**DMP 项目需求**

# 指标含义

|  |  |
| --- | --- |
| **指标** | **定义** |
| **参与竞价数** | 本日收到 ADX 平台发来的竞价请求并成功响应次数 |
| **竞价成功数** | 在本日内成功竞价的次数 |
| **竞价成功率** | 竞价成功率=竞价成功数/参与竞价数 |
| **展示量（曝光）** | 广告在终端被显示的数量 |
| **点击量** | 广告展示后被终端用户点击的数量 |
| **点击率** | 点击率=点击量/展示量 |
| **ECPC** | eCPC=成本/点击量 |
| **ECPM** | eCPM=成本/展示量\*1000 |
| **消费** | 收取广告主支付的用于广告投放的费用 |
| **成本** | 广告花费在渠道与媒体上的费用 |
| **毛利** | 毛利=消费-成本 |

1. **日志字段属性说明**

|  |  |  |
| --- | --- | --- |
| **序号** | **属性名称** | **描述** |
| **1** | sessionid: String, | 会话标识 |
| **2** | advertisersid: Int, | 广告主 id |
| **3** | adorderid: Int, | 广告 id |
| **4** | adcreativeid: Int, | 广告创意 id ( >= 200000 : dsp) |
| **5** | adplatformproviderid: Int, | 广告平台商 id (>= 100000: rtb) |
| **6** | sdkversion: String, | sdk 版本号 |
| **7** | adplatformkey: String, | 平台商 key |
| **8** | putinmodeltype: Int, | 针对广告主的投放模式,1：展示量投放 2：点击量投放 |
| **9** | requestmode: Int, | 数据请求方式（1:请求、2:展示、3:点击） |
| **10** | adprice: Double, | 广告价格 |
| **11** | adppprice: Double, | 平台商价格 |
| **12** | requestdate: String, | 请求时间,格式为：yyyy-m-dd hh:mm:ss |
| **13** | ip: String, | 设备用户的真实 ip 地址 |
| **14** | appid: String, | 应用 id |
| **15** | appname: String, | 应用名称 |

|  |  |  |
| --- | --- | --- |
| **16** | uuid: String, | 设备唯一标识 |
| **17** | device: String, | 设备型号，如 htc、iphone |
| **18** | client: Int, | 设备类型 （1：android 2：ios 3：wp） |
| **19** | osversion: String, | 设备操作系统版本 |
| **20** | density: String, | 设备屏幕的密度 |
| **21** | pw: Int, | 设备屏幕宽度 |
| **22** | ph: Int, | 设备屏幕高度 |
| **23** | long: String, | 设备所在经度 |
| **24** | lat: String, | 设备所在纬度 |
| **25** | provincename: String, | 设备所在省份名称 |
| **26** | cityname: String, | 设备所在城市名称 |
| **27** | ispid: Int, | 运营商 id |
| **28** | ispname: String, | 运营商名称 |
| **29** | networkmannerid: Int, | 联网方式 id |
| **30** | networkmannername:  String, | 联网方式名称 |
| **31** | iseffective: Int, | 有效标识（有效指可以正常计费的）(0：无效 1：有效 |
| **32** | isbilling: Int, | 是否收费（0：未收费 1：已收费） |
| **33** | adspacetype: Int, | 广告位类型（1：banner 2：插屏 3：全屏） |
| **34** | adspacetypename: String, | 广告位类型名称（banner、插屏、全屏） |
| **35** | devicetype: Int, | 设备类型（1：手机 2：平板） |
| **36** | processnode: Int, | 流程节点（1：请求量 kpi 2：有效请求 3：广告请求） |
| **37** | apptype: Int, | 应用类型 id |
| **38** | district: String, | 设备所在县名称 |
| **39** | paymode: Int, | 针对平台商的支付模式，1：展示量投放(CPM) 2：点击 |
| **40** | isbid: Int, | 是否 rtb |
| **41** | bidprice: Double, | rtb 竞价价格 |
| **42** | winprice: Double, | rtb 竞价成功价格 |
| **43** | iswin: Int, | 是否竞价成功 |
| **44** | cur: String, | values:usd|rmb 等 |
| **45** | rate: Double, | 汇率 |
| **46** | cnywinprice: Double, | rtb 竞价成功转换成人民币的价格 |
| **47** | imei: String, | Imei（移动设备识别码） |
| **48** | mac: String, | Mac（苹果设备） |
| **49** | idfa: String, | Idfa（广告标识符） |
| **50** | openudid: String, | Openudid（UDID的第三方解决方案） |
| **51** | androidid: String, | Androidid（安卓设备的唯一id） |
| **52** | rtbprovince: String, | rtb 省 |
| **53** | rtbcity: String, | rtb 市 |
| **54** | rtbdistrict: String, | rtb 区 |
| **55** | rtbstreet: String, | rtb 街 道 |
| **56** | storeurl: String, | app 的市场下载地址 |
| **57** | realip: String, | 真实 ip |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **58** | | | isqualityapp: Int, | | | | 优选标识 | | | |
| **59** | | | bidfloor: Double, | | | | 底价 | | | |
| **60** | | | aw: Int, | | | | 广告位的宽 | | | |
| **61** | | | ah: Int, | | | | 广告位的高 | | | |
| **62** | | | imeimd5: String, | | | | imei\_md5 | | | |
| **63** | | | macmd5: String, | | | | mac\_md5 | | | |
| **64** | | | idfamd5: String, | | | | idfa\_md5 | | | |
| **65** | | | openudidmd5: String, | | | | openudid\_md5 | | | |
| **66** | | | androididmd5: String, | | | | androidid\_md5 | | | |
| **67** | | | imeisha1: String, | | | | imei\_sha1 | | | |
| **68** | | | macsha1: String, | | | | mac\_sha1 | | | |
| **69** | | | idfasha1: String, | | | | idfa\_sha1 | | | |
| **70** | | | openudidsha1: String, | | | | openudid\_sha1 | | | |
| **71** | | | androididsha1: String, | | | | androidid\_sha1 | | | |
| **72** | | | uuidunknow: String, | | | | uuid\_unknow tanx 密文 | | | |
|  | **73** |  |  | decuuidunknow: String, | |  |  | 解密的 tanx 明文 | |  |
| **74** | | | userid: String, | | | | 平台用户 id | | | |
|  | **75** |  |  | reqdate: String, |  | |  | 日期 |  | |
|  | **76** |  |  | reqhour: String, |  | |  | 小时 |  | |
| **77** | | | iptype: Int, | | | | 表示 ip 类型 | | | |
| **78** | | | initbidprice: Double, | | | | 初始出价 | | | |
| **79** | | | adpayment: Double, | | | | 转换后的广告消费 | | | |
| **80** | | | agentrate: Double, | | | | 代理商利润率 | | | |
| **81** | | | lomarkrate: Double, | | | | 代理利润率 | | | |
| **82** | | | adxrate: Double, | | | | 媒介利润率 | | | |
| **83** | | | title: String, | | | | 标题 | | | |
| **84** | | | keywords: String, | | | | 关键字 | | | |
| **85** | | | tagid: String, | | | | 广告位标识(当视频流量时值为视频 ID 号) | | | |
| **86** | | | callbackdate: String, | | | | 回调时间 格式为:YYYY/mm/dd hh:mm:ss | | | |
| **87** | | | channelid: String, | | | | 频道 ID | | | |
| **88** | | | mediatype: Int | | | | 媒体类型：1 长尾媒体 2 视频媒体 3 独立媒体 默认:1 | | | |

1. **需求**

## 日志转Parquet 文件

1. 要求一： 将数据转换成 parquet 文件格式
2. 要求二： 序列化方式采用 KryoSerializer 方式
3. ~~要求三： parquet 文件采用 Snappy 压缩方式~~

## 报表

* + 1. **统计各省市数据量分布情况**

要求一：

将统计的结果输出成 json 格式，并输出到磁盘目录。

{"ct":943,"provincename":"内蒙古自治区","cityname":"阿拉善盟"}

{"ct":578,"provincename":"内蒙古自治区","cityname":"未知"}

{"ct":262632,"provincename":"北京市","cityname":"北京市"}

{"ct":1583,"provincename":"台湾省","cityname":"未知"}

{"ct":53786,"provincename":"吉林省","cityname":"长春市"}

{"ct":41311,"provincename":"吉林省","cityname":"吉林市"}

{"ct":15158,"provincename":"吉林省","cityname":"延边朝鲜族自治州"}

要求二：

将结果写到到 mysql 数据库。

要求三：

用 spark 算子的方式实现上述的统计，存储到磁盘。

* + 1. **地域分布**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **省市/城市** | **原 始 请**  **求** | **有效请求** | **广告请求** | **参与竞价**  **数** | **竞价成功**  **数** | **竞价成功**  **率** | **展示**  **量** | **点击**  **量** | **点击**  **率** | **广 告 成**  **本** | **广 告 消**  **费** |
| **-A 省** |  |  |  | 1000 | 600 | 30.56 | 800 | 800 | 2 | 15.87 | 15.87 |
| **B 市** | 200 | 100 | 30.56 | 178 | 178 | 2 | 15.87 | 15.87 |
| **C 市** | 100 | 50 | 30.56 | 78 | 78 | 2 | 15.87 | 15.87 |
| **D 市** | 400 | 200 | 30.56 | 324 | 324 | 2 | 15.87 | 15.87 |
| **F 市** | 300 | 250 | 30.56 | 167 | 167 | 2 | 15.87 | 15.87 |
| **+B 省** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |

* + 1. **终端设备**

运营商：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **运营商** | **总请求** | **有效请求** | **广告请求** | **参与竞价**  **数** | **竞价成功**  **数** | **竞价成功**  **率** | **展示**  **量** | **点击**  **量** | **点击**  **率** | **广 告 成**  **本** | **广 告 消**  **费** |
| **电信** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **联通** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **移动** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **其他** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |

网络类型：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **网 络 类**  **型** | **总请求** | **有效请求** | **广告请求** | **参与竞价**  **数** | **竞价成功**  **数** | **竞价成功**  **率** | **展示**  **量** | **点击**  **量** | **点击**  **率** | **广 告 成**  **本** | **广 告 消**  **费** |
| **2G** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **3G** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **4G** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **WiFi** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **其他** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |

设备类型：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **设 备 类**  **型** | **总请求** | **有效请求** | **广告请求** | **参与竞价**  **数** | **竞价成功**  **数** | **竞价成功**  **率** | **展示**  **量** | **点击**  **量** | **点击**  **率** | **广 告 成**  **本** | **广 告 消**  **费** |
| **手机** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **平板** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **其他** |  |  |  |  |  |  |  |  |  |  |  |

操作系统：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **操 作 系**  **统** | **总请求** | **有效请求** | **广告请求** | **参与竞价**  **数** | **竞价成功**  **数** | **竞价成功**  **率** | **展示**  **量** | **点击**  **量** | **点击**  **率** | **广 告 成**  **本** | **广 告 消**  **费** |
| **Andrio**  **d** |  |  |  | 1000 | 600 | 30.56 | 800 | 800 | 2 | 15.87 | 15.87 |
| **IOS** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **其他** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |

* + 1. **媒体分析**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **媒 体 类**  **别** | **总请求** | **有效请求** | **广告请求** | **参与竞价**  **数** | **竞价成功**  **数** | **竞价成功**  **率** | **展示**  **量** | **点击**  **量** | **点击**  **率** | **广 告 成**  **本** | **广 告 消**  **费** |
| **爱奇艺** |  |  |  | 1000 | 600 | 30.56 | 800 | 800 | 2 | 15.87 | 15.87 |
| **腾 讯 新**  **闻** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |
| **PPTV** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |

* + 1. **渠道报表**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **渠 道 名**  **称** | **总请求** | **有效请求** | **广告请求** | **参与竞价**  **数** | **竞价成功**  **数** | **竞价成功**  **率** | **展示**  **量** | **点击**  **量** | **点击**  **率** | **广 告 成**  **本** | **广 告 消**  **费** |
| **Tanx** |  |  |  | 1000 | 600 | 30.56 | 800 | 800 | 2 | 15.87 | 15.87 |
| **搜狐** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **芒 果**  **RTB** |  |  |  | 1000 | 600 | 23.23 | 700 | 323 | 4 | 98.21 | 50.81 |

## 数据标签化

过滤数据，将每天数据打上以下标签：

1. 广告位类型（标签格式： LC03->1 或者 LC16->1）xx 为数字，小于 10 补 0，把广告位类型名称，LN 插屏->1
2. App 名称（标签格式： APPxxxx->1）xxxx 为 App 名称，使用缓存文件 appname\_dict 进行名称转换；APP 爱奇艺->1
3. 渠道（标签格式： CNxxxx->1）xxxx 为渠道 ID(adplatformproviderid)
4. 设备：
   1. (操作系统 -> 1)
   2. (联网方 -> 1)
   3. (运营商 -> 1)

设备操作系统

1 Android D00010001

2 IOS D00010002

3 WinPhone D00010003

\_ 其 他 D00010004

设 备 联 网 方 式WIFI D00020001 4G D00020002

3G D00020003

2G D00020004

\_ D00020005

设备运营商方式

移 动 D00030001 联 通 D00030002 电 信 D00030003

\_ D00030004

1. 关键字（标签格式：Kxxx->1）xxx 为关键字，关键字个数不能少于 3 个字符，且不能

超过 8 个字符；关键字中如包含‘‘|’’，则分割成数组，转化成多个关键字标签

1. 地域标签（省标签格式：ZPxxx->1, 地市标签格式: ZCxxx->1）xxx 为省或市名称
2. 商圈标签
3. 上下文标签： 读取日志文件，将数据打上上述 6 类标签，并根据用户 ID 进行当前文件的合并，数据保存格式为：userId K 青云志:3 D00030002:1 ……

## 统一用户识别

利 用 Spark Graphx 进 行 统 一 用 户 的 识 别 ， 参 考 示 例 ： [https://github.com/apache/spark/blob/master/examples/src/main/scala/org/apache/sp](https://github.com/apache/spark/blob/master/examples/src/main/scala/org/apache/spark/examples/graphx/ConnectedComponentsExample.scala)

ark/examples/graphx/ConnectedComponentsExample.scala 样例数据：

<https://github.com/apache/spark/tree/master/data/graphx>

图计算后数据保存格式：UID:8931238,mac:xxxxx,idfa:xxsadxx……

## 上下文标签合并

将统一用户数据和上下文标签数据进行合并(join 操作)，产生当天最终的用户标签数据。