# 用户行为挖掘项目预研

### 1.背景及目标

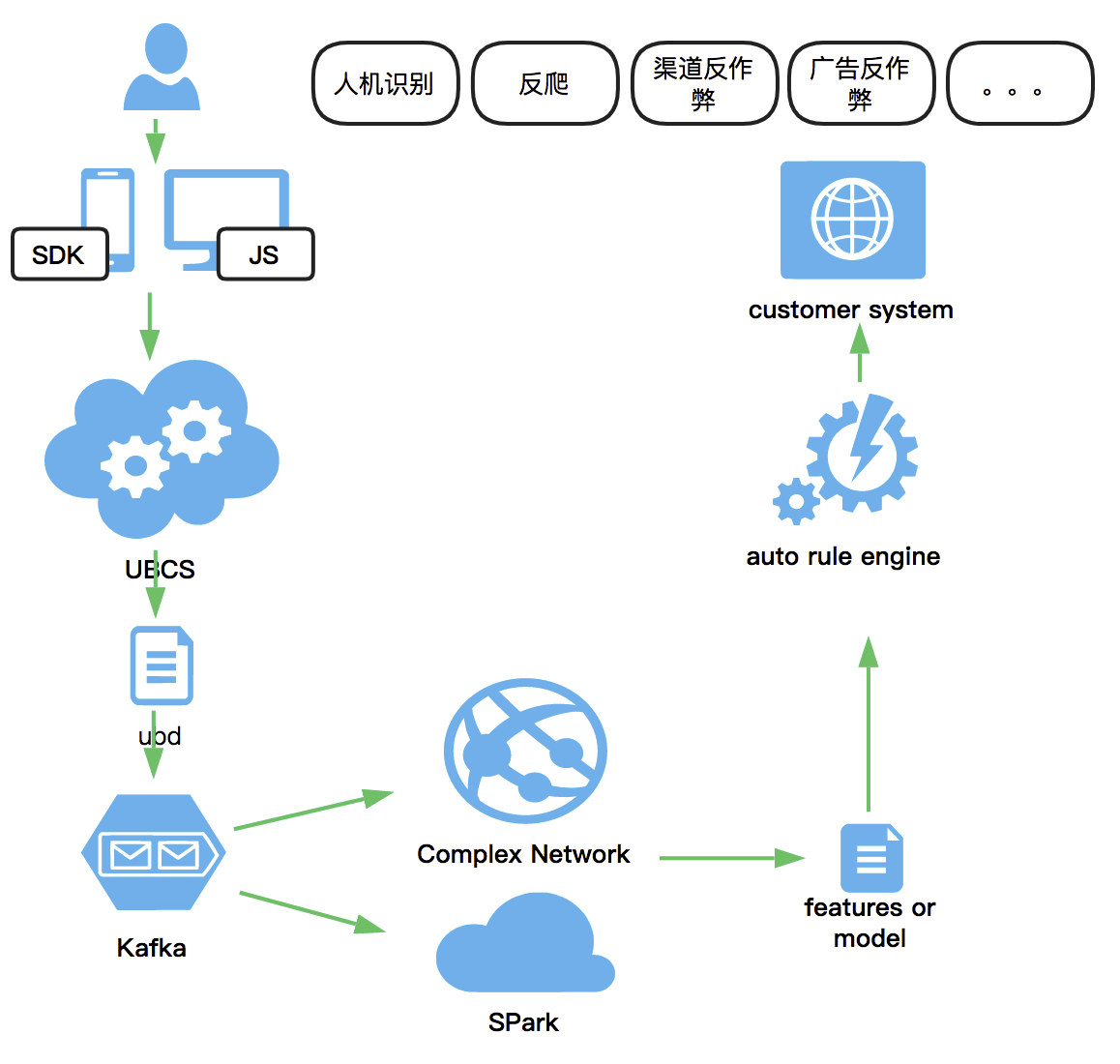
背景：同盾目前的产品体系及解决方案涉及设备指纹数据和事件数据，从基于用户行为角度来挖掘欺诈行为方向来说，用户行为数据显不足；有部分竞争对手使用本地化部署方案中，该方案使用流量镜像的方式可以采集到更多的用户行为数据。

目标：采集用户行为数据，并联合设备指纹数据、事件数据，通过机器学习算法挖掘出特征、训练出模型；产出或优化人机识别、反爬、渠道反作弊、广告反作弊等解决方案，同时也可做漏斗分析，帮助客户优化业务。

### 2.涉及模块

WEB采集模块、Android SDK、 IOSSDK、用户行为采集服务、ETL、用户行为挖掘建模、自动规则引擎

### 3.系统架构



### 4.用户行为采集字段

**采集字段列表**

|  |  |  |  |
| --- | --- | --- | --- |
|  | web | android | ios |
| 设备基础信息 [应用层] | * UA * 语言 * Accept * 字体列表 * 插件列表 * 浏览器权限   1. 是否启用cookie   2. 是否启用flash * refer | * SDK/APK 签名 * 字体列表 | * idfv/idfa/uuid * 应用版本 * 签名 |
| 设备基础信息 [操作系统层] | * 操作系统 | * Android设备API版本 * Android发行版本 * ROM主机/ROM标签 * 开机时间/运行时间/当前时间/休眠时间 * 是否ROOT * 语言/屏幕亮度/充电状态 * 基带版本/Linux内核版本 * HOOK信息/模拟器特征/运营包/安装包 | * 操作系统版本 * 开机时间/运行时间/当前时间/休眠时间 * 语言/屏幕亮度/充电状态 * 是否越狱 * 语言/时区 * 环境变量/动态链接库 * HOOK信息 |
| 设备基础信息 [网络层] | * TCP/IP协议栈   1. IP * 代理信息   1. 地址/端口/类型 | * TCP/IP协议栈 * WIFI * 移动网络   1. sim卡序列号   2. 运营商/网络   3. 基站信息 * 代理信息   1. 地址/端口/类型 | * TCP/IP协议栈 * WIFI * 移动网络   1. 运营商   2. 网络 * 代理信息   地址/端口/类型 |
| 设备基础信息[硬件层] | * 屏幕分辨率 * 显卡[帆布指纹] * 声卡 * 视频/音频设备数 | * IMEI/IMSI/MAC/serialNo * 品牌/型号/出厂名称/硬件平台 * 屏幕分辨率 * CPU架构/频率/平台 * 内存大小/硬盘大小 | * 屏幕分辨率 |
| 设备基础信息 [位置信息] | * IP 定位 | * IP 定位 * 基站定位 * WIFI定位 * GPS定位 | * IP 定位 * WIFI定位 * GPS定位 |
| 设备指纹相关信息 | * 版本 * 合作方 * 混淆模式 * 加密模式 * 校验模式 * 校验值 | * 版本 * 合作方 * 混淆模式 * 加密模式 * 校验模式 * 校验值 | * 版本 * 合作方 * 混淆模式 * 加密模式 * 校验模式 * 校验值 |
| 业务字段采集 | * 页面加载时间 * 点击控件名 * 点击控件坐标 * 点击控件时间 | * 页面加载时间 * 点击控件名 * 点击控件坐标 * 点击控件时间 * 传感器状态 | * 页面加载时间 * 点击控件名 * 点击控件坐标 * 点击控件时间 * 传感器状态 |

### 5.用户行为采集关键技术

* web JS 无痕采集方案：
* 页面加载ucb.js
* onReady()事件中对按需对页面的元素做事件绑定
* 用户行为发生时，采集用户行为数据发送到UBCS.
* Android无痕采集方案
* 参考：ActivityLifecycleCallbacks注入ClickTracker
* IOS 无痕采集方案
  + 。。。

### 6.数据处理

Kafka+交换平台+Spark

### 7.数据模型

人机识别模型：

反爬模型：

渠道反作弊模型：

广告反作弊模型：

### 8.困难点和挑战

页面和客户端改版