**数据分析结构体系文档**

（本文档只适用于中科云智内部交流）

**文档信息**

|  |  |
| --- | --- |
| **项目名称** | 数据分析结构体系文档 |
| **版本** | Version 1.0 |
| **日期** | 2015年3月11日 |
| **作者** | 朱文文 李琦 |

**修改历史记录**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **日期** | **版本号** | **修改说明** | **修改人** | **核准人** |
|  |  |  |  |  |

目录

[1. 引言 4](#_Toc414553294)

[2. 总体架构图 4](#_Toc414553295)

[3. 功能模块 5](#_Toc414553296)

### 引言

本文档旨在介绍数据分析现有的结构体系，各个模块之间的联系及实现过程。

#### 编写目的

本文档的阅读群体是数据分析组的全部成员，目的是为了让所有成员对公司的业务体系有个深刻的了解，同时也便于日后工作的开展。

#### 设计原则

从分光点每天采集数据，对原始数据进行预处理，然后生成统一的数据格式提取出各种功能数据，到功能数据应用等模块，分别进行描述。

#### 名词解释

clk数据：是指用户真实的点击数据

access数据：是指通过向客户端插入js提取到的具有title等信息的网页数据

nstat数据：是指采集到的全量数据

#### 概述

本文档共包括六个部分，

### 总体架构图

Figure 1为数据分析总体架构图，主要包括数据采集、数据预处理、数据处理、数据存储、数据应用等模块。

格式：流程（输入输出路径）、处理过程 实现

存放路径，格式，输入输出，处理流程 补充细节部分，具体实现步骤方法

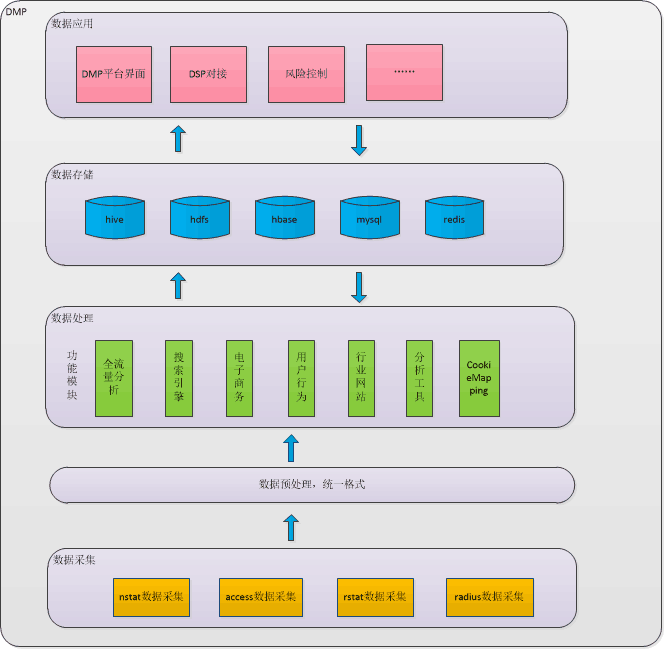


Figure 1 总体架构图

### 功能模块

#### 数据采集

##### nstat数据

模块描述

nstat数据是指直接在分光点网卡上采集到的数据，包括PC端和移动端两种类型的用户访问日志。

接口实现

nstat数据共有11个字段，分隔符为”\001”, 示例：

t6401\0012015-03-12 10:59:17\001DEFAULT\_NSTAT\001115.231.156.59\001124.238.253.102\001news.china.com\001/xml/important\_news\_day.json?b1426129160734=1\001http://military.china.com/zh\_cn/\001uid=128690323.651980766135674400.1417337948964.555\001浙江省杭州市\001Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0)

Table 1 字段格式

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **名称解释** |
| 1 | area | 地区代号 |
| 2 | atime | 时间 |
| 3 | pname | 数据类型 |
| 4 | sip | 源IP |
| 5 | dip | 目的IP |
| 6 | host | 主机名 |
| 7 | url | url |
| 8 | refer | 来路 |
| 9 | cookie | 用户标识符 |
| 10 | loc | 用户IP归属地 |
| 11 | user\_agent | user\_agent |

设计实现

在分光机器上/ipush下，存放了采集nstat数据所需要的进程文件，在n\*.sh中配置好采集比例、地区代码等参数，确保b\*.sh已运行，执行n\*.sh就可以进行数据采集。将push\_nstat.sh加入crontab, 并确保本机到目的地址机器的无密码登录，数据会自动拷贝到分析机器上。

流程图



Figure nstat数据采集流程图

安装部署

nstat采集程序部署在分光机器上，在路径/ipush/下，直接运行sh n.sh就可以开始数据采集，然后同步到分析机器上。

##### access数据

模块描述

access数据是指通过向客户端插入js提取到的具有title等信息的网页数据。

接口实现

access数据分隔符为空格，示例：

60.190.148.234 js.3tip.com - [10/Mar/2015:23:57:01 +0800] "GET /re/re.php?src=t6401&t=&ci=1019122922&r= HTTP/1.1" 200 0 "http://my.tv.sohu.com/us/232423049/78815806.shtml?lcode=AAAARpiVK1t7\_PU3RD8OneDxnhRwUEL9FGNRevctYNKyRZEfjfSInIqsUW7VydQm3iHXcVhlZ2DMoVDIyn8Aaqy\_PFVOstv6KKyko6..d4y&lqd=14042" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)" "uaid=57081b7e5c92dd938fc2ca91004c5845030"

Table 2 字段格式

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **名称解释** |
| 1 | proxy\_ip | 地区代号 |
| 2 | domain | 域名 |
| 3 | time | 时间 |
| 4 | GET | GET信息 |
| 5 | user\_agent | user\_agent |
| 9 | cookie | 用户标识符 |

设计实现

在广告投放服务器上，配置采集策略及比例，并将采集到的数据以天为单位ftp拷贝到分析服务器，分析机器需要部署vsftp, 并设置传输路径、用户名及密码等。

流程图



Figure access数据采集流程图

安装部署

access采集程序部署在广告投放服务器上，然后拷贝到分析机器上。

##### radius数据

模块描述

radius数据是指用户拨号上网在Radius认证服务器产生的数据。

RADIUS是运营商用来验证用户身份、动态分配ip使用的协议。每个用户有唯一的RADIUSID。从RADIUS原始报文里可以提取出RADIUSID、IP、STATUS字段。STATUS 有三种1 上线，2 下线，3 心跳。依次记录用户每次会话原始报文里的这三个字段以及这次会话的时间戳，就可以知道一个RADIUSID在特定时间使用的IP是多少或者一个IP在特定时间是被分配到哪个RADIUSID。

RADIUS数据采集负责从SIG RDS或RADIUS CENTER读取上述三个字段。SIG RDS或RADIUS CENTER发出的数据流是从原始报文实时解析得到的，所以可以把读取某条数据的时间作为该条数据的时间戳。每条数据包括RADIUSID、IP、STATUS三个字段。

接口实现

RADIUS数据采集对外提供的数据有两种，实时数据、离线数据。当前有RADIUS数据的省份有两个浙江、江苏。虽然浙江发出RADIUS数据流的是SIG RDS，江苏发出RADIUS数据流的是RADIUS CENTER。但是对外提供的数据接口基本上是一致的。

浙江的实时RADIUS存储在Memcache中，江苏的实时RADIUS存储在Redis中，以RADIUSID作为key，可以查出此时该RADIUSID使用的IP，以IP为key，可以查出此时该IP被分配到哪个RADIUSID。

离线radius数据存储在radius log中，分隔符为”\t”, 字段格式为IP, 状态，radiusID, 时间，示例：

100.66.136.217\t2\tnUTV2IO6dHer8wQ6Mjl9sCQTyGa+QAmvCifag27Ziio==\t2015-02-02 07:37:05

Table 3 字段格式

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **名称解释** |
| 1 | IP | 用户IP |
| 2 | status | 心跳状态 |
| 3 | radiusID | radius认证服务器端的用户标识 |
| 4 | time | 时间 |

设计实现

浙江从SIG RDS发出RADIUS数据流，江苏从RADIUS CENTER发出RADIUS数据流。两个数据流的格式不同，而且SIG RDS和RADIUS CENTER认证数据流接收端的方式也不相同。所以两个地方的RADIUS数据接收端是使用不同的程序实现。



Figure

1. **浙江Radius**

浙江RADIUS接收端通过TCP协议连接SIG RDS，发送验证信息。SIG RDS确认客户端身份后，向接收端发送应答信息并开始发送Radius数据流。

每隔5秒，RADIUS接收端向SIG RDS发送一个KA信息，以通知接收端正常存在，SIG RDS定时检查KA是否更新，如果超过20s未更新则发送超时关闭信息，并断开链接。

RADIUS接收端收到RADIUSINFO消息，解析出RADIUSID、IP、STATUS字段，更新memcache里RADIUSID对应的IP以及IP对应的RADIUSID。并把当前时间作为时间戳，按照”IP地址\t 在线\离线状态\t 用户ID\t 接收时间戳”的格式写入日志文件。



1. **江苏Radius**

江苏RADIUS接收端通过TCP协议连接RADIUS CENTER，身份认证是通过接收端的IP来进行。RADIUS CENTER维持一份允许的IP列表，只支持这些IP地址的接收端读取数据。RADIUS CENTER在确认接收端连接后，向接收端发送RADIUS数据流。该RADIUS数据流是RADIUS CENTER转发的原始IP报文，包括报文头和报文体。



Figure radius数据采集流程图

安装部署

1. **浙江Radius**

接收端程序和memcahe位于两台机器上，其中安装程序及源码位于183.136.168.67/root/thz\_work/radius\_client上，memcahe位于183.136.168.68上。

1. **江苏Radius**

接收端程序位于180.96.26.203/home/lq下，实时解析程序位于180.96.26.204/home/lq下。

##### rstat数据

模块描述

rstat数据是指在网卡上实时采集到的数据，同nstat数据。

接口实现

rstat数据和nstat数据格式相同。

设计实现

安装部署

#### 数据预处理

###### 模块描述

数据预处理是对原始数据做第一步清洗，过滤掉格式有误、字段不满足要求的数据，然后再将其转化为统一的数据格式。

###### 接口描述

存放路径，格式，输入输出，处理流程

1. **nstat数据输入**

分光机器采集到原始数据后，会同步到分析机器nstat程序处理节点/hadoop/data下（通常会选择分析机器第一个节点），按天来存放。预处理时，会将数据put到hdfs上。输入数据有三种：hdfs上原始数据/user/clickwise/nstat/date/raw/, hdfs上运营商提供的IP地址归属地文件/user/clickwise/bigdata/udf/ZJ\_QQWry.dat, radius\_server服务 进程/home/clickwise/click\_ana/nstat/radius\_server.

1. **access数据输入**

广告投放服务器采集到原始数据后，会同步到分析机器nstat程序处理节点/var/ftp/pub/addata/下，按天来存放。预处理时，会将数据put到hdfs上。输入数据为/user/clickwise/astat/date/raw.

1. **输出**

nstat数据输出结果在hdfs上/user/clickwise/nstat/date/preprocess/下，access数据输出结果在hdfs上/user/clickwise/astat/date/preprocess/下，各字段之间均用\001分隔。

Table 2 预处理输出字段格式

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **名称解释** |
| 1 | time | 时间 |
| 2 | domain\_id | 域名 |
| 3 | proxy\_ip | 代理IP |
| 4 | area | 地区代号 |
| 5 | user\_id | 用户标识符，比如uid=radiusID\*\*\* |
| 6 | user\_domain | 用户标识符类型，比如radiusID=\*\*\* |
| 7 | host | 主机名 |
| 8 | url | url |
| 9 | title | 网页title信息 |
| 10 | user\_agent | 上网设备信息 |
| 11 | sip | 源IP |
| 12 | dip | 目的IP |
| 13 | user\_loc | 用户IP归属地 |
| 14 | refer | 来路 |
| 15 | data\_type | 数据类型 |
| 16 | Of1 | 预留字段 |
| 17 | Of2 | 预留字段 |
| 18 | Of3 | 预留字段 |
| 19 | Of4 | 预留字段 |

###### 设计实现

补充细节部分，具体实现步骤方法

在分析机器nstat程序处理节点上，接收采集到的用户日志数据和radius数据，通过调用radius\_server服务、运营商提供的IP归属地文件，将原始数据转化为统一格式的19个字段。

在分析机器nstat程序处理节点上，路径/home/clickwise/click\_ana/nstat下，存放nstat数据处理程序，预处理程序调用方式为：hadoop jar preprocess.jar preprocess <data\_type> <inputfile> <outputfile> [<cookie> |<baidu> |<radius> <radius\_server\_host:port>], 各参数解释如下：

Table 3 预处理输出字段格式

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **名称解释** |
| 1 | preprocess | 处理名称 |
| 2 | data\_type | 数据类型，比如nstat,access |
| 3 | inputfile | hdfs输入文件路径 |
| 4 | outputfile | hdfs输出文件路径 |
| 5 | <cookie> |<baidu> |<radius> <radius\_server\_host:port> | 用户标识符类型，比如cookie, baidu, radius, 如果用radius, 要同时给出radius\_server所在IP及端口 |

下图为处理流程图。



Figure 3 数据预处理流程图

###### 安装部署

1. **nstat数据**

在分析机器nstat程序处理节点上，启动定时任务

/home/clickwise/click\_ana/nstat/sync\_radius.sh,/home/clickwise/click\_ana/nstat/nstat\_app.php, 程序会每天自动处理前一天的数据。

1. **access数据**

在分析机器nstat程序处理节点上，启动定时任务

/home/clickwise/click\_ana/astat/astat\_app.php, 程序会每天自动处理前一天的数据。

#### 数据处理

数据处理是整个流程的核心，主要是在现有数据的基础上，统计分析出预定的主题，由于算法的复杂度和数据的海量程度，整个过程都要在hadoop集群上进行分布式计算。包括以下几个分析功能模块：全流量分析、搜索引擎、电子商务、用户行为、行业网站、网站分析工具（包括网站分类、共现词和相似网站）等。

添加整体流程图

##### 全流量分析

模块描述

流量统计模块对应DMP分析平台中的全流量样本部分，包括主机链接分析和流量分布分析，前者主要统计全网host的排名，以及每个host的url和排名，每个url的refer和排名，每个host的地域分布和排名情况等，并对host进行一级分类和二级分类；后者主要统计各地区和各数据采集点的独立IP数和用户数。

接口实现

1. **输入数据**

输入数据为nstat hive表，存放在hdfs上/user/hive/warehouse/clickwise.db/nstat/dt=date

1. **输出数据**

输出数据存放在hdfs上，/user/clickwise/nstat/date/ host\_stat\_output下，结果数据有6个文件。

Table 4 全流量分析输出文件

|  |  |  |
| --- | --- | --- |
| **序号** | **结果路径** | **文件解释** |
| 1 | /user/clickwise/nstat/date/host\_cate\_res | 各个host的类别及标签 |
| 2 | /user/clickwise/nstat/date/host\_stat\_output/host\_stat | 各个host的独立IP数和PV数统计 |
| 3 | /user/clickwise/nstat/date/host\_stat\_output/area\_stat | 各个host在全省及各地市的独立IP数和PV数统计 |
| 4 | /user/clickwise/nstat/date/host\_stat\_output/pvip\_stat | 各地区和各数据采集点的独立IP数和用户数 |
| 5 | /user/clickwise/nstat/date/host\_stat\_output/url\_stat | 各个host下的热点URL独立IP数和PV数统计 |
| 6 | /user/clickwise/nstat/date/host\_stat\_output/urlr\_stat | 各个host下的热点URL和refer的独立IP数和PV数统计 |

设计实现

首先从格式化数据中提取出host对其分类，如果该host不存在已有的数据库中，则使用网络爬虫工具爬取host内容，并对其分类；然后采用海量数据统计技术对host, url, refer等字段进行pv数和独立IP数、每个地区的pv总数和独立IP总数等数据进行统计。

流程图



Figure 4. 全流量分析流程图

安装部署

##### 搜索引擎

模块描述

搜索引擎模块对应DMP分析平台中的搜索引擎部分，包括百度、谷歌和360等常用搜索引擎的搜索词排名，各搜索词的点击链接，各搜索词在最近10天内的搜索量；以及百度热门搜索中的电影、综艺、影视剧、小说和新闻资讯等类别下的搜索词和点击链接等。

接口实现

1. **预处理输入数据**

输入数据为nstat hive表，存放在hdfs上/user/hive/warehouse/clickwise.db/nstat/dt=date

1. **预处理输出数据**

输出数据存放在hdfs上，/user/clickwise/nstat/date/raw\_se下，结果数据有4个文件为baidu, mbaidu, s360, google, 分别表示百度数据、百度移动数据、360数据、google数据。然后，将此类数据作为输入数据。

1. **搜索引擎输出数据**

输出数据存放在hdfs上，包括/user/clickwise/nstat/date/se\_\*, 其中\*代表baidu,360,google.

1. **百度热门搜索输入数据**

输入数据为hdfs上/user/clickwise/nstat/date/se\_baidu/keywords文件

1. **百度热门搜索输出数据**

输出数据存放在hdfs上，/user/clickwise/nstat/date/jbkw/sort\_nv\_output下

Table 5 搜索引擎输出文件

|  |  |  |
| --- | --- | --- |
| **序号** | **结果路径** | **文件解释** |
| 1 | /user/clickwise/nstat/date/se\_\*/keywords | 搜索词、搜索量对比分析 |
| 2 | /user/clickwise/nstat/date/se\_\*/links | 搜索词点击分析、搜索量对比分析 |
| 3 | /user/clickwise/nstat/date/se\_\*/m\_keywords | 移动端搜索词、搜索量对比分析 |
| 4 | /user/clickwise/nstat/date/se\_\*/user\_keywords | 用户搜索词、搜索量对比分析 |
| 5 | /user/clickwise/nstat/date/se\_\*/user\_links | 用户搜索词点击分析、搜索量对比分析 |
| 6 | /user/clickwise/nstat/date/jbkw/sort\_nv\_output | 百度热门搜索数据，包括小说、电影、综艺、影视剧、新闻资讯分析 |

设计实现

首先从格式化数据中提取百度，谷歌和360等搜索引擎的相关数据，然后再分别提取每个搜索引擎的搜索词和点击链接，再从百度搜索数据中提取出电影、综艺、影视剧、小说和新闻资讯等热点数据。

流程图



Figure 5 搜索引擎流程图

安装部署

##### 电子商务

模块描述

电子商务模块对应DMP分析平台中的电子商务部分。该模块选取京东、天猫、苏宁、易讯和亚马逊等五大热门电商，从网站流量、页面浏览量、销售情况、搜索情况等四个维度对电商行业进行综合分析，根据分析产生的结果，绘制对比图，展现热点单品数据。

接口实现

1. **预处理输入数据**

输入数存放在hdfs上，路径/user/clickwise/nstat/date/preprocess下

1. **预处理输出数据**

输出数据存放在hdfs上，/user/clickwise/nstat/date/raw\_ec下，并load到电商hive表monitor\_data中，然后，将此类数据作为输入数据。

1. **输入数据**

输入数据为hive表，存放在hdfs上/user/hive/warehouse/clickwise.db/ monitor\_data/dt=date

1. **输出数据**

输出数据存放在hdfs上，/user/clickwise/nstat/ec\_date/下和hive表中。

Table 6 电子商务输出文件

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **类型** | **结果路径** | **文件解释** |
| 1 | hive表 | ec\_whole\_pvs\_ips | 全站流量及浏览量数据 |
| 2 | ec\_channel\_pvs\_ips | 频道页流量及浏览量 |
| 3 | ec\_product\_pvs\_ips | 四级页面流量及浏览量数据 |
| 4 | ec\_refer\_pvs\_ips | 站外流量来源数据 |
| 5 | ec\_area\_pvs\_ips | 流量来源城市 |
| 6 | ec\_homepage\_pvs\_ips | 首页浏览量数据 |
| 7 | ec\_leimu\_pvs\_ips | 类目页面浏览量数据 |
| 8 | ec\_products\_pvs\_ips | 四级页面浏览量TOP |
| 9 | ec\_cart\_pvs\_ips | 购物车页面浏览量 |
| 10 | ec\_search\_keywords | 站内搜索关键词 |
| 11 | ec\_search\_pvs\_ips | 电商站内搜索使用人数 |
| 12 | ec\_pay\_pvs\_ips | 支付方式数据 |
| 13 | hdfs文件 | /user/clickwise/ec\_date/ec\_sale\_tp\_ds | 单品销售数据 |
| 14 | /user/clickwise/nstat/date/user\_ec\_product\_day | 用户电商单品浏览数据 |
| 15 | /user/clickwise/nstat/date/user\_ec\_search\_day | 用户电商站内搜索数据 |

设计实现

首先分析得出每个电商需要分析的各种类型数据的url链接和请求数据包的特征，再从格式化数据中提取出五大电商的浏览、购买、搜索、推荐等数据，然后爬取电商商品页面的title等信息，并提取出流量数据、浏览量数据、销售情况数据和搜索词数据等，对这些数据进行各大电商的综合分析。

流程图



Figure 6 电子商务流程图

安装部署

##### 用户行为

模块描述

用户行为模块对应DMP分析平台中的用户行为部分。主要对host分类、访问主机、搜索词和点击链接等维度对用户画像，从用户访问网站的时间、访问内容、点击内容等这些零散的信息中，提炼和整理出用户在网站内部的行为，进而分析出用户的个性化标签。

接口实现

1. **输入数据**

输入数据有多个，包括nstat hive表，以及搜索引擎中分析的用户数据

1. **输出数据**

输出数据存放在hdfs上，/user/clickwise/nstat/date/user\_\*下，结果数据有7个文件。

Table 7 用户行为输出文件

|  |  |  |
| --- | --- | --- |
| **序号** | **结果路径** | **文件解释** |
| 1 | /user/clickwise/nstat/date/user\_day | 用户基本属性数据 |
| 2 | /user/clickwise/nstat/date/user\_cate\_day | 用户分类属性数据 |
| 3 | /user/clickwise/nstat/date/user\_host\_day | 用户访问主机数据 |
| 4 | /user/clickwise/nstat/date/user\_ec\_product\_day | 用户电商单品浏览数据 |
| 5 | /user/clickwise/nstat/date/user\_ec\_search\_day | 用户电商搜索数据 |
| 6 | /user/clickwise/nstat/date/user\_se\_keywords\_day | 用户百度搜索关键词数据 |
| 7 | /user/clickwise/nstat/date/user\_se\_links\_day | 用户百度搜索点击链接数据 |

设计实现

首先从格式化数据中识别出真实用户的标识，然后根据真是用户的表示从海量存储中提取出这部分用户最近一周内的浏览、点击、搜索引擎、电商日志，对日志进行统计和分析，得出用户的兴趣偏好和浏览点击等行为动机等多角度的用户画像。

流程图



Figure 7 用户行为流程图

安装部署

##### 行业网站

模块描述

行业网站是对各大行业的用户进行分析。统计各行业中的用户行为规律，统计用户的共同行为，为市场决策等做指导。当前针对旅游、视频、门户网站、房产、游戏、金融六个行业，对给定行业的热门站点的用户进行分析。

接口实现

1. **输入数据**

输入数据存放在hdfs上，/user/clickwise/nstat/date/raw下（调整）

1. **输出数据**

输出数据存放在hdfs上，结果数据有13个文件。

Table 7 行业网站输出文件

|  |  |  |
| --- | --- | --- |
| **序号** | **结果路径** | **文件解释** |
| 1 | /user/clickwise/nstat/date//estate/visit\_statis\_hdfs/ | 房产网数据 |
| 2 | /user/clickwise/nstat/date/video/video\_statis\_info | 视频网站数据 |
| 3 | /user/clickwise/nstat/date/game/pv\_gameName\_day | 游戏类型数据 |
| 4 | /user/clickwise/nstat/date/game/pv\_gameWebsite\_day | 游戏网站数据 |
| 5 | /user/clickwise/nstat/date/travel/travel\_hdfs\_statis/qunar\_pair | 去哪儿旅游网双向城市数据 |
| 6 | /user/clickwise/nstat/date/travel/travel\_hdfs\_statis/qunar\_single | 去哪儿旅游网单向城市数据 |
| 7 | /user/clickwise/nstat/date/travel/travel\_hdfs\_statis/ctrip\_pair | 携程旅游网双向城市数据 |
| 8 | /user/clickwise/nstat/date /travel/travel\_hdfs\_statis/ctrip\_single | 携程旅游网单向城市数据 |
| 9 | /user/clickwise/nstat/date /portal/pv\_host\_day | 门户网一级频道数据 |
| 10 | /user/clickwise/nstat/date /portal/pv\_subhost\_day | 门户网二级频道数据 |
| 11 | /user/clickwise/nstat/date /finance/finance\_statis\_all\_hdfs | 金融类数据 |
| 12 | /user/clickwise/nstat/date /finance/finance\_statis\_p2p\_hdfs | 金融类ptp数据 |
| 13 | /user/clickwise/nstat/date /finance/finance\_statis\_at\_hdfs | 金融类支付网数据 |

设计实现

首先需要总结各类型网站的链接特征，从原始数据中提取出各种数据，然后根据预定的主题进行统计，分析出各行业网站下的功能数据并存放在hdfs上。

流程图



Figure 7 行业网站流程图

安装部署

##### 分析工具

模块描述

网站分析工具不同于以上DMP平台功能，是为了方便用户日常分析工作而提供的。分析工具模块对应DMP分析平台中的网站分类、共现词和相似网站等部分。网站分类方便用户查询某个host的类别及在一段时间内的统计量，共现词帮助用户查询某个词的相关词，相似网站方便用户查询某个host有哪些相似host。

接口实现

1. **输入数据**

输入数据为nstat hive表

1. **输出数据**

输出数据存放在hdfs上

Table 4 分析工具输出文件

|  |  |  |
| --- | --- | --- |
| **序号** | **结果路径** | **文件解释** |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |

设计实现

网站分类是对host对应网站的内容进行自动分类，并统计出一段时间的所有host访问量及其排名，方便查看全网内的各类别网站的流量排名情况；对于共现词，首先要整理出一个词典，用相关词挖掘技术找出每个词的相关词；相似网站和共现词类似，它利用收集的海量数据进行分析，首先分析给定host的内容关键词，然后利用相关词挖掘技术找出和这些关键词相类似的带权重的关键词集合，然后通过计算关键词之间的相似度来寻找和其相似的网站host。

流程图

安装部署

##### cookiemapping

模块描述

接口实现

设计实现

安装部署

#### 数据存储

数据存储是DMP平台专门负责数据读写的模块。对外提供一个抽象的读写接口。数据处理模块和数据应用模块的程序可能会调用数据存储中已有的数据，而这两个模块的输出结果部分也会写入数据存储模块中。

数据存储单独作为一个模块的目的是：1.不同的数据应用共享同一数据源，避免同一数据源的冗余存储。 2. 更新数据源时，保证不同的应用中版本一致，例如知识库的统一管理。3. 便于和任务平台隔离，以免因为存储模块消耗内存过多而使任务终端或平台崩溃。4. 底层使用分布式存储，根据数据量大小灵活添加或移除机器，使数据存储具有良好的扩展性。5. 使开发人员能专注于开发高效的实时读写服务，为用户定向、广告匹配等提供服务。6. 便于提供一系列简单的调用接口，简化数据处理模块和数据应用模块程序的调用。

###### 接口描述

数据存储的接口分为http接口和java接口两种。Http接口简单直接，便于跨语言调用。Java接口满足开发高效的数据应用。Http接口规范旨在指定请求url和响应的格式。Java 接口旨在指定该模块api。

Http接口

提供如下REST风格的HTTP接口：

http://<ip>:<port>/<method>?{query\_string}

query\_string由系统级参数部分和具体Open API调用参数部分组成，以key1=value1&key2=value2&…表示，对于采用POST请求的Open API，query\_string部分则是在POST请求体里。所有查询类的Open API接口既支持POST，也支持GET方式，提交类的OPEN API接口仅支持POST方式。

* 表格1 系统级参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **类型** | **是否必需** | **描述** |
| method | string | 是 | 采取userinfo.query这样的命名空间方式制定方法 |
| call\_id | int | 否 | 时间戳，系统时间的秒值,同个应用的不同api请求的time值应该是递增的, 用于防replay攻击 |
| format | string | 否 | 响应包格式，可以是xml（默认）或json |
| ie | string | 否 | API调用请求包的编码类型，支持UTF-8和GBK |
| ip | string | 是 | 提供存储查询服务的ip |
| port | int | 是 | 提供存储查询服务的端口 |

* 用户数据查询接口
* 参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **类型** | **是否必需** | **描述** |
| port | int | 是 | 6579用户数据查询 |
| method | string | 是 | userinfo.query 查询用户特征 |
| uid | string | 是 | 用户标识，例如RADIUSID、COOKIE |
| feattype\* | int | 否 | 特征类型，0 百度搜索  1 百度搜索关键词  2 电商title  3 电商title关键词  4 电商搜索  5 电商搜索关键词 |
| starttime | long | 否 | 起始时间 |
| endtime | long | 否 | 结束时间 |

注：\*表示可以有0-n个

* 返回值
* Response XML 示例

<?xmlversion=*"1.0"*encoding=*"UTF-8"*?>

<clickwise\_userinfo\_responsexmlns=[*http://clickwise.cn*](http://clickwise.cn)xmlns:xsi=*"http://www.w3.org/2001/XMLSchema-instance"*>

<user>

<uid>7d1cf3105a7388753da162a802d78055009</uid>

<infolist>

<feat>

<feattype>bdsearch</feattype>

<conent>植物大战僵尸小游戏</conent>

</feat>

<feat>

<feattype>bdkeywords</feattype>

<conent>植物大战僵尸小游戏</conent>

</feat>

<feat>

<feattype>ectitle</feattype>

<conent>春秋平底平跟流苏单鞋唐卡豆豆鞋牛筋底磨砂真牛皮豹纹小白鞋女鞋</conent>

</feat>

</infolist>

</user>

</clickwise\_userinfo\_response>

* Json格式的字符串内容是XML输出数据所对应的PHP数组的标准JSON字符串
* CookieMap接口
* 参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **类型** | **是否必需** | **描述** |
| port | int | 是 | 6578cookie map |
| method | string | 是 | cookiemap.map映射用户id |
| src\_uid | string | 是 | 源用户标识，例如RADIUSID、COOKIE |
| src\_type | int | 是 | 源用户标识类型  0 RADIUSID  1 COOKIE  2 Email |
| dest\_type+ | int | 是 | 目标用户标识类型  0 RADIUSID  1 COOKIE  2 Email |

注：+表示可以有1-n个

* 返回值
* Response XML 示例

<?xmlversion=*"1.0"*encoding=*"UTF-8"*?>

<clickwise\_cookiemap\_responsexmlns=[*http://clickwise.cn*](http://clickwise.cn)xmlns:xsi=*"http://www.w3.org/2001/XMLSchema-instance"*>

<src\_uid>

<uid\_type>cookie</uid\_type>

<uid\_value>7d1cf3105a7388753da162a802d78055009</uid\_value>

</src\_uid>>

<dest\_uids>

<uid\_item>

<uid\_type>radius</uid\_type>

<uid\_value>854f8dddc446a9822572aa47f8f75500e</uid\_value>

</uid\_item>

<uid\_item>

<uid\_type>email</uid\_type>

<uid\_value>test@126.com</uid\_value>

</uid\_item>

</dest\_uids>

</clickwise\_cookiemap\_response>

* Java API

*To do*

###### 设计实现

数据存储共分为数据存储介质、数据存储驱动、存储查询服务三个子层次。



数据存储介质子层是实际存储数据的各种类型的数据库或文件系统。根据响应时间分为非实时存储介质，如文本文件、hdfs、hive、mysql等，和实时存储介质，如hbase、redis、ardb、memcache等。根据可扩展性分为本地存储介质，如文本文件、mysql、redis等，和分布式存储介质,如hdfs、hive、hbase等。根据所要存储的数据量大小、数据格式以及响应时间选择实际的存储介质。一些查询速度快的介质，如redis、内存等可以用来做本地缓存。

数据存储驱动子层是数据存储介质子层上面封装的一系列接口，该子层的存在目的是：1. 统一管理已有数据，能够方便的浏览已有的表、数据库以及各自的格式 2. 记录读写日志，以分析数据的质量情况、读写的速度和稳定性。3. 将高并发实时查询分配到多台机器上，平衡各机器之间的负载。4. 随机的热点数据的本地缓存，以提高查询效率。5. 将连接数据库、读写数据库等过程封装成统一接口，便于存储查询服务中调用。

存储查询子层服务提供不同数据源的读写接口，例如用户数据、CookieMap、知识库等。该子层根据不同数据源的数据特性，调用数据存储子层的一系列接口，对数据应用层和数据加工层提供读写服务，对外提供Http接口和Java API。

###### 安装部署

*To do*

#### 数据应用

##### DSP平台界面

##### DSP对接

DMP平台为DSP平台提供数据和计算支持。DMP平台利用已有的用户数据、计算能力、实现算法，对特定上下文中的某一用户优化广告投放，从而提高广告的点击率和转化率。DMP平台的用户数据来自本身的数据采集层和广告投放日志。为了制定广告投放策略，DMP平台需要维持并迅速更新本地广告库，使本地广告库和DSP平台广告库同步。DMP平台的DSP对接模块依托数据存储层的用户和广告数据，首先根据业务需求生成候选广告集合，然后调用已实现算法选择出点击率和转化率可能比较高的广告。DSP对接模块还负责跟踪广告投放过程，统计当前的广告投放效果，收集广告投放日志等。

DMP平台主要分两部分内容，一是业务逻辑，包括同步广告库、广告投放策略实现、跟踪广告投放日志、统计广告投放效果；二是广告算法，该部分根据广告投放日志和用户数据，不断优化广告投放效果。

接口描述

该模块和DSP平台进行交互，所以提供http接口，从而使DSP平台可以直接调用。同样提供如下REST风格的HTTP接口：

http://<ip>:<port>/<method>?{query\_string}

用户广告匹配接口

参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **类型** | **是否必需** | **描述** |
| port | int | 是 | 6577ad match |
| method | string | 是 | dsp.admatch对用户选择广告 |
| uid | string | 是 | 用户标识，例如RADIUSID、COOKIE |
| uid\_type | int | 是 | 用户标识类型  0 RADIUSID  1 COOKIE  2 Email |
| host | string | 否 | 用户正在访问的网站域名 |
| uri | string | 否 | 用户正在浏览的网页 |
| title | string | 否 | 用户正在浏览的网页title |
| vtime | long | 否 | 用户访问时间 |
| adloc | double[] | 否 | 广告位的位置 |
| adform | int | 否 | 广告位的形式 |
| adsize | double[] | 否 | 广告位的尺寸 |
| adnum | int | 否 | 返回广告的最大数目 |

* 返回值
* Response XML 示例

<?xmlversion=*"1.0"*encoding=*"UTF-8"*?>

<clickwise\_admatch\_responsexmlns=[*http://clickwise.cn*](http://clickwise.cn)xmlns:xsi=*"http://www.w3.org/2001/XMLSchema-instance"*>

<matchinfo>

<uid>7d1cf3105a7388753da162a802d78055009</uid>

<uid\_type>cookie</uid\_type>

<matchads>

<matchad>

<aid>1001</aid>

<score>0.7</score>

</matchad>

<matchad>

<aid>1002</aid>

<score>0.5</score>

</matchad>

</matchads>

</matchinfo>

</clickwise\_admatch\_response>

* 广告日志接口
* 参数

|  |  |  |  |
| --- | --- | --- | --- |
| **参数名** | **类型** | **是否必需** | **描述** |
| port | int | 是 | 6576 ad log |
| method | string | 是 | dsp.adlog 对用户选择广告 |
| uid | string | 是 | 用户标识，例如RADIUSID、COOKIE |
| uid\_type | int | 是 | 用户标识类型  0 RADIUSID  1 COOKIE  2 Email |
| aid | string | 是 | 广告id ,能够唯一确定广告库里的某一广告 |
| host | string | 否 | 用户正在访问的网站域名 |
| uri | string | 否 | 用户正在浏览的网页 |
| title | string | 否 | 用户正在浏览的网页title |
| vtime | long | 否 | 用户访问时间 |
| adloc | double[] | 否 | 广告位的位置 |
| adform | int | 否 | 广告位的形式 |
| adsize | double[] | 否 | 广告位的尺寸 |
| click | boolean | 是 | 是否点击 |

* 返回值
* Response XML 示例

<?xmlversion=*"1.0"*encoding=*"UTF-8"*?>

<clickwise\_status\_responsexmlns=[*http://clickwise.cn*](http://clickwise.cn)xmlns:xsi=*"http://www.w3.org/2001/XMLSchema-instance"*>

<error\_code>0</error\_code>

<error\_msg>success</error\_msg>

</clickwise\_status\_response>

设计实现

DSP对接模块主要有三个子模块，广告匹配服务、业务逻辑及广告算法、用户及广告数据更新。



广告匹配服务是DSP对接模块与其它模块交互的窗口。广告匹配服务提供广告匹配http接口，DSP程序可以方便的调用。并提供存储广告日志的http接口，方便DSP程序记录广告日志。广告匹配服务要考虑高并发的匹配请求，并且保证每个广告匹配的实时性。这就要求业务逻辑实现及广告算法要高效，并且用户和广告的查询速度要满足实时要求。

业务逻辑及广告算法，业务逻辑是广告主制定的广告匹配策略，例如对特定host或特定地域的人群投放，时间选择、黑名单、白名单等。通过业务逻辑的过滤，形成候选广告集合。对候选广告集合的广告根据预测点击率进行排序。例如从用户广告点击日志建立LR模型。

用户和广告数据更新在后台进行，DSP对接模块每隔特定时间更新用户和广告的数据库。用户数据的更新间隔时间可以长一些，例如两小时更新一次，每次更新的数据量也比较大。广告数据库要更新的数据库相对较小，更新间隔时间要短一些，例如五分钟更新一次。

安装部署