DSP系统日志ETL格式

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 版本 | 作者 | 说明 |
| 2016.6.12 | V1.0 | set\_daemon@126.com | 初版，日志格式版本为V1.0，只记录单个广告位的数据，注意：对于被横线划掉的字段，表示数据不用填写，但是保持该字段 |
|  |  |  |  |

目录

[1 前言 3](#_Toc453491901)

[2 技术规范 3](#_Toc453491902)

[3 ETL日志 3](#_Toc453491903)

[3.1 投放数据整合表 3](#_Toc453491904)

1 前言

本文主要定义原始日志经过整合后，生成清洗可用的不同中间数据，以满足不同数据业务需求。

2 技术规范

以离线的方式经过清洗、转换后存储在hadoop系统中。

3 ETL日志

3.1 投放数据整合表

依赖：ad\_request、ad\_push、ad\_win、ad\_show、ad\_click，以ad\_push的push id以及竞价id串联这些日志数据，即未发生push的request将不会生成该日志数据。

|  |  |  |
| --- | --- | --- |
| **字段名** | **字段类型** | **说明** |
| **日志系统段（8段）** | | |
| log\_name | string | etl\_push\_log |
| log\_ver | string | 该日志类型的版本号：V1.0 |
| push\_id | string | 广告推送id，用于标识一个投放事务 |
| ad\_request | int | 0或者1，表示存在该类型的日志 |
| ad\_push | int | 0或者1，表示存在该类型的日志 |
| ad\_win | int | 0或者1，表示存在该类型的日志 |
| ad\_show | int | 0或者1，表示存在该类型的日志 |
| ad\_click | int | 0或者1，表示存在该类型的日志 |
| log\_time | string | 格式“2016.12.03 10:12:12.1234”。生成本日志的的时间 |
| log\_src\_ip | string | Ipv4格式，例如“192.168.1.1”。生产日志的服务器ip |
| log\_src\_serv | string | 生产日志的服务名称 |
| request\_time | string | 格式“2016.12.03 10:12:12.1234”。请求时间 |
| push\_time | string | 格式“2016.12.03 10:12:12.1234”。投放时间 |
| win\_time | string | 格式“2016.12.03 10:12:12.1234”。竞价成功时间 |
| show\_time | string | 格式“2016.12.03 10:12:12.1234”。曝光时间 |
| click\_time | string | 格式“2016.12.03 10:12:12.1234”。点击时间 |
|  |  |  |
| **流量资源相关** | | |
| flow\_source | string | 流量来源，如“BES”。 |
| flow\_type | string | 流量类型，如”pc”、”mob”、”ott” |
| flow\_detail | string | 流量详情，如“browser”、”app”，对于PC流量分别指”浏览器“和”客户端“，对于mob流量分别指”wap“和”app“ |
| flow\_src\_ip | string | Ipv4格式，例如“192.168.1.1”。流量来源ip，这里指adx服务器的ip。 |
| deal\_type | string | 交易类型：rtb/pdb/pa/pd |
| bid\_id | string | 竞价请求id |
| imp\_id | string | 曝光id |
| slot\_id | string | 广告位id |
| slot\_size | string | 广告位尺寸“宽x高” |
| slot\_type | string | 广告位类型 |
| slot\_billtype | string | 广告位计费类型：通常是cpm |
| slot\_floor | float | 广告位底价，单位为元 |
| slot\_pos | integer | 广告位所处屏（对浏览器而言） |
| **受众信息** | | |
| user\_id | string | Dsp的用户id |
| third\_user\_id | string | 第三方（ADX）的用户id |
| user\_ip | string | Ipv4格式，例如“192.168.1.1”。受众ip。 |
| user\_area | string | 4字节十六进制格式。地域id |
| user\_geo | string | 地理位置，”经度,纬度”，其中经度和纬度均为浮点数 |
| user\_timezone | string | 受众所处时区 |
| third\_user\_tags | string | 第三方的受众标签数据，以逗号分隔 |
| third\_user\_interest | string | 第三方的受众兴趣数据，以逗号分隔 |
|  |  |  |
|  |  |  |
| **网站信息** | | |
| site\_name | string | 网站名称 |
| site | String | 网站地址 |
| page\_url | string | 受众访问的页面地址 |
| ref\_url | string | 受众访问的页面来源地址 |
| page\_keywords | string | 页面关键词，以逗号分隔 |
| page\_cat | string | 第三方的页面分类，以逗号分隔 |
| page\_channel | string | 第三方的页面频道（通常在视频网站分的比较清楚） |
|  |  |  |
| **app信息** | | |
| app\_name | string | App名称 |
| App\_ver | String | App版本 |
| App\_pkg | String | App包名 |
| App\_down\_url | String | App下载地址 |
| App\_publisher | String | App发布商 |
| App\_paid | String | App是否为付费软件：y或者n |
| App\_keywords | string | App关键词，以逗号分隔 |
| App\_cat | String | App分类，以逗号分隔 |
|  |  |  |
| **设备信息** | | |
| dev\_type | string | 设备类型：pc/iphone/mac/ipad/pad/phone |
| dev\_brand | string | 设备品牌 |
| Dev\_model | string | 设备型号 |
| Dev\_os | string | 设备系统 |
| Dev\_os\_ver | string | 设备系统版本 |
| Dev\_network | string | 设备联网类型 |
| Dev\_isp | string | 设备运营商 |
| Dev\_size | string | 设备屏幕尺寸 |
| Dev\_px\_ratio | string | 设备屏幕分辨率 |
| Dev\_ua | string | 客户端（browser/app）的user agent |
| idfa | string | 设备的idfa，格式：”编码方案,数据“ |
| mac | string | 设备的mac地址，格式：”编码方案,数据“ |
| imei | string | Imei号，格式：”编码方案,数据“ |
| android\_id | string | Android系统的id，格式：”编码方案,数据“ |
|  |  |  |
| **广告信息（11）** |  |  |
| order\_id | string | 订单id |
| plan\_id | string | 计划id |
| group\_id | string | 策略组id |
| ad\_id | string | 广告id |
| sponsor\_id | string | 广告主id |
| agency\_id | string | 代理商id |
| discount | float | 折扣 |
| ad\_cost | float | 广告成本价格：元 |
| bid\_price | float | 竞价价格：元 |
| pred\_ctr | float | 预估ctr值 |
| win\_price | float | 竞价成功价格 |
|  |  |  |