中国联通

大数据公司-生产多租户

生产平台多租户模型部署说明文档

用户画像项目合作

## 项目背景

通过分析用户DPI数据，评估用户在单位时间内的访问网站行为分析，相当于用户画像.（**此项目用于营销程序**）

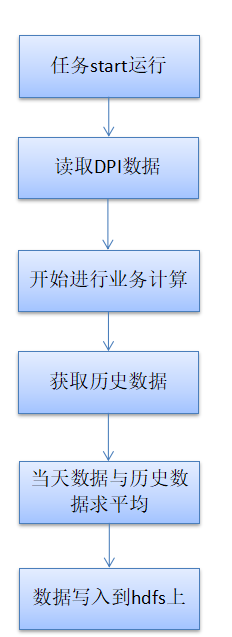
## 项目目标

给全量用户进行打标签，并统计出标签的打开次数。从而获取用户的浏览行为

## 模型业务逻辑说明

1. 读取全量用户数据（关于北京，上海地区），并进行数据初步计算。
2. 对计算结果的url进行标签化处理
3. 进行业务逻辑计算

## 模型运行流程图



## 硬件及软件环境要求

1. 硬件要求：目前尚不清楚，后期进行商谈；
2. 软件环境：如hdfs，spark；

生产环境软件版本声明：

Hadoop 2.6.0-cdh5.5.4  
 hive   1.1.0  
 Python 2.7.5   
 HBase1.0.0-cdh5.5.4  
 spark 1.5.0    
 scala 2.10.4    
 jdk1.7

## 生产平台模型数据输入

原始的dpi 数据

1. 原始流量上网日志表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 中文名 | 字段名 | 位置 | 加密说明 | 类型 | 备注 |
| 1 | 手机号 | device\_number | 0 | MD5加密 | String |  |
| 2 | 位置区编码 | lac | 1 | MD5加密 | String |  |
| 3 | CI号码 | ci | 2 | MD5加密 | String |  |
| 4 | 终端类型 | imei | 3 | MD5加密 | String | IMEI |
| 5 | 流量类型 | DPI\_TYPE | 4 |  | String |  |
| 6 | 开始时间 | START\_TIME | 5 |  | String | YYYY-MM-DD HH:MM:SS.1234567，精确到0.1微妙 |
| 7 | 结束时间 | END\_TIME | 6 |  | String |  |
| 8 | 持续时长 | CDR\_DURA | 7 |  | String | 以秒为单位 |
| 9 | 上行流量 | UP\_FLOW | 8 |  | String | 以bytes为单位 |
| 10 | 下行流量 | DOWN\_FLOW | 9 |  | String | 以bytes为单位 |
| 11 | 总流量 | TOTAL\_FLOW | 10 |  | String | 以bytes为单位 |
| 12 | 网络类型 | RAT\_TYPE | 11 |  | String |  |
| 13 | 终端IP | TERMINAL\_IP | 12 |  | String | 用户每次请求和应答的IP地址 |
| 14 | 访问IP | TARGET\_IP | 13 |  | String | 通信对端IP地址 |
| 15 | 状态码 | STATUS\_CODE | 14 |  | String |  |
| 16 | User Agent | USER\_AGENT | 15 |  | String |  |
| 17 | APN | APN | 16 |  | String |  |
| 18 | IMSI | IMSI | 17 | MD5加密 | String |  |
| 19 | SGSN IP | SGSN\_IP | 18 |  | String |  |
| 20 | GGSN IP | GGSN\_IP | 19 |  | String |  |
| 21 | 内容类型 | CONTENT\_TYPE | 20 |  | String |  |
| 22 | 源端口 | SOURCE\_PORT | 21 |  | String |  |
| 23 | 目的端口 | TARGET\_PORT | 22 |  | String |  |
| 24 | 记录标识 | RECORD\_MEGTYPE | 23 |  | String |  |
| 25 | 合并记录数 | merge\_records | 24 |  | String |  |
| 26 | 网址/特征信息 | url | 25 |  | String |  |

注：涉及根据租户提供host进行过滤用户上网日志的数据存储目录，以/prov\_id /day\_id /\*.gz结尾为统一存储路径，day\_id为日期，prov\_id为三位省份代码。例如/011/ 20161001/\*.gz

## 合作伙伴模型数据输入

无

## 模型输出

模型的输出是按天进行输出，输出的路径目前是 hdfs 上的文件夹，目前没有表结构的说法

1. 用户画像项目，输出介绍

key|value

key：imei号

value ：对数据计算进行一个拼接

（数字 表示的就是数字，字母表示的就是小数）

## 程序部署说明

主要程序部署说明及执行说明

1.部署说明：

程序以jar的形式提交，需要依赖配置文件（文件名），依赖第三方jar包（jar包名和版本）。

1. 执行说明：

①关于hadoop的执行语句样例如下：

hadoop jar xxx.jar 主类参数或者配置文件 input output

参数1 xxx.jar 部署程序的jar包

参数2 主类 可以在外部指定，也可在jar包中指定

参数3 参数或者配置文件配置文件可以在外部指定，也可在jar包中指定

参数4 input 输入路径

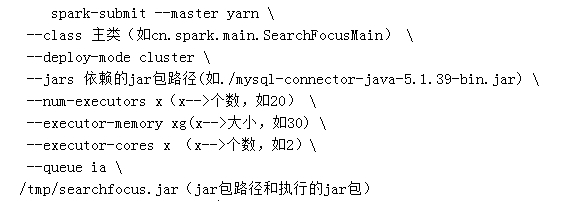
参数5 output 输出路径

**注意：(1) mapreduce的queuename设置为由部署人员指定，请设置成参数**

**(2)job\_name请写成参数，由部署人员填写**

**(3)数据输入输出路径请写成参数,由部署人员进行指定**

②关于spark-submit的执行语句样例如下，可根据需要进行增减：

**

1. 模型脚本规范：

部署人员不再对租户脚本进行修改，采用脚本外传入参数的形式操作：

sh 租户模型脚本.sh 账期 省份 资源队列名 任务名称 输入目录 输出目录 配置文件目录(没有置空) 分区类型(没有置空) 其余可添加参数

参数说明如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 参数名 | 格式 | 说明 |
| 1 | 账期 | yyyyMMdd | 8位时间账期：20171111 |
| 2 | 省份 | 0XX | 三位省份代代码，如北京011；全国为099 |
| 3 | yarn资源队列名 |  | 由部署人员指定 |
| 4 | 任务名称 |  | 由部署人员指定 |
| 5 | 输入目录 |  | 多个输入目录间逗号分割，程序中写死可置空 |
| 6 | 输出目录 |  | 必须指定最终数据目录或者表名 |
| 7 | 配置文件目录 |  | 多个配置目录间逗号分割，没有可置空 |

## 程序部署时间计划

主要说明程序部署时间计划；

附件：

|  |  |  |
| --- | --- | --- |
| 预估项 | 个数或者大小 | 备注 |
| 模型最终结果大小 | XX G/天 |  |
| 模型最终结果文件个数 | XX个/天 |  |
| 模型中间结果大小 | XX G/月 |  |
| 模型中间结果文件个数 | XX个/月 |  |
| 缓存结果大小 | XX G/天 |  |
| 缓存结果文件个数 | XX 个/天 |  |