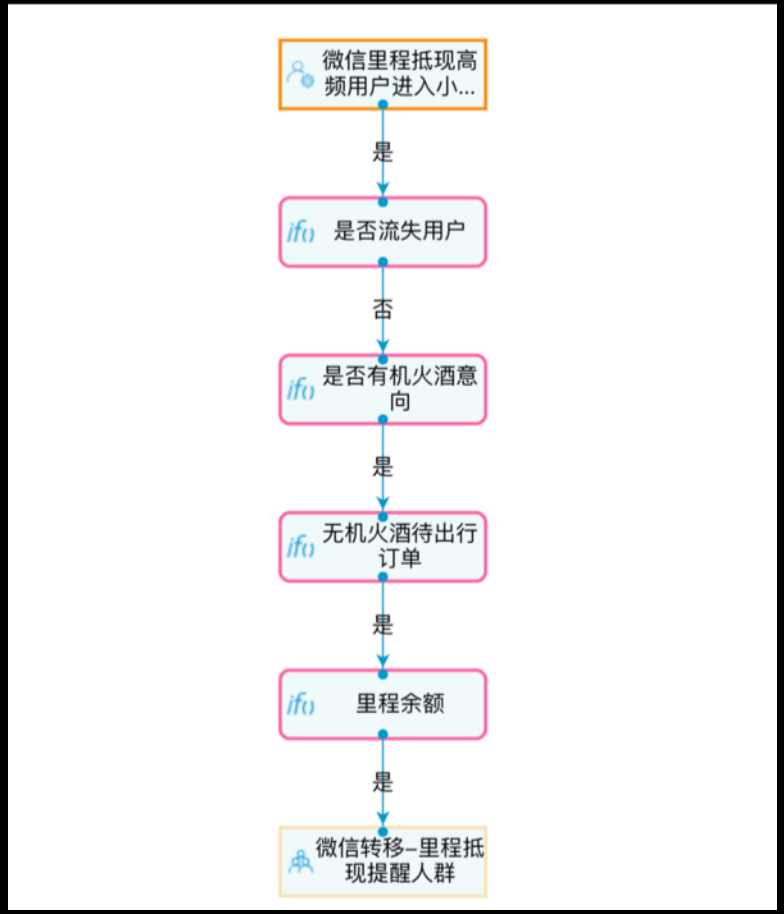
先知流程-【离线】当天生日祝福短信



Case2:

1. 当天过生日的会员
2. 根据性别不同年龄段发短信

先知流程-【延迟】里程抵现权益使用提醒短信



Case3:

1. 打开火车票tab页30分钟后
2. 不是流失用户
3. 浏览过机票火车票酒店页面
4. 没有待出行订单
5. 有里程余额

为什么设计先知

同程艺龙

先知提供一站式解决方案

先知优势



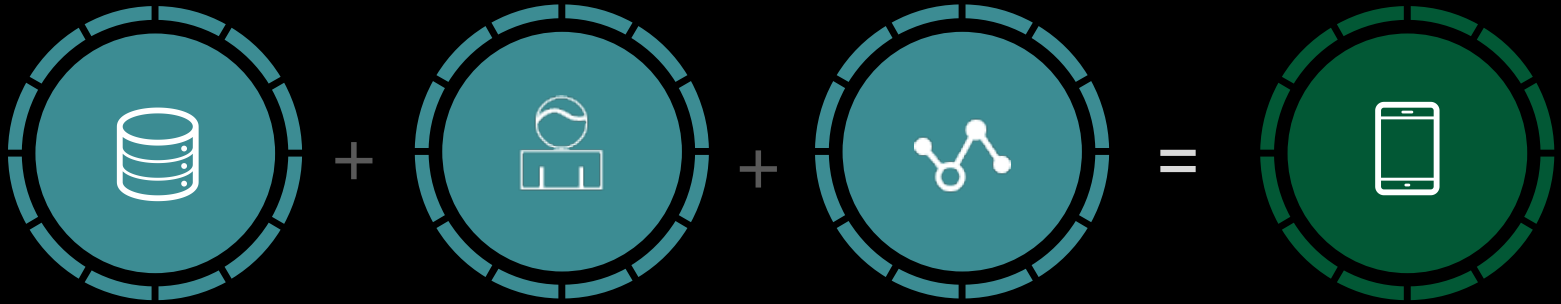
同程艺龙

1. 配置简单，上线快速
2. 数据全，扩展方便
3. 接入平台多，实现自动化
4. 智能化，支持算法模型
5. 性能强，数据处理秒级
6. 输出丰富，满足多种需求

同程艺龙

目录

先知-主要组成部分



先知-数据源设计

同程艺龙

数据源特点

低延迟

- 实时数据源全流程高效低延迟
- 订单冷热数据分离，保障查询效率

规范化

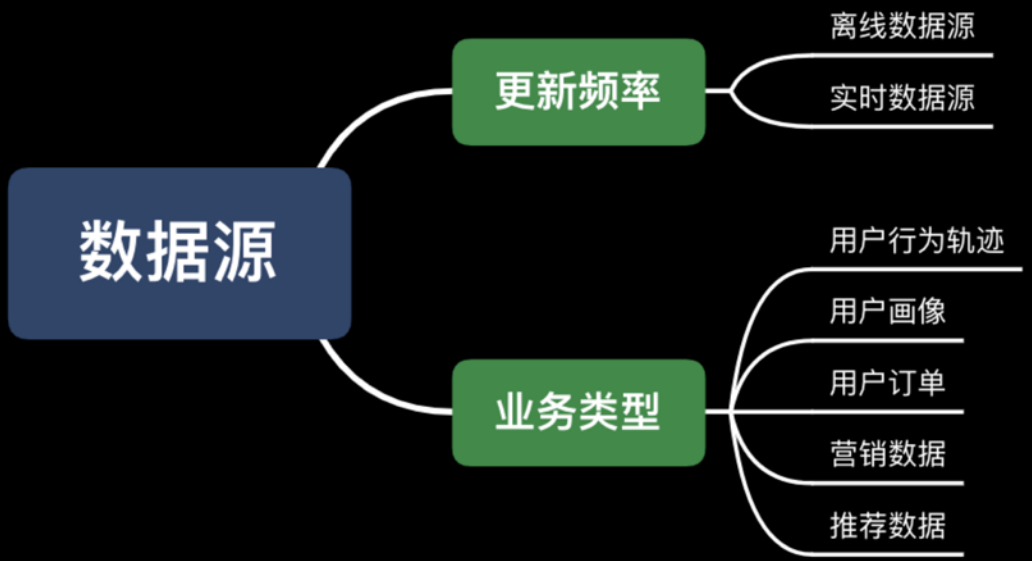
- 参照OneData标准，制定数据源规范，统一数据口径

自动化

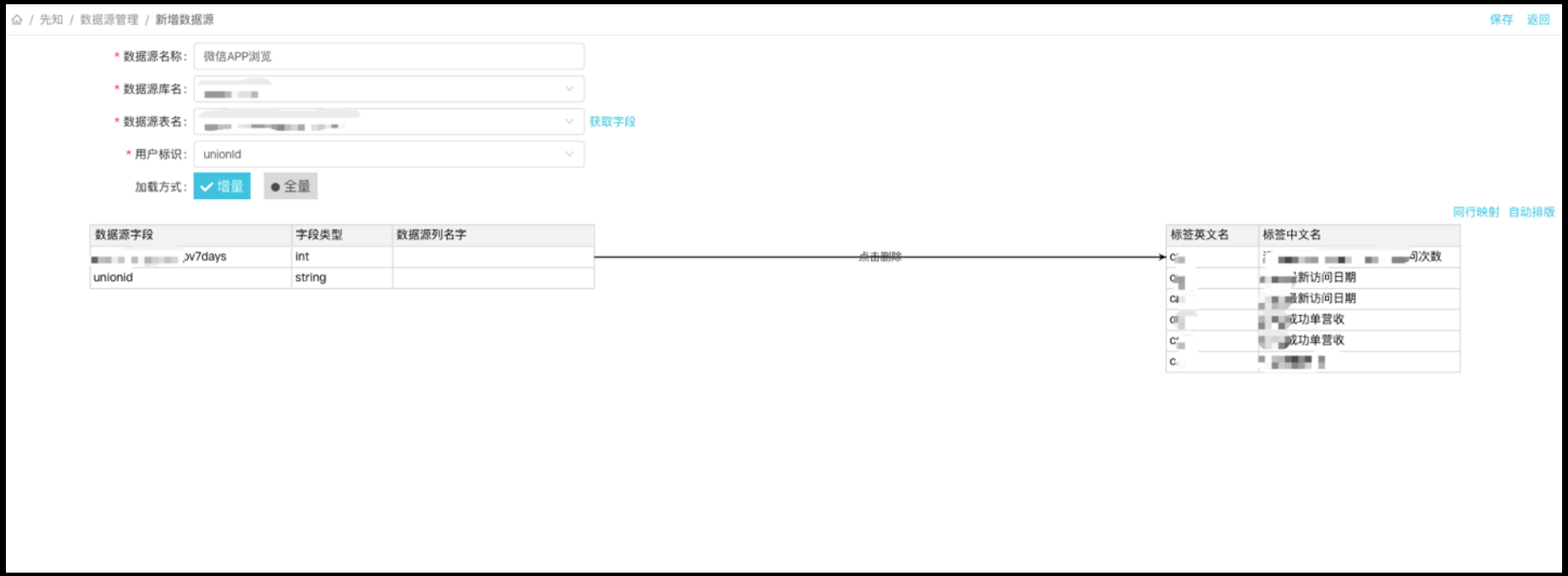
- 实时数据源接入自动化
- 离线数据源清洗完后导入自动化

数据源是先知的基础，稳定高效的数据源为整个系统提供了强有力的保障

先知-数据源分类



先知-数据源


/ 先知 / 数据源管理 / 新增数据源
保存
返回
* 数据源名称：
微信APP浏览
* 数据源库名：
* 数据源表名：
获取字段
* 用户标识：
unionid
加载方式：
增量
全量
同行映射
自动排版
数据源字段
字段类型
数据源列名字
v7days
int
unionid
string
点击删除
标签英文名
标签中文名
问次数
新访问日期
最新访问日期
成功单营收
成功单营收

先知-标签设计

同程艺龙

标签特点

高效率

- 实时标签配置即生成

规范化

- 参照OneData标准，规范分类
- 标签复用性高，避免重复造轮子

准确性

- 每条标签数据版本管理，保证数据的准确性

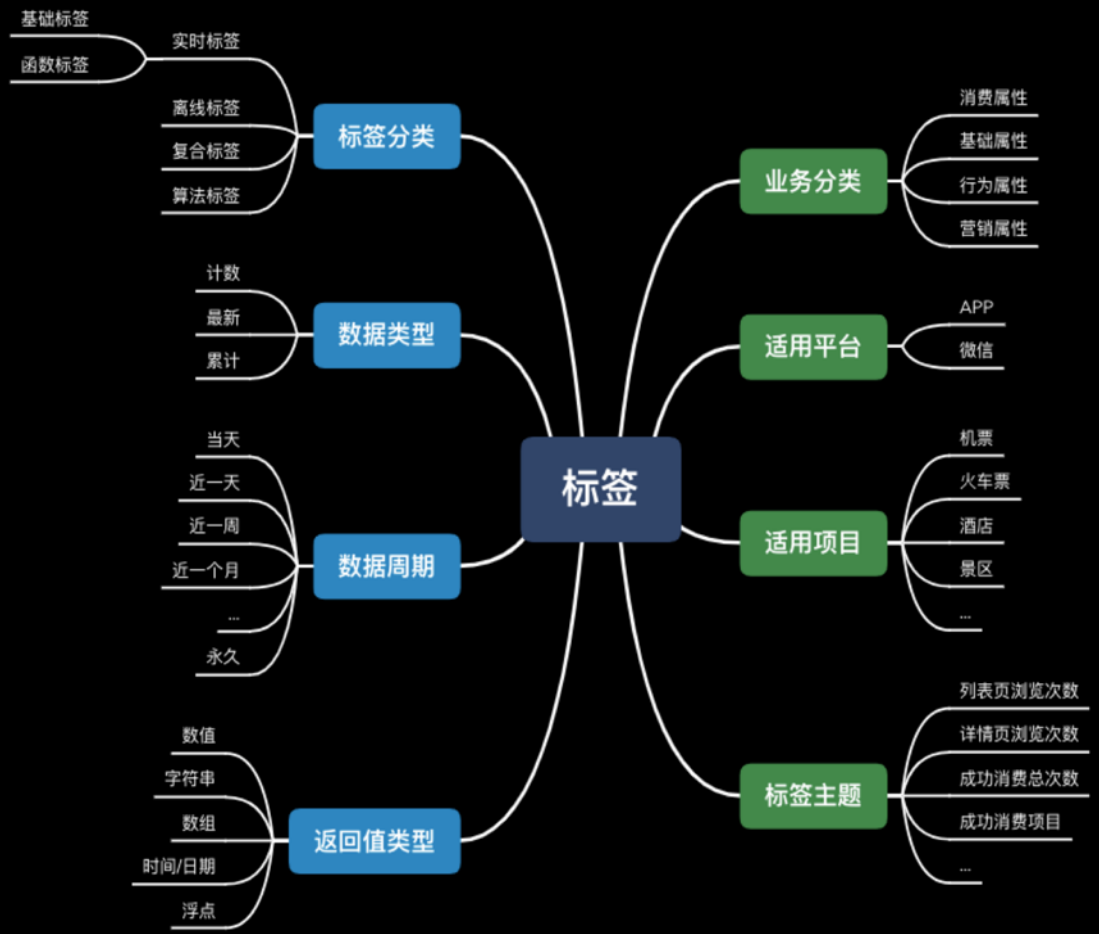
智能化

- 支持PMML算法标签
- 支持外部ETL标签导入

标签：

- 先知系统的核心模块，经过数据源的抽取计算得到。
- 业务范围覆盖流量，订单，会员，活动，红包，里程 等。

先知-标签分类体系


标签
标签分类
基础标签
实时标签
函数标签
离线标签
复合标签
算法标签
数据类型
计数
最新
累计
数据周期
当天
近一天
近一周
近一个月
...
永久
返回值类型
数值
字符串
数组
时间/日期
浮点
业务分类
消费属性
基础属性
行为属性
营销属性
适用平台
APP
微信
适用项目
机票
火车票
酒店
景区
...
标签主题
列表页浏览次数
详情页浏览次数
成功消费总次数
成功消费项目
...

先知-标签

同程艺龙

基础信息

标签ID：

标签英文名：co

* 标签分类：行为信息

创建人：

过滤条件　提取规则　函数配置

1、输入函数公式、测试参数、期望结果后，点击【测试】，对输入内容进行验证；
2、点击函数列表中函数名称，自动填入函数示例；
3、验证通过的函数会随标签自动保存；验证失败的不会随标签保存；
4、已配置函数的标签，函数不可删除，只可修改；

* 输入公式：

```
1  subResult(getByIndex(getByDateLatestOrEarlist(dateFilterByPeriodAndIndex(filterList
       ,3,generateDatetime(      ,generateDatetime(      ,"yyyy
     -MM-dd HH:mm:ss"),8,"latest","yyyy-MM-dd HH:mm:ss"),0),10)
```

请输入计算公式,测试通过时才可生效(请先参考函数及函数参数说明)

测试参数：

```
1
```

请输入测试参数(支持string,boolean,数组类型；有的函数无需测试参数)

期望结果：期望结果(目前期望结果支持string、boolean、list类型)

测试

与固定时间比较函数
英文名称：compareDateTimeByFix　返回值类型：class java.lang.Boolean　查看参数说明

与当前时间前(后)x天比较函数
英文名称：compareDateTimeByToday　返回值类型：class java.lang.Boolean　查看参数说明

按日期过滤函数
英文名称：dateFilterByPeriodAndIndex　返回值类型：interface java.util.List　查看参数说明

按时间戳过滤函数
英文名称：timestampFilterByPeriodAndIndex　返回值类型：interface java.util.List　查看参数说明

按序号删除
英文名称：cutString　返回值类型：class java.lang.Object　查看参数说明

按时间过滤函数
英文名称：timeFilterByStartAndEnd　返回值类型：interface java.util.List　查看参数说明

先知-场景设计

同程艺龙

场景特点

高效率

- 场景标签配置审核后立即生效

功能齐全

- 支持判断，ab分流，人群等多种节点
- 支持大于，小于，等于，时间比较，模糊匹配多种操作符
- 支持定制自定义函数

输出丰富

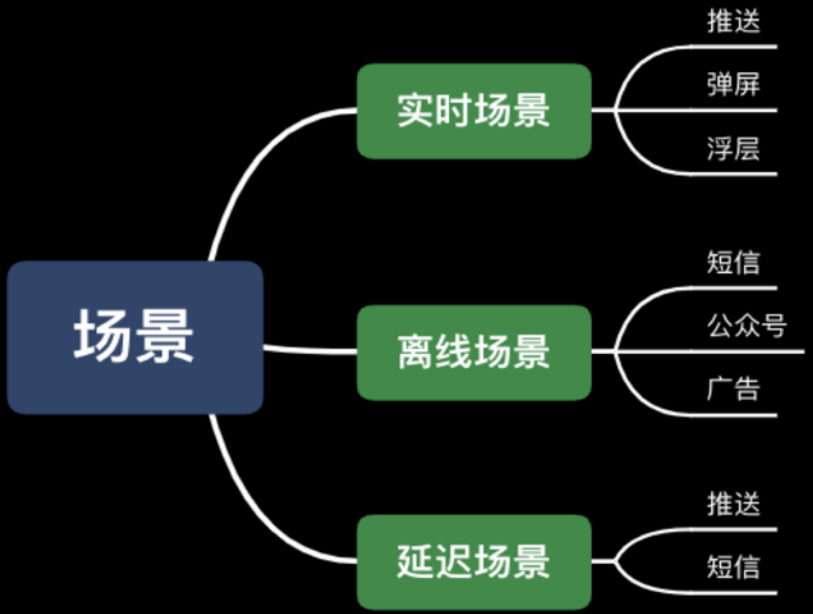
- 以接口，kafka，hdfs的方式输出人群包，可快速扩展
- 支持广告，公众号等多种输出场景，一站式解决用户分群、精准营销需求

场景：

- 场景即营销流程，场景也就是将各个营销节点进行个性化的组合。
- 标签组成节点，节点组成流程。

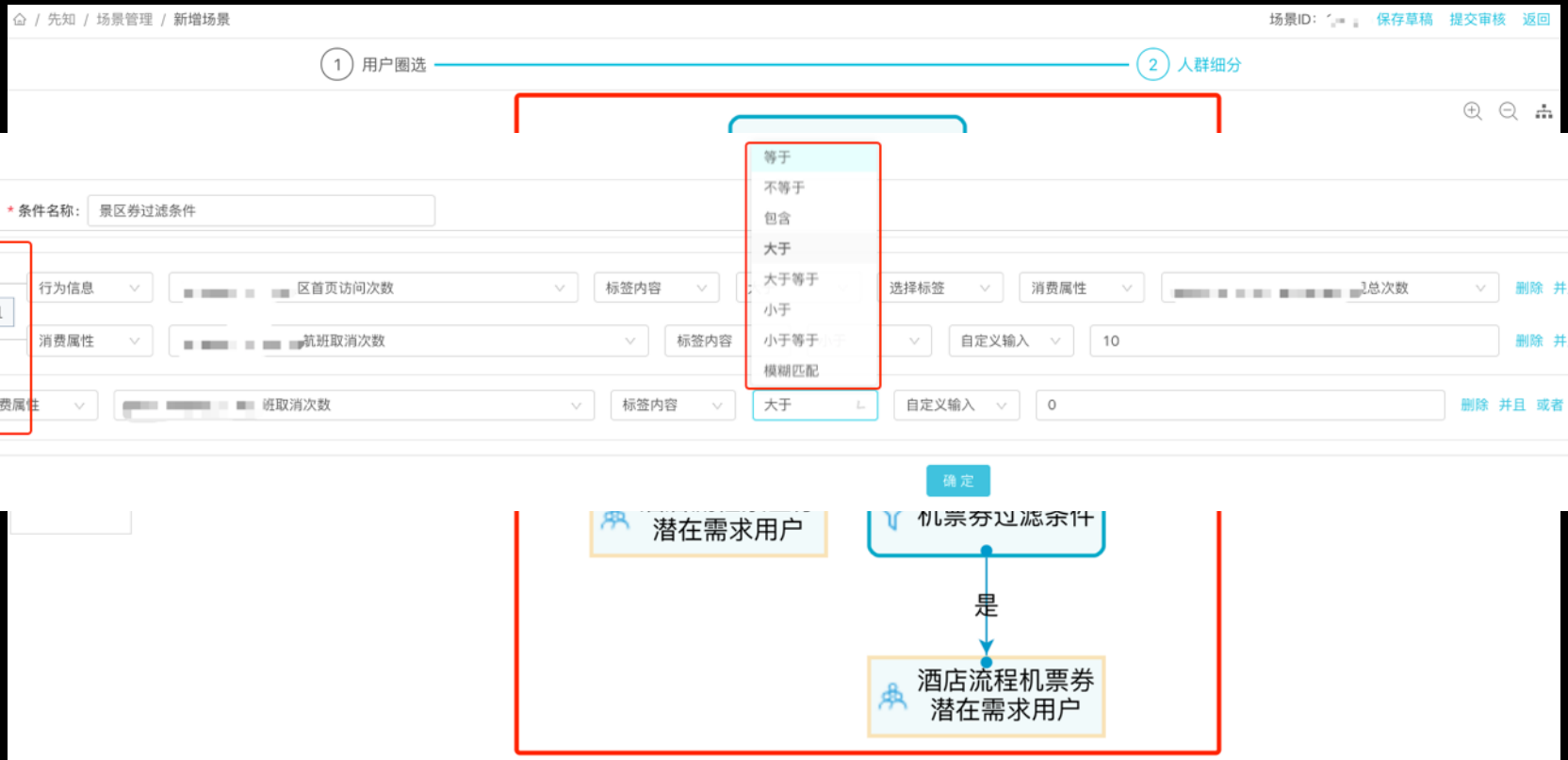
先知-场景分类

同程艺龙



先知-场景

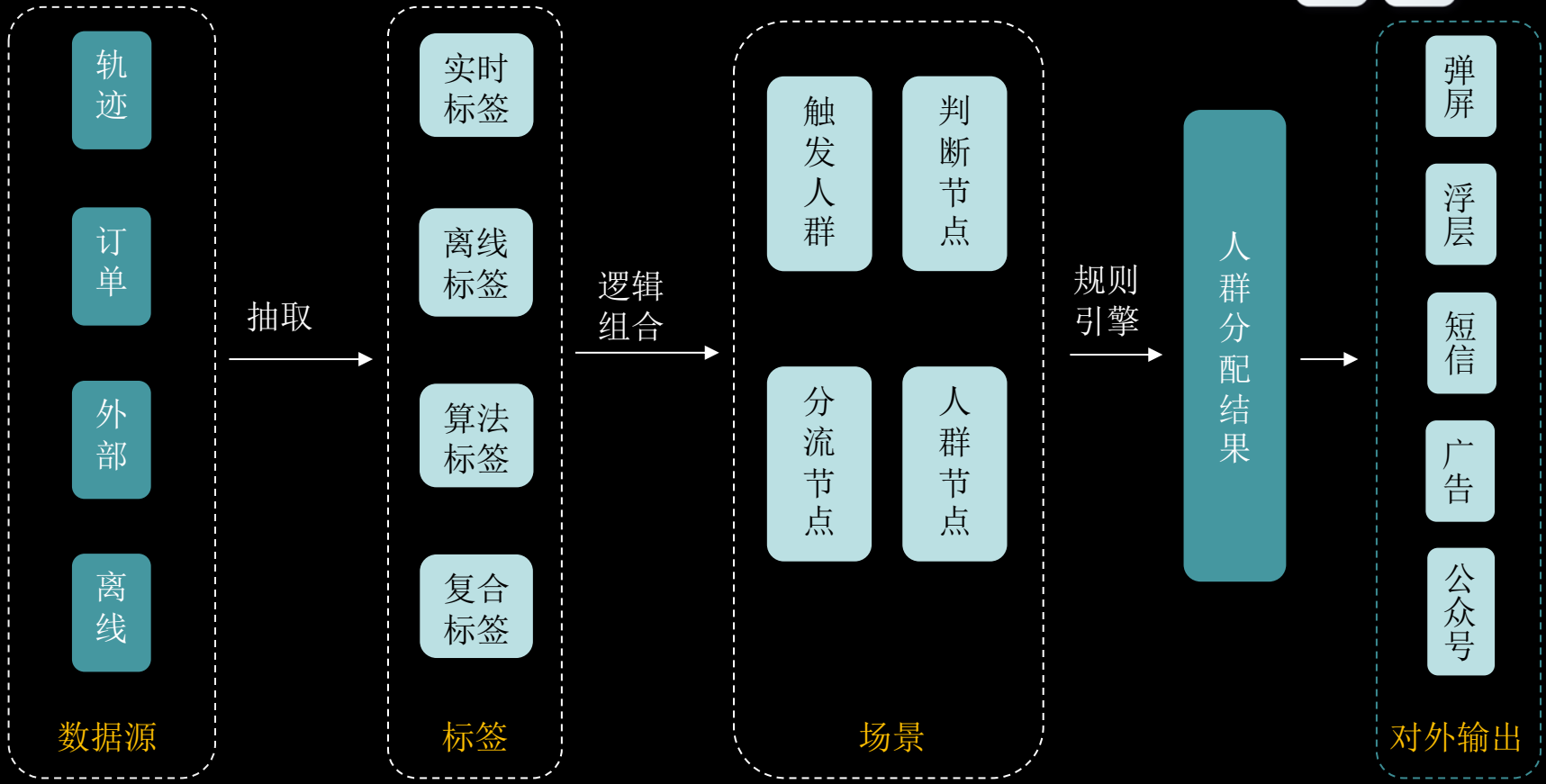
同程艺龙


先知 / 场景管理 / 新增场景
场景ID:
保存草稿
提交审核
返回
1 用户圈选
2 人群细分
节点信息
* 条件名称: 景区券过滤条件
或
且
行为信息
区首页访问次数
标签内容
选择标签
消费属性
总次数
删除 并且 或者
消费属性
班取消次数
标签内容
自定义输入
10
删除 并且 或者
消费属性
班取消次数
标签内容
大于
自定义输入
0
删除 并且 或者
等于
不等于
包含
大于
大于等于
小于
小于等于
模糊匹配
确 定
潜在需求用户
是
酒店流程机票券
潜在需求用户

先知-监控


酒店订单平均时延
对延实时监控
标签名称
唯一标识
今日增量
增量占比
增量环比昨日
用户标识覆盖量
用户标识覆盖率
操作
deviceid
78354
0.0%
0.0%
105482044
0.0%
增量曲线
微信_全项目_
增量波动曲线
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
增量值:3662079
占比:0.0%
增量值:3366423
占比:0.0%
增量值:3206150
占比:6.1%
增量值:3447704
占比:6.4%
增量值:2753057
占比:5.0%
增量值:2783588
占比:5.5%
增量值:2742963
占比:5.4%
增量值:2242073
占比:28.1%
增量值:2684748
占比:0.0%
2019-06-22
2019-07-14
2019-08-05
2019-08-27
2019-09-18
2019-10-12
2019-11-04
2019-11-27
2019-12-19
取消
确定
bus_order_app
scenic_isnew
usecar_order_app

先知-组成关系


轨迹
订单
外部
离线
数据源
抽取
实时标签
离线标签
算法标签
复合标签
标签
逻辑组合
触发人群
判断节点
分流节点
人群节点
场景
规则引擎
人群分配结果
弹屏
浮层
短信
广告
公众号
对外输出

目录

系统架构设计-技术框架选择

同程艺龙

计算引擎

Flink/Spark

计算引擎

- 标签计算，场景计算。
- 数据抽取，数据处理。
- 规则计算。

存储引擎

Redis/Hbase/hdfs/rocksdb

标签存储引擎

- 实时标签存储。
- 标签缓存信息，降低hbase的压力。
- 离线标签数据存储。

数据中转

Kafka

数据中转

- 标签的数据源。
- 标签和场景之间的数据缓冲。
- 场景计算结果和场景输出之间的数据缓冲。

结果存储

ES/HDFS

场景结果存储

- 场景结果输出，动态配置实时生效。
- 实时标签计算结果输出。

系统架构设计-规则引擎选择



同程艺龙

组件名称	优点	缺点
Drools	功能较为完善，自带系统监控、操作平台等功能	• 学习曲线陡峭，其引入的DRL语言较复杂，独立的系统很难进行二次开发 • 无法有效支持定时触达（如用户在浏览发生后30分钟触达支付条件判断）
Flink CEP	Flink自身支持	• 依靠定义Pattern实现，需要硬编码，无法实现动态规则 • 无法用在spark或者web项目中
Mvel	动态执行，更灵活	只是动态表达式，需要手动实现规则引擎

V1遇到的问题和解决思路

同程艺龙

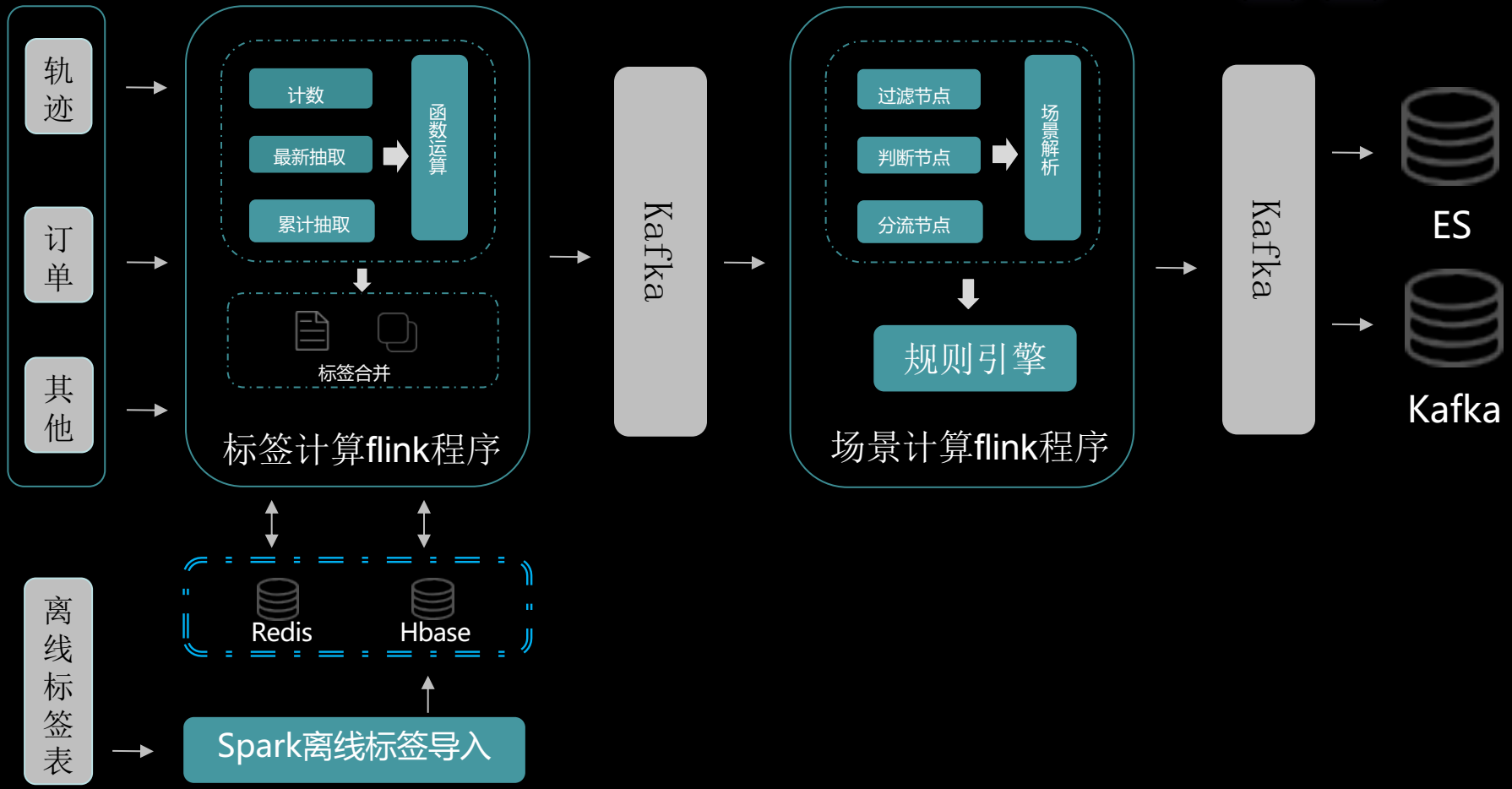
数据乱序

- 版本号：数据源，标签，场景均用时间戳作为版本号，全流程保证准确性
- kafka：同一用户发送到相同partition中
- Es&hbase：写入es/hbase数据带有版本号，只有高版本数据可以写入

数据结构设计

- 标签设计：尽量保证标签原子性，计数标签分天存储，灵活组合
- 场景设计：es数据一个场景对应一个索引存储，方便查询

系统架构-V1


轨迹
订单
其他
计数
最新抽取
累计抽取
函数运算
标签合并
标签计算flink程序
Kafka
过滤节点
判断节点
分流节点
场景解析
规则引擎
场景计算flink程序
Kafka
ES
Kafka
Redis
Hbase
离线标签表
Spark离线标签导入

V1 → V2遇到的问题和解决思路

同程艺龙

- 数据源数量翻倍
- 标签数量翻倍
- 标签初始化数据量过大

数据源过多，flink反压

■数据源拆分：按类型拆分，在上游标签计算发送下游进行标签合并，解决反压引起其他数据消费延迟

标签翻倍，存储有压力

■优化存储：String压缩；自动生成标签短英文名；实时标签拆分成基础标签和函数标签，只存储基础标签

■业务过滤：过滤无效流量，抽取不到标签的数据不发往下游

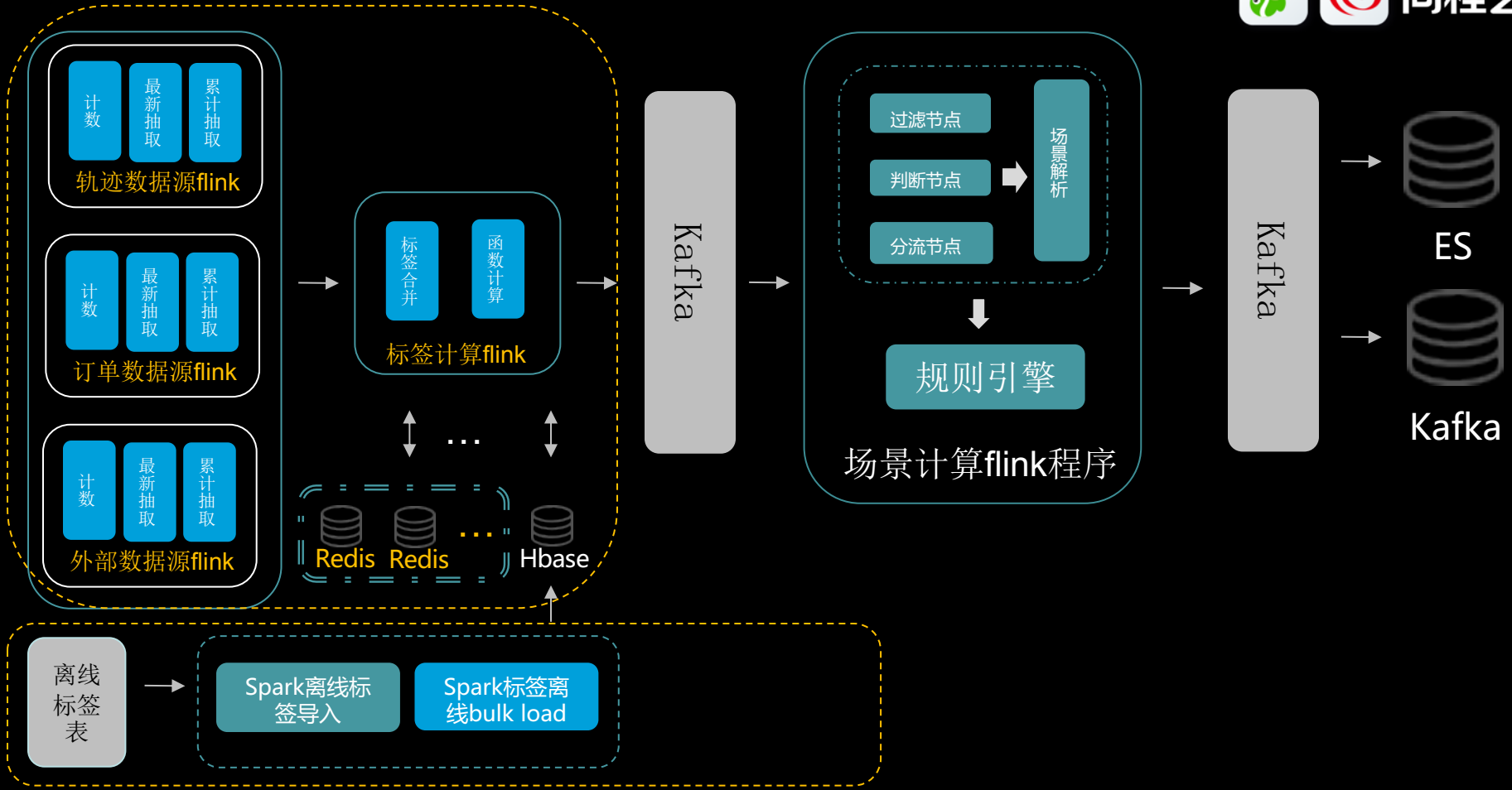
■扩充存储：采用多redis集群组存储，打样tikv代替redis

离线标签初始化过慢

■Bulkload：离线标签存储到hbase，bulkload的方式代替put，提高插入速度

■平台化：自动化操作初始化流程，无需手动跑任务

先知系统架构-V2


计数
最新抽取
累计抽取
轨迹数据源flink
计数
最新抽取
累计抽取
订单数据源flink
计数
最新抽取
累计抽取
外部数据源flink
标签合并
函数计算
标签计算flink
Redis
Redis
Hbase
Kafka
过滤节点
判断节点
分流节点
场景解析
规则引擎
场景计算flink程序
Kafka
ES
Kafka
离线标签表
Spark离线标签导入
Spark标签离线bulk load

V2 → V3遇到的问题和解决思路

同程艺龙

- 场景数量翻倍
- 新增离线场景
- 新增延迟触达

场景数量翻倍，写入查询性能下降

- 优化数据结构：优化场景es数据数据结构，按场景切分索引合并数据，降低写入和索引数量的同时提高查询性能
- 业务流量过滤：对人群结果没有变化的数据进行过滤，对变化的标签没有被对应场景使用的数据进行过滤

离线标签延迟导致离线场景延迟

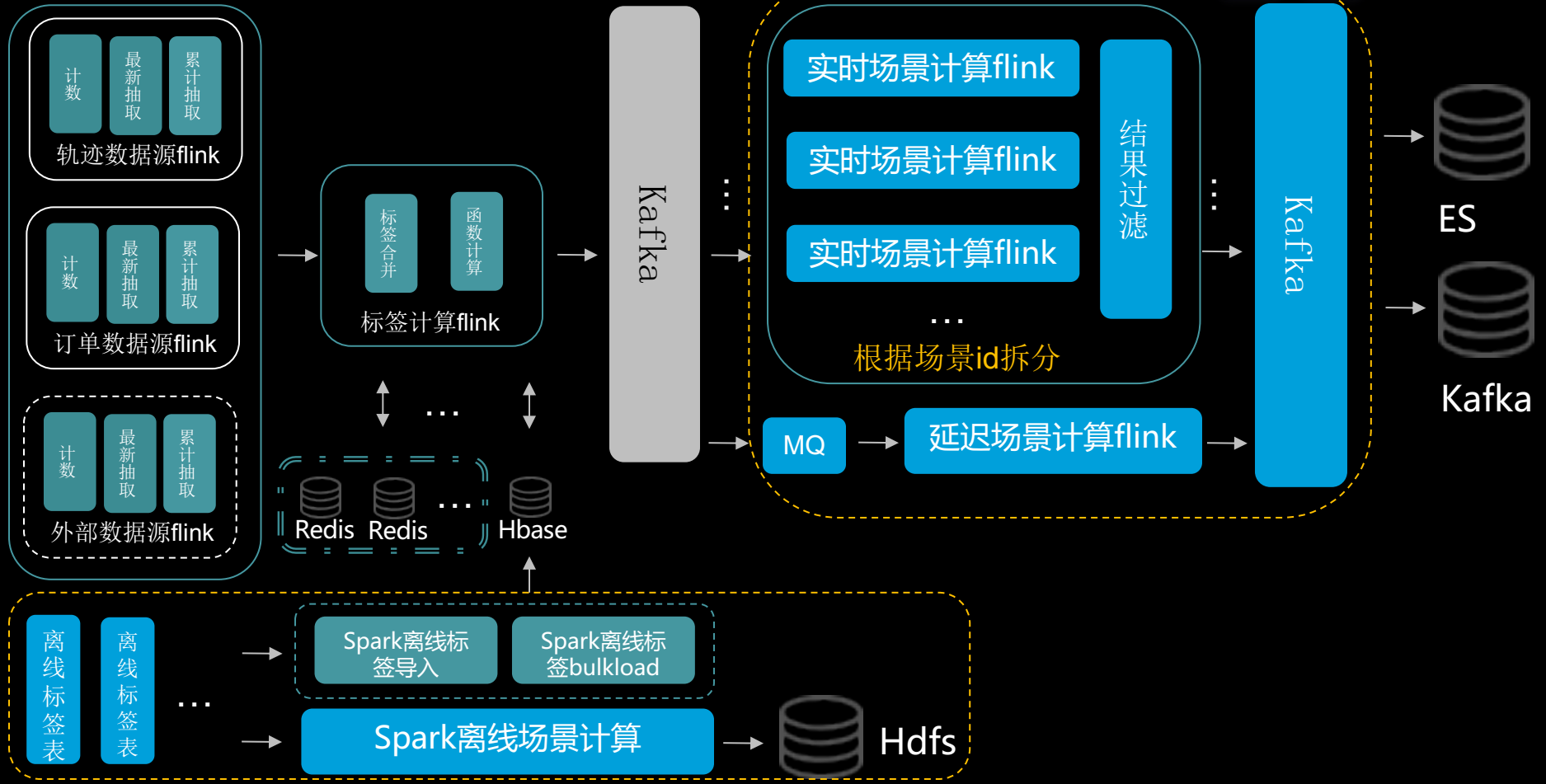
- 大表拆分：离线标签由一张大表拆分成多张小表，一张表etl完成后回调场景运行接口，判断当前可以运行的场景，按队列控制场景并发运行的数量

实现延迟触达

- 延迟队列：轨迹类延迟采用turboMq，对延迟数据和最新数据进行比较
- 定时job：订单类延迟采用定时job
- 特殊处理：对N分钟没有回来的场景，进行特殊处理

先知系统架构-V3

同程艺龙



小结与展望



同程艺龙

先知系统从研发到上线稳定运行已经近一年，取得了较好的效果：

1. 日均处理流量50亿+
2. 触达计算延迟秒级
3. 对外接口QPS5000，99线在30ms
4. 标签数500+，场景数200+
5. 建立了一套大而全的标签体系，平台化系统化地解决了智能营销方面的问题
6. 帮助业务线开展丰富运营活动，对转化率，GMV，拉新等指标进行了强力的支持
7. 未来系统规划主要往更智能化，自动化，全面化的方向发展

Q&A

谢谢聆听